



AFRL-RY-WP-TR-2022-0036

**APPLICATION OF ANALOGICAL REASONING FOR USE
IN VISUAL KNOWLEDGE EXTRACTION**

**Kara Lian Combs
Wright State University**

**FEBRUARY 2022
Final Report**

DISTRIBUTION STATEMENT A. Approved for public release; distribution is unlimited.

See additional restrictions described on inside pages

STINFO COPY

**AIR FORCE RESEARCH LABORATORY
SENSORS DIRECTORATE
WRIGHT-PATTERSON AIR FORCE BASE, OH 45433-7320
AIR FORCE MATERIEL COMMAND
UNITED STATES AIR FORCE**

REPORT DOCUMENTATION PAGE

PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ORGANIZATION.

1. REPORT DATE February 2022	2. REPORT TYPE Thesis	3. DATES COVERED	
		START DATE 21 December 2021	END DATE 21 December 2021
4. TITLE AND SUBTITLE APPLICATION OF ANALOGICAL REASONING FOR USE IN VISUAL KNOWLEDGE EXTRACTION			
5a. CONTRACT NUMBER N/A	5b. GRANT NUMBER N/A	5c. PROGRAM ELEMENT NUMBER N/A	
5d. PROJECT NUMBER N/A	5e. TASK NUMBER N/A	5f. WORK UNIT NUMBER N/A	
6. AUTHOR(S) Kara Lian Combs			
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Wright State University 3640 Colonel Glenn Hwy Dayton, OH 45435			8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) Air Force Research Laboratory Sensors Directorate Wright-Patterson Air Force Base, OH 45433-7320 Air Force Materiel Command United States Air Forces		10. SPONSOR/MONITOR'S ACRONYM(S) AFRL/RYPAR	11. SPONSOR/MONITOR'S REPORT NUMBER(S) AFRL-RY-WP-TR-2022-0036
12. DISTRIBUTION/AVAILABILITY STATEMENT DISTRIBUTION STATEMENT A. Approved for public release; distribution is unlimited.			
13. SUPPLEMENTARY NOTES PAO case number AFRL-2021-4518, Clearance Date 21 December 2021. A thesis submitted in partial fulfillment of the requirements for the degree of Master of Science in Industrial and Human Factors Engineering. The U.S. Government is joint author of this work and has the right to use, modify, reproduce, release, perform, display, or disclose the work. Report contains color.			
14. ABSTRACT There is a continual push to make Artificial Intelligence (AI) as human-like as possible; however, this is a difficult task because of its inability to learn beyond its current comprehension. Analogical reasoning (AR) has been proposed as one method to achieve this goal. Current literature lacks a technical comparison on psychologically-inspired and natural-language-processing-produced AR algorithms with consistent metrics on multiple-choice word-based analogy problems. Assessment is based on "correctness" and "goodness" metrics. There is not a one-size-fits-all algorithm for all textual problems. As contribution in visual AR, a convolutional neural network (CNN) is integrated with the AR vector space model, Global Vectors (GloVe), in the proposed, Image Recognition Through Analogical Reasoning Algorithm (IRTARA). Given images outside of the CNN's training data, IRTARA produces contextual information by leveraging semantic information from GloVe. IRTARA's quality of results is measured by definition, AR, and human factors evaluation methods, which saw consistency at the extreme ends. The research shows the potential for AR to facilitate more a human-like AI through its ability to understand concepts beyond its foundational knowledge in both a textual and visual problem space.			
15. SUBJECT TERMS analogy, artificial intelligence, text analytics			
16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT SAR
a. REPORT Unclassified	b. ABSTRACT Unclassified	c. THIS PAGE Unclassified	
			18. NUMBER OF PAGES 220
19a. NAME OF RESPONSIBLE PERSON Trevor Bihl			19b. PHONE NUMBER (Include area code) (937) 713-8116

APPLICATION OF ANALOGICAL REASONING FOR USE IN VISUAL
KNOWLEDGE EXTRACTION

A thesis submitted in partial fulfillment
of the requirements for the degree of
Master of Science in Industrial and Human Factors Engineering

By

Kara Lian Combs

B.S., Wright State University, 2021

WRIGHT STATE UNIVERSITY

GRADUATE SCHOOL

November 17, 2021

I HEREBY RECOMMEND THAT THE THESIS PREPARED UNDER MY SUPERVISION BY Kara Lian Combs ENTITLED Application of Analogical Reasoning for Use in Visual Knowledge Extraction BE ACCEPTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF Master of Science in Industrial and Human Factors Engineering

Subhashini Ganapathy, Ph.D.
Thesis Co-Director

Trevor Bihl, Ph.D.
Thesis Co-Director

Committee on Final Examination:

Subhashini Ganapathy, Ph.D.

Assaf Harel, Ph.D.

Trevor J. Bihl, Ph.D.

Barry Milligan, Ph.D.
Dean of the Graduate School

Subhashini Ganapathy Ph.D.
Chair, Department of Biomedical,
Industrial and Human Factors Engineering

ABSTRACT

Combs, Kara Lian. M.S.I.H.E. Department of Biomedical, Industrial and Human Factors Engineering, Wright State University, 2021. Application of analogical reasoning for use in visual knowledge extraction.

There is a continual push to make Artificial Intelligence (AI) as human-like as possible; however, this is a difficult task because of its inability to learn beyond its current comprehension. Analogical reasoning (AR) has been proposed as one method to achieve this goal. Current literature lacks a technical comparison on psychologically-inspired and natural-language-processing-produced AR algorithms with consistent metrics on multiple-choice word-based analogy problems. Assessment is based on “correctness” and “goodness” metrics. There is not a one-size-fits-all algorithm for all textual problems. As contribution in visual AR, a convolutional neural network (CNN) is integrated with the AR vector space model, Global Vectors (GloVe), in the proposed, Image Recognition Through Analogical Reasoning Algorithm (IRTARA). Given images outside of the CNN’s training data, IRTARA produces contextual information by leveraging semantic information from GloVe. IRTARA’s quality of results is measured by definition, AR, and human factors evaluation methods, which saw consistency at the extreme ends. The research shows the potential for AR to facilitate more a human-like AI through its ability to understand concepts beyond its foundational knowledge in both a textual and visual problem space.

TABLE OF CONTENTS

TABLE OF CONTENTS	IV
1 INTRODUCTION	17
1.1 Technical Motivation	21
1.2 Operation Motivation	22
1.3 Research Contribution.....	23
1.4 Research Objectives	25
2 BACKGROUND	27
2.1 Machine Intelligence	30
2.2 Algorithmic Analogical Reasoning Approaches	33
2.2.1 Text-based Analogy Methods	36
2.2.2 Image-based Analogy Methods	66
2.3 Image recognition and context	69
2.3.1 Artificial Neural Networks (ANNs).....	69
2.3.2 Convolutional Neural Networks (CNNs).....	72
2.3.3 Meaning Making for Understanding Images	76

3	SYSTEMATIC COMPARISON OF ANALOGICAL REASONING ALGORITHMS.....	79
3.1	Algorithm Selection.....	79
3.2	Comparative Example	82
3.2.1	Representative Example Data	82
3.2.2	Performance Metrics	83
3.3	Results	86
3.3.1	Correctness Metric Results	87
3.3.2	Goodness Metric Results	88
3.4	Summary and Conclusion	89
4	IMAGE-BASED ANALOGICAL REASONING.....	91
4.1	Visual Data Set Selection	91
4.2	Algorithm Framework.....	92
4.2.1	Image Classification	94
4.2.2	Creation of Class Name Word Vectors.....	97
4.2.3	Application of Analogical Reasoning & Knowledge Extraction	99
4.2.4	Evaluation.....	102
4.3	Example Result Walkthroughs	106
4.3.1	Creation of Class Name Word Vectors.....	106

4.3.2	Bad vs. Good Walkthrough.....	107
4.4	Mystery Class Walkthrough.....	119
4.5	Evaluation Methods & Metrics	124
4.5.1	Definition Evaluation Method	125
4.5.2	AR Evaluation Method	128
4.5.3	Human Factors Evaluation Method	132
4.6	Results	136
5	DISCUSSIONS.....	144
5.1	Definition Evaluation	144
5.2	Analogical Reasoning Evaluation	146
5.3	Human Factors Evaluation	147
5.4	Overall Summary	152
5.5	Placing IRTARA in Analogical Reasoning Literature.....	157
6	CONCLUSIONS AND FUTURE WORK.....	159
6.1	Problems With Current AI Methods.....	159
6.2	Contributions to research.....	160
6.2.1	Text-based Analogical Reasoning Evaluation	160

6.2.2 Image-based Analogical Reasoning Algorithm & Evaluation162

6.2.3 Future Work.....164

7 REFERENCES 167

8 APPENDIX A..... 178

9 APPENDIX B..... 184

10 APPENDIX C 193

11 APPENDIX D..... 197

12 APPENDIX E 203

13 APPENDIX F 217

LIST OF FIGURES

Figure	Page
Figure 1-1. Knowns and Unknowns Matrix	17
Figure 1-2. Fireworks Images	19
Figure 1-3. Fireworks Image Decomposed by Google Cloud Vision AI (Google, 2021)	19
Figure 2-1. Visible Action of an Example Copycat Analogy, adapted from (Hofstadter & Mitchell, 1995).....	28
Figure 2-2. Bird to Nest as Dog is to ? Analogy, from (Goswami & Brown, 1990).....	29
Figure 2-3. Three Stages of Scientific Inquiry, adapted from (Flach & Kakas, 2000).....	31
Figure 2-4. Relationship Between Abduction, Abstraction, Deduction, and Induction, adapted from (Weigand & Hartung, 2012).....	32
Figure 2-5. Structure-mapping-based Analogy Process Visual, from (Gentner & Smith, 2012)	34
Figure 2-6. AR Models in the Context of AI Schools of Thought, from (Combs, Bihl, Ganapathy, & Staples, 2022)	39
Figure 2-7. Analytical Shift in Animal Location Comparing "Size" and "Ferocity" from shifting from Henley's to Rumelhart's Model, from (Rumelhart & Abrahamson, 1973)	40
Figure 2-8. Example Slipnet Potential Slippages, adapted from (Bolland, 2004).....	43
Figure 2-9. Steps for STAR's Simple Analogical Reasoning, adapted from (Halford, et al., 1994)	46

Figure 2-10. Visualization of Water/heat-flow Analogy, from (Falkenhainer & Forbus, 1989)	48
Figure 2-11. Heat/water-flow Textual Representation, adapted from (Eliasmith & Thagard, 2001)	49
Figure 2-12. LISA's Representation of Propositions, adapted from (Hummel & Holyoak, 2005)	51
Figure 2-13. Visualization of LISA's Mapping Process, adapted from (Hummel & Holyoak, 1997)	52
Figure 2-14. HRR superimposed in a 3D space, from (Eliasmith & Thagard, 2001)	54
Figure 2-15. CAB's Representation of "Jim loves Betty", adapted from (Larkey & Love, 2003)	56
Figure 2-16. SAT Question Used as Example in LRA Walkthrough, adapted from (Turney, 2006)	58
Figure 2-17. Pair-pattern Frequency Matrix Example, adapted from (Turney, 2006)	59
Figure 2-18. Original and Alternative Pairs Cosines, adapted from (Turney, 2006).....	60
Figure 2-19. Timeline and Relationships Between Analogical Reasoning Algorithms, from (Combs, Bihl, Ganapathy, & Staples, 2022).....	64
Figure 2-20. Geometric-analogy Problem, from (Evans, 1964)	67
Figure 2-21. ANN Families and Types, from (Bihl, Young, & Weckman, 2018)	70

Figure 2-22. Types of artificial neural networks (ANNs), from (Bihl, Young, & Frimel, 2022)	72
Figure 2-23. Convolutional Operation, adapted from (Wu, 2017)	73
Figure 2-24. Convolutional Layer Applied to the Lenna Image, from (Wu, 2017)	74
Figure 2-25. Example Pooling Application	75
Figure 2-26. OAS Application – Weight Scale, from (Farhadi, et al., 2010)	76
Figure 2-27. OAS Application – Horse Rider, from (Farhadi, et al., 2010)	77
Figure 3-1. Example of Textual Data, from (Combs, Bihl, Ganapathy, & Staples, 2022) using data from (Sternberg & Nigro, 1980; Morrison, et al., 2004)	83
Figure 3-2. Conceptualization of Textual Evaluation Steps with an Example	84
Figure 3-3. Percent Correctness Metric Result	87
Figure 4-1. Image Recognition Through Analogical Reasoning Algorithm (IRTARA) Framework	93
Figure 4-2. IRTARA CNN Architecture Visualization	97
Figure 4-3. AR Algorithm Selection Process	98
Figure 4-4. Relationship Diagram Between Definition Words	104
Figure 4-5. Mystery Class Word Cloud	122
Figure 4-6. Tile View of Mystery Class Images from the Caltech-256 dataset (Griffin, Holub, & Perona, 2007)	124
Figure 4-7. Term Frequency List and Definition Words Overlap for Mars	127

Figure 4-8. Primary and Second AR Words for Mars 129

Figure 4-9. Primary and Secondary AR Words for Chandelier 129

Figure 5-1. Differences vs. Rankings for Evaluation Results..... 156

Figure 9-1. Primary and Secondary AR Words for Mars 196

Figure 9-2. Primary and Second AR Words - Identifying Duplicate Primary AR Words
(Blue) 196

LIST OF TABLES

Table	Page
Table 1-1. Google Cloud Vision AI Label Predictions.....	20
Table 1-2. Relational Mapping Between Previous Technical Contribution(s) and Current Research Contributions (Denoted by X in the “Focus” Columns)	24
Table 2-1. Comparison of Analogical Reasoning Algorithms.....	65
Table 3-1. VSM Similarity Equations	82
Table 3-2. Goodness Metric Averages.....	89
Table 3-3. Textual Analysis Overall Results	90
Table 4-1. CNN Average Results for Loss and Accuracy	95
Table 4-2. CNN Architecture.....	96
Table 4-3. Stop Words List.....	101
Table 4-4. Classes with Additional Removed Words/Phrases from Primary and Secondary AR Word Lists	106
Table 4-5. Sample Images of Mars and Chandelier.....	109
Table 4-6. CNN Results for Mars and Chandelier Image 002.....	111
Table 4-7. Confidence-threshold Comparison for Mars and Chandelier Image 002.....	111
Table 4-8. Top AR Words for Chandelier and Mars Image 002	112
Table 4-9. Top AR Word's Definition and Formatted Definition Words for Mars Image 002.....	116

Table 4-10. Top AR Word's Definition and Formatted Definition Words for Chandelier Image 002.....	117
Table 4-11. Top Ranking CNN Classes for Chandelier and Mars	118
Table 4-12. Mystery Class IRTARA CNN Classes for First Five Images	119
Table 4-13. Top IRTARA CNN Class Results for All Mystery Images	120
Table 4-14. Top AR Words for First Five Mystery Images	120
Table 4-15. Top AR Words for Mystery Images	121
Table 4-16. Definition Words for Mars and Chandelier.....	125
Table 4-17. Comparison Between Term Frequency List and Definition Words for Chandelier and Mars	126
Table 4-18. Definition Evaluation Results for Mars and Chandelier	127
Table 4-19. Base Word Comparison for Mars' Primary AR Words	130
Table 4-20. Term Frequency and AR Word List Comparison for Mars	132
Table 4-21. AR Evaluation Results for Mars and Chandelier	132
Table 4-22. HF Evaluation Comparison for Mars	133
Table 4-23. HF Evaluation Comparison for Chandelier.....	134
Table 4-24. Overall Score from HF Evaluation Ranking for Chandelier and Mars	135
Table 4-25. Definition Evaluation Results for Select Classes	138
Table 4-26. AR Evaluation Results for Select Classes	140
Table 4-27. HF Evaluation Results for Select Classes	142

Table 5-1. Definition Evaluation Results Rankings	146
Table 5-2. AR Evaluation Results Rankings	147
Table 5-3. HF Evaluation Results Rankings.....	152
Table 5-4. Overall Rankings for Evaluation Methods	154
Table 5-5. Overall Rankings for Evaluation Methods Ordered.....	155
Table 7-1: Goodness Metric Results.....	178
Table 8-1: Caltech-256 Classes, Number of Images, Word Vector Representation, and PyDictionary Representation	184
Table 8-2. Classes with Definitions from Lexico	191
Table 9-1. Top 100 Words in Term Frequency List for Chandelier and Mars.....	193
Table 10-1. Definition and Definition Words for First Five Mystery Images.....	197
Table 10-2. First 100 Words in Term Frequency List for Mystery Class.....	202
Table 11-1. Term Frequency List for AK-47, Cactus, and Fireworks.....	203
Table 11-2. Term Frequency List for Floppy Disk, Frog, and Galaxy.....	206
Table 11-3. Term Frequency List for Iguana, Penguin, and People	209
Table 11-4. Term Frequency List for Sheet Music, Skyscraper, and Swiss Army Knife	212
Table 11-5. Term Frequency List for T-shirt and Waterfall.....	215
Table 12-1. Additional Runs Definition Evaluation Results	217
Table 12-2. Addition Runs AR Evaluation Results	217

Table 12-3. Addition Runs Modified HF Evaluation Results..... 218

ACKNOWLEDGMENTS

There are many places, groups, and individuals that have helped drive me to where I am and who I am today. I am so glad for the Wright State community I've been a member of for 4.5 years. First most, I want to thank my committee members, Dr. Assaf Harel, and two co-advisors, Dr. Subhashini Ganapathy and Dr. Trevor Bihl, for their guidance on my thesis and throughout my college career. Second, I am very appreciative of my church family at Northridge Freewill Baptist Church for sharing their spiritual wisdom with me. Third, I am grateful for all the friends I have met along the way. Regardless of the number of "friendship points" they have (or lack thereof), each of them has a special place in my heart. Next, I have the utmost respect for my family. Though in heaven, I know Poppy (Combs) and Papaw Slone would have been immensely proud of my accomplishments. Of course, I owe so much to my three main pillars of support, my grandma, dad, and mom. I cannot imagine where I would be without their unfathomable love. Lastly, I want to express my admiration for my favorite virtual classmate and frequent collaborator, my 15-year-old Shih-Tzu, Crystal. I am still trying to figure out exactly what her contributions were, but nevertheless, I am sure this would not have been possible without her.

1 INTRODUCTION

Shown throughout the entertainment world is the idea that robots, embodiments of artificial intelligence (AI), can recognize and detect objects almost instantly. However, the reality is significantly different for AI today. Operational AI is trained to understand, recognize, or act upon several known instances; however, like humans, it's not feasible to train AI on every scenario it may encounter, so it has some number of unknown scenarios, hence the rows of Figure 1-1. When placed into practice, the AI can observe or come into contact with something (a situation, object, etc.) that it either knows or does not know. The result is that the AI interaction involves one of four categories of possible results as shown in Figure 1-1 based on whether the entities are known (in-library) or not (out-of-library) ranging from correct classification (known knows), misclassifications (unknown knows), or various out of library situations (known unknowns and unknown unknowns) (Situ, Friend, Bauer, & Bihl, 2016).

		Predicted	
		Known	Unknown
Actual	Known	Known Knowns	Unknown Knowns (Mistakes)
	Unknown	Known Unknowns (Multiple classifications)	Unknown Unknowns

Figure 1-1. Knowns and Unknowns Matrix

In three of the categories of Figure 1-1, at least one portion is known, however, there is a significant amount of interest in exploring how to “learn” the unknown unknowns. Unknown unknowns would be exemplified by an attempt to recognize an object that a machine learning (ML) algorithm was not previously trained on. The motivation to explore this area includes the constant growth in automated systems and the inability to produce the number of models that can evaluate the problem in a known-knowns context (Bihl & Talbert, 2020).

The modern entertainment industry presents AI as being capable to solve the problem of the unknown unknowns almost instantly as shown in 2004 and 2008 films, *iRobot* and *Wall-E*. While both films take place later in the future compared to the present day, they leave the impression of AI being much more self-efficient than what it truly is. In both movies, AI can recognize an immensely broad array of objects and situations with seemingly minimal time needed for observation. This task is intrinsically complex and involves multiple AI processes, including image recognition, identification and classification of unknowns, and sophisticated reasoning logic. AI used in this context colloquially includes many methods and domains which involve pattern recognition, or ML; while ML is a subset of AI, colloquially AI/ML can be used to include many capabilities, ranging from classification and image processing to fully machine conscious computers.

To better exemplify the state of AI in the context of image recognition, the image shown in Figure 1-2.a was evaluated by a human (i.e., the author), and Google Cloud’s Vision AI. As shown in Figure 1-2.b, a human would easily identify many fireworks in the

sky and then, the water beneath the display. It is clear to a human observer that this image contains multiple objects; however, Vision AI struggled with this conclusion.



Figure 1-2.a. Original Fireworks Image, from (Griffin, Holub, & Perona, 2007)

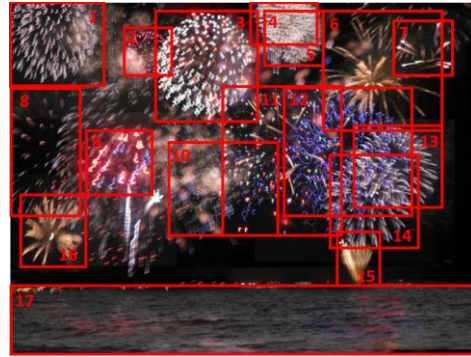


Figure 1-2.b. Fireworks Image Decomposed by a Human

Figure 1-2. Fireworks Images

Vision AI includes Vision API, which classifies, identifies, and detects a variety of objects/characteristics within an image (Google, 2021). Using their web demo of the tool, the same image shown in Figure 1-2.a was passed through and was evaluated in two different contexts, object recognition, and image labeling. Vision AI only identifies on the object, denoted in the green box in Figure 1-3, as lightning with a score of 51% (where the “score” is a value ranging from no confidence, 0%, to high confidence, 100% (Google, 2021)).

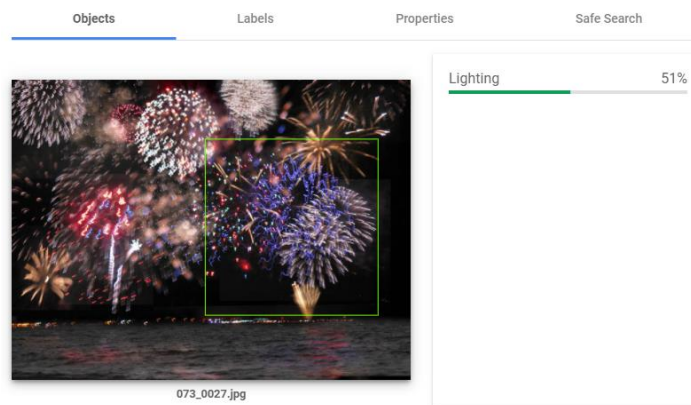


Figure 1-3. Fireworks Image Decomposed by Google Cloud Vision AI (Google, 2021)

However, Vision AI improves its prediction significantly when attempting to only label the image as a whole instead of searching for specific objects. These results, the ranking, label, and score, are shown in Table 1-1. Rankings denoted with a “t-“ at the beginning represent a tie in the score. At the top of the rankings, these labels seem appropriate for the image especially since “fireworks” appears at the top with a score of 96%. Several labels stir curiosity regarding how the algorithm works. Despite having a score of 77%, “landmark” and “space” are inaccurate if taking the image at face value. Several labels would seemingly be difficult to generally visualize such as “midnight,” “event,” and “holiday.” Lastly, some labels may or may not be accurate based on the context of the label’s usage (ex. homophones such as “light” in the sense of brightness or light-weight, both of which happen to be appropriate here) as well as the context in which the picture was taken (ex. “New Year’s Eve,” “Diwali,” and “Chinese New Year”).

Table 1-1. Google Cloud Vision AI Label Predictions

Ranking	Label	Score	Ranking	Label	Score
1	Fireworks	96%	17	Lake	69%
2	Water	93%	t-18	New Year’s Eve	68%
3	Light	91%	t-18	New Year	68%
4	Nature	90%	t-18	Public Event	68%
5	Entertainment	86%	t-21	Reflection	67%
6	Sky	85%	t-21	Festival	67%
8	World	84%	t-23	Diwali	65%
9	Pink	83%	t-23	New Years Day	65%
10	Midnight	80%	25	Night	64%
11	Landmark	77%	26	Horizon	62%
t-12	Darkness	75%	27	Spectacle	61%
t-12	Event	75%	28	Chinese New Year	60%
t-14	Electric Blue	73%	t-29	Recreation	59%
t-14	Space	73%	t-29	City	59%
16	Holiday	71%	31	Pollution	57%

Taking a step back, this is likely a known-known situation; however, looking beyond the “fireworks” label in Table 1-1, the remaining top classifications (score greater than or equal to 90%) are on classes that do not describe the image, e.g., “water,” “light,” or “nature”. This is where image classification offers very narrow results due to its limitations to the classes/labels that it's aware of. Being able to accurately explain or identify these unknowns is of great interest to the current literature. One proposal to solve the unknown unknowns is through the application of analogical reasoning (AR), thereby reasoning/learning through analogies.

1.1 TECHNICAL MOTIVATION

Many images classification algorithms were created for the ImageNet Large Scale Visual Recognition Challenge (ILSVRC), which ran from 2010 – 2018 (Russakovsky, et al., 2015; Stanford Vision Lab, 2020). The ILSVRC primarily looked at three different tasks: image classification, single-object localization, and object detection (with some variations between individual years) (Russakovsky, et al., 2015). The data set consists of 1000 different classes with over a million training, 50,000 validation, and 100,000-150,000 test images (Russakovsky, et al., 2015). The winners in 2010 and 2011 used “shallow” artificial neural networks (ANNs); however, starting in 2012, the competition saw its first entry using deep ANNs, which remained popular through the lifespan of the competition (Russakovsky, et al., 2015). These deep ANNs are successful in the image-classification realm but require a significant amount of time and high-performing computational resources. The algorithms, such as ANNs applied to the ILSVRC, are trained on a certain number of familiar instances and thus handle known knowns. However, such algorithms

are either entirely incapable or perform poorly when posed with an unexpected query, i.e., a new image class that was not presented in the initial release.

Such issues are where AR has great potential to improve AI results. AR can extract information from an unexpected query based on information the algorithm already knows. Mimicking how humans use analogies to learn, an algorithm can do the same without the need for additional training scenarios, more computational resources, and/or unreasonably extending the runtime needed. Of interest is thus the different types of AR algorithms in existence and how they have been or can be integrated with current state-of-the-art image recognition programs.

Many AR algorithms exist that focus on various tasks from both the verbal and visual realms. However, these are often limited to either verbal or visual problems, with little overlap on leveraging information from both. In addition to that, many visual AR algorithms are focused on geometric-based problems, c.f. (Polya, 1990; Sadeghi, Zitnick, & Farhadi, 2015), which do not apply to image-classification problems as posed above. Thus, of interest is using AR in an image-recognition context to handle problems involving unknowns.

1.2 OPERATION MOTIVATION

Image recognition is just a small portion of AI research; however, it has one of the greatest impacts on everyday life. Some examples include facial ID recognition used to unlock our phones, image-to-text automated caption generators, self-driving cars, among many others. The consequences for inaccuracies and unknowns in these scenarios largely vary from being a mild inconvenience (i.e., having to manually unlock a phone) to a

potentially life-threatening event (i.e., a self-driving car does not detect a pedestrian). As the day-to-day uses of AI increase and the consequences scale up, the need for accurate AI which can handle unknowns also increases.

Specifically looking at the self-driving car scenario, there are many different things, objects, and/or people the image recognition algorithm would need to recognize and it is increasingly impossible to collect data for all possible real-world situations. Consider a stop sign for instance, in viewing one a variety of factors can change its representation, such as glare, lighting, obscuration, damage, sun-angle, background, paint quality, look-angle, mounting height, and more. Since it is impossible to collect data for every possible one of these scenarios, let alone for other objects, being able to reason by analogy that an observed stop sign with faded paint is similar to what known stop signs look like and then decide this is likely a stop sign and then direct the car to stop.

1.3 RESEARCH CONTRIBUTION

Since its start with the work of Polya in 1954, algorithmic AR approaches have been developed first with Evan's ANALOGY program in 1964 (Polya, 1990). Since then, many avenues of AR have been explored. The technical areas most relevant to the author's contributions are listed in

Table 1-2, with example references of recent prior work (2000 and later) as well as the research conducted by the author in this thesis (Combs, 2021) or a separate article (Combs, Bihl, Ganapathy, & Staples, 2022).

Table 1-2. Relational Mapping Between Previous Technical Contribution(s) and Current Research Contributions (Denoted by X in the “Focus” Columns)

Technical Area	Prior Work		Current Work	
	Focus	Example References	Focus	References
AR Textual Models/ Algorithms	X	(Eliasmith & Thagard, 2001); (Doumas, et al., 2008); (Lu, et al., 2012); (Mikolov, et al., 2013); (Levy & Goldberg, 2014); (Drozd, et al., 2016); (Speer, et al., 2017)	X	(Combs, 2021)
AR Algorithm Comparisons	X	(French R. M., 2002); (Kokinov & French, 2003); (Leech, et al., 2008); (Genter & Forbus, 2010); (Rogers, et al., 2017); (Chen, et al., 2017); (Mikolov, et al., 2018); (Peterson, et al., 2020)	X	(Combs, 2021); (Combs, et al., 2022)
Interdisciplinary AR Comparison with Metrics			X	(Combs, 2021); (Combs, et al., 2022)
Image-based AR	X	(Yaner & Goel, 2006); (Doumas & Hummel, 2010); (Hwang, Grauman, & Sha, 2013); (Sadeghi, Zitnick, & Farhadi, 2015); (Reed, Zhang, Zhang, & Lee, 2015);		
Image and text to AR	X	(Lu, Liu, Ichien, Yuille, & Holyoak, 2019)		
Image-text to AR			X	(Combs, 2021)
AR Algorithm Taxonomy			X	(Combs, 2021); (Combs, et al., 2022)
AR Comparison Metrics	X	(Leech, et al., 2008); (Genter & Forbus, 2010); (Rogers, et al., 2017); (Chen, et al. 2017); (Mikolov, et al., 2018); (Peterson, et al., 2020)	X	(Combs, 2021); (Combs, et al., 2022)

Correctness Metric	X	(Morrison, et al., 2004)	X	(Combs, 2021); (Combs, et al., 2022)
Goodness Metric			X	(Combs, 2021); (Combs, et al., 2022)
Contextual Metrics			X	(Combs, 2021)

1.4 RESEARCH OBJECTIVES

Understanding the technical and operation motivations to better attempt an unexpected query, the objective of this thesis aims to improve image recognition in the presence of unknown unknowns through the development of an AR-augmenting framework. There are many ways in which image recognition is being developed; however, they are limited in their ability to interpret beyond the “known” corpus. With its structure around familiar and unfamiliar scenarios, AR has previously been used and will be used to generate information from previously unfamiliar scenarios. To meet these objectives, the research and development processes were broken down into four sections.

Firstly, in Chapter 2, a comprehensive understanding of AR algorithms, both centered around textual and visual problems, is needed to understand the current state of AR. Since this is taking place in the context of an image-classification problem, a brief portion dedicated to research in image recognition and convolutional neural networks (CNNs) is also found here. Secondly, in Chapter 3, due to the varieties of AR algorithms in the literature, a broad comparison is developed to select the best of breed in AR for further use in the image-based problem. Six text-based AR algorithms, from both the hybrid and connectionist families, are compared on two metrics evaluating correctness and goodness. Next, in Chapter 4, a new AR-integrated algorithm for image classification of unknown unknowns is described in detail. This section talks about the data set used to test

the algorithm, how the algorithm works (a technical description and 3 step-by-step walkthroughs), and lastly, the results generated by the algorithm. Finally, in Chapter 5, the two automated methods used to evaluate the results as well as a third human-based analysis to use as a baseline are discussed in the context of selected “unknowns.” Chapter 6 concludes the thesis with a general discussion of the novelty of the research in the context of an image classification problem and future work on how AR can be used within other unknown unknown situations.

2 BACKGROUND

Analogical reasoning (AR) is a technique where one learns through analogies. Analogies are unique figure-of-speeches that map two different objects or scenarios based on their similar individual elements (Gentner & Maravilla, 2018; Bailer-Jones, 2002). AR and understanding analogies extend beyond low-abstraction based approaches, such as sentiment analysis and word usage (Bihl & Bauer Jr., 2017), which are subjective in nature, to understand the semantics of concepts and their meanings through low-level features that provide in-depth information (Bihl & Bauer Jr., 2017). An analogy also differs from generalization and specialization in how a polygon is a generalization of a triangle, an equilateral triangle is a specialization of a triangle, but pyramids and tetrahedrons are analogous to triangles (Polya, 1990).

To create an analogy, one takes a familiar situation (called the “base”) and attempts to identify parallels with an unfamiliar or incomplete scenario (called the “target”) (Gentner & Smith, 2012). One of the reasons analogies are of interest to researchers is because they cannot be created through a strict formula or method. Some analogies may need to be interpreted more than others or require more background knowledge (Khatena, 1972). In one of Khatena’s examples where participants were asked to provide a graphic analogy given a stimulus, the word “jingle” yielded the responses “a mesh of fishhooks” and “crickets in harmony.” In addition to variation depending on the individual, there are also different levels of abstraction of an analogy depending on how simple or complex it is (Mitchell, 1993). Analogies can only occur at a “high level” of perception, through the usage of concepts, relationships, and situations; whereas “low-level” perception is based purely on raw information gathered by the senses (Chalmers, French, & Hofstadter, 1991).

These can be paralleled to abstraction levels in textual analysis where high-abstraction is general and concrete in interpretation (such as word count) contrasted with low-abstraction, which is specific and more fluid in interpretation (such as sentiment analysis) (Bihl & Bauer Jr., 2017). Likewise, to high-level perception and/or low-level abstraction, a complex analogy is built around concepts coming from one’s long-term memory and identifying parallels to a base situation that may or may not use synaptic cues (Mitchell, 1993). In Figure 2-1, the author(s) of Copycat, a hybrid and prominent AR algorithm presented the following analogy problem using alphabetic characters.

	$ABC \rightarrow ABD$	
	$PQR \rightarrow ???$	
A. PQS	B. PQD	C. PQR

Figure 2-1. Visible Action of an Example Copycat Analogy, adapted from (Hofstadter & Mitchell, 1995)

Copycat identifies three potential solutions and their corresponding rules to the problem displayed in Figure 2-1:

1. PQS – replace the rightmost letter with its successor
2. PQD – replace the rightmost letter with D
3. PQR – replace all C’s with D’s (Hofstadter, 1984).

Though all the rules may be argued as correct, how does a computer determine which one is “most correct” or most likely to correspond with a human’s answer? How “similarity” should be measured is not always clear (Potts, 1978). In the early days, where the majority of analogical reasoning work was done through human trials, there were two approaches where the body of information was predefined by the researcher and one where it was not (Potts, 1978). As the researcher instills “artificial” information, in the sense that it was not there before the experiment, in the modern era, a programmer would have to follow a

similar process to create a sense of “memory” for a machine (Potts, 1978). At its start, new AI only has unknown unknowns similar to a young child, but yet a child is eventually able to learn intricate ideas through methods such as AR.

Initially, a child’s ability to use analogical reasoning was measured in two ways: IQ tests and problem-solving tasks (Goswami, 1991). Due to inaccuracies and inconsistencies with children’s answers (such as when applied in a computer program), there has been substantial work done regarding how this reasoning develops in children (Goswami, 1991). The first study looking at how children use analogical reasoning was conducted by Jean Piaget (Goswami, 1992). Similar to Piaget’s theory of cognitive development, he argued that children are not able to truly reason by analogy until in the third stage called “formal operational” (Goswami, 1991). In the experiment shown in Figure 2-2 looking at the analogy, *Bird:Nest::Dog:?*, in a visual sense, children as young as 4 were able to identify *D*, the Doghouse, as the correct answer (Goswami, 1992).

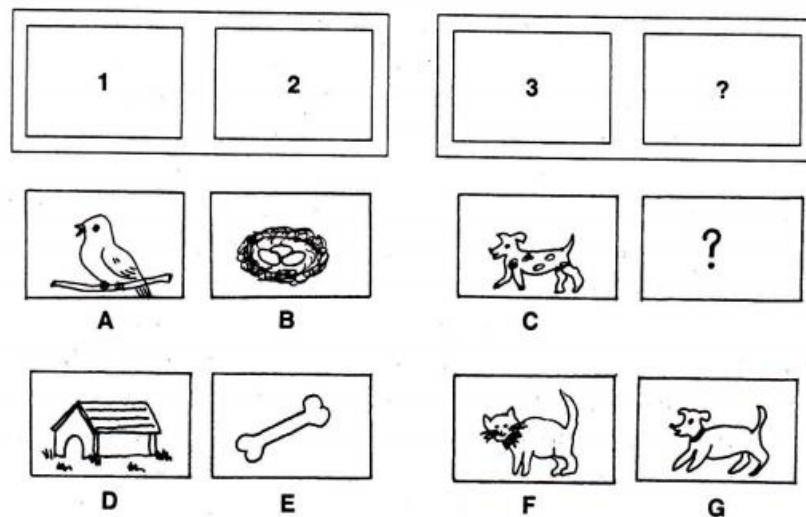


Figure 2-2. Bird to Nest as Dog is to ? Analogy, from (Goswami & Brown, 1990)

In one study, children who were introduced to true analogies before being given an incomplete analogy problem were able to more accurately select the correct response than children who were not introduced to true analogies (Levinson & Carpenter, 1974). In another study with infants between the ages of ten and thirteen months, where they needed to remove a barrier, pull on a cloth, and then pull on a string to get to the toy attached to the other end in three different, but similar scenarios (Chen, Sanches, & Campbell, 1997). Though some children needed their parents to show them the initial process, they were able to reasonably apply it to the remaining scenarios (Chen, Sanches, & Campbell, 1997). In the 1980s, Gentner and her colleagues developed the Structure-mapping Theory (SMT) which introduced the idea of a “Relational Shift” due to children’s focus on a base and target’s similar elements or characteristics (Goswami, 1991). This phenomenon was tested in several studies (Genter & Toupin, 1986; Gentner, 1988), which showed children’s ability to use it to an extent in story mapping and interpreting metaphors (Goswami, 1991). This is just a small portion of the domains in which analogical reasoning among children has been conducted. This strategy has been applied to assist understanding of biological principles, “Piagetian” tasks, pairs of relations, transitive mapping, and class inclusion among many more experiments that are too many to number (Goswami, 1992). AR’s proof of concept in children gives the potential for its impact in the world of AI.

2.1 MACHINE INTELLIGENCE

Mimicking human reasoning is a difficult task for AI researchers and there is a psychological debate about how the human mind thinks. Reasoning is a precursor to intelligence, and despite being a complex process, human minds can do it rather

automatically (Evans, Newstead, & Byrne, 1993). First identified by Charles Sanders Peirce in the late 1800s and early 1900s, three types of reasonings were identified: abduction, deduction, and induction based on the scientific method (Flach & Kakas, 2000). These types are not alternatives to one another but rather, they work together to create a logical conclusion. In a theoretical situation shown in Figure 2-3, a person will make observations and create a hypothesis (abduction). Using the hypothesis, they will make predictions about the real world (through deduction). Finally, these predictions are validated and assumed to be true to reality (via induction).

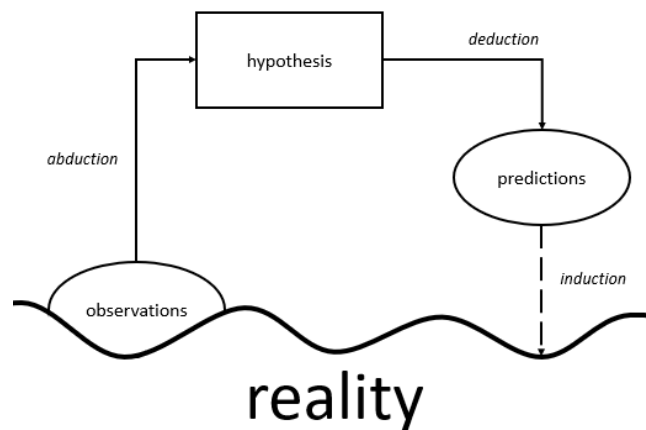


Figure 2-3. Three Stages of Scientific Inquiry, adapted from (Flach & Kakas, 2000)

All of these types of reasoning are very easy and commonly done by humans; however, mimicking this process in computers has proven to be arduous. The rational paradigm believes that computer science is most similar to mathematics due to the use of deductive reasoning to confirm how “correct” a program is (Eden, 2007). Since there is no inherent logic structure, such as the human brain, computers’ “logic” is built through “if”, “and”, and “or” statements (Johnson-Laird, 2010). These logic statements were the building blocks for the initial artificial neural network (ANNs) designs for reasoning aligned with McCulloch and Pitts’ 1943 theory (Stenning & Van Lambelgen, 2012;

McCulloch & Pitts, 1943). The initial models were logical programs, which have been transformed into more sophisticated models such as recurrent neural networks with considerations for other types of reasoning (Stenning & Van Lambelgen, 2012).

This general understanding is extended in Figure 2-4 whereby the relationship between abduction, deduction, and induction is presented as one large feedback loop and with the incorporation of abstraction. Furthermore, the diagnostic hypothesis node has an individual loop that looks to structure the diagnostic space of the problem. As more information is learned about the problem, the hypothesis and the problem, in general, can likely be narrowed or re-focused. Similarly, the observed/expected data node is reinforced by a new data request loop. Again, as more information is discovered, more information is probably needed to help direct research efforts in the intended direction.

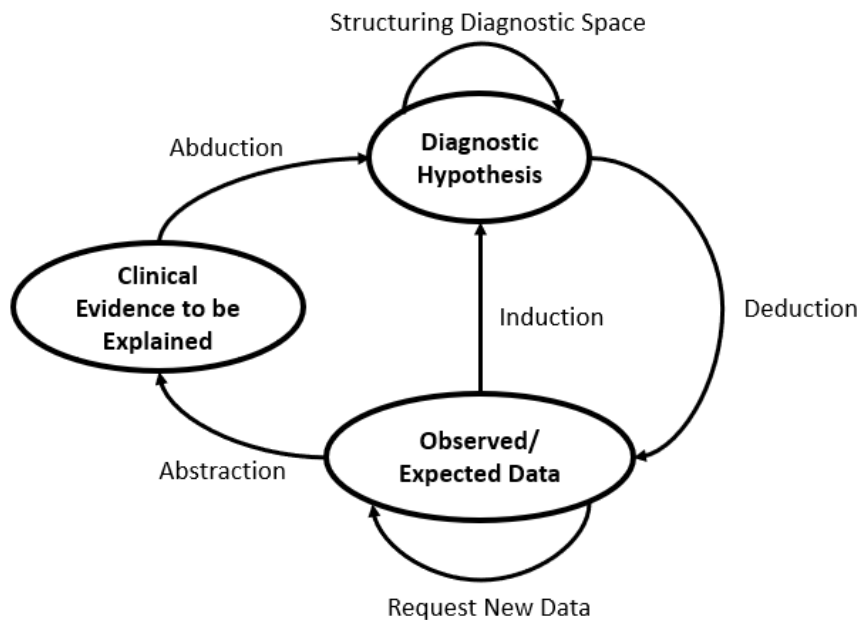


Figure 2-4. Relationship Between Abduction, Abstraction, Deduction, and Induction, adapted from (Weigand & Hartung, 2012)

Current approaches in AI/ML involve primarily the inductive reasoning aspect(s) of learning whereby large amounts of data are used to train predictive (classification or

regression) algorithms to represent data in a lower-dimensional form (Duda, Hart, & Stork, 2000). Such approaches, when taken to the extreme current state of the art in deep learning, are constantly learning and adapting to new information during training, but become static once training is completed and inflexible. For a machine to become human-like it must have the following properties (Summers-Stay, 2017):

“ - Be capable of associational, analogical, inductive, abductive, and deductive reasoning;
- when exact answers can't be found, guess at an approximate answer;
- be aware of the strength or weakness of its arguments;
- creatively find connections that were not deliberately given, and
- find arguments that add up to a whole, rather than find strictly linear connections”.

To bridge the gap between current AI systems and the principles outlined above, AR is one potential key whereby it could allow computers to identify possible and accurate alternatives despite not having an exact answer for a given scenario (Summers-Stay, 2017).

2.2 ALGORITHMIC ANALOGICAL REASONING APPROACHES

Reasoning by analogies is one of the many ways to understand a new topic by portraying information in a different, sometimes indirect, manner. However, analogies do not have a uniform style in terms of the language used or the sentence structure. Despite this seeming disconnect, the human brain is theorized to have a common process of creating them: (i) retrieval, (ii) mapping, and (iii) evaluation (Gentner & Smith, 2012). An example of the mapping process is shown in Figure 2-5. As depicted in Figure 2-5, an analogy is

broken into the “base” and “target” portions, consisting of objects (squares/rectangles) and relations (circles). The base is mapped to the target using the orange lines shown in Figure 2-5(a). Next, a “candidate inference” is made regarding a known portion of the base and the likely relationship placement shown on the target (orange arrow in Figure 2-5(b)). Lastly based on some measure of confidence, through abstraction, the target accepted the candidate inference to be truly shown in Figure 2-5.c.

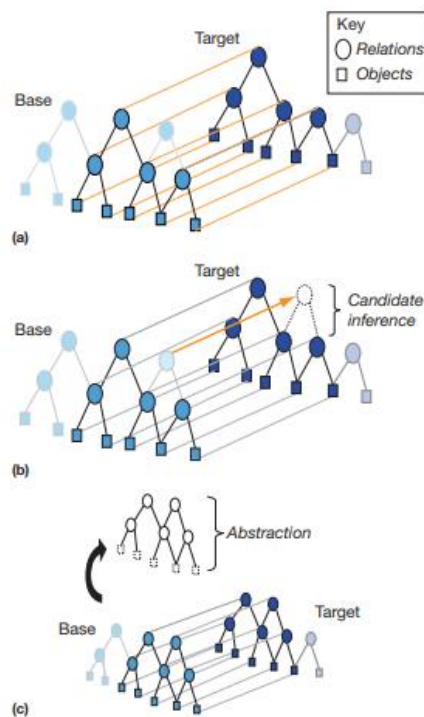


Figure 2-5. Structure-mapping-based Analogy Process Visual, from (Gentner & Smith, 2012)

An additional fourth step to analogy-creation is considered to be “subsequent learning” (Holyoak & Thagard, 1989; Thagard, Holyoak, Nelson, & Gochfield, 1990); which requires the structure to be re-created to account for the new knowledge it has gained. Taken together, analogical reasoning thus can be considered through four steps: (i) recognition of the source, (ii) elaboration where there is a mapping between the source and

target, (iii), evaluation of the mapping(s), and (iv) consolidation regarding the analogy result (Hall, 1989).

Learning by analogy is further primarily focused on three functions: generalizing, contrasting, or re-representation (Gentner & Maravilla, 2018). In the generalization function, also called “schema abstraction,” an analogy’s purpose is to pull out the common elements between the target and base (Gentner & Maravilla, 2018). The contrasting type also called “difference detection” does the exact opposite of generalizing, but points out how the target and base are different and to what extent (Gentner & Maravilla, 2018). Finally, the re-representation function allows an analogy to be formed despite having “nonidentical conceptual relations” such as with the phrases, “Attila burned the fort” and “Napoleon torched the castle” (Gentner & Maravilla, Analogical Reasoning, 2018).

Additionally, the breadth of analogical reasoning problems is quite large. They range from simple $A:B::C:D$, i.e. the words, “A is to B as C is to D,” to complicated story problems (Ichien, Lu, & Holyoak, 2020). In the $A:B::C:D$ category of analogies, five primary types are identified (Ichien, Lu, & Holyoak, 2020):

1. Evaluation - valid vs not a valid analogy, e.g. “Is Man:Boy::Woman:Girl a valid analogy?”
2. One-term Generative - A, B, and C are given, but D is missing and left open-ended, e.g. “Man is to boy as woman is to what?”
3. Two-term Generative - A and B are given, but C and D are missing and left open-ended, e.g. “Man is to boy as what other word pair?”

4. Multiple-Choice - A, B, and C are given, but D is missing but options for D are given, e. g. “Man is to boy as woman is to what? (A) girl, (B), mother, (C) female, or (D) daughter”,
5. Matrix - given many $A:B$ analogies, need to identify the two that are most alike, e.g. which words pair best with one another to form an analogy? (A) man:boy, (B) woman:girl, (C) man:woman, (D) man:father, etc.

Word problems as they pertain to analogy come in four main types: (i) Retrieval (recall elements from the original (“target stories”) while reading “cue stories,” (ii) Generative (given a story re-create similar story with different elements), (iii) Problem Solving (using a similar story to solve a problem), (iv) Extended Mapping (A and B are given and given multiple C words or phrases, their corresponding D’s are selected from the options) (Ichien, Lu, & Holyoak, 2020). Considering the breadth in problem types, there is a significant number of algorithms each optimized to focus on a specific portion of AR or specific AR problems.

2.2.1 Text-based Analogy Methods

In the early explorations in AR methods, the analogies evaluated were primarily textual rather than visual, i.e. images. Originally most text-based algorithms were considered to be symbolic due to representing the source and target as objects and the relations between them (top-down); however, now there is a focus on connectionist algorithms due to its increased ability to consider the similarity between the source and target from the bottom-up (French, 2002).

At a high level, artificial AR is an AI approach, and understanding it requires a general knowledge of the AI schools of thought: symbolist, connectionist, and dynamicist

(Eliasmith, 1997; Zhang, 2008). These schools of thought differ largely on how intelligence is understood and conceptualized through artificial means. Briefly, symbolism considers the mind to be a computer/logic system, connectionism considers the mind to be a neural network, and dynamicism considers the mind a watt governor (Eliasmith, 1997). Given that biological mental processes likely follow a combination of these approaches (or something yet to be discovered), hybrid AI paradigms are also of interest as discussed by Eliasmith (2013).

AR algorithms, similarly, are structured according to these paradigms, particularly: symbolist and connectionist (with some models being hybrids) (Kokinov & French, 2003) (Gentner & Forbus, 2010). In AR applications, symbolist approaches consider each element of an analogy to be separate and independent from one another similar to a top-down approach (Kokinov & French, 2003). Originally, the first AR methods were symbolic, beginning with Evan's 1963 ANALOGY model for visual AR problems (Kokinov & French, 2003). Later in 1989, Gentner's word-based structure mapping theory (SMT) would be turned into the influential AR algorithm, the structure mapping engine (SME) (part of the Many Are Called but Few Are Chosen (MAC/FAC) program) (Forbus, Gentner, & Law, 1995; Holyoak & Thagard, 1989).

Though AR's origins started with symbolist algorithms, currently, there is a push toward connectionist ones (Kokinov & French, 2003). These models are characterized by elements that are associated using a bottom-up approach; many do this in a distributed fashion. The first connectionist algorithm was Holyoak and Thagard's 1989 Analogical Constraint Mapping Engine (ACME), though its methods followed symbolist ideals more so than today's standard for connectionism (Holyoak & Thagard, 1989). However, some

more recent algorithms include Structure Tensor Analogical Reasoning (STAR) (Halford, et al., 1994; Wilson, Halford, Gray, & Philips, 2001), Learning and Inference with Schemas and Analogies (LISA) (Hummel & Holyoak, 1997), Discovery Of Relations by Analogy (DORA) (Doumas, Hummel, & Sandofer, 2008), and Bayesian Analogy with Relational Transformations (BART) (Lu, Chen, & Holyoak, 2012; Lu, Wu, & Holyoak, 2019). STAR is a tensor-product-based parallel distributed processing model embedded in a neural network (Halford, et al., 1994), a framework popular for many AR algorithms to come. LISA uses a neural network to process analogies while modeling a human's short-term and long-term memory (Hummel & Holyoak, 1997). DORA focuses on improving and incorporating self-supervised learning (SSL) into LISA (Doumas, Hummel, & Sandofer, 2008). SSL has enabled role-fillers to fire asynchronously; whereas in LISA once fired, all corresponding semantic units are activated (Doumas, Hummel, & Sandofer, 2008). Additionally, VSMs have been included in the connectionist paradigm due to operating in a distributed fashion. Latent Relation Analysis (LRA) was one of the first VSMs created in 2006 (Turney, Similarity of Semantic Relations, 2006); however, since then, the creation of Word2vec, Global Vectors (GloVe), 3CosAvg, and LRCos, as well as many others, has been accomplished.

Considering the benefits of both the symbolist and connectionist algorithms, some research has investigated hybrid algorithms that incorporate the best of both (Kokinov & French, 2003). The first hybrid algorithm was Copycat which had a unique domain of nonsensical strings (example: $ABC:ABD::PQR:\{PQS, PQD, \text{ or } PQR\}$) (Hofstadter & Mitchell, 1995). Copycat later inspired the creation of an action-based analogy program called Tabletop (French R. M., 1995). The first generally accepted word/sentence-based

hybrid algorithm was created in 1994, called the Associative Memory-Based Reasoning (AMBR) model (Kokiov, 1994) (Petrov, 1997), which was followed by Distributed Representation Analogy Mapper (DRAMA) (Eliasmith & Thagard, Integrating structure and meaning: A distributed model of analogical mapping, 2001). Few hybrid algorithms exist due to their complexity compared to the number of symbolist and connectionist algorithms (Gentner & Forbus, 2010).

Following this reasoning, a general taxonomy of AR algorithms appears in Figure 2-6. While no known dynamicist AR algorithms exist to date, this paradigm of AI is included for completeness.

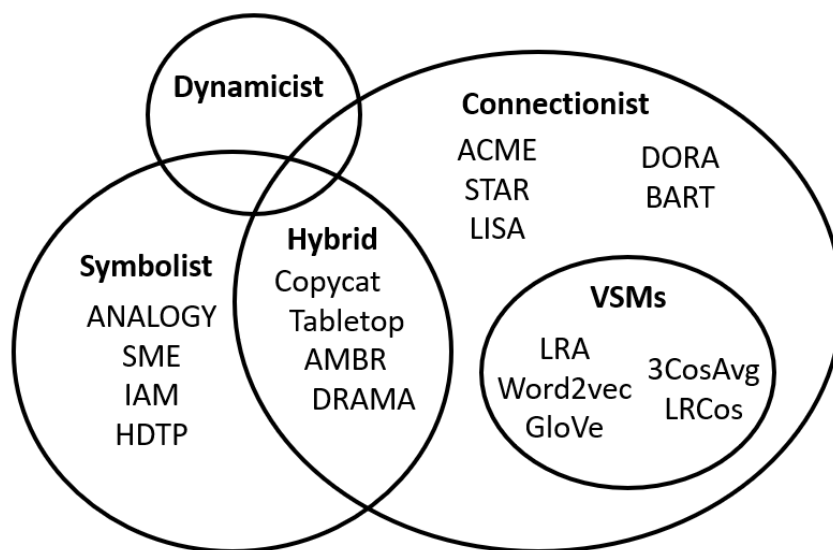


Figure 2-6. AR Models in the Context of AI Schools of Thought, from (Combs, Bihl, Ganapathy, & Staples, 2022)

2.2.1.1 Rumelhart's Model

In Rumelhart's model, the authors created a 3-dimensional Euclidean space that mapped all the semantic relationships among, in this case, animals. Rather than being a repeated set of steps and based on human results, Rumelhart's AR result(s) is considered to be a model rather than an algorithm. The semantic mapping originated from Henley's

prior work which compiled data from human test subjects who mapped animals on a 4-quadrant grid based on two dimensions (Henley, 1969). Henley's test subjects were asked to participate in five experiments: listing, pair rating, triad rating, associations, and pair-associated learning; however, Rumelhart asked participants to rank the options from best to worst in regards to how to complete the given analogy. Based on the participant's answers, Rumelhart introduced a new shift in the location of some animals on Henley's graph which is shown in Figure 2-7. Notably, this method may become overly cluttered and messy if used in broader scenarios, and it was still strongly reliant on human data (Rumelhart & Abrahamson, 1973).

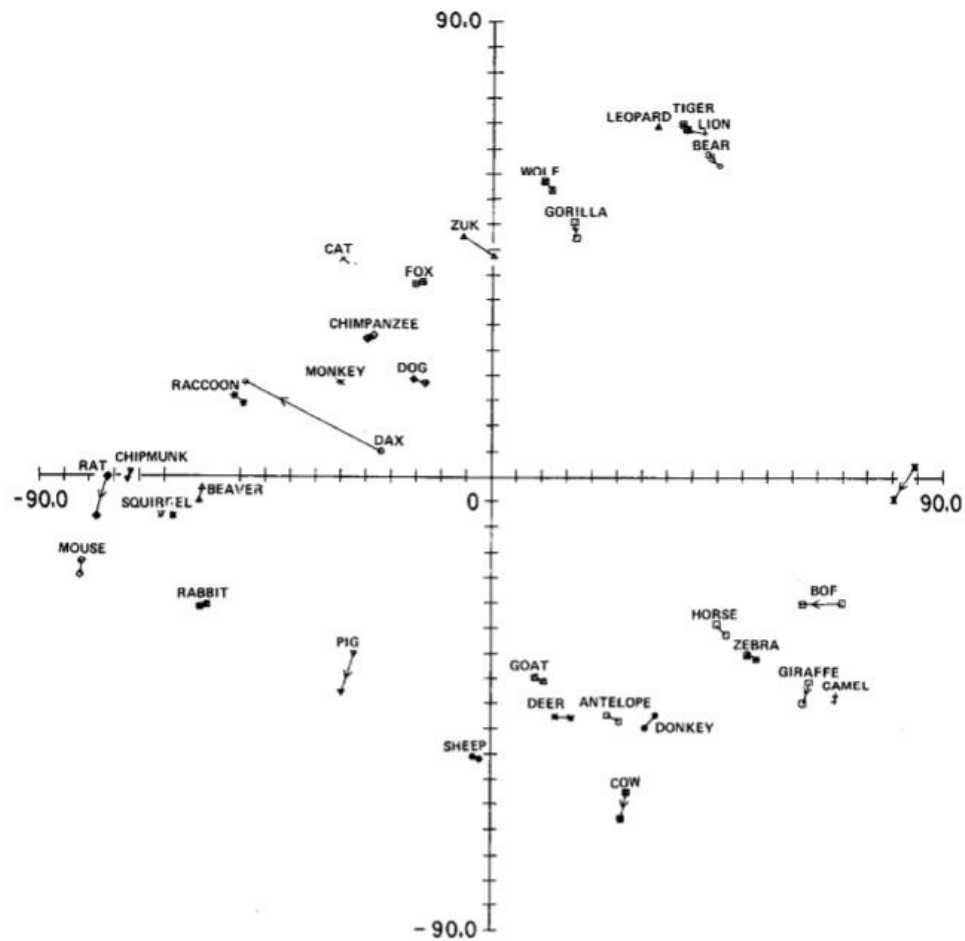


Figure 2-7. Analytical Shift in Animal Location Comparing "Size" and "Ferocity" from shifting from Henley's to Rumelhart's Model, from (Rumelhart & Abrahamson, 1973)

2.2.1.2 *Structure-Mapping Engine (SME)*

Touched on briefly earlier in the introductory section of this chapter, Section 2.2.1, the structure-mapping theory (SMT) was the first prominent analogical reasoning algorithm, which revolved around objects and their relationships (Leech, Mareschal, & Cooper, 2008). The SMT was uniquely proposed due to the idea that (i) the relations are what should be mapped (as opposed to the object's attributes) and (ii) analogies can only be formed from a collection of overarching knowledge (Gentner, 1983). SMT has been used as the basis for an analogy-creating program called the Structure-mapping Engine (SME), which breaks down the analogy creation into three subprocesses: (i) access, (ii) mapping and inference, and (iii) evaluation and use (Falkenhainer & Forbus, 1989). SME was also incorporated into the larger "Many Are Called but Few Are Chosen" (MAC/FAC) model (Forbus, Gentner, & Law, 1995). MAC identifies items in long-term memories via a non-structural matcher (that notable uses "content vectors" in its first stage (Forbus, Gentner, & Law, 1995)). The second stage, FAC, involves SME identifying structural matches among the items found in the earlier stage (Forbus, Gentner, & Law, 1995). SME identifies the similarities between the base and target analogies and identifies the best "match" based on its structural similarity, validity with the real world, and relevance as a test of usefulness (Falkenhainer & Forbus, 1989). One of the benefits of SME is its ability to produce multiple interpretations for a given analogy; however, the quantity is strongly reliant on how similar the base and target is, the identical constraint is limiting, and its inability to solve certain analogies in a reasonable amount of time (Falkenhainer & Forbus, 1989). Written in Lisp, SME has continuously been expanded with the most recent, version 4, being published in 2017 (Forbus, Ferguson, Lovett, & Gentner, 2017).

2.2.1.3 *Analogical Constraint Mapping Engine (ACME)*

Structure-mapping theory can be contrasted with the Analogical Constraint Mapping Engine (ACME), which emphasizes the semantic similarity between the elements of the base and target (Holyoak & Thagard, 1989). ACME incorporates three constraints that it considers before creating an analogy which is as follows (i) isomorphism (one-to-one mapping), (ii) semantic similarity, and (iii) pragmatic centrality (practicality). These are “soft” constraints, which are not required for every analogy, but good analogies are assumed to have a balance among the three (Holyoak & Thagard, 1989). Though ACME assists in the mapping process of analogy formation, it requires the companion program, Analog Retrieval by Constraint Satisfaction (ARCS), to assist with how potential bases are selected from memory (Holyoak & Thagard, 1989). ARCS retrieves elements from memory using similar constraints as ACME but does not compare or measure the quantity of the chosen elements (Thagard, Holyoak, Nelson, & Gochfield, 1990).

2.2.1.4 *Copycat*

Though first theorized in 1984 (Hofstadter, 1984), Copycat’s internal programs were not finalized until the early 1990s (Hofstadter & Mitchell, 1995; Mitchell, 1993). At its earliest installment, Copycat consisted of two programs: (i) Jumbo, a stochastic search program, and (ii) Slipnet, a concept network navigated by Jumbo (Hofstadter, 1984). Rather than word-based analogies as used in the previously mentioned literature, Copycat’s domain only looks at “codelets” which are the letters, “A” through “Z” of the English alphabet due to its inherent structure and relationships between “tokens” (aka “token letters”) (Hofstadter, 1984). Figure 2-1 shows the transformation from ABC (the “prototype”) to ABD (the “result”) and given PQR (the “target”), asking what three letters

should appear for the “goal.” In the next iteration of Copycat, Jumbo was removed and replaced by the two new programs called the Workspace and Coderack (Hofstadter & Mitchell, 1995). The Slipnet still acts as Copycat’s long-term memory; however, its Workspace is its short-term memory and the Coderack is its stochastic “agenda” tool (Hofstadter & Mitchell, 1995).

In comparison to SME, Copycat allows for semantics considerations, “conceptual similarity” called “slippage,” and removes the assumption of predicate logic (Mitchell, 1993). Specifically, Copycat is unique due to its ability of “conceptual slippages” which find relationships between non-identical entities, in other words, the “letters” (Bolland, 2004), this is shown in Figure 2-8.

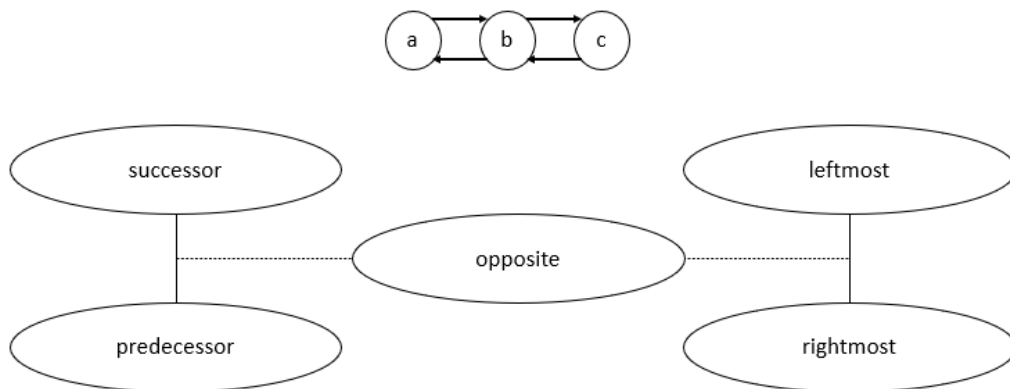


Figure 2-8. Example Slipnet Potential Slippages, adapted from (Bolland, 2004)

However, the Slipnet changes the “length” (activation levels for a given link) which directly affects the likelihood for different slippages to occur partly due to being coded was the global rather than local level (Bolland, 2004). In comparison with ACME, Copycat can attempt all possibilities given the syntax, relaxes the need for descriptions to be hard-coded by a programmer, and introduce a new element of randomness in the construction of the analogies (Mitchell, 1993). Though Copycat’s use of alphabetical strings led it to be subject

to criticism due to lacking a “real-world” context (Mitchell, 1993). An additional limitation was the lack of memory regarding previous “paths” traveled by Copycat, which allowed the program to arrive at the same answer despite multiple runs (Bolland, 2004).

Though more of an offshoot from Copycat’s original intentions, Tabletop attempts to apply the same abstract methods to actions, specifically in the scenario where two people are sitting at a dinner table with one mimicking the other’s movements (French R. M., 1995). Later, the limitation regarding Copycat’s memory of the paths travel was addressed by the program, Metacat, which consists of the programs, “Thespace,” “Temporal Trace,” and Episodic Memory (Marshall, 1999). Additionally, Metacat can recognize when it is “stuck” and unable to create an acceptable solution, called “joosting” (Marshall, 2002). A more recent expansion of Copycat, the Fluid Analogies Engine (FAE), reduced the need to code “conceptual slippage” but rather allows it to occur through a local means and allowed for information transformation through looking at the differences between the base analogy rather than just using rules (Bolland, 2004).

2.2.1.5 Associative Memory-Based Reasoning (AMBR)

Similar to Copycat’s hybrid nature, AMBR was created with considerations for both the symbolist and connectionist theories begun in 1988 and consists of two iterations AMBR1 (referred to simply as AMBR for the remainder of this paper) (1994) and AMBR2 (1998) (Kokinov & French, 2003). It is important to note that AMBR is reliant on the cognitive architecture, DUAL, created previously by the same authors (Kokinov & French, 2003). DUAL assumed that all the tasks are to be completed by “coalition(s)” of “microagents” whose interaction ultimately drives the result (Kokinov & French, 2003). There are five components within AMBR’s architecture: retrieval, mapping, transfer,

evaluation, and learning (Kokiov, 1994). The retrieval portion, unlike most algorithms, computes a relevance score for the piece of information that it locates and identifies the “focus,” which is the piece of knowledge with the highest relevance score (Kokiov, 1994). The mapping process acts as a typical algorithm comparing the different “foci” and the original input (called the “goal”) (Kokiov, 1994). In the transfer stage, elements called “inferences” are added to the target description based on what was initially given (Kokiov, 1994). In the last two stages, the new inferences are evaluated for accuracy and then, the algorithm “learns” in hopes of improving its performance on future problems, respectively (Kokiov, 1994). These processes run in parallel to one another, yet another feat unseen in other algorithms created at this time (Kokiov, 1994).

Initially, one may suspect AMBR is closely related to ACME; however, many key elements separate the two algorithms such as (i) more realistic working-memory requirements, (ii) parallel and interacting mapping and memory processes, (iii) dynamically-constructed hypotheses, (iv) semantic similarity is dynamic and context-dependent, and (v) ability to handle n number of arguments (similar to LISA) (Kokinov & Petrov, 2000). AMBR2 allows more flexibility regarding “episodes,” which act as the program’s memories (Kokinov & Petrov, 2000). By enacting this form of “recollection” AMBR2 can create more connections regarding the original input (Kokinov & French, 2003).

2.2.1.6 *Structural Tensor Analogical Reasoning (STAR)*

Seeing the greater potential in distributed representations (rather than local), such as in the ACME algorithm, STAR is a tensor-product-based parallel distributed processing (PDP) model embedded in a neural network (Halford, et al., Connectionist implications for

processing capacity limitations in analogies, 1994). The authors identified four levels of analogical mappings:

1. element mappings – singular relationships are mapped, i.e. mapping that occurs between elements of a metaphor, (e.g. “Dan’s house is a pig’s pen”),
2. relational mappings – binary relationships are mapped, i.e. mapping that occurs in $A:B::C:D$ form (e.g. “woman:baby::mare:foal”),
3. system mappings – tertiary relationships are mapped, i.e. individual elements may not need to match as long as their relations in general do, and
4. multiple system mappings – quaternary relationships are mapped, which are largely theoretical due to their complexity.

In its first iteration, STAR primarily used element and relations mappings. PDP is a computation algorithm that attempts to model capacity and the memory technique of “chunking” information (Halford, et al., 1994). Using the analogy *woman:baby::mare:?* (symbolically mapped as $A:B::C:D$), STAR assumes the predicate (connecting *woman* and *baby*) is MOTHER-OF as shown in Figure 2-9 (Halford, et al., 1994).

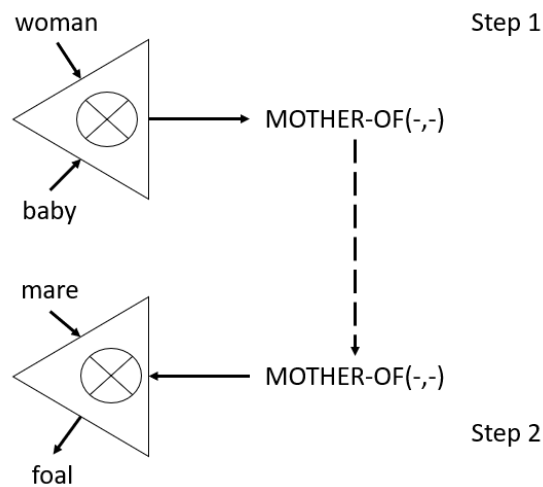


Figure 2-9. Steps for STAR's Simple Analogical Reasoning, adapted from (Halford, et al., 1994)

However, STAR is also able to consider other predicates such as LOVES, FEEDS, or LARGER-THAN, which are all characteristics of the mother relationship, so a “bundle” is created, BUNDLE-MOTHER-OF (Halford, et al., 1994). STAR identifies pairs of arguments (and not necessarily including the original arguments) that satisfy the individual predicates found in the bundle which is summed to create the tensor product representation, T (Wilson, Halford, Gray, & Philips, 2001). The dot product of the tensor product, T , and each proposed pair of arguments is taken to identify the “relation-symbol bundle,” denoted B_p , which is the sum of every relationship between the pair of arguments (Wilson, Halford, Gray, & Philips, 2001). The relation-symbol bundle, B_p , is mapped to the “C” part of the analogy (in this case, “mare”) and then, combined with T through the dot product, creates a system of “weights” for the potential matches for “D” in the analogy (Wilson, Halford, Gray, & Philips, 2001). The highest weight is selected as the “best” D since more propositions are true in this case (Wilson, Halford, Gray, & Philips, 2001). In the case of the example, the best word to complete the analogy is “foal.”

In the earlier algorithms, e.g. SME, ACME, and Copycat (the version from 1993), the base (e.g., mother and baby) and target (e.g., mare and foal) would be created and then, have their similarities mapped to each of the entities in the base and target (Halford, et al., 1994). STAR, rather, maps the base and target on the same elements (in this case the predicates/relationships) within a neural network (Halford, et al., 1994). For example, in evaluating the object-color analogy, *chair:brown::table:brown*, previous algorithms would create two instances of “brown” since it appears in the base (*chair:brown*) and the target (*table:brown*) for a total of 4 instances; however, STAR would only map three instances, since it recognizes *brown* as being the same for the base and the target. Though, the main

new element of STAR is its consideration of human mental capacity and its ability to process multiple items at the same time (Halford, et al., 1994). However, STAR is limited in its ability to understand “hierarchically structured knowledge representations,” which STAR-2 was able to successfully expand upon (Wilson, Halford, Gray, & Philips, 2001).

2.2.1.7 Structural Tensor Analogical Reasoning 2 (STAR-2)

As mentioned previously in the STAR section, STAR-2 builds upon the original algorithm to allow hierarchically-structured analogies (Wilson, Halford, Gray, & Philips, 2001). STAR-2 still factors in the mental capacity constraints by only allowing one pair of up to four propositions at a given time (Wilson, Halford, Gray, & Philips, 2001). The infrastructure of STAR-2 consists of the Focus Selection Network, the Argument Mapping Network, and the information storage structures (Wilson, Halford, Gray, & Philips, 2001). Since the heat/water-flow analogy (Gentner, 1983; Falkenhainer & Forbus, 1989) was unable to be solved by STAR, it is used as an example to explain how STAR-2 works (Wilson, Halford, Gray, & Philips, 2001). For context, the heat/water-flow analogy, visually shown in Figure 2-10, draws upon the statement “heat is like water” and originated from Shawn Buckley’s *Sun Up to Sun Down* book about solar energy (Falkenhainer & Forbus, 1989).

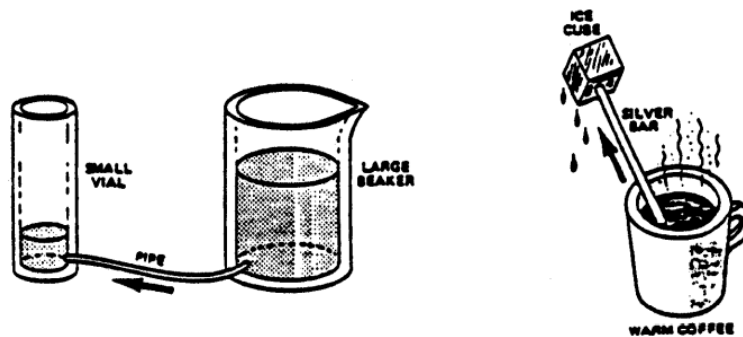


Figure 2-10. Visualization of Water/heat-flow Analogy, from (Falkenhainer & Forbus, 1989)

Looking deeper at Figure 2-10, water will flow via the pipe from the (large) beaker to the (small) vial if the pressure in the beaker is greater than the pressure in the vial. In a similar situation, if the temperature of the (warm) coffee is greater than the temperature of the ice cube, heat will flow via the bar from the coffee to the ice cube. Using the explanation of Figure 2-10 and Structure Mapping Theory to translate it into a textual representation resulted in Figure 2-11 (Falkenhainer & Forbus, 1989).

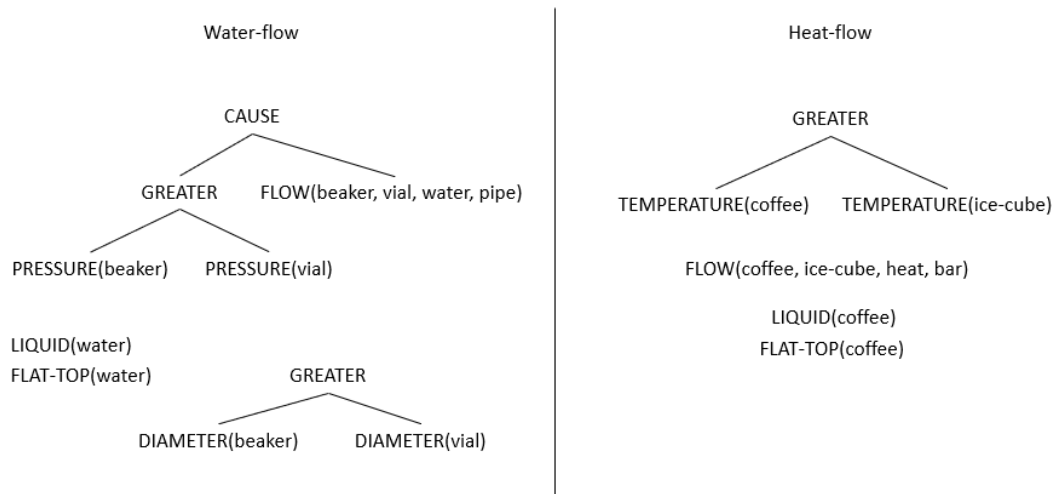


Figure 2-11. Heat/water-flow Textual Representation, adapted from (Eliasmith & Thagard, 2001)

Understanding the context of the analogy, STAR-2 follows the steps outlined below:

1. elements of the analogy (such as similarity, item-types, etc.) are given to the information storage structures via user input,
2. the focus selection network identifies the “highest” commonality between the base and target, the water-flow and heat-flow, respectively (in this case, it is the “cause” “relation symbol”),
3. the argument mapping network finds mappings between the propositions of the analogies (in the example, the propositions are GREATER_PRESSURE(water-

flow) and GREATER_TEMPERATURE(heat-flow) and the resulting action, FLOW, for both the water-flow and heat-flow),

4. the focus selection network moves on to the next “highest” mapping that has not previously been a “focus,” which is now the GREATER_PRESSURE and GREATER_TEMPERATURE propositions,
5. the argument mapping network creates new mappings for the analogy given the “GREATER” proposition identified in Step #4 and records it in the map storing network,
6. these steps are repeated until the program hits a stopping value (Wilson, Halford, Gray, & Philips, 2001).

In a broad sense, STAR-2 combines “thinking” in series and parallel by selecting corresponding elements of the analogy in an ordinal sense but then mapping these elements in parallel in an attempt to expand the algorithm into hierarchical analogies (Wilson, Halford, Gray, & Philips, 2001).

2.2.1.8 Learning and Inference with Schemas and Analogies (LISA)

LISA was the first AR algorithm that attempts to incorporate both semantics and structure in analogical reasoning (Hummel & Holyoak, 1997). LISA differentiates itself from previous algorithms due to its basis in multi-constraint theory (not related to ACME) and the addition of cognitive constraints to better mimic the human mind (Hummel & Holyoak, 1997). LISA breaks down a proposition (ovals) into objects (circles) and predicates (triangles), which are connected to semantic units (smaller circles) as shown in Figure 2-12 (Hummel & Holyoak, 2005).

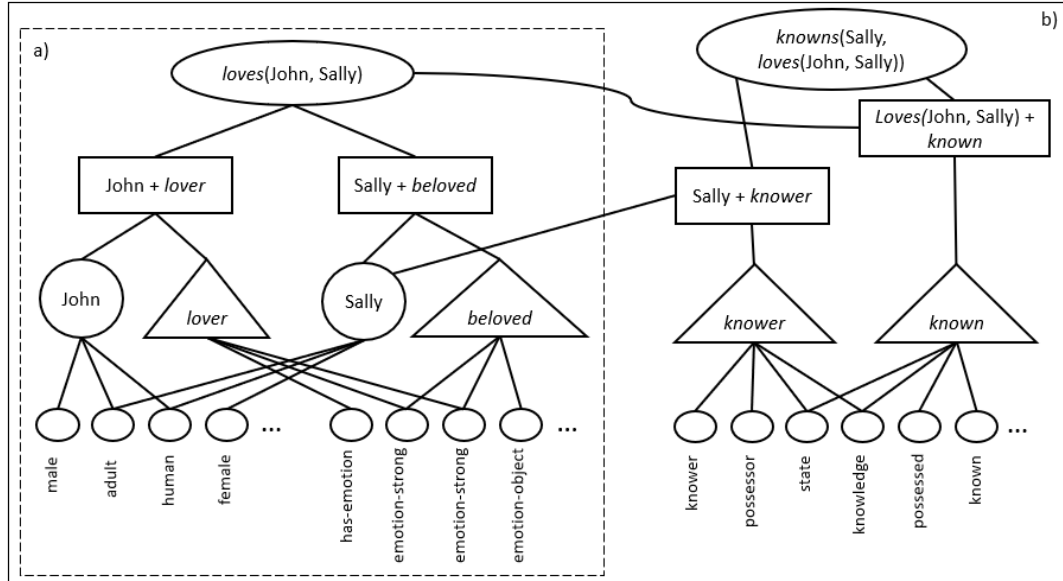


Figure 2-12. LISA's Representation of Propositions, adapted from (Hummel & Holyoak, 2005)

Similar to Copycat, LISA can retrieve information from long-term memory and perform the mapping process within one program (Hummel & Holyoak, 2005). Focusing on the mapping problem, LISA breaks an analogy into the driver and the recipient, which is connected by similar semantic units (Hummel & Holyoak, 2005). One important element when breaking down statements is what LISA calls “role fillers,” which holds an object (the role) constant given the action (the predicates) (Hummel & Holyoak, 1997). In Figure 2-12, an example of a role filler would be “John+lover,” “Sally+beloved,” and their combination *loves(John,Sally)+known* (Hummel & Holyoak, 2005).

In the example shown in Figure 2-13, LISA uses the statement, “John loves Mary”, as the driver to map to the recipient statements, “Bill likes Susan” and “Peter fears Beth” (Hummel & Holyoak, 2005). The color in Figure 2-13 represents how “active” the semantic units are based on how connected the base/target (ovals), broken down into objects (circles) and agents (triangles) in the driver and recipient phrases are (Hummel & Holyoak, 2005).

The more ways in which two objects may be connected will cause more activity in the semantic unit (row of small circles in the middle of Figure 2-13) (Hummel & Holyoak, 2005).

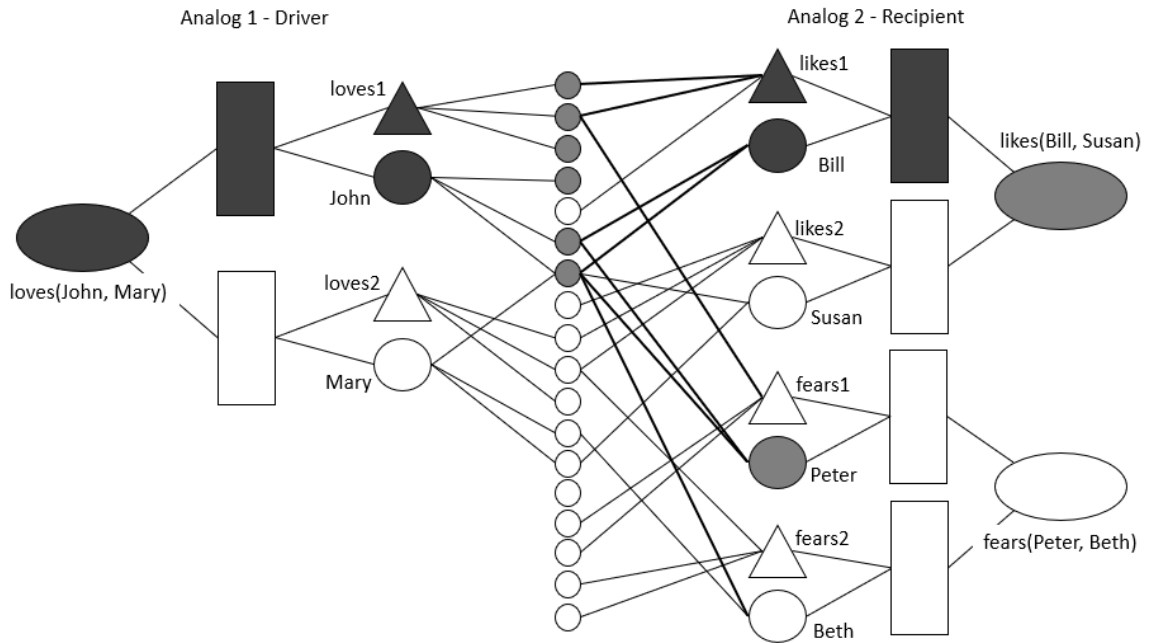


Figure 2-13. Visualization of LISA's Mapping Process, adapted from (Hummel & Holyoak, 1997)

LISA improves upon both SME and ACME due to its ability to interpret concepts (though ACME's semantic network is similar but inferior) and its infrastructure based on the human mind (Hummel & Holyoak, 1997). Though there are several parallels between LISA and STAR (such as both being neural networks), the latter algorithm's mental capacity is based on limiting the number of firings done synchronously (Hummel & Holyoak, 2005); whereas, STAR reduces the number of "chunks" that can be evaluated simultaneously (Halford, et al., 1994). LISA's limitations (the ability to only fire three propositions at once (Hummel & Holyoak, 2005)) could potentially represent the constraints on a human's short-term memory compared to other algorithms (Eliasmith &

Thagard, 2001). LISA is partly biased due to relational concepts being hard-coded into the program and its inability to learn new predicates (Lu, Chen, & Holyoak, 2012). However, DORA attempts to expand the functionality of LISA by allowing new predicates to be learned (Doumas, Morrison, & Richland, 2009).

2.2.1.9 *Distributed Representation Analogy Mapper (DRAMA)*

DRAMA looks at analogies from a “soft”-constraint point-of-view in terms of structure, similarity, and purpose (Eliasmith & Thagard, Integrating structure and meaning: A distributed model of analogical mapping, 2001). DRAMA is a distributed algorithm because it considers both structure and the underlying meaning (semantics) of an analogy via holographic reduced representations (HRRs) (Eliasmith & Thagard, Integrating structure and meaning: A distributed model of analogical mapping, 2001). HRRs allow for easy interpretation of structure due to their application of circular convolution and superposition of “predicate-like objects” (Plate, 1994). The basis of this theory is described in the following example created by Eliasmith and Thagard using A, B, and C to represent HRRs:

If $C = A \otimes B$ (C equals A convolved with B), then

$C \# A \approx B$ (C correlated with A approximately equals B) and

$C \# B \approx A$ (C correlated with B approximately equals A) (2001).

The superposition of these statements is shown on the three-dimensional grid in Figure 2-14. Vector D represents the superposition of A and B.

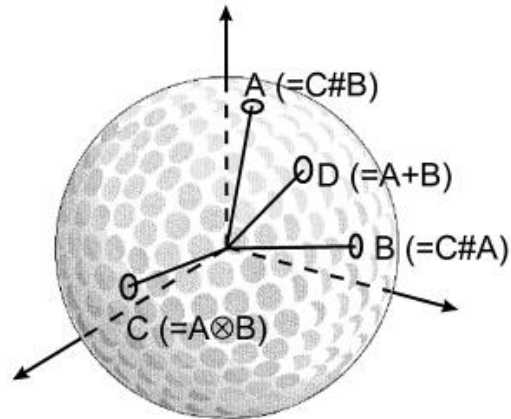


Figure 2-14. HRR superimposed in a 3D space, from (Eliasmith & Thagard, 2001)

DRAMA uses ACME as the basis for the mapping process considering the latter's success with a multi-constraint theory but incorporates HRRs to assist with weight attributed to an analogy's semantics-portion (Eliasmith & Thagard, 2001). Using the structure described in the example above, DRAMA identifies the relation (action, verb, etc.), object (the cause or primary focus of the statement), and agent (what the relation is applied to, or the effect) and convolves each with a corresponding word or phrase from the analogy (Eliasmith & Thagard, Integrating structure and meaning: A distributed model of analogical mapping, 2001). DRAMA improves on the methods in ACME by (i) producing smaller mapping networks, (ii) stochastic representations of the analogy, the incorporation of semantic consideration, and (iii) more in-depth representations (Eliasmith & Thagard, Integrating structure and meaning: A distributed model of analogical mapping, 2001). However, DRAMA is limited due to the (i) potential loss of information considering the HRRs has to be "cleaned-up" to be recognized by the system, (ii) inability to comprehend asymmetrical analogies (Part A [base] of an analogy map well to Part B [target], but not well if the roles were reverse), (iii) difficulty incorporating a "learning" process based on previously identified analogies, and (iv) though minor, the lack of an application and

retrieval process. In comparison to the only other claimed distributed algorithm, LISA, DRAMA proves to be the more advanced algorithm considering its (i) ability to compare and contrast multiple propositions (and therefore, analogies), (ii) able to handle multi-structural levels within its working memory at once, and (iii) consistently with the encoding of structural and semantic information (which also reduces bias from the programmer's code) (Eliasmith & Thagard, 2001).

2.2.1.10 Discovery Of Relations by Analogy (DORA)

DORA is considered to be an extension of Hummel and Holyoak's LISA algorithm for AR that incorporates learning (Domas, Morrison, & Richland, 2009). The interworking parts of DORA are almost identical to LISA except for its ability to allow role-filler firing to occur asynchronously (Domas, Hummel, & Sandofer, 2008). This is important due to two unique features of DORA: (i) the ability to learn a new predicate from an object and (ii) combine role-filler pairs into one relation which better copies the human mind (Domas, Hummel, & Sandofer, 2008). DORA allows learning from unlabeled examples and does this via a logical intersection of several examples (Lu, Chen, & Holyoak, Bayesian analogy with relationship transformations, 2012). One negative to DORA's learning process is the resulting decreased working memory limitations (Domas, Hummel, & Sandofer, 2008).

2.2.1.11 Connectionist Analogy Builder (CAB)

Yet another connectionist algorithm is the Connectionist Analogy Builder (CAB) which primarily works through comparison with a focus on structure and introduces a systematic constraint (Larkey & Love, 2003; Genter & Forbus, 2010). CAB emphasizes how correspondences between elements are made within the human brain for tasks such as

AR (Larkey & Love, 2003). When looking at the phrase “Jim loves Betty,” a predicate calculus format is applied and yields Figure 2-15. Nodes are objects and entities considered to be the building blocks of an analogy (Larkey & Love, 2003). Node weights, denoted by arrows in Figure 2-15 vary via a “learning rule,” which identifies elements that would successfully be mapped with one another in a given scenario (Larkey & Love, 2003). The predicate nodes (shown in the “top” row in Figure 2-15) are identified, linked to the argument nodes (shown in the “bottom” row in Figure 2-15) and then, CAB creates an entity and value associated with each predicate. CAB also uses arrows to represent “links” within the framework (Larkey & Love, 2003). The authors specifically speak to CAB’s ability to identify “alignable” (analogs have suppllicated “dimensions”, such as with gender in Figure 2-15) and “non-alignable” differences (one analog has a certain predicate, which another analog does not) (Larkey & Love, 2003). The representation also allows “unambiguous representations” of the predicates and arguments involved (Larkey & Love, 2003). Each link has a direction associated with it with is important when “traveling” between nodes (Larkey & Love, 2003).

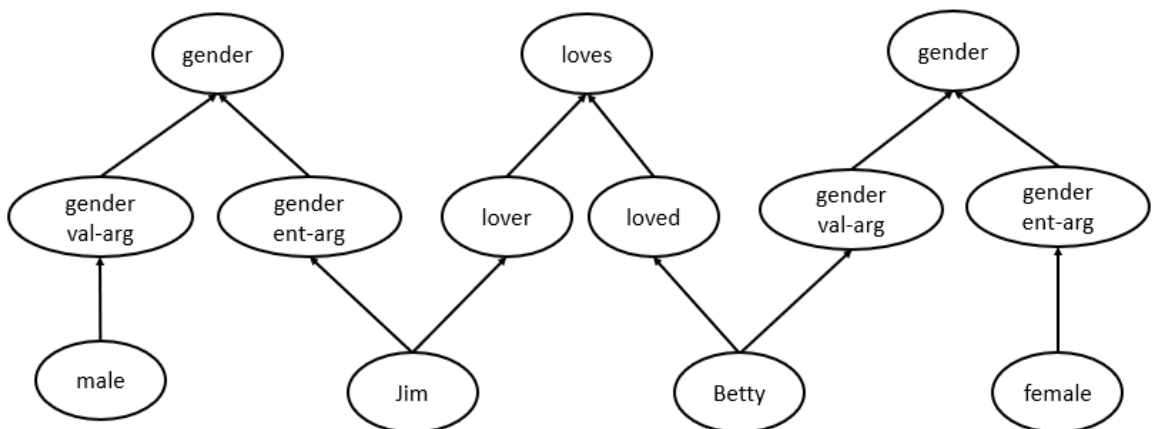


Figure 2-15. CAB's Representation of "Jim loves Betty", adapted from (Larkey & Love, 2003)

CAB uses one-to-one mappings between the analogs, *A* and *B*, such as Jim and Betty in Figure 2-15 (Larkey & Love, 2003). Based on the number of commonalities between nodes, CAB will calculate the number of mapping weights associated with the node (Larkey & Love, 2003). These weights will increase for parallel structures between two analogs (Larkey & Love, 2003). The relationships with shorter links have a larger weight than those that are more distant, and the tendency for CAB to create these more distance relationships is affected by a parameter to mimic working memory in humans (Larkey & Love, 2003). An evidence constraint attempts to control how much these weights grow through the one-to-one mapping process until the weights eventually equate to 0 or 1 (Larkey & Love, 2003).

When compared to SME, CAB creates temporary analog mappings every iteration, incorporates a capacity constraint, and is more helpful if the goal is estimating response time (Larkey & Love, 2003). Looking at ACME, CAB can only perform one-to-one mappings, requires nodes to be identical when looking at compatibilities, and incorporates a parameter that mimics working memory capacity (Larkey & Love, 2003). CAB is significantly more computationally conservative than LISA by only being built on 4 equations and parameters compared to LISA's 21 equations and 22 parameters, respectively (Larkey & Love, 2003).

2.2.1.12 Latent Relational Analysis (LRA)

Taking a step back to Rumelhart's model, where the analogy questions lacked greater content, but the model was asked to identify the best of four choices that complete the general analogy of *A:B::C:?*, LRA attempts to quantify the similarity between the word pairs *A:B* and *C:D* (Turney, Similarity of semantic relations, 2006). In the example shown

in Figure 2-16, quart (A) and volume (B) is given and the goal of LRA is to identify the best C and D pair among the five options (C* and D* will be used to refer to the actual best option, mile, and distance, respectively) (Turney, Similarity of semantic relations, 2006).

Stem: quart:volume
Choices: (a) day:night
 (b) mile:distance
 (c) decade:century
 (d) friction:heat
 (e) part:whole
Solution: (b) mile:distance

Figure 2-16. SAT Question Used as Example in LRA Walkthrough, adapted from (Turney, 2006)

A large part of LRA is the incorporation of a Vector Space Model (VSM) with the addition of Singular Value Decomposition (SVD) to help smooth the created vectors (Turney, Similarity of semantic relations, 2006). A VSM is a special case of a connectionist algorithm that can only consider $A:B::C:D$ word pair analogies.

Collecting from various sources, the LRA finds the cosine between two vectors consisting of the “features” of the word pairs, R_1 (with words A and B) and R_2 (with words C and D) (Turney, Similarity of semantic relations, 2006). In this scenario, “features” are relationships between phrases that use the two original words in R_1 (A and B) and then, R_2 (C and D for each option) (Turney, 2006). Given two sets of word pairs, LRA takes the following high-level steps:

1. replace both words in each pair with their respective synonyms to create alternative pairs,
2. remove all alternative pairs from consideration if they are not “near analogies,” (meaning A and C have high attribute similarity as do B and D),

3. given the initial word pairs and alternative pairs, search for phrases that begin with the first word and end with the second word,
4. create patterns by replacing words identified in step three with “wild cards” which allows a certain number of words in the phrase (depending on the original length) to be replaced by another unidentified word,
5. build the “pair-pattern frequency matrix,” shown in Figure 2-17, whose rows are the corresponding combinations of the original pair (“A P B, then, “B P A”) and the column is the “patterns” (P) identified in Step #4 (matrix values are the frequency of the number of phrases that have the format identified in the row with the corresponding pattern in the column; this is typically reduced to a sparse matrix; note that an asterisk (*) in the “patterns” identified in the first row of Figure 2-17 represents a “wild card” that can be replaced by any word),

	P = "in"	P = "* of"	P = "of *"	P = "* *"
freq("quart P volume")	4	1	5	19
freq("volume P quart")	10	0	2	16

Figure 2-17. Pair-pattern Frequency Matrix Example, adapted from (Turney, 2006)

6. transform the pair-pattern frequency matrix via log and entropy calculations and pass it onto the SVD, which simplifies LRA’s matrix calculations,
7. looking at the rows with the original word pairs (A and B and C and D) in Figure 2-18 the cosine of the row vectors is calculated, which is repeated for each combination of the original pairs and alternative pairs and the alternative pairs with one another found within the matrix; Figure 2-18 shows a visual of this process given the original example in Figure 2-16,

Word Pairs	Cosine	Cosine >= Original Pairs
quart:volume::mile:distance	0.525	Yes (Original Pair)
quart:volume::feet:distance	0.464	
quart:volume::mile:length	0.634	Yes
quart:volume::length:distance	0.499	
liter:volume::mile:distance	0.736	Yes
liter:volume::mile:length	0.687	Yes
liter:volume::mile:length	0.745	Yes
liter:volume::length:distance	0.576	Yes
gallon:volume::mile:distance	0.763	Yes (Highest Cosine)
gallon:volume::feet:distance	0.71	Yes
gallon:volume::mile:length	0.781	
gallon:volume::length:distance	0.615	
pumping:volume::mile:distance	0.412	
pumping:volume::feet:distance	0.439	
pumping:volume::mile:length	0.446	
pumping:volume::length:distance	0.491	

Figure 2-18. Original and Alternative Pairs Cosines, adapted from (Turney, 2006)

8. average all the cosines calculated in Step #7 and compare to the average cosine for all the other potential *C* and *D* pairs. The highest average is the “best” corresponding word pair (*C** and *D**) for the original inputs (*A* and *B*) (Turney, 2006).

Though LRA appears to perform at par with a human on similar SAT analogy questions, there is potential that the error is too high for a computerized algorithm (Turney, *Similarity of semantic relations*, 2006). One criticism of LRA is its reliance on inputs with “relational vocabulary” rather than concepts (Lu, Chen, & Holyoak, 2012).

2.2.1.13 Bayesian Analogy with Relational Transformations (BART)

Similar to LISA and DORA, BART considers the semantics of an analogy while also incorporating a learning process (Lu, Chen, & Holyoak, 2012). BART attempts to learn from object concepts which removes the “relational vocabulary” needed by LRA (Lu,

Chen, & Holyoak, 2012). After receiving the inputs, BART beings its two processes, First-order relation learning and importance-guided relation mapping (Lu, Chen, & Holyoak, 2012).

In first-order relation learning, BART creates weights for pairs of words (from their features) and uses this to determine whether they represent a relation (Lu, Chen, & Holyoak, 2012). In importance-guided relation mapping (that allows for higher-order relations), BART attempts to map $A:B$ to $C:D$ in a way that minimizes the distance while factoring in the weights created in first-order relation learning (Lu, Chen, & Holyoak, 2012). BART specifically attempts to answer how relational representations are theorized, particularly to children (Lu, Chen, & Holyoak, 2012). The model is special because it attempts to learn from its inputs (Lu, Chen, & Holyoak, 2012) while not requiring them to have a relational element (Lu, Wu, & Holyoak, 2019). Instead of mapping between individual predicates within an analogy, BART uses the predicate's features to derive the relations which the authors consider to be "subsymbolic" (Lu, Chen, & Holyoak, 2012).

When compared to neural network models BART's weight system is more advanced due to being at the "features" level (Lu, Chen, & Holyoak, 2012). BART is most similar to DORA due to being based on a bottom-up approach; however, they differ in a few key ways (Lu, Chen, & Holyoak, 2012). BART requires labeled examples and its regression algorithm allows BART's expansion into Leuven and topics vector inputs, the latter of which gives it an advantage over DORA (Lu, Chen, & Holyoak, 2012). Compared to the infrastructure of LISA (separate pool of features for each predicate) and DORA (one pool of features for all elements involved), BART uses a vector to present features with the relations displayed through the weight distributions DORA allows learning from unlabeled

examples and does this via a logical intersection of several examples (Lu, Chen, & Holyoak, 2012). BART was extended to general inferences in a model called BART-g in 2017 (Chen, Lu, & Holyoak, 2017). After being used in conjunction with ResNet50-A to classify images, BART was applied to images as well (Lu, Liu, Ichien, Yuille, & Holyoak, 2019).

2.2.1.14 LRCos

Several recent developments in analogical reasoning have focused on a VSM approach as used in Turney's LRA model (Rogers, Drozd, & Li, 2017). It is important to note that the VSMs are primarily concerned with measuring word similarity, which some consider synonymous with AR (Rogers, Drozd, & Li, 2017). Despite VSMs having the word "model" in its name, it is still considered to be an algorithm by the standards herein and will be called likewise throughout the remainder of the document. Some of the recent algorithms include Word2Vec (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013; Mikolov, Yih, & Zweig, 2013), GloVe (Pennington, Socher, & Manning, 2014), 3CosAvg (Drozd, Gladkova, & Matsuoka, 2016) and LRCos (Drozd, Gladkova, & Matsuoka, 2016). Unlike word2vec and GloVe, 3CosAvg and LRCos are not complete VSMs, but rather a specific equation for calculating the cosine distance between two word vectors (Drozd, Gladkova, & Matsuoka, 2016). Due to being the most promising and the most recent VSM model (Rogers, Drozd, & Li, 2017), LRCos has been selected for an in-depth look here though the rest are discussed further in Chapter 3.1.

As with other models, VSMs look at analogies in the form $a:b::c:d$ where d is unknown and needs to be discovered given the context between a , b , and c (Drozd, Gladkova, & Matsuoka, 2016). LRCos' predecessor, 3CosAdd, represented d as the vector,

$$\operatorname{argmax}_{\text{dev}}(\operatorname{sim}(d, c-a+b)), \quad (2-1)$$

as defined in (Mikolov, Yih, & Zweig, 2013). *Sim* is a similarity measure incorporating the cosine of the parameters, u and v , two vectors such that it yielded the following equation,

$$\operatorname{sim}(u,v) = \cos(u,v) = \frac{u \cdot v}{\|u\| \|v\|}. \quad (2-2)$$

LRCos incorporates a measure of similarity such as previous VSMs and calculates the chances of a potential answer belonging to the target class, which it uses to determine the “correct” solution (Drozd, Gladkova, & Matsuoka, 2016). One way in which LRCos distinguishes itself from 3CosAdd is being able to perform equally as well when the a , b , and c words are excluded from the search space as when they are included (Rogers, Drozd, & Li, 2017).

2.2.1.15 Summary of Analogical Reasoning Algorithms

Ever since the late 1980s, research into text-based AR has been growing due to a need to understand how to leverage our understanding of biological learning mechanisms. Many AR algorithms were created by building off a previous biological/psychology foundation or learning or were created in response to limitations in another AR algorithm. Most AR algorithms were continuous work-in-progresses that develop over time and lack a specific year to pinpoint their final iteration. However, based on the literature and as described in (Combs, Bihl, Ganapathy, & Staples, 2022), the best approximated years were used and visually represented in Figure 2-19. It is worth noting that Rumelhart’s research was a study of human AR processes, i.e., a model, so it was excluded from Figure 2-19.

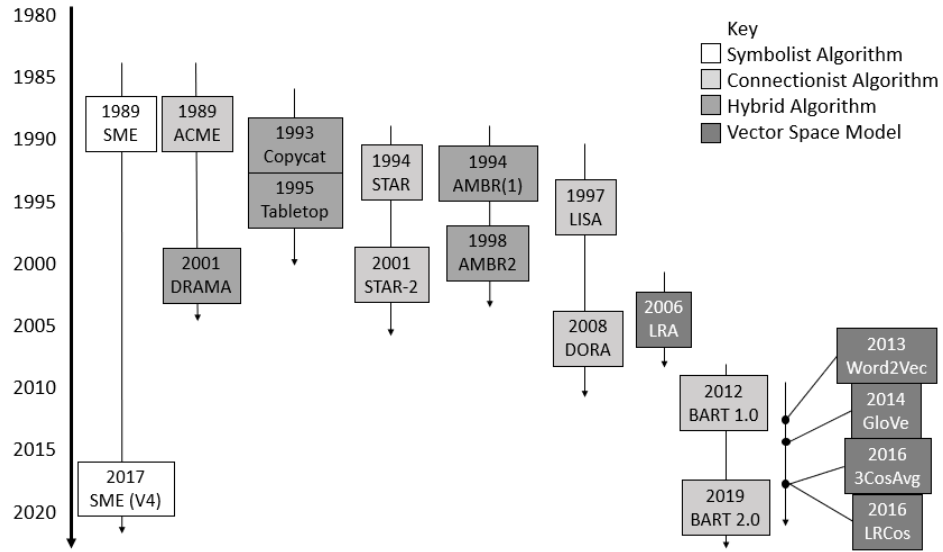


Figure 2-19. Timeline and Relationships Between Analogical Reasoning Algorithms, from (Combs, Bihl, Ganapathy, & Staples, 2022)

A high-level overview of the previously-described algorithms is shown in Table 2-1. It considers which process and considerations that the algorithms are primarily concerned with. As mentioned in the introduction, analogy making is concerned with four processes: retrieval, mapping, evaluation, and learning (Holyoak & Thagard, 1989; Thagard, Holyoak, Nelson, & Gochfield, 1990; Gentner & Smith, 2012). All algorithms needed an evaluation portion, but the remaining processes were the focus of different algorithms. Table 2-1 compares the AR algorithms based on the remaining three AR processes (retrieval, mapping, and learning), algorithm type (following the three AI schools of thought, see Section 2.2.1), and a brief description of how they work or how they are unique from the others

Table 2-1. Comparison of Analogical Reasoning Algorithms

Algorithm	Process(es)			Algorithm Type	Methods
	Retrieval	Mapping	Learning		
Rumelhart		X		N/A (Study/Model)	3D Euclidean mappings based on human subjects; Not a computer program
SME (Part of FAC)		X		Symbolist	Creation of pairwise matches whose quality is measured on the Structural Evaluation Score (SES)
MAC	X				
ACME		X		Connectionist	Emphasis on 3 “soft” constraints: isomorphism, semantic similarity, and pragmatic centrality
ARCS	X				
Copycat/ Tabletop/ Metacat/ FAE	X	X		Hybrid	Abstract modeling method with short and long-term memory components
AMBR(1)/ AMBR2	X	X	X	Hybrid	Based on DUAL cognitive architecture for information representation with parallel processes
STAR		X		Connectionist	ANN using Tensor Product with Parallel Distributed Processing (PDP)
STAR-2		X		Connectionist	ANN using Tensor Product with PDP; expanded for serial selection of propositions
LISA	X	X		Connectionist/ Hybrid	ANN with Dynamic Binding and Role-filler Synchrony
DORA	X	X	X	Hybrid	ANN with Dynamic Binding and Role-filler Asynchrony
DRAMA		X		Hybrid	Analogies evaluated as Holographic Reduced Representation (HRRs)
CAB		X		Connectionist	Structuring analogies based on element correspondences
LSA		X		Connectionist	Vector Space Model (VSM) with Singular Value Decomposition (SVD)
BART		X	X	Connectionist	Bayesian Inference through bootstrapping
LRCos		X		Connectionist	VSM that combines a similarity measure and prediction of a word’s class

As with any field, many branches and algorithms exist beyond that were reviewed. However, the discussion above was constructed as a high-level list to give a broad background on the history of text-based AR algorithms. This discussion also focused on AR methods that were considered to be the foundation of AR at its start or more recent models introduced into the literature that had yet to be compared in a review. Due to the abundance of AR algorithms and the focus on modern ones, many earlier algorithms prominent before the rise of connectionist ones could not be evaluated in-depth such as CARL (Burstein, 1983), Environmental Model of Analogy (EMMA) (Ramscar & Pain, 1996), Heuristic-Driven Theory Projection (HDTP) (Gust, Kuhnberger, & Schmid, 2006), Incremental Analogy Machine (IAM) (Keane & Brayshaw, 1988), Similarity, Interactive Activation, and Mapping (SIAM) (Golstone, 1994), and Aligning Between Systems using Relations Derived Inside Systems for Translations (ABSURDIST) (Goldstone & Rogosky, 2002) are some more notable symbolist algorithms. However, some of these are compared to one another in earlier studies, e.g., (Gentner & Forbus, 2010).

2.2.2 Image-based Analogy Methods

Work in analogies has also involved image-based concepts. This begins with the geometric work of Polya and further builds on the advances of text-based analogical work (1990). As with text-based analogies, image-based analogies can also be symbolically represented as $A:B::C:D$, where A and B is the original and transformed image, respectively, C is the target image, and D being the newly transformed image of C , using the same technique that transformed A into B . Additionally, given the underlying understanding that children learn through analogies and do so before being able to understand text, such an extension is natural; however, in AR implementations, image-

based analogies have proven to be a challenge due to a lack of obvious semantic structure in images. Despite these issues, several algorithms attempt to tackle this problem.

2.2.2.1 ANALOGY

At its earliest theorization, image-based analogies were derived in the form of logic puzzles, such as the one in Figure 2-20. ANALOGY was created before the earliest text-based model by Rumelhart and is often classified as a symbolist method despite dealing with images (Kokinov & French, 2003). Considering the question, “*Figure A* is to *Figure B* as *Figure C* is to which of the given figures,” one is expected to select the image (denoted by a number) that “fits best.”

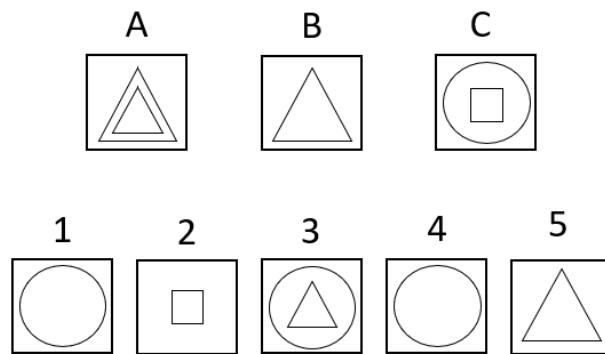


Figure 2-20. Geometric-analogy Problem, from (Evans, 1964)

This task was given to a “geometric-analogy problem”, ANALOGY, which decomposed the images to identify similar “objects” (Evans, 1964). Each “object” is described appropriately as a dot (DOT), simple closed curves (SCC), or other (REG) with coordinates corresponding to the shape’s vertices (or origin in the case of DOT) (Evans, 1964). Next, the program evaluates the properties of the shapes (scale factor, rotation angle, etc.) and relationships between objects (location relative to another object, similarity between the figures, etc.) in the output (Evans, 1964). Given this, the *solve* process begins to map how *Figures A* and *B* are similar as well as how *Figure C* and all the potential

answers are; these mappings are called “rules” (Evans, 1964). The goal of *solve* is to identify the “rule” between *Figure C* and the answer that deviates as little as possible from the *Figure A to B* rule (Evans, 1964). In the event of a tie, a new method is attempted but ANALOGY will always result in a selection (Evans, 1964). This is considered the most prominent analogical reasoning model from the 1960s (French, 2002).

2.2.2.2 Other Image-based Algorithms

Similar to ANALOGY, Proteus decomposes largely geometrical images into individual elements, their characteristics, and relationships one to one another in a semantic network where the subprocesses, Geminus, identifies which images it matches from memory and Galatea, considers their common elements in comparison with the source case (Yaner & Goel, 2006). Analogy-making similar to Copycat (but with a larger focus on the domain rather than inference) was the subject of an image-analogy program that used a Boltzmann Machine (Memisevic & Hinton, 2010). The previously mentioned algorithm, DORA, has been successfully used to break down three-dimensional shapes called “geons” and apply this knowledge to similar image-based analogy problems (Doumas & Hummel, 2010). The Analogy-preserving Semantic Embedding (ASE) method uses an “analogical parallelogram” consisting of learned vectors to assist with image categorization based on the parallelogram’s analogies (Hwang, Grauman, & Sha, 2013). In 2013, Zero-shot learning was performed on the CIFAR-100 dataset, which involved attempting to identify classes without being trained on any images from the “unknown” class (Socher, Ganjoo, & Manning, 2013). Image analogy task phrased such that $A:B:C:?$ was solved via a Siamese ConvNet in the 2015 program, “Visalogy” (Sadeghi, Zitnick, & Farhadi, 2015). Visual analogy-making via deep learning methods was explored as an option to create more

advanced pictorial transformations such as the rotation of 2D multi-colored video game characters and 3D car models (Reed, Zhang, Zhang, & Lee, 2015). Returning to text-based analogy roots with Gentner's SME, in 2017, it was used as the basis for a visual analogy program that solves Raven's Progressive Matrices problems through contrastive, descriptive, holistic, and component patterns (Lovett & Forbus, 2017). As mentioned earlier, BART has also been successfully expanded with an image recognition program to select the correct pictorial solution using semantic information about the options (Lu, Liu, Ichien, Yuille, & Holyoak, 2019). Despite having early roots, there is still a lot of research occurring in this area. These advancements in image analogies are beneficial when converting from one image to another image, but research is still lacking in how to interpret images using analogies.

2.3 IMAGE RECOGNITION AND CONTEXT

Image recognition is a subset of AI and machine learning (ML) that uses algorithms to detect, classify, segment, or otherwise process images. Many image recognition programs have an ANNs structure due to the advances made in image processing through the combined feature extraction and classification from deep convolutional ANNs (Ball, Anderson, & Chan, 2017).

2.3.1 Artificial Neural Networks (ANNs)

Inspired by their biological namesake, neurons, an ANN has three main layers, input, hidden, and output (Bihl, Young, & Weckman, 2018). The hidden layer is similar to a black box, as the ANN assigns unknown weights to the inputs which are used in the activation function that yields the resulting output (Bihl, Young, & Weckman, 2018).

Within the hidden layer, the number of hidden nodes is determined by the human in such a way that balances having an accurate model and overfitting (Bihl, Young, & Weckman, 2018). There may also be multiple hidden layers, called multi-layer perceptrons (Bihl, Young, & Weckman, 2018). How information is passed between nodes and whether the outputs can influence the model further expands the vastness of ANN model types, a small subset can be seen in Figure 2-21 (Bihl, Young, & Weckman, 2018).

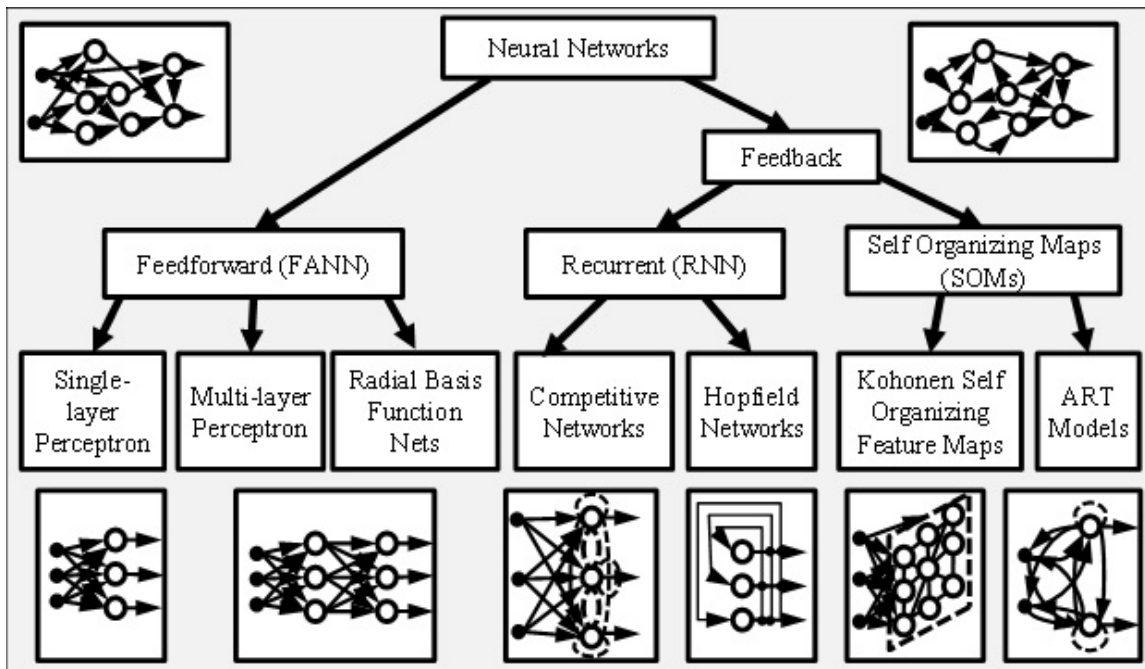


Figure 2-21. ANN Families and Types, from (Bihl, Young, & Weckman, 2018)

It is important to have a broad understanding of previous work done in the field of computer vision (CV) specifically regarding ANNs and deep learning (DL). CV involves a machine analyzing an image or picture (Ball, Anderson, & Chan, 2017). DL extends upon ANNs by allowing larger scaled architectures and, often, automated feature extraction processes which result in highly processed and normalized data features (Ball, Anderson, & Chan, 2017). When DL is applied to CV, many feature extraction layers have been found to provide highly accurate results (Ball, Anderson, & Chan, 2017). While many DL

methods exist, these can largely be grouped into the types as shown in Figure 2-22: autoencoders (AE), convolutional neural networks (CNN), deep belief networks (DBNs), recurrent ANNs (RNN), and deconvolutional ANNs (DeconvNet) (Ball, Anderson, & Chan, 2017; Bihl, Young, & Frimel, 2022). A DBN often has many hidden layers and nodes between its input and output layers, a CNN is typically used for image classification tasks, and an RNN uses a feedback loop to influence previous layers/nodes with the architecture (Bihl, Young, & Frimel, 2022). In general:

1. Autoencoders (AEs) are ANNs used for unsupervised data exploration
2. Deep Belief Networks (DBMs) are probabilistic graph models which leverage graph theory and ANN architecture constructs for data processing
3. Convolutional neural networks (CNN), which incorporate many layers for filtering and feature extraction through convolutions, pooling, and nonlinear functions to highly normalize data for, primarily, computer vision.
4. Recurrent ANNs (RNNs) are temporal approaches with connections forming across sampled cycles (Bihl, Young, & Frimel, 2022).

Beyond these approaches, in the CV and the object recognition field, several additional types of ANNs has proven to be successful: feed-forward, self-organizing feature map (SOM), hopfield, adaptive resonance theory (ART), associative memories (and Random Access Memory (RAM)), neocognition, higher-order network, fuzzy neural/neuro-fuzzy system, and more (Egmont-Petersen, de Ridder, & Handels, 2002).

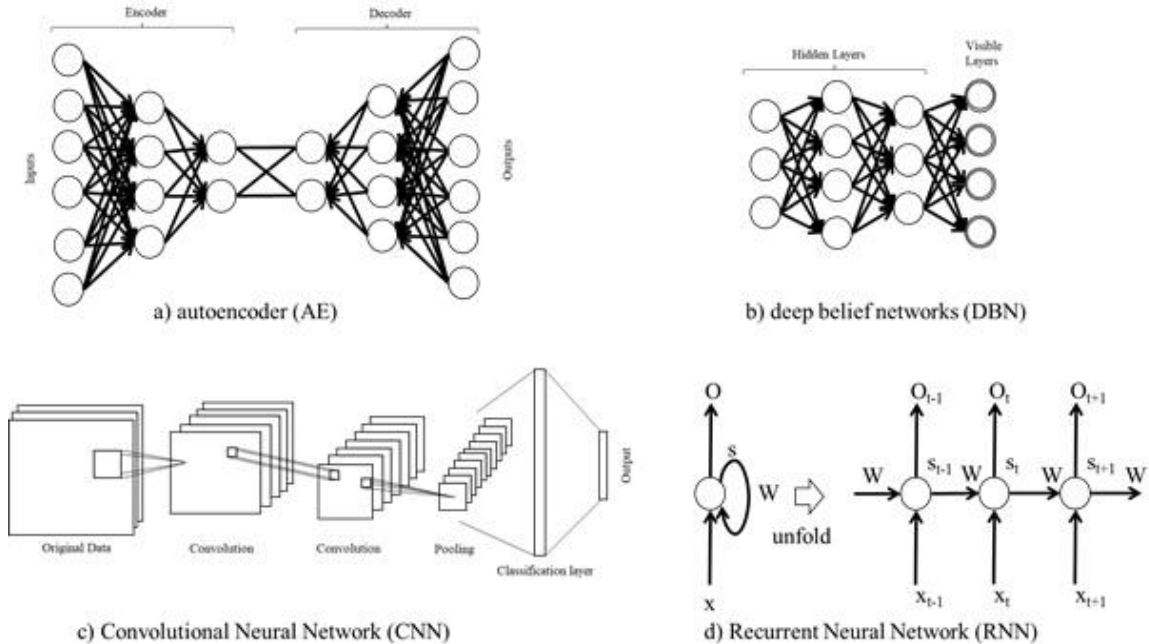


Figure 2-22. Types of artificial neural networks (ANNs), from (Bihl, Young, & Frimel, 2022)

2.3.2 Convolutional Neural Networks (CNNs)

Specific interest herein is on CNNs considering their successful application with image recognition (Wu, 2017). CNNs are feedforward and used almost exclusively to identify patterns among images (O'Shea & Nash, 2015). When used in a CNN, an image has three inputs, associated with its height, width, and channels (associated with the colors used, e.g. grayscale, binary (black/white), true-color (red, green, blue), or multispectral) (Wu, 2017). Within a CNN there are several different layers, but across all CNNs there will be convolutional, pooling, and fully-connected layers (O'Shea & Nash, 2015). The layers specific to CNNs are the convolutional and pooling layers, which are applied before the fully-connected layers as found within all ANNs (O'Shea & Nash, 2015). A simple CNN architecture is shown in Figure 2-22.c.

2.3.2.1 Convolutional Layers

A convolutional layer is most often incorporated with a Rectified Linear Unit (ReLU), which doesn't change the image's size, but rather increases nonlinearity already found within the image (Wu, 2017). This is different from when it's used in the context of Keras' activation functions, the "Relu" option outputs the maximum value between 0 and the input value from the image matrix (Keras, 2020). These layers all involve a pre-determined kernel, which is a small matrix that "slides" throughout the entirety of the original image represented in matrix form visualized in Figure 2-23 (Wu, 2017). In this particular example, the kernel size is 2x2 with each cell having a value of 1 that is applied via matrix multiplication to yield the leftmost matrix in Figure 2-23.

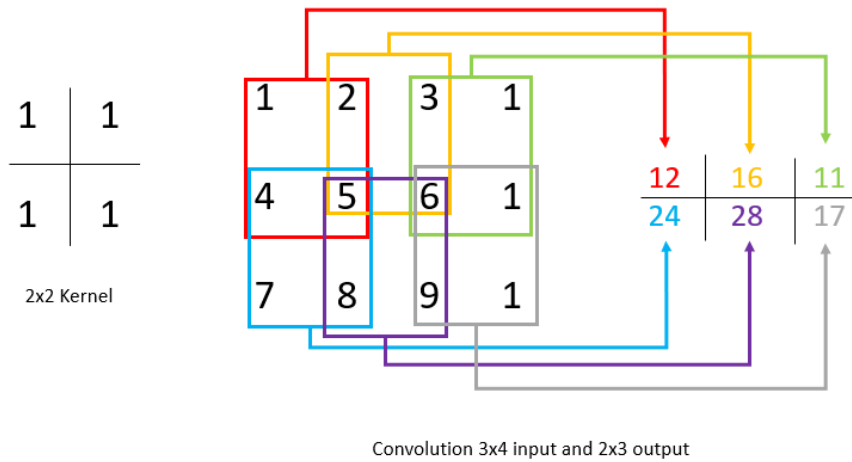


Figure 2-23. Convolutional Operation, adapted from (Wu, 2017)

When applied to an image, the output looks significantly different than what's shown in Figure 2-23. Kernel sizes are typically 3x3 and the ReLU element creates an arbitrary outline of the main elements of an image called "edge detection features" (Wu, 2017). Depending on the convolution operation used such as brighter pixel detection in the

horizontal versus vertical directions, a convolutional layer will yield images such as those in Figure 2-24.b or Figure 2-24.c from Figure 2-24.a (Wu, 2017).

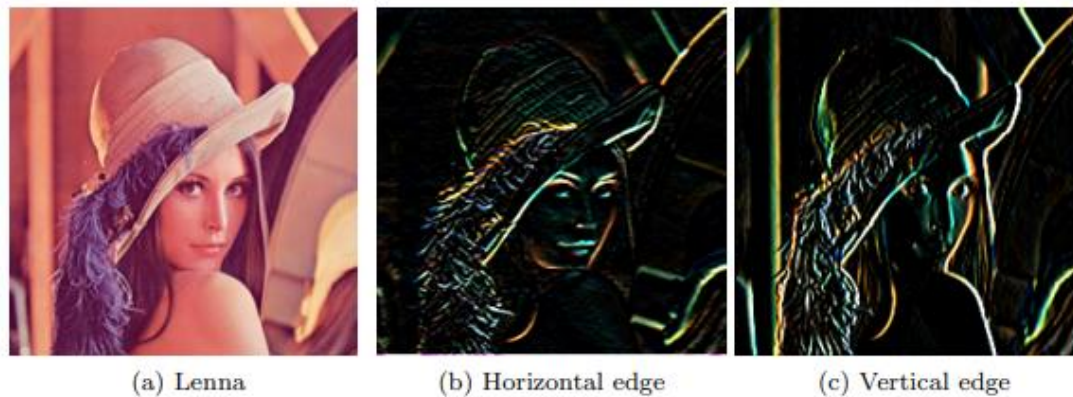


Figure 2-24. Convolutional Layer Applied to the Lenna Image, from (Wu, 2017)

It is also common to specify three parameters within a convolutional layer: depth, stride, and padding (O'Shea & Nash, 2015). Similar across all ANNs, the depth is concerned with how interconnected the neurons within the hidden layer are (O'Shea & Nash, 2015). The stride refers to the “steps” taken when moving the kernel around the input image matrix and controls how much overlapping occurs within the output matrix (O'Shea & Nash, 2015). Padding, also called zero-padding, is concerned with the dimensions of the input and in particular, the “border” of the image (O'Shea & Nash, 2015). CNNs can also make use of parameter sharing, which reduces the number of total parameters during the backpropagation stage (O'Shea & Nash, 2015).

2.3.2.2 Pooling Layers

After a convolutional layer is applied, it is followed by a pooling layer, which aims at reducing model complexity (O'Shea & Nash, 2015). Similar to the kernel associated with the convolutional layer, there is a pooling kernel (typically a 2x2 matrix) that also moves across the matrix that results from the previous convolutional layer (O'Shea & Nash, 2015).

Instead of applying matrix multiplication, this kernel (typically) identifies the maximum value among the cells it is considering and has a stride equivalent to its dimensions (O'Shea & Nash, 2015). Figure 2-25 illustrates how max-pooling works on a 4x4 matrix, other types exist, but max pooling is the most popular.

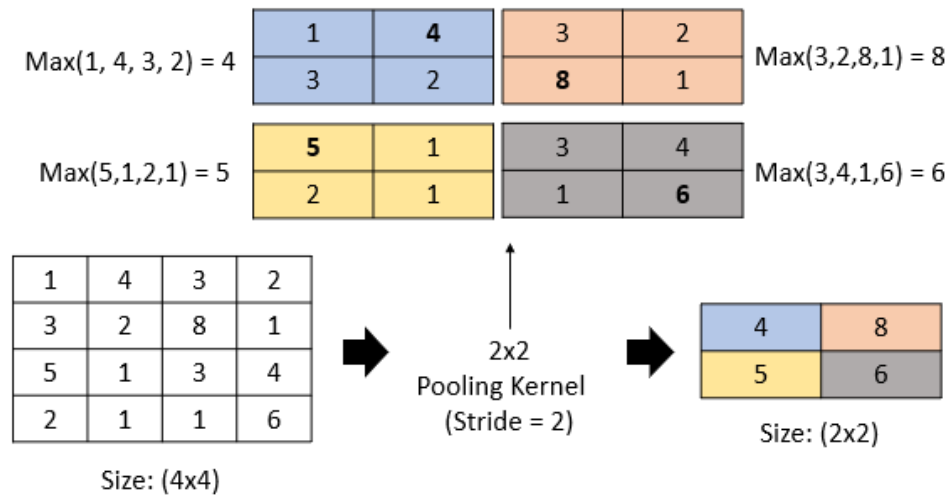


Figure 2-25. Example Pooling Application

2.3.2.3 Other Layers

A CNN will conclude with a fully-connected layer similar to the architecture of a regular ANN (O'Shea & Nash, 2015). However, in the creation of a CNN via the Python library, Keras, the model needs to add flatten and dense (aka fully-connected) layer(s).

A flatten layer simply resizes the resulting matrix into an $(n, 1)$ matrix where n represents the batch. Only one flatten layer needs to be added and simply multiplies the shape of the matrix to create one value, n . A CNN with shape, $(None, 1, 10, 64)$, will yield the shape, $(None, 640)$, after the flatten layer is applied (Keras, 2020).

The fully-connected/dense layer resizes the previous resulting matrix given the output size, *units*, and activation function (typically 'relu') (Keras, 2020). The dense

function uses the chosen activation function to reduce the size of the matrix into the size pre-determined by the *units* parameter (Keras, 2020).

2.3.3 Meaning Making for Understanding Images

Sometimes an image alone is not enough for human understanding of its meaning or context. For example, in Figure 2-26, a man figures out his weight as Barack Obama steps on the scale, thus increasing the weight without the man's knowledge. This figure provides an example of uncaptured context, for instance, the person stepping on the scale happens to be the President of the United States and the other people in the room are obviously in on the joke while the man on the scale is seemingly oblivious to the events. If only looking at the weight value, it is skewed from the man's true weight because he is missing context, i.e., the additional force added to the scale; if only looking at the picture through an algorithm, all of the contexts would be seemingly lost.



Figure 2-26. OAS Application – Weight Scale, from (Farhadi, et al., 2010)

One application that attempts to address understanding in images is the (*object, action scene*) (OAS) model (Farhadi, et al., 2010). The example presented in Figure 2-27 employs OAS for the contextually complex image of a horse and rider in motion. OAS selects the most accurate objective, action, and scene of the given image, based on the pre-coded potential options (Farhadi, et al., 2010). The OAS model uses a linear support vector machine (SVM) to determine the most appropriate nodes, that is the object, action, and scene labels, for a given image (Farhadi, et al., 2010). Then, OAS also uses similarity measures to determine how well the three-node labels fit to form a sentence (Farhadi, et al., 2010).



Figure 2-27. OAS Application – Horse Rider, from (Farhadi, et al., 2010)

Realizing that every object, action, and scene could not be identified and embedded into the program, an “Out of Vocabulary” extension was discussed (Farhadi, et al., 2010). By training the model on Tree-F1 (accurate and specificity) and BLUE (validity), sentences were constructed and featured unknown objects, actions, and scenes not in the original space (Farhadi, et al., 2010). Using a similar but different infrastructure of objects,

attributes, and relationships, computer vision models have been able to build complex sentences based on image recognition (Krishna, et al., 2017). Similarity, discovering and learning about out-of-library (OOL) concepts, i.e., “unknown unknowns,” is an important goal for AI today (Situ, Friend, Bauer, & Bihl, 2016). AR is proposed as one of the potential solutions to identifying and more accurately describing OOL objects, which is explored throughout the remainder of this document.

3 SYSTEMATIC COMPARISON OF ANALOGICAL REASONING

ALGORITHMS

To achieve a better understanding regarding text-based analogical reasoning, several algorithms from the discussion in Section 2.2.1 were selected for an apples-to-apples comparison. The approach to select AR algorithms involved selected based on its recency, its ability to work with $A:B::C:D$ word pair analogies, and its previous success in regards to other state-of-the-art algorithms. The algorithms selected were Distributed Representation Analogy Mapper (DRAMA), Bayesian Analogy with Relational Transformations (BART 1.0 and 2.0 versions), Word to Vector (Word2Vec), Global Vectors (GloVe), 3 Cosine Average (3CosAvg), and Linear Regression Cosine (LRCos). Using the AI schools of thought in which these algorithms fall (as shown in Figure 2-6), DRAMA was the most recent hybrid algorithm, BART was the most recent connectionist algorithm (excluding vector space models (VSMs)), Word2Vec and GloVe are considered to be the general standard for VSMs, and 3CosAvg and LRCos are more recent VSMs that has shown the most promising results.

3.1 ALGORITHM SELECTION

Upon evaluating different model types, in Chapter 2.2.1, this analysis leans more toward word-based analogies rather than “sentence-based analogies” due to VSMs being limited to analogies only in the form $A:B::C:D$. The algorithms discussed in the background primarily fall in the latter category hence why this analysis uses primarily VSM methods. The algorithms specifically tested in this study in addition to some of the other analogies mentioned above are divided according to the 3 AI Schools of Thought:

Symbolist, Connectionist, and Dynamicist in Figure 2-6. To date, there is yet to be a Dynamicist model with most falling in the remaining two categories. VSMS were considered to be a sub-section of the connectionist algorithms due to their basis in LRA. DRAMA and BART are psychologically-inspired AR models; whereas, the VSMS came from the natural language processing (NLP) field.

DRAMA is the only “full” algorithm that can be altered to understand analogies of the form $A:B:C:D$, while still staying true to the model’s interworking steps. In other words, DRAMA is capable of handling word, sentence, and story-based analogies.

Though BART is not a VSM, it was initially tested on $A:B::C:D$ -like analogies, and thus, appropriate for our analysis (Lu, Chen, & Holyoak, 2012). The remaining algorithms are all related and VSM-based. Since their creation, word2vec and GloVe have been used as the industry standard, making them essential to include. LRCos is described above more in-depth, but in the same paper the authors also suggest the “3CosAvg” method, so are also considering it (Drozd, Gladkova, & Matsuoka, 2016). It is important to note that Word2vec and GloVe are completely packaged programs; however, 3CosAvg and LRCos’s primary contribution is a new similarity measure method rather than a program as a whole. See

Table 3-1 for the VSM's similarity measurement equations based on the assumption of the analogy form, $a:a':b:b'$ (which is equivalent to $A:B::C:D$, but better shows relationships between the pairs), where V represents all words in the vector space.

Table 3-1. VSM Similarity Equations

Name	Equations
3CosAdd (Part of word2vec)	$\arg \max_{b' \in V} (\cos(b', b - a + a')) \quad (3-1)$
3CosMul (Part of GloVe)	$\arg \max_{b' \in V} (\cos(b', b) - \cos(b', a) + \cos(b', a')) \quad (3-2)$
3CosAvg (Corrected equation found in (Kafe, 2019))	$\arg \max_{b' \in V} (\cos(b', b + avg_offset)) \quad (3-3)$
	<p>Where,</p> $avg_offset = \frac{\sum_{i=0}^m a'_i}{m} - \frac{\sum_{i=0}^n a_i}{n} \quad (3-4)$
LRCos	$\arg \max_{b' \in V} P(b' \in target_class) \cos(b', b) \quad (3-5)$

3.2 COMPARATIVE EXAMPLE

Understanding some of the core differences in the theoretical structure of the selected models, the models were tested and compared using the same dataset and metrics.

3.2.1 Representative Example Data

As mentioned previously, e.g. Chapter 2.2, there are several different types of analogy problems (Ichien, Lu, & Holyoak, 2020). For the apples-to-apples comparison, the Sternberg-Nigro dataset (originally used in (Sternberg & Nigro, 1980)) was selected; however, due to availability, a modified version will be used. The modified Sternberg dataset was initially used in (Morrison, et al., 2004) and this dataset was modified to provide two choices rather than the previous standard of four answer choices found in the original.

In the modified Sternberg dataset, we call the two options: D , the “correct” option, and D' , the “distractor.” At first glance, one can see how C is related to both D and D' , but given the context of $A:B$, it is clearer as to why option D is better than D' . The modified Sternberg-Nigro dataset consists of 197 analogies divided into five relationships: synonym,

antonym, category, functional, and linear ordering (Sternberg & Nigro, 1980; Morrison, et al., 2004). Analogies are almost evenly split among the five categories with 40 synonyms, antonym, and category analogies, 41 functional, and 36 linear ordering analogies. The category relationship was further split into subordinate (35) and superordinate (5) analogies. Examples of each analogy relationship with its respective choices are shown in Figure 3-1.

Relationship	Example				
	A	B	C	D (Correct)	D' (Distractor)
Antonym	STOP	GO	EAST	WEST	DIRECTION
Synonym	NEAR	CLOSE	FIX	MEND	TAPE
Category - Subordinate	LION	ANIMAL	CHRISTMAS	HOLIDAY	EASTER
Category - Superordinate	DAY	SUNDAY	CLOTHES	SHOES	WEAR
Functional	BIRD	FLY	RABBIT	HOP	BUNNY
Linear Ordering	JANUARY	FEBRUARY	FIRST	SECOND	LAST

Figure 3-1. Example of Textual Data, from (Combs, Bihl, Ganapathy, & Staples, 2022) using data from (Sternberg & Nigro, 1980; Morrison, et al., 2004)

3.2.2 Performance Metrics

As discussed in (Combs, Bihl, Ganapathy, & Staples, 2022), two performance metrics were identified for evaluation in this study a correctness and “goodness” metric. The correctness metric looks at the accuracy of the algorithms when selecting between the correct option and the “distractor” for each analogy. The goodness metric evaluations how well the correct analogy pair, $C:D$, aligns with the given analogy, $A:B$. To calculate the values necessary, the following steps take place given an analogy, $A:B::C:[D,D']$:

1. calculate the similarity score between
 - a. A and B , $simAB$
 - b. C and D , $simCD$
 - c. C and D' , $simCD'$,

- 2 calculate the similarity ratio between
 - a. sim_{AB} and sim_{CD} ,

$$sim_{AB}/sim_{CD} = sim_{RD} \tag{3-6}$$
 - b. sim_{AB} and sim_{CD}' ,

$$sim_{AB}/sim_{CD}' = sim_{RD}', \tag{3-7}$$
- 3 take the absolute difference between the similarity ratios identified in Step #2 and an “ideal” analogy, 1. The formulas for the similarity ratios are

$$| sim_{RD} - 1 | \tag{3-8}$$
 and

$$| sim_{RD}' - 1 |, \tag{3-9}$$
 respectively,
- 4 compare the resulting values found in Step #3 and select the lower of the two as the option the algorithm would have selected as the “correct” answer.

This methodology is shown visually in Figure 3-2 in a conceptual example with the actual values given by the Word2Vec algorithm for the first analogy.

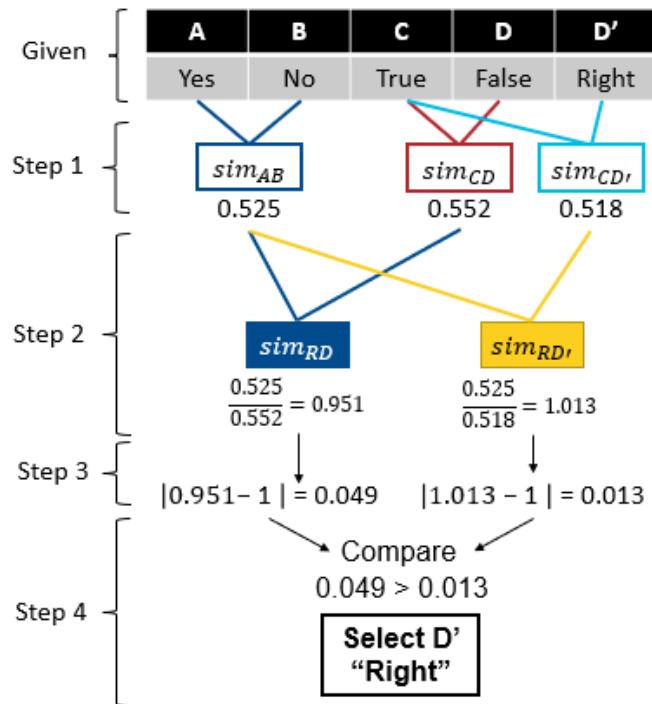


Figure 3-2. Conceptualization of Textual Evaluation Steps with an Example

3.2.2.1 Correctness Metric

In general, correctness is a percent looking at how many times the algorithm correctly selected D (over D') divided by the total (also called “raw”) or adjusted number of analogies. The raw percentage correct is formulated as

$$\begin{aligned} \text{Raw Percent Correct (RPC)} \\ = \frac{\text{Number of times } D \text{ was selected over } D'}{\text{Total number of analogies}}, \end{aligned} \quad (3-10)$$

where the number of times the correct answer was selected is divided by the total number of analogies. The adjusted percentage correct is also formulated as

$$\begin{aligned} \text{Adjusted Percent Correct (APC)} \\ = \frac{\text{Number of times } D \text{ was selected over } D'}{\text{Adjusted number of analogies}}, \end{aligned} \quad (3-11)$$

where the number of times the correct answer was selected is divided by the adjusted number of analogies, i.e., those which the algorithm could attempt to solve. The algorithm’s selection between D and D' is based on a comparison of their similarity metric explained in the next paragraph. The “raw” values look at the total number of analogies in the overall set for a given relationship, and the “adjusted” values are the number of analogies that the given algorithm has the potential to answer correctly. Several instances involved the model not knowing the A , B , and/or C words, which makes the remainder of the analysis impossible. With that begin said, the overall algorithm should not be penalized for this; however, if an algorithm does not understand many words, it is also not ideal. While the APC is a fairer comparison, it is important to consider the difference between the RPC and APC values since if there is a large difference, this suggests that an algorithm lacks vital “vocabulary” or the basic knowledge needed to solve the given analogy. An

ideal algorithm would be able to identify every word so that it can at least attempt every analogy.

3.2.2.2 Goodness Metric

The similarity metric is a continuous value that measures how similar two words are. When calculating this, DRAMA uses the dot product between word vectors, \vec{v}_1 and \vec{v}_2 (symbolized, $\vec{v}_1 \cdot \vec{v}_2$); whereas, BART, Word2vec, GloVe, 3CosAvg, and LRCos use cosine similarity to compare the potential solution space. DRAMA’s similarity scale ranges from [-1,1] instead of [0,1]; to create a level playing field, DRAMA’s similarity scores were modified per:

$$sim_{DRAMA} = \frac{\vec{v}_1 \cdot \vec{v}_2}{2} + \frac{1}{2}, \quad (3-12)$$

where \vec{v}_1 = vector representing word 1

\vec{v}_2 = vector representing word 2.

This will be called the “similarity” metric and be evaluated alongside the other metrics.

The analogy goodness score is given by the equation:

$$Goodness = 1 - sim_{RD} = 1 - \frac{sim_{AB}}{sim_{CD}}, \quad (3-13)$$

where sim_{RD} = similarity ratio for a given analogy, $A:B::C:D$

sim_{AB} = similar metric between words, A and B

sim_{CD} = similar metric between words, C and D .

3.3 RESULTS

The results from the correctness and goodness metrics yielded different results. DRAMA performed overwhelmingly well with 78.7% accuracy on the correctness metric; however, LRCos had the best average goodness metric of 0.055.

3.3.1 Correctness Metric Results

Figure 3-3 presents each model’s performance within each analogical relationship type, as described in (Combs, Bihl, Ganapathy, & Staples, 2022). Figure 3-3 is split into two sections looking at the raw percentage correct (RPC) on the left and the adjusted percentage correct (APC) on the right; furthermore, the columns beneath the metric denote the algorithms evaluated. The leftmost column of Figure 3-3, shows the different analogy relationships captured by the analogies within the Sternberg dataset. Highlighting in Figure 3-3 is used to denote the highest performance for each relationship type as well as for the dataset as a whole.

The results in Figure 3-3 show that DRAMA had the best overall performance and outperformed the other algorithms on the synonym, category, and linear ordering relationships. However, BART 2.0 tied DRAMA’s performance on functional analogies and had a slight advantage on those with an antonym relationship. DRAMA also had the highest performance for subordinate category problems; however, for the superordinate, BART 2.0 and LRCos tied one another. Since some of BART 1.0 and all of DRAMA’s mappings require hand-coding to identify the words within the analogies, their RPC and APC correctness scores are the same. All of the models were trained with enough vocabulary to attempt at least 188 of the total 197 analogies.

Analogy Relationship	Correctness Metrics													
	Raw Percent Correctness (RPC)							Adjusted Percent Correctness (APC)						
	DRAMA	BART 1.0	BART 2.0	Word2vec	GloVe	3CosAvg	LRCos	DRAMA	BART 1.0	BART 2.0	Word2vec	GloVe	3CosAvg	LRCos
Antonym	72.5%	42.5%	75.0%	42.5%	72.5%	40.0%	42.5%	72.5%	42.5%	75.0%	42.5%	72.5%	42.1%	44.7%
Synonym	80.0%	47.5%	76.3%	37.5%	55.0%	47.5%	50.0%	80.0%	47.5%	76.3%	41.7%	55.0%	50.0%	52.6%
Category	82.5%	42.5%	57.5%	47.5%	50.0%	57.5%	67.5%	82.5%	42.5%	57.5%	50.0%	51.3%	59.0%	69.2%
Subordinate	85.7%	42.9%	54.3%	54.3%	48.6%	57.1%	65.7%	85.7%	42.9%	54.3%	55.9%	48.6%	58.8%	67.6%
Superordinate	60.0%	40.0%	80.0%	0.0%	60.0%	60.0%	80.0%	60.0%	40.0%	80.0%	0.0%	75.0%	60.0%	80.0%
Functional	78.0%	58.5%	78.0%	56.1%	53.7%	61.0%	41.5%	78.0%	58.5%	78.0%	57.5%	55.0%	64.1%	43.6%
Linear Ordering	80.6%	63.9%	71.4%	52.8%	63.9%	47.2%	38.9%	80.6%	63.9%	71.4%	55.9%	65.7%	50.0%	41.2%
All	78.7%	50.8%	71.6%	47.2%	58.9%	50.8%	48.2%	78.7%	50.8%	71.6%	49.5%	59.8%	53.2%	50.5%

Figure 3-3. Percent Correctness Metric Result

Overall, from Figure 3-3 and based on having the highest percent correct metrics for both the RPC and APC, DRAMA was the best algorithm for the modified Sternberg dataset, followed by BART 2.0 and GloVe, respectfully, with the remaining algorithms having a similar performance around the 50% mark. At the top level, there was not a large difference in results between the RPC and APC scores; however, there was some shifting among the lower-ranking algorithms such as with 3CosAvg and LRCos. However, despite DRAMA's exceptional performance, there is not a "one size fits all" algorithm regarding the different analogy relationships tested. Though valuable, overall correctness may not be appropriate for studies that consider a large number of potential answers for D , an area where VSMs perform better.

3.3.2 Goodness Metric Results

In a comparison of the similarity metric, a heatmap of the analogy goodness measure scores for all of the considered data is shown in Table 8-1. In the table, an analogy goodness measure of 0.000 indicates that the given $A:B::C:D$ is equivalent to an "ideal" analogy as discussed in Chapter 3.2.2.2 and shown in Figure 3-2. An "average" analogy was determined to be 0.251 based on an average of the goodness score across all the algorithms. Anything with a score equal to or greater than 1.000 was considered a "poor" analogy. As mentioned earlier, the VSMs (Word2vec, GloVe, 3CosAvg, and LRCos) and BART 2.0 were not trained on certain words, and a goodness score could not be calculated; these instances were denoted in black. The table uses white to represent an "ideal" analogy, light grey to represent an "average" analogy, dark grey for a "poor" analogy, and black to denote analogies that could not be attempted by the given algorithm.

Looking at the average shown in the bottom row of Table 8-1, the algorithms rank as follows based on the goodness metric:

1. LRCos(0.055)
2. 3CosAvg(0.078)
3. BART 1.0 (0.107)
4. BART 2.0 (0.0220)
5. Word2Vec (0.417)
6. DRAMA (0.434)
7. GloVe (0.445).

When doing a broad visual overview, 3CosAvg and LRCos appear to be roughly tied followed by BART 1.0, BART 2.0, and the remaining models, which were tied on a different scale. In summary, LRCos provided the best possible comparison between analogies; however, it was followed relatively closely by 3CosAvg and BART 1.0, respectively.

Table 3-2. Goodness Metric Averages

	DRAMA	BART 1.0	BART 2.0	Word2Vec	GloVe	3CosAvg	LRCos
Average	0.434	0.107	0.220	0.417	0.445	0.078	0.055
Number of Analogies Unable to Attempt	0	0	3	19	6	11	11

3.4 SUMMARY AND CONCLUSION

This chapter presented a review and analysis of analogical reasoning algorithms for word-based analogies. This review focused on 6 algorithms: DRAMA (Eliasmith & Thagard, 2001), BART 1.0 (Lu, Chen, & Holyoak, 2012) & 2.0 (Lu, Wu, & Holyoak, 2019), Word2vec (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013), GloVe (Levy & Goldberg, 2014), 3CosAvg (Drozd, Gladkova, & Matsuoka, 2016), and LRCos (Drozd, Gladkova, & Matsuoka, 2016), which encompasses the general state of the art in the field today. Previous comparisons, see (Rogers, Drozd, & Li, 2017) (Kokinov & French, 2003)

(Gentner & Forbus, 2010) (Hall, 1989) (French R. M., 2002), only considered a small subset of these algorithms. In addition to providing a broad review of algorithms and their capabilities, the authors further provided comparison metrics and a consistent dataset for analysis. In a broad sense, it appears that psychological models currently have a slight advantage over VSMS based on our defined metrics, correctness, and analogy goodness. When concerned with the selection of the correct answer, DRAMA is the best overall model (78.7% correctness); however, the “best” model may depend on the relationship of a given analogy. When comparing models based on how “good” the similarity of an analogy is, LRCos has a small advantage over the other models (goodness score of 0.055).

When combining the results of the two evaluation metrics, the results in Table 3-3 can be computed. Here, in Table 3-3, each algorithm was ranked based on its performance on the adjusted correctness and goodness, with those ranking averaged in the third row. Based on the average, third row of Table 3-3, an overall ranking was assigned. Many ties were observed given the discrete nature of the rankings. Though close, and tied with 3CosAvg, BART 2.0 came out on top when considering both metrics.

Table 3-3. Textual Analysis Overall Results

Rankings	DRAMA	BART 1.0	BART 2.0	Word2Vec	GloVe	3CosAvg	LRCos
(Adj.) Correctness	1	5	2	7	3	4	6
Goodness	6	3	4	5	7	2	1
Average	3.5	4	3	6	5	3	3.5
Overall	T-3	5	T-1	7	6	T-1	T-3

4 IMAGE-BASED ANALOGICAL REASONING

The analysis of the state of the art in textual analogical reasoning (AR), Chapters 2 and 3, allowed for an exploration of the literature. Thus, of interest relative to handling unknown unknowns in image recognition, this can be extended into image-based analogies. Realistically, it is impossible to train an image recognition system on every possible object class that it might come across; of interest is a possibility to use AR to help derive unknown class labels when presented with the results of attempting to classify unknown image data. This chapter develops a process for this concept and the following sections are split into visual data set, algorithm framework, evaluation metrics, and results.

4.1 VISUAL DATA SET SELECTION

For the problem at hand, twelve datasets were considered: Caltech-101/256 (Griffin, Holub, & Perona, 2007), Canadian Institute for Advanced Research (CIFAR)-10/100 (Krizhensky, 2009), CINIC-10 (Darlow, Crowley, Antoniou, & Storkey, 2018), ImageNet (Russakovsky, et al., 2015), LabelMe (Russell, Torralba, Murphy, & Freeman, 2008), Microsoft Common Objects in COntext (COCO) (Lin, et al., 2015), Open Images Dataset V6 (Krasin, et al., 2020), Tiny Images (Torralba & Freeman, 2007), Visual Analogy Question Answer (VAQA) (Sadeghi, Zitnick, & Farhadi, 2015), and Visual Genome (Krishna, et al., 2016). Due to a limited number of classes (8-10 total), CIFAR-10, CINIC-10, and LabelMe were excluded from consideration. Tiny Images was removed in June 2020 due to controversial and offensive labeling (Birhane & Prabhu, 2021; Torralba, Fergus, & Freeman, 2020). VAQA had little to no other applications outside of the study it was created and evaluated in (Sadeghi, Zitnick, & Farhadi, 2015). The COCO,

Open Images Dataset V6, and Visual Genome datasets consisted of “noisy” images pre-annotated and typically used for multi-object recognition. CIFAR-100’s classes were limited in the sense they primarily pertained to animals and vehicles. Caltech-256 included and expanded upon Caltech-101, so it was selected due to its breadth of pre-labeled classes depicting singular objects (Griffin, Holub, & Perona, 2007).

Caltech-256 was an extension of the Caltech-101 dataset created in 2004, through the addition of over 150 classes, an increase in the minimum number of images per class, and previous images affected by rotation excluded (Fei-Fei, Fergus, & Perona, 2004; Griffin, Holub, & Perona, 2007). Caltech-256 has 256 object classes with a total of over 30,000 images with each class having at least 80 images (Griffin, Holub, & Perona, 2007). The images presented had varying length and height dimensions, which were standardized to be 128 by 128 pixels and a uniform color scheme. Though the vast majority of the pictures appear to be true color RGB (which would equalate to three channels) in the case that there were grayscale images, the images were condensed to one channel. In addition to the 256 classes, the original dataset technically has a 257th class called “clutter” “for testing background rejection” (Griffin, Holub, & Perona, 2007), which was ignored for this study. The different classes and their respective number of images are shown in Table 9-1 of Appendix B (101 in the name denotes a class originally in the Caltech-101 dataset).

4.2 ALGORITHM FRAMEWORK

The overall framework developed follows the routine of (1) image classification, (2) creation of class name word vectors, (3) application of AR & knowledge extraction, and (4) evaluation, though processes 1 and 2 (image classification and creation of class

name word vectors) can be done simultaneously or in reverse order, so as long as both processes are completed before process 3, application of AR & knowledge extraction. This framework is split primarily into four main sections as shown in Figure 4-1 with colors used to better separate each section. The sections developed are Image Classification (cyan), Creation of Class Name Word Vectors (Green), Application of AR & Knowledge Extraction (blue), and Evaluation (red). In Figure 4-1, and semi-consistent with general flow chart (e.g. (Lucid, 2021)), cylinders represent the data used, circles/ovals represent a program/algorithm/tool utilized, and squares/rectangles represent the entities yielded from the previous program/algorithm/tool(s). Consistent coloring showing the various as a whole, the framework will be called the Image Recognition Through Analogical Reasoning Algorithm (IRTARA).

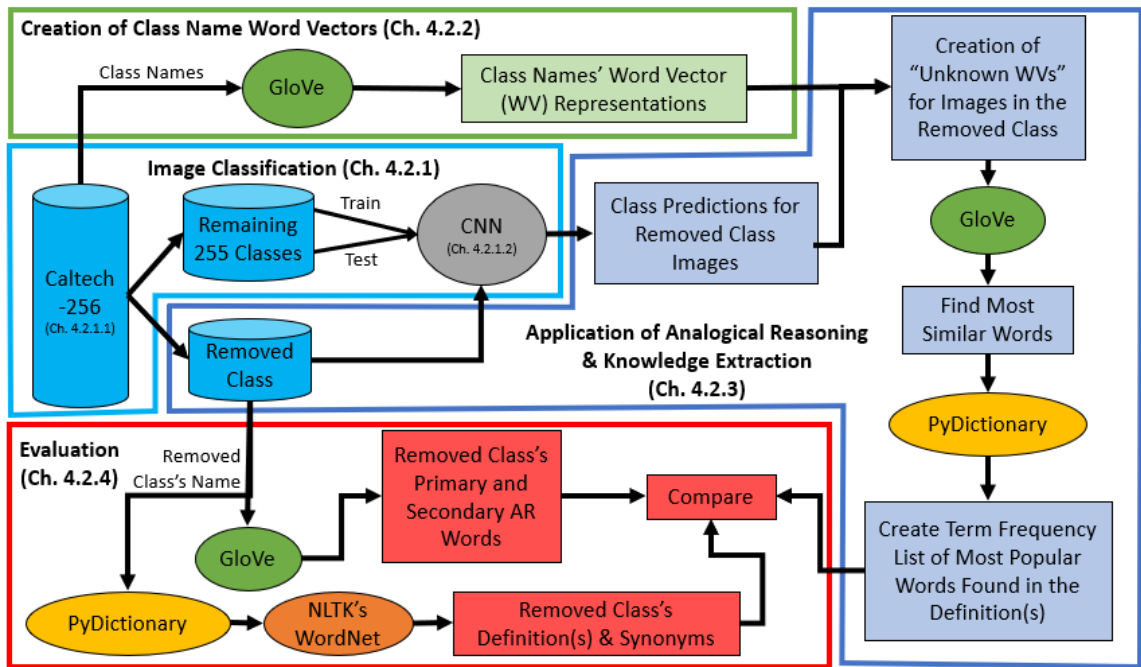


Figure 4-1. Image Recognition Through Analogical Reasoning Algorithm (IRTARA) Framework

4.2.1 Image Classification

The first step of the process shown in Figure 4-1 is to pose it as a standard image classification problem. The important elements of this process involve the initialization of data and the creation of the CNN architecture. This process includes two sub-steps of data processing and image recognition algorithm selection.

4.2.1.1 Data Processing

In the data processing step, and unlike in standardized supervised learning, a decision is made on which class to exclude from the classifier training. This is because IRTARA is intended to make predictions on an unknown class. Since there are not truly unknowns in this dataset, to assess performance, the Caltech-256 was used and removed one of the classes (called the “removed class”) before training, testing, and evaluating the model. Two important sections of the data are the images and the labels (or as we call them class names). Both the images and labels are sent to develop the CNN, but only the labels are used in the AR word vector creation section.

4.2.1.2 Image Recognition Algorithm Selection

Posed as a standard image classification problem, an image recognition algorithm is needed to provide a baseline for the AR algorithm to use. Though any image recognition could be integrated at this point, a CNN was selected for use in the overall framework with selection based on a combination of stated performance on Caltech-256, due to the promising results yielded from deep CNNs on 1000 classes from the ImageNet data set as part of the ImageNet Large Scale Visual Recognition Challenges (Russakovsky, et al., 2015).

Initially, IRTARA was thought to be integrated with one of Keras’ pretrained CNN architectures, which was initially trained on ImageNet dataset. Despite being pretrained on a different dataset, it could be used in conjunction with other image data sets through transfer learning. However, after preliminary trials with low accuracy, a simpler 11-layer CNN with similar accuracy to the Deep CNNs was integrated within the IRTARA framework.

The hyperparameters of the IRTARA CNN looking at all 256 classes included the following: *optimizer = adam, batch_size = 32, epochs = 10, and validation_split = 0.1*. The input shape was (128,128,1) for all images. On the Caltech-256 dataset, the IRTARA CNN showed an average accuracy rate of 22.1% and a loss of 0.0524 across 10 trials, which can be seen in

Table 4-1.

Table 4-1. CNN Average Results for Loss and Accuracy

Trial	Loss	Accuracy
1	0.0834	21.02%
2	0.0192	24.65%
3	0.0335	21.49%
4	0.0570	20.00%
5	0.0707	19.14%
6	0.0823	19.56%
7	0.0186	27.23%
8	0.0317	23.17%
9	0.0554	23.07%
10	0.0719	21.69%
Average	0.0524	22.10%
Standard Deviation	0.025	2.5%

After observing the IRTARA CNN’s performance on all 256 classes, it was adjusted to be compatible with only 255 classes since one class would need to be the “removed” class for a given trial. This was accomplished by having the second fully

connected layer have 255 nodes instead of 256. The IRTARA CNN is re-trained every time a new trial is run so that it is compatible with whichever 255 classes are known to the algorithm and so that it will not be inadvertently biased towards a removed class. Depending on the removed class, there may be multiple classes excluded; however, a superclass is not formed. Examples of when the exclusion of multiple classes include highly similar ones to the removed class such as “frog” and “toad” or “airplane” and “fighter-jet.” The architecture is shown in its tabular form in Table 4-2.

Table 4-2. CNN Architecture

Layer	Output Size	Filter Shape	Activation Function
Convolution	126 x 126 x 16	3 x 3 x 16	Relu
Max Pooling	63 x 63 x 16	2 x 2	Relu
Convolution	61 x 61 x 32	3 x 3 x 32	Relu
Max Pooling	30 x 30 x 32	2 x 2	Relu
Convolution	28 x 28 x 64	3 x 3 x 64	Relu
Max Pooling	14 x 14 x 64	2 x 2	Relu
Convolution	12 x 12 x 128	3 x 3 x 128	Relu
Max Pooling	6 x 6 x 128	2 x 2	Relu
Fully Connected		128	Relu
Fully Connected		255	Softmax

Visually, the IRTARA CNN is also shown in Figure 4-2 using a consistent color scheme. Blue represents the original image, yellow represents the result of a convolutional layer, green represents the result of a max-pooling layer, and grey represents fully connected layers.

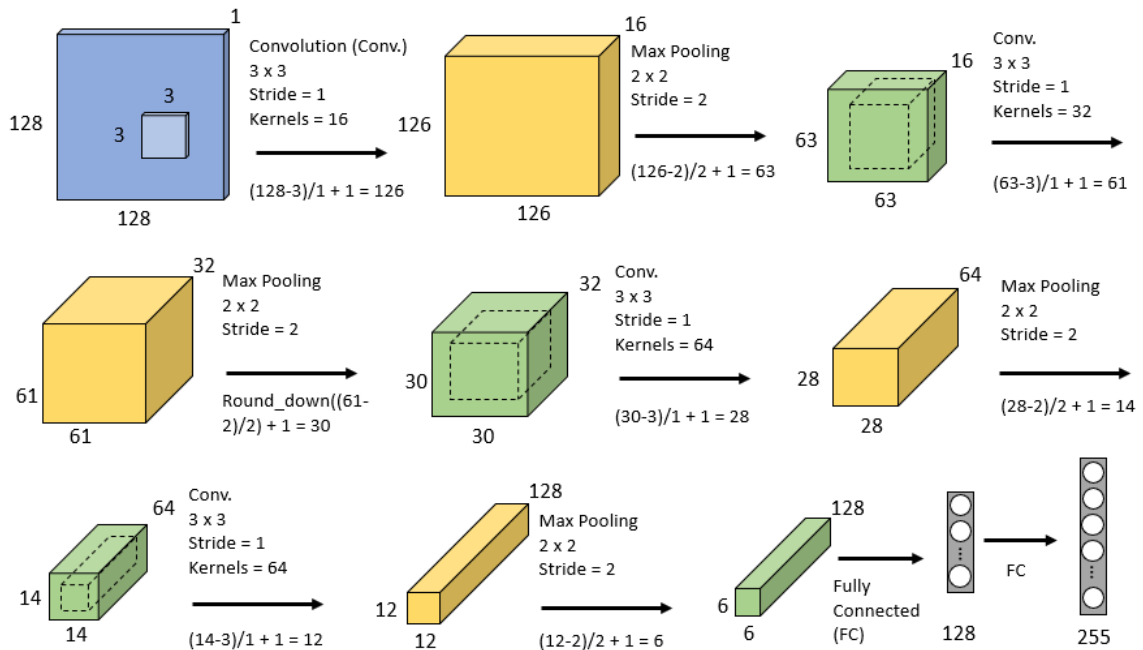


Figure 4-2. IRTARA CNN Architecture Visualization

The IRTARA CNN produced predictions on images within the unknown class based on the classes it has been trained on. It is important to understand that the prediction probabilities do not sum up to 100%, but rather is a confidence measure regarding whether one of the unknown images truly belongs to the identified class.

4.2.2 Creation of Class Name Word Vectors

This portion of the IRTARA, the green section of Figure 4-1, involves the incorporation of the GloVe algorithm (Levy & Goldberg, 2014). While the results in Table 3-3 showed that BART 2.0 and 3CosAvg performed best, additional requirements were required for the process in Figure 4-1. Looking at all six algorithms tested three (DRAMA, 3CosAvg, and LRCos) were incompatible with identifying a “most similar” function used to identify a specific word given a WV. The WV needed to represent individual words rather than phrases, which eliminated Word2Vec, and finally, BART 2.0 (shown simply as

“BART” in Figure 4-3) was likely compatible, but could not be integrated within the given time frame. Thus, GloVe was selected over the other AR algorithms tested in Chapter 0 to use within the IRTARA framework. GloVe is implemented in the Python library, Gensim (Řehůřek & Sojka, 2010).

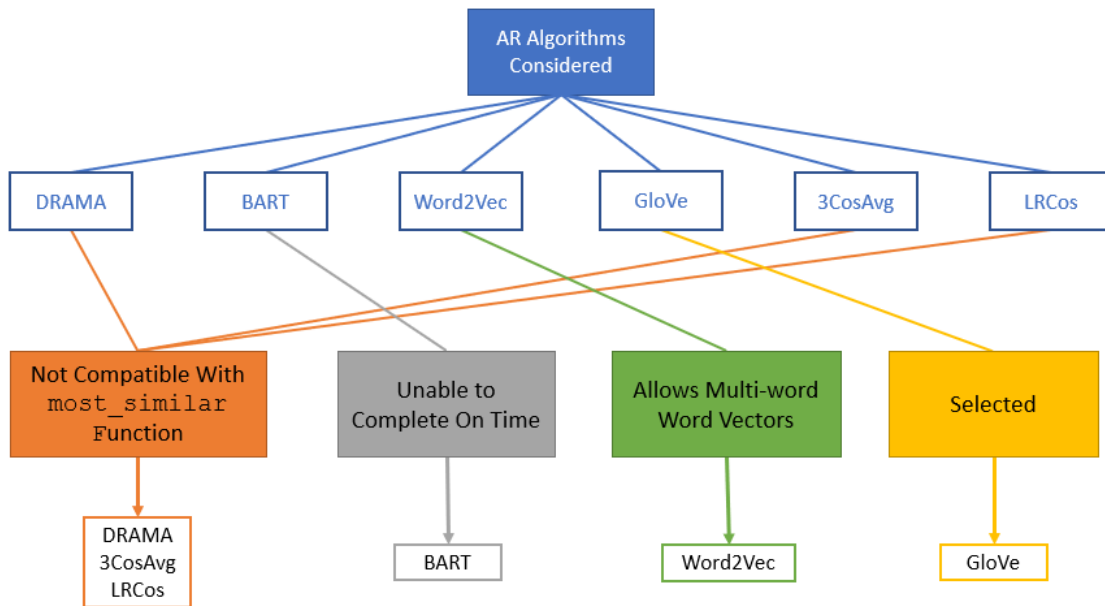


Figure 4-3. AR Algorithm Selection Process

In parallel with the IRTARA CNN creation, the creation of the class name word vectors process can occur. Using the class names as inputs, a user determines whether the algorithm “knows” a given word or phrase. One-word classes such as “backpack,” “dog,” and “kayak” are trivial for the algorithm; however, two-word or multi-word phrases require the need for new word vectors to be created. To create this “new” word vector, the class name is decomposed into its words and those vectors are summed. For example:

$$WV_{\text{coffee_mug}} = WV_{\text{coffee}} + WV_{\text{mug}} \quad (4-1)$$

Where,

$WV_{\text{coffee_mug}}$ = calculated word vector of the phrase, “coffee mug”

wv_{coffee} = word vector associated with “coffee”

wv_{mug} = word vector associated with “mug.”

There are some special cases where the class name cannot be accurately decomposed without human intervention due to:

1. Loss of important semantical meaning or potential misinterpretation (ak47, horseshoe crab, sheet music, etc.)
2. Lack of knowledge of (the decomposed) word(s) (Eiffel tower, triceratops, etc.)
3. “Noisy” or “uncommon” class names (self-propelled lawnmower, car side, etc.)

For these cases, a new word vector that’s associated with a different word is used in place of the label’s name. In total, this affected 93 classes and is shown in Appendix C’s Table 9-1.

4.2.3 Application of Analogical Reasoning & Knowledge Extraction

Using the class prediction probabilities produced from the IRTARA CNN and the AR class name word vectors, AR can be applied to each image to make a better-educated guess at the image’s true class. For one image belonging to the removed class, the IRTARA CNN produces probabilities associated with the known 255 classes; however, the algorithm selects a user-specified top m class to be sent to GloVe to create the “unknown” WV (uWV), where m can range from 1 to the remaining number of known classes. Looking at each class’s confidence, if it is equal to or above a user-defined threshold, α , it will influence the uWV . Looking at the top m classes with a confidence greater than the threshold, α , the product of class’s WV and its confidence is summed for each of the eligible m classes. For the IRTARA, $m = 5$ classes and $\alpha = 5\%$, the meaning of the top 5 classes if their confidences were greater than or equal to 5%, will influence the uWV . In

the scenario that none of the top m classes were at least α , the uWV was only influenced by the top class's WV. The uWV ideally “represents” the removed class for a given image, but likely does not correspond directly to an actual word.

Using this uWV, GloVe identifies the top k words that it has the highest cosine similarity with; k can almost have an infinite range. However, it is recommended that $k < 10$ considering that this process is repeated for each image within the removed class. Using the top- k words identified via GloVe, each word's corresponding definition (if available) is pulled from the PyDictionary library (geekpradd, 2020). In cases where a word is unknown, it is skipped over and the remaining words are analyzed. If a word has multiple definitions, they are all considered for our analysis. The definitions undergo preprocessing in which punctuation marks were removed, all words were transformed into lowercase versions, and “stop words” were removed. All the 197 default stop words identified by the Natural Language ToolKit (NLTK) corpus' English list are excluded as well 30 additional words as shown in Table 4-3 (NLTK, 2021).

Table 4-3. Stop Words List

Nltk Stopwords	A, about, above, after, again, against, ain, all, am, an, and, any, are, aren, aren't, as, at, be, because, been, before, being, below, between, both, but, by, can, couldn, couldn't, d, did, didn, didn't, do, does, doesn, doesn't, doing, don, don't, down, during, each, few, for, from, further, had, hadn, hadn't, has, hasn, hasn't, have, haven, haven't, having, he, her, here, hers, herself, him, himself, his, how, i, if, in, into, is, isn, isn't, it, it's, its, itself, just, ll, no, nor, not, now, m, ma, me, mightn, mightn't, more, most, mustn, mustn't, my, myself, needn, needn't, o, of, off, on, once, only, or, other, our, out, ours, ourselves, over, own, re, s, same, shan, shan't, she, she's, should, shouldn, should've, shouldn't, so, some, such, t, than, that, that'll, the, their, theirs, them, themselves, then, there, these, they, this, those, through, to, too, under, until, ve, very, up, was, wasn, wasn't, we, were, weren, weren't, what, when, where, which, while, who, whom, why, will, with, won, won't, wouldn, wouldn't, y, you, you'd, you'll, you're, you've, your, yours, yourself, yourselves
Additional Words	Adjective, adverb, awareness, cause, consisting, easily, especially, form, get, having, like, made, make, noun, object, often, one, put, resembling, someone, something, start, take, together, two, used, usually, various, verb, within

This process is repeated for all the definitions corresponding to the k words produced by the AR algorithm for each test image within the removed class. At this point, IRTARA looks at all the remaining words from the definitions and constructs a term frequency list across all the test images. The j highest-ranking words (ties not broken), were selected; however, if there were less than j words in the original term frequency list, the entire original list was considered.

For our analysis, $j = 100$ but j could range from 1 to the length of the original term frequency list, and for our analysis demonstrated in Chapter 4.3, $m = 5$ and $k = 5$ for simplicity and consideration for computational power. Though $m = k$ in our analysis, this is not an assumption of the IRTARA.

4.2.4 Evaluation

The evaluation block, outlined in red in Figure 4-1, describes two automated methods to quantitatively assess the results of the IRTARA framework. These methods are described briefly again in Chapter 4.5.

4.2.4.1 *Definition Evaluation*

The top term frequency words are compared to the words which are found in the removed class's definition(s). These definitions were primarily retrieved from PyDictionary using the class label with three exceptions:

1. the class consists of two or more words – a definition is constructed by either combining those belonging to each word (e.g., “beer mug” becomes a combination of “beer” and “mug”) or using one of the words that provide predominant meaning (e.g., “hot tub” becomes “tub”),
2. unknown context of words – a definition may be constructed using an alternative word (e.g., “Swiss army knife” becomes “penknife”) or by using an alternative dictionary (e.g., the South Park character, “Cartman,” does not have any reasonable synonyms, so an external definition was used),
3. poor quality definition – a definition was constructed using an alternative dictionary (e.g., PyDictionary's “galaxy” definition was simply, “(astronomy”).

Many of the words had multiple definitions due to having a variety of contexts. If a replacement word or definition had to be constructed, it is denoted in Table 9-1. For example, “floppy” (which represented “floppy-disk”), as defined by PyDictionary, can be used as an adjective or noun as shown:


```
{'Adjective': ['hanging limply'],
 'Noun': ['a small plastic magnetic disk enclosed in a
          stiff envelope with a radial slit; used to store
          data or programs for a microcomputer]}.
```

It is clear that in our context, the second definition, which corresponds to the noun version is the correct one (we are calling this the “Original Word” (OW) because it corresponds to the original word form). However, since the AR algorithm, GloVe, does not distinguish between these meanings, we will consider all of the words within all the definitions for an object/concept, called “All Words” (AW). All in the sense that it includes all the words within the definition(s) of the removed class; however, if there is only one definition, this simplifies to the OW. Also considering PyDictionary lacks understanding of synonyms or the plural forms of words, a separate list is created that includes these words. For synonyms, nltk’s WordNet corpus is utilized, especially its `lemmas()` function to identify words with a similar meaning to those passed in. WordNet utilizes underscores to connect two words such as (“credit_card”), but since all of the definitions are broken down to a word-by-word basis, these synonyms were removed completely.

Next, the plural form of the words within the original definition and its synonyms are pluralized using `inflect’s plural()` function. When applied to the correct definition’s original words, these are called “Synonym Words” (SW) and when applied to all of the words in all of the definitions (aka AW) they are called “All Synonym Words” (ASW). It would be ideal if the term frequency words were found within the OW list, but the next best would be the SW, with AW and ASW coming in third and fourth, respectively. The OW list is a subset of the AW similar to how the SW is a subset of the ASW; however,

OW/AW and the SW/ASW lists are mutually exclusive from one another (ex. a word cannot be in one of the definitions, but also a synonym). Using a condensed version of the “floppy” definition results mentioned previously, these relationships are shown in Figure 4-4 (for reference, there were no synonyms found for “limply”).

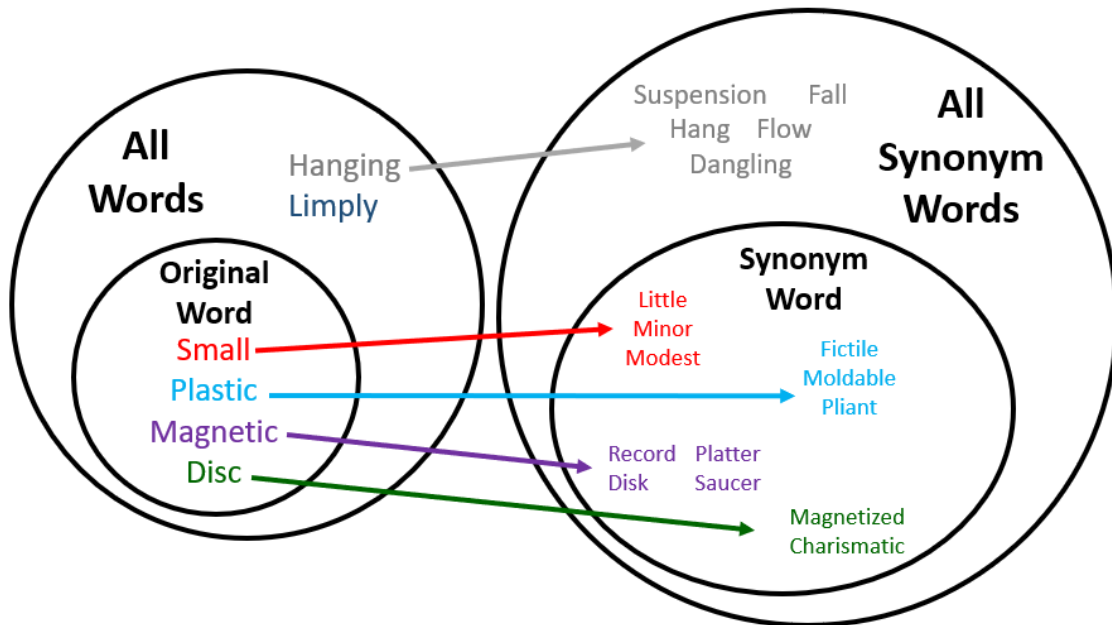


Figure 4-4. Relationship Diagram Between Definition Words
 These definition words are compared to the words found in the term frequency list.

A Boolean value representing each of the different “types” of words found in the definition (OW, SW, AW, and ASW) is associated with each term frequency word regarding whether that specific word appears in the various forms of the definition. These values are then, totaled to provide an overall summary.

4.2.4.2 Analogical Reasoning Evaluation

To implement this evaluation method, the primary and secondary AR words associated with the removed class are explored. Primary AR words are those that are the most similar ones to the removed class, which are identified using gensim’s most_similar function on the removed class. Secondary AR words follow a similar

pattern except that the `most_similar` function is used on the primary AR words. The number of primary AR words can be set to any number, but this will directly influence how many secondary AR words exist (and therefore, computation time). For this analysis, the top p primary AR words are pulled, and then, the top s words are associated with those primary AR words. Therefore, there will be a maximum of (ps) secondary AR words; however, this is usually reduced after duplicates and word variations are removed.

The initial primary AR words are converted to their base word, if necessary, using NLTK's WordNet Interface function, `morph` (Bird, Klein, & Loper, 2009) (`morph`'s functionality is specifically discussed in (NLTK, 2021)). Base words vary from case to case but usually reduce a word down to its root with exceptions allow if the root word is of another part of speech than the original ("robot" and "robotic" are treated as separate base words because one is a noun and the other is an adjective). Additionally, if the class name itself appears in this list (which is often the case because we pass through the WV representation rather than the actual class name), it will need to be removed. In the case the removed class has multiple words, each word will be removed as well as all combinations of the words. For example, "ice-cream-cone" will have the following words removed (if they appear): "ice," "cream," "cone," "icecream," "icecone," "creamcone," and "icecreamcone." If the removed class is represented by another word, that word is also removed; this only affects a handful of classes shown in Table 4-4.

Table 4-4. Classes with Additional Removed Words/Phrases from Primary and Secondary AR Word Lists

Class	Additionally Removed Word/Phrase
Top-hat	Top-hat
Triceratops	Dinosaurs
Washing-machine	Washer
Horseshoe-crab	Limulus
Palm-pilot	PDA
Stained-glass	Stained-glass
Swiss-army-knife	Penknife

4.3 EXAMPLE RESULT WALKTHROUGHS

To better show how the results are generated, the next sections will show two walkthroughs. The first will look step-by-step at the process for a class that had good results and another one with bad results. The second one will walk through the process with the removed class not revealed at the end to see how the results can infer the identity of the removed class.

4.3.1 Creation of Class Name Word Vectors











The creation of the class name WV, denoted in the green portion of Figure 4-1, does not necessarily have to be the first step; however, since it only needs to be completed once for the entire data set, it may be easiest to complete first. For both walkthroughs, the same class name WVs can be used. These were previously determined and listed in Table 9-1. The process by which they were created is explained in-depth in the previous Chapter 4.2.2. Gensim’s implementation of GloVe was used for this process (Řehůřek & Sojka, 2010). The class names or their representations must be recognizable by GloVe, which must be figured out manually for the classes that may not directly align with the original form.

4.3.2 Bad vs. Good Walkthrough

To better show how IRTARA works, a step-by-step walkthrough looking at a “good” and “bad” example is demonstrated through this section. There are 156 images within the mars class and 106 images within the chandelier class, a sample of both are shown in

Table 4-5. For an illustration of general performance, irrelevant to final results since the IRTARA CNN never trains on them, using the same CNN architecture but expanded to be able to classify 256 classes rather than 255, chandelier had an accuracy of 54.2% and mars had an accuracy of 81.0%. As mentioned in Chapter 4.1, all the images were resized to a standard 128 x 128 grid with one color channel in case any images were black and white. This step corresponds to the image classification portion of Figure 4-1 in which the removed class has been identified and the CNN is trained on the remaining 255 classes.

Table 4-5. Sample Images of Mars and Chandelier

Image Index	Mars	Chandelier
001		
002		
003		
004		
005		

Once the IRTARA CNN is trained on the remaining classes, all the images from the removed class are fed into the IRTARA CNN and their top n classes, i.e. 5 in our analysis, classes and their confidences are identified. The top 5 classes are shown in Table 4-6, for the mars and chandelier picture with image index 002 (see

Table 4-5). Any number of classes could be considered up to 255; however, this would introduce a significant amount of noise into the process.

Table 4-6. CNN Results for Mars and Chandelier Image 002

Ranking	Mars		Chandelier	
	Class	Confidence	Class	Confidence
1	Lightning	0.3558	Ladder	0.2214
2	Comet	0.1781	Screwdriver	0.1142
3	Saturn	0.0809	Golden-gate-bridge	0.0529
4	Galaxy	0.0788	Swiss-army-knife	0.0499
5	Umbrella	0.0392	Airplane	0.0335

Going back to the single image example started in Table 4-6, each class's confidence is compared to a minimum threshold (α) value as shown in Table 4-7. Any confidence below the threshold is considered to be noise and does not further impact the rest of the analysis. After analyzing the results, it was decided that $\alpha = 5\%$. These results are the beginning of the Application of AR & Knowledge Extraction process in Figure 4-1.

Table 4-7. Confidence-threshold Comparison for Mars and Chandelier Image 002

Mars			Chandelier		
Class	Confidence	Conf. > α ?	Class	Confidence	Conf. > α ?
Lightning	0.3558	Yes	Ladder	0.2214	Yes
Comet	0.1781	Yes	Screwdriver	0.1142	Yes
Saturn	0.0809	Yes	Golden-gate-bridge	0.0529	Yes
Galaxy	0.0788	Yes	Swiss-army-knife	0.0499	No
Umbrella	0.0392	No	Airplane	0.0335	No

If the confidence is above α , then, that class influences the “unknown word vector” (uWV). The uWV is “unknown” because it does not directly correspond to a word, but ideally (depending on the quality of the CNN results), it would “represent” the removed class. The uWV is constructed by taking the product of the CNN class's WV (which was predetermined and found in Table 9-1) and its confidence, and then, summing these values

for all that meet the threshold minimum. Using the data in Table 4-7, the uWV for mars and chandelier image 002 would be as follows:

$$uWV_m = WV['Lightning'] * 0.3558 + WV['Comet'] * 0.1781 + WV['Saturn'] * 0.0809 + WV['Galaxy'] * 0.0788 \quad (4-2)$$

and

$$uWV_c = WV['Ladder'] * 0.2214 + WV['Screwdriver'] * 0.1142 + WV['Golden-gate-bridge'] * 0.0529, \quad (4-3)$$

where $WV[<class\ name>]$ is the word vector representation of the class corresponding to the word or phrase found in Table 9-1, which had already been predetermined in the Creation of Class Name Word Vectors process shown in green in Figure 4-1.

At this point, IRTARA uses Gensim's `most_similar` function with GloVe's corpus to apply AR to the uWV (Řehůřek & Sojka, 2010). The `most_similar` function finds the `topn` closest words given either a single WV or multiple ones. In this analysis, a singular uWV is sent into the positive parameter and `topn = m` to retrieve the top m closest words (again, $m = 5$ in this analysis) based on their cosine similarity, which is being called the "top AR words." Going back to the example, the top 5 AR words and their cosine similarity are shown in Table 4-8. When comparing the CNN results in Table 4-7 and the AR results in Table 4-8, there are several repeated words or word variations. Since those words influenced the creation of the uWV, it makes sense that their cosine similarity would be relatively high compared to other words; however, occasionally an unseen word will enter at this point such as "meteor" for mars and "rungs" for chandelier.

Table 4-8. Top AR Words for Chandelier and Mars Image 002

	Mars		Chandelier	
Ranking	Top AR Word	Cosine Similarity	Top AR Word	Cosine Similarity

1	Lightning	0.8356	Ladder	0.8841
2	Comet	0.6269	Ladders	0.6041
3	Galaxy	0.5007	Rope	0.5495
4	Meteor	0.4915	Rungs	0.5017
5	Saturn	0.4879	Screwdriver	0.5013

AR introduces some new terms that (sometimes) help guide us toward the removed class; however, we wanted to find more context that may bring us closer. To do this, the definition corresponding to all the top AR words are found via the PyDictionary library, utilized in the final parts of the Application of AR & Knowledge Extraction process in Figure 4-1, and formatted appropriately so that it only works with some semantic meanings are left. These words' definitions and their formatted “definition words” are found in

Table 4-9 and

Table 4-10 for mars and chandelier, respectively.

Table 4-9. Top AR Word's Definition and Formatted Definition Words for Mars Image 002

Mars		
Top AR Word	Definition	Definition Words
Lightning	{'Noun': ['abrupt electric discharge from cloud to cloud or from cloud to earth accompanied by the emission of light', 'the flash of light that accompanies an electric discharge in the atmosphere (or something resembling such a flash)]}	Abrupt, electric, discharge, cloud, cloud, cloud, earth, accompanied, emission, light, flash, light, accompany, electric, discharge, atmosphere, resembling, flash
Comet	{'Noun': ['(astronomy)]}	Astronomy
Galaxy	{'Noun': ['a splendid assemblage (especially of famous people', 'tufted evergreen perennial herb having spikes of tiny white flowers and glossy green round to heart-shaped leaves that become coppery to maroon or purplish in fall', '(astronomy)]}	Splendid, assemblage, famous, people, tufted, evergreen, perennial, herb, spikes, tiny, white, flowers, glossy, green, round, heart-shaped, leaves, coppery, maroon, purplish, fall, astronomy
Meteor	{'Noun': ['(astronomy', "a streak of light in the sky at night that results when a meteoroid hits the earth's atmosphere and air friction causes the meteoroid to melt or vaporize or explode"]}	Astronomy, streak, light, sky, night, results, meteoroid, hits, earth, atmosphere, air, friction, meteoroid, melt, vaporize, explode
Saturn	{'Noun': ['a giant planet that is surrounded by three planar concentric rings of ice particles; the 6th planet from the sun', '(Roman mythology)]}	Giant, planet, surrounded, three, planar, concentric, rings, ice, particles, planet, sun, roman, mythology

Table 4-10. Top AR Word's Definition and Formatted Definition Words for Chandelier Image 002

Chandelier		
Top AR Word	Definition	Definition Words
Ladder	{'Noun': ['steps consisting of two parallel members connected by rungs; for climbing up or down', 'ascending stages by which somebody or something can progress', 'a row of unravelled stitches'], 'Verb': ['come unraveled or undone as if by snagging']}	Steps, consisting, two, parallel, members, connected, rungs, climbing, up, down, ascending, stages, somebody, progress, row, unravelled, stitches, come, unraveled, undone, snagging
Ladders	{'Noun': ['steps consisting of two parallel members connected by rungs; for climbing up or down', 'ascending stages by which somebody or something can progress', 'a row of unravelled stitches'], 'Verb': ['come unraveled or undone as if by snagging']}	Steps, consisting, two, parallel, members, connected, rungs, climbing, up, down, ascending, stages, somebody, progress, row, unravelled, stitches, come, unraveled, undone, snagging
Rope	{'Noun': ['a strong line', 'street names for flunitrazepan'], 'Verb': ['catch with a lasso', 'fasten with a rope']}	Strong, line, street, names, flunitrazepam, catch, lasso, fasten, rope
Rungs	{'Noun': ['a crosspiece between the legs of a chair', 'one of the crosspieces that form the steps of a ladder']}	Crosspiece, legs, chair, one, crosspieces, form, steps, ladder
Screwdriver	{'Noun': ['a hand tool for driving screws; has a tip that fits into the head of a screw', 'a cocktail made with vodka and orange juice']}	Hand, tool, driving, screws, tip, fits, head, screw, cocktail, vodka, orange, juice

A list that combines all the definition words for all top AR words for each image within a class is constructed. For sake of space, this list will not be shown since there are over 1,000 words in the mars list and over 400 in the chandelier list. For each word in the list, a count is created that denotes how many times that word appears. The term needs to appear at least j times to be included in this list; for our purposes $j=10$. Since this list is a compilation of data from the entire class, the remainder of this walkthrough will be

focusing on the class as a whole rather than a specific image. The top-100 words in the list, or the entire list if there are less than 100 words, will be used for the remainder of the analysis, and this version is dubbed the “term frequency list.” The term frequency list is shown in Table 10-1. Looking at the words in Table 10-1, it is clear the words within the mars list better align with mars rather than chandelier’s words do.

4.3.2.1 Results Variation Explanation

Why this happens likely boils down to two main reasons: (1) sparsity of CNN results and (2) homogeneousness of the class names. This is evident when looking at the totaled top-ranking CNN classes and their respective counts for both classes are shown in Table 4-11.

Table 4-11. Top Ranking CNN Classes for Chandelier and Mars

	Mars		Chandelier	
Ranking	CNN Class	Count	CNN Class	Count
1	Brain	56	Tower-pisa	9
2	Saturn	36	Microscope	8
3	Comet	21	T-shirt	8
4	Galaxy	6	Fireworks	7
t-5	Bowling-ball	5	Teapot	6
t-5			Watch	6

Starting with the first reason, sparsity of the IRTARA CNN results, when looking at the results in Table 4-11, the data is focused on three main classes for mars, with there being a steep drop off after the third class. However, the results are low and significantly more distributed when looking at chandelier’s. Now shifting to the second reason, homogeneousness of the class names, when looking at the top classes, we realize that mars’ top 2, 3, and 4 classes all fit in the general field of “astronomy” or “space.” However, chandelier’s results lack an encompassing subject with one another and with the word

“chandelier” outside of a general “household items” category shared by t-shirt, teapot, and watch.

4.4 MYSTERY CLASS WALKTHROUGH

With an understanding of how the results are developed and constructed, this walkthrough will go through all portions of the process without prior knowledge of what the removed class is, hence the name “mystery class.” This is to mimic how IRTARA evaluates an unknown unknown scenario.

For the first 5 images, their IRTARA CNN classifications are shown in Table 4-12. For sake of space, the gray-shaded classes correspond to having confidence greater than or equal to the threshold of 0.05.

Table 4-12. Mystery Class IRTARA CNN Classes for First Five Images

Image Number	Class 1	Class 2	Class 3	Class 4	Class 5
1	Lightning	Light-house	Smokestack	Fireworks	Pyramids
2	Lightning	Boxing Glove	Light-house	Bathtub	Mushroom
3	Light-house	Lightning	Lightbulb	Smokestack	Bathtub
4	Light-house	Lightning	Minaret	Smokestack	Windmill
5	Lightning	Light-house	Bathtub	Swan	Ibis

To provide an encompassing overview of all the mystery images, Table 4-13 shows IRTARA CNN classes that appear at least 10 times within the top five classes. Similar to what we saw with the mars example earlier, the classification is condensed into predominantly two classes for Class 1, light-house and lightning, and overall, four classes, light-house, lightning, bathtub, and smokestack. In an attempt to keep this portion neutral, the class themes will not be discussed until after the class is revealed.

Table 4-13. Top IRTARA CNN Class Results for All Mystery Images

Class	Class 1	Class 2	Class 3	Class 4	Class 5	Total
Light-house	35	27	10	4	5	81
Lightning	37	24	5	2	3	71
Bathtub	11	7	9	8	11	46
Smokestack	3	4	16	11	7	41
Minaret	0	2	14	7	9	32
Mattress	4	6	2	8	7	27
Golden-gate-bridge	1	3	2	4	5	15
Lightbulb	0	1	4	7	2	14
Windmill	0	2	1	4	5	12
Comet	0	2	3	4	2	11
Blimp	0	0	2	4	5	11
Flashlight	0	3	5	1	1	10

After sending the uWV into GloVe, the top five AR words for the first five images are shown in Table 4-14. From there, we can see some duplicate words from the IRTARA CNN classes as well as a few new words, which are shaded in gray. We are introduced to words such as thunder, Tampa, flames, storms, night, white, the, and lights.

Table 4-14. Top AR Words for First Five Mystery Images

Image No.	AR Word 1	AR Word 2	AR Word 3	AR Word 4	AR Word 5
1	Lightning	Thunder	Tampa	Flames	Storms
2	Light	Boxing	Glove	Lightning	Night
3	Light	House	White	The	Night
4	Light	House	Night	The	Night
5	Lightning	Light	House	Night	Lights

To provide another encompassing overview, the top AR words which appear at least ten times for all the mystery images are shown in Table 4-15. Again, the parallels between the IRTARA CNN classes and the top AR words are clear; however, this method also introduces new words into the algorithm. Similar to the IRTARA CNN classes, we see a significant drop in overall totals after the light, house, and lightning words.

Table 4-15. Top AR Words for Mystery Images

	AR Word 1	AR Word 2	AR Word 3	AR Word 4	AR Word 5	Total
Light	22	33	0	1	0	56
House	14	21	9	1	1	46
Lightning	35	1	7	2	0	45
The	0	0	3	17	5	25
White	0	0	17	2	4	23
Thunder	0	15	5	1	1	22
Night	0	0	5	7	10	22
Flames	0	2	7	5	3	17
Lights	0	0	5	3	8	15
Houses	0	0	0	4	10	14
Bathtub	10	1	1	0	1	13
Tampa	0	0	8	3	2	13

Going back to the first five images and their AR words, their definitions are pulled from PyDictionary and had the words with semantic meaning pulled out for the next analysis. Since the first five images share many of the same AR words, rather than showing the process for each image, the AR words' definitions and their words are shown in Table 11-1. The term frequency list was constructed and the first hundred entries are shown in

Table 11-2. However, by looking at the top twenty most frequent words, we gain some insights into what the mystery class may be as shown in Figure 4-5. A larger size denotes words that appeared more frequently. Light appeared the most with 604 instances and several words appeared more than 200 times including little, illumination, fire, united, and states.



Figure 4-5. Mystery Class Word Cloud

Though the words may not immediately direct us toward the mystery class, once the class is revealed, there are clear words that correspond to or might be associated with the mystery class. Given the term frequency list or the word cloud, what appears to be common themes regarding the context of the words? Without knowing what the mystery class is, the reader is experiencing an “unknown unknown,” it is difficult to cipher the meaning behind the word cloud. As with anything, there is noise, which can be difficult to identify in an unknown unknown because it eliminates the ability to use context clues. Just within the first twenty words, there are clear indications of light in various forms – illumination, fire, color, visual, etc. Grouping these into a “light” category, we can also ground a “science” category by including cloud, living, physics, discharge, and device. The

remaining words – children, members, great, states, united, little, divine, and building – do not necessarily fit neatly in an obvious category, so they will be assumed to be noise at the moment. The two categories identified were “light” and “science,” which will steer us toward perhaps a tangible device like a lightbulb or something intangible such as the sun.

As for the reveal, the mystery class in this instance “rainbow.” Perhaps this was not the immediate thought when looking at the words in Figure 4-5, but several words probably align with most people’s mental model of a rainbow such as light, color, and visual. Other words might be semi-related such as cloud (rainbows often appear in the sky with and like clouds) and building (a rainbow shape resembles a structure of some sort). The list is vague with noise, but a decent description of a rainbow if one were unfamiliar with it. While trivial since a human could look at the pictures shown in Figure 4-6, this contextual result provides value since future objects explored in this construct may not have meaningful human meanings or too numerous in quantity to individually view. Though a human may be able to arbitrarily be able to tell whether a vague list of words accurately describes the removed class, automated metrics are needed to be able to assign a quantitative score to the results we get for each class.

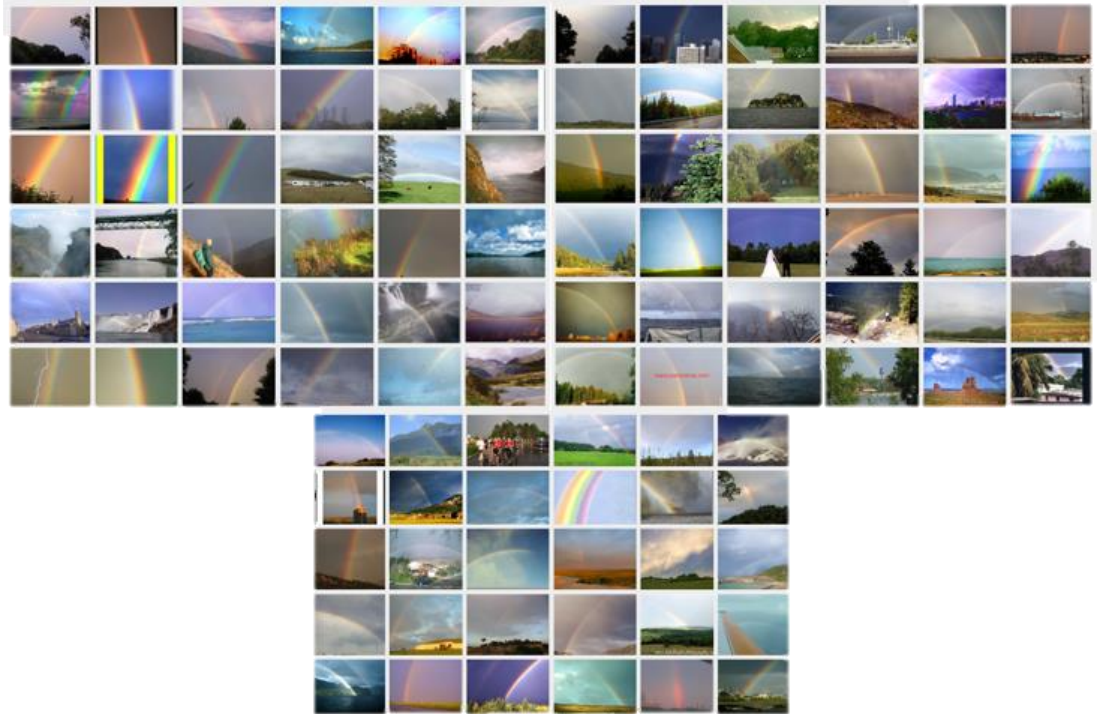


Figure 4-6. Tile View of Mystery Class Images from the Caltech-256 dataset (Griffin, Holub, & Perona, 2007)

4.5 EVALUATION METHODS & METRICS

While the results in Chapter 4.3 show qualitative value, a quantitative understanding of the results is needed. To do so, the results from the term frequency lists are evaluated based on three different methods: (1) definition evaluation, (2) AR evaluation, and (3) human factors (HF) evaluation. The definition and AR evaluations are not necessarily supposed to be compared to each other, but rather two different outlooks on the same data to provide additional insights. However, they can each be compared to the HF results to see how well the automated quantitative methodologies match with a human's qualitative evaluation.

4.5.1 Definition Evaluation Method

As discussed in Chapter 4.2.4.1, the definition evaluation compares the words found in the term frequency list to the removed class’s definition(s) through four ways:

1. the original word (OW) – words found in the removed class’s true definition,
2. synonym word (SW) – synonyms of the words found in the removed class’s true definition,
3. all words (AW) – words found in all the removed class’s definition(s), if multiple due to multiple meanings,
4. all synonym words (ASW) – synonyms of the words found in all of the removed class’s definition(s).

Returning to our good and bad example, Mars and chandelier, from above, we can see how the results from this method provide a qualitative value to rank how well the term frequency list represents (or doesn’t represent) the removed class. The definition and its respective words for chandelier and mars are shown in Table 4-16. Since chandelier only has one definition, the words found in its OW list are the same as its AW list and the same goes for its SW list and ASW lists. Since there are over 100 SW and ASW for both classes, a small subset has been included to save space.

Table 4-16. Definition Words for Mars and Chandelier

	Mars	Chandelier
PyDictionary Definition(s)	{'Noun': ['the month following February and preceding April', "a mark or flaw that spoils the appearance of something (especially on a person's body", 'a small reddish planet that is the 4th from the sun and is periodically visible to the naked eye; minerals rich in iron cover its surface and are responsible for its characteristic color', '(Roman mythology)'], 'Verb': ['make imperfect',	{'Noun': ['branched lighting fixture; often ornate; hangs from the ceiling']}

	'destroy or injure severely']}]	
True Definition	'a small reddish planet that is the 4th from the sun and is periodically visible to the naked eye; minerals rich in iron cover its surface and are responsible for its characteristic color'	'branched lighting fixture; often ornate; hangs from the ceiling'
Original Word	Small, reddish, planet, sun, periodically, visible, naked, eye, minerals, rich, iron, cover, surface, responsible, characteristic, color	Branched, lighting, fixture, ornate, hangs, ceiling
Subset of Synonym Word	Small: little, minor, modest, ... Reddish: red, ruddy, carmine, Color: colour, coloring, colouring, ...	Branched: ramify, branch, fork, furcate, separate, ... Lighting: light, ignition, firing, kindling, Ceiling: roof, cap
All Words	Month, following, february, preceding, April, mark, flaw, spoils, appearance, body, small, reddish, planet, sun, periodically, visible, naked, eye, minerals, rich, iron, cover, surface, responsible, characteristic, color, Roman, mythology, imperfect, destroy, injure, severely	Same as Original Word
Subset of All Synonym Words	Month: None Following: followers, pursuit, chase, Severely: badly, gravely, seriously, ...	Same as Subset of Synonym Word

The list of definition words associated with the removed class (or its representation) can be compared to the words found in the term frequency list. The words which have overlap between the lists can be found in Table 4-17. One of Mars' words is "th" which is because its original definition included the term "4th," however, non-alphabetic were stripped from the definitions for the analysis, hence leaving "th." Mars has several words that appear (visually shown in Figure 4-7), but chandelier is very clear since there is only one word in the term frequency list.

Table 4-17. Comparison Between Term Frequency List and Definition Words for Chandelier and Mars

	Term Frequency Word	
	Chandelier	Mars
Original Word		Planet, sun, responsible, th
Synonym Word	Light	Satellites, satellite, centers
All Words		Planet, sun, responsible, th, mythology, body, roman

All Synonym Words	Light	Satellites, satellite, centers, someones, people, person
-------------------	-------	--

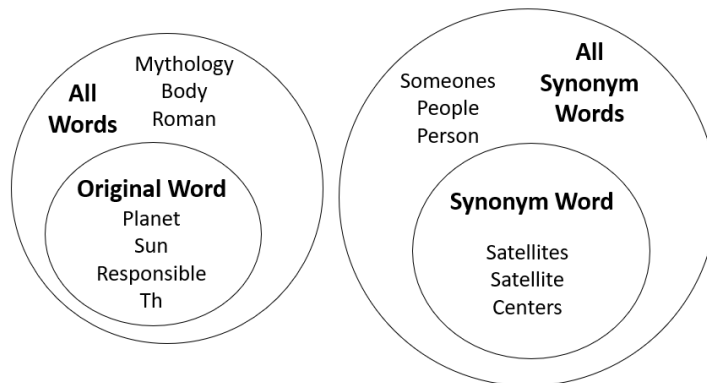


Figure 4-7. Term Frequency List and Definition Words Overlap for Mars

The quality of the term frequency list can be measured by the number of words found in both and through a ratio that takes into consideration the solution possibility space. The ratio takes the number of words found in the term frequency list and definition words (whether it be OW, SW, AW, or ASW) divided by the total number of definition words. The ratio better reflects how many words are found because the more words in the definition, the greater the possibility that those words are also in the term frequency list. The ratio is given by:

$$\text{ratio} = \frac{\text{Number of Words Found in Term Frequency List and Definition Word List}}{\text{Total Number of Definition Words}} \quad (4-4)$$

Table 4-18. Definition Evaluation Results for Mars and Chandelier

Word Type	Value Description	Mars	Chandelier
	# of Definitions	6	1
Original Word (OW)	# of OW in term frequency list	4	0
	Total Words in OW	17	6
	OW Ratio	23.5%	0.0%
Synonym Word (SW)	# of SW in term frequency list	3	1
	Total Words in SW	276	116
	SW Ratio	1.1%	0.8%
All Words (AW)	# of AW in term frequency list	7	0
	Total Words in AW	34	6
	AW Ratio	20.6%	0.0%

All Synonym Words (ASW)	# of ASW in term frequency list	6	1
	Total Words in ASW	634	116
	ASW Ratio	0.9%	0.8%

These metrics help quantify why chandelier’s term frequency list performs poorly compared to mars rather than simply having a human give an arbitrary, subjective score. However, there were some portions of the remove class’s context that this portion lacks, such as similar words that do not appear in the definition, which can be better evaluated in the AR method.

4.5.2 AR Evaluation Method

Though the removed class’s definition provides some context as to its meaning, it lacks more general terms or similar objects better captured by its AR words. These are not always synonyms, but usually, other similar terms that are about something are often associated with the target word. This method, described in-depth in Chapter 4.2.4.2, looks at the most similar AR words associated with the removed class (or its WV representation as denoted in Table 9-1). The primary AR words are the top p words with the highest cosine similar to the removed class, and the secondary AR words are the top s words with the highest cosine similarity to the primary AR words. For this evaluation, $p=20$ words and $s=10$ words/(primary word), for a maximum of $(ps) = 20*10 = 200$ secondary AR words. As mentioned earlier the actual number is usually less after duplicate and word variations are removed.

To better demonstrate how this evaluation is useful, take Mars’ definition, which does not include the word “space” or “Jupiter;” however, many would agree that these words point us in the direction of Mars since it is located in space and is a planet like

Jupiter. AR can draw on other, broader concepts that are still related to the removed class, despite not potentially appearing in its definition nor its synonyms. Returning to the good vs. bad example, the primary (blue), secondary (green), and duplicate (grey) AR words are shown for Mars and chandelier in Figure 4-8 and Figure 4-9, respectively

Input: mars		Secondary AR Words									
Primary AR Words		1	2	3	4	5	6	7	8	9	10
1	martian	earth	lunar	meteorite	lander	planet	spacecraft	extraterrestrial	moon	microbe	interplanetary
2	spacecraft	spaceship	orbit	soyuz	orbiter	nasa	astronaut	shuttle	space	satellite	unman
3	lander	orbiter	spacecraft	pathfinder	beagle	martian	lunar	rover	nasa	polar	planetary
4	moon	lunar	earth	ki	spacecraft	planet	orbit	sun	jupiter	saturn	apollo
5	orbiter	spacecraft	lander	shuttle	i	orbit	nasa	atlantis	lunar	astronaut	endeavour
6	earth	planet	orbit	moon	spacecraft	martian	universe	space	planetary	surface	gravity
7	planet	earth	planet	orbit	jupiter	pluto	universe	asteroid	extrasolar	moon	dwarf
8	pathfinder	lander	rover	spacecraft	sojourner	orbiter	spaceship	jpl	martian	pathfinder	odyssey
9	orbit	orbit	spacecraft	orbital	satellite	geostationary	earth	geosynchronous	astronaut	planet	space
10	lunar	moon	spacecraft	martian	orbiter	orbit	lander	eclipse	earth	apollo	solar
11	jupiter	moon	planet	uranus	saturn	neptune	comet	io	orbit	ganymede	asteroid
12	nasa	shuttle	astronaut	spacecraft	space	endeavour	i	atlantis	hubble	aeronautics	orbit
13	rover	jaguar	pathfinder	lander	spacecraft	sojourner	bmw	lr3	carmaker	freelander	volvo
14	astronaut	astronaut	cosmonaut	i	nasa	foale	spacewalk	shuttle	mir	spacecraft	thagard
15	comet	asteroid	halley	comet	spacecraft	meteor	jupiter	tempel	meteorite	orbit	bopp
16	spaceship	spacecraft	soyuz	spaceship	atlantis	shenzhou	astronaut	orbit	unman	capsule	man
17	planetary	planet	earth	orbit	astronomer	atmosphere	asteroid	astronomical	celestial	spacecraft	astronomy
18	galileo	galilei	spacecraft	cassini	copernicus	huygens	ganymede	rosetta	satellite	flyby	telescope
19	mission	mission	peacekeeping	u.n.	nasa	spacecraft	force	un	mandate	space	missionary
20	robotic	robot	spacecraft	unman	submersible	humanoid	robotics	astronaut	underwater	cybernetic	orbiter

Figure 4-8. Primary and Second AR Words for Mars

Input: chandelier		Secondary AR Words									
Primary AR Words		1	2	3	4	5	6	7	8	9	10
1	earring	earring	pendant	necklace	ponytail	choker	brooch	blouse	headband	hairdo	rhinestone
2	sconce	sconce	revetment	elissa	66.21	gopura	jehoash	afp03	shifra	kalonymus	wolfsberg
3	pendant	necklace	earring	brooch	pendant	jewel	locket	bracelet	crucifix	ornament	amulet
4	ornate	gild	adorn	ornament	opulent	decorate	elegant	marble	rococo	elaborately	decorative
5	gild	sentimentalize	tarnish	dramatize	kishline	scold	misbehavin	deface	prettify	recalibrate	re-arrange
6	skylight	skylight	atrium	stained-glass	cupola	staircase	glass	stairway	clerestory	arch	plexiglas
7	staircase	stairway	staircase	stairs	stairwell	foyer	hallway	walkway	stair	balcony	doorway
8	fireplace	fireplace	chimney	patio	stove	hearth	mantel	porch	staircase	kitchen	marble
9	dangling	dangle	hanging	protrude	earring	rope	wire	strap	wrist	ribbon	ankle
10	candlestick	petco	foxboro	dalymount	comiskey	49ers	racetrack	ballpark	croke	kingdome	fenway
11	ceiling	ceiling	wall	roof	floor	coffered	above	tile	window	hanging	plaster
12	candelabra	candelabrum	menorah	candlestick	candelabra	candleholders	torch	filigree	pendant	sconce	gold-plate
13	panelling	panelling	wainscoting	panel	mahogany	flooring	carpeting	cabinetry	molding	ceiling	varnish
14	adorn	adorn	decorate	plaster	mural	festoon	poster	statue	grace	carving	facade
15	drapery	drapery	upholstery	drape	embroidery	diaphanous	curtain	swag	wallpaper	multicolored	moulding
16	panelling	panelling	plasterwork	panel	reredos	wainscoting	iconostasis	stonework	cabinetry	fireplace	balustrade
17	bead	beads	necklace	pendant	bead	earring	embroidery	bracelet	bangle	caulk	brooch
18	crucifix	crucifix	altar	rosary	pendant	necklace	altarpiece	adorn	statue	chalice	earring
19	swarovski	rhinestone	crystal	encrust	beading	rhombohedral	bead	sequin	necklace	lace	baccarat
20	glitter	glam	glitz	sparkle	sequin	glamour	glittery	glitter	rhinestone	garish	glamor

Figure 4-9. Primary and Secondary AR Words for Chandelier

An in-depth look at how the primary and secondary AR words were identified for Mars is described below. First, the initial top 23 primary words for Mars are shown in Table 4-19. There are four instances where the initial primary AR words have a different

base word as shown with Words 10, 15, 16, and 17. Only the base word is kept. Since the base word “orbit” appears for Word 10, “orbits,” and Word 16, “orbiting,” the first instance remains, but the second one is removed to prevent duplicates. Words 8 and 15 have the same base word, “planet,” so the second instance will be replaced in addition to “mars” being the top primary AR word. The reason why “Mars” is the top AR word is that the representative class name WV has to be passed through to be compatible with all classes (e.g. a class with two words in its name such as “coffee mug”), but GloVe does not realize that the WV passe through directly matches “Mars.” This means that words 21-23 (“Galileo,” “Mission,” and “Robotic”) will be included in the 20 primary AR words used to determine the secondary AR words.

Table 4-19. Base Word Comparison for Mars' Primary AR Words

Number	Initial Top Primary AR Words	Base Word (if blank same as initial)
1	Mars	
2	Martian	
3	Spacecraft	
4	Lander	
5	Moon	
6	Orbiter	
7	Earth	
8	Planet	
9	Pathfinder	
10	Orbits	Orbit
11	Lunar	
12	Jupiter	
13	Nasa	
14	Rover	
15	Planets	Planet
16	Orbiting	Orbit
17	Astronauts	Astronaut
18	Comet	
19	Spaceship	
20	Planetary	

21	Galileo	
22	Mission	
23	Robotic	

Secondary AR words are the most similar words to the primary AR words. However, like the primary AR words, they are all converted to their base form. As long as the base word is not already a secondary AR word (determined by a previous primary AR word), it is kept in the list. This process is repeated until 10 non-duplicate words are found. However, if a secondary AR base word is a primary AR base word, it is kept as a secondary word, but ignored in the remainder of the analysis. They were not outright excluded, because this led to obscure words being a part of the secondary AR words list. In Appendix D, Figure 10-1 shows all the primary and secondary AR words for Mars, Figure 10-2 identifies any primary AR words that is listed as a secondary AR word (via blue coloring), and Figure 4-8 identifies any duplicate secondary AR words leftover (via gray coloring). The words found in green text in Figure 4-8 are the secondary AR words, which totals 81 unique words. The same process occurred for chandelier, but resulted in 145 secondary AR words which are identified in green in Figure 4-9.

Similar to the definition evaluation method, the term frequency list is compared to the list of primary and secondary AR words to see which word(s) if any appears in both. In Table 4-20, the words that appear in either the AR word list or the term frequency list are shown for Mars below. When compared to Mars' definition words identified earlier in

Table 4-9, only two words captured by this analysis appeared in its OW or AW list (SW and ASW were excluded due to their size). Chandelier’s term frequency list did not have any overlap with its AR word lists.

Table 4-20. Term Frequency and AR Word List Comparison for Mars

Word	AR Words		Definition Words	
	Primary	Secondary	OW	AW
Planet	Y		Y	Y
Moon	Y			
Earth	Y			
Sun		Y	Y	Y
Astronomy		Y		
Satellite		Y		
Atmosphere		Y		

The results for this analysis for both evaluation methods are shown in Table 4-21. Results for more instances can be found in Section 4.6. Similar to the ratio used for the evaluation method, a ratio can be used to describe the results of the secondary AR words analysis since the number varied depending on what the removed class was.

Table 4-21. AR Evaluation Results for Mars and Chandelier

Type of AR Words	Value Description	Mars	Chandelier
Primary AR Words	Number of Primary AR Words Found	3	0
	Total Primary AR Words	20	20
Secondary AR Words	Number of Secondary AR Words Found	4	0
	Total Secondary AR Words	81	145
	Ratio of Secondary AR Words	5.0%	0.0%

4.5.3 Human Factors Evaluation Method

A subjective human factors (HF) evaluation method was implemented to compare how well the quantitative metrics correspond with a human’s rating. This is considered separate from the Evaluation (red) portion of the IRTARA framework described in Figure

4-1. The top 21 words from the term frequency list for each removed class was presented to 25 collegiate students and the author (for a total of 26) where they were asked whether each word “by itself, in combination of another listed word or its characteristics, describe or could be associated with [the removed class]?” These scores were aggregated by determining whether at least 75% (20+) or 50% (13+) of the respondents believed the word was associated with the removed class. Afterward, they assigned an overall 1-5 ranking based on how well the identified words describe the removed class. An average was computed from all of the scores.

Returning to our good and bad examples, both HF evaluations are shown for the top 21 words in the term frequency list as well as the definition and AR evaluation results for chandelier and Mars in

Table 4-23 and Table 4-22, respectively. For Mars, the method showed at least 75% (therefore also at least 50%) of respondents felt that 4 words (planet, sun, mythology, and mass) described the removed class well. For chandelier, the HF method showed at least 75% of respondents felt that 2 words (light and look) and 50% of respondents felt that 5 words (observe, light, look, eyes, and building) described the removed class well. When comparing these results to our quantitative methods, we see more overlap with the words in the high-performing trial, Mars, compared to chandelier.

Table 4-22. HF Evaluation Comparison for Mars

	Evaluation Method		
	Definition	AR	HF

Rank	Word	O.W.	S.W.	A.W.	A.S.W.	Primary	Secondary	# of Yes's	=>75% Yes	=>50% Yes
1	Brain		X					4		
2	Skull							3		
3	Nervous							6		
4	Ability							3		
5	Planet	X				X		26	X	X
6	Part							9		
7	Hit							3		
8	Certain							4		
9	Ones							1		
10	Sun	X					X	23	X	X
11	Ball			X				1		
12	Mythology							20	X	X
13	Meat							2		
14	Central							6		
15	Feelings							5		
16	Exceptional							12		
17	Mass							23	X	X
18	Cord							2		
19	Spinal							1		
20	Head							3		
21	Continuous							7		

Table 4-23. HF Evaluation Comparison for Chandelier

		Evaluation Method								
		Definition				AR		HF		
Rank	Word	OW	SW	AW	ASW	Primary	Secondary	# of Yes'	=>75% Yes	=>50% Yes
1	Small		X					8		

2	Observe							18		X
3	Person							12		
4	Determine							3		
5	Light							26	X	X
6	Ones							5		
7	Watch							10		
8	Sound							11		
9	Base							12		
10	Look							22	X	X
11	Members							2		
12	Travel							5		
13	Quickly							2		
14	Period							8		
15	Move							11		
16	Eyes							17		X
17	Building							18		X
18	Making							9		
19	Plural							4		
20	Living							10		
21	Body							4		

When looking at overall rankings from the HF methods from the results shown in Table 4-24. The group was less impressed with the overall results for Mars given that its overall score was effectively the same as chandelier. However, at least 75% of respondents identified more words in the term frequency accurate describes Mars (4 words) more so than chandelier (2 words). It is also worth noting that the HF analysis only looked at the top 21 words rather than all 100 words within the term frequency list as the other two evaluation methods consider.

Table 4-24. Overall Score from HF Evaluation Ranking for Chandelier and Mars

	Mars	Chandelier
HF Evaluation Average Overall Score ± Standard Deviation	2.85 ± 1.14	2.52 ± 0.95

4.6 RESULTS

As shown in the previous sections, the methodology was applied to several different classes taking the place of the “removed class.” The term frequency lists from these additional trials are found in Appendix E except for Mars (see Table 10-1), chandelier (see Table 10-1), and rainbow (see Table 11-1). At a high level, these results are shown in

Table 4-25. Split between original word (OW), synonym word (SW), all words (AW), and all synonym words (ASW) of the definition(s), the table displays the number of words found in the given category, the total number of words possible to be found, and then the ratio of the number found divided by the total. In instances where the all words and all synonym words are blacked-out means that the removed class only had one definition, it's the true one.

Table 4-25. Definition Evaluation Results for Select Classes

Removed Class	Original Word			Synonym Word			All Words			All Synonym Words		
	# Found	Total	Ratio (%)	# Found	Total	Ratio (%)	# Found	Total	Ratio (%)	# Found	Total	Ratio (%)
Ak-47	0	4	0	0	42	0						
Cactus	0	11	0	2	196	1.0						
Chandelier	0	6	0	1	116	0.9						
Fireworks	1	8	12.5	2	250	0.8						
Floppy Disk	0	13	0	0	20	0	1	15	6.7	0	244	0
Frog	1	10	10	1	84	3	3	20	15	2	210	1.0
Galaxy	0	9	0	1	239	0.4						
Iguanas	5	16	31.3	2	212	0.9						
Mars	4	17	23.5	3	275	1.1	7	34	20.6	6	634	0.9
Penguin	0	12	0	3	147	2.0						
People	0	3	0	2	61	3.3	3	18	16.7	6	321	1.9
Rainbow	1	9	11.1	5	222	2.3	1	11	9.1	5	234	2.1
Sheet Music	0	13	0	1	146	0.7	4	58	6.9	9	784	1.1
Skyscraper	2	4	50.0	1	58	1.7						
Swiss Army Knife	1	6	16.7	1	150	0.7						
T-shirt	2	5	40.0	0	80	0	2	6	33.3	0	82	0
Waterfall	1	4	25.0	0	80	0						

Looking at the AR evaluation for the selected classes,

Table 4-26 was created. The total number of primary words is always 20, so the raw number found can be compared directly one-to-another. Since the number of secondary AR words varies, a ratio is computed to better compare the results between classes.

Table 4-26. AR Evaluation Results for Select Classes

Removed Class	# of Primary Words Found	# of Secondary Words Found	Total # of Secondary Words	Secondary Words Ratio (%)
Ak-47	0	0	60	0
Cactus	0	0	126	0
Chandelier	0	0	113	0
Fireworks	0	2	116	1.7
Floppy Disk	0	1	116	0.9
Frog	0	3	126	2.4
Galaxy	3	5	133	3.8
Iguanas	0	3	142	2.1
Mars	3	4	77	5.2
Penguin	0	0	132	0
People	1	2	90	2.2
Rainbow	0	3	165	1.8
Sheet Music	0	1	102	1.0
Skyscraper	3	3	112	2.7
Swiss Army Knife	0	0	145	0
T-shirt	1	1	67	1.5
Waterfall	0	1	129	0.8

The HF analysis result was compiled into

Table 4-27 for both the individual and group analysis methods. For the most part, there are negligible differences between the classes for this evaluation method. One class that stood out as exceptional was the “galaxy” class with a majority of the words being relevant and an average overall score of 4.5/5.

Table 4-27. HF Evaluation Results for Select Classes

Removed Class	$\geq 75\%$ Agreed on Relevancy	$\geq 50\%$ Agreed on Relevancy	Average Overall Score (\pm Standard Deviation)
Ak-47	2	7	2.04 ± 0.77
Cactus	6	10	2.92 ± 0.8
Chandelier	2	5	2.52 ± 0.95
Fireworks	6	14	3.62 ± 0.85
Floppy Disk	3	9	2.81 ± 1.02
Frog	5	11	2.92 ± 0.9
Galaxy	15	18	4.5 ± 0.81
Iguanas	8	13	3.15 ± 0.89
Mars	4	4	2.58 ± 1.14
Penguin	6	7	2.69 ± 1.01
People	7	9	2.69 ± 1.05
Rainbow	6	9	3.42 ± 0.81
Sheet Music	1	7	2.08 ± 1.06
Skyscraper	6	8	3.35 ± 0.89
Swiss Army Knife	5	9	2.65 ± 0.85
T-shirt	9	12	3.27 ± 1.12
Waterfall	3	5	2.27 ± 0.96

More results that were not generated with sufficient time to undergo the HF evaluation method are shown in Chapter 0. Some of these classes were repeated from the initial study to show how variation may occur between iterations. Since the same HF evaluation could not be completed, a modified version conducted solely by the author was included in

Table 13-3. An overall score of 1-5 was assigned based on the quality of the first 20 words of the term frequency list and the total number of words found within the list that the author considered to accurately describe or characterize the removed class.

5 DISCUSSIONS

The three evaluation methods described in Chapter 4.6 produced the results shown through sections of this chapter. Further shown later in Chapter 5.4, there is a sense of commonalities among the classes that perform at the extremes; however, it is difficult to see consistency among those that fall in between. Thus, this chapter aims to better quantify and interpret the results and the IRTARA framework as a whole.

5.1 DEFINITION EVALUATION

When ranking all the removed classes by their ratios, shown in percentage-form, for the various metrics,

Table 5-1 was constructed. Using the ratios from the definition evaluations for OW, SW, AW, and ASW, an average ratio was computed, from which the rankings were based. Microsoft Excel's `rank.avg()` function was used, which breaks ties by using the average; however, it was not needed for this portion. The ratios across OW, SW, AW, and ASW ranged from 0% to 50% in regards to the number of definition words that appeared in the term frequency list. The average ratio ranged from 0% to 25.86%, which was used to determine the rankings. Based on the rankings, the top 5 classes are skyscraper, t-shirt, iguana, waterfall, and mars, respectively. This infers that the term frequency lists corresponding to these categories can pick up the contextual elements/components of the removed class.

Table 5-1. Definition Evaluation Results Rankings

Removed Class	OW Ratio	SW Ratio	AW Ratio	ASW Ratio	Average Ratio	Overall Ratio Ranking
Ak-47	0.0%	0.0%	0.0%	0.0%	0.00%	17
Cactus	0.0%	1.0%	0.0%	1.0%	0.51%	14
Chandelier	0.0%	0.9%	0.0%	0.9%	0.43%	15
Fireworks	12.5%	0.8%	12.5%	0.8%	6.65%	8
Floppy Disk	0.0%	0.0%	6.7%	0.0%	1.67%	12
Frog	10.0%	1.2%	15.0%	1.0%	6.79%	7
Galaxy	0.0%	0.4%	0.0%	0.4%	0.21%	16
Iguanas	31.3%	0.9%	31.3%	0.9%	16.10%	3
Mars	23.5%	1.1%	20.6%	0.9%	11.54%	5
Penguin	0.0%	2.0%	0.0%	2.0%	1.02%	13
People	0.0%	3.3%	16.7%	1.9%	5.45%	10
Rainbow	11.1%	2.3%	9.1%	2.1%	6.15%	9
Sheet Music	0.0%	0.7%	6.9%	1.1%	2.18%	11
Skyscraper	50.0%	1.7%	50.0%	1.7%	25.86%	1
Swiss Army Knife	16.7%	0.7%	16.7%	0.7%	8.67%	6
T-shirt	40.0%	0.0%	33.3%	0.0%	18.33%	2
Waterfall	25.0%	0.0%	25.0%	0.0%	12.50%	4

5.2 ANALOGICAL REASONING EVALUATION

The classes were ranked based on their raw primary words (since this was consistently out of 20 for all classes) and their secondary words ratio, displayed as percentages, in Table 5-2. Since the results were more condensed for the AR evaluation, there are several ties in the lower portion of the rankings, which were again, generated by the same `rank.avg()` function used in Chapter 5.1. Opposed to the previous definition evaluation discussion in Chapter 5.1, the top five classes based on the AR evaluation are Mars, galaxy, skyscraper, people, and t-shirt. Since the AR evaluation is comparing the term frequency list to a different set of words, i.e. the primary and secondary AR words, we expect some results to be different from the definition evaluation. High performance, which is a high ratio average, in this section suggests that rather than the term frequency

list picking up on the characteristics/qualities of the removed class as seen with the definition evaluation, the list better represents similar/associated concepts to the removed class.

Table 5-2. AR Evaluation Results Rankings

Removed Class	Primary Ranking	Secondary Ratio Ranking	Average Ratio	Overall Ranking
Ak-47	0.0%	0.0%	0.0%	15
Cactus	0.0%	0.0%	0.0%	15
Chandelier	0.0%	0.0%	0.0%	15
Fireworks	0.0%	1.7%	0.9%	9
Floppy Disk	0.0%	0.9%	0.4%	11
Frog	0.0%	2.4%	1.2%	6
Galaxy	15.0%	3.8%	9.4%	2
Iguanas	0.0%	2.1%	1.1%	7
Mars	15.0%	5.2%	10.1%	1
Penguin	0.0%	0.0%	0.0%	15
People	5.0%	2.2%	3.6%	4
Rainbow	0.0%	1.8%	0.9%	8
Sheet Music	0.0%	1.0%	0.5%	10
Skyscraper	15.0%	2.7%	8.8%	3
Swiss Army Knife	0.0%	0.0%	0.0%	15
T-shirt	5.0%	1.5%	3.2%	5
Waterfall	0.0%	0.8%	0.4%	12

5.3 HUMAN FACTORS EVALUATION

The human factors (HR) evaluation was presented as a class assignment to 25 students enrolled at Wright State University's Introduction to Human Factors Engineering class. The class consisted of a mix of graduate and undergraduate students. Specifics regarding this evaluation method can be found in Chapter 4.5.3. However, the results reflect 26 respondents with the author's results posing as the additional respondent. The HF results, specifically discussed in Chapter 4.5.3, were quite different from the previous two methods as seen in

Table 5-1 and Table 5-2. As mentioned previously in Chapter 4.5.3, this evaluation aimed to see how the definition and AR evaluations aligned with a human’s evaluation. The HF evaluation included three metrics: (1) average overall score, (2) number of words at least 75% of respondents thought were “good” (dubbed “# of Good Words, 75%” in

Table 5-3), and (3) number of words at least 50% of respondents thought were “good”
(dubbed “# of Good Words, 50%” in

Table 5-3). Since the average score had a range of 1-5 and the number of good words (for both 75% and 50%) could range from 1-21, a ranking was calculated for each using the `rank.avg()` function, which were then average in the “Average Ranking” column in

Table 5-3.

The overall ranking was only based on the average ranking score. Rankings that include decimals are due to a multi-way tie. The top five also shift more with galaxy, t-shirt, rainbow, skyscraper, and iguanas appearing at the top simply due to being evaluated in another way. This evaluation method was designed to compare how well the previous two methods align with a human's analysis rather than an overall test of how well the term frequency list describes the removed class as a whole. This is discussed more in Section 5.4.

Table 5-3. HF Evaluation Results Rankings

Removed Class	Metric Rankings			Average Ranking	Overall Ranking
	Average Overall Score Ranking	# of Good Words, 75%	# of Good Words, 50%		
Ak-47	17	15.5	14.5	15.67	17
Cactus	7.5	10.5	4.5	7.5	7.5
Chandelier	13	15.5	14.5	14.3	13.5
Fireworks	2	5	11	6	6
Floppy Disk	9	13	11	11	11
Frog	7.5	8.5	7	7.67	9
Galaxy	1	1	1	1	1
Iguanas	6	8.5	2.5	5.67	5
Mars	14	13	16	14.33	13.5
Penguin	10.5	5	11	8.83	10
People	10.5	5	7	7.5	7.5
Rainbow	3	5	4.5	4.17	3
Sheet Music	16	17	11	14.67	15
Skyscraper	4.5	5	7	5.5	4
Swiss Army Knife	12	10.5	11	11.17	12
T-shirt	4.5	2	2.5	3	2
Waterfall	15	13	17	15	16

5.4 OVERALL SUMMARY

The rankings from the three evaluation methods are compared to one another in Table 5-4. These were taken from the analysis shown in

Table 5-1, Table 5-2, and

Table 5-3 earlier in this chapter. The average ranking is the average of the “Evaluation Methods Rankings” columns in Table 5-4. The average ranking, which was calculated using the `rank.avg()` function, was used to determine the overall ranking.

Table 5-4. Overall Rankings for Evaluation Methods

Removed Class	Evaluation Methods Rankings			Average Ranking	Overall Ranking
	Definition	Analogical Reasoning	Human Factors		
Ak-47	17	15	17	16.33	17
Cactus	14	15	7.5	12.17	14
Chandelier	15	15	13.5	14.5	16
Fireworks	8	9	6	7.67	9
Floppy Disk	12	11	11	11.33	12
Frog	7	6	9	7.33	8
Galaxy	16	2	1	6.33	4
Iguanas	3	7	5	5	3
Mars	5	1	13.5	6.5	5
Penguin	13	15	10	12.67	15
People	10	4	7.5	7.17	7
Rainbow	9	8	3	6.67	6
Sheet Music	11	10	15	12	13
Skyscraper	1	3	4	2.67	1
Swiss Army Knife	6	15	12	11	11
T-shirt	2	5	2	3	2
Waterfall	4	12	16	10.67	10

As stated previously in Chapter 4.5, the definition and AR evaluation methods are two ways of measuring the same data, rather than being compared to one another; however, ideally, these rankings would be close to the HF evaluation rankings. However, looking at

the overall rankings, there appears to be more consistency with the top and bottom classes in terms of their scores across all three evaluation rankings. For ease of viewing Table 5-4 has been re-arranged in order of highest overall rank (1) to lower (17) in Table 5-5 with two additional columns showing the difference between the Definition and HF rankings and the difference between the AR and HF rankings.

Table 5-5. Overall Rankings for Evaluation Methods Ordered

Overall Ranking	Removed Class	Evaluation Methods Rankings		
		Definition	Analogical Reasoning	Human Factors
1	Skyscraper	1	3	4
2	T-shirt	2	5	2
3	Iguanas	3	7	5
4	Galaxy	16	2	1
5	Mars	5	1	13.5
6	Rainbow	9	8	3
7	People	10	4	7.5
8	Frog	7	6	9
9	Fireworks	8	9	6
10	Waterfall	4	12	16
11	Swiss Army Knife	6	15	12
12	Floppy Disk	12	11	11
13	Sheet Music	11	10	15
14	Cactus	14	15	7.5
15	Penguin	13	15	10
16	Chandelier	15	15	13.5
17	Ak-47	17	15	17

The IRTARA and its evaluation methods, definition and AR, performed relatively consistently when compared to one another in Table 5-5. The largest difference between the two rankings is 14 for Galaxy (which is likely an outlier) followed by 9 for Swiss Army

Knife. However, these evaluation methods did not always align with the results from the human factors study as shown in Figure 5-1; the difference between the rankings for the definition and HF valuations are black circles, and the same difference between the AR and HF evaluations are grey triangles. Some of these discrepancies may be due to the relatively small values being dealt with in the definition and AR evaluation results are shown above in Chapters 5.1 and 5.2.

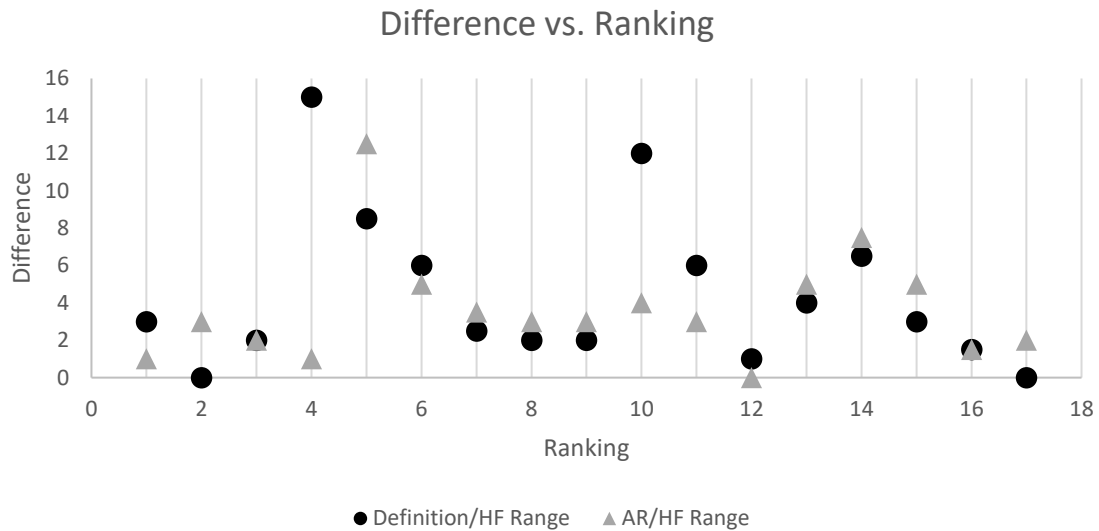


Figure 5-1. Differences vs. Rankings for Evaluation Results

Ultimately, IRTARA provides a repeatable process with automated measurable results for an unknown-unknowns image classification process. The IRTARA is capable of producing the context for an unfamiliar image, without the need for human intervention through analogical reasoning. Though the quality of results for IRTARA varies between classes, there are valuable insights presented in the top-performing classes.

5.5 PLACING IRTARA IN ANALOGICAL REASONING LITERATURE

Unlike the systematical comparison in Chapter 3, the image-based analysis does not compare images such that the typical $A:B::C:D$ comparison is made. Instead of being labeled as AR, it is rather an application utilizing AR to show the benefits of utilizing such semantic-deriving methods to improve the results of predicting unknown unknown scenarios in an image recognition context. This application is separate from simply text-based augmentation because it uses the word vectors produced by the AR algorithm, GloVe, to yield the results found in the term frequency list. The benefits of utilizing AR in addition to the typically (deep) CNN include reduction in data required and reasonable training time vs quality of results trade-off. CNNs, in general, but more specifically the deep ones often require a significant amount of training data to be able to accurately predict the outcome. Considering the amount of training data, this process often takes a long time such as hours or even days depending on the computation power of the machine. IRTARA can utilize a relatively small data set (minimum total number of classes is 80) on a mediocre CNN (average 22% accuracy on all 256 classes), which greatly assists with the time it takes to run in comparison to deep-CNN-based image recognition programs. Though IRTARA does not outright produce the word corresponding to the image, the term frequency list does a reasonable job attempting to describe the removed class given the reduction in training time and resources typically needed.

The practical application of the IRTARA is for image recognition software that may interact with “unknown” objects. This can be used in image caption generation or labeling unfamiliar objects in sensors such as in self-driving cars. Ultimately it would be beneficial to be able to use IRTARA in an automated decision-making process to assist

and further evolve AI. IRTARA may be able to produce results quicker than current state-of-the-art algorithms through a decrease in training data needed and processing time to produce results.

6 CONCLUSIONS AND FUTURE WORK

Identifying, interpreting, and understanding unexpected queries are a challenge to artificial intelligence (AI) algorithms. Many of these algorithms are taught to recognize specific objects and items; however, when a new concept is introduced to them, they perform poorly if they can generate results at all. For this reason, research regarding how to handle “unknown unknowns” is of great interest to the AI community. This research explores using analogical reasoning (AR) to better comprehend how to handle unfamiliar concepts, first by reviewing AR in Chapters 2 and 3, then by investigating AR within image classification in Chapters 4 and 5. To achieve the original research objectives, a review of AR algorithms was conducted and tested and then, a new algorithm, Image Recognition Through Analogical Reasoning (IRTARA), was then proposed in Chapter 4 which integrated an AR algorithm with a convolutional neural network (CNN) to extract data from an unknown-unknowns problem.

6.1 PROBLEMS WITH CURRENT AI METHODS

As discussed throughout Chapter 1, current AI is relatively weak and struggles to produce quality results outside of its original domain. When given an “unknown unknown” entity such as in image recognition, AI struggles to understand and characterize the entity because it is limited to what it has been trained on. One option to solve this issue is to retrain the AI to be able to identify the aforementioned “unknown unknown,” but this is not a feasible solution considering the amount of data and process time needed. In regards to the issue with data, the AI would need to train on images of every single object it may come into contact with. Objects are also not always observed consistently, e.g., the viewing

angle may cause distortion or there may be a glare if looking at a reflective surface. To collect and train on each potential object and variations of how they are perceived, would require a significant amount of data. With the problem of data also comes the issue of time. As the amount of data grows, the amount of time needed to train the AI and then, use it also grows based on the available computational resources. AI used to aid decision-makers often needs to be quick or else the window allotted to make the decision risks closing before any additional intelligence is gathered. With these two motivations, the idea of using AR to assist with AI was suggested and recommended.

6.2 CONTRIBUTIONS TO RESEARCH

This research was divided into two portions. The first portion looked at text-based AR, arguably the predominant field of AR research. The second portion was focused on how to intertwine current AR algorithms within visual data, specifically framing an image recognition scenario.

6.2.1 Text-based Analogical Reasoning Evaluation

First, a general understanding of AR research was needed, which was achieved through an in-depth literature review of popular algorithms discussed in Chapter 2. The algorithms gathered were further classified by the AI paradigm they were best associated with (symbolist, connectionist, or dynamicist) in a chronological taxonomy, which is a new contribution to AR research.

From this, six algorithms were selected for an apples-to-apples comparison detailed in Section 3 based on their AL paradigm and recency. The six AR algorithms selected were: Distributed Representation Analogy Mapper (DRAMA), Bayesian Analogy with

Relational Transformations (BART), Word2Vec, Global Vectors (GloVe), 3 Cosine Average (3CosAvg), and Linear Regression Cosine (LRCos). To provide a fair assessment of the selected AR algorithms, the modified Sternberg-Nigro dataset was used, consistent with Morrison et al.'s study (2004). This dataset consists of 197 word-based analogies from five different relationships (synonym, antonym, category, functional, and linear ordering) in the form, $A:B::C:[D \text{ or } D']$. Each algorithm produced a similarity score for the word pairs: $A:B$, $C:D$, and $C:D'$; next a similarity ratio was computed between the similarity scores for $A:B$ and $C:D$ as well as $A:B$ and $C:D'$. The difference between the two similarity ratios and the "ideal" similarity ratio of one was computed and compared to determine whether the algorithm selected D or D' to complete the given analogy. The algorithm's results were evaluated by two developed metrics, correctness, and a goodness metric.

The correctness metric calculated the number of times the algorithm selected the correct answer over the "distractor" option. Overall, DRAMA came out on top with a 78.7% accuracy, with the highest individual correctness scores in all the categories except for functional and antonym categories, which BART 2.0 tied and outperformed, respectively. The goodness metric looked at the difference between the algorithm's similarity ratio and an "ideal" analogy similarity ratio of 1, for the correct pairing of $A:B::C:D$. A similarity ratio of 1 indicates that the semantic similarity between $A:B$ aligns exactly with $C:D$, which is expected since $A:B::C:D$ forms a proper analogy. However, the algorithms did not always yield this result, so this metric looks at how closely the algorithm predicted an "ideal" analogy for those in the data set. This difference was averaged across all the attempted analogies for each algorithm, which showed LRCos followed closely by 3CosAvg on top with DRAMA and GloVe at the bottom, respectively.

In conclusion, there is not a “one-size-fits-all” algorithm for all types of AR problems. The best algorithm will vary based on the task at hand the given relationship within the analogy. This portion helped identify psychologically and natural language processing (NLP)-produced AR algorithms and compare them with consistent data and metrics, which had yet to appear in the literature before (Combs, Bihl, Ganapathy, & Staples, 2022).

6.2.2 Image-based Analogical Reasoning Algorithm & Evaluation

With a better understanding of the AR algorithms’ advantages and limitations, the next step was to intertwine AR with an unknown unknown scenario, which was posed as an image classification problem. The ultimate goal was to extract previously unknown information about a particular item/object/etc. via AR.

The primary framework is built around a convolutional neural network (CNN), an artificial neural network architecture used primarily for analyzing images. The IRTARA CNN was trained on the Caltech-256 data set (Griffin, Holub, & Perona, 2007), which had 256 classes and yielded an average accuracy of 22.1%. To modify this as an unknown unknowns situation, one of the classes was excluded from the training portion (i.e. the “removed” class) and the IRTARA CNN was trained on the remaining 255 classes. The images belonging to the removed class were evaluated by the IRTARA CNN, which yielded a confidence percentage regarding how likely a given image belonged to each of the classes. At this point, AR was integrated into the framework. GloVe was selected as the AR algorithm to integrate with this framework due to its compatibility with the type of AR problem, feasible time implementation, and ability to represent singular words (Levy & Goldberg, 2014). GloVe is used to construct a representative word vector (WV) (called

the “unknown” WV) for each image that ideally represents the removed class and maps it to the closest WVs that correspond to English words found in GloVe’s corpus of known words. The top five AR words for each image were identified and had their definitions pulled from PyDictionary (geekpradd, 2020). A running total of the top 100 words, called the term frequency list, with semantic meaning (excluded particles, some prepositions, and select other words) found in the definition(s) were compiled, which is the final result of the new framework. The term frequency list can be thought of as a list of words for context regarding the removed class, some of which aligned and some of which did not as measured by our three evaluation methods.

The two automated methods were the definition and AR evaluations. The definition method compared the words found in the term frequency list to those found in the removed class’s definition(s). The AR method compared the words found in the term frequency list to the removed class’s primary and secondary AR words. Primary AR words were the top 20 words that were most similar to the removed class; whereas, secondary AR words were the top 10 words that were most similar to the primary AR words, with duplicates and word variations removed. These two methods provided an automated method of calculating how “good” the term frequency list was at describing the removed class. Ideally, these results would mimic a human, so a human factors (HF) evaluation looking at the top 21 words was also deployed to compare to the automated scoring methods. This method asked respondents to note whether a given word (in the term frequency list) “by itself, in combination [with] another listed word, or its characteristics describe or could be associated with [the removed class]?” This data was transformed into two metrics, which asked for a given work in the term frequency list whether at least 50% and/or at least 75%

of respondents said “yes.” For the same remove classes term frequency list, looking at the words the respondent said “yes” to in the previous question, they were asked to give a ranking between 1-5 about how well the identified word(s) describe the removed class with 1 being poor and 5 being excellent. An average across all 26 respondents was taken for the overall quality rating. The HF method was created to see how well the automated methods, the definition, and AR evaluations, align with a human’s perspective. As shown in Chapter 5, doing well in one evaluation method did not necessarily constitute doing well in the remaining ones; however, consistently high or low scores were observed at the extreme ends. However, in-between the extremes, the classes vary in their rankings compared to one another across all three metrics.

In conclusion of this section, a repeatable and measurable method for applying AR to an unknown unknowns scenario was then developed and presented in this section. A framework, IRTARA, was created and deployed to evaluate “unknown unknowns” in image recognition. Three assessments were developed to evaluate the quality of the term frequency lists produced by IRTARA, the definition, AR, and human factors evaluation methods. These methods showed consistent results for removed classes that performed exceptionally well or poorly, but ambiguity for those in-between. This contribution successfully mimics an “unknown unknown” scenario that can be quantitatively evaluated. In addition to the promising results, there are several directions future work could explore.

6.2.3 Future Work

Looking at the current framework, several aspects could yield better improvements. First, an in-depth analysis solely dedicated to the image recognition algorithm could prove to be useful, so initial image classification is higher. Second, the use of a different

dictionary from which the definitions were extracted could affect the results since there seems to be extraneous noise within the definitions found in PyDictionary (e.g. such as the listing of historical figures for the definition of “white”). Third, the use of a different vector space model (VSMs) has the potential to improve results as well based on its representations of the words. Specifically of interest would be a VSM or AR algorithm that can understand the context and separate homophones (such as “fly” the insect vs “fly” the verb). Fourthly, perhaps the easy suggestion to implement would be using the cosine similarity between the unknown WV and the words generated to create a penalty system at various stages in the process that would theoretically point us closer to the removed class. Finally, the ability to test IRTARA’s results on a true unknown scenario, meaning the removed class is never revealed, is needed. This requires new metrics that allow the results to be compared to no ground truth.

Outside of IRTARA, in a new successor framework, several areas could be explored. By using a different CNN that can detect multiple objects in a given image, it would be more similar to real-world images, such as that observed by a self-driving car. With the ability for multi-object recognition, image context can be generated, which has the potential to improve results. In conjunction with multi-objective recognition or a separate venture, a focus on zero- or few-shot learning could also be another direction. The benefit of this is the reduction in the information needed, and hopefully not at the cost of accuracy. This direction also poses the question of when to classify an image as a known class compared to deciding it belongs in its separate class. However, it further builds upon being able to feature extraction such as learning what a head or a leg in general looks like

for a variety of images. Overall, IRTARA shows the benefits of AR-integrated AI and paves the way for future research fusing the two.

7 References

- Bailer-Jones, D. M. (2002). Models, metaphors, and analogies. In P. Machamer, & M. Silberstein, *The Blackwell guide to the philosophy of science* (pp. 108-127). Oxford: Blackwell Publishers Ltd.
- Ball, J. E., Anderson, D. T., & Chan, C. S. (2017). Comprehensive survey of deep learning in remote sensing: theories, tools, and challenges for the community. *Journal of Applied Remote Sensing*, 11(4), 042609-1-042609-54.
- Bihl, T. J., & Bauer Jr., K. W. (2017). Statistical analysis of high-level features from State of the Union addresses. *International Journal of Information Systems and Social Change (IJISSC)*, 8(2), 50-73.
- Bihl, T. J., Young, W. A., & Frimel, S. (2022). *Artificial neural networks and data science*. white paper.
- Bihl, T. J., Young, W. A., & Weckman, G. R. (2018). Artificial neural networks and their applications in business. In *Encyclopedia of Information Science and Technology*. Khosrow-Pour, Mehdi.
- Bihl, T., & Talbert, M. (2020). Analytics for autonomous C4ISR within e-government: A research agenda. *Proceedings of the 53rd Hawaii International Conference on System Sciences* (pp. 2218-2227). Wailea: University of Hawaii at Manoa.
- Bird, S., Klein, E., & Loper, E. (2009). *Natural Language Processing with Python*. O'Reilly Media Inc. Retrieved from <http://nltk.org/book>
- Birhane, A., & Prabhu, V. U. (2021). Large image datasets: A pyrrhic win for computer vision? *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)* (pp. 1537-1547). IEEE.
- Bolland, S. W. (2004). *F AE: The Fluid Analogies Engine A Dynamic Hybrid Model of Perception and Mental Deliberation*. Brisbane, Australia: University of Queensland.
- Burstein, M. H. (1983). A model of learning by incremental analogical reasoning and debugging. *National Conference on Artificial Intelligence*, (pp. 45-48).
- Chalmers, D. J., French, R. M., & Hofstadter, D. R. (1991). *High-level perception, representation, and analogy: A critique of artificial intelligence methodology*. Indiana University. Bloomington: Center for Research on Concepts and Cognition. Retrieved from <http://www.consc.net/papers/highlevel.pdf>

- Chalmers, D. J., French, R. M., & Hofstadter, D. R. (1992). High-level perception, representation, and analogy: A critique of artificial intelligence methodology. *Journal of Experimental & Theoretical Artificial Intelligence*, 4(3), 185-211.
- Chen, D., Lu, H., & Holyoak, K. J. (2017). Generative interferences based on learned relations. *Cognitive Science*, 41(5), 1062-1092.
- Chen, D., Peterson, J. C., & Griffiths, T. L. (2017). Evaluating vector-space models of analogy. *arXiv preprint*.
- Chen, Z., Sanches, R. P., & Campbell, T. (1997). From Beyond to Within Their Grasp: The Rudiments of Analogical Problem Solving in 10- and 13-Month-Olds. *Developmental Psychology*, 33(5), 790-801.
- Clark, A., & Toribio, J. (1994). Doing without representing? *Synthese*, 101(3), 401-431.
- Combs, K., Bihl, T. J., Ganapathy, S., & Staples, D. (2022). Analogical reasoning: An algorithm comparison for natural language processing. *Proceedings of the 55th Hawaii International Conference on System Sciences*. Hawaii International Conference on System Sciences (HICSS).
- Darlow, L. N., Crowley, E. J., Antoniou, A., & Storkey, A. J. (2018). *CINIC-10 Is Not ImageNet or CIFAR-10*. University of Edinburgh, School of Informatics. Edinburgh: Institute for Adaptive and Neural Computation. Retrieved from <https://arxiv.org/pdf/1810.03505.pdf>
- Doumas, L. A., & Hummel, J. E. (2010). A computational account of the development of the generalization of shape information. *Cognitive Science*, 34, 698-712.
- Doumas, L. A., Hummel, J. E., & Sandofer, C. M. (2008). A theory of the discovery and predication of relational concepts. *Psychological Review*, 115(1), 1-43.
- Doumas, L. A., Morrison, R. G., & Richland, L. E. (2009). The development of analogy: Working memory in relational learning and mapping. *Proceedings of the Thirty-First Annual Conference of The Cognitive Science Society*, 34, 1-6.
- Drozd, A., Gladkova, A., & Matsuoka, S. (2016). Word embeddings, analogies, and machine learning: Beyond king - man + woman = queen. *26th International Conference on Computational Linguistics: Technical Papers* (pp. 3519-3530). Osaka: COLING 2016 Organizing Committee.
- Duda, R., Hart, P., & Stork, D. (2000). *Pattern classification*. Hoboken: Wiley.
- Eden, A. H. (2007). Three paradigms of computer science. *Minds & Machines*, 17, 135-167.

- Egmont-Petersen, M., de Ridder, D., & Handels, H. (2002). Image processing with neural networks--a review. *Pattern Recognition*, 35(10), 2279-2301.
- Eliasmith, C. (1997). Computational and dynamical models of mind. *Minds and Machines*, 7, 531-541.
- Eliasmith, C. (2013). *How to build a brain: A neural architecture for biological cognition*. Oxford University Press.
- Eliasmith, C., & Thagard, P. (2001). Integrating structure and meaning: A distributed model of analogical mapping. *Cognitive Science*, 25(2), 245-286.
- Evans, J. S., Newstead, S. E., & Byrne, R. M. (1993). *Human reasoning: The psychology of deduction*. East Sussex: Lawrence Erlbaum Associates Ltd.
- Evans, T. (1964, April). A heuristic program to solve geometric-analogy problems. *Proceedings of the 1964 Spring Joint Computer Conference*, 327-338.
- Falkenhainer, B., & Forbus, K. D. (1989). The structure-mapping engine: Algorithm and examples. *Artificial Intelligence*, 41(1), 1-63.
- Farhadi, A., Hejrati, M., Sadeghi, M. A., Young, P., Rashtchian, C., Hockenmaier, J., & Forsyth, D. (2010). Every picture tells a story: Generating sentences from images. *European Conference on Computer Vision* (pp. 15-29). Berlin: Springer.
- Fei-Fei, L., Fergus, R., & Perona, P. (2004). Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. *2004 Conference on Computer Vision and Pattern Recognition Workshop* (pp. 178-178). Washington DC: IEEE. doi:10.1109/CVPR.2004.383
- Flach, P. A., & Kakas, A. C. (2000). Abductive and inductive reasoning: Background and issues. In *Abduction and induction* (pp. 1-27). Dordrecht: Springer.
- Forbus, K. D., & Lovett, A. (2021). Same/different in visual reasoning. *Behavioral Sciences*, 37, 63-68. doi:10.1016/j.cobeha.2020.09.008
- Forbus, K. D., Ferguson, R. W., Lovett, A., & Gentner, D. (2017). Extending SME to Handle Large-Scale Cognitive Modeling. *Cognitive Science*, 41, 1152-1201.
- Forbus, K. D., Gentner, D., & Law, K. (1995). MAC/FAC: A model of similarity-based retrieval. *Cognitive Science*, 19(2), 141-205.
- Forbus, K. D., Gentner, D., Markman, A. B., & Ferguson, R. W. (1998). Analogy just looks like high level perception: Why a domain-general approach to analogical mapping is right. *Artificial Intelligence*, 10, 231-257.

- French, R. M. (1995). *The subtlety of sameness: A theory and computer model of analogy-making*. Cambridge: MIT Press.
- French, R. M. (2002). The computational modeling of analogy-making. *TRENDS in Cognitive Sciences*, 6(5), 200-205.
- geekpradd. (2020, July 8). *PyDictionary 2.0.1*. Retrieved from PyPI: <https://pypi.org/project/PyDictionary/>
- Genter, D., & Forbus, K. D. (2010). Computational models of analogy. *WIREs Cognitive Science*, 2, 266-276.
- Genter, D., & Toupin, C. (1986). Systematicity and surface similarity in the development of analogy. *Cognitive Science*, 10(3), 277-300.
- Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, 7(2), 155-170.
- Gentner, D. (1988). Metaphor as structure mapping: The relational shift. *Child Development*, 59(1), 47-59.
- Gentner, D., & Maravilla, F. (2018). Analogical reasoning. In L. J. Ball, & V. A. Thompson, *International Handbook of Thinking & Reasoning* (pp. 186-203). New York: Psychology Press.
- Gentner, D., & Markman, A. B. (1997). Structure mapping in analogy and similarity. *American Psychologist*, 52(1), 45-56.
- Gentner, D., & Smith, L. (2012). Analogical reasoning. In V. S. Ramachandran, *Encyclopedia of Human Behavior* (pp. 130-136). Academic Press.
- Goldstone, R. L., & Rogosky, B. J. (2002). Using relations within conceptual systems to translate across conceptual systems. *Cognition*, 84, 295-320.
- Golstone, R. (1994). Similarity, interactive activation, and mapping. *Experimental Psychology: Learning, Memory, and Cognition*, 20(1), 3-28.
- Google. (2021, November). *Detect Labels*. Retrieved from Google Cloud: https://cloud.google.com/vision/docs/labels#detect_labels_in_a_remote_image
- Google. (2021, November). *Vision AI*. Retrieved from Google Cloud: <https://cloud.google.com/vision/?cloudshell=false#section-2>
- Goswami, U. (1991). Analogical reasoning: What develops? A review of research and theory. *Child Development*, 62(1), 1-22.
- Goswami, U. (1992). *Analogical reasoning in children*. Lawrence Erlbaum Associates.

- Goswami, U., & Brown, A. L. (1990). Higher-order structure and relational reasoning: Contrasting analogical and thematic relationship. *Cognition*, 36, 207-226.
- Greiner, R. (1988). Learning by understanding analogies. *Artificial Intelligence*, 35(1), 81-125.
- Griffin, G., Holub, A., & Perona, P. (2007). *Caltech-256 object category dataset*. Pasadena: Caltech. Retrieved from http://www.vision.caltech.edu/Image_Datasets/Caltech256/
- Gust, H., Kuhnberger, K.-U., & Schmid, U. (2006). Metaphors and heuristic-driven theory projection (HDTP). *Theoretical Computer Science*, 354(1), 98-117.
- Halford, G. S., Wilson, W. H., Guo, J., Gayler, R. W., Wiles, J., & Stewart, J. E. (1994). Connectionist implications for processing capacity limitations in analogies. *Advances in connectionist and neural computation theory*, 2, 363-415.
- Hall, R. P. (1989). Computational approaches to analogical reasoning: A comparative analysis. *Artificial Intelligence*, 39, 39-120.
- Henley, N. M. (1969). A psychological study of the semantics of animal terms. *Journal of Verbal Learning and Verbal Behavior*, 8(2), 176-184.
- Hertzmann, A., Jacobs, C., Oliver, N., Curless, B., & Salesin, D. (2001). Image Analogies. *SIGGRAPH Conference Proceedings*, 1-14.
- Hofstadter, D. R. (1984). *The Copycat Project: An Experiment in Nondeterminism and Creative Analogies*. Massachusetts Institute of Technology, Artificial Intelligence Lab, Cambridge.
- Hofstadter, D. R., & Mitchell, M. (1995). The copycat project: A model of mental fluidity and analogy-making. *Advances in connectionist and neural computation theory*, 2, 205-267.
- Holyoak, K. J. (1995). Problem solving. In E. E. Smith, & D. N. Osherson (Eds.), *An Invitation to Cognitive Science* (2 ed., Vol. 3, pp. 267-296). The MIT Press.
- Holyoak, K. J., & Thagard, P. (1989). Analogical mapping by constraint satisfaction. *Cognitive Science*, 13, 295-355.
- Hummel, J. E., & Holyoak, K. J. (1997). Distributed representations of structure. *Psychological Review*, 104(3), 427-466.
- Hummel, J. E., & Holyoak, K. J. (2005). Relational reasoning in a neurally plausible cognitive architecture: An overview of the LISA project. *Current Directions in Psychological Science*, 14(3), 153-157.

- Hwang, S. K., Grauman, K., & Sha, F. (2013). Analogy-preserving semantic embedding for visual object categorization. *International Conference on Machine Learning*, 639-647.
- Ichien, N., Lu, H., & Holyoak, K. J. (2020). Verbal analogy problem sets: An inventory of testing materials. *Behavior Research Methods*, 53, 1803--1816.
- Johnson-Laird, P. (2010). Deductive reasoning. *Wiley Interdisciplinary Reviews: Cognitive Science*, 1(1), 8-17.
- Kafe, E. (2019). Fitting semantic relations to word embeddings. *10th Global WordNet Conference*. Wroclaw.
- Keane, M. T., & Brayshaw, M. (1988). The incremental analogy machine: A computational model of analogy. *Third European Working Session on Machine Learning*, (pp. 53-62). San Mateo.
- Keras. (2020, December). *Keras API Reference*. Retrieved from Layer Activation Function: <https://keras.io/api/layers/activations/>
- Keras. (2020, December). *Keras API Reference*. Retrieved from Flatten Layer: https://keras.io/api/layers/reshaping_layers/flatten/
- Keras. (2020, December). *Keras API Reference*. Retrieved from Dense Layer: https://keras.io/api/layers/core_layers/dense/
- Khatena, J. (1972). The use of analogy in the production of original verbal images. *The Journal of Creative Behavior*, 9(3), 209-213.
- Kokinov, B. N., & Petrov, A. A. (2000). Integration of memory and reasoning in analogy-making. In D. Genter, K. Holyoak, & B. Kokinov, *The analogical mind: Perspectives from Cognitive Science* (pp. 59-124). Cambridge: MIT Press.
- Kokinov, B., & French, R. M. (2003). Computational models of analogy-making. *Encyclopedia of Cognitive Science*, 1, 113-118.
- Kokinov, B. (1994). A hybrid model of reasoning by analogy. In K. J. Holyoak, & J. A. Barnden, *Advances in connectionist and neural computation theory* (Vol. 2, pp. 247-318). Norwood, NJ: Ablex.
- Korogodina, O., Karpik, O., & Klyshinsky, E. (2020). Evaluation of vector transformations for Russian Word2Vec and FastText embeddings. *Proceedings of the 30th International Conference on Computer Graphics and Machine Vision*. Saint Petersburg: ITMO University.
- Krasin, I., T. D., Alldrin, N., Ferrari, V., Abu-El-Haija, S., Kuznetsova, A., . . . Murphy, K. (2020, February). *OpenImages: A public dataset for large-scale multi-label and*

multi-class image classification. Retrieved from Open Images Dataset V6: <https://storage.googleapis.com/openimages/web/index.html>

- Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., . . . Bernstein, M. S. (2017). Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, *123*(1), 32-73.
- Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., . . . Fei-fei, L. (2016). *Visual Genome: Connecting language and vision using crowdsourced dense image annotations*. Visual Genome. Retrieved from <https://arxiv.org/abs/1602.07332>
- Krizhensky, A. (2009). *Learning Multiple Layers of Features from Tiny Images*.
- Kuehne, S., Forbus, K., Gentner, D., & Quinn, B. (2000). SQL: Category learning as progressive abstraction using structure mapping. *22nd annual meeting of the cognitive science society*, (pp. 770-775).
- Larkey, L. B., & Love, B. C. (2003). CAB: Connectionist analogy builder. *Cognitive Science*, *27*, 781-794.
- Leech, R., Mareschal, D., & Cooper, R. P. (2008). Analogical as relational priming: A developmental and computational perspective on the origins of a complex cognitive skill. *Behavioral and Brain Sciences*, *31*, 357-414.
- Levinson, P. J., & Carpenter, R. L. (1974). An analysis of analogical reasoning in children. *Child Development*, *45*(3), 857-861.
- Levy, O., & Goldberg, Y. (2014). Linguistic regularities in sparse and explicit word representations. *Eighteenth Conference on Computational Natural Language Learning* (pp. 171-180). Ann Arbor: Association for Computational Linguistics.
- Lin, T.-y., Maire, M. B., Bourdev, L., Girshick, R., Hayes, J., Perona, P., . . . Dollar, P. (2015). Microsoft COCO: Common objects in context. Retrieved from <https://arxiv.org/pdf/1405.0312.pdf>
- Lovett, A., & Forbus, K. (2017). Modeling visual problem solving as analogical reasoning. *Psychological Review*, *124*(1), 60-90.
- Lu, H., Chen, D., & Holyoak, K. (2012). Bayesian analogy with relationship transformations. *Psychological Review*, *119*(3), 617-648.
- Lu, H., Liu, Q., Ichien, N., Yuille, A. L., & Holyoak, K. J. (2019). Seeing the meaning: Vision meets semantics in solving pictorial analogy problems. *41st Annual Meeting of the Cognitive Science Society*. Austin: Cognitive Science Society.
- Lu, H., Wu, Y. N., & Holyoak, K. J. (2019). Emergence of analogy from relation learning. *Proceedings of the National Academy of Science*, *116*(10), 4176-4181.

- Lucid. (2021). *Flowchart Symbols and Notation*. Retrieved from Lucidchart: <https://www.lucidchart.com/pages/flowchart-symbols-meaning-explained>
- Marshall, J. B. (1999, November). Metacat: A self-watching cognitive architecture for analogy-making and high-level perception. Bloomington, Indiana: Indiana University.
- Marshall, J. B. (2002). Metacat: A self-watching cognitive architecture for analogy-making. *Cognitive Science Society*, 24. Fairfax.
- McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4), 115-133.
- Memisevic, R., & Hinton, G. E. (2010). Learning to represent spatial transformations with factored higher-order Boltzmann machines. *Neural Computation*, 22, 1473-1492.
- Mikolov, T., Grave, E., Bojanowski, P., Puhersch, C., & Joulin, A. (2018). Advances in pre-training distributed word representations. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki: European Language Resources Association (ELRA).
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, 3111-3119.
- Mikolov, T., Yih, W.-t., & Zweig, G. (2013). Linguistic regularities in continuous space word representations. *2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technology* (pp. 746-751). Atlanta: Association for Computational Linguistics.
- Mitchell, M. (1993). *Analogy-making as perception: A computer model*. Cambridge: MIT Press.
- Morrison, R. G., Krawczyk, D. C., Holyoak, K. J., Hummel, J. E., Chow, T. W., Miller, B. L., & Knowlton, B. J. (2004). A neurocomputational model of analogical reasoning and its breakdown in frontotemporal lobar degeneration. *Journal of Cognitive Neuroscience*, 16(2), 260-271.
- NLTK. (2021, August). *WordNet Interface*. Retrieved from NLTK: <http://www.nltk.org/howto/wordnet.html>
- Oppy, G., & Dowe, D. (2020). The Turing test. In E. N. Zalta, *The Standard Encyclopedia of Philosophy*. Stanford: Metaphysics Research Lab.
- O'Shea, K., & Nash, R. (2015). An introduction to convolutional neural networks. *arXiv preprint*, 1-11.

- Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for word representation. *2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1532-1543). Doha: Association for Computational Linguistics.
- Peterson, J. C., Chen, D., & Griffiths, T. L. (2020). Parallelograms revisited: Exploring the limitations of vector space models for simple analogies. *Cognition*, *205*, 1-15. doi:10.1016/j.cognition.2020.104440
- Petrov, A. A. (1997). *Extensions fo DUAL and AMBR*. Sofia: New Bulgarian University, Cognitive Science Department.
- Plate, T. A. (1994). Distributed representations and nested compositional structure. Toronto, Ontario, Canada: University of Toronto, Department of Computer Science.
- Polya, G. (1990). *Mathematics and plausible reasoning: Induction and analogy in mathematics* (Vol. 1). Princeton: Princeton University Press.
- Potts, G. R. (1978). The role of inference in memory for real and artificial information. In R. Revlin, & R. E. Mayer, *Human Reasoning* (pp. 139-161). Washington D.C.: V. H. Winston & Sons.
- Ramscar, M. J., & Pain, H. G. (1996). Can a real distinction be made between cognitive theories of analogy and categorization? *18th Annual Conference of the Cognitive Science Society*. San Diego.
- Reed, S. E., Zhang, Y., Zhang, Y., & Lee, H. (2015). Deep visual analogy-making. *Advanced in neural information processing systems*, 1252-1260.
- Řehůřek, R., & Sojka, P. (2010). Software framework for topic modeling with large corpora. *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks* (pp. 45-50). Valletta, Malta: ELRA. Retrieved from <http://is.muni.cz/publication/884893/en>
- Roads, B. D., & Love, B. C. (2020). Learning as the unsupervised alignment of conceptual systems. *Nature Machine Intelligence*, *2*(1), 76-82. doi:10.1038/s42256-019-0132-2
- Rogers, A., Drozd, A., & Li, B. (2017). The (too many) problems of analogical reasoning with word vectors. *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics*, 135-148.
- Rumelhart, D. E., & Abrahamson, A. A. (1973). A model for analogical reasoning. *Cognitive Psychology*, *5*(1), 1-28.

- Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., . . . Fei-fei, L. (2015). ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, *115*, 211-252. doi:10.1007/s11263-015-0816-y
- Russell, B. C., Torralba, A., Murphy, K. P., & Freeman, W. T. (2008). LabelMe: A database and web-based tool for image annotation. *International Journal of Computer Vision*, *77*(1-3), 157-173. Retrieved from <http://people.csail.mit.edu/brussell/research/AIM-2005-025-new.pdf>
- Sadeghi, F., Zitnick, C. L., & Farhadi, A. (2015). Visalogy: Answering visual analogy questions. *Advances in Neural Information Processing Systems*, *28*, 1882-1890.
- Situ, J. X., Friend, M. A., Bauer, K. W., & Bihl, T. J. (2016). Contextual features and Bayesian belief networks for improved synthetic aperture radar combat identification. *Military Operations Research*, *21*(1), 89-106.
- Skorstad, J., Gentner, D., & Medin, D. (1988). Abstraction processes during concept learning: A structural view. *Tenth Annual Conference of the Cognitive Science Society*, (pp. 419-425).
- Socher, R., Ganjoo, M., & Manning, C. D. (2013). Zero-shot learning through cross-modal transfer. *Proceedings of the 26th International Conference on Neural Information Processing Systems. 1*, pp. 935-943. Sydney: Curran Associates Inc.
- Speer, R., Chin, J., & Havasi, C. (2017). Conceptnet 5.5: An open multilingual graph of general knowledge. *Proceedings of the AAAI Conference on Artificial Intelligence. 31*. San Francisco: Association for the Advancement of Artificial Intelligence (AAAI).
- Stanford Vision Lab. (2020). *ImageNet*. (Stanford University & Princeton University) Retrieved from ImageNet Large Scale Visual Recognition Challenge (ILSVRC): <https://image-net.org/challenges/LSVRC/index.php>
- Stenning, K., & Van Lambelgen, M. (2012). *Human reasoning and cognitive science*. Cambridge: MIT Press.
- Sternberg, R. J., & Nigro, G. (1980). Developmental patterns in the solution of verbal analogies. *Child Development*, *51*(1), 27-38.
- Summers-Stay, D. (2017). Deductive and analogical reasoning on a semantically embedded knowledge graph. *International Conference on Artificial General Intelligence* (pp. 1-10). Cham: Springer.
- Thagard, P., Holyoak, K., Nelson, G., & Gochfield, D. (1990). Analog retrieval by constraint satisfaction. *Artificial Intelligence*, *46*(3), 259-310.

- Torralba, A. F., & Freeman, W. T. (2007). *Tiny images*. Cambridge: MIT. Retrieved from http://people.csail.mit.edu/torralba/publications/TR_tiny_images.pdf
- Torralba, A., Fergus, R., & Freeman, B. (2020, June 29). *Tiny Images*. Retrieved from <https://groups.csail.mit.edu/vision/TinyImages/>
- Turney, P. D. (2006). Similarity of semantic relations. *Computational Linguistics*, 32(3), 379-416.
- Weigand, K. A., & Hartung, R. (2012). Abduction's role in reverse engineering software. *IEEE National Aerospace and Electronics Conference (NAECON)* (pp. 57-62). Dayton: IEEE.
- Wilson, W. H., Halford, G. S., Gray, B., & Philips, S. (2001). The STAR-2 model for mapping hierarchically structured analogs. (D. Gentner, K. J. Holyoak, & B. N. Kolkinov, Eds.) *The analogical mind*, 125-60.
- Wu, J. (2017). *Introduction to convolutional neural networks*. Nanjing: National Key Lab for Novel Software Technology.
- Yaner, P. W., & Goel, A. K. (2006). Visual analogy: Viewing analogical retrieval and mapping as constraint satisfaction problems. *Applied Intelligence*, 25(1), 91-105.
- Zhang, B. (2008). Hypernetworks: A molecular evolutionary architecture for cognitive learning and memory. *IEEE computational intelligence magazine*, 3(3), 49-63.
- Zhang, C., Gao, F., Jia, B., Zhu, Y., & Zhu, S. (2019). RAVEN: A dataset for relational and analogical visual reasoning. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5317-5327.

8 APPENDIX A

Table 8-1: Goodness Metric Results

Key									
	0.000	Perfect Analogy		0.251	Average Analogy		1.000+	Poor Analogy	Unable to Attempt
Analogy	DRAMA	BART 1.0	BART 2.0	Word2vec	GloVe	3CosAvg	LRCos		
YES:NO::TRUE:FALSE	0.057	0.221	0.226	0.049	0.000	0.019	0.023		
COOL:WARM::BLACK:WHITE	0.013	0.029	0.159	0.497	0.065	0.041	0.042		
OFTEN:SELDOM::HAPPY:SAD	0.392	0.145	0.159	1.345	0.031	0.008	0.000		
LOVE:HATE::HIT:MISS	0.495	0.098	0.234	0.804	0.415	0.020	0.125		
STOP:GO::EAST:WEST	0.533	0.085	0.170	0.422	0.308	0.001	0.002		
NARROW:WIDE::QUESTION:ANSWER	0.765	0.027	0.216	0.063	0.376	0.048	0.004		
WILD:TAME::HARD:SOFT	0.008	0.068	0.178	0.210	0.287	0.081	0.001		
STRAIGHT:BENT::FIND:LOSE	0.010	0.060	0.246	0.606	0.478				
UP:DOWN::POOR:RICH	0.068	0.147	0.250	0.058	0.807	0.003	0.004		
EMPTY:FULL::BETTER:WORSE	0.057	0.062	0.308	0.154	0.460	0.106	0.002		
WIN:LOSE::ABOVE:BELOW	0.101	0.222	0.208	0.276	0.315	0.008	0.002		
LIKE:DISLIKE::WARM:COOL	0.098	0.120	0.239	0.400	0.680	0.047	0.051		
RIGHT:WRONG::CALM:STORMY	0.390	0.026	0.291	1.617	1.195	0.114	0.088		
FAST:SLOW::ON:OFF	0.120	0.010	0.166	0.173	0.212	0.032	0.031		
FOOLISH:WISE::EARLY:LATE	0.051	0.121	0.217	0.542	0.376				
COME:GO::YOUNG:OLD	0.450	0.094	0.158	0.077	0.932	0.008	0.000		
BEGIN:END::DARK:LIGHT	0.653	0.036	0.247	0.295	0.073	0.048	0.060		
OPEN:CLOSE::SHALLOW:DEEP	0.119	0.210	0.222	0.034	0.294	0.194	0.072		
BLACK:WHITE::BAD:GOOD	0.119	0.165	0.100	0.864	0.107	0.385	0.373		
QUIET:LOUD::CATCH:THROW	0.020	0.186	0.240	0.086	0.305	0.042	0.090		
JOHN:NAME::DINNER:MEAL	0.716	0.064	0.264	0.326	0.562	0.017	0.007		
ROSE:FLOWER::CHURCH:BUILDING	0.542	0.124	0.309	0.042	0.474	0.049	0.075		
BREAD:FOOD::BLACK:COLOR	0.185	0.063	0.222	0.118	0.127	0.097	0.094		
RED:COLOR::HIT:ACTION	0.078	0.077	0.183	0.169	0.936	0.112	0.048		
BEE:INSECT::LOVE:FEELING	0.110	0.056	0.228	0.623	0.316	0.014	0.010		
NOON:TIME::WEST:DIRECTION	0.643	0.086	0.238	0.148	0.107	0.006	0.000		
ENGLISH:LANGUAGE::BASEBALL:GAME	0.094	0.229	0.300	0.416	0.174	0.101	0.034		
LION:ANIMAL::CHRISTMAS:HOLIDAY	0.019	0.156	0.211	0.459	0.571	0.053	0.005		
CITY:NEW::YORK:FISH:GUPPY	0.706	0.204	0.196			0.071	0.029		
ANIMAL:CALF::CAT:SIAMESE	0.058	0.099	0.227		0.080	0.100	0.084		
MOTHER:WOMAN::STREET:THOROUGHFARE	0.611	0.214	0.255		0.765	0.114	0.130		

DAY:SUNDAY::CLOTHES:SHOES	0.586	0.149	0.242	0.200	0.047	0.057	0.123
TABLE:FURNITURE::APPLE:FRUIT	0.058	0.042	0.304	0.179	0.315	0.152	0.151
UNCLE:MAN::DOG:ANIMAL	0.552	0.022	0.208	0.576	0.251	0.001	0.013
SONG:MUSIC::PAINTING:ART	0.675	0.065	0.148	0.449	0.064	0.145	0.040
GOLD:METAL::SLOW:SPEED	0.086	0.071	0.211	0.569	0.400	0.019	0.020
RIVER:WATER::MOUNTAIN:LAND	0.704	0.125	0.212	0.239	0.713	0.033	0.032
CHAIR:FURNITURE::YOUNG:AGE	0.597	0.166	0.301	0.462	0.310		
TROUT:FISH::PIG:ANIMAL	0.814	0.250	0.231	0.028	0.199	0.418	0.449
G:LETTER::SEVEN:NUMBER	0.059	0.029	0.357	0.907	0.666	0.281	0.020
BIRD:FLY::RABBIT:HOP	0.548	0.071	0.118	0.048	3.344	0.186	0.112
HORN:PLAY::HORSE:RIDE	0.566	0.199	0.256	0.802	0.612	0.000	0.025
ROAD:CAR::SKY:PLANE	0.069	0.048	0.168	0.114	0.200	0.045	0.034
HOT:FURNACE::COOL:FAN	0.763	0.084	0.332	0.432	0.018	0.027	0.011
PLAY:GAME::GIVE:PARTY	0.681	0.075	0.216	0.300	1.598	0.132	0.020
WET:RAIN::DRY:SUN	0.634	0.146	0.194	0.174	0.917	0.043	0.031
BOOKS:READ::TOYS:PLAY	0.120	0.077	0.183	0.141	2.218	0.029	0.000
TASTE:MOUTH::TOUCH:HANDS	0.798	0.136	0.281	0.177	0.382		
SCHOOL:LEARN::STORE:BUY	0.013	0.002	0.189	0.329	0.201	0.123	0.055
PEOPLE:FOOD::CARS:GAS	0.305	0.120	0.277	0.104	0.104	0.002	0.017
BED:SLEEP::CHAIR:SIT	0.590	0.329	0.172	1.599	0.171	0.255	0.260
MORNING:BREAKFAST::EVENING:DINNER	0.394	0.058	0.095	0.109	0.199	0.081	0.117
FIRE:BURN::WIND:BLOW	0.060	0.104	0.204	0.128	0.209	0.065	0.045
STORY:TELL::SONG:SING	0.431	0.198	0.133	0.007	0.160	0.004	0.008
AIRPLANE:SKY::SHIP:OCEAN	0.434	0.045	0.178	0.000	0.543	0.001	0.001
BIRD:NEST::HORSE:STABLE	0.559	0.341	0.185	0.008	0.232	0.273	0.262
PLATE:EAT::CUP:DRINK	0.792	0.059	0.215	0.180	0.220	0.018	0.085
COOK:KITCHEN::BUY:STORE	0.434	0.009	0.164	0.740	0.011	0.285	0.004
CAR:DRIVE::BOAT:SAIL	0.178	0.223	0.140	0.174	0.115	0.065	0.090
READ:NEWSPAPER::SIT:CHAIR	0.660	0.197	0.205	2.247	0.186	0.010	0.120
MONTH:YEAR::INCH:FOOT	0.785	0.005	0.190	0.536	1.019	0.075	0.026
JANUARY:FEBRUARY::FIRST:SECOND	0.560	0.013	0.188	0.443	0.135	0.125	0.125
PALACE:CABIN::ROCK:PEBBLE	0.668	0.227	0.258	0.334	0.470	0.109	0.008
GOOD:BETTER::BAD:WORSE	0.427	0.098	0.078	0.141	0.218	0.119	0.021
WORSE:WORST::LOWER:LOWEST	0.663	0.032	0.102	0.260	0.173	0.007	0.000
PUPPY:DOG::CUB:BEAR	0.451	0.011	0.222	2.902	0.527	0.015	0.020
WARM:HOT::COOL:COLD	0.071	0.064	0.327	0.004	0.109	0.098	0.099
EGG:CHICKEN::SEED:FLOWER	0.804	0.050	0.258	0.414	0.702	0.029	0.069
SMALL:SMALLER::APPLE:FRUIT	0.034	0.035	0.253	0.234	0.612	0.138	0.136
BIRTH:LIFE::DAWN:DAY	0.747	0.033	0.254	0.243	0.036		
BREAKFAST:LUNCH::LUNCH:DINNER	0.004	0.025	0.232	0.021	0.061	0.097	0.041

MORNING:AFTERNOON::AFTERNOON:EVENING	0.067	0.027	0.082	0.052	0.141	0.012	0.003
INCH:FOOT::YARD:MILE	0.124	0.199	0.297	0.145	0.111	0.006	0.019
PRINCE:KING::PRINCESS:QUEEN	0.621	0.135	0.105	0.339	0.025	0.011	0.011
PERSON:CROWD::DROP:PUDDLE	0.817	0.219	0.323	1.037	1.978	0.044	0.003
NONE:SOME::MOST:ALL	0.678	0.060	0.129	0.101	0.203	0.029	0.035
RUN:WALK::SHOUT:TALK	0.331	0.181	0.269	0.387	1.074	0.122	0.056
FOURTH:FIFTH::APRIL:MAY	0.544	0.005	0.396	0.047	0.367	0.171	0.074
CAR:BUS::KNIFE:CLEAVER	0.415	0.020	0.191	0.465	0.680	0.148	0.006
EQUAL:SAME::QUIET:STILL	0.642	0.129	0.247	0.034	0.383	0.011	0.015
NEARLY:ALMOST::CAR:AUTO	0.662	0.011	0.263	1.038	0.528	0.066	0.058
EASY:SIMPLE::SHUT:CLOSE	0.528	0.172	0.256	0.144	0.629	0.017	0.008
END:FINISH::BROOK:STREAM	0.617	0.010	0.202	0.101	0.040	0.003	0.012
NEAR:CLOSE::FIX:MEND	0.087	0.222	0.258	2.243	0.730	0.051	0.051
CERTAIN:SURE::SHIP:BOAT	0.725	0.261	0.165	0.259	0.243	0.087	0.001
STEAL:ROB::PULL:DRAG	0.586	0.104	0.159	0.584	0.036	0.021	0.008
HUGE:ENORMOUS::PAMPHLET:BOOKLET	0.197	0.270	0.102		0.446		
QUARREL:FIGHT::BUILD:MAKE	0.730	0.063	0.290		0.603	0.004	0.019
THROW:PITCH::BEGIN:START	0.643	0.105	0.264	0.711	0.399	0.057	0.002
BIG:LARGE::WEAK:FEEBLE	0.454	0.201	0.151		0.007	0.382	0.484
FOREST:WOODS::STREET:ROAD	0.380	0.121	0.249	0.127	0.221	0.095	0.027
HAVE:POSSESS::HARD:DIFFICULT	0.684	0.007	0.186	0.030	0.499	0.051	0.008
LIBERTY:FREEDOM::FATHER:DAD	0.146	0.089	0.155	0.182	0.025	0.084	0.115
HELP:AID::HAT:CAP	0.612	0.002	0.272	0.402	0.380	0.011	0.001
MERRY:GAY::INTELLIGENT:SMART	0.053	0.075	0.333	0.353	0.925	0.004	0.002
CORRECT:RIGHT::OLD:AGED	0.708	0.122	0.240	0.273	0.133	0.009	0.001
REMAIN:STAY::SPEAK:TALK	0.429	0.136	0.145	0.396	0.142	0.047	0.029
RICH:WEALTHY::FAT:ROTUND	0.492	0.015	0.215		3.426	0.059	0.070
FAST:SPEEDY::DOCTOR:PHYSICIAN	0.882	0.055	0.151	0.267	0.387	0.056	0.000
ENTER:LEAVE::FLOAT:SINK	0.291	0.177	0.202	0.317	0.396	0.014	0.018
DIFFERENT:SAME::SHORT:LONG	0.726	0.146	0.169	0.166	0.041	0.147	0.146
LOUD:SOFT::FAT:THIN	0.461	0.112	0.210	0.324	0.185	0.020	0.019
START:FINISH::FAR:NEAR	0.127	0.014	0.215	0.706	0.400	0.011	0.004
BACK:FRONT::WET:DRY	0.699	0.014	0.184	0.354	0.354	0.006	0.004
NEVER:ALWAYS::LOWEST:HIGHEST	0.814	0.197	0.132	0.001	0.038	0.063	0.054
FEW:MANY::NOISY:QUIET	0.070	0.225	0.230	0.080	0.986	0.004	0.040
ADD:SUBTRACT::BEST:WORST	0.713	0.042	0.186	0.341	0.137	0.020	0.039
INSIDE:OUTSIDE::OVER:UNDER	0.717	0.008	0.165	0.337	0.537	0.014	0.013
GOOD:BAD::NEW:OLD	0.539	0.062	0.184	0.261	0.753	0.191	0.181
HUGE:TINY::CLEAN:DIRTY	0.869	0.172	0.171	0.220	0.198	0.314	0.311
SLOWLY:QUICKLY::WORK:PLAY	0.603	0.086	0.223	0.408	0.494		

HEAVY:LIGHT::LONG:SHORT	0.667	0.226	0.268	0.153	0.327	0.007	0.005
ASLEEP:AWAKE::LARGE:SMALL	0.692	0.127	0.269	0.476	0.227	0.019	0.009
BEFORE:AFTER::AFRAID:BRAVE	0.193	0.135	0.253	0.649	1.175	0.005	0.002
SUCCEED:FAIL::REMEMBER:FORGET	0.079	0.197	0.212	0.475	0.398	0.132	0.026
WEAK:STRONG::GROW:WITHER	0.011	0.001	0.290		1.142	0.211	0.012
DANGER:SAFETY::CRY:LAUGH	0.003	0.027	0.203	0.595	0.307	0.072	0.006
FORWARD:BACKWARD::FUTURE:PAST	0.716	0.038	0.218	0.022	0.055	0.016	0.004
HIRE:FIRE::START:FINISH	0.020	0.229	0.306	0.064	0.699	0.029	0.012
EGG:FOOD::SOUTH:DIRECTION	0.064	0.061	0.297	0.097	0.011	0.135	0.094
QUEEN:RULER::CIRCLE:SHAPE	0.323	0.087	0.231	1.633	0.051	0.016	0.032
MONTH:MAY::SEASON:WINTER	0.423	0.127	0.208	0.201	0.031	0.065	0.063
CORN:VEGETABLE::DOLLAR:MONEY	0.126	0.127	0.251	0.032	0.120	0.015	0.014
HAMMER:TOOL::SHORT:SIZE	0.107	0.299	0.340	0.212	0.045	0.061	0.058
ONE:NUMBER::BALL:TOY	0.037	0.024	0.375	0.192	2.767	0.001	0.002
FRUIT:ORANGE::DOG:POODLE	0.750	0.037	0.173	0.333	0.001	0.167	0.212
SILVER:METAL::PRINCE:NOBILITY	0.053	0.126	0.237	0.583	0.567	0.030	0.032
ARITHMETIC:SUBJECT::HEAVY:WEIGHT	0.041	0.112	0.310		0.756		
HOUSE:BUILDING::SMALL:SIZE	0.717	0.238	0.165	0.177	0.048	0.153	0.078
DOCTOR:TITLE::PRIVATE:RANK	0.008	0.243	0.271	0.200	0.386	0.027	0.064
KITCHEN:ROOM::NEAR:DISTANCE	0.413	0.015	0.254	0.185	0.925	0.047	0.017
PINE:TREE::DARK:ILLUMINATION	0.668	0.085	0.236	0.086	1.563	0.006	0.010
NEST:HOME::OUNCE:WEIGHT	0.486	0.103	0.225	0.602	0.359	0.009	0.009
FOOTBALL:GAME::BIG:SIZE	0.802	0.040	0.303	0.033	0.273	0.065	0.082
ROBIN:BIRD::JUNE:MONTH	0.760	0.077	0.219	0.086	0.759	0.012	0.047
KING:RULER::BOY:MALE	0.014	0.127	0.280	0.168	0.540	0.098	0.071
SWIMMING:SPORT::GUN:WEAPON	0.360	0.162	0.228	0.351	0.248	0.001	0.005
PACIFIC:OCEAN::KITCHEN:ROOM	0.237	0.087	0.256	0.108	0.022	0.003	0.004
ZINC:MINERAL::SHORT:HEIGHT	0.450	0.300	0.238	0.059	0.495	0.031	0.024
CLIMB:HILL::DIG:HOLE	0.850	0.099	0.196	1.248	0.215	0.000	0.005
EAR:HEAR::EYE:SEE	0.749	0.192	0.082	0.096	0.215		
DOOR:OPEN::STAIRS:CLIMB	0.083	0.146	0.204	0.129	0.077	0.003	0.018
ZOO:ANIMALS::FOREST:TREES	0.527	0.050	0.152	0.071	0.121	0.082	0.064
CUT:KNIFE::MIX:SPOON	0.218	0.000	0.184	0.136	0.074	0.013	0.012
SHOES:FEET::HAT:HEAD	0.043	0.138	0.312	0.595	0.469	0.036	0.081
NAIL:BUIL::DPEN:WRITE	0.094	0.269	0.240			0.002	0.012
WARM:SUN::WET:CLOUDS	0.515	0.133	0.173	0.054	0.124	0.010	0.005
CLOCK:TIME::NEWSPAPER:NEWS	0.042	0.131	0.282	0.264	0.290	0.005	0.003
CRY:HURT::SMILE:HAPPY	0.173	0.006	0.269	0.449	0.451	0.037	0.002
MOON:NIGHT::SUN:DAY	0.547	0.069	0.218	0.458	0.004	0.017	0.075
GLASS:DRINK::PAN:COOK	0.705	0.009	0.162	6.516	0.197	0.042	0.036

SLED:SNOW::CAR:ROAD	0.650	0.338	0.243	0.629	0.342	0.042	0.118
RAIN:SUMMER::SNOW:WINTER	0.217	0.030	0.135	0.185	0.437	0.066	0.059
FIGHT:SOLDIER::HELP:DOCTOR	0.545	0.000	0.151	0.407	0.043	0.023	0.005
BICYCLE:RIDE::BASEBALL:THROW	0.562	0.090	0.237	0.082	0.535	0.004	0.004
APPLE:EAT::ROSE:SMELL	0.431	0.132	0.278	0.172	2.719	0.006	0.008
DIRT:SOAP::PAIN:PILL	0.114	0.004	0.218	0.758	0.476	0.008	0.007
PLAY:GAME::WORK:JOB	0.883	0.097	0.134	0.130	0.199	0.145	0.054
CAT:PET::COW:MILK	0.377	0.077	0.155	0.322	0.221	0.481	0.138
BASEBALL:MARBLE::BUCKET:GLASS	0.819	0.143	0.226	0.340	0.607	0.056	0.006
CRAWL:WALK::WALK:RUN	0.109	0.151	0.267	0.121	0.116	0.141	0.005
LITTLE:TINY::LARGE:HUGE	0.116	0.019	0.264	0.025	0.359	0.377	0.372
BETTER:BEST::TALLER:TALLEST	0.742	0.119	0.177		0.397	0.146	0.023
TAP:STRIKE::BREAK:DESTROY	0.570	0.070	0.282	0.709	0.676		
YESTERDAY:TODAY::BEFORE:NOW	0.706	0.027	0.089	0.042	0.202	0.010	0.012
NEVER:SOMETIMES::OFTEN:ALWAYS	0.832	0.004	0.185	0.180	0.178	0.299	0.001
PAGES:BOOK::LETTERS:WORD	0.090	0.214	0.259	0.622	0.265	0.133	0.010
A LOT:A LITTLE::WET:MOIST	0.397	0.031	0.216			0.213	0.011
KITTEN:CAT::CALF:COW	0.714	0.036	0.214	0.249	0.014	0.074	0.075
ONE:DOZEN::PENNY:DOLLAR	0.055	0.026	0.238	0.030	1.373	0.002	0.056
FUTURE:PRESENT::PRESENT:PAST	0.951	0.049	0.286	0.053	0.436	0.420	0.005
LESS:LEAST::MORE:MOST	0.505	0.069	0.138	0.395	0.110	0.050	0.065
LAKE:OCEAN::BIG:BIGGER	0.315	0.202	0.222	0.177	0.418	0.118	0.111
CITY:COUNTY::STATE:COUNTRY	0.941	0.058	0.249	0.189	0.045	0.074	0.074
MODERATE:HEAVY::RAIN:DOWNPOUR	0.876	0.179	0.188		0.590	0.124	0.002
MOSTLY:SOMEWHAT::OVERCAST:PARTLY CLOUDY	0.795	0.105				0.071	0.066
EAGLE:HAWK::LARGE:MODERATE	0.904	0.182	0.207	0.142	0.791	0.040	0.028
STRANGE:ODD::DISH:PLATE	0.809	0.154	0.265	1.287	0.875	0.250	0.056
PERHAPS:MAYBE::CHAIR:SEAT	0.033	0.013	0.272	1.985	0.498	0.084	0.056
ALLOW:LET::CRY:WEEP	0.572	0.058	0.194	0.041	0.343	0.048	0.024
HURRY:RUSH::SAD:UNHAPPY	0.426	0.020	0.226	1.584	0.140	0.219	0.137
UNDER:BENEATH::PAIN:HURT	0.502	0.159	0.259	0.310	0.097	0.018	0.016
ENJOY:LIKE::FALL:AUTUMN	0.058	0.025	0.217	0.010	0.174	0.278	0.271
SLENDER:THIN::SICK:ILL	0.696	0.015	0.256		0.231	0.102	0.101
NOTICE:SEE::MURDER:HOMICIDE	0.237	0.059	0.161	0.445	0.419	0.111	0.024
OVER:ABOVE::PANTS:TROUSERS	0.040	0.161	0.174	0.454	0.572	0.001	0.005
FOOLISH:SILLY::HANDGUN:PISTOL	0.554	0.012	0.199	0.128	0.113	0.125	0.025
MOTOR:ENGINE::MIDDLE:CENTER	0.896	0.190	0.215	0.107	0.401	0.051	0.023
DISCOVER:FIND::DONKEY:ASS	0.664	0.202	0.264	0.231	0.614	0.031	0.041
HAPPY:CHEERFUL::TABLET:PAD	0.494	0.206	0.300	2.412	1.517	0.005	0.240
PART:PIECE::MOTHER:MOM	0.839	0.158	0.247	0.343	0.356	0.231	0.231

SOIL:EARTH::HAPPINESS:JOY	0.750	0.000	0.238	0.070	0.340	0.048	0.077
DEFRAUD:CHEAT::CHILD:KID	0.232	0.132	0.256		0.038	0.096	0.042
GIFT:PRESENT::FOOT:12 INCHES	0.473	0.149				0.018	0.054
DIFFICULT:HARD::COUGAR:MOUNTAIN LOIN	0.875	0.047				0.031	0.002
ILLEGAL:UNLAWFUL::HINDER:IMPEDE	0.733	0.078	0.140		0.325		
STRANGE:UNUSUAL::FEMALE:WOMAN	0.448	0.305	0.212	0.024	0.054	0.021	0.087
Average	0.434	0.107	0.220	0.417	0.445	0.078	0.055

9 APPENDIX B

Table 9-1: Caltech-256 Classes, Number of Images, Word Vector Representation, and PyDictionary Representation

Category Name	Number	WV Representation	PyDictionary Representation
ak47	98		Kalashnikov
american-flag	97	flag	flag
backpack	151		
baseball bat	127	(baseball + bat)	bat
baseball glove	148	(baseball + glove)	glove
basketball hoop	90	(basketball + hoop)	hoop
bat	106		
bathub	232		
bear	102		
beer-mug	94	(beer+mug)	(beer+mug)
billiards	278		
binoculars	216		*
birdbath	98		
blimp	86		
bonsai 101	122		
boom box	91	boombox	*
bowling ball	104	(bowling+ball)	(bowling+ball)
bowling pin	101		pin
boxing glove	124	(boxing+glove)	glove
brain 101	83		
breadmaker	142	(bread+maker)	*
Buddha 101	97		
bulldozer	110		
butterfly	112		
cactus	114		
cake	106		
calculator	100		
camel	110		
cannon	103		
canoe	104		
car tire	90	tire	tire
cartman	101		*
cd	102		
centipede	100		

cereal box	87	(cereal + box)	(cereal + box)
chandelier 101	106		
chess board	120	chessboard	chessboard
chimp	110		
chopsticks	85		
cockroach	124		
coffee mug	87	(coffee + mug)	(coffee + mug)
coffin	87		
coin	124		
comet	121		*
computer keyboard	85	keyboard	keyboard
computer monitor	133	monitor	monitor
computer mouse	94	mouse	mouse
conch	103		
cormorant	106		
covered wagon	97	wagon	wagon
cowboy hat	114	(cowboy + hat)	(cowboy + hat)
crab 101	85		
desk globe	82	globe	globe
diamond ring	118	ring	ring
dice	98		
dog	103		
dolphin 101	106		
doorknob	93		
drinking straw	83	straw	straw
duck	87		
dumb bell	102	dumbbell	dumbbell
eiffel tower	83	eiffel	(eiffel+tower)
electric guitar 101	122	guitar	guitar
elephant 101	131		
elk	101		
ewer 101	83		
eyeglasses	83		
fern	110		
fighter jet	99	jet	jet
fire extinguisher	84	extinguisher	extinguisher
fire hydrant	99	hydrant	hydrant
fire truck	118	firetruck	(fire+truck)
fireworks	100		firecrackers
flashlight	115		

floppy disk	83	(floppy+disk)	floppy
football helmet	84	helmet	helmet
french horn	92	horn	horn
fried egg	90	egg	egg
frisbee	99		
frog	116		
frying pan	95	pan	pan
galaxy	81		*
gas pump	95	pump	pump
giraffe	84		
goat	112		
golden gate bridge	80	bridge	bridge
goldfish	93		
golf ball	98	golfball	(golf+ball)
goose	110		
gorilla	212		
grand piano 101	95	piano	piano
grapes	201		
grasshopper	112		
guitar pick	104	pick	pick
hamburger	86		
hammock	285		
harmonica	89		
harp	100		
harpsichord	80		
hawksbill 101	93		
head phones	138	headphones	headphones
helicopter	88		
hibiscus	111		*
homer simpson	97	(homer + simpson)	*
horse	270		
horseshoe crab	87	Limulus	*
hot air balloon	89	balloon	balloon
hot dog	85	hotdog	hotdog
hot tub	156	hottub	tub
hourglass	85		
house fly	84	housefly	housefly
human skeleton	84	skeleton	skeleton
hummingbird	116		
ibis 101	120		

ice cream cone	88	icecream	(icecream+cone)
iguana	107		
iPod	121		*
iris	108		
Jesus Christ	87	jesus	jesus
joy stick	130	joystick	joystick
kangaroo 101	82		
kayak	103		
ketch 101	111		
killer whale	91	whale	whale
knife	101		
ladder	242		
laptop 101	128		
lathe	105		
leopards	190		
license plate	91	(license + plate)	(license + plate)
lightbulb	92		
light house	190	lighthouse	lighthouse
lightning	136		
llama 101	119		
mailbox	93		
mandolin	93		
mars	156		
mattress	192		
megaphone	86		
menorah	89		candelabra
microscope	117		
microwave	107		
minaret	130		
minotaur	82		*
motorbikes 101	798		
mountain bike	82	mountainbike	bike
mushroom	202		
mussels	174		
necktie	103		
octopus	111		
ostrich	109		
owl	120		
palm pilot	93	palmpilot	pda
palm tree	103	tree	tree

paperclip	92		
paper shredder	96	shredder	shredder
PCI card	105	PCI	card
penguin	149		
people	209		
pez dispenser	83	pez	(candy+dispenser)
photocopier	103		
picnic table	91	table	table
playing card	90	card	card
porcupine	101		
pram	88		
praying mantis	92	mantis	mantis
pyramid	86		
raccoon	140		
radio telescope	92	radiotelescope	(radio+telescope)
rainbow	102		
refrigerator	84		
revolver 101	99		
rifle	106		
rotary phone	84	phone	phone
roulette wheel	83	roulette	wheel
saddle	110		
saturn	96		
school bus	98	schoolbus	bus
scorpion 101	80		
screwdriver	102		
segway	100		*
self propelled lawn mower	120	lawnmower	mower
sextant	100		
sheet music	84	music	(sheet+music)
skateboard	103		
skunk	81		
skyscraper	95		
smokestack	88		
snail	119		
snake	112		
sneaker	111		
snowmobile	112		
soccer ball	174	soccerball	(soccer+ball)

socks	112		
soda can	87	can	can
spaghetti	104		
speed boat	100	speedboat	motorboat
spider	109		
spoon	105		
stained glass	100	stained-glass	(stained+glass)
starfish 101	81		
steering wheel	97	wheel	wheel
stirrups	91		
sunflower 101	80		
superman	87		*
sushi	95		
swan	115		
swiss army knife	109	knife	penknife
sword	102		
syringe	111		
tambourine	95		
teapot	136		
teddy bear	101	(toy+bear)	(toy+bear)
teepee	139		
telephone box	84	(telephone + box)	(telephone + room)
tennis ball	98	(tennis + ball)	(tennis + ball)
tennis court	105	(tennis + court)	(tennis + court)
tennis racket	81	racket	racket
theodolite	84		
toaster	94		
tomato	103		
tombstone	91		
top hat	80	top-hat	hat
touring bike	110	(touring + bike)	(touring + bike)
tower pisa	90	pisa	(pisa + tower)
traffic light	99	(traffic + light)	(traffic + light)
treadmill	147		
triceratops	95	dinosaur	dinosaur**
tricycle	95		
trilobite 101	94		
tripod	112		
t shirt	358	shirt	shirt
tuning fork	100	tuning	*

tweezer	122		
umbrella 101	114		
unicorn	97		
VCR	90		
video projector	97	projector	projector
washing machine	84	washer	washer
watch 101	201		
waterfall	95		
watermelon	93		
welding mask	90	mask	mask
wheelbarrow	91		
windmill	91		
wine bottle	101	bottle	bottle
xylophone	92		
yarmulke	84		
yo yo	100	yoyo	*
zebra	96		
airplanes 101	800		
car side 101	116	car	car
faces easy 101	435	face	face
greyhound	95		
tennis shoes	103	(tennis+shoes)	*
toad	108		

* PyDictionary was unaware of class' true context or provided an insufficient definition; therefore, the definition of the class from Lexico was used ([Lexico's website](#))

** Triceratops was not found in PyDictionary and Lexico's definition was too specific, so the definition of "dinosaur" was used in its place

Table 9-2. Classes with Definitions from Lexico

Class	PyDictionary's Definition	Lexico's Definition
Binoculars	{'Noun': ['(plural)']}	'an optical instrument with a lens for each eye used for viewing distant objects'
Boom Box	N/A	'a portable sound system typically including radio and cassette or CD player capable of powerful sound'
Breadmaker	N/A	'an electric counter appliance designed specifically for making bread and baking it'
Cartman	N/A	'an elementary school student who lives with his mother in the fictional town of South Park Colorado, where he routinely has extraordinary experiences atypical of a small town'
Comet	{'Noun': ['(astronomy)']}	'a celestial object consisting of a nucleus of ice and dust and when near the sun a tail of gas and dust particles pointing away from the sun'
Galaxy	{'Noun': ['a splendid assemblage (especially of famous people', 'tufted evergreen perennial herb having spikes of tiny white flowers and glossy green round to heart-shaped leaves that become coppery to maroon or purplish in fall', '(astronomy)']}	'a system of millions or billions of stars together with gas and dust held together by gravitational attraction'
Hibiscus	{'Noun': ['any plant of the genus Hibiscus']}	'a plant of the mallow family grown in warm climates for its large brightly colored flowers or for products such as fiber or timber'
Homer-Simpson	N/A	'a fictional character and one of the main characters of the American animated sitcom The Simpsons'
Horseshoe-crab	N/A	'a large marine arthropod with a domed horseshoe-shaped shell a long tail-spine and ten legs little changed since the Devonian'
	{'Limulidae': 'Noun': ['horseshoe crabs']}	
Ipod	{'Noun': ['(trademark)']}	'a portable electronic device for playing and storing digital audio and video files'

Minotaur	{'Noun': [(Greek mythology)]}	'a creature who was half man and half bull the offspring of Pasiphae and a bull with which she fell in love'
Segway	{'Noun': [(trademark)]}	'a motorized personal vehicle consisting of two wheels mounted side by side beneath a platform that the rider stands on while holding on to handlebars controlled by the way the rider distributes their weight'
Superman	{'Noun': ['a person with great powers and abilities', 'street name for lysergic acid diethylamide']}	'a US cartoon TV and film character having great strength the ability to fly and other extraordinary powers'
Tuning-fork	N/A	'a two pronged steel device used by musicians which vibrates when struck to give a note of specific pitch'
Yo-yo	N/A	'a toy consisting of a pair of joined discs with a deep groove between them in which string is attached and wound which can be spun alternately downward and upward by its weight and momentum as the string unwinds and rewinds'
Tennis-shoes	N/A	'a light canvas or leather soft-soled shoe suitable for tennis or casual wear'

10 APPENDIX C

Table 10-1. Top 100 Words in Term Frequency List for Chandelier and Mars

Ranking	Chandelier		Mars	
	Word	Count	Word	Count
1	small	91	brain	263
2	observe	68	skull	227
3	person	60	nervous	226
4	determine	53	ability	226
5	light	49	planet	224
6	ones	48	part	189
7	watch	45	hit	170
8	sound	40	certain	164
9	base	40	ones	160
10	look	38	sun	151
11	members	37	ball	144
12	travel	35	mythology	142
13	quickly	35	meat	142
14	period	35	central	142
15	move	35	feelings	140
16	eyes	35	exceptional	140
17	building	35	mass	127
18	making	34	cord	122
19	plural	33	spinal	121
20	living	33	head	120
21	body	33	continuous	120
22	guard	32	body	119
23	effort	31	seat	117
24	worn	30	system	116
25	rapidly	30	enclosed	116
26	objects	30	responsible	115
27	long	30	th	114
28	image	30	satellites	114
29	attention	29	includes	114
30	sides	28	animals	114
31	looking	28	thoughts	113
32	large	28	someones	113
33	act	28	smashing	113
34	wheels	27	reason	113
35	baseball	27	originality	113
36	animal	27	mental	113
37	vigilant	26	kill	113
38	vehicle	26	intellectual	113

39	use	26	higher	113
40	see	26	functions	113
41	religious	26	faculty	113
42	mind	26	conscious	113
43	lookout	26	centers	113
44	little	26	game	94
45	learn	26	astronomy	84
46	inquiry	26	roman	83
47	follow	26	new	82
48	find	26	satellite	76
49	certainty	26	rather	76
50	careful	26	many	76
51	bird	26	move	75
52	attentively	26	light	75
53	area	26	bat	75
54	provide	25	spherical	74
55	illumination	24	playing	73
56	great	24	teams	72
57	away	24	lavish	72
58	garment	23	formal	72
59	come	23	dance	72
60	air	23	particles	71
61	children	22	ice	71
62	tower	21	giant	71
63	somebody	21	largest	70
64	music	21	united	67
65	fly	21	states	67
66	boxing	21	serves	64
67	triangular	20	rounded	63
68	time	20	round	62
69	stock	20	sky	61
70	steps	20	greek	61
71	step	20	born	61
72	shape	20	people	60
73	paper	20	natural	60
74	handle	20	moon	60
75	foot	20	way	59
76	float	20	reproductive	58
77	fire	20	tissue	53
78	device	20	earth	53
79	airplane	20	purpose	52
80	surface	19	night	52
81	portable	19	abnormal	52

82	metal	19	played	51
83	magnifier	19	objects	51
84	hit	19	games	51
85	event	19	players	50
86	wings	18	star	49
87	tuscany	18	atmosphere	49
88	theatrical	18	pitch	45
89	site	18	person	45
90	purposes	18	produce	43
91	playing	18	brightest	43
92	leaning	18	surrounded	42
93	food	18	requiring	42
94	famous	18	cloud	42
95	employed	18	three	41
96	city	18	spermatozoa	41
97	vessel	17	secrete	41
98	timepiece	17	program	41
99	state	17	piece	41
100	plant	17	male	41

Input: mars		Secondary AR Words									
Primary AR Words	1	2	3	4	5	6	7	8	9	10	
1	martian	earth	lunar	meteorite	lander	planet	spacecraft	extraterrestrial	moon	microbe	interplanetary
2	spacecraft	spaceship	orbit	soyuz	orbiter	astronaut	nasa	shuttle	space	satellite	unman
3	lander	orbiter	spacecraft	pathfinder	beagle	martian	lunar	rover	nasa	polar	planetary
4	moon	lunar	earth	ki	spacecraft	planet	orbit	sun	jupiter	saturn	apollo
5	orbiter	spacecraft	lander	shuttle	i	orbit	nasa	atlantis	lunar	astronaut	endeavour
6	earth	planet	orbit	moon	spacecraft	martian	universe	space	planetary	surface	gravity
7	planet	earth	planet	orbit	jupiter	pluto	universe	asteroid	extrasolar	moon	dwarf
8	pathfinder	lander	rover	spacecraft	sojourner	orbiter	spaceship	jpl	martian	pathfinder	odyssey
9	orbit	orbit	spacecraft	orbital	satellite	geostationary	earth	geosynchronous	astronaut	planet	space
10	lunar	moon	spacecraft	martian	orbiter	orbit	lander	eclipse	earth	apollo	solar
11	jupiter	moon	planet	uranus	saturn	neptune	comet	io	orbit	ganymede	asteroid
12	nasa	shuttle	astronaut	spacecraft	space	endeavour	i	atlantis	hubble	aeronautics	orbit
13	rover	jaguar	pathfinder	lander	spacecraft	sojourner	bmw	lr3	carmaker	freelander	volvo
14	astronaut	astronaut	cosmonaut	i	nasa	foale	spacewalk	shuttle	mir	spacecraft	thagard
15	comet	asteroid	halley	comet	spacecraft	meteor	jupiter	tempel	meteorite	orbit	bopp
16	spaceship	spacecraft	soyuz	spaceship	atlantis	shenzhou	astronaut	orbit	unman	capsule	man
17	planetary	planet	earth	orbit	astronomer	atmosphere	asteroid	astronomical	celestial	spacecraft	astronomy
18	galileo	galilei	spacecraft	cassini	copernicus	huygens	ganymede	rosetta	satellite	flyby	telescope
19	mission	mission	peacekeeping	u.n.	nasa	spacecraft	force	un	mandate	space	missionary
20	robotic	robot	spacecraft	unman	submersible	humanoid	robotics	astronaut	underwater	cybernetic	orbiter

Figure 10-1. Primary and Secondary AR Words for Mars

Input: mars		Secondary AR Words									
Primary AR Words	1	2	3	4	5	6	7	8	9	10	
1	martian	earth	lunar	meteorite	lander	planet	spacecraft	extraterrestrial	moon	microbe	interplanetary
2	spacecraft	spaceship	orbit	soyuz	orbiter	astronaut	nasa	shuttle	space	satellite	unman
3	lander	orbiter	spacecraft	pathfinder	beagle	martian	lunar	rover	nasa	polar	planetary
4	moon	lunar	earth	ki	spacecraft	planet	orbit	sun	jupiter	saturn	apollo
5	orbiter	spacecraft	lander	shuttle	i	orbit	nasa	atlantis	lunar	astronaut	endeavour
6	earth	planet	orbit	moon	spacecraft	martian	universe	space	planetary	surface	gravity
7	planet	earth	planet	orbit	jupiter	pluto	universe	asteroid	extrasolar	moon	dwarf
8	pathfinder	lander	rover	spacecraft	sojourner	orbiter	spaceship	jpl	martian	pathfinder	odyssey
9	orbit	orbit	spacecraft	orbital	satellite	geostationary	earth	geosynchronous	astronaut	planet	space
10	lunar	moon	spacecraft	martian	orbiter	orbit	lander	eclipse	earth	apollo	solar
11	jupiter	moon	planet	uranus	saturn	neptune	comet	io	orbit	ganymede	asteroid
12	nasa	shuttle	astronaut	spacecraft	space	endeavour	i	atlantis	hubble	aeronautics	orbit
13	rover	jaguar	pathfinder	lander	spacecraft	sojourner	bmw	lr3	carmaker	freelander	volvo
14	astronaut	astronaut	cosmonaut	i	nasa	foale	spacewalk	shuttle	mir	spacecraft	thagard
15	comet	asteroid	halley	comet	spacecraft	meteor	jupiter	tempel	meteorite	orbit	bopp
16	spaceship	spacecraft	soyuz	spaceship	atlantis	shenzhou	astronaut	orbit	unman	capsule	man
17	planetary	planet	earth	orbit	astronomer	atmosphere	asteroid	astronomical	celestial	spacecraft	astronomy
18	galileo	galilei	spacecraft	cassini	copernicus	huygens	ganymede	rosetta	satellite	flyby	telescope
19	mission	mission	peacekeeping	u.n.	nasa	spacecraft	force	un	mandate	space	missionary
20	robotic	robot	spacecraft	unman	submersible	humanoid	robotics	astronaut	underwater	cybernetic	orbiter

Figure 10-2. Primary and Second AR Words - Identifying Duplicate Primary AR Words (Blue)

11 APPENDIX D

Table 11-1. Definition and Definition Words for First Five Mystery Images

Word(s)	PyDictionary Definition	Definition Words
Boxing	{'Noun': ['fighting with the fists', 'the enclosure of something in a package or box'], 'Verb': ['put into a box', 'hit with the fist', 'engage in a boxing match']}	Fighting, fists, enclosure, package, box, box, hit, fist, engage, boxing, match
Flames	{'Noun': ['the process of combustion of inflammable materials producing heat and light and (often)', 'Verb': ['shine with a sudden light', 'be in flames or aflame', 'criticize harshly, usually via an electronic medium']}	Process, combustion, inflammable, materials, producing, heat, light, shine, sudden, light, flames, aflame, criticize, harshly, electronic, medium
Glove	{'Noun': ['the handwear used by fielders in playing baseball', 'handwear: covers the hand and wrist', 'boxing equipment consisting of big and padded coverings for the fists of the fighters; worn for the sport of boxing']}	Handwear, fielders, playing, baseball, handwear, covers, hand, wrist, boxing, equipment, big, padded, coverings, fists, fighters, worn, sport, boxing
House	{'Noun': ['a dwelling that serves as living quarters for one or more families', 'the members of a business organization that owns or operates one or more establishments', 'the members of a religious community living together', 'the audience gathered together in a theatre or cinema', 'an official assembly having legislative powers', 'aristocratic family line', 'play in which children take the roles of father or mother or children and pretend to interact like adults', '(astrology)', 'the management of a gambling house or casino', 'a social unit living together', 'a building where theatrical performances or motion-picture shows can be presented',	Dwelling, serves, living, quarters, more, families, members, business, organization, owns, operates, more, establishments, members, religious, community, living, together, audience, fathered, theatre, cinema, official, assembly, legislative, powers, aristocratic, family, line, play, children, roles, father, mother, children, pretend, interact, adults, astrology, management, gambling, house, casino, social, unit, living, building, theatrical, performances, motion-picture, shows, presented, building, sheltered, located,

	'a building in which something is sheltered or located', 'Verb': ['contain or cover', 'provide housing for']	contain, over, provide, housing
Light / Lights	{'Adjective': ['of comparatively little physical weight or density', '(used of color', 'of the military or industry; using (or being', 'not great in degree or quantity or number', 'psychologically light; especially free from sadness or troubles', 'characterized by or emitting light', '(used of vowels or syllables', 'easily assimilated in the alimentary canal; not rich or heavily seasoned', '(used of soil', '(of sound or color', 'moving easily and quickly; nimble', 'demanding little effort; not burdensome', 'of little intensity or power or force', '(physics, chemistry', 'weak and likely to lose consciousness', 'very thin and insubstantial', 'marked by temperance in indulgence', 'less than the correct or legal or full amount often deliberately so', 'having little importance', 'intended primarily as entertainment; not serious or profound', 'silly or trivial', 'designed for ease of movement or to carry little weight', 'having relatively few calories', 'or lite', 'or light', '(of sleep', [Explicit], 'or light'], 'Adverb': ['with few burdens'], 'Noun': ['(physics', 'any device serving as a source of illumination', 'a particular perspective or aspect of a situation',	Comparatively, little, physical, weight, density, color, military, industry, using, great, degree, quantity, number, psychologically, light, free, sadness, troubles, characterized, emitting, light, vowel, syllables, assimilated, alimentary, canal, rich, heavily, seasoned, soil, sound, color, moving, quickly, nimble, demanding, little, effort, burdensome, little, intensity, power, force, physics, chemistry, weak, lose, consciousness, thin, insubstantial, marked, temperance, indulgence, less, correct, legal, full, amount, deliberately, little, importance, intended, primarily, entertainment, serious profound, silly, trivial, designed, ease, movement, carry, little, weight, having, relatively, few, calories, lite, light, sleep, light, few, burdens, physics, device, serving, source, illumination, particular, perspective, aspect, situation, quality, luminous, emitting, reflecting, light, illuminated, area, condition, spiritual awareness, divine, illumination, visual, effect, illumination, objects,

	<p>'the quality of being luminous; emitting or reflecting light', 'an illuminated area', 'a condition of spiritual awareness; divine illumination', 'the visual effect of illumination on objects or scenes as created in pictures', 'a person regarded very fondly', 'having abundant light or illumination', 'mental understanding as an enlightening experience', 'merriment expressed by a brightness or gleam or animation of countenance', 'public awareness', 'a divine presence believed by Quakers to enlighten and guide the soul', 'a visual warning signal', 'a device for lighting or igniting fuel or charges or fires', 'Verb': ['make lighter or brighter', 'begin to smoke', 'to come to rest, settle', 'cause to start burning; subject to fire or great heat', 'fall to somebody by assignment or lot', 'alight from (a horse)', 'start or maintain a fire in']}]</p>	<p>scenes, created, pictures, person, regarded, fondly, abundant, light, illumination, mental, understanding, enlightening, experience, merriment, expressed, brightness, gleam, animation, countenance, public, awareness, divine, presence, believed, Quakers, enlighten, guide, soul, visual, warning, signal, device, lightning, igniting, fuel, charges, fires, lighter, brighter, begin, smoke, come, rest, settle, start, burning, subject, fire, great, heat, fall, somebody, assignment, lot, alight, horse, start, maintain, fire</p>
Lightning	<p>{'Noun': ['abrupt electric discharge from cloud to cloud or from cloud to earth accompanied by the emission of light', 'the flash of light that accompanies an electric discharge in the atmosphere (or something resembling such a flash)']}</p>	<p>Abrupt, electric discharge, cloud, cloud, cloud, earth, accompanied, emission, light, flash, light, accompanies, electric, discharge, atmosphere, resembling, flash</p>
Night	<p>{'Noun': ['the time after sunset and before sunrise while it is dark outside', 'a period of ignorance or backwardness or gloom', 'the period spent sleeping', 'the dark part of the diurnal cycle considered a time unit', 'darkness', 'a shortening of nightfall', 'the time between sunset and midnight',</p>	<p>Time, sunset, before, sunrise, dark, outside, period, ignorance, backwardness, gloom, period, spent, sleeping, dark, part, diurnal, cycle, considered, time, unit, darkness, shortening, nightfall, time, sunset, midnight, roman, goddess,</p>

	'Roman goddess of night; daughter of Erebus; counterpart of Greek Nyx']}	night, daughter, Erebus, counterpart, Greek, Nyx
Storms	{'Noun': ['a violent weather condition with winds 64-72 knots (11 on the Beaufort scale', 'a violent commotion or disturbance', 'a direct and violent assault on a stronghold'], 'Verb': ['behave violently, as if in state of a great anger', 'take by force', 'rain, hail, or snow hard and be very windy, often with thunder or lightning', 'blow hard', 'attack by storm; attack suddenly']}]	Violent, weather, condition, winds, knots, Beaufort, scale, violent, commotion, disturbance, direct, violent, assault, stronghold, behave, violently, state, great, anger, take, force, rain, hail, snow, hard, windy, thunder, lightning, blow, hard, attack, storm, attack, suddenly
Tampa	{'Noun': ['a resort city in western Florida; located on Tampa Bay on the Gulf of Mexico', 'plug of cotton or other absorbent material; inserted into wound or body cavity to absorb exuded fluids (especially blood)]}	Resort, city, western, Florida, located, Tampa, bay, gulf, Mexico, plug, cotton, other, absorbent, material, inserted, wound, body, cavity, absorb, exuded, fluids, blood
The	N/A	
Thunder	{'Noun': ['a deep prolonged loud noise', 'a booming or crashing noise caused by air expanding along the path of a bolt of lightning', 'street names for heroin'], 'Verb': ['move fast, noisily, and heavily', 'utter words loudly and forcefully', 'be the case that thunder is being heard', 'to make or produce a loud noise']}]	Deep, prolonged, loud, noise, booming, crashing, noise, caused, air, expanding, along, path, bolt, lightning, street, name, heroin, move, fast, noisily, heavily, utter, words, loudly, forcefully, case, thunder, heard, produce, loud, noise
White	{'Adjective': ['being of the achromatic color of maximum lightness; having little or no hue owing to reflection of almost all incident light', 'of or belonging to a racial group having light skin coloration', 'free from moral blemish or impurity; unsullied', 'marked by the presence of snow', 'restricted to whites only', 'glowing white with heat', 'benevolent; without malicious intent', '(of a surface',	Achromatic, color, maximum, lightness, little, hue, owing, reflection, incident, light, belonging, racial, group, light, skin, coloration, free, moral, blemish, impurity, unsuited, marked, presence, snow, restricted, whites, only, flowing, white, heat, benevolent, malicious, intent, surface, coffee, hair, anemic,

	<p>'(of coffee', '(of hair', 'anemic looking from illness or emotion', 'of summer nights in northern latitudes where the sun barely sets'], 'Noun': ['a Caucasian', 'the quality or state of the achromatic color of greatest lightness (bearing the least resemblance to black', 'United States jurist appointed chief justice of the United States Supreme Court in 1910 by President Taft; noted for his work on antitrust legislation (1845-1921', 'Australian writer (1912-1990', 'United States political journalist (1915-1986', 'United States architect (1853-1906', 'United States writer noted for his humorous essays (1899-1985', 'United States educator who in 1865 (with Ezra Cornell', '1832-1918', 'a tributary of the Mississippi River that flows southeastward through northern Arkansas and southern Missouri', 'the white part of an egg; the nutritive and protective gelatinous substance surrounding the yolk consisting mainly of albumin dissolved in water', '(board games', '(usually in the plural'], 'Verb': ['turn white']}]</p>	<p>looking, illness, emotion, summer, nights, northern, latitudes, sun, sets, Caucasian, quality, state, achromatic, color, greatest, lightness, bearing, least, resemblance, black, United, States, jurist, appointed, chief, justice, United States, Supreme Court, President, Taft, noted, work, antitrust, legislation, Australian, writer, United, States, political, journalist, United, States, architect, United, States, writer, noted humorous, essays, United, States, educator, Ezra, Cornell, tributary, Mississippi, river, flows, southeastward, northern, Arkansas, southern, Missouri, white, part, egg, nutritive, protective, gelatinous, substance, surrounding, yolk, mainly, albumin, dissolved, water, board, games, plural, turn, white</p>
--	---	---

Table 11-2. First 100 Words in Term Frequency List for Mystery Class

Word	Count	Word	Count
light	604	part	84
little	321	time	82
illumination	299	person	80
fire	242	intensity	80
united	219	heavily	79
states	219	source	78
living	195	located	78
color	191	somebody	77
great	175	rest	77
building	167	horse	77
heat	153	understanding	76
device	150	serving	76
visual	148	public	76
divine	148	gleam	76
burning	142	experience	76
discharge	139	effect	76
cloud	135	come	76
children	132	brightness	76
physics	131	particular	75
members	131	force	75
emitting	131	believed	75
fuel	120	warning	74
weight	114	subject	74
white	110	spiritual	74
quality	110	soul	74
presence	110	situation	74
relatively	102	signal	74
electric	100	settle	74
provide	99	scenes	74
marked	97	regarded	74
maintain	96	reflecting	74
free	93	quakers	74
unit	91	pictures	74
flash	90	perspective	74
area	89	objects	74
condition	88	merriment	74
water	85	mental	74
smoke	85		

12 APPENDIX E

Table 12-1. Term Frequency List for AK-47, Cactus, and Fireworks

AK-47		Cactus		Fireworks	
Word	Count	Word	Count	Word	Count
long	67	large	103	large	81
small	61	long	97	small	54
move	59	small	84	long	54
person	48	fungi	75	cloud	49
played	42	edible	67	light	48
baseball	41	body	66	body	43
ball	41	numerous	64	terrestrial	37
travel	39	fleshy	54	noise	36
search	35	food	49	astronomy	34
playing	35	America	48	person	32
body	35	terrestrial	42	played	30
bat	35	central	42	ornamental	30
instrument	32	person	37	electric	29
certain	32	light	37	descent	29
printed	30	head	37	shrub	28
horse	30	fur	37	numerous	28
back	29	sign	36	flash	28
rectangular	28	mushrooms	36	discharge	28
air	28	nocturnal	33	legs	27
six	27	bird	33	food	27
part	27	related	31	loud	26
place	26	north	31	instrument	26
piece	26	limbs	31	grown	26
ones	26	africa	31	french	26
legs	26	horns	30	flowers	26
wood	25	tropical	29	tree	25
thin	25	thin	29	family	25
music	25	shell	29	air	25
hand	25	flesh	29	pot	24
game	25	etc	29	plant	24
paper	24	wood	28	leaves	24
metal	24	plants	28	fungi	24
handle	24	move	28	limbs	23
water	23	grow	28	forests	23
provide	23	genus	28	world	22
go	23	gather	28	leaping	22
firearm	23	American	27	hind	22
games	22	tail	26	ball	22

quickly	21	legs	26	Africa	22
wings	20	end	26	tower	21
tool	20	shaped	25	tiny	21
state	20	hunt	25	shallow	21
lowest	20	fronds	25	evergreen	21
animal	20	coat	25	water	20
sound	19	aquatic	25	tray	20
shoulder	19	upward	24	round	20
shaped	19	eyes	23	move	20
musical	19	common	23	hunt	20
device	19	back	23	horns	20
airplane	19	arboreal	23	dwarfed	20
strings	18	marine	22	deer	20
short	18	hind	22	atmosphere	20
set	18	African	22	upward	19
play	18	wild	21	thunder	19
horses	18	tailless	21	tailless	19
held	18	stoutbodied	21	stoutbodied	19
head	18	species	21	species	19
fly	18	semiaquatic	21	semiaquatic	19
fight	18	reproduce	21	perennial	19
family	18	native	21	lightning	19
cut	18	mushroom	21	larger	19
blade	18	mammal	21	antlers	19
away	18	leaping	21	amphibians	19
written	17	inedible	21	along	19
way	17	human	21	white	18
stiff	17	forests	21	produce	18
relatively	17	cap	21	plants	18
position	17	amphibians	21	people	18
living	17	vascular	20	fleshy	18
flat	17	uncurl	20	fast	18
electric	17	TRUE	20	famous	18
another	17	spread	20	decorative	18
use	16	spores	20	central	18
holding	16	sound	20	tufted	17
fish	16	seedless	20	trees	17
direction	16	roots	20	sky	17
composition	16	rhizome	20	north	17
unauthorized	15	name	20	many	17
time	15	living	20	green	17
strike	15	herbivorous	20	earth	17
someones	15	flowerless	20	chiefly	17

plural	15	come	20	building	17
pass	15	rising	19	splendid	16
member	15	relatively	19	spikes	16
large	15	pick	19	river	16
hit	15	metal	19	purplish	16
edible	15	hawklike	19	maroon	16
board	15	flowers	19	largest	16
belongings	15	beak	19	herb	16
act	15	water	18	heartshaped	16
vehicle	14	underground	18	glossy	16
turn	14	toadstools	18	game	16
support	14	toadstool	18	fall	16
rapidly	14	subdivision	18	edible	16
portable	14	stem	18	coppery	16
may	14	sky	18	coat	16
lamp	14	several	18	branched	16
insect	14	rubble	18	become	16
food	14	puffballs	18	assemblage	16
float	14	nuclear	18	ape	16

Table 12-2. Term Frequency List for Floppy Disk, Frog, and Galaxy

Floppy Disk		Frog		Galaxy	
Word	Count	Word	Count	Word	Count
small	60	large	114	planet	153
ball	49	body	97	sun	135
body	48	fungi	92	mythology	92
long	47	small	80	th	80
device	44	edible	72	small	71
material	43	fleshy	71	roman	63
large	42	person	70	satellite	59
sound	41	numerous	62	satellites	55
instrument	39	move	49	around	51
wood	38	related	48	light	47
worn	37	mushrooms	46	particles	45
use	35	long	45	ice	44
metal	34	shaped	42	giant	44
played	33	america	42	natural	43
game	32	common	41	largest	43
computer	31	ones	40	many	42
open	29	nocturnal	40	numerous	41
pointed	28	mammal	40	moon	39
pierce	28	horse	39	astronomy	39
piece	28	gather	39	sky	38
trademark	27	holding	38	born	38
move	26	people	37	body	36
foot	26	head	37	mars	34
fasten	26	north	36	period	33
equipment	26	legs	36	three	31
container	26	etc	36	night	31
implement	25	certain	35	move	31
place	24	shell	34	metal	31
part	24	marine	34	person	30
light	24	grow	33	color	30
less	24	genus	32	greek	29
water	23	end	32	earth	29
vehicle	23	light	31	surrounded	28
teams	23	african	31	composed	28
rectangular	23	hold	30	rings	27
person	23	food	30	planar	27
hold	23	Africa	30	orbits	27
hand	23	open	29	concentric	27
signals	22	horns	29	celestial	27
shape	22	forests	29	rock	26
holding	22	turtle	28	ones	26
back	22	living	28	objects	26
thin	21	herbivorous	28	mainly	26
objects	21	fast	28	jupiter	26

mass	21	high	27	bodies	26
informal	21	fur	27	united	25
electronic	21	coat	27	states	25
cover	21	back	27	shaped	25
bird	21	water	26	outer	25
spherical	20	omnivorous	26	craft	25
plural	20	central	26	brightest	25
people	20	american	26	fungi	24
lavish	20	tropical	25	cloud	24
hit	20	sky	25	atmosphere	24
formal	20	rising	25	expose	22
dance	20	native	25	edible	22
bat	20	mushroom	25	traveling	21
players	19	inedible	25	technically	21
length	19	cap	25	space	21
leg	19	black	25	capable	21
head	19	stem	24	way	20
games	19	spread	24	successive	19
cloth	19	sea	24	star	19
boxing	19	quadruped	24	religious	19
portable	18	pick	24	new	19
machine	18	oysters	24	musings	19
handwear	18	including	24	moons	19
food	18	heat	24	listless	19
electric	18	great	24	leader	19
become	18	wood	23	Korea	19
surface	17	use	23	idle	19
set	17	unpleasant	23	fantasies	19
relatively	17	underground	23	dreamy	19
playing	17	toadstools	23	dreamlike	19
garment	17	toadstool	23	days	19
covering	17	subdivision	23	buttocks	19
wheels	16	rubble	23	awake	19
wheel	16	puffballs	23	surface	18
wash	16	nuclear	23	large	18
vessel	16	name	23	instrument	18
try	16	mycelium	23	fleshy	18
terms	16	morels	23	ring	17
stick	16	explosion	23	eye	17
illumination	16	dust	23	bluegreen	17
end	16	coral	23	played	16
solid	15	contrasting	23	water	15
rounded	15	cloud	23	plural	15
pad	15	bomb	23	music	15
decorate	15	basidiomycota	23	english	15
contact	15	arising	23	ball	15
communication	15	agaric	23	sign	14

come	15	fruit	22	shell	14
certain	15	vehicle	21	predict	14
blade	15	reptiles	21	persons	14
united	14	relatively	21	newtons	14
states	14	provide	21	motion	14
sounds	14	horses	21	long	14
saddle	14	come	21	laws	14
provide	14	chiefly	21	iron	14
order	14	catch	21	heaven	14

Table 12-3. Term Frequency List for Iguana, Penguin, and People

Iguana		Penguin		People	
Word	Count	Word	Count	Word	Count
long	118	small	128	large	186
small	106	person	106	body	160
large	85	large	101	ball	148
coat	65	body	93	move	143
genus	61	long	64	small	138
legs	59	move	57	person	109
africa	58	light	52	long	98
living	57	water	49	certain	84
vitis	56	head	49	face	78
clusters	56	forests	47	horse	75
horns	55	building	46	vehicle	74
water	53	ape	45	surface	74
terrestrial	52	ones	41	game	74
body	52	vessel	39	ones	73
tropical	48	larger	39	turn	70
person	48	ball	39	use	69
neck	48	horse	38	part	69
leaping	47	Africa	38	bat	69
hind	47	animal	37	metal	67
edible	43	food	36	head	66
limbs	41	vehicle	35	played	65
move	40	wood	34	informal	65
African	40	slender	34	holding	63
quadruped	39	legs	34	catch	61
wine	38	game	34	unpleasant	59
tailless	38	quantity	33	edible	57
stoutbodied	38	tall	32	animal	57
species	38	central	32	without	53
semiaquatic	38	act	32	vessel	53
numerous	38	travel	30	teams	53
fruit	38	signals	30	shaped	53
amphibians	38	hunt	30	fungi	53
tallest	37	back	30	somebody	52
spotted	37	great	29	wheels	51
savannahs	37	whales	28	provide	51
food	37	provide	28	light	51
purple	34	area	28	legs	51
green	34	terrestrial	27	playing	50
open	32	neck	27	genus	50
forests	32	living	27	garment	50

bearing	32	go	27	term	49
shot	31	container	27	moving	49
nocturnal	31	box	27	common	49
woody	30	trees	26	wood	48
travel	30	short	26	woman	48
skins	30	largest	26	open	48
grow	30	upper	25	hold	48
fired	30	sound	25	water	47
black	30	part	25	liquids	46
white	29	numerous	25	domesticated	45
hunt	29	members	25	building	45
vines	28	african	25	players	44
vehicle	28	wheels	24	gun	44
projectiles	28	structure	24	fight	44
produce	28	shaped	24	rear	43
juicy	28	plural	24	horses	43
hail	28	play	24	games	43
cluster	28	marine	24	slender	42
cannon	28	liquids	24	man	42
berries	28	holding	24	front	42
spots	25	glass	24	shot	41
french	25	fish	24	padded	41
feline	25	etc	24	area	41
vessel	24	bird	24	worn	40
tawny	24	bat	24	white	40
pelt	24	without	23	numerous	40
marine	24	west	23	hit	40
leopard	24	vegetarian	23	fleshy	39
descent	24	use	23	shape	38
bird	24	open	23	direction	38
asian	24	children	23	back	38
loop	23	arboreal	23	typically	37
insect	23	anthropoid	23	manner	37
fur	23	worn	22	male	37
frogs	23	top	22	wheel	36
decorative	23	somewhat	22	trained	36
cord	23	somebody	22	neck	36
braid	23	smoke	22	living	36
animal	23	party	22	family	36
wheels	22	metal	22	etc	36
mammal	22	mammals	22	baseball	36
hold	22	introduced	22	another	36
head	22	intelligent	22	united	35

central	22	herbivorous	22	states	35
three	21	fire	22	cover	35
tail	21	equatorial	22	boxing	35
heat	21	converts	22	supports	34
eyes	21	come	22	opposite	34
several	20	warm	21	mass	34
ride	20	size	21	horseback	34
claws	20	rectangular	21	horizontally	34
holding	19	qualities	21	herbivorous	34
bed	19	quadruped	21	great	34
ones	18	public	21	glass	34
north	18	little	21	circular	34
great	18	illumination	21	american	34
prey	17	horses	21	wheeled	33
america	17	gases	21	ride	33
use	16	domesticated	21	prevent	33
shell	16	device	21	marine	33

Table 12-4. Term Frequency List for Sheet Music, Skyscraper, and Swiss Army Knife

Sheet Music		Skyscraper		Swiss Army Knife	
Word	Count	Word	Count	Word	Count
small	137	light	84	small	93
rectangular	88	tower	71	ball	76
area	86	small	52	instrument	71
glass	85	building	47	body	71
box	76	little	43	device	60
plural	67	large	43	played	59
quantity	65	illumination	40	move	53
place	65	slender	36	back	51
food	60	balconies	35	long	49
hit	56	fire	30	person	48
container	56	members	28	large	46
grain	49	great	28	water	43
metal	48	united	27	signals	43
hand	46	states	27	game	42
field	46	ball	27	computer	40
ball	45	move	26	use	39
boxing	44	somebody	25	head	38
people	43	ones	25	sound	37
separate	40	color	25	pad	37
group	39	device	24	part	36
contained	39	come	24	pointed	33
seat	38	central	24	games	33
foot	38	body	24	electronic	33
engage	38	person	23	ones	32
body	38	living	23	tool	31
watch	37	area	23	converts	31
several	37	use	22	container	31
public	37	players	22	handle	30
drawing	37	game	21	rectangular	29
device	37	metal	20	playing	28
blow	37	forms	20	numerous	28
batter	37	equipment	20	sounds	27
transparent	36	burning	20	play	27
trees	35	act	20	blade	27
sound	35	upper	19	teams	26
lens	35	tuscany	19	sign	26
holding	35	steps	19	sharp	26
theater	34	site	19	open	26
starchy	34	objects	19	direct	26
skillful	34	leaning	19	around	26

shrubs	34	heat	19	recording	25
private	34	famous	19	mass	25
prepared	34	city	19	bat	25
predicament	34	bc	19	vehicle	24
positioned	34	bat	19	surface	24
performance	34	visual	18	place	24
partitioned	34	signals	18	food	24
match	34	public	18	electric	24
impossible	34	playing	18	edge	24
grandstand	34	material	18	relatively	23
grains	34	italy	18	lavish	23
graceful	34	hoop	18	held	23
fist	34	divine	18	act	23
evergreen	34	cross	18	spherical	22
escape	34	boxing	18	rounded	22
ear	34	worn	17	piece	22
drivers	34	water	17	formal	22
designated	34	played	17	equipment	22
coaches	34	place	17	dance	22
coach	34	physics	17	plural	21
catcher	34	marked	17	wash	20
areas	34	long	17	vessel	20
water	33	lighting	17	try	20
game	33	emitting	17	thin	20
sugar	31	connected	17	television	20
material	31	climbing	17	shot	20
furnish	31	children	17	shape	20
eye	31	weight	16	players	20
mirror	30	vehicle	16	hit	20
large	30	tall	16	go	20
piece	28	sound	16	front	20
brake	28	row	16	bed	20
use	27	relatively	16	activity	20
plate	27	provide	16	wood	19
hold	26	progress	16	slender	19
whose	24	people	16	ride	19
sweet	24	part	16	communication	19
played	24	forests	16	certain	19
liquids	23	etc	16	wheels	18
hard	23	back	16	travel	18
set	22	ascending	16	tails	18
part	22	vessel	15	strike	18
held	22	quantity	15	recorder	18

flat	22	games	15	received	18
structure	21	fuel	15	portable	18
muscles	21	four	15	human	18
light	21	designed	15	end	18
coat	21	circular	15	electrical	18
vision	20	bridge	15	ears	18
open	20	baseball	15	bottom	18
facial	20	activity	15	typically	17
defective	20	unravelled	14	transmitted	17
correcting	20	unraveled	14	solid	17
ankle	20	undone	14	several	17
playing	19	teams	14	set	17
plants	19	structure	14	screen	17
games	19	stitches	14	round	17
drinking	19	stages	14	rodents	17
cereal	19	source	14	rats	17
card	19	snagging	14	magnetic	17

Table 12-5. Term Frequency List for T-shirt and Waterfall

T-shirt		Waterfall	
Word	Count	Word	Count
ball	383	fungi	157
light	289	large	133
game	270	fleshy	117
face	253	body	103
small	243	edible	93
body	224	mushrooms	78
played	217	long	69
covering	209	numerous	63
activity	194	small	59
cover	187	person	55
person	180	related	51
players	172	open	47
teams	171	fast	47
instrument	166	etc	46
conceal	161	common	46
worn	159	cloud	46
people	158	shaped	44
playing	147	mushroom	44
ones	145	inedible	44
mask	144	gather	44
metal	142	cap	44
appearance	133	end	43
sound	131	sky	41
part	131	move	41
equipment	127	including	41
move	126	water	40
hit	126	subdivision	40
bat	126	dust	40
mass	125	basidiomycota	40
games	124	underground	39
material	120	toadstools	39
act	118	toadstool	39
boxing	117	stem	39
circular	115	spread	39
use	113	rubble	39
spherical	109	rising	39
large	109	puffballs	39
lavish	106	pick	39
dance	106	nuclear	39
formal	105	name	39

try	104	mycelium	39
protective	103	morels	39
head	100	grow	39
great	100	explosion	39
reproductive	98	coral	39
hand	98	contrasting	39
handwear	96	bomb	39
shape	95	arising	39
surface	94	agaric	39
sport	92	heat	36
food	91	native	35
illumination	90	use	34

13 APPENDIX F

Table 13-1. Additional Runs Definition Evaluation Results

Removed Class	Original Word			Synonym Word			All Words			All Synonym Words		
	# Found	Total	Ratio (%)	# Found	Total	Ratio (%)	# Found	Total	Ratio (%)	# Found	Total	Ratio (%)
Ak-47	0	4	0	0	42	0						
Chandelier	0	6	0	1	116	0.9						
Fireworks	1	8	12.5	4	250	1.6						
Mars	4	17	23.5	3	276	1.1	7	34	20.6	6	634	0.9
Pyramid	0	7	0	1	127	0.8	2	53	3.8	10	1284	0.8
Skyscraper	1	4	25	0	58	0						
Superman	0	11	0	2	211	0.9						
Swiss Army Knife	1	6	16.7	1	150	0.7						
Teddy Bear	2	13	15.4	1	156	0.6	5	50	10	10	780	1.3
Tricycle	5	6	83.3	5	130	3.8						

Table 13-2. Addition Runs AR Evaluation Results

Removed Class	# of Primary Words Found	# of Secondary Words Found	Total # of Secondary Words	Secondary Words Ratio (%)
Ak-47	0	0	71	0
Chandelier	0	0	145	0
Fireworks	1	2	127	1.6
Mars	3	4	81	4.9
Pyramid	1	2	129	1.6
Skyscraper	2	2	112	1.8
Superman	0	0	105	0
Swiss Army Knife	0	0	150	0
Teddy Bear	0	3	136	2.2
Tricycle	1	3	79	3.8

Table 13-3. Addition Runs Modified HF Evaluation Results

Removed Class	Overall Rating (1-5)	# of "Good" Words in Top 20
Ak-47	2	7
Chandelier	1	4
Fireworks	5	9
Mars	5	6
Pyramid	2	6
Skyscraper	3	4
Superman	2	3
Swiss Army Knife	3	8
Teddy Bear	1	3
Tricycle	4	8