MACHINE TRANSLATION WITH IMAGE CONTEXT FROM MANDARIN CHINESE TO ENGLISH

THESIS

Brooke E. Johnson, Second Lieutenant, USAF

AFIT-ENG-MS-19-M-035

DEPARTMENT OF THE AIR FORCE
AIR UNIVERSITY

AIR FORCE INSTITUTE OF TECHNOLOGY

Wright-Patterson Air Force Base, Ohio

The views expressed in this thesis are those of the author and do not reflect the official policy or position of the United States Air Force, Department of Defense, or the United States Government. This material is declared a work of the U.S. Government and is not subject to copyright protection in the United States.

AFIT-ENG-MS-19-M-035

MACHINE TRANSLATION WITH IMAGE CONTEXT FROM MANDARIN CHINESE TO ENGLISH

THESIS

Presented to the Faculty

Department of Electrical and Computer Engineering

Graduate School of Engineering and Management

Air Force Institute of Technology

Air University

Air Education and Training Command

In Partial Fulfillment of the Requirements for the

Degree of Master of Science in Electrical Engineering

Brooke E. Johnson, BS

Second Lieutenant, USAF

March 2019

AFIT-ENG-MS-19-M-035

MACHINE TRANSLATION WITH IMAGE CONTEXT FROM MANDARIN CHINESE TO ENGLISH

Brooke E. Johnson, BS

Second Lieutenant, USAF

Committee Membership:

Dr. Brett J. Borghetti
Chair

Lieutenant Colonel Alan C. Lin
Member

Captain Richard P. Uber
Member

AFIT-ENG-MS-19-M-035

Abstract

Despite ongoing improvements in machine translation, machine translators still lack the capability of incorporating context from which source text may have been derived. Machine translators use text from a source language to translate it into a target language without observing any visual context. This work aims to produce a neural machine translation model that is capable of accepting both text and image context as a multimodal translator from Mandarin Chinese to English. The model was trained on a small multimodal dataset of 700 images and sentences, and compared to a translator trained only on the text associated with those images. The model was also trained on a larger text only corpus of 21,000 sentences with and without the addition of the small multimodal dataset. Notable differences were produced between the text only and the multimodal translators when trained on the small 700 sentence and image dataset, however no observable discrepancies were found between the translators trained on the larger text corpus. Further research with a larger multimodal dataset could provide more results clarifying the utility of multimodal machine translation.

Acknowledgments

I wish to express my gratitude my thesis committee for their active support in my research. To Dr. Borghetti for his research guidance, to Lt Col Lin for initial research support, hiring intern assistance, and his continued involvement despite the distance, and to Captain Uber for his support in language, working with me to hire another intern, and maintaining contact with my sponsor organization.

I would also like to extend my thanks to the two wonderful interns whose assistance provided much help. To Ms. Leong for the creation of the multimodal dataset used in this research, and to Mr. Slater for his hard work and thorough programming.

Finally, I would like to thank my family for their care and encouragement, and my cat, Eugene for his company throughout my study at AFIT.

Brooke E. Johnson

Table of Contents

Page

List of Figures

List of Tables

NEURAL MACHINE TRANSLATION WITH IMAGE CONTEXT FROM MANDARIN CHINESE TO ENGLISH

I. Introduction

1.1 Importance and Motivation

Language translation is important for many facets of life. From personal travel and businesses, to government and military affairs, communication across language barriers happens everywhere. Because of this, research to develop machine translators has been conducted for decades [1]. However, due to the difficulty of natural language processing, many machine translators still perform worse than humans. This is especially true of Mandarin Chinese due to grammar differences, complexity of the writing system, cultural idioms, and textual context.

One of the main differences between human translation and machine translation in regards to is that current translation technology relies solely on text input during translation. Humans, on the other hand, are capable of processing many other features. In any naturally occurring environment, humans have access to more modes than text, unlike a machine translator. Many polyglots consider visual context a very important modality when using a foreign language. In multimodal translation studies, human translation significantly drops in quality when lacking any image context [2]. Consider a situation in which homographs or unknown words have been used. With image or other situational context, despite the unknown words, humans are capable of understanding the information being conveyed. For example, in English a sentence using the homograph, “bank” could cause confusion if incorrectly translated. (Is someone talking about the ground at the edge of a river or a financial institution?) This uncertainty in translation brings attention to a potential need for machine translation improvement.

1.2 Problem Statement

There are still many translations that are ambiguous or impossible to translate without situational context. This situational context can be input through an extra mode, while still including the available text [3]. Multimodal machine translation could provide the next improvement to machine translation quality, approaching human level parity.

While multimodality within machine translation has only been applied to bilingual image captioning, there is a significant lack of research regarding multimodal translation of naturally occurring scenes [4]. With the capabilities of deep learning combined with the advantages of situational context in translation, there is a need for true multimodal machine translation.

Consider a machine translator that is capable of more than a text modality. With access to the same context and information that a human might have, as well as training using this information, the performance of a machine translator ought to be improved to approach an improved level of translation quality. With better translation quality in machine translators, many currently ambiguous translations could be made certain or even corrected.

1.3 Research Contributions

This research produces a multimodal machine translation system from Mandarin Chinese to English. Currently, as there are no machine translators that make use of further modes than text in natural settings, there is a large gap in regards to translation. This work provides a preliminary approach to fill the gap through the incorporation of image context modality into neural machine translation. This work produces and evaluates a multimodal translator model using neural machine learning methods.

Machine translation specifically fits into the domain of natural language processing. By producing a machine translator that is more capable of translating from Chinese to English, this could save a significant number of man-hours required for hand translation of text. This work is of interest to the Air Force and the Department of Defense as due to the necessity of international communications. Not only is translation useful to the government, but research in this area in any language could also improve business and personal endeavors of people all over the world.

1.4 Research Questions

The goal of this research is to produce and evaluate the differences between neural machine translation models that can accept only text and a translator that can accept both text and image input. The comparison will be conducted with a nonbiased machine translation scoring mechanism. The score comparisons between each translation, as well as the multimodal architecture to be developed will provide insight to the uses and performance of machine translation using image context. To evaluate the success of this objective, the following research questions will be explored.

1) Does addition of image tags/labels improve translator performance, and if so, by how much?

Determining the answer to this question was conducted by comparing the performance of a multimodal translator to the text-only translator. A statistically significant translation improvement provides evidence suggesting the usefulness of image classification tags for translation.

2) Is there a repeated topic/structure/word composition with respect to sentences improved through multimodal translation?

This question is answerable through a by-hand analysis of the sentences translated from the multimodal translation model verses the baseline text-only translator. Checking what homographs have been improved through translation could provide evidence that an uncertain translation was corrected via image context. The number of improved sentences, as well as the percentage of improvement provides proof that translations were corrected by the inclusion of image context. Sentences will be considered for grammar, word count, use of uncommon words, and inclusion of homographs.

3) Do certain image label topics/part of speech/repetition improve translation performance?

This question is answered through a by-hand analysis of the image label context used to derive the translated sentences in the test set. Translated sentences with and without variation between the test set without image labels and the test set with image labels will be considered for the possible effects of the image labels. Image tags associated with those sentences will be considered for their effect on sentences for the length of tag words, accuracy, and use of identical words or synonyms to words used in a sentence translation.

1.5 Methodology

Machine translation is typically approached as a sequence to sequence problem. This thesis introduces a novel approach to machine translation through incorporation of image context within a sequence to sequence framework. The neural machine translation architecture is modeled after a widely accepted approach to neural machine translation with attention by Bahdanau *et al.* [5]. Bahdanau's translation model is a common starting point for research in text translation, employing a sequence to sequence modeling approach using recurrent neural networks with attention. A new translator model was produced by extending Bahdanau's work by training with a text dataset that contains image tags or labels of items and events within an image. This multimodal dataset produced a resulting multimodal neural machine translation model that will be tested for its performance against a text-only translation model. The comparison between the models may provide evidence that neural machine translation is a good approach to improve Chinese-to-English translators. Translation evaluations have been conducted by comparing the image translator to the text only translator, also comparing both to Google Translate as a baseline. The translation scoring was conducted using BLEU, (Bilingual Evaluation Understudy) [6].

As bilingual image data is fairly lacking, the visual input has been derived from the automatic image tagging system, Google Image Tagger API [7]; image tag generation is not part of this research. This work direction is not only novel in the field of machine translation but could provide much-needed improvement in interpreting such a context rich language as Mandarin Chinese. The use of image tags in translation is novel in machine translation in general and could provide much needed clarifications specifically for Mandarin Chinese.

1.5.1 Datasets

Two datasets were used to train the multimodal translator. The first dataset was a standard text-only parallel corpus from Tatoeba Project containing 21,000 identical-meaning sentences translated in both Mandarin Chinese and English [8]. This dataset was used for the initial training of the translator model. A novel dataset was produced and then used to train the translator on image context: a collection of 700 images containing naturally occurring Chinese text or audio. The transcribed Chinese text associated with each image was translated to English, producing a multimodal parallel corpus. Training for the text-only dataset was conducted on the majority of the 21,000 sentence pairs, with the exception of 200 sentences removed for a test set. The translation model for the image dataset was trained on the same text corpus plus the inclusion of 600 bilingual Chinese and English text associated images. The multimodal translation model was evaluated by a test set containing 100 images and their associated text.

1.6 Assumptions and Limitations

The relevance of this research rests upon the usability and translation quality of the datasets upon which the translation models were trained. The datasets contain Chinese and English text. The Chinese text is assumed to be naturally occurring Mandarin Chinese sentences and phrases produced by humans. Naturally occurring text is anything translated from Chinese to English, that occurred during some real-world event rather than machine created text. It is also assumed that the English text is correctly translated from the source Chinese. A limitation of the translation is that even correctly translated text is ambiguous for any grammatically-incorrect text of Chinese origin. Consider translating a familiar or grammatically incorrect word or phrase, such as, “gonna” to another language. Should it first be converted into a proper English future tense “going to,” which could be translated into “travel to” or “headed towards” in the target language, or should it be translated to an idiomatically similar but grammatically incorrect verbal abbreviation suggesting a future state (“I’m gonna be mad”; “That’s gonna break”)? This type of translation is left to the discretion of human translators. Because the target use of the language often depends upon a known audience, humans who translate informal text or dialogue tend to be consistent with each respective situational translation. The assumption must then be made that the machine translator will also perform consistently, with the limitation that the machine translator cannot direct any translations at a known target audience. While the machine translator may approach human translation quality, machine-learning-based translator performance is limited by the quality of the available datasets. Aside from the limitations of the correctness of the datasets, the image dataset is also a restrictive size for the machine learning task, because small datasets sometimes result in poor machine learning models, and larger datasets typically correlated with better performing models. The small dataset of images may limit the quality of the multimodal translation model.

1.6.1 Assumptions

1) The text dataset contains natural Chinese words and sentences.

2) The image dataset contains images with naturally occurring Chinese language.
3) Both the text and image dataset contain correctly translated English text.

1.6.2 Limitations

1) Translations of informal or grammatically incorrect Chinese text cannot be directed at a known target audience due to the deterministic nature of a machine translator. (As mentioned above, translation of informal speech is often directed at a target audience by a human translator, with respect to the regional specifics of the audience.)
2) The translation models are based on the correctness of the datasets.
3) The size of the image dataset is very restrictive for the machine learning task, as the quality of a translation model is based on the amount of available data.

1.7 Organizational Preview

This chapter motivated a current translation problem, presented several research questions, and described important terminology regarding natural language processing and Chinese writing. Background information on machine translation and various machine learning tools and techniques are found in Chapter 2. Following that is a description of the developed Chinese to English translation model, including text preprocessing steps and the multimodal machine learning model. The architecture of the sequence-to-sequence model, as well as a description of each layer can be found in Chapter 3. In Chapter 4, results of translation performance evaluations using BLEU provide information regarding the usefulness of multimodal neural machine translation when translating from Chinese to English. Finally, conclusions and future work can be found in Chapter 5, with detailed recommendations to future approaches on the problem of multimodal machine translation.

II. Literature Review

2.1 Chapter Purpose

While some research evaluates the performance of multilingual machine image captioning systems, almost no research has been executed in regards to translation with image context. Despite that, what multimodal research exists still provides understanding of natural language processing with deep learning. For better understanding of the experiment, the following subsections contain explanations of Chinese language terminology, as well as translation methods pertaining to neural machine learning. Selected works also detail the current capabilities of translation of image description captions, with the purpose of understanding how neural machine translation works in a multimodal setting.

2.2 Natural Language Processing

Natural language refers to language that has developed naturally as it is used by humans. Natural language appears in everyday conversations, books, street signs, or any other use of language by a human. Natural language processing aims to imitate, or otherwise analyze language in areas of speech recognition, translation, and other language uses.

2.2.1 Chinese Specific and Translation Terminology

For understanding of natural language processing of Mandarin Chinese, Table 1 shows the composition of Chinese writing, and is included in addition to several terms pertaining Chinese and general language that are explained as follows:

- Granularity

  In machine translation, during encoding, granularity refers to the separation and tokenization of the text to be encoded. For most languages, the granularity split is along each word. For Chinese (and other languages without spaces,) this can be split along a phrase, word, subword, or character [9].

- Word

  While the separation of words is very straightforward in English text, in Chinese, words are not separated, and are sometimes a combination of characters, or just a character itself.

- Character

  A Chinese character is a single logogram representing a word or a portion of a word. A full Chinese character can be made up of one or more radicals.

- Radical

  A radical is a combination of strokes, which in turn are used to make a full Chinese character. A large portion of radicals are a full character on their own.

- Stroke

  A stroke is a basic symbol used to form a Chinese radical. Many strokes do not make up a full radical or have represent any kind of meaning until combined into a radical; however, in some rare cases they can.

Stroke	**Radical**	**Character**	**Word**
亅	子 Child	学 Learn	学生 Student
一 One	生 Part of: raw, health, unripe, crude, etc.	生 (Same as radical)	生的 Raw

Table 1: Depiction of written Chinese structure with English translations

2.2.2 Modeling and Evaluation of Natural Language

Some form of grading is necessary to evaluate the quality of machine translations. While this task used to be completed by hand, there are tools now that are capable of machine trading by comparison of a reference to a machine translated candidate sentence. One of such tools is BLEU (Bilingual Evaluation Understudy), an automatic and thus low cost method for evaluation of Machine Translation. BLEU uses a modified *n*-gram precision metric for such measurements [10]. Both concepts can be seen in the following subsections.

2.2.2.1 *N*-Gram Model

In natural language processing, an *n*-gram is a contiguous subsequence of a sequence typically made up of words in a sentence. Each *n*-gram contains a set of *n* items of text [11]. The purpose of *n*-gram representation of text is to provide a model in which sentences can be systematically divided for mathematical applications such precision scoring [12]. The terminology for *n*-gram follows as: (n = 1), bigram (n = 2), trigram (n = 3), etc.

2.2.2.2 BLEU Scoring

BLEU is useful because it eliminates the need for human evaluation of translations by providing automatic translation scoring. It is used by comparing a candidate sentence to one or more reference sentences [6]. In the case of machine translation, the candidate sentence is the machine translated sentence to test, and the reference is a human translated comparison.

A portion of the scoring metric by BLEU is a modified *n*-gram precision computed over blocks of text [6]. The score evaluation is completed by comparing reference sentences to candidate sentences over a corpus of text. *N*-gram precision is found by counting maximum number of *n*-gram matches in a single reference translation for each *n*-gram for each candidate sentence. (The denominator of the equation contains a tic-mark representing the counts only of the candidate sentence.) Then the total number of matches of a candidate *n*-gram is clipped for each *n*-gram for each candidate sentence by the maximum reference match. All clipped matches over all candidate sentences are added for each n-gram over all candidate sentences in the corpus. Finally, the total number of unclipped candidate *n*-gram counts in the corpus are divided for each *n*-gram. [13]. The ranking system using *n*-gram precision has been shown to differentiate between human and machine translations with very strong differentiation on 4-gram precision [6]. The equation for this over an entire text corpus, as well as an example for unigram precision can be found below.

$$p_n = \frac{\sum_{c\in\{\text{candidates}\}} \sum_{\text{n-gram}\in c} \text{count}_{\text{clip}}(\text{n-gram})}{\sum_{c'\in\{\text{candidates}\}} \sum_{\text{n-gram}'\in c'} \text{count}_{(}\text{n-gram}')} \tag{1}$$

Table 2: Unigram Precision Example

Candidate: the the the the the the the
Reference: The dog is in the yard.
Unigram precision: 7/7
Modified unigram precision: 2/7

The bigram precision in the reference sentence would be 0 because it does there are no two directly sequential words in the candidate sentence that matches the reference sentence. Table 3 contains a new example for bigram precision with an incorrect but better candidate sentence. Using 1, the *n*-gram precision can be found over a single example for the bigrams in the example below. Taking only part of the summation, the number of bigram matches can be found. In this example the bigram match count is 5, but due to repetition the number of clipped bigrams is 4, because there are only four unrepeated word pairs that can be found in both the candidate sentence and reference sentence. This means that the modified bigram precision is then the clipped bigram count divided by the number of bigrams in the candidate, so the bigram precision is p_2 = 0.667. This equation can be used for each *n*-gram variation to find the modified *n*-gram precision in every sentence in a text corpus to score.

Table 3: Bigram Precision Example

Candidate: The dog the dog is the yard.
Reference: The dog is in the yard.
Bigrams in candidate sentence:
the dog, dog the, the dog, dog is, is the, the yard
Bigram matches:
the dog, the dog, dog is, the yard
Clipped bigrams in candidate sentence:
the dog, dog is, the yard
Bigram precision: 5/6
Modified bigram precision: 4/6

The overall translation score of a translated text corpus is from the geometric mean of the corpus' modified precision scores, then multiplied by an exponential brevity penalty factor. First the geometric mean of the modified *n*-gram precisions, p_n is found, using *n*-grams with maximum length of *N* and the positive weights w_n that sum to one. The positive weights represent the graded preference of which *n*-grams to consider the most. Conventionally, BLEU-4 scores are found, using weights of ¼ for each *n*-gram up to a 4-gram. Then the length of the candidate translation *c* and the effective reference corpus length *r* are used to compute a brevity penalty BP as shown in 2. The brevity penalty is used as an adjustment factor to reduce translation scores for sentences that are too short. If the machine translation candidate is longer than the reference sentence, then there is no penalty. The length of the translation is most often calculated using the word count for languages that use phonetic alphabets, however, character count can be used for languages with differing writing styles.

$$\mathrm{BP} = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r/c)} & \text{if } c \leq r \end{cases} \tag{2}$$

The total BLEU score is then found by multiplying the brevity penalty by the sum of the positive weights to *N*, by the log of the modified *n*-gram precisions as seen in 3. The overall BLEU score is always a number from 0 to 1, often represented as a percentage [6].

$$\text{B}\textsc{leu}= \mathrm{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \tag{3}$$

While BLEU scores provide an established ability to properly rank translation accuracy, note that few sentences receive a perfect translation score of 1 because few translations are entirely identical to their reference. With more reference sentences most translations receive improved scores due to the higher flexibility of word order or synonym use. On a test corpus containing nearly 500 sentences, a human translator received BLEU scores of 0.3468 against four references, but only 0.2571 against two references [6]. Note that this is not necessarily because the human translations were wrong, but caused by the variation in language for sentences that may carry the same meaning.

2.3 Machine Learning

Neural networks can be used to predict patterns based on training data. Deep neural networks are powerful machine learning models that can provide tremendous results on even difficult machine learning tasks. It can be applied to any problems that can be encoded into vectors with fixed or variable dimensionality [14]. There are many subtopics within machine learning. The following topics which pertain to translation using machine learning will be discussed throughout this section of Chapter II: embeddings, encoder-decoder, sequence to sequence models, recurrent neural networks, gated recurrent units, teacher forcing, and attention.

2.3.1 Embeddings

Embeddings are used to represent data in a more efficient space than one hot encoding. One hot encoding which converts categorical values into integers, while simple can take up very high dimensional spaces using up memory and slowing down processes. Embeddings alleviate the problem of high dimensionality and space usage due to their low dimensionality and use of floating-point vectors over binary classification [15]. Word embeddings are learned from data and are meant to map human language into a geometric space. They are learned from starting with random word vectors and are then sorted through use of a neural network [16].

Embedding is a problem of optimizing the loss function between a pair of examples for each point x_i to find an embedding $f(x_i)$. This is shown in the

4, where L represents the loss function, α represents learning parameters subject to a balancing constraint, W represents weights as a matrix of similarity between examples x_i and x_j. By minimizing this summation, the embedding vector $f(x_i)$ is found [16].

$$\sum_{i,j=1}^{U} L(f(x_i,\alpha), f(x_j,\alpha), W_{ij}) \tag{4}$$

The loss function L is found using multidimensional scaling in order to preserve distance between points and embeds them into a low dimensional space, as shown in 5.

$$L(f_i, f_j, W_{ij}) = (||f_i - f_j|| - W_{ij})^2 \qquad 5$$

Embedding vectors for learning neural networks are made by learning a model with layers of non-linear mappings with N layers of hidden units that give a C-dimensional output vector as shown in 6 [16]. In this equation, $f_i(x)$ represents the embedding vector to be learned, x represents an example, M and U are the total number of examples x_i.

$$\sum_{i=1}^{M} \ell(f(x_i), y_i) + \lambda \sum_{i,j=1}^{M+U} L(f^k(x_i), f^k(x_j), W_{ij}) \qquad 6$$

Overall, embeddings can simply be thought of as a vector representation of some token. In the case of word embeddings, this is often done by segmentation of the text into words wherein each word is transformed into a vector. The embedding layer is then a dictionary that can map integer indices to dense vectors, saving dimensionality from the alternative approach of using one-hot vectors [15].

2.3.2 Encoder-Decoder

Neural machine translation models are built with an encoder to decoder architecture. The purpose of an encoder is to extract a fixed-length representation from some input sequence of variable length. The fixed-length representation is often smaller than the initial input sequence. The fixed-length representation is then forwarded to the decoder network to generate its own sequence output. The decoder's purpose is the generation of a variable length translation from this fixed-length representation [17]. This allows for the translation from one language to the other by input of a source language into the encoder, and then output from the decoder into the target language. A simple diagram of this applied to neural machine translation can be seen below in Figure 1: Basic Encoder-Decoder Architecture.

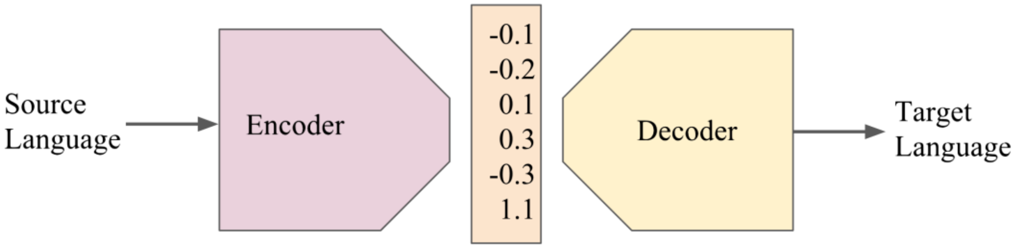


Figure 1: Basic Encoder-Decoder Architecture for Machine Translation

2.3.3 Sequence to Sequence

Sequence to sequence models are useful for many applications of problems in machine learning because they can be used on problems containing vectors with unknown dimensionality [14]. Sequence to sequence models typically have an encoder and a decoder part, that are separate neural network models combined for the purpose of one problem. At the core, sequence to sequence learning makes use of recurrent neural networks in order to map variable-length input sequences to variable length output sequences. Typically, a recurrent neural network layer acts as an encoder, processing input and returning own internal state, and another recurrent neural network acts as the decoder, predicting next portion of target sequence. (Recurrent neural networks are described in

Section 202.3.4.) Sequence to sequence models are useful for machine translation, image captioning, constituency parsing, and other tasks [18].

The aim of sequence to sequence learning is to directly model the conditional probability of mapping an input sequence into an output sequence using an encoder-decoder framework [18]. From the input representation, the output sequence is generated one unit at a time by the decoder. The conditional probability is defined in

7, where *y* represents an output sequence, *x* represents an input sequence, and ***s*** represents the fixed-length encoder representation of an input sequence [18].

$$\log p(y|x) = \sum\nolimits_{j=1}^{m} \log p\left(y_j | y_{<j}, x, \boldsymbol{s}\right) \qquad 7$$

2.3.4 Recurrent Neural Networks

Recurrent neural networks (RNNs) are a type of neural network with the ability to process sequential data. The RNN is a generalization of a feedforward neural network to sequences [19]. RNNs can scale to much longer sequences than networks without flexible sequence-based specialization [20]. Many other neural networks are limited to accepting only fixed-sized vectors as input [21]. Recurrent neural networks can accept sequences of information as input, output, or both.

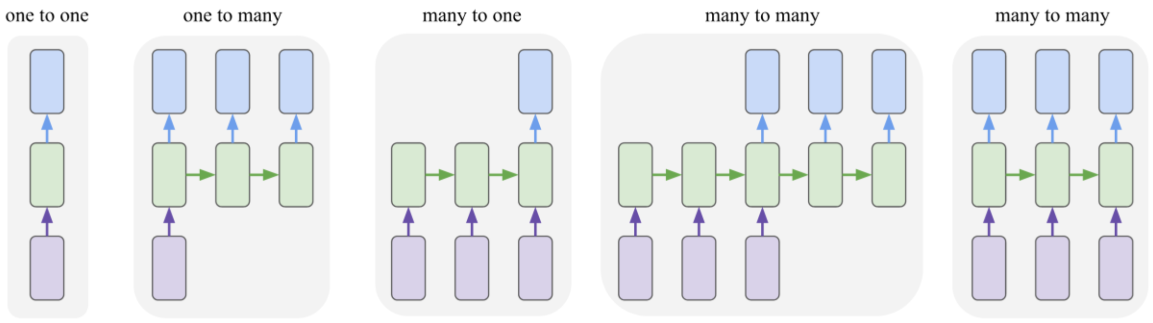


Figure 2: Flexibility of Recurrent Neural Networks

In Figure 2, the flexibility offered by RNNs can be seen. In this image, each rectangle represents a vector, and each arrow represents a function. Input vectors are shown in purple, output vectors are in blue, and internal states are represented by green arrows. An RNN combines the input vector with the state vector and a predetermined function to produce a new state vector [21]. RNNs are able to handle variable-length sequences through use of a recurrent hidden state, where the hidden state activation is dependent on the previous states. Given a sequence $x = (x_1, x_2, \ldots, x_T)$, and a recurrent hidden state h_t, an RNN can compute a sequence of outputs $y = (y_1, y_2, \ldots, y_T)$ through use of the following equations [19]. In Equation 8 the function g is a smooth, bounded function, often a logistic sigmoid, or hyperbolic tangent function [22].

$$h_t = g(W^{hx}x_t + W^{hh}h_{t-1}) \qquad 8$$

$$y_t = W^{yh}h_t \qquad 9$$

Without modification, simple RNNs are not able to solve problems effectively [22]. To counter this, generative models can be made. A generative RNN allows for the output of a probability distribution over the next element of the sequence, given the current state h_t. The generative model is able to capture a distribution over variable length sequences. This provides a sequence probability, which can be seen in $p(x_T|x_1, \ldots, x_{T-1})$

10, where the last element represents an end-of-sequence value or token [22].

$$p(x_1, \ldots, x_T) = p(x_1)p(x_2|x_1)p(x_3|x_1, x_2) \ldots p(x_T|x_1, \ldots, x_{T-1}) \qquad 10$$

2.3.4.1 Gated Recurrent Units

A Gated Recurrent Unit (GRU) is a special type of generative recurrent neural network that is capable of adaptively capturing dependencies along different time scales. The GRU has gating units that allow for the flow of information to modulate. GRU models are capable of capturing long-term dependencies by alleviating problems with a vanishing or exploding gradient [22].

The GRU works using a linear interpolation at time *t* for the activation h_j^t between the previous activation and the candidate activation as shown in 11. An update gate z_t^j determines the level of update of activation or content. This is found as shown in 12. The candidate activation is shown in 13, where r_t is a set of reset gates which is in turn computed using 14. A diagram of the GRU can be seen in Figure 3 [22].

$$h_t^j = (1 - z_t^j)h_{t-1}^j + z_t^j \tilde{h}_t^j \qquad 11$$

$$z_t^j = \sigma\left(W_z \mathbf{x}_t + U_z \mathbf{h}_{t-1}\right)^j \qquad 12$$

$$\tilde{h}_t^j = \tanh\left(W \mathbf{x}_t + U\left(\mathbf{r}_t \odot \mathbf{h}_{t-1}\right)\right)^j \qquad 13$$

$$r_t^j = \sigma\left(W_r \mathbf{x}_t + U_r \mathbf{h}_{t-1}\right)^j \qquad 14$$

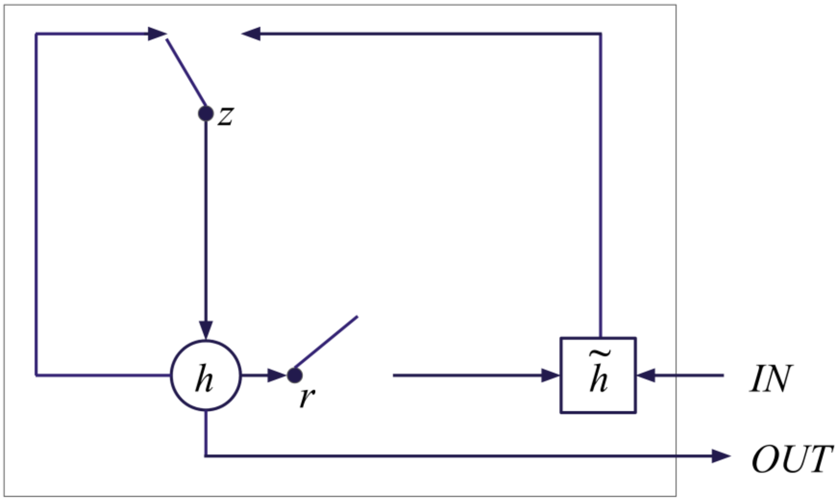


Figure 3: Gated Recurrent Unit

2.3.5 Teacher Forcing

Teacher forcing is a useful technique that allows recurrent neural network models to use output from a prior time step as direct input. Through this technique, slow convergence and instability in model training can be resolved. This is done through replacement during training of the actual output of a unit by something called a teacher signal [23]. (This is useful for language training because one word or character can be input for the next portion of the sequence, allowing the model to learn the sequence with the recursive output-as-input sequence.) Teacher forcing is incorporated into backward propagation through time through a backpropagation computation from later times being blocked at any unit in the network with an output set to a target value. Any unit with an external target value at a specific time step also should not be given error for that time step. This technique can also provide results where approximation is involved as well as training of continually operating networks. As information along the network moves, there is gradient information which can indicate the direction in which the network weights must be changed. Upon reaching steady-state behavior, this information disappears. This usefulness of teacher forcing is because the network weights as well as initial conditions can determine the behavior of the network, however by using desired values that can partially reset the network state at the current time it helps control the initial conditions for better subsequent training [23].

2.3.6 Attention

Although recurrent neural networks using a GRU model with teacher forcing can ensure better performance, there is still a problem with the need to encode an entire sentence into a single fixed-length vector representation. This single fixed length vector is all that the decoder has access to when generating a translation. While encoding an entire sentence into a single vector is still possible, better results can be created when also implementing an attention mechanism in the translator. Attention is devised so that the decoder is able to access every hidden state generated by the encoder during all time steps [24]. This attention mechanism is implemented as a multi-layer perceptron. A single-layer feed-forward network can be used to compete an expected alignment between each hidden vector that represents a source word or character, as well as the target word at the current time step. This allows a normalized alignment matrix between each source hidden vector to be created with trained model parameters. 15 and 16 show the calculation of this, where ***h*** represents a hidden vector, ***d*** represents the decoder state, *e* represents the expected alignment, ***v*** is the fixed-length vector of the encoder embedding, and ***U*** and ***W*** are trained model parameters. The final calculation output *a* is the normalized alignment matrix [24].

$$e_{t,i} = (\boldsymbol{v}_a)^T \tanh(\boldsymbol{U}_a \boldsymbol{d}_{t-1} + \boldsymbol{W}_a \boldsymbol{h}_i) \qquad 15$$

$$a_{t,i} = \frac{\exp(e_{t,i})}{\sum_{j=1}^{N} \exp(e_{t,j})} \qquad 16$$

A final calculation of a time-dependent source context vector is also computed. It is a weighted sum over the source hidden vectors, and each vector is weighted by the normalized alignment matrix attention weight. The context vector $\boldsymbol{c}_t$ is computed for each time step *t* of the decoder. It replaces the fixed-length vector $\boldsymbol{v}$ that was initially used in the encoder-decoder framework [24]. The equation for the context vector can be seen in

17.

$$\boldsymbol{c}_t = \sum_{i=1}^{T_x} a_{t,i} \boldsymbol{h}_i \qquad 17$$

2.3.7 Multimodal Machine Learning

In a survey of multimodal machine learning by Baltrusaitis et al. several more important concepts and terminology were explained [25]. Representation of data is a fundamental problem in machine learning, and even more complex to manage when representing multimodal data while properly exploiting its complementarity and redundancy. Mapping from modality to modality is another challenge of multimodal machine learning, and is referred to as translation. Alignment is the challenge of identifying relations between elements of different modalities. Fusion is the joining of information from more than one modality to make a prediction; this is an important challenge for multimodal natural language processing using text and images. Finally, this paper described the challenge of co-learning, which manages the learning from one modality in order to train on another modality. Meaningful representation of data is an important part of machine learning, and is especially difficult when dealing with multiple modalities. Recent work on most multimodal representations has often been a simple concatenation of unimodal data, however, new changes have involved joint and coordinated representations. Joint representations combine unimodal signals into a single representation space, and are best suited for situations in which all modalities are present during inference. Coordinated representations enforce similarity constraints into what is called a coordinated space while processing unimodal signals separately, they exist in their own space but are coordinated through a structure constraint or similarity. Coordinated representations are most suited to applications in which only a single modality is present at test time [25].

Sequential representation through recurrent neural networks and their variants can be used to represent data of varying sequence length. The task of an RNN encoder-decoder is representation of such a sequence, and is not limited to unimodal data. Currently a popular problem solved using multimodal machine learning is visual scene description such as image captioning. For this problem, when translating natural (human) languages, mapping from one modality to another is important. Within this problem there are two types of models, namely example-based and generative. A dictionary is used when translating (mapping) between modalities of example-based models, and they are restricted by their training data. Generative models generate sequences of symbols and are more complicated than example-based models. A challenge with these multimodal methods is the difficulty of their evaluation. This is especially true of human language translation, as there are many translations deemed correct and the evaluation of the best translation is often subjective. This is why there are many methods for translation evaluation including costly human evaluation, or various less accurate but more affordable machine evaluators [25].

Fusion of multimodal data integrates information from each modality in an attempt to predict an outcome. Fusion can provide robust predictions, complimentary information, and is operable even when a modality is missing. One of the modal-based approaches to fusion is the use of neural networks. Modern neural architectures described in this paper allow for end-to-end training of both the fusion and multimodal representation components of a model. Despite the impressive performance of neural networks on multiple modalities, a drawback to them is the difficulty of interpreting which features predictions rely on. Along with fusion, co-learning can be used to allow a modality to influence the training of another modality, which allows for complementary information sharing. This task of co-learning is independent and can improve fusion, mapping, and alignment models [25].

2.3.8 Machine Translation

Machine translation has been significantly improved through the incorporation of deep machine learning. The leap from statistical machine translation to neural machine translation was a great breakthrough in the field of machine translation [5][26][27]. Neural machine translation is a sequence to sequence task completed with an encoder and decoder model using recurrent neural networks [17]. Training is done using a parallel corpus of source language text and correct translations of a target language text. This method of machine translation assigns a fixed length encoding to a variable length sequence of input text in a source language using what is called the encoder. From there, the fixed length encoding vector can be decoded into the target language as a variable length text sequence. The translation model contains recurrent neural networks which are effectively used to learn a distribution over the variable length source and target language texts [17]. A general diagram for this model can be seen in Figure 4: Basic NMT encoder-decoder model.

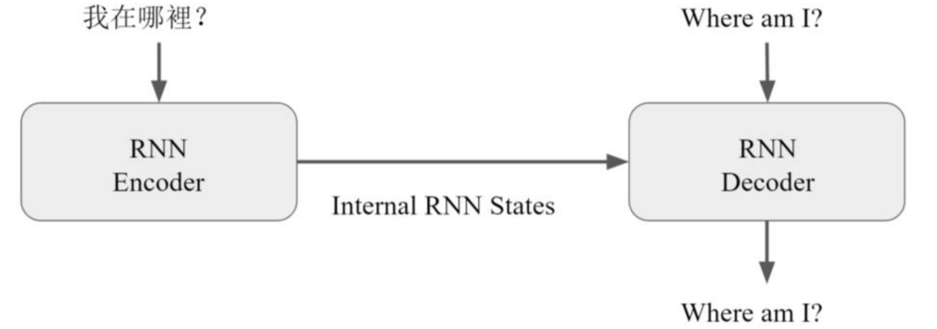


Figure 4: Basic NMT encoder-decoder model

More recent methods of this sometimes include an attention network as a feedforward neural network that is jointly trained with the rest of the translation system. This attention mechanism is used to compute a soft alignment, allowing for backpropagation of the cost function [5]. Local attention networks have been shown to provide better results when translating longer text sequences compared to systems without attention or that only contain global attention [26]. Attention is used in general for translation as it typically provides significant score increases from translators without attention [5][26][28].

2.4 Current Research

2.4.1 Chinese Text Granularity in Neural Machine Translation

Many current NMT systems are limited to a moderate vocabulary size, and are rarely capable of translating rare words due to lack of proper training in Chinese. To improve the modelling of words, the varying granularity methods of separating Chinese words has been assessed for performance. NMT allows for the freedom of choice with regards to token units and sentence segmentation [9]. Four granularities commonly discussed are character level separation, hybrid separation of word-characters, and two forms of subword level separation.

The first granularity discussed was Character Level separation, in which a sentence is split into a sequence of characters. For Chinese, this means that each Chinese character is separated (even if some characters in Chinese represent a whole word.) The character model for English is somewhat challenging as a sentence in English contains 300-1000 characters typically, making the state space very large. Taking that into account, this work only separated Chinese sentences by character split.

The second granularity discussed was a hybrid separation of word-characters. In this split method each character in a word has its location designated as beginning, middle, or end.

The third granularity method is a method of subword level separation called byte pair encoding. This is a compression method that iteratively replaces most frequently used pairs of bytes in a sequence with an unused byte. This means that characters and character sequences merge.

The last granularity in the paper was another subword level separation with a different approach. This granularity, called wordpiece model is a deterministic data-driven segmentation method for any sequence of characters. Through this model, a special symbol is prepended to the beginning of words and the rest of the word is split into subword components by character [9].

2.4.2 Multilingual Image Captioning

A significant amount of multimodal machine translation research falls into the category of multilingual automatic image captioning. These image captioning systems take an image and its source language caption, and then generate a caption in the target language, while using both the source language text and image as input. While this is a different problem from translation using image context, there is still valuable insight to be gained from said research for multimodal translation. Several approaches in recent literature shall be described in this subsection.

2.4.2.1 Multimodal Image Caption Translation Compared to Statistical Machine Translation

The primary translation task completed in a paper written by Caglayan et al. was to multimodally translate English image descriptions into German [2]. The baseline system that was used for this task was built using a statistical translation system called Moses. The Moses pipeline was trained using minimum error training rate (MERT). The task described in this paper was also completed through use of continuous space language model with auxiliary features support, which allowed the use of sentence-level features. The auxiliary features used for training were image features extracted from a layer of the image network using the FC7 layer of the VGG-19 network [29] and sentence text representation vectors. In order to test against the statistical phrase-based machine translation system, a neural MT system was also built. The model built was an attention-based encoder-decoder with GRU for a recurrent decoder shown in Figure 5. One attention mechanism was implemented for a fully-connected feed-forward neural network, that determined the decoder's initial hidden state by receiving the mean annotation vector. The dataset on which both of the systems were trained is the multilingual Flickr30k provided by WMT. The visual data of the images was trained through convolutional neural networks based off previous research using ResNet-50 features [2].

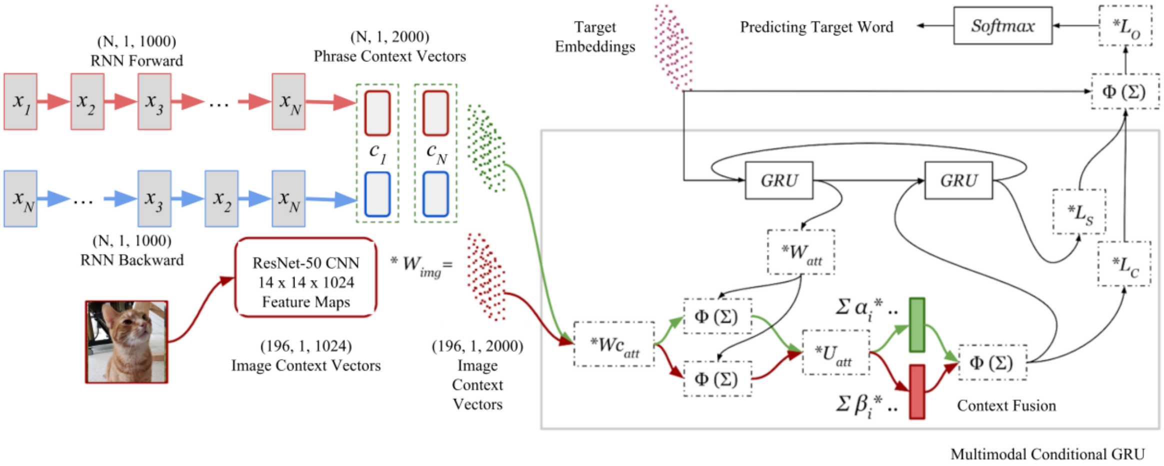


Figure 5: Multimodal Image Caption Translation Model by Caglayan *et al.* [2]

The multimodal translation model was the combination of the translation and image evaluation systems using two GRU layers and an attention mechanism. A shared attention layer consisting of a fully-connected feed-forward network was used for computation of a set of attention coefficients along each timestep. The second GRU generated hidden states from intermediate representations of them along with the context vector. The results of the multimodal system vs the monomodal system were reported with the monomodal system performing better than the multimodal system for both BLEU and another language scoring metric reported scores. This was explained by noting that either their image and text representations were not integrated well in their system, or the images contained too much irrelevant information [2].

2.4.2.2 Multimodal Neural Machine Translation of Image Captions using a Doubly-Attentive Decoder

A notable paper by Calixto, Liu, and Campbell described their novel approach to multimodal neural machine translation on image descriptions [30]. Detailed in this paper is a machine translation model built to incorporate spatial visual features of images to aid in translation of image captions. (The captions to the images are descriptions of the image content.) This paper’s contributions to the field of multi-modal neural machine translation include use of attention-based models to incorporate the image information, as well as the proof that images provide useful information to a neural machine translation model for the application of translating the datasets provided by the Conference of Machine Translation.

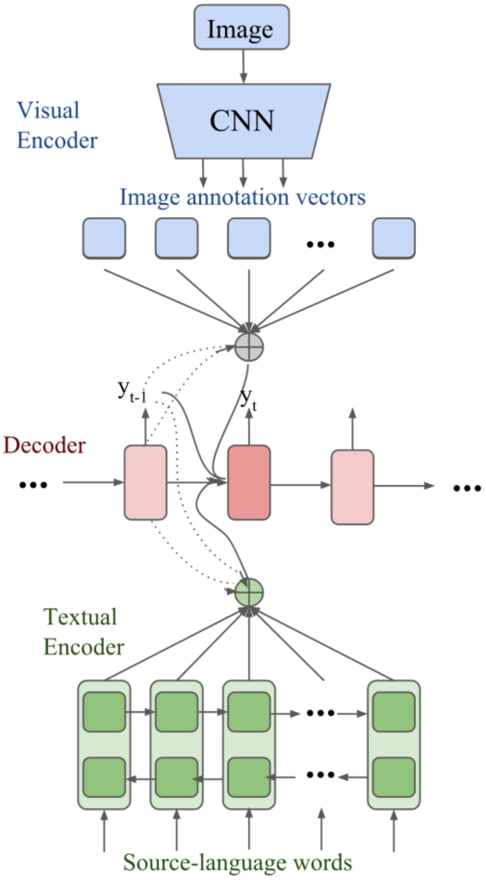


Figure 6: Calixto's doubly-attentive decoder attends to image and language features independently [30]

The translation model described in this paper is a doubly attentive model based off prior researchers' work. The paper describes the previous work as having an encoder and decoder as two recurrent neural networks with one attention mechanism implemented by a multilayer perceptron. The encoder is descried as a bidirectional recurrent neural network with gated recurrent unit, while the decoder has a conditional gated recurrent unit. The overall diagram of the model created by Calixto *et al.* can be seen in Figure 6. Their use of the previous translator work was extended in this research through the incorporation of multimodal attention based neural machine translation using spatial visual features as well as text. The spatial visual features were extracted using the 50-layer Residual network [31]. To generate translations, the doubly-attentive decoder independently attends to the image features and the source-language words. To start training the translator, a text only phrase-based system was built, and then used along with a pretrained image processing system. A notable limitation of the model is that it was only trained to translate sentences up to a certain length as anything in the dataset exceeding the length limit was removed. The dataset used for this research is a multi-lingual version of the Flickr30k dataset [31][32]. (The Flickr30k dataset contains thirty thousand images with five human generated captions per image written in English. The variation of this dataset used for the multimodal project was the M30k, which contains each image with one of the five English translations having been human translated to German [32].)

The performance of the model described in this paper exceeded other comparable models that were trained on the same dataset. Notably, it purportedly outperforms the other models with an improvement of +1.4 BLEU. The doubly-attentive model is noted as having improvements in recall and precision-oriented metrics due to the incorporation of images in the model. Performance of the model when pre-trained on different datasets is also noted in the paper [30].

In a later publication about multimodal translation by Calixto *et al.* a similar multimodal translator from English to German using neural machine translation with convolutional neural networks was created [33]. The extracted image features used in translation by Calixto et al. were incorporated on several levels, including in the source sentence, in the encoder, and also in the decoder. Their work evaluated differences between the three uses of image features, and confirmed previous research conclusions, stating that using image features directly in the decoder caused model overfitting [33]. This overfitting was later described in another publication by Calixto *et al.* as being avoidable by incorporating visual features to ground translations rather than use image features directly along each time step in the decoder thus improving translation quality [24].

2.4.2.3 Multimodal Image Caption Translation using Image Region Bounding

A different approach to use of image context in image caption translation has been proposed by Huang et al. to translate English image captions into German. Their method used a neural machine translation framework including extensive research into bounding image regions into separate feature fields [34]. The structure of their method allowed association between text and image features. Their incorporation of image context was shown to outperform translations using text only [34].

Research has also been done to translate image captions from English to Czech multimodally using neural machine translation by Helcl [35]. The research covered more modes than just including image input.

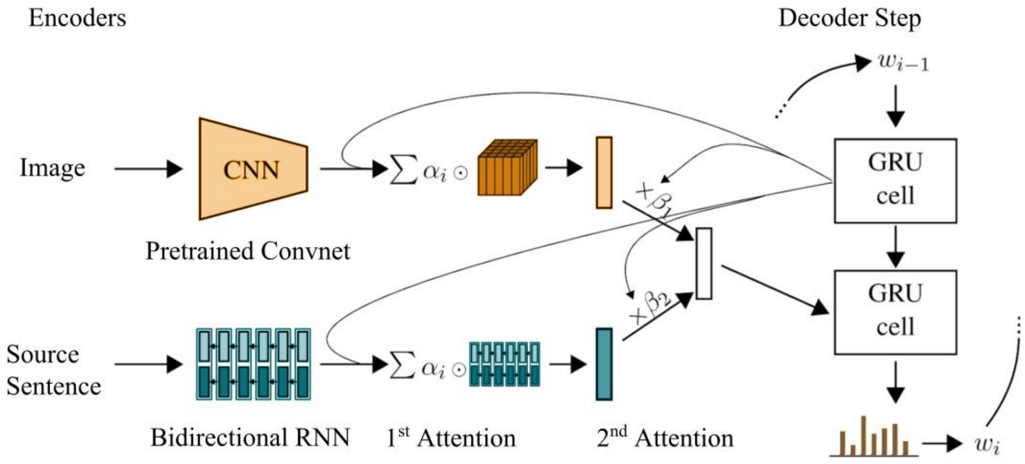


Figure 7: Overall model using hierarchical attention by Helcl [35]

To translate an image description from one language to the other, Helcl used a neural translation model including a convolutional neural network encoder for the image with two textual encoders for the image text and translation. (Note that while the main research was to improve Czech translation, the research for image caption translation used German as the target language) [35].

2.5 Research Gaps

As noted before, nearly all of the multimodal machine translation research applies only to image caption translation, there is an entire field of translation research yet to be explored. Compounded with the fact that it is all to and from German and English, research in any other language pair would provide new insight to the problem of multimodal translation.

In an earlier publication by Calixto using image context for statistical machine translation was surveyed, including a report on the usefulness of images [36]. That survey aimed to evaluate if visual information could alleviate text ambiguity of unknown words in translation, and looked to determine what ways image clues could be used in statistical machine translation systems, wherein Calixto et al. used similar images and their predetermined textual information. This short publication did not approach an identification of methods answering *how* to use image information as it only evaluated potential use of images in translation [36]. That research asked questions about using image context in translation; however, it did not provide insight or architecture for the incorporation of visual context.

An important area of research which lacks current solutions is incorporation of image context into machine translation. Research that would open up this unexplored avenue of machine translation could provide meaningful input to the topic area of machine translation. This work aims to do so, while using a state-of-the-art neural machine translation model with attention and incorporating image context in the form of image tags into that model.

III. Methodology

3.1 Chapter Overview

This section describes the both the implementation of an NMT with attention built from the architecture provided by Bahdanau et al [5], and an image context incorporation to the text translator. There were several datasets used containing an extensive parallel text corpus of human translated Chinese and English sentences as well as a parallel corpus of images and associated text in Chinese and English. Performance of the translator was evaluated using the machine translation scorer, Bilingual Evaluation Understudy. The architecture and implementation of the machine learning model for text translation and multimodal translation are presented throughout the following sections.

The goal of this research is to provide answers to questions regarding the differences between neural machine translation models that can accept text only against a translator with both text and image label input, comparing the two by using a nonbiased machine translation scoring mechanism. The score comparisons between each translation, as well as the multimodal architecture to be developed will provide insight to the uses and performance of machine translation using image context. To evaluate the completion of this objective, the following research questions will be evaluated.

1) Does addition of image tags/labels improve translator performance, and if so, by how much?

Determining the answer to this question was conducted by evaluating the level of improvement the multimodal translator had over the text-only translator. Any statistically significant number of improved translations provides evidence for a positive answer to this research question. This question is answerable through measuring whether translation evaluation scores for the multimodal translation model are improved from the baseline text-only translator.

2) Is there a repeated topic/structure/word composition with respect to sentences improved through multimodal translation?

This question is answered through a by-hand analysis of the translated sentences in the test set. Sentences have been considered for grammar, word count, use of uncommon words, and inclusion of homographs. Checking what homographs have been improved through translation provides evidence that an uncertain translation was corrected via image context. The number of improved sentences, as well as the percentage of improvement provides proof that translations were corrected by the inclusion of image context.

3) Do certain image label topics/part of speech/repetition improve translation performance?

This can be answered through a comparison of attention weights for tags. The visualization of attention plots provides insight to the features the translator viewed as most important. A by-hand assessment of length and use of image tags has been considered to find a pattern in what kind of image tags were most important. Patterns being considered include part of speech, nouns referring to objects or people, and homographs clarified by an image label.

This chapter describes the datasets used to train text only and multimodal translators, as well as the preprocessing steps necessary to use each dataset. Following that is a description of the translator model architecture and how it is applied for text only and a multimodal translation. Model fitting is described for each translation model, followed by evaluation and analysis methods for measuring the performance of each trained translator.

3.2 Dataset and Preprocessing

Two datasets were used for training each translator model, a text dataset, and a dataset containing both text and images. Before training the model, the text required reformatting. To avoid out of memory problems, any sentence exceeding a maximum character count of 100 characters was removed from the datasets for either Chinese or Roman alphabetic characters. Separation and tokenization of the sentences in each language was also conducted. The English granularity within the model was separated by word; this is a very simple but effective approach to granularity and works well for the majority of languages containing spaces between each word. Its high effectiveness and simplicity makes it common to work at the word level for many text sequence problems [9][15]. The Chinese granularity for this translation model was separated by character. This is because separation by character is more intuitive, and produces acceptable results. (To understand why separation between words is impractical for Chinese, refer to Section 2.2.1, where it is described that characters and radicals may represent words or only portions of words depending on the other surrounding characters.) A word index and reverse word index was created for the purpose of mapping from word to index and back. The parallel corpora of sentence pairs provided the necessary data to conduct sequence to sequence machine learning. While sentences do not contain specific features that vary in translational relevance, it is expected that more training information will produce better results.

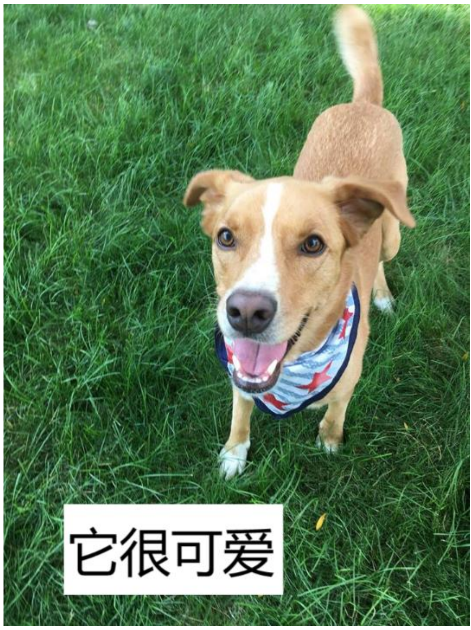
3.2.1 Text Dataset

The dataset containing only text consists of over 21,000 sentences written in English and in Mandarin Chinese. Within the thousands of sentences/observations, there is a variable count of words/features. The parallel sentence corpus was written and translated by humans. Note that human translation, despite variability, is the best way to conduct language translation. The list of sentences was freely available under a Creative Commons license by the Tatoeba Project, and also includes select sentences from various Chinese literature, as well as a dictionary of Chinese characters. The character dictionary was included to allow for training on rare words.

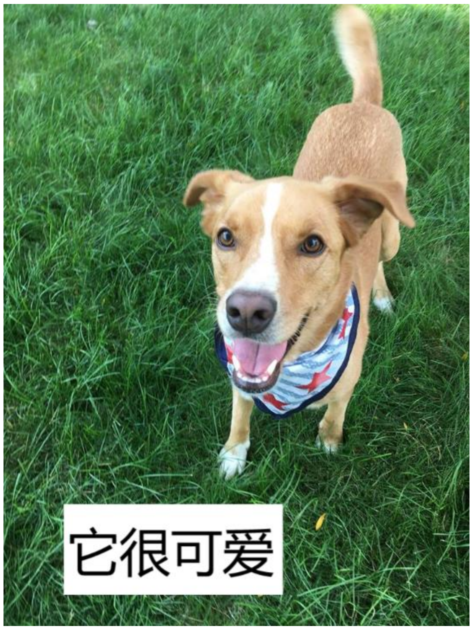
The following sections provide detailed explanation of the datasets used to train the models, as well as the translation model architecture. Model architecture is described for both text translator and the image context input, starting with the training and execution of the text translator. Finally, evaluation of the translators is discussed, along with methods used to answer the research questions.

3.2.2 Image Dataset

The dataset for use with the image translator is composed of 700 images with an associated sentence drawn from Chinese media. Each Chinese sentence has been translated into English by a human experienced with Chinese-to-English translation. The images were pulled from several sources that included several types of textual formats.



One of the use cases is ‘natural text in the wild,’ which includes text on storefronts, billboards, and other real-world writing. Another portion of the dataset is made from screen caps of subtitled videos. An example of one of these images, and its associated information can be seen in Figure 8. The videos used are from Weibo Videos [37], a Chinese video streaming site similar to YouTube that is very popular in China, as well as YouTube [38] itself, BBC News [39], and other similar sources [40][41][42]. (Because the database does not involve ordinary reading or viewing of the processed works, the use is non-consumptive and is therefore considered fair use.) This video source was chosen because it contains a wide variety of videos all commonly watched and understood by speakers of Mandarin Chinese, which makes it a good general representative of Chinese language used in videos. Due to the plethora of dialects in China, videos almost always contain subtitles. (The subtitles are technically in Mandarin Chinese, but due to the logographic nature of the language, each word is written with the same character in every Chinese dialect and can be understood by any literate Chinese reader.) For the dataset, the subtitles and other naturally occurring Chinese text in other image data were translated to English. The English translations were completed by native speakers of Mandarin Chinese and English. The preprocessing required for the image dataset includes the same granularity separation of Chinese text as described in Chapter 2, as well as the retrieval of image tags. The image tags will be used to provide visual context rather than developing an image interpretation model.



Chinese Text	它很可爱
English Text	It's so cute
Image Labels/Tags	dog breed, dog, grass, snout, companion dog, beagle, dog like mammal, plummer terrier, hound

Figure 8: Multimodal Dataset Example

3.3 Translation Model Architecture

The overall approach to multimodal translation is to combine previous research of image classification with text translation. The input to the multimodal translator contains text input like any standard translator, as well as image input in the form of image classification tags. The goal of this multimodal translator is to produce a translator that can incorporate extra contextual information into a translation, thus producing more accurate results even on sentences lacking verbal context clues. The conceptual level design of the overall multimodal translator can be seen in

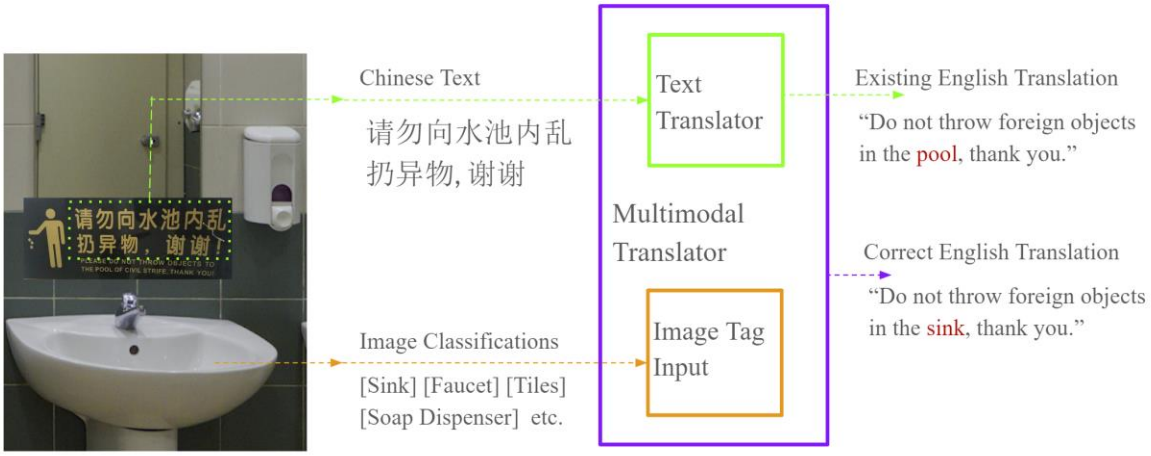


Figure 9.

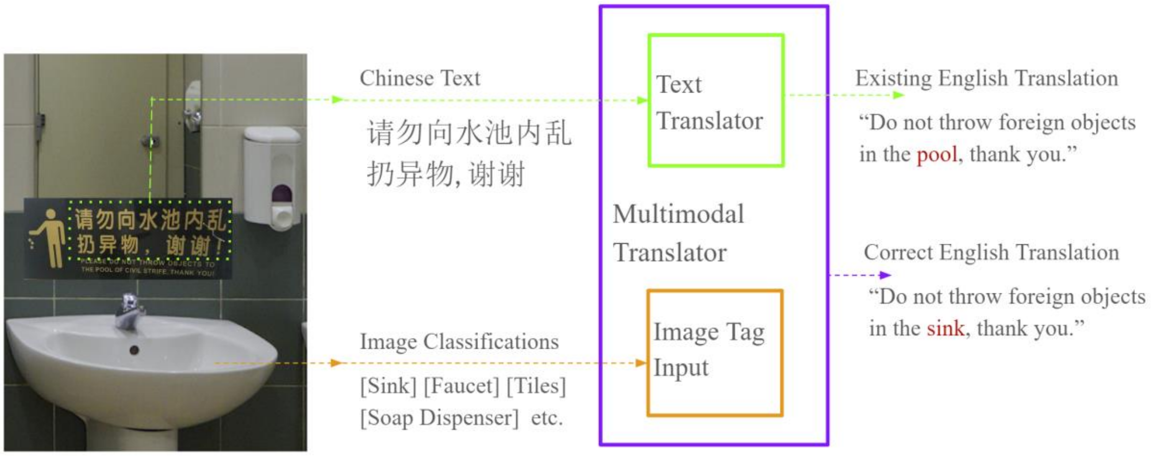


Figure 9: Top Level Conceptual Diagram of Multimodal Translator

Language translation is a sequence to sequence problem. Before a multimodal translator can be completed, it is important to have an established text translation model. To do this, a neural machine translator with an attention mechanism has been modified to accomplish translation from Chinese to English. This architecture provides support for the multimodal translation model. The input sentence is put through an encoder model that in turn gives the encoder output and the encoder hidden state. Both encoder and decoder models are composed of a gated recurrent unit (GRU) recurrent neural network (RNN).

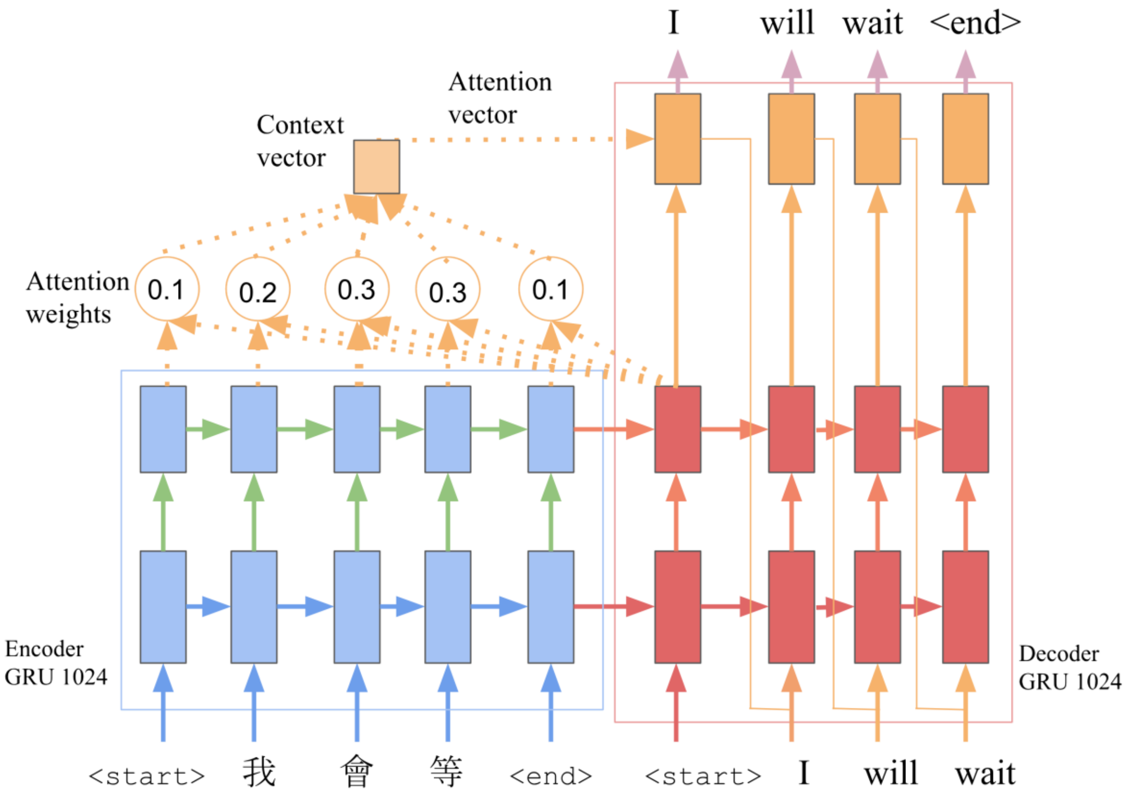
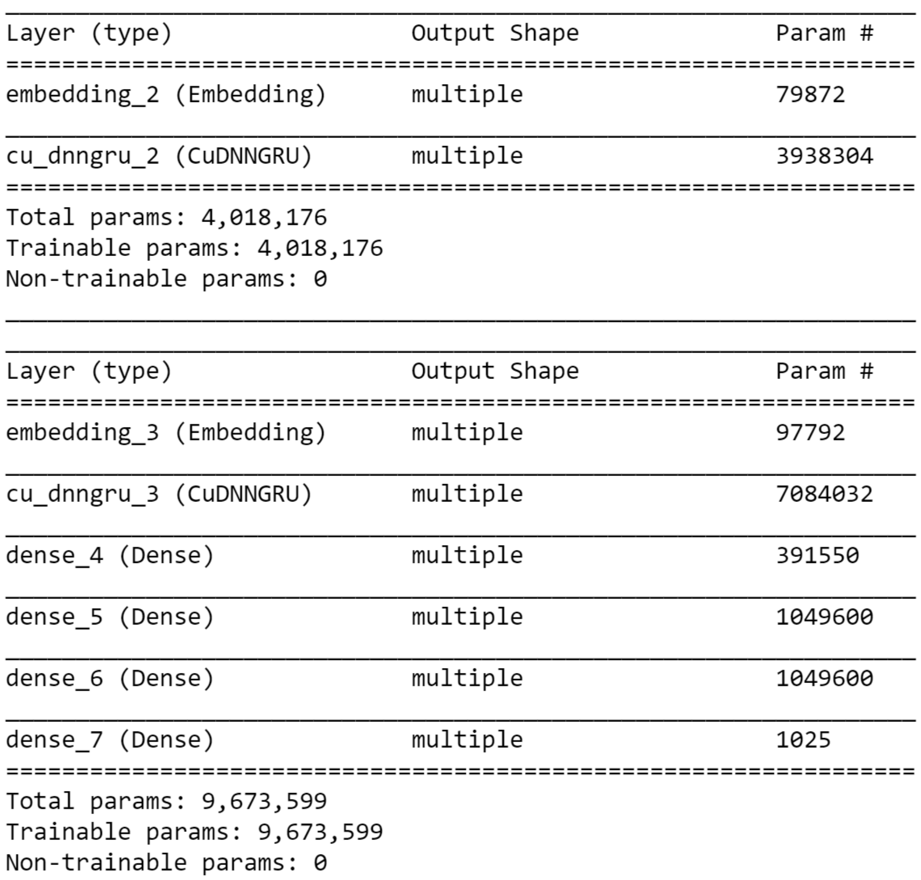


Figure 10: Neural Machine Translator with Attention Architecture

The encoder consists of an embedding layer with an output dimension of 256 followed by three GRU layers with 1024 units and a batch size of 64. The GRU layers output a sequence and a state, with embedded source language text Chinese sent in with start and end tokens. The decoder also has an embedding layer and three GRU layers containing 1024 units and a batch size of 64, as well as two fully connected layers, and six fully connected layers for attention. The sequence and state are passed to the fully connected layers. The attention weights are calculated through application of a softmax activation function to the output of the fully connected layers. Once the attention weights are calculated, the context vector is found by calculating the dot product of the attention weights with the encoder output. The training of the model is competed using a sparse softmax cross entropy with logits loss function, which computes the cross entropy between the log probabilities and labels providing the softmax cross entropy loss [43]. During training, prior time steps are used as input to the model as teacher forcing.

The model parameterizations and structure were based off of Bahdanau’s model architecture [5], while the code implemented contains hyperparameters that were pre-tuned by the TensorFlow authors [44]. The parameters of the model can be seen in Table 4. (Output shape is multiple because each layer returns multiple items, an output as well as a state.)



```
_________________________________________________________________
Layer (type)                 Output Shape              Param #   
=================================================================
embedding_2 (Embedding)      multiple                  79872     
_________________________________________________________________
cu_dnngru_2 (CuDNNGRU)       multiple                  3938304   
=================================================================
Total params: 4,018,176
Trainable params: 4,018,176
Non-trainable params: 0
_________________________________________________________________
_________________________________________________________________
Layer (type)                 Output Shape              Param #   
=================================================================
embedding_3 (Embedding)      multiple                  97792     
_________________________________________________________________
cu_dnngru_3 (CuDNNGRU)       multiple                  7084032   
_________________________________________________________________
dense_4 (Dense)              multiple                  391550    
_________________________________________________________________
dense_5 (Dense)              multiple                  1049600   
_________________________________________________________________
dense_6 (Dense)              multiple                  1049600   
_________________________________________________________________
dense_7 (Dense)              multiple                  1025      
=================================================================
Total params: 9,673,599
Trainable params: 9,673,599
Non-trainable params: 0
```

Table 4: Model Parameters

3.3.1 Text Translator Model

The text translator model uses the model description and parameter count described in the previous section. The model reads through source words beginning with a `<start>` token, until it reaches the `<end>` token. Then it emits one target word at a time. Let $X^k = (x_1^k, x_2^k, \dots, x_{N_k}^k)$ represent the corresponding word embeddings in a sentence S^k containing word indices $\{w_1^k, w_2^k, \dots, w_{N_k}^k\}$ for a language L_k. The encoder reads each word from left to right in X^k thus generating a sequence of annotation vectors.

Each training example is a sentence pair, a tuple of sentences S^k in L_k. Along each training sample, the embeddings $X^k = (x_1^k, x_2^k, \ldots, x_{N_k}^k)$ are retrieved for each sentence S^k. For the English sentences there is a separate word embedding $\{w_1^k, w_2^k, \ldots, w_{N_k}^k\}$ for each word, while the Chinese is represented through individual character embedding in each sentence S^k. A top-level diagram of the model architecture can be seen in Figure 10.

3.3.2 Image Classification Acquisition

After the development of the bilingual image dataset, each image was assigned machine-generated image tags. The image label tags were made from Google Cloud Vision API [7]. (The image tags were checked to be reasonable but were not modified as the purpose of this research is to evaluate the usefulness of machine generated image tags in multimodal machine translation.) The image tags contain classifications of objects and actions perceived in the image.

3.3.3 Image Classification in Multimodal Translator Model

The model reads through source words one word at a time, as well as image classification words. The input to the encoder is then `<start> Chinese Text <image> Image Labels <end>` while the English text input through the decoder remains unchanged during training. Then it emits one target word at a time. Let $X^k = (x_1^k, x_2^k, \ldots, x_{N_k}^k)$ represent the corresponding word embeddings in a sentence S^k containing word indices $\{w_1^k, w_2^k, \ldots, w_{N_k}^k\}$ for a language L_k. Let $I^k = (i_1^k, i_2^k, \ldots, i_{N_k}^k)$ represent the image classification labels. A source language sentence S^k is now composed of X^k as well as I^k. The encoder reads each word from left to right in X^k followed by I^k

thus generating a sequence of annotation vectors containing both Chinese source language characters and English image labels. Each training example is a sentence pair, a tuple of sentences S^k in L_k with Chinese source language sentences with image labels, and English target language sentences. Along each training sample, the embeddings $X^k = (x_1^k, x_2^k, \dots, x_{N_k}^k)$ through $I^k = (i_1^k, i_2^k, \dots, i_{N_k}^k)$ are retrieved for each sentence S^k. For the English sentences there is a separate word embedding $\{w_1^k, w_2^k, \dots, w_{N_k}^k\}$ for each word, while the Chinese is represented through individual character embedding in each sentence S^k. A top-level diagram of the model architecture can be seen in Figure 11 with the additional input of image labels.

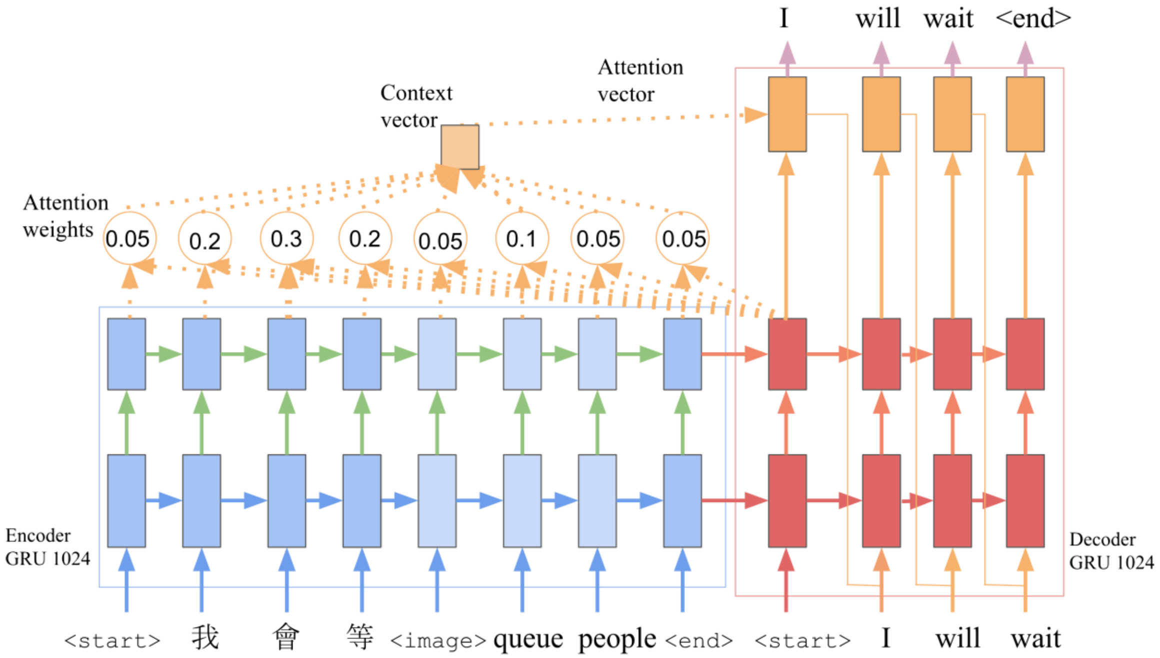


Figure 11: Incorporation of Image Classification Input to Translator

3.4 Model Fitting Details

Training on a parallel sentence corpus is supervised learning with sequential data. The model was trained for three separate evaluation types. Once, for text only with a training set containing 21,000 English and Chinese sentences, and a test set containing 200 sentences. The test set is made up of 200 sentences in English and Chinese, that were randomly removed from the original text corpora.

The translator was also trained using the image dataset, once with text and image label tags, and once with only text. The size of the image dataset is 700 images and associated sentences. The Training set contains 650 samples, and the test set contains 50 samples. In order to analyze the images, Google Image Tagger API was used to derive image classification labels. The image tags were then used to represent context for situation setting, items in an image, or other visually identifiable contexts as plain text.

Finally, the model was also trained with a large number of sentences, and also the available multimodal observations of images with text. The large text dataset is a 20,000 sentence parallel corpus containing Chinese and English. The image dataset added to this is the same size for both training and test, so the total number of training observations is 20,000 sentences plus 650 sentences and image labels. This was trained both with and without the image labels for comparison between the monomodal and multimodal translator models.

Because many translators are trained on enormous training sets, the test sets are typically under 3% the size of the training set [45][5][30]. (Note that the reported number of the training set is the sentence count after the test set was removed.) The written text within each parallel corpora is formatted in UTF-8, which is then converted to ASCII before being trained on or translated. Input to the translator is a sentence from the dataset in the source language, and output is a sentence in the target language making the input and output many to many. The model has been trained with a varying number of epochs for model tuning, using validation based early stopping, with a delta of 0.0001 and a patience of ten epochs.

3.5 Model Evaluation and Analysis

For performance evaluation, the NMT models with and without image context have been compared to each other as well as to Google Translate through use of BLEU translation score evaluator. This is because Google Translate provides a baseline of competitiveness, while the text only model provides a comparison to the effectiveness of the translation model using image context. The translation scores were evaluated through BLEU on varying *n*-gram levels.

BLEU, or Bilingual Evaluation Understudy is used to generate performance scores for machine translations. BLEU works by comparing one or more human translated truth translation to each equivalent machine translation. This comparison is made by counting position independent word token, or *n*-gram matches [6]. Scores generated by BLEU represent a percentage of similarity, thus an identical sentence would receive a one, and a sentence without any matching words would receive a zero. BLEU is a simple algorithm that calculates scores base off of word matches and order, and does not consider synonyms. To gauge how human translator scores compare to machine translation scores, note that humans typically score between 0.20-0.35 depending on the number of references. Note that word order is also taken into account, and has a smaller score deduction than missing or incorrect words. For a full explanation of BLEU scoring, refer to Section 2.2.2.2.

While BLEU can be used for evaluating machine translated sentences against more than one human translated reference, [6] the Chinese-English sentence and image corpora only contain one sentence translation each, and so only one sentence will be used as a reference for all purposes of scoring. Performance comparisons for each model can be seen in Chapter IV. For some of the translations, the translation between the text only and multimodal models may produce equal output.

Along with BLEU scoring, performance evaluations will be conducted for translations in the test set of models trained using the multimodal dataset by hand. These human evaluated sentences will be considered for homographs, as well as context of the image itself regarding the image tags. This evaluation by hand should provide insight to the similarity between translation using context of a human vs machine, and insight to whether this multimodal neural machine translator successfully mimics human translation. The human analysis of the translations includes consideration for the image tag labels, as well as the contents of the image. The Chinese sentences that have been translated to English without image context will be considered for the possibility of human interpretation given an accompanying image. Special consideration will be taken for translations of identical sentences that have been improved with image tag context from its counterpart without context.

IV. Analysis and Results

4.1 Chapter Overview

This chapter contains the results and the analysis produced by testing the performance of the translation models varying on the training data used. Key results include BLEU scores and homographical analysis for the model trained on each dataset.

The first section of this chapter evaluates the performance of the resulting translator from training the model using image classification tags along with text on the small 700 sentence multimodal dataset. Two comparative translators have been made, including a multimodal translator using all of the available information, and a text only translator. This text only translator should provide clarity as a baseline model for comparison.

Then evaluation was completed with the addition of a large text only parallel corpus from Chinese to English. This larger corpus contains 21,000 sentences used to train a text only translator. Another text only translator was trained using the 21,000 sentences as well as the 700 sentences contained within the multimodal dataset. Finally, a multimodal translator was trained using the 21,000 sentences in addition to the text and image context of the smaller dataset containing 700 multimodal observations of sentences and image tags.

4.2 Text Only Versus Context-and-Text Models Small Dataset

Image tag context incorporation was conducted using a dataset containing 700 Images whose context was derived using an automatic label generator. The NMT translation model with image incorporation was trained on the sentences associated with the images in Chinese, their English translations, and the image labels. The NMT translation model with text only was trained on the same Chinese and English sentences, however, the image labels were excluded from training. The test sets for both models contain 50 samples. The text of both datasets is identical Chinese and English sentences, however, the test set also includes image labels for the translation model trained with image context labels.

4.2.1 Analysis

Table 5 contains the average BLEU-4 scores for both of the translators trained on text only as well as text plus image labels using the neural machine translation model described in Section 3.3, Translation Model Architecture. The translator that was trained with image context incorporation achieved a BLEU score of 0.006076 at best performance for an epoch count of 20. That puts it slightly higher than the best performing text only model at 60 epochs.

Table 5: Average BLEU-4 Score for Translation Model With and Without Image Context

Epoch Count	With Image Tags	Without Image Tags
10	0.001764	0.002715
20	0.006076	0.003789
30	0.004219	0.003745
40	0.003844	0.003228
50	0.003606	0.004127
60	0.003765	0.005576
70	0.004278	0.003695
80	0.004647	0.003483

90	0.004282	0.004730
100	0.003623	0.004637

A visualization of the information found in the table above can be seen in Figure 12: BLEU Scores by Epoch Variation, showing the score of the translator along each Epoch from 10-100. Despite the appearance of potential score increase near the end, training was stopped at 100 epochs because of overfitting.

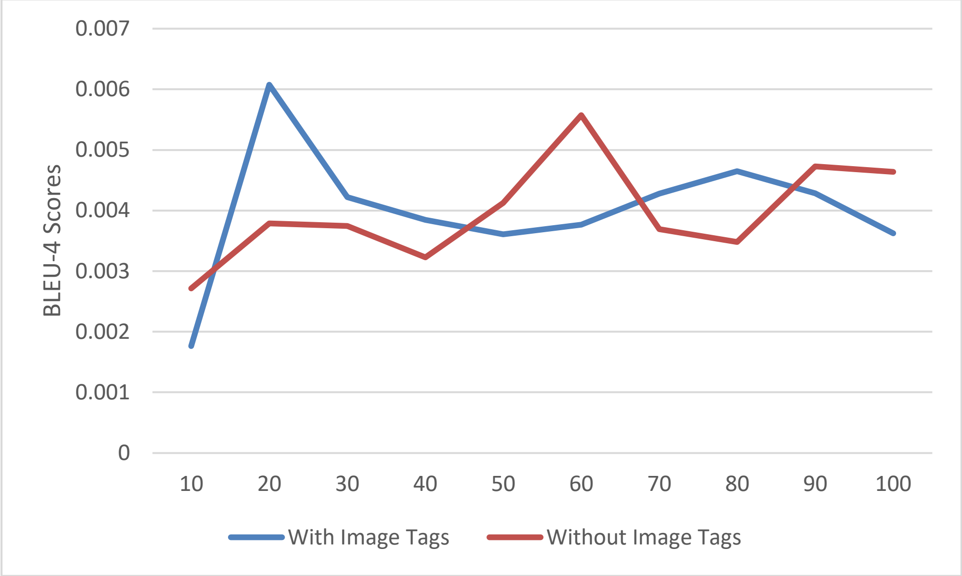


Figure 12: BLEU Scores by Epoch Variation

As a comparison to the overall translational performance of both models, Google Translate scores have also been included for performance on the test set sentences in Table 6.

Table 6: Average BLEU Scores of Best Performance vs. Google Translate

Translator	**BLEU-4 Score**
No Image Tags 20 Epochs	0.006076
With Image Tags 60 Epochs	0.005576
Google Translate	0.170538

Along with the BLEU scores accomplished by the translator, some examples can be found in Table 7: Translation Examples at 20 Epochs. The full test set sentences and individual scores at 20 epochs can be found in Appendix A. As can be seen in the following table, despite receiving positive scores, nearly every translation makes very little sense.

Table 7: Translation Examples at 20 Epochs

		Using Image Labels		**Without Image Labels**	
Reference Sentence	**Image Labels**	**Translated Sentence**	**BLEU-4**	**Translated Sentence**	**BLEU-4**
living independently is also a kind of training	girl	The handicraft of the forest	0.00524	This is a day	0.0062
look upwards	community conversation event fun	I have red	0.00550	I have a year	0.0045
mainly teaches badminton class	ball centre competition event game	the shape of snow has arrived	0.00229	This is the middle of the middle of the middle of the middle of the	0.0007
middle cut	fish cook food cuisine animal fat	Cut the morning	0.02554	I have a year	0.0045
put in the napa cabbage	bakeware Chinese cooking cookware	Cut the romaine lettuce into strips	0.00725	This is the middle of the middle of the middle of the middle of the	0.0023
it already has a history of over 1500 years	area city hill land lot	The cultural reputation is till strong tropical grasslands	0.00137	This is a day	0.0040

4.2.2 Results

The scores of both NMT models were significantly lower than what Google Translate has to offer. Because the multimodal training dataset was so small, none of the translations were correct along any epoch count for either translator trained using this available data. Although the BLEU scores reported imply that some of the translations may have made sense, the translations were entirely insensible. A portion of BLEU scores contained positive numbers simply due to repetition of simple common words like “the” or “is”. This is the reason that the average BLEU scores were higher for the model without image context along some epochs. Looking at the sentences produced by each resulting translator, for the model trained on the dataset with image tags, the translations were more diverse and human readable than for the translator trained without image tags.

In the homograph analysis, a sensible homographical discrepancy was found. The model trained with image context produced an interesting result, matching cabbage to lettuce in the translation test of the translator with image context, but not in the translation model not using image context. This match was formed by the training set containing a sentence containing the word lettuce, which has part of the same Chinese characters in it as the word for cabbage. This overlap alone was not enough for the translation model without image labels to draw a connection; only the translation model with image labels, and a matching hind of the word “cookware” brought the two words together. This translation result is very interesting, containing the Chinese character 菜 that is part of the words 生菜: lettuce, and 白菜: cabbage, along with the matching image context word. As the translation identically matches the sentence from the training set, it brings the possibility of overfitting into concern. Another concern brought by this is that rather than clarifying and correcting the meaning of a homograph, the image context instead caused a similar but incorrect translation due to the 菜 character.

Table 8: Translation Results Regarding Chinese Homographs

Training Set English	Chinese	Image Labels		
Cut the romaine lettuce into strips	生菜切片	food cookware cuisine dish	**Machine Translation**	
Test set English			**With Image Labels**	**No Image Labels**
put in the napa cabbage	将白菜放入	bakeware Chinese cooking cookware	Cut the romaine lettuce into strips	This is the middle of the middle of the middle of the middle

A notable consistent difference between the two translators was the lack of repetition produced by the translator using image context when compared to the continuously repeated sentences output by the model not using image context. It is notable that the translator with image context scored only a few percentage points different than the text only translator. The overall low translation scores are expected for translator models trained on such a small dataset, as neural machine translation is a task that requires large training sets in order to perform well, so it is unsurprising that Google Translate produced significantly better results.

4.3 Text Only Compared to Context and Text with Addition of Large Dataset

The model described in this section was trained for a text only analysis and as well as text and incorporation of image context classifications. This translator model was trained with three varying dataset differences.

First, it was trained using 21,000 Chinese and English sentences for the text only analysis of the model. Then, it was trained for the same text only corpus with the addition of the text in the small 700 sentence dataset. Finally, the model was trained using all available data of the 21,000 sentences + 700 sentences with both text and image label context. BLEU-4 score comparisons and a human analysis of the translation of the test set, as well as the implications of the results can be seen throughout the following sections.

The test sets for these translators is composed of 200 sentences held out from the text only dataset, and 50 sentences from the small multimodal dataset. The monomodal translators used only text, while the multimodal translator trained using the text and image context was validated using 200 text only sentences, as well as 50 sentences that also contained image context.

4.3.1 Analysis

The BLEU scores for the translators trained on the large text corpora of 21,000 sentences can be seen in Table 9. The highest average BLEU scores were achieved by the text only translators with the highest scoring translator varying between the translator trained on 21,000 text only and the +700 without image context depending on the particular number of epoch upon each translator was trained. The consistently lowest scoring translator was the one trained using image context incorporation.

Table 9: BLEU-4 Scores of Translation Models

Epoch Count	+700 With Image Tags	+700 Without Image Tags	21,000 Text Only
10	0.008875118	0.031355872	0.029259704
20	0.013642770	0.031276284	0.035750026
30	0.015233238	0.031742756	0.027756914
40	0.014634604	0.031080695	0.035674637
50	0.016404786	0.033939655	0.023021317
60	0.017768224	0.029486692	0.036272446
70	0.018497228	0.036152980	0.038712413

80	0.020695962	0.041531918	0.037470628
90	0.014483491	0.027564053	0.036803192
100	0.018309572	0.031896373	0.036318036

In Figure 13, the BLEU scores can be seen for each translator that the model produced when trained on the variations of the data. The green line represents the translator trained on the entire text only dataset of 21,000 parallel corpus sentences. The red line represents the translator trained on the entire text only dataset with the addition of 700 sentences from the small multimodal dataset. This translator is still a text only translator. The blue line represents the resulting multimodal translator trained on the text only dataset as well as the 700 sentences and image context labels. While the scores appear to improve near the end, overfitting occurred causing training to cease at 100 epochs.

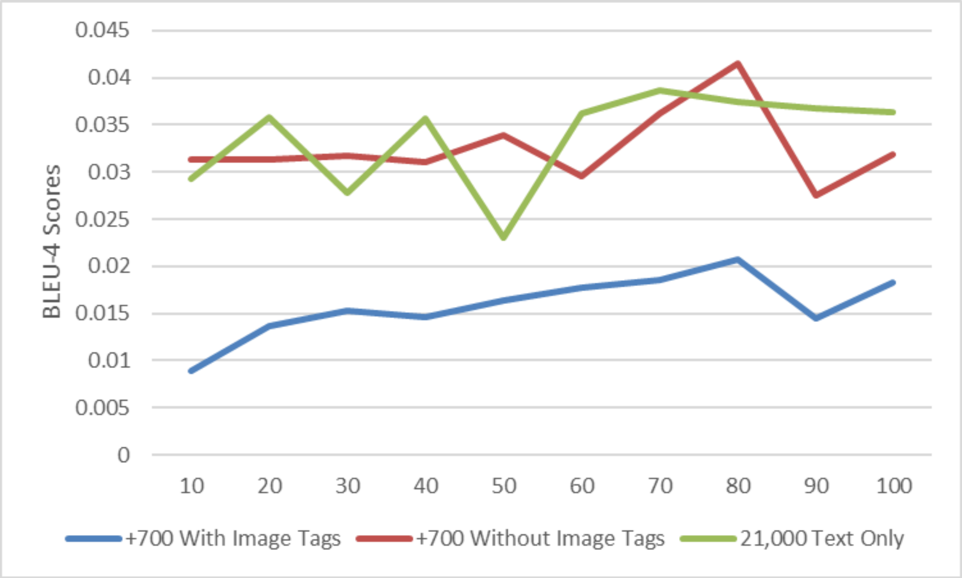


Figure 13: BLEU-4 Scores by Epoch

The highest average score for the multimodal translator was achieved at 80 epochs of training. While it produced the highest BLEU scores at 80 epochs, the resulting score was still very low with nonsensical translations. Some examples of the translations produced by this translator can be seen in Table 10.

Table 10: Multimodal Translation Examples at 80 Epochs

Reference Sentence	**Multimodal Machine Translation**	**BLEU-4**
look upwards	why	0.003679
mainly teaches badminton class	American-Israeli relations grow closer by the type of affairs	0.001349
middle cut	Even with the students happy life	0.002296
put in the napa cabbage	Lie down and agility	0.003519
slow slow slow	this color you the doctor called him to	0.001562
stretch until thin	like to welcome the romaine lettuce into strips	0.001562
the tower at the highest elevation	The poor young man of the world is free of a bad influence on the world is free of a serious illness intervenes.	0.001751
this day is a holy day	This is a day of days ago .	0.026269
this will be a long and dangerous journey	This is the right to be a long stories .	0.071718
what's the reason?	What has it?	0.005503

4.3.2 Results

The three translators produced with the additional training of a 21,000 sentence parallel corpus produced fairly poor results with BLEU-4 scores not exceeding 0.045 at any epoch count. The sentences produced by these translators made very little sense for the most part with examples of the translations produced by the multimodal translator provided in the previous section. The machine translations were evaluated for any homographical clarification and elimination of repeated phrases, however, nothing was found in those regards. Every translator trained on the large text set was found to have few repeated translations on every epoch count.

The lowest scoring translator was the one the multimodal translator, which could have been caused by several reasons. The low multimodal score could simply be because the majority of the data that the multimodal translator was trained on was not actually multimodal. Only 700 of the 21,700 sentences in the dataset contained image tags as context. This lack of consistency of training examples could have produced confounding issues in the resulting translator. Another potential cause of low scores could be because the input of the image context was in the form of text. The image tags associated with each sentence could have caused problems with the translator attempting to translate English image tags into English from a Chinese sentence. This problem is similar in results to the lack of a large entirely multimodal training corpus; however, the form of image context input could have been a specific cause of the low scores.

4.4 Discussion

Overall, the translators produced using the model described in Chapter 3 were all quite low performing and worse than Google Translate. The translators trained only on the 700 sentence corpus produced interesting results with regards to Chinese homographs for the multimodal translator. The translator with image context produced a result about cooking with cabbage, for a sentence that actually mentioned cooking with lettuce, while the translator without any image context merely produced nonsense. The translators trained on larger datasets, while producing slightly better scores, did not produce any results differing because of homographs. Another notable result of the translators is that the ones trained on the small dataset had a difference between repetition of results. The multimodal translator trained on the 700 sentence corpus produced a wide variety of sentences, while the monomodal translator produced repetitive identical sentences until very late epoch counts. These noteworthy differences between the multimodal and monomodal translators were only observed in the translators trained exclusively on parallel corpora of the designated translator mode. Due to the small size of the multimodal training set, it is difficult to draw any conclusions from these results.

V. Conclusions and Recommendations

5.1 Conclusions of Research

This research provides a first glimpse of the use of image labels as a context for multimodal machine translation. It can be seen that the overall results of the translator model produced results on par with current state of the art translators. Answering the research questions, an overall glimpse of the novelty provided by the multimodal neural machine translator can be seen.

1) Are image classification tags/labels a useful feature for improving translation scores?

Through the results found in Chapter IV the answer to this is inconclusive. While there was not a significant difference in BLEU scoring for the translations with and without image context, there was significant difference in the kinds of sentences produced with regards to the model trained on only 700 sentences. The diversity of the translation output implies that there was some kind of difference produced by the addition of image tags; however, no improvement can be determined from these results. On the small dataset, the translator trained only with text produced output that was entirely nonsense, while the translator trained with image context produced a result about cooking with cabbage, for a sentence that instead mentioned cooking with lettuce. While there was only one significant homograph influenced translation, this still provides merit to the practice of including image tags for translation. The translators trained on larger datasets, while producing slightly better scores, did not produce any results differing because of homographs.

2) What kind of sentences are improved through multimodal translation?

As none of the sentences had significant improvement, an answer to this question cannot be provided by this research. The test set for each dataset did have different results, but despite these differences, significant improvement was not made with regards to grammar, sentence structure, or vocabulary used. A sentence resulting from the translation of a homograph and that contained an identical image label produced output identical to a sentence in the training set. The one example of homographical influence in the small dataset analysis could provide evidence that sentences containing similar words to those used in the training set could be corrected or identically copied through inclusion of image context.

3) What kind of image labels improve translation performance?

Similar to the answer of question number two, the lack of improvement leaves this question difficult to provide a meaningful answer from the results produced. The one notable homograph translation was associated with an image tag that was contained in the training set. Because of this image tag match, it is suspected that there is an association between sentences containing image tags in the training set that match image tags in a test set.

5.2 Significance of Research

Research in language translation is important for personal, business, and government oriented endeavors. Better translation allows for better international exchanges without the cost of a professional translator. This particular research is significant in the field of machine translation because it is an entirely novel approach to translation. Never before have researchers used image labels (or raw image features) as an input for a translator outside of the context of automatic image captioning. This use of image context in translating naturally occurring sentences associated with real world scenes provides a first study into a new field of multimodal neural machine translation.

5.3 Recommendations for Future Research

While the research described here provided some insight into the problem of translation with image labels as context, there are many other approaches that could be taken to further alleviate the problem of poorly performing machine translators. From a standpoint of incorporating image features as context, the following section describes machine learning model variations, and data augmentation possibilities recommended for future researchers.

5.3.1 Model Changes

5.3.1.1 Separate Input Technique for Raw Image Features

There are two main different approaches to using image context within a multimodal neural machine translator.

The first approach is to use image features from pretrained network like Google Image Tagger, while incorporating the image labels into a separate input space. This would produce a two input sequence to sequence model, with separate input of sentence text and image label text. The model would have two inputs and one output instead of one extended input and one output.

Another approach would be to train a convolutional neural network on multilingually captioned images. This would be a very difficult approach with the current dataset as image training for such a complex topic as language association would require far more training data than the 700 image sentence pairs that were used for this research. Research along this line could potentially be more difficult because it requires expertise with both image processing and with language translation. One would also be required to consider what kinds of problems arise from removing the simplicity of the current model proposed in this work.

5.3.1.2 Different Model Approach than Sequence to Sequence

Instead of approaching multimodal translation with the accepted translation model of sequence to sequence, it would also be possible to implement a newly published approach to translation using a transformer network [46]. This model architecture relies solely on attention and provides promising results. With a new type of translation architecture, it is possible that it could produce better results than a sequence to sequence model with separate input types. The use of the transformer model could be used with the data and approach in this research, or with either proposed approach mentioned above.

5.3.2 Data Changes

5.3.2.1 More Data

Machine learning tasks often produce better results with larger datasets. As languages contain an infinite possibility of word arrangements and length, a dataset containing fewer than several thousand multimodal samples cannot train a translator model suitable for real world use. A multimodal dataset of similar size to an accepted text only dataset would approach a more reasonable amount of data in producing a translator model suitable for commercial use.

5.3.2.2 Data Augmentation

Building a multimodal dataset by gathering images and sentences is a tedious task done by hand. As the output of an image labeler provides objects within an image the majority of the time, these results could be artificially created for sentences not associated with a real image. For example, a sentence that says something about apples could have an image tag “apple” artificially created without having a real image associated with the sentence. This type of augmentation could be completed with special consideration for clarifying the use of Chinese homographs in short sentences that may lack significant verbal context.

5.3.2.3 Different Languages

The majority of ideas presented in this work are not exclusive to Mandarin Chinese. While potentially less reliant on context, most languages contain homographs and other context dependent phrases. With a similarly constructed dataset, and tokenization alteration suitable for each language, the same experimentation could be run, producing a multimodal translator for any other language.

5.4 Summary

This work provides a novel approach to neural machine translation through incorporation of image context in the form of text image tags. Previous chapters included thorough explanation of machine learning and neural machine translation background. Methodology description included details about the translator model with explanations about each of the datasets used for training. Results provided BLEU scores and a by-hand analysis of translation results with discussion regarding mistakes made and differences between the monomodal text only verses the multimodal model. This research provided a few answers to the research questions posed, providing some amount of insight to the use of image context in translation of sentences associated with images containing natural scenes. The analysis of the translation results does not provide significant evidence for the usefulness of image context in translation, however, as the results were produced using a very small there is still possibility that further research in this area could provide more clear results. Further research is needed to completely assess the worth of image context in machine translation.

Bibliography

[1] S. Chand, “Empirical survey of machine translation tools,” *Res. Comput. Intell. Commun. Networks*, vol. Second Int, pp. 181–185, 2016.

[2] O. Caglayan *et al.*, “Does Multimodality Help Human and Machine for Translation and Image Captioning?,” *arXiv Prepr. arXiv1605.09186*, 2016.

[3] I. Calixto, D. Stein, E. Matusov, P. Lohar, S. Castilho, and A. Way, “Using Images to Improve Machine-Translating E-Commerce Product Listings,” *Short Pap.*, vol. 2, no. 2016, pp. 637–643, 2017.

[4] L. Specia, S. Frank, K. Sima’an, and D. Elliott, “A Shared Task on Multimodal Machine Translation and Crosslingual Image Description,” *Proc. First Conf. Mach. Transl.*, vol. 2, pp. 543–553, 2016.

[5] D. Bahdanau, K. Cho, and Y. Bengio, “Neural Machine Translation by Jointly Learning to Align and Translate,” CoRR, vol. abs/1409.0, pp. 1–15, 2015.

[6] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, “BLEU: a Method for Automatic Evaluation of Machine Translation,” *Proc. 40th Annu. Meet. Assoc. Comput. Linguist. - ACL ’02*, pp. 311–318, 2002.

[7] Google, “Google Cloud Vision API Services.” [Online]. Available: https://cloud.google.com/vision/. [Accessed: 02-Nov-2018].

[8] “Tatoeba Project: English-Chinese (Mandarin) Sentence Pairs,” 2018. [Online]. Available: https://tatoeba.org/eng/?lang=eng. [Accessed: 03-Jul-2018].

[9] Y. Wang, L. Zhou, J. Zhang, and C. Zong, “Word, subword or character? an empirical study of granularity in Chinese-English NMT,” *Commun. Comput. Inf. Sci.*, vol. 787, pp. 30–42, 2017.

[10] Y. Zhang, S. Vogel, and A. Waibel, “Interpreting BLEU/NIST Scores: How Much Improvement Do We Need to Have a Better System?,” *Lang. Resour. Eval.*, 2004.

[11] J. M. Cavnar, William B. and Trenkle, “N-gram based text categorization,” *Proc. Symp. Doc. Anal. Inf. Retr.*, pp. 161–175, 1994.

[12] C.-Y. Lin and E. Hovy, “Automatic evaluation of summaries using N-gram co-occurrence statistics,” *Proc. 2003 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. - NAACL ’03*, vol. 1, pp. 71–78, 2003.

[13] A. Sokolov, “Evaluation of SMT systems : BLEU,” 2015.

[14] I. Sutskever, O. Vinyals, and Q. V. Le, “Sequence to Sequence Learning with Neural Networks,” *Adv. Neural Inf. Process. Syst.*, pp. 3104–3112, 2014.

[15] F. Chollet, *Deep learning with Python*. Manning Publications Co., 2017.

[16] J. Weston, F. Ratle, H. Mobahi, and R. Collobert, “Deep learning via semi-supervised embedding,” *Neural Networks: Tricks of the Trade*, no. Springer, Berlin, Heidelberg, pp. 639–655, 2012.

[17] K. Cho, B. van Merrienboer, D. Bahdanau, and Y. Bengio, “On the Properties of Neural Machine Translation: Encoder-Decoder Approaches,” *Syntax. Semant. Struct. Stat. Transl.*, p. 103, 2014.

[18] M. Luong, “Multi-task sequence to sequence learning,” *arXiv Prepr. arXiv1511.06114*, 2015.

[19] I. Sutskever, O. Vinyals, and Q. V. Le, “Sequence to Sequence Learning with Neural Networks,” *Adv. Neural Inf. Process. Syst.*, pp. 3104–3112, 2014.

[20] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.

[21] A. Karpathy, “The Unreasonable Effectiveness of Recurrent Neural Networks,”

Karpathy.Github.Io, pp. 1–28, 2015.

[22] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, “Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling,” *CoRR*, vol. abs/1412.3, 2014.

[23] R. J. Williams and D. Zipser, “Gradient-based learning algorithms for recurrent connectionist networks and their computational complexity,” *Backpropagation Theory, Archit. Appl.*, vol. 1, pp. 433–486, 1995.

[24] I. Calixto, C. Nick, and Q. Liu, “Incorporating Visual Information into Neural Machine Translation,” no. August, 2017.

[25] T. Baltrusaitis, C. Ahuja, and L. P. Morency, “Multimodal Machine Learning: A Survey and Taxonomy,” *IEEE Trans. Pattern Anal. Mach. Intell.*, pp. 1–20, 2018.

[26] M.-T. Luong, H. Pham, and C. D. Manning, “Effective Approaches to Attention-based Neural Machine Translation,” *Proc. 2015 Conf. Emperical Methods Nat. Lang. Process.*, pp. 1412–1421, 2015.

[27] Y. Wu *et al.*, “Google’s Neural Machine Translation System: Bridging the Gap between Human and Machine Translation,” *arXiv Prepr. arXiv1609.08144*, 2016.

[28] O. Firat, K. Cho, and Y. Bengio, “Multi-Way, Multilingual Neural Machine Translation with a Shared Attention Mechanism,” 2016.

[29] K. Simonyan and A. Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recognition,” *arXiv Prepr. arXiv1409.1556*, 2014.

[30] I. Calixto, Q. Liu, and N. Campbell, “Doubly-Attentive Decoder for Multi-modal Neural Machine Translation,” *arXiv Prepr. arXiv1702.01287*, 2017.

[31] I. Calixto, Q. Liu, and N. Campbell, “Doubly-Attentive Decoder for Multi-modal Neural Machine Translation,” *Proc. 55th Annu. Meet. Assoc. Comput. Linguist.*

(Volume 1 Long Pap., vol. 1, pp. 1913–1924, 2017.

[32] B. Plummer *et al.*, “Flickr30k Entities: Collecting Region-to-Phrase Correspondences for Richer Image-to-Sentence Models,” *Proc. IEEE Int. Conf. Comput. Vis.*, pp. 2641–2649, 2015.

[33] I. Calixto, Q. Liu, and N. Campbell, “Incorporating Global Visual Features into Attention-Based Neural Machine Translation,” 2017.

[34] P.-Y. Huang, F. Liu, S. Shiang, J. Oh, and C. Dyer, “Attention-based Multimodal Neural Machine Translation,” *Wmt-2016*, vol. 2, pp. 639–645, 2016.

[35] J. Helcl, “Improving Neural Machine Translation with External Information Thesis Proposal.”

[36] I. Calixto and L. Specia, “Images as Context in Statistical Machine Translation ∗,” no. November, pp. 2–3, 2011.

[37] “YouTube,” 2018. [Online]. Available: Youtube.com. [Accessed: 21-Dec-2018].

[38] “主页 - BBC News 中文,” *BBC News*, 2018. [Online]. Available: https://www.bbc.com/zhongwen/simp. [Accessed: 21-Dec-2018].

[39] “Weibo - 微博北美站,” 2018. [Online]. Available: Weibo.com. [Accessed: 21-Dec-2018].

[40] “Youku 优酷-这世界很酷.” [Online]. Available: Youku.com. [Accessed: 21-Dec-2018].

[41] “Yew York Times - 纽约时报中文网 国际纵览,” 2018. [Online]. Available: cn.nytimes.com. [Accessed: 21-Dec-2018].

[42] "Baidu News 百度新闻——全球最大的中文新闻平台," 2018. [Online]. Available: News.baidu.com. [Accessed: 23-Dec-2018].

[43] M. Abadi *et al.*, "TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems," 2015. [Online]. Available: Software available from tensorflow.org.

[44] B. Lamberta, Y. Katariya, R. Chada, and R. Yang, "Neural Machine Translation with Attention," *Licensed under the Apache License, Version 2.0,* 2018. [Online]. Available: https://github.com/tensorflow/tensorflow/blob/master/tensorflow/contrib/eager/python/examples/nmt_with_attention/nmt_with_attention.ipynb. [Accessed: 06-Jul-2018].

[45] M.-T. Luong, H. Pham, and C. D. Manning, "Effective Approaches to Attention-based Neural Machine Translation," *arXiv Prepr. arXiv1508.04025,* 2015.

[46] A. Vaswani *et al.*, "Attention Is All You Need," *Adv. Neural Inf. Process. Syst.,* pp. 5998–6008, 2017.

Appendix

Comparison_20_Epochs	**With Labels/ With Labels 20 epochs**		**No Labels/ No Labels 20 epochs**	
Reference Sentences	**Candidate Sentences**	**BLEU-4 Scores**	**Candidate Sentences**	**BLEU-4 Scores**
20 minutes later the weather turned sunny	The icefields on the Crested Ibis	0.00614528404	This is a day	0.002134156817
A little more	A radio	0.04288819425	This is the middle of the middle of the middle of the middle of the	0.0007432998185
According to tradition	I am thirsty	0.005503212081	This is the middle of the middle of the middle of the middle of the	0.0007432998185
Can they succeed?	The icefields on the scenery for free while working	0.001348511186	This is the middle of the middle of the middle of the middle of the	0.0007432998185
China's development	The spring plowing this year old	0.002295748847	This is the middle of the middle of the middle of the middle of the	0.0007432998185
Chinese cleavers are a multi-use tool	The spring plowing this inspection tea .	0.002295748847	This is the middle of the middle of the middle of the middle of the	0.0007432998185
Each pig of ours requires 5kg of water per day	The largest automobile manufacturer of error of the glass shards	0.004463236138	This is a day	0.003187905703
Even I don't know how much I've run	the air	0.0003520477366	This is a day	0.001662083001

Comparison_20_Epochs	**With Labels/ With Labels 20 epochs**		**No Labels/ No Labels 20 epochs**	
Reference Sentences	**Candidate Sentences**	**BLEU-4 Scores**	**Candidate Sentences**	**BLEU-4 Scores**
Every detail is important	The largest automobile manufacturer of infrared cameras	0.0018575058	This is the middle of the middle of the middle of the middle of the	0.002350520411
Good morning teacher	the morning	0.04288819425	This is the middle of the middle of the middle of the middle of the	0.0007432998185
Good, eat it	I am a dog	0.004518010018	I have a year	0.004518010018
Good, enough, you can stand up now	The spring plowing this year old	0.001943309444	This is the middle of the middle of the middle of the middle of the	0.0007432998185
How old are you?	the morning	0.002601300475	This is the middle of the middle of the middle of the middle of the	0.0007432998185
I can't do that kind	the morning	0.001577768493	This is the middle of the middle of the middle of the middle of the	0.0007432998185
I really enjoy life and work in China	the best	0.0003520477366	This is a day	0.001662083001
I'll be home to eat in a bit	I have a new kindergarten student	0.005201870634	This is the middle of the middle of the middle of the middle of the	0.0007432998185
I've been living in Beijing for 3 years already	the morning	0.0002135277459	This is a day	0.001294431542

Comparison_20_Epochs	**With Labels/ With Labels 20 epochs**		**No Labels/ No Labels 20 epochs**	
Reference Sentences	**Candidate Sentences**	**BLEU-4 Scores**	**Candidate Sentences**	**BLEU-4 Scores**
It needs a home	The spring plowing this year old	0.002295748847	This is the middle of the middle of the middle of the middle of the	0.0007432998185
It's so cute	the morning	0.004288819425	This is the middle of the middle of the middle of the middle of the	0.0007432998185
It's straight that way	the morning	0.002601300475	This is the middle of the middle of the middle of the middle of the	0.0007432998185
Look how beautiful the color is	This woman is Tibetan	0.008665626145	This is the middle of the middle of the middle of the middle of the	0.002795255596
Safely land in the middle of the lake below	The UAV contionusly rocks to the best	0.00524932594	This is a day	0.001294431542
So many people	the morning	0.004288819425	This is the middle of the middle of the middle of the middle of the	0.0007432998185
This one is a smaller size	The entire village is clear , [this man] began mountain-climbing	0.003753119269	This is a day	0.0360645288
This year I'm 23. I come from India	The researchers installed a year old	0.005201870634	This is a day	0.005255967942

Comparison_20_Epochs	**With Labels/ With Labels 20 epochs**		**No Labels/ No Labels 20 epochs**	
Reference Sentences	**Candidate Sentences**	**BLEU-4 Scores**	**Candidate Sentences**	**BLEU-4 Scores**
Using/use local traditional activities	The spring plowing this year old	0.002295748847	This is the middle of the middle of the middle of the middle of the	0.0007432998185
Wait, Don't eat it	the morning	0.002601300475	This is the middle of the middle of the middle of the middle of the	0.0007432998185
Watch my body movements	The spring plowing this year old	0.002295748847	This is the middle of the middle of the middle of the middle of the	0.0007432998185
Weather is fickle on plateaus	the yellow	0.001577768493	This is the middle of the middle of the middle of the middle of the	0.002350520411
Your human smell is different from its smell	the best	0.0003520477366	This is a day	0.005255967942
a world filled with green	A radio	0.01577768493	This is the middle of the middle of the middle of the middle of the	0.0007432998185
a young bird	This woman is eating temple vegetarian food	0.0018575058	I have a year	0.01428720215
dip in some chili peppers	This woman is the morning	0.003021375397	This is the middle of the middle of the middle of the middle of the	0.0007432998185

Comparison_20_Epochs	**With Labels/ With Labels 20 epochs**		**No Labels/ No Labels 20 epochs**	
Reference Sentences	**Candidate Sentences**	**BLEU-4 Scores**	**Candidate Sentences**	**BLEU-4 Scores**
go	Cut the romaine lettuce into the romaine lettuce into the romaine lettuce into the romaine	0.0007432998185	I have a year	0.004518010018
grab	A radio	0.007071067812	I have a year	0.004518010018
handmade iron woks	the pot to the morning	0.003021375397	This is the middle of the middle of the middle of the middle of the	0.0007432998185
hello classmate	The most beautiful bride and through	0.002295748847	I have a year	0.004518010018
in the midst of a frozen land	the morning	0.005804285916	This is the middle of the middle of the middle of the middle of the	0.002795255596
it already has a history of over 1500 years	The cultural reputation is till strong tropical grasslands	0.001378433659	This is a day	0.004093351949
living independently is also a kind of training	The handicraft of the forest	0.00524358122	This is a day	0.006250434473
look upwards	I have red	0.005503212081	I have a year	0.004518010018
mainly teaches badminton class	the shape of snow has arrived	0.002295748847	This is the middle of the middle of the middle of the middle of the	0.0007432998185
middle cut	Cut the morning	0.02554364775	I have a year	0.004518010018

Comparison_20_Epochs	**With Labels/ With Labels 20 epochs**		**No Labels/ No Labels 20 epochs**	
Reference Sentences	**Candidate Sentences**	**BLEU-4 Scores**	**Candidate Sentences**	**BLEU-4 Scores**
put in the napa cabbage	Cut the romaine lettuce into strips	0.007259795291	This is the middle of the middle of the middle of the middle of the	0.002350520411
slow slow slow	A house	0.004288819425	This is the middle of the middle of the middle of the middle of the	0.0007432998185
stretch until thin	A radio	0.004288819425	This is the middle of the middle of the middle of the middle of the	0.0007432998185
the tower at the highest elevation	The icefields on the forrest canyon	0.008633400214	This is the middle of the middle of the middle of the middle of the	0.002795255596
this day is a holy day	the wild	0.0009569649651	This is a day	0.03875385825
this will be a long and dangerous journey	The cultural reputation is till very strong , ok ?	0.001348511186	This is a day	0.006250434473
what's the reason?	Cut the morning	0.02554364775	This is the middle of the middle of the middle of the middle of the	0.002350520411
	Average Score:	0.00607579376	**Average Score:**	0.003788772839

REPORT DOCUMENTATION PAGE	*Form Approved* *OMB No. 074-0188*

1. REPORT DATE *(DD-MM-YYYY)* 21-03-2019	**2. REPORT TYPE** Master’s Thesis	**3. DATES COVERED** *(From – To)* September 2017 – March 2019

TITLE AND SUBTITLE

Machine Translation with Image Context from Mandarin Chinese to English

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Johnson, Brooke E., Second Lieutenant, USAF

5d. PROJECT NUMBER
JONs 18G906 (AFOSR) and 18G271D, 19G271D (FRC)

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAMES(S) AND ADDRESS(S)

Air Force Institute of Technology
Graduate School of Engineering and Management (AFIT/EN)
2950 Hobson Way, Building 640
WPAFB OH 45433-8865

8. PERFORMING ORGANIZATION REPORT NUMBER

AFIT-ENG-MS-19-M-035

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Air Force Office of Scientific Research
875 N. Randolph, Ste.325
Arlington Virginia, 22203
Email: info@us.af.mil Phone: 703-696-7797
Grant# F4FGA08087J001

10. SPONSOR/MONITOR’S ACRONYM(S)
AFOSR

11. SPONSOR/MONITOR’S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

DISTRUBTION STATEMENT A. APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

13. SUPPLEMENTARY NOTES

This material is declared a work of the U.S. Government and is not subject to copyright protection in the United States.

14. ABSTRACT

Despite ongoing improvements in machine translation, machine translators still lack the capability of incorporating context from which source text may have been derived. Machine translators use text from a source language to translate it into a target language without observing any visual context. This work aims to produce a neural machine translation model that is capable of accepting both text and image context as a multimodal translator from Mandarin Chinese to English. The model was trained on a small multimodal dataset of 700 images and sentences, and compared to a translator trained only on the text associated with those images. The model was also trained on a larger text only corpus of 21,000 sentences with and without the addition of the small multimodal dataset. Notable differences were produced between the text only and the multimodal translators when trained on the small 700 sentence and image dataset, however no observable discrepancies were found between the translators trained on the larger text corpus. Further research with a larger multimodal dataset could provide more results clarifying the utility of multimodal machine translation.

15. SUBJECT TERMS

Neural machine translation, multimodal translation, natural language processing, machine learning

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON** Dr. Brett J. Borghetti, AFIT/ENG
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			**19b. TELEPHONE NUMBER** *(Include area code)* (937) 255-6565, ext 4612 brett.borghetti@afit.edu
U	**U**	**U**	**UU**	91	

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std. Z39-18