AFRL-RI-RS-TR-2023-135

LATERAL: LEARNING AUTOMATIC, TRANSFER- ENHANCED, AND RELATION-AWARE LABELS

IBM THOMAS J. WATSON RESEARCH CENTER

JULY 2023

FINAL TECHNICAL REPORT

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED

STINFO COPY

AIR FORCE RESEARCH LABORATORY
INFORMATION DIRECTORATE

■ **AIR FORCE MATERIEL COMMAND** ■ **UNITED STATES AIR FORCE** ■ **ROME, NY 13441**

NOTICE AND SIGNATURE PAGE

Using Government drawings, specifications, or other data included in this document for any purpose other than Government procurement does not in any way obligate the U.S. Government. The fact that the Government formulated or supplied the drawings, specifications, or other data does not license the holder or any other person or corporation; or convey any rights or permission to manufacture, use, or sell any patented invention that may relate to them.

This report is the result of contracted fundamental research deemed exempt from public affairs security and policy review in accordance with SAF/AQR memorandum dated 10 Dec 08 and AFRL/CA policy clarification memorandum dated 16 Jan 09. This report is available to the general public, including foreign nations. Copies may be obtained from the Defense Technical Information Center (DTIC) (http://www.dtic.mil).

AFRL-RI-RS-TR-2023-135 HAS BEEN REVIEWED AND IS APPROVED FOR PUBLICATION IN ACCORDANCE WITH ASSIGNED DISTRIBUTION STATEMENT.

FOR THE CHIEF ENGINEER:

/ S /
PETER A. JEDRYSIK
Work Unit Manager

/ S /
JULIE BRICHACEK
Chief, Information Systems Division
Information Directorate

This report is published in the interest of scientific and technical information exchange, and its publication does not constitute the Government's approval or disapproval of its ideas or findings.

REPORT DOCUMENTATION PAGE

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED	
		START DATE	END DATE
JULY 2023	FINAL TECHNICAL REPORT	AUGUST 2019	MAY 2023

4. TITLE AND SUBTITLE

LATERAL: LEARNING AUTOMATIC, TRANSFER- ENHANCED, AND RELATION-AWARE LABELS

5a. CONTRACT NUMBER	5b. GRANT NUMBER	5c. PROGRAM ELEMENT NUMBER
FA8750-19-C-1001	N/A	61101E
5d. PROJECT NUMBER	**5e. TASK NUMBER**	**5f. WORK UNIT NUMBER**
		R2TW

6. AUTHOR(S)

Assaf Arbelle, Graeme Blackwood, Leonid Karlinsky, Aadarsh Sahoo, Joseph Schtok, Rogerio Feris

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

IBM Thomas J. Watson Research Center
1101 Kitchawan Rd
Yorktown Heights NY 10598

8. PERFORMING ORGANIZATION REPORT NUMBER

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

Air Force Research Laboratory/RISB
525 Brooks Road
Rome NY 13441-4505

DARPA/I2O
675 N. Randolph St.
Arlington VA 22203-2114

10. SPONSOR/MONITOR'S ACRONYM(S)

AFRL/RI & DARPA

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

AFRL-RI-RS-TR-2023-135

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for Public Release; Distribution Unlimited. This report is the result of contracted fundamental research deemed exempt from public affairs security and policy review in accordance with SAF/AQR memorandum dated 10 Dec 08 and AFRL/CA policy clarification memorandum dated 16 Jan 09.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Our team obtained excellent results in all of the official evaluations of the Darpa Learning with Less Labels (LwLL) program. We participated in the image classification, object detection, and machine translation tasks. For many of the checkpoints, we achieved the very best results across all performers, significantly outperforming the baseline provided by the JPL team. A key lesson learned to achieve these strong results was to focus on finding a good feature embedding, so that new tasks can be learned with just a few examples. In this vein, we proposed novel approaches for representation learning (across domains and modalities, using real and synthetic data), explored new model architectures, as well as transfer learning techniques. These research works have been published in top-tier AI conferences and several of them integrated into our high-performing systems delivered to Darpa for the official program evaluations.

15. SUBJECT TERMS

Learning with Limited Labels, Representation Learning, Transfer Learning

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES
a. REPORT	b. ABSTRACT	C. THIS PAGE		
U	U	U	SAR	37

19a. NAME OF RESPONSIBLE PERSON	19b. PHONE NUMBER *(Include area code)*
PETER A. JEDRYSIK	N/A

Table of Contents

List of Figures

EXECUTIVE SUMMARY

Our team obtained excellent results in all of the official evaluations of the DARPA Learning with Less Labels (LwLL) program. We participated in the image classification, object detection, and machine translation tasks. For many of the checkpoints, we achieved the very best results across all performers, significantly outperforming the baseline provided by the JPL team.

A key lesson learned to achieve these strong results was to focus on finding a good feature embedding, so that new tasks can be learned with just a few examples. When the LwLL program first started, the status quo in few-shot learning was largely dictated by sophisticated meta-learning techniques. Our ECCV 2020 papers "Rethinking few-shot classification: is a good embedding all you need?" [1] and "A broader study of cross-domain few-shot learning" [2] have conveyed to the community that using a good learned embedding model can be more effective than sophisticated meta-learning algorithms. In addition, when evaluated in our proposed cross-domain few-shot learning benchmark, SOTA meta-learning methods at the time were outperformed in relation to simple fine-tuning by 12.8% average accuracy.

As a result of these findings, we focused our research efforts on techniques to identify a good feature embedding, including self-supervised representation learning and novel model architectures. Specifically, to facilitate learning with less data (in addition to less labels), and to mitigate important shortcomings related to privacy, ethics, and copyright attribution, we explored representation learning based on synthetic data generated by graphics simulators for vision, and procedural tasks for machine translation. We have also proposed novel approaches for representation learning across domains, and across different modalities such as vision, audio, and language, all of which have strong potential to be applied in real-world use cases within the scope of the LwLL program. A good feature embedding also depends on the model architecture. We have proposed novel transformer models for both image and video classification, with strong performance in standard datasets and capabilities to transfer across a wide range of downstream tasks. Finally, we have conducted research on other topics related to LwLL, including transfer learning and generative data augmentation. Overall, our research has led to more than 20 publications in the most prestigious AI conferences around the world.

We would like to highlight that several of the research works described above have been integrated into our high-performing systems delivered to DARPA for the official program evaluations. For example, we have integrated our RegionViT architecture [3] (ICLR 2022), trained our models using both real and synthetic data [4] (CVPR 2022), and performed representation learning using both labeled and unlabeled data (similar to our Dynamic Distillation Network [5], NeurIPS 2021, and Task-Adaptive Feature Sub-Space Learning [6], ECCV 2020). Our proposed cross-domain few-shot learning benchmark [2] (ECCV 2020) has been used by other LwLL performers and the external community.

Beyond research accomplishments and system deliverables, we have organized a tutorial at ICCV 2019 and a workshop at CVPR 2020 on visual learning with limited labeled data. Both events received quite significant attention by the community and involved the participation of several performers of the DARPA LwLL program. We have also developed interactive demos to showcase our research to transition partners in the final PI meeting.

1. INTRODUCTION: IS A GOOD EMBEDDING ALL YOU NEED?

When we started our participation in the LwLL program, meta-learning was the prevailing approach for few-shot learning. The idea was to design "learning to learn" algorithms that could quickly adapt to test time tasks with limited data and low computational cost.

While significant progress has been made along this direction, we have shown that a very simple baseline based on learning a supervised or self-supervised representation on the meta-training set, followed by training a linear classifier on top of this representation (Figure 1), outperformed state-of-the-art few-shot learning methods (Figure 2). An additional boost was achieved by self-distillation. This demonstrates that a good learned embedding model can be more effective than sophisticated meta-learning algorithms. More details can be found in our ECCV 2020 paper [1], which was the basis for our Phase I image classification system.

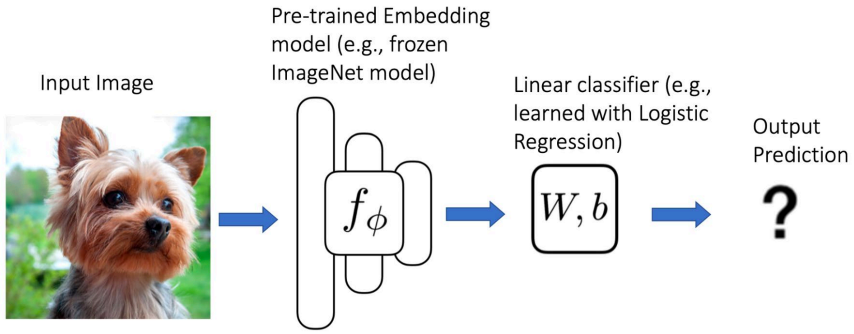


Figure 1. We use a simple feature extractor based on a pre-trained embedding model and then train a linear model using just a few labeled images.

		miniImageNet 5-way		tieredImageNet 5-way		
model	**backbone**	**1-shot**	**5-shot**	**1-shot**	**5-shot**	
MAML [12]	32-32-32-32	48.70 ± 1.84	63.11 ± 0.92	51.67 ± 1.81	70.30 ± 1.75	Complex SOTA Meta-Learning Methods
Matching Networks [55]	64-64-64-64	43.56 ± 0.84	55.31 ± 0.73	-	-	
IMP [2]	64-64-64-64	49.2 ± 0.7	64.7 ± 0.7	-	-	
Prototypical Networks† [46]	64-64-64-64	49.42 ± 0.78	68.20 ± 0.66	53.31 ± 0.89	72.69 ± 0.74	
TAML [21]	64-64-64-64	51.77 ± 1.86	66.05 ± 0.85	-	-	
SAML [15]	64-64-64-64	52.22 ± n/a	66.49 ± n/a	-	-	
GCR [27]	64-64-64-64	53.21 ± 0.80	72.34 ± 0.64	-	-	
KTN(Visual) [35]	64-64-64-64	54.61 ± 0.80	71.21 ± 0.66	-	-	
PARN[60]	64-64-64-64	55.22 ± 0.84	71.55 ± 0.66	-	-	
Dynamic Few-shot [14]	64-64-128-128	56.20 ± 0.86	73.00 ± 0.64	-	-	
Relation Networks [48]	64-96-128-256	50.44 ± 0.82	65.32 ± 0.70	54.48 ± 0.93	71.32 ± 0.78	
R2D2 [3]	96-192-384-512	51.2 ± 0.6	68.8 ± 0.1	-	-	
SNAIL [29]	ResNet-12	55.71 ± 0.99	68.88 ± 0.92	-	-	
AdaResNet [32]	ResNet-12	56.88 ± 0.62	71.94 ± 0.57	-	-	
TADAM [34]	ResNet-12	58.50 ± 0.30	76.70 ± 0.30	-	-	
Shot-Free [41]	ResNet-12	59.04 ± n/a	77.64 ± n/a	63.52 ± n/a	82.59 ± n/a	
TEWAM [37]	ResNet-12	60.07 ± n/a	75.90 ± n/a	-	-	
MTL [47]	ResNet-12	61.20 ± 1.80	75.50 ± 0.80	-	-	
Variational FSL [64]	ResNet-12	61.23 ± 0.26	77.69 ± 0.17	-	-	
MetaOptNet [26]	ResNet-12	62.64 ± 0.61	78.63 ± 0.46	65.99 ± 0.72	81.56 ± 0.53	
Diversity w/ Cooperation [11]	ResNet-18	59.48 ± 0.65	75.62 ± 0.48	-	-	
Fine-tuning [9]	WRN-28-10	57.73 ± 0.62	78.17 ± 0.49	66.58 ± 0.70	85.55 ± 0.48	
LEO-trainval† [44]	WRN-28-10	61.76 ± 0.08	77.59 ± 0.12	66.33 ± 0.05	81.44 ± 0.09	
Ours-simple	ResNet-12	62.02 ± 0.63	79.64 ± 0.44	69.74 ± 0.72	84.41 ± 0.55	Simple Baseline (embedding + linear model)
Ours-distill	ResNet-12	**64.82 ± 0.60**	**82.14 ± 0.43**	**71.52 ± 0.69**	**86.03 ± 0.49**	

Figure 2. Average few-shot classification accuracies with 95% confidence intervals on miniImageNet and tieredImageNet

In parallel to this work, we have proposed a new benchmark for cross-domain few-shot learning [2], analyzing the transferability of models pre-trained on ImageNet to domains of varying dissimilarity from natural images, including medical, aerial, and agriculture domains. We have

conducted a comprehensive experimental analysis on the proposed benchmark to evaluate state-of-the-art meta-learning approaches, transfer learning approaches, and newer methods for cross-domain few-shot learning. The surprising results demonstrate that state-of-art meta-learning methods are outperformed by earlier meta-learning approaches, and all meta-learning methods underperformed in relation to simple fine-tuning by 12.8% average accuracy. In some cases, meta-learning even underperformed networks with random weights.

These results served as motivation to adjust our research direction to focus on finding a good feature embedding by exploring novel representation learning methods and model architectures. In particular, we proposed novel methods for representation learning based on synthetic data (section 2.1), across domains (section 2.2), and across modalities (section 2.3). We have also proposed novel model architectures with a goal of finding a good feature embedding for LwLL (section 2.4) and conducted research on transfer learning (section 2.5). These works have been published in the most prestigious Artificial Intelligence conferences around the world and several of them have been integrated into our system deliverables, as part of the DARPA LwLL evaluations.

1.1 Performance on DARPA LwLL evaluations

Our team participated in the machine translation, image classification, and object detection evaluations. Figure 3 compares the performance of our system submissions for the final evaluation of the DARPA LwLL program against the baseline system provided by JPL. We have achieved quite significant improvements over the baseline, as can be seen in the plots.

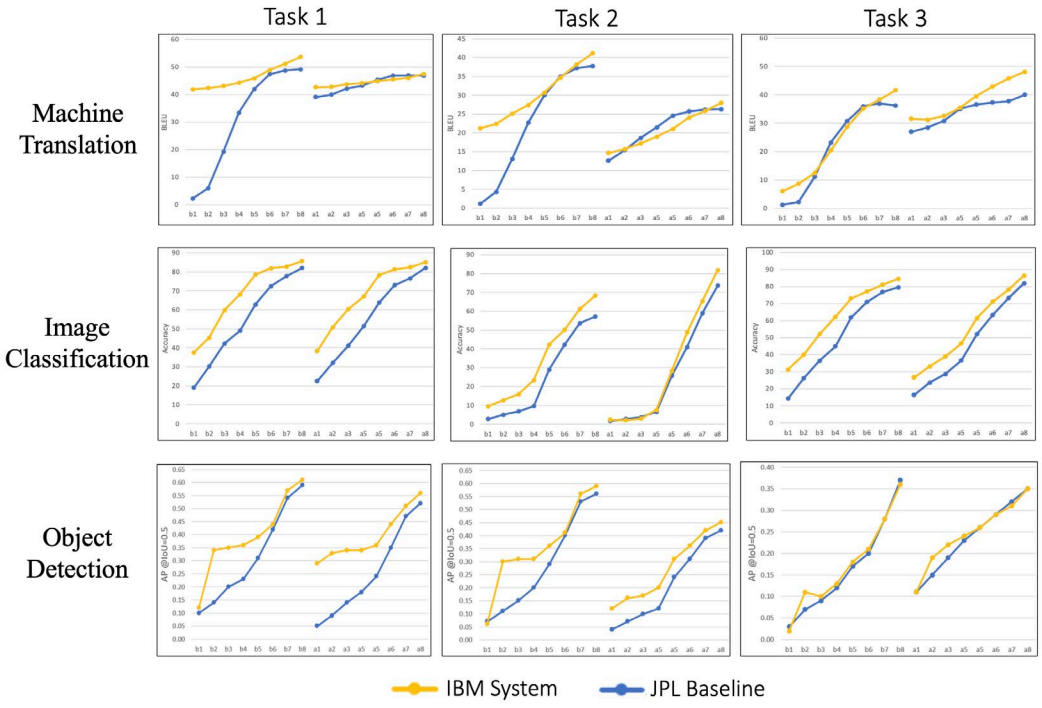


Figure 3. Performance of our system in comparison with JPL baseline (final evaluation)

We would like to emphasize that several of our papers published in top conferences during the LwLL program have been integrated into the systems we delivered for the DARPA evaluations [1][3][4][5]. See section 3 for more details about our system deliverables.

2. METHODS, ASSUMPTIONS, AND PROCEDURES

2.1 Representation Learning based on Synthetic Data

Learning a feature embedding from synthetic data is important for many reasons. Firstly, it allows us to learn with "less data", since we assume synthetic data can be generated on the fly by a simulator. Secondly, leveraging synthetic data can mitigate many of the shortcomings inherent in real data, including privacy, bias, data protection, and copyright attribution. Finally, embeddings from synthetic data can be used in tandem with features learned from real data, as demonstrated in our image classification system delivered to DARPA. In this section, we discuss our research on synthetic tasks and data for both image classification and machine translation.

2.1.1. Synthetic Data Pre-training for Image Classification

We study, for the first time, the transferability of pre-trained models based on synthetic data generated by graphics simulators to downstream tasks from very different domains. In using such synthetic data for pre-training, we find that downstream performance on different tasks is favored by different configurations of simulation parameters (e.g. lighting, object pose, backgrounds, etc.), and that there is no one-size-fits-all solution. For the best performance, it is thus better to tailor synthetic pre-training data to a specific downstream task. To this end, we propose Task2Sim, a unified model that maps downstream task representations to optimal simulation parameters to generate synthetic pre-training data for them. Task2Sim learns this mapping by training to find the set of best parameters on a set of "seen" tasks. Once trained, it can then be used to predict best simulation parameters for novel "unseen" tasks in one shot, without requiring additional training.

Figure 4 shows more details about our approach. During training, a batch of "seen" tasks is provided as input. Their task2vec vector representations are fed as input to Task2Sim, which is a parametric model (shared across all tasks) mapping these downstream task2vecs to simulation parameters, such as lighting direction, amount of blur, background variability, etc. These parameters are then used by a data generator (in our implementation, built using the Three-D-World platform) to generate a dataset of synthetic images. A classifier model then gets pre-trained on these synthetic images, and the backbone is subsequently used for evaluation on specific downstream tasks. The classifier's accuracy on this task is used as a reward to update Task2Sim's parameters.

Given a budget in number of images per class, our extensive experiments with 20 diverse downstream tasks (see our CVPR 2022 paper [4] for details) show Task2Sim's task-adaptive pre-training data results in significantly better downstream performance than non-adaptively choosing

simulation parameters on both seen and unseen tasks. It is even competitive with pre-training on real images from ImageNet.

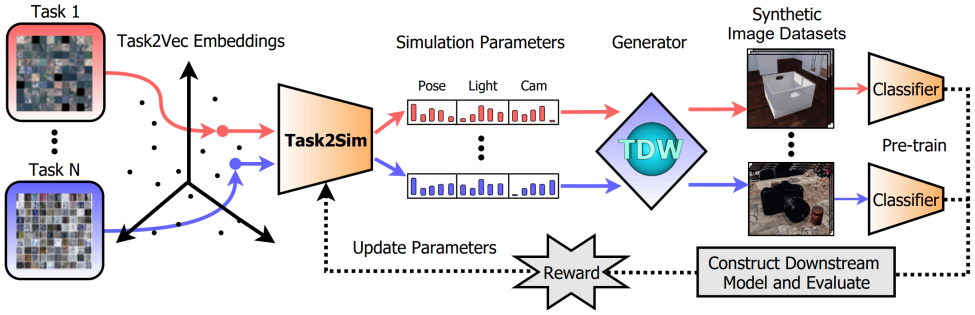


Figure 4. Overview of our Task2Sim approach for pre-training models with synthetic data, which is adaptively generated based on the input task to maximize downstream task performance.

We have also explored synthetic data pre-training for video classification [7] and visual question-answering [8]. In addition, we have shown that it is possible to pre-train models with short procedural programs (shaders) [9], which offers better scalability and efficiency compared to data generation based on graphics simulators.

2.1.2. Synthetic Pre-Training Tasks for Neural Machine Translation

As motivated above, pre-training models with web-scale crawled corpora can lead to issues of toxicity and bias, as well as copyright and privacy concerns. A promising way of alleviating such concerns is to conduct pre-training with synthetic tasks and data, since no real-world information is ingested by the model. Our ACL 2023 paper [10] proposes several novel approaches to pre-training NMT models with different levels of lexical and structural knowledge, including (i) generating obfuscated data from a large parallel corpus, (ii) concatenating phrase pairs extracted from a small word-aligned corpus, and (iii) generating synthetic parallel data without real human language corpora. Experiments on multiple language pairs show that pre-training benefits can be realized even with high levels of obfuscation or purely synthetic parallel data. We give a brief overview of two of our synthetic pre-training tasks below.

Our first task (pb-trees) generates purely synthetic parallel sentence pairs from permuted aligned binary trees of "nonsense" symbols. The tree structure is intended to model some aspects of the processes that occur during natural language translation, e.g. the reordering of contiguous spans. An example pb-trees synthetic sentence pair is shown in Figure 5.

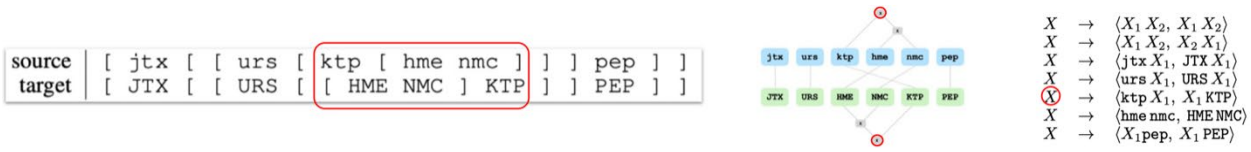


Figure 5. Synthetic sentence pair created by permuting aligned binary trees (pb-trees).

Our second task (phrase-cat) first extracts a set of aligned phrase pairs from a small fine-tuning parallel corpus. We then generate new synthetic parallel data by a simple concatenation of uniformly sampled phrase pairs. Although the boundaries between phrases are frequently ungrammatical, useful information for pre-training NMT models such as local word order and

alignments are preserved within each phrase pair. Figure 6 shows an example synthetic sentence pair created by concatenation of phrases extracted from a small Indonesian-to-English parallel corpus.

src	[sejak Wright] [sambil seringkali] [kami] [50 juta mengingat]
trg	[from Wright] [in most times] [we] [50 millions as]

Figure 6. Synthetic sentence pair created by concatenating aligned phrases (phrase-cat).

We evaluate our NMT synthetic pre-training tasks on three low-resource language pairs with relatively small fine-tuning set sizes: Burmese-to-English (`my-en`; 18.0k), Indonesian-to-English (`id-en`; 24.5k), and Turkish-to-English (`tr-en`; 207.7k). Figure 7 shows improvements in BLEU scores for both the pb-trees and phrase-cat synthetic pre-training tasks compared to a naïve baseline that trains a randomly initialized model using only the fine-tuning data. We call attention to the fact that these gains over the baseline are achieved without any additional monolingual or parallel labeled data.

Pre-Training	my-en	id-en	tr-en
random-init	4.1	18.2	14.7
pb-trees	11.4	23.1	14.4
phrase-cat	14.0	27.3	16.5

Figure 7. Synthetic pre-training tasks vs. fine-tuning a randomly initialized model (BLEU).

2.2 Representation Learning across Domains

One of the more powerful qualities of human cognition is the ability to apply our sensory signals in other previously unseen domains with little or no labeled data. This quality of "cross-domain generalization" is a necessary trait in many real-world applications. In our series of works on this topic, we have covered numerous aspects of this important and impactful research problem, including: (i) self-supervised domain adaptation and generalization without any paired data (CVPR 2022) [11]; (ii) cross-domain few-shot learning allowing leveraging other domains pre-training (ECCV 2020, NeurIPS 2021) [2][5]; and (iii) enhancing cross-domain transferability via contrastive learning (ICCV 2021) [12].

In our paper "Unsupervised domain generalization by learning a bridge across domains" [11] we aim to find the best representation for images across domains by mapping all domains to a joint "bridge" domain in which the inter-domain variance is minimal. In contrast to most cross-domain

papers which utilize the source domain for supervision, we apply a relatively new and very practical methodology of unsupervised domain generalization (UDG), where we have no training supervision in the source nor in the target domains. We learn the representation by leveraging our Bridge Across Domains (BrAD) which is an auxiliary domain accompanied by a set of semantics preserving visual mappings to BrAD from each of the training domains. The BrAD and mappings to it are learned jointly with a contrastive self-supervised representation model that semantically aligns each of the domains to its BrAD-projection, and hence implicitly drives all the domains, including unseen ones, to semantically align to each other (see Figure 8). In this work, we show how using an edge-regularized BrAD achieves significant gains across multiple benchmarks and a range of tasks, including UDG, few-shot unsupervised domain adaptation, and unsupervised generalization across multi-domain datasets, including generalization to unseen domains and classes.

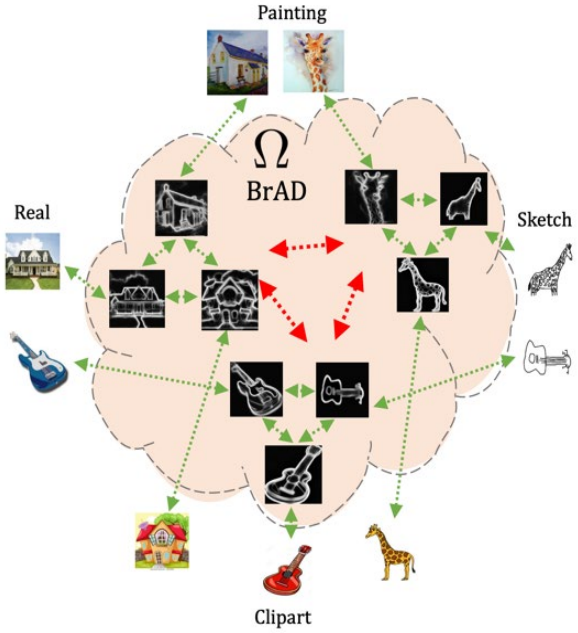


Figure 8. An illustration of our Bridge Across Domains (BrAD) approach.

In another work "Dynamic distillation network for cross-domain few-shot recognition with unlabeled data" [5], we tackle the problem of cross-domain few-shot learning where there is a large shift between the base and target domain. The problem of cross-domain few-shot recognition with unlabeled target data is largely unaddressed in the literature (see Figure 9). STARTUP was the first method that tackles this problem using self-training. However, it uses a fixed teacher pretrained on a labeled base dataset to create soft labels for the unlabeled target samples. As the base dataset and unlabeled dataset are from different domains, projecting the target images in the class-domain of the base dataset with a fixed pretrained model might be sub-optimal. We propose a simple dynamic distillation-based approach to facilitate unlabeled images from the novel/base dataset. We impose consistency regularization by calculating predictions from the weakly-augmented versions of the unlabeled images from a teacher network and matching it with the strongly augmented versions of the same images from a student network. The parameters of the teacher network are updated as exponential moving average of the parameters of the student network. We show that the proposed network learns representation that can be easily adapted to the target domain even though it has not been trained with target-specific classes during the pretraining phase. Our model outperforms the current state-of-the art method by 4.4% for 1-shot and 3.6% for 5-shot classification in the BSCD-FSL benchmark, and also shows competitive performance on traditional in-domain few-shot learning task.

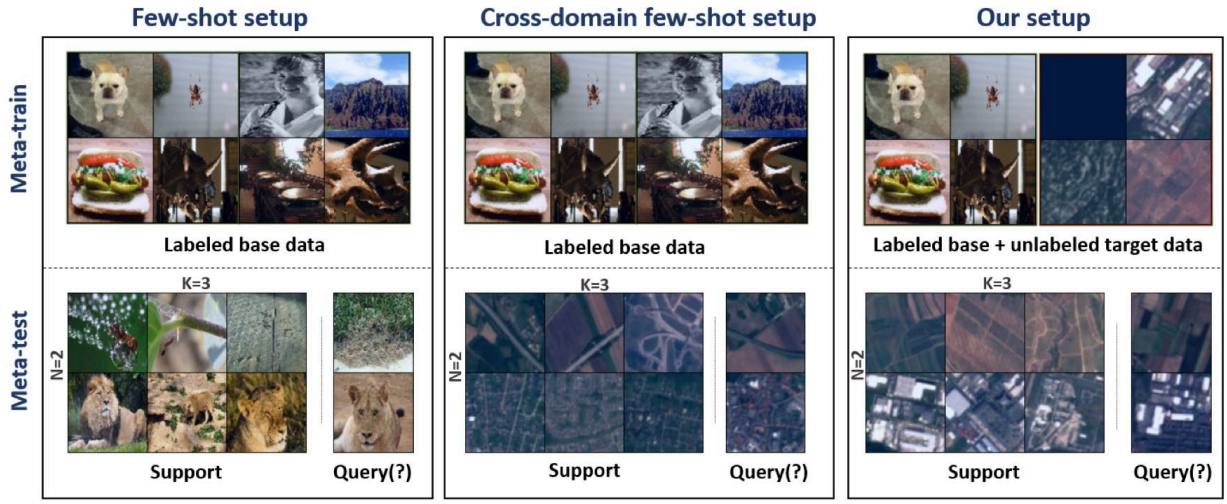


Figure 9. Our cross-domain few-shot learning setting with additional unlabeled images.

2.3 Representation Learning across Modalities

Recent studies popularized the use of massive scale free-form text as a means of cross-modal supervision, unlocking use cases such as zero-shot classification, which is very relevant within the scope of the LwLL program. We have done research work in the domain of cross-modal representation learning: (i) using self-training, multiple-instance learning, and language side expansion for improved zero-shot recognition in videos [13]; (ii) multiple-instance learning for improved expert domains cross-modal alignment (NeurIPS 22) [14]; (iii) continual, data-free, improvement without forgetting of vision & language models (CVPR 23) [15]; (iv) text-side augmentation (CVPR 2023) [16] and synthetic data for enhancing zero-shot compositional reasoning capabilities in Vision & Language models [17]; and (v) effective cross-modal grounding (ICCV 21) [18]. Bellow we describe in more detail two of our representative publications along this direction.

Foundation Models (FMs) have demonstrated unprecedented capabilities including zero-shot learning, high fidelity data synthesis, and out of domain generalization. However, as we show in our paper "FETA: Towards Specializing Foundation Models for Expert Task Applications" [14], FMs still have poor out-of-the-box performance on expert tasks (e.g. retrieval of car manuals technical illustrations from language queries), data for which is either unseen or belonging to a long-tail part of the data distribution of the huge datasets used for FM pre-training. This underlines the necessity to explicitly evaluate and finetune FMs on such expert tasks, arguably ones that appear the most in practical real-world applications. In our work, we propose a first of its kind FETA benchmark built around the task of teaching FMs to understand technical documentation, via learning to match their graphical illustrations to corresponding language descriptions. Our FETA benchmark focuses on text-to-image and image-to-text retrieval in public car manuals and sales catalogue brochures. FETA is equipped with a procedure for completely automatic annotation extraction, allowing easy extension of FETA to more documentation types and application domains in the future. Our automatic annotation leads to an automated performance metric shown to be consistent with metrics computed on human-curated annotations (also released). We provide

multiple baselines and analysis of popular FMs on FETA leading to several interesting findings that we believe would be very valuable to the FM community, paving the way towards real-world application of FMs for practical expert tasks currently "overlooked" by standard benchmarks focusing on common objects.

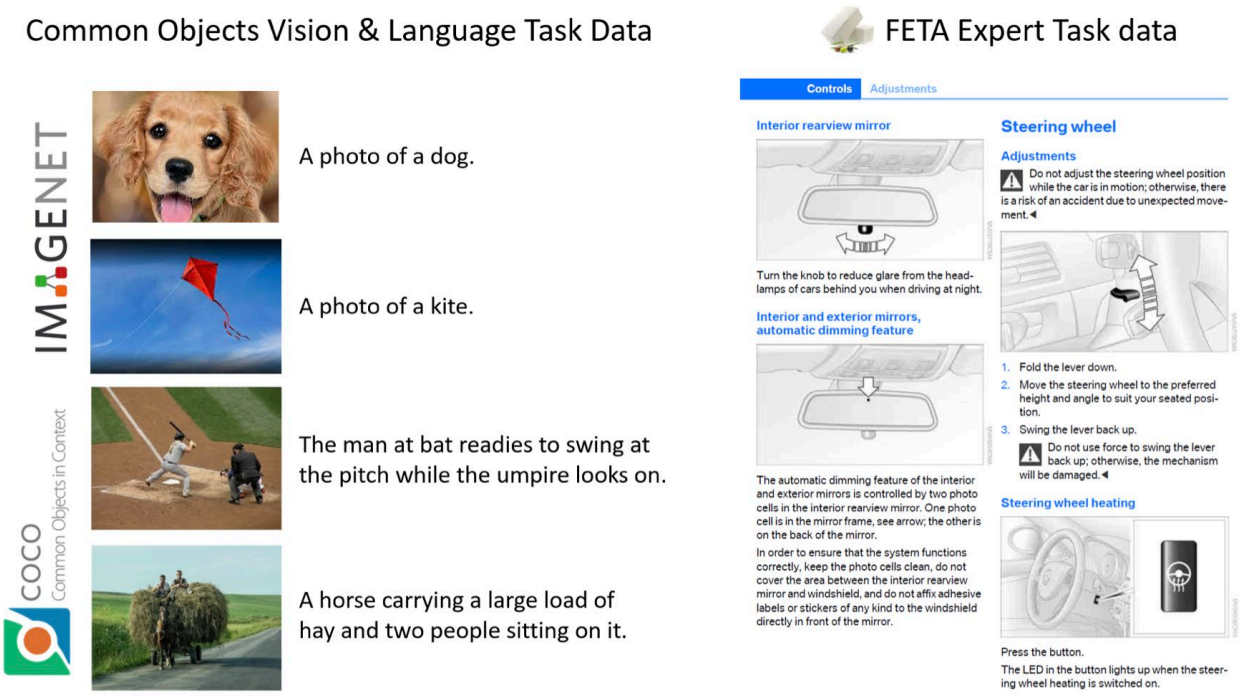


Figure 10. A comparison of the FETA dataset to common Vision and Language datasets.

In another work [16], we examine the abilities of recent Foundation Models understanding of complex concepts. Our paper "Teaching Structured Vision & Language Concepts to Vision & Language Models" introduces the collective notion of Structured Vision & Language Concepts (SVLC) which includes object attributes, relations, and states which are present in the text and visible in the image. While Vision and Language (VL) models demonstrated remarkable zero-shot performance in a variety of tasks, some aspects of complex language understanding remain a challenge. Recent studies have shown that even the best VL models struggle with SVLC. A possible way of fixing this issue is by collecting dedicated datasets for teaching each SVLC type, yet this might be expensive and time-consuming. Instead, we propose a more elegant data-driven approach for enhancing VL models' understanding of SVLCs that makes more effective use of existing VL pre-training datasets and does not require any additional data. While automatic understanding of image structure still

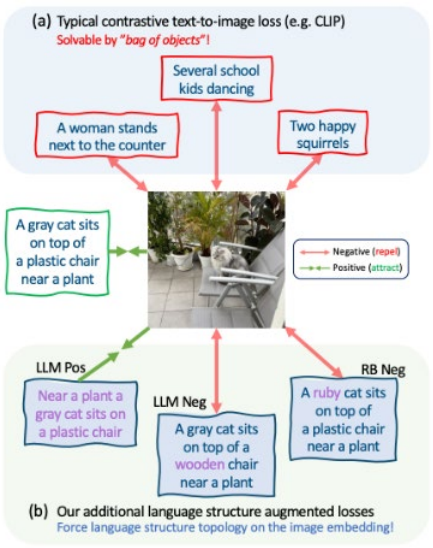


Figure 11. An illustration of our method for text augmentations to improve SVLC understanding.

remains largely unsolved, language structure is much better modeled and understood, allowing for its effective utilization in teaching VL models. We propose various techniques based on language structure understanding that can be used to manipulate the textual part of off-the-shelf paired VL datasets (see details in [16]). VL models trained with the updated data exhibit a significant improvement of up to 15% in their SVLC understanding with only a mild degradation in their zero-shot capabilities both when training from scratch or fine-tuning a pre-trained model.

2.4 Novel Model Architecture: RegionViT

Vision transformer (ViT) has recently shown its strong capability in achieving comparable results to convolutional neural networks (CNNs) on image classification. However, vanilla ViT simply inherits the same architecture from the natural language processing directly, which is often not optimized for vision applications. For example, the transformer has an isotropic network structure with a fixed number of tokens and unchanged embedding size, which loses the capability to model the context with different scales and allocates computations at different scales. Another critical bottleneck of the transformer is that the self-attention module has a quadratic cost in memory and computation with regard to the sequence length (i.e., the number of tokens). This issue is even worse in ViT as images are 2-D, suggesting a quadratic relationship between the number of tokens and image resolution. As a result, ViT indicates a quadruple complexity w.r.t image resolution. The highly compute- and memory-intensive self-attention makes it challenging to train vision transformer models at fine-grained patch sizes. Motivated by this, we propose a new architecture that adopts the pyramid structure and employ novel regional-to-local attention rather than global self-attention in vision transformers. Specifically, our approach first divides the input image into a group of non-overlapping patches of large size (e.g., 28×28), on which regional tokens are computed via linear projection. Similarly, local tokens are created for each region using a smaller patch size (e.g., 4×4). We then use a standard transformer to process regional and local tokens separately. To enable communication between the two types of tokens, we first perform self-attention on regional tokens (regional attention) and then jointly attend to the local tokens of each region including their associated regional token (local attention). By doing so, regional tokens pass global contextual information to local tokens efficiently while being able to effectively learn from local tokens themselves. Therefore, even though local self-attention confines the scope in a local region but it can still receive global information. For clarity, we represent this two-stage attention mechanism as Regional-to-Local Attention, or R2L attention for short (see Figure 12 for an illustration). Since both regional and local attention involve much fewer tokens, our R2L attention requires substantially less memory than regular global self-attention used in vision transformers. For example, in our default setting, the memory saving using R2L attention can be up to as much as 73%. We demonstrate the effectiveness of our approach on image classification and several downstream vision tasks including object detection and action recognition. We employ RegionViT in our image classification system which is described later in Section 4.1. For more details on RegionViT, check our ICLR 2022 paper [3].

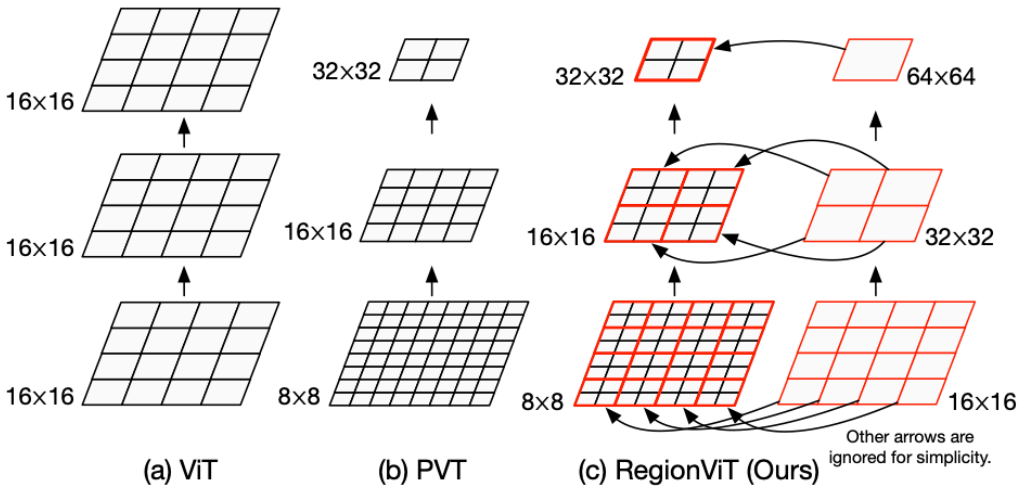


Figure 12. Regional-to-Local Attention for Vision Transformers.

2.5 Transfer Learning

Currently, transfer learning methods rely on manual decisions for questions such as: which pre-training model (feature embedding) to select? Which layers to freeze or fine-tune? Which features to share across tasks? Throughout the program, we have conducted research to automate these decisions.

Specifically, we proposed an adaptive fine-tuning approach, called SpotTune, which finds the optimal fine- tuning strategy per instance for the target data. In SpotTune, given an image from the target task, a policy network is used to make routing decisions on whether to pass the image through the fine-tuned layers or the pre-trained layers (see Figure 13). We conducted extensive experiments to demonstrate the effectiveness of our proposed approach. Our method outperforms the traditional fine-tuning approach on 12 out of 14 standard datasets. We also compared SpotTune with other state-of-the-art fine-tuning strategies, showing superior performance. On the Visual Decathlon datasets, our method achieves the highest score across the board without bells and whistles. For more details about SpotTune and results, see our CVPR 2019 paper [19].

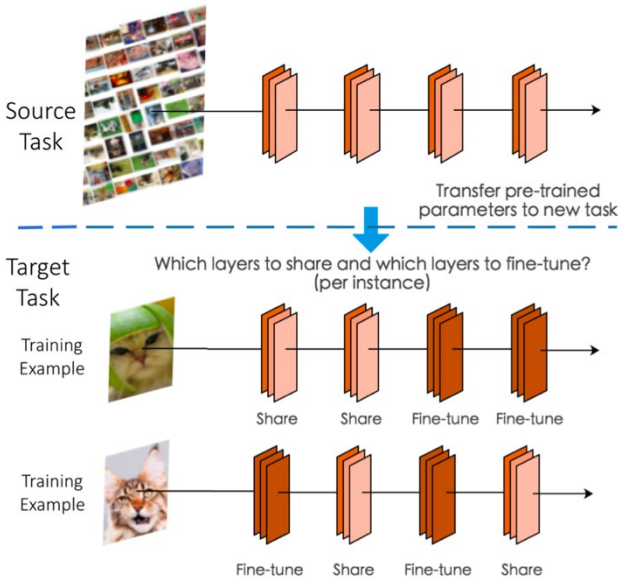


Figure 13. Our proposed method, SpotTune, automatically decides which layers of the network model should be fine-tuned to improve transfer learning performance.

We have also extended SpotTune to multi-task learning (see Adashare, NeurIPS 2020 [20]) and conducted a comprehensive analysis related to the impact of contrastive representations in transfer learning performance [12].

Finally, we have conducted research on pre-trained model selection. Figure 14 shows an analysis of the effectiveness of model selection considering the size of the source training set and similarity to the target domain. See more details in [21]. Our work on model selection has been integrated into our image classification and objected detection systems delivered to DARPA, as detailed later.

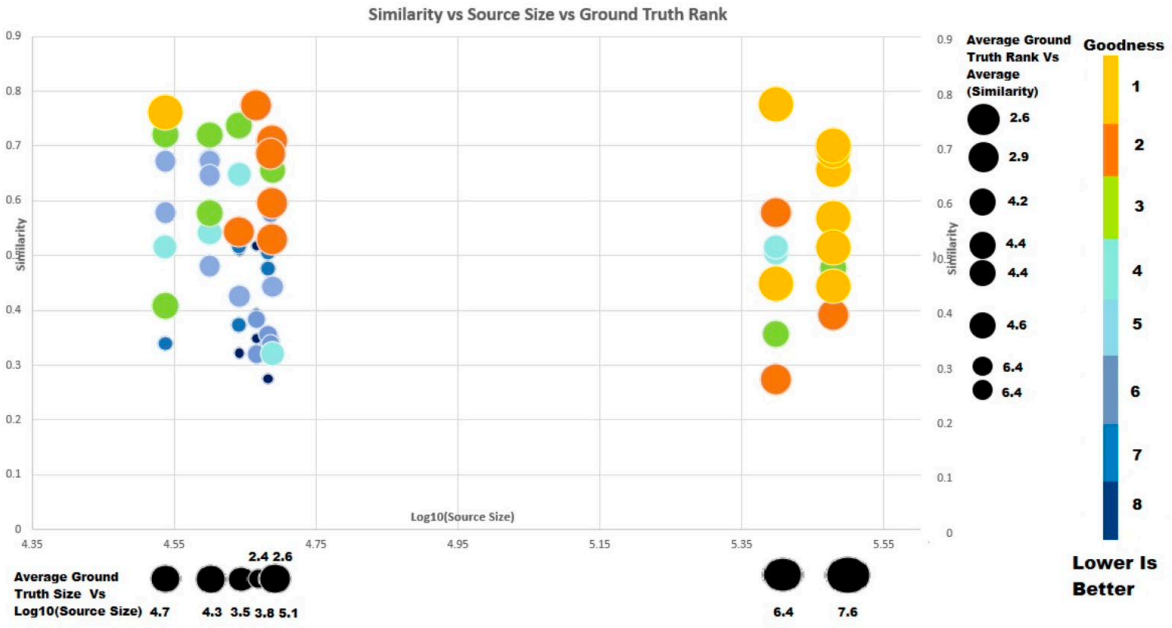


Figure 14. Relationship of performance of the target model to size of the source dataset (X-axis) and its similarity with source dataset (Y -axis) for 9 targets over 8 sources

3. SYSTEM DELIVERABLES, RESULTS, AND DISCUSSION

3.1 Image Classification

In this section, we will describe our phase-3 image classification system. Our design choices have followed the research outcomes we obtained in Section 2. First, we will discuss the various architectural details of our system, then we will show how we leverage synthetic data in our system and other training details. Finally, we show the performance on the development datasets with analysis through ablation studies. We mainly focus on our motivation of choosing a good embedding and try to resonate with it through innovations in model architecture and training data.

3.1.1 System Overview

In Figure 15, we show the overview of the image classification system. The system backbone consists of an ensemble group of four different model architectures, namely, from left-to-right, – M1: EfficientNet-V2-S [22] M2: IBM RegionViT-S [3], M3: SwinV2 Transformer [23], M4: EfficientNet-V2-S [22]. The advantages of convolutional neural networks (CNNs) over vision transformers and vice versa has been well studied in the computer vision literature [24]. While the sliding-window strategy of training CNNs can learn feature embeddings with information like corners and lines, self-attention in transformers can capture information across distant image locations. In order to capture the best of both worlds and harness complementary information from different models, our system is comprised of state-of-the-art architectures belonging to both categories. We follow a two-step training process in deploying our system: (1) *representation learning through pretraining* and (2) *finetuning for few-shot classification*. As shown in the figure, for pretraining, we consider all the available stage data (base/adapt) as unlabeled and pretrain models M1, M2, and M3 using the self-supervised DINO loss [25]. On the other hand, we use synthetic data to pretrain the model M4 with full supervision, as we already have the labels for all the images. Once we have all the 4 models pretrained, for the first 3 checkpoints, we freeze the backbone of each model, which makes each of the models act as a feature extractor. Once, we have the features extracted, for few-shot classification, we employ a linear support vector classifier on top of the extracted features and obtain the predictions for each of the four models. We then average the predictions from all the models to obtain the final ensemble prediction. In the following sections, we will describe each of the components in detail.

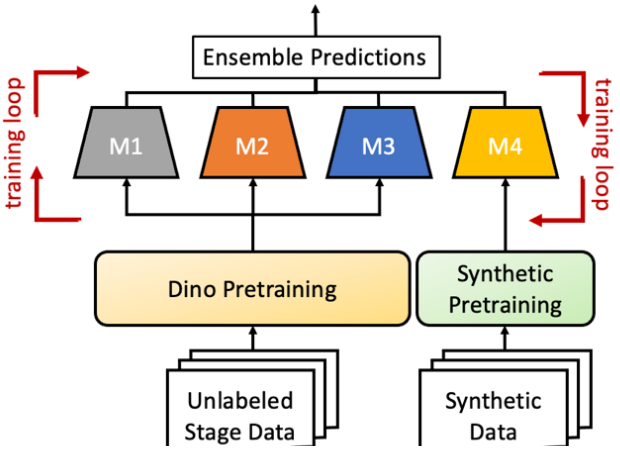


Figure 15. Overview of the image classification system.

3.1.2 Leveraging Synthetic Data

In our system we explore the use of synthetic data generated using the Task2Sim framework described in Section 2.1. We use the best parameters from Task2Sim optimization and pretrain an EfficientNet-V2 model in a fully-supervised fashion using cross-entropy loss as shown in Figure 16. The pretraining is performed as an offline step and once done, the Synthetic model is used as a feature extractor in the further fine-tuning steps.

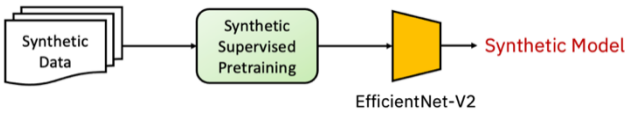


Figure 16. Overview of the synthetic pretraining step.

3.1.3 Self-supervision through DINO for better initialization

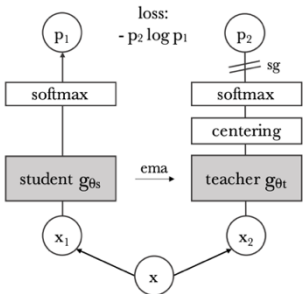
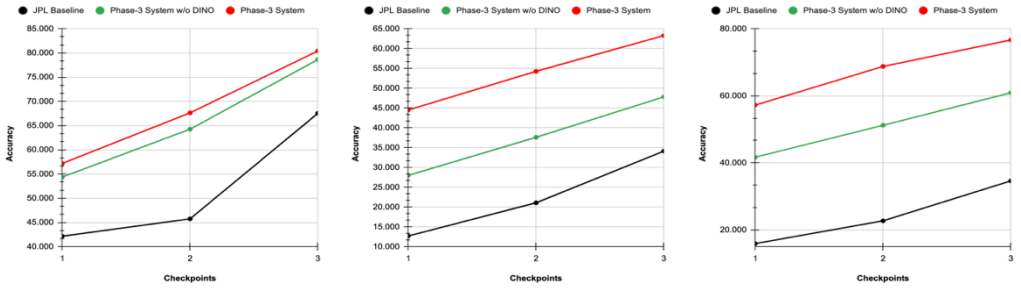


Figure 17. Overview of self-supervised pretraining using DINO. Figure borrowed from Caron et. al. [25].

As mentioned above, in the pretraining step, we employ DINO loss for self-supervision pretraining of all the models. As shown in Figure 17, DINO leverages two different views or augmentations of the same image and passes them through a student and a teacher network of the same architecture. The similarity of the output predictions is measured using a cross-entropy loss. The output of the teacher network is centered and gradients are not backpropagated through it; instead, exponential moving average is used for updating the parameters of the teacher network. For more details, please refer to the DINO paper by Caron et. al. [25]. We pretrain models M1, M2, and M3 using DINO considering all the available stage data as unlabeled. E.g. in the base stage, we take all the images and pretrain the models using DINO, while in the adapt stage we take all the images in adapt as well as base (only if the class overlap between base and adapt is 75%, else we consider only adapt data) and pretrain the models using DINO. Once pretrained, we use the weights as initialization for the few-shot finetuning ahead. Figure 18 shows the effect of using DINO pretraining on the three development tasks. As can be seen, DINO helps boost the performance considerably for all the three datasets and hence provides a robust initialization for few-shot finetuning.



(a) Mars Surface *(b) DomainNet Clipart* *(c) CIFAR-100*

Figure 18. Comparison of performance with and without self-supervised pretraining with DINO.

3.1.4 Dynamic Model Selection for Later Checkpoints

We use the four-model ensemble group only for the first three checkpoints where the labeled data available to us is very low. As we progress to the later checkpoints, the amount of labeled data increases and hence, instead of freezing the backbones of all the model architectures, in order to achieve good performance, we need to leverage full-finetuning of these models. But, with a given time constraint, finetuning all the four models is not feasible; hence, we adopt a dynamic model selection strategy for the checkpoints 4 and beyond. We start with 4 models: 2 copies of EfficientNet-V2 (one initialized with DINO and one with ImageNet weights), and 2 copies of RegionViT (one initialized with DINO and one with ImageNet weights). We then perform a logistic regression prediction using all the 4 models and select the best 2 models for further

training. After selection, we unfreeze the backbone and train the models on the available labeled data for the checkpoint. In addition to this, we also incorporate an adaptive learning rate selector, which basically performs a few steps of training on these models, gets validation accuracy and chooses the learning rate values providing the best validation accuracy. This design is important as we do not know what type of data the model gets to see, and as different distributions of data are sensitive to different learning rate values.

3.1.5 Performance on Development Datasets

We extensively validate our system on multiple development datasets covering a wide variety of data. From high domain shift datasets like Mars Surface, Mars Curiosity, to large-scale datasets like DomainNet, in Figure 19, we report the performance of our system and compare it with the JPL baseline and our Phase-2 system. As can be observed, the Phase-3 system provides superior accuracy as compared to the Phase-2 system and the JPL baseline. We mainly attribute this to the complementary information from each of the models in the ensemble group which is harnessed for improved classification performance on all the datasets.

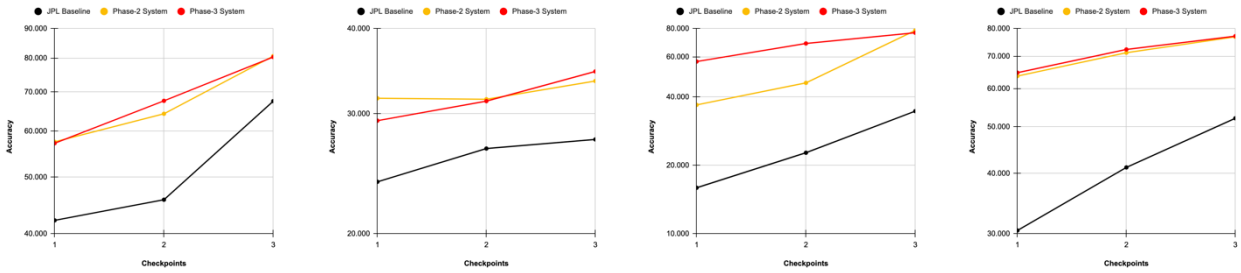


Figure 19. Performance on development datasets. Plots show the accuracy of the JPL baseline, Phase-2 system, and Phase-3 system on, from left-to-right, Mars Surface, Mars Curiosity, CIFAR-100, and DomainNet-Real datasets.

3.1.6 Ablation Studies

In order to analyze the contribution of each of the models in the ensemble group, in Figure 20 and Figure 21, we remove a given model from the system and then observe the change in performance

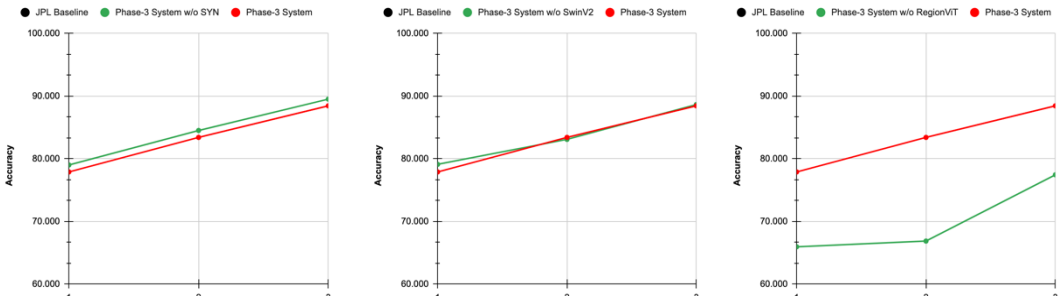


Figure 20. The above plots, from left to right, show accuracy by removing the Synthetic model, the SwinV2 transformer, and the RegionViT model, respectively, on CIFAR-100 dataset.

in CIFAR-100 and EuroSAT dataset, respectively. As can be seen from Figure 20, for CIFAR-100 dataset, removing the Synthetic model and the SwinV2 transformer results in a considerable drop in accuracy, while removing the RegionViT model shows slight decrease in the performance. This implies that CIFAR-100 benefits greatly from the Synthetic model and the SwinV2 transformer as compared to the RegionViT model.

On the other hand, as can be observed from Figure 21, for EuroSAT dataset, removing the Synthetic model and the SwinV2 transformer slightly reduce the performance, while removing the RegionViT model from the system drastically drops the accuracy. This tells us the importance of the RegionViT model for the EuroSAT dataset.

These observations tell us that different models can help different datasets in different magnitudes and corroborate our motivation of finding a good embedding by hitchhiking an ensemble of strong models.

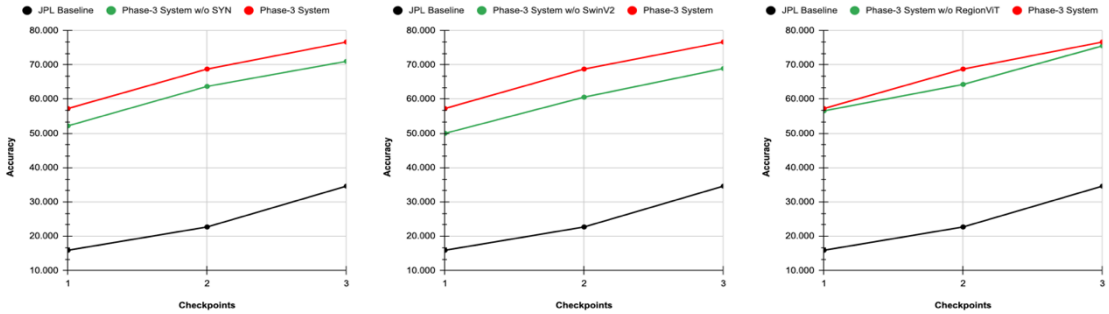


Figure 21. The above plots, from left to right, show accuracy by removing the Synthetic model, the SwinV2 transformer, and the RegionViT model, respectively, on EuroSAT dataset.

3.2 Object Detection

3.2.1. Overview

Our object detection system (Figure 22) is based on an ensemble of models, including those pretrained on the whitelist datasets. In the recurrent training scheme of the LwLL program, our flow consists of (1) initial performance evaluation of the models, (2) training on the available labels, with data-dependent hyper-parameters, in a semi-supervised setting, and (3) merging the predictions of the models on test images for final output.

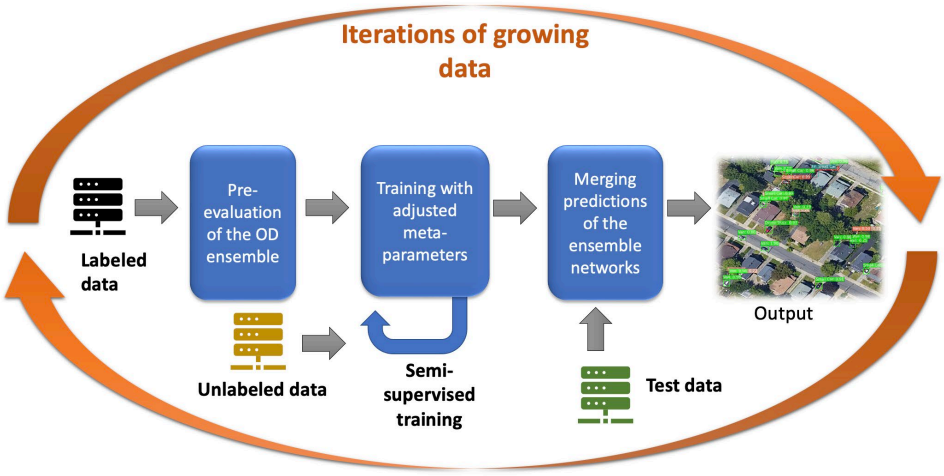


Figure 22. Overview of our object detection system.

3.2.2 Main features of the system

Dynamic model selection

Our dynamic model selection mechanism consists of performance evaluation of the ensemble models prior to the training at each checkpoint, using the new available labels as validation data. Post-training, this evaluation is used to weight the models for predictions merging, in the simplest case just selecting the best model for the current checkpoint. Furthermore, the implemented logic accounts for possible model divergence, stopping its training and removing from the ensemble, and also trims the ensemble towards the last (longest) checkpoints, retaining only the best-performing models.

Automatic hyper-parameter selection

The following parameters are critical for any object detection system performance: (1) learning rate of the optimizer, (2) image sizes used for training/inference, (3) number of iterations for the training. These parameters depend on the dataset, but properties of the target datasets are a-priori unknown. We perform automatic selection of these hyperparameters during the training.

Learning rate, (LR): a few copies of a model with 10x-factor difference in LR are used in initial (fast) checkpoints, and the best version is retained.

Image sizes: Batch size (per GPU) depends on the image resolution used. So, if the original image sizes are large, we decrease the batch size in order to maintain the resolution and still fit into GPU memory.

Number of iterations: it usually depends linearly on the dataset size, but in case of the LwLL regime with exponentially increasing data sizes, such approach will lead to significantly high number of iterations. Therefore, we use a sublinear regime for the number of iterations as a function of the number of samples in each checkpoint.

Semi-supervised learning (SSL)

Initially, we have attempted to deploy the standard Unbiased Teacher (UBT) algorithm [26] for model training at early checkpoints (where it is still feasible to use SSL, due to time limitations), but this did not lead to notable improvements. In Phase 3, we have implemented an ad-hoc version of SSL training, based on ideas of UBT and specialized image augmentation, featured in recent literature [27][28]. Specifically, pseudolabels, generated with a fully-supervised model are used to augment the original training data by the mixup technique, gluing in the object crops from unlabeled images. This operation is followed by affine transformations and color-space augmentations. While such operation is computationally expensive to apply for all LwLL checkpoints, it yields good improvements in first few (low-labels) checkpoints (Figure 23).

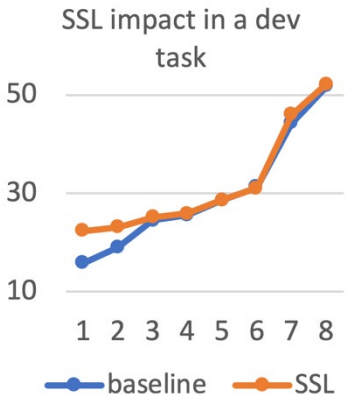


Figure 23. Performance improvement due to SSL training

Additional architectures

In Phase 3, we added the transformer-based SWIN architecture, in addition to ResNet and EfficientNet backbones as used before. This has enabled higher performance across the datasets (Figure 24).

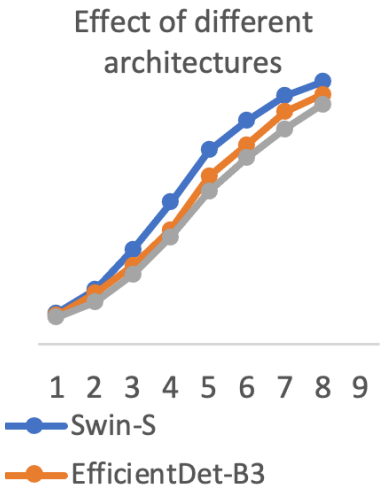


Figure 24. Performance with backbones of different architectures.

Support for aerial data

During the system development, in the period of evaluation by transition partners, we have extended the system support to aerial data, which has proved challenging for the phase 2 system. Augmentation by rotation of the training images by 90-degrees steps has led to performance improvements (such augmentation is natural for aerial data). We also found that longer training protocols and validation-based selection of best training checkpoint are valuable for such data.

3.2.3. Performance of Phase-2 vs Phase-3 systems

In the final evaluation, our system has shown a strong performance superior to the JPL baseline, which has justified the steps taken in the development. For most development datasets (provided by the program and our own), we observe performance increase relative to the previous version of the system (Figure 25). Here the blue graphs correspond to previous phase 2 system and orange graphs are for the latest system. In the first checkpoints, the SSL mechanism proves more effective, while stronger backbone and longer protocols have provided improvements along all checkpoints. Some performance drops, observed for aerial datasets in some checkpoints, are prevented by in-training validation for best iteration selection. However, for some checkpoints of some datasets,

the phase-2 system yields better results. We observe gains due to stronger backbone (SWIN) across the checkpoints, and gains for aerial data due to rotation-based augmentation. Overall, we have developed a strong and robust object detection system that can handle different dataset types with reduced expert supervision.

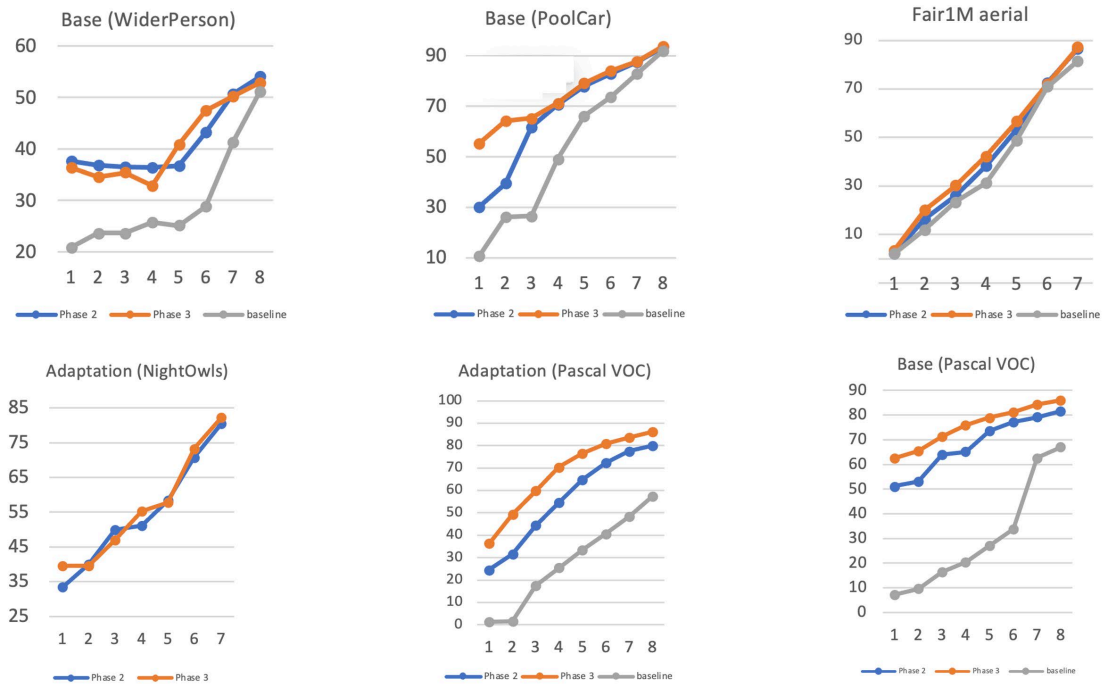


Figure 25. Performance graphs for phases 2, 3 of our system and the baseline (JPL) system, for some of the development datasets.

3.3 Machine Translation

For our Phase 3 MT submission, we developed a two-stage pre-training with translation task customization pipeline. The first stage consists of large-scale monolingual language model (LM) pre-training. We use the mBART-50 model which is trained as a sequence-to-sequence denoising autoencoder using the BART [29] objective on a large corpus of monolingual data covering 50 individual languages [30]. In the second stage of our pipeline, we customize the pre-trained mBART model for the task of translation using parallel data from the LwLL whitelist. The mBART model uses source and target language code embeddings to support translation between any two languages. Any language pairs excluded from the whitelist (which includes the languages of the development and evaluation tasks) represent zero-shot directions at the beginning of a LwLL session. When parallel data is presented to the model we tokenize and prepend a language code prefix as shown in Figure 26 (left). The mBART-50 model includes pre-trained language code embeddings for only 50 languages, but our MT system must be able to handle any source language. For example, Figure 26 (right) shows that 31 of the languages in the LwLL whitelist do not have a corresponding language code in mBART-50. For those languages, we borrow as a default the language code embedding from French.

mBART-50 Tokenizer		
[src-lang-code]	s_1 s_2 s_3...	[eos]
[trg-lang-code]	t_1 t_2 t_3...	[eos]

LwLL Whitelist Coverage	Yes	No
mBART-25	11	44
mBART-50	24	31

Figure 26. Tokenization with language codes (left) and whitelist coverage by model (right).

The main advantage of our two-stage pipeline is that it achieves strong transfer learning from both monolingual pre-training and translation task customization. For many language pairs, our model obtains good translation quality even in zero-shot settings.

3.3.1 MT System Evaluation and Analysis

Figure 27 shows BLEU scores for the three LwLL development tasks. We compare the performance of our Phase 2 (orange) and Phase 3 (gold) submissions against the JPL baseline (blue). Our P3 system significantly outperformed our P2 system on all three tasks. Our P2 system, a 55-languages-to-English sequence-to-sequence transformer trained using only the whitelisted parallel data and therefore without the advantage of large-scale monolingual pre-training, struggled on the `sin-eng` task because there were no useful related languages in the LwLL whitelist. Our P3 system obtains much better BLEU scores on this task because it has the advantage of pre-training on a large collection of monolingual data that includes Sinhala allowing it to learn useful representations and embeddings.

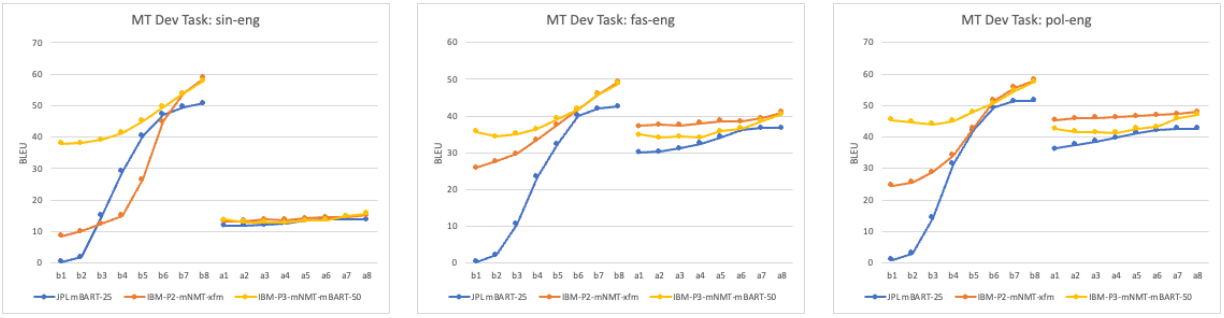


Figure 27. IBM MT progress: P3 vs. P2 vs. JPL baseline on LwLL development tasks.

Figure 28 shows that the early checkpoints of the base stage most clearly demonstrate the advantage of our pre-training + translation task customization approach. For example, our P3 system improves by an average of over 27 BLEU vs. our P2 system on the `sin-eng` task.

base	P2 vs. P3 mean ΔBLEU	
Task	Chkpts 1-4	Chkpts 5-8
sin-eng	+27.62	+5.58
fas-eng	+6.26	+0.32
pol-eng	+16.61	+0.63

Figure 28. Mean change in BLEU for early vs. later checkpoints on LwLL development tasks.

Figure 29 compares mBART-25 and mBART-50 against our P3 model customized for the translation task on the first four checkpoints of each development task. We first compare the performance of models without parallel data customization: mBART-25 and mBART-50. For `sin-eng`, which is covered by both mBART-25 and mBART-50, the models have similar performance. For `fas-eng` and `pol-eng`, which are included in mBART-50 but not mBART-25, there is a clear performance gap. Monolingual pre-training language coverage is thus an important factor that determines how effectively the model can learn to translate when presented with parallel data for a previously unseen language pair. Comparing the monolingually trained models against our P3 multilingual NMT pre-trained model (mBART-50-mNMT) shows that parallel data from other language pairs can also provide strong transfer learning capabilities. Our P3 model has much higher BLEU scores in the early checkpoints compared to standard off-the-shelf mBART pre-trained models.

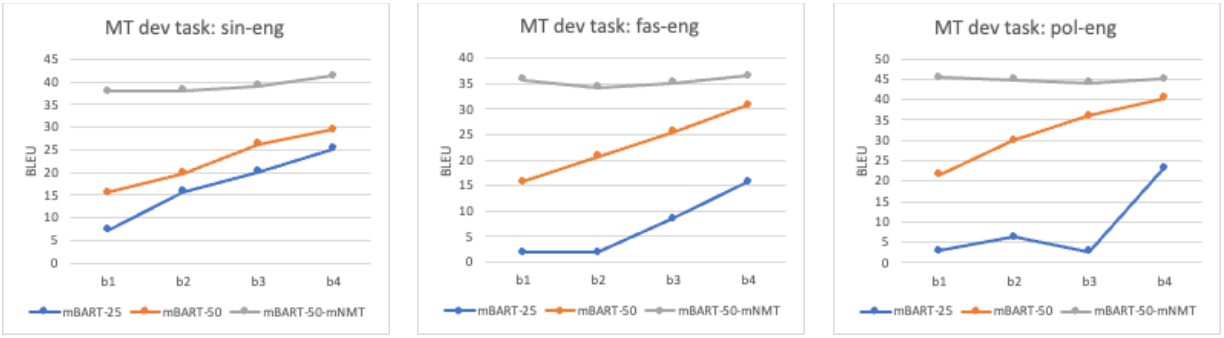


Figure 29. Pure monolingual pre-training vs. translation task customization.

Figure 30 shows the gains made by our P3 system over our P2 system on the official evaluation tasks. We see BLEU score improvements on all three tasks. Task 2, which was very difficult for our P2 system, shows the most improvement. Task 3 is still difficult for our model: although our P3 system did improve, we still lag the baseline in some checkpoints. In the PI meeting, the source language for Task 3 was revealed to be Hungarian. This explains the difficulty of this task for our model: Hungarian is not included in mBART-50 pre-training. Performance on this task could have

been improved by extending the model to cover additional languages, or by choosing a pre-trained model such as mT5 [31] that includes more languages in its pre-training.

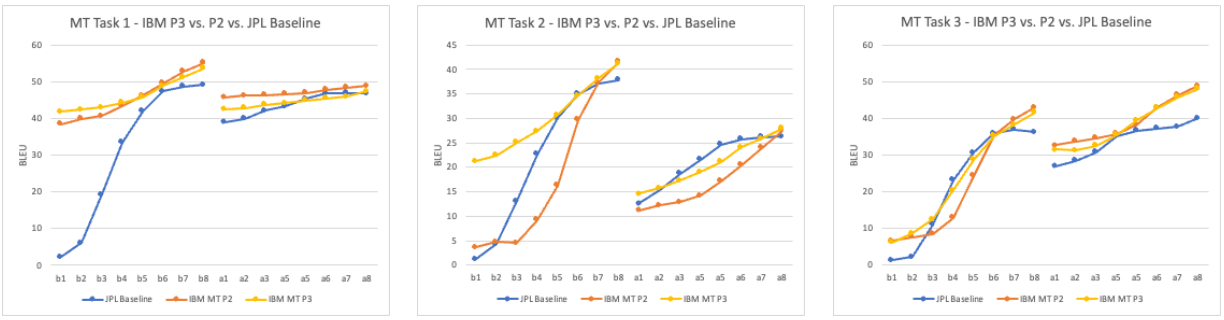


Figure 30. IBM MT progress: P3 vs. P2 vs. JPL baseline on LwLL evaluation tasks.

3.3.2 MT Zero-Shot Decoding Performance

Zero-shot decoding results for our model are shown in Figure 31. We evaluate zero-shot performance using the FLORES-dev [32] set to avoid any possible train/test bias in the LwLL development task data. We see that using off-the-shelf versions of mBART-25 and mBART-50 without translation task customization is ineffective: the zero-shot BLEU scores are close to zero for all three development tasks. These two models do not yet know how to translate. Our P2 55-languages-to-English transformer trained using only parallel data from the LwLL whitelist does better on the `fas-eng` and `pol-eng` tasks. However, it performs poorly on the `sin-eng` task. This is because Sinhala has an unusual script and there are no related languages in the LwLL whitelist that help via transfer learning. Our P3 model, that includes large-scale monolingual pre-training on all three of the development task languages, achieves good BLEU scores for a zero-shot setting on all three language pairs.

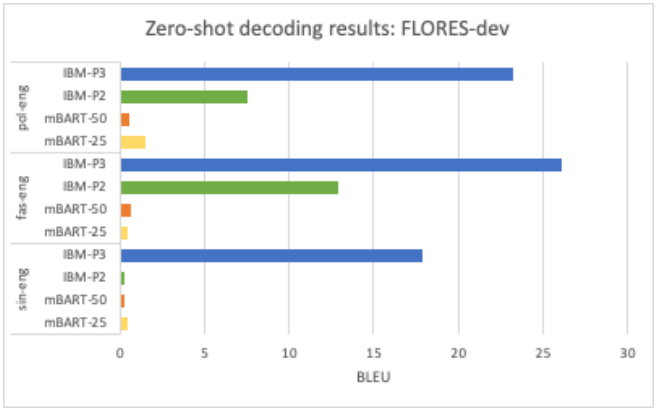


Figure 31. Zero-shot decoding on FLORES-dev for LwLL development task language pairs.

3.3.3 MT Implementation Details

For each checkpoint of a LwLL session, our P3 system is fine-tuned with target language translations returned though the API up to the specified checkpoint budget. All data is tokenized using the mBART `sentencepiece` model with a vocabulary size of 250k. We fine-tune the pre-trained model on a merged data set created by combining labels from the current checkpoint and all previous checkpoints for each stage. In the base stage, fine-tuning starts from the original

pre-trained model. In the adaptation stage, fine-tuning starts from the fine-tuned model obtained in the final checkpoint of the base stage. We do not prioritize label requests and simply sample labels uniformly at random from the training set. At the start of each checkpoint, we reset the optimizer and learning rate. We adjust the number of warm-up steps according to the total label budget, maximum number of epochs, and batch size. Ten percent of the labeled sentence pairs are reserved for validation data. Minimal length- and length-ratio-based filters are applied to the returned labels. All pre-training and fine-tuning is implemented using the transformers library from HuggingFace[1]. The model has 630m parameters and we optimized the training schedule, learning rate, and warmup ratio on the development tasks.

4. PUBLICATIONS, ACTIVITIES, AND DEMOS

4.1 Publications

Throughout the program, we have published our work in the most prestigious artificial intelligence conferences in the world (CVPR, ICCV, ECCV, NeurIPS, ICLR, ACL, AAAI, and more). Below we include selected publications per topic:

- Is a good embedding all you need? [1], [2]
- Pre-training and transfer from synthetic data: [4], [7], [8], [9], [10], [33]
- Representation Learning: [5], [11], [34], [35]
- Transfer Learning: [12], [19], [20], [21]
- Generative Data Augmentation: [36], [37], [38]
- Domain Adaptation: [39] (Best paper award, Honorable mention, WACV 2023)

[1] https://huggingface.co/docs/transformers/index

4.2 Tutorials

We organized a tutorial at ICCV 2019 on visual learning with limited labeled data. Our tutorial was very successful, and very well attended (see Figure below). The room was packed and overflowing, with a much larger number of attendees compared to other tutorials and workshops on that day.

Figure 32. Our team organized a tutorial on visual learning with limited labeled data at ICCV 2019.

4.3 Workshops

We organized the Visual Learning with Limited Labels (VL3) workshop at CVPR 2020. The workshop received a lot of attention, with 60 paper submissions. We accepted 8 papers as oral presentations, and 39 papers as posters. In addition, as part of this workshop, we had a challenge on cross-domain few-shot classification, with the participation of 9 teams. Our invited speakers included several performers of the DARPA LwLL program.

4.4 Demos for Transition Partners

We presented two demos of our systems during the PI meeting. The first demo covers our image classification and object detection systems that were submitted for the final evaluation of the program. The demos illustrated a use-case of satellite imagery coarse classification, followed by detailed object detection. The figures below show screenshots of the demo screen:

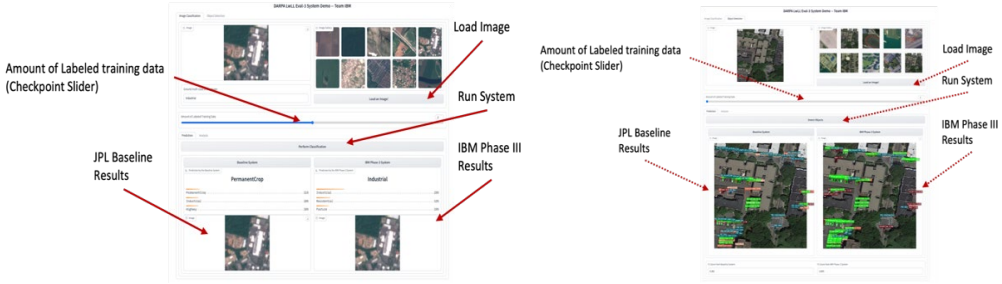


Figure 33. Screenshot from the image classification demo (left) and the object detection demo (right).

The second demo that we presented was our unsupervised cross-domain system that accompanies our paper [11]. This demo is openly available at the following location:

https://mitibmdemos.draco.res.ibm.com/cdr/.

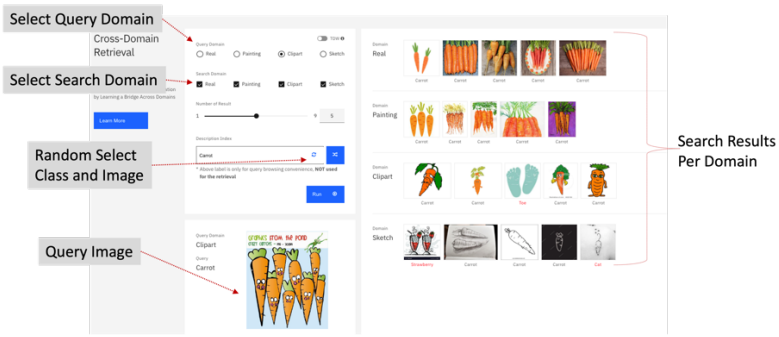


Figure 34. A screenshot from the "Bridge Across Domains" demo.

5. CONCLUSIONS AND RECOMMENDATIONS

Throughout the more than three years of the DARPA LwLL program, our team has made significant advances in the field of learning with less labels and related topics such as cross-domain and multi-modal learning. We have contributed over 20 published research papers to top ranking venues such as CVPR, NeurIPS, ICCV/ECCV, among others. We have also organized workshops and a tutorial promoting the field and the goals of the DARPA LwLL program in the scientific community. In addition, we achieved excellent results in the competitive DARPA LwLL evaluations, surpassing the baselines and (frequently) the results of all other performers.

More specifically, our team has contributed several novel techniques that seek to find a good embedding: representation learning; model architectures; and pre-trained model selection. Many of these techniques were integrated into our system submissions for the official DARPA LwLL evaluations. We have shown our contributions to generalize well to all three problem domains we participated in, including image classification, object detection, and machine translation, leading to strong and highly competitive results on our part.

Looking to the future of the LwLL field, although a lot of progress has been made, many interesting research questions remain open. As with most of the AI research community, our current belief is that future methods will mostly rely on (multi-modal) foundation models as the basis of all AI solutions, including those of open-vocabulary (natural language input) low-shot and zero-shot learning tasks that have been the focus of DARPA LwLL. We believe that language modeling as well as cross modal learning are key to these future solutions. We (and others in the scientific

community) have already made some interesting advances along this relatively new paradigm, and more exciting advances are just around the corner. We also believe that pre-training models on synthetic data is another very important direction, not only for addressing the problem of learning with less labels (and less data), but also to mitigate issues inherent with real-world datasets related to privacy, copyright, data protection, bias, and ethics.

REFERENCES

[1] Y. Tian, Y. Wang, D. Krishnan, J. Tenenbaum, and P. Isola. “Rethinking Few-Shot Image Classification: a Good Embedding Is All You Need?”, ECCV 2020.

[2] Y. Guo, N. Codella, L. Karlinsky, J. Codella, J. Smith, K. Saenko, T. Rosing, and R. Feris. “A Broader Study of Cross-Domain Few-Shot Learning”, ECCV 2020.

[3] Chen, Chun-Fu, Rameswar Panda, and Quanfu Fan. "Regionvit: Regional-to-local attention for vision transformers", ICLR 2022.

[4] S. Mishra, R. Panda, C. Phoo, C. Chen, L. Karlinsky, K. Saenko, V. Saligrama, and R. Feris. “Task2Sim: Towards Effective Pre-training and Transfer from Synthetic Data”, CVPR 2022.

[5] A. Islam. C. Chen, R. Panda, L. Karlinsky, R. Feris, and R. Radke. “Dynamic Distillation Network for Cross-Domain Few-Shot Recognition with Unlabeled Data”, NeurIPS 2021.

[6] M. Lichtenstein, P. Sattigeri, R. Feris, R. Giryes, L. Karlinsky. “Tafssl: Task-adaptive feature sub-space learning for few-shot classification”, ECCV 2020.

[7] Y. Kim, S. Mishra, S. Jin, R. Panda, H. Kuehne, L. Karlinsky, V. Saligrama, K. Saenko, A. Oliva, and R. Feris, R. “How Transferable are Video Representations Based on Synthetic Data?”, NeurIPS 2022.

[8] P. Cascante-Bonilla, H. Wu, L. Wang, R. Feris, and V. Ordonez. “SimVQA: Exploring Simulated Environments for Visual Question Answering”, CVPR 2022.

[9] M. Baradad, R. Chen, J. Wulff, T. Wang, R. Feris, A. Torralba, and P. Isola “Procedural Image Programs for Representation Learning”, NeurIPS 2022.

[10] Z. He, G. Blackwood, R. Panda, J. McAuley and R. Feris, "Synthetic Pre-Training Tasks for Neural Machine Translation", ACL 2023.

[11] S. Harary, E. Schwartz, A. Arbelle, P. Staar, S. Abu-Hussein, E. Amrani, R. Herzig, A. Alfassy, R. Giryes and H. Kuehne, "Unsupervised domain generalization by learning a bridge across domains", CVPR 2022.

[12] A. Islam, R. Chen, R. Panda, L. Karlinsky, R. Radke, and R. Feris. “A Broad Study on the Transferability of Visual Representations with Contrastive Learning”, ICCV 2021.

[13] W. Lin, L. Karlinsky, N. Shvetsova, H. Possegger, M. Kozinski, R. Panda, R. Feris, H. Kuehne, H. Bischof. “MAtch, eXpand and Improve: Unsupervised Finetuning for Zero-Shot Action Recognition with Language Knowledge”, ArXiv 2023.

[14] A. Alfassy, A. Arbelle, O. Halimi, S. Harary, R. Herzig, E. Schwartz, R. Panda, M. Dolfi, C. Auer, P. Staar, K. Saenko, R. Feris, L. Karlinsky. “FETA: Towards Specializing Foundational Models for Expert Task Applications”, NeurIPS 2022.

[15] J.S. Smith, P. Cascante-Bonilla, A. Arbelle, D. Kim, R. Panda, D. Cox, D. Yang, Z. Kira, R. Feris, L. Karlinsky “ConStruct-VL: Data-Free Continual Structured VL Concepts Learning”, CVPR 2023.

[16] S. Doveh et al. “Teaching Structured Vision&Language Concepts to Vision&Language Models”, CVPR 2023.

[17] P. Cascante-Bonilla et al, “Going Beyond Nouns With Vision & Language Models Using Synthetic Data”, arXiv:2303.17590.

[18] Arbelle, S. Doveh, A. Alfassy, J. Shtok, G. Lev, E. Schwartz, H. Kuehne, H. Levi, P. Sattigeri, R. Panda, R. Chen, A. Bronstein, K. Saenko, S. Ullman, R. Giryes, R. Feris, and L. Karlinsky. “Detector-Free Weakly Supervised Grounding by Separation”, ICCV 2021.

[19] Y. Guo, H. Shi, A. Kumar, K. Grauman, T. Rosing, and R. Feris. “SpotTune: Transfer Learning through Adaptive Fine-tuning”, CVPR 2019.

[20] X. Sun, R. Panda, R. Feris, and K. Saenko. “AdaShare: Learning What to Share for Efficient Deep Multi-Task Learning”, NeurIPS 2020.

[21] B. Bhattacharjee, J. Kender, M. Hill, P. Dube, S. Huo, M. Glass, B. Belgodere, S. Pankanti, N. Codella, P. Watson. “P2L: Predicting Transfer Learning for Images and Semantic Relations”, arXiv:1908.07630, 2019.

[22] Tan, Mingxing, and Quoc Le. "Efficientnetv2: Smaller models and faster training", ICML 2021.

[23] Liu, Ze, et al. "Swin transformer v2: Scaling up capacity and resolution", CVPR 2022.

[24] Liu, Zhuang, et al. "A convnet for the 2020s", CVPR 2022.

[25] Caron, Mathilde, et al. "Emerging properties in self-supervised vision transformers", ICCV 2021.

[26] Yen-Cheng Liu, Chih-Yao Ma, Zijian He, Chia-Wen Kuo, Kan Chen, Peizhao Zhang, Bichen Wu, Zsolt Kira, Peter Vajda, “Unbiased Teacher for Semi-Supervised Object Detection”, ICLR 2021.

[27] Q. Zhou and C. Yu and Z. Wang and Q. Qian and H. Li, “Instant-Teaching: An End-to-End Semi-Supervised Object Detection Framework”, CVPR 2021.

[28] Xu, Mengde, Zheng Zhang, Han Hu, Jianfeng Wang, Lijuan Wang, Fangyun Wei, Xiang Bai, and Zicheng Liu. "End-to-end semi-supervised object detection with soft teacher", ICCV 2021.

[29] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov and L. Zettlemoyer, "BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension", ACL 2020.

[30] Y. Liu, J. Gu, N. Goyal, X. Li, S. Edunov, M. Ghazvininejad, M. Lewis and L. Zettlemoyer, "Multilingual Denoising Pre-training for Neural Machine Translation," in *Transactions of the Association for Computational Linguistics*, 2020.

[31] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua and C. Raffel, "mT5: A massively multilingual pre-trained text-to-text transformer," ACL 2021.

[32] N. Goyal, C. Gao, V. Chaudhary, P.-J. Chen, G. Wenzek, D. Ju, S. Krishnan, M. Ranzato, F. Guzmán and A. Fan, "The FLORES-101 Evaluation Benchmark for Low-Resource and Multilingual Machine Translation," in *Transactions of the Association for Computational Linguistics*, 2022.

[33] Y. Li, R. Panda, Y. Kim, C. Chen, R. Feris, D. Cox, and N. Vasconcelos. “VALHALLA: Visual Hallucination for Machine Translation”, CVPR 2022.

[34] G. Bukchin, E. Schwartz, K. Saenko, O. Shahar, R. Feris, R. Giryes, and L. Karlinsky. “Fine-grained Angular Contrastive Learning with Coarse Labels”, CVPR 2021.

[35] A. Singh, O. Chakraborty, A. Varshney, R. Panda, R. Feris, K. Saenko, and A. Das. “Semi-Supervised Action Recognition with Temporal Contrastive Learning”, CVPR 2021.

[36] Z. Tang, Y. Gao, P. Sattigeri, L. Karlinsky, R. Feris, and D. Metaxas. “OnlineAugment: Online Data Augmentation with Less Domain Knowledge”, ECCV 2020.

[37] A. Sahoo, A. Singh, R. Panda, R. Feris, and A. Das. “Mitigating Dataset Imbalance via Joint Generation and Classification”, ECCV Workshop on Imbalance Problems in Computer Vision, 2020.

[38] A. Alfassy, L. Karlinsky, A. Aides, J. Shtok, S. Harary, R. Feris, R. Giryes, and A. Bronstein. “LaSO: Label-Set Operations Networks for Multi-label Few-shot Learning”, CVPR 2019.

[39] A. Sahoo, R. Panda, R. Feris, K., and A. Das. “Select, Label, and Mix: Learning Discriminative Invariant Feature Representations for Partial Domain Adaptation”, WACV 2023, Best Paper Award Honorable Mention.

[40] T. Kudo and J. Richardson, "SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing", EMNLP 2018.

LIST OF SYMBOLS, ABBREVIATIONS, AND ACRONYMS

AI	Artificial Intelligence
BrAD	Bridge Across Domains
BSCD-FSL	Broader Study of Cross-domain Few-shot Learning
CNN	Convolutional Neural Network
CVPR	Computer Vision and Pattern Recognition Conference
DARPA	Defense Advanced Research Projects Agency
DINO	DIstillation with NO labels
ECCV	European Conference on Computer Vision
FETA	Foundation Models for Expert Task Applications
FM	Foundation Model
ICCV	International Conference on Computer Vision
ICLR	International Conference on Learning Representations
JPL	Jet Propulsion Laboratory
LwLL	Learning with Less Labels
NeurIPS	Neural Information Processing Systems Conference
NMT	Neural Machine Translation
SSL	Semi-Supervised Learning
SVLC	Structured Vision & Language Concepts
UBT	Unbiased Teacher
UDG	Unsupervised domain Generalization
ViT	Vision Transformer
VL	Vision and Language

IBM TEAM

Team members (in alphabetical order): Amit Alfassy, Assaf Arbelle, Bishwaranjan(Bhatta) Bhattacharjee, Graeme Blackwood, Richard Chen, Sivan Doveh, Rogerio Feris, Roei Herzig, Leonid Karlinsky, Hilde Kühne, Rameswar Panda, Aadarsh Sahoo, Eli Schwartz, Joseph Shtok