



AFRL-RI-RS-TR-2023-198

LEARNING TO LEARN AND ADAPT WITH LESS LABELS

MONASH UNIVERSITY

NOVEMBER 2023

FINAL TECHNICAL REPORT

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED

STINFO COPY

**AIR FORCE RESEARCH LABORATORY
INFORMATION DIRECTORATE**

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-2023-198 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 /

MICHELE R. GRIECO
signed for: 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	
NOVEMBER 2023		FINAL TECHNICAL REPORT		START DATE AUGUST 2019	END DATE MAY 2023
4. TITLE AND SUBTITLE LEARNING TO LEARN AND ADAPT WITH LESS LABELS					
5a. CONTRACT NUMBER FA8750-19-2-0501		5b. GRANT NUMBER N/A		5c. PROGRAM ELEMENT NUMBER 61101E	
5d. PROJECT NUMBER		5e. TASK NUMBER		5f. WORK UNIT NUMBER R2VS	
6. AUTHOR(S) Gholamreza Haffari, Ehsan Abbasnejad, Mehrtash Harandi, Islam Nassar, Mahsa Ghorbanali, Chengyao Qian, Samitha Herath, Hamid Rezaatofighi, Trang Vu					
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Monash University Wellington Rd, Clayton VIC 3800 Australia				8. PERFORMING ORGANIZATION REPORT NUMBER Adelaide University Adelaide, South Australia 5005 Australia	
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) Air Force Research Laboratory/RISB 525 Brooks Road Rome NY 13441-4505			10. SPONSOR/MONITOR'S ACRONYM(S) DARPA/I2O 675 N Randolph St Arlington VA 22203-2114 AFRL/RI & DARPA		11. SPONSOR/MONITOR'S REPORT NUMBER(S) AFRL-RI-RS-TR-2023-198
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 This document summarizes our research and technical contributions in the DARPA LwLL (Learning with Less Labels) program to reduce labeled data required in training and adapting machine learning models. Based on transfer learning, our approach leverages inductive biases acquired from various sources, including datasets, knowledge resources, models and architectures, existing tasks, and training algorithms. It integrates the aforementioned approaches into a coherent framework, aligned with the LwLL evaluation protocol, to learn in various learning conditions with limited supervision. Capitalizing on the recent advances in generative AI, this work studies novel and effective ways of injecting proper inductive biases to push the state of the art in image classification, object detection, video classification, and machine translation. The outcomes include extensive publications in top AI venues, workshop organization, and the training of postdocs and PhD students. The team has also excelled in LwLL evaluations, with the object detection system selected for NGA transition task. Furthermore, the knowledge and skills gained from the LwLL program, have been applied to our team's participation in the DARPA CCU program, where only 20% of the data is labeled.					
15. SUBJECT TERMS Semi-supervised learning, meta-learning, active learning, domain adaptation, machine translation, object detection, video classification, image classification.					
16. SECURITY CLASSIFICATION OF:				17. LIMITATION OF ABSTRACT	
a. REPORT U	b. ABSTRACT U	c. THIS PAGE U	SAR		18. NUMBER OF PAGES 90
19a. NAME OF RESPONSIBLE PERSON PETER JEDRYSIK					19b. PHONE NUMBER (Include area code) N/A

TABLE OF CONTENTS

LIST OF FIGURES.....	III
LIST OF TABLES	IV
1 EXECUTIVE SUMMARY.....	1
2 INTRODUCTION	2
2.1 LWLL OVERVIEW	3
2.1.1. Terminology.....	3
2.1.2. Evaluation Protocol.....	4
2.2 LALA OVERVIEW	5
2.2.1. Inductive Biases in the Vision Systems.....	5
2.2.2. Inductive Biases in the Machine Translation Systems.....	6
2.2.3. Final Overview Presentation	8
3 METHODS, ASSUMPTIONS AND PROCEDURES.....	9
3.1 IMAGE CLASSIFICATION	9
3.1.1. Overview	9
3.1.2. Whitelist Pretraining	10
3.1.3. Model selection at evaluation time.....	11
3.1.4. Semi-supervised Transfer Learning	13
3.1.5. Active Learning.....	16
3.2 VIDEO CLASSIFICATION.....	17
3.2.1. Overview	17
3.2.2. Model Architecture.....	17
3.2.3. Pre-evaluation Preparation.....	18
3.2.4. System Evaluation	19
3.3 OBJECT DETECTION	21
3.3.1. Overview	21
3.3.2. Components	22
3.4 MACHINE TRANSLATION.....	25
3.4.1. Overview	25
3.4.2. Model Architecture.....	26
3.4.3. Pre-evaluation Preparation.....	29
3.4.4. System Evaluation	31
3.5 SUMMARY	35
4. RESULTS AND DISCUSSION	38
4.1 IMAGE CLASSIFICATION.....	38
4.1.1. Development Results.....	38
4.1.2. Evaluation Results.....	39
4.1.3. Discussion.....	39
4.2 VIDEO CLASSIFICATION	41
4.2.1. Development Results.....	41
4.2.2. Evaluation Results	42
4.3 OBJECT DETECTION	44
4.3.1. Development Results.....	44
4.3.2. Evaluation Results.....	48
4.4 MACHINE TRANSLATION.....	49
4.4.1. Development Results.....	49
4.4.2. Evaluation Results	50

5	CONCLUSION	52
5.1	SUMMARY OF CONTRIBUTIONS	52
5.2	FUTURE DIRECTIONS	53
5.3	ACKNOWLEDGMENTS	54
6	REFERENCES	55
APPENDIX A	LIST OF PAPERS ACCORDING TO THE AREAS	60
A.1	VISION	60
A.1.1	<i>Few Shot Learning</i>	60
A.1.2	<i>Semi-supervised Learning</i>	60
A.1.3	<i>Active Learning</i>	61
A.1.4	<i>Domain Adaptation</i>	61
A.1.5	<i>Others</i>	61
A.2	LANGUAGE	62
A.2.1	<i>Machine Translation/NLP tasks</i>	62
A.2.2	<i>Transfer Learning: Domain Adaptation</i>	63
A.2.3	<i>Semi-supervised Learning</i>	64
A.2.4	<i>Active Learning</i>	65
A.2.5	<i>Multi-task Learning</i>	65
A.2.6	<i>Other</i>	65
APPENDIX B	LALA TEAM'S FINAL OVERVIEW PRESENTATION	67
	LIST OF SYMBOLS, ABBREVIATIONS AND ACRONYMS	83

LIST OF FIGURES

Figure 1 - Learning Problem in LwLL Protocol.	4
Figure 2 - Publication grouped by LwLL Problem Setting and Inductive Bias. Here, the x-axis shows the sources of inductive biases, including related tasks, human, and related data. The y-axis shows the machine learning setting, including supervised learning, domain adaptation, semi-supervised learning, active learning, few-shot learning and unsupervised learning.	7
Figure 3 - A conceptual diagram of SemCo [5].	13
Figure 4 - LAVA Method Overview [4].	15
Figure 5 - Class & Domain collapse. We display two query images from clipart target (left), and their 4 nearest neighbors, from clipart and ImageNet, in LAVA's representation space (middle) and an ImageNet-supervised learner's (right) before training on target instances. ImageNet instances are bordered in blue.	15
Figure 6 - A overview of our architecture [6].	18
Figure 7 - Pretrained dataset comparison.	19
Figure 8 - Few-labeled action recognition approach [6].	20
Figure 9 - Activate learning method [7].	21
Figure 10 – Samples from various domains.	22
Figure 11 — Number of images and classes in the Development Dataset.	23
Figure 12 – Number of Images in the Evaluation Dataset.	23
Figure 13 – Number of Classes in the Evaluation Dataset.	24
Figure 14 – Overview of Unbiased Teacher [23].	25
Figure 15 - Adapter architecture and its added position in a Transformer layer.	27
Figure 16 - Hyper adapter architecture [9].	28
Figure 17 - Number of Mined Sentences by Source Language before and after filtering.	30
Figure 18 - Development results.	38
Figure 19 - Our performance in Phase 1 of the program against the baseline and the best performer.	40
Figure 20 - Our performance in Phase 2 and 3 of the program against the baseline and the best performer.	41
Figure 21 - Performance on HMDB and UCF101.	42
Figure 22 - JPL evaluation results on base dataset and adaptation dataset.	43
Figure 23 – Effect of pretrained model on two Domains as example.	44
Figure 24 – Comparison between different backbone models.	45
Figure 25 – Supervised vs Semi-supervised Learning Methods.	46
Figure 26 – Brief description of different systems.	47
Figure 27 – Comparative analysis of the object detection system.	48
Figure 28 - JPL Evaluation Results.	49
Figure 29 - Our development result on development task 3 (Sin-Eng).	50
Figure 30 - Performance of our NMT system in the LwLL evaluation 3.	51

LIST OF TABLES

Table 1 - Our Semantic Grouping of the Whitelist Datasets.....	10
Table 2 - A subset of the dataset used to train our model selection decision tree.....	12
Table 3 – Datasets of Different Domains.....	23
Table 4 - The donor language for each source language.	32
Table 5 - NMT evaluation task description.	50

1 EXECUTIVE SUMMARY

DARPA's Learning with Less Labels (LwLL) program aims to make the process of training machine learning models more efficient by reducing the amount of labeled data required to train models and adapt models to new environments. The LALA (Learning and Adapt with Less Labels) team consists of personnel over two institutions, Monash University (the Primary Contractor) and Adelaide University, as the subcontractor.

We provide a unified treatment of learning from less labels which is based on transfer learning. It aims to learn inductive biases in order to learn new tasks more effectively. In our work, these inductive biases are injected/transferred through datasets, knowledge resources, models and architectures, existing tasks, and training algorithms. Our approaches are integrated into a coherent framework, aligned with the LwLL evaluation protocol, to learn in various learning conditions with limited supervision that an agent may encounter in the real world. We note that learning for fully autonomous agents in various learning conditions is under-explored in the research community. The agents considered in our work continually encounter new problems and tasks, while operating in an environment with dynamic change, in a manner that may not be foreseen by the agents' designer.

We have witnessed significant progress in the research community towards learning with less labels throughout the LwLL program. Shortly after the program started in 2019, large-scale pre-training of large neural networks proved to be an effective approach to boost state-of-the-art in many vision and language benchmarks and reduce the labeled data requirements for various real-life machine learning tasks. The other turning point has been the introduction of instruction-following large language models (LLMs) in late 2022 towards the end of the LwLL program, which has led to significant improvement in learning with scarce labeled data. Capitalizing on these advances, in our work, we study novel and effective ways of injection of proper inductive biases to push the state of the art in image classification, object detection, video classification, and machine translation.

Our team's outcomes can be summarized as follows. We have extensively published the results of our work in top and leading AI venues, including (i) Language Papers: 10 EMNLP, 4 ACL, 4 COLING, 2 EACL, 1 SIGIR, 1 INTERSPEECH, 1 NAACL, and (ii) Vision Papers: 16 CVPR, 4 NeurIPS, 4 ICCV, 2 ECCV, 4 WACV, 3 TPAMI, 2 ICLR, 1 IJCAI, 1 ICML, 1 ACCV. We further co-organised the workshop "Deep Learning for Low Resource NLP (DeepLo)" co-located with NACCL 2022. Our participation in LwLL has led to the training of 4 Postdocs and 8 PhD students. We have submitted strong systems in all evaluations throughout the LwLL program, where our object detection system was selected for the NGA transition task. We further transferred the knowledge and skills from this program to our team performing in the DARPA CCU program, where only 20% of the data is labeled and the rest is unlabeled.

2 INTRODUCTION

Humans, and many animals, are able to learn a new concept by only a handful of examples of that concept, which allows rapid adaptation to a new environment. State-of-the-art deep neural networks, however, used to often require a lot of labeled data to learn the target concept of interest, mainly due to the high learning capacity of these models. Nonetheless, recent advances in Generative AI and instruction-following large language models (LLMs) have opened up new possibilities for rapid learning and adaptation of machine learning systems when labeled data is scarce. **In this project, our approach is to learn how to effectively learn and adapt with less labels by transferring knowledge from some tasks to others, aka transfer learning.**

A desiderata for an approach for learning from less labeled data is that it should not be specific to a particular task formulated for the learning of a concept. Roughly speaking, a task consists of a dataset and an evaluation metric for learning a specific concept, e.g., the ImageNet challenge. Ad-hoc and heuristic-based approaches for learning from less labels are undesirable as they cannot consistently work well across varying data and task distributions. **In this project, we design effective strategies to learn and/or adapt to new concepts by exploiting inductive biases required for generalization.** These inductive biases are “transferred” to the target tasks of interest using different mechanisms under the broad umbrella of transfer learning. We study the generalization of a broad range of deep neural networks, including state-of-the-art models in image classification, object detection, video classification, and machine translation.

As mentioned, we achieve sample efficiency in learning with less labels (LwLL) through the injection of **inductive biases**. It is worth clarifying the concept of “inductive bias” in learning machines, as used in this report. “Learning involves the ability to generalize from past experience in order to deal with new situations that are “related to” this experience. The *inductive leap* needed to deal with new situations seems to be possible only under certain *biases* for choosing one generalization of the situation over another” [1]. Our working definition of *bias* in this report is “to refer to any basis for choosing one *generalization [or hypothesis]* over another, other than strict consistency with the observed training instances” [1].

Our use of “transfer learning” is broad and general, and includes recent advances such as (i) pretrain-then-finetune paradigm, and (ii) zero/few shot learning using instruction-following large language models (LLMs). The former consists of distilling knowledge and inductive biases through the use of self-supervised learning with massive amounts of unlabeled naturally-occurring web-scale data with diverse distributions. One may ask, fundamentally, how the transfer of such inductive biases occur and how they lead to more effective and efficient learning with less labels. We provide an intuition based on the compression argument [2]. The “compression” of large amount of data, using a compression algorithm found with gradient-based search in a space of functions defined by a big neural network, leads to capturing regularities that can be then exploited in a target supervised learning task of interest, e.g., through informative and predictive automatically crafted features.¹

¹ These ideas are formalized in “No Free Lunch Theorem, Kolmogorov Complexity, and the Role of Inductive Biases in Machine Learning” (Goldblum et al., 2023).

The latter relies on “mobilizing” the distilled knowledge and inductive biases learned in a large neural network in the pre-training phase, for the benefit of the target task of interest. “Instruction-following” is achieved by exposing a large pre-trained LLM to a diverse set of instructions so that the network can understand and follow novel human instructions, enabling access to the massive knowledge distilled in the network. These instructions can, for example, be a description of how labels for a classification task should be produced, similar to the instructions given to human annotators when curating a labeled dataset. This indeed turns the LLM to a “noisy annotator”, or in other words, a classifier in the “zero-shot” manner.

Capitalizing on these advances, throughout our work in DARPA LwLL, we have studied and explored further injection of proper inductive biases, in a variety of learning settings, to push the state of the art in image classification, object detection, video classification, and machine translation. We aim to ‘automatically specialize’ generic pre-trained models for a learning setting of interest that a generic intelligent agent encounters in the deployment phase. This requires a holistic treatment of various challenges and learning conditions, including adoption of appropriate backbone pre-trained architectures (model selection), identifying the knowledge gap of the agent and enquiring the label of proper unlabeled data points from human annotators (active learning), algorithms for learning from a combination of the labeled and unlabeled data (semi-supervised learning), and adaptation to a new domain where the data distribution changes (domain adaptation). We emphasize that the research community typically focuses on ‘one’ of the aforementioned machine learning settings in a published work. However, an autonomous intelligent agent ought to be able to handle all of these settings as they arise during the deployment of the system, and this is the core problem addressed in our work, aligning with the evaluation protocol of the LwLL program. Furthermore, our consistent usage of the Transformer architecture [3] in the vision and language problems provides a unified treatment of these different modalities, enabling the seamless exploration of various ideas.

This report is organized as follows. We start by describing the evaluation protocol in the DARPA LwLL program, and describe our main framework to tackle the challenges involved in learning with less labels. This is then followed by a methodology section which describing the adaptation of our framework to the four problems considered in the LwLL evaluation: image classification, object detection, video classification, and machine translation. We then report our evaluation results and findings in the Result and Discussion section. Finally, we conclude the report by the summary of our contributions and thoughts about the future of learning with less labels.

2.1 LwLL Overview

2.1.1. Terminology

- **Whitelist datasets:** A collection of labeled datasets that we can use to supervisedly (pre)train our models. We are allowed access to the pre-downloaded White-list datasets during the evaluation. Any pretrained weights that are publicly available on the Whitelist data are fair to be used.
- **Development datasets:** A collection of datasets that we can use for developing our algorithms. These datasets are integrated to the **interactive API**. As such, the

performers are able to emulate the evaluation process on these development datasets. JPL is open for receiving suggestions and help for broadening these development datasets.

- **Evaluation dataset:** The collection of datasets that our system submission will be evaluated. These are unseen datasets. However, meta-data of evaluation dataset such as number of classes, size will be provided to the performers.
- **Checkpoint:** A point in the evaluation where the performers submit the predictions for the test data.
- **Stage:** The two stages of the evaluation Base and Adaptation.

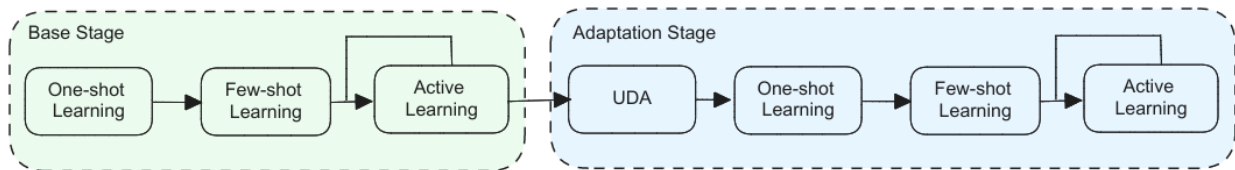


Figure 1 - Learning Problem in LwLL Protocol.

2.1.2. Evaluation Protocol

An evaluation task consists of two stages, Base and Adaptation, with two different datasets Base-Dataset and Adaptation-Dataset. At the beginning of the evaluation, we are provided with meta-data (e.g., number of channels of images, train set size, labeling budget for checkpoints, class names as strings) of the datasets. The dataset for each evaluation stage (e.g., Base and Adaptation) has its own test set to which we have access during training. However, transductive learning, i.e using a test set for training, is not allowed. Depending on the task size (small, medium or large), there are constraints in terms of GPUs and time resources, up to 8 GPUs and 7 days.

The evaluation process starts with the Base stage. We are given the unlabeled training sets for both the Base-Dataset and the Adaptation-Dataset. During evaluation, we will query labels for unlabeled data up to a predefined labeling budget via an interactive API. The model's performance is evaluated on a sequence of checkpoints when the labeling budget of that checkpoint is exhausted, 8 checkpoints for each stage.

While each system may differ on the detailed implementation of the labeling annotation, checkpoints are generally categorized in terms of the learning problem as follows (see Figure 1):

- **One-shot checkpoint**, i.e., checkpoint 1, where we are provided labels for one data point per class.
- **Few-shot checkpoints** where we are provided an equal number of labeled examples from each class. Hence, these checkpoints are equivalent to the class-balanced few-shot learning problem.
- **Active-learning checkpoints** where we are provided labels for unlabeled instances we select. These checkpoints are the last few checkpoints of a stage. Each checkpoints will have a predefined labeling budget increasing on a log scale. The final checkpoint (checkpoint 8) will have enough budget to cover the full

labeled data. At a particular checkpoint, queries do not need to be done at once. In general, there is no restriction on the query size and training strategy. When the labeling budget is exhausted, the prediction needs to be submitted and the evaluation will be moved to the next checkpoint.

- **Unsupervised domain adaptation (UDA)** checkpoint - only present in the Adaptation stage. This is the checkpoint before the Seed-Label checkpoint in the Adaptation Stage. There is no assumption on the class overlap between Base-Dataset and Adaptation-Dataset. It can be complete (closed-set) UDA where some base classes may be missing in adapt, open-set UDA where some adaptation classes are missing in the base classes, or a combination of partial-open.

Prior to the evaluation, the performers are provided with a set of whitelist datasets and pretrained models for preparing the system submission. Beyond the constraints on labeled dataset and pretrained models, generally, there are no limits in terms of unlabeled data and computational resources during preparation.

2.2 LALA Overview

In what follows, we look into the inductive biases used in our vision and language systems throughout the LwLL program. We then provide a categorization of our LwLL publications and work through the lens of inductive biases.

2.2.1. Inductive Biases in the Vision Systems

We demonstrate the importance of inductive bias in image classification, video classification, and object detection. By leveraging pre-training models, selecting appropriate backbones/detectors, and incorporating visual semantic prior, we inject proper inductive biases via transfer learning. Furthermore, by utilizing semi-supervised and active learning techniques, we aim to improve the performance and generalization capabilities of these systems using unlabeled data and human supervision.

The inductive biases in self-supervised pre-training include the human-provided objectives for learning that lead to more general representations not tied to predefined classes or domains, improving generalization [4]. Self-supervised pre-training also addresses the limitations of supervised pre-training, such as representation collapse when transferred to a target dataset with class or domain shift, and the semantic-free nature of class labels used in supervised pre-training. Additionally, self-supervised learning yields more adaptable and generalizable representations compared to supervised learning. The prior obtained from knowledge graphs and language embeddings encourages improved training process and visual similarities learned. Pretraining multiple base models on semantically grouped whitelist datasets also accounts for possible tasks encountered during evaluation. Our models are pretrained on whitelist datasets and then fine-tuned on partially labeled datasets in subsequent checkpoints.

Inductive biases in semi-supervised learning refer to the prior knowledge or assumptions that are incorporated into the learning algorithm to guide the model's behavior when dealing with unlabeled data. These biases help the model make informed decisions and generalize from limited labeled data to unlabeled data [5]. In our work, inductive biases are used in various ways. For example, in the video classification system,

we use the priors obtained from a pre-trained language model to capture and then encourage semantic similarities into video representations. These biases help the model learn, requiring less data to improve its performance [6]. Additionally, in the object detection system, the Teacher-Student Mutual Learning stage involves the Teacher generating pseudo-labels to train the Student, incorporating the bias that the Teacher's knowledge is valuable for the Student's learning. These inductive biases contribute to the overall performance and effectiveness of the semi-supervised learning algorithms.

Inductive biases in active learning refer to the assumptions or prior knowledge that are built into the learning algorithm to guide the selection of informative instances for labeling. These biases help the algorithm make decisions about which instances to query for labels in order to improve the model's performance. In the context of active learning, inductive biases can include assumptions about the distribution of the data, the complexity of the model, or the relevance of certain features [7]. These biases are important because they influence the effectiveness of the active learning process in selecting informative instances and reducing the need for large amounts of labeled data.

Lastly, our investigation delved into the inductive biases within the model architectures. As anticipated, working with smaller datasets typically demands a greater need for prior biases to effectively train a model. For instance, a convolutional neural network incorporates priors concerning local pixel properties in an image, which a transformer architecture designed for vision lacks. Interestingly, our findings reveal that, even with this distinction, larger datasets coupled with pre-training empower transformers to acquire more effective priors from data [6]. This notably translates into improved outcomes. Particularly, employing self-supervised pre-training with transformers demonstrates enhanced performance in vision tasks, especially when data availability is limited.

2.2.2. Inductive Biases in the Machine Translation Systems

Our approaches to machine translation involve leveraging inductive bias to address the challenges of multilinguality, efficiency, and learning from noisy data. The models are designed to handle a wide range of languages and adapt to their specific characteristics. Efforts are made to optimize the training process and improve efficiency through data cleaning, mining parallel data, and language transferability analysis. Additionally, we incorporate linguistic features and performance metrics to guide the selection of languages for transfer learning, enhancing the quality of the machine translation models.

To tackle the challenge of multilinguality, we employ cross-lingual transfer to unseen languages to improve performance in low-resource languages. We note that unseen languages may vary in available data, scripts, and other linguistic properties. This suggests that the models are designed to handle a wide range of languages and adapt to their specific characteristics. Upon encountering a new domain with only source language data, we collect target language data using an advanced data filtering based on cross-lingual cross-domain transfer [8]. We further utilize other techniques such as domain adaptation with hypernetworks [9] and active domain adaptation.

Efficiency is another important aspect that we consider in machine translation. We aim to efficiently learn a good neural machine translation (NMT) model on large scale

data within time and resource constraints. To achieve this, we expand the state-of-the-art pretrained sequence-to-sequence model mBART [10] to cover all languages of interest. We utilize adapters and hyper-adapters for knowledge sharing among the related languages, while minimizing the memory footprints of our models. We also perform data cleaning and mining parallel data to improve the quality of the training data. These strategies help optimize the training process and improve efficiency.

We conduct language transferability analysis to identify top donors for each recipient language, reducing negative interference. We employ a ranker with BLEU scores [11] and multiple data and linguistic features to train a model that predicts the language transferability between donor and recipient languages. This allows them to complete the missing cells in the ranking matrix and select the most effectively transferable languages as the final donor languages. By incorporating linguistic features and performance metrics, we introduce a form of bias that guides the selection of languages for transfer learning, improving the overall quality of the machine translation models.

Different AL algorithms perform well at different stages during model training. We meta-learn the AL acquisition policy from the related parallel data by imitation learning [12], [13] to effectively keep human-in-the-loop for annotation. Our approach considers various linguistic and statistical factors, such as lexical coverage and token entropy, to score potential data points. Furthermore, the effectiveness of AL can be hindered by the overhead of pseudo-label computation and acquisition score estimation for a large pool of unlabeled data. We employ length-based curriculum learning to subsample unlabeled candidates. These AL strategies result in an efficient and effective training of our NMT systems in the low-resource checkpoints.

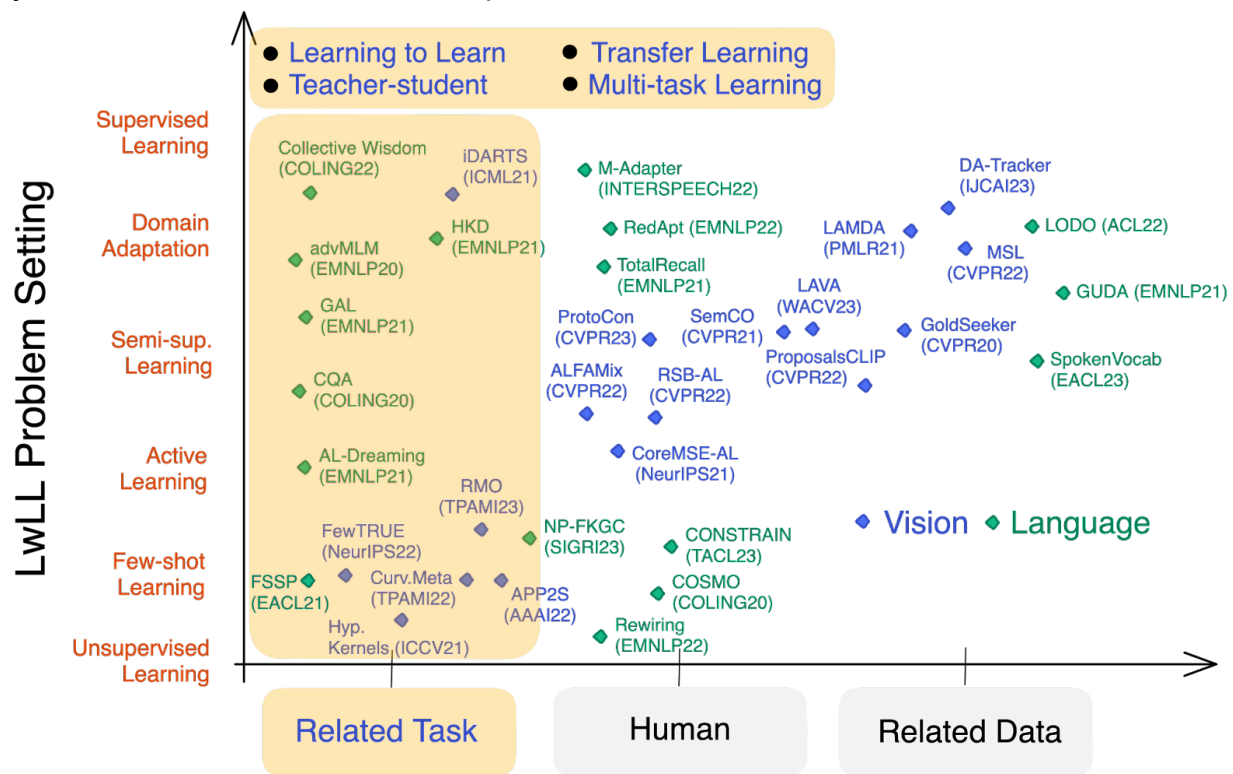


Figure 2 - Publication grouped by LwLL Problem Setting and Inductive Bias. Here, the

x-axis shows the sources of inductive biases, including related tasks, human, and related data. The y-axis shows the machine learning setting, including supervised learning, domain adaptation, semi-supervised learning, active learning, few-shot learning and unsupervised learning.

2.2.3. Final Overview Presentation

We have organized our papers based on a 2D space (see Figure 2) where each paper is given a coordinate based on the source of its inductive biases and the machine learning setting that it considers.

For completeness, **Appendix A** presents the list of papers out of our LwLL participation according to the areas that they correspond. Furthermore, **Appendix B** provides the final presentation of the LALA Team in DARPA LwLL. It compactly summarizes the main findings, contributions, breakthroughs, outcomes, and thoughts of the LALA team on learning with less labels, based on the efforts in the LwLL program.

3 METHODS, ASSUMPTIONS AND PROCEDURES

3.1 Image Classification

3.1.1. Overview

Under the LwLL protocol, the Image Classification (IC) task consists of two stages: the Base stage and the Adaptation stage. These stages use different but related datasets. The Base dataset classes may have some class and/or domain overlap with that of the adaptation dataset, but not necessarily. The goal is to progressively train our system over 8 checkpoints and make predictions at the end of each checkpoint on a given test set.

During the active-learning checkpoints, our system selects images from the unlabeled pool to ask for their labels, but it must adhere to a specific budget for each checkpoint. At the last checkpoint of each stage, our system has access to the complete corresponding labeled dataset. However, in the first seven checkpoints, our system only has access to a subset of the labels along with the rest of unlabeled images.

Furthermore, there are predefined whitelisted datasets available for each task that we can use to pretrain an initial model. In the evaluation phase, our whole system is tested from the first to the eighth checkpoint in a dataset unknown to us. This blind evaluation ensures generalizability of our model.

In light of the above protocol, our IC system is built around three main components:

1. **Whitelist Pretraining.** Due to the blind evaluation protocol used in LwLL, we lack prior knowledge about the specific properties of each task, such as the visual domain of the dataset (e.g., aerial, natural images, places) or whether it involves coarse or fine-grained classification. Consequently, it becomes challenging to decide which whitelist datasets to use for training our models. Additionally, pretraining a large model on all whitelist datasets and using it for all tasks is not possible since the whitelist datasets are task-specific and only disclosed at the start of the evaluation task (as part of the task meta-data). Moreover, due to limited GPU and time resources in the evaluation environment, it is infeasible to pretrain our models during evaluation time.
Hence, our first design strategy is to pretrain different base models on semantically grouped whitelist datasets and then develop a model selection algorithm to efficiently decide which model to employ for a given evaluation task.
2. **Semi-supervised Transfer Learning.** In all but the last checkpoint of the LwLL protocol, we are provided with partially labeled datasets. In the initial checkpoint, we choose (i.e., transfer learn) a model pretrained on the whitelist datasets in the Base stage or the Base dataset in the Adaptation stage. In subsequent checkpoints, we transfer from a model trained on the data from the previous checkpoint. Since the number of labeled examples are generally small, one major aspect of our strategy is to utilize unlabeled data to improve our models.
Therefore, we adopt semi-supervised transfer learning as the primary paradigm to develop our algorithms. In the course of the LwLL program, we have developed and published three algorithms corresponding to each phase of the program, which we elaborate on in the following subsection.
3. **Active Learning.** For the active learning (AL) checkpoints, we make use of our novel AL method, called ALFA-Mix. We propose a novel active learning approach

called ALFA-Mix as part of this project that is used in both image and video classification tasks. After the AL queries and expanding the labeled training set, we make use of semi-supervised learning to re-train the model using expanded labeled data and the unlabeled data. We will delve into the details later in this section.

The remainder of this section provides a comprehensive description of our IC system, detailing our approach to each of the design areas mentioned above.

3.1.2. Whitelist Pretraining

The blind evaluation protocol in LwLL leads to a lack of task-specific knowledge, making it challenging to choose appropriate whitelist datasets for model training. To tackle this challenge, our approach was to pretrain multiple base models on semantically grouped whitelist datasets, then develop a model selection method to efficiently decide which model to use at evaluation time. In summary, our approach can be summarized in the following steps:

- **Semantic grouping of whitelist datasets.** First, we split our whitelist datasets into groups (Table 1) based on their visual domain based on human-provided prior. The intuition here is to account for the possible tasks we would encounter during the evaluation. Note that this might be a sub-optimal categorization but in our preliminary experiments this led to the best performance on the development tasks.
-

Table 1 - Our Semantic Grouping of the Whitelist Datasets.

Natural Images	Domain Net domains	Aerial Images
imagenet_1k coco2014 domain_net-real	domain_net-quickdraw domain_net-clipart domain_net-sketch domain_net-painting domain_net-infograph	pool_car google_open_image (subset)

- **Pretrain a self-supervised model on each group of datasets.** We train three self-supervised models, one on each of the group’s combined dataset images. Even though, we have access to the labels of all the whitelist datasets, we opt for using the self-supervised learning algorithm (DINO) [14] for this stage due to the observations made in our paper [4] where we found that representations learnt by DINO transfers better to new classes and visual domains. Supervision with labels, more often than not, is too eager to learn specialized features which can successfully discriminate the source classes/domains but fails to generalize beyond them. We argue that its other limitation is the semantic-free nature of the labels used. Labels are represented using one-hot vectors explicitly encouraging to ignore label semantics. For example, the

learner is encouraged to treat “bus” and “school bus” as two unrelated classes. Self-supervised learning, on the other hand, does not use class labels to supervise learning but rather it employs auxiliary tasks as a source of supervision. For example, the objective used in DINO encourages the model to learn representations invariant to spatial augmentations and encourages the model to produce similar representations when seeing whole concepts (e.g., a car) or parts thereof (e.g., a wheel). Accordingly, this much richer supervision signal leads to learning more general representations which are not specifically tied to a set of predefined classes, hence improving generalization. We refer the reader to [6] for a more elaborate discussion.

3.1.3. Model selection at evaluation time

At the beginning of a given evaluation task, we are provided with the unlabeled images for both the Base and Adaptation datasets together with the task meta-data (e.g., class names as strings, train set size, labeling budget for checkpoints, etc.). Hence, in this step, we aim to select one of the 3 pretrained models to use as our initial model in the first Base checkpoint. This is to ensure we have chosen the model trained on the most similar dataset and has the most suitable biases for the task at hand. We formulate the model selection problem as follows: given our 3 pretrained models, source class names, target dataset unlabeled images, target class names, and the 1-shot labels available at the first checkpoint (i.e., 1 image per class in the Base dataset), what is the best model to use?

Intuitively, we would like to select the model which is pretrained on the most relevant images to solve the task-at-hand. To achieve this, we calculate the following metrics based on each of the 3 pretrained models:

- **Class overlap** between source classes and target classes: which denotes the semantic similarity between the classes on which the model has been self-supervisedly pre-trained and the target classes on which it will be further trained and evaluated. We use a language model to capture the semantic similarity to account for cases when there is no one-to-one string match.
- **Clustering quality** aims to measure how well the target unlabeled images are clustered in the representation space of each of 3 pretrained models. The rationale here is that the more well clustered the images are in a given model’s embedding space, the better that model would be able to discriminate between the classes. Hence, for each of the 3 pretrained models, we first embed the target unlabeled images (this is cheap as it does not involve any training), then we use K-means clustering to cluster the image embeddings into K clusters (where we set K to the number of classes in the target dataset), then we calculate the following clustering quality metrics: Silhouette, and Davies-Bouldin score [15].
- **Average Entropy** based on 1-NN classification. In this step, we make use of the 1-shot labeled data that is available in the first checkpoint. Using this single labeled instance per class, we can calculate prediction probability on the entire target dataset using weighted K-nearest neighbors. For a given unlabeled image, the probability of belonging to class k is the ratio between the relative distance of the image to the image with label k and the sum of distances with all the other k-1

labeled image embeddings. Knowing the prediction probability of each image enables us to then calculate the average entropy which is the measure of uncertainty of predictions using a given model. Intuitively, we would like to use the model which produces the least average entropy.

Now the question is: “based on the above calculated metrics specific for the given target dataset, how can we select the best model among the three to deploy at evaluation time?”. Our answer was to learn the importance of each metric via a simple decision tree model which is trained by emulating the LwLL protocol on 12 datasets. To elaborate, prior to the evaluation time, we selected 12 distinct target datasets spanning different visual domains, namely, BU-101 action dataset, the 6 domains of DomainNet dataset (real, clipart, sketch, infograph, quickdraw, painting), stanford-40 action dataset, caltech-256 dataset, mars surface proprietary dataset, MSL curiosity proprietary dataset, and mini-imagenet dataset. For each of those datasets, we calculate the above metrics based on each of our three pretrained models besides a randomly-initialized model. Additionally, we calculate the accuracy on the test set of the corresponding target dataset by using KNN on the labeled training set of that dataset. This results in a table of 48 entries, i.e., 4 source models X 12 target datasets (a subset is shown in Table 2). This, then, enables us to learn the importance of each of those metrics towards the resulting accuracy by training a simple decision tree model.

Subsequently, at evaluation time, we use our decision tree model, together with the calculated metrics for the target dataset, to select one of our 3 initial pretrained models.

Table 2 - A subset of the dataset used to train our model selection decision tree.

source_model	target_dataset	class_overlap	silhouette	db_index	entropy	knn_accuracy
domainnet	bu_101	0.00990	0.05281	0.04015	2.71267	76.87461
googleimages	bu_101	0.00990	0.07385	0.04195	2.35899	80.27725
imagenet	bu_101	0.03960	0.07354	0.04237	2.54578	85.12917
random_init	bu_101	0.00000	0.15217	0.02014	2.72059	15.79500
.....						
domainnet	caltech_256	0.25681	0.07157	0.02015	3.33460	78.38942
googleimages	caltech_256	0.26848	0.06696	0.01495	3.18303	73.19503
imagenet	caltech_256	0.18677	0.09497	0.01785	3.22282	84.59654
random_init	caltech_256	0.00000	0.12797	0.00972	3.63545	8.55929

3.1.4. Semi-supervised Transfer Learning

3.1.4.1. SemCo: Enhancing Semi-supervision via Semantic Grouping and Co-training

We make use of externally existing knowledge bases to improve sample efficiency in semi-supervised pseudo-labeling methods (see Figure 3).

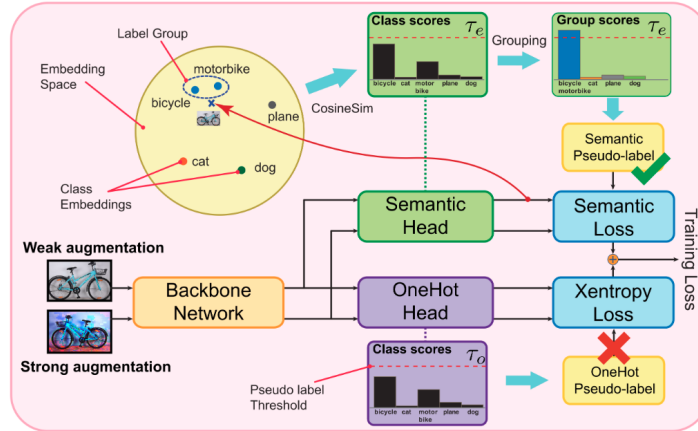


Figure 3 - A conceptual diagram of SemCo [5].

Motivation. Traditional pseudo-labeling methods use an iterative approach to leverage unlabeled samples during training: in each training iteration, the model is used to generate pseudo-labels for a batch of unlabeled samples which are then included as additional targets to train the model. To ensure that only reliably-labeled samples are used for training, confidence-based methods only select samples with high classification confidence to include in the pseudo-labeled data pool. In phase 1 of the program, we developed SemCo², a method which integrates knowledge bases with neural models to improve confidence-based pseudo-labeling. The key novelty of our method is to not only rely on the model's confidence to select pseudo-labels, but to also consider prior knowledge about the visual similarity among the classes. Our key insight is that visually similar classes (e.g., bike and motorbike) more often are associated with low confidence scores leading to their samples being less selected. This degrades the quality of pseudo-labeling as it poorly represents visually similar classes in the pool of pseudo-labeled data. Accordingly, we show that the visual similarity among the target classes is a particularly essential prior to improving those systems.

The main contributions of SemCo are:

- We show how knowledge graphs can be combined with language distributional statistics to obtain a prior on visual similarity in forms of label embeddings.
- We present a novel semi-supervised method to incorporate the visual semantic prior via a co-training procedure.

² Enhancing Semi-supervision via **S**emantic Grouping and **C**o-training

- We demonstrate state-of-the-art performance on standard benchmarks and strong performance on LwLL development datasets.

Visual Similarity Prior. We leverage the ConceptNet knowledge graph [16] and GloVe language embeddings [17] to obtain our visual similarity priors. ConceptNet is a multilingual knowledge graph that connects words of natural language with labeled, weighted relations including relations which convey visual similarity such as: *SimilarTo*, *IsA*, etc. GloVe is a prominent set of word embeddings trained on 840 billion words of the Common Crawl dataset to capture the distributional similarity among the different words but do not necessarily capture visual similarity. Combining the distributional embeddings with the ConceptNet graph allows us to address this problem: we follow a retrofitting procedure to align the distributional embeddings with the knowledge graph structure. Retrofitting [18] is a process which adjusts a word embedding matrix based on a knowledge graph by optimizing an objective function which tries to find for each term a new vector close to the vector's original value but also close to the term neighbors in the graph.

Model Training. In an LwLL evaluation task, before training our model, and once the class labels become known, we use a density-based clustering method in conjunction with our pretrained visual embeddings to group the class labels into visually similar groups using a carefully tuned predefined threshold. This way, we can control using an adaptive confidence threshold for visually similar classes to allow them to be selected more often for pseudo-labeling. Our model comprises a shared visual encoder (e.g., ResNet50) with two classifier heads. The first classifier (one-hot classifier) uses the one-hot view of the labels as targets and disregards any potential similarity among the classes, while the other (semantic classifier) uses the visual embeddings of the labels as targets and consider the label groups for pseudo-labeling. We then jointly co-train the two classifiers to learn based on their disagreements over pseudo-labels. Additionally, we follow previous work [19] and employ a consistency regularization approach by using weak and strong augmentations: we obtain pseudo-labels based on a weakly augmented view of the images and we train the model to predict a consistent pseudo-label for the strongly augmented view. Figure 3 depicts a high-level overview of our method.

3.1.4.2. LAVA: Leveraging self-supervised learning to improve semi-supervised transfer learning

Motivation. In phase 2 and 3 of the program, we focused on self-supervised learning and studied how to best employ it to improve semi-supervised low-shot transfer. At the time, self-supervised methods were starting to show promising results in learning more generic task-independent visual representations [14].

Due to the richer training signal they employ, those methods excel in learning class-independent features that allow for better performance on low-shot recognition and transfer tasks across classes and visual domains.

We hence chose to study the generalizability of self-supervised pretrained models using the LwLL evaluation protocol on the development datasets. This led to novel insights which we translated to LAVA (Figure 4), a method for low-shot transfer learning which excels in cross-domain generalization. Our study showed that supervised pretrained models result in representation collapse when transferred to a target dataset with class or domain shift. Since supervised models are trained on a predefined set of

classes, more often than not, they are too eager to learn specialized features which can successfully discriminate the source classes/domains but fails to generalize beyond them.

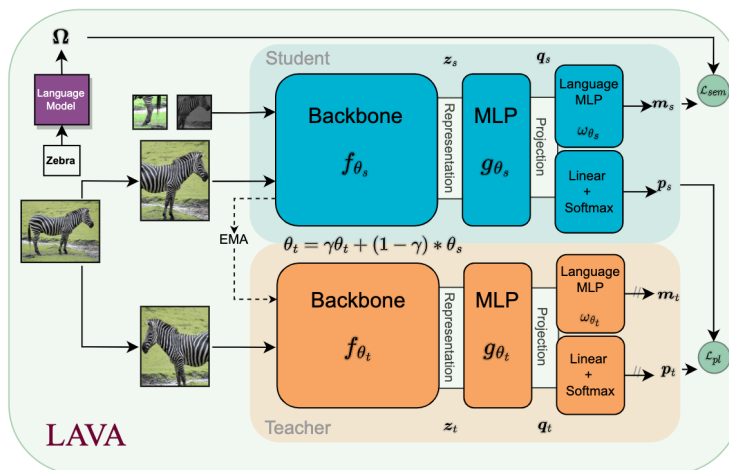


Figure 4 - LAVA Method Overview [4].

We argue that the other limitation of supervised pretraining is the semantic-free nature of the class labels used. Labels are represented using one-hot vectors explicitly encouraging to ignore label semantics. For example, the learner is encouraged to treat “bus” and “school_bus” as two unrelated classes. In Figure 5, we use DomainNet clipart dataset to qualitatively demonstrate two artifacts associated with supervised pretrained representations: class collapse, whereby the pretrained representations collapse into incorrect source classes just because they share superficial similarities with target classes; and domain collapse, where the class semantics are preserved but the visual domain information is disregarded.



Figure 5 - Class & Domain collapse. We display two query images from clipart target (left), and their 4 nearest neighbors, from clipart and ImageNet, in LAVA's representation space (middle) and an ImageNet-supervised learner's (right) before training on target instances. ImageNet instances are bordered in blue

Method Details. LAVA's first ingredient is introducing a two-fold approach to address the demonstrated collapse problem: 1) self-supervised source pretraining to learn task-independent features leading to better transfer; and in the second phase of the program, we additionally use the language modality to ground the self-supervised representations to an independent semantic space: during pretraining, LAVA uses source class labels to learn a mapping between the visual representations of the source instances and the language semantic representation of their class labels. At transfer time, that mapping is used to infer relations between "seen" and "unseen" classes by virtue of their foreknown semantic similarities. LAVA's second ingredient is to leverage unlabeled target data to improve transfer performance. For that, we employ multi-crop augmentation which was originally proposed to encourage learning representations that are invariant to spatial augmentations for self-supervised contrastive learning [20]. We extend the idea to semi-supervised learning exploiting the observation that images can often contain multiple semantic classes and hence using a single label per image can hurt the performance. Hence, we calculate pseudo-labels based on multiple local and global views of images to account for images containing multiple semantic concepts. We show that applying a pair-wise aggregate loss across multiple views enhances the quality of pseudo-labels.

3.1.5. Active Learning

We propose a novel active learning approach called ALFA-Mix [7] as part of this project that is used in both image and video classification tasks. The proposed approach identifies unlabeled instances with distinct features by detecting inconsistencies in model predictions resulting from interventions on their representations. It achieves this by constructing interpolations between representations of labeled and unlabeled instances and analyzing the predicted labels. Our work demonstrates that inconsistencies in these predictions highlight features that the model struggles to recognize in the unlabeled instances. The efficient implementation of ALFA-Mix is achieved based on a closed-form solution that optimally determines the interpolation causing changes in predictions.

Existing AL strategies often face challenges when applied to deep neural networks, high-dimensional data, and scenarios with limited data. ALFA-Mix addresses these challenges by efficiently identifying distinctive features within unlabeled instances and outperforming other approaches in various experimental setups.

Extensive experimentation in our paper [7] on 12 white benchmarks covering images, videos, and non-visual data demonstrates the superiority of ALFA-Mix over various state-of-the-art AL methods in 30 different settings. Notably, the proposed approach showcases significant improvements, particularly in scenarios with limited available data and when applied to self-trained vision transformers. ALFA-Mix outperforms existing methods in 59% and 43% of the experiments for these cases, respectively. Our work underscores the importance of active learning in enhancing machine learning applications by iteratively selecting valuable instances for labeling, especially in situations where high-quality annotations are scarce or expensive.

3.2 Video Classification

3.2.1. Overview

Video Classification (VC) is a challenging task in computer vision that automatically classifies videos into specific predefined classes or labels. Unlike other vision tasks, VC encounters distinct complexities due to the temporal dimensions of the video data. Videos are sequences of frames presented over time, incorporating both temporal and spatial dimensions. This is different with traditional vision tasks that primarily handle two-dimensional images containing only spatial information.

As a result of the aforementioned characteristics, VC faces specific challenges that demand careful consideration in its implementation:

1. **Efficiency:** One of the challenges is handling the large data size associated with video content. Videos containing numerous frames, making video analysis computationally intensive tasks. The fine labeling for video is also expensive, which needs to label each frame of video.
2. **Sparse Target Features:** Another difficulty arises due to the sparsity of target features in video data. Not all frames in a video contain relevant information for classification.

To address these challenges and enhance the effectiveness of our VC system, we have dedicated efforts to upgrading the state-of-the-art VC model. Our system can better capture various features and improve overall performance. We carefully choose a pretrained dataset with various types of actions, enhancing the ability of the backbone networks in extracting meaningful features. Moreover, we meticulously select the optimal sampling ratio for the evaluation dataset to select appropriate frames for classification. Addressing unlabeled data, we also propose a new approach to work across different learning regimes using varying amounts of labeled and unlabeled data. This approach significantly enhances the effectiveness of our method on datasets with limited labels. Furthermore, our activate-learning method can enhance the ability of models to learn more novel features of video to reduce the requirement of labeled data. Through these improvements, our system aims to achieve more accurate and reliable video classification results, making it suitable for various real-world applications.

3.2.2. Model Architecture

The architecture of our system is shown in Figure 6. We adopt the MViT model as our feature extraction backbone and introduce an additional semantic encoder to generate semantic embeddings. Leveraging the linguistic semantic information of the label, we incorporate the inductive bias to facilitate low-shot transfer learning.

To create the semantic embeddings, we harness a pretrained language model to extract linguistic features from the labels. These linguistic features serve as valuable guidance for the model during low-shot transfer learning. Our objective is to encourage the semantic embeddings to be closely aligned with the linguistic features of samples from the same class while being distinctly separated from linguistic features belonging to other classes.

For labeled inputs, we adopt a combination of the semantic loss and the cross-

entropy loss function. The semantic loss helps to create meaningful embeddings that capture the underlying semantic information, while the cross-entropy loss drives accurate predictions for labeled samples.

By combining these innovative strategies, our system excels in low-shot transfer learning scenarios, where labeled data is limited. The integration of MViT [21] for feature extraction, semantic encoders for meaningful embeddings, and self-distillation for knowledge transfer empowers our system to achieve state-of-the-art results in various datasets.

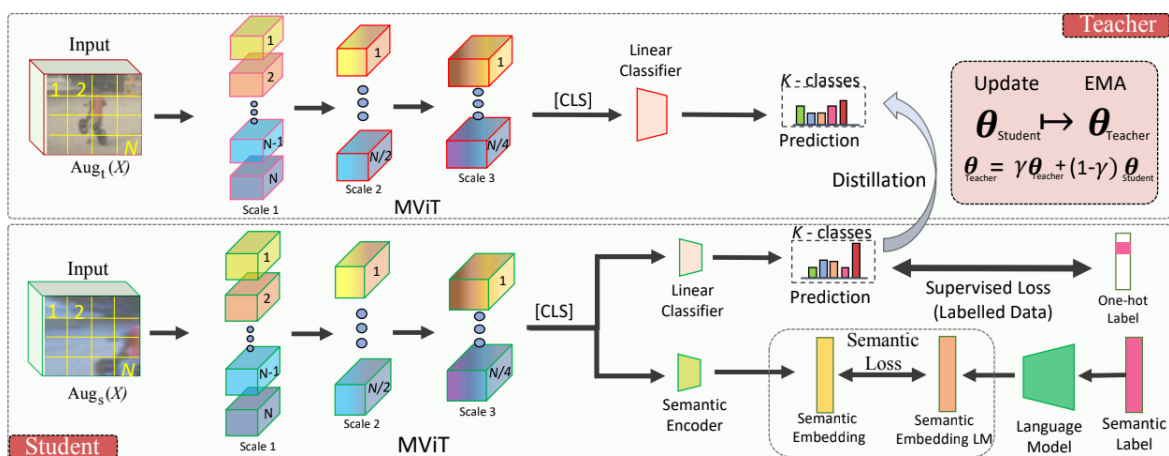


Figure 6 - A overview of our architecture [6].

3.2.3. Pre-evaluation Preparation

3.2.3.1. Data Preparation

We use various augmentation methods to enrich the dataset and increase the model's performance. These augmentation techniques encompass mixup, color jitter, and random augmentations, including Random Rotation, Random Shear, and Random Sharpness. Through these augmentation methods, we significantly augment the training dataset space, empowering the model to acquire more robust and generalized features.

We also need to pick up a fixed number of frames from a video as the input size of the model is fixed. The sample ratio we set is 2 indicating that we pick up one frame from every two frames in the video. We select 32 frames from each video with the set sample ratio.

3.2.3.2. Pretrained Model Selection

Using a model pretrained on a large dataset can significantly increase the

performance. It initializes the model's weight which already learns various features on the pretraining dataset. We use Kinetics-600 to pretrain our model. The Kinetics-600 dataset collects 480,000 videos distributed across 600 action categories. Each video in the Kinetics-600 dataset comprises a 10-second clip, annotated to capture the precise moment of action from the raw YouTube videos. Compared with Kinetics 400, Kinetics 600 has more than 200 classes, enhancing the capacity of our backbone network to learn more class-specific features. Notably, the training set of Kinetics 600 comprises approximately 100,000 more videos than Kinetics 400. Although Kinetics 700 has 100 more classes than Kinetics 600, it contains 200,000 more training videos. In order to make a trade-off between training time and dataset diversity, we select kinetics 600 as our pretrain dataset. In Figure 7, we compare the performance of our method with different pretrained datasets.

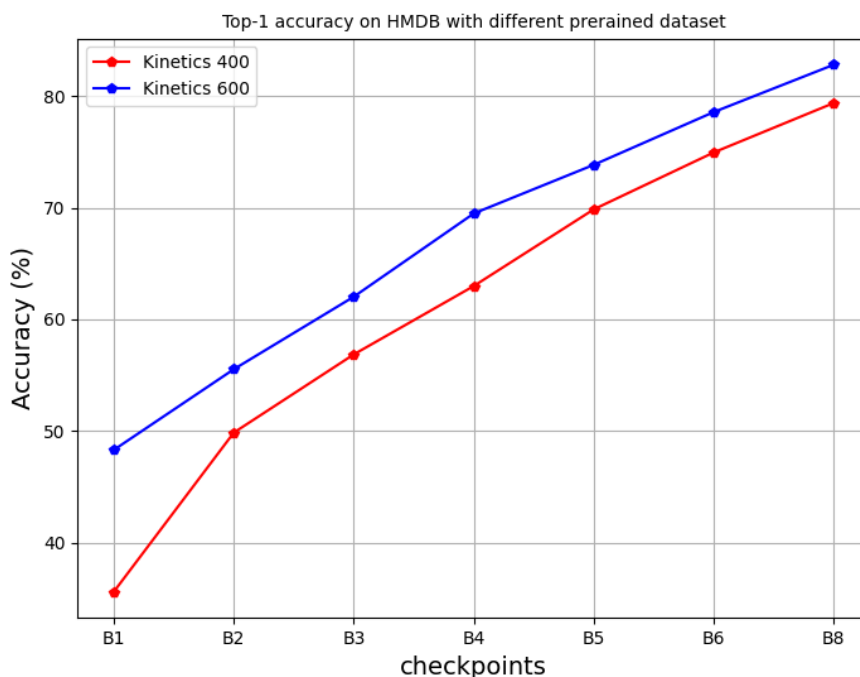


Figure 7 - Pretrained dataset comparison.

3.2.4. System Evaluation

We adopt the same strategies for both base stage and adaptation stage.

3.2.4.1. Seed-Label checkpoint

For the first four checkpoints, we will get one or a few labels for each class, resulting in many unlabeled videos. To address this challenge, we introduce a novel approach, as shown in Figure 8. When processing these unlabeled videos, we adopt a slow-changing network that acts as a "teacher", while ensuring the alignment of outputs between this teacher network and the primary network for the unlabeled videos. This approach allows the network to exploit those unlabeled data.

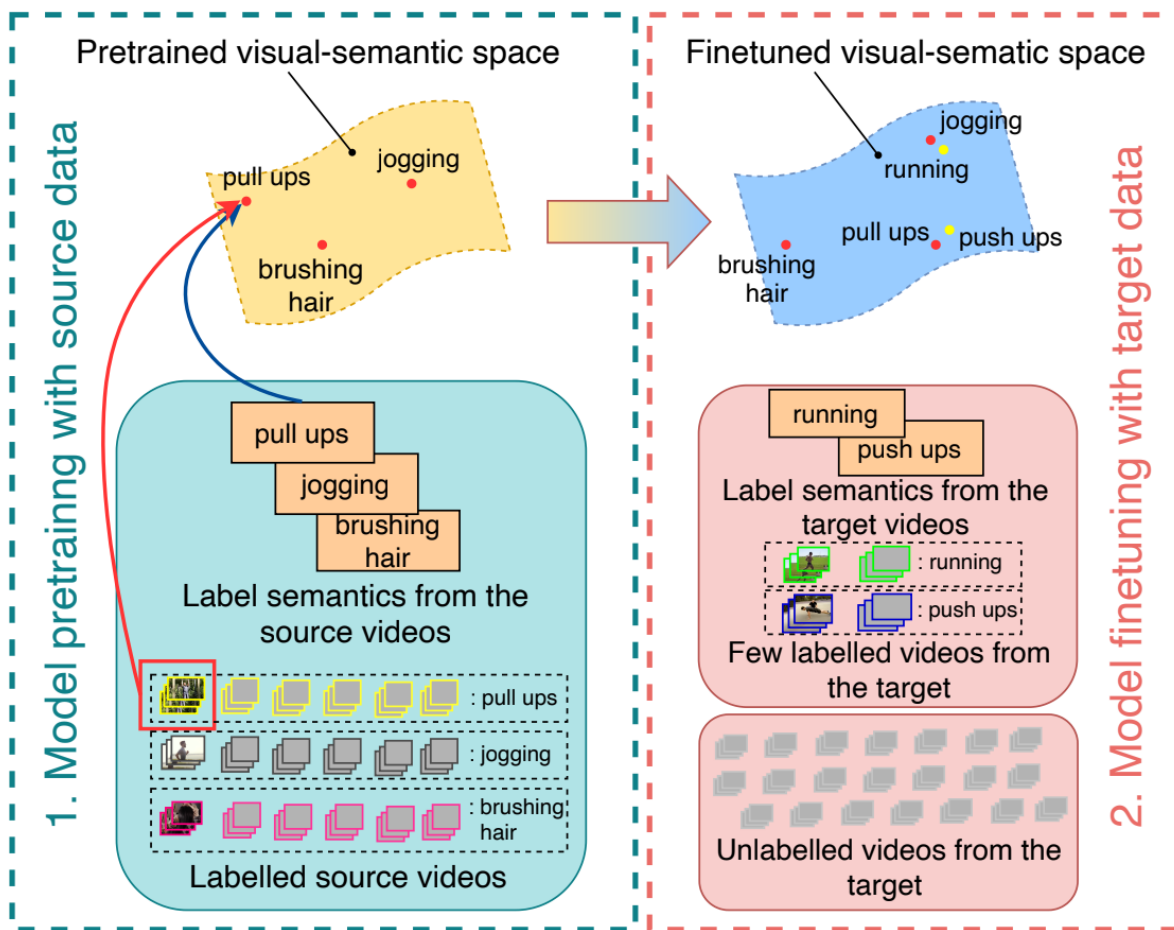


Figure 8 - Few-labeled action recognition approach [6].

3.2.4.2. Active Learning checkpoint

The active learning strategy employed in our system revolves around enabling the model to learn novel and informative features, depicted in Figure 9. When dealing with unlabeled data, we interpolate the features of unlabeled videos with the features of different classes. By mixing these features, we create synthetic features. Next, we leverage the current model to predict the class for these synthetic features. If the model's predictions remain consistent and align with prediction of the corresponding unlabeled videos, it suggests that the features contained in unlabeled video already exist in the knowledge captured by the model. However, if the model's predictions for the synthetic features diverge significantly from the corresponding unlabeled videos, it indicates that these features are novel and distinct from what the model has learned. In such instances, we identify these unlabeled videos as potential candidates for the next round of labeling. By this strategy, we allow the model to progressively learn new representations from unlabeled data.

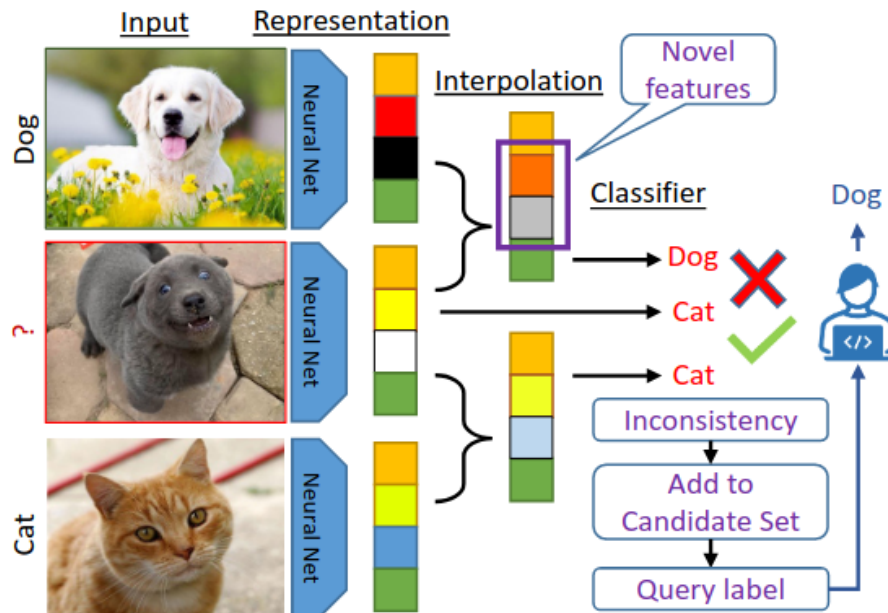


Figure 9 - Activate learning method [7].

3.3 Object Detection

3.3.1. Overview

Object detection is a computer vision task focused on identifying and localizing multiple objects within an image or video by predicting bounding boxes and associating corresponding class labels. Overcoming challenges in object detection is crucial for creating a reliable system capable of detecting objects accurately in various scenarios. In line with the JPL evaluation protocol, our aim is to develop an object detection system with high accuracy, especially when dealing with limited data. To achieve this goal, we have made two main contributions.

- Beyond the LWLL API we created a comprehensive benchmark dataset by gathering datasets from different domains. This benchmark covers a wide range of objects from various domains and categories, ensuring that our system's performance can be tested and applied to various situations.
- We developed a system that finely tuned with optimal parameter setting using our benchmark dataset. We tried out different detectors and backbones to see which ones work better for our evaluation. Then, we adopted a self-supervised training strategy by generating noisy pseudo labels from the existing state-of-the-art solutions and using these pseudo labels along with the annotated labels in each checkpoint to train the model. Lastly, after trying different state-of-the-art semi-supervised strategies we chose the one with high accuracy based on what we learned from our benchmark dataset experiments.

In the following sections, we will outline the fundamental architecture of our system and delve into its key components. Furthermore, we will detail the process of selecting suitable detectors and backbones. Finally, we will present the outcomes of our evaluation based on the JPL assessment. Our results demonstrate that we consistently outperformed the baseline JPL system and secured a position among the top-performing teams.

3.3.2. Components

3.3.2.1. Benchmark Dataset

To construct a robust object detection system within a context of limited labeled data, capable of universally identifying all object types across diverse application domains, it is imperative to access a benchmark using datasets sourced from diverse domains. To this end, we dedicated an effort to construct such a benchmark dataset from diverse domains such as satellite, construction, natural, farming, fashion, and underwater.

To be consistent with the JPL evaluation protocol, we attempted to pair four datasets for each domain—two designated for development and two for evaluation. In cases where obtaining four datasets for specific domains proves challenging, a pragmatic approach involves employing two datasets for evaluation while utilizing other available datasets for the developmental phase.



Figure 10 – Samples from various domains.

Table 3 and Figure 10, respectively, represent the list of our datasets and a few examples from each dataset in our benchmark. The chosen datasets introduce a broad spectrum of diversity in terms of domain-specific challenges, object characteristics, camera resolutions, detection class numbers, and dataset scales. Figure 11, Figure 12 and Figure 13 demonstrate graphs providing information about the dataset size and the number of classes in each dataset.

Table 3 – Datasets of Different Domains.

Satellite	Natural	Farming	Construction	Fashion
SIMD ISAID Dior DOTA	PascalVoc Exdark Open Images (Short Version) Object365 (Short Version)	ACFR+ WGISD VegImages	MOCS SODA	DeepFashion2 Fashionpedia

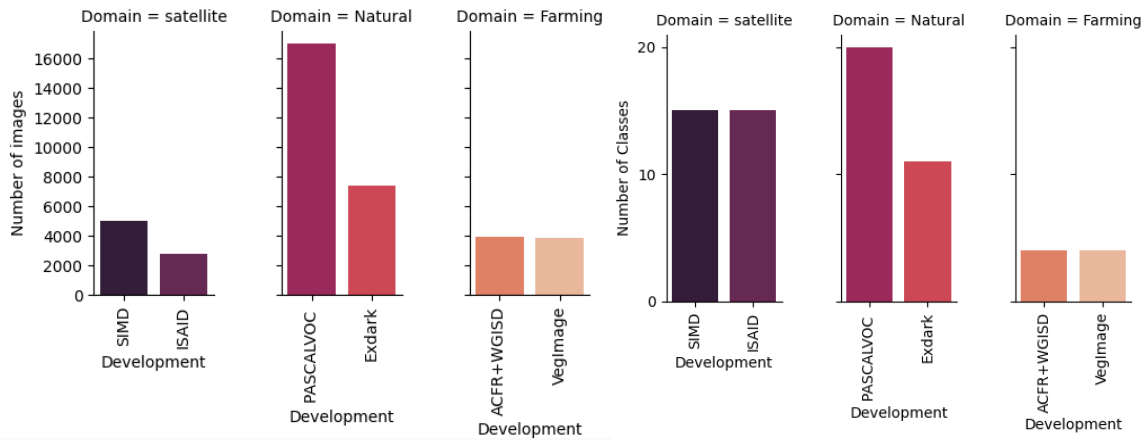


Figure 11 — Number of images and classes in the Development Dataset.

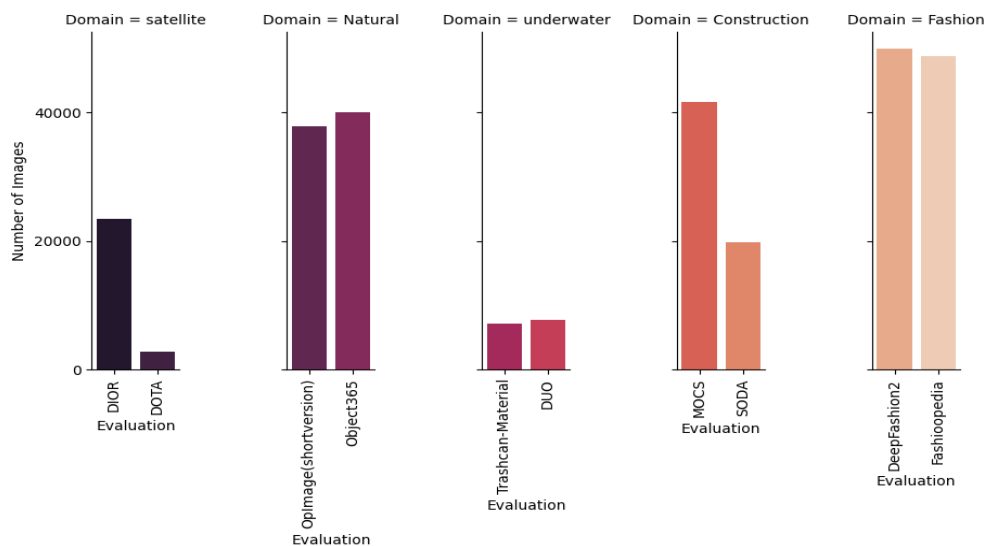


Figure 12 – Number of Images in the Evaluation Dataset.

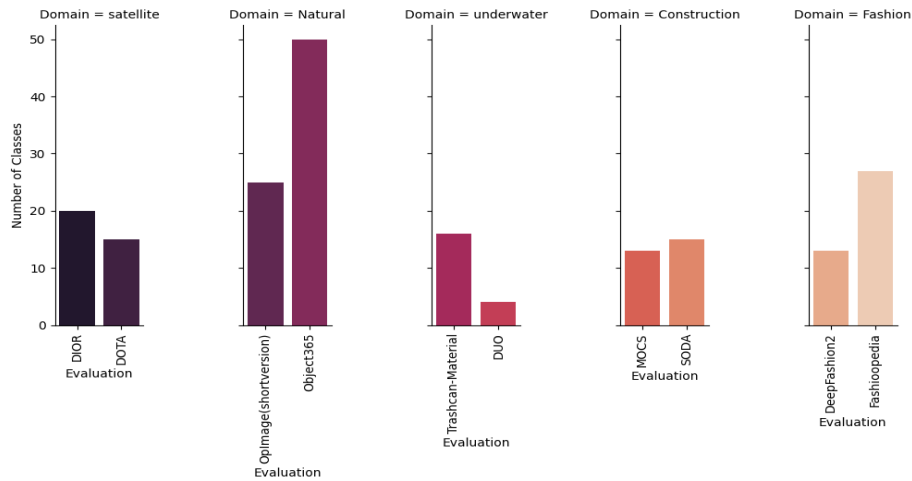


Figure 13 – Number of Classes in the Evaluation Dataset.

3.3.2.2. A highly tuned parameters system

The benchmark dataset discussed in Section 3.3.2.1 played a crucial role in guiding our decisions regarding optimal design choices, training strategies, detector models, parameter settings, and more. In the following, we provide a brief overview of our system components and the reasons behind our choices.

- **Parameter setting & model decision.** At the beginning, we opted for the most effective baseline design, which involved a supervised model with meticulously adjusted parameters. We assessed detectors like RetinaNet and Faster R-CNN, using various backbones like ResNet50, ResNet101, ResNeXt101, Swin Transformer, and PTV-v2. During training, we tested different parameter configurations to attain the best outcomes. We tweaked parameters such as learning rate, weight decay, batch size, and iterations per checkpoint to optimize performance. Based on these tests, we ultimately settled on the detector type, its backbones, and hyperparameters.
- **Self-supervised pre-training strategy.** After training the supervised model on white-list datasets, our aim was to make the trained model adaptable to the target domain (evaluation) dataset through an unsupervised pre-training strategy. For this purpose, we employed the advanced Open-Vocabulary Panoptic Segmentation with Text-to-Image Diffusion Models (ODISE) framework [22]. This involved using a list of class names and images from the evaluation dataset as input for the ODISE framework. Consequently, ODISE produced a set of somewhat noisy pseudo labels from these inputs. Finally, we refined the supervised baseline further using these pseudo labels, enabling the model to adjust to the evaluation dataset.

Note: Since the last LwLL evaluation, we have incorporated this new pre-training strategy, which isn't reflected in the results presented in Section 4.3.2. This strategy has shown promising results in enhancing our outcomes substantially. However, it has been partially assessed using our benchmark dataset and its

comprehensive evaluation has been left for future. For a detailed comparison between our system's performance with and without this stage, please refer to section 4.3.1.

- Semi-supervised learning.** Our system operates on a semi-supervised architecture, specifically employing an Unbiased teacher for semi-supervised object detection. This unbiased teacher model consists of two key phases: the burn-in stage, where the object detector is trained using provided annotated data at each checkpoint, followed by the teacher-student mutual learning stage. Within the Teacher-Student Mutual Learning stage, two steps unfold: student learning and teacher refinement. This design facilitates a reciprocal learning process between the Teacher and Student models through a mutual learning mechanism (as depicted in Figure 14). The Teacher generates pseudo-labels to train the Student, reciprocally updating its knowledge to the Teacher, thereby enhancing the quality of pseudo-labels used for Student training. In the teacher refinement stage, exponential moving averages (EMAs) on network weights enable the transfer of knowledge from the Student to the slowly evolving Teacher [23]. We have embraced this framework in our system, markedly enhancing its performance through optimized backbones and parameters.

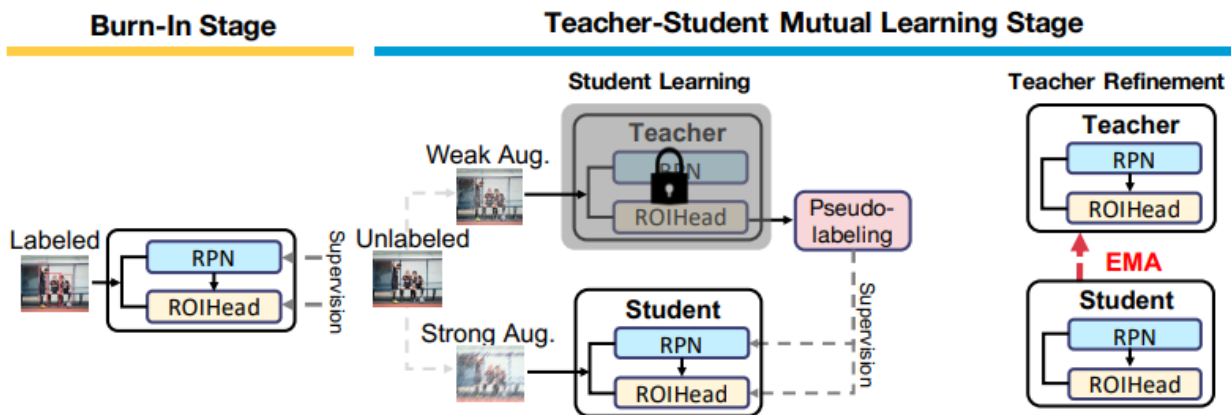


Figure 14 – Overview of Unbiased Teacher [23].

- Final system.** Finally, we employed teacher-student learning methods that used pre-trained models as an initialization. Then our models are trained through ODISE pseudo labels, combined with annotated data at each checkpoint.

3.4 Machine Translation

3.4.1. Overview

The Machine Translation (MT) evaluation includes translation tasks from unknown source languages to English. The submitted MT system should only make use of monolingual data from Oscar and English-centric parallel data for 5x whitelist languages.

Oscar is a collection of web crawled data and contains 166 languages [24]. The main challenge of MT tasks is to accommodate a large set of possible languages which may belong to different language families with different writing systems and syntactic structures.

Unlike computer vision tasks in previous sections, Machine Translation is not a classification task but a structure prediction task. Hence, the evaluation protocol of the MT task is slightly different from other vision tasks. All the checkpoints in the MT task are active learning checkpoints where we have the flexibility to choose unlabeled data points (i.e., monolingual sentences) to ask for annotation (i.e., translation). Additionally, the MT annotation budget is at the word level, not sentence (data point) level. In addition to the challenges of LwLL learning problems such as active learning and domain adaptation, the MT evaluation protocol raises multiple interesting research challenges, including:

- **Multilinguality:** How to accommodate multiple languages without the negative interference? Although multilingual pretrained language models are giving hope to many low-resource languages, we still have a long way to go to support other thousands of languages [25]. Cross-lingual transfer to unseen languages provides a solution to the curse of multilinguality, where adding more languages to models delivers a performance gain in low-resource languages [26]. The main challenge is that the unseen languages may vary in the amount of available data and their scripts, as well as other linguistic properties.
- **Efficiency:** How to efficiently learn a good NMT model on the large scale of data within the time and resource constraints?
- **Learn from noisy data:** Leveraging pseudo-labeled data through back translation or mining parallel data from monolingual text are common approaches to overcome the challenge of limited labeled data in NMT. However, these approaches often yield noisy datasets. How to train an NMT model that is robust to noisy training datasets and capable of producing high-quality translation?

We expand the capability of the state-of-art pretrained sequence-to-sequence model mBART to cover all languages of interests. We also perform data cleaning for monolingual data and mining parallel data to the continue-pre training step. Additionally, to address the negative interference, we conduct language transferability analysis to identify top **donors** for each **recipient** language.

The rest of this section describes our NMT system and approach to address the aforementioned challenges.

3.4.2. Model Architecture

Our NMT system is carefully designed for training efficiency and storage optimization. To achieve these goals, we have adopted the Hyper-Adapters architecture [9] - a variant of the Transformer model [3] with hypernetwork [27] to generate adapters on-the-fly. Adapters are specialized modules integrated into a pretrained network, allowing quick adaptation to new tasks without updating the entire backbone; hence, reducing training time [28]. However, maintaining a separate adapter set for each language, especially to support around 100 languages, is impractical and not scalable. The **hypernetwork** resolves this by dynamically generating adapters, resulting in significant reduction of storage requirements for handling multiple translation languages.

Adapters. An adapter module is inserted into every transformer layer of the pretrained NMT model to capture language-specific or task-specific features. For a layer l , the adapter module includes a down-projection W_{down}^l followed by an up-projection W_{up}^l to project the hidden state h^l to a lower-dimension space (bottleneck size) then

project back to high-dimension space $h^l = h^l + f(h^l W_{down}^l) W_{up}^l$ where $f(\cdot)$ is a nonlinear activation function. We denote $\omega = \{W_{down}^l, W_{up}^l\}_{l=1}^L$ as adapter parameters. While freezing the encoder and decoder parameters, the adapter module can be learned for either a given translation task X-Eng, a language X, or a domain. Figure 15 illustrates the adapter architecture and its added position in a pretrained Transformer layer.

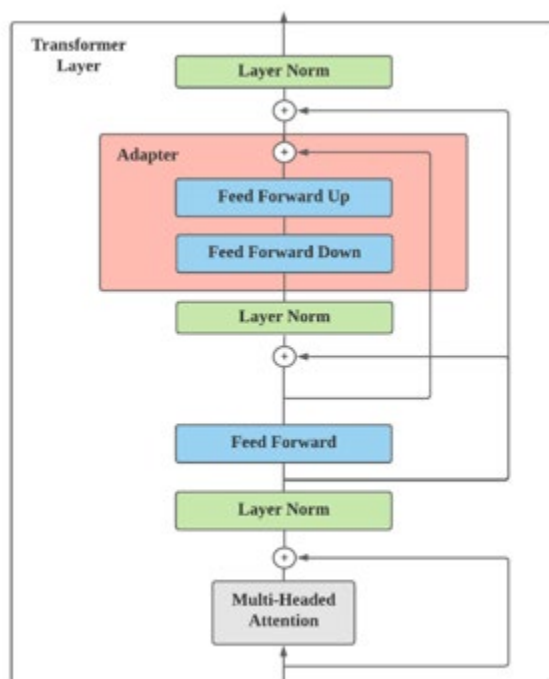


Figure 15 - Adapter architecture and its added position in a Transformer layer.

Hyper-Adapters. They utilize hyper-networks to generate adapters based on language and layer embeddings. A hypernetwork is a type of networks that generates the weights of another network [27]. In this case, a single hypernetwork is employed to generate adapters for all languages and layers by conditioning on the source and target language and the layer id. Each encoder and decoder transformer layer in the NMT model is assigned a unique id, such as enc6.

The hyper-adapter contains embedding layers, an encoder and a projection layer, as shown in Figure 16. We first embed source language s and target language t with a shared language embedding layer. We use a separate embedding layer for the layer-id

embedding for all encoder and decoder layers. The language and layer-id embeddings are concatenated then projected via a feedforward layer and a non-linear activation function to obtain hyper-network hidden-representation h_0 . We then pass h_0 through an encoder containing N residual blocks $h_{i+1} = W_2(\phi(W_1 LN(h_i))) + h_i$ where W_1 and W_2 are the trainable weights and $\phi(\cdot)$ is the activation function.

The final representation h is fed to separate projection heads to obtain the up-projection and down-projection matrix for the adapter. The hypernetwork is trained together with the backbone during pretraining step. We will describe the training detail in the next section.

Model Configuration. We use mBART-large as our backbone. mBART-large is a sequence-to-sequence Transformer with 12 layers of encoder and 12 layers of decoder, hidden state of 1024 and 16 attention heads. For the adapter, we set the bottleneck size to 128. For the hyper-network, we set the size of language and layer embeddings to 50, hidden state dimension to 204 and use 2 layers in the encoder.

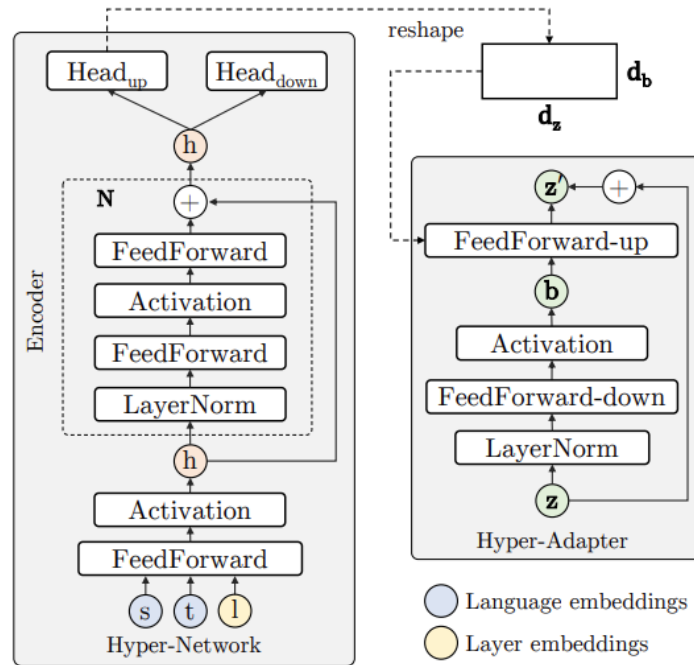


Figure 16 - Hyper adapter architecture [9].

3.4.3. Pre-evaluation Preparation

3.4.3.1. Data Preparation

To augment our training data, we extracted parallel sentence pairs related to our system from OPUS³ [29]. OPUS provides both monolingual corpora and parallel corpora in multiple languages, but we focused solely on the monolingual text for our work. Specifically, we mined pairs in the form of <English, X>, where X represents a language allowed by OPUS and covered in their dataset.

To obtain high-quality data, we followed the guidelines outlined in [30]. Initially, we cleaned the monolingual text and extracted bitext using the prescribed methodology. From the mined bitext, we then carefully selected a subset of data that met our quality criteria. Given that the monolingual text in OPUS contains a considerable amount of noise, we employed language identification (LID) techniques to remove non-target text. Our text cleaning process involved a series of rule-based filters and cleaning techniques, followed by deduplication of the corpora.

To ensure better alignment pairs during the mining process, we prioritized higher-quality English data. Therefore, we applied a language model exclusively to the English text. This step aimed to improve the alignment quality of the mined pairs.

During the mining process, we leveraged the LASER toolkit⁴ to generate sentence embeddings. Prior to computing pairwise sentence similarity scores, we made careful choices regarding the LASER encoders we employed. Specifically, we utilized LASER 3 encoders [31] for the languages of interest covered by those encoders. For the remaining languages, we either tried to find the closest language covered by the LASER 3 language list, or utilized the multilingual LASER 2 encoder. Note that due to the limitations of our university-level hardware resources, we partitioned the corpora into subsets. This enabled us to efficiently and concurrently complete the entire mining process.

The resulting mined corpora have a total number of ~1.8 billion sentence pairs, which can be computationally expensive to use. For this reason, we further filtered the data based on a set of filtering criteria (e.g., we accepted the top p% of pairs based on the alignment scores, and rejected pairs whose source-target length ratio fell out of a certain range). We report the data statistics before and after the filtering stage in Figure 17. Ultimately, we kept ~490M sentence pairs.

³ <https://opus.nlpl.eu>

⁴ <https://github.com/facebookresearch/LASER>

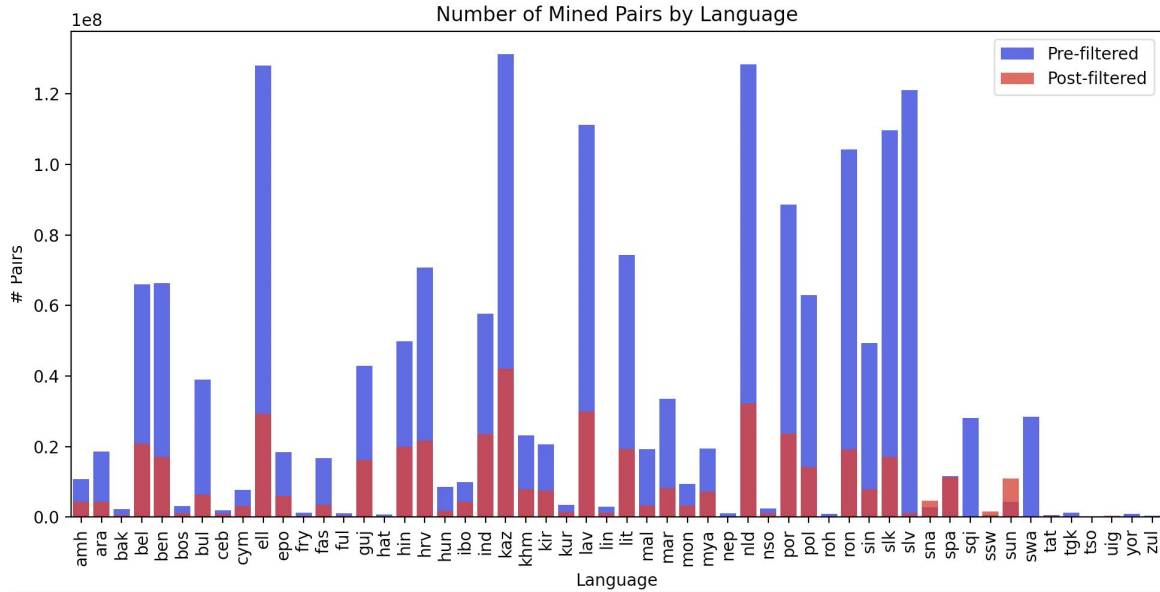


Figure 17 - Number of Mined Sentences by Source Language before and after filtering.

3.4.3.2. Pretraining

We extend the mBART model [10] to support all whitelist and evaluation languages.

Further Pretraining

For training, we further finetune the mBART model on our own dataset. Following [32], we firstly rescale the dataset size across languages with the temperature = 3. After rescaling, we spend more computation on the low-resourced languages. Besides rescaling, we mainly employ three strategies to perturb the input text for preventing the overfitting following [33]:

- **Random Word Masking:** We randomly mask out 20% of the tokens of both source and target sentences.
- **Random Word Replacement:** We randomly replace 20% of the tokens of both source and target sentences with other random words in the vocabulary.
- **Code Switching:** We randomly swap the positions of 20% of the tokens of both source and target sentences.

For each batch, we randomly choose one of these three strategies. We implement the strategies described above and further finetune the mBART model using Fairseq [34].

Hyperadapter Pretraining

During the preparation phase, we also pretrain the hyper-adapter together with the Transformer backbone on the whitelist parallel data and mined bitext obtained in the previous section. We use Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.98$, $\epsilon = 10^{-6}$ and regularize models with label smoothing with $\alpha = 0.1$. We train the models with a learning rate of 1e-3 for 1M steps on 8 V100 GPUs for 10 days.

3.4.3.3. Donor Language Selection

We employ a collection of 100 languages sourced from OPUS100 bilingual corpora for identifying potential donor languages. These 100 languages are partitioned into two subsets. The first subset comprises 50 high-resource languages, denoted as donor languages, whereas the latter consists of a combination of 50 mid- and low-resource languages, referred to as recipient languages.

For each language XX in the donor subset, we train a bilingual MT model from XX to English utilizing our enhanced mBART model. Then, for each language YY in the recipient subset, we apply fine-tuning to the YY-English data on each bilingual model derived from the donor set. Although we can use this method to generate a 50x50 ranking matrix between the donor and recipient languages to find the most transferable donor languages, due to the computational constraint, we can only obtain the partial results for the matrix. To complete the missing cells, we follow [35] and use multiple data and linguistic features to train a ranker with BLEU scores on hand. Then we can predict the rank of the missing cells of the transferability matrix. The complete set of dataset features includes the following:

- **Dataset Size:** The number of data entries used for training the recipient language.
- **Word/Subword Vocabulary Size:** The number of word/subword types for the recipient language.
- **Average Sentence Length:** The average length for the recipient language.
- **Word/Subword overlap** for the recipient language
- **Type-Token Ratio (TTR):** The ratio between the number of types and number of tokens of the training for the recipient language.
- **The genetic, syntactic, phonological, geographical feature** distance between the donor and recipient languages.

Now we can complete this language transferability matrix based on the predicted result. Finally, we select the most effectively transferable languages to be the final donor languages.

3.4.4. System Evaluation

3.4.4.1. Initialization

At the start of the evaluation process, we receive details about the source languages being evaluated, along with the source-side data from both the Base Task and Adaptation task. When available, we utilize mined parallel data for the source language as a noisy training dataset to enhance the performance of the evaluation. In cases where mined parallel data is unavailable, we use parallel data from donor languages. Moreover, we initialize the source language embedding with the embedding from the corresponding donor language. The donor language for each evaluation language is outlined in Table 4. We found that 9 languages Italia (ita), Hebrew (heb), Catalan (cat), Turkish (tur), Serbian (srp), Macedonian (mkd), French (fra), Russian (rus), Danish (dan) are top donors for other evaluation languages. For languages not covered in the Table, we use Turkish (tur) as the default language.

Word embeddings constitute a significant portion of the model parameters. However, during evaluation, our primary focus is on a bilingual translation task from language X to English. As a result, the subword units related to other languages become

redundant. To address this, we eliminate the redundant subword units from the vocabulary, leading to a reduction in model parameters of up to 40% during evaluation. To prepare the vocabularies for each language, we utilize monolingual data that has already been tokenized using the mBART sentence-piece model. Depending on the size of the monolingual data, we select the most frequent tokens to construct the vocabularies, with the aim of maintaining a vocabulary size of approximately 50K subwords for each language. During evaluation, we augment the vocabularies with additional subwords from both the Base-Dataset and the Adaptation-Dataset's provided unlabeled data. This step is essential to enhance vocabulary coverage and prevent encountering unknown tokens during the evaluation process.

Table 4 - The donor language for each source language.

Src	Donor	Src	Donor	Src	Donor	Src	Donor	Src	Donor
amh	ita	ara	heb	bel	ita	ben	tur	bos	mkd
bul	cat	cym	srp	ell	cat	epo	tur	fas	tur
guj	tur	hin	srp	hrv	srp	hun	ita	ind	tur
kaz	cat	khm	srp	kir	srp	kur	dan	lav	srp
lit	heb	mar	tur	mya	rus	nep	tur	nld	dan
por	cat	pol	ita	ron	cat	sin	tur	slk	rus
slv	srp	sqi	mkd	swe	ita	tat	mkd	tkg	dan
tha	tur	uig	cat	yor	cat	zul	mkd		

3.4.4.2. Active Learning

Active learning (AL) is a potential approach to efficiently improve the performance of machine translation (MT) models by intelligently selecting the most informative data for annotation. Unlike other computer vision tasks, annotation for the MT task is not at data-point level (sentence-level), but at the token-level.

Candidate Sampling: Despite the benefit of having an informative labeled training dataset, the gain of AL can be diminished due to the overhead of computing pseudo labels and estimating acquisition scores for unlabeled data, especially when the unlabeled pool is sufficiently large [36]. Effective candidate selection is pivotal to balance the benefit of AL and the computation overhead. We perform random selection to reduce the size of unlabeled candidate set at each AL round. Additionally, we employ a simple curriculum learning strategy with length bucketing to have models trained on short sentences first. More specifically, we split the unlabeled pool into 5 buckets according to the sentence length $[m/6, m/3, m/2, 3m/4, m]$ where m is the max sentence length. The length bucketing strategy is not only a simple curriculum to have the model learn on easy examples first, i.e., short sentences, but also plays as a heuristic to optimize the number of training sentences in the early checkpoints. In the first checkpoint where the budget is only 5000

tokens, we only sample the unlabeled candidate from the shortest bucket. We then increase the probability of sampling from longer buckets when more annotation budget is available.

Combined AL Strategies: Different AL algorithms perform well at different stages during model training. Uncertainty methods often fall behind other diversity methods in the first few AL iterations as the model is poorly calibrated; hence, the confidence and uncertainty scores are unreliable. While there isn't a one-size-fits-all approach to active learning, combining multiple strategies can yield a more robust labeled dataset. Additionally, we often select data in batch because single instance selection is inefficient, especially when neural network training is often sensitive to small batch size and time-consuming. Simply selecting the top-k uncertainty data points without considering their dependency is sub-optimal and may result in selecting those clustered in a single uncertain area of the decision boundary. To encourage diversity within a batch, we cluster data points based on their gradient embeddings [37] and select an example in each cluster to balance uncertainty and diversity.

We implement the following AL strategies for our NMT system:

- **Maximize lexical coverage strategy** aims to select the sentence which can increase the unique n-gram counts in the labeled dataset the most. The AL acquisition score is computed as how many new unigram, bigram, trigram and four-gram are added to the labeled dataset if the corresponding sentence is labeled. We give higher weight to the phrase with longer length. The weights for unigram, bigram, trigram and four-gram are 0.1, 0.2, 0.3 and 0.4, respectively.
- **Total-token entropy** [38] is the sum of token entropy for each token in the predicted translation. We select sentences with highest total-token entropy to ask for annotation.
- **Data-driven AL** [12], [13] learns an AL query policy using imitation learning. The pool-based AL problem is formalized as a Markov decision process [39], where a state consists of the labeled and unlabeled datasets paired with the parameters of the current NMT model, an action corresponds to the selection of a query data point, and the reward function is the performance gain of the NMT model. The policy network is a two-layer feed-forward network taking state representation as input and outputting the score for the given unlabeled data point. The input to the policy network is the concatenation of
 - the BERT-pooling representation of the candidate sentence and its predicted translation
 - the bag-of-word representation of sentences in the sample pool of unlabeled data and labeled data
 - the entropy statistics of the candidate sentence
 - the n-gram statistics of labeled dataset, unlabeled dataset and the new n-gram score of the candidate sentence
- **BADGE** [37] measures uncertainty as the gradient embedding with respect to parameters in the output (pre-softmax) layer and then chooses an appropriately diverse subset by sampling via k-means++ [40].
- We also **combine multiple AL strategies** by first clustering the candidates to k clusters and applying each above strategy in k/4 clusters.

Practical Consideration: Due to time constraints, we do not retrain the NMT model after each AL round. We only retrain the model at the end of the checkpoint. When retraining, we re-initialize the NMT model from the pretrained model instead of continuing from the best NMT model from the previous checkpoint. As the benefit of AL diminishes when the labeled data gets larger, e.g. >100K sentences, we use random sampling for checkpoint 6-8 to reduce computation overhead.

3.4.4.3. Semi-supervised Learning

Semi-supervised learning (SSL) is an effective learning algorithm to make the most of the available labeled data while also leveraging unlabeled data. In this section, we discuss the semi-supervised learning methods implemented in our NMT system, including back translation and jointly training with unlabeled data.

Back translation (Eval 1) [41] is a simple and effective semi-supervised learning method of NMT. It requires training a reverse NMT model Eng-X which translates from English to the source language. It also assumes the availability of monolingual text in the target language, i.e., English in the LwLL setting. The reverse NMT model provides the translation of target text, resulting in a pseudo-labeled dataset with pseudo-source and clean target text. In the context of LwLL evaluation, we have access to unlabeled training data in the source language and abundant English monolingual data. This English monolingual data tends to be noisy. It is not feasible to back-translate this whole English data. We apply the cross-lingual data selection [8] to get a relevant English text for the base and adaptation stage, then back-translate this relevant English text to the source language. While effective, back-translation incurred substantial inference costs during the evaluation. Therefore, we employ this technique in Evaluation 1, when maximum computation resources are provided to the system (8GPUs - 7 days), and discontinue it in the subsequent evaluations.

SSL with mined bitext and labeled data from related tasks: As back-translation becomes the bottleneck at evaluation time, we prepare mined parallel data at the pretraining time (Section 3.4.3.2) and train our NMT system on this pseudo-parallel data. Additionally, we also make use of the parallel data from the donor language to English for further improving the NMT system at checkpoint 1-4. As the labeled data get larger, the gain from the mined bitext and donor parallel data decrease, we train the NMT system on the clean labeled data only.

3.4.4.4. Domain Adaptation

Unsupervised Domain Adaptation (Eval 1): On the transition from the base stage to the adaptation stage, one can perform unsupervised domain adaptation (UDA) to adapt the NMT model from the source domain (the base task) to the target domain (the adaptation task) without labeled data in the target domain. The common assumption of UDA for NMT is the availability of either non-parallel texts of both languages or only the target-language monolingual text in the new domain to adapt the NMT model. However, in the LwLL evaluation, we only have access to the source-language monolingual text. To address this novel setting, we propose a cross-lingual data selection method to extract in-domain English sentences from the large English corpus. Having a well-aligned shared representation space among languages is the key success of such data retrieval tasks. We propose a contrastive-based alignment method to improve the representation

alignment between two languages and preserve their domain-clustered properties, which are crucial for the selection step. We train a domain classifier to distinguish if a sentence in the source language is from the base or adaptation domain on the aligned representation space. The domain classifier then is used to select relevant English sentences for the base and adaptation domain. We use the selected English text for both back translation (Section 3.4.3) and UDA. For UDA, the NMT model jointly learns to translate with the translation model jointly learns to translate with the translation loss on pseudo bitext, and captures the characteristics of the adaptation domain by the domain discrimination loss on data from the base and adaptation domains.

Active Domain Adaptation: For AL checkpoints in the adaptation phase, we employ the similar combined AL strategy described in Section 3.4.4.2. That is, we implicitly consider the active domain adaptation problem via the AL score estimated by the current NMT model.

Domain Adaptation with Hypernetwork (Eval 3): As described above, our NMT model in the Evaluation 3 is a Transformer with hypernetwork to generate the adapter on-the-fly by conditioning on the source language tag, target language tag and layer ID. To adapt this NMT model to the adaptation task, we add a new source language tag `src_adapt` and target language tag `tgt_adapt` to learn a new adapter for the adaptation task. The embeddings of `src_adapt` and `tgt_adapt` tags are initialized from the current source and target language tags. When the labeled dataset of the adaptation task is limited in the early checkpoints, we continue training the base NMT model on both the base and adaptation parallel data. Additionally, the adaptation parallel data is oversampled to address the imbalance between the base and adaptation data.

3.5 Summary

In this section, we summarize our contributions of each system in LwLL.

Image Classification. Our contributions can be summarized as follows:

- We developed three novel semi-supervised learning methods for Image Classification improving over state-of-the-art with significant margins.
- We showed the potential of integrating curated symbolic knowledge sources (such as knowledge graphs) to the neural methods to improve sample efficiency. We provided one possible implementation via our method SemCo.
- We thoroughly studied the generalizability of self-supervised pretraining compared to the supervised counterpart. Our insights from that study were translated into a novel low-shot transfer method (LAVA) which excels in semi-supervised and few-shot transfer.
- We explored possible extensions of LAVA by leveraging language grounding: by mapping visual representations to a pretrained language model’s semantic space, we improve zero/low-shot transfer capability by virtue of language semantics.

Video Classification. In our system, we introduced a novel method to address the challenges of video classification. Recognizing the cost and difficulty of obtaining labeled video data, we deployed an auxiliary semantic head in the VC network to leverage the features of unlabeled videos effectively. Additionally, we selected an appropriate pretrained dataset and optimized the sampling ratio for video classification, leading to a

substantial reduction in the dependency on labeled videos. Furthermore, we proposed a novel active learning approach that enhances the process of selecting unlabeled instances with sufficiently-distinct features. This approach helps the neural network to learn novel features from videos, thereby enhancing its ability in video classification. Our contributions are outlined as follows:

- Introduction of an auxiliary semantic head in the video classification system which allows the network to utilize the feature of unlabeled videos [6].
- Optimization of pre-trained dataset selection and sampling ratio. Novel active learning strategy that selects unlabeled with sufficiently-distinct features [7].

Object Detection. Due to the blind nature of LwLL's evaluation, we were initially unaware of the exact dataset employed for assessment. To effectively tackle various possibilities, we required a system capable of adaptation to any dataset and varying data quantities and qualities. Our contributions encompassed two key facets. Firstly, we devised a benchmark dataset incorporating data from disparate domains, offering generalizability across all conceivable objects. Secondly, a system was constructed featuring finely tuned parameters and architectural configurations. This system used the benchmark dataset to iteratively optimize these parameters. Subsequently, through systematic parameter tuning, an optimal backbone, detector, and learning method were used. Together, these enhancements fostered a broad applicability spanning the entire range of evaluation datasets.

Machine Translation. In this section, we have presented our approaches to address the challenge of multilinguality and other learning with limited label problems such as active learning and domain adaptation in our submitted NMT system for LwLL evaluation. We have collected monolingual data and further trained the state-of-the-art sequence-to-sequence model to equip the model with the understanding ability of additional languages. Our NMT system is a Transformer with a hyper network that can improve parameter efficiency in training and reduce disk storage by generating weights of the adapter layers on the fly. We combine multiple AL methods to take into account various linguistic and statistical factors of unlabeled data to have a better score for potential data points. Thanks to these approaches and parallel data from donor languages, our NMT system significantly outperforms the baseline in the low-resource checkpoints for all evaluation tasks. Throughout the development of our NMT system, we have made contributions and findings that enhance our understanding of the effective learning strategies for learning with limited labeled data as follows:

- **Active Learning:** Our investigation into AL has revealed the effectiveness of acquisition functions that extend beyond hand-crafted heuristics. We have explored the learning to AL approaches that can benefit from related data to train the AL policy for the target task of interest [12], [13].
- **Data:** When both labeled and unlabeled data can be scarce, we can take advantage of synthetic data generated by generative models or publicly available data from the web. Generative models can synthesize task-specific unlabeled data which can improve performance of self-training, knowledge distillation and few-shot learning [42]. On the other hand, web data tends to be very noisy. Proper data cleaning such as filtering, deduplication are crucial [43].

- **Pretrained models:** Pretrained representations inject key inductive biases for efficient and effective training on downstream tasks. However, it is sensitive to the distribution shift such as domain-shift and language shift. We have addressed this challenge by modifying representation to bridge domain-shift by adversarial training [44]. We propose multilingual representation alignment via contrastive learning to address the language-shift and facilitate the cross-lingual transfer [8].
- **Parameter-efficient training:** Fine-tuning large pretrained language models is computationally expensive. Additionally, the fine-tuned model may overfit on a small training dataset. We have shown that injecting adapters to each Transformer layer, typically one adapter for each language, can enable efficient fine-tuning as the backbone parameters are kept frozen [45]. Additionally, adapters are also useful to bridge the modality gap, for example, speech and text in speech translation [46], or combined to improve model generalization across domains [47]. To support a large number of languages, one can train a hypernetwork to generate adapter weights on-the-fly, hence saving space for model storage.

4. RESULTS AND DISCUSSION

4.1 Image Classification

4.1.1. Development Results

In this section, we present our development results throughout the three phases of the LwLL program. We developed our methods using the LwLL development tasks via the provided API end-point. Specifically, we focused on two tasks: 1) DomainNet Clipart to DomainNet Sketch to verify efficacy of our methods when transferring across different domains with no class shift; 2) DomainNet Real to CIFAR-100 to evaluate the transfer performance when there is a domain and class shift (note that DomainNet Real has higher resolution images than CIFAR-100).

Figure 18 summarizes the performance of our developed methods in the three phases of the programs against the baseline provided by JPL. First, we recap the models we used in each year of the program. In phase 1, we used our method SemCo (section 3.1.3, denoted as EVAL 1 in Figure 18) which combines FixMatch semi-supervised method with our novel co-training label grouping mechanism based on knowledge graph embeddings. In phase 2, we enhanced our model by leveraging self-supervised learning and multi-crop-based pseudo-labeling by introducing our method LAVA (denoted as EVAL 2 in Figure 18). Finally, in the last phase of the program, we improved LAVA by employing language grounding where we learned a functional mapping to ground the visual representations into a shared semantic space as that of a frozen language model to help zero shot transfer (denoted as EVAL 3 in Figure 18).

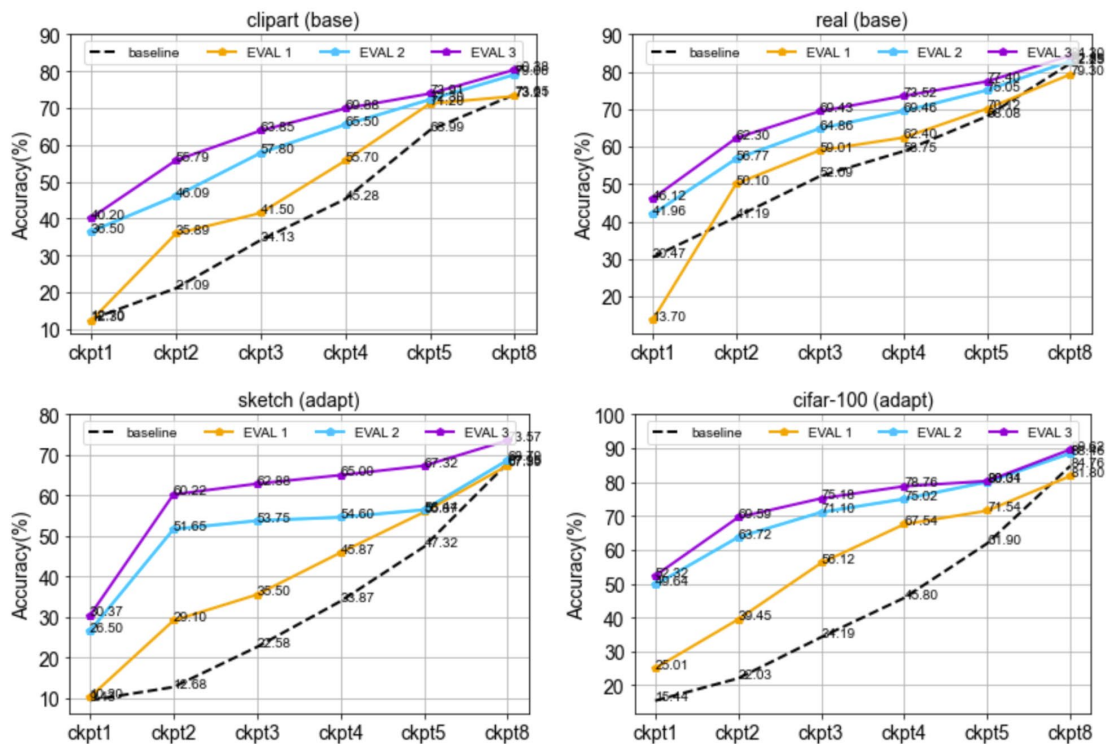


Figure 18 - Development results.

As we observe in Figure 18, our system consistently improves in subsequent years across all the development datasets. Especially in the low-data regime, we observe significant gains in early checkpoints.

4.1.2. Evaluation Results

In this section, we delve into the Evaluation results for our system in each of the three phases. The Image Classification evaluation comprises three tasks to evaluate different aspects of LwLL objectives. Before we present our results, we give a brief description of each of the three tasks.

- **Task 1** uses the *places* dataset which contains images from subset classes that are within the same super-hierarchical class. The base data consists of classes with a large number of images coming from the same superclass (e.g., lecture room, and classroom both coming from superclass “schoolroom”), while the adaptation dataset contains classes which belong to the same superclass but with fewer number of images (e.g., kindergarten room).
- **Task 2** uses the *Hotels50K* dataset to assess the sensor shift scenario. The dataset comprises 1 million images from 50,000 different hotels around the world. Importantly, some of the images are coming from the travel websites (hence it is usually taken with professional cameras and are high resolution), while others come from the travelers’ feedback websites and hence are usually taken by phone cameras and are lower quality. Accordingly, task 2 is set to use the high-res images of given classes in the base stage and the corresponding low-res images of the same classes in the adaptation stage.
- **Task 3** uses the Google Landmarks dataset v1 and v2 which is primarily designed to assess instance recognition and retrieval tasks. The base dataset leverages a subset of the version 1 landmarks classes whereas the adaptation dataset uses a subset of version 2.

4.1.3. Discussion

In phase 1 of the program, there was another task which was discontinued in phase 2 and phase 3 and was replaced by the Google Landmarks task. Hence, we present our figures with dataset names besides Task IDs to avoid confusion.

The results are shown in Figure 19 and Figure 20. Our reflections can be summarized as follows:

- Our systems are better than baseline in almost all cases.
- Our model’s improvement compared to the baseline is more significant in the low-data regime
- We see consistent improvements in subsequent phases of the program.
- Our solution (SemCo) is not well suited for sensor-shift scenarios as evident in the Hotels50K task in phase 1. Unfortunately, due to technical issues, our system did not run on the Hotels task in phase 2 and 3 of the programs and hence we have no reported results.
- Our systems are always among the top 4 performers.
- Self-supervised pretraining significantly helps low-shot learning even more than supervised whitelist pretraining.
- Semantics extracted from language helps improve low-shot learning even without

relying on large-scale models such as CLIP. Simple language embeddings can go a long way.

- Data augmentation, in general, is a useful technique to improve generalization and to avoid overfitting. We show via our method LAVA, that a specific type of augmentations i.e., multi-cropping when combined with pseudo-labeling significantly improves the ability to learn with limited labels.

Future Work. They can be summarized as follows:

- **Neuro-Symbolic Classification via Hybrid Attention Mechanisms:** We aim to improve our current neuro-symbolic method, Semco. In its current state, the symbolic knowledge representations are learned separately (via word and graph embedding techniques), which provides relevant, yet weak, similarity priors. To improve upon this, we are currently working on designing cross-attention mechanisms that dynamically switch between focusing on image features and relevant symbols from knowledge bases. This allows both the image and symbolic representations to be jointly learned aiming to improve the explainability, sample efficiency and robustness of image classification by grounding it with external symbolic knowledge.
- **Multi-Modal Semi-Supervised Learning:** With the huge success of multi-modal text-image foundation models such as CLIP [48], we see a huge potential in to improve semi-supervised learning; we are exploring self-training methods which employ such models to generate highly accurate pseudo-labels for unlabeled target data, minimizing the need for extensive human annotation.

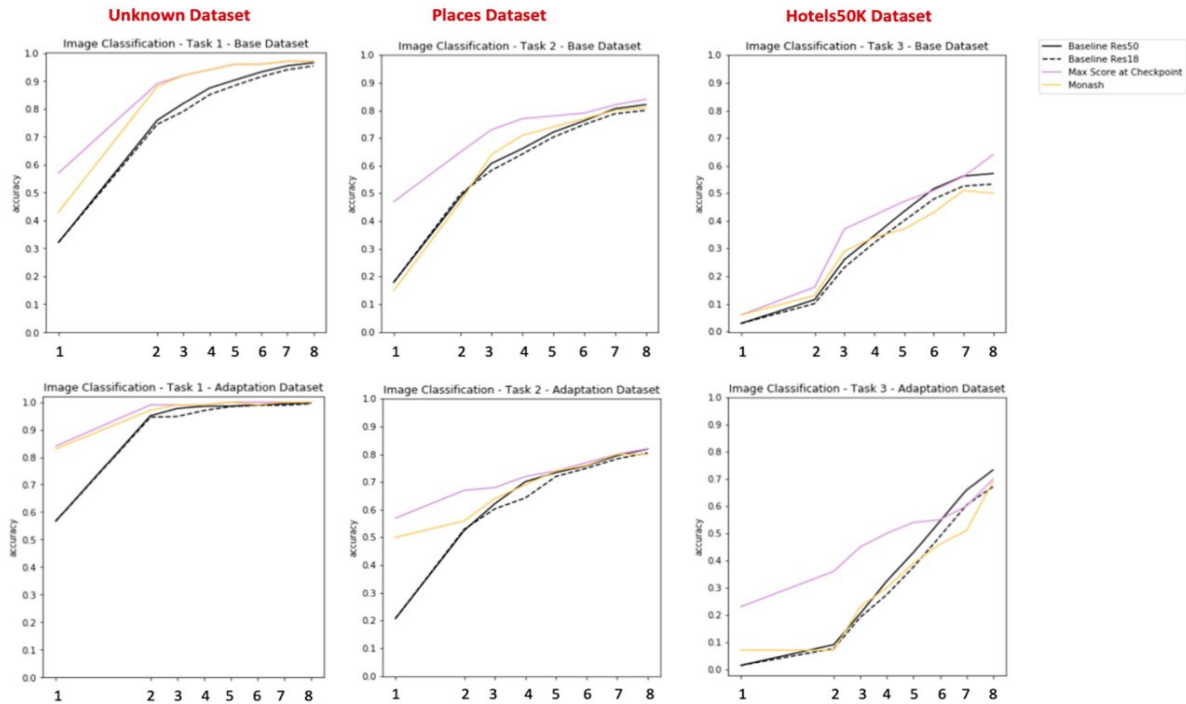


Figure 19 - Our performance in Phase 1 of the program against the baseline and the best performer.

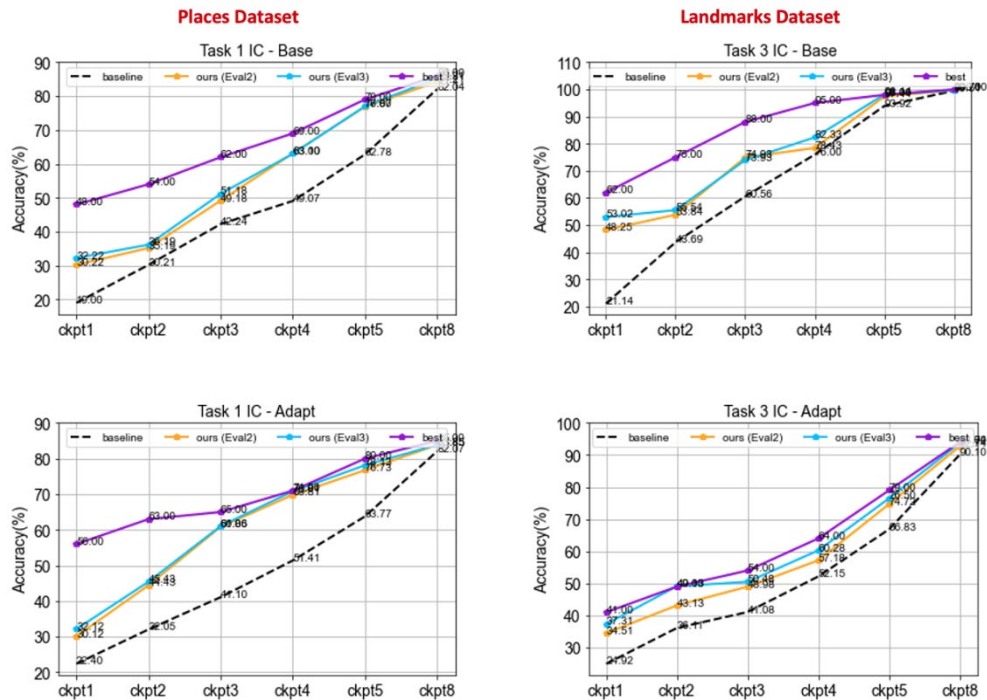


Figure 20 - Our performance in Phase 2 and 3 of the program against the baseline and the best performer.

4.2 Video Classification

4.2.1. Development Results

During development, we use the HDMD and UCF101 datasets to evaluate our method. The accuracy rates for these two datasets are shown in Figure 20. We compare our method with the vanilla MVIT. As shown in Figure 21, our method achieves a significant improvement on both evaluation datasets. At the first two checkpoints, we improve around 3% on both datasets. At the fifth checkpoint, deploying our activate learning method, our method obtains about 2% improvement.

4.2.2. Evaluation Results

For evaluation, the base dataset is EPIC-KITCHENS-55, which contains 432 videos of kitchens involving 125 actions. The adaptation dataset is EPIC-KITCHENS-100, which contains 700 videos and 90K actions. The adaptation dataset is more complex and difficult than the base dataset. It contains more fine-grained actions. Our results are shown in Figure 22 labeled by orange dash lines. We are better than most other teams.

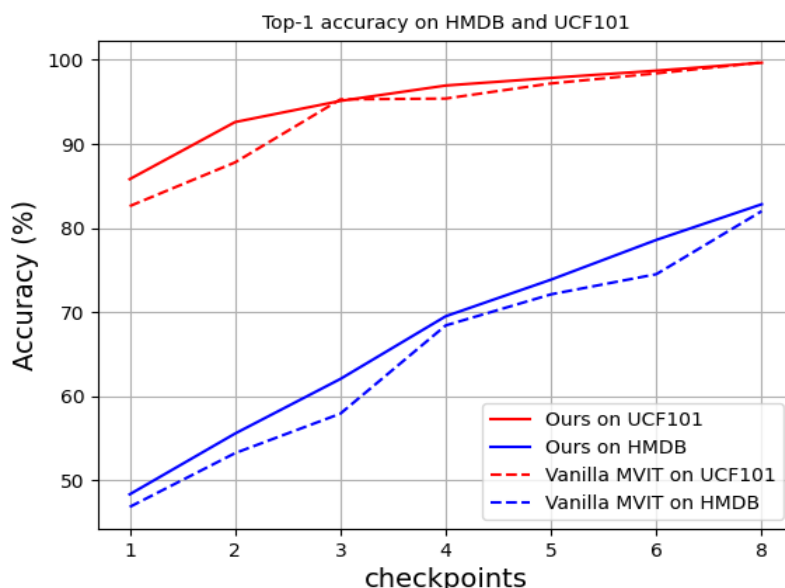


Figure 21 - Performance on HMDB and UCF101.

As shown Figure 22, at the initial checkpoint, where only one labeled video is available per class, we are the second among all participating teams. This result validates our selection of suitable pretraining datasets and the effectiveness of our approach to unlabeled data. At the fifth checkpoint where our active learning method is deployed, our performance soars to the top position on the base dataset. This achievement demonstrates the effectiveness of our active learning approach.

For the third checkpoint, even though we get more labeled data, our performance degrades. This is our shortage on few-labeled datasets.

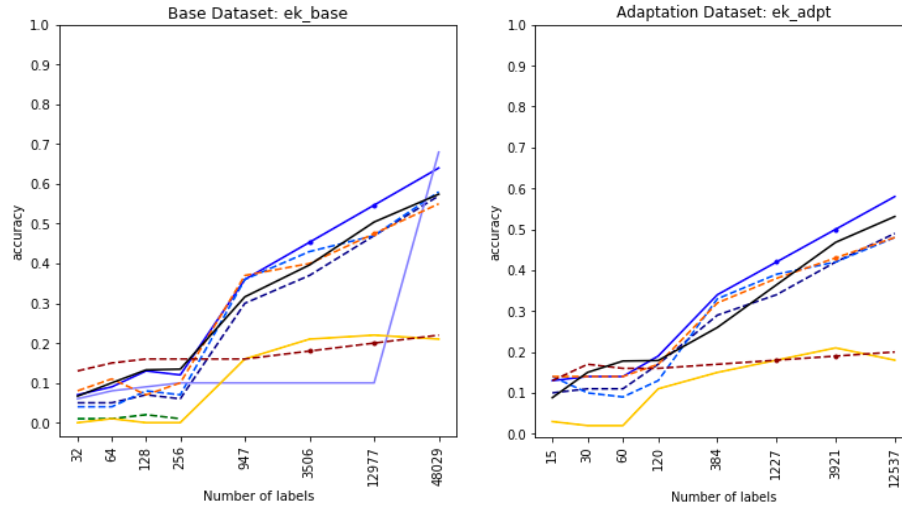


Figure 22 - JPL evaluation results on base dataset and adaptation dataset.

Future work. We propose the following directions:

- Addressing the underperformance of our system on checkpoints with limited labeled data is a priority. We should enhance the system's capability on the few-labeled dataset.
- We should develop one-shot or few-shot approaches to improve our system on the adaptive dataset.

4.3 Object Detection

4.3.1. Development Results

- **Effect of pre-training model.** We used both the COCO and Imagenet pretrained model as an initialization to train all domains except the Satellite Domain. For satellite domain we train Fair1m dataset and using COCO as an initialization then used Fair1m model as an initialization and actually a pretrained model to train satellite datasets. Results indicate that employing the Fair1m model as a pre-trained model led to an enhancement in mean Average Precision (mAP) within the Satellite domain. Moreover, the outcomes demonstrated that utilizing the COCO pre-trained model, when compared to the Imagenet pre-trained model, resulted in an increase in mAP. You can see the results in Figure 23.

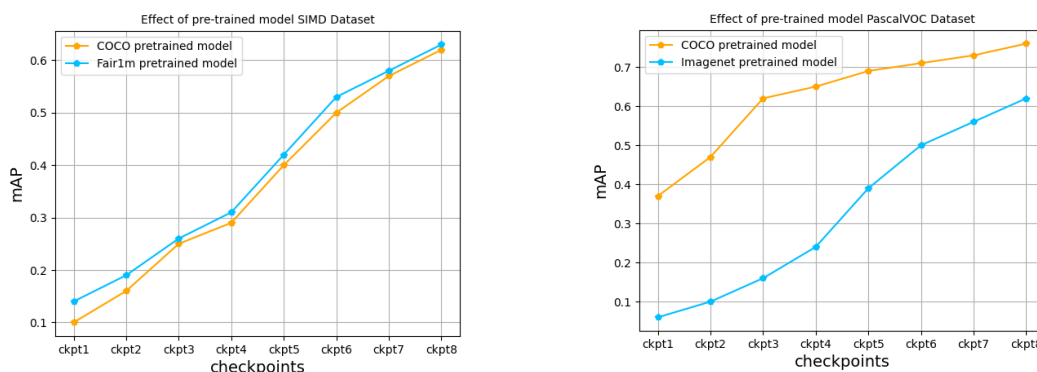


Figure 23 – Effect of pretrained model on two Domains as example.

- **Backbone and detector selection.** We employ the development datasets for both training and testing purposes, aiming to identify the optimal architecture for our system. We opted for FasterRCNN as our fixed detector and employed resnext101 as the backbone of our system. Subsequently, we evaluated our system based on this setup. The outcomes are illustrated in Figure 24. Based on the results which you see in the graphs we use resnext101 as the backbone of our system.
- **Supervised vs Semi-supervised.** We conducted a comparison between supervised and semi-supervised approaches (without self-supervised pretraining) using our development dataset. The graphs in Figure 25 provide a visual representation of this analysis, highlighting the enhancements achieved through the unbiased teacher method. The detector in all of the experiments is FasterRCNN and the backbone is renext-101.

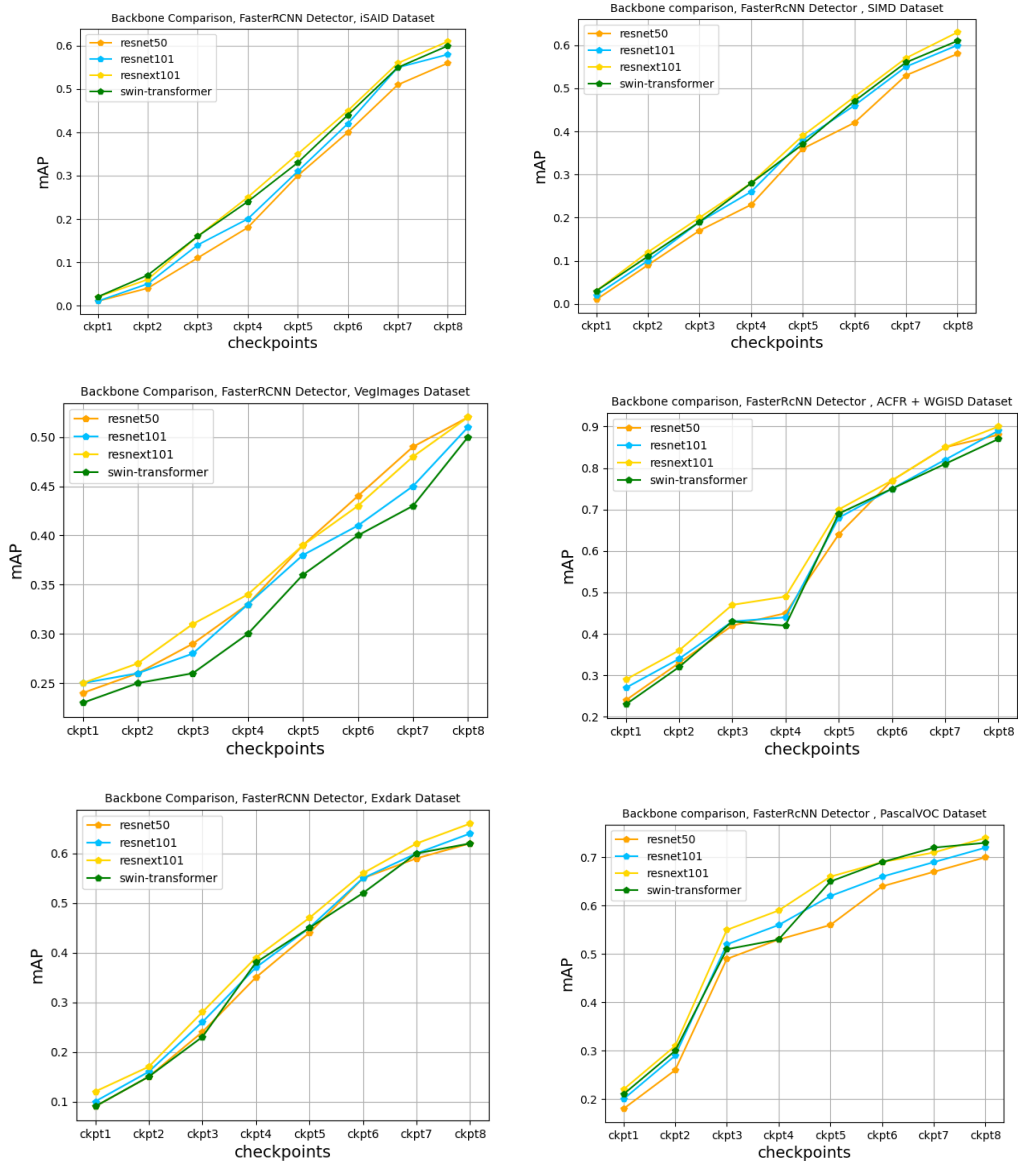


Figure 24 – Comparison between different backbone models.

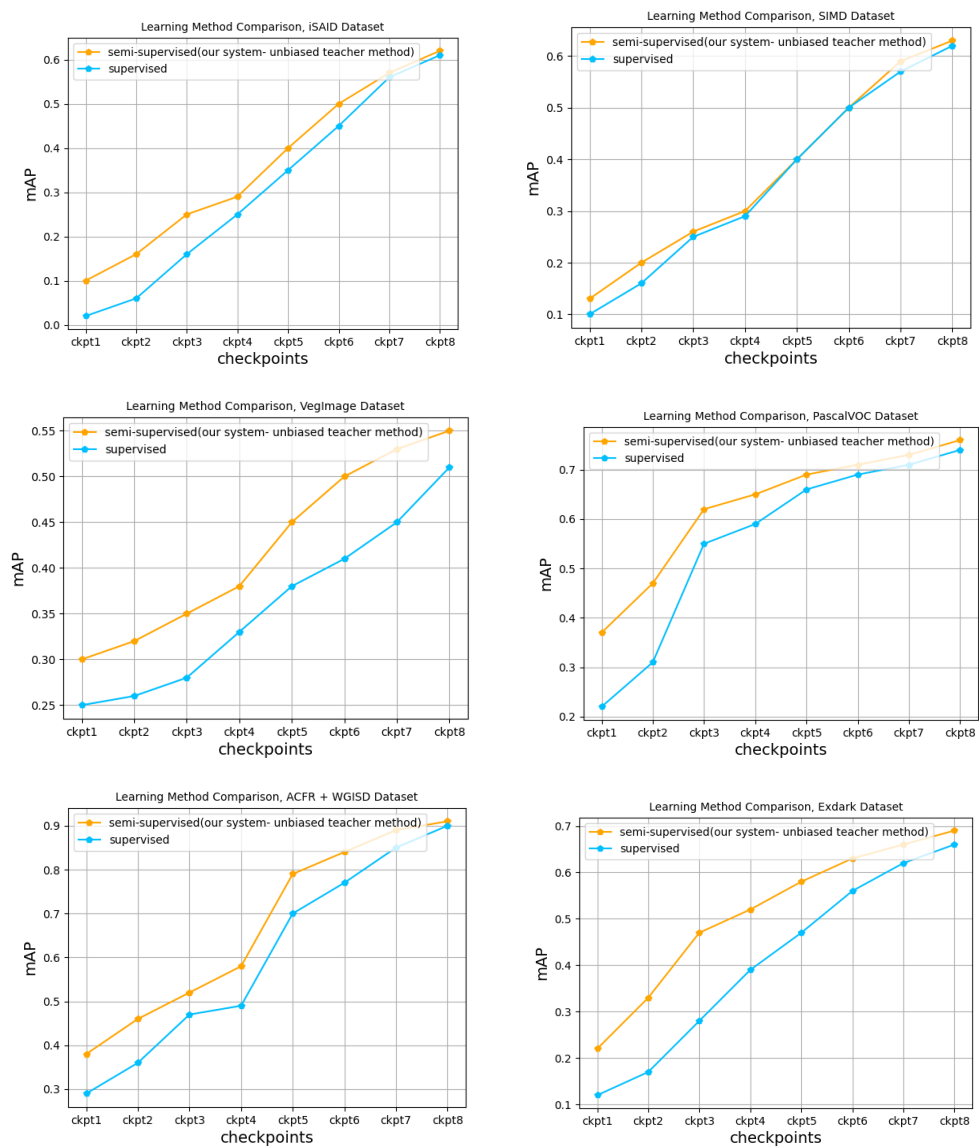
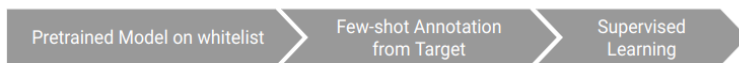


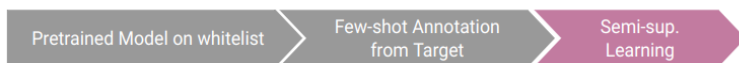
Figure 25 – Supervised vs Semi-supervised Learning Methods.

- Effect of Self-supervised pre-training** We employed a pre-trained model on a whitelist dataset and then used ODISE to generate pseudo labels for unlabeled target (evaluation) dataset. We addressed the following challenges to assess the efficacy of our self-supervised pre-training strategy in enhancing our detection system: 1) The baseline model's training can involve images from a different domain (whitelist) compared to the target (evaluation) dataset domain, 2) The pseudo-labels generated by ODISE for the target (evaluation) data may contain considerable noise. To overcome this, we fine-tune the pre-trained model on the target data using these noisy pseudo labels. This can be seen as an additional preparatory stage before integrating the model into our semi-supervised framework. Once we receive a few labeled data points from the target data (equivalent to Checkpoint 1 of LwLL evals), we integrate our semi-supervised system and train it by combining annotated labels and carefully filtered pseudo labels at each evaluation checkpoint.

Supervised (Baseline):



Semi-Supervised (Our system in **Final** submission):



Self supervised + semi-supervised:

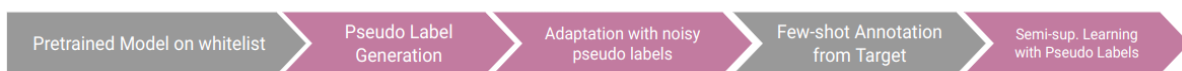


Figure 26 – Brief description of different systems.

Figure 26 presents a schematic illustrating the distinctions between the supervised (baseline), semi-supervised, and self-supervised pre-training+semi-supervised models. Figure 27 illustrates a comparison of the mentioned systems, with their performance evaluated based on their MAP (Mean Average Precision). This visualization is supplemented by a comparison of the outcomes of these three approaches using a satellite image dataset. The results demonstrate that our final model achieved the highest performance. Further in-depth analysis of our benchmark remains a focus for future investigations.

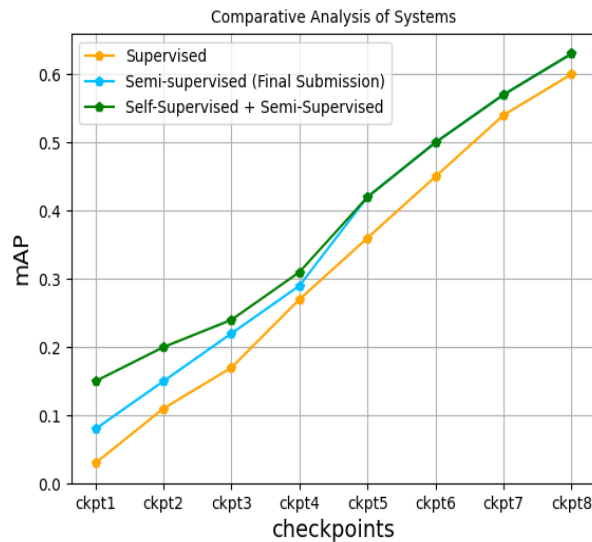


Figure 27 – Comparative analysis of the object detection system.

4.3.2. Evaluation Results

In the subsequent sections, we present the performance outcomes of our system on the most recent JPL evaluation results (Evaluation 3). It's important to note that these results are based on the initial version of our system, wherein the self-supervised pre-training strategy had not yet been incorporated.

- **Task1:** In this task the base dataset is BDD100K- all. The images of the dataset derived from 100K driving videos (40 seconds each) collected from more than 50K rides. The scene types are city streets, residential areas, and highways. Recorded in diverse weather conditions at different times of the day. They have 720p resolution. It has 70,000 images for training, 10,000 images for validation and 20,000 images for testing. The total number of objects are 1.8M. The adaptation dataset is IDD. This Dataset consists of images obtained from a front facing camera attached to a car from drive sequences on Indian roads. It has 10K Images from 182 drive sequences.
- **Task2:** The base dataset in this task is BDD100K-day and the adaptation dataset is BDD 100k-night.
- **Task3:** The base dataset is AU-AIR. It is the first multi-modal UAV dataset for object detection. It has Multi-modal data from different on-board sensors and pushes forward the development of computer vision and robotic algorithms targeted at autonomous aerial surveillance. The images extracted from 2h raw videos. It consists of 32,823 labeled frames and 132,034 object instances. There are 8 object classes such as human, car, van, truck, bike, motorbike, bus and trailer in this dataset. The adaptation dataset is Wilddash2. It is the Dataset and benchmark for panoptic, semantic, and instance segmentation. It has Variety of data sources, with different scenarios (e.g., rain, road coverage, darkness, overexposure) and camera characteristics (noise, compression artifacts, distortion).

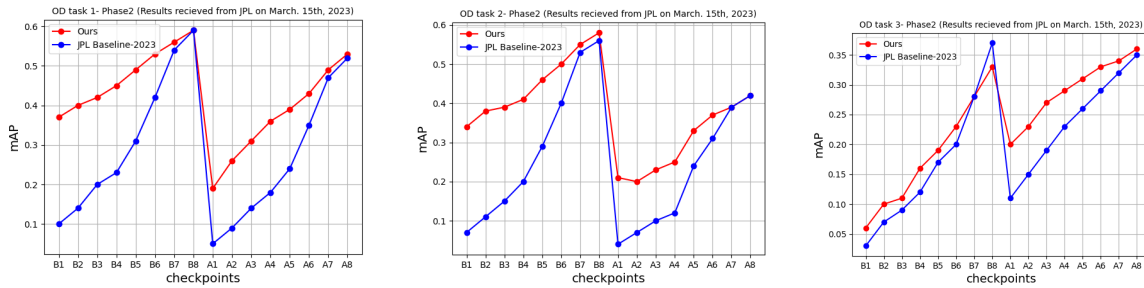


Figure 28 - JPL Evaluation Results.

Figure 28 illustrates a side-by-side comparison between the achievements of our system and the JPL baseline system across all three tasks. Particularly noteworthy is our system's robust performance in the initial checkpoints. Our development findings indicate that employing the semi-supervised learning approach (unbiased teacher) along with our refined backbone and well-tuned hyperparameters contributes to enhanced performance, especially when data availability is limited. Based on the graphs that JPL provided, it's evident that our performance consistently outperforms the baseline. Across all tasks, our systems consistently rank among the top performers.

Future Works. In the future, we could do a few things to make our system better. First, we might include a special type of detector called DETR, which is based on transformers. Second, we could try out a method called Grounding DINO, which enhances object detection by enabling the system to learn from unlabeled data. Third, we could use Active Learning that selects the most informative examples for the system to learn from. Fourth, we could have Transfer Learning. Transfer learning in object detection means using knowledge gained from one task to improve the performance of the system on a different but related task. All these ideas could help us make our system even better!

4.4 Machine Translation

4.4.1. Development Results

There are 3 NMT development tasks which cover 3 different translation directions: Farsi -> English (Fas-Eng), Polish -> English (Pol-Eng), and Sinhala -> English (Sin-Eng). All development tasks have the base task from cc_aligned dataset and the adaptation task from wikimatrix dataset. All tasks are medium size with the resource constraints of 8 GPUs over 4 days.

The baseline at each checkpoint is the mBART large with 610 parameters and trained on a single GPUs for 8 days with the default hyperparameters provided in the Fairseq library.

Figure 29 shows the BLEU score of our NMT system on the development task 3 (Sin-Eng). Our NMT systems significantly outperformed the baseline in the checkpoint 1-4 - low-resource regimes with less than 10K labeled data. When having a reasonable amount of labeled data, our NMT system is slightly lagged behind the baseline.

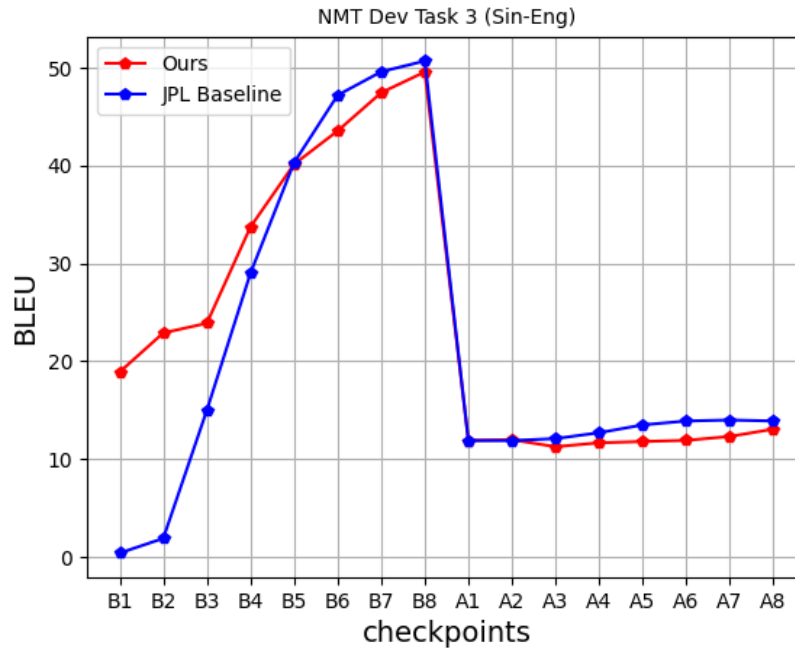


Figure 29 - Our development result on development task 3 (Sin-Eng).

4.4.2. Evaluation Results

Table 5 provides the description of 3 NMT evaluation tasks which covers 3 different translation directions: Portuguese -> English (Por-Eng), Arabic -> English (Ara-Eng) and Hungarian -> English (Hun-Eng). Each task aims to evaluate the ability of NMT models in handling different distribution shifts between the base and adaptation tasks:

Table 5 - NMT evaluation task description.

	Language	Stage	Dataset	Domain	Size (#tokens)	Resources
Task 1	Por-Eng	Base	Paracrawl	general	30M	Medium
		Adaptation	Medline	medical	3.4M	
Task 2	Ara-Eng	Base	MSA news	news	22.6M	Medium
		Adaptation	Egyptian Arabic	forum	2.9M	
Task 3	Hun-Eng	Base	Hunglish 1.0	web 2006	39.8M	Medium
		Adaptation	Paracrawl	web 2021	27.9M	

- Domain shift (Task 1): The base task is in the generic domain and the adaptation task is in the medical domain.
- Shift in dialect and genre (Task 2): The base task is text written in modern standard arabic (MSA) while the adaptation task is written in Egyptian Arabic. There is also a domain shift between the base and adaptation task: news -> web forum.
- Diachronic shift (Task 3): This task aims to evaluate how the NMT models handle languages evolving over time. The base task is text crawled from the web in 2006 while the adaptation task is crawled in 2021.

All the evaluation tasks are medium size with the resource constraints of 8 GPUs over 4 days.

Figure 30 reports the performance of our NMT system in the evaluation 3. Thanks to the further pretraining and the use of parallel data from donor languages, our NMT system achieve a decent translation performance and significantly outperforms the baseline in the first checkpoint with the annotation budget of only 5000 words. However, the benefit of donor languages and hyper-adaptor diminishes as we have more labeled data. Our NMT system lags behind this strong baselines in the high-resource checkpoints (checkpoint 6-8). While having lower performance at the last checkpoint of the base stage, our NMT system achieve competitive BLEU score at the first checkpoint of the adaptation phase. It demonstrates the effectiveness of our unsupervised domain adaptation methods. In conclusion, our NMT systems are very competitive in the low-resource regimes.

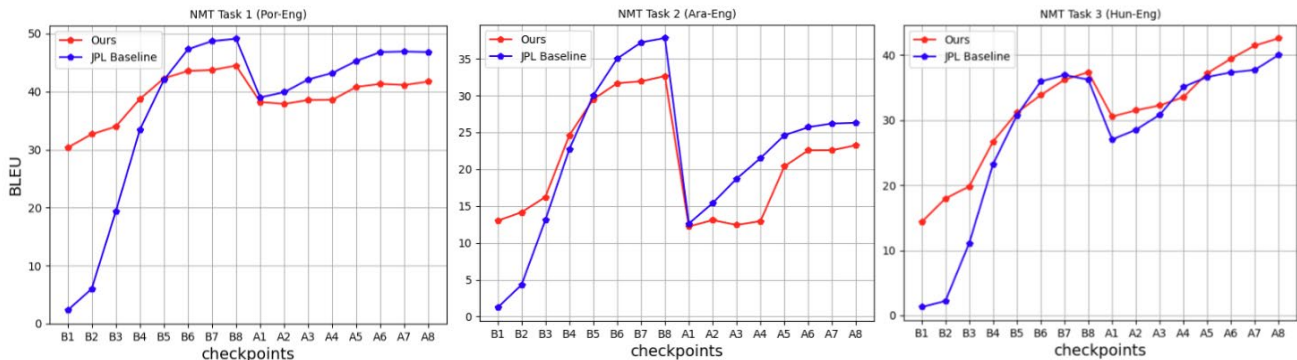


Figure 30 - Performance of our NMT system in the LwLL evaluation 3.

Future Works. Looking forward, our findings lay the groundwork for several future directions:

- We acknowledge the underperformance of parameter-efficient models in high-resource scenarios. Further research should pay attention to enhancing the performance while maintaining its efficiency.
- Additionally, we recognize the evolving landscape of low-resource NMT in the modern generative AI era. Given significant advances in instruction-following agents, for example, chatGPT and GPT4, will LwLL still be a challenge? How to effectively leverage modern generative AI for LwLL.

5 CONCLUSION

5.1 Summary of Contributions

Image Classification. We have developed three novel semi-supervised learning methods for Image Classification improving over state-of-the-art with significant margins. We showed the potential of integrating curated knowledge sources (such as knowledge graphs) to the neural methods to improve sample efficiency. We provided one possible implementation via our method SemCo. We thoroughly studied the generalizability of self-supervised pretraining compared to the supervised counterpart. Our insights from that study was translated into a novel low-shot transfer method (LAVA) which excels in semi-supervised and few-shot transfer. We explored possible extensions of LAVA by leveraging language grounding: by mapping visual representations to a pretrained language model's semantic space, we improve zero/low-shot transfer capability by virtue of language semantics.

Video Classification. We have introduced a novel method to address the challenges of video classification. Recognizing the cost and difficulty of obtaining labeled video data, we deployed an auxiliary semantic head in the VC network to leverage the features of unlabeled videos effectively. Additionally, we selected an appropriate pretrained dataset and optimized the sampling ratio for video classification, leading to a substantial reduction in the dependency on labeled videos. Furthermore, we proposed a novel active learning approach that enhances the process of selecting unlabeled instances with sufficiently-distinct features. This approach helps the neural network to learn novel features from videos, thereby enhancing its ability in video classification.

Object Detection. Due to the blind nature of LwLL's evaluation, we were initially unaware of the exact dataset employed for assessment. To effectively tackle various possibilities, we required a system capable of adaptation to any dataset and varying data quantities and qualities. Our contributions encompassed two key facets. Firstly, we devised a benchmark dataset incorporating data from disparate domains, offering generalizability across all conceivable objects. Secondly, a system was constructed featuring finely tuned parameters and architectural configurations. This system used the benchmark dataset to iteratively optimize these parameters. Subsequently, through systematic parameter tuning, an optimal backbone, detector, and learning method were used. Together, these enhancements fostered a broad applicability spanning the entire range of evaluation datasets.

Machine Translation. We have presented our approaches to address the challenge of multilinguality and other learning with limited label problems such as active learning and domain adaptation in our submitted NMT system for LwLL evaluation. We have collected monolingual data and further trained the state-of-the-art sequence-to-sequence model to equip the model with the understanding ability of additional languages. Our NMT system is a Transformer with a hyper network that can improve parameter efficiency in training and reduce disk storage by generating weights of the adapter layers on the fly. We combine multiple AL methods to take into account various linguistic and statistical factors of unlabeled data to have a better score for potential data points. Thanks to these approaches and parallel data from donor languages, our NMT system significantly outperforms the baseline in the low-resource checkpoints for all evaluation tasks.

5.2 Future Directions

We outline the strategic directions and key research areas for advancing the research program goals, i.e., Learning with limited human supervision. This plan focuses on harnessing recent techniques such as multimodal learning, causality theory, and neuro-symbolic architectures to enhance sample efficiency, robustness, and performance in scenarios with scarce labeled data.

Expanding Capabilities through Multi-Modal Learning. Integrating information from multiple sources, such as text, images, audio, and video, can provide complementary insights. This can be especially valuable when dealing with limited labeled data in one modality, as information from other modalities can help bridge the gap. A direction we are currently exploring is to leverage pre-trained models in one modality to bootstrap learning in another modality so as to improve sample efficiency and generalization. For example, pre-trained large language models can provide context for understanding images or videos with minimal labeled examples in conjunction with pretrained off-the-shelf vision models.

Another direction is Multi-Modal Data Augmentation, whereby the aim is to generate augmented data by combining different modalities to enrich the training dataset. For example, with the eminence of text-to-image generative models (e.g., stable diffusion), one can use text to generate images-of-interest to further train visual models.

Incorporating Causality Theory to Enhance Robustness. Causality theory provides a principled framework for understanding the causal relationships between variables. By integrating causality into deep learning, we can enhance model robustness. For instance, causal inference regularization can help regularize the learning process by making the model less sensitive to noisy or biased human annotations. This can lead to improved generalization and robustness even when labeled data is limited.

Another direction we are pursuing is CounterFactual Learning. By introducing counterfactual reasoning into the training process, we enable the model to reason about alternative scenarios, aiding in handling situations where the available labeled data might be scarce or biased.

Exploring Neuro-Symbolic Approaches for Learning with Limited Labels. Neuro-symbolic architectures combine the uncertainty-aware deep learning methods and the robustness guarantees provided by symbolic reasoning methods, making them particularly well-suited for scenarios with limited labeled data. One method we pursued which falls under that category is SemCo whereby prior symbolic knowledge from knowledge graphs is infused into neural networks to provide valuable constraints and guidance, reducing the need for extensive labeled data. We aim to explore more joint-learning neuro-symbolic methods, whereby knowledge is represented using predicates, logical operators, and quantifiers, and neural networks are self-supervised trained to generate useful representations. Subsequently, by aligning both symbols and representations via a symbolic reasoning module, this approach allows the model to learn complex tasks with limited labeled data by inferring answers based on both neural-encoded context and logical rules, enabling robustness, interpretability, and few-shot learning capabilities.

5.3 Acknowledgments

We would like to thank all members of the Monash/Adelaide team, who worked tirelessly throughout the DARPA LwLL program. This list includes Trang Vu, Samitha Herath, Islam Nassar, Christian Simon, Miao Zhang, Xuanli He, Minghao Wu, Jinming Zhao, Amin Parvaneh, Changlin Lin, Mahsa Ghorbanali, Munawar Hayat, Dinh Phung, Geoff Webb, Jeff Riley, Wray Buntine, Christian Rodriguez, and Anton van den Hengel. This work is supported by DARPA's Learning with Less Labels (LwLL) program under agreement FA8750-19-2-0501.

6 REFERENCES

- [1] T. M. Mitchell, 'The Need for Biases in Learning Generalizations', 1980.
- [2] M. Goldblum, M. Finzi, K. Rowan, and A. G. Wilson, 'The No Free Lunch Theorem, Kolmogorov Complexity, and the Role of Inductive Biases in Machine Learning'. arXiv, Apr. 11, 2023. Accessed: Aug. 18, 2023. [Online]. Available: <http://arxiv.org/abs/2304.05366>
- [3] A. Vaswani *et al.*, 'Attention is All you Need', in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2017. Accessed: Aug. 21, 2023. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html
- [4] I. Nassar, M. Hayat, E. Abbasnejad, H. Rezatofighi, M. Harandi, and G. Haffari, 'LAVA: Label-efficient Visual Learning and Adaptation', in *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, Waikoloa, HI, USA: IEEE, Jan. 2023, pp. 147–156. doi: 10.1109/WACV56688.2023.00023.
- [5] I. Nassar, S. Herath, E. Abbasnejad, W. Buntine, and G. Haffari, 'All Labels Are Not Created Equal: Enhancing Semi-supervision via Label Grouping and Co-training', presented at the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE Computer Society, Jun. 2021, pp. 7237–7246. doi: 10.1109/CVPR46437.2021.00716.
- [6] I. Nassar *et al.*, 'Revisiting Low-shot Action Recognition: A Method and a Benchmark'. CoRR, 2022.
- [7] A. Parvaneh, E. Abbasnejad, D. Teney, R. Haffari, A. Van Den Hengel, and J. Q. Shi, 'Active Learning by Feature Mixing', in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, New Orleans, LA, USA: IEEE, Jun. 2022, pp. 12227–12236. doi: 10.1109/CVPR52688.2022.01192.
- [8] T.-T. Vu, X. He, D. Phung, and G. Haffari, 'Generalised Unsupervised Domain Adaptation of Neural Machine Translation with Cross-Lingual Data Selection', in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 3335–3346. doi: 10.18653/v1/2021.emnlp-main.268.
- [9] C. Baziotis, M. Artetxe, J. Cross, and S. Bhosale, 'Multilingual Machine Translation with Hyper-Adapters', in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 1170–1185. doi: 10.18653/v1/2022.emnlp-main.77.
- [10] Y. Liu *et al.*, 'Multilingual Denoising Pre-training for Neural Machine Translation', *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 726–742, 2020, doi: 10.1162/tacl_a_00343.
- [11] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, 'Bleu: a Method for Automatic Evaluation of Machine Translation', in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, Jul. 2002, pp. 311–318. doi: 10.3115/1073083.1073135.

- [12] M. Liu, W. Buntine, and G. Haffari, 'Learning to Actively Learn Neural Machine Translation', in *Proceedings of the 22nd Conference on Computational Natural Language Learning*, Brussels, Belgium: Association for Computational Linguistics, Oct. 2018, pp. 334–344. doi: 10.18653/v1/K18-1033.
- [13] T.-T. Vu, M. Liu, D. Phung, and G. Haffari, 'Learning How to Active Learn by Dreaming', in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 4091–4101. doi: 10.18653/v1/P19-1401.
- [14] M. Caron *et al.*, 'Emerging Properties in Self-Supervised Vision Transformers', in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, Montreal, QC, Canada: IEEE, Oct. 2021, pp. 9630–9640. doi: 10.1109/ICCV48922.2021.00951.
- [15] S. V. Petrovic, 'A Comparison Between the Silhouette Index and the Davies-Bouldin Index in Labelling IDS Clusters', 2006. Accessed: Sep. 27, 2023. [Online]. Available: <https://www.semanticscholar.org/paper/A-Comparison-Between-the-Silhouette-Index-and-the-Petrovic/b2db00f73fc6b97ebe12e97cfdaefbb2fefc253b>
- [16] R. Speer, J. Chin, and C. Havasi, 'ConceptNet 5.5: an open multilingual graph of general knowledge', in *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, in AAAI'17. San Francisco, California, USA: AAAI Press, Feb. 2017, pp. 4444–4451.
- [17] J. Pennington, R. Socher, and C. Manning, 'GloVe: Global Vectors for Word Representation', in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 1532–1543. doi: 10.3115/v1/D14-1162.
- [18] M. Faruqui, J. Dodge, S. K. Jauhar, C. Dyer, E. Hovy, and N. A. Smith, 'Retrofitting Word Vectors to Semantic Lexicons', in *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Denver, Colorado: Association for Computational Linguistics, May 2015, pp. 1606–1615. doi: 10.3115/v1/N15-1184.
- [19] K. Sohn *et al.*, 'FixMatch: Simplifying Semi-Supervised Learning with Consistency and Confidence', in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2020, pp. 596–608. Accessed: Aug. 21, 2023. [Online]. Available: <https://papers.nips.cc/paper/2020/hash/06964dce9addb1c5cb5d6e3d9838f733-Abstract.html>
- [20] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, 'Unsupervised Learning of Visual Features by Contrasting Cluster Assignments', in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2020, pp. 9912–9924. Accessed: Aug. 21, 2023. [Online]. Available: <https://proceedings.neurips.cc/paper/2020/hash/70feb62b69f16e0238f741fab228fec2-Abstract.html>
- [21] H. Fan *et al.*, 'Multiscale Vision Transformers', in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, Montreal, QC, Canada: IEEE, Oct. 2021, pp. 6804–6815. doi: 10.1109/ICCV48922.2021.00675.
- [22] J. Xu, S. Liu, A. Vahdat, W. Byeon, X. Wang, and S. D. Mello, 'Open-Vocabulary Panoptic Segmentation With Text-to-Image Diffusion Models', in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 2955–2966.

- [23] Y.-C. Liu *et al.*, ‘UNBIASED TEACHER FOR SEMI-SUPERVISED OBJECT DETECTION’, in *International Conference on Learning Representation*, 2021.
- [24] P. J. O. Suárez, B. Sagot, and L. Romary, ‘Asynchronous Pipeline for Processing Huge Corpora on Medium to Low Resource Infrastructures’, presented at the 7th Workshop on the Challenges in the Management of Large Corpora, 2019.
- [25] P. Joshi, S. Santy, A. Budhiraja, K. Bali, and M. Choudhury, ‘The State and Fate of Linguistic Diversity and Inclusion in the NLP World’, in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online: Association for Computational Linguistics, Jul. 2020, pp. 6282–6293. doi: 10.18653/v1/2020.acl-main.560.
- [26] A. Conneau *et al.*, ‘Unsupervised Cross-lingual Representation Learning at Scale’, in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online: Association for Computational Linguistics, Jul. 2020, pp. 8440–8451. doi: 10.18653/v1/2020.acl-main.747.
- [27] D. Ha, A. M. Dai, and Q. V. Le, ‘HyperNetworks’, presented at the International Conference on Learning Representations, Oct. 2016. Accessed: Aug. 21, 2023. [Online]. Available: <https://openreview.net/forum?id=rkpACe1lx>
- [28] N. Houlsby *et al.*, ‘Parameter-Efficient Transfer Learning for NLP’, in *Proceedings of the 36th International Conference on Machine Learning*, PMLR, May 2019, pp. 2790–2799. Accessed: Aug. 21, 2023. [Online]. Available: <https://proceedings.mlr.press/v97/houlsby19a.html>
- [29] J. Tiedemann, ‘Parallel Data, Tools and Interfaces in OPUS’, in *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC’12)*, Istanbul, Turkey: European Language Resources Association (ELRA), May 2012, pp. 2214–2218. Accessed: Aug. 21, 2023. [Online]. Available: http://www.lrec-conf.org/proceedings/lrec2012/pdf/463_Paper.pdf
- [30] NLLB Team *et al.*, ‘No Language Left Behind: Scaling Human-Centered Machine Translation’, no. arXiv:2207.04672. arXiv, Aug. 25, 2022. doi: 10.48550/arXiv.2207.04672.
- [31] K. Heffernan, O. Çelebi, and H. Schwenk, ‘Bitext Mining Using Distilled Sentence Representations for Low-Resource Languages’, in *Findings of the Association for Computational Linguistics: EMNLP 2022*, Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 2101–2112. doi: 10.18653/v1/2022.findings-emnlp.154.
- [32] N. Arivazhagan *et al.*, ‘Massively Multilingual Neural Machine Translation in the Wild: Findings and Challenges’. arXiv, Jul. 11, 2019. doi: 10.48550/arXiv.1907.05019.
- [33] P. Li, L. Li, M. Zhang, M. Wu, and Q. Liu, ‘Universal Conditional Masked Language Pre-training for Neural Machine Translation’, in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 6379–6391. doi: 10.18653/v1/2022.acl-long.442.
- [34] M. Ott *et al.*, ‘fairseq: A Fast, Extensible Toolkit for Sequence Modeling’, in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 48–53. doi: 10.18653/v1/N19-4009.

- [35] M. Xia, A. Anastasopoulos, R. Xu, Y. Yang, and G. Neubig, 'Predicting Performance for Natural Language Processing Tasks', in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online: Association for Computational Linguistics, Jul. 2020, pp. 8625–8646. doi: 10.18653/v1/2020.acl-main.764.
- [36] G. Citovsky *et al.*, 'Batch Active Learning at Scale', presented at the Advances in Neural Information Processing Systems, Nov. 2021. Accessed: Aug. 21, 2023. [Online]. Available: <https://openreview.net/forum?id=zzdf0CirJM4>
- [37] J. T. Ash, C. Zhang, and A. Krishnamurthy, 'DEEP BATCH ACTIVE LEARNING BY DIVERSE, UNCERTAIN GRADIENT LOWER BOUNDS', 2020.
- [38] B. Settles and M. Craven, 'An Analysis of Active Learning Strategies for Sequence Labeling Tasks', in *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, Honolulu, Hawaii: Association for Computational Linguistics, Oct. 2008, pp. 1070–1079. Accessed: Aug. 21, 2023. [Online]. Available: <https://aclanthology.org/D08-1112>
- [39] P. Bachman, A. Sordoni, and A. Trischler, 'Learning Algorithms for Active Learning', in *Proceedings of the 34th International Conference on Machine Learning*, PMLR, Jul. 2017, pp. 301–310. Accessed: Aug. 21, 2023. [Online]. Available: <https://proceedings.mlr.press/v70/bachman17a.html>
- [40] D. Arthur and S. Vassilvitskii, 'k-means++: the advantages of careful seeding', in *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, in SODA '07. USA: Society for Industrial and Applied Mathematics, Jan. 2007, pp. 1027–1035.
- [41] R. Sennrich, B. Haddow, and A. Birch, 'Improving Neural Machine Translation Models with Monolingual Data', in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 86–96. doi: 10.18653/v1/P16-1009.
- [42] X. He, I. Nassar, J. Kiros, G. Haffari, and M. Norouzi, 'Generate, Annotate, and Learn: NLP with Synthetic Text', *Transactions of the Association for Computational Linguistics*, vol. 10, pp. 826–842, 2022, doi: 10.1162/tacl_a_00492.
- [43] T.-T. Vu, X. He, G. Haffari, and E. Shareghi, 'Koala: An Index for Quantifying Overlaps with Pre-training Corpora'. arXiv, Mar. 26, 2023. doi: 10.48550/arXiv.2303.14770.
- [44] T.-T. Vu, D. Phung, and G. Haffari, 'Effective Unsupervised Domain Adaptation with Adversarially Trained Language Models', in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online: Association for Computational Linguistics, Nov. 2020, pp. 6163–6173. doi: 10.18653/v1/2020.emnlp-main.497.
- [45] J. Zhao, H. Yang, G. Haffari, and E. Shareghi, 'RedApt: An Adaptor for wav2vec 2 EncodingFaster and Smaller Speech Translation without Quality Compromise', in *Findings of the Association for Computational Linguistics: EMNLP 2022*, Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 1960–1967. doi: 10.18653/v1/2022.findings-emnlp.142.
- [46] J. Zhao, H. Yang, G. Haffari, and E. Shareghi, 'M-Adapter: Modality Adaptation for End-to-End Speech-to-Text Translation', in *Interspeech 2022*, ISCA, Sep. 2022, pp.

111–115. doi: 10.21437/Interspeech.2022-592.

- [47] T.-T. Vu, S. Khadivi, D. Phung, and G. Haffari, ‘Domain Generalisation of NMT: Fusing Adapters with Leave-One-Domain-Out Training’, in *Findings of the Association for Computational Linguistics: ACL 2022*, Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 582–588. doi: 10.18653/v1/2022.findings-acl.49.
- [48] A. Radford *et al.*, ‘Learning Transferable Visual Models From Natural Language Supervision’, in *Proceedings of the 38th International Conference on Machine Learning*, PMLR, Jul. 2021, pp. 8748–8763. Accessed: Aug. 28, 2023. [Online]. Available: <https://proceedings.mlr.press/v139/radford21a.html>

APPENDIX A LIST OF PAPERS ACCORDING TO THE AREAS

A.1 Vision

A.1.1 Few Shot Learning

- Z. Gao, Y. Wu, X. Fan, M. Harandi, and Y. Jia, 'Learning to Optimize on Riemannian Manifolds', *IEEE Trans Pattern Anal Mach Intell*, vol. 45, no. 5, pp. 5935–5952, May 2023, doi: [10.1109/TPAMI.2022.3215702](https://doi.org/10.1109/TPAMI.2022.3215702).
- M. Hiller, R. Ma, M. Harandi, and T. Drummond, 'Rethinking Generalization in Few-Shot Classification', *Advances in Neural Information Processing Systems*, vol. 35, pp. 3582–3595, Dec. 2022.
- R. Ma, P. Fang, T. Drummond, and M. Harandi, 'Adaptive Poincaré Point to Set Distance for Few-Shot Classification', *AAAI*, vol. 36, no. 2, pp. 1926–1934, Jun. 2022, doi: [10.1609/aaai.v36i2.20087](https://doi.org/10.1609/aaai.v36i2.20087).
- C. Simon, P. Koniusz, and M. Harandi, 'Meta-Learning for Multi-Label Few-Shot Classification', in *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, Waikoloa, HI, USA: IEEE, Jan. 2022, pp. 346–355. doi: [10.1109/WACV51458.2022.00042](https://doi.org/10.1109/WACV51458.2022.00042).
- Cheraghian *et al.*, 'Synthesized Feature based Few-Shot Class-Incremental Learning on a Mixture of Subspaces', in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2021, pp. 8641–8650. doi: [10.1109/ICCV48922.2021.00854](https://doi.org/10.1109/ICCV48922.2021.00854).
- Z. Gao, Y. Wu, Y. Jia, and M. Harandi, 'Curvature Generation in Curved Spaces for Few-Shot Learning', presented at the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 8691–8700. Accessed: Sep. 29, 2023. [Online]. Available: https://openaccess.thecvf.com/content/ICCV2021/html/Gao_Curvature_Generation_in_Curved_Spaces_for_Few-Shot_Learning_ICCV_2021_paper.html
- J. Hong *et al.*, 'Reinforced Attention for Few-Shot Learning and Beyond', presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 913–923. Accessed: Sep. 29, 2023. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2021/html/Hong_Reinforced_Attention_for_Few-Shot_Learning_and_Beyond_CVPR_2021_paper.html

A.1.2 Semi-supervised Learning

- I. Nassar, M. Hayat, E. Abbasnejad, H. Rezatofighi, M. Harandi, and G. Haffari, "LAVA: Label-efficient Visual Learning and Adaptation," in *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, Waikoloa, HI, USA: IEEE, Jan. 2023, pp. 147–156. doi: [10.1109/WACV56688.2023.00023](https://doi.org/10.1109/WACV56688.2023.00023).
- I. Nassar, M. Hayat, E. Abbasnejad, H. Rezatofighi, and G. Haffari, "Protocon: Pseudo-Label Refinement via Online Clustering and Prototypical Consistency for Efficient Semi-Supervised Learning," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Vancouver, BC, Canada:

IEEE, Jun. 2023, pp. 11641–11650. doi: [10.1109/CVPR52729.2023.01120](https://doi.org/10.1109/CVPR52729.2023.01120).

- I. Nassar, S. Herath, E. Abbasnejad, W. Buntine, and G. Haffari, “All Labels Are Not Created Equal: Enhancing Semi-supervision via Label Grouping and Co-training,” presented at the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE Computer Society, Jun. 2021, pp. 7237–7246. doi: [10.1109/CVPR46437.2021.00716](https://doi.org/10.1109/CVPR46437.2021.00716).
- I. Nassar *et al.*, “Revisiting Low-shot Action Recognition: A Method and a Benchmark.” CoRR, 2022.

A.1.3 Active Learning

- W. Tan, L. Du, and W. Buntine, “Diversity Enhanced Active Learning with Strictly Proper Scoring Rules,” in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2021, pp. 10906–10918. Accessed: Sep. 30, 2023. [Online]. Available: <https://proceedings.neurips.cc/paper/2021/hash/5a7b238ba0f6502e5d6be14424b20ded-Abstract.html>
- A. Parvaneh, E. Abbasnejad, D. Teney, R. Haffari, A. Van Den Hengel, and J. Q. Shi, “Active Learning by Feature Mixing,” in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, New Orleans, LA, USA: IEEE, Jun. 2022, pp. 12227–12236. doi: [10.1109/CVPR52688.2022.01192](https://doi.org/10.1109/CVPR52688.2022.01192).
- P. Munjal, N. Hayat, M. Hayat, J. Sourati, and S. Khan, “Towards Robust and Reproducible Active Learning using Neural Networks,” in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, New Orleans, LA, USA: IEEE, Jun. 2022, pp. 223–232. doi: [10.1109/CVPR52688.2022.00032](https://doi.org/10.1109/CVPR52688.2022.00032).

A.1.4 Domain Adaptation

- X. Liu, B. Hu, X. Liu, J. Lu, J. You, and L. Kong, “Energy-constrained Self-training for Unsupervised Domain Adaptation.” arXiv, Jan. 01, 2021. doi: [10.48550/arXiv.2101.00316](https://doi.org/10.48550/arXiv.2101.00316).
- C. Simon *et al.*, “On Generalizing Beyond Domains in Cross-Domain Continual Learning,” in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, New Orleans, LA, USA: IEEE, Jun. 2022, pp. 9255–9264. doi: [10.1109/CVPR52688.2022.00905](https://doi.org/10.1109/CVPR52688.2022.00905).
- C. S, H. Z, H. M, and Y. X, “Domain Neural Adaptation,” *IEEE transactions on neural networks and learning systems*, vol. PP, Aug. 2022, doi: [10.1109/TNNLS.2022.3151683](https://doi.org/10.1109/TNNLS.2022.3151683).
- T. Le, T. Nguyen, N. Ho, H. Bui, and D. Phung, “LAMDA: Label Matching Deep Domain Adaptation,” in *Proceedings of the 38th International Conference on Machine Learning*, PMLR, Jul. 2021, pp. 6043–6054. Accessed: Sep. 30, 2023. [Online]. Available: <https://proceedings.mlr.press/v139/le21a.html>

A.1.5 Others

- N. Rosa, T. Drummond, and M. Harandi, “A Differentiable Distance Approximation for Fairer Image Classification,” in *Computer Vision – ACCV*

2022, vol. 13846, L. Wang, J. Gall, T.-J. Chin, I. Sato, and R. Chellappa, Eds., in Lecture Notes in Computer Science, vol. 13846. , Cham: Springer Nature Switzerland, 2023, pp. 218–234. doi: [10.1007/978-3-031-26351-4_14](https://doi.org/10.1007/978-3-031-26351-4_14).

- X. Cheng *et al.*, “Implicit Motion Handling for Video Camouflaged Object Detection,” in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, New Orleans, LA, USA: IEEE, Jun. 2022, pp. 13854–13863. doi: [10.1109/CVPR52688.2022.01349](https://doi.org/10.1109/CVPR52688.2022.01349).
- Z. Gao, Y. Wu, Y. Jia, and M. Harandi, “Hyperbolic Feature Augmentation via Distribution Estimation and Infinite Sampling on Manifolds,” *Advances in Neural Information Processing Systems*, vol. 35, pp. 34421–34435, Dec. 2022.
- D. Teney, E. Abbasnejad, S. Lucey, and A. Van Den Hengel, “Evading the Simplicity Bias: Training a Diverse Set of Models Discovers Solutions with Superior OOD Generalization,” in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, New Orleans, LA, USA: IEEE, Jun. 2022, pp. 16740–16751. doi: [10.1109/CVPR52688.2022.01626](https://doi.org/10.1109/CVPR52688.2022.01626).
- S. Xue, R. Wang, B. Zhang, T. Wang, G. Guo, and D. Doermann, “IDARTS: Interactive Differentiable Architecture Search,” in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, Montreal, QC, Canada: IEEE, Oct. 2021, pp. 1143–1152. doi: [10.1109/ICCV48922.2021.00120](https://doi.org/10.1109/ICCV48922.2021.00120).
- L. V. Jospin, H. Laga, F. Boussaid, W. Buntine, and M. Bennamoun, “Hands-On Bayesian Neural Networks—A Tutorial for Deep Learning Users,” *IEEE Comput. Intell. Mag.*, vol. 17, no. 2, pp. 29–48, May 2022, doi: [10.1109/MCI.2022.3155327](https://doi.org/10.1109/MCI.2022.3155327).
- C. Simon, P. Koniusz, and M. Harandi, “On Learning the Geodesic Path for Incremental Learning,” in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Nashville, TN, USA: IEEE, Jun. 2021, pp. 1591–1600. doi: [10.1109/CVPR46437.2021.00164](https://doi.org/10.1109/CVPR46437.2021.00164).
- M. Zhang, H. Li, S. Pan, X. Chang, and S. Su, “Overcoming Multi-Model Forgetting in One-Shot NAS With Diversity Maximization,” presented at the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE Computer Society, Jun. 2020, pp. 7806–7815. doi: [10.1109/CVPR42600.2020.00783](https://doi.org/10.1109/CVPR42600.2020.00783).
- C. Li *et al.*, “Block-Wisely Supervised Neural Architecture Search With Knowledge Distillation,” in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA: IEEE, Jun. 2020, pp. 1986–1995. doi: [10.1109/CVPR42600.2020.00206](https://doi.org/10.1109/CVPR42600.2020.00206).

A.2 Language

A.2.1 Machine Translation/NLP tasks

- T.-T. Vu, S. Khadivi, D. Phung, and G. Haffari, “Domain Generalisation of NMT: Fusing Adapters with Leave-One-Domain-Out Training,” in *Findings of the Association for Computational Linguistics: ACL 2022*, Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 582–588. doi:

[10.18653/v1/2022.findings-acl.49](https://doi.org/10.18653/v1/2022.findings-acl.49).

- H. Yang, J. Zhao, G. Haffari, and E. Shareghi, “Self-supervised Rewiring of Pre-trained Speech Encoders: Towards Faster Fine-tuning with Less Labels in Speech Processing,” in *Findings of the Association for Computational Linguistics: EMNLP 2022*, Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 1952–1959. doi: [10.18653/v1/2022.findings-emnlp.141](https://doi.org/10.18653/v1/2022.findings-emnlp.141).
- M. Wu, Y. Li, M. Zhang, L. Li, G. Haffari, and Q. Liu, “Uncertainty-Aware Balancing for Multilingual and Multi-Domain Neural Machine Translation Training,” in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 7291–7305. doi: [10.18653/v1/2021.emnlp-main.580](https://doi.org/10.18653/v1/2021.emnlp-main.580).
- F. Saleh, W. Buntine, G. Haffari, and L. Du, “Multilingual Neural Machine Translation: Can Linguistic Hierarchies Help?,” in *Findings of the Association for Computational Linguistics: EMNLP 2021*, Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 1313–1330. doi: [10.18653/v1/2021.findings-emnlp.114](https://doi.org/10.18653/v1/2021.findings-emnlp.114).
- D. Kanojia, R. Dabre, S. Dewangan, P. Bhattacharyya, G. Haffari, and M. Kulkarni, “Harnessing Cross-lingual Features to Improve Cognate Detection for Low-resource Languages,” in *Proceedings of the 28th International Conference on Computational Linguistics*, Barcelona, Spain (Online): International Committee on Computational Linguistics, Dec. 2020, pp. 1384–1395. doi: [10.18653/v1/2020.coling-main.119](https://doi.org/10.18653/v1/2020.coling-main.119).
- F. Saleh, W. Buntine, and G. Haffari, “Collective Wisdom: Improving Low-resource Neural Machine Translation using Adaptive Knowledge Distillation,” in *Proceedings of the 28th International Conference on Computational Linguistics*, Barcelona, Spain (Online): International Committee on Computational Linguistics, Dec. 2020, pp. 3413–3421. doi: [10.18653/v1/2020.coling-main.302](https://doi.org/10.18653/v1/2020.coling-main.302).
- X. He, G. Haffari, and M. Norouzi, “Dynamic Programming Encoding for Subword Segmentation in Neural Machine Translation,” in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online: Association for Computational Linguistics, Jul. 2020, pp. 3042–3051. doi: [10.18653/v1/2020.acl-main.275](https://doi.org/10.18653/v1/2020.acl-main.275).

A.2.2 Transfer Learning: Domain Adaptation

- J. Zhao, H. Yang, G. Haffari, and E. Shareghi, “RedApt: An Adaptor for wav2vec 2 Encoding Faster and Smaller Speech Translation without Quality Compromise,” in *Findings of the Association for Computational Linguistics: EMNLP 2022*, Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 1960–1967. doi: [10.18653/v1/2022.findings-emnlp.142](https://doi.org/10.18653/v1/2022.findings-emnlp.142).
- Z. Li, L. Qu, and G. Haffari, “Total Recall: a Customized Continual Learning Method for Neural Semantic Parsers,” in *Proceedings of the 2021 Conference*

on *Empirical Methods in Natural Language Processing*, Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 3816–3831. doi: [10.18653/v1/2021.emnlp-main.310](https://doi.org/10.18653/v1/2021.emnlp-main.310).

- T.-T. Vu, X. He, D. Phung, and G. Haffari, “Generalised Unsupervised Domain Adaptation of Neural Machine Translation with Cross-Lingual Data Selection,” in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 3335–3346. doi: [10.18653/v1/2021.emnlp-main.268](https://doi.org/10.18653/v1/2021.emnlp-main.268).
- X. Guo, Y.-F. Li, and G. Haffari, “Understanding Unnatural Questions Improves Reasoning over Text,” in *Proceedings of the 28th International Conference on Computational Linguistics*, Barcelona, Spain (Online): International Committee on Computational Linguistics, Dec. 2020, pp. 4949–4955. doi: [10.18653/v1/2020.coling-main.434](https://doi.org/10.18653/v1/2020.coling-main.434).
- T.-T. Vu, D. Phung, and G. Haffari, “Effective Unsupervised Domain Adaptation with Adversarially Trained Language Models,” in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online: Association for Computational Linguistics, Nov. 2020, pp. 6163–6173. doi: [10.18653/v1/2020.emnlp-main.497](https://doi.org/10.18653/v1/2020.emnlp-main.497).
- T. Feng, L. Qu, and G. Haffari, “Less is More: Mitigate Spurious Correlations for Open-Domain Dialogue Response Generation Models by Causal Discovery,” *Transactions of the Association for Computational Linguistics*, vol. 11, pp. 511–530, 2023, doi: [10.1162/tacl_a_00561](https://doi.org/10.1162/tacl_a_00561).
- T.-T. Vu, S. Khadivi, D. Phung, and G. Haffari, “Domain Generalisation of NMT: Fusing Adapters with Leave-One-Domain-Out Training,” in *Findings of the Association for Computational Linguistics: ACL 2022*, Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 582–588. doi: [10.18653/v1/2022.findings-acl.49](https://doi.org/10.18653/v1/2022.findings-acl.49).

A.2.3 Semi-supervised Learning

- L. Luo, Y.-F. Li, G. Haffari, and S. Pan, “Normalizing Flow-based Neural Process for Few-Shot Knowledge Graph Completion,” in *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, in SIGIR ’23. New York, NY, USA: Association for Computing Machinery, Jul. 2023, pp. 900–910. doi: [10.1145/3539618.3591743](https://doi.org/10.1145/3539618.3591743).
- F. Shiri, T. Wu, Y. Li, and G. Haffari, “TCG-Event: Effective Task Conditioning for Generation-based Event Extraction,” in *Proceedings of the The 20th Annual Workshop of the Australasian Language Technology Association*, Adelaide, Australia: Australasian Language Technology Association, Dec. 2022, pp. 22–30. Accessed: Sep. 30, 2023. [Online]. Available: <https://aclanthology.org/2022.alt-1.4>
- Z. Li, L. Qu, Q. Xu, T. Wu, T. Zhan, and G. Haffari, “Variational Autoencoder with Disentanglement Priors for Low-Resource Task-Specific Natural Language Generation,” in *Proceedings of the 2022 Conference on Empirical Methods in*

Natural Language Processing, Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 10335–10356. doi: [10.18653/v1/2022.emnlp-main.706](https://doi.org/10.18653/v1/2022.emnlp-main.706).

- X. He, I. Nassar, J. Kiros, G. Haffari, and M. Norouzi, “Generate, Annotate, and Learn: NLP with Synthetic Text,” *Transactions of the Association for Computational Linguistics*, vol. 10, pp. 826–842, 2022, doi: [10.1162/tacl_a_00492](https://doi.org/10.1162/tacl_a_00492).
- Z. Li, L. Qu, S. Huang, and G. Haffari, “Few-Shot Semantic Parsing for New Predicates,” in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, Online: Association for Computational Linguistics, Apr. 2021, pp. 1281–1291. doi: [10.18653/v1/2021.eacl-main.109](https://doi.org/10.18653/v1/2021.eacl-main.109).
- T. Y. Zhuo et al., “On Robustness of Prompt-based Semantic Parsing with Large Pre-trained Language Model: An Empirical Study on Codex,” in *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, Dubrovnik, Croatia: Association for Computational Linguistics, May 2023, pp. 1090–1102. doi: [10.18653/v1/2023.eacl-main.77](https://doi.org/10.18653/v1/2023.eacl-main.77).

A.2.4 Active Learning

- T.-T. Vu, M. Liu, D. Phung, and G. Haffari, “Learning How to Active Learn by Dreaming,” in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 4091–4101. doi: [10.18653/v1/P19-1401](https://doi.org/10.18653/v1/P19-1401).

A.2.5 Multi-task Learning

- P. Zaremoondi and G. Haffari, “Learning to Multi-Task Learn for Better Neural Machine Translation.” arXiv, Jan. 09, 2020. doi: [10.48550/arXiv.2001.03294](https://doi.org/10.48550/arXiv.2001.03294).
- P. Zaremoondi and G. Haffari, “Adaptively Scheduled Multitask Learning: The Case of Low-Resource Neural Machine Translation,” in *Proceedings of the 3rd Workshop on Neural Generation and Translation*, Hong Kong: Association for Computational Linguistics, Nov. 2019, pp. 177–186. doi: [10.18653/v1/D19-5618](https://doi.org/10.18653/v1/D19-5618).

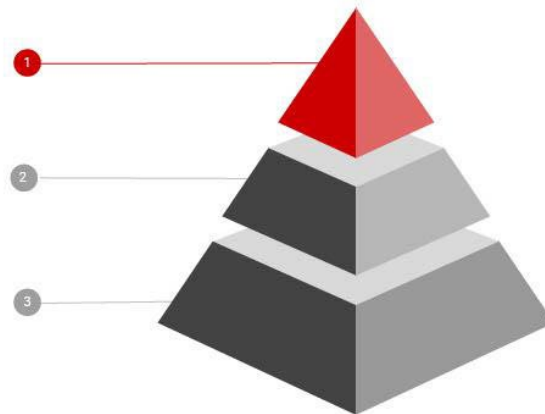
A.2.6 Other

- F. Moghimifar, L. Qu, Y. Zhuo, M. Baktashmotlagh, and G. Haffari, “CosMo: Conditional Seq2Seq-based Mixture Model for Zero-Shot Commonsense Question Answering,” in *Proceedings of the 28th International Conference on Computational Linguistics*, Barcelona, Spain (Online): International Committee on Computational Linguistics, Dec. 2020, pp. 5347–5359. doi: [10.18653/v1/2020.coling-main.467](https://doi.org/10.18653/v1/2020.coling-main.467).
- F. Moghimifar, G. Haffari, and M. Baktashmotlagh, “Domain Adaptive Causality Encoder,” in *Proceedings of the The 18th Annual Workshop of the Australasian Language Technology Association*, Virtual Workshop: Australasian Language Technology Association, Dec. 2020, pp. 1–10. Accessed: Sep. 30, 2023. [Online]. Available: <https://aclanthology.org/2020.alt-1.1>

- P. Arthur, D. Ryu, and G. Haffari, “Multilingual Simultaneous Neural Machine Translation,” in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, Online: Association for Computational Linguistics, Aug. 2021, pp. 4758–4766. doi: [10.18653/v1/2021.findings-acl.420](https://doi.org/10.18653/v1/2021.findings-acl.420).
- J. Zhao, P. Arthur, G. Haffari, T. Cohn, and E. Shareghi, “It Is Not As Good As You Think! Evaluating Simultaneous Machine Translation on Interpretation Data,” in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 6707–6715. doi: [10.18653/v1/2021.emnlp-main.537](https://doi.org/10.18653/v1/2021.emnlp-main.537).
- J. Zhao, H. Yang, G. Haffari, and E. Shareghi, “M-Adapter: Modality Adaptation for End-to-End Speech-to-Text Translation,” in *Interspeech 2022*, ISCA, Sep. 2022, pp. 111–115. doi: [10.21437/Interspeech.2022-592](https://doi.org/10.21437/Interspeech.2022-592).
- N. Askarian, E. Abbasnejad, I. Zukerman, W. Buntine, and G. Haffari, “Inductive Biases for Low Data VQA: A Data Augmentation Approach,” in *2022 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW)*, Waikoloa, HI, USA: IEEE, Jan. 2022, pp. 231–240. doi: [10.1109/WACVW54805.2022.00029](https://doi.org/10.1109/WACVW54805.2022.00029).
- T. Wu *et al.*, “Towards relation extraction from speech,” in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 10751–10762. doi: [10.18653/v1/2022.emnlp-main.738](https://doi.org/10.18653/v1/2022.emnlp-main.738).
- J. Zhao, G. Haffari, and E. Shareghi, “Generating Synthetic Speech from SpokenVocab for Speech Translation,” in *Findings of the Association for Computational Linguistics: EACL 2023*, Dubrovnik, Croatia: Association for Computational Linguistics, May 2023, pp. 1975–1981. doi: [10.18653/v1/2023.findings-eacl.147](https://doi.org/10.18653/v1/2023.findings-eacl.147).
- H. Xu, L. Liu, and E. Abbasnejad, “Progressive Class Semantic Matching for Semi-supervised Text Classification,” in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Seattle, United States: Association for Computational Linguistics, Jul. 2022, pp. 3003–3013. doi: [10.18653/v1/2022.naacl-main.219](https://doi.org/10.18653/v1/2022.naacl-main.219).

Outline

- Executive Summary
- Reflections and Future Outlook
- Our Approaches to LwLL



Executive Summary

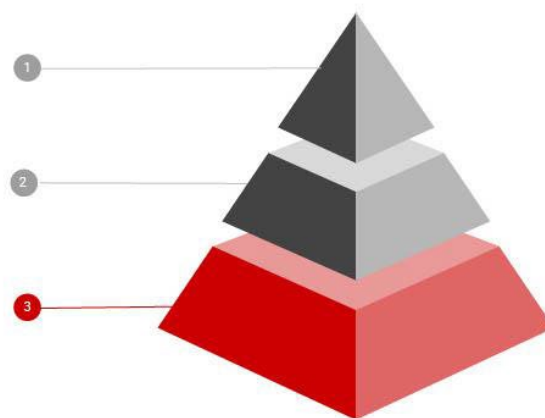
- Rich set of publications/codebase on LwLL in Vision and NLP
 - Papers:
 - i. Language: 10 EMNLP, 4 ACL, 4 COLING, 2 TACL, 2 EACL, 1 SIGIR, 1 INTERSPEECH, 1 NAACL
 - ii. Vision: 16 CVPR, 4 NeurIPS, 4 ICCV, 2 ECCV, 4 WACV, 3 TPAMI, 2 ICLR, 1 JCAI, 1 ACCV, 1 ICML
 - Codebase/Software:
 - i. Language: 20 github repos
 - ii. Vision: 16 github repos
- Strong systems submitted for LwLL evaluations
 - Object detection NGA transition task
- Training researchers
 - Trained 4 postdocs and 8 PhD students

Executive Summary

- Workshops on “Deep Learning for Low-resource NLP” (DeepLo)
 - Collocated with NAACL 2022
- Transition to Knowledge/Technology to other DARPA projects
 - In CCU, only 20% is labeled and the rest 80% is unlabeled
- Application to projects with Defence Science & Technology (DST), Australia
 - Maritime surveillance and security, based on incident reports/news articles/intelligence reports

Outline

- Executive Summary
- Reflections and Future Outlook
- Our Approaches to LwLL



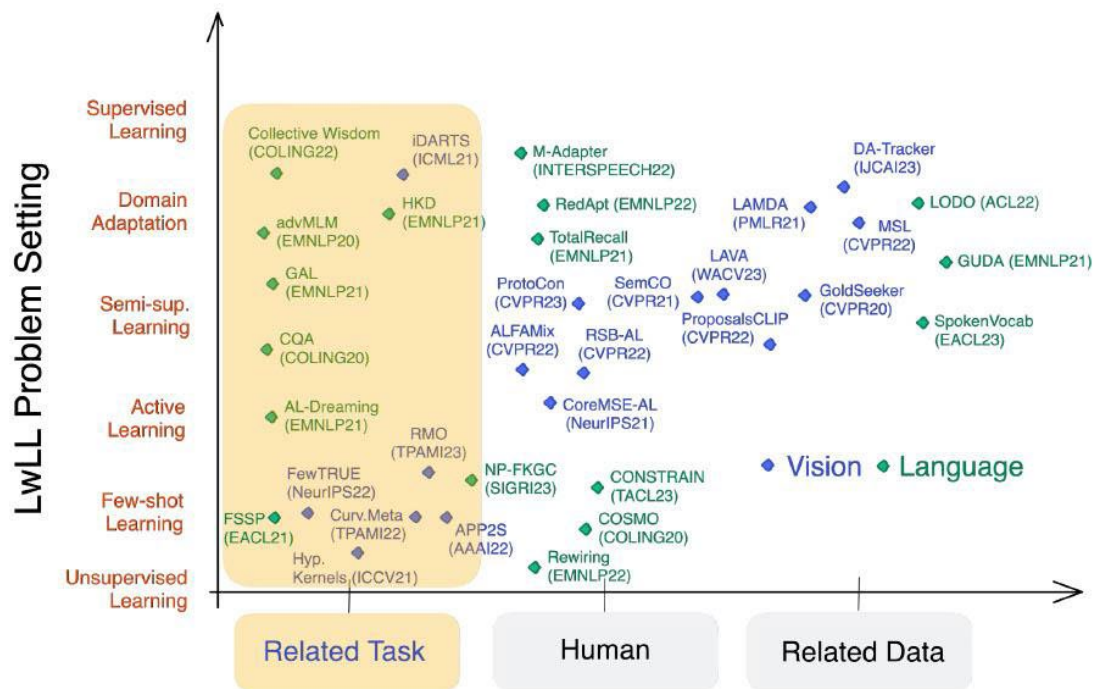
Our Approach to LwLL

- Achieve sample efficiency by injecting **inductive biases**
- **Identify** such inductive biases applicable to LwLL settings

Related Task

Human

Related Data



Meta-Learning

- Enables models to learn from related tasks and quickly adapt to new tasks with limited data
- Challenges
 - **Limited** availability of **effective related tasks**
 - **Limitations** in **meta-optimization** algorithms
- Our Approaches
 - **Synthesize related tasks** that can help the learning to learn, eg Learning to Actively Learn [1]
 - **Improved meta-optimization** algorithms with **Riemannian** meta-optimizers [2] for constrained optimization

[1] Yu, Liu, Phung, Haffari. "Learning how to active learn by dreaming". ACL2019

[2] Gao, Wu, Fan, Harandi, Jia. "Learning to Optimize on Riemannian Manifolds". TPAMI2022

Multitask Learning

- To **learn** inductive biases from related **auxiliary** tasks **simultaneously**
- Challenges
 - Effective architectures for **knowledge sharing** among tasks
 - Effective **training strategy** of the MTL architecture
- Our Approaches:
 - **Training scheduler to adaptively** identify tasks relevant to the main task [1,2]
 - Encourages/Prevents positive/negative transfer
 - **Adaptive knowledge sharing** via router blocks [3]

[1] Zareemoodi, Haffari. "Adaptively scheduled multitask learning: The case of low-resource neural machine translation." NTG2019.

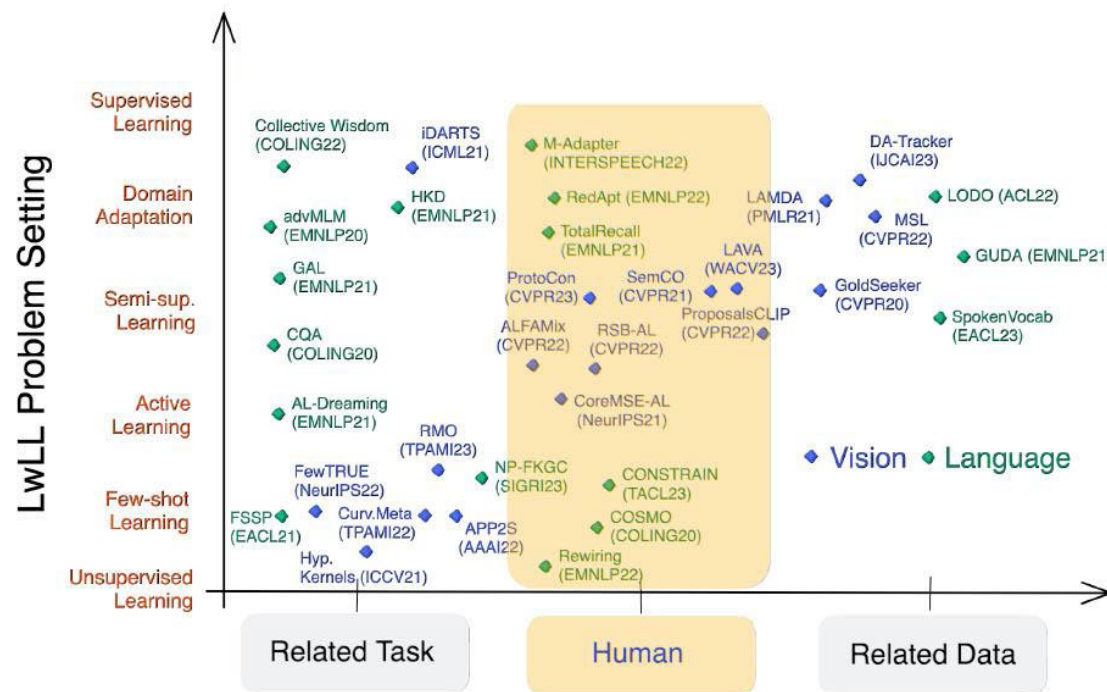
[2] Zareemoodi, Haffari. "Learning to multi-task learn for better neural machine translation." (preprint) 2020.

[3] Zareemoodi, Buntine, Haffari. "Adaptive Knowledge Sharing in Multitask Learning: Improving Low-Resource Neural Machine translation", ACL 2018.

Supervision from Teacher Models

- Transferring knowledge/features from teacher to student networks, hence reducing the need of labeled data
- Challenge: Effectively learning from **multiple teachers**?
 - They may have different strengths
 - They maybe behind the firewall
- Our Approaches
 - **Collective Knowledge Distillation (KD) [1]:** Dynamically adjust the contribution of the teacher models during distillation process
 - **Selective KD [2]:** Leveraging hierarchical teachers
 - **Imitation Attack [3]:** To learn from teacher models behind firewalls

[1] Saleh, Buntine, Haffari, "Collective Wisdom: Improving Low-resource Neural Machine Translation using Adaptive Knowledge Distillation", COLING2020
 [2] Saleh, Du, Haffari, Buntine, "Multilingual Neural Machine Translation: Can Linguistic Hierarchies Help?", EMNLP Findings 2021
 [3] Xu, He, Lyu, Qu, Haffari, "Student Surpasses Teacher: Imitation Attack for Black-Box NLP APIs." COLING2022.



Supervision from Human

- Mostly studied under active learning
- Challenges
 - Uncertainty-Diversity trade-off
 - Effective AL acquisition functions beyond hand-crafted heuristics
- Our approaches
 - We show *AL-heuristics* are not effective when strong regularization is used in sup. learning [1]
 - AL approaches beyond heuristics that can benefit from related data [2] and data geometry [4]
 - Theoretical analysis of uncertainty estimates [3]
 - Make informed judgments about uncertainty-diversity trade-off for batch AL

[1] Munjal, Hayat, [Hayat](#), Sourati, Khan. "Towards Robust and Reproducible Active Learning Using Neural Networks". CVPR2022

[2] [Yu](#), Liu, [Phung](#), [Haffari](#). "Learning how to active learn by dreaming". ACL2019

[3] Tan, Du, [Buntine](#). "Diversity Enhanced Active Learning with Strictly Proper Scoring Rules". NeurIPS2021

[4] [Parvaneh](#), [Abbasnejad](#), Teney, [Haffari](#), van den Hengel, Qinfeng Shi. "Active Learning by Feature Mixing". CVPR2022

Geometric Priors

- We identify **geometric priors** as structural assumptions applied on:
 - **Data manifold** [5, 6, 7]
 - **Model parameter** [1]
 - **Learnt representations** [2, 3, 4]to inject extra knowledge to reduce the amount of supervision
- Challenges
 - **Identification** of the appropriate geometry
 - **Efficient ways to realize/use** such geometries

[1] Gao, Wu, Fan, [Harandi](#), Jia. "Learning to Optimize on Riemannian Manifolds". TPAMI2022

[2] Ma, Fang, Drummond, [Harandi](#). "Adaptive Poincaré Point to Set Distance for Few-Shot Classification". AAAI2022

[3] [Simon](#), Koniusz, [Harandi](#). "On Learning the Geodesic Path for Incremental Learning without Forgetting". CVPR2021

[4] Cheraghian, Rahman, Ramasinghe, Fang, [Simon](#), Petersson, [Harandi](#). "Synthesized Feature Based Few-Shot Class-Incremental Learning on a Mixture of Subspaces". ICCV2021

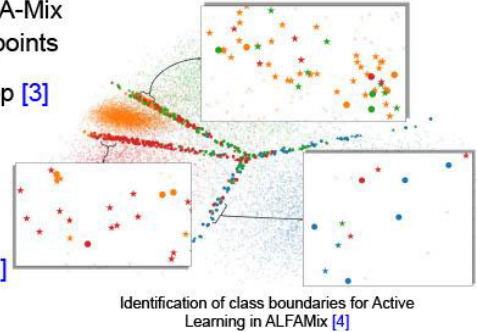
[5] [Parvaneh](#), [Abbasnejad](#), Teney, [Haffari](#), van den Hengel, Qinfeng Shi. "Active Learning by Feature Mixing". CVPR 2022

[6] [Nassar](#), [Hayat](#), [Abbasnejad](#), [Rezatofoghi](#), [Harandi](#), [Haffari](#). "LAVA: A Label-resilient Language-guided Visual Learner". WACV2023

[7] [He](#), [Nassar](#), Kiros, [Haffari](#), Norouzi. "Generate, Annotate, and Learn: NLP with Synthetic Text". TACL2022.

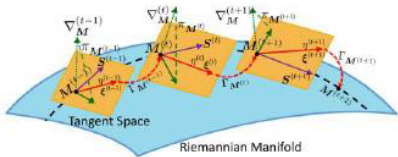
Geometric Priors

- Using **data manifold smoothness**:
 - Improved **prototypical spaces** (PROTOCON [1], ALFA-Mix [4]) to model the data manifold with few labelled data points
 - Improved image augmentation methods *e.g.*, multi-crop [3]
 - Data synthesis techniques [5]
- Our developments show performance gains in:
 - Pseudo labelling for Semi-supervised learning [1,2,3,5]
 - Effective uncertainty estimate for Active Learning [4]



- [1] Nassar, Hayat, Abbasnejad, Rezatofighi, Haffari. "ProtoCon: Pseudo-label Refinement via Online Clustering and Prototypical Consistency for Efficient Semi-supervised Learning". CVPR2023
- [2] Nassar, Herath, Abbasnejad, Buntine, Haffari. "All Labels Are Not Created Equal: Enhancing Semi-supervision via Label Grouping and Co-training". CVPR2021
- [3] Nassar, Hayat, Abbasnejad, Rezatofighi, Harandi, Haffari. "LAVA: A Label-resilient Language-guided Visual Learner". WACV2023
- [4] Parvaneh, Abbasnejad, Teney, Haffari, van den Hengel, Qinfeng Shi. "Active Learning by Feature Mixing". CVPR2022
- [5] He, Nassar, Kiros, Haffari, Norouzi. "Generate, Annotate, and Learn: NLP with Synthetic Text". TACL2022.

Geometric Priors



An illustration of optimization with our Riemannian optimizer [1].

- Effective ways to realize **geometric constraints** & add inductive bias to a **model parameters**
 - Riemannian optimization [1]: To realize *Orthogonality* and *Positive-definiteness*
- Geometry for learning **representations**
 - **Hyperbolic embedding spaces** [2]: Hierarchical embeddings of structured data for few-shot learning
 - **Subspace assumptions** [3, 4]: Sample-efficient low dimensional representation learning

- [1] Gao, Wu, Fan, Harandi, Jia. "Learning to Optimize on Riemannian Manifolds." TPAMI2022
- [2] Ma, Fang, Drummond, Harandi. "Adaptive Poincaré Point to Set Distance for Few-Shot Classification". AAAI2022
- [3] Simon, Koniusz, Harandi. "On Learning the Geodesic Path for Incremental Learning without Forgetting." CVPR2021
- [4] Cheraghian, Rahman, Ramasinghe, Fang, Simon, Petersson, Harandi. "Synthesized Feature Based Few-Shot Class-Incremental Learning on a Mixture of Subspaces". ICCV2021

Casual Learning for LwLL

- Current ML requires large datasets because we need to **learn associations** of frequent patterns
- Just associations are not sufficient for **strong generalisation** [1]
 - *Knowing causal (in)dependencies leads to better transferability, generalisation, and data efficiency*
- Using causal reasoning, we can **target, analyse, design and understand** where the biases are needed [2,3,4]



Car
Invariant to background
color, orientation, speed

[1] Bengio. "From system 1 deep learning to system 2 deep learning." NeurIPS 2019.

[2] Abbasnejad, Teney, Parvaneh, Shi, can den Hengel, "Counterfactual vision and language learning", CVPR 2020.

[3] Teney, Abbasnejad, van den Hengel, "Unshuffling data for improved generalization in visual question answering", ICCV 2021

[4] Zhang, Dong, Liu, Abbasnejad, Gong, Zhang, Shi, "Truncated Matrix Power Iteration for Differentiable DAG Learning", NeurIPS 2022.

Casual Learning for LwLL

- Identifying direct causes of responses in open-domain conversations [1]
 - Mitigate spurious correlations and improve relevance, informativeness and fluency
- Disentangled priors help to separate contents and labels in the text [2]
 - Combining content representation from seen tasks with novel class representations
 - Synthetic labelled data generation for style transfer and data augmentation
- First use of counterfactual representations for vision-and-language tasks [4]
- Learning the causal structure of the dependencies between contents [3,5]

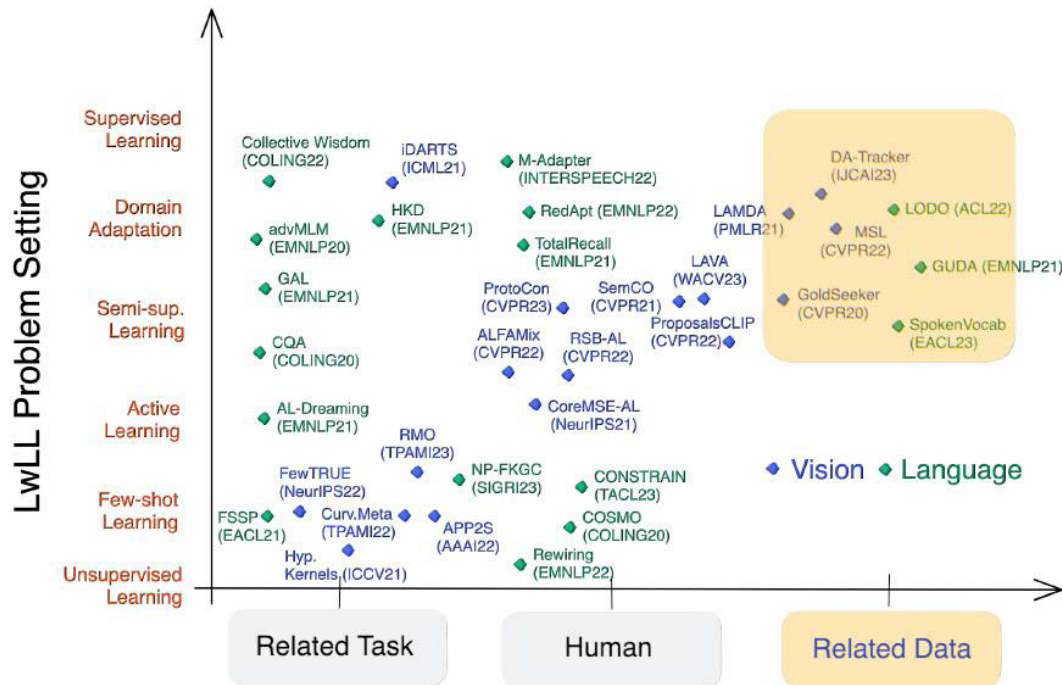
[1] Feng, Lizhen, Haffari, "Less is More: Mitigate Spurious Correlations for Open-Domain Dialogue Response Generation Models by Causal Discovery". TACL2023

[2] Li, Qu, Xu, Zhan, Haffari, "Variational Autoencoders with Disentanglement Priors for Low-Resource Task-specific Natural Language Generation." EMNLP2022

[3] Teney, Abbasnejad, Lucey, van den Hengel, "Evading the simplicity bias: Training a diverse set of models discovers solutions with superior ood generalization", CVPR 2022.

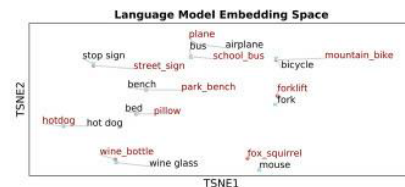
[5] Abbasnejad, Teney, Parvaneh, Shi, can den Hengel, "Counterfactual vision and language learning", CVPR 2020.

[4] Zhang, Dong, Liu, Abbasnejad, Gong, Zhang, Shi, "Truncated Matrix Power Iteration for Differentiable DAG Learning", NeurIPS 2023.



Supervision with Extra Modalities

- Related **semantic knowledge** can provide additional supervision to ground **visual features** to improve the supervision signal for LwLL
- Challenges
 - Semantic **gap** between the visual and semantic features
 - Limited **coverage** of semantic KGs
- **Co-training** [1] to make **image and semantic classifiers** learn from each other.
- Rich semantic resources can be used:
 - ConceptNet, GloVE, Word2Vec in SemCo [1]
 - BERT in LAVA [2]
 - CLIP in ProposalsCLIP [3]

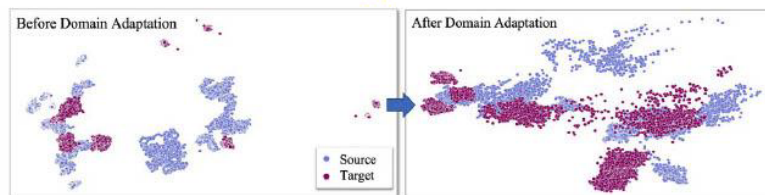


Visualization of language model embeddings for COCO and ImageNet classes used in LAVA [2].

[1] Nassar, Herath, Abbasnejad, Buntine, Haffari. "All Labels Are Not Created Equal: Enhancing Semi-supervision via Label Grouping and Co-training". CVPR2021
 [2] Nassar, Hayat, Abbasnejad, Rezaatoughi, Harandi, Haffari. "LAVA: A Label-resilient Language-guided Visual Learner". WACV2023
 [3] Shi, Hayat, Wu, Cai. "ProposalCLIP: Unsupervised Open-Category Object Proposal Generation via Exploiting CLIP Cues". CVPR2022

Distribution Matching

- Distribution matching enables using related labelled source data for a **less labelled target task** in domain adaptation
- We **generalize distribution disparity** metrics (*i.e.*, *Wasserstein metric*) to both **data and concept shifts** (e.g., LAMDA [2])
- Application to **challenging computer vision/machine learning tasks**
 - multi-object tracking [1]
 - cross-domain continual learning [3]

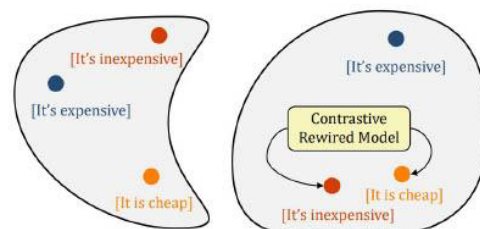


Feature distributions before and after applying our domain adaptation for multi-object tracking [1]

[1] Abeyasinghe, Reid, Rezatofighi, Meyer. "Tracking Different Ant Species: An Unsupervised Domain Adaptation Framework and a Dataset for Multi-object Tracking". IJCAI2023
 [2] Le, Nguyen, Ho, Bui, Phung. "LAMDA: Label Matching Deep Domain Adaptation". PMLR2021
 [3] Simon, Faraki, Tsai, Yu, Schuller, Suh, Harandi, Chandraker. "On Generalizing Beyond Domains in Cross-Domain Continual Learning". CVPR2022

Distribution Matching

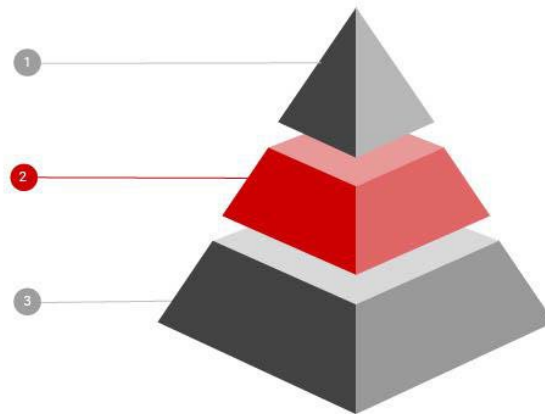
- **Pretrained representations** inject key inductive biases for efficient and effective training on downstream tasks
- Challenges: Sensitive to distribution shift, e.g. domain-shift, language-shift
- Our approaches:
 - Modifying representation to bridge domain-shift [1], enforce content similarity [2]
 - Aligning multilingual representations to enable cross-lingual transferring [3]



[1] Vu, Phung, Haffari. "Effective Unsupervised Domain Adaptation with Adversarially Trained Language Models". EMNLP2020
 [2] Yang, Zhao, Haffari, Shareghi. "Self-supervised Rewiring of Pre-trained Speech Encoders: Towards Faster Fine-tuning with Less Labels in Speech Processing." EMNLP2022
 [3] Vu, He, Phung, Haffari. "Generalised unsupervised domain adaptation of neural machine translation with cross-lingual data selection." EMNLP2021

Outline

- Executive Summary
- Reflections and Future Outlook
- Our Approaches to LwLL

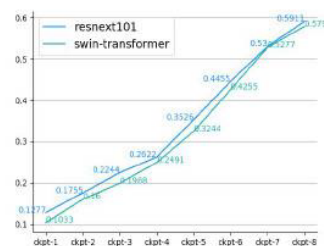


Reflection: Model Architecture in LwLL

- Effective NN architecture choices depends on the task
- Case study:
 - Classification Tasks (e.g., Image/Video): [Vision Transformers](#)
 - Structure prediction tasks (e.g., Object Detection): [CNNs](#)



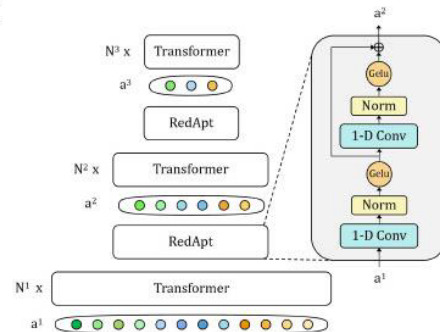
ResNet (JPL Baseline) and MVIT (Our) performance comparisons for Video Classification.



Our comparisons on CNN and Swin-transformer performance for object detection.

Reflection: Model Architecture in LwLL

- Fine-tuning large pretrained language models:
 - Computationally expensive
 - May overfit on small training dataset
- Inject adapters for efficient fine-tuning & ability to handle multiple domains and languages [1,2,3]



[1] Zhao, Yang, Haffari, Shareghi. "RedApt: An Adaptor for wav2vec 2 Encoding Faster and Smaller Speech Translation without Quality Compromise". Findings EMNLP2022
 [2] Zhao, Yang, Shareghi, Haffari. "M-adapter: Modality adaptation for end-to-end speech-to-text translation". INTERSPEECH2022
 [3] Vu, Khadivi, Phung, Haffari. "Domain generalisation of NMT: Fusing adapters with leave-one-domain-out training". Findings ACL2022

Reflection: Training Algorithm in LwLL

- LwLL methods are very sensitive to training hyper-parameters
- Should continue training from last checkpoint or reinitialize with pretrained model?
 - Re-initialize from the whitelisted pretrained model is better than initializing from the previous checkpoint weights
- How should we use our limited data for training, validation, early stopping and model selection?
 - Sub-optimal to define a validation set

Reflection: (Synthetic and Real) Data in LwLL

- Both **labeled** AND **unlabeled** data can be scarce
- Our data strategy:
 - Generative models can synthesize task-specific unlabeled data [1,2]
 - Improves performance of self-training, knowledge distillation and few-shot learning
 - Collecting data, eg from the web
 - Proper data cleaning such as filtering, deduplication are crucial [3]

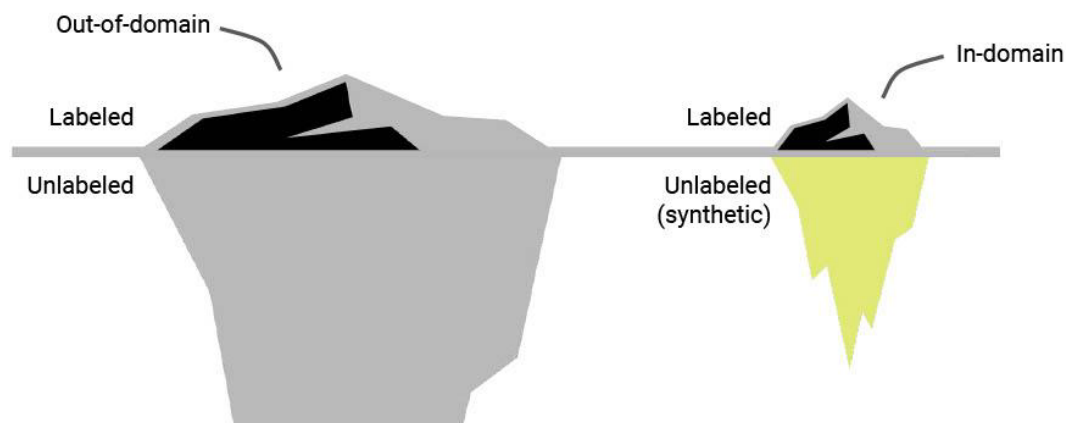
[1] He, Nassar, Kiros, Haffari, Norouzi. "Generate, Annotate, and Learn: NLP with Synthetic Text". TACL2022

[2] Zhuo, Li, Huang, Li, Wang, Haffari, Shiri. "On Robustness of Prompt-based Semantic Parsing with Large Pre-trained Language Model: An Empirical Study on Codex". EACL2023

[3] Wu, He, Haffari, Shareghi. "Koala: An Index for Quantifying Overlaps with Pre-training Corpora". (preprint) 2023

Reflection: (Synthetic and Real) Data in LwLL

- Both **labeled** AND **unlabeled** data can be scarce



He, Nassar, Kiros, Haffari, Norouzi. "Generate, Annotate, and Learn: NLP with Synthetic Text". TACL2022

Reflection and Outlook: Evaluation in LwLL

- Evaluation benchmarks in the community have (big) gaps to real-life
- We appreciate DARPA LwLL's evaluation strategy
 - It reduces the gap between research and reality
 - Future evaluation framework may extend it to **continual learning**
- It provides a solid basis for evaluation benchmarks in the research community
 - To extend DeepLo with shared tasks based on the LwLL evaluation protocol
 - DeepLo was collocated with ACL2018, EMNLP2019, NAACL 2022

Reflection and Outlook: Beyond LwLL Evaluations

- A benchmarking project for Object Detection
- Extend the LwLL protocol with object detection datasets (*16 diverse datasets*) covering a range of domains
 - Aerial Image datasets (4)
 - Underwater datasets (2)
 - Agriculture datasets (4)
 - Natural image datasets (4)
 - Construction datasets (2)
- Extensive study on metrics, model architectures (e.g. ResNet, Swin-Transformer, PVT etc) and detection formulations (e.g. Faster R-CNNs, YOLO & DETR families)



Outlook: LwLL in the Modern Generative AI Era

- Given significant percent advances in instruction-following agents (eg chatGPT, GPT4), is LwLL going to be **still** a challenge?
- How to **effectively** leverage modern GAI for LwLL?

Outlook: Is LwLL still a Challenge in the GAI Era?

- It is unclear how effective they are for tasks for which there is no/limited related data in the pre-training
- Case study: Building an MT system between Kashmiri and English
 - chatGPT has much lower quality translations compared to specialised MT models
- We conjecture there are many tasks, in the **long-tail distribution of tasks**, for which there is **no or little** related publicly available data
 - Consider tasks in organisations with sensitive, confidential, and private data

Outlook: Is LwLL still a Challenge in the GAI Era?

- Challenges in using such **large** models directly
 - Due to memory footprint on edge devices and speed concerns in real-time tasks
- Existing models have limited coverage of data modalities
 - There exists considerable variation in data modalities, eg various sensor data
- A way forward: To build specialised GAI models for organisations to help tackling their LwLL tasks
 - Addressing the constraints on latency, memory, compute, and data modality and governance

Outlook: How to Effectively Leverage Modern GAI?

- Direct approaches:
 - To prompt the definition of the tasks and possibly a few examples (zero/few-shot learning)
 - To fine-tune backbones with (little) data of the target task
- Indirect approach:
 - To generate synthetic data to train specialised models
 - Needs proper instructions/prompts & proper quality-control checks
- Leveraging modern GAI to **assist** the system designers in LwLL?
 - Suggest a blueprint of the system (eg ML approach, training datasets, architecture)

LIST OF SYMBOLS, ABBREVIATIONS AND ACRONYMS

ACCV	Asian Conference on Computer Vision
ACL	Annual Meeting of the Association for Computational Linguistics
AI	Artificial Intelligence
AL	Active Learning
API	Application Programming Interface
API	Application Programming Interface
BLEU	Bilingual Evaluation Understudy
CCU	Cross-Culture Understanding
COLING	International Conference On Computational Linguistics
CVPR	Computer Vision and Pattern Recognition Conference
DARPA	Defense Advanced Research Projects Agency
EACL	Conference of the European Chapter of the Association for Computational Linguistics
ECCV	European Conference on Computer Vision
EMNLP	Conference on Empirical Methods in Natural Language Processing
GPU	Graphics Processing Unit
IC	Image Classification
ICCV	International Conference on Computer Vision
ICLR	International Conference on Learning Representations
ICML	International Conference on Machine Learning
IJCAI	International Joint Conference on Artificial Intelligence
INTERSPEECH	Conference of the International Speech Communication Association
JPL	Jet Propulsion Laboratory
LALA	L earning and A dapt with L ess I Labels
LLM	Large Language Model
LwLL	Learning with Less Labels
NAACL	Annual Conference of the North American Chapter of the Association for Computational Linguistics
NeurIPS	Conference on Neural Information Processing Systems
NGA	National Geospatial-Intelligence Agency
NLP	Natural Language Processing
NMT	Neural Machine Translation
OD	Object Detection
SIGIR	Special Interest Group on Information Retrieval
SSL	Semi-supervised Learning
TPAMI	Transactions on Pattern Analysis and Machine Intelligence
UDA	Unsupervised Domain Adaptation
VC	Video Classification
WACV	Winter Conference on Applications of Computer Vision