AFRL-RI-RS-TR-2024-016

COMBINED REPRESENTATION FOR ADEPT LEARNING (CORAL)

UNIVERSITY OF SOUTHERN CALIFORNIA

FEBRUARY 2024

FINAL TECHNICAL REPORT

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED

STINFO COPY

AIR FORCE RESEARCH LABORATORY
INFORMATION DIRECTORATE

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

NOTICE AND SIGNATURE PAGE

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

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

AFRL-RI-RS-TR-2024-016 HAS BEEN REVIEWED AND IS APPROVED FOR PUBLICATION IN ACCORDANCE WITH ASSIGNED DISTRIBUTION STATEMENT.

FOR THE CHIEF ENGINEER:

/ S /
CARLOS MERLOS
Work Unit Manager

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

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

REPORT DOCUMENTATION PAGE

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED	
		START DATE	END DATE
FEBRUARY 2024	FINAL TECHNICAL REPORT	AUGUST 2019	AUGUST 2023

4. TITLE AND SUBTITLE

COMBINED REPRESENTATION FOR ADEPT LEARNING (CORAL)

5a. CONTRACT NUMBER	5b. GRANT NUMBER	5c. PROGRAM ELEMENT NUMBER
N/A	FA8750-19-1-1000	61101E
5d. PROJECT NUMBER	**5e. TASK NUMBER**	**5f. WORK UNIT NUMBER**
		R2TY

6. AUTHOR(S)

Wael Abd-Almageed

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

University of Southern California
Information Sciences Institute
4676 Admiralty Way
Marina del Rey CA 90292

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)	11. SPONSOR/MONITOR'S REPORT NUMBER(S)
Air Force Research Laboratory/RISC 525 Brooks Road Rome NY 13441-4505	AFRL/RI	AFRL-RI-RS-TR-2024-016

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

In this report, we present a summary of the most recent contributions of the USC Information Sciences Combined Representation for Adept Learning (CORAL) team, under DARPA Learning with Less Labels (LwLL) program.

15. SUBJECT TERMS

Machine Learning, Reinforcement Learning, Transfer Learning

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES
a. REPORT	b. ABSTRACT	C. THIS PAGE		
U	U	U	SAR	13

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

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 SUMMARY

In this report, we present a summary of the most recent contributions of the USC Information Sciences Combined Representation for Adept Learning (CORAL) team, under DARPA Learning with Less Labels (LwLL) program.

2 INTRODUCTION

This report is organized as follows. Section 3 briefly describes three contributions of the CORAL team for (1) generalized few-shot object detection, (2) long term dynamics prediction using computer vision and (3) a novel architecture for few-shot machine translation. Experimental evaluations of these methods are presented in Section 4. Section 5 provides some conclusions for the proposed work.

3 METHODS, ASSUMPTIONS AND PROCEDURES

3.1 Method A: Generalized Few-shot Object Detection

3.1.1 Process A1.

We studied the training strategy for generalized few-shot object detection and explained the challenges of existing detection precision trade off between base and novel classes. First, as the novel instances are extremely limited during training, it is hard to capture the representative visual information of novel classes and adapt the knowledge learned from base classes to novel classes. As a result, the model cannot distinguish between the novel classes, which weakens the few-shot adaptation. Secondly, balanced training strategies such as down-sampling fail to utilize the diverse training samples from base set. Thus, it is hard to pre-serve the complete knowledge of base classes, which leads to overfitting and further decreases the detection scores.

The objective is to learn Discriminative Geometry-aware features via inter-class separation and intra-class compactness. For inter-class separation, we expect the class centers to be well distinct from each other. As shown in Figure 1, motivated by the symmetric geometry of simplex equiangular tight frame (ETF), we proposed to use ETF as classifier to guide the separation of features. To be specific, we derive an offline ETF whose weights are maximally and equivalently separated (i.e., independent from the training data distribution) and are assigned as fixed centers for all classes.

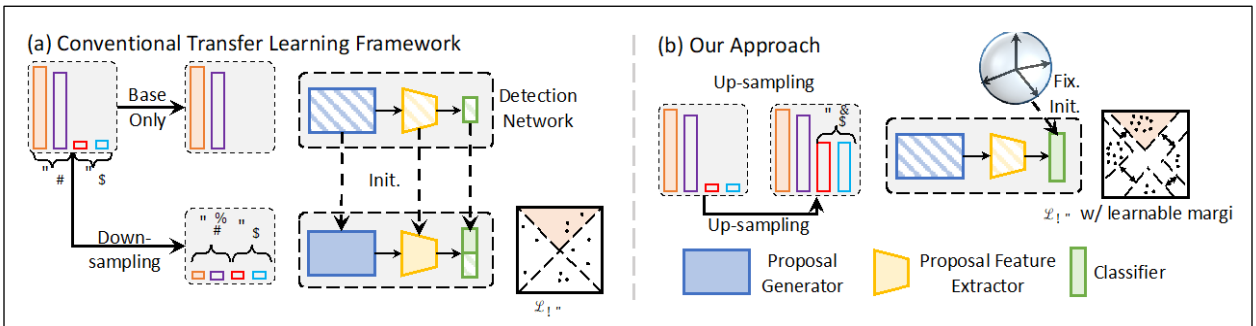

Figure 1: Comparison of training frameworks. (a) Conventional approaches first pre-train a base detector among the *base* set and then finetune on the union of novel set and down-sampled subset of base classes. (b) Instead, we choose to up-sample novel classes and directly train the detector on the full set. We derive a fixed classifier offline with maximally and equally separated weights and learn the adaptive margins to tighten the feature clusters. The margins are estimated from priors of instance distribution and learned though self-distillation. The block with shading means training from scratch. We use the same design for localization and omit it for simplicity.

For intra-class compactness, we expect the features to be closed to the class centers for a clear decision boundary. In practice, we add class-specific margins to output logits during training to push the features close to the class centers. The margins are based on instance distribution prior and are then adaptively adjusted though self-distillation. Meanwhile, even though the margins are added during training, the extreme imbalance between base set and novel set still makes the detector favors more on base set. Considering this limitation and the challenge that the number of novel classes is very limited to provide the gradients for network updating, we proposed to up-sample the images containing annotations of novel classes.

3.2 Method B: Long Term Dynamics Prediction

3.2.1 Process B1.

We identified that the *Cross-Domain* challenge can significantly degrades the performance of the state-of-the-art vision-based long-term dynamics prediction model, Region Proposal Convolutional Interaction Network (RPCIN). The literature postulates that the visual appearance and the dynamic properties of the object are disentangled and should be separately modeled. However, this postulate narrows their discussion to be on the state space of the objects, where the inputs are semantic properties of objects rather than raw images, while the environment characteristics are ignored. For addressing *Cross-Domain* challenge, we extend this postulate and argue that we should seek a common intermediate representation space for both object and environment. Inspired by prior work, where the data is simply provided as semantic masks, we argue that the semantic segmentation space can serve as the intermediate space, where a visual observation model first maps raw images to semantic segmentation masks, and then, the masks, along with the static information of objects, are used for predicting the dynamics.

3.3 Method C: Few-shot Machine Translation

3.3.1 Process C1.

We developed the *moving average equipped gated attention* mechanism, named **Mega**. Mega is a simple, efficient and effective neural architecture which can be used as a drop-in replacement for regular multi-head attention. By leveraging the classic exponential moving average (EMA) approach, Mega is capable of incorporating stronger inductive biases into the attention mechanism. Moreover, the EMA approach enables the design of **Mega**-chunk, an efficient variant of Mega with linear complexity. On five sequence modeling tasks across various data types, **Mega** achieves impressive improvements over a variety of strong baselines, including previous state-of-the-art systems.

Limitations of the attention mechanism – Attention provides the key mechanism that captures contextual information from the entire sequence by modeling pairwise interactions between the inputs at every timestep. However, there are two common drawbacks in the design of attention mechanism: i) *weak inductive bias*; and ii) *quadratic computational complexity*. First, the attention mechanism does not assume prior knowledge of the patterns of dependencies between tokens (e.g., positional inductive bias), instead learning to predict the pairwise attention weights directly from data. Second, the cost to compute and store the attention weights is quadratic in the length of the input sequences. Recent studies have shown the limitations of applying Transformers to long sequence tasks, with respect to both accuracy and efficiency.

Moving Average Gated Attention – Figure 2 illustrates the model architecture of Mega. The key idea is to incorporate inductive biases into the attention mechanism across the timestep dimension, by leveraging the classic exponential moving average (EMA) approach. EMA captures local dependencies that exponentially decay over time and has been widely used in time series data modeling. We introduce a multi-dimensional damped form of EMA with learnable coefficients, and subsequently develop the moving average equipped gated attention mechanism by integrating the EMA with a variant of the single-head gated attention. Theoretically, we show that the single-head gated attention is as expressive as the most commonly used multi-head attention.

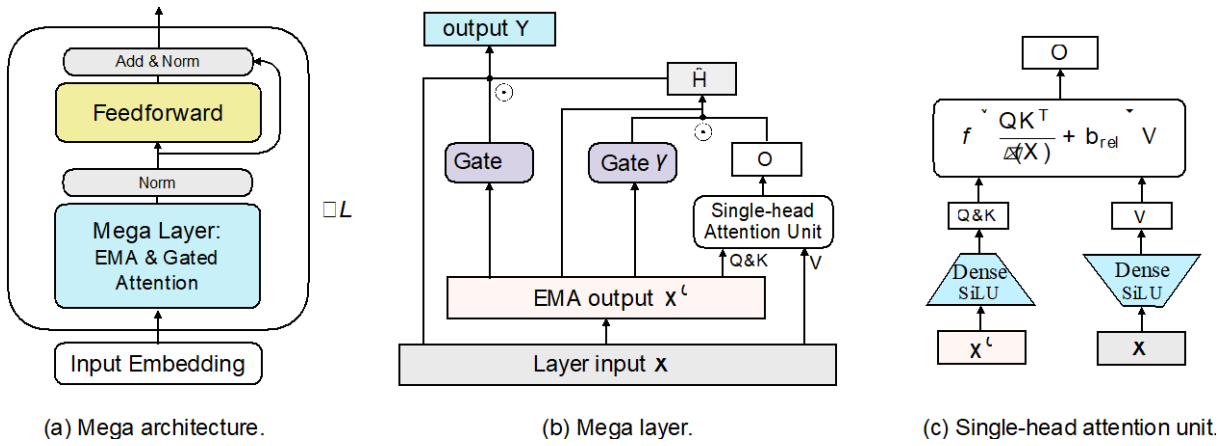

Figure 2: Mega – model architecture. Figure (a) shows the overall architecture of each Mega block. Figure (b) illustrates the gated attention sub-layer equipped with EMA, while Figure (c) displays the details of a single-head attention unit.

4 RESULTS

4.1 Method A: Generalized Few-shot Object Detection

4.1.1 Process A1.

Table 1 shows a comprehensive ablation study of the proposed fixed weight ETF showing improved novel class detection, in particular, the nAP$_{50}$ is boosted from 12.2 to 35.8, which shows that the inter-class separation is essential for distinguishing objects in generalized few shot object detection.

Table 1: Ablation study of the performance of the proposed ETF for generalized few shot object detection

Idx.	ETF	Margin	RFS	AP_{50}	bAP_{50}	nAP_{50}
1				60.6	80.8	0.0
2	X			61.8	81.2	3.6
3		X		63.9	81.1	12.2
4			X	62.8	80.8	9.3
5	X	X		69.6	81.0	35.8
6	X		X	65.2	81.2	17.0
7	X	X	X	74.9	81.0	56.4

4.2 Method B: Long Term Dynamics Prediction

4.2.1 Process B1.

We conducted extensive evaluations on the proposed dataset with *Cross-Domain* setup. Our experiments show that the performance by using ground-truth masks as input to RPCIN can significantly exceed the performance by using raw image as input. Even in the case that the ground-truth mask is absent, sub-optimal masks, which can be obtained via self-supervised learning, can also dramatically mitigate the *Cross-Domain* challenge, as shown in Tables 2 and 3.

Table 2: Performance of **SimB** domain trained RPCIN model on *Cross-Domain* challenge with various types of input. BN is used for all segmentation mask input training. In addition to the BN baseline, which takes RGB image as input, we also list other normalization method results as baselines for a comprehensive comparison and further demonstrating the advantage of unifying the visual domains with semantic masks. Aligned and Cross results are the same for GT-Mask trained model because they share exactly the same data. **Bold** highlights the best results and underline highlights the second best results.

Source Dataset	SimB-Border				SimB-Split			
Target Dataset	SimB-Border (Aligned)		BlenB-Border (Cross)		SimB-Split (Aligned)		BlenB-Split (Cross)	
Eval Period	P1	P2	P1	P2	P1	P2	P1	P2
Raw RGB Image Input								
RGB-BN	1.131 ± 0.011	9.568 ± 0.121	6.185 ± 2.206	23.564 ± 4.307	0.913 ± 0.019	7.732 ± 0.208	8.622 ± 2.313	27.511 ± 4.3
RGB-IN	1.102 ± 0.045	9.426 ± 0.446	2.754 ± 0.188	15.863 ± 1.073	0.945 ± 0.082	7.641 ± 0.549	2.127 ± 0.381	11.281 ± 1.3
RGB-GN	1.117 ± 0.058	9.323 ± 0.346	1.637 ± 0.106	11.507 ± 0.425	**0.899 ± 0.042**	7.632 ± 0.386	2.315 ± 0.447	12.335 ± 0.7
RGB-LN	**1.085 ± 0.033**	**9.165 ± 0.145**	3.114 ±1.086	16.074 ± 3.716	0.922 ± 0.042	7.433 ± 0.237	3.648 ± 1.199	15.662 ± 3.4
Segmentation Mask Input								
GT-Mask	1.091 ± 0.044	9.358 ± 0.465	**1.091 ± 0.044**	**9.358 ± 0.465**	0.916 ± 0.005	7.431 ± 0.511	**0.916 ± 0.005**	7.431 ± 0.51
Sup-Mask	1.093 ± 0.021	9.396 ± 0.285	1.093 ± 0.021	9.397 ± 0.286	0.971 ± 0.011	**7.372 ± 0.089**	0.981 ± 0.012	**7.422 ± 0.0**
Self-Mask	1.119 ± 0.037	9.604 ± 0.300	1.132 ± 0.035	9.614 ± 0.291	0.911 ± 0.025	7.837 ± 1.334	0.959 ± 0.020	8.017 ± 1.30

Table 3: Performance of **BlenB** domain trained RPCIN model on *Cross-Domain* challenge with various types of input. BN is used for all segmentation mask input training. In addition to the BN baseline, which takes RGB image as input, we also list other normalization method results as baselines for a comprehensive comparison and demonstrating the advantage of unifying visual domains with semantic masks. Aligned and Cross results are the same with GT-Mask trained model because they share exactly the same data. **Bold** highlights the best results and underline highlights the second best results.

Source Dataset	BlenB-Border				BlenB-Split			
Target Dataset	BlenB-Border (Aligned)		SimB-Border (Cross)		BlenB-Split (Aligned)		SimB-Split (Cross)	
Eval Period	P1	P2	P1	P2	P1	P2	P1	P2
Raw RGB Image Input								
RGB-BN	1.084 ± 0.023	9.713 ± 0.554	261.768 ± 75.655	233.977 ± 34.460	0.918 ± 0.037	7.478 ± 0.079	171.750 ± 84.274	187.635 ± 53.9
RGB-IN	1.103 ± 0.041	9.471 ± 0.443	25.600 ± 15.781	58.599 ± 21.827	0.906 ± 0.062	7.368 ± 0.168	24.460 ± 17.239	42.613 ± 18.1
RGB-GN	1.075 ± 0.045	9.560 ± 0.145	3.899 ± 0.526	20.425 ± 1.913	0.931 ± 0.039	7.641 ± 0.165	5.033 ± 0.575	18.970 ± 1.3
RGB-LN	**1.064 ± 0.020**	**9.345 ± 0.350**	12.969 ± 1.508	46.113 ± 1.157	0.892 ± 0.027	7.687 ± 0.321	9.518 ± 2.024	31.364 ± 3.7
Segmentation Mask Input								
GT-Mask	1.091 ± 0.044	9.358 ± 0.465	**1.091 ± 0.044**	9.358 ± 0.465	0.916 ± 0.005	7.431 ± 0.511	0.916 ± 0.005	7.431 ± 0.51
Sup-Mask	1.122 ± 0.036	9.353 ± 0.268	1.121 ± 0.036	**9.353 ± 0.268**	**0.891 ± 0.027**	**7.317 ± 0.273**	**0.889 ± 0.027**	**7.313 ± 0.27**
Self-Mask	1.136 ± 0.024	9.945 ± 0.563	1.136 ± 0.024	9.943 ± 0.560	0.914 ± 0.022	7.539 ± 0.227	0.944 ± 0.023	7.650 ± 0.22

4.3 Method C: Few-shot Machine Translation

4.3.1 Process C1.

To evaluate **Mega**, we conduct experiments on five benchmark sequence modeling tasks across various data types, comparing with current state-of-the-art models on each task. Specifically, through five sequence modeling tasks across various data types, including long-context sequence modeling, neural machine translation, auto-regressive language modeling, and image and speech classification, we demonstrate that **Mega** significantly outperforms a variety of strong baseline models, in terms of both effectiveness and efficiency (see Table 4). These improvements illustrate the importance of modeling long- and short-term dependencies via different patterns of inductive biases.

Table 4: Experimental results of Transformer (XFM), S4 and **Mega** on five sequence modeling benchmarks of different types of data, including long range arena (LRA), machine translation (WMT14 en-de), language modeling (WikiText-103), image classification (ImageNet-1k), raw speech classification (SC-Raw).

	LRA (Acc. ")	WMT14 (BLEU ")	WT103 (PPL. #)	ImageNet (Acc. ")	SC (Acc.
XFM	59.24	27.68	18.66	81.80	7
S4	85.86	–	20.95	–	97.50
Mega	88.21	29.01	18.07	82.35	97.30

5 CONCLUSIONS

We presented a summary of the most recent contributions of the USC Information Sciences Combined Representation for Adept Learning (CORAL). For detecting objects from novel categories, we presented a novel generalized few-shot learning method that uses a new representation approach, from the perspective of discriminative feature geometry. The proposed method alleviates the problem of forgetting base classes. In terms of predicting long term dynamics from visual information, we identified the problem of cross-context and cross-domain challenges and introduced a novel method for predicting long term dynamics using semantic segmentation as an intermedia symbolic representation, which shows robustness to the cross-domain problem. Finally, for few-shot machine translation, we presented Mega, a novel transformer architecture that uses exponential moving average principals to alleviate the challenges of long-term sequence modeling. Mega is more computationally efficient than state of the art transformer-based machine translation architectures and provides state of the art translation performance.

6 REFERENCES

7 ACRONYMS

AE AutoEncoder.

AI Artificial Intelligence.

AUC Area under the Curve.

Celeb-DF Celebrity DeepFake Dataset.

CNN Convolutional Neural Network.

COCO Common Objects in Context.

DCNN Deep Convolutional Neural Network.

DCT Discrete Cosine Transform.

DFDC Deefake Detection Challenge Preview Dataset.

DSP-FWA Dual Spatial Pyramid for Exposing Face Warp Artifacts.

FAR False Acceptance Rate.

FF++ FaceForensics++.

GAN Generative Adversarial Network.

JPEG Joint Photographic Experts Group.

LoG Laplacian of Gaussian.

log(wP) Log-Weighted Precision.

LSTM Long Short-Term Memory.

MesoNet Mesoscopic Analysis Neural Network.

MISLNet Multimedia and Information Security Lab Neural Network.

OIS OpenImages Splices.

pAUC Partial Area Under the Curve.

PRNU Photo-Response Non-Uniformity.

R-CNN Region-based Convolutional Neural Network.

ResNet Residual Neural Network.

ROC Receiver Operating Characteristic.

SNR Signal-to-Noise Ratio.

SVDD Support Vector Data Description.

SVM Support Vector Machine.

t-SNE t Distributed Stochastic Neighbor Embedding.

TAR True Acceptance Rate.

tAUC Truncated AUC.

Xception Extreme Inception Neural Network