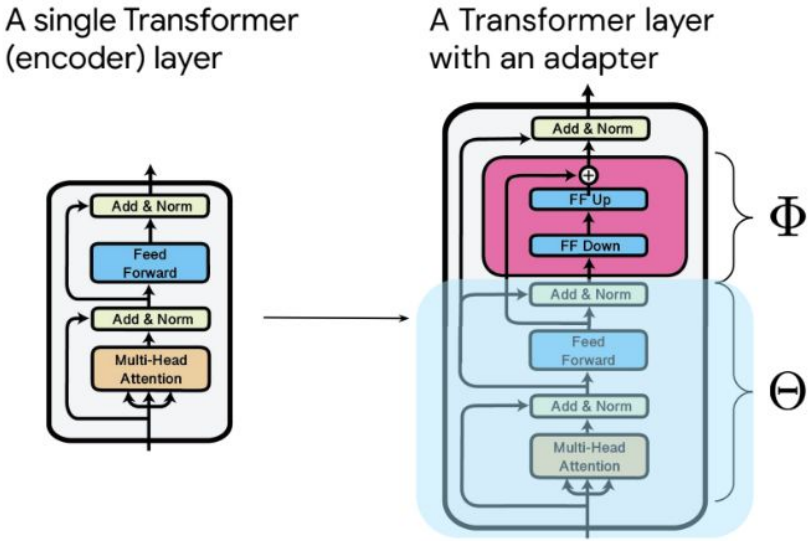
Weak language supervision fine-tuning of vision encoders

Diganta Misra

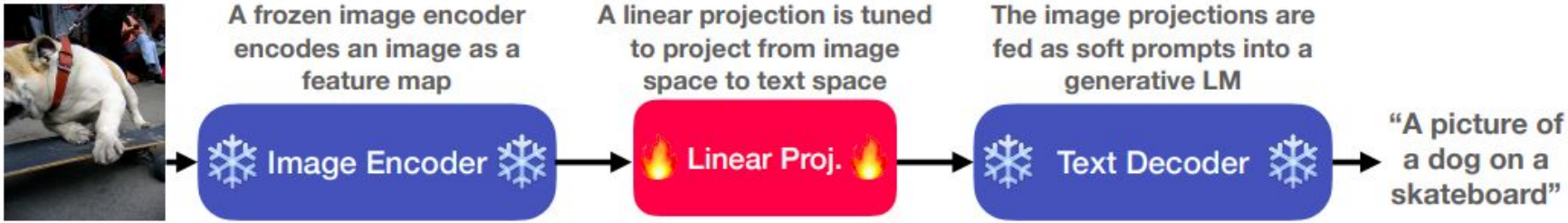
Background: Adapters

- Allocate additional capacity for to enable transfer learning on incoming downstream tasks without training a new model for every task using adapters
- Small bottleneck layers inserted between a pre-trained model's weights
- Adapter parameters are **encapsulated** between transformer layers with parameters which are frozen



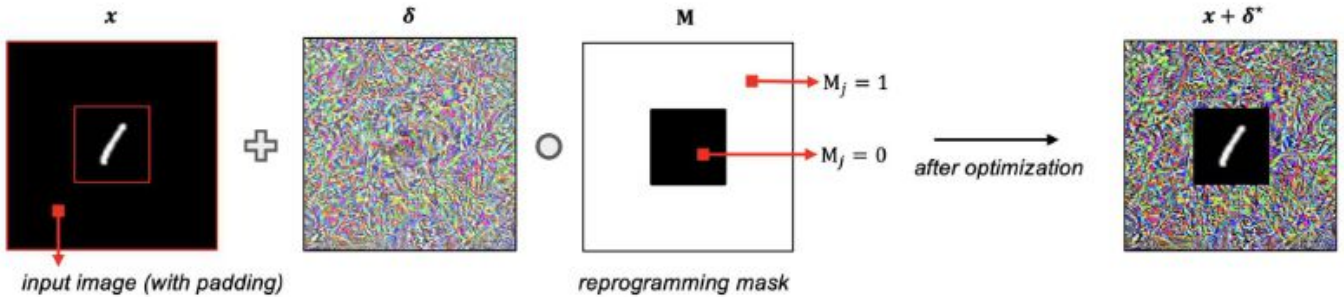
Background: Linear Mapping Image to Text

- Learns a simple mapping (similar to adapters) to bridge pre-trained image models and pre-trained LM
- Allows deep analysis of individually-trained visual encoders and transfer of their features

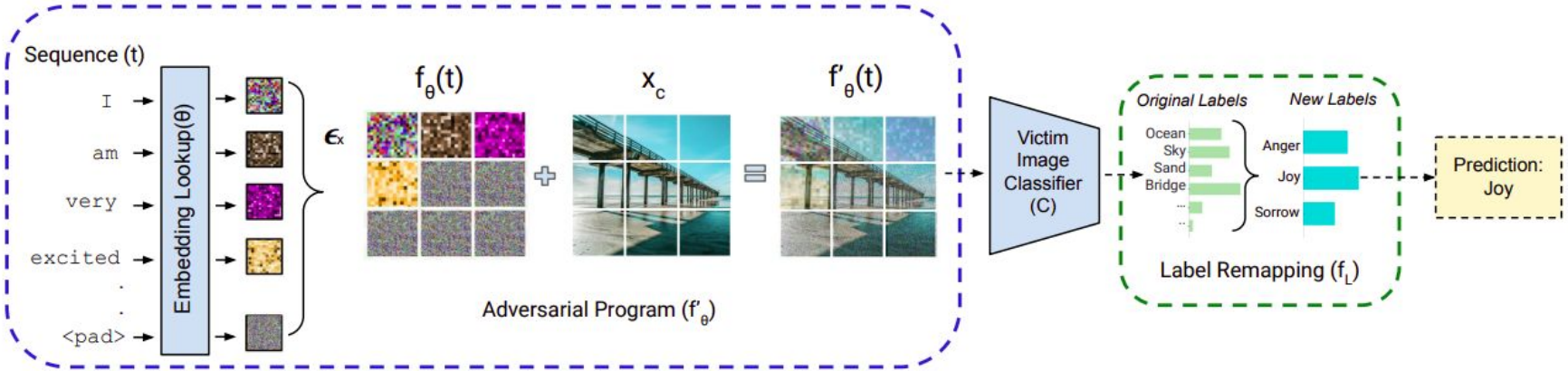


Background: Adversarial Reprogramming

- Repurpose the pretrained model to perform a new task through input space transformation
- Similar to Adapters but does the transformation at input and output level
- Computationally cheaper

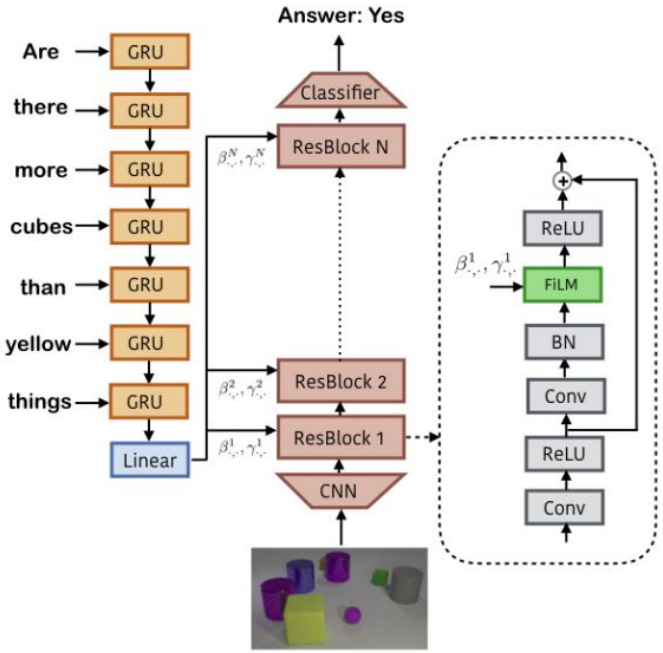


Background: Cross Modal Reprogramming

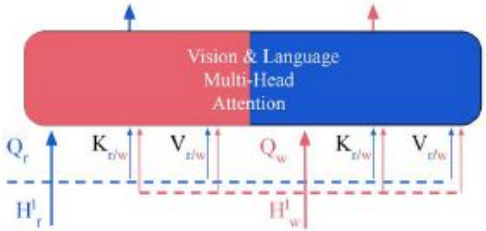


Background: FiLM

- Feature-wise affine transformation based on conditioning information.
- Highly effective for visual reasoning tasks that require multi-step, high-level processing, a challenge for standard deep learning methods that don't explicitly model reasoning.

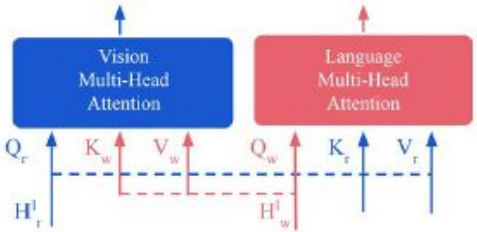


Background: Modality Fusion Strategies



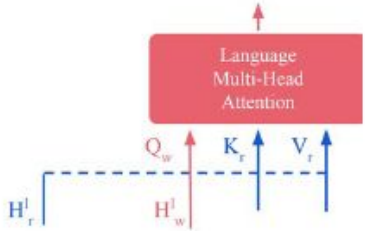
Merged attention

Each modality *attends* to **both** modalities.



Coattention

Each modality *attends* **only** to the other modality (two asymmetric attentions).

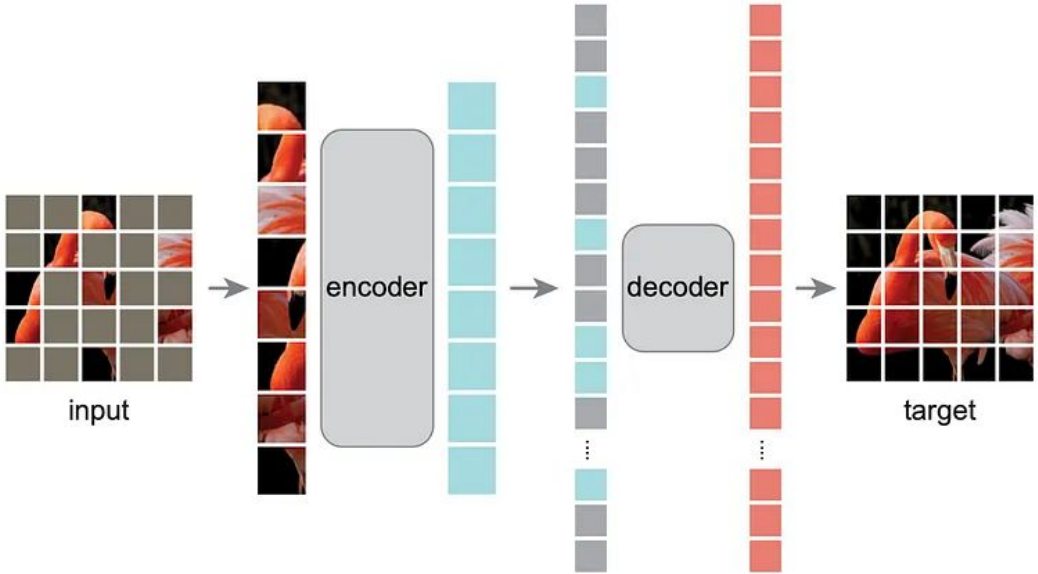


Asymmetric attention

Only one modality (*e.g.*, language) *attends* to the **other** modality (*e.g.*, image).

Background: Masked Autoencoders (MAE)

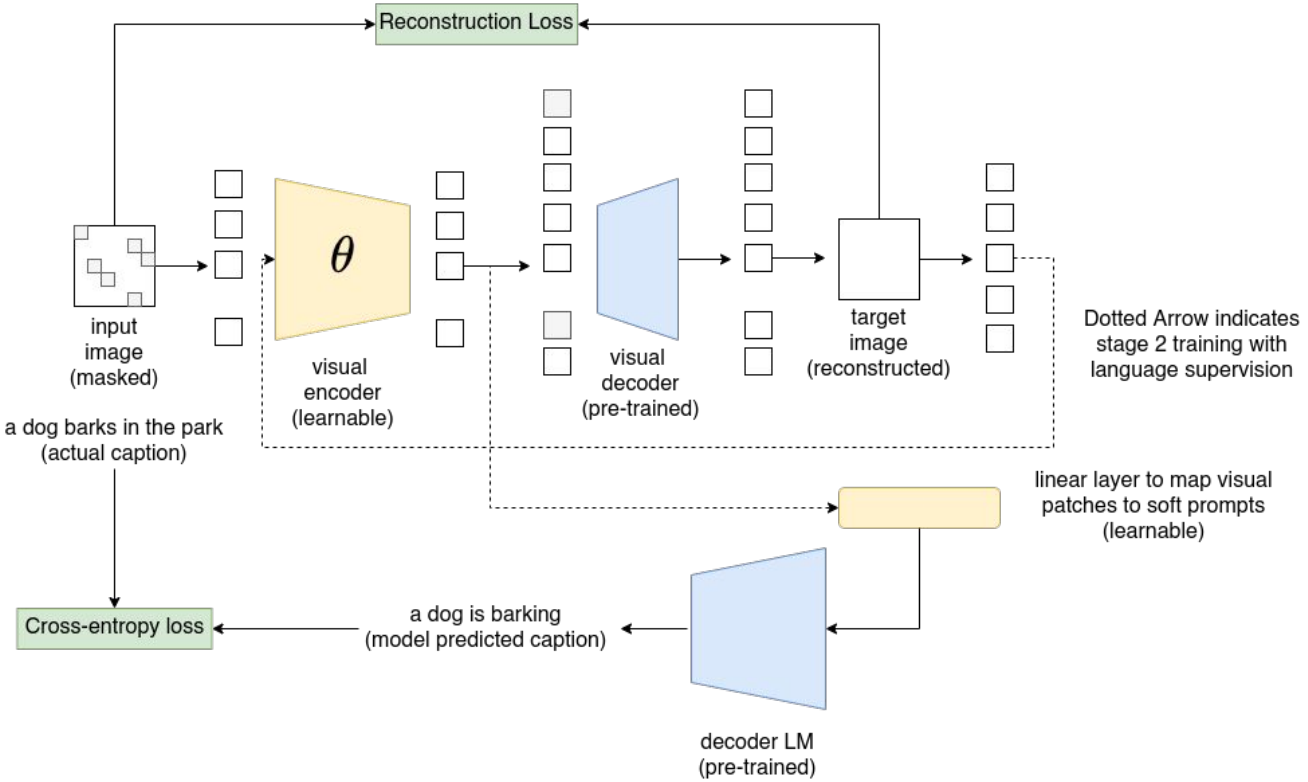
- Self-supervised pre-training of vision models (similar to BERT in language processing)



Motivation

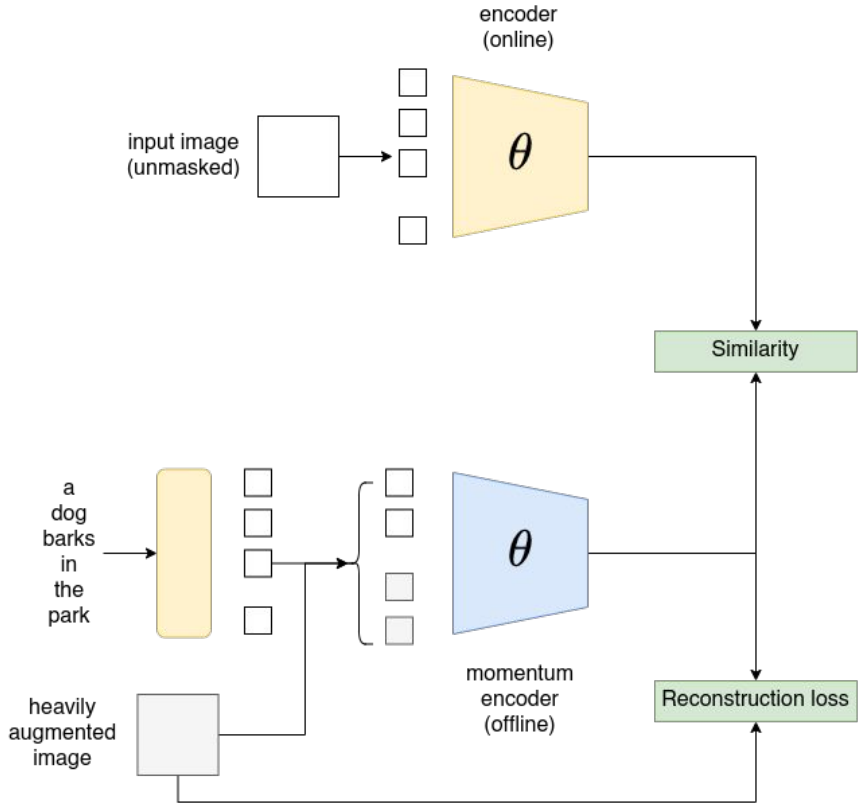
- Current SOTA vision models have achieved impressive results on various visual recognition tasks. However, these models still fail to generalize to out-of-distribution (OOD) settings.
- Prior work such as CLIP has demonstrated strong zero-shot performance on several downstream tasks by leveraging image-caption pairs. However, collecting large amounts of such pairs can be expensive and impractical for certain domains, such as medical imaging or remote sensing.
- Can we adapt existing vision models with minimal language supervision by using a limited number of image-caption pairs?

Approach 1: MAE + LM

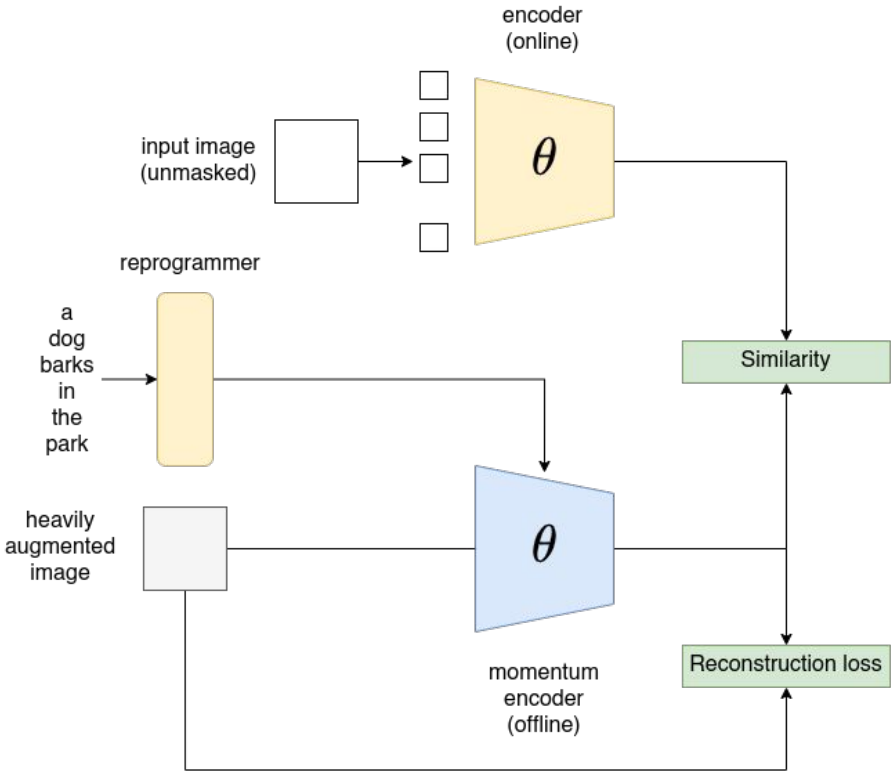


Approach 2: MAE + CL

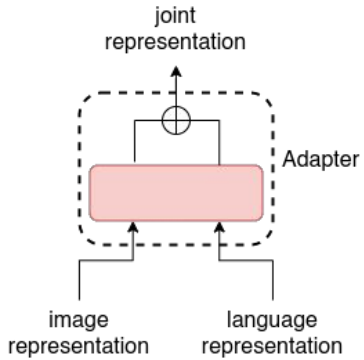
- Reformulate the approach 1 to reduce the computational resource requirements by substituting contrastive learning instead of language modeling objective.



Approach 2: MAE + CL + Adapters

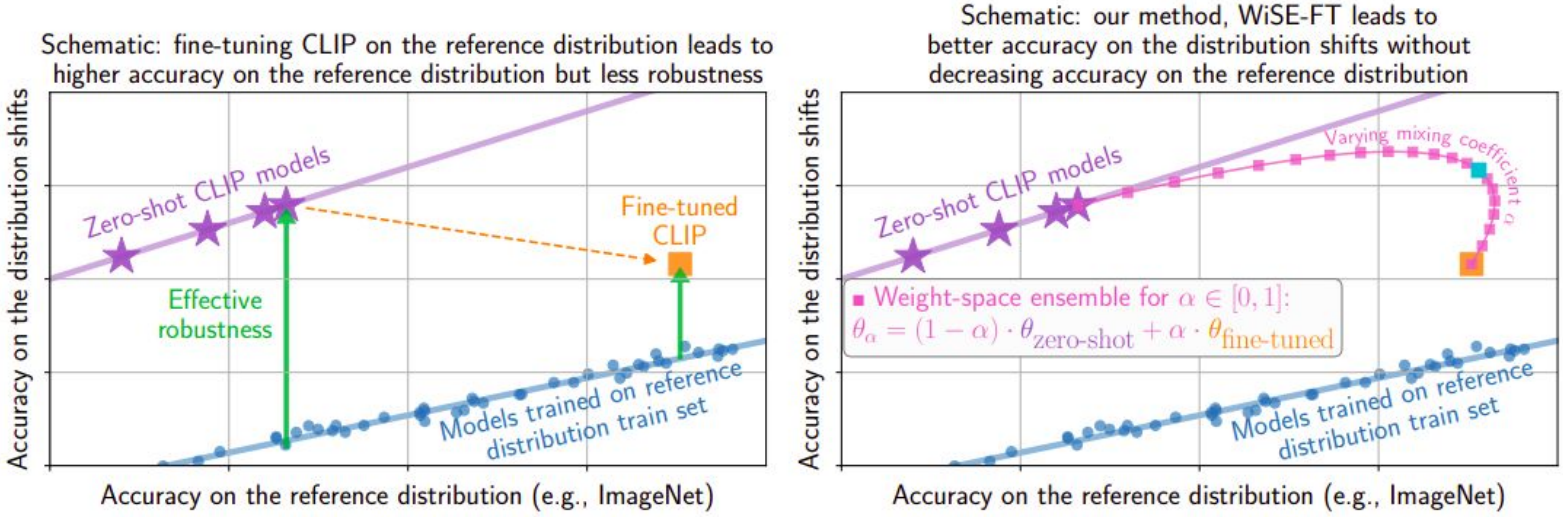


- Introduces adapters for information fusion from visual and textual modality as a means of grounding to embed language concepts in the vision models.



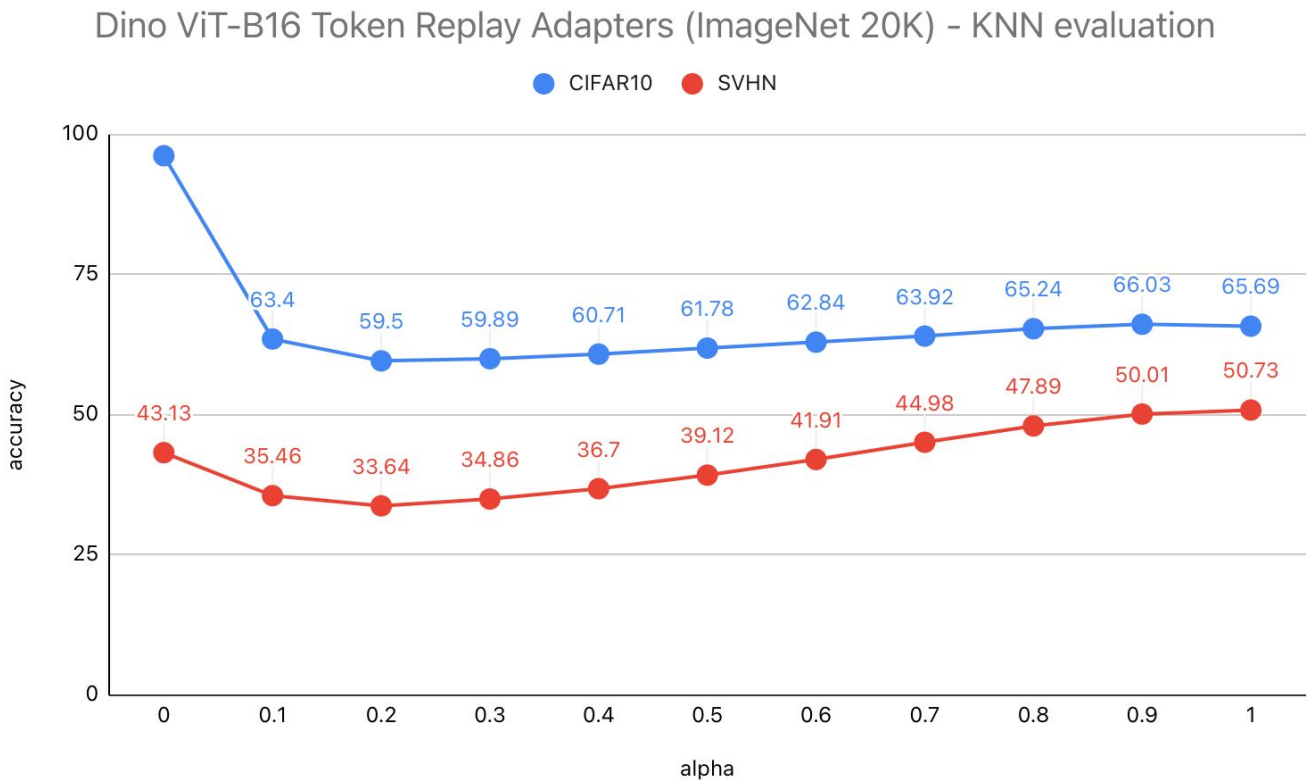
Background: Wise-FT

- Fine-tuning pretrained models on a target distribution often leads to reduce robustness to distribution shifts
- Can zero-shot models be fine-tuned without reducing accuracy under distribution shift?

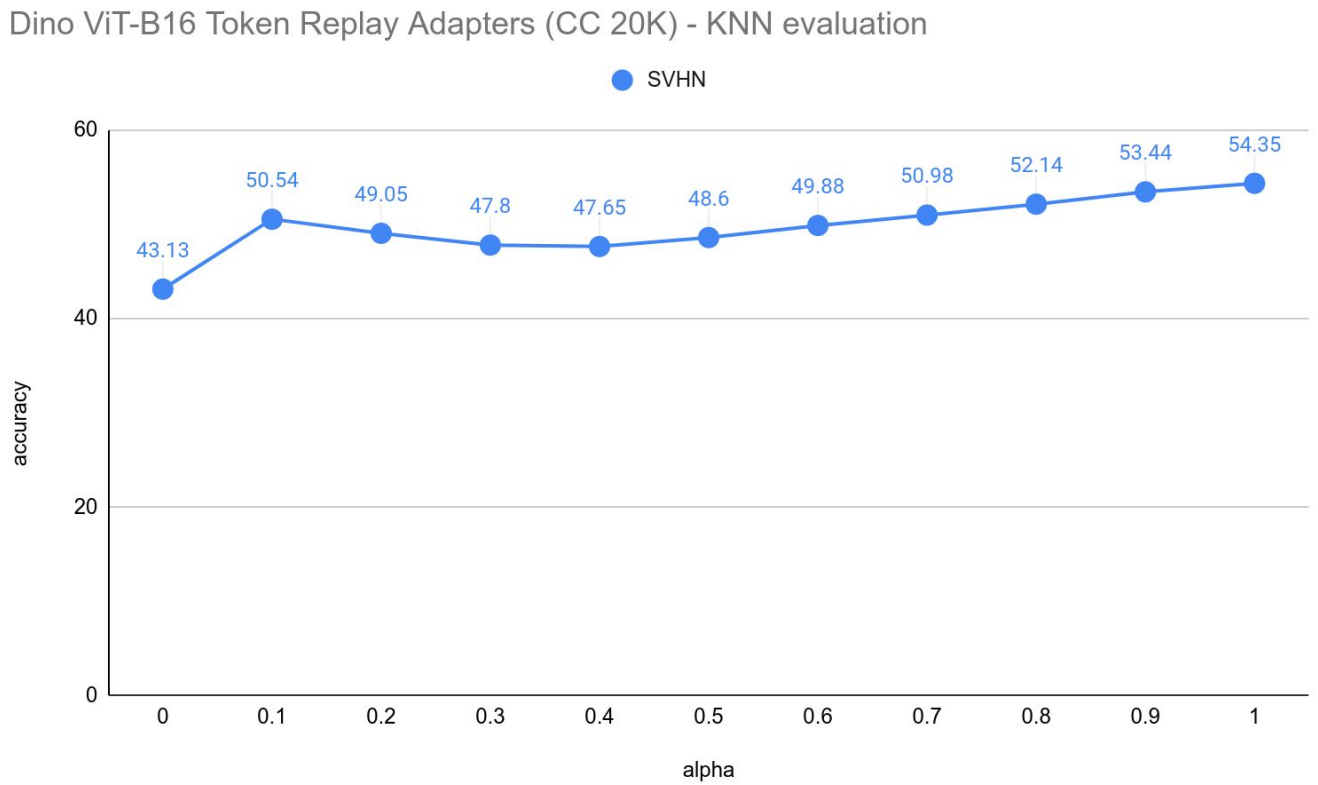


```
alpha x finetuned model + (1 - alpha) x pretrained model
```

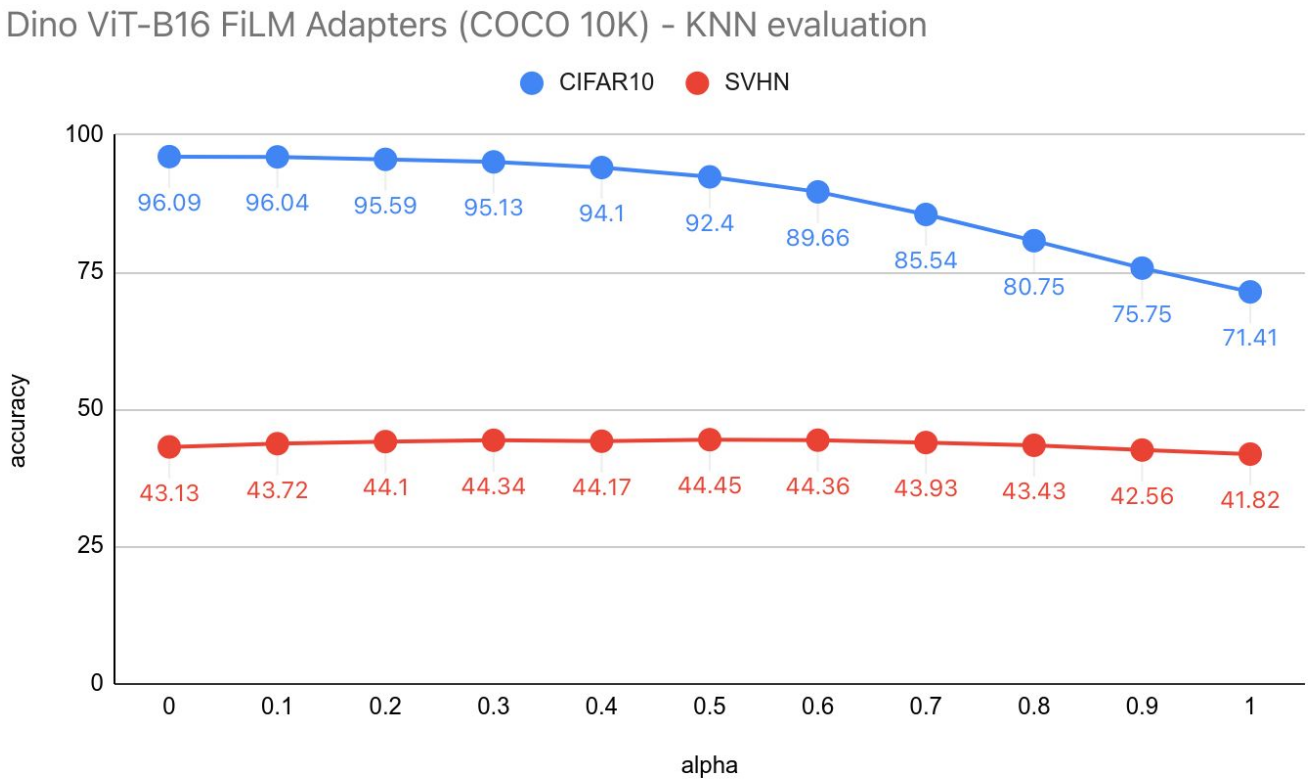
Results



Results


Dino ViT-B16 Token Replay Adapters (CC 20K) - KNN evaluation
SVHN
60
40
20
0
accuracy
43.13
50.54
49.05
47.8
47.65
48.6
49.88
50.98
52.14
53.44
54.35
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
alpha

Results

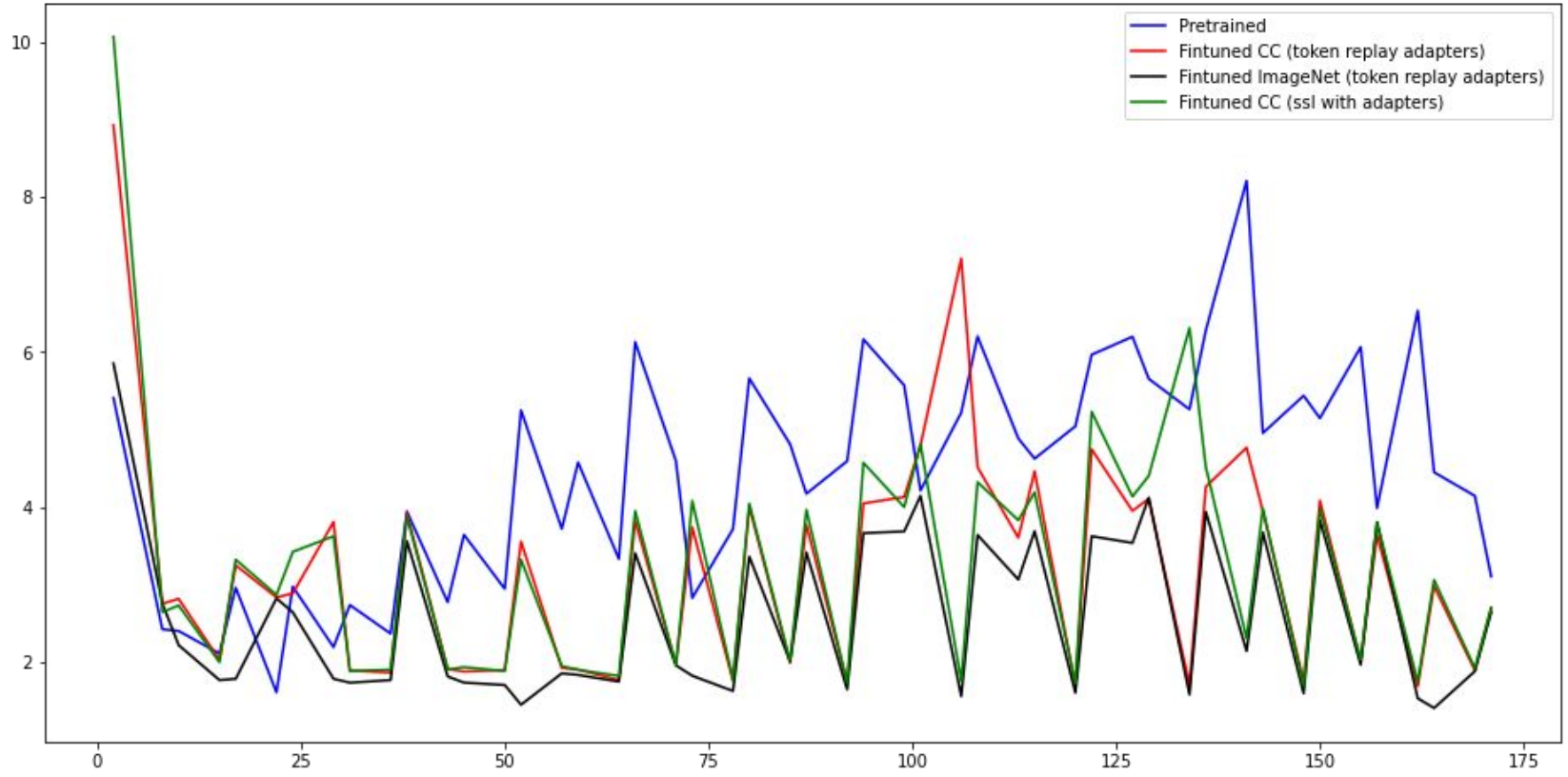

Dino ViT-B16 FiLM Adapters (COCO 10K) - KNN evaluation
CIFAR10
SVHN
100
75
50
25
0
accuracy
96.09
96.04
95.59
95.13
94.1
92.4
89.66
85.54
80.75
75.75
71.41
43.13
43.72
44.1
44.34
44.17
44.45
44.36
43.93
43.43
42.56
41.82
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
alpha

Findings

- We observe a decrease in performance after finetuning models on CC-3M and COCO using our proposed approaches on in-domain sets such as CIFAR compared to pretrained models on ImageNet, which is a dataset that is closer in distribution to CIFAR.
- Our method sometimes outperforms pretrained models on datasets such as SVHN, EuroSAT, DTD, which is not closer in distribution to ImageNet.
- Some regularization is needed to ensure that the performance on in-domain sets is not affected while adapting the models to out-of-domain sets.
- Future direction might be to do Selective k-finetuning approaches or approaches similar to LST (Ladder Side Tuning) to prevent ID performance collapse due to shift in distribution.

Generalization Metrics: WeightWatcher

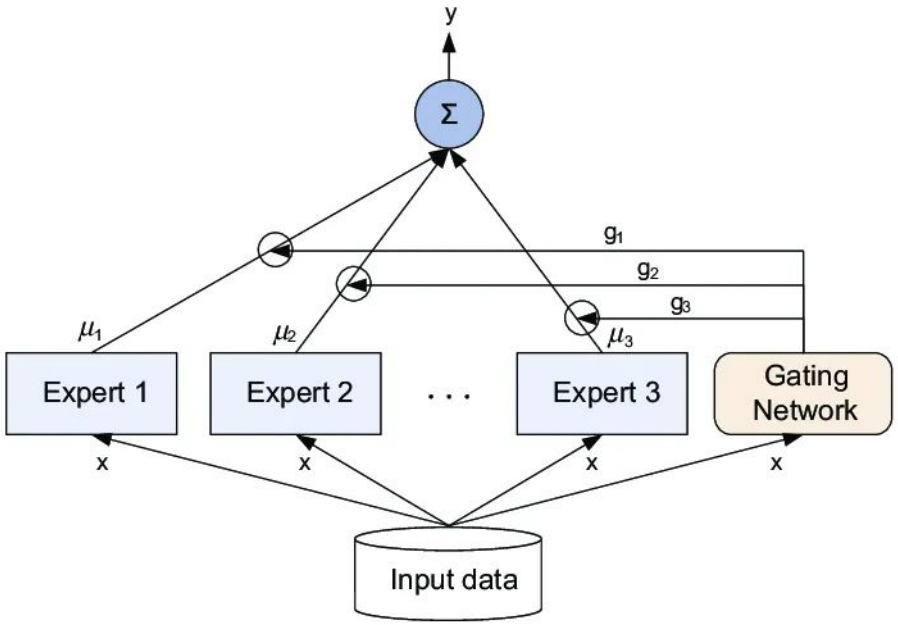
- Analyzes the weight matrices of deep neural networks to provide insights into their generalization performance.
- Calculates an "alpha" value, which is a measure of the level of overfitting in the model by comparing the distribution of the singular values of the weight matrices of the model to the expected distribution of the singular values for a random matrix of the same dimensions.


Pretrained
Fintuned CC (token replay adapters)
Fintuned ImageNet (token replay adapters)
Fintuned CC (ssl with adapters)
10
8
6
4
2
0
25
50
75
100
125
150
175

Back to the drawing board

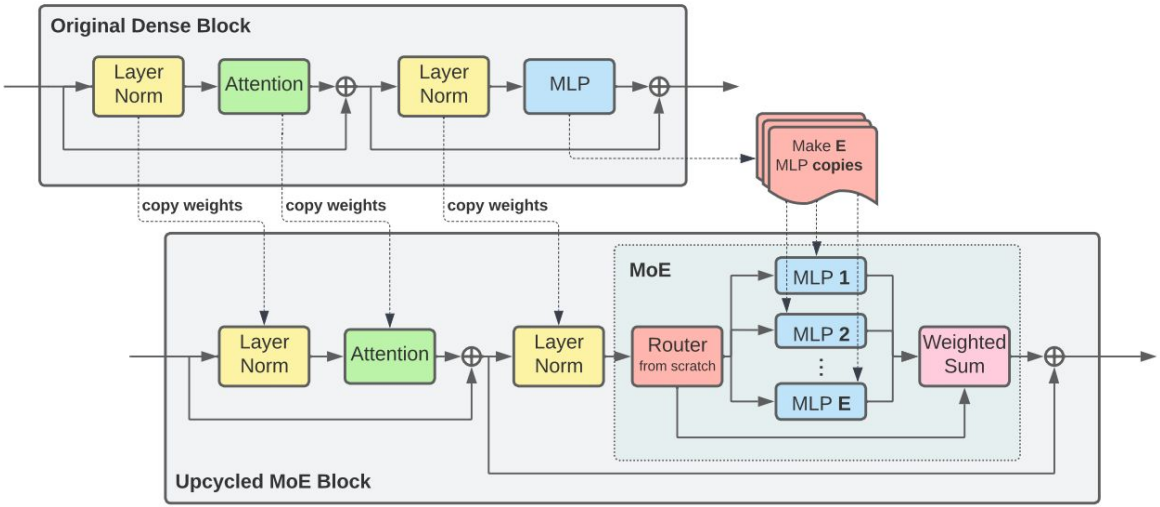
Background: Mixture of Experts (MoE)

- Combines multiple models (“experts”) to form a single, more powerful predictive model
- Each expert is responsible for modeling a specific sub-task and model's output is a weighted combination of the outputs from each expert
- MoE are sparse and have an exponential growth rate in functional space



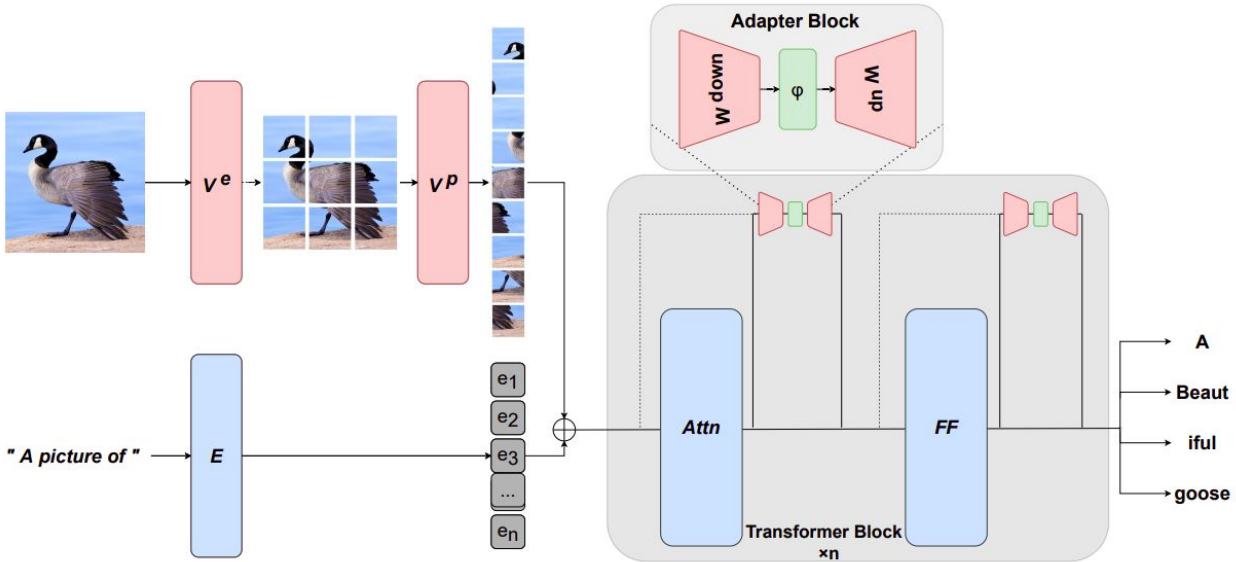
Background: Sparse Upcycling

- MoEs requires data and compute at scale due to sparse nature.
- Sparse upcycling reuses the dense checkpoints as a good initialization to reduce the training cost.



Background: MAGMA

- Augments generative language models with additional modalities using adapter-based finetuning.



λ

λ (Image)

If learnable then concate

MoE

Router

FFN 1

FFN 2

...

FFN N

Weighted Sum

"A picture of "

E

e1

e2

e3

...

en

Attn

MoE

Transformer Block x N

a beautiful goose

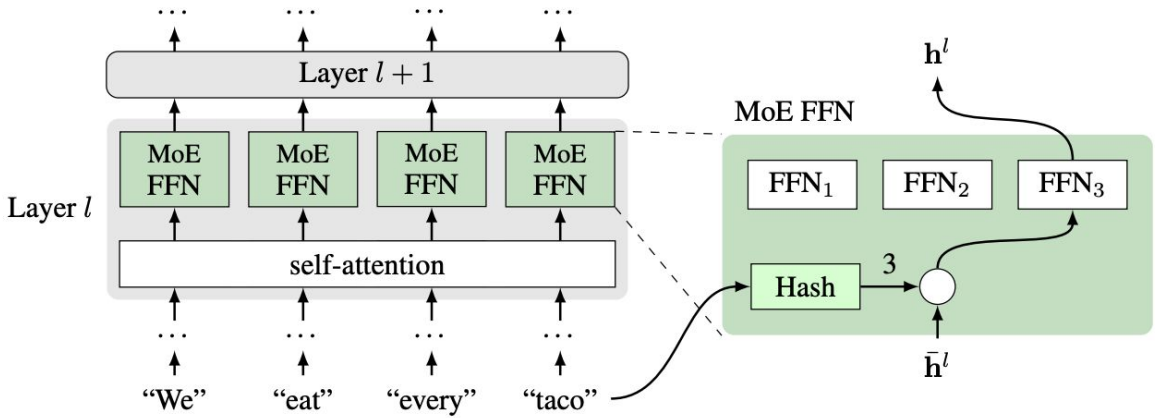


Figure 1: **Overview of the Hash Layer.** Tokens are routed to fixed expert modules based on their hash.

Our approach

- Improves MAGMA by adding a learnable MoE for better representation learning
- Next Steps:
  - Test the efficacy of the proposed approach on image-captioning datasets such as CC-3M. (In progress)
  - Explore the effect of sparsity by introducing different gating mechanism in routing for trade-off between compute efficiency and generalization.
  - Resolve MoE instability issues by optimizing hyper-params related to scale of init.
  - Scale to Summit HPC.

Summit HPC

Specifications and Features

Processor: IBM POWER9™ (2/node)

GPUs: 27,648 NVIDIA Volta V100s (6/node)

Nodes: 4,608

Node Performance: 42TF

Memory/node: 512GB DDR4 + 96GB HBM2

NV Memory/node: 1600GB

Total System Memory: >10PB DDR4 + HBM + Non-volatile

Interconnect Topology: Mellanox EDR 100G InfiniBand, Non-blocking Fat Tree

Peak Power Consumption: 13MW

Takeaways/ Issues

1. Cross modal reprogramming from language to visual tokens is hard and unstable due to the unstructured nature of natural language and many to one mapping.
2. We faced numerous challenges with getting hyper-params to work to get a converging loss profile when training adapter based reprogrammers. We hypothesized that this is due to the fact that the early reprogrammed input to the adapter might be introducing pure noise that the randomly initialized adapters fail to be conditioned on properly.
3. An additional issue was the fact that the only ID dataset which had captions was ImageNet on which numerous image models were available however the ImageNet captions are quite poor and do not necessarily help in learning.
4. Lack of diverse weights for MAE models. We only had one checkpoint for a encoder+ decoder pretrained MAE.
5. [Current] Solving instability issues with large sparse MoE models.
6. [Current] Scaling to Summit HPC has challenges in parallelization and acceleration due to IBM Power9 System constraints.

Questions?