Diversifying Vision Transformers for Out-of-Distribution Generalization

Armand Nicolicioiu[1], Andrei Nicolicioiu[2], Bogdan Alexe[1]
[1]University of Bucharest, [2]Max Planck Institute for Intelligent Systems, Tübingen, Germany

Introduction

- Deep learning models **generalize** well in unseen scenarios. However, **out-of-distribution** data brings difficulties.
- Simple **spurious correlations** in training data can act as **shortcuts** used instead of relying on the **causal features**.

Computer Vision

In CI-MNIST [2] the training label is the parity of the digit. The background color can be correlated with the label or it can be random.

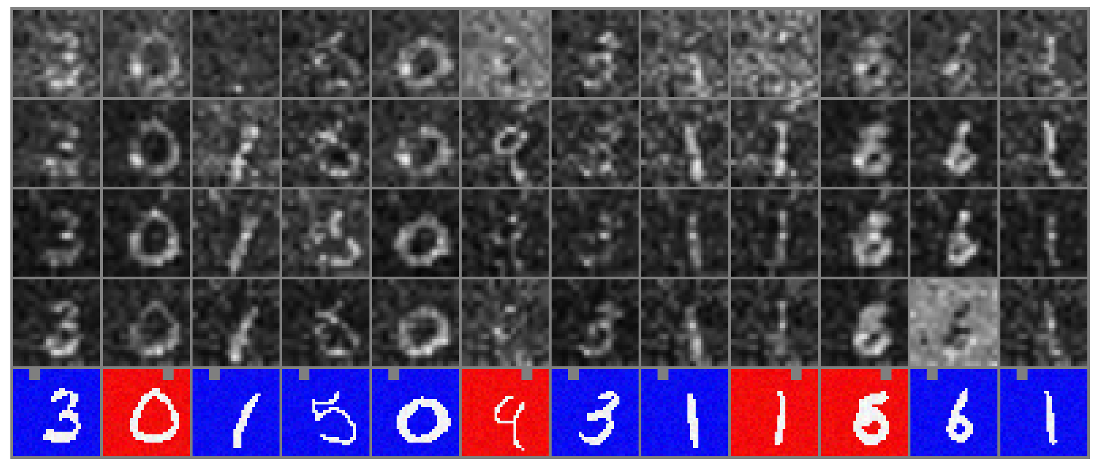


Figure 7: Unbiased training (random background)

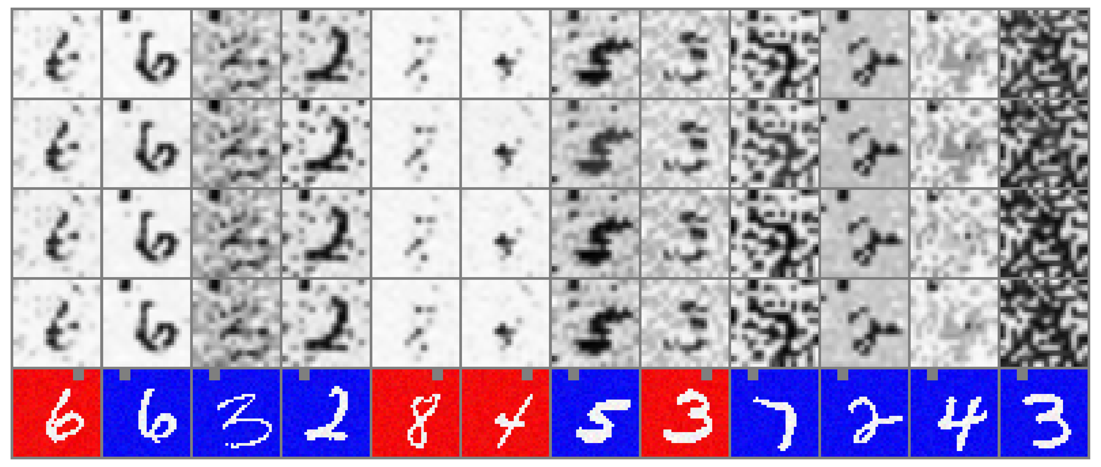


Figure 8: Biased training (correlated background)

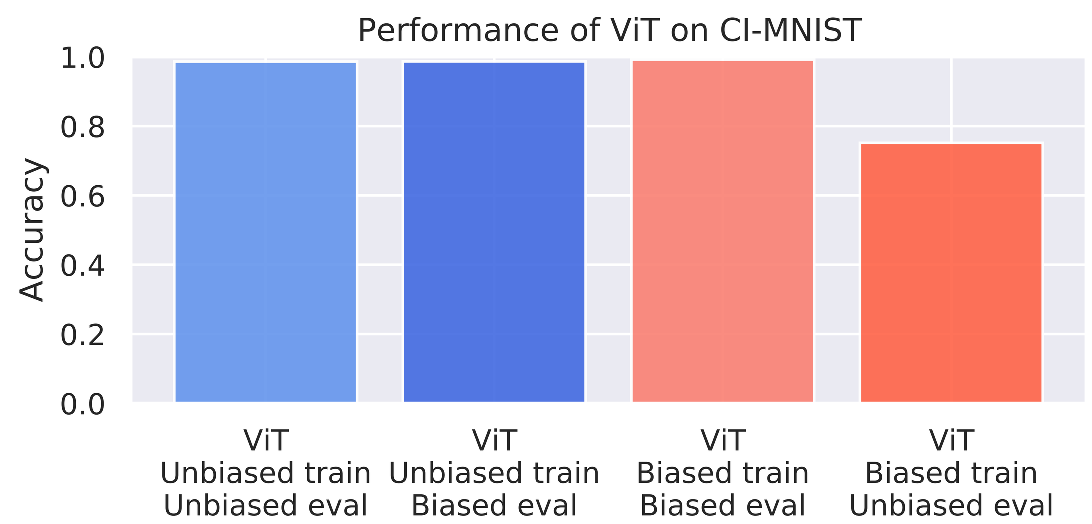


Figure 9: ViT performance for ERM training.

Contributions

- Investigated ViTs [1] inductive bias for modularity and designed a **head selection** method that improves OOD performance.
- Proposed a **head diversification** method based on orthogonality of head influence, leading to better head specialization.

Reinforcement Learning

- Novel RL environment based on CartPole.
- A green dot is overlayed on on the left or right side, according to the optimal action. It can act as a "shortcut" during training.

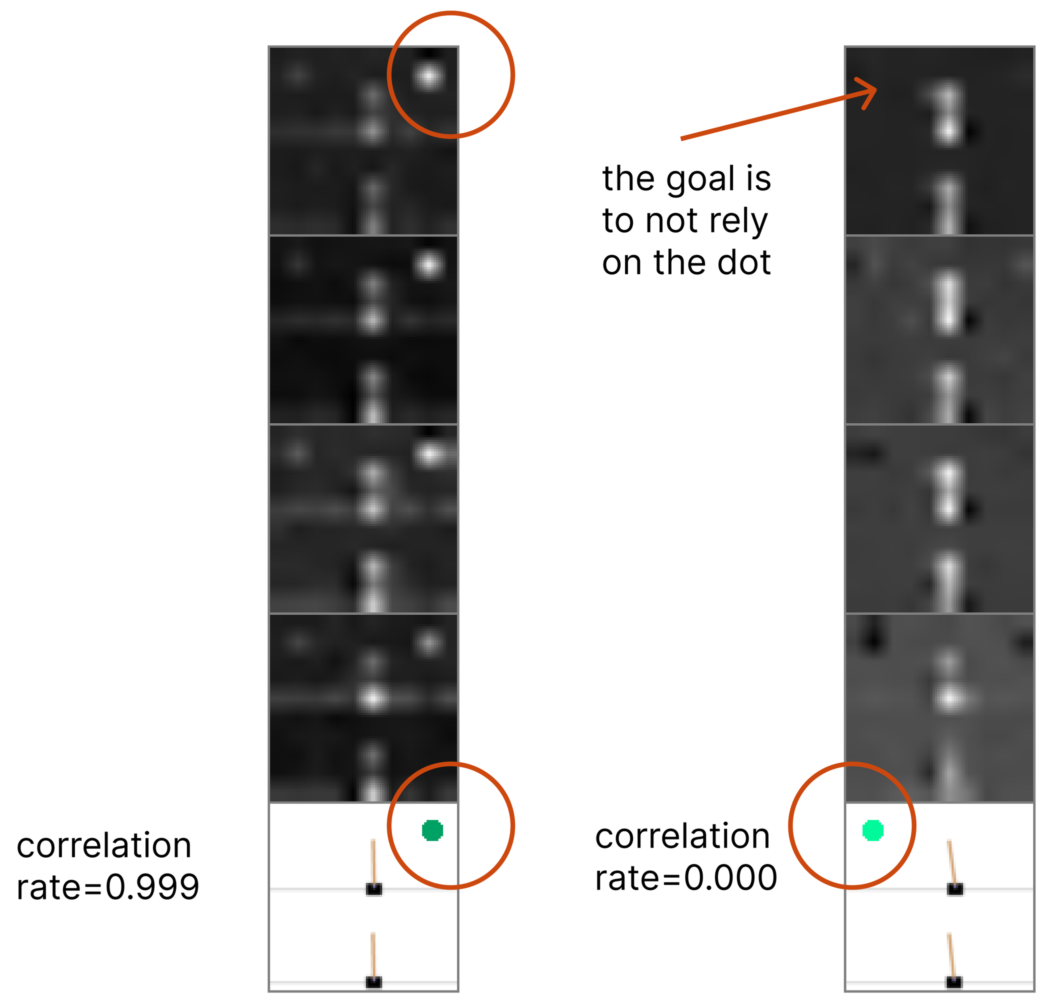


Figure 10: DQN with ViT backbone trained in a biased setting (left) and in an unbiased setting (right).

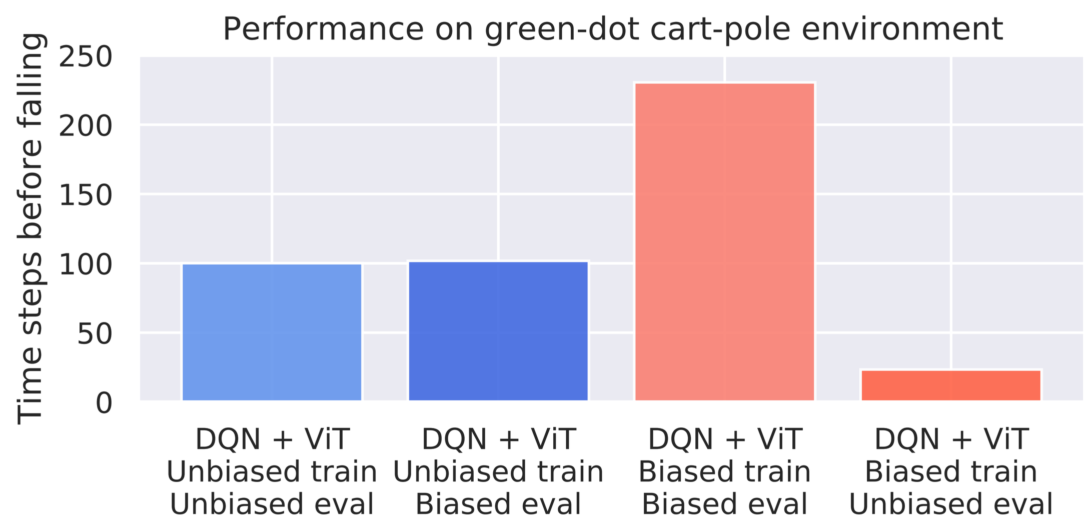
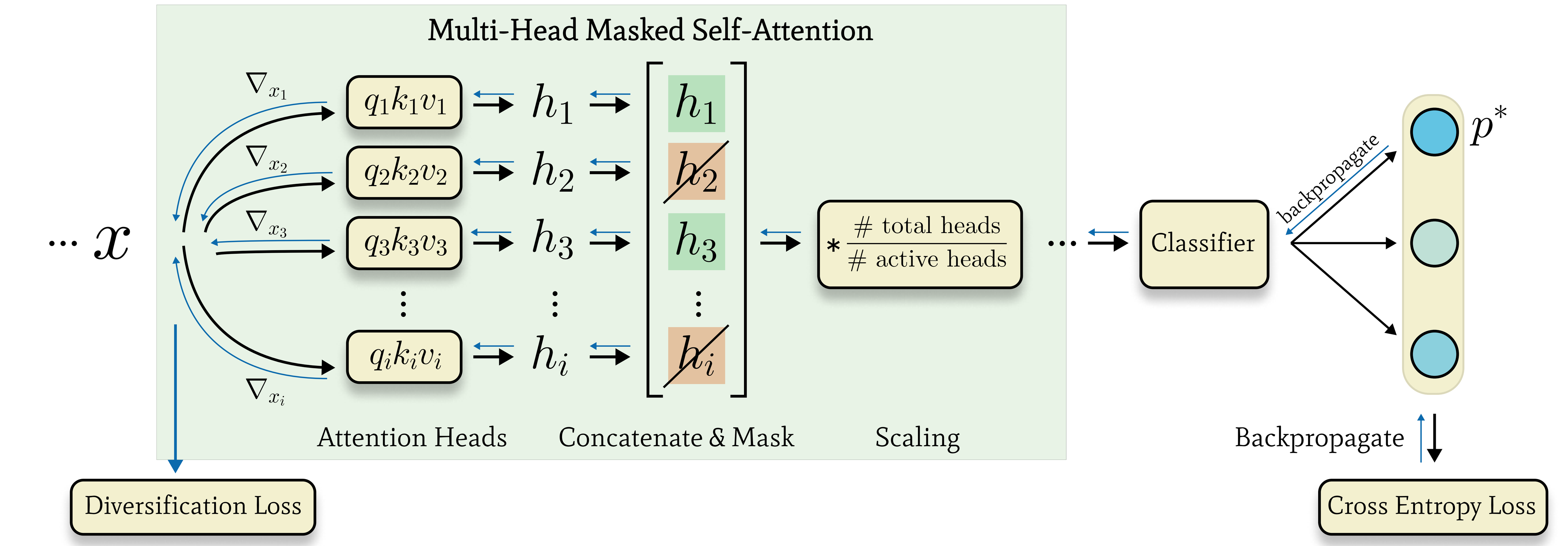


Figure 11: DQN + ViT performance for ERM training.

Proposed Method



Head selection:

1. Compute QKV self-attention:

$$h_i = \text{softmax}\left(\frac{W_{q_i}x(W_{k_i}x)^T}{\sqrt{d_k}}\right)W_{v_i}x$$

2. Mask a subset of the heads (change value to 0) and concatenate results.
3. Scale output to compensate masked heads.

Diversity Loss:

1. Compute Input Gradient (similar to [3]) through each attention head, defined as the gradient of the top prediction p^* w.r.t the shared input x.
2. Add orthogonality of input gradients to the training objective, to promote head specialization.

Mathematical details:

$$\nabla_{x_i} = \frac{\partial p*}{\partial x} \in \mathbb{R}^{N \times D}$$

$$c_{n,i,j} = \nabla_{x_{i,n}}^T \nabla_{x_{j,n}} \in \mathbb{R}$$

$$\mathcal{L}_{IG} = \frac{1}{N} \sum_{i \neq j} \sum_{n=1}^{N} c_{n,i,j}^2$$

$$\mathcal{L} = \mathcal{L}_{ERM} + \lambda \mathcal{L}_{IG}$$

Experimental Results

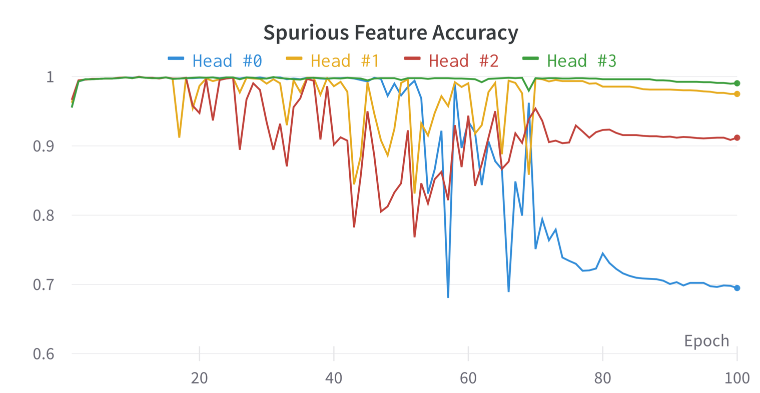


Figure 1: Spurious feature accuracy for single head (ERM training).

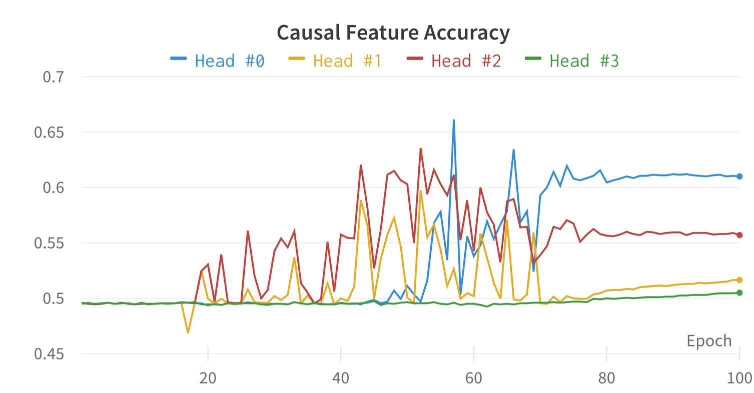


Figure 2: Causal feature accuracy for single head (ERM training).

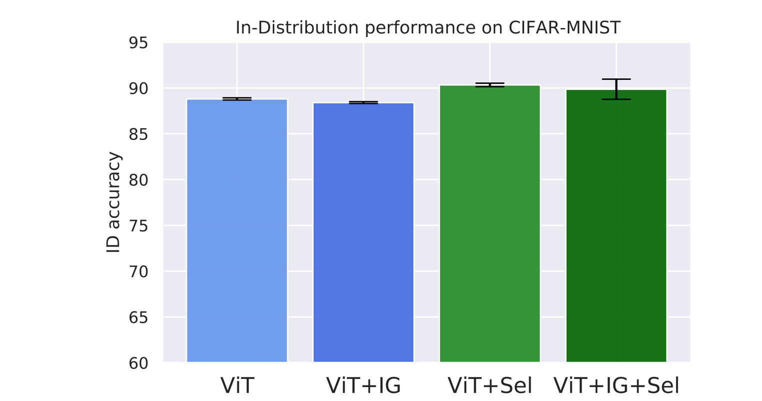


Figure 3: In-Distribution performance comparison.

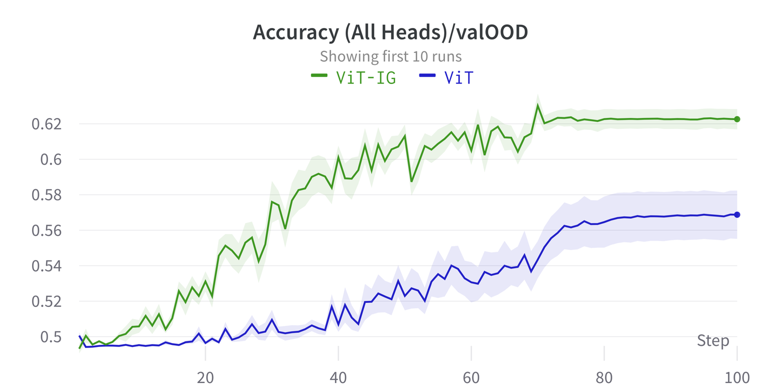


Figure 4: OOD benefits of diversification objective.

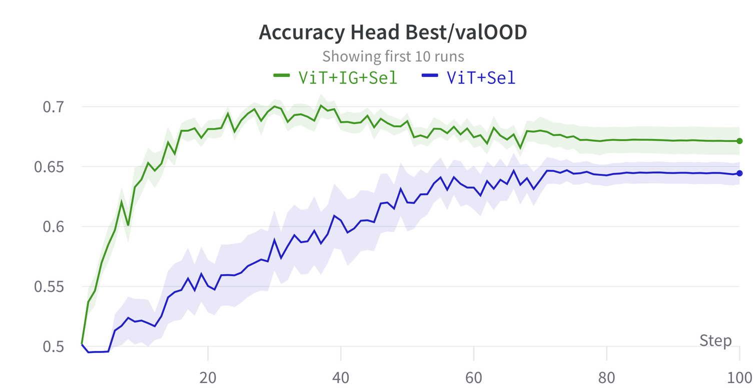


Figure 5: OOD benefits of best head selection and diversification.

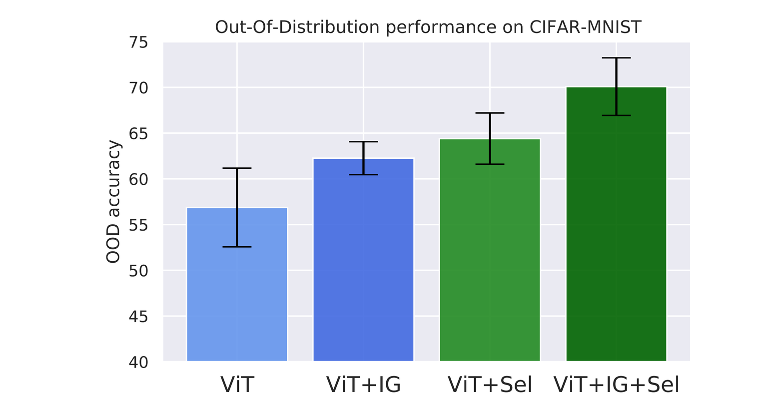


Figure 6: Out-of-Distribution performance comparison.

References

[1] Dosovitskiy et al. 2020,
[2] Reddy et al. NeurIPS 2021,
[3] Teney et al. CVPR 2022
[4] Lee, Yao & Finn 2022

Contact

Email: armand.nicolicioiu@gmail.com
Website: armandnicolicioiu.github.io