Learning Diverse Features in Vision Transformers for Improved Generalization

Armand Nicolicioiu[1] Andrei Nicolicioiu[2] Bogdan Alexe[3] Damien Teney[1]

[1]Idiap Research Institute [2]Mila - Quebec AI Institute [3]University of Bucharest

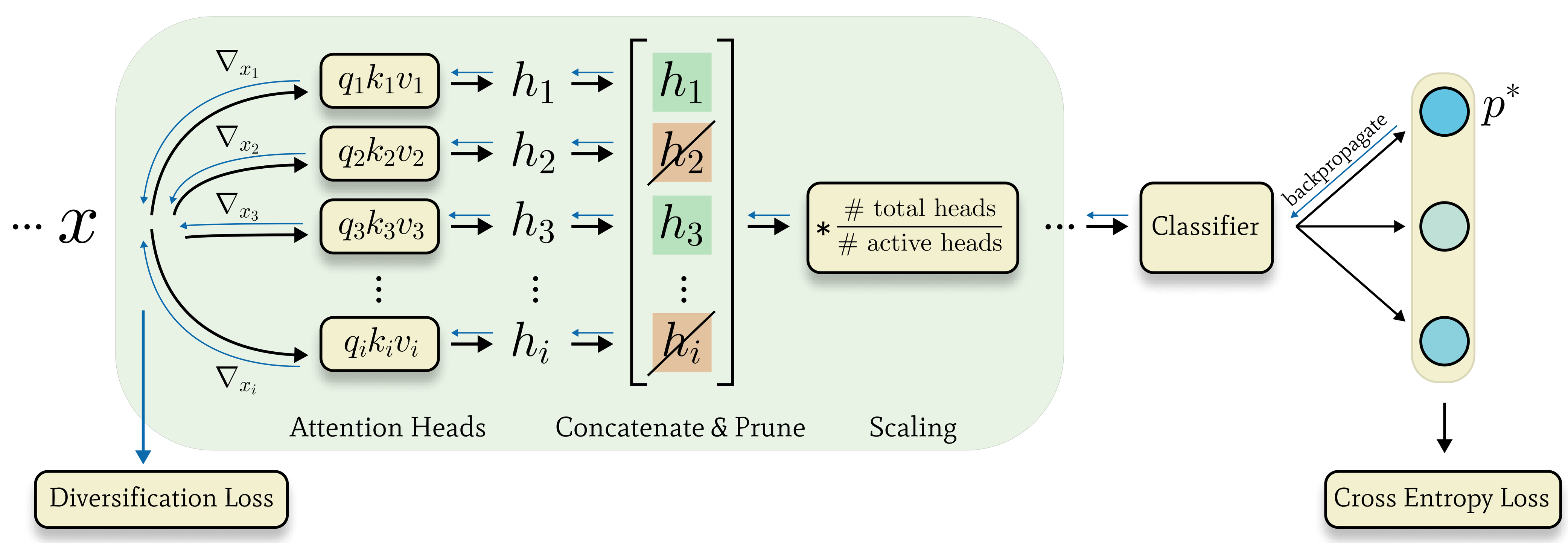
Introduction

- Deep learning models can be accurate on standard in-domain data while generalizing poorly **out of distribution (OOD)**.
- Training data oftens contains a mixture of **robust** and **spurious** predictive features.
- To improve OOD **generalization**, we want to better control over the features a model learns and relies on.

Contributions

- We investigate ViTs' [1] inherent property for modularity in the features learned by each attention head.
- We show that "oracle selection" of attention heads (pruning those corresponding to spurious features) can significantly improve OOD performance.
- We propose a **head diversification** method based on orthogonality of head influence, leading to better head specialization.

Proposed method



Head selection:

1. Compute QKV self-attention:

$$h_i = \mathrm{softmax}\left(\frac{Q_i x (K_i x)^T}{\sqrt{d_K}}\right) V_i x$$

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

Diversity regularizer:
(Applying [2] to ViTs' heads)

1. Compute the input gradient 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:

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

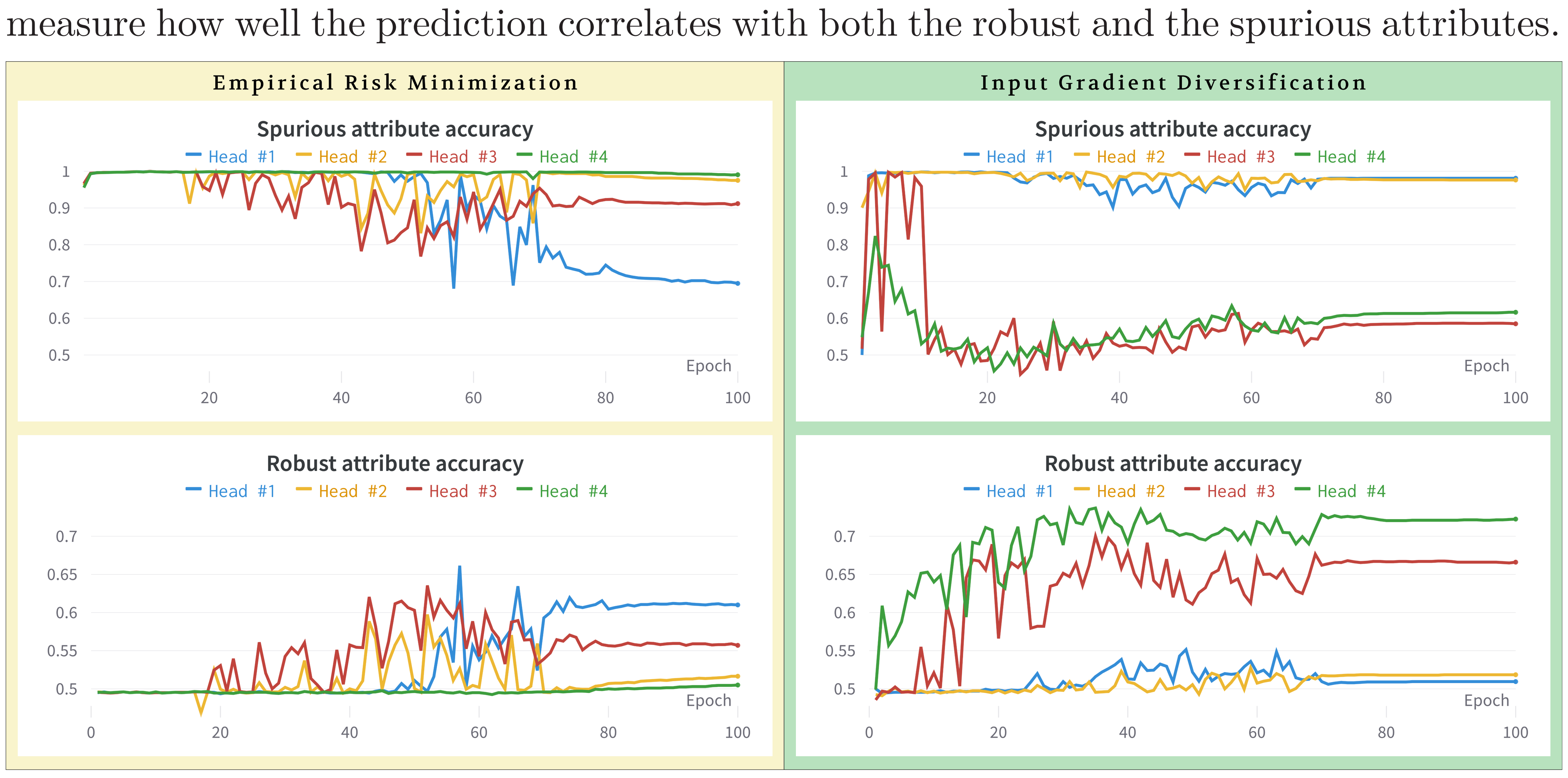
$$c_{n,i,j} = \nabla_{x_{i,n}}^T \nabla_{x_{j,n}} \in \mathbb{R}$$

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

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

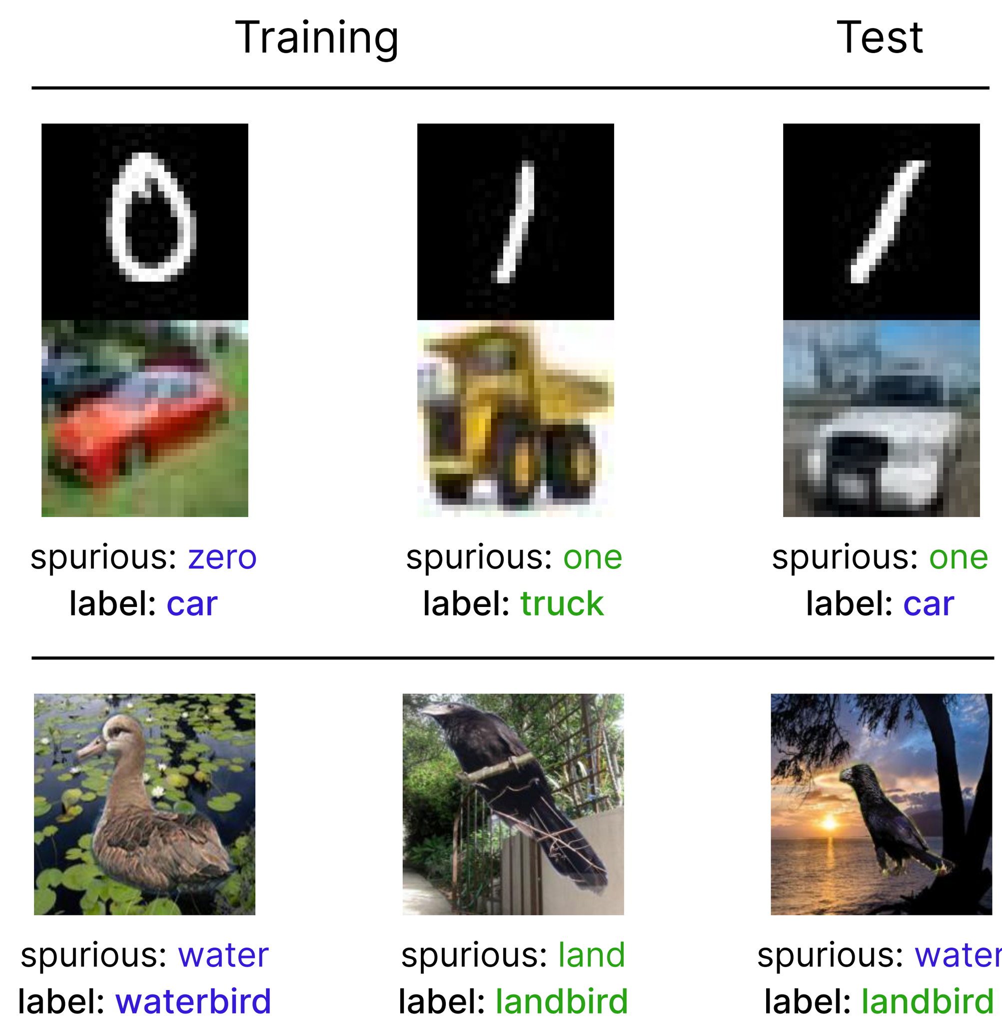
Diversity of the learned features

We perform an analysis to diagnose the diversity of the features learned by each head. We measure how well the prediction correlates with both the robust and the spurious attributes.



Per-head performance comparison on MNIST-CIFAR. The heads predicting well on the robust attribute are predicting poorly on the spurious one, and vice-versa.

Datasets: MNIST-CIFAR, Waterbirds



Results

Table 1: Results on MNIST-CIFAR.

Method	ID Acc.	OOD Acc.
ViT+ERM	88.80 ± 0.1	56.87 ± 4.3
ViT+Div	88.40 ± 0.1	62.26 ± 1.8
ViT+ERM+Sel	**90.33** ± 0.1	64.40 ± 2.8
ViT+Div+Sel	89.86 ± 1.1	**70.08** ± 3.1

Table 2: Results on Waterbirds.

Method	ID Acc.	OOD Acc.
ViT+ERM	96.55 ± 0.2	83.37 ± 0.4
ViT+Div	96.99 ± 0.1	83.87 ± 0.7
ViT+ERM+Sel	96.50 ± 0.5	85.70 ± 1.6
ViT+Div+Sel	**96.99** ± 0.1	**87.96** ± 0.1

Diversity and pruning of Vision Transformer's attention heads improve generalization.

Take-aways

- **The attention heads learn distinct features.**
- **Pruning heads corresponding to spurious features can significantly improve OOD performance.**
- **We still need OOD information to select the heads post-training.**

Contact

E-mail: armand.nicolicioiu@gmail.com

Implementation publicly available:

[1] Alexey Dosovitskiy et al. "An image is worth 16x16 words: Transformers for image recognition at scale". ICLR 2021.
[2] Damien Teney et al. "Evading the Simplicity Bias: Training a Diverse Set of Models Discovers Solutions with Superior OOD Generalization". CVPR 2022.