Learning Diverse Features in Vision Transformers for Improved Generalization

Armand Mihai Nicolicioiu [1] **Andrei Liviu Nicolicioiu** [2 3] **Bogdan Alexe** [4 5] **Damien Teney** [1]

Abstract

Deep learning models often rely only on a small set of features even when there is a rich set of predictive signals in the training data. This makes models brittle and sensitive to distribution shifts.

In this work, we first examine vision transformers (ViTs) and find that they tend to extract robust and spurious features with distinct attention heads. As a result of this modularity, their performance under distribution shifts can be significantly improved at test time by pruning heads corresponding to spurious features, which we demonstrate using an "oracle selection" on validation data.

Second, we propose a method to further enhance the diversity and complementarity of the learned features by encouraging orthogonality of the attention heads' input gradients. We observe improved out-of-distribution performance on diagnostic benchmarks (MNIST-CIFAR, Waterbirds) as a consequence of the enhanced diversity of features and the pruning of undesirable heads.

1. Introduction

State-of-the-art models in machine learning show their limits when faced with distribution shifts. Even though they can perform remarkably well when training and test data are drawn from the same distribution, their predictive performance can degrade dramatically otherwise. The reason is that features that are predictive in the training data may be spurious and misleading at test time. Out-of-distribution (OOD) generalization is the capability of a model to maintain its predictive performance in the face of such shifts.

[1]Idiap Research Institute, Martigny, Switzerland [2]Mila, Montreal, Canada [3]University of Montreal, Canada [4]University of Bucharest, Romania [5]Gheorghe Mihoc - Caius Iacob Institute of Mathematical Statistics and Applied Mathematics of the Romanian Academy. Correspondence to: Armand Mihai Nicolicioiu <armand.nicolicioiu@gmail.com>.

Published at the ICML 2023 Workshop on Spurious Correlations, Invariance, and Stability. Honolulu, Hawaii, USA.

Prior work has shown that deep learning models often rely only on a small set of predictive features (Geirhos et al., 2020). If any of these features are spurious and affected by a distribution shift, chances are high that a model's performance will be affected. A recent line of work seeks to increase the diversity of the learned features, either as an objective when training a predictive model (Ross et al., 2020; Teney et al., 2022a; Lee et al., 2022) or as a step prior to the application of invariance-learning methods (Chen et al., 2023b; Zhang et al., 2022).

This paper focuses on computer vision tasks and vision transformers (ViT) (Dosovitskiy et al., 2020). We apply a regularizer based on input gradients (Ross et al., 2020; Teney et al., 2022a) to a ViTs' attention heads to diversify the features learned across these heads. This encourages different parts of the model to rely on different aspects of the data and to discover additional predictive patterns. In contrast to methods that diversify functional behaviour in prediction space (Lee et al., 2022; Chen et al., 2023a), our approach operates in feature space and does not require any OOD data (even unlabeled) during training.

This paper presents early experiments on standard OOD benchmarks (MNIST-CIFAR, Waterbirds). First, we find that **ViTs already have an inherent property for modularity**: their attention heads rely each on different features, such that they can be pruned selectively to discard spurious ones and improve generalization. Second, we show that **the proposed regularizer can further increase the diversity and complementarity of the learned features**. Our method, *DiverseViT* (see Figure 1), leads to improvements in a standard OOD evaluation setting, and even more so when we allow pruning attention heads at test time using for selection the highest accuracy obtained on a labeled validation set from the target distribution.[1]

Summary of contributions.

- We evaluate off-the-shelf ViTs on diagnostic OOD benchmarks and find inherent modularity in their representations, such that OOD generalization can be improved by pruning the attention heads that rely on spurious features.

[1]This setting provides an upper bound on the performance achievable with ideal model selection heuristics.

- We propose a simple regularizer to increase the diversity of features learned by ViTs.
- We evaluate models trained with our method (DiverseViT) and observe increased diversity and complementarity of the learned features. They show better OOD performance in standard evaluations, and yet much higher performance when pruning specific attention heads at test time.

2. Related Work

Diversity of solutions in machine learning. A range of methods have been proposed to train models that are diverse in properties such as OOD generalization (Lee et al., 2022; Teney et al., 2022b), interpretability (Chen et al., 2023c; Semenova et al., 2022), or fairness. These diversification methods are relevant in cases of underspecification (D'Amour et al., 2020) when the standard ERM objective (empirical risk minimization) does not constrain the solution space to a unique one.

Increasing diversity. Most diversification methods train a set of multiple *entire* models. By contrast, our approach seeks to increase diversity within a single transformer. Existing methods train multiple models in parallel or sequentially. They encourage diversity in **feature space** (Heljakka et al., 2022; Yashima et al., 2022), **prediction space** (Pagliardini et al., 2022; Lee et al., 2022), or **gradient space** (Ross et al., 2018; 2020; Teney et al., 2022a;b). Our method applies a gradient-based approach similar to that of Ross et al. (2020); Teney et al. (2022a) to the attention heads of ViTs.

Vision transformers (ViTs). We focus on ViTs because they archieve state-of-the-art performance for multiple tasks in computer vision (Khan et al., 2022). Multiple works have sought to understand the features learned by these models (Bhojanapalli et al., 2021; Naseer et al., 2021; Zhou et al., 2022). Our findings are complementary. We also seek to nudge the (generally beneficial) inductive biases of ViTs. In particular, we seek to overcome the general "simplicity bias" of deep learning models (Shah et al., 2020).

3. Proposed method

We describe a simple method to diversify the features learned by ViTs trained for image classification. We propose a regularizer that encourages orthogonality of the input gradients corresponding to their attention heads. We show empirically that this provides a better inherent robustness to distribution shifts. Moreover, this allows further improvements in OOD performance with test-time pruning of the attention heads that correspond to spurious features, while retaining those necessary for robust classification.

3.1. Background: ViTs and attention heads

The input image to a ViT is partitioned into a grid of small patches. A sequence of tokens is formed by combining the patches with positional embeddings. The main operation to aggregate a sequence of tokens is the multi-head self-attention, defined as:

$$h_i = \text{softmax}\left(\frac{xW_q(xW_k)^T}{\sqrt{D}}\right)xW_v \tag{1}$$

where $x \in \mathbb{R}^{N\times D}$ represents the set of N input tokens of the self-attention layer and $W_q, W_k, W_v \in \mathbb{R}^{D\times D}$ are learnable parameters of the layer.

The output of all heads are concatenated and the result is projected with a linear mapping:

$$y = [h_1, ..., h_H]W_o \tag{2}$$

with $W_o \in \mathbb{R}^{D\times H*D}$ learnable parameters.

3.2. Encouraging feature diversity with input gradients

Our diversification method is a regularizer term added to the optimization objective when training a ViT for a supervised task with standard ERM. This approach increases the diversity of features within a single transformer, which contrasts with many diversification methods that train multiple entire models. The motivation for our approach is (1) its lower computational cost and (2) leveraging the existing inductive biases of vision transformers to avoid the type of "adversarial" solutions to the gradient-based regularizer that were described in Teney et al. (2022b).

Input gradients. To determine how much each dimension of a feature vector contributes to the prediction of the model, we look at the gradient of the prediction with respect to this vector. Concretely, we compute the gradient of the top predicted score p^* with respect to the input x:

$$\nabla_x = \frac{\partial p^*}{\partial x} \in \mathbb{R}^{N\times D}. \tag{3}$$

Influence of each head. The outputs of all attention heads are concatenated and projected (Equation 2), so ∇_x considers all attention heads' effects simultaneously. As we are interested in diversifying the effect of each individual head, we want to capture their individual contributions. For this, we backpropagate the gradient of the top prediction (Equation 3) H times, each time through a single element h_i of Equation 2 while ignoring the rest. We obtain a set of H input gradient:

$$\Big\{\nabla_{x_i} \in \mathbb{R}^{N\times D} \mid i \in \{1 \cdots H\}\Big\}. \tag{4}$$

Each element in this set represents the importance of the input features to the trop prediction for a specific head, with $\sum_{i=1}^{H} \nabla_{x_i} = \nabla_x$.

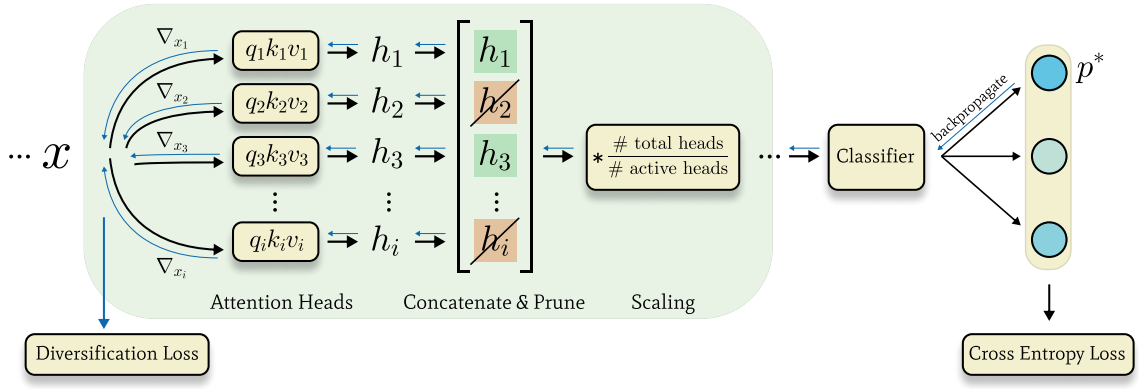

Figure 1. **Overview of the proposed DiverseViT method.** We depict a single-layer multi-head attention model.

Diversity regularizer. To promote diversity across the heads, we define an orthogonality regularizer over the input gradients. We first compute the cosine similarity between the nth token in heads i and j by normalizing over the channel dimension D and taking the dot product between all tokens:

$$c_{n,i,j} = \nabla^T_{x_{i,n}} \nabla_{x_{j,n}} \in \mathbb{R}. \tag{5}$$

The orthogonality regularizer is then defined as the average squared similarity across all tokens and pairs of heads:

$$\mathcal{L}_{Div} = \frac{1}{N} \sum_{i \neq j} \sum_{n=1}^{N} c^2_{n,i,j}. \tag{6}$$

The regularizer, weighted by a hyperparameter λ, is added to the standard cross-entropy classification loss to form the complete training objective:

$$\mathcal{L} \;=\; \mathcal{L}_{ERM} + \lambda \mathcal{L}_{Div}. \tag{7}$$

3.3. Pruning attention heads

Different attention heads of a transformer can attend to different features which can be more or less robust (i.e. useful across distribution shifts) or spurious. Therefore, using only a subset of heads is a simple technique for ignoring undesirable features and improving OOD performance.

To prune a subset of the heads, we multiply their output h_i by zero in Equation 2. To compensate for the missing heads, we scale the remaining ones to remain in the same range after the projection. This can be done at test time with no need for further adaptation of the model.

4. Experiments

We present experiments on two popular diagnostic datasets: MNIST-CIFAR and Waterbirds. See Appendix A for details. Our experiments answer the following questions:

1. Does diversifying the learned features of the self-attention heads lead to better OOD performance compared with ERM training? **Yes.**
2. Can we improve generalization by pruning heads associated with spurious features with a standard ERM-trained ViT (i.e. without our diversification method)? **Yes.**
3. Does our diversification method amplify the distinction between spurious and robust features, improving post-pruning performance even more? **Yes.**

The **baseline** experiment (VIT+ERM) trains a ViT with ERM. In both datasets used, a spurious feature is strongly correlated with the label during training. This correlation is reversed at test time. A model that relies on this spurious feature will therefore perform poorly at test time.

In the **diversification** experiment (VIT+DIV), we add our diversity regularizer to the training objective. Without any changes in the architecture, we obtain better generalization, possibly as an ensemble effect of a more diverse set of features captured collectively by all attention heads.

For the **head selection** experiments (SEL), we perform inference at test time using a single attention head. We select the head with the highest accuracy on the OOD validation data. This therefore requires access to labeled OOD examples. The pruning is a test-time procedure and requires no further training of the model. In all experiments, we select hyperparameters for highest OOD validation accuracy.

Main results. In Tables 1–2 we report accuracy on both ID (in-domain) and OOD (out-of-distribution) test sets. We observe that the ERM baseline already exhibits a degree of separation between spurious and robust features among the self-attention heads, thus allowing the head selection (*ViT+ERM+Sel*) to improve the OOD accuracy. The diversification method (*ViT+Div*) is effective on its own (i.e. even without head selection). This indicates a benefit from learning diverse features. The combination of diverse features and proper head selections (*ViT+Div+Sel*) is especially powerful, leading to our best result.

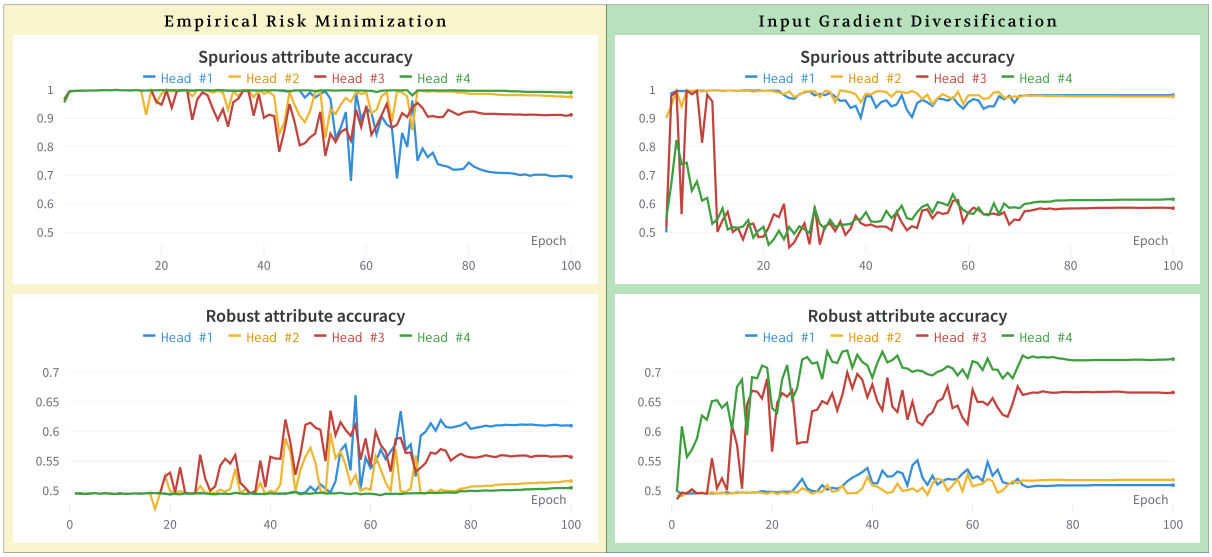

Figure 2. **Per-head performance comparison between standard ERM and our Diversification method on MNIST-CIFAR.** We observe an inherent modularity of ViT's attention heads, such that the heads predicting well on the robust attribute are predicting poorly on the spurious one, and vice-versa. With standard ERM, the gap between the two attributes predictions is not as clear for most heads, indicating a higher overlap in the information captured by each head. With the proposed diversification method, we observe that most heads predict either, but not both of the robust and spurious features. This shows a high level of specialization, which is desirable since this allows pruning undesirable heads without losing information relevant to robust predictions.

Table 1. Results on MNIST-CIFAR.

METHOD	ID ACCURACY	OOD ACCURACY
VIT+ERM	88.80 ± 0.12	56.87 ± 4.30
VIT+DIV	88.40 ± 0.10	62.26 ± 1.80
VIT+ERM+SEL	**90.33** ± 0.19	64.40 ± 2.80
VIT+DIV+SEL	89.86 ± 1.10	**70.08** ± 3.15

Table 2. Results on Waterbirds.

METHOD	ID ACCURACY	OOD ACCURACY
VIT+ERM	96.55 ± 0.22	83.37 ± 0.44
VIT+DIV	96.99 ± 0.11	83.87 ± 0.79
VIT+ERM+SEL	96.50 ± 0.58	85.70 ± 1.64
VIT+DIV+SEL	**96.99** ± 0.12	**87.96** ± 0.14

Understanding the learned features To gain insight into the learned features, we evaluate the models on a balanced split of MNIST-CIFAR with no correlation between the robust and the spurious attributes. We measure the correlation between the model's predictions and either the robust or the spurious attribute. How well each attribute can be predicted with each head indicates how much of each attribute is captured by each head. Figure 2 shows that diversification leads to a higher level of specialization of the heads. This is advantageous, allowing us to keep only the heads containing information relevant to robust predictions.

5. Conclusions and future work

We have shown that ViT have an inherent tendency to capture distinct features in their attention heads, and that this property can be improved with a simple regularizer. This directly improves the robustness of ViTs on several diagnostic benchmarks for out-of-distribution generalization, without changing the architecture. These improvements come "for free" as an ensembling effect from the increased diversity of the learned features.

We also empirically showed that these diverse features have little overlap and are complementary. Therefore, pruning selected attention heads at test time is an effective technique to improve OOD performance, using some information that can identify which heads correspond to spurious features (e.g. an OOD validation set).

Future work. Our methods should be evaluated on larger-scale real-world datasets in vision, but also language and reinforcement learning. The head pruning procedure was evaluated using labeled OOD data, which is only meant to provide an upper bound on the performance achievable in an ideal setting. The approach should be evaluated with the recent heuristics proposed for OOD model selection (Baek et al., 2022; Deng et al., 2023; Garg et al., 2022; Liu et al., 2023; Lu et al., 2022). Other options include various forms of human feedback and unsupervised objectives to enable test-time adaptation.

Acknowledgements

Bogdan Alexe was funded by UEFISCDI under Project EEA-RO-2018-0496. Damien Teney was partially supported by an Amazon Research Award (ARA).

References

Baek, C., Jiang, Y., Raghunathan, A., and Kolter, J. Z. Agreement-on-the-line: Predicting the performance of neural networks under distribution shift. *Advances in Neural Information Processing Systems*, 35:19274–19289, 2022.

Bhojanapalli, S., Chakrabarti, A., Glasner, D., Li, D., Unterthiner, T., and Veit, A. Understanding robustness of transformers for image classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10231–10241, 2021.

Chen, A. S., Lee, Y., Setlur, A., Levine, S., and Finn, C. Project and probe: Sample-efficient domain adaptation by interpolating orthogonal features. *arXiv preprint arXiv:2302.05441*, 2023a.

Chen, Y., Huang, W., Zhou, K., Bian, Y., Han, B., and Cheng, J. Towards understanding feature learning in out-of-distribution generalization. *arXiv preprint arXiv:2304.11327*, 2023b.

Chen, Z., Zhong, C., Seltzer, M., and Rudin, C. Understanding and exploring the whole set of good sparse generalized additive models. *arXiv preprint arXiv:2303.16047*, 2023c.

D'Amour, A., Heller, K., Moldovan, D., Adlam, B., Alipanahi, B., Beutel, A., Chen, C., Deaton, J., Eisenstein, J., Hoffman, M. D., et al. Underspecification presents challenges for credibility in modern machine learning. *arXiv preprint arXiv:2011.03395*, 2020.

Deng, W., Suh, Y., Gould, S., and Zheng, L. Confidence and dispersity speak: Characterising prediction matrix for unsupervised accuracy estimation. *arXiv preprint arXiv:2302.01094*, 2023.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Garg, S., Balakrishnan, S., Lipton, Z. C., Neyshabur, B., and Sedghi, H. Leveraging unlabeled data to predict out-of-distribution performance. *arXiv preprint arXiv:2201.04234*, 2022.

Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., and Wichmann, F. A. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

Heljakka, A., Trapp, M., Kannala, J., and Solin, A. Representational multiplicity should be exposed, not eliminated. *arXiv preprint arXiv:2206.08890*, 2022.

Khan, S., Naseer, M., Hayat, M., Zamir, S. W., Khan, F. S., and Shah, M. Transformers in vision: A survey. *ACM computing surveys (CSUR)*, 54(10s):1–41, 2022.

Lee, Y., Yao, H., and Finn, C. Diversify and disambiguate: Learning from underspecified data. *arXiv preprint arXiv:2202.03418*, 2022.

Liu, Y., Tian, C. X., Li, H., Ma, L., and Wang, S. Neuron activation coverage: Rethinking out-of-distribution detection and generalization. *arXiv preprint arXiv:2306.02879*, 2023.

Lu, W., Wang, J., Wang, Y., Ren, K., Chen, Y., and Xie, X. Towards optimization and model selection for domain generalization: A mixup-guided solution. *arXiv preprint arXiv:2209.00652*, 2022.

Naseer, M. M., Ranasinghe, K., Khan, S. H., Hayat, M., Shahbaz Khan, F., and Yang, M.-H. Intriguing properties of vision transformers. *Advances in Neural Information Processing Systems*, 34:23296–23308, 2021.

Pagliardini, M., Jaggi, M., Fleuret, F., and Karimireddy, S. P. Agree to disagree: Diversity through disagreement for better transferability. *arXiv preprint arXiv:2202.04414*, 2022.

Ross, A., Pan, W., Celi, L., and Doshi-Velez, F. Ensembles of locally independent prediction models. In *Proc. Conf. AAAI*, 2020.

Ross, A. S., Pan, W., and Doshi-Velez, F. Learning qualitatively diverse and interpretable rules for classification. *arXiv preprint arXiv:1806.08716*, 2018.

Sagawa, S., Koh, P. W., Hashimoto, T. B., and Liang, P. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019.

Semenova, L., Rudin, C., and Parr, R. On the existence of simpler machine learning models. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1827–1858, 2022.

Shah, H., Tamuly, K., Raghunathan, A., Jain, P., and Netrapalli, P. The pitfalls of simplicity bias in neural networks. *arXiv preprint arXiv:2006.07710*, 2020.

Teney, D., Abbasnejad, E., Lucey, S., and van den Hengel, A. Evading the simplicity bias: Training a diverse set of models discovers solutions with superior OOD generalization. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2022a.

Teney, D., Peyrard, M., and Abbasnejad, E. Predicting is not understanding: Recognizing and addressing underspecification in machine learning. *arXiv preprint arXiv:2207.02598*, 2022b.

Yashima, S., Suzuki, T., Ishikawa, K., Sato, I., and Kawakami, R. Feature space particle inference for neural network ensembles. *arXiv preprint arXiv:2206.00944*, 2022.

Zhang, J., Lopez-Paz, D., and Bottou, L. Rich feature construction for the optimization-generalization dilemma. In *International Conference on Machine Learning*, pp. 26397–26411. PMLR, 2022.

Zhou, D., Yu, Z., Xie, E., Xiao, C., Anandkumar, A., Feng, J., and Alvarez, J. M. Understanding the robustness in vision transformers. In *International Conference on Machine Learning*, pp. 27378–27394. PMLR, 2022.

A. Datasets

MNIST-CIFAR The MNIST-CIFAR dataset (Shah et al., 2020) contains collages of an image from MNIST (digits 0 and 1) and an image from CIFAR (a car or a truck) vertically concatenated. The true label comes from the vehicle type, and the digit is a spurious attribute highly correlated with the label during training. In our experiments, 90% of examples containing the digit 0 pair it with a car, and 90% of examples containing the digit 1 pair it with a truck. During the evaluation, this correlation no longer holds, and the images are evenly distributed.

Waterbirds The Waterbirds dataset (Sagawa et al., 2019) is constructed by adding a segmented image of a bird from the Caltech-UCSD Birds dataset over a background image from the Places dataset. The true label is the bird type (waterbird or landbird) and the spurious attribute is the background (water or land). In the training data, there are 3498 waterbirds on a water background and 1057 landbirds on a land background, while the minority groups contain only 184 waterbirds on a land background and 56 landbirds on a water background.

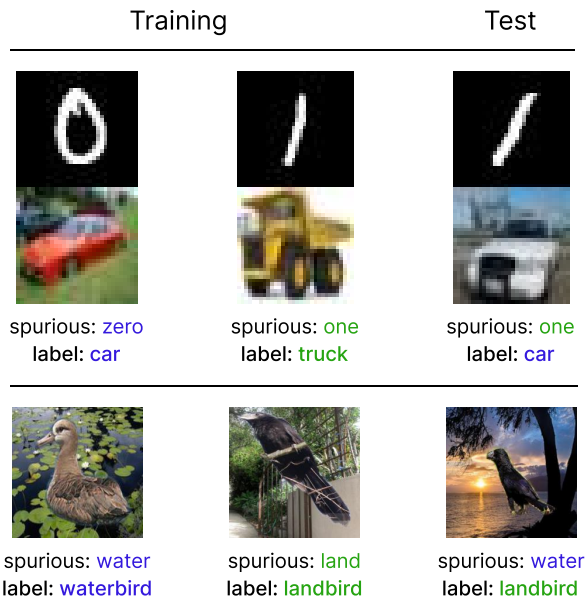

Figure 3. Samples from MNIST-CIFAR dataset (top) and Waterbirds (bottom).