Are Vision Transformers Robust to Spurious Correlations ?

Anonymous Authors[1]

Abstract

Deep neural networks may be susceptible to learning spurious correlations that hold on average but not in atypical test samples. As with the recent emergence of vision transformer (ViT) models, it remains underexplored how spurious correlations are manifested in such architectures. In this paper, we systematically investigate the robustness of vision transformers to spurious correlations on three challenging benchmark datasets and compare their performance with popular CNNs. Our study reveals that when pre-trained on a sufficiently large dataset, ViT models are more robust to spurious correlations than CNNs. Key to their success is the ability to generalize better from the examples where spurious correlations do not hold.

1. Introduction

A key challenge in building robust image classification models is the existence of *spurious correlations*: misleading heuristics imbibed within the training dataset that are correlated with majority examples but do not hold in general. Prior works have shown that convolutional neural networks (CNNs) can rely on spurious features to achieve high average test accuracy. Yet, such models lead to low accuracy on rare and untypical test samples lacking those heuristics (Geirhos et al., 2019; Goel et al., 2021; Sagawa et al., 2020; Tu et al., 2020). In Figure 3 (Appendix), we illustrate a model setup that exploits the spurious correlation between the `water background` and label `waterbird` for prediction. While the robustness of CNNs has been widely studied, it remains underexplored how spurious correlation is manifested in the recent development of vision transformers (ViT) (Dosovitskiy et al., 2021). As with the paradigm shift to attention-based architectures, it becomes increasingly critical to understand their behavior under ill-conditioned data. From a network architecture perspective, ViTs lack the inductive bias in CNNs, such as translational equivariance and spatial locality, and may be more prone to overfitting (Dosovitskiy et al., 2021). For this reason, one may expect the fully-connected dependencies in ViT models may exacerbate capturing the spurious correlations in the training data. In this paper, we seek to answer the following question: *Are Vision Transformers more robust to spurious correlations compared to CNNs*? Motivated by the question, we systematically investigate how and when ViT models exhibit robustness to spurious correlations on challenging benchmarks. Our findings reveal that for transformers, larger models and more pre-training data yield a significant improvement in robustness to spurious correlations. The key reason for success can be attributed to the ability to generalize better from those examples where spurious correlations do not hold, while fine-tuning. However, despite better generalization capability, ViT models suffer high errors on challenging benchmarks when these counterexamples are scarce in the training set. On the other hand, when pre-trained on a relatively smaller dataset such as ImageNet-1k, the performance of transformer-based models are much worse as compared to CNN counterparts. This indicates that in smaller pre-training data regimes, transformers have a higher propensity to overfit the spurious features and are less robust than CNNs of comparable size.

Our key contributions are summarized below:

(1) To the best of our knowledge, we provide first systematic study on robustness of Vision Transformers when learned on datasets containing spurious correlations.

(2) We perform extensive experiments and ablations to understand effect of model architectures, model capacity, pre-training dataset, data imbalance, fine-tuning, etc.

(3) We provide insights on ViT's robustness by analyzing the attention matrix, which encapsulates important information about the interaction among image patches. We hope that our work will inspire future research on further understanding the robustness of ViT models.

2. Preliminaries

2.1. Spurious Correlations

Spurious features refer to statistically informative features that work for majority of training examples but do not capture essential cues related to the labels (Geirhos et al., 2019;

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the SCIS workshop.

Model	Average Acc.	Worst-Group Acc.
ViT-B/16	**96.75**$_{\pm 0.05}$	**89.30**$_{\pm 1.95}$
ViT-S/16	96.30$_{\pm 0.51}$	85.45$_{\pm 1.16}$
ViT-Ti/16	89.50$_{\pm 0.05}$	71.65$_{\pm 0.16}$
BiT-M-R50x3	94.90$_{\pm 0.05}$	80.51$_{\pm 1.02}$
BiT-M-R101x1	94.05$_{\pm 0.07}$	77.50$_{\pm 0.50}$
BiT-M-R50x1	92.05$_{\pm 0.05}$	75.10$_{\pm 0.62}$

Table 1. Average and worst-group accuracies over test set for different models when finetuned on Waterbirds (Sagawa et al., 2020). All models are pre-trained on ImageNet-21k.

Goel et al., 2021; Sagawa et al., 2020; Tu et al., 2020). Formally, we consider a training set, $\mathcal{D}^{\text{train}}$, consisting of N training samples: $\{\mathbf{x}_i, y_i\}_{i=1}^N$, where samples are drawn independently from a probability distribution: $\mathcal{P}_{X,Y}$. Here, $X \in \mathcal{X}$ is a random variable defined in the pixel space, and $Y \in \mathcal{Y} = \{1, \ldots, K\}$ represents its label. We further assume that the data is sampled from a set of E environments $\mathcal{E} = \{e_1, e_2, \cdots, e_E\}$. The training data has spurious correlations, if the input $\mathbf{x}_i$ is generated by a combination of invariant features $\mathbf{z}_i^{inv} \in \mathbb{R}^{d_{inv}}$, which provides essential cues for accurate classification, and environmental features $\mathbf{z}_i^e \in \mathbb{R}^{d_e}$ dependent on environment e:

$$\mathbf{x}_i = \rho(\mathbf{z}_i^{inv}, \mathbf{z}_i^e).$$

Here ρ represents a function transformation from the feature space $[\mathbf{z}_i^{inv}, \mathbf{z}_i^e]^T$ to the pixel space $\mathcal{X}$. Under the data model, we form groups $g = (y, e) \in \mathcal{Y} \times \mathcal{E}$ that are jointly determined by the label y and environment e. For this study, we consider the binary setting where $\mathcal{E} = \{1, -1\}$ and $\mathcal{Y} = \{1, -1\}$, resulting in four groups. The concrete meaning for each environment and label will be instantiated in corresponding tasks, which we describe in Section 3.

2.2. Model Zoo

In this study, we aim to understand the robustness of ViT models when trained on a dataset containing spurious correlations and how they fare against popular CNNs. We contrast ViT with Big Transfer (BiT) models (Kolesnikov et al., 2020) that are primarily based on the ResNet-v2 architecture. For both ViT and BiT models, we consider different variants that differ in model capacity and pre-training dataset, as summarized in Table 4 (Appendix). Specifically, we use model variants pre-trained on both ImageNet-1k (Russakovsky et al., 2015) and on ImageNet-21k (Deng et al., 2009) datasets. To indicate input patch size in ViT models, we append "/x" to model names. We prepend -B, -S, -Ti to indicate `Base`, `Small` and `Tiny` version of the corresponding architecture.

3. Robustness to Spurious Correlation

In this section, we systematically measure the robustness performance of ViT models when trained on datasets containing spurious correlations, and compare how their robustness fares against popular CNNs. For evaluation benchmarks, we adopt the same setting as in (Sagawa et al., 2020). Specifically, we consider the following three classification datasets to study the robustness of ViT models in a spurious correlated environment: Waterbirds (Section 3.1), CelebA (Section 3.2), and ColorMNIST. Due to space constraints, results on ColorMNIST are in the Appendix.

3.1. Waterbirds

Introduced in (Sagawa et al., 2020), this dataset contains spurious correlation between background features and target label $y \in \{\texttt{waterbird}, \texttt{landbird}\}$. The dataset is constructed by selecting bird photographs from the Caltech-UCSD Birds-200-2011 (CUB) (Wah et al., 2011) dataset and then superimposing on either of $e \in \mathcal{E} = \{\texttt{water}, \texttt{land}\}$ background selected from the Places dataset (Zhou et al., 2017).

Results and insights on generalization performance. Table 1 compares worst-group test accuracies of different models when fine-tuned on Waterbirds (Sagawa et al., 2020) using empirical risk minimization. Note that all the compared models are pre-trained on ImageNet-21k. This allows us to isolate the effect of model architectures, in particular, ViT vs. BiT models. The worst-group test accuracy reflects the model's generalization performance for groups where the correlation between the label y and environment e does not hold. A high worst-group accuracy is indicative of less reliance on the spurious correlation in training. Our results suggest that: (1) ViTs are relatively more robust to spurious associations between background feature and target label than convolution-based BiTs. Interestingly, ViT-B/16 attains a significantly higher worst-group test accuracy (89.3%) than BiT-M-R50x3 despite having a considerably smaller capacity (86.1M vs. 211M). (2) Furthermore, these results reveal a correlation between generalization performance and model capacity. With an increase in model capacity, both ViTs and BiTs tend to generalize better, measured by both average accuracy and worst-group accuracy. The relatively poor performance of ViT-Ti/16 can be attributed to its failure to learn the intricacies within the dataset due to its compact capacity.

Results and insights on robustness performance. We now delve deeper into the robustness of ViT models. In particular, we investigate the robustness in model prediction under varying background features. Our key idea is to compare the predictions of image pairs $(\mathbf{x}_i, \bar{\mathbf{x}}_i)$ with the same foreground object yet different background features (i.e., `water` vs. `land` background). We define *Consistency Measure* of a model as the average number of consistent predictions on the evaluation dataset given the predictions are correct, i.e., $\frac{1}{N}\sum_{i=1}^N \mathbb{I}\{\hat{f}(\mathbf{x}_i) = \hat{f}(\bar{\mathbf{x}}_i) \mid \hat{f}(\mathbf{x}_i) = y_i\}$, where y_i denotes the target label. To generate the image pairs $(\mathbf{x}_i, \bar{\mathbf{x}}_i)$, we first take a foreground bird photograph

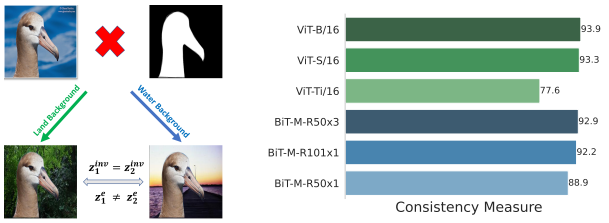

Figure 1. **Consistency Measure.** In Waterbirds dataset, y $\in$ {waterbird, landbird} is correlated with environment e $\in$ {water, land}. **Left**: Visual illustration of the experimental setup for measuring model consistency. Ideally, changing the spurious features ($\mathbf{z}^e$) should have no impact on model prediction. **Right**: Evaluation results quantifying consistency for models of different architectures and varying capacity.

using the pixel-level segmentation masks from the CUB dataset (Wah et al., 2011). We then place it on the top of water and land background images from the Places dataset (Zhou et al., 2017). We generate multiple such pairs to form the evaluation dataset $\{(\mathbf{x}_i, \bar{\mathbf{x}}_i)\}_{i=1}^N$ and use this dataset to quantify the robustness performance. For this study, the evaluation dataset consists of $N = 11788$ paired samples.

Figure 1 provides a visual illustration of the experimental setup (**left**), along with the evaluation results (**right**). Our operating hypothesis is that a robust model should predict same class label $\hat{f}(\mathbf{x}_i)$ and $\hat{f}(\bar{\mathbf{x}}_i)$ for a given pair $(\mathbf{x}_i, \bar{\mathbf{x}}_i)$, as they share exactly the same foreground object (i.e., invariant feature). Our results in Figure 1 show that ViT models achieve overall higher consistency measures than BiT counterparts. For example, the best model ViT-B/16 obtains consistent predictions for 93.9% of image pairs. Overall, using ViT pre-trained models yields strong generalization and robustness performance on Waterbirds.

3.2. CelebA

Beyond background spurious features, we further validate our findings on a different type of spurious feature. Here, we investigate the behavior of machine learning models when learned on training samples with spurious associations between target label and demographic information such as gender. Following (Ming et al., 2022), we use CelebA dataset, consisting of celebrity images with each image annotated using 40 binary attributes. We have the label space $\mathcal{Y} = \{\texttt{gray hair}, \texttt{nongray hair}\}$ and gender as the spurious feature, $\mathcal{E} = \{\texttt{male}, \texttt{female}\}$. The training data consists of 4010 images with label grey hair, out of which 3208 are male, resulting in spurious association between gender attribute male and label grey hair. Formally, $\mathbb{P}(e = \texttt{grey hair} | y = \texttt{male}) \approx \mathbb{P}(e = \texttt{non-grey hair} | y = \texttt{female}) \approx 0.8$.

Results. We see from Table 2 that ViT models achieve higher test accuracy (both average and worst-group) as opposed to BiTs. In particular, ViT-B/16 achieves +**4.3**% higher worst-group test accuracy than BiT-M-R50x3, despite having a considerably smaller capacity (86.1M vs. 211M). These findings along with our observations in Section 3.1 demonstrate that ViTs are not only more robust when there are strong associations between the label and background features, but also avoid learning spurious correlations between demographic features and target label.

Model	Average Acc.	Worst-Group Acc.
ViT-B/16	**97.40**$\pm$0.42	**94.10**$\pm$0.51
ViT-S/16	96.26$\pm$0.66	91.50$\pm$1.56
ViT-Ti/16	96.71$\pm$0.18	88.60$\pm$3.92
BiT-M-R50x3	97.31$\pm$0.05	89.80$\pm$0.42
BiT-M-R101x1	97.20$\pm$0.08	89.33$\pm$0.78
BiT-M-R50x1	96.82$\pm$1.20	87.72$\pm$1.56

Table 2. Average and worst-group accuracies over test set for different models when finetuned on CelebA (Liu et al., 2015). All models are pre-trained on ImageNet-21k.

4. Discussion: A Closer Look at ViT Under Spurious Correlation

In this section, we perform extensive ablations and experiments to understand the role of ViT models under spurious correlations. For consistency, we present the analyses below based on the Waterbirds dataset.

4.1. How does the size of the pre-training dataset affect robustness to spurious correlations?

In this section, we aim to understand the role of large-scale pre-training on the model's robustness to spurious correlations. To understand the importance of the pre-training dataset, we compare models pre-trained on ImageNet-1k (1.3 million images) and ImageNet-21k (12.8 million images). We report results for transformer-based models and BiT models in Table 3. For detailed ablation results on other benchmark datasets, please refer to the Appendix. Based on these results, we highlight the following observations:

(1) First, large-scale pre-training improves the performance of the models on challenging benchmarks. For transformers, larger models (base and small) and more pre-training data (ImageNet-21k) yields a significant improvement in all reported metrics. Hence, larger pre-training data and increasing model size play a crucial role in improving model robustness to spurious correlations. We also see a similar trend in the case of BiT models.

(2) Second, when pre-trained on a relatively smaller dataset such as ImageNet-1k, the performance of transformer-based DeiT models are much worse as compared to BiT-S models. Interestingly, although increasing size of DeiT models leads to improved average test accuracy but suffers high error on worst-group samples. This indicates that in smaller pre-

	Model	Test Accuracy		Consistency Measure↑
		Avg.	Worst-Group	
ImageNet-21k	ViT-B/16	96.8	89.3	93.9
	ViT-S/16	96.3	85.5	93.3
	ViT-Ti/16	89.5	71.7	77.6
ImageNet-1k	DeiT-B/16	85.9	44.6	71.9
	DeiT-S/16	84.5	46.7	74.3
	DeiT-Ti/16	83.4	41.8	71.1
ImageNet-21k	BiT-M-R50x3	94.9	80.5	92.9
	BiT-M-R101x1	94.1	77.5	92.2
	BiT-M-R50x1	92.1	75.1	88.9
ImageNet-1k	BiT-S-R50x3	87.0	60.3	77.8
	BiT-S-R101x1	87.3	64.9	80.8
	BiT-S-R50x1	86.3	63.5	78.7

Table 3. Investigating effect of large-scale pre-training on model robustness to spurious correlations. All models are fine-tuned on Waterbirds (Sagawa et al., 2020). Pre-training on ImageNet-21k provides better performance.

training data regimes, transformers have a higher propensity of memorizing training samples and are less robust compared to CNNs of comparable size.

4.2. Investigating model performance under data imbalance

Recall that model robustness to spurious correlations is correlated with its ability to generalize from the training examples where spurious correlations do not hold. We hypothesize that this generalization ability varies depending on the inherent data imbalance. In this section, we investigate the effect of data imbalance on the model's performance. In the extreme case, the model only observes 5 samples from the underrepresented group.

Setup. Considering the problem of `waterbird` vs `landbird` classification, these examples correspond to those in the groups: `waterbird` on `land` background and `landbird` on `water` background. We refer to these examples that do not include spurious associations with label as minority samples. For this study, we remove varying fraction of minority samples from the smallest group(`waterbird` on `land` background), while fine-tuning. We measure the effect based on the worst-group test accuracy and model consistency defined in Section 3.1.

Takeaways. In Figure 2, we report results for ViT-S/16 and BiT-M-R50x1 model when finetuned on Waterbirds dataset. We find that as more minority samples are removed, there is a graceful degradation in the generalization capability of both ViT and BiT models. However, the decline is more prominent in BiTs with the model performance reaching near-random when we remove 90% of minority samples. From this experiment, we conclude that additional robustness of ViT models to spurious associations stems from their better generalization capability from minority samples. However, they still suffer from spurious correlations when minority examples are scarce.

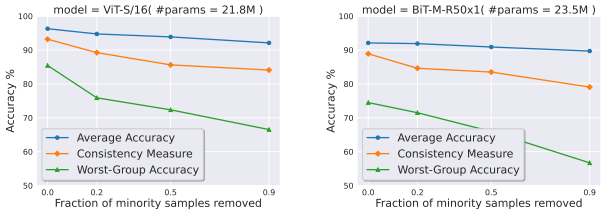

Figure 2. **Data Imbalance.** We investigate the effect of data imbalance on different model architectures. Our findings reveal that both ViT and BiT models suffers from spurious correlations when minority samples are scarce in fine-tuning dataset.

We provide additional ablations and experimental results in Appendix C, Appendix D and Appendix E.

5. Related Works

Since the introduction of transformers by Vaswani *et al.* (Vaswani et al., 2017) in 2017, there has been a deluge of studies adopting the attention-based transformer architecture for solving various problems in natural language processing (Dai et al., 2019; Radford et al., 2018; 2019; Yang et al., 2019). In the domain of computer vision, Dosovitskiy *et al.* (Dosovitskiy et al., 2021) first introduced the concept of Vision Transformers (ViT) by adapting the transformer architecture in (Vaswani et al., 2017) for image classification tasks. Naseer *et al.* (Naseer et al., 2021) provides a comprehensive understanding of the working principle of ViT architecture through extensive experimentation. Some notable findings in (Naseer et al., 2021) reveal that transformers are highly robust to severe occlusions, perturbations, and distributional shifts. Recently, performance of ViT models in the wild has been extensively studied (Bai et al., 2021; Bhojanapalli et al., 2021; Park & Kim, 2022; Paul & Chen, 2021; Tian et al., 2022; Zhang et al., 2021) using a set of robustness generalization benchmarks, e.g., ImageNet-C (Hendrycks & Dietterich, 2019), Stylized-ImageNet (Geirhos et al., 2019), ImageNet-A (Hendrycks et al., 2021), etc. Different from prior works, in this paper, we provide a first systematic study on the robustness of vision transformers when learned on datasets containing spurious correlations. Refer Appendix J for detailed discussion on related works.

6. Conclusion

In this paper, we investigate robustness of ViT models when learned on datasets containing spurious associations between target label and environmental features. Our findings can be summarized as: 1) ViTs are more robust to spurious correlations than CNNs under large-scale pre-training data regime. However, when the pre-training dataset is relatively small, transformer models perform worse as compared to CNNs of comparable size; 2) Improved robustness of ViTs can be attributed to better generalization capability from the counterexamples where spurious correlations do not hold.

References

Abnar, S. and Zuidema, W. Quantifying attention flow in transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4190–4197. Association for Computational Linguistics, 2020. 8, 12

Bai, Y., Mei, J., Yuille, A., and Xie, C. Are transformers more robust than cnns? In *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021. 4, 14

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. 4, 8, 14

Dai, Z., Yang, Z., Yang, Y., Carbonell, J., Le, Q., and Salakhutdinov, R. Transformer-XL: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 2978–2988, 2019. 4, 13

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE/CVF International Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. 2

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pp. 4171–4186, 2019. 9, 13, 14

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 1, 4, 7, 13

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., and Brendel, W. Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations*, 2019. 1, 4, 14

Goel, K., Gu, A., Li, Y., and Ré, C. Model patching: Closing the subgroup performance gap with data augmentation. In *International Conference on Learning Representations*, 2021. 1, 2

He, H., Zha, S., and Wang, H. Unlearn dataset bias in natural language inference by fitting the residual. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, pp. 132–142. 14

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016. 13

Hendrycks, D. and Dietterich, T. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019. 4, 14

Hendrycks, D., Lee, K., and Mazeika, M. Using pre-training can improve model robustness and uncertainty. In *International Conference on Machine Learning*, pp. 2712–2721. PMLR, 2019. 13

Hendrycks, D., Zhao, K., Basart, S., Steinhardt, J., and Song, D. Natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15262–15271, 2021. 4, 14

Kolesnikov, A., Beyer, L., Zhai, X., Puigcerver, J., Yung, J., Gelly, S., and Houlsby, N. Big Transfer (BiT): General Visual Representation Learning. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2020. 2, 13

Liu, W., Wang, X., Owens, J., and Li, Y. Energy-based out-of-distribution detection. 2020. 10

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. 13

Liu, Z., Luo, P., Wang, X., and Tang, X. Deep learning face attributes in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3730–3738, 2015. 3, 7, 10, 11

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10012–10022, 2021. 13

McCoy, R. T., Pavlick, E., and Linzen, T. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. *arXiv preprint arXiv:1902.01007*, 2019. 14

Ming, Y., Yin, H., and Li, Y. On the impact of spurious correlation for out-of-distribution detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022. 3, 10, 11

Naseer, M., Ranasinghe, K., Khan, S., Hayat, M., Khan, F., and Yang, M.-H. Intriguing properties of vision transformers. In *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021. 4, 13

Park, N. and Kim, S. How do vision transformers work? In *International Conference on Learning Representations*, 2022. 4, 14

Paul, S. and Chen, P.-Y. Vision transformers are robust learners. *arXiv preprint arXiv:2105.07581*, 2021. 4, 14

Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I. Improving language understanding by generative pre-training. 2018. 4, 13

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. Language models are unsupervised multitask learners. *OpenAI blog*, pp. 9, 2019. 4, 13

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763. PMLR, 2021. 13

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, pp. 211–252, 2015. 2

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. In *International Conference on Learning Representations (ICLR)*, 2020. 1, 2, 4, 7, 9, 10, 12

Steiner, A., Kolesnikov, A., Zhai, X., Wightman, R., Uszkoreit, J., and Beyer, L. How to train your vit? data, augmentation, and regularization in vision transformers. *arXiv preprint arXiv:2106.10270*, 2021. 7, 13

Tian, R., Wu, Z., Dai, Q., Hu, H., and Jiang, Y.-G. Deeper insights into vits robustness towards common corruptions. *arXiv preprint arXiv:2204.12143*, 2022. 4, 14

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., and Jégou, H. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pp. 10347–10357. PMLR, 2021a. 13

Touvron, H., Cord, M., Sablayrolles, A., Synnaeve, G., and Jégou, H. Going deeper with image transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 32–42, 2021b. 13

Tu, L., Lalwani, G., Gella, S., and He, H. An Empirical Study on Robustness to Spurious Correlations using Pre-trained Language Models. *Transactions of the Association for Computational Linguistics*, pp. 621–633, 2020. 1, 2, 9, 14

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 6000–6010, 2017. 4, 13

Wah, C., Branson, S., Welinder, P., Perona, P., and Belongie, S. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011. 2, 3

Wang, W., Xie, E., Li, X., Fan, D.-P., Song, K., Liang, D., Lu, T., Luo, P., and Shao, L. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 568–578, 2021. 13

Xue, F., Shi, Z., Wei, F., Lou, Y., Liu, Y., and You, Y. Go wider instead of deeper. *arXiv preprint arXiv:2107.11817*, 2021. 13

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., and Le, Q. V. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 2019. 4, 13

Yuan, L., Chen, Y., Wang, T., Yu, W., Shi, Y., Jiang, Z.-H., Tay, F. E., Feng, J., and Yan, S. Tokens-to-token vit: Training vision transformers from scratch on imagenet. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 558–567, 2021. 13

Zeiler, M. D. and Fergus, R. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision*, pp. 818–833, 2014. 13

Zhai, X., Kolesnikov, A., Houlsby, N., and Beyer, L. Scaling vision transformers, 2021. 13

Zhang, C., Zhang, M., Zhang, S., Jin, D., Zhou, Q., Cai, Z., Zhao, H., Yi, S., Liu, X., and Liu, Z. Delving deep into the generalization of vision transformers under distribution shifts. *arXiv preprint arXiv:2106.07617*, 2021. 4, 14

Zhang, T., Wu, F., Katiyar, A., Weinberger, K. Q., and Artzi, Y. Revisiting few-sample bert fine-tuning. *arXiv preprint arXiv:2006.05987*, 2020. 9

Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., and Torralba, A. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, pp. 1452–1464, 2017. 2, 3, 10

A. Implementation Details

1. **Transformers.** For both ViT and DeiT models, we obtain the pre-trained checkpoints from the `timm` library[1]. For downstream fine-tuning on Waterbirds and CelebA dataset, we scale up the resolution to 384 × 384 by adopting 2D interpolation of the pre-trained position embeddings proposed in (Dosovitskiy et al., 2021). Note, for CMNIST we keep the resolution as 224×224 during fine-tuning. We fine-tune models using SGD with a momentum of 0.9 with an initial learning rate of 3e-2. As described in (Steiner et al., 2021), we use a fixed batch size of 512, gradient clipping at global norm 1 and a cosine decay learning rate schedule with a linear warmup. We fine-tune `tiny` & `small` versions of models (*i.e.,* ViT-Ti/16 and ViT-S/16) for 1000 steps, whereas `base` version (*i.e.,* ViT-B/16) is fine-tuned for 2000 steps.

2. **BiT.** We obtain the pretrained checkpoints from the official repository[2]. For downstream fine-tuning, we use SGD with an initial learning rate of 0.003, momentum 0.9, and batch size 512. We fine-tune models with various capacity for 500 steps, including BiT-M-R50x1, BiT-M-R50x3, and BiT-M-R101x1.

B. Representative Examples

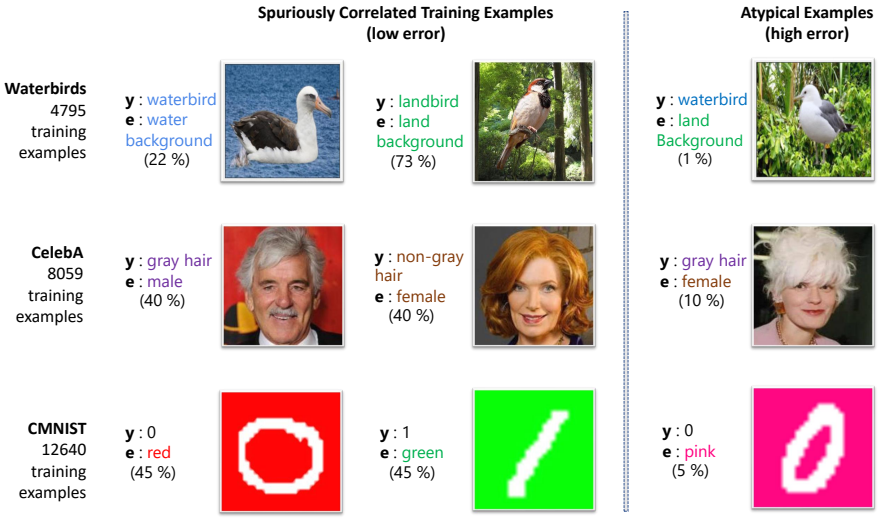

Figure 3. **Representative Examples.** We study three image datasets Waterbirds (Sagawa et al., 2020), CelebA (Liu et al., 2015) and CMNIST. The label y is spuriously correlated with environment e in majority of training samples. The frequency of each group in training data is denoted by (%). Figure is adapted from (Sagawa et al., 2020).

Pretraining Dataset							
ImageNet-21k	**Model**	ViT-B	ViT-S	ViT-Ti	BiT-M-R50x3	BiT-M-R101x1	BiT-M-R50x1
	#Params	86.1M	21.8M	5.6M	211M	42.5M	23.5M
ImageNet-1k	**Model**	DeiT-B	DeiT-S	DeiT-Ti	BiT-S-R50x3	BiT-S-R101x1	BiT-S-R50x1
	#Params	86.1M	21.8M	5.6M	211M	42.5M	23.5M

Table 4. Different model architectures used in our experiments along with number of trainable parameters and pre-training dataset. Note that the DeiT architecture is identical to ViT variant of comparable size with the only difference lying in pre-training dataset and data augmentations used during pre-training.

[1] https://github.com/rwightman/pytorch-image-models/tree/master/timm
[2] https://github.com/google-research/big_transfer

C. Extension: Understanding role of self-attention mechanism for improved robustness in ViT models

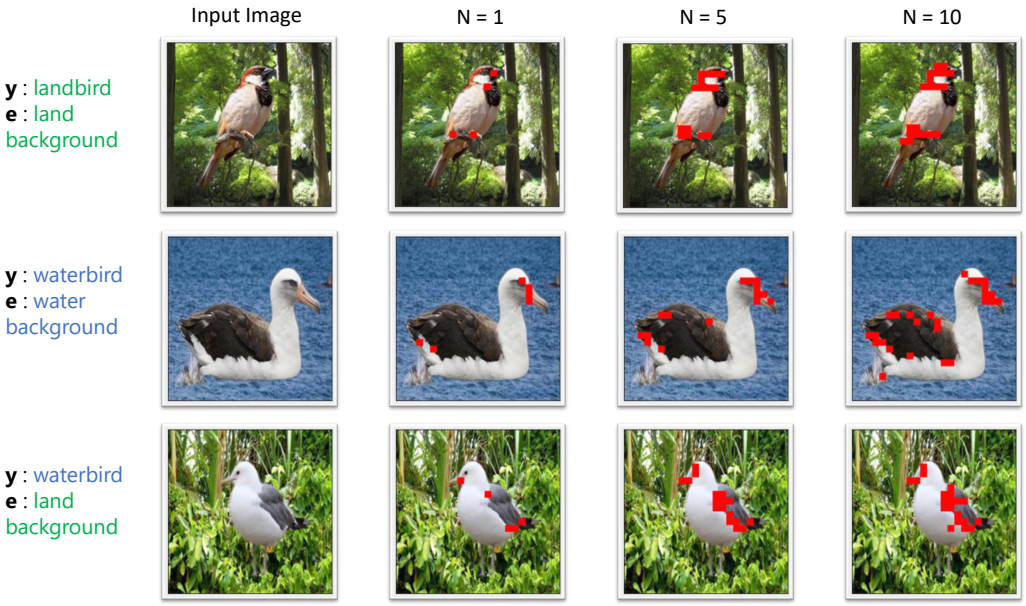

Figure 4. Visualization of the top N patches receiving the highest attention (marked in red). Investigating the attention matrix, we find that all image patches—irrespective of spatial location—provides maximum attention to the patches representing essential cues for accurately identifying the foreground object such as claw, beak and fur color. See text for details. See Supplementary for visualizations on other datasets and models.

Given the results above, a natural question arises: what makes ViT particularly robust in the presence of spurious correlations? In this section, we aim to understand the role of ViT by looking into the self-attention mechanism. The attention matrix in ViT models encapsulates crucial information about the interaction between different image patches.

Latent pattern in attention matrix. To gain insights, we start by analyzing the attention matrix, where each element in the matrix $a_{i,j}$ represents attention values with which an image patch i focuses on another patch j. For example: consider an input image of size 384×384 and patch resolution of 16×16, then we have a 576×576 attention matrix (excluding the `class token`). To compute final attention matrix, we use Attention Rollout (Abnar & Zuidema, 2020) which recursively multiplies attention weight matrices in all layers below. Our analysis here is based on the ViT-B/16 model fine-tuned on Waterbirds. Intriguingly, we observe that each image patch, irrespective of its spatial location, provides maximum attention to the patches representing essential cues for accurately identifying the foreground object.

Figure 4 exhibits this interesting pattern, where we mark (in red) the top $N = \{1, 5, 10\}$ patches being attended by every image patch. To do so, for every image patch i, where $i \in \{1, \cdots, 576\}$, we find the top N patches receiving the highest attention values and mark (in red) on the original input image. This would give us $576 \times N$ patches, which we overlay on the original image. Note that different patches may share the same top patches, hence we observe the sparse pattern. In Figure 4, we can see that the patches receiving the highest attention represent important signals such as the shape of the beak, claw, and fur color—all of which are essential for the classification task `waterbird` vs `landbird`. It is particularly interesting to note the last row in Figure 4, which is an example from the minority group (`waterbird` on `land` background). This is a challenging case where the spurious correlations between y and e do not hold. A non-robust model would utilize the background environmental features for predictions. In contrast, we notice that each patch in the image correctly attends to the foreground patches.

Masked attention. The attention matrix in ViT models encapsulates crucial information about the interaction between different image patches resulting in access to more global information. Inspired by (Bhojanapalli et al., 2021), we use a

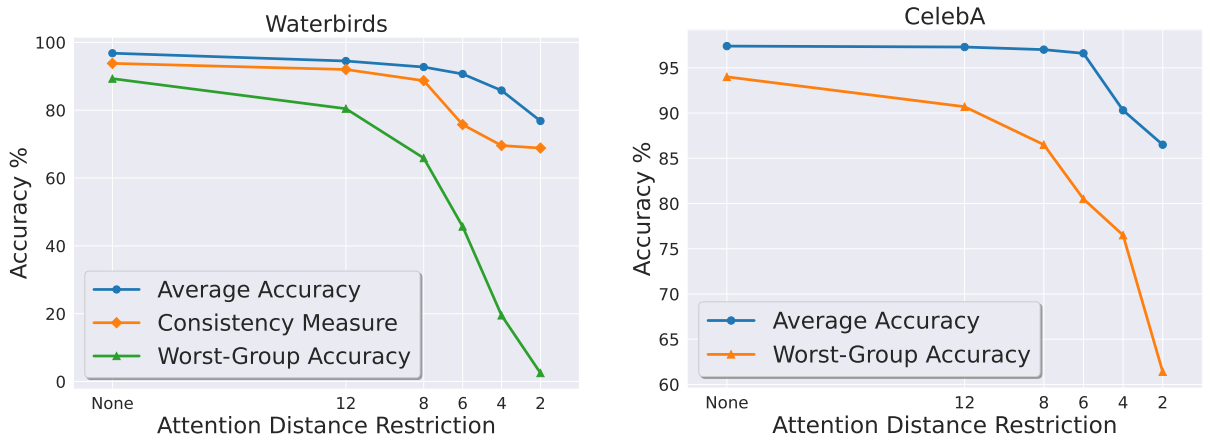

Figure 5. **Masked Attention.** We study the role of global attention in ViT models in providing improved robustness to spurious correlations. We observe that constraining the attention to be local results in degradation of model performance on spuriously correlated datasets such as Waterbirds (**left**) and CelebA (**right**).

spatial mask to study the effect of restricting image patches to attend only those lying within a certain distance. However, the `class token` is allowed to interact and attend to all other image patches. Note, while fine-tuning we do not use any spatial mask and allow the model to leverage information from the complete attention matrix. Masking is done only during inference time. Figure 5 depicts the results of our study on ViT-B/16 when fine-tuned on Waterbirds (**left**) and CelebA (**right**). For both datasets, we see a monotonic decrease in worst-group test accuracy and Consistency Measure, as we increase the restriction on allowable attention distance. In the extreme case, when the constrained attention distance equals 2, the model completely fails to correctly classify the test images in the smallest group indicating high reliance on spurious features while making the prediction. In other words, limiting the attention to be local results in degradation of model robustness to spurious correlations. Thus, we conclude that global attention in ViT models indeed plays a crucial role in providing additional robustness to spurious correlations.

D. Extension: Does longer fine-tuning in ViT improve robustness to spurious correlations?

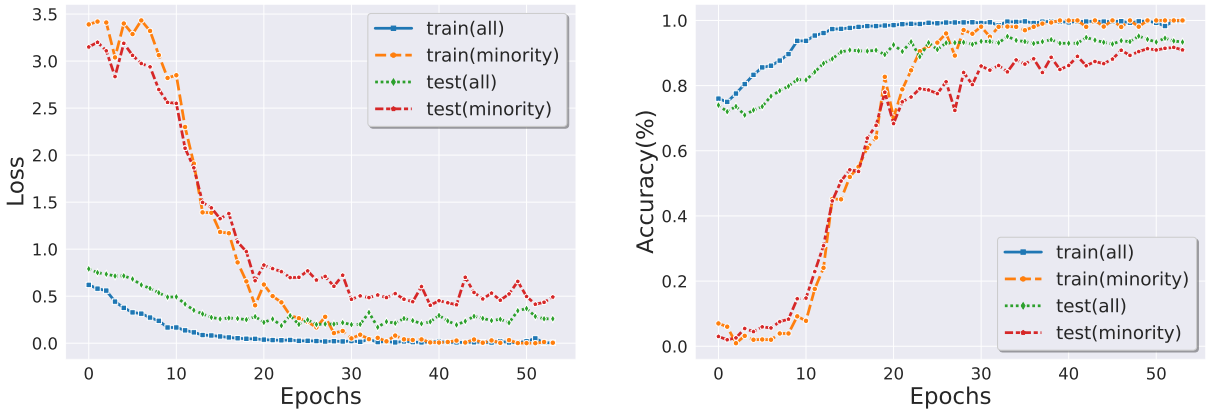

Figure 6. **Longer Fine-tuning.** We study the effect of longer fine-tuning on performance of ViT models. We report loss and accuracy for ViT-S/16 model finetuned on Waterbirds (Sagawa et al., 2020) at each epoch of fine-tuning. Investigating further we observe that although fine-tuning for more epochs provide no additional gain in average test accuracy, but it improves model performance on minority samples.

Recent studies in the domain of natural language processing (Tu et al., 2020; Zhang et al., 2020) have shown that the performance of BERT (Devlin et al., 2019) models on smaller datasets can be significantly improved through longer

fine-tuning. In this section, we investigate if longer fine-tuning also plays a positive role in the performance of ViT models in spuriously correlated environments.

Takeaways. Figure 6 reports the loss (**left**) and accuracy (**right**) at each epoch for ViT-S/16 model fine-tuned on Waterbirds dataset (Sagawa et al., 2020). To better understand the effect of longer fine-tuning on worst-group accuracy, we separately plot the model loss and accuracy on all examples and minority samples. From the loss curve, we observe that the training loss for minority examples decreases at a much slower rate as compared to the average loss. Specifically, the average train loss takes 20 epochs of fine-tuning to reach near-zero values, while training loss on minority group plateaus after 40 epochs. Similarly, we see that although the average test accuracy of the model stops increasing after 30 epochs, the accuracy of minority samples reaches a stationary state after 50 epochs of fine-tuning. These results reveal two key observations: (1) While longer fine-tuning does not benefit the average test accuracy, it plays a positive role in improving model performance on minority samples, and (2) ViT models do not overfit with longer fine-tuning.

E. Extension: Spurious Out-of-Distribution Detection

Model	Waterbirds (Sagawa et al., 2020)		CelebA (Liu et al., 2015)		CMNIST	
	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
ViT-B/16	**56.8**	**91.0**	**60.5**	**88.4**	**7.4**	**98.8**
ViT-S/16	62.2	87.0	61.3	86.7	8.7	97.7
ViT-Ti/16	79.5	71.6	94.3	72.7	16.4	96.7
BiT-M-R50x3	96.0	59.0	63.8	85.3	45.9	84.1
BiT-M-R101x1	95.5	59.5	70.3	85.6	44.5	81.4
BiT-M-R50x1	95.1	63.4	69.7	85.7	30.0	88.4

Table 5. **Spurious OOD evaluation.** OOD detection performance of ViT and BiT models when finetuned on Waterbirds (Sagawa et al., 2020), CelebA (Liu et al., 2015) & CMNIST. We use energy score (Liu et al., 2020) for calculating AUROC and FPR95. We observe that ViT models are more robust to spurious OOD examples as compared to BiTs.

In this section, we study the performance of ViT models in out-of-distribution setting. Introduced in (Ming et al., 2022), spurious out-of-distribution (OOD) data is defined as samples that do not contain the invariant features $\mathbf{z}^{inv}$ essential for accurate classification, but contain the spurious features $\mathbf{z}^{e}$. Hence, these samples are denoted as $\mathbf{x}_{ood} = \rho(\mathbf{z}^{\bar{y}}, \mathbf{z}^{e})$ where $\bar{y}$ is an out-of-class label, such that $\bar{y} \notin \mathcal{Y}$. In the problem of `waterbird` vs `landbird` classification, an image of a person standing in forest would be an example of spurious OOD, since it contains different semantic class `person` $\notin \{$`waterbird`, `landbird`$\}$, yet has the environmental features of land background. A non-robust model relying on the background feature may classify such OOD data as an in-distribution class with high confidence. Hence, we aim to understand if self-attention based ViT models can mitigate this problem and if so, to what extent.

Setup. To investigate the performance of different models against spurious OOD examples, we use the setup introduced in (Ming et al., 2022). Specifically, for Waterbirds (Sagawa et al., 2020) we test on subset of images of land and water sampled from the Places dataset (Zhou et al., 2017). Considering, CelebA (Liu et al., 2015) as in-distribution, our test suite consists of images of `bald male` as spurious OOD, since they contain environmental features (`gender`) without invariant features (`hair`). For CMNIST, the in-distribution data contains digits $\mathcal{Y} = \{0, 1\}$ and the background colors, $\mathcal{E} = \{$`red, green, purple, pink`$\}$. We use digits $\{5, 6, 7, 8, 9\}$ with background color `red` and `green` as test OOD samples.

Takeaways. We report our findings in Table 5. Clearly, ViT models achieve better OOD evaluation metrics as compared to BiTs. Specifically, ViT-B/16 achieves $+$**32**% higher AUROC than BiT-M-R50x3, considering Waterbirds (Sagawa et al., 2020) as in-distribution.

	Model	Test Accuracy	
		Average Acc.	Worst-Group Acc.
ImageNet-21k	ViT-B/16	**97.4**	**94.0**
	ViT-S/16	97.0	91.5
	ViT-Ti/16	96.5	84.6
ImageNet-1k	DeiT-B/16	96.4	88.0
	DeiT-S/16	96.1	87.1
	DeiT-Ti/16	94.9	85.7
ImageNet-21k	BiT-M-R50x3	97.3	89.8
	BiT-M-R101x1	97.2	89.8
	BiT-M-R50x1	96.8	87.7
ImageNet-1k	BiT-S-R50x3	96.4	88.3
	BiT-S-R101x1	96.5	90.2
	BiT-S-R50x1	96.3	90.9

Table 6. Investigating the effect of large scale pre-training on model robustness to spurious correlations when finetuned on CelebA (Liu et al., 2015).

F. Extension: How does the size of pre-training dataset affect robustness to spurious correlations?

In this section, to further validate our findings on the importance of large-scale pre-training dataset, we show results on CelebA (Liu et al., 2015) dataset. We report our findings in Table 6. We also observe a similar trend for this setup that larger model capacity and more pre-training data yields significant improvement in worst-group accuracy for ViT models. Further, when pre-trained on a relatively smaller dataset such as ImageNet-1k, the performance of transformer-based DeiT models are poor as compared to the corresponding CNN counterpart.

Also, compared to BiT models, *the robustness of ViT models benefits more with a large pre-training dataset*. For example, compared to ImageNet-1k, fine-tuning ViT-B/16 pre-trained on ImageNet-21k improves the worst-group accuracy by **6%**. On the other hand, for BiT models, fine-tuning with a larger pre-trained dataset yields marginal improvement. Specifically, BiT-M-R50x3 only improves the worst-group accuracy by 1.5% with ImageNet-21k.

G. Extension : Color Spurious Correlation

To further validate our findings beyond natural background and gender as spurious (*i.e.* environmental) features, we provide additional experimental results with the ColorMNIST dataset, where the digits are superimposed on coloured backgrounds. Specifically, it contains spurious correlation between the target label and the background color. Similar to the setup in (Ming et al., 2022), we fix the classes $\mathcal{Y} = \{0, 1\}$ and the background colors, $\mathcal{E} = \{\texttt{red, green, purple, pink}\}$. For this study, label $y = 0$ is spuriously correlated with background color $\{\texttt{red}, \texttt{purple}\}$, and similarly, label $y = 1$ has spurious associations with background color $\{\texttt{green}, \texttt{pink}\}$. Formally, we have $\mathbb{P}(e = \texttt{red}|y = 0) = \mathbb{P}(e = \texttt{purple}|y = 0) = \mathbb{P}(e = \texttt{green}|y = 1) = \mathbb{P}(e = \texttt{pink}|y = 1) = 0.45$ and $\mathbb{P}(e = \texttt{green}|y = 0) = \mathbb{P}(e = \texttt{pink}|y = 0) = \mathbb{P}(e = \texttt{red}|y = 1) = \mathbb{P}(e = \texttt{purple}|y = 1) = 0.05$. Note that, while fine-tuning the models, we fix the foreground color of digits as `white`.

Results and insights on robustness performance. We compare model predictions on samples with same class label but different background & foreground colors. Given a data point $(\mathbf{x}_i, y_i)$, we modify the background and foreground color of $\mathbf{x}_i$ randomly to generate a new test image $\bar{\mathbf{x}}_i$ with the constraint of having the same semantic label. During evaluation, the background color is chosen uniform-randomly from the set of colors: {`#ecf02b, #f06007, #0ff5f1, #573115, #857d0f, #015c24, #ab0067, #fbb7fa, #d1ed95, #0026ff`} and the foreground color is selected randomly from the set $\{\texttt{black}, \texttt{white}\}$. For evaluation purpose, we form a dataset consisting of 2100 samples and the results reported are averaged over 50 random runs. Figure 7 depicts the distribution of training samples in CMNIST dataset (**left**) and few representative examples after transformation (**right**).

We report our findings in Figure 8. Our operating hypothesis is that a robust model should predict same class label $\hat{f}(\mathbf{x}_i)$ and $\hat{f}(\bar{\mathbf{x}}_i)$ for a given pair $(\mathbf{x}_i, \bar{\mathbf{x}}_i)$, as they share exactly the same target label (i.e., the invariant feature is approximately the same). We can observe from Figure 8 that the best model ViT-B/16 obtains consistent predictions for 100% of image

pairs. After extensive experimentation over all combinations, we find that setting the foreground color as `black` and the background as `white` caused the models to be most vulnerable. We see a significant decline in model consistency when the foreground color is set as `black` and the background as `white` (indicated as **BW**) as compared to random setup.

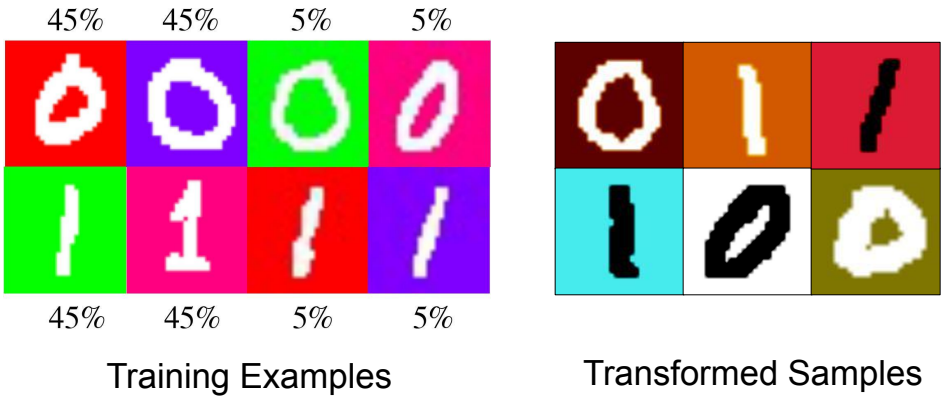

Figure 7. **CMNIST.** Distribution of training samples in CMNIST dataset(**left**) and few representative examples after transformation(**right**) as defined in Section G.

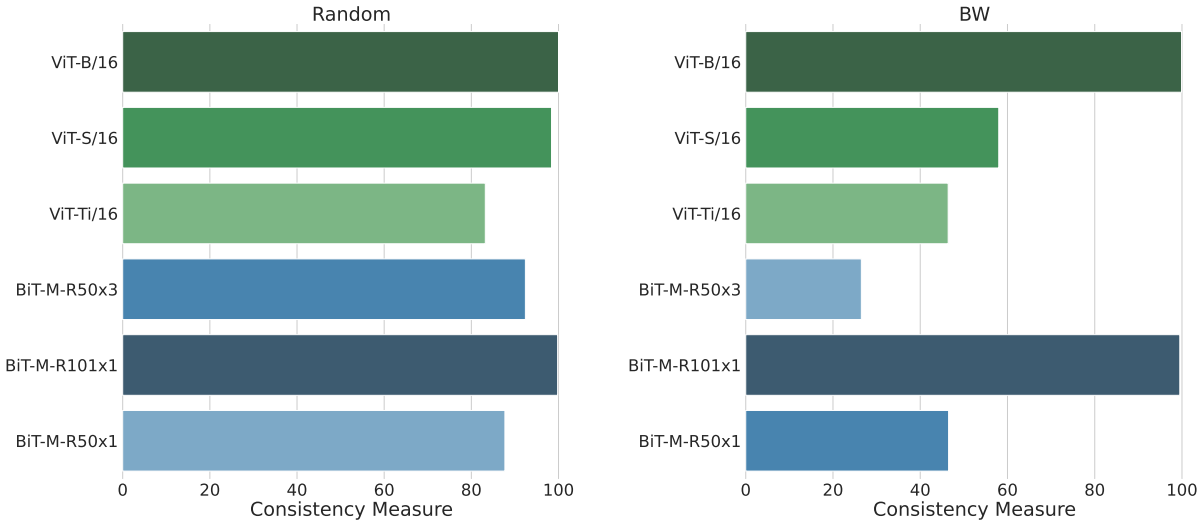

Figure 8. **Consistency Measure.** Evaluation results quantifying consistency for models of different architectures and varying capacity. We indicate the setup when the foreground color is set as `black` and the background as `white` using **BW**(**right**). **Random** represents setting both the foreground and background color randomly(**left**).

H. Visualization

H.1. Attention Map

In Figure 9, we visualize attention maps obtained from ViT-B/16 model for some samples images from Waterbirds (Sagawa et al., 2020) and CMNIST dataset. We use Attention Rollout (Abnar & Zuidema, 2020) to obtain the attention matrix. We can observe that the model successfully attends spatial locations representing invariant features while making predictions.

H.2. The Attention Matrix of CMNIST

In the main text, we provide visualizations in which each image patch, irrespective of its spatial location, provides maximum attention to the patches representing essential cues for accurately identifying the foreground object. In Figure 10, we show

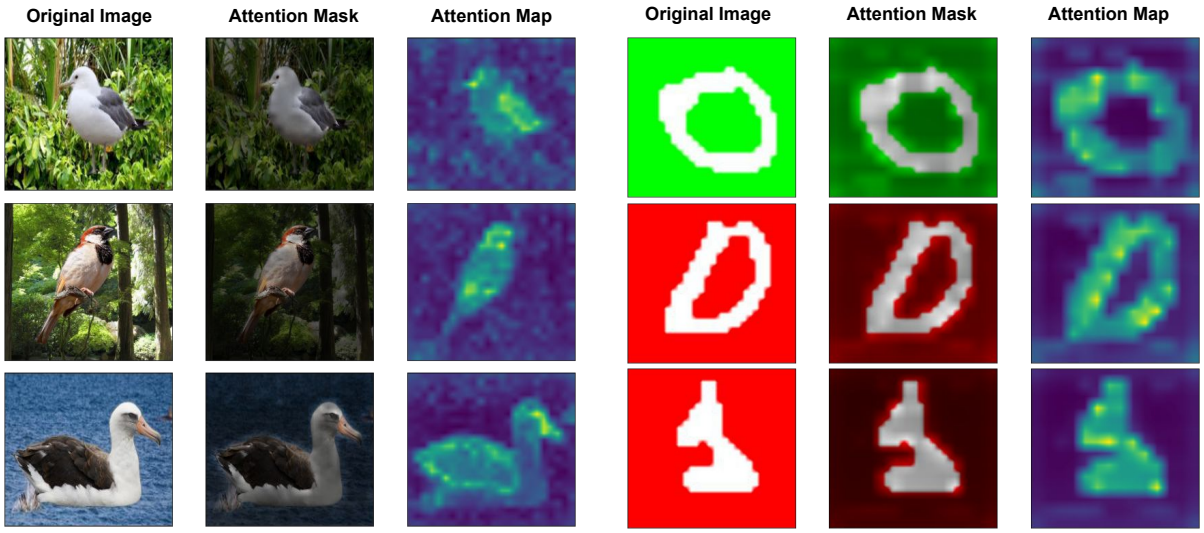

Figure 9. **Attention Map.** Visual illustration of attention map obtained from ViT-B/16 model for few representative images.

visualizations for ViT-B/16 fine-tuned on CMNIST dataset to further validate our findings.

I. Software and Hardware

We run all experiments with Python 3.7.4 and PyTorch 1.9.0 using Nvidia Quadro RTX 5000 GPU.

J. Extension: Related Works

Pre-training and robustness. Recently, there has been an increasing amount of interest in studying the effect of pre-training (Devlin et al., 2019; Kolesnikov et al., 2020; Liu et al., 2019; Radford et al., 2021). Specifically, when the target dataset is small, generalization can be significantly improved through pre-training and then finetuning (Zeiler & Fergus, 2014). Findings of Hendrycks *et al.* (Hendrycks et al., 2019) reveal that pre-training provides significant improvement to model robustness against label corruption, class imbalance, adversarial examples, out-of-distribution detection, and confidence calibration. In this work, we focus distinctly on robustness to *spurious correlation*, and how it can be improved through large-scale pretraining.

Vision transformer. Since the introduction of transformers by Vaswani *et al.* (Vaswani et al., 2017) in 2017, there has been a deluge of studies adopting the attention-based transformer architecture for solving various problems in natural language processing (Dai et al., 2019; Radford et al., 2018; 2019; Yang et al., 2019). In the domain of computer vision, Dosovitskiy *et al.* (Dosovitskiy et al., 2021) first introduced the concept of Vision Transformers (ViT) by adapting the transformer architecture in (Vaswani et al., 2017) for image classification tasks. Subsequent studies (Dosovitskiy et al., 2021; Steiner et al., 2021) have shown that when pre-trained on sufficiently large datasets, ViT achieves superior performance on downstream tasks, and outperforms state-of-art CNNs such as residual networks (ResNets) (He et al., 2016) of comparable sizes. Since coming to the limelight, multiple variants of ViT models have been proposed. Touvron *et al.* (Touvron et al., 2021a) showed that it is possible to achieve comparable performance in small pre-training data regimes using extensive data augmentation and novel distillation strategy. Further improvements on ViT include enhancement in tokenization module (Yuan et al., 2021), efficient parameterization for scalability (Touvron et al., 2021b; Xue et al., 2021; Zhai et al., 2021) and building multi-resolution feature maps on transformers (Liu et al., 2021; Wang et al., 2021). In this paper, we provide a first systematic study on the robustness of vision transformers when learned on datasets containing spurious correlations.

Robustness of transformers. Naseer *et al.* (Naseer et al., 2021) provides a comprehensive understanding of the working principle of ViT architecture through extensive experimentation. Some notable findings in (Naseer et al., 2021) reveal that transformers are highly robust to severe occlusions, perturbations, and distributional shifts. Recently, performance of ViT

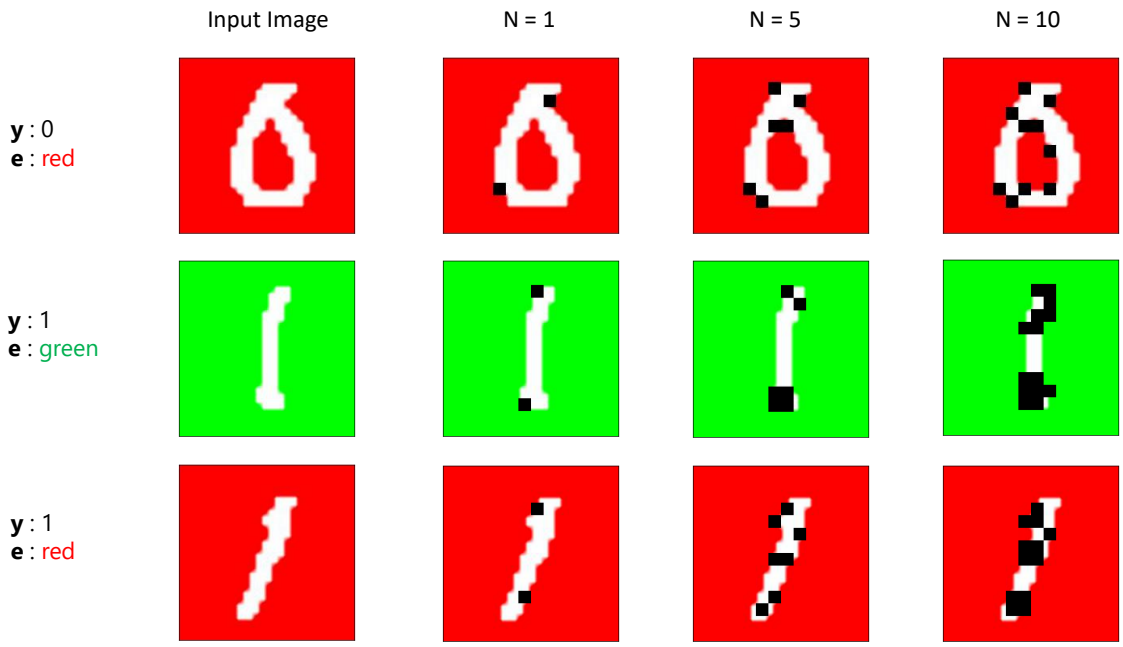

Figure 10. Visualization of the top N patches receiving the highest attention (marked in **black**) for ViT-B/16 fine-tuned on CMNIST. Investigating the attention matrix, we find that all image patches—irrespective of spatial location—provides maximum attention to the patches representing essential cues

models in the wild has been extensively studied (Bai et al., 2021; Bhojanapalli et al., 2021; Park & Kim, 2022; Paul & Chen, 2021; Tian et al., 2022; Zhang et al., 2021) using a set of robustness generalization benchmarks, e.g., ImageNet-C (Hendrycks & Dietterich, 2019), Stylized-ImageNet (Geirhos et al., 2019), ImageNet-A (Hendrycks et al., 2021), etc. Different from prior works, we focus on robustness performance on challenging datasets, which are designed to expose spurious correlations learned by the model. Our analysis reveals that pre-training improves robustness by better generalizing on examples from under-represented groups. Our findings are also complementary to robustness studies (He et al.; McCoy et al., 2019; Tu et al., 2020) in the domain of natural language processing, which reported that transformer-based BERT (Devlin et al., 2019) models improve robustness to spurious correlations.