Are Vision Transformers Robust to Spurious Correlations?

Anonymous authors
Paper under double-blind review

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 unexplored how spurious correlations are manifested in such architectures. In this paper, we systematically investigate the robustness of different transformer architectures 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. Further, we perform extensive ablations and experiments to understand the role of the self-attention mechanism in providing robustness under spuriously correlated environments. We hope that our work will inspire future research on further understanding the robustness of ViT models to spurious correlations.

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 (Sagawa et al., 2020a; Geirhos et al., 2019; Goel et al., 2021; Tu et al., 2020). In Figure 1, we illustrate a model setup that exploits the spurious correlation between the `water background` and label `waterbird` for prediction. Consequently, a model that relies on spurious features can perform poorly on test samples where the correlation no longer holds, such as `waterbird` on `land background`. Recent work by Sagawa et al. (2020b) has shown that over-parameterization in CNNs exacerbates robustness to spurious correlations.

As with the paradigm shift to attention-based architectures, it becomes increasingly critical to understand their behavior under ill-conditioned data. Particularly, in this paper, we provide a first study on the following question: *Are Vision Transformers robust to spurious correlations*? Although the behavior of ViTs against occlusions, perturbations, and distributional shifts have been extensively studied in the literature (Bhojanapalli et al., 2021; Bai et al., 2021; Zhang et al., 2022; Paul & Chen, 2022; Park & Kim, 2022), it remains unexplored how spurious correlation is manifested in the recent development of vision transformers. Unlike prior works, we specifically focus on robustness performance on challenging datasets designed to expose spurious correlations learned by the model. The learning task is challenging as the model is more likely to learn those spurious associations while training to achieve high training accuracy and fail on those samples where the spurious associations do not hold at test time. In contrast, previous works on ViT robustness did not consider this problem and focused on robustness to perturbations and data shifts during inference. Thus, the nature of the "robustness" problem studied in our paper is fundamentally different from the prior art.

Motivated by the question, we systematically investigate how and when transformer-based models exhibit robustness to spurious correlations on challenging benchmarks. We base our findings after studying a variety of transformer architectures including ViT (Dosovitskiy et al., 2021), Swin transfomer (Liu et al., 2021b), Pyramid Vision Transformer (Wang et al., 2021) and recently proposed Robust Vision Transformer (RVT) (Shi et al., 2022). Our findings reveal that: (1) For transformers, larger models and more pre-training data

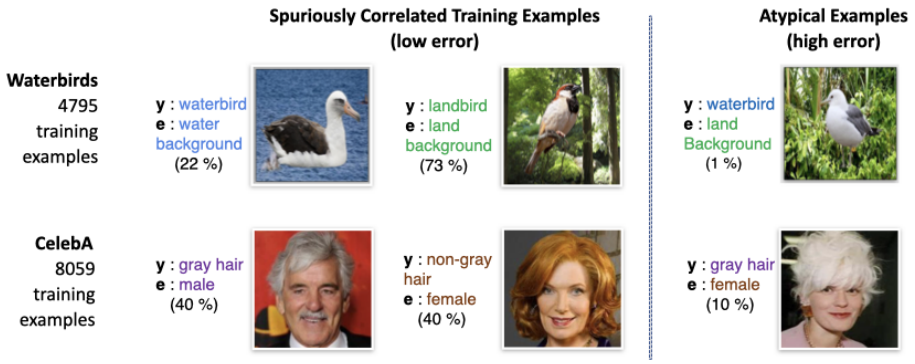

Figure 1: **Representative Examples.** We visualize samples from Waterbirds (Sagawa et al., 2020a) and CelebA (Liu et al., 2015) dataset. The label y is spuriously correlated with environment e in the majority of training samples. The frequency of each group in training data is denoted by (%). Figure is adapted from (Sagawa et al., 2020a).

yield a significant improvement in robustness to spurious correlations. On the other hand, when pre-trained on a relatively smaller dataset such as ImageNet-1k, transformer-based models have a higher propensity of memorizing training samples and are less robust to spurious correlations compared to the ImageNet-21k counterpart. (2) Interestingly, our linear probing experiment shows that the improvement in robustness does not simply stem from strong pre-trained features. (3) 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. (4) Finally, our study also reveals that, under the small pre-training regime, modern transformer architectures (Mao et al., 2022) despite being designed to be robust against common corruptions, out-of-distribution and adversarial attacks perform poorly when fine-tuned on datasets containing spurious associations further highlighting the importance of this study.

Going beyond, we perform extensive ablations to understand the role of the self-attention mechanism in providing robustness to ViT models. Our findings reveal that the self-attention mechanism in ViTs plays a crucial role in guiding the model to focus on spatial locations in an image which are essential for accurately predicting the target label. Our key contributions are summarized below:

1. To the best of our knowledge, we provide a first systematic study on the robustness of Vision Transformers when learned on datasets containing spurious correlations. Our work sheds light on the effectiveness of pre-training on ViT's robustness to spurious correlations.

2. We perform extensive experiments and ablations to understand the impact of model architectures, model capacity, pre-training dataset, data imbalance, fine-tuning, etc. These ablations lead to previously unknown findings on ViT's robustness to spurious correlations.

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 when fine-tuned on datasets containing spurious correlations.

2 Related Works

Pre-training and robustness. Recently, there has been an increasing amount of interest in studying the effect of pre-training (Kolesnikov et al., 2020; Devlin et al., 2019; Radford et al., 2021; Liu et al., 2019; Shi et al., 2022). 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. (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. Recent work (Liu et al., 2021a) has also shown that self-supervised learning makes CNNs more immune to dataset imbalance in pre-training datasets. In this work, we focus distinctly on robustness to *spurious correlation*, and how it can be improved through large-scale pre-training. To the best of our knowledge, this problem has not been investigated in prior literature on pre-training.

Vision transformer. Since the introduction of transformers by 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 (Radford et al., 2018; 2019; Yang et al., 2019; Dai et al., 2019). In the domain of computer vision, 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. showed that it is possible to achieve comparable performance in small pre-training data regimes using extensive data augmentation and a 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., 2021b; Wang et al., 2021). Recent studies (Zhou et al., 2022; Mao et al., 2022) propose enhanced ViT architectures to provide additional robustness against adversarial attacks (Mao et al., 2022) and image occlusion/corruptions (Mao et al., 2022; Zhou et al., 2022) during inference. While we also consider these architectures, our focus significantly differs from prior works — *we provide a first systematic study on the robustness of vision transformers when the training data itself contains inherent biases.*

Robustness of transformers. Naseer et al. 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 (Bhojanapalli et al., 2021; Zhang et al., 2022; Paul & Chen, 2022; Bai et al., 2021; Tian et al., 2022; Park & Kim, 2022) 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: (1) Robustness against occlusions, perturbations, and distribution shifts does not necessarily guarantee robustness against spurious correlations. (2) Large scale pre-training helps to mitigate this problem by better generalizing on examples from under-represented groups. Our findings are also complementary to robustness studies (Tu et al., 2020; He et al.; McCoy et al., 2019) in the domain of natural language processing, which reported that transformer-based BERT (Devlin et al., 2019) models improve robustness to spurious correlations.

3 Preliminaries

3.1 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 (Sagawa et al., 2020a; Geirhos et al., 2019; Goel et al., 2021; Tu et al., 2020). We illustrate a few examples in Figure 1. In `waterbird` vs `landbird` classification problem, majority of the training images has the target label (`waterbird` or `landbird`) spuriously correlated with the background features (`water` or `land` background). Sagawa et al. showed that deep neural networks can rely on these statistically informative yet spurious features to achieve high test accuracy on average, but fail significantly on groups where such correlations do not hold such as `waterbird` on `land background`.

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

Table 1: 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.

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

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 $\mathcal{X}$. Considering the example of `waterbird` vs `landbird` classification, invariant features $\mathbf{z}_i^{inv}$ would refer to signals which are essential for classifying $\mathbf{x}_i$ as y_i, such as the feather color, presence of webbed feet, and fur texture of birds, to mention a few. Environmental features $\mathbf{z}_i^e$, on the other hand, are cues not essential but correlated with target label y_i. For example, many waterbird images are taken in water habitats, so water scenes can be considered as $\mathbf{z}_i^e$. 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 4.

3.2 Transformers

Similar to the Transformer architecture in Vaswani et al. (2017), ViT model expects the input as a 1D sequence of token embeddings. An input image is first partitioned into non-overlapping fixed-size square patches of resolution $P \times P$, resulting in a sequence of flattened 2D patches. Following Devlin et al. (2019), ViT prepends a learnable embedding (`class` token) to the sequence of embedded patches, and this `class` token is used as image representation at the output of the transformer. To imbibe relative positional information of patches, position embeddings are further added to the patch embeddings.

The core architecture of ViT mainly consists of multiple stacked encoder blocks, where each block primarily consists of: (1) multi-headed self-attention layers, which learn and aggregate information across various spatial locations of an image by processing interactions between different patch embeddings in a sequence; and (2) a feed-forward layer. See an expansive discussion in related work (Section 2).

3.3 Model Zoo

In this study, we aim to understand the robustness of transformer-based models when trained on a dataset containing spurious correlations and how they fare against popular CNNs. In particular, 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 1. Specifically, we use model variants pre-trained on both ImageNet-1k (IN-1k) (Russakovsky et al., 2015) and on ImageNet-21k (IN-21k) (Deng et al., 2009) datasets. Table 1 summarizes the size and pre-training dataset of different models used in our study. Note that the DeiT architecture is identical to ViT variant of comparable size with the only difference lying in the pre-training

Table 2: Average and worst-group accuracies over train and test set for different models when finetuned on Waterbirds. Both ViT-B/16 and ViT-S/16 attain better test worst-group accuracy as compared to BiT models. All models are pre-trained on ImageNet-21k. Results (mean and std) are estimated over 3 runs for each setting.

Model	Train		Test	
	Average Acc.	Worst-Group Acc.	Average Acc.	Worst-Group Acc.
ViT-B/16	100	100	**96.75**$_{\pm 0.05}$	**89.30**$_{\pm 1.95}$
ViT-S/16	100	100	96.30$_{\pm 0.51}$	85.45$_{\pm 1.16}$
ViT-Ti/16	95.7	81.6	89.50$_{\pm 0.05}$	71.65$_{\pm 0.16}$
BiT-M-R50x3	100	100	94.90$_{\pm 0.05}$	80.51$_{\pm 1.02}$
BiT-M-R101x1	100	100	94.05$_{\pm 0.07}$	77.50$_{\pm 0.50}$
BiT-M-R50x1	100	100	92.05$_{\pm 0.05}$	75.10$_{\pm 0.62}$

dataset and data augmentations. For completeness, we also provide detailed ablations on other state-of-art architectures including Swin transformer (Liu et al., 2021b), Robust Vision Transformer (RVT) (Mao et al., 2022), Pyramid Vision Transformer (PVT) (Wang et al., 2021), and ConvNeXt (Liu et al., 2022) models.

Notation: 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. For instance: ViT-B/16 implies the `Base` variant with an input patch resolution of 16×16. In this paper, we use a 16×16 input patch size for computational simplicity.

4 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. (2020a). Specifically, we consider the following three classification datasets to study the robustness of ViT models in a spurious correlated environment: Waterbirds (Section 4.1), CelebA (Section 4.2), and ColorMNIST. Refer to Appendix, for results on ColorMNIST.

4.1 Waterbirds

Introduced in Sagawa et al. (2020a), this dataset contains spurious correlation between the 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). The spurious correlation is injected by pairing `waterbirds` on `water` background and `landbirds` on `land` background more frequently, as compared to other combinations. The dataset consists of $n = 4795$ training examples, with the smallest group size of 56 (i.e., waterbird on the land background).

Results and insights on generalization performance. Table 2 compares worst-group accuracies of different models when fine-tuned on Waterbirds (Sagawa et al., 2020a) 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

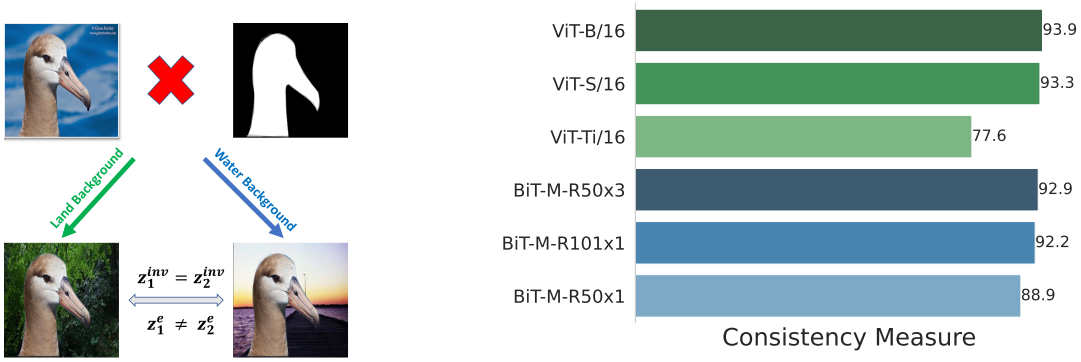

Figure 2: **Consistency Measure.** In Waterbirds dataset, y $\in$ {`waterbird`, `landbird`} is correlated with environment e $\in$ {`water`, `land`}. **Left**: Visual illustration of the experimental setup for measuring model consistency. Ideally, changing the spurious features ($\mathbf{z}^e$) should have no impact on model prediction. **Right**: Evaluation results quantifying consistency for models of different architectures and varying capacity.

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 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 2 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 2 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.

4.2 CelebA

Beyond background spurious features, we further validate our findings on a different type of spurious feature based on gender attributes. 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 3 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

Table 3: Average and worst-group accuracies over train and test set for different models when finetuned on CelebA. Both ViT-B/16 and ViT-S/16 attain better worst-group accuracy as compared to BiT models. All models are pre-trained on ImageNet-21k. Results (mean and std) are estimated over 3 runs for each setting.

Model	Train		Test	
	Average Acc.	Worst-Group Acc.	Average Acc.	Worst-Group Acc.
ViT-B/16	100	100	**97.40**$_{\pm 0.42}$	**94.10**$_{\pm 0.51}$
ViT-S/16	100	100	96.26$_{\pm 0.66}$	91.50$_{\pm 1.56}$
ViT-Ti/16	97.9	93.3	96.71$_{\pm 0.18}$	88.60$_{\pm 3.92}$
BiT-M-R50x3	100	100	97.31$_{\pm 0.05}$	89.80$_{\pm 0.42}$
BiT-M-R101x1	100	100	97.20$_{\pm 0.08}$	89.33$_{\pm 0.78}$
BiT-M-R50x1	100	100	96.82$_{\pm 1.20}$	87.72$_{\pm 1.56}$

with our observations in Section 4.1 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.

5 Discussion: A Closer Look at ViT Under Spurious Correlation

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

5.1 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. Specifically, we compare pre-trained models of different capacities, architectures, and sizes of pre-training data. 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 different transformer- and convolution-based models in Table 4. Specifically, we report for ViT & Swin (Liu et al., 2021b) models from the transformer family and BiT (Kolesnikov et al., 2020) & ConvNeXt (Liu et al., 2022) architectures based on convolutions. For detailed ablation results on other transformer architectures: RVT (Mao et al., 2022) and PVT (Wang et al., 2021), readers can refer to Table 8 (Appendix). Based on the results in Table 4, 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 convolution-based model family: BiT (Kolesnikov et al., 2020) and ConvNeXt (Liu et al., 2022) models.

2. Second, when pre-trained on a relatively smaller dataset such as ImageNet-1k, both vision transformers and convolution-based models have a higher propensity of memorizing training samples and are less robust to spurious correlations compared to the ImageNet-21k counterpart. In particular, previous work (Sagawa et al., 2020b) theoretically showed that overparameterization in CNNs—increasing model size well beyond the point of zero training error—can hurt test error in minority groups. Our study further provides new insights by showing that this trend of memorizing training samples can be alleviated using large-scale pre-training data. In the extreme case without any pre-training, we show in Appendix I that both ViT and BiT models severely overfit the training dataset to attain 100% training accuracy and significantly fail on test samples.

Table 4: Investigating effect of large-scale pre-training on model robustness to spurious correlations. All models are fine-tuned on Waterbirds. Pre-training on ImageNet-21k (IN-21k) provides better performance.

		Model	Params	Train Accuracy		Test Accuracy	
				Average	Worst-Group	Average	Worst-Group
Transformer Models	IN-21k	ViT-B/16	86.1M	100	100	96.8	89.3
		ViT-S/16	21.8M	100	100	96.3	85.5
		ViT-Ti/16	5.6M	95.7	81.6	89.5	71.7
	IN-1k	DeiT-B/16	86.1M	100	100	85.9	44.6
		DeiT-S/16	21.8M	100	100	84.5	46.7
		DeiT-Ti/16	5.6M	94.9	74.6	83.4	41.8
	IN-21k	Swin-Base	86.7M	100	100	95.7	87.5
		Swin-Small	48.8M	100	100	94.8	83.2
		Swin-Tiny	27.5M	100	100	93.9	78.5
	IN-1k	Swin-Base	86.7M	100	100	92.3	61.7
		Swin-Small	48.8M	100	100	93.1	62.5
		Swin-Tiny	27.5M	100	100	91.3	50.7
Convolutional Models	IN-21k	BiT-M-R50x3	211M	100	100	94.9	80.5
		BiT-M-R101x1	42.5M	100	100	94.1	77.5
		BiT-M-R50x1	23.5M	100	100	92.1	75.1
	IN-1k	BiT-S-R50x3	211M	100	100	87.0	60.3
		BiT-S-R101x1	42.5M	100	100	87.3	64.9
		BiT-S-R50x1	23.5M	100	100	86.3	63.5
	IN-21k	ConvNeXt-B	87.5M	100	100	93.1	76.7
		ConvNeXt-S	49.5M	100	100	92.5	74.2
		ConvNeXt-T	27.9M	100	100	90.3	69.6
	IN-1k	ConvNeXt-B	87.5M	100	100	87.3	66.1
		ConvNeXt-S	49.5M	100	100	86.4	67.6
		ConvNeXt-T	27.9M	100	100	85.1	63.2

5.2 Is linear probing sufficient for robustness to spurious correlation?

One may hypothesize that the robustness to spurious correlation can benefit from the strong pre-trained features. To verify this, we conduct the "linear probing" experiments by freezing all the learned parameters from pre-training, and only training a linear classifier on top. Note that this preserves the information entirely from pre-training. From Table 5, we observe that the model fails to learn the essential cues necessary for accurate classification on the Waterbirds (Sagawa et al., 2020a) dataset. This is evident from both poor train and test accuracy over the worst-group samples (where spurious correlations do not hold). Interestingly, simply preserving the pre-training distribution does not sufficiently provide ViT's robustness to spurious correlation. Further, Table 8 (Appendix) summarizes the robustness of RVT (Mao et al., 2022) models when fine-tuned on Waterbirds. Note, RVT architecture is specifically modified to provide additional robustness against adversarial attacks, common corruptions, and out-of-distribution inputs. Our results indicate that despite these additional architectural modifications, ImageNet-1k pre-trained RVT models perform poorly when fine-tuned on Waterbirds, indicating a lack of robustness to spurious correlations. However, we do notice that RVT displays stronger robustness to spurious correlation than ViT when pre-trained on ImageNet-1k. In this regard, we believe our work will trigger a new round of thinking in the community about the problem of spurious correlations and ways to mitigate it, from an architectural design perspective.

5.3 Understanding role of self-attention mechanism for improved robustness in ViT 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

Table 5: Investigating the effect of linear probing pre-trained models on Waterbirds (Sagawa et al., 2020a).

	Model	Params	Train Accuracy		Test Accuracy	
			Average	Worst-Group	Average	Worst-Group
IN-21k	ViT-B/16	86.1M	98.6	80.3	89.5	68.3
	ViT-S/16	21.8M	95.6	39.3	86.2	35.9
	ViT-Ti/16	5.6M	88.5	13.2	82.3	0.1
IN-21k	BiT-M-R50x3	211M	100	100	90.5	68.3
	BiT-M-R101x1	42.5M	99.9	98.2	88.5	64.9
	BiT-M-R50x1	23.5M	99.8	98.0	88.6	62.1

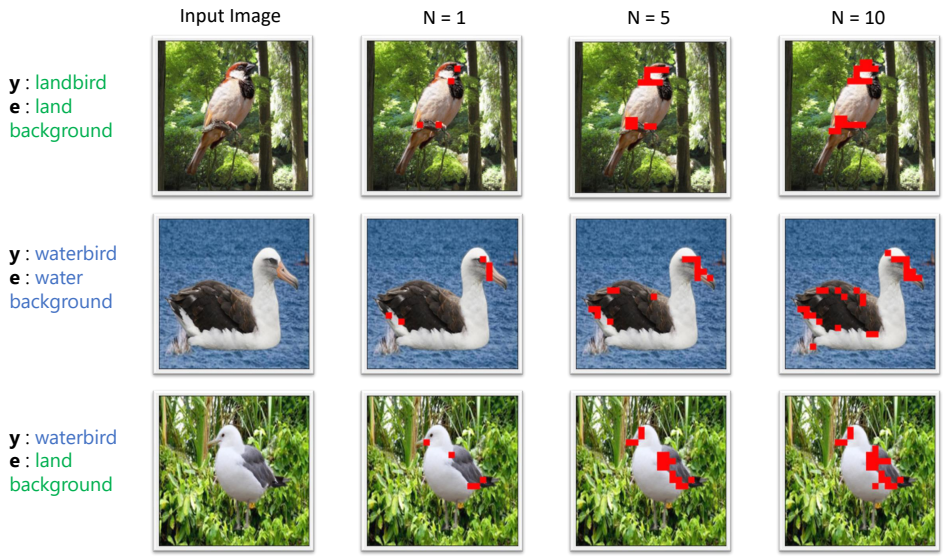

Figure 3: Visualization of the top N patches receiving the highest attention (marked in red). Investigating the attention matrix, we find that all image patches—irrespective of spatial location— provide maximum attention to the patches representing essential cues for accurately identifying the foreground object such as claw, beak, and fur color. See text for details. See Supplementary for visualizations for other datasets and models.

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 the 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 3 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 3, 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

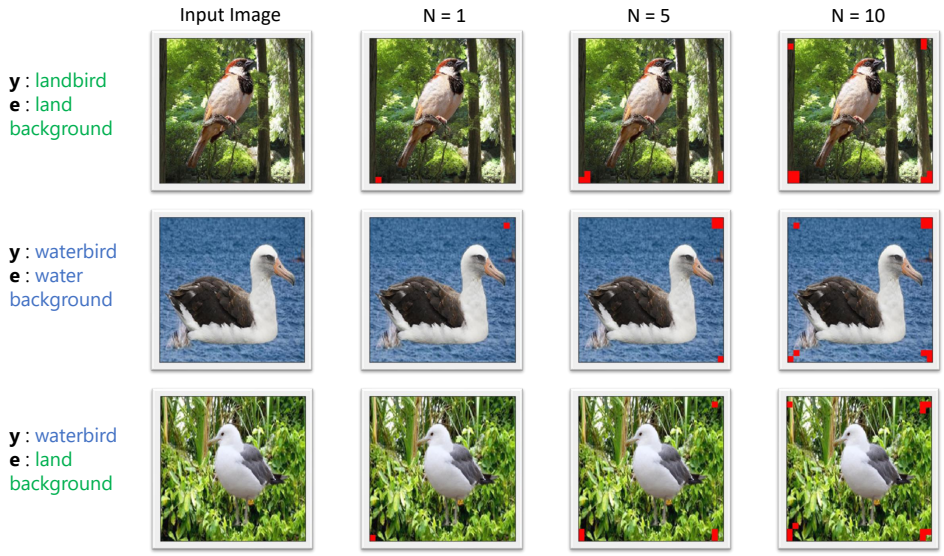

Figure 4: Visualization of the top N patches receiving the highest attention (marked in red) for DeiT-B/16 model finetuned on Waterbirds (Sagawa et al., 2020a) dataset.

which are essential for the classification task `waterbird` vs `landbird`. It is particularly interesting to note the last row in Figure 3, 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.

Further, we show similar visualizations for the DeiT-B/16 model fine-tuned on the waterbirds dataset in Figure 4. We observe that the model completely focuses on background patches for making predictions. This indicates that the model has learned the spurious associations between background features and target labels for making classification, and hence explains the low robustness accuracy of the model. We also provide observations for ViT-Ti/16 model in Appendix G. Specifically, for ViT-Ti/16, we notice: (1) The model correctly attends to patches responsible for accurate classification in images belonging to the majority groups, i.e., `waterbird` on `water` background and `landbird` on `land` background. (2) For images belonging to minority groups such as `waterbird` on `land` background, the model provides maximum attention to the environmental features, exhibiting a lack of robustness to spurious correlations. These experiments further help us to observe a direct correlation between model robustness and patches being attended.

5.4 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 inherent data imbalance. In this section, we investigate the effect of data imbalance on the model's performance. In an 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 4.1.

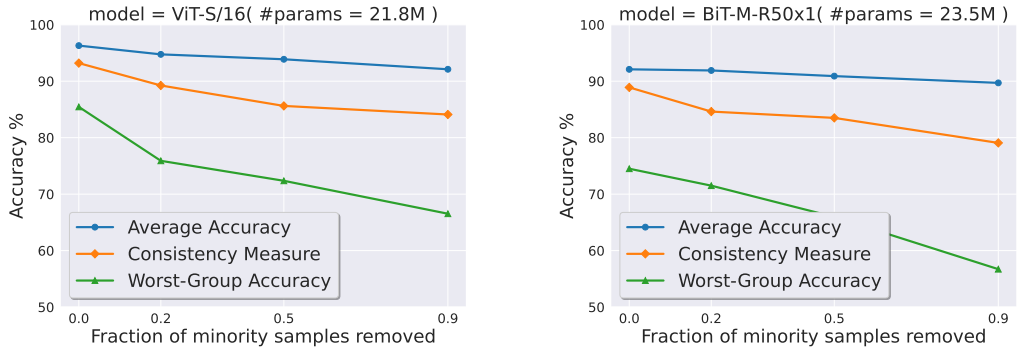

Figure 5: **Data Imbalance.** We investigate the effect of data imbalance on different model architectures. Our findings reveal that both ViT and BiT models suffer from spurious correlations when minority samples are scarce in the fine-tuning dataset.

Takeaways. In Figure 5, we report results for ViT-S/16 and BiT-M-R50x1 model when finetuned on Waterbirds dataset (Sagawa et al., 2020a). 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 observe that for a fixed fraction of minority samples ViTs are more robust than BiT models. Thus, we conclude that the additional robustness of ViT models to spurious associations stems from their better generalization capability from minority samples. However, both ViTs and BiTs still suffer from spurious correlations when minority examples are scarce.

5.5 Does longer fine-tuning in ViT improve robustness to spurious correlations?

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., 2020a). 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.

6 Spurious Out-of-Distribution Detection

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 $\texttt{person} \notin \{\texttt{waterbird}, \texttt{landbird}\}$, 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.

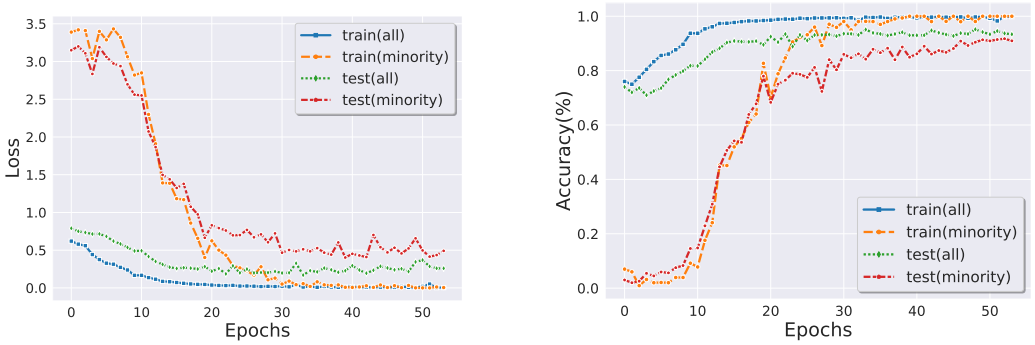

Figure 6: **Longer Fine-tuning.** We study the effect of longer fine-tuning on the performance of ViT models. We report loss and accuracy for ViT-S/16 model finetuned on Waterbirds (Sagawa et al., 2020a) at each epoch of fine-tuning. Investigating further we observe that although fine-tuning for more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

Table 6: **Spurious OOD evaluation.** OOD detection performance of ViT and BiT models when finetuned on Waterbirds

Model	Waterbirds		CelebA		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

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., 2020a) 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. We report our findings in Table 6. 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., 2020a) as in-distribution.

7 Conclusion

In this paper, we investigate the robustness of ViT models when learned on datasets containing spurious associations between target labels and environmental features. Our study lead to a series of new findings including: 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, both vision transformers and CNNs display limited robustness to spurious correlation; 2) Under spurious correlation, self-attention mechanism in ViTs plays a crucial role in guiding the model to focus on spatial locations in an image which are essential for accurately predicting the target label. 3) Improved robustness of ViT models can be attributed to better generalization capability from the counterexamples where spurious correlations do not hold. However, when such samples become scarce ViT models tend to overfit to spurious associations. We hope that our work will inspire future research on understanding the robustness of ViT models in the presence of spurious correlation.

References

Samira Abnar and Willem Zuidema. 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.

Yutong Bai, Jieru Mei, Alan Yuille, and Cihang Xie. Are transformers more robust than cnns? In *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021.

Srinadh Bhojanapalli, Ayan Chakrabarti, Daniel Glasner, Daliang Li, Thomas Unterthiner, and Andreas Veit. Understanding robustness of transformers for image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10231–10241, 2021.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc Le, and Ruslan Salakhutdinov. 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.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pp. 4171–4186, 2019.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations*, 2019.

Karan Goel, Albert Gu, Yixuan Li, and Christopher Ré. Model patching: Closing the subgroup performance gap with data augmentation. In *International Conference on Learning Representations*, 2021.

He He, Sheng Zha, and Haohan Wang. 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.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019.

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

Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15262–15271, 2021.

Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. 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.

Hong Liu, Jeff Z HaoChen, Adrien Gaidon, and Tengyu Ma. Self-supervised learning is more robust to dataset imbalance. *arXiv preprint arXiv:2110.05025*, 2021a.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10012–10022, 2021b.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11976–11986, 2022.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3730–3738, 2015.

Xiaofeng Mao, Gege Qi, Yuefeng Chen, Xiaodan Li, Ranjie Duan, Shaokai Ye, Yuan He, and Hui Xue. Towards robust vision transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12042–12051, 2022.

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

Yifei Ming, Hang Yin, and Yixuan Li. On the impact of spurious correlation for out-of-distribution detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022.

Muzammal Naseer, Kanchana Ranasinghe, Salman Khan, Munawar Hayat, Fahad Khan, and Ming-Hsuan Yang. Intriguing properties of vision transformers. In *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021.

Namuk Park and Songkuk Kim. How do vision transformers work? In *International Conference on Learning Representations*, 2022.

Sayak Paul and Pin-Yu Chen. Vision transformers are robust learners. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 2071–2081, 2022.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. 2018.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, pp. 9, 2019.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763. PMLR, 2021.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, pp. 211–252, 2015.

Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *International Conference on Learning Representations (ICLR)*, 2020a.

Shiori Sagawa, Aditi Raghunathan, Pang Wei Koh, and Percy Liang. An investigation of why overparameterization exacerbates spurious correlations. In *International Conference on Machine Learning*, pp. 8346–8356. PMLR, 2020b.

Yuge Shi, Imant Daunhawer, Julia E Vogt, Philip HS Torr, and Amartya Sanyal. How robust are pre-trained models to distribution shift? *arXiv preprint arXiv:2206.08871*, 2022.

Andreas Steiner, Alexander Kolesnikov, Xiaohua Zhai, Ross Wightman, Jakob Uszkoreit, and Lucas Beyer. How to train your vit? data, augmentation, and regularization in vision transformers. *arXiv preprint arXiv:2106.10270*, 2021.

Rui Tian, Zuxuan Wu, Qi Dai, Han Hu, and Yu-Gang Jiang. Deeper insights into vits robustness towards common corruptions. *arXiv preprint arXiv:2204.12143*, 2022.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pp. 10347–10357. PMLR, 2021a.

Hugo Touvron, Matthieu Cord, Alexandre Sablayrolles, Gabriel Synnaeve, and Hervé Jégou. Going deeper with image transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 32–42, 2021b.

Lifu Tu, Garima Lalwani, Spandana Gella, and He He. An Empirical Study on Robustness to Spurious Correlations using Pre-trained Language Models. *Transactions of the Association for Computational Linguistics*, pp. 621–633, 2020.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 6000–6010, 2017.

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

Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. 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.

Fuzhao Xue, Ziji Shi, Futao Wei, Yuxuan Lou, Yong Liu, and Yang You. Go wider instead of deeper. *arXiv preprint arXiv:2107.11817*, 2021.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 2019.

Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Zi-Hang Jiang, Francis EH Tay, Jiashi Feng, and Shuicheng Yan. 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.

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

Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers, 2021.

Chongzhi Zhang, Mingyuan Zhang, Shanghang Zhang, Daisheng Jin, Qiang Zhou, Zhongang Cai, Haiyu Zhao, Xianglong Liu, and Ziwei Liu. Delving deep into the generalization of vision transformers under distribution shifts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7277–7286, 2022.

Tianyi Zhang, Felix Wu, Arzoo Katiyar, Kilian Q Weinberger, and Yoav Artzi. Revisiting few-sample bert fine-tuning. *arXiv preprint arXiv:2006.05987*, 2020.

Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, pp. 1452–1464, 2017.

Daquan Zhou, Zhiding Yu, Enze Xie, Chaowei Xiao, Animashree Anandkumar, Jiashi Feng, and Jose M Alvarez. Understanding the robustness in vision transformers. In *International Conference on Machine Learning*, pp. 27378–27394. PMLR, 2022.

A Societal Impact

Our project studies robustness and generalization of Vision Transformer (ViT) models when fine-tuned on data containing spurious associations between target label and environmental features. Our study has positive societal impacts and will further benefit understanding robustness of ViT models when used in domains related to fairness and AI safety, such as gender classification and medical imaging. Our study does not involve any human subjects or violation of legal compliance. We do not anticipate any potentially harmful consequence of our work. Through our study and releasing our code, we hope to raise stronger research and societal awareness towards the problem of spurious correlation and its mitigation.

B 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 $\times$ 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 $\times$ 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.

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.

C 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 7. 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.

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

Table 7: Investigating the effect of large scale pre-training on model robustness to spurious correlations when finetuned on CelebA

	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

$1) = \mathbb{P}(e = \texttt{purple}|y = 1) = 0.05$. 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.

E Visualization

E.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., 2020a) 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.

E.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 fore-

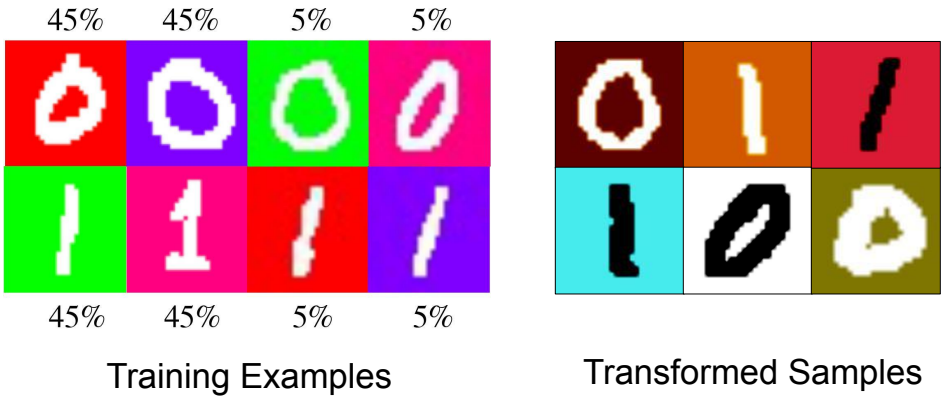

Figure 7: **CMNIST.** Distribution of training samples in CMNIST dataset(**left**) and few representative examples after transformation(**right**) as defined in Section D.

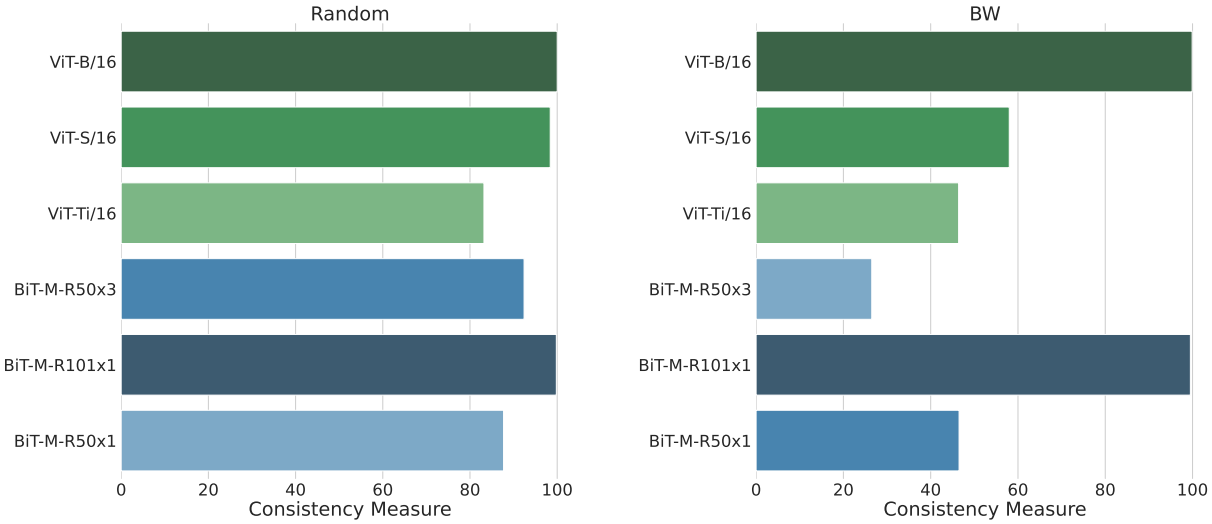

Figure 8: **Consistency Measure.** Evaluation results quantifying consistency for models of different architectures and varying capacity. We indicate the setup when the foreground color is set as `black` and the background as `white` using **BW**(**right**). **Random** represents setting both the foreground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

ground object. In Figure 10, we show visualizations for ViT-B/16 fine-tuned on CMNIST dataset to further validate our findings.

F Software and Hardware

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

G Extension: Pattern in Attention Matrix

In this section, we provide visualizations for top N patches receiving highest attention values for ViT-Ti/16 (Figure 11) fine-tuned on Waterbirds dataset (Sagawa et al., 2020a) on various test images.

Takeaways. Figure 11 shows the top N patches receiving highest attention values for ViT-Ti/16 model fine-tuned on Waterbirds dataset. We observe: (1) The model correctly attends to patches responsible for

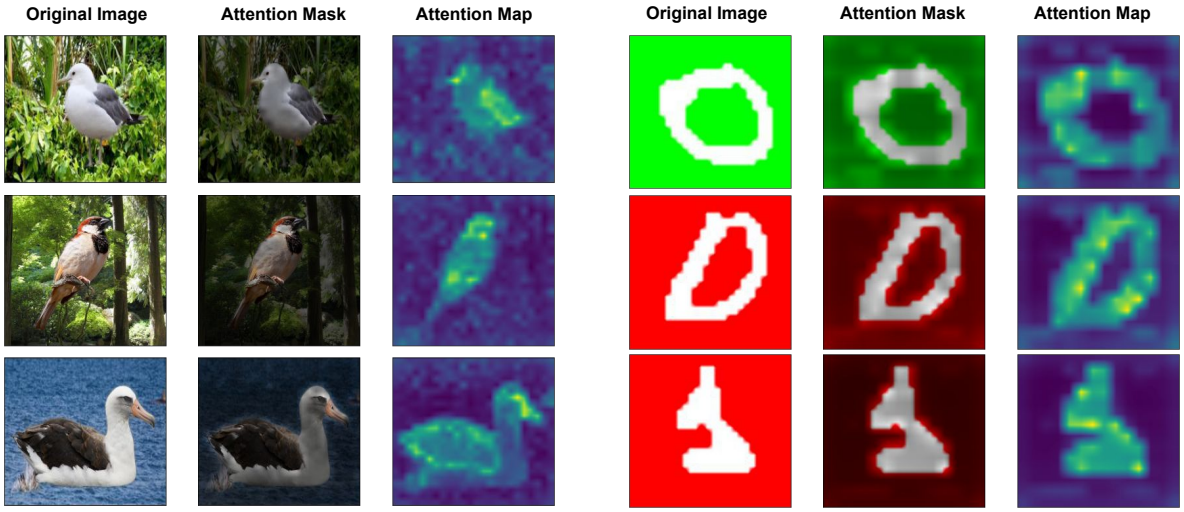

Figure 9: **Attention Map.** Visual illustration of attention map obtained from ViT-B/16 model for few representative images.

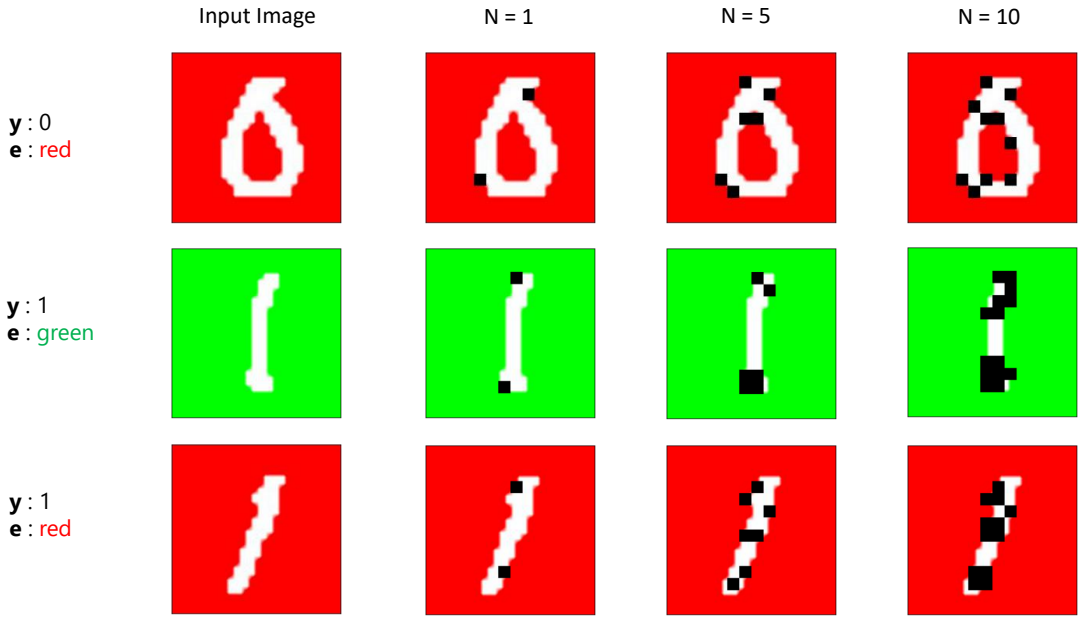

Figure 10: Visualization of the top N patches receiving the highest attention (marked in **black**) for ViT-B/16 fine-tuned on CMNIST. Investigating the attention matrix, we find that all image patches—irrespective of spatial location—provides maximum attention to the patches representing essential cues

accurate classification in images belonging to the majority groups, i.e, `waterbird` on `water` background and `landbird` on `land` background. (2) For images belonging to minority groups (3rd and 4th row in Figure 11) such as `waterbird` on `land` background, the model provides maximum attention to the environmental features, exhibiting lack of robustness to spurious correlations.

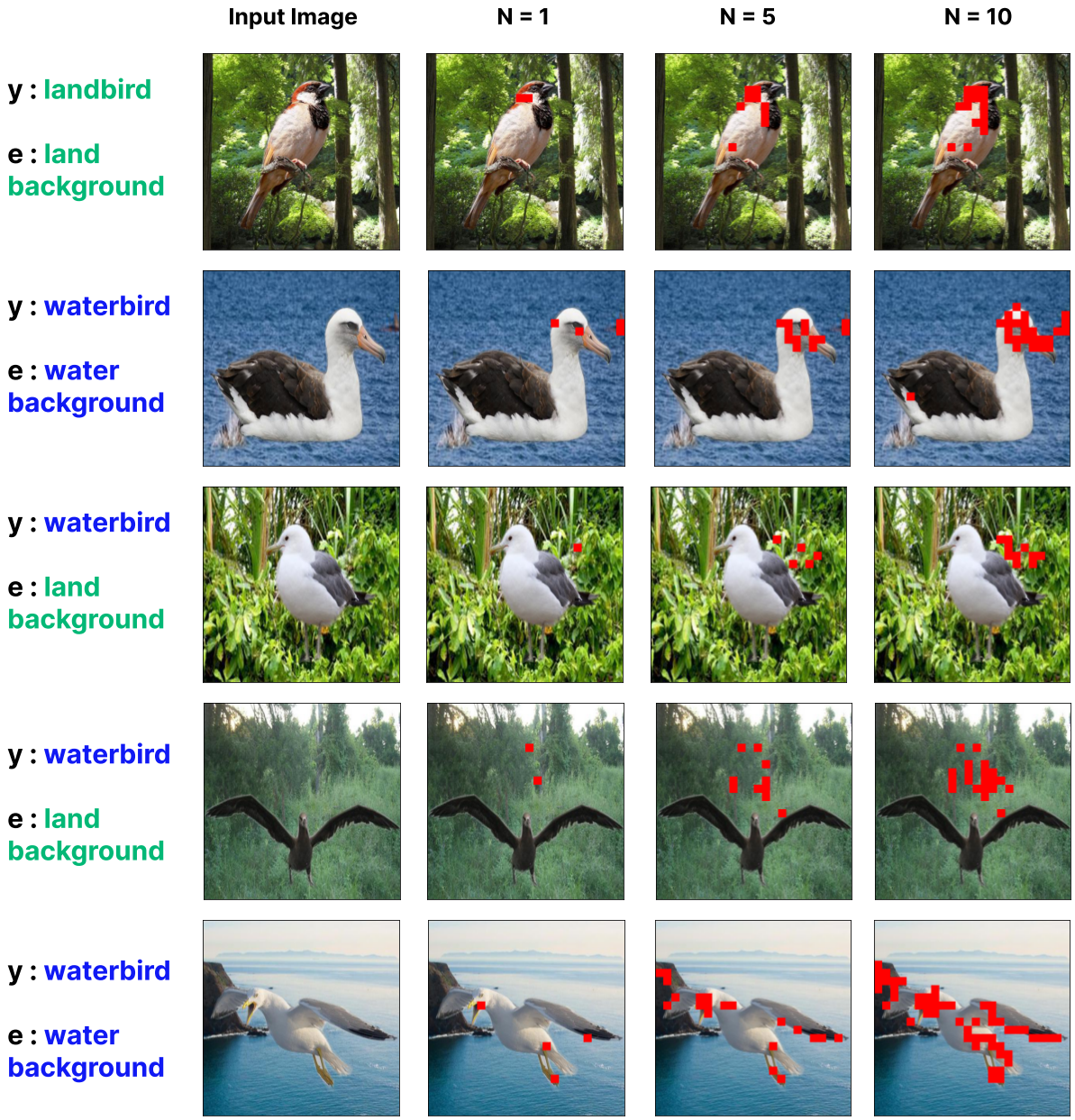

Figure 11: Visualization of the top N patches receiving the highest attention (marked in red) for ViT-Ti/16 model finetuned on Waterbirds (Sagawa et al., 2020a) dataset.

H Extension: Results highlighting importance of pre-training dataset on additional architectures

In this section, to further validate our findings on the importance of large-scale pre-training dataset we show results for additional transformer architectures: RVT (Mao et al., 2022) and PVT (Wang et al., 2021).

Takeaways: We report the results of finetuning different model architectures on Waterbirds (Sagawa et al., 2020a) dataset in Table 8. Note, RVT architecture is specifically modified to provide additional robustness against adversarial attacks, common corruptions and out-of-distribution inputs. Despite these modifications, RVT models when pre-trained on ImageNet-1k perform poorly when fine-tuned on Waterbirds indicating

Table 8: Investigating the effect of large scale pre-training on model robustness to spurious correlations when finetuned on Waterbirds

	Model	Params	Train Accuracy		Test Accuracy	
			Average	Worst-Group	Average	Worst-Group
IN-1k	RVT-Base	85.5M	100	100	92.7	66.5
	RVT-Small	21.8M	100	100	90.2	67.8
	RVT-Tiny	8.9M	100	100	87.8	48.2
IN-1k	PVT-Large	60.9M	100	100	89.9	56.9
	PVT-Medium	43.7M	100	100	89.3	57.5
	PVT-Small	23.9M	100	100	88.8	49.2
	PVT-Tiny	12.7M	100	100	85.1	34.5

lack of robustness to spurious correlations. However, we do notice that RVT displays stronger robustness to spurious correlation than ViT when pre-trained on ImageNet-1k.

I Experiments on training models from scratch

In section, we train both ViT and BiT models from scratch on Waterbirds(Sagawa et al., 2020a) dataset. We observe that without any pre-training both ViT and BiT models severely overfit the training dataset to attain 100% training accuracy and significantly fail on test samples. This observation indicates that without pre-training, both transformers and CNNs have a high propensity of memorizing training samples (along with their inherent bias).

Table 9: Investigating the effect of training models on Waterbirds

	Model	Params	Train Accuracy		Test Accuracy	
			Average	Worst-Group	Average	Worst-Group
IN-21k	ViT-B/16	86.1M	100	100	72.9	0.1
	ViT-S/16	21.8M	100	100	63.9	0.02
	ViT-Ti/16	5.6M	100	100	63.3	0.01
IN-21k	BiT-M-R50x3	211M	100	100	58.6	0.09
	BiT-M-R101x1	42.5M	100	100	58.4	0.03
	BiT-M-R50x1	23.5M	100	100	54.2	0.01