Vision Transformers are Robust Learners

Sayak Paul*
PyImageSearch
s.paul@pyimagesearch.com

Pin-Yu Chen*
IBM Research
pin-yu.chen@ibm.com

Abstract

Transformers, composed of multiple self-attention layers, hold strong promises toward a generic learning primitive applicable to different data modalities, including the recent breakthroughs in computer vision achieving state-of-the-art (SOTA) standard accuracy with better parameter efficiency. Since self-attention helps a model systematically align different components present inside the input data, it leaves grounds to investigate its performance under model robustness benchmarks. In this work, we study the robustness of the Vision Transformer (ViT) [1] against common corruptions and perturbations, distribution shifts, and natural adversarial examples. We use six different diverse ImageNet datasets concerning robust classification to conduct a comprehensive performance comparison of ViT [1] models and SOTA convolutional neural networks (CNNs), Big-Transfer [2]. Through a series of six systematically designed experiments, we then present analyses that provide both quantitative and qualitative indications to explain why ViTs are indeed more robust learners. For example, with fewer parameters and similar dataset and pre-training combinations, ViT gives a top-1 accuracy of 28.10% on ImageNet-A which is 4.3x higher than a comparable variant of BiT. Our analyses on image masking, Fourier spectrum sensitivity, and spread on discrete cosine energy spectrum reveal intriguing properties of ViT attributing to improved robustness. Code for reproducing our experiments is available here: git.io/J3VO0.

1 Introduction

Transformers [3] are becoming a preferred architecture for various data modalities. This is primarily because they help reduce inductive biases that go into designing network architectures. Moreover, Transformers have been shown to achieve tremendous parameter efficiency without sacrificing predictive performance over architectures that are often dedicated to specific types of data modalities. Attention, in particular, self-attention is one of the foundational blocks of Transformers. It is a computational primitive that allows us to quantify pairwise entity interactions thereby helping a network learn the hierarchies and alignments present inside the input data [4, 3]. These are desirable properties to eliminate the need for carefully designed inductive biases to a great extent.

Although Transformers have been used in prior works [5, 6] it was only until 2020, the performance of Transformers were on par with the SOTA CNNs on standard image recognition tasks [7, 8, 1]. Attention has been shown to be an important element for vision networks to achieve better empirical robustness [9]. Since attention is a core component of ViTs (and Transformers in general) a question that naturally gets raised here is - *could ViTs be inherently more robust?* If so, *why ViTs are more robust learners?* In this work, we provide an affirmative answer to the first question and provide empirical evidence to reason about the improved robustness of ViTs.

Various recent works have opened up the investigation on evaluating the robustness of ViTs [10–12] but with a relatively limited scope. We build on top of these and provide further and more comprehensive analyses to understand why ViTs provide better robustness for semantic shifts, common

*The authors contributed equally to this work.

Preprint. Under review.

corruptions and perturbations, and natural adversarial examples to input images in comparison to SOTA CNNs like Big Transfer (BiT) [2]. Through a set of carefully designed experiments, we first verify the enhanced robustness of ViTs to common robustness benchmark datasets [13, 14, 9, 15]. We then provide quantitative and qualitative analyses to help understand the reasons behind this enhancement. In summary, we make the following contributions:

- We use 6 diverse ImageNet datasets concerning different types of robustness evaluation and conclude that ViTs achieve significantly better performance than BiTs.
- We design 6 experiments, including robustness to masking, energy and loss landscape analysis, and sensitivity to high-frequency artifacts to reason about the improved robustness of ViTs.
- Our analysis provides novel insights for robustness attribution of ViT. Moreover, our robustness evaluation and analysis tools are generic and can be used to benchmark and study future image classification models. Code for reproducing our experiments is available here: git.io/J3VO0.

2 Related Work

To the best of our knowledge, [16] first explored the use of Transformers [3] for the task of image super-resolution [17–19] which essentially belongs to the category of image generation. Image-GPT [6] used Transformers for unsupervised pre-training from pixels of images. However, the transfer performance of the pre-training method is not on par with supervised pre-training methods. ViT [1] takes the original Transformers and makes very minimal changes to make it work with images. In fact, this was one of the primary objectives of ViT i.e. to keep the original Transformer architecture as original as possible and then examining how that pans out for image classification in terms of large-scale pre-training. As noted in [1], because of the lesser number of inductive biases, ViT needs to be pre-trained on a relatively larger dataset (such as ImageNet-21k [20]) for achieving reasonable downstream performance.

Multiple variants of Transformers have been proposed to show that it is possible to achieve comparable performance on ImageNet-1k *without* using additional data. DeIT [8] introduces a novel distillation strategy [21] to learn a student Transformers-based network from a well-performing teacher network based on RegNets [22]. With this approach, DeIT achieves 85.2% top-1 accuracy on ImageNet-1k without any external data. T2T-ViT [23] proposes a novel tokenization method enabling the network to have more access to local structures of the images. For the Transformer-based backbone, it follows a deep-narrow network topology inspired by [24]. With proposed changes, T2T-ViT achieves 83.3% top-1 accuracy on ImageNet-1k. LV-ViT [25] introduces a new training objective namely token labeling and also tunes the structure of the Transformers. It achieves 85.4% top-1 accuracy on ImageNet-1k. In this work, we only focus on ViT [1].

Concurrent to our work, there are a few recent works that study the robustness of ViTs from different perspectives. In what follows, we summarize their key insights and highlight the differences from our work. [11] showed that ViTs has better robustness than CNNs against adversarial input perturbations. The major performance gain can be attributed to the capability of learning high-frequency features that are more generalizable and the finding that convolutional layers hinder adversarial robustness. [10] studied improved robustness of ViTs over ResNets [26] against adversarial and natural adversarial examples as well as common corruptions. Moreover, it is shown that ViTs are robust to the removal of almost any single layer. [12] studied adversarial robustness of ViTs through various white-box, black-box and transfer attacks and found that model ensembling can achieve unprecedented robustness without sacrificing clean accuracy against a black-box adversary. This paper shows novel insights that are fundamentally different from these works: (**i**) we benchmark the robustness of ViTs on a wide spectrum of ImageNet datasets (see Table 2), which are the most comprehensive robustness performance benchmarks to date; (**ii**) we design six new experiments to verify the superior robustness of ViTs over BiT and ResNet models.

3 Robustness Performance Comparison on ImageNet Datasets

3.1 Preliminaries

In this section, we will provide an overview of ViT. The main components of ViT are as follows.

Multi-head Self Attention (MHSA). Central to ViT's model design is self-attention [4]. Here, we first compute three quantities from linear projections ($X \in \mathbb{R}^{N \times D}$): **(i)** **Q**uery = XW_{Q}, **(ii)** **K**ey = XW_{K} , and **(iii)** **V**alue = XW_{V}, where W_{Q}, W_{K}, and W_{V} are linear transformations. The linear projections (X) are computed from batches of the original input data. Self-attention takes these three input quantities and returns an output matrix ($N \times d$) weighted by attention scores using (1):

$$\mathrm{Attention}(Q, K, V) = \mathrm{Softmax}\left(QK^{\top}/\sqrt{d}\right)V \tag{1}$$

This form of attention is also popularly referred to as the "scaled dot-product attention" [3]. One important aspect of self-attention is that it operates between all pairs of elements within an input. In summary, a single attention layer tries to find out how to best align the keys to the queries and quantifies this finding in the form of attention scores. These scores are then multiplied with the values to obtain the final output. To enable feature-rich hierarchical learning, h self-attention layers (or so-called "heads") are stacked together producing an output of $N \times dh$. This output is then fed through a linear transformation layer that produces the final output of $N \times d$ from MHSA. MHSA then forms the core Transformer block.

Transformer block. A single transformer block is composed of MHSA, Layer Normalization (LN) [27], feed-forward network (FFN), and skip connections [26]. It is implemented using (2):

$$\mathbf{z}'_{\ell} = \mathrm{MHSA}\left(\mathrm{LN}\left(\mathbf{z}_{\ell-1}\right)\right) + \mathbf{z}_{\ell-1};\ \mathbf{z}_{\ell} = \mathrm{FFN}\left(\mathrm{LN}\left(\mathbf{z}'_{\ell}\right)\right) + \mathbf{z}'_{\ell};\ \mathbf{y} = \mathrm{LN}\left(\mathbf{z}^{0}_{L}\right), \tag{2}$$

where $\ell \in \{0, 1, \ldots, L\}$ is the layer index and L is the number of hidden layers.

The FFN is comprised of two linear layers with a GELU non-linearity [28] in between them. We refer the reader to Figure 1 of [1] for a pictorial overview of the Transformer block. Next, we discuss the class-token (learned version of which is represented as z^0_L in (2)) and how images are fed to a Transformer block with patch encoding.

Class token and encoded patches of images. Inspired by BERT [29], a class token is prepended to the image patches and it flows through the entirety of ViT. It is initialized as z^0_0 and serves as the final representation of the image patches which is then passed to task head. Transformers can only process sequences of inputs. Consider an image of $N \times N$ shape. If we were to extract patches of shape $P \times P$ then the total number of patches would be $(N/P)^2$ (see Appendix A for more details).

A Transformer block processes these patches in parallel which makes it invariant to the order in which the patches would appear. Since locality is not just desirable but also is necessary especially in images, a learnable position encoder is used to get linear projections of the positions of the image patches. These projections are combined with the linear projections of the actual image patches and are then fed to the subsequent Transformer blocks. In [1], the authors also investigate hybrid models wherein the patch encoding is applied on feature maps computed using a CNN. However, in this work, we do not consider those.

3.2 Performance Comparison on Diverse ImageNet Datasets for Robustness Evaluation

In this work, our baseline is a ResNet50V2 model [30] pre-trained on the ImageNet-1k dataset [31] except for a few results where we consider ResNet-50 [26][2]. To study how ViTs hold up with the SOTA CNNs we consider BiT [2]. At its core, BiT networks are scaled-up versions of ResNets with Group Normalization [32] and Weight Standardization [33] layers added in place of Batch Normalization [34]. Since ViT and BiT share similar pre-training strategies (such as using larger datasets like ImageNet-21k [20] and JFT-300 [35], longer pre-training schedules, and so on) they are excellent candidates for our comparison purposes. So, a question, central to our work is:

Where does ViT stand with respect to BiT in terms of robustness under similar parameter, pre-training, and data regimes, and how to attribute their performance difference?

Table 1 reports the parameter counts of the different ViT and BiT models that are publicly available along with their top-1 accuracy[3] on the ImageNet-1k dataset [31]. It is clear that different variants of ViT are able to achieve comparable performance to BiT but with lesser parameters.

[2]In these cases, we directly referred to the previously reported results with ResNet-50.

[3]Figure 4 of [2] and Table 5 of [1] were used to collect the top-1 accuracy scores.

Table 1: Parameter counts and top-1 accuracy (%) of different variants of ViT and BiT. All the reported variants were pre-trained on ImageNet-21k and then fine-tuned on ImageNet-1k.

Variant	# Parameters (Million)	ImageNet-1k Top-1 Acc
ResNet50V2	25.6138	76
BiT m-r50x1	25.549352	80
BiT m-r50x3	217.31908	84
BiT m-r101x1	44.54148	82.1
BiT m-r101x3	387.934888	84.7
BiT m-r152x4	936.53322	85.39
ViT B-16	86.859496	83.97
ViT B-32	88.297192	81.28
ViT L-16	304.715752	85.15
ViT L-32	306.63268	80.99

Table 2: Summary of the datasets used in this work and their purpose.

Dataset	Purpose
ImageNet-C [13]	Common corruptions
ImageNet-P [13]	Common perturbations
ImageNet-R [14]	Semantic shifts
ImageNet-O [9]	Out-of-domain distribution
ImageNet-A [9]	Natural adversarial examples
ImageNet-9 [15]	Background dependence

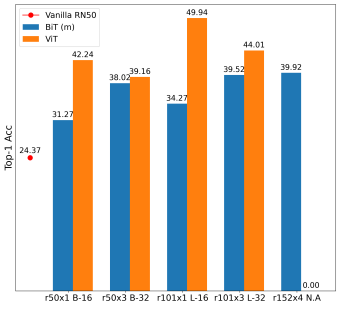

Figure 1: Mean top-1 accuracy scores (%) on the ImageNet-C dataset as yielded by different variants of ViT and BiT.

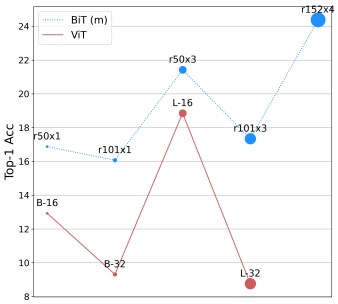

Figure 2: Top-1 accuracy (%) of ViT and BiT for contrast corruption (with the highest severity level) on ImageNet-C [13].

Table 3: mCEs (%) of different models and methods on ImageNet-C (lower is better). Note that Noisy Student Training incorporates additional training with data augmentation for noise injection.

Model / Method	mCE
ResNet-50	76.7
BiT `m-r101x3`	58.27
DeepAugment+AugMix [14, 37]	53.6
ViT L-16	45.45
Noisy Student Training [38]	28.3

In what follows, we compare the performance of ViT and BiT on six robustness benchmark datasets [13, 14, 9], as summarized in Table 2. These datasets compare the robustness of ViT, BiT and the baseline ResNet50V2 in different perspectives, including (**i**) common corruptions, (**ii**) semantic shifts, (**iii**) natural adversarial examples, and (**iv**) out-of-distribution detection. A summary of the datasets and their purpose is presented in Table 2 for easier reference.

Notably, in these datasets ViT exhibits significantly better robustness than BiT of comparable parameter counts. The attribution analysis of improved robustness in ViT is given in Section 4.

ImageNet-C. The ImageNet-C dataset [13] consists of 15 types of algorithmically generated corruptions, and each type of corruption has five levels of severity. Along with these, the authors provide additional four types of general corruptions making a total of 19 corruptions. We consider all the 19 corruptions at their highest severity level (5) and report the mean top-1 accuracy in Figure 1 as yielded by the variants of ViT and BiT. We consistently observe a better performance across all the variants of ViT under different parameter regimes. Note that BiT `m-r50x1` and `m-r101x1` have lesser parameters than the lowest variant of ViT (`B-16`) but for other possible groupings, variants of ViT have lesser parameters than that of BiT. Overall, we notice that ViT performs consistently better across different corruptions except for *contrast*. In Figure 2, we report the top-1 accuracy of ViT and BiT on the highest severity level of the contrast corruption. This observation leaves grounds for future research to investigate why this is the case since varying contrast factors are quite common in real-world use-cases. Based on our findings, contrast can be an effective but unexplored approach to studying ViT's robustness, similar to the study of human's attention mechanism [36].

In [13], mean corruption error (mCE) is used to quantify the robustness factors of a model on the ImageNet-C dataset. Specifically, the top-1 error rate is computed for each of the different corruption (c) types ($1 \leq c \leq 15$) and for each of the severity (s) levels ($1 \leq s \leq 5$). When error rates for all

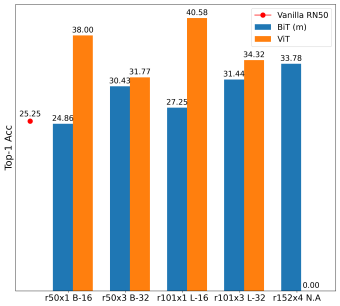

Figure 3: Top-1 accuracy scores (%) on ImageNet-R dataset [14].

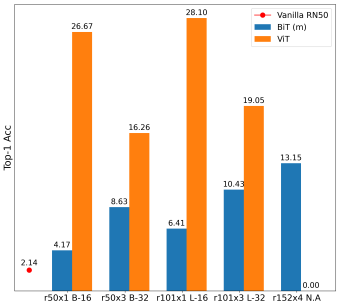

Figure 4: Top-1 accuracy scores (%) on ImageNet-A dataset [9].

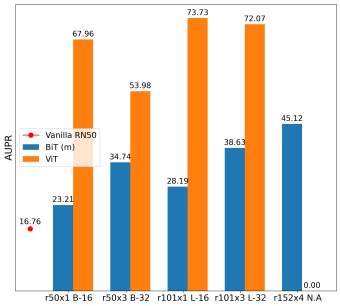

Figure 5: AUPR (higher is better) on ImageNet-O dataset [9].

the severity levels are calculated for a particular corruption type, their average is stored. This process is repeated for all the corruption types and the final value is an average over all the average error rates from the different corruption types. The final score is normalized by the mCE of AlexNet [39].

We report the mCEs for BiT-m `r101x3`, ViT L-16, and a few other models in Table 3. The mCEs are reported for 15 corruptions as done in [13]. We include two additional models/methods in Table 3 because of the following: (**a**) Noisy Student Training [38] uses external data and training choices (such as using RandAugment [40], Stochastic Depth [41], etc.) that are helpful in enhancing the robustness of a vision model, (**b**) DeepAugment and AugMix [14, 37] are designed explicitly to improve the robustness of models against corruptions seen in ImageNet-C. So, to provide a fair ground to see where BiT and ViT stand in comparison to state-of-the-art we add these two models. It is indeed interesting to notice that ViT is able to outperform the combination of DeepAugment and AugMix which are specifically designed to provide robustness against the corruptions found in ImageNet-C. As we will discuss in Section 4, this phenomenon can be attributed to two primary factors: (**a**) longer pre-training with a larger dataset and (**b**) self-attention. It should also be noted that Noisy Student Training [38] incorporates various factors during training such as an iterative training procedure, strong data augmentation transformations from RandAugment for noise injection, test-time augmentation, and so on. These factors largely contribute to the improved robustness gains achieved by Noisy Student Training.

ImageNet-P. The ImageNet-P dataset [13] has 10 types of common perturbations. Unlike the common corruptions, the perturbations are subtly nuanced spanning across fewer number of pixels inside images. As per [13] mean flip rate (mFR) and mean top-5 distance (mT5D) are the standard metrics to evaluate a model's robustness under these perturbations. They are reported in Table 4. Since the formulation of mFR and mT5D are more involved than mCE and for brevity, we refer the reader to [13] for more details on these two metrics. We find ViT's robustness to common perturbations is significantly better than BiT as well as AugMix.

Table 4: mFRs (%) and mT5Ds (%) on ImageNet-P dataset (lower is better).

Model / Method	mFR	mT5D
ResNet-50	58	82
BiT-m `r101x3`	49.99	76.71
AugMix [37]	37.4	NA
ViT L-16	33.064	50.15

ImageNet-R. The ImageNet-R dataset [14] contains images labelled with ImageNet labels by collecting renditions of ImageNet classes. It helps verify the robustness of vision networks under semantic shifts under different domains. Figure 3 shows that ViT's treatment to domain adaptation is better than that of BiT.

ImageNet-A. The ImageNet-A dataset [9] is comprised of natural images that cause misclassifications. One of the major reasons behind this is in multi-class image classification problems, images that have multiple objects get associated with single discrete categories [9]. Other reasons include the texture bias of CNNs [42]. In Figure 4, we report the top-1 accuracy of ViT and BiT on the ImageNet-A dataset [9]. In [9], self-attention is noted as an important element to tackle these problems. This may help explain why ViT performs significantly better than BiT in this case.

ImageNet-O. The ImageNet-O dataset [9] consists of images that belong to different classes not seen by a model during its training and are considered as *anomalies*. For these images, a robust model is expected to output low confidence scores. We follow the same evaluation approach of using *area*

Table 5: Top-1 accuracy (%) of ImageNet-9 dataset and its different variants. "BG-Gap" is the between "Mixed-Same" and "Mixed-Rand". It measures how impactful background correlations are in presence of correct-labeled foregrounds.

Model	Original	Mixed-Same	Mixed-Rand	BG-Gap
BiT-m `r101x3`	94.32	81.19	76.62	4.57
ResNet-50	95.6	86.2	78.9	7.3
ViT L-16	96.67	88.49	81.68	6.81

Table 6: Performance on detecting vulnerable image foregrounds from ImageNet-9 dataset.

Model	Challenge Accuracy (%)
BiT-m `r101x3`	3.78
ViT L-16	20.02
ResNet-50	22.3

under the precision-recall curve (AUPR) as [9] for this dataset. In Figure 5, we report the AUPR of the different ViT and BiT models on the ImageNet-O dataset [9]. ViT demonstrates superior performance in anomaly detection than BiT.

ImageNet-9. ImageNet-9, proposed in [15] helps to verify the background-robustness of vision models. In most cases, the foregrounds of images inform our decisions on what might be present inside images. Even if the backgrounds change, as long as the foregrounds stay intact, these decisions should not be influenced. However, do vision models exhibit a similar kind of treatment to image foregrounds and backgrounds? It turns out that the vision models may break down when the background of an image is changed [15]. It may suggest that the vision models may be picking up unnecessary signals from the image backgrounds. In [15] it is also shown that background-robustness can be important for determining models' out of distribution performance. So, naturally, this motivates us to investigate if ViT would have better background-robustness than BiT. We find that is indeed the case (refer to Table 6). Additionally, in Table 6, we report how well BiT and ViT can detect if the foreground of an image is vulnerable[4]. It appears that for this task also, ViT significantly outperforms BiT. Even though we notice ViT's better performance than BiT but it is surprising to see ViT's performance being worse than ResNet-50. We suspect this may be due to the simple tokenization process of ViT to create small image patches that limits the capability to process important local structures [23].

4 Attributions for Improved Robustness in ViTs

In this section, we systematically design and conduct six experiments to identify the sources of improved robustness in ViTs from both qualitative and quantitative standpoints.

4.1 Longer Pre-training Schedule and Larger Pre-training Dataset Improve Robustness

As pointed out in [43, 2, 44], a longer pre-training schedule on a larger pre-training dataset is transferring effectively to downstream tasks including few-shot learning. It is also shown in [2] that having a larger pre-training dataset adds a regularization effect during downstream tasks. To this end, we raise the question - *could a longer pre-training schedule and a larger pre-training dataset be also helpful for enhanced robustness?* To investigate this further, we conduct the following experiment:

Table 7: Relative improvements as achieved by `BiT-m` and ViT variants. BiT `m-r101x3` is comparable to ViT L-16, and ViT L-32 with respect to the number of model parameters (refer to Table 1).

BiT-s Variant	BiT-s Top-1 Acc (%)	BiT-m − BiT-s	ViT − BiT-m
s-r50x1	2.6	+1.57	+22.5 (ViT B-16)
s-r50x3	3.2	+5.43	+7.63 (ViT B-32)
s-r101x1	3.11	+3.3	+21.69 (ViT L-16)
s-r101x3	4.29	+6.14	+8.62 (ViT L-32)
s-r152x4	4.64	+8.51	NA

- We take all the images from the ImageNet-A dataset [9]. This dataset is chosen because it has many properties that are desirable for testing a model's robustness capabilities: (**a**) Objects of interest present in many of the images in the dataset are not centrally oriented, (**b**) Multiple images have multiple objects present inside them which makes it harder for a model to associate the images with discrete individual categories, (**c**) Different images have varying amount of textures that can act as spurious correlations for neural nets to produce misclassifications [42], and (**d**) These traits are not very uncommon to catch in a large portion of real-world images.

[4]For details, we refer the reader to the official repository of the background robustness challenge: git.io/J3TUj.

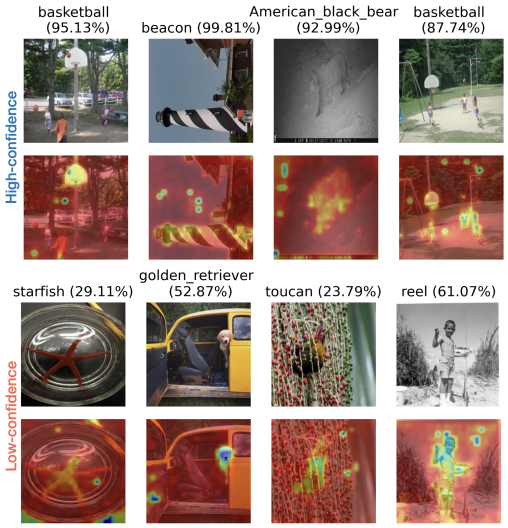

Figure 6: Visualization of the attention maps of ViT on images (top rows) from ImageNet-A.

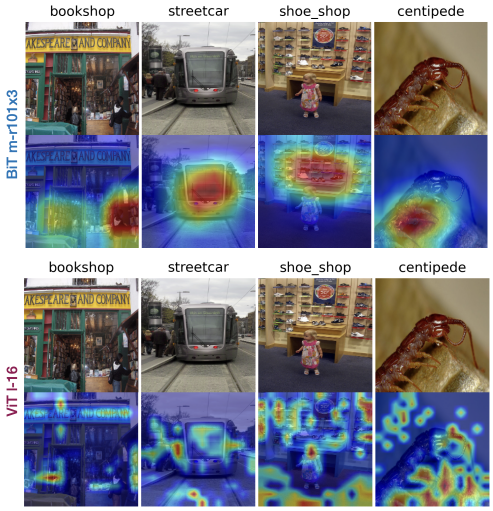

Figure 7: Grad-CAM results for the images where both BiT and ViT give correct predictions.

- Run the images through different variants of BiT and ViT and record the top-1 accuracies. For this experiment, we also include the `BiT-s` variants that are pre-trained on the ImageNet-1k dataset.

Our findings for the above-described experiment are summarized in Table 7. When a longer pre-training schedule is coupled with a larger pre-training dataset it can be helpful in improving a model's performance on the ImageNet-A dataset. Another noticeable trend is that as the model capacity increases the performance also improves. Among all the variants of the different models we present in Table 7, ViT variants consistently perform better than the BiT variants (comparable ones).

4.2 Attention is key to Robustness

We hypothesize that the performance gain discussed in Section 4.1 can be attributed to the use of the attention mechanism. To verify this, we make use of Attention Rollout [45] to visualize the attention maps for two different cases: (**a**) where ViT yields high-confidence *correct* predictions and (**b**) where ViT yields low-confidence *correct* predictions. Figure 6[5] presents a few visualizations of this study.

It is surprising to see that even under such dark lighting conditions, ViT is able to make the correct predictions for the "American Black Beer" class (second-last plot from Figure 6 (a)). On the other hand, for the low-confidence cases, although ViT is still able to produce the correct predictions it is not very clear where it is putting its focus. For example, consider the last plot from Figure 6 (b). ViT draws all its attention to the individual that is standing and *not* on the reel they are holding.

To further investigate the representations learned by ViT and to better understand the spread of the attention span of ViT, we apply Grad-CAM [46] and compare the results to that of BiT. For BiT, we take the last convolutional block for computing the gradients with respect to the target class. Our comparative results are presented in Figure 7. But we cannot apply these steps directly to ViT because of its structure. Hence, we follow the implementation of [47] to compute the gradients of the last attention block reshaped to fit the computations of Grad-CAM.

From Figure 7, it can be noticed that ViT tries to maintain a global context in order to generate the predictions while the explanations for BiT are more local and central. For example, consider the image predicted as "bookshop" in Figure 7. We can observe that ViT uses information from different parts of the image to determine the target class. Objects of interest inside images may not be always centrally aligned with the respect of the objects of images. Besides, capturing long-range dependencies is desirable when dealing with tasks like object detection and segmentation [7]. This is why we hypothesize that ViT should be able to perform well even when some seemingly attentive

[5]For this study, we used the ViT L-16 variant as it yields the best performance on the ImageNet-A dataset (refer to Figure 4).

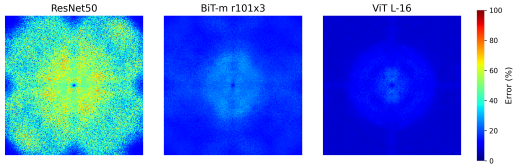

Figure 8: Sensitivity heatmap of 2D discrete Fourier transform spectrum [50]. The low-frequency/high-frequency components are shifted to the center/corner of the spectrum.

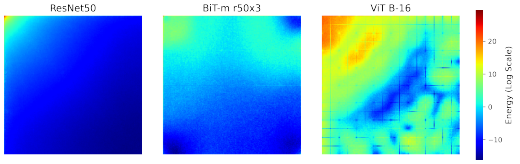

Figure 9: Spectral decomposition of adversarial perturbations generated using DeepFool [51]. The top-left/bottom-right quadrants denote low-frequency/high-frequency regions.

regions of an image are masked out. We study this hypothesis in Section 4.3. It should also be noted that there are some spurious attention regions that are not very explanatory (refer to the image predicted as "centipede") and may lead to future research[6].

4.3 ViT Has Better Robustness to Image Masking

In order to further establish that attention indeed plays an important role for the improved robustness of ViTs, we conduct the following experiment:

- Randomly sample a common set of 1000 images from the ImageNet-1k validation set.
- Apply Cutout [49] at four different levels: {5,10,20,50}% and calculate the mean top-1 accuracy scores for each of the levels with BiT (`m-r101x3`) and ViT (`L-16`)[7]. In Cutout, square regions from input images are randomly masked out. It was originally proposed as a regularization technique.

Table 8: Mean top-1 accuracy (%) of BiT (`m-r101x3`) and ViT (`L-16`) with different masking factors.

Masking Factor	Top-1 Acc (BiT)	Top-1 Acc (ViT)
0	79	83
0.05	76	82.3
0.1	75	81.4
0.2	72.4	77.9
0.5	52	60.4

Table 8 reports that ViT is able to consistently beat BiT when square portions of the input images have been randomly masked out. Randomness is desirable here because ViT can utilize global information. If we fixate the region of masking it may be too restrictive for a ViT to take advantage of its ability to utilize global information. It should be noted that the ViT variant (`L-16`) we use in this experiment is shallower than the BiT variant (`m-r101x3`). This may suggest that attention indeed is the strong force behind this significant gain.

4.4 Fourier Spectrum Analysis Reveals Low Sensitivity

A common hypothesis about vision models is that they can easily pick up the spurious correlations present inside input data that may be imperceptible and unintuitive to humans [52–54, 13]. To measure how ViT holds up with this end of the bargain, we conduct a Fourier analysis [50] of ViT, BiT, and our baseline ResNet-50. The experiment goes as follows:

Table 9: Different percentiles (P) of the error matrix computed from Fourier analysis (Figure 8).

	ResNet50	BiT-m r101x3	ViT L-16
P=10	21.8	13.9	6.7
P=25	30.2	14.8	7
P=50	40.4	16.4	7.6
P=90	58.9	23	13.1
P=95	63.6	24.9	15.1

- Generate a Fourier basis vector with varying frequencies.
- Add the basis vector to 1000 randomly sampled images from the ImageNet-1k validation set.
- Record error-rate for every perturbed image and generate a heatmap of the final error matrix.

For additional details on this experiment, we refer the reader to [50]. In Figure 8, it is noticed that both ViT and BiT stay robust (have low sensitivity) to most of the regions present inside the perturbed images while the baseline ResNet50V2 loses its consistency in the high-frequency regions. The value at location (i, j) shows the error rate on data perturbed by the corresponding Fourier basis noise.

The low sensitivity of ViT and BiT may be attributed to the following factors: (**a**) Both ViT and BiT are pre-trained on a larger dataset and then fine-tuned on ImageNet-1k. Using a larger dataset during pre-training may be acting as a regularizer here [2]. (**b**) Evidence also suggests that increased network width has a positive effect on model robustness [13, 9]. To get a deeper insight into the

[6]DINO [48] shows when ViT is trained with self-supervised objective on a larger data corpus, its self-attention span is made even more pronounced than their self-supervised counterparts. With the virtue of self-supervision, DINO is able to perfectly segment objects of interest from an image *without* any supervision.

[7]We use these two variants because they are comparable with respect to the number model parameters.

heatmaps shown in Figure 8, in Table 9, we report error-rate percentiles for the three models under consideration. For a more robust model, we should expect to see lower numbers across all the five different percentiles reported in Table 9 and we confirm that is indeed the case. This may also help explain the better behavior of BiT and ViT in this experiment.

4.5 Adversarial Perturbations of ViT Has Wider Spread in Energy Spectrum

In [55], it is shown that small adversarial perturbations can change the decision boundary of neural networks (especially CNNs) and that adversarial training [56] exploits this sensitivity to induce robustness. Furthermore, CNNs primarily exploit discriminative features from the low-frequency regions of the input data. Following [55], we conduct the following experiment on 1000 randomly sampled images from the ImageNet-1k validation set with ResNet-50, BiT-m `r50x3`, and ViT B-16[8]:

- Generate small adversarial perturbations (δ) with DeepFool [51] with a step size of 50[9].
- Change the basis of the perturbations with discrete cosine transform (DCT) to compute the energy spectrum of the perturbations.

With this experimental setup, we aim to confirm that ViT's perturbations will spread out the whole spectrum, while perturbations of ResNet-50 and BiT will be centered only around the low-frequency regions. This is primarily because ViT has the ability to better exploit information that is only available in a global context. Figure 9 shows the energy spectrum analysis. It suggests that to attack ViT, (almost) all frequency spectrum needs to be affected, while it is less so for BiT and ResNet-50.

4.6 ViT Has Smoother Loss Landscape to Input Perturbations

One way to attribute the improved robustness of ViT over BiT is to hypothesize ViT has a smoother loss landscape with respect to input perturbations, contributing to enhanced robustness for classification. Here we explore the loss landscapes of ViT and BiT based on a common set of 100 ImageNet-1k validation images that are correctly classified by both models. We apply the multi-step projected gradient descent (PGD) attack [56] with an ℓ_∞ perturbation budget of $\epsilon = 0.002$ when normalizing the pixel value range to be between $[-1, 1]$ on this common set[10] (refer to Appendix G for details on hyperparameters). Figure 10 shows that the classification loss (cross entropy) of ViT increases at a much slower rate than that of BiT as one varies the attack steps, validating our hypothesis of smoother loss landscape to input perturbations.

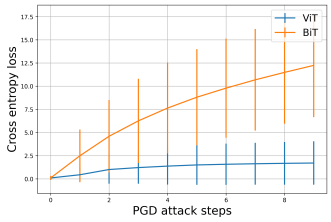

Figure 10: Loss progression (mean and standard deviation) ViT (L-16) and BiT-m (`r101x3`) during PGD attacks [56].

In summary, in this section, we broadly verify that ViT, primarily due to attention, can yield improved robustness (even with fewer parameters in some cases). This indicates that the use of Transformers can be orthogonal to the known techniques to improve the robustness of vision models [57, 58, 38].

5 Conclusion and Future Works

Robustness is an important aspect to consider when deploying deep learning models into the wild. This work provides a comprehensive robustness performance assessment of ViTs using 6 different ImageNet datasets and concludes that ViT significantly outperforms its CNN counterpart (BiT) and the baseline ResNet50V2 model. We further conducted 6 new experiments to verify our hypotheses of improved robustness in ViT, including the use of large-scale pre-training and attention module, the ability to recognize randomly masked images, the low sensibility to Fourier spectrum domain perturbation, and the property of wider energy distribution and smoother loss landscape under adversarial input perturbations. Our analyses and findings show novel insights toward understanding the source of robustness and can shed new light on robust neural network architecture design. Future works can build on top of the findings of our work and develop specific methods to probe the representations learned by vision transformers and other architectures that make use of self-attention.

[8] For computational constraints we used smaller BiT and ViT variants for this experiment.

[9] Rest of the hyperparameters are same as what is specified here.

[10] We follow the PGD implementation from here.

Acknowledgements

We are thankful to the Google Developers Experts program[11] (specifically Soonson Kwon and Karl Weinmeister) for providing Google Cloud Platform credits to support the experiments. We also thank Justin Gilmer (of Google), Guillermo Ortiz-Jimenez (of EPFL, Switzerland), and Dan Hendrycks (of UC Berkeley) for fruitful discussions.

References

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

[2] A. Kolesnikov, L. Beyer, X. Zhai, J. Puigcerver, J. Yung, S. Gelly, and N. Houlsby, "Big transfer (bit): General visual representation learning," in *European Conference on Computer Vision*, 2020, pp. 491–507.

[3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[4] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *International Conference on Learning Representations*, 2015.

[5] T. H. Trinh, M.-T. Luong, and Q. V. Le, "Selfie: Self-supervised pretraining for image embedding," *arXiv preprint arXiv:1906.02940*, 2019.

[6] M. Chen, A. Radford, R. Child, J. Wu, H. Jun, D. Luan, and I. Sutskever, "Generative pretraining from pixels," in *International Conference on Machine Learning*, vol. 119, Jul 2020, pp. 1691–1703.

[7] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European Conference on Computer Vision*. Springer, 2020, pp. 213–229.

[8] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," *arXiv preprint arXiv:2012.12877*, 2020.

[9] D. Hendrycks, K. Zhao, S. Basart, J. Steinhardt, and D. Song, "Natural adversarial examples," *Conference on Computer Vision and Pattern Recognition*, 2021.

[10] S. Bhojanapalli, A. Chakrabarti, D. Glasner, D. Li, T. Unterthiner, and A. Veit, "Understanding robustness of transformers for image classification," *arXiv preprint arXiv:2103.14586*, 2021.

[11] R. Shao, Z. Shi, J. Yi, P.-Y. Chen, and C.-J. Hsieh, "On the adversarial robustness of visual transformers," *arXiv preprint arXiv:2103.15670*, 2021.

[12] K. Mahmood, R. Mahmood, and M. Van Dijk, "On the robustness of vision transformers to adversarial examples," *arXiv preprint arXiv:2104.02610*, 2021.

[13] D. Hendrycks and T. G. Dietterich, "Benchmarking neural network robustness to common corruptions and perturbations," in *International Conference on Learning Representations*, 2019.

[14] D. Hendrycks, S. Basart, N. Mu, S. Kadavath, F. Wang, E. Dorundo, R. Desai, T. Zhu, S. Parajuli, M. Guo, D. Song, J. Steinhardt, and J. Gilmer, "The many faces of robustness: A critical analysis of out-of-distribution generalization," *arXiv preprint arXiv:2006.16241*, 2020.

[15] K. Xiao, L. Engstrom, A. Ilyas, and A. Madry, "Noise or signal: The role of image backgrounds in object recognition," *International Conference on Learning Representations*, 2021.

[11] Google Developers Experts

[16] N. Parmar, A. Vaswani, J. Uszkoreit, L. Kaiser, N. Shazeer, A. Ku, and D. Tran, "Image transformer," in *International Conference on Machine Learning*, vol. 80, Jul 2018, pp. 4055–4064.

[17] W. Shi, J. Caballero, F. Huszár, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1874–1883.

[18] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, "Photo-realistic single image super-resolution using a generative adversarial network," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 105–114.

[19] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. C. Loy, "ESRGAN: Enhanced super-resolution generative adversarial networks," in *European Conference on Computer Vision Workshops*, L. Leal-Taixé and S. Roth, Eds. Springer, 2019, pp. 63–79.

[20] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *IEEE Conference on Computer Vision and Pattern Recognition,*, 2009, pp. 248–255.

[21] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," in *NeurIPS Deep Learning and Representation Learning Workshop*, 2015.

[22] I. Radosavovic, R. Kosaraju, R. Girshick, K. He, and P. Dollar, "Designing network design spaces," in *IEEE Conference on Computer Vision and Pattern Recognition*, jun 2020, pp. 10 425–10 433.

[23] L. Yuan, Y. Chen, T. Wang, W. Yu, Y. Shi, F. E. Tay, J. Feng, and S. Yan, "Tokens-to-token vit: Training vision transformers from scratch on imagenet," *arXiv preprint arXiv:2101.11986*, 2021.

[24] S. Zagoruyko and N. Komodakis, "Wide residual networks," *arXiv preprint arXiv:1605.07146*, 2016.

[25] Z. Jiang, Q. Hou, L. Yuan, D. Zhou, X. Jin, A. Wang, and J. Feng, "Token labeling: Training a 85.5% top-1 accuracy vision transformer with 56m parameters on imagenet," *arXiv preprint arXiv:2104.10858*, 2021.

[26] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[27] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *arXiv preprint arXiv:1607.06450*, 2016.

[28] D. Hendrycks and K. Gimpel, "Gaussian error linear units (gelus)," *arXiv preprint arXiv:1606.08415*, 2016.

[29] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Jun. 2019, pp. 4171–4186.

[30] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *European Conference on Computer Vision*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Springer, 2016, pp. 630–645.

[31] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[32] Y. Wu and K. He, "Group normalization," in *European Conference on Computer Vision*, September 2018, pp. 3–19.

[33] S. Qiao, H. Wang, C. Liu, W. Shen, and A. Yuille, "Micro-batch training with batch-channel normalization and weight standardization," *arXiv preprint arXiv:1903.10520*, 2019.

[34] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International Conference on Machine Learning*, vol. 37, Jul 2015, pp. 448–456.

[35] C. Sun, A. Shrivastava, S. Singh, and A. Gupta, "Revisiting unreasonable effectiveness of data in deep learning era," in *IEEE International Conference on Computer Vision*, 2017, pp. 843–852.

[36] B. M. 't Hart, H. C. E. F. Schmidt, I. Klein-Harmeyer, and W. Einhäuser, "Attention in natural scenes: contrast affects rapid visual processing and fixations alike," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 368, no. 1628, p. 20130067, 2013.

[37] D. Hendrycks*, N. Mu*, E. D. Cubuk, B. Zoph, J. Gilmer, and B. Lakshminarayanan, "Augmix: A simple method to improve robustness and uncertainty under data shift," in *International Conference on Learning Representations*, 2020.

[38] Q. Xie, M.-T. Luong, E. Hovy, and Q. V. Le, "Self-training with noisy student improves imagenet classification," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10 684–10 695.

[39] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, vol. 25, 2012.

[40] E. D. Cubuk, B. Zoph, J. Shlens, and Q. V. Le, "Randaugment: Practical automated data augmentation with a reduced search space," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 3008–3017.

[41] G. Huang, Y. Sun, Z. Liu, D. Sedra, and K. Q. Weinberger, "Deep networks with stochastic depth," in *European conference on computer vision*. Springer, 2016, pp. 646–661.

[42] R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F. A. Wichmann, and W. Brendel, "Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness," in *International Conference on Learning Representations*, 2019.

[43] D. Mahajan, R. Girshick, V. Ramanathan, K. He, M. Paluri, Y. Li, A. Bharambe, and L. Van Der Maaten, "Exploring the limits of weakly supervised pretraining," in *European Conference on Computer Vision*, 2018, pp. 181–196.

[44] T. Chen, S. Kornblith, K. Swersky, M. Norouzi, and G. E. Hinton, "Big self-supervised models are strong semi-supervised learners," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 22 243–22 255.

[45] S. Abnar and W. Zuidema, "Quantifying attention flow in transformers," in *Annual Meeting of the Association for Computational Linguistics*, Jul. 2020, pp. 4190–4197.

[46] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *IEEE International Conference on Computer Vision*, 2017, pp. 618–626.

[47] J. Gildenblat and contributors, "pytorch-cam," https://github.com/jacobgil/pytorch-grad-cam, 2021.

[48] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," *arXiv preprint arXiv:2104.14294*, 2021.

[49] T. DeVries and G. W. Taylor, "Improved regularization of convolutional neural networks with cutout," *arXiv preprint arXiv:1708.04552*, 2017.

[50] D. Yin, R. Gontijo Lopes, J. Shlens, E. D. Cubuk, and J. Gilmer, "A Fourier perspective on model robustness in computer vision," in *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[51] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "Deepfool: a simple and accurate method to fool deep neural networks," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2574–2582.

[52] G. R. Lanckriet, L. E. Ghaoui, C. Bhattacharyya, and M. I. Jordan, "A robust minimax approach to classification," *Journal of Machine Learning Research*, vol. 3, no. null, p. 555–582, Mar. 2003.

[53] A. Globerson and S. Roweis, "Nightmare at test time: robust learning by feature deletion," in *International Conference on Machine learning*, 2006, pp. 353–360.

[54] J. Jo and Y. Bengio, "Measuring the tendency of cnns to learn surface statistical regularities," *arXiv preprint arXiv:1711.11561*, 2017.

[55] G. Ortiz-Jimenez, A. Modas, S.-M. Moosavi, and P. Frossard, "Hold me tight! influence of discriminative features on deep network boundaries," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 2935–2946.

[56] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," *International Conference on Learning Representations*, 2018.

[57] Y. Balaji, T. Goldstein, and J. Hoffman, "Instance adaptive adversarial training: Improved accuracy tradeoffs in neural nets," *arXiv preprint arXiv:1910.08051*, 2019.

[58] Y. Carmon, A. Raghunathan, L. Schmidt, J. C. Duchi, and P. S. Liang, "Unlabeled data improves adversarial robustness," in *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[59] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *International Conference on Learning Representations*, 2015.

Appendix

A Image Patches

In ViT, the input images are divided into small patches as depicted in Figure 11. Here, the original image is of 224×224 shape and each patch is of 16×16 shape. This gives us a total of 196 patches. Since we are dealing with RGB images here, we also need to consider the channel dimension. So, in total, these 16×16 patches (with 3 channels) are flattened into a dimension of 768 ($16 \times 16 \times 3$) using a linear transformation. The spatial information of the patches gets lost due to this and to mitigate that position encoding is used. For visual depictions of how the patch encodings operate with each other, we refer the reader to the Figure 7 of [1].

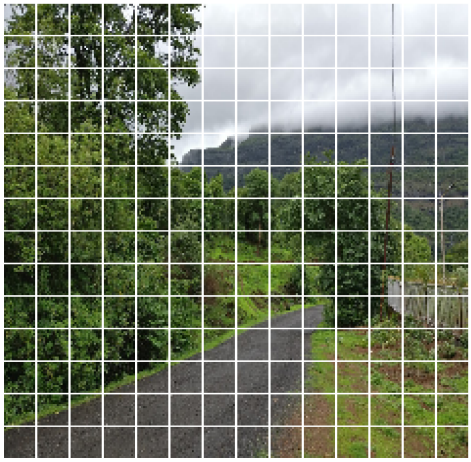

Figure 11: A sample image divided into patches.

B Additional Examples with Grad-CAM

To make our arguments in Section 4.2 more concrete, we provide additional results from Grad-CAM in Figure B. All the original images are from ImageNet-1k validation set.

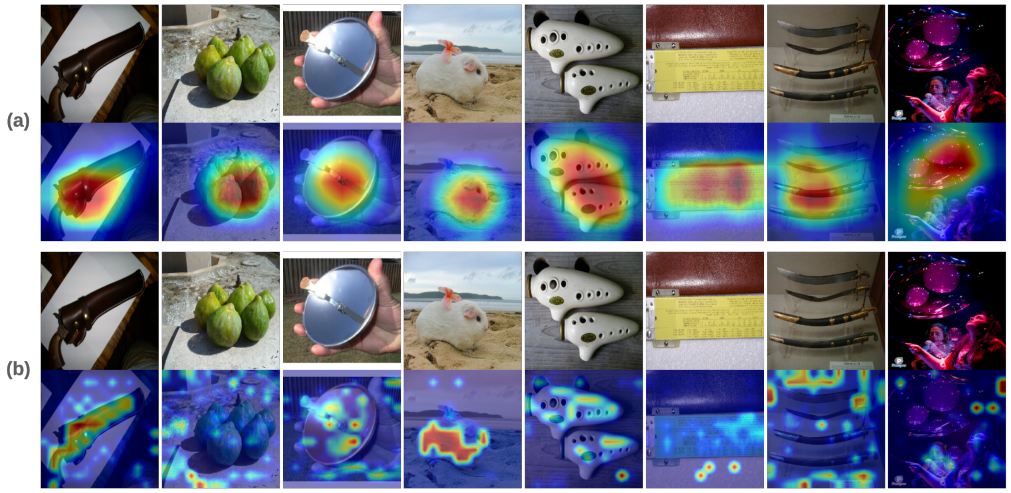

Figure 12: Additional Grad-CAM results. Predictions are (left-to-right) "holster", "fig", "solar_dish", "guinea_pig", "ocarina", "slide_rule", "scabbard", and "bubble". All the predictions are correct. (**a**) BiT (`m-r101x3`). (**b**) ViT L-16.

C Additional Results on ImageNet-C

Scores on individual corruptions of ImageNet-C. In Table 10, we provide the individual top-1 accuracy scores for the 15 different corruption types of ImageNet-C. Note that for this we only consider the severity level of 5. As mentioned in Section 3.2, ViT particularly performs poorly on the "contrast" corruption.

Table 10: Individual top-1 accuracy scores (%) on all the corruption types of ImageNet-C.

	Noise			**Blur**				**Weather**				**Digital**			
Model	Gauss	Shot	Impulse	Defocus	Glass	Motion	Zoom	Snow	Frost	Fog	Bright	Contrast	Elastic	Pixelate	JPEG
BiT-m r50x1	26.25	26.81	27.65	21.7	12.44	25.94	26.32	28.24	36.38	27.77	58.94	16.88	19.08	49.22	49.77
BiT-m r50x3	35.96	36.69	37.62	31.56	16.7	30.42	31.93	30.65	43.02	31.89	64.35	21.42	22.39	53.83	57.43
BiT-m r101x1	31.6	32.58	33.16	26.34	16.59	29	29.82	28.53	37.56	28.03	60.43	16.07	21.39	48.79	52.91
BiT-m r101x3	36.25	36.48	38.18	33.8	16.25	38.52	36.9	36.51	43.81	26.64	64.56	17.34	27.48	56.88	56.08
BiT-m r152x4	35.08	35.38	37.44	34.66	18.66	35.75	39.82	33.98	43.42	31.03	63.62	24.38	27.13	56.66	55.35
ViT B-16	23.27	26.32	25.43	31.65	28.8	43.36	39.85	46.1	45.99	42.05	71.96	12.93	41.91	60.77	57.52
ViT B-32	23.45	25.43	25.39	31.71	28.97	38.39	33.67	34.83	42.52	33.44	68	9.31	44.71	60.8	54.79
ViT L-16	39.1	39.95	42.1	36.71	36.3	49.69	47.65	52.82	51.53	47.99	74.53	18.85	49.87	70.05	63.45
ViT L-32	33.56	35.06	35.07	36.06	33.86	44.5	40.73	39.53	45.54	38.28	68.93	8.76	49.94	62.17	59.31

Next, in Table 11, we report the individual unnormalized corruption errors (not scaled using the AlexNet errors) on the same 15 different corruptions as given by BiT-m `r101x3` and ViT L-16.

Table 11: Individual unnormalized corruption errors (%) on 15 different corruption types of ImageNet-C.

	Noise			**Blur**				**Weather**				**Digital**			
Model	Gauss	Shot	Impulse	Defocus	Glass	Motion	Zoom	Snow	Frost	Fog	Bright	Contrast	Elastic	Pixelate	JPEG
BiT-m `r101x3`	44.45	45.88	46.7	48.18	64.34	44.7	50.03	51.18	46.3	51.48	28.48	45.49	49.3	34.31	36.53
ViT L-16	35.05	35.98	35.88	41.78	42.45	34.59	40.73	40.47	40.42	39.61	22.64	43.57	32.97	24.31	27.79

Stability under common corruptions. As reported in Section 3.2, ViT significantly outperforms BiT when exposed to common corruptions. To better understand if ViT is able to hold its attention-span under those corruptions, in Figure 13, we provide Grad-CAM results for a few images from ImageNet-C sampled from different levels of severity.

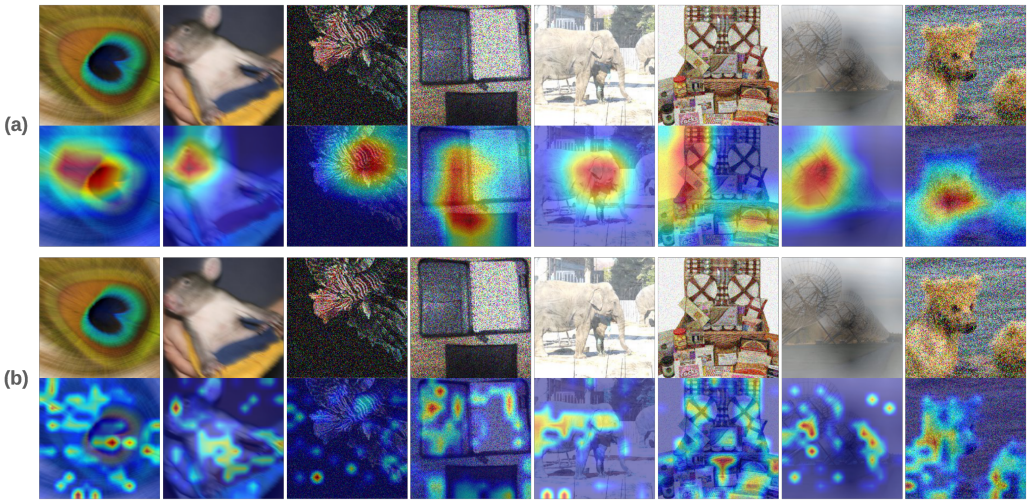

Figure 13: Grad-CAM results on ImageNet-C. For a fair comparison, we only consider the case where BiT makes wrong predictions whereas ViT predictions are still correct. The predictions (with confidence scores) are (left-to-right): "quill" (98.02%), "wombat" (92.54%), "lionfish" (99.55%), "binder" (80.29%), "Indian_elephant" (59.60%), "hamper" (99.82%), "radio_telescope" (97.93%), and "brown_bear" (79.02%). (**a**) BiT (`m-r101x3`). (**b**) ViT L-16.

While the Grad-CAM results for BiT may seem more appealing but the slight corruptions are enough to make its predictions flip. However, that is not the case for ViT. By utilizing more global context it is able to perform with strong confidence even under these corruptions.

D Additional Results on ImageNet-P

In Table 12 and Table 13, we report the flip rates and top-5 distances of the individual perturbation types of ImageNet-P respectively with BiT and ViT. These scores are unnormalized meaning that they were not scaled using corresponding AlexNet scores.

Table 12: Individual unnormalized flip rates (%) on all the perturbation types of ImageNet-P.

	Noise		Blur		Weather		Digital			
Model	Gauss	Shot	Motion	Zoom	Snow	Bright	Translate	Rotate	Tilt	Scale
BiT-m r101x3	11.932	13.686	4.594	3.665	5.523	3.23	4.315	6.121	3.893	9.273
ViT L-16	7.363	8.263	2.388	1.945	2.969	2.031	3.63	5.085	2.788	8.434

Table 13: Individual unnormalized top-5 distances (%) on all the perturbation types of ImageNet-P.

	Noise		**Blur**		**Weather**		**Digital**			
Model	Gauss	Shot	Motion	Zoom	Snow	Bright	Translate	Rotate	Tilt	Scale
BiT-m r101x3	3.8449	4.2462	1.44212	1.20941	1.75846	1.1756	1.59866	2.1012	1.45261	2.75042
ViT L-16	2.49186	2.68482	0.7974	0.63717	0.92836	0.69184	1.27774	1.67676	1.0207	2.39877

E Random Masking with Cutout

In Figure 14, we show how the predictions of BiT (`m-r101x3`) and ViT (L-16) change as a function of the masking factor in Cutout [49]. We provide these results for giving a clearer sense of the experiments conducted in Section 4.3. These results should not be treated as anything conclusive.

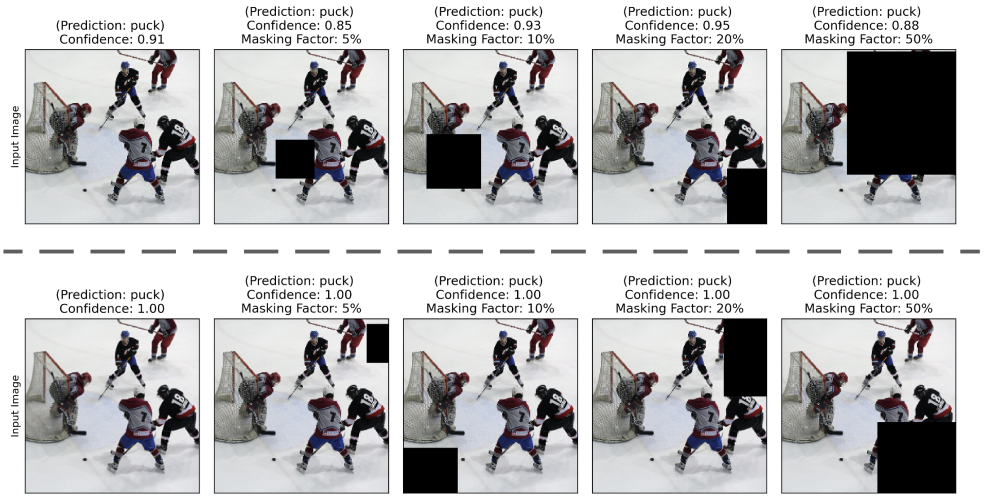

Figure 14: Change of prediction confidence as a function of masking. First row is from BiT-m `r101x3` and second row is from ViT L-16.

F Magnitude Spectrum and High-Frequency Components

Since we use Fourier analysis in Section 4.4, in the interest of comprehensiveness, we provide visualizations of the magnitude spectrum of frequency components as well as the raw high-frequency components of natural images in Figure 15.

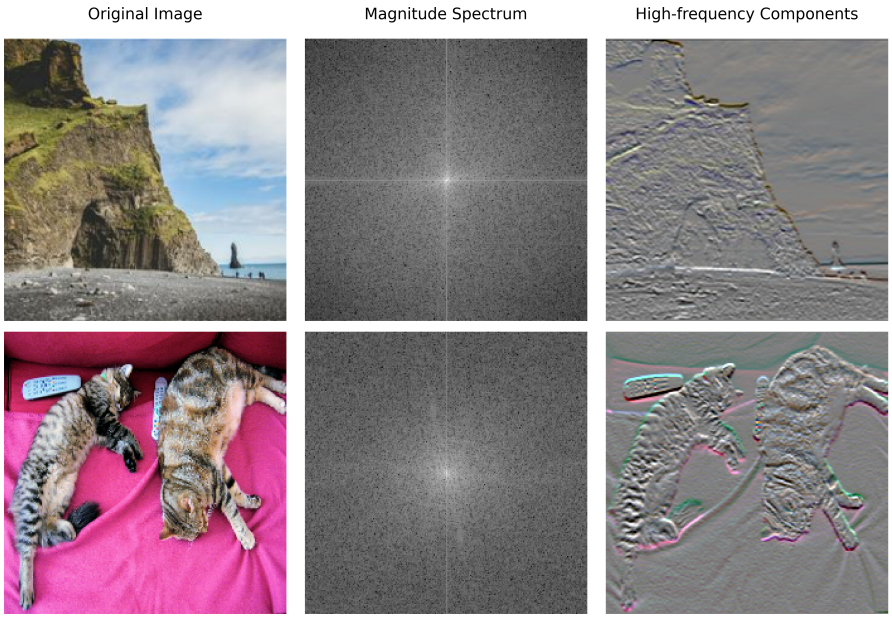

Figure 15: Visualization of the magnitude spectrum in the frequency domain and raw high-frequency components.

G Adversarial Perturbations

Peaking into the adversarial perturbations. In Figure 16, we visualize the perturbations as learned by BiT-m `r101x3` and ViT L-16. We use Adam [59] as optimizer here to implement PGD attack (see Section 4.6 for implementation details) with a learning rate of 1e-3. Generally, we find that the perturbations are smoother in case of ViT.

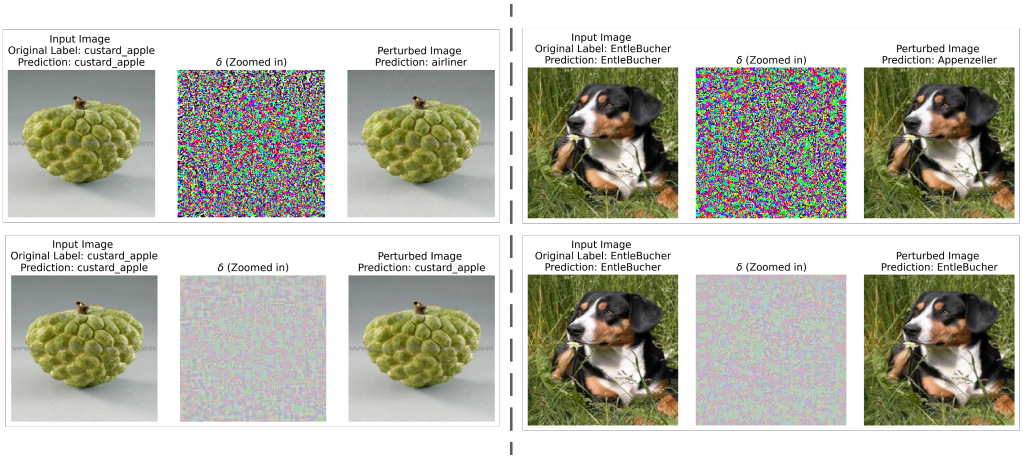

Figure 16: Visualization of the adversarial perturbations. First row is from BiT-m `r101x3` and second row is from ViT L-16.

Loss landscape of individual examples. In Figure 17, we show PGD loss plots from five individual ImageNet-1k validation images used in Section 4.6. These examples are not cherry-picked and provided to better isolate the results shown in Figure 10.

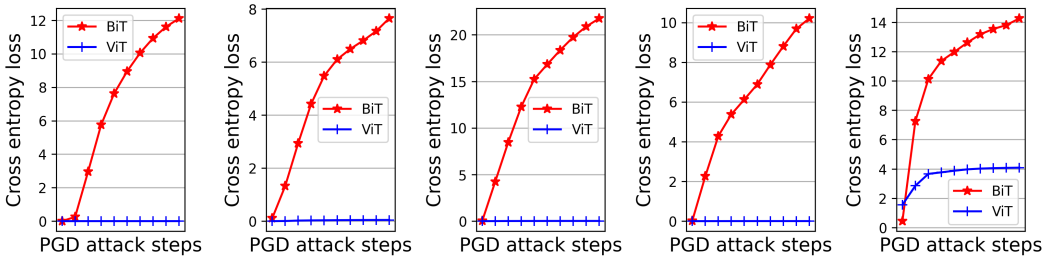

Figure 17: PGD loss plots of individual examples from the ImageNet-1k validation set.