Are Transformers More Robust Than CNNs?

Yutong Bai[1] **Jieru Mei**[1] **Alan Yuille**[1] **Cihang Xie**[2]
[1]Johns Hopkins University [2] University of California, Santa Cruz
{ytongbai, meijieru, alan.l.yuille, cihangxie306}@gmail.com

Abstract

Transformer emerges as a powerful tool for visual recognition. In addition to demonstrating competitive performance on a broad range of visual benchmarks, recent works also argue that Transformers are much more robust than Convolutions Neural Networks (CNNs). Nonetheless, surprisingly, we find these conclusions are drawn from unfair experimental settings, where Transformers and CNNs are compared at different scales and are applied with distinct training frameworks. In this paper, we aim to provide the first **fair & in-depth** comparisons between Transformers and CNNs, focusing on robustness evaluations.

With our unified training setup, we first challenge the previous belief that Transformers outshine CNNs when measuring *adversarial robustness*. More surprisingly, we find CNNs can easily be as robust as Transformers on defending against adversarial attacks, if they properly adopt Transformers' training recipes. While regarding *generalization on out-of-distribution samples*, we show pre-training on (external) large-scale datasets is not a fundamental request for enabling Transformers to achieve better performance than CNNs. Moreover, our ablations suggest such stronger generalization is largely benefited by the Transformer's self-attention-like architectures per se, rather than by other training setups. We hope this work can help the community better understand and benchmark the robustness of Transformers and CNNs. The code and models are publicly available at `https://github.com/ytongbai/ViTs-vs-CNNs`.

1 Introduction

Convolutional Neural Networks (CNNs) have been the widely-used architecture for visual recognition in recent years [21, 35, 37, 15, 20]. It is commonly believed the key to such success is the usage of the convolutional operation, as it introduces several useful inductive biases (*e.g.*, translation equivalence) to models for benefiting object recognition. Interestingly, recent works alternatively suggest that it is also possible to build successful recognition models without convolutions [31, 55, 2]. The most representative work in this direction is Vision Transformer (ViT) [11], which applies the pure self-attention-based architecture to sequences of images patches and attains competitive performance on the challenging ImageNet classification task [32] compared to CNNs. Later works [25, 44] further expand Transformers with compelling performance on other visual benchmarks, including COCO detection and instance segmentation [22], ADE20K semantic segmentation [56].

The dominion of CNNs on visual recognition is further challenged by the recent findings that Transformers appear to be *much more robust* than CNNs. For example, Shao *et al*. [34] observe that the usage of convolutions may introduce a negative effect on models' adversarial robustness, while migrating to Transformer-like architectures (*e.g.*, the Conv-Transformer hybrid model or the pure Transformer) can help secure models' adversarial robustness. Similarly, Bhojanapalli *et al*. [3] report that, if pre-trained on sufficiently large datasets, Transformers exhibit considerably stronger robustness than CNNs on a spectrum of out-of-distribution tests (*e.g.*, common image corruptions [16], texture-shape cue conflicting stimuli [12]).

35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia.

Though both [3] and [34] claim that Transformers are preferable to CNNs in terms of robustness, we find that such conclusion cannot be strongly drawn based on their existing experiments. Firstly, Transformers and CNNs are not compared at the same model scale, *e.g.*, a small CNN, ResNet-50 (∼25 million parameters), by default is compared to a much larger Transformer, ViT-B (∼86 million parameters), for these robustness evaluations. Secondly, the training frameworks applied to Transformers and CNNs are distinct from each other (*e.g.*, training datasets, number of epochs, and augmentation strategies are all different), while little efforts are devoted on ablating the corresponding effects. In a nutshell, due to these inconsistent and unfair experiment settings, *it remains an open question whether Transformers are truly more robust than CNNs*.

To answer it, in this paper, we aim to provide the first benchmark to fairly compare Transformers to CNNs in robustness evaluations. We particularly focus on the comparisons between Small Data-efficient image Transformer (DeiT-S) [40] and ResNet-50 [15], as they have similar model capacity (*i.e.*, ∼22 million parameters *vs*. ∼25 million parameters) and achieve similar performance on ImageNet (*i.e.*, 76.8% top-1 accuracy *vs*. 76.3% top-1 accuracy[1]). Our evaluation suite accesses model robustness in two ways: 1) adversarial robustness, where the attackers can actively and aggressively manipulate inputs to approximate the worst-case scenario; 2) generalization on out-of-distribution samples, including common image corruptions (ImageNet-C [16]), texture-shape cue conflicting stimuli (Stylized-ImageNet [12]) and natural adversarial examples (ImageNet-A [17]).

With this unified training setup, we present a completely different picture from previous ones [34, 3]. Regarding adversarial robustness, we find that Transformers actually are no more robust than CNNs—if CNNs are allowed to properly adopt Transformers' training recipes, these two types of models will attain similar robustness on defending against both perturbation-based adversarial attacks and patch-based adversarial attacks. While for generalization on out-of-distribution samples, we find Transformers can still substantially outperform CNNs even without the needs of pre-training on sufficiently large (external) datasets. Additionally, our ablations show that adopting Transformer's self-attention-like architecture is the key for achieving strong robustness on these out-of-distribution samples, while tuning other training setups will only yield subtle effects here. We hope this work can serve as a useful benchmark for future explorations on robustness, using different network architectures, like CNNs, Transformers, and beyond [39, 23].

2 Related Works

Vision Transformer. Transformers, invented by Vaswani *et al*. in 2017 [43], have largely advanced the field of natural language processing (NLP). With the introduction of self-attention module, Transformer can effectively capture the non-local relationships between all input sequence elements, achieving the state-of-the-art performance on numerous NLP tasks [50, 9, 4, 10, 28, 29].

The success of Transformer on NLP also starts to get witnessed in computer vision. The pioneering work, ViT [11], demonstrates that the pure Transformer architectures are able to achieve exciting results on several visual benchmarks, especially when extremely large datasets (*e.g.*, JFT-300M [36]) are available for pre-training. This work is then subsequently improved by carefully curating the training pipeline and the distillation strategy to Transformers [40], enhancing the Transformers' tokenization module [51], building multi-resolution feature maps on Transformers [25, 44], designing parameter-efficient Transformers for scaling [53, 42, 48], *etc*. In this work, rather than focusing on furthering Transformers on standard visual benchmarks, we aim to provide a fair and comprehensive study of their performance when testing out of the box.

Robustness evaluations. Conventional learning paradigm assumes training data and testing data are drawn from the same distribution. This assumption generally does not hold, especially in the real-world case where the underlying distribution is too complicated to be covered in a (limited-sized) dataset. To properly access model performance in the wild, a set of robustness generalization benchmarks have been built, *e.g.*, ImageNet-C [16], Stylized-ImageNet [12], ImageNet-A [17], *etc*. Another standard surrogate for testing model robustness is via adversarial attacks, where the attackers deliberately add small perturbations or patches to input images, for approximating the worst-case evaluation scenario [38, 13]. In this work, both robustness generalization and adversarial robustness are considered in our robustness evaluation suite.

[1]Here we apply the general setup in [41] for the ImageNet training. We follow the popular ResNet's standard to train both models for 100 epochs. Please refer to Section 3 for more training details.

Concurrent to ours, both Bhojanapalli *et al*. [3] and Shao *et al*. [34] conduct the robustness comparisons between Transformers and CNNs. Nonetheless, we find their experimental settings are unfair, *e.g*., models are compared at different capacity [3, 34] or are trained under distinct frameworks [34]. In this work, our comparison carefully align the model capacity and the training setups, which draws completely different conclusions from the previous ones.

3 Settings

3.1 Training CNNs and Transformers

Convolutional Neural Networks. ResNet [15] is a milestone architecture in the history of CNN. We choose its most popular instantiation, *ResNet-50* (with ∼25 million parameters), as the default CNN architecture. To train CNNs on ImageNet, we follow the standard recipe of [14, 30]. Specifically, we train all CNNs for a total of 100 epochs, using momentum-SGD optimizer; we set the initial learning rate to 0.1, and decrease the learning rate by 10× at the 30-th, 60-th, and 90-th epoch; no regularization except weight decay is applied.

Vision Transformer. ViT [11] successfully introduces Transformers from natural language processing to computer vision, achieving excellent performance on several visual benchmarks compared to CNNs. In this paper, we follow the training recipe of DeiT [40], which successfully trains ViT on ImageNet without any external data, and set *DeiT-S* (with ∼22 million parameters) as the default Transformer architecture. Specifically, we train all Transformers using AdamW optimizer [26]; we set the initial learning rate to 5e-4, and apply the cosine learning rate scheduler to decrease it; besides weight decay, we additionally adopt three data augmentation strategies (*i.e*., RandAug [8], MixUp [54] and CutMix [52]) to regularize training (otherwise DeiT-S will attain significantly lower ImageNet accuracy due to overfitting [5]).

Note that different from the standard recipe of DeiT (which applies 300 training epochs by default), we hereby train Transformers only for a total of 100 epochs, *i.e*., same as the setup in ResNet. We also remove {Erasing, Stochastic Depth, Repeated Augmentation}, which were applied in the original DeiT framework, in this basic 100 epoch schedule, for preventing over-regularization in training. Such trained DeiT-S yields 76.8% top-1 ImageNet accuracy, which is similar to the ResNet-50's performance (76.3% top-1 ImageNet accuracy).

3.2 Robustness Evaluations

Our experiments mainly consider two types of robustness here, *i.e*., robustness on adversarial examples and robustness on out-of-distribution samples.

Adversarial examples, which are crafted by adding human-imperceptible perturbations or small-sized patches to images, can lead deep neural networks to make wrong predictions. In addition to the very popular PGD attack [27], our robustness evaluation suite also contains: A) AutoAttack [7], which is an ensemble of diverse attacks (*i.e*., two variants of PGD attack, FAB attack [6] and Square Attack [1]) and is parameter-free; and B) Texture Patch Attack (TPA) [49], which uses a predefined texture dictionary of patches to fool deep neural networks.

Recently, several benchmarks of **out-of-distribution samples** have been proposed to evaluate how deep neural networks perform when testing out of the box. Particularly, our robustness evaluation suite contains three such benchmarks: A) ImageNet-A [18], which are real-world images but from some challenging scenarios (*e.g*., occlusion, fog scene) for recognition; B) ImageNet-C [16], which is designed for measuring model robustness against 75 distinct common image corruptions; and C) Stylized-ImageNet [12], which creates texture-shape cue conflicting stimuli by removing local texture cues from images while retaining their global shape information.

4 Adversarial Robustness

In this section, we investigate the robustness of Transformers and CNNs on defending against adversarial attacks, using ImageNet validation set (with 50,000 images). We consider both perturbation-based attacks (*i.e*., PGD and AutoAttack) and patch-based attacks (*i.e*., TPA) for robustness evaluations.

4.1 Robustness to Perturbation-Based Attacks

Following [34], we first report the robustness of ResNet-50 and DeiT-S on defending against AutoAttack. We verify that, when applying with a small perturbation radius $\epsilon = 0.001$, DeiT-S indeed achieves higher robustness than ResNet-50, *i.e.*, 22.1% *vs*. 17.8% as shown in Table 1.

However, when increasing the perturbation radius to 4/255, a more challenging but standard case studied in previous works [33, 45, 46], *both models will be circumvented completely*, *i.e.*, 0% robustness on defending against AutoAttack. This is mainly due to that both models are not adversarially trained [13, 27], which is an effective way to secure model robustness against adversarial attacks, and we will study it next.

Table 1: Top-1 accuracy of ResNet-50 and DeiT-S on defending against AutoAttack, using ImageNet validation set. We note both models will broken completely when setting perturbation radius to 4/255.

	Clean	Perturbation Radius	
		0.001	4/255
ResNet-50	76.9	17.8	0.0
DeiT-S	76.8	22.1	0.0

4.1.1 Adversarial Training

Adversarial training [13, 27], which trains models with adversarial examples that generated on-the-fly, aims to optimize the following min-max framework:

$$\arg\min_{\theta} \mathbb{E}_{(x,y)\sim\mathbb{D}}\Big[\max_{\epsilon\in\mathbb{S}} L(\theta, x+\epsilon, y)\Big], \tag{1}$$

where $\mathbb{D}$ is the underlying data distribution, $L(\cdot,\cdot,\cdot)$ is the loss function, θ is the network parameter, x is a training sample with the ground-truth label y, ϵ is the added adversarial perturbation, and $\mathbb{S}$ is the allowed perturbation range. Following [47, 45], the adversarial training here applies *single-step PGD* (PGD-1) to generate adversarial examples (for lowering training cost), with the constrain that maximum per-pixel change $\epsilon = 4/255$.

Adversarial Training on Transformers. We apply the setup above to adversarially train both ResNet-50 and DeiT-S. However, surprisingly, this default setup works for ResNet-50 but will collapse the training with DeiT-S, *i.e.*, the robustness of such trained DeiT-S is merely ∼4% when evaluating against PGD-5. We identify the issue is over-regularization—when combining strong data augmentation strategies (*i.e.*, RangAug, Mixup and CutMix used in DeiT-S) with adversarial attacks, the yielded training samples are too hard to be learnt.

(a) Epoch = 0 (b) Epoch = 4 (c) Epoch = 9

Figure 1: The illustration of the proposed augmentation warm-up strategy. At the beginning of adversarial training (from epoch=0 to epoch=9), we progressively increase the augmentation strength.

To ease this observed training difficulty, we design a curricular of the applied augmentation strategies. Specifically, as shown in Figure 1, at the first 10 epoch, we progressively enhance the augmentation strength (*e.g.*, gradually changing the distortion magnitudes in RandAug from 1 to 9) to warm-up the training process. Our experiment verify it enables a successful adversarial training with DeiT-S—such trained DeiT-S now attains ∼44% robustness (boosted from ∼4%) on defending against PGD-5.

Transformers with CNNs' Training Recipe. Interestingly, an alternative way to address the observed training difficulty is directly adopting CNN's recipe to train Transformers [34], *i.e.*, applying SGD with step learning rate schedule and using no strong data augmentation strategies (like Mixup). Though this setup can stabilize the adversarial training process, it significantly hurts the overall performance of DeiT-S—the clean accuracy drops to 59.9% (**-6.6%**), and the robustness on defending against PGD-100 drops to 31.9% (**-8.4%**).

One reason for this degenerated performance is that strong data augmentation strategies are not included in CNNs' recipe, therefore Transformers will be easily overfitted during training [5]. Another key factor here is the incompatibility between the SGD optimizer and Transformers. As explained in [24], compared to SGD, adaptive optimizers (like AdamW) are capable of assigning different learning rates to different parameters, resulting in consistent update magnitudes even with unbalanced gradients. This property is crucial for enabling successful training of Transformers, given the gradients of attention modules are highly unbalanced.

CNNs with Transformers' Training Recipe. We now explore the opposite choice—adopting Transformers' recipe to train CNNs. Interestingly, if we directly apply AdamW to ResNet-50, the adversarial training process becomes very unstable, sometime even collapses completely. We also explore the possibility of adversarially train ResNet-50 with strong data augmentation strategies (*i.e.*, RandAug, Mixup and CutMix). However, we find ResNet-50 will be overly regularized in adversarial training, leading to significantly lower clean accuracy and PGD robustness.

Though Transformers' optimizer and augmentation strategies cannot improve CNNs' adversarial training, we find Transformers' activation function matters. According to the experiments in [46], activation function plays an important role in adversarial training. DeiT uses GELU as an activation function, while ResNet uses ReLU. In Table 2, experiments show that after properly applying GELU as activation function to ResNet-50, ResNet's adversarial training can lead to both higher clean accuracy and robustness compared with same settings with ReLU.

As shown in Table 2, after adversarial training, ResNet-50 can outperform DeiT-S on ImageNet clean accuracy by 0.9% while have exactly the same robustness on defending against PGD attacks and AutoAttacks under the same 4/255 perturbation radius, challenging previous conclusions that Transformers are more robust than CNNs. We also re-evaluate their robustness to AutoAttack under radius 0.001 on adversarially trained model. After changing to GELU, ResNet-50's defense success rate increases from 62.2% to 65.7% (+2.5%), surpass DeiT-S's 64.8% robustness by 0.9%.

Table 2: After aligning ResNet with DeiT's activation function GELU, ResNet-50 shows better accuracy (%) on ImageNet and while maintains the same robustness under adversarial attacks.

	Activation	Clean Acc	PGD-5	PGD-10	PGD-50	PGD-100	AutoAttack-4/255
ResNet-50	ReLU	66.77	38.70	34.19	32.47	32.26	26.41
	GELU	67.38	44.01	40.98	40.28	40.27	35.51
DeiT-S	GELU	66.50	43.95	41.03	40.34	40.32	35.50

4.2 Robustness to Patch-Based Attacks

Since Autoattack is composed with one patch based attack, we further study the robustness with a more advanced patch attackers, named Texture Patch Attack (TPA) [49]. For TPA, we set the number of patches to 4 and the total area to 10% of the image area by default. With the number of patch increases, the difficulty of the TPA grows as well. We mainly evaluate non-targeted attacks of TPA.

Table 3 shows both clean accuracy on ImageNet and the defense success rate to TPA of ResNet-50 under different augmentation techniques used in DeiT. Compared to the original ResNet-50, all augmentation improve the robustness to patch attacks. But interestingly, we note observe that CutMix is extremely effective to improve the robustness of TPA, *i.e.*, CutMix only can subtantially improve model robustness against TPA by 29.4%.

The strongest performance for defending against TPA can be achieved by training with both CutMix and RandAug. In Table 4, we provide the comparison of the robustness to stronger patch attacks (*i.e.*, by increasing patch number from 4 to 10) between ResNet-50 and DeiT-S under this setting. We can observe that ResNet is more robust to different levels of TPA compared with DeiT.

Table 3: Defense success rate (%) of ResNet-50 with different augmentations against Patch Attacks. We also show clean accuracy on ImageNet(%).

Augmentations			ImageNet ↑	Patch Attack ↑
RandAug	MixUp	CutMix		
✗	✗	✗	78.9	19.7
✓	✓	✓	76.4	48.6 (+28.9)
✓	✓	✗	75.7	31.7 (+12.0)
✓	✗	✓	76.7	**52.4 (+32.7)**
✗	✓	✓	77.1	39.8 (+20.1)
✓	✗	✗	77.5	24.3 (+4.6)
✗	✓	✗	75.9	31.5 (+11.8)
✗	✗	✓	77.2	49.1 (+29.4)

Table 4: Defense success rate (%) of ResNet-50 and DeiT-S against Patch Attacks with different patch number. We also show clean accuracy(%).

Architecture	ImageNet↑	Patch Attack↑	
		P=4	P=10
ResNet-50	76.7	**52.4**	**34.5**
DeiT-S	76.8	47.7	28.9

5 Generalization on Out-of-distribution Samples

In this section we evaluate the robustness to out-of-distribution samples of DeiT to ResNet models. As aforementioned, the training frameworks applied to Transformers and CNNs varies in many aspects, *e.g.*, training epochs, optimizers, learning rate schedule, augmentations *etc*. To capture different aspects of robustness, we rely on different, specialized benchmarks including ImageNet-A, ImageNet-C, and Stylized ImageNet.

5.1 Aligned Setting Comparison

We first provide a direct comparison between DeiT and ResNet in their original setting. In addition, for a more fair comparison, we fully align to ResNet-50 DeiT's training strategies, whose details can be found in Section 3.1. We denote this model as ResNet-50*.

Experiments in Table 5 show that DeiT-S has better generalization on the out-of-distribution samples, compared with both ResNet-50's original setting and aligned setting ResNet-50*.

Table 5: DeiT-S shows better generalization to ImageNet-C, ImageNet-A and Stylized-ImageNet. For ImageNet, ImageNet-A and Stylized ImageNet, results show Top-1 accuracy (%). For ImageNet-C, results show mCE (%, mean corruption error, lower is better).

Architecture	ImageNet ↑	ImageNet-A ↑	ImageNet-C ↓	Stylized-ImageNet ↑
ResNet-50	76.9	3.2	57.9	8.3
ResNet-50*	76.3	4.5	55.6	8.2
DeiT-S	76.8	**12.2**	**48.0**	**13.0**

5.2 Best Setting Comparison

Considering the fair comparison and the more advanced training strategies used in DeiT, we first provide the results on fully aligned settings. However, this strategy crafted for DeiT may not benefit ResNet-50 the most. Therefore, we study those settings step-by-step and provide a best setting for ResNet-50 and DeiT-S as another comparison.

5.2.1 Optimizers and Learning Rate Scheduler

In previous studies of the robustness comparison between ResNet and DeiT, optimizers and learning rate scheduler are not aligned. Here we study which is best for ResNet and under its best setting, if ResNet is comparable with DeiT.

Experiment Setting ResNet uses SGD in the standard training, while DeiT used AdamW with cosine learning rate. We report the generalization results on both those settings.

Main Results Table 6 shows with SGD and cosine learning rate scheduler, ResNet-50 has better generalization to those datasets. Therefore, in later experiments, without extra mentions, we'll keep this setting. But even under the best setting, DeiT-S still leads by a large margin.

Table 6: The generalization of ResNet-50 under different optimizer and learning rate scheduler compared with DeiT-S.

	Optimizer	ImageNet ↑	ImageNet-A ↑	ImageNet-C ↓	Sty-ImageNet ↑
ResNet-50	SGD-Step	76.9	3.2	57.9	8.3
	SGD-Cosine	77.4	3.3	56.9	8.4
	AdamW	76.4	3.1	59.3	8.1
DeiT-S	AdamW	76.8	**12.2**	**48.0**	**13.0**

5.2.2 Augmentations

Experiments show that augmentations are beneficial to the robustness on out-of-distribution samples, such as ImageNet-A and ImageNet-C. DeiT uses many augmentations, including RandAug, MixUp and CutMix, which may potentially increase its robustness. Therefore, we apply the same augmentations to ResNet as well.

Experiment Setting Here we stick to SGD as optimizer with cosine learning rate schedule, which works the best for ResNet-50 as aforementioned, and apply different combinations of the augmentation strategies used in DeiT.

Main Results As is shown in Table 7, after applying the augmentations, the generalization ability of ResNet-50 on distorted ImageNet datasets has improved a lot. In general, RandAug with MixUp shows the best improvements. However, there still exist some gap compared with DeiT-S.

Table 7: DeiT shows better generalization compared with ResNet under different augmentations.

Architecture	Augmentations			ImageNet ↑	ImageNet-A ↑	ImageNet-C ↓	Sty-ImageNet↑
	RandAug	MixUp	CutMix				
ResNet-50	✗	✗	✗	77.4	3.3	56.9	8.4
	✓	✓	✓	76.4	5.5	54.0	9.1
	✓	✓	✗	75.7	**6.3**	**52.3**	**10.8**
	✓	✗	✓	76.7	6.3	56.3	7.1
	✗	✓	✓	77.1	6.1	55.1	8.8
DeiT-S	✓	✓	✓	**76.8**	**12.2**	**48.0**	**13.0**

5.2.3 Results of Best Setting Comparison

Therefore, after fully analyzing those settings step-by-step, we are able to provide a comparison between these two architectures under their best setting. DeiT-S shows better generalization on out-of-distribution data in all those three dataset by a large margin under the best setting. The best setting we find for ResNet is SGD optimizer with cosine learning rate scheduler, applied with RandAug and MixUp. And we'll later refer to it as ResNet-Best.

Table 8: DeiT shows better generalization ability compared with ResNet-50 under best settings.

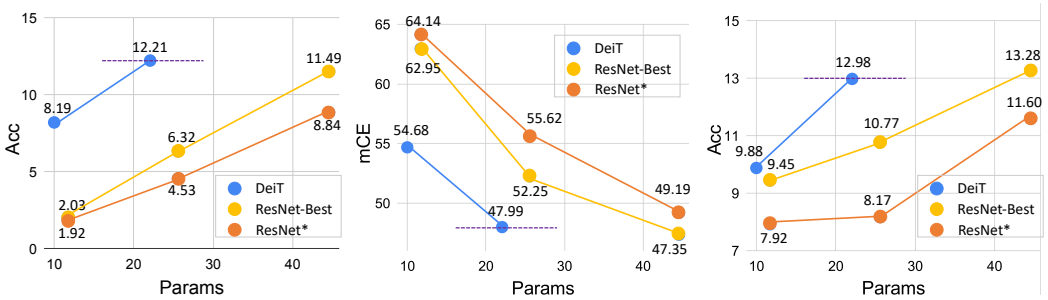
Architecture	ImageNet ↑	ImageNet-A ↑	ImageNet-C ↓	Stylized-ImageNet ↑
ResNet-50	75.7	6.3	52.3	10.8
DeiT-S	76.8	**12.2**	**48.0**	**13.0**

5.3 The Model Size

In order to better understand and compare DeiT and ResNets, we evaluate a range of models from each architecture family to study this two archtectures' robustness under different model size.

Experiment Setting As aforementioned, ResNet-50 and DeiT-S has similar parameters. However, for a smaller model ResNet-18, the original DeiT model family do not contain one with similar parameters. Therefore, in order to make a fair comparison with , we follow the designing rule of DeiT, and set up a model with similar parameters as ResNet-18, namely DeiT-Mini, whose embedding dimension is 256 and number of head is 4, between DeiT-Tiny and DeiT-S.

Main Results From Figure 2 we can observe that both DeiT models show better robustness compared with ResNet models with similar model size under their best settings. Moreover, we further increase ResNet's model size to ResNet-101. After the capacity of ResNet grows to twice its size, it is now compatible with DeiT-S.



(a) ImageNet-A (b) ImageNet-C (c) Stylized-ImageNet

Figure 2: DeiT models show better generalization ability compared with ResNet models with different model sizes on ImageNet-A, ImageNet-C and Stylized-ImageNet.

5.4 Distillation

In the previous sections, we make comparison of ResNet-50 and DeiT-S under aligned setting and best setting seperately. Here, we provide another comparison with the help of knowledge distillation. By analyzing the generalization difference between teacher model and student model, which learns from the passed knowledge its teacher model, we can further explore the generalization difference between these two different architectures.

Experiment Setting We design two sets of experiments. In the first set, we use ResNet-50 as a teacher model to distill DeiT-S; in the second, DeiT-S as the teacher model to distill ResNet-50. We perform soft distillation following[19], which minimizes the Kullback-Leibler(KL) divergence between the softmax of the teacher and the softmax of the student model. During distillation, our training strategy remains consistent with DeiT-S setting, same as the setting in Section 5.1.

Main Results In the first set, after distilling with ResNet-50's knowledge, DeiT-S achieves higher generalization on ImageNet-A, ImageNet-C and Stylized-ImageNet, by a large margin. However, in the second set, it is counter-intuitive that after using DeiT-S as teacher model, the student model, distilled ResNet-50, is actually performing worse. This suggest that the generalization robustness of DeiT is rooted in the structure of Transformer that cannot be passed to ResNet via knowledge distillation.

Table 9: Generalization ability of ResNet-50, DeiT-S and their distilled models.

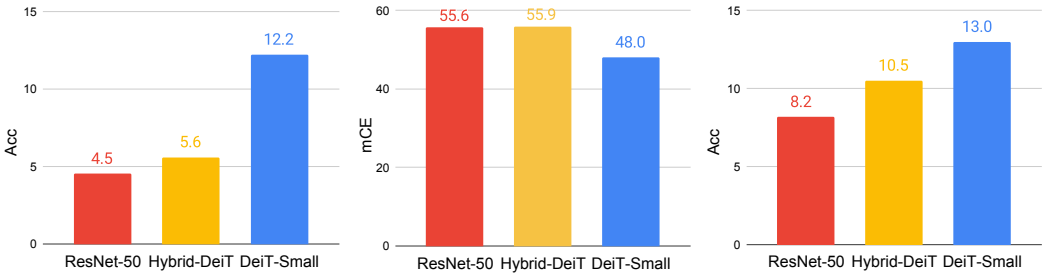
Distillation	Architecture	ImageNet ↑	ImageNet-A ↑	ImageNet-C ↓	Stylized-ImageNet ↑
Teacher	ResNet-50*	76.3	4.5	55.6	8.2
Student	DeiT-S-Distill	76.2	10.9 (+6.4)	49.3 (-6.3)	11.9 (+3.7)
Teacher	DeiT-S	76.8	12.2	48.0	13.0
Student	ResNet-50*-Distill	76.7	5.2 (-7.0)	54.2 (+6.2)	9.8 (-3.2)

5.5 Hybrid Architecture

As discussed above, DeiT architecture shows great generalization ability on out-of-distribution data by a large margin compared to ResNet models. In order to further validate whether this robustness is brought by the architecture of Transformers, we introduce Transformer block to ResNet structure as hybrid model and analyze its generalization ability.

Experiment Setting We follow the basic rule of ViT's design for hybrid model, in which the patch embedding projection is applied to patches extracted from a CNN feature map. In our case, we take the output of stage 4 of a regular ResNet-18 and then feed the intermediate feature maps into DeiT-Mini. The choice of those two architectures are meant to keep the model size of original and the hybrid at the same level. Under our setting, hybrid-DeiT contains ~21 million parameters, which is in the same scale with ResNet-50 (~25 million parameters) and DeiT-S (~22 million parameters). Noe that all those models are trained following the DeiT recipe, same as the setting in Section 5.1.

Main Results Figure 3 shows the generalization on out-of-distribution samples of ResNet-50, DeiT-S and Hybrid-DeiT on ImageNet-A, ImageNet-C and Stylized ImageNet, respectively. We note that, with the introduction of Transformer block, this hybrid stucture's robustness increases in general compared with ResNet-50—with 1.1% growth on ImageNet-A top-1 accuracy and 2.5% growth on Stylized-ImageNet, while maintains similar mCE on ImageNet-C. However, if compared with DeiT-S, which is composed of purely Transformer blocks, hybrid structure's generalization robustness is still lower by a large margin. Therefore, we can conclude that the Transformer block do imply more generalization robustness.



(a) ImageNet-A (b) ImageNet-C (c) Stylized-ImageNet

Figure 3: Transformer blocks show better generalization compared with ResNet structures.

5.6 300-Epoch Training

As aforementioned, 100-epoch setting is able to provide a fair comparison under similar number of parameters and a similar top-1 accuracy on ImageNet classification with less computational budget. Nevertheless, we are also interested in the longer training set-up, to further understand and compare the robustness of these two structures.

From Table 10, though we can still observe that the Transformer shows much stronger generalization than the CNN on out-of-distribution samples, we argue that such setup is less interesting and a little bit unfair—DeiT-S shows much better accuracy than ResNet-50 on ImageNet classification, therefore it is expected that DeiT-S will also show stronger robustness generalization than ResNet-50 on out-of-distribution samples.

Table 10: DeiT-S generalizes better under 300-epoch setting compared with ResNet-50.

Architecture	ImageNet ↑	ImageNet-A ↑	ImageNet-C ↓	Stylized-ImageNet ↑
ResNet-50	78.1	8.8	50.3	9.5
DeiT-S	79.9	**19.2**	**42.8**	**15.1**

To make the setup fairer (i.e., comparing the robustness of models that have similar accuracy), we now compare DeiT-S to the much larger ResNet-101 (*i.e.*, ~22 million parameters *vs*. ~45 million

parameters). The results are shown in Table 11. We can see that though both models achieve similar accuracy on ImageNet, DeiT-S demonstrates much stronger robustness generalization than ResNet-101. This observation can also holds for bigger Transformers and CNNs, *e.g.*, DeiT-B can consistently outperforms the much larger ResNet-200 on ImageNet-A, ImageNet-C and Stylized-ImageNet, suggesting Transformers are truly more robust than CNNs on out-of–distribution samples.

Table 11: DeiTs show better generalization ability under similar accuracy with ResNets.

Architecture	ImageNet ↑	ImageNet-A ↑	ImageNet-C ↓	Stylized-ImageNet ↑
ResNet-101	80.2	17.6	45.8	11.9
DeiT-S	79.9	**19.2**	**42.8**	**15.1**
ResNet-200	82.1	23.8	40.8	13.6
DeiT-B	81.8	**27.9**	**38.0**	**17.9**

In summary, in this 300 training epoch setup, the same conclusion is drawn as 100-epoch setup that Transformers are much more robust than CNNs on out-of-distribution samples.

6 Conclusion

With the recent success of Vision Transformer on recognition tasks, researchers begin to study its robustness compared with traditional CNNs. While many recent works suggest that Transformers are much more robust than CNNs, their comparisons are not fair in many aspects, *e.g.*, training datasets, model scales, training strategies, *etc*. Therefore, we believe that such conclusion cannot be drawn based on these settings, which motivates us to provide a fair and in-depth comparisons between these two architectures. With our unified training setup, we found that Transformers are no more robust than CNNs regarding adversarial robustness. And if properly aligned, CNNs can achieve similar robustness as Transformers on defending against both perturbation-based adversarial attacks and patch-based adversarial attacks. While regarding generalization on out-of-distribution samples (*e.g.*, ImageNet-A, ImageNet-C and Stylized ImageNet), we find pre-training on (external) large-scale datasets is not a necessary condition for letting Transformers show strong performance than CNNs. Additionally, we find adopting self-attention-like architectures is crucial for generalization robustness. We hope this work would shed lights on the understanding of Vision Transformer and the robustness between Transformers and CNNs.

Acknowledgements

This work was partially supported by the ONR N00014-20-1-2206, ONR N00014-18-1-2119 and Institute for Assured Autonomy at JHU with Grant IAA 80052272. Cihang Xie was supported by a gift grant from Open Philanthropy.

References

[1] Maksym Andriushchenko, Francesco Croce, Nicolas Flammarion, and Matthias Hein. Square attack: a query-efficient black-box adversarial attack via random search. In *European Conference on Computer Vision*, pages 484–501. Springer, 2020.

[2] Irwan Bello, Barret Zoph, Ashish Vaswani, Jonathon Shlens, and Quoc V Le. Attention augmented convolutional networks. In *ICCV*, 2019.

[3] Srinadh Bhojanapalli, Ayan Chakrabarti, Daniel Glasner, Daliang Li, Thomas Unterthiner, and Andreas Veit. Understanding robustness of transformers for image classification. *arXiv preprint arXiv:2103.14586*, 2021.

[4] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

[5] Xiangning Chen, Cho-Jui Hsieh, and Boqing Gong. When vision transformers outperform resnets without pretraining or strong data augmentations. *arXiv preprint arXiv:2106.01548*, 2021.

[6] Francesco Croce and Matthias Hein. Minimally distorted adversarial examples with a fast adaptive boundary attack. In *International Conference on Machine Learning*, pages 2196–2205. PMLR, 2020.

[7] Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International Conference on Machine Learning*, pages 2206–2216. PMLR, 2020.

[8] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical data augmentation with no separate search. *arXiv preprint arXiv:1909.13719*, 2019.

[9] Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V Le, and Ruslan Salakhutdinov. Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv preprint arXiv:1901.02860*, 2019.

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[12] 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 *ICLR*, 2018.

[13] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *ICLR*, 2015.

[14] Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[16] Dan Hendrycks and Thomas G Dietterich. Benchmarking neural network robustness to common corruptions and surface variations. *arXiv preprint arXiv:1807.01697*, 2018.

[17] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. *arXiv preprint arXiv:1907.07174*, 2019.

[18] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. *CVPR*, 2021.

[19] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[20] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *CVPR*, 2017.

[21] Alex Krizhevsky, Ilya Sutskever, and Geoff Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.

[22] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *ECCV*. 2014.

[23] Hanxiao Liu, Zihang Dai, David R So, and Quoc V Le. Pay attention to mlps. *arXiv preprint arXiv:2105.08050*, 2021.

[24] Liyuan Liu, Xiaodong Liu, Jianfeng Gao, Weizhu Chen, and Jiawei Han. Understanding the difficulty of training transformers. *arXiv preprint arXiv:2004.08249*, 2020.

[25] 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. *arXiv preprint arXiv:2103.14030*, 2021.

[26] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[27] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *ICLR*, 2018.

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

[29] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[30] Ilija Radosavovic, Raj Prateek Kosaraju, Ross Girshick, Kaiming He, and Piotr Dollár. Designing network design spaces. In *CVPR*, 2020.

[31] Prajit Ramachandran, Niki Parmar, Ashish Vaswani, Irwan Bello, Anselm Levskaya, and Jonathon Shlens. Stand-alone self-attention in vision models. *arXiv preprint arXiv:1906.05909*, 2019.

[32] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *IJCV*, 2015.

[33] Ali Shafahi, Mahyar Najibi, Amin Ghiasi, Zheng Xu, John P. Dickerson, Christoph Studer, Larry S. Davis, Gavin Taylor, and Tom Goldstein. Adversarial training for free! In *NeurIPS*, 2019.

[34] Rulin Shao, Zhouxing Shi, Jinfeng Yi, Pin-Yu Chen, and Cho-Jui Hsieh. On the adversarial robustness of visual transformers. *arXiv preprint arXiv:2103.15670*, 2021.

[35] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.

[36] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *ICCV*, 2017.

[37] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *CVPR*, 2015.

[38] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *ICLR*, 2014.

[39] Ilya Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Daniel Keysers, Jakob Uszkoreit, Mario Lucic, et al. Mlp-mixer: An all-mlp architecture for vision. *arXiv preprint arXiv:2105.01601*, 2021.

[40] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. *arXiv preprint arXiv:2012.12877*, 2020.

[41] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers distillation through attention. *arXiv preprint arXiv:2012.12877*, 2020.

[42] Hugo Touvron, Matthieu Cord, Alexandre Sablayrolles, Gabriel Synnaeve, and Hervé Jégou. Going deeper with image transformers. *arXiv preprint arXiv:2103.17239*, 2021.

[43] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, 2017.

[44] 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. *arXiv preprint arXiv:2102.12122*, 2021.

[45] Eric Wong, Leslie Rice, and J Zico Kolter. Fast is better than free: Revisiting adversarial training. In *ICLR*, 2020.

[46] Cihang Xie, Mingxing Tan, Boqing Gong, Alan Yuille, and Quoc V Le. Smooth adversarial training. *arXiv preprint arXiv:2006.14536*, 2020.

[47] Qizhe Xie, Eduard Hovy, Minh-Thang Luong, and Quoc Le. Self-training with noisy student improves imagenet classification. *arXiv preprint arXiv:1911.04252*, 2019.

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

[49] Chenglin Yang, Adam Kortylewski, Cihang Xie, Yinzhi Cao, and Alan Yuille. Patchattack: A black-box texture-based attack with reinforcement learning. In *European Conference on Computer Vision*, pages 681–698. Springer, 2020.

[50] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. *arXiv preprint arXiv:1906.08237*, 2019.

[51] Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Zihang Jiang, Francis EH Tay, Jiashi Feng, and Shuicheng Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet. *arXiv preprint arXiv:2101.11986*, 2021.

[52] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *ICCV*, 2019.

[53] Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. *arXiv preprint arXiv:2106.04560*, 2021.

[54] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *ICLR*, 2018.

[55] Hengshuang Zhao, Jiaya Jia, and Vladlen Koltun. Exploring self-attention for image recognition. In *CVPR*, 2020.

[56] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *CVPR*, 2017.