Trade-off between Robustness and Accuracy of Vision Transformers

Yanxi Li, Chang Xu
School of Computer Science, Faculty of Engineering, The University of Sydney, Australia
yali0722@uni.sydney.edu.au, c.xu@sydney.edu.au

Abstract

*Although deep neural networks (DNNs) have shown great successes in computer vision tasks, they are vulnerable to perturbations on inputs, and there exists a trade-off between the natural accuracy and robustness to such perturbations, which is mainly caused by the existence of robust non-predictive features and non-robust predictive features. Recent empirical analyses find Vision Transformers (ViTs) are inherently robust to various kinds of perturbations, but the aforementioned trade-off still exists for them. In this work, we propose Trade-off between Robustness and Accuracy of Vision Transformers (**TORA-ViTs**), which aims to efficiently transfer ViT models pretrained on natural tasks for both accuracy and robustness. TORA-ViTs consist of two major components, including a pair of accuracy and robustness adapters to extract predictive and robust features, respectively, and a gated fusion module to adjust the trade-off. The gated fusion module takes outputs of a pretrained ViT block as queries and outputs of our adapters as keys and values, and tokens from different adapters at different spatial locations are compared with each other to generate attention scores for a balanced mixing of predictive and robust features. Experiments on ImageNet with various robust benchmarks show that our TORA-ViTs can efficiently improve the robustness of naturally pretrained ViTs while maintaining competitive natural accuracy. Our most balanced setting (TORA-ViTs with $\lambda = 0.5$) can maintain 83.7% accuracy on clean ImageNet and reach 54.7% and 38.0% accuracy under FGSM and PGD white-box attacks, respectively. In terms of various ImageNet variants, it can reach 39.2% and 56.3% accuracy on ImageNet-A and ImageNet-R and reach 34.4% mCE on ImageNet-C.*

1. Introduction

In the past few decades, deep neural networks (DNNs) have been well developed to achieve or even surpass the performance of humans on computer vision tasks [13,23,24,54, 55]. However, a fatal drawback of them is that they are vulnerable to perturbations on inputs [8, 14, 15, 17, 32], which will cause dramatically drop in their accuracy. Furthermore, recent studies demonstrate that there exists a trade-off between natural accuracy and adversarial robustness [48, 57], which means improving the robustness of a network typically leads to a decrease in accuracy on natural samples.

A popular theory explains this trade-off by the existence of two kinds of different features [22, 48, 53]. The first kind of feature is moderately correlated to the task and robust to attacks, while the second kind of feature is weakly correlated to the task and therefore non-robust. It is unfortunate that those moderately correlated and robust features only have limited contributions to accurate predictions (*robust* and *non-predictive*), and further improving the accuracy heavily relies on those weakly related and non-robust features (*predictive* and *non-robust*) [48]. Therefore, this trade-off is usually considered an inherent characteristic of DNNs. Although there are many efforts that aim to control or improve this trade-off [26, 40, 41, 50, 57], it is still hard to efficiently and effectively improve it on large-scale datasets such as ImageNet [3].

Recently, a new family of vision models, namely Vision Transformers (ViTs) [6, 44, 47], has outperformed CNNs on various kinds of tasks. There are many subsequent works that discuss diverse variants of ViTs to improve their performance. TNT [11] divides patches in ViTs into smaller sub-patches and applies a transformer-in-transformer architecture with an additional inner transformer. T2T-ViT [56] introduces local feature aggregation to boost local information. Swin [29] performs local attention within various windows, and a shifted window partitioning approach is introduced for cross-window connections.

However, the aforementioned works mainly focus on the natural accuracy on clean data. Although empirical analyses have demonstrated that ViTs demonstrates robustness against various kinds of perturbations [1, 33, 37], there are only a limited number of works [9, 19, 34, 60] focus on improving the robustness. Besides, how to boost a naturally pretrained ViT for robustness has been ignored by existing methods. A pretrained ViT has high utility because it can extract predictive features to ensure high accuracy on downstream tasks. Despite the high utility, it also has low reliability, because its non-robust features are vulnerable to perturbations. Therefore, it is worth studying how to obtain a useful and reliable ViT.

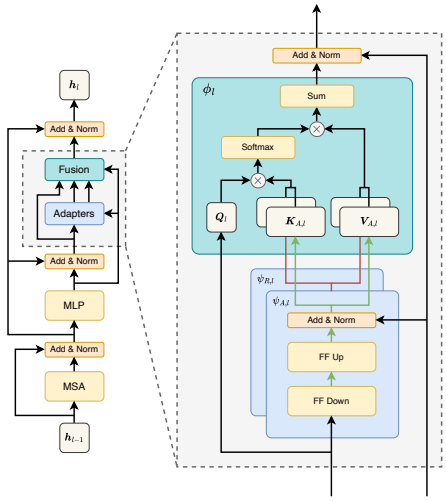


Figure 1. The overall architecture of our TORA-ViTs. The TORA-ViTs consists of two major components, including a pair of an accuracy adapter $\psi_{A,l}$ to extract predictive features and a robust adapter $\psi_{R,l}$ to extract robust features, and a gated fusion module to combine those features as inputs for next ViT block. TORA-ViTs is inserted after the MLP layer in each ViT block.

Furthermore, fine-tuning of ViTs is very computationally intensive [6]. Considering the high overheads of adversarial training [8, 32, 57], it is even more expensive to train both accurate and robust ViT models on a large-scale dataset. To reduce the intractable cost of training and fine-tuning large-scale Transformer models on various tasks, adapter [20] proposes to add and fine-tune only a few parameters per task in the field of NLP. AdapterFusion [38] further supports multi-task transfer learning by using multiple adapters in parallel and combining their outputs with an attention-based gated fusion module.

In this work, we propose Trade-off between Robustness and Accuracy of Vision Transformers (**TORA-ViTs**) for utility and reliability at the same time. TORA-ViTs transfer naturally pretrained models with low computational demands to improve their adversarial robustness while maintaining competitive natural accuracy. Based on the theory of robust non-predictive and non-robust predictive features, we add two kinds of adapter modules after the MLP layer of an existing ViT block, including an accuracy one $\psi_{A,l}$ and a robust one $\psi_{R,l}$ to extract predictive and robust features, respectively. Then, a gated fusion module (ϕ_l) is introduced to combine the extracted features in a trade-off-aware manner utilizing the attention mechanism. In the gated fusion module, features extracted by pretrained ViT blocks are used as queries, and features extracted by the newly added accuracy and robustness adapters are used as keys and values. Then, the softmax function is applied adapter-wise as a gate to combine the two kinds of features. The overall architecture of our TORA-ViTs is shown in Fig. 1.

The TORA-ViTs are optimized in a two-phase manner. In the first phase, the accuracy and robustness adapters are optimized alternately along with the gated fusion module. When each of them reaches a proper performance, they are frozen, and the gated fusion module is optimized with a joint objective of accuracy and robustness with a trade-off ratio λ. Experiments on ImageNet with various robust benchmarks, including white-box adversarial attacks (FGSM and PGD), natural adversarial example (ImageNet-A), out-of-distribution data (ImageNet-R), and common corruptions (ImageNet-C), show that our TORA-ViTs can efficiently improve the robustness of naturally pretrained ViTs. Meanwhile, the natural accuracy is still competitive with or even better than the models pursuing accuracy. Our most balanced setting (TORA-ViTs with $\lambda = 0.5$) can maintain 83.7% accuracy on clean ImageNet and reach 54.7% and 38.0% accuracy under FGSM and PGD white-box attacks, respectively. In terms of various ImageNet variants, it can reach 39.2% and 56.3% accuracy on ImageNet-A and ImageNet-R and reach 34.4% mCE on ImageNet-C.

2. Related Works

2.1. ViTs and Adapter

Inspired by the success of multi-head self-attention [49] (MSA) in natural language processing (NLP), there are many attempts to apply this family of models to computer vision tasks. Image Transformer [36] proposes to use Transformers for image generation tasks in a sequence modeling formulation. To modify Transformers for image classification, Hu *et al*. [21] and Zhao *et al*. [59] design local multi-head dot-product self-attention blocks. Ramachandran *et al*. [42] further expands self-attention for both classification and object detection. Vision Transformers (ViTs) [6] propose a novel embedding method, which splits images as sequences of non-overlapping patches. Although ViTs reach state-of-the-art accuracy, they demand costly pretraining on large-scale datasets, such as ImageNet-21K [3] and JFT-300M [45]. DeiT [47] proposes a method to efficiently distill ViTs via an additional distillation token, which ensures that the student learns from the teacher through attention. Naseer *et al*. [35] modify DeiT and introduce a shape token to encode shape-information. MAE [12] proposes a masked autoencoder (MAE) built with ViTs for self-supervised learning, which masks random patches of input images.

To reduce the training and fine-tuning cost of Transformers, in the field of NLP, adapter [20] proposes to add and fine-tune only a few trainable parameters to pretrained BERT Transformers [4]. This enables efficient transfer learning among a large number of diverse text classification tasks. To support multi-task learning (MTL) in NLP, AdapterFusion [38] uses multiple adapters in parallel in a

two-stage manner, which consists of a knowledge extraction stage and a knowledge composition stage. In the knowledge composition stage, the attention mechanism is used to combine the set of adapters.

2.2. Robust Vision Model

Several studies investigate the robustness of convolutional neural networks (CNNs), with adversarial training being a popular method to enhance resistance to adversarial attacks. The Fast Gradient Sign Method (FGSM) [8] generates adversarial perturbations using one-step gradient ascent. Madry *et al*. [32] proposes a stronger multi-step variant using projected gradient descent (PGD). However, TRADES [48] reveals a trade-off between natural accuracy and adversarial robustness. The authors analyze this trade-off and propose a boundary error to guide defense design. Kim *et al*. [22] propose to distill robust and non-robust features in intermediate feature space by employing Information Bottleneck. Yang *et al*. [53] propose a disentanglement network for robust and non-robust following the framework of the autoencoder.

More recently, perturbations beyond adversarial attacks are gaining increasing interests. ImageNet-C [15] considers common corruptions, which applies a series of 19 common visual corruptions in 5 categories to images. ImageNet-A [17] considers natural adversarial examples, which places objects in unusual contexts or orientations. ImageNet-R [14] considers out-of-distribution data, which contains abstract or rendered versions of objects.

Researchers find that the robustness of neural networks is also dependent on their architectures. There are several efforts [5, 10, 26] to enhance the adversarial robustness of neural networks by Neural Architecture Search [25, 28, 61]. Additionally, ConvNeXt [30] introduces a neural architecture, which is fine-tuned manually and demonstrates robustness [39]. Croce *et al*. [2] also find that small modifications to traditional ResNet-50 architecture lead to substantial improvements in robustness against adversarial attacks.

As an emerging family of new architectures for vision models, there are several empirical studies [1, 33, 37] find that ViTs are robust against various kinds of perturbations. To improve the robustness of ViTs, Robust Vision Transformer (RVT) [34] redesigned the building blocks of ViTs and propose two plug-and-play techniques called position-aware attention scaling and patch-wise augmentation. In the contrast, pyramid adversarial training (PyramidAT) [19] does not modify the network architecture but proposes pyramid attacks to generate adversarial examples by perturbing the input image at multiple scale. FAN [60] examines the role of self-attention in learning robust representations in ViTs and proposes a family of fully attentional networks (FANs) that improve mid-level representations. Gu *et al*. [9] explore the robustness of ViTs against patch-wise perturbations. They find that ViTs are more robust to natural corrupted patches than CNNs, but more vulnerable to adversarial patches. Therefore, a temperature-scaling based method is proposed to improve the robustness of ViTs against adversarial patches.

3. Methodology

3.1. Preliminary

Given the input image x and its relevant label y in a training set $\mathcal{D}$, a common supervised training objective of vision transformers can be written as:

$$\mathcal{L}_{\text{ACC}}(f;\mathcal{D}) = \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}}[\ell_{\text{CE}}\left(f(\boldsymbol{x}),y\right)], \tag{1}$$

where ℓ_{CE} is a cross-entropy function, and f stands for the vision transformer.

To improve the adversarial robustness of model against perturbations on inputs, adversarial training is a common method, where perturbations are used to attack a target model, and the target model is optimized under such attacks. The perturbations are generated with gradient ascent to maximize the classification objective. An adversarial example $\boldsymbol{x}'$ with perturbations is typically limited in a l_p ball $\mathcal{B}_p(\boldsymbol{x},\varepsilon) = \{\boldsymbol{x}' : \|\boldsymbol{x}-\boldsymbol{x}'\|_p \leq \varepsilon\}$ around its corresponding natural example $\boldsymbol{x}$, where ε defines the scale of allowed perturbations, and $\|\cdot\|_p$ is a l_p normalization. Then, adversarial training can be formed into a min-max problem, whose objective function is defined as

$$\mathcal{L}_{\text{ROB}}(f;\mathcal{D}) = \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}}\left[\max_{\boldsymbol{x}'\in\mathcal{B}_p(\boldsymbol{x},\varepsilon)} \ell_{\text{CE}}\left(f(\boldsymbol{x}'),y\right)\right], \tag{2}$$

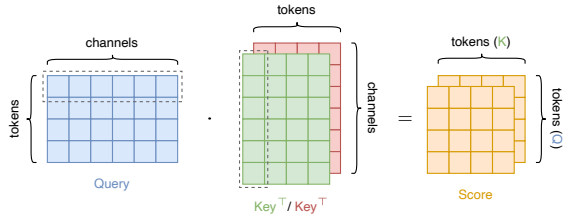
where f is a deep model, $\mathcal{D}$ is the distribution of the natural example $\boldsymbol{x}$ and the corresponding label y, and ℓ_{CE} is a cross-entropy function.

3.2. Robustness and Accuracy Adapters

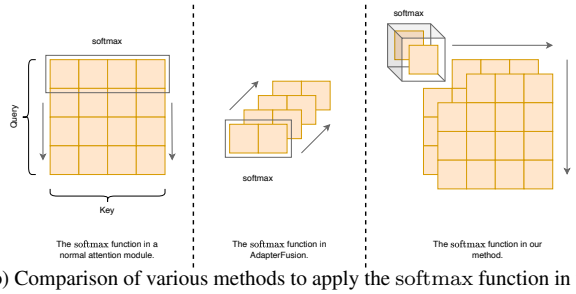
Off-the-shelf vision transformers [6, 11, 29, 56] are often well trained to pursue the natural accuracy through Eq. 1. To enhance the adversarial robustness of these well-trained vision transformers, a standard fine-tuning process can be executed by tuning the weights with the objective of Eq. 2. However, the adversarial robustness as a new objective for vision transformers may lead to a completely different set of weights, which are far away from the initialization and degrade the accuracy of vision transformers.

Taking both natural accuracy and adversarial robustness into consideration, we insert two adapters into an existing ViT block, including an accuracy adapter $\psi_{A,l}$ for predictive features and a robustness adapter $\psi_{R,l}$ for robust features. Given the feature $\boldsymbol{z}_l \in \mathbb{R}^{N\times d_{\text{m}}}$ output by the MLP layer in the $1 \leq l \leq L$ block with N tokens of d_m dimensions, we have the accuracy adapter

$$\boldsymbol{a}_l = \psi_{A,l}(\boldsymbol{z}_l), \tag{3}$$



(a) The dot products between query and keys.



(b) Comparison of various methods to apply the softmax function in the attention mechanism.

Figure 2. The dot-product attention and softmax function in our gated fusion module.

and the robust adapter

$$\boldsymbol{r}_l = \psi_{R,l}(\boldsymbol{z}_l), \tag{4}$$

where $\boldsymbol{a}_l, \boldsymbol{r}_l \in \mathbb{R}^{N \times d_\text{m}}$ are the predictive and robust features, respectively.

For the architecture of adapters, we use two feed-forward layers with a bottleneck and a residual connection following Houlsby *et al*. [20]. We insert adapters right after the MLP layer of an existing ViT block and do not insert adapters after the multi-head attention (MSA). The overall architecture of a block in our TORA-ViTs is as shown in Fig. 1.

3.3. Attention-based Gated Fusion

To combine the predictive and robust features extracted by the accuracy and robustness adapters in a trade-off-aware manner, we propose an attention-based gated fusion module. We first calculate the dot-product attention score matrices between the features from the ViT blocks and adapters. Then, a softmax function is applied adapter-wise to the score matrices. The softmax results are used as a **weighted gate** to fuse the predictive and robust features.

The feature $\boldsymbol{z}_l$ output by the ViT block is used to generate the **query**, and the features $\boldsymbol{a}_l$ and $\boldsymbol{r}_l$ output by adapters are used to generate **keys**. The dot products between the two Q-K pairs are calculated as

$$\boldsymbol{s}_{A,l} = (\boldsymbol{z}_l \cdot \boldsymbol{w}_{Q,l}) \cdot (\boldsymbol{a}_l \cdot \boldsymbol{w}_{K,l})^\top, \tag{5}$$

$$\boldsymbol{s}_{R,l} = (\boldsymbol{z}_l \cdot \boldsymbol{w}_{Q,l}) \cdot (\boldsymbol{r}_l \cdot \boldsymbol{w}_{K,l})^\top, \tag{6}$$

where $\boldsymbol{s}_{A,l}$, $\boldsymbol{s}_{R,l} \in \mathcal{R}^{N \times N}$ are the attention score matrices for the accuracy adapter and robustness adapter, and $\boldsymbol{w}_{Q,l}$, $\boldsymbol{w}_{K,l} \in \mathbb{R}^{d_\text{m} \times d_\text{q}}$ are projection parameters for query and key matrices. The projection parameters are share among adapters.

Then, the softmax function is applied to the attention score matrices adapter-wise instead of token-wise by:

$$\boldsymbol{s}'_{A,l,m,n} = \frac{\exp(\boldsymbol{s}_{A,l,m,n})}{\sum_{k \in \{A,R\}} \exp(\boldsymbol{s}_{k,l,m,n})} \tag{7}$$

$$\boldsymbol{s}'_{R,l,m,n} = \frac{\exp(\boldsymbol{s}_{R,l,m,n})}{\sum_{k \in \{A,R\}} \exp(\boldsymbol{s}_{k,l,m,n})}. \tag{8}$$

In this manner, if a token in one adapter corresponds to a larger attention score than the other, it will be assigned a larger weight, and vice versa. It can act as a gate to select which kind of features can be forwarded to the next block at a larger scale. Similar to keys, the **values** are also generated from the features from adapters. By applying the weights, we can calculate the output of attention module for each adapter as

$$\boldsymbol{o}_{A,l} = \boldsymbol{s}'^\top_{A,l} \cdot (\boldsymbol{a}_l \cdot \boldsymbol{w}_{V,l}) \tag{9}$$

$$\boldsymbol{o}_{R,l} = \boldsymbol{s}'^\top_{R,l} \cdot (\boldsymbol{r}_l \cdot \boldsymbol{w}_{V,l}), \tag{10}$$

where $\boldsymbol{o}_{A,l}$, $\boldsymbol{o}_{R,l} \in \mathbb{R}^{N \times d_m}$, and $\boldsymbol{w}_{V,l} \in \mathbb{R}^{d_\text{m} \times d_\text{v}}$ is the projection parameter for value matrices. The comparison of our method to other previous methods is as shown in Fig. 2.

Because $\boldsymbol{o}_{A,l}$ and $\boldsymbol{o}_{R,l}$ has already been multiplied with weight matrices $\boldsymbol{s}'_{A,l}$ and $\boldsymbol{s}'_{R,l}$ from softmax function, we can directly sum them adapter-wise for the final output:

$$\boldsymbol{o}_l = \sum_{k \in \{A,R\}} \boldsymbol{o}_{k,l}, \tag{11}$$

which ensures $\boldsymbol{o}_l$ to have same dimensions as $\boldsymbol{z}_l$. Finally, we add a residual connection from the output of the ViT block and have the final layer output $\boldsymbol{h}_l$ as

$$\boldsymbol{h}_l = \boldsymbol{z}_l + \text{LN}(\boldsymbol{o}_l), \tag{12}$$

where $\text{LN}(\cdot)$ is a layer norm.

3.4. Two-Phase Trade-off Training

Dosovitskiy *et al*. [6] introduces an extra randomly initialized classification token [CLS] to the embedded patch tokens in ViTs following BERT. This token is later used for the classification task. Similarly, we add an accuracy token and a robustness token for our trade-off training. The original class token is at the first dimension of the output (i.e., [CLS] $:= \boldsymbol{z}_{l,1,:}$). To add our new tokens, we replace $\boldsymbol{z}_{l,1,:}$ with the accuracy token $[\text{ACC}]_{l-1}$ and the robust token $[\text{ROB}]_{l-1}$ to form the adapters inputs. Then, Eqs. 3 and 4 become:

$$\boldsymbol{a}_l = \psi_{A,l}\left(\text{Concat}([\text{ACC}]_{l-1}, \boldsymbol{z}_{l,2:,:})\right) \tag{13}$$

$$\boldsymbol{r}_l = \psi_{R,l}\left(\text{Concat}([\text{ROB}]_{l-1}, \boldsymbol{z}_{l,2:,:})\right). \tag{14}$$

After the adapter, we can have $[\text{ACC}]_l = \boldsymbol{a}_{l,1,:}$ and $[\text{ROB}]_l = \boldsymbol{r}_{l,1,:}$. To make the final classification, an accuracy classification head f_{ACC} and a robustness classification head f_{ROB} are added, and their predictions are averaged:

$$\hat{y} = \frac{1}{2} f_{\text{ACC}}([\text{ACC}]_L) + \frac{1}{2} f_{\text{ROB}}([\text{ROB}]_L). \quad (15)$$

To optimize our TORA-ViTs, we use a two-phase training strategy. We first optimize each adapter independently with their specific objective. In this phase, the fusion module is also optimized. Then, the two adapters are frozen, and the fusion module is optimized with the joint robustness and accuracy objective. During the entire training process, the pretrained ViT is always frozen. In the first phase, Eqs. 1 and 2 are used to optimize the corresponding adapter along with the gated fusion:

$$\min_{\Psi_R, \Phi} \quad \mathcal{L}_{\text{ROB}(F;\mathcal{D})}, \quad (16)$$

$$\min_{\Psi_A, \Phi} \quad \mathcal{L}_{\text{ACC}(F;\mathcal{D})}, \quad (17)$$

where $F = \{f, \Psi_R, \Psi_A, \Phi\}$ with $\Psi_R = \{\psi_{R,l} | 1 \leq l \leq L\}$, $\Psi_A = \{\psi_{A,l} | 1 \leq l \leq L\}$ and $\Phi = \{\phi_l | 1 \leq l \leq L\}$ represents the TORA-ViT model with adapters and gated fusion. In Eqs. 16 and 17, the trade-off ratio λ is temporatly omitted, because each objective is optimized alternately. In the second phase, we use a joint objective to optimize Φ with λ:

$$\min_{\Phi} \quad \lambda \mathcal{L}_{\text{ROB}}(F;\mathcal{D}) + (1-\lambda)\mathcal{L}_{\text{ACC}}(F;\mathcal{D}). \quad (18)$$

Because the fusion module Φ can be easily biased to the current object in the previous phase, this phase aims to adjust Φ with joint optimization and make the trade-off to be better correlated with the demand ratio λ.

4. Experiments

4.1. Settings

Pretrained ViTs. We consider the vanilla ViT architecture proposed by Dosovitskiy *et al*. [6] in our experiments. The ViT-B/16 with 224×224 input size, 16×16 patch size, 768-dimension embedding, and 12 layers is used. We initialize the network with pretrained parameters provided by Steiner *et al*. [44].

Training. The existing ViT blocks are frozen during training. The adapters are optimized with AdamW [31] optimizer on ImageNet-1K [3] with a 0.0001 initial learning rate and step decay with a rate of 0.97. For adversarial training, we use the single-step FGSM [8] with $\varepsilon = 1/255$ to generate adversarial examples. The model is trained for 9 epochs in total. In the first 6 epochs, the alternate optimization in Eqs. 16 and 17 is performed, which means each objective is optimized for 3 epochs. In the last 3 epochs, the joint optimization in Eq. 18 is performed.

Evaluation. For white-box attacks, we use single-step FGSM [8] and multi-step PGD [32] on ImageNet-1K. We follow Mao *et al*. [34] and use $\varepsilon = 1/255$, PGD with 5 steps, and step size $0.5/255$. For natural adversarial examples, we use ImageNet-A [17], which places the ImageNet objects in unusual contexts or orientations. For out-of-distribution data, we use ImageNet-R [14], which contains abstract or rendered versions of the object. For common corruption, we use ImageNet-C [15], which applies 19 common corruptions in 5 categories (e.g., motion blur, Gaussian noise, fog, JPEG compression, etc.).

4.2. Comparison to SOTA Methods

In Table 1, we compare our TORA-ViT (categorized as "robust adapters") to 4 categories of state-of-the-art (SOTA) methods, including naturally trained CNNs, robust CNNs, naturally trained ViTs, and robust ViTs. We report three different trade-offs of TORA-ViT in Table 1, including a most balanced setting, which outperforms all the baselines, a setting for good natural accuracy, and a setting for extremely high robustness against adversarial attacks. Other ratios are reported and discussed with more details in Table 2.

Our method performs well on both clean and robust tasks with $\lambda = 0.5$. It outperforms previous works on all metrics. This is our **most balanced** setting. Comparing to previous best SOTA methods, it improves natural accuracy on *clean data* by 0.3% than Swin-B; improves accuracy under *FGSM* by 1.7%, accuracy under *PGD* by 8.1% and accuracy on *ImageNet-R* by 7.6% than RVT-B*; improves accuracy on *ImageNet-A* by 2.8% than PyramidAT-384; and reduce mCE on *ImageNet-C* by 10.6% than PyramidAT.

The model mainly focuses on natural accuracy when $\lambda = 0.1$. Comparing to naturally trained ViTs, it improves the natural accuracy by 0.7% comparing to the previous best model, i.e., Swin-B. Besides, in terms of robustness, it also reaches better performance than Swin-B under PGD attack (2% higher accuracy) and on all ImageNet variants (10.7% and 11.0% higher accuracy on ImageNet-A and -R, respectively, and 22.7% lower mCE on ImageNet-C). Is is only slightly lower than Swin-B under FGSM by 0.8%. Comparing to robust ViTs, it is better than all of them in terms of natural accuracy and accuracy on ImageNet variants. Although the performance under adversarial attacks is lower than robust ViTs, considering this is a model trading robustness for accuracy, it is still remarkable to reach the best natural accuracy and the best robustness on ImageNet variants among our settings, which also outperforms all the previous SOTA methods, by sacrificing some robustness against adversarial attacks.

The adversarial robustness becomes the main target if we set $\lambda = 0.9$. This setting improve accuracy under FGSM by 21.2% and accuracy under PGD by 27.6% comparing to the previously best RVT-B*, which is a surprisingly large im-

Categories	Models	Clean	Attacks		ImageNet Variants		
			FGSM	PGD	A	R	C(↓)
CNNs	ResNet-50 [13]	76.1	12.2	0.9	0.0	36.1	76.7
	ResNeXt50-32x4d [52]	79.8	34.7	13.5	10.7	41.5	64.7
	EfficientNet-B4 [46]	83.0	**44.6**	**18.5**	26.3	47.1	71.1
	ConvNeXt-B [30]	**83.8**	-	-	**36.7**	**51.3**	**46.8**
Robust CNNs	ANT [43]	76.1	17.8	3.1	1.1	39.0	63.0
	AugMix [16]	77.5	20.2	3.8	3.8	41.0	65.3
	Debiased CNN [27]	76.9	20.4	5.5	3.5	40.8	67.5
	DeepAugment [14]	75.8	27.1	9.5	3.9	**46.7**	**53.6**
	Anti-Aliased CNN [58]	**79.3**	**32.9**	**13.5**	**8.2**	41.1	68.1
ViTs	ViT-B/16 [6]	72.8	-	-	8.0	27.1	74.8
	ViT-B/16 + CutMix [6]	75.5	-	-	14.8	28.5	64.1
	ViT-B/16 + MixUp [6]	77.8	-	-	12.2	34.9	61.8
	ViT-B/16 + AugReg [44]	79.9	-	-	17.5	38.2	52.5
	ViT-B/16-384 + AugReg [44] †	81.4	-	-	26.2	38.2	58.2
	PVT-Large [51]	81.7	33.1	7.3	26.6	42.7	59.8
	ConViT-B [7]	82.4	45.4	20.8	29.0	**48.4**	**46.9**
	DeiT-B/16 [47]	82.0	46.4	21.3	27.4	44.9	48.5
	T2T-ViT_t-24 [56]	82.6	46.7	17.5	28.9	47.9	48.0
	Swin-B [29]	**83.4**	49.2	21.3	**35.8**	46.6	54.4
	PiT-B [18]	82.4	**49.3**	**23.7**	33.9	43.7	48.2
Robust ViTs	PyramidAT [19]	81.7	-	-	23.0	47.7	45.0
	PyramidAT-384 [19] †	83.3	-	-	**36.4**	46.7	47.8
	RVT-B [34]	82.5	52.3	27.4	27.7	48.2	47.3
	RVT-B* [34]	82.7	**53.0**	**29.9**	28.5	48.7	46.8
	MAE-ViT-B [12]	83.6	-	-	35.9	48.3	51.7
	FAN-L-ViT [60]	**83.9**	-	-	34.2	**53.1**	**43.3**
Robust Adapters (ours)	TORA-ViT-B/16 ($\lambda = 0.1$)	**84.1**	48.4	23.3	**46.5**	**57.6**	**31.7**
	TORA-ViT-B/16 ($\lambda = 0.5$)	83.7	54.7	38.0	39.2	56.3	34.4
	TORA-ViT-B/16 ($\lambda = 0.9$)	80.3	**74.2**	**57.5**	22.2	53.7	41.6

Table 1. Performance on ImageNet-1K and variants. For performance on clean ImageNet-1K, under adversarial attacks, on ImageNet-A, and on ImageNet-R, the top-1 accuracy is reported. For performance on ImageNet-C, the mean Corruption Error (mCE) is reported, which is the smaller the better (marked by ↓).

†: "ViT-B/16-384 + AugReg" and "PyramidAT-384" use 384×384 inputs, and other models use 224×224 inputs.

provement on robustness against adversarial attacks. This improvement sacrifices performance on clean data and ImageNet variants. Comparing to the most balanced setting with $\lambda = 0.5$, its natural accuracy drops 3.4%, and its accuracy on ImageNet-A drops 17%. Although its robustness on ImageNet-R and -C is also the worst among our settings, it is still better than previous SOTA methods.

Another interesting observation about our method is that *robustness on the three ImageNet variants have positive correlation with natural accuracy and negative correlation with robustness under adversarial attacks*. This phenomenon also exists for other SOTA methods, although the correlations are not as strong as those in our method. For example, PiT reaches the best robustness against adversarial attacks among ViTs, but its performance under other kinds of perturbations is not the best; Anti-Aliased CNN reaches the best robustness against adversarial attacks among robust CNNs, but its robustness on ImageNet-R and -C is worse than DeepAugment. This also demonstrates the importance of controlling the trade-off when applying adversarial training, because robustness to adversarial attacks is only one aspect of robustness, and the robustness to other kinds of perturbations is not always improved together with it.

λ	Head	Clean	Attacks		ImageNet Variants		
			FGSM	PGD	A	R	C(↓)
0.1	Acc.	**84.15**	47.96	22.08	45.75	56.79	32.61
	Rob.	83.89	**48.54**	**24.89**	**46.33**	**57.38**	**31.89**
	Joint	*84.10*	*48.44*	*23.26*	*46.73*	*57.64*	*31.69*
0.3	Acc.	**83.79**	50.42	32.42	42.05	56.17	33.77
	Rob.	83.36	**53.73**	**35.62**	**42.32**	**56.49**	**33.19**
	Joint	*84.03*	*51.85*	*33.84*	*42.45*	*56.72*	*32.91*
0.5	Acc.	**83.38**	53.41	36.58	**38.93**	55.80	35.29
	Rob.	83.01	**56.19**	**39.78**	38.85	**56.12**	**34.73**
	Joint	*83.66*	*54.75*	*37.99*	*39.23*	*56.27*	*34.44*
0.7	Acc.	**80.80**	63.70	49.89	**23.64**	**54.09**	42.27
	Rob.	80.37	**67.37**	**52.23**	23.59	54.04	**42.13**
	Joint	*81.11*	*65.75*	*50.99*	*23.68*	*54.29*	*41.55*
0.9	Acc.	**80.66**	70.02	56.10	**22.69**	**53.64**	42.30
	Rob.	80.04	**74.24**	**58.34**	22.37	53.39	**42.11**
	Joint	*80.34*	*74.19*	*57.50*	*22.21*	*53.67*	*41.56*

Table 2. Performance of different heads and their joint prediction with different λ.

λ	Tuning	FLOPs (G)	Params (M)	GPU Hours	Clean	Attacks		ImageNet Variants		
						FGSM	PGD	A	R	C(↓)
0.1	Head only	17.6	88.1	15.55	80.2	41.1	15.5	22.1	42.0	56.9
	Single adapter	17.8	88.3	15.55	82.5	40.9	15.1	36.9	48.3	46.2
	AdapterFusion	24.9	111.2	19.63	82.2	46.2	22.6	36.4	52.2	35.5
	TORA-ViT	26.0	111.2	19.82	84.1	48.4	23.3	46.5	57.6	31.7
0.9	Head only	17.6	88.1	15.55	79.0	42.0	16.3	12.9	40.2	62.5
	Single adapter	17.8	88.3	15.55	72.3	53.1	30.1	3.1	21.4	78.7
	AdapterFusion	24.9	111.2	19.69	79.5	66.2	55.3	20.4	51.7	42.9
	TORA-ViT	26.0	111.2	19.83	80.3	74.2	57.5	22.2	53.7	41.6

Table 3. Comparison of different tuning methods.

4.3. Classification Heads and Trade-off Ratios

Our TORA-ViT uses two kinds of adapters and tokens to extract different features, and each token corresponds to a corresponding classification head. To decide the final predictions, we use the average outputs of them for joint prediction. To better understand the behaviors of the two kinds of heads, the performance of each head along with the joint prediction with different λ are reported in Table 2.

Firstly, the natural accuracy and robustness against adversarial attacks are well correlated to the trade-off ratio λ. Furthermore, the accuracy head and robustness head also performs consistently on this two kinds of metrics. To be specific, the accuracy head always outperforms the robustness head on clean data, and the robustness always outperforms the accuracy head under attacks.

However, the behaviors change under other perturbations. As aforementioned in Section 4.2, the robustness against other kinds of perturbations is not always positively correlated with robustness against adversarial attacks. When $\lambda < 0.5$, the robust head can still consistently outperforms the accuracy head on all 5 kinds of perturbations. When $\lambda = 0.5$, the accuracy head outperforms robust head on ImageNet-A (natural adversarial examples). When $\lambda > 0.5$, the accuracy head outperforms robust head on ImageNet-A and ImageNet-R (out-of-distribution data). The robustness head is consistently performs better than accuracy head on ImageNet-C (common corruptions).

Overall, we can conclude that when the natural accuracy is high, adversarial training indeed contributes more to the robustness against perturbations other than adversarial attacks. However, if the natural accuracy drops, the contribution of adversarial training to those kinds of robustness also reduces. Therefore, from this point of view, controlling the trade-off between robustness and accuracy is also crucial for the overall robustness against various kinds of perturbations.

4.4. Tuning Methods

As we design a new tuning methods for ViTs, which considers the trade-off between robustness and accuracy via leveraging the robust non-predictive and predictive non-robust features, it is meaningful to compare our method with existing methods that are agnostic to this characteristic of features. In Table 3, we compare our TORA-ViT with three other tuning methods, including tuning a new classification head only, tuning a single new adapter for robustness, and tuning two new adapters with AdapterFusion. Our model have similar number of FLOPs and parameters with AdapterFusion, and our method only requires 0.16 more GPU hours to train, which is approximately 10 minutes. Although the heads only and single adapter tuning are very lightweight, their performance are not as good as our method and AdapterFusion.

In terms of performance, the weakest method is tuning a new head only. Although it is easy for the new head to maintain competitive accuracy, it is hard to improve its robustness. Because the entire model except the head are frozen, the extracted features cannot be changed. It is hard to train a robust classification head on top of non-robust features. When using a single adapter, it's hard to control the trade-off. For exaple, when λ, the natural accuracy of the single adapter drops dramatically to only 72.3%, which is the lowest among all the four methods, but its performance under adversarial attacks is only better than using a new head only. Besides, its performance on the three ImageNet variants are also poor. AdapterFusion is the strongest among the three baselines, but it only has attention at adapter level, which is agnostic of the robust and predictive features. In the contrast, our TORA-ViTs reaches the best performance with the trade-off-aware patches-level attention, which can distinguish robust and predictive features. We will further demonstrate the ability of TORA-ViTs to distinguish the two kinds of features via visualization in Section 4.5.

4.5. Visualization of Attention Maps

We visualize attentions for different adapters in Fig. 3. We extract attention scores after the softmax in the gated fusion. They are $\mathbb{R}^{2\times N\times N}$ tensors, where the first dimension is 2 corresponding to 2 adapters, and the remaining dimensions are N corresponding to the number of tokens (including accuracy/robustness tokens and patch tokens). We average scores for each token to get a $\mathbb{R}^{2\times N}$ matrix. Average scores in all blocks are multiplied to get accumulated

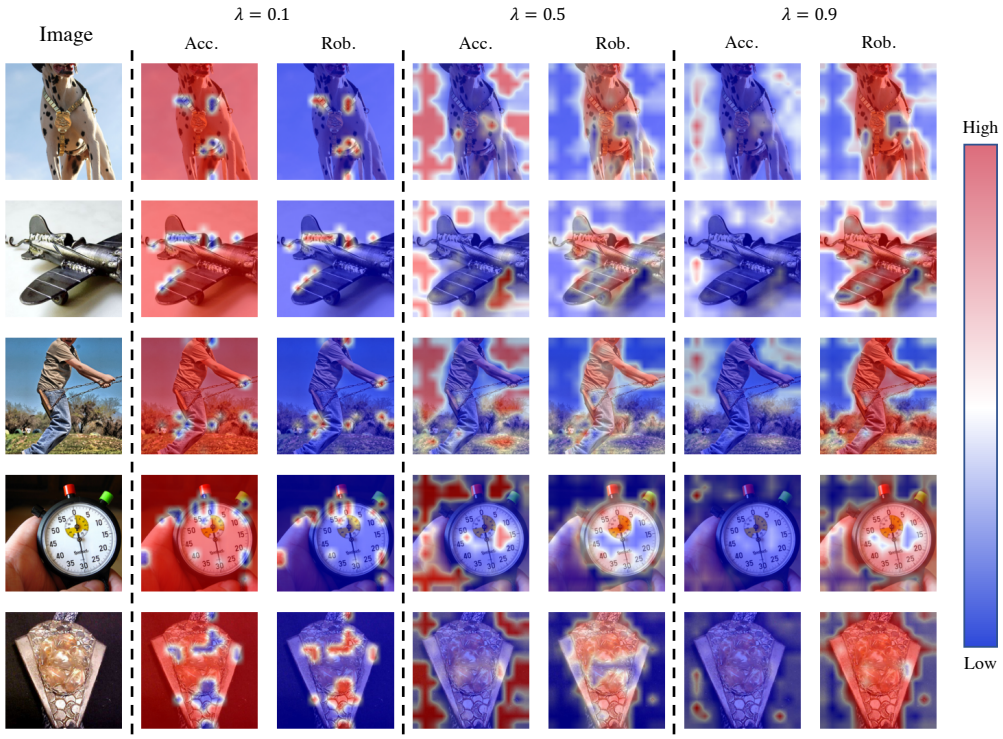


Figure 3. Visualization of the attentions for different adapters in the gated fusion module with various ratio λ. The blue-white-red color map is used, where red represents high attention, and blue represents low attention. As can be seen, the features yielded by the accuracy adapter focus more on **context**, and in the contrast, the features yielded by the robustness adapter focus more on the main **object** to be classified. This is consistent with the theory of robust non-predictive and predictive non-robust features.

attention maps of the entire network.

We find the accuracy adapter focus more on context and the accuracy adapter focus more on the object to be classified. When $\lambda = 0.1$ and the model focuses on accuracy, the features yield by the accuracy adapter have higher attention, and the robustness adapter only have a few highlights in attentions. However, we can still see those highlights are mainly falls in the region of the main object. When $\lambda = 0.9$ and the model focuses on robustness, the features yield by the robustness adapter have higher attentions, and those attentions overlaps the main object. In this case, the accuracy adapter only has a few attentions on the context. If we consider a more balanced trade-off ratio, i.e., $\lambda = 0.5$, we can find this phenomenon is clearer. Attentions of the robustness adapter and the accuracy adapter have a similar amount, but distributed in different regions, and the accuracy adapter focuses more on the context and the robustness adapter focus more on the object.

If we also take into account Table 2, we can find the high attentions on context of the accuracy adapter won't reduce its accuracy. The accuracy on clean data of the accuracy adapter consistently outperforms the robustness adapter, even when $\lambda = 0.9$ and it only have a few attentions on the context. In the contrast, we can find such focuses on context makes it non-robust under adversarial attacks.

5. Conclusion

In this work, we propose Trade-off between Robustness and Accuracy of Vision Transformers (**TORA-ViTs**). TORA-ViTs is inspired by the theory of predictive non-robust and robust non-predictive features. By introducing two different adapters, including an accuracy adapter and a robustness adapter, TORA-ViTs is able to extract both predictive and robust features. To combine the two kinds of features in a trade-off-aware manner, an attention-based gated fusion module is further proposed. It takes the outputs of ViT blocks as queries and utilizes attention mechanism to combine features. Experiments on ImageNet with various robust benchmarks demonstrate that our TORA-ViTs can efficiently improve the robustness of naturally pretrained ViTs while maintaining competitive natural accuracy. Visualization of the attention map in the gated fusion module empirically proves the theory of robust non-predictive features and predictive non-robust features.

Acknowledgments

This work was supported in part by the Australian Research Council under Project DP210101859 and the University of Sydney Research Accelerator (SOAR) Prize.

References

[1] Srinadh Bhojanapalli, Ayan Chakrabarti, Daniel Glasner, Daliang Li, Thomas Unterthiner, and Andreas Veit. Understanding robustness of transformers for image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10231–10241, 2021. 1, 3

[2] Francesco Croce and Matthias Hein. On the interplay of adversarial robustness and architecture components: patches, convolution and attention. *arXiv preprint arXiv:2209.06953*, 2022. 3

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 1, 2, 5

[4] 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. 2

[5] Minjing Dong, Yanxi Li, Yunhe Wang, and Chang Xu. Adversarially robust neural architectures. *arXiv preprint arXiv:2009.00902*, 2020. 3

[6] 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. 1, 2, 3, 4, 5, 6

[7] Stéphane d'Ascoli, Hugo Touvron, Matthew L Leavitt, Ari S Morcos, Giulio Biroli, and Levent Sagun. Convit: Improving vision transformers with soft convolutional inductive biases. In *International Conference on Machine Learning*, pages 2286–2296. PMLR, 2021. 6

[8] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014. 1, 2, 3, 5

[9] Jindong Gu, Volker Tresp, and Yao Qin. Are vision transformers robust to patch perturbations? In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XII*, pages 404–421. Springer, 2022. 1, 3

[10] Minghao Guo, Yuzhe Yang, Rui Xu, Ziwei Liu, and Dahua Lin. When nas meets robustness: In search of robust architectures against adversarial attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 631–640, 2020. 3

[11] Kai Han, An Xiao, Enhua Wu, Jianyuan Guo, Chunjing Xu, and Yunhe Wang. Transformer in transformer. *Advances in Neural Information Processing Systems*, 34:15908–15919, 2021. 1, 3

[12] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022. 2, 6

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1, 6

[14] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8340–8349, 2021. 1, 3, 5, 6

[15] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *Proceedings of the International Conference on Learning Representations*, 2019. 1, 3, 5

[16] Dan Hendrycks, Norman Mu, Ekin D Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. Augmix: A simple data processing method to improve robustness and uncertainty. *arXiv preprint arXiv:1912.02781*, 2019. 6

[17] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. *CVPR*, 2021. 1, 3, 5

[18] Byeongho Heo, Sangdoo Yun, Dongyoon Han, Sanghyuk Chun, Junsuk Choe, and Seong Joon Oh. Rethinking spatial dimensions of vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11936–11945, 2021. 6

[19] Charles Herrmann, Kyle Sargent, Lu Jiang, Ramin Zabih, Huiwen Chang, Ce Liu, Dilip Krishnan, and Deqing Sun. Pyramid adversarial training improves vit performance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13419–13429, 2022. 1, 3, 6

[20] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pages 2790–2799. PMLR, 2019. 2, 4

[21] Han Hu, Zheng Zhang, Zhenda Xie, and Stephen Lin. Local relation networks for image recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3464–3473, 2019. 2

[22] Junho Kim, Byung-Kwan Lee, and Yong Man Ro. Distilling robust and non-robust features in adversarial examples by information bottleneck. *Advances in Neural Information Processing Systems*, 34:17148–17159, 2021. 1, 3

[23] Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551, 1989. 1

[24] Yann LeCun, LD Jackel, Léon Bottou, Corinna Cortes, John S Denker, Harris Drucker, Isabelle Guyon, Urs A Muller, Eduard Sackinger, Patrice Simard, et al. Learning algorithms for classification: A comparison on handwritten digit recognition. *Neural networks: the statistical mechanics perspective*, 261:276, 1995. 1

[25] Yanxi Li, Minjing Dong, Yunhe Wang, and Chang Xu. Neural architecture search via proxy validation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 3

[26] Yanxi Li, Zhaohui Yang, Yunhe Wang, and Chang Xu. Neural architecture dilation for adversarial robustness. *Advances in Neural Information Processing Systems*, 34:29578–29589, 2021. 1, 3

[27] Yingwei Li, Qihang Yu, Mingxing Tan, Jieru Mei, Peng Tang, Wei Shen, Alan Yuille, and Cihang Xie. Shape-texture debiased neural network training. *arXiv preprint arXiv:2010.05981*, 2020. 6

[28] Hanxiao Liu, Karen Simonyan, and Yiming Yang. Darts: Differentiable architecture search. In *International Conference on Learning Representations*, 2018. 3

[29] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 1, 3, 6

[30] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11976–11986, 2022. 3, 6

[31] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *7th International Conference on Learning Representations*, 2019. 5

[32] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017. 1, 2, 3, 5

[33] Kaleel Mahmood, Rigel Mahmood, and Marten Van Dijk. On the robustness of vision transformers to adversarial examples. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7838–7847, 2021. 1, 3

[34] Xiaofeng Mao, Gege Qi, Yuefeng Chen, Xiaodan Li, Ranjie Duan, Shaokai Ye, Yuan He, and Hui Xue. Towards robust vision transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12042–12051, 2022. 1, 3, 5, 6

[35] Muhammad Muzammal Naseer, Kanchana Ranasinghe, Salman H Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Intriguing properties of vision transformers. *Advances in Neural Information Processing Systems*, 34:23296–23308, 2021. 2

[36] Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Lukasz Kaiser, Noam Shazeer, Alexander Ku, and Dustin Tran. Image transformer. In *International conference on machine learning*, pages 4055–4064. PMLR, 2018. 2

[37] Sayak Paul and Pin-Yu Chen. Vision transformers are robust learners. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2071–2081, 2022. 1, 3

[38] Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. Adapterfusion: Non-destructive task composition for transfer learning. *arXiv preprint arXiv:2005.00247*, 2020. 2

[39] Francesco Pinto, Philip HS Torr, and Puneet K. Dokania. An impartial take to the cnn vs transformer robustness contest. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XIII*, pages 466–480. Springer, 2022. 3

[40] Rahul Rade and Seyed-Mohsen Moosavi-Dezfooli. Helper-based adversarial training: Reducing excessive margin to achieve a better accuracy vs. robustness trade-off. In *ICML 2021 Workshop on Adversarial Machine Learning*, 2021. 1

[41] Rahul Rade and Seyed-Mohsen Moosavi-Dezfooli. Reducing excessive margin to achieve a better accuracy vs. robustness trade-off. In *International Conference on Learning Representations*, 2021. 1

[42] Prajit Ramachandran, Niki Parmar, Ashish Vaswani, Irwan Bello, Anselm Levskaya, and Jon Shlens. Stand-alone self-attention in vision models. *Advances in Neural Information Processing Systems*, 32, 2019. 2

[43] Evgenia Rusak, Lukas Schott, Roland S Zimmermann, Julian Bitterwolf, Oliver Bringmann, Matthias Bethge, and Wieland Brendel. A simple way to make neural networks robust against diverse image corruptions. In *European Conference on Computer Vision*, pages 53–69. Springer, 2020. 6

[44] Andreas Steiner, Alexander Kolesnikov, Xiaohua Zhai, Ross Wightman, Jakob Uszkoreit, and Lucas Beyer. How to train your vit? data, augmentation, and regularization in vision transformers. *arXiv preprint arXiv:2106.10270*, 2021. 1, 5, 6

[45] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the IEEE international conference on computer vision*, pages 843–852, 2017. 2

[46] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019. 6

[47] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pages 10347–10357. PMLR, 2021. 1, 2, 6

[48] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. *arXiv preprint arXiv:1805.12152*, 2018. 1, 3

[49] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2

[50] Haotao Wang, Tianlong Chen, Shupeng Gui, TingKuei Hu, Ji Liu, and Zhangyang Wang. Once-for-all adversarial training: In-situ tradeoff between robustness and accuracy for free. *Advances in Neural Information Processing Systems*, 33:7449–7461, 2020. 1

[51] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 568–578, 2021. 6

[52] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep

neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017. 6

[53] Shuo Yang, Tianyu Guo, Yunhe Wang, and Chang Xu. Adversarial robustness through disentangled representations. In *AAAI*, pages 3145–3153, 2021. 1, 3

[54] Xingyi Yang, Zhou Daquan, Songhua Liu, Jingwen Ye, and Xinchao Wang. Deep model reassembly. In *Advances in Neural Information Processing Systems*, 2022. 1

[55] Xingyi Yang, Jingwen Ye, and Xinchao Wang. Factorizing knowledge in neural networks. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXIV*, pages 73–91. Springer, 2022. 1

[56] Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Zi-Hang Jiang, Francis EH Tay, Jiashi Feng, and Shuicheng Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 558–567, 2021. 1, 3, 6

[57] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In *International conference on machine learning*, pages 7472–7482. PMLR, 2019. 1, 2

[58] Richard Zhang. Making convolutional networks shift-invariant again. In *International conference on machine learning*, pages 7324–7334. PMLR, 2019. 6

[59] Hengshuang Zhao, Jiaya Jia, and Vladlen Koltun. Exploring self-attention for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10076–10085, 2020. 2

[60] Daquan Zhou, Zhiding Yu, Enze Xie, Chaowei Xiao, Animashree Anandkumar, Jiashi Feng, and Jose M Alvarez. Understanding the robustness in vision transformers. In *International Conference on Machine Learning*, pages 27378–27394. PMLR, 2022. 1, 3, 6

[61] Barret Zoph, Vijay Vasudevan, Jonathon Shlens, and Quoc V Le. Learning transferable architectures for scalable image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8697–8710, 2018. 3