Adversarial Robustness Comparison of Vision Transformer and MLP-Mixer to CNNs

Philipp Benz
phibenz@gmail.com
Soomin Ham
smham@kaist.ac.kr
Chaoning Zhang
chaoningzhang1990@gmail.com
Adil Karjauv
mikolez@gmail.com
In So Kweon
iskweon77@kaist.ac.kr

Robotics and Computer Vision Lab
Korea Advanced Institute of Science
and Technology (KAIST)
Daejeon, South Korea

*Equal contribution

Abstract

Convolutional Neural Networks (CNNs) have become the de facto gold standard in computer vision applications in the past years. Recently, however, new model architectures have been proposed challenging the status quo. The Vision Transformer (ViT) relies solely on attention modules, while the MLP-Mixer architecture substitutes the self-attention modules with Multi-Layer Perceptrons (MLPs). Despite their great success, CNNs have been widely known to be vulnerable to adversarial attacks, causing serious concerns for security-sensitive applications. Thus, it is critical for the community to know whether the newly proposed ViT and MLP-Mixer are also vulnerable to adversarial attacks. To this end, we empirically evaluate their adversarial robustness under several adversarial attack setups and benchmark them against the widely used CNNs. Overall, we find that the two architectures, especially ViT, are more robust than their CNN models. Using a toy example, we also provide empirical evidence that the lower adversarial robustness of CNNs can be partially attributed to their shift-invariant property. Our frequency analysis suggests that the most robust ViT architectures tend to rely more on low-frequency features compared with CNNs. Additionally, we have an intriguing finding that MLP-Mixer is extremely vulnerable to universal adversarial perturbations. Code: https://github.com/phibenz/robustness_comparison_vit_mlp-mixer_cnn.

1 Introduction

Convolutional Neural Networks (CNNs) [37] have been the gold standard architecture in computer vision. In Natural Language Processing (NLP), however, attention-based transformers are the dominant go-to model architecture [13, 55, 56]. Various attempts have been made to apply such transformer architectures to computer vision tasks [8, 10, 53, 58].
A breakthrough moment was achieved with the advent of the Vision Transformer (ViT) [15], presenting a transformer architecture achieving comparable performance to state-of-the-art CNN architectures. Recently, another MLP-Mixer model architecture [66] which does not rely on convolutions or self-attention, has been presented competing with CNN and ViT. Following [66], MLP-Mixer is termed Mixer in the remainder of this work for simplicity.

Despite the success of CNNs, they are widely known to be vulnerable to adversarial examples [18, 65] whose small additive perturbations of the input cause the CNN to misclassify a sample. This vulnerability causes serious concerns in security-sensitive applications, and thus it is also important to know whether the recently proposed ViT and Mixer are also vulnerable to adversarial attacks. This work sets out to evaluate the adversarial vulnerability of ViT and Mixer architectures and compare their robustness against the CNN models. Therefore, a wide range of adversarial attack methods has been adopted for a comprehensive study. Specifically, first, the performance of the different architectures is compared under the white-box attack, where an adversary has full knowledge of the model parameters to attack. Overall, the two newly proposed architectures, especially ViT, exhibit significantly higher robustness than CNNs against adversarial examples. We further compare their robustness under both query-based and transfer-based black-box attacks. In both cases, we observe a similar trend that ViT is the most robust architecture while CNN is the least robust.

To facilitate the understanding of why CNN is more vulnerable, we design a toy task of binary classification where each class is only represented by a single image. The image from each class has either a vertical or horizontal black stripe in the center. We find that the adversarial example for a CNN exhibits repetitive stripes over the image, while that of an FC network mainly exhibits a single stripe in the center. This observation indicates that the vulnerability of a CNN might be partially attributed to the fact that a CNN, which exploits local connections and shared weights by convolving kernels, is shift-invariant [38, 83]. We also attempt to provide an analysis from the perspective of frequency, investigating whether the different model architectures are biased toward learning more high-frequency or low-frequency features. We find that the ViT seems to learn more low-frequency features, while the CNN is biased towards high-frequency features. Finally, we also investigate their robustness against common corruptions [26] and universal adversarial perturbations [47].

2 Related Work

Beyond CNNs for vision applications. In Natural Language Processing (NLP), transformers [70], solely based on the attention mechanisms, are the predominant model architecture [13, 55, 56]. In contrast, CNNs have been the de facto standard in deep learning for vision applications, while the application of transformers to vision tasks is an emerging trend [8, 10, 53, 58]. The Vision Transformer (ViT) [15] was recently introduced, demonstrating that transformers can achieve state-of-the-art performance, by sequencing the images into patches and pre-training the model on large amounts of data. To address the data issue, DeiT [67] introduced a teacher-student strategy specific to transformers and trained a transformer architecture only on the ImageNet-1K dataset. Concurrently, the T2T-ViT had been proposed [78] introducing an advanced Tokens-to-Tokens strategy. Further works are attempting to extend the ViT architecture to increase the efficiency and performance of transformer architectures [11, 22, 41, 74]. ViTs have further been explored beyond the task of image classification [7, 25, 34, 51, 71]. Tolstikhin et al. [66] challenge the status-quo of convolutions and attention in current computer vision models and proposes MLP-Mixer, a
pure Multi-Layer Perceptron (MLP)-based architecture.

**Adversarial attacks and robustness.** CNNs are known to be vulnerable to adversarial examples \([18, 36, 65]\). Depending on the accessibility to the target model, adversarial attacks can be divided into white box ones \([6, 18, 42, 46]\) that require full access to the target model, query-based black-box attacks \([9, 19, 30, 31, 52, 57, 62, 69]\), and transfer-based black-box attacks \([14, 21, 33, 40, 68, 73, 75]\). Adversarial attacks can be divided into image-dependent ones \([6, 18, 42, 46, 60]\) and universal ones \([3, 47, 48, 79, 80, 82]\). Specifically, contrary to image-dependent attacks, a single perturbation, \(i.e\). universal adversarial perturbation (UAP) exists to fool the model for most images \([82]\). The vulnerability of transformers in the context of NLP has also been investigated \([16, 23, 27, 29, 35, 39, 63]\). This work focuses on the empirical robustness evaluation of CNN, ViT, and Mixer in image classification.

**Concurrent works.** Recently, also other works \([2, 4, 20, 44, 45, 49, 50, 54, 61, 72]\) investigated transformers from the perspective of adversarial robustness. Specifically, \([2, 4, 44, 49, 54, 61]\) concurrently compare the robustness of transformers to CNNs and independently derive conclusions resembling each other. Their main conclusions, ignoring their nuance difference, can be summarized as *vision transformers are more robust than CNNs*. Not surprisingly, our work also comes to the same main take-away message but differs in multiple aspects, such as co-analysis of MLP-Mixer, perturbation-minimization (C&W and DeepFool) results in white-box setting, evaluation under universal attack. Our works also leads to some additional insight, such that MLP-Mixer shows an increased vulnerability to universal attacks. In another parallel line, \([45, 50]\) have investigated how to improve the adversarial robustness of vision transformers.

## 3 Research Goal and Experimental Setup

**Research goal and scope.** CNNs have achieved dominant success in numerous vision applications in the last few years, however, they are also vulnerable to adversarial attacks. Such vulnerability causes serious concern in security-sensitive applications, such as autonomous driving. This concern has motivated an extensive study on model robustness against various attack methods. With the recent popularity of ViT and Mixer as alternatives to CNNs, it is vital for the community to understand their adversarial robustness and to benchmark them against the widely used CNNs. To this end, this work *empirically* investigates the adversarial robustness of the three architectures. In other words, this work has no intention to understand the reason behind why a certain architecture is more or less robust. Note that there is still no consensus \([1]\) on the explanation of CNN being sensitive to adversarial examples despite a large body of works in this field. As an early attempt to investigate the adversarial robustness of ViT and Mixer, our work focuses on the empirical evaluation and it is out of the scope of this work to theoretically understand why they might be vulnerable. Nonetheless, our work attempts to provide better understandings of the robustness gap between models from a shift-invariance perspective and a frequency perspective. Admittedly, our attempt for explanation is limited and future work is needed for better understanding.

**Models and dataset.** In our experiments, we compare the ViT \([15]\) models, MLP-Mixer \([66]\) and CNN architectures \([24]\). Note that they all adopt shortcut \([24]\) in their architecture design. For the ViT models, we consider ViT-B/16 and ViT-L/16, where B and L stand for “base” and “large”, respectively, while 16 indicates the patch size. The considered ViT models were pre-trained on ImageNet-21K and fine-tuned on ImageNet-1K \([64]\). We also evaluate ViT models that are directly trained on ImageNet-1K from \([64]\) (indicated by
“ImageNet-1K”). Correspondingly, the ViT models, we also investigated Mixer-B/16 and Mixer-L/16 [32] which are trained on ImageNet-1K. We further consider CNN architectures, ResNet-18 and ResNet-50 [32] trained on ImageNet-1K as well as the semi-weakly supervised (SWSL) variant [27], which is pre-trained on IG-1B-Targeted [33] with associated hashtags from 1,000 ImageNet-1K classes followed by fine-tuning on ImageNet-1K dataset. To evaluate adversarial attacks, if not otherwise mentioned we evaluate different adversarial attacks in the untargeted setting on an ImageNet-compatible dataset. This dataset was originally introduced in the NeurIPS 2017 adversarial challenge1. We compare different architectures for the corresponding most widely used models, such as ResNet-18/50 for CNN, and ViT and Mixer for B/16 and L/16. However, we also note that many other factors other than the architecture itself might also have a role in influencing the robustness. Given the publicly available models, it is nearly impossible to rule out all other factors.

4 Experiment Results

4.1 Robustness Against White-Box Attacks

We first investigate the robustness under white-box attacks. Particularly, we deploy PGD [32] and FGSM [18]. For both attacks we consider $\varepsilon = \{d/255 \mid d \in \{0.1, 0.3, 0.5, 1, 3\}\}$ for images in range $[0, 1]$. For the PGD attack, we set the number of iterations to 20 and keep the other parameters as the default settings of Foolbox [33]. For these two attacks, we report the attack success rate (ASR), i.e., the percentage of samples which were classified differently from the ground-truth class. Additionally, we evaluate the models on the $\ell_2$-variants of the C&W attack [16] and DeepFool [44]. These two attacks have the objective to minimize the perturbation magnitude given the ASR of 100%. Hence, we report the $\ell_2$-norm of the adversarial perturbation and the results are available in Table 1. Overall a trend can be observed that compared with CNN architectures, the ViT and Mixer models have a lower attack success rate, suggesting they are more robust than CNN architectures. The increased robustness of ViT and Mixer models is further supported by a higher $\ell_2$-norm for the C&W and DeepFool attacks. However, when the perturbation magnitude is very small, the opposite phenomenon can be observed. For example for Mixer-L/16 for the PGD or FGSM with $\varepsilon = 0.1$ the model can exhibit decreased robustness compared to the CNN models.

<table>
<thead>
<tr>
<th>Model</th>
<th>ImageNet</th>
<th>NeurIPS</th>
<th>PGD ($\ell_2$)</th>
<th>FGSM ($\ell_2$)</th>
<th>C&amp;W ($\ell_2$)</th>
<th>DeepFool ($\ell_2$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ViT-B/16</td>
<td>81.4</td>
<td>54.7</td>
<td>22.7</td>
<td>63.6</td>
<td>85.5</td>
<td>97.5</td>
</tr>
<tr>
<td>ViT-L/16</td>
<td>82.9</td>
<td>89.3</td>
<td>22.8</td>
<td>60.1</td>
<td>80.9</td>
<td>95.8</td>
</tr>
<tr>
<td>ViT-B/16 (ImageNet-1K)</td>
<td>76.7</td>
<td>86.6</td>
<td>29.2</td>
<td>68.3</td>
<td>87.1</td>
<td>96.8</td>
</tr>
<tr>
<td>ViT-L/16 (ImageNet-1K)</td>
<td>72.8</td>
<td>79.2</td>
<td>44.9</td>
<td>77.0</td>
<td>90.6</td>
<td>97.1</td>
</tr>
<tr>
<td>ResNet-18/16</td>
<td>67.5</td>
<td>86.2</td>
<td>29.5</td>
<td>63.4</td>
<td>82.0</td>
<td>96.2</td>
</tr>
<tr>
<td>ResNet-50/16</td>
<td>71.8</td>
<td>80.0</td>
<td>41.1</td>
<td>67.3</td>
<td>80.4</td>
<td>92.1</td>
</tr>
<tr>
<td>ResNet-18/SWSL</td>
<td>73.3</td>
<td>90.4</td>
<td>47.9</td>
<td>93.7</td>
<td>98.7</td>
<td>99.5</td>
</tr>
<tr>
<td>ResNet-50/SWSL</td>
<td>81.2</td>
<td>96.3</td>
<td>39.4</td>
<td>90.2</td>
<td>97.0</td>
<td>98.4</td>
</tr>
<tr>
<td>ResNet-18</td>
<td>69.8</td>
<td>83.7</td>
<td>46.1</td>
<td>90.0</td>
<td>97.8</td>
<td>99.9</td>
</tr>
<tr>
<td>ResNet-50</td>
<td>76.1</td>
<td>93.0</td>
<td>55.8</td>
<td>86.3</td>
<td>97.9</td>
<td>99.5</td>
</tr>
</tbody>
</table>

Table 1: White-box attacks on benchmark models with different epsilons. We report the clean accuracy on the ImageNet and NeurIPS dataset. The attack success rate (%) of PGD and FGSM under $\ell_\infty$ distortion, and the $\ell_2$-norm of C&W and DeepFool, respectively are reported for the NeurIPS dataset. All models are trained with an image size of 224. A model with a lower ASR↓ or higher $\ell_2$-norm↑ is considered to be more robust.

1https://github.com/rwightman/pytorch-nips2017-adversarial
Class-wise Robustness. To provide a more detailed robustness evaluation, we perform a class-wise robustness study. We perform the class-wise robustness study on the ImageNet validation dataset, where for each class 50 validation images are present and attack the models with the $\ell_\infty$-PGD attack ($\epsilon = 0.3$). Figure 1 (left) shows every 50th class from the most robust classes to the least robust classes for ViT-L/16, where the class “screen” exhibits no robustness, while the class “yellow lady’s slipper” exhibits robustness of 100%. This indicates an imbalance in the class-wise robustness between different classes. We further calculate the similarity of the class-wise accuracies between different models. We treat the different class-wise accuracies as a vector and calculate the cosine similarity between the class-wise accuracies of different models. From Figure 1 (center), it can be observed that the ViT and Mixer models exhibit relatively high similarity values, but also that ResNet18 and ResNet50 are similar in their class-wise robust accuracies. Finally, we examine the relative class-wise robustness of the models, by calculating how many classes of one model are more robust than those of another model. These results are presented in Figure 1 (right). Here it can be observed that the ViT and Mixer models exhibit higher class-wise accuracies than the CNNs, with consistently more than 945 classes being more robust than the examined CNNs.

4.2 Robustness Against Black-Box Attacks

We evaluate and compare the robustness of different model architectures for the black-box attacks in two setups: query-based black-box attack and transfer-based black-box attack.

<table>
<thead>
<tr>
<th>Boundary ($\ell_2$)</th>
<th>ViT-B</th>
<th>ViT-L</th>
<th>Mixer-B</th>
<th>Mixer-L</th>
<th>RN18 (SWSL)</th>
<th>RN50 (SWSL)</th>
<th>RN18</th>
<th>RN50</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.980</td>
<td>7.408</td>
<td>1.968</td>
<td>1.951</td>
<td>1.403</td>
<td>1.846</td>
<td>1.468</td>
<td>1.740</td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Boundary attack on benchmark models. We use 100 randomly selected images from NeurIPS dataset against decision-based attack, and the $\ell_2$-norm of adversarial perturbation is presented.

Query-based black-box attacks. Query-based black-box attacks work by evaluating a sequence of perturbed images through the model. We adopt a popular decision-based attack, the Boundary Attack [3], which only requires a model’s final decision (i.e., class label) and aims to minimize the perturbation while remaining adversarial. As with the white-box attack, a trend can be observed in the black-box attack that the ViT and Mixer models are more robust, indicated by the relatively higher $\ell_2$-norm of the adversarial perturbation (see Table 2).
We further test and compare the models using a more recently proposed approach \cite{31} that reduces the query cost. To achieve a higher success rate with fewer queries, Bandits$_{TD}$ integrates the prior information about the gradient (i.e., time and data) using bandit algorithm to reduce the query number. Following the setting in \cite{31}, the maximum $\ell_\infty$-norm of perturbations is set to 0.05 and the others are set to default as well. As shown in Table 3, overall we observe that ViT and Mixer require a larger average number of queries with a lower average ASR, suggesting ViT and Mixer are more robust than their CNN counterparts.

**Transfer-based black-box attacks.** Transfer-based black-box attacks exploit the transferable property of adversarial examples, i.e., the adversarial examples generated on a source model transfer to another unseen target model. For the source model, we deploy the I-FGSM \cite{36} attack with 7 steps and evaluate the transferability on the target model. From the result in Table 4, we have two major observations. First, adversarial examples from the same family (or similar structure) exhibit higher transferability, suggesting models from the same family learn similar features. Second, when a different model architecture is used as the source model, there is also a trend that CNNs are relatively more vulnerable (i.e., transfer poorly toward foreign architectures). For example, the transferability from CNN to ViT is often lower than 20%, while the opposite scenario is higher.

### Table 3: Results of Bandits$_{TD}$ (with time and data-dependent priors) attack on benchmark models.

ASR is the attack success rate and #Q denotes the average number of queries of the successful attacks. We test on the total NeurIPS dataset with a query limit of 10,000.

### Table 4: Transfer-based black-box attacks on benchmark models.

We report the attack success rate (%) and a model with a lower ASR is considered to be more robust. All models are trained with an image size of 224, and attacked with a maximum $\ell_\infty$ perturbation of $\varepsilon = 16$.

#### 4.3 Toy Example

ViT and Mixer are more robust to adversarial attacks than conventional CNNs. In other words, CNN tends to be the least robust in most setups. To facilitate the understanding of the mechanisms, we design a toy example of binary classification where each class is represented by a single image with a size of 224. The two images shown in Figure 2 consist of a single black stripe on a grey background, differing in the stripe orientation, namely a vertical and a horizontal stripe.
We train a Fully Connected network (FC), Convolution Neural Network (CNN), and Vision Transformer (ViT) on the images. Note that we designed the networks to be of relatively similar small capacity (< 5M). We evaluate the adversarial robustness of these models with the commonly used $\ell_2$ attacks C&W [6] and DDN [60]. We report the $\ell_2$-norm of the adversarial perturbation in Table 5. We observe that the CNN is also less robust than the FC and the ViT in this toy example setup.

<table>
<thead>
<tr>
<th></th>
<th>C&amp;W ($\ell_2$)</th>
<th>DDN ($\ell_2$)</th>
<th>#params</th>
</tr>
</thead>
<tbody>
<tr>
<td>CNN</td>
<td>25.06</td>
<td>25.39</td>
<td>4.82M</td>
</tr>
<tr>
<td>FC</td>
<td>27.82</td>
<td>59.99</td>
<td>4.59M</td>
</tr>
<tr>
<td>ViT</td>
<td>59.99</td>
<td>4.88M</td>
<td></td>
</tr>
</tbody>
</table>

Table 5: $\ell_2$-norm of the adversarial perturbation on our toy example.

Figure 3: Adversarial examples and perturbations generated with C&W (top) and PGD (bottom) using different architectures trained on our toy example.

**Explanation from a shift-invariance perspective.** Recently, [17] has shown that the shift-invariance property of CNNs can be one cause for its vulnerability to adversarial attacks. Their conclusion is derived with the quantitative analysis and theoretical proof, while we focus on an intuitive qualitative analysis. The qualitative results of adversarial perturbations generated by the attacks are shown in Figure 3. For the ViT, one phenomenon can be observed that the adversarial perturbation consists of square patches. This is likely due to the division of the input image into patches in the ViT architecture. Without this split process on the image, we observe clear stripes but with different patterns for CNN and FC. While the CNN model generates perturbations with repeated stripes, the FC model generates perturbations with only a single stripe centered in the image. While the CNN model generates perturbations with repeated stripes, the FC model generates perturbations with only a single stripe centered in the image. It should be noted that perturbations are generated toward the adversary, i.e., in the direction of the opposite class’ stripe. The observation that the CNN model yields stripes all over the image can be attributed to the shift-invariant property of the CNN model. From the perspective of shift-invariance, the CNN model recognizes features, i.e., horizontal or vertical stripe in this setup, regardless of the position of the features on the image. Thus, it is somewhat expected that the perturbation has stripes in a different direction all over the image. For the FC model without the shift-invariant property, it only recognizes the stripes in the center; thus, the resulting perturbation mainly has the stripe in the center. Since our toy example only consists of two samples, the ASR has only limited informative value about the model’s robustness. However, the qualitative results can still be observed. For the $\ell_2$-PGD attack, we choose a sufficiently high $\varepsilon$ of 40, such that both samples are misclassified. The qualitative results of the PGD attack in
Figure 3 (bottom) resemble those of the C&W attack. These qualitative results provide an interesting insight into the possible link between shift-invariant property and CNN vulnerability. Admittedly, this link is vague and future work is needed to establish a more concrete link between them.

4.4 Frequency Analysis

We further attempt to explain the lower robustness of CNN from a frequency perspective [77, 81]. We apply a low-pass filter and a high-pass filter to the input images before feeding them to the model. We then evaluate the Top-1 accuracy of images from the NeurIPS dataset by applying low-pass or high-pass filtering, and the results are shown in Figure 4. For the low-pass filtering, a sharper decline of the CNN architectures can be observed than for the ViT and Mixer, indicating that the CNN architectures are more reliant on the high-frequency features compared to other models. For example, ResNet-50 (SWSL) has higher accuracy than ViT-L/16 when filtered with large bandwidths; however, with smaller bandwidths, the accuracy of ViT-L/16 becomes higher than the ResNet. Additionally, the Mixers have a sharper decline than the ViTs at low-pass bandwidths from 60 to 180. For the high-pass filtering, the ViT models show the steepest decline among the models, indicating that the ViT models rely more on low-frequency features. Note that non-robust features tend to have high-frequency properties [32, 77, 81] and explain the decreased model robustness. This demonstrates why ViT models are more robust than CNN architectures from the frequency perspective. Comparing the results from low-pass and high-pass filtering, we observe that Mixers, regardless of their absolute value of accuracy, exhibit a similar trend to CNNs rather than ViTs.

Figure 4: Top-1 accuracy across a range of frequency bandwidths from low/high-pass filtering (The right direction of x-axis is more oriented to the low and high frequency, respectively). Left: Low-pass filtering. Right: High-pass filtering.

5 Additional Investigations

5.1 Robustness against common corruptions

Additional to our investigation of the adversarial vulnerability of ViT, Mixer, and CNN models, we examine the robustness of these models to common, i.e. naturally occurring, corruptions [26]. ImageNet-C was proposed by [26] benchmarking neural network robustness
against these common corruptions. In essence, ImageNet-C is a perturbed version of the original ImageNet validation dataset that has 1000 classes and 50 images per class. ImageNet-C has 15 test corruptions, and four hold-out corruptions, each consisting of 5 severities.

Following [28], we evaluate on 15 test corruptions and the results are shown in Table 6. First, there is a clear trend that models that have higher accuracy on the original (clean) ImageNet tend to also have higher accuracy on ImageNet-C, which is somewhat expected. Second, with comparable accuracy on the original (clean) ImageNet, ViT, and Mixer architectures tend to have higher robustness against corruptions. For example, ViT-B/16 has similar accuracy as RN50(SWSL), i.e. 81.43% vs. 81.18%, but the robustness of ViT-B/16 is notably higher than that of RN50(SWSL), i.e. 58.85% vs. 52.03%. A similar phenomenon can be observed for Mixer-B/16 and RN50. More detailed corruption-wise results are shown in Figure 5. We find that under some corruptions, such as zoom blur and snow, the superiority of ViT is more significant than other corruptions, such as Gaussian noise.

Table 6: Evaluation of benchmark models on the ImageNet-C dataset. Accuracy the higher the better, mCE the lower the better.

<table>
<thead>
<tr>
<th>Model</th>
<th>ImageNet</th>
<th>ImageNet-C</th>
<th>mCE</th>
</tr>
</thead>
<tbody>
<tr>
<td>ViT-B/16</td>
<td>81.43</td>
<td>58.85</td>
<td>51.98</td>
</tr>
<tr>
<td>ViT-L/16</td>
<td>82.89</td>
<td>64.11</td>
<td>45.46</td>
</tr>
<tr>
<td>Mixer-B/16</td>
<td>76.47</td>
<td>47.00</td>
<td>67.35</td>
</tr>
<tr>
<td>Mixer-L/16</td>
<td>71.77</td>
<td>40.47</td>
<td>75.84</td>
</tr>
<tr>
<td>RN18 (SWSL)</td>
<td>73.29</td>
<td>37.84</td>
<td>78.30</td>
</tr>
<tr>
<td>RN50 (SWSL)</td>
<td>81.18</td>
<td>52.03</td>
<td>60.49</td>
</tr>
<tr>
<td>RN18</td>
<td>69.76</td>
<td>32.92</td>
<td>84.67</td>
</tr>
<tr>
<td>RN50</td>
<td>76.13</td>
<td>39.17</td>
<td>76.70</td>
</tr>
</tbody>
</table>

5.2 Robustness against Universal Adversarial Perturbations

We further compare the robustness of the different model architectures against Universal Adversarial Perturbations (UAPs). UAPs have been proposed in [47], however, their algorithm has been identified to be relatively less effective yet slow. In this work, we adopt the state-of-the-art UAP algorithm in [48]. Following [47, 80], we set the image size to 224 and constrain the UAP with an $l_\infty$-norm of $\varepsilon = 10/255$. The white-box and the corresponding transferability results are shown in Table 7. Several observations can be made. First, MLP-Mixer models are very vulnerable to UAPs in both white-box and black-box scenarios. Especially in the black-box scenario, the attack success rate is always higher than 95% regardless of the surrogate model architecture. On the contrary, ViT models and CNN models are more robust against UAPs, particularly in the black-box scenarios. Second, comparing ViT and CNN models, there is no obvious robustness gap with an ASR of around 90% for all models. However, in the more challenging black-box scenarios, the ViT models are more robust than their CNN counterparts. This trend aligns well with our previous finding that ViT models are more robust than CNN models. We visualize amplified versions of the resulting UAPs in Figure 6. It is noticeable that for the ViT and Mixer architectures, a tile pattern is observable, which is caused by the operation of dividing the images into patches as tokens. Another interesting observation is that the UAPs of Mixer tend to be less locally smooth.
than those generated on ViT and CNN. The ViT-L/16 with the highest robustness against UAP seems to have the most locally smooth patterns. Overall, it is interesting that Mixer is extremely vulnerable to UAPs and the finding from a qualitative result that they have locally non-smooth patterns. However, we have no clear explanation for the observed phenomenon since the Mixer is still a very recent architecture. An important message to the community is that the adversarial threat to Mixer cannot be ignored since it is vulnerable in the practical transferable universal attack scenario.

Table 7: Universal adversarial attacks on benchmark models. We report the attack success rate (%) and a model with a lower ASR is considered to be more robust. All models are trained with an image size of 224, and evaluated on the NeurIPS dataset. The bold font indicates white-box attacks.

<table>
<thead>
<tr>
<th>Source model</th>
<th>ViT-B/16</th>
<th>ViT-L/16</th>
<th>Mixer-B/16</th>
<th>Mixer-L/16</th>
<th>RN18 (SWSL)</th>
<th>RN50 (SWSL)</th>
<th>RN18</th>
<th>RN50</th>
</tr>
</thead>
<tbody>
<tr>
<td>ViT-B/16</td>
<td>93.9</td>
<td>32.7</td>
<td>95.7</td>
<td>96.0</td>
<td>43.3</td>
<td>29.2</td>
<td>46.4</td>
<td>37.2</td>
</tr>
<tr>
<td>ViT-L/16</td>
<td>42.1</td>
<td><strong>84.8</strong></td>
<td>95.6</td>
<td>95.6</td>
<td>39.0</td>
<td>21.2</td>
<td>41.6</td>
<td>31.7</td>
</tr>
<tr>
<td>Mixer-B/16</td>
<td>16.6</td>
<td>14.9</td>
<td><strong>98.6</strong></td>
<td>96.7</td>
<td>35.4</td>
<td>19.6</td>
<td>43.3</td>
<td>30.6</td>
</tr>
<tr>
<td>Mixer-L/16</td>
<td>15.2</td>
<td>13.7</td>
<td>96.3</td>
<td><strong>98.9</strong></td>
<td>30.3</td>
<td>15.5</td>
<td>38.7</td>
<td>26.0</td>
</tr>
<tr>
<td>RN18 (SWSL)</td>
<td>12.6</td>
<td>12.1</td>
<td>95.4</td>
<td>95.6</td>
<td><strong>91.2</strong></td>
<td>49.2</td>
<td>56.6</td>
<td>56.5</td>
</tr>
<tr>
<td>RN50 (SWSL)</td>
<td>11.8</td>
<td>11.9</td>
<td>95.3</td>
<td>95.6</td>
<td>50.4</td>
<td><strong>87.1</strong></td>
<td>46.8</td>
<td>46.6</td>
</tr>
<tr>
<td>RN18</td>
<td>13.5</td>
<td>11.6</td>
<td>95.7</td>
<td>95.7</td>
<td>66.2</td>
<td>46.1</td>
<td><strong>93.7</strong></td>
<td>63.3</td>
</tr>
<tr>
<td>RN50</td>
<td>12.2</td>
<td>12.0</td>
<td>95.4</td>
<td>95.6</td>
<td>63.3</td>
<td>57.0</td>
<td>63.4</td>
<td><strong>91.9</strong></td>
</tr>
</tbody>
</table>

Figure 6: Visualization of the amplified UAPs generated on different model architectures.

6 Conclusion

Our work performs an empirical study on the adversarial robustness comparison of ViT and MLP-Mixer to the widely used CNN on image classification. Our results show that ViT is significantly more robust than CNN in a wide range of white-box attacks. A similar trend is also observed in the query-based and transfer-based black-box attacks. Our toy task of classifying two simple images with a vertical or horizontal black stripe in the image center provides interesting insight on the possible link between shift-invariant property and CNN vulnerability and future work is necessary for further investigating this link. Our analysis from the feature perspective further suggests that ViTs are more reliant on low-frequency (robust) features while CNNs are more sensitive to high-frequency features. We also investigate the robustness of the newly proposed MLP-Mixer and find that its robustness generally locates in the middle of ViT and CNN. We have also performed additional investigations on robustness against common corruptions and UAPs. One very intriguing finding is that Mixer is extremely vulnerable to UAPs, even in transfer-based black-box attacks. Future work is needed for a better understanding of the reported empirical results.
References


[32] Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. In NeurIPS, 2019.


