Robustness Comparison of Vision Transformer and MLP-Mixer to CNNs

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

Korea Advanced Institute of Science and Technology (KAIST)

Abstract

Convolutional Neural Networks (CNNs) have become the de facto gold standard in computer vision applications for several years. However, new model architectures have recently been proposed challenging the status quo. The Vision Transformer (ViT) relies solely on attention modules, while the Mixer architecture substitutes the self-attention modules with Multi-Layer Perceptrons (MLPs). Despite their great success, CNNs have been shown vulnerable to adversarial examples. This work sets out to investigate the adversarial vulnerability of the recently introduced ViT and MLP-Mixer architectures and compare their performance with CNNs. Our results on white-box and black-box attacks suggest that ViT and MLP-Mixer architectures are more robust to adversarial examples. Using a toy example, we also provide empirical evidence that the lower adversarial robustness of CNNs can be attributed to their shift-invariant property. With a frequency study, we further analyze the distribution of frequencies learned from different model architectures.

1. Introduction

Convolutional Neural Networks (CNNs) [24] 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 [10, 34, 35]. Various attempts have been made to apply such transformer architectures to computer vision tasks [7, 33, 36, 6]. A breakthrough moment was achieved with the advent of the Vision Transformer (ViT) [11], presenting a transformer architecture achieving comparable performance to state-of-the-art CNN architectures. Recently, another alternative model architecture has been presented competing with CNN and ViT. The MLP-Mixer architecture, which does not rely on convolutions or self-attention, has been proposed in [43].

Despite the success of CNNs, they remain vulnerable to adversarial perturbations [41, 13], small input perturbations causing the CNN to misclassify a sample. Due to the rather recent introduction of the ViT and Mixer architecture, the adversarial vulnerability of these novel architectures has not been well studied yet. This work sets out to explore and analyze the adversarial vulnerability of ViT and Mixer architectures and compare the findings against the CNN models. Therefore, previously proven attacks on CNN architectures are used. 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. We find that overall, ViT and Mixer (especially ViT) architectures exhibit greater robustness against adversarial examples than CNNs. We further compare their robustness under both query-based and transfer-based black-box attacks. In both cases, we observe the same trend that among the three explored architectures, 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 middle. We find that CNN yields adversarial stripes all over the images, while an FC network mainly attacks the stripe in the middle. This observation indicates that the vulnerability of CNN can be partially attributed to the fact that CNN, which exploits local connections and shared weights by convolving kernels, has a shift-invariance [56, 25]. Finally, we 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. The high-frequency and low-frequency features are
commonly considered to be more non-robust and robust, respectively [47]; therefore, ViT which is more reliant on the robust (low-frequency) features, tends to be more robust.

2. Related Work

Vision Transformers. In Natural Language Processing (NLP) Transformers [45], which are solely based on the attention mechanisms, are the predominant model architecture [10, 34, 35]. While CNNs have been the de facto standard in deep learning for computer vision, also the application of transformers has been explored for vision tasks [7, 33, 36, 6]. Recently, the Vision Transformer (ViT) [11] was introduced, demonstrating that transformers can achieve state-of-the-art performance. ViTs have further been explored beyond the task of image classification [46, 5, 21, 32, 17].

MLP-Mixer. Tolstikhin et al. [43] challenge the status quo of convolutions and attention in the current computer vision models and proposes MLP-Mixer, a pure Multi-Layer Perceptron (MLP)-based architecture. Pre-trained on large datasets, MLP-Mixer achieves comparable performance with ViT. The main idea behind the Mixer architecture is to separate the per-location operations and cross-location operations, which are both realized through MLPs. Additionally, the Mixer architecture relies on several advances in CNNs over the past years, such as skip-connections [16], dropout [40], layer norm [1], etc.

Robustness. CNNs are commonly known to be vulnerable to adversarial examples [41, 13, 23], which has prompted numerous studies on both image-dependent [13, 30, 4, 28, 38] and universal attacks [29, 31, 52, 53, 2, 55]. The vulnerability of transformers in the context of NLP tasks has also been investigated [19, 22, 39, 18, 26, 12, 15]. In this work we set out to investigate and compare the ViT and Mixer architectures from an adversarial robustness standpoint with existing attack methods with a comparison against CNNs.

3. Methodology

Models and Dataset. In our experiments, we mainly compare the ViT [11] models, MLP-Mixer [43] and ResNet architectures [16]. 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 [9]. Corresponding to the ViT models, we also investigated Mixer-B/16 and Mixer-L/16 [43], except that these models were directly trained on the ImageNet-1K without additional pre-training. We further consider the ResNet-18 and ResNet-50 [16] architectures trained on ImageNet-1K as well as the semi-supervised (SSL) variant [49], which is pre-trained on a subset of unlabeled YFCC100M [42] public image dataset and fine-tuned with the ImageNet-1K, and the semi-weakly supervised (SWSL) variant [49] which are pre-trained on 940 million public images with 1.5K hashtags matching with 1,000 ImageNet-1K synsets, followed by fine-tuning on ImageNet-1K dataset.

To evaluate adversarial attacks, we evaluate different adversarial attacks in the untargeted setting on an ImageNet-compatible dataset (composed of 1,000 images in 430 classes). This dataset was originally introduced in the NeurIPS 2017 adversarial challenge\(^1\).

4. Experiment Results

4.1. Robustness Against White-Box Attacks

We first investigate the robustness under white-box attacks. Particularly, we deploy PGD [28] and FGSM [13]. For both attacks we consider \(\epsilon = \{d/255 \mid d \in \{0.1, 0.3, 0.5, 1, 3\}\} \) 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 [37]. For these two attacks, we report the attack success rate (ASR), meaning 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 [4] and DeepFool [30]. 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 architecture, the ViT and Mixer models have a lower attack success rate, suggesting they are more robust than CNN architectures. This is further confirmed by finding that CNN requires a relatively lower \(\ell_2\)-norm for the C&W and DeepFool attacks. One exception to this observation is the Mixer model, which appears to exhibit increased vulnerability to very small perturbations, being as vulnerable as the CNN models for an \(\epsilon = 0.1\).

4.2. Robustness Against Black-Box Attacks

For the black-box attacks, we evaluate and compare their robustness in two common setups: query-based black-box

\(^1\)https://github.com/rwightman/pytorch-nips2017-adversarial
Table 1: White-box attacks on benchmark models with different epsilons. We report the clean accuracy on 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. All models were trained with an image size of 224, and a model with a lower ASR or higher $\ell_2$-norm is considered to be more robust.

<table>
<thead>
<tr>
<th>Model</th>
<th>Clean ImageNet</th>
<th>NeurIPS</th>
<th>PGD ($\ell_\infty$)</th>
<th>FGSM ($\ell_\infty$)</th>
<th>C&amp;W ($\ell_2$)</th>
<th>DeepFool ($\ell_2$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ViT-B/16</td>
<td>18.1</td>
<td>90.7</td>
<td>22.6 63.6 86.5 97.5 99.9</td>
<td>19.1 38.7 52.8 66.3 79.7</td>
<td>0.468</td>
<td>0.425</td>
</tr>
<tr>
<td>ViT-L/16</td>
<td>22.9</td>
<td>93.2</td>
<td>23.6 60.1 80.9 95.8 100</td>
<td>19.5 35.9 44.9 57.9 67.3</td>
<td>0.459</td>
<td>0.548</td>
</tr>
<tr>
<td>Mixer-B/16</td>
<td>75.1</td>
<td>86.2</td>
<td>29.5 63.4 82.0 96.2 100</td>
<td>27.7 49.3 59.5 69.3 78.0</td>
<td>0.375</td>
<td>0.339</td>
</tr>
<tr>
<td>Mixer-L/16</td>
<td>76.6</td>
<td>81.0</td>
<td>41.1 67.3 80.4 92.1 99.4</td>
<td>36.7 51.8 56.9 61.6 67.4</td>
<td>0.297</td>
<td>0.377</td>
</tr>
<tr>
<td>ResNet-18 (SWSL)</td>
<td>75.3</td>
<td>90.4</td>
<td>47.9 93.7 98.7 99.5 99.9</td>
<td>38.0 76.5 89.9 96.2 97.6</td>
<td>0.295</td>
<td>0.132</td>
</tr>
<tr>
<td>ResNet-50 (SWSL)</td>
<td>73.8</td>
<td>96.3</td>
<td>39.4 90.2 97.0 98.4 99.4</td>
<td>26.3 60.9 73.0 83.8 87.5</td>
<td>0.380</td>
<td>0.149</td>
</tr>
<tr>
<td>ResNet-18 (SSL)</td>
<td>72.6</td>
<td>90.5</td>
<td>42.3 93.8 98.8 99.8 99.8</td>
<td>34.3 75.1 88.9 96.6 97.9</td>
<td>0.312</td>
<td>0.142</td>
</tr>
<tr>
<td>ResNet-50 (SSL)</td>
<td>74.2</td>
<td>95.5</td>
<td>39.5 91.8 97.6 99.5 99.9</td>
<td>26.3 60.5 75.2 85.8 89.5</td>
<td>0.372</td>
<td>0.149</td>
</tr>
<tr>
<td>ResNet-18</td>
<td>69.8</td>
<td>83.7</td>
<td>46.1 90.0 97.8 99.9 100</td>
<td>42.0 75.2 88.5 95.7 98.2</td>
<td>0.302</td>
<td>0.237</td>
</tr>
<tr>
<td>ResNet-50</td>
<td>75.1</td>
<td>93.0</td>
<td>35.8 86.3 97.9 99.5 100</td>
<td>27.5 63.1 77.6 89.4 93.9</td>
<td>0.371</td>
<td>0.287</td>
</tr>
</tbody>
</table>

Table 2: 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 were trained with an image size of 224, and attacked with a maximum $\ell_\infty$ perturbation of $\epsilon = 16$.

Table 3: Query-based black-box attack on benchmark models. We test 100 random samples from NeurIPS dataset, and the $\ell_2$-norm of adversarial perturbation is presented.

<table>
<thead>
<tr>
<th>Source model</th>
<th>Variant</th>
<th>Target model</th>
<th>Source model</th>
<th>Variant</th>
<th>Target model</th>
<th>Source model</th>
<th>Variant</th>
<th>Target model</th>
</tr>
</thead>
<tbody>
<tr>
<td>ViT-L/16</td>
<td>I-FGSM</td>
<td>ResNet-18 (SWSL)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mixer-B/16</td>
<td>I-FGSM</td>
<td>ResNet-18 (SWSL)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mixer-L/16</td>
<td>I-FGSM</td>
<td>ResNet-18 (SWSL)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ResNet-18 (SWSL)</td>
<td>I-FGSM</td>
<td>ResNet-18 (SWSL)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ResNet-50 (SWSL)</td>
<td>I-FGSM</td>
<td>ResNet-18 (SWSL)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ResNet-18</td>
<td>I-FGSM</td>
<td>ResNet-18 (SWSL)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ResNet-50</td>
<td>I-FGSM</td>
<td>ResNet-18 (SWSL)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

4.3. Toy Example

In the previous white-box attack, we observed that ViT and MLP-Mixer are more robust to adversarial examples than conventional CNNs. To facilitate the understanding of the mechanisms, we design a toy example of binary classification where each class is only represented by a single image with a size of 224. The two images consist of a single black stripe on a grey background, differing only in the orientation of the stripe, namely a vertical and a horizontal stripe. The two images used for training are shown in Figure 1. We then train a Fully Connected network (FC), a Convolution Neural Network (CNN), and a Vision Transformer (ViT) on the images. Note that we designed the networks to be of relatively small capacity ($< 5M$), due to the simplicity of the task and to constrain that the networks have around the same number of parameters. We evaluate the adversarial robustness of these models with the commonly used $\ell_2$ attacks C&W and DDN. We report the $\ell_2$-norm of the adversarial perturbation in Table 4. It can be observed that the CNN is also less robust than the FC and the ViT in this toy example setup.

**Explanation from the perspective of shift-invariance.** The qualitative results of adversarial perturbations gener-
Adversarial examples and perturbations generated against C&W attack using different architectures trained on toy example.

Table 4: Results for the $\ell_2$-norm of adversarial perturbation on our toy example.

<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>12.55</td>
<td>13.91</td>
<td>4.50M</td>
</tr>
<tr>
<td>FC</td>
<td>25.06</td>
<td>25.39</td>
<td>4.82M</td>
</tr>
<tr>
<td>ViT</td>
<td>27.82</td>
<td>59.99</td>
<td>4.88M</td>
</tr>
</tbody>
</table>

The observation that the CNN model yields stripes all over the image can be naturally attributed to the shift-invariant property of the CNN model. From the perspective of shift-invariance, 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 middle; thus, the resulting perturbation mainly has the stripe in the middle. This qualitative result suggests that the reason for CNN being more vulnerable can be partially attributed to its shift-invariance. Future work is needed to further establish the link between shift-invariant property and model vulnerability to adversarial attack.

4.4. Frequency Analysis

We further attempt to explain the lower robustness of CNN from the perspective of frequency [50, 54]. Following the practice in [50, 54], we 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 3. For the low-pass filtering, a sharper decline of the CNN architectures can be observed than for the ViT, indicating that the CNN architectures are more reliant on the high-frequency features. 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 [50, 54, 20], and attribute to decreased model robustness. This indicates why ViT models are more robust than CNN architectures. When results from both low-pass and high-pass filtering are compared, it is observed that Mixers, regardless of their absolute value of accuracy, exhibit a similar trend to CNNs rather than ViTs.

5. 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 vertical or horizontal black stripe in the middle indicates that the lower robustness of CNN can be partially attributed to the shift-invariant property of CNNs. Our analysis from the feature perspective further suggests that ViTs are more reliant on low-frequency (robust) features while CNNs are more sensitive to the 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. Future work is needed for better understanding.
References

[20] Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. In NeurIPS, 2019. 4