Robustness Comparison of Vision Transformer and MLP-Mixer to CNNs

Philipp Benz* pbenz@kaist.ac.kr

Chaoning Zhang* chaoningzhang1990@gmail.com

Soomin Ham* smham@kaist.ac.kr

Adil Karjauv mikolez@gmail.com

In So Kweon 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

*Equal contribution

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, by sequencing the images into patches and pre-training the model on large amounts of data. To address the data issue, DeiT [44] 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 [51] 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 [8, 14, 27, 48]. 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 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 architecture 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 ℓ_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 ℓ_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 ℓ_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 ℓ_∞ distortion, and the ℓ_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 ℓ_2-norm is considered to be more robust.

	Clean		PGD (ℓ_∞)					FGSM (ℓ_∞)					C&W (ℓ_2)	DeepFool (ℓ_2)
Model	ImageNet	NeurIPS	0.1	0.3	0.5	1	3	0.1	0.3	0.5	1	3		
ViT-B/16	81.4	90.7	**22.6**	63.6	86.5	97.5	99.9	**19.1**	38.7	52.8	66.3	79.7	**0.468**	0.425
ViT-L/16	82.9	89.3	22.8	**60.1**	80.9	95.8	100	19.5	**35.9**	**44.9**	**57.9**	**67.3**	0.459	**0.548**
Mixer-B/16	76.5	86.2	29.5	63.4	82.0	96.2	100	27.7	49.3	59.5	69.3	78.0	0.375	0.339
Mixer-L/16	71.8	80.0	41.1	67.3	**80.4**	**92.1**	**99.4**	36.7	51.8	56.9	61.6	67.4	0.297	0.377
ResNet-18 (SWSL)	73.3	90.4	47.9	93.7	98.7	99.5	99.6	38.0	76.3	89.9	96.2	97.6	0.295	0.132
ResNet-50 (SWSL)	81.2	96.3	39.4	90.2	97.0	98.4	**99.4**	26.3	60.9	73.0	83.8	87.5	0.380	0.149
ResNet-18 (SSL)	72.6	90.5	42.3	93.2	98.8	99.8	99.8	34.3	75.1	88.9	96.6	97.9	0.312	0.142
ResNet-50 (SSL)	79.2	95.3	39.5	91.8	97.6	99.5	99.9	26.3	60.5	75.2	85.8	89.5	0.372	0.149
ResNet-18	69.8	83.7	46.1	90.0	97.8	99.9	100	42.0	75.2	88.5	95.7	98.2	0.302	0.237
ResNet-50	76.1	93.0	35.8	86.3	97.9	99.5	100	27.5	63.1	77.6	89.4	93.9	0.371	0.287

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 ℓ_∞ perturbation of $\epsilon = 16$.

		Target model									
Source model	**Variant**	**ViT-B/16**	**ViT-L/16**	**Mixer-B/16**	**Mixer-L/16**	**ResNet-18 (SWSL)**	**ResNet-50 (SWSL)**	**ResNet-18 (SSL)**	**ResNet-50 (SSL)**	**ResNet-18**	**ResNet-50**
ViT-B/16	I-FGSM	100	84.7	48.8	50.5	32.0	20.5	34.3	23.4	40.9	31.7
ViT-L/16	I-FGSM	90.9	99.9	45.7	48.0	30.4	22.2	34.4	23.6	40.8	30.9
Mixer-B/16	I-FGSM	33.9	25.3	100	89.1	30.6	20.5	34.5	23.3	40.8	32.0
Mixer-L/16	I-FGSM	27.7	20.1	80.3	99.7	27.7	17.0	31.5	17.5	38.2	28.4
ResNet-18 (SWSL)	I-FGSM	16.2	13.6	24.8	29.5	99.6	57.1	80.2	58.0	73.5	63.4
ResNet-50 (SWSL)	I-FGSM	15.3	13.5	23.6	29.9	56.5	99.5	51.6	69.1	49.4	51.0
ResNet-18 (SSL)	I-FGSM	17.7	13.7	28.6	34.4	84.4	54.6	99.9	65.4	78.2	66.8
ResNet-50 (SSL)	I-FGSM	18.1	15.0	26.4	32.3	58.9	73.3	64.7	100	54.7	62.2
ResNet-18	I-FGSM	18.2	14.7	28.9	35.6	84.6	49.9	85.3	60.4	100	81.6
ResNet-50	I-FGSM	17.7	13.6	28.4	34.5	73.9	63.9	74.3	74.7	80.6	100

attack and transfer-based black-box attack.

Query-based Black-box Attacks. We adopt one popular Boundary Attack [3] and the results are available in Table 3. As with the white-box attack, a trend can be observed in the black-box that the ViT and Mixer models are more robust, indicated by the relatively higher ℓ_2-norm of the adversarial perturbation.

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 [23] attack with 7 steps and evaluate the transferability on the target model. From the result in Table 2, 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 much higher.

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 ℓ_2 attacks C&W [4] and DDN [38]. We report the ℓ_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.

Table 3: **Query-based black-box attack on benchmark models**. We test 100 random samples from NeurIPS dataset, and the ℓ_2-norm of adversarial perturbation is presented.

	ViT-B	ViT-L	Mix-B	Mix-L	RN18 (SWSL)	RN50 (SWSL)	RN18 (SSL)	RN50 (SSL)	RN18	RN50
Boundary (ℓ_2)	3.980	7.408	1.968	1.951	1.403	1.846	1.434	1.780	1.468	1.740

Explanation from the perspective of shift-invariance. The qualitative results of adversarial perturbations gener-

Figure 1: Images for our binary classification toy example.

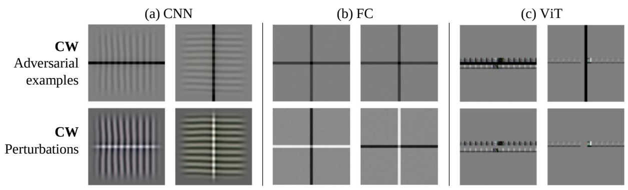

Figure 2: Adversarial examples and perturbations generated against C&W attack using different architectures trained on toy example.

Table 4: Results for the ℓ_2-norm of adversarial perturbation on our toy example.

	C&W (ℓ_2)	DDN (ℓ_2)	$\#params$
CNN	12.55	13.91	$4.59M$
FC	25.06	25.39	$4.82M$
ViT	27.82	59.99	$4.88M$

ated by the attacks are shown in Figure 2. 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 in the middle. 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 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.

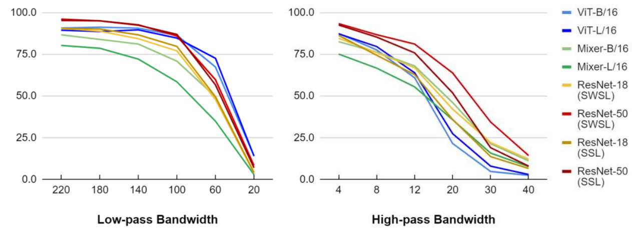

Figure 3: Top-1 accuracy across a range of frequency bandwidths from low/high-pass filtering. **Left:** Low-pass filtering. **Right:** High-pass filtering.

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

[1] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 2

[2] Philipp Benz, Chaoning Zhang, Tooba Imtiaz, and In So Kweon. Double targeted universal adversarial perturbations. In *ACCV*, 2020. 2

[3] Wieland Brendel, Jonas Rauber, and Matthias Bethge. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models. *arXiv preprint arXiv:1712.04248*, 2017. 3

[4] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *Symposium on Security and Privacy (SP)*, 2017. 2, 3

[5] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021. 2

[6] Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In *ICML*, 2020. 1, 2

[7] Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. Generating long sequences with sparse transformers. *arXiv preprint arXiv:1904.10509*, 2019. 1, 2

[8] Xiangxiang Chu, Bo Zhang, Zhi Tian, Xiaolin Wei, and Huaxia Xia. Conditional positional encodings for vision transformers. *arXiv preprint arXiv:2102.10882*, 2021. 2

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR*, 2009. 2

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 1, 2

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 1, 2

[12] Siddhant Garg and Goutham Ramakrishnan. Bae: Bert-based adversarial examples for text classification. *arXiv preprint arXiv:2004.01970*, 2020. 2

[13] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *ICLR*, 2015. 1, 2

[14] Kai Han, An Xiao, Enhua Wu, Jianyuan Guo, Chunjing Xu, and Yunhe Wang. Transformer in transformer. *arXiv preprint arXiv:2103.00112*, 2021. 2

[15] Yaru Hao, Li Dong, Furu Wei, and Ke Xu. Self-attention attribution: Interpreting information interactions inside transformer. *arXiv preprint arXiv:2004.11207*, 2020. 2

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 2

[17] Shuting He, Hao Luo, Pichao Wang, Fan Wang, Hao Li, and Wei Jiang. Transreid: Transformer-based object re-identification. *arXiv preprint arXiv:2102.04378*, 2021. 2

[18] Dan Hendrycks, Xiaoyuan Liu, Eric Wallace, Adam Dziedzic, Rishabh Krishnan, and Dawn Song. Pretrained transformers improve out-of-distribution robustness. *arXiv preprint arXiv:2004.06100*, 2020. 2

[19] Yu-Lun Hsieh, Minhao Cheng, Da-Cheng Juan, Wei Wei, Wen-Lian Hsu, and Cho-Jui Hsieh. On the robustness of self-attentive models. In *Annual Meeting of the Association for Computational Linguistics*, 2019. 2

[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

[21] Yifan Jiang, Shiyu Chang, and Zhangyang Wang. Transgan: Two transformers can make one strong gan. *arXiv preprint arXiv:2102.07074*, 2021. 2

[22] Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. Is bert really robust? a strong baseline for natural language attack on text classification and entailment. In *AAAI*, 2020. 2

[23] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial machine learning at scale. In *ICLR*, 2017. 2, 3

[24] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 2015. 1

[25] Karel Lenc and Andrea Vedaldi. Understanding image representations by measuring their equivariance and equivalence. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 991–999, 2015. 1

[26] Linyang Li, Ruotian Ma, Qipeng Guo, Xiangyang Xue, and Xipeng Qiu. Bert-attack: Adversarial attack against bert using bert. *arXiv preprint arXiv:2004.09984*, 2020. 2

[27] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *arXiv preprint arXiv:2103.14030*, 2021. 2

[28] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *ICLR*, 2018. 2

[29] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Universal adversarial perturbations. In *CVPR*, 2017. 2

[30] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *CVPR*, 2016. 2

[31] Konda Reddy Mopuri, Utkarsh Ojha, Utsav Garg, and R. Venkatesh Babu. Nag: Network for adversary generation. In *CVPR*, 2018. 2

[32] Daniel Neimark, Omri Bar, Maya Zohar, and Dotan Asselmann. Video transformer network. *arXiv preprint arXiv:2102.00719*, 2021. 2

[33] Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Lukasz Kaiser, Noam Shazeer, Alexander Ku, and Dustin Tran. Image transformer. In *ICML*, 2018. 1, 2

[34] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding with unsupervised learning. 2018. 1, 2

[35] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI blog*, 2019. 1, 2

[36] Prajit Ramachandran, Niki Parmar, Ashish Vaswani, Irwan Bello, Anselm Levskaya, and Jonathon Shlens. Stand-alone self-attention in vision models. In *NeurIPS*, 2019. 1, 2

[37] Jonas Rauber, Roland Zimmermann, Matthias Bethge, and Wieland Brendel. Foolbox native: Fast adversarial attacks to benchmark the robustness of machine learning models in pytorch, tensorflow, and jax. *Journal of Open Source Software*, 2020. 2

[38] Jérôme Rony, Luiz G Hafemann, Luiz S Oliveira, Ismail Ben Ayed, Robert Sabourin, and Eric Granger. Decoupling direction and norm for efficient gradient-based l2 adversarial attacks and defenses. In *CVPR*, 2019. 2, 3

[39] Zhouxing Shi, Huan Zhang, Kai-Wei Chang, Minlie Huang, and Cho-Jui Hsieh. Robustness verification for transformers. *arXiv preprint arXiv:2002.06622*, 2020. 2

[40] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *JMLR*, 2014. 2

[41] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013. 1, 2

[42] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 2016. 2

[43] Ilya Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Daniel Keysers, Jakob Uszkoreit, Mario Lucic, and Alexey Dosovitskiy. Mlp-mixer: An all-mlp architecture for vision. *arXiv preprint arXiv:2105.01601*, 2021. 1, 2

[44] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. *arXiv preprint arXiv:2012.12877*, 2020. 2

[45] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 2

[46] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. *arXiv preprint arXiv:2102.12122*, 2021. 2

[47] Zifan Wang, Yilin Yang, Ankit Shrivastava, Varun Rawal, and Zihao Ding. Towards frequency-based explanation for robust cnn. *arXiv preprint arXiv:2005.03141*, 2020. 2

[48] Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. Cvt: Introducing convolutions to vision transformers. *arXiv preprint arXiv:2103.15808*, 2021. 2

[49] I Zeki Yalniz, Hervé Jégou, Kan Chen, Manohar Paluri, and Dhruv Mahajan. Billion-scale semi-supervised learning for image classification. *arXiv preprint arXiv:1905.00546*, 2019. 2

[50] Dong Yin, Raphael Gontijo Lopes, Jon Shlens, Ekin Dogus Cubuk, and Justin Gilmer. A fourier perspective on model robustness in computer vision. In *NeurIPS*, 2019. 4

[51] Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Francis EH Tay, Jiashi Feng, and Shuicheng Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet. *arXiv preprint arXiv:2101.11986*, 2021. 2

[52] Chaoning Zhang, Philipp Benz, Tooba Imtiaz, and In-So Kweon. Cd-uap: Class discriminative universal adversarial perturbation. In *AAAI*, 2020. 2

[53] Chaoning Zhang, Philipp Benz, Tooba Imtiaz, and In-So Kweon. Understanding adversarial examples from the mutual influence of images and perturbations. In *CVPR*, 2020. 2

[54] Chaoning Zhang, Philipp Benz, Adil Karjauv, and In So Kweon. Universal adversarial perturbations through the lens of deep steganography: Towards a fourier perspective. *AAAI*, 2021. 4

[55] Chaoning Zhang, Philipp Benz, Chenguo Lin, Adil Karjauv, Jing Wu, and In So Kweon. A survey on universal adversarial attack. *IJCAI*, 2021. 2

[56] Richard Zhang. Making convolutional networks shift-invariant again. In *International Conference on Machine Learning*, pages 7324–7334. PMLR, 2019. 1