Adversarial Robustness towards different Vision Transformer

Yuan Ma, Apr.28, 2023

Introduction

Machine learning systems have become integral to a wide array of applications, necessitating a critical examination of their security. Adversarial attacks emerge as a key concern by introducing input data meticulously perturbed to cause incorrect model outputs. While it is hard to be noticed by human, these adversarial examples exploit the model's weaknesses, revealing a discrepancy between machine processing and the ground truth. The discovery of such vulnerabilities has led to a distinction between attack types based on the attacker's knowledge —white-box attacks, with full model awareness, and black-box attacks, with access only to model outputs. The adversarial intent also varies, with targeted attacks aiming for specific errors and non-targeted attacks for any incorrect output. The implications of these vulnerabilities are significant as adversarial examples can be used to mislead systems in various applications. In image recognition, for example, an adversary could manipulate traffic signs in a way that autonomous vehicles misinterpret them.

In 2017, Attention is all you need[2] proposes fundamental model called transformer, provokes evolutions in natural language processing. Based on attention mechanism, transformer is proof to be a powerful understanding machine due to its ability to capture long term dependencies. Later on, computer vision community also developed their own transformer model: Vision Transformer(Vit). The inspiration of this work is following exactly the same design as transformer in NLP to prove that pure attention based architecture can achieve competitive results as Convolutional Neural Network(CNN), which dominates computer vision for years. Although Vit has a lot strength over its counterpart CNN such as global context awareness, scalability with data size and powerful learning ability, it has been found that while ViTs demonstrate some inherent robustness benefits due to their attention mechanisms, they are not immune to adversarial attacks. Nowadays, there are numerous ongoing research focus on robustness of vision transformer, we can divided them in several areas: comparing the robustness of Vit to CNN, the effect of exist attack method on Vit and also theoretical researches on adversarial attack over Vit.

However, there are few researches investigate which specific architecture of vision transformer will be more robust to adversarial attack given that there is a lot different structures in vision transformer family. This work aimed to bridge this gap. In order to explore robustness of different vision transformer against adversarial attack, we argue that there is four main categories of them for convenience: Data-efficient model[10], multi-scale structure, transformer with local operation, transformer with efficient attention mechanism. We will introduce more details in method part.

In this paper, we will do both adversarial trainings and adversarial attacks on different categories of Vit, and study how these difference can impacts the robustness of vision transformer. Adversarial training is a technique designed to improve the robustness of machine learning models by explicitly training them with adversarial examples. These adversarial examples are inputs to the machine learning model that have been intentionally perturbed in a way that causes the model to make a mistake. The idea is to simulate an adversarial environment during training so that the model can learn to withstand such conditions during

inference. The attack method we choose here is Projected Gradient Descent attack, which is one the most popular and powerful attack at presents. Introduced by Madry et al. in 2017, it is aimed at finding the worst-case perturbations within a specified neighborhood of a given input. We will discuss the method later in related work section. After conduct attacks on different Vit models, it can be found that the multi-scale structure helps model improve their robustness significantly, which again and again, proves the invaluable roles of multi-scale features in vision model. And also local operation may not make model more vulnerable towards adversarial attack, which is opposite from the observation from [4] On the Adversarial Robustness of vision transformer. And efficient attention mechanism can also gives us a more robust model than standard attention.

Related Work

PGD Attack The Projected Gradient Descent (PGD) attack[3] stands out as a powerful and widely used method for evaluating the robustness of machine learning models, particularly in the context of deep learning. The PGD attack is an iterative method that generates adversarial examples, which are slightly modified inputs designed to deceive machine learning models into making incorrect predictions.

PGD operates by performing small steps in the direction that maximizes the loss of the model with respect to the input, thus leading the model towards misclassification. After each step, the perturbation is projected back onto a set of allowable perturbations (typically an norm ball around the original input) to ensure that the adversarial example remains perceptually similar to the original input.

The process is formulated as a constrained optimization problem:

$$maximize J(\theta, x', y)$$
$$subject\ to\ \ x' - x\|_p \leq \epsilon$$

where x' is the adversarial example, x is the original input, y is the true label, J is the loss function of the model with parameters θ, and ϵ is the magnitude of the allowed perturbation.
The PGD attack is considered one of the strongest first-order adversarial attacks due to its iterative nature and is commonly used as a benchmark to test the robustness of models against adversarial threats. Models that are robust to PGD are considered to have a strong stance against a wide array of adversarial tactics.

Vision transformer and its variances The structure of vision transformer is very simple. It only contain transformer encoder with an additional head for classification. It view picture in as little 2-D patches, then project these patches into patch embeddings to feed into encoders. One of the main advantages of ViTs is their ability to model long-range interactions between pixels, irrespective of their spatial proximity, allowing for a more deeper understanding of the image. This capability becomes increasingly significant as the scale of data and model size grows, where ViTs have been shown to outperform CNNs, especially in datasets with a large number of training samples.

However, comparing to CNNs, VIT do not have inductive bias so that they are data-hungry, requiring substantial amounts of labeled data to reach their full potential. This has led to the adoption of techniques like transfer learning and pre-training on large datasets, followed by

fine-tuning on specific tasks. Moreover, recent advancements have introduced variations in the architecture and training procedures to improve the data efficiency and performance of ViTs.

One of the representative work among those advancements is Shift Window Transformer. While Vision transformer is focus on stay the original architecture with NLP, Swin is a specialist in CV. Unlike the flat architecture of the standard ViT, which processes a sequence of non-overlapping image patches, the Swin Transformer operates at multiple scales. It begins with small patches and progressively merges them, enabling the model to capture features at various resolutions. This hierarchical structure makes it more efficient and adaptable to different image sizes, an advantage over traditional transformers that require fixed-size inputs. Another distinctive feature of the Swin Transformer is the shifted window approach to self-attention. In contrast to global self-attention, which is computationally expensive and not feasible for high-resolution images, Swin applies self-attention within local windows that are shifted across the image in a way that allows for cross-window connections. This method significantly reduces the computational burden and allows for efficient processing of larger images. Except Swin Transformer, there are a lot of other works aimed to take a further investigation in vision transformer, in order to analyze how different structure will influence robustness of vision transformers, we better categories them in different groups.

On the Adversarial Robustness of Vision Transformer Tested on various white-box and transfer attack settings, it can be found that ViTs possess better adversarial robustness when compared to CNN. Through frequency analysis and feature visualization, this paper summarize the following main observations in robustness of ViTs: 1) Features learned by ViTs contain less high-frequency patterns that have spurious correlation, which helps explain why ViTs are less sensitive to high-frequency perturbations than CNNs and MLP-Mixer, and there is a high correlation between how much the model learns high-frequency features and its robustness against different frequency-based perturbations. 2) Introducing convolutional or tokens-to-token blocks for learning high-frequency features in ViTs can improve classification accuracy but at the cost of adversarial robustness. 3) Modern CNN designs that borrow techniques from ViTs including activation function, layer norm, larger kernel size to imitate the global attention, and patchify the images as inputs, etc., could help bridge the performance gap between ViTs and CNNs not only in terms of performance, but also certified and empirical adversarial robustness. Moreover, it shows adversarial training is also applicable to ViT for training robust models, and sharpness-aware minimization can also help improve robustness, while pre-training with clean images on larger datasets does not significantly improve adversarial robustness. Although this paper provide us a pipeline to have a adversarial training towards vision transformer, it do not investigate different structure. Based on training pipeline provided in this work, we can perform adversarial attack towards various vision transformer.

Towards Different Categories of Vision Transformer

To our best knowledge, this is the first work that investigate different vision transformer to see which architecture is good for robustness while the other design is harmful for robustness. But prior to that, we firstly make a classification on different vision transformer. Here we argue that there are mainly four types of Vision Transformers. Because of the data efficient vision transformer is not involved with change of model structure, so here we don't dive deeper in this type of vision transformer.

ViT with locality property As we talked in ViT, a noticeable weakness of ViT is its lack of ability to capture local information, which happens to be the key success regard to CNN. So there are a lot of researchers focus on locality to improve performance of ViT.

GLiT is such a model that combines local operation and global attention into one structure. Authors of this work argue that it can not be optimal if we directly use structure from NLP (which is exactly ViT doing), however instead of design a specific structure by hand, it propose a search space and searching algorithm to generate a optimal structure which contain both local convolution and global attention. Specifically, it replace some self-attention head with convolutional head, and use search algorithm to find out the best proportion of convolution and self-attention head, the structure is shown in Fig.5

Another example of work taking advantage of local information is LocalViT. It's more simple comparing to GLiT but also very effective. Instead of replace the self-attention with convolution, it view the last layers in ViT encoder module, which is feed forward network, as a 1*1 Convolution, given that we perform sequence to image operation first to output token. Then it argue that such 1*1 convolution do not have enough ability to fuse local information together. So a straightforward solution is that we modify this 1*1 convolution as 3*3 convolution. Specifically, we can change the FFN structure into one shown in Fig.

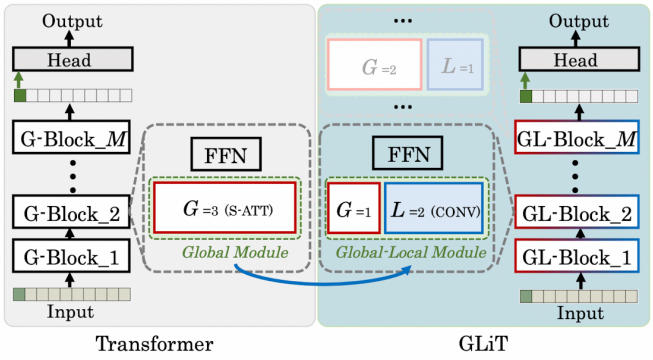

Fig.5 Refined Global-Local module

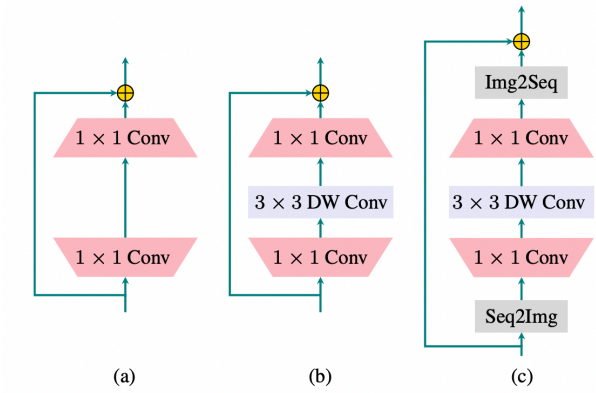

Fig.6 Comparing FFN in ViT and LocalViT

Vision Transformer with multi-scale structure Pyramid structure is widely regarded as a valuable component for multiple vision tasks such as object detection and semantic segmentation since 1960 (Receptive fields of optic nerve fibres in the spider monkey, Hubel and Wiesel) . Success of Feature Pyramid Network(FPN) and its variance such as Path Aggregation Network(PAN) are also strong proof of importance of multi-scale structure. The multi-scale structure approach processes visual information at different resolutions. It mimics the human visual system's ability to perceive fine details when focusing on an object closely and broader shapes and structures from a distance. By incorporating multiple scales, computer vision systems can capture both granular details and global context. However, vanilla ViT do not hold this good property, so there are a lot of following work aimed to solve this issues. Here we introduce several representative works to give a flavor of how does Vision Transformer involved with multi-scale features.

One of the first multi-scale vision transformer is proposed in , as we can see PVT view image not as 16*16 token but image, in other word, it convert token to image again after each stage. In first stage, it separate image as 4*4 token, after the self-attention, it still take token as pixel, then apply same operation on new “image”. In this case, although the number of toke remain the same, but it actually achieve a pyramid structure.

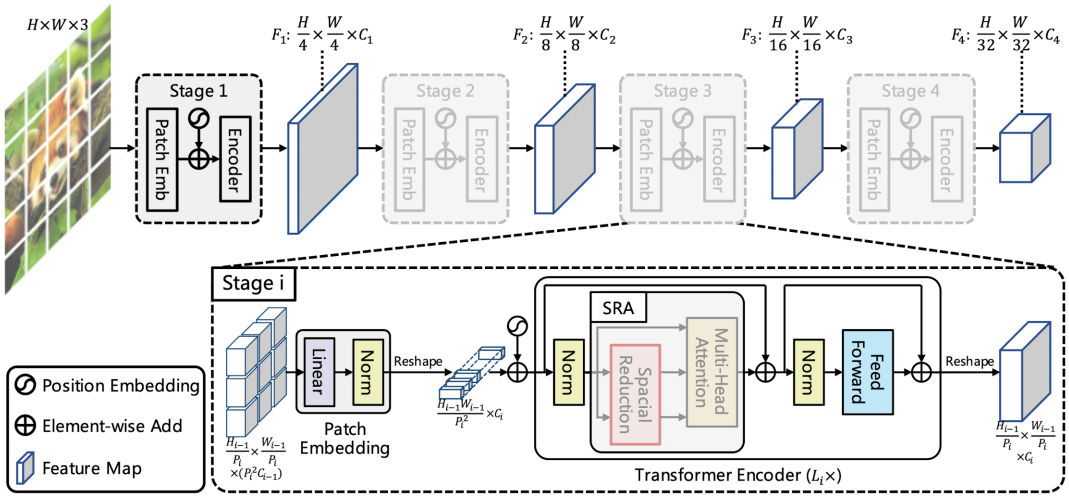

Fig.3 The structure of PVT

Swin transformer one of the most importance variances of Vit, it achieves state-of-art performance on image classification, objects detection and so on. Unlike standard ViTs that process an image as a sequence of flat patches, the Swin Transformer introduces a novel hierarchical design. It processes images through a series of stages, each operating at a different scale, similar to how convolutional neural networks (CNNs) reduce spatial resolution while increasing the depth of feature maps.A core innovation in the Swin Transformer is the use of shifted windows. This design allows the model to compute self-attention in local windows, but in a manner that the windows are shifted from one block to the next. This shift enables cross-window connections and reduces the computational complexity of self-attention from quadratic to linear, as it limits the number of interactions to be within local neighborhoods while still capturing global information.

Another notable work who exploit multi scale information is MViT (Multi-scale Vision Transformer). And the idea of this work is quite simple. Just like CNN network, it uses pooling layers to reduce spatial dimension of features, just as shown in Fig.4.

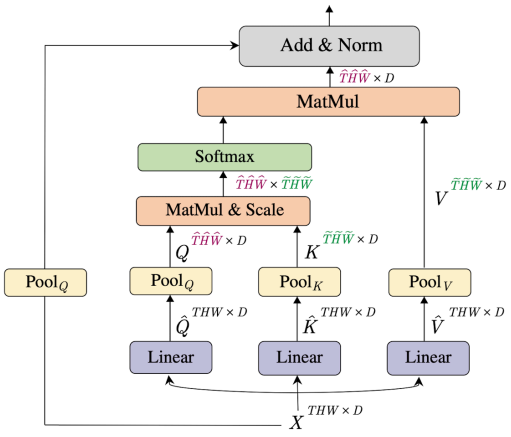

Fig.4 pooling operation in MVT

Vision transformer with efficient self-attention The rising of transformer in both NLP and CV proves ability of self-attention operation. However, self-attention is also viewed as high-computational complexity which increases quadratically with number of tokens. As a result, it can be extremely hard for vision transformer to deal with high resolution image, however such high quality image is critical for some down-stream tasks such as object detection and so on. Swin Transformer solve computational problem by local self-attention, but it might do harm to model ability to capture long-term dependency. Some researchers aim to find an adapted more efficient attention mechanism to solve computation issues.

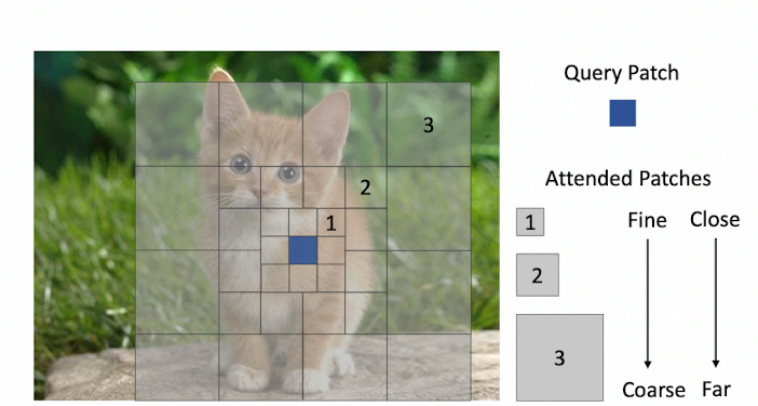

Fig.7 Focal attention mechanism

Focal transformer perform fine-grained self-attention only in local region while the coarse-grained attention globally, as shown in Fig.7 In practice, it suffers from a high memory cost if we need to store fine-grained tokens around each pixel. It proposed a window-wise focal attention to solve this issue.

And another example is Vision transformer with deformable attention[ ]. This deformable mechanism is very effective in CNN, which is known as Deformable Convolutional Networks. Inspired by this work, authors propose a novel deformable attention. Notice that the concept of deformable is nothing new and deformable attention is used in DEtection TRansformer (DETR),

however, none of this issue can be directly use in vision transformer. So in this paper, author developed a novel deformable attention module which is designed to fit to the natural of vision transformer. It takes uniform sample points and then learns offsets by a lightweight CNN. Then perform multi-head attention with respect to sample points. With such efficient attention mechanism, we can not only reduce the computation complexity of transformer, but also increase the generality of model.

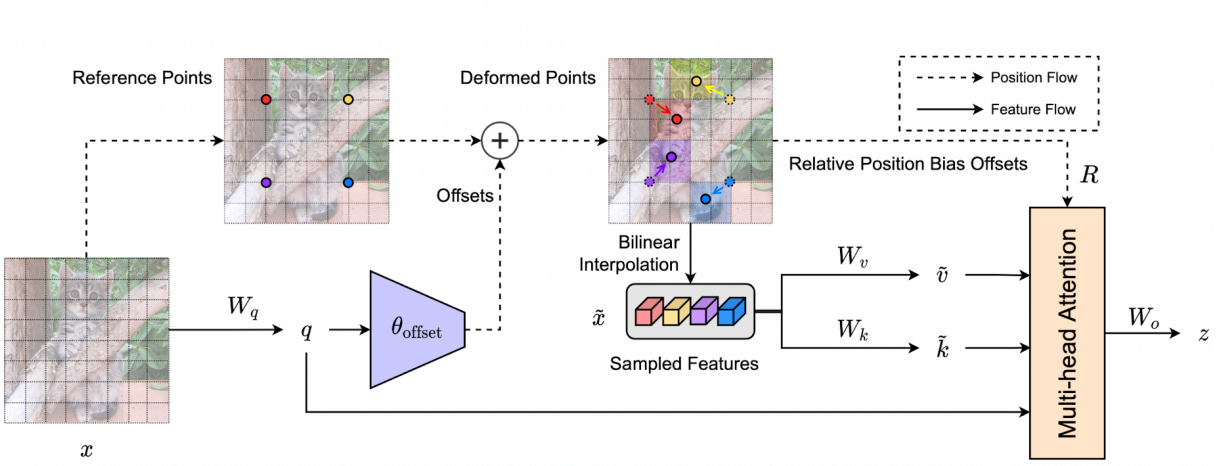

Fig.8 Deformable attention mechanism

Adversarial Training On Various Vision Transformers

The adversarial robustness vary a lot in different adversarial training setting, so that we need training on a uniform training setting. In this article we follow the training processing as [], in which they trained model pre-trained on imagenet1k and train Vision Transformer on different data set such as cifar-10, imagenette and imagenet. It apply L-norm infinite attack and set theta as 8/255. Then they use 20-steps PGD attack[3] with 2/255 step size to evaluating model robustness.

Imagenette is a subset of ImageNet that includes a selection of 10 easily classified classes from the original dataset. The dataset is much smaller and more manageable compared to the full ImageNet, making it suitable for our experimentation scale. For L-norm infinite, it can be formulated as an optimization problem.

$$maximize\ L(f(x'), y)$$
$$subject\ to\quad \|x' - x\|_\infty \leq \theta$$

$(L(f(x'), y))$ is a loss function that measures the error between the classifier's prediction for (x') and the true label y. The goal of the attack is to maximize this loss, meaning that the adversarial image x' is classified incorrectly with high confidence. $\|x' - x\|_\infty$ represents the L_∞ norm of the perturbation, which is the maximum change over all pixels in the image.
θ is the maximum magnitude of the perturbation allowed for any pixel. If pixel values are in the range [0, 1].

Because of different purpose and computational resource, this paper makes following changes: (1) We still use pre-trained model but perform adversarial training only on imagenette dataset because of limited resources. (2) Here we will perform adversarial training on various Vision transformer to compare the compact of different structure. Specifically, we training three different kind of structure:vanilla ViT, Vision transformer with multi-scale structure, Vision transformer with efficient attention. Notice that we didn't train Vision transformer with local operation because lack of pre-trained model. For vanilla vision transformer, we follow the same setting we talked before, however, in PGD attack we switch batches to 2 because of memory constrain. And then we trained vision transformer with multi-scale structure: Swin Transformer small and PVT small transformer. Authors of PVT provide a lot pre-trained model in different scale, and do not involve a lot of complex design, so it is a very good structure to explore multi-scale feature's impact on model robustness. 3) Also we can conduct more ablation experiment in single structure or between different structure to give us a concrete results. For example, in multi-scale vision transformers, we can reduce the number of multi-scale features to observe model robustness. Also in GLiT we can change the ratio between local operation and global operation to obverse the different natural accuracy and robust accuracy. (4) Given the fact that even single category of Vision Transformer have multiple models (ViT-small, ViT-base, ViT-large), we only do experiment on relative small transformer. However, it is fair to give more experiment on more model in different size and compare the results within similar scale.

Results

	Natural accuracy	PGD-20	PGD100
ViT (small)	92.5	63.2	62.8
Swin Transformer(small)	95.6	73.0	73.0
PVT(small)	73.0	48.6	48.6
DAT	83.7	62.1	61.8

In this paper we perform experiment on 4 different structure of Vision transformer in 3 different category to give us some general sense their impact on model robustness. As we can see, the pyramid structure of Vision transformer is good for model robustness: For vanilla ViT the accuracy drops by around 30% after PGD attack. And for Swin and PVT, the accuracy drops by around 20% and 25% respectively. Additionally, efficient attention like Deformable transformers is also good for model robustness: the performance drops from 83.7 to 62.1 after adversarial attack. Here we argue that the ability for model generality and ability to capture high complexity context information is important to improve model robustness. In multi-scale structure, transformers can learn much highly semantic information with the resolution decrease with layers, in the top layer, tranformers can learning very good context information so that is hard to trick model. Additionally, efficient attention like deformable attention always reduce the number of parameters and operation of neural network, so that it will give the model better generality to deal with adversarial attack. Another inside is that efficient attention like deformable attention only focus on small key points of features maps and it is likely to avoid the toxic pixel generated by adversarial attack.

Conclusion

In this paper we conduct several experiments towards the robustness of vision transformer. We can observe form the training results we can conclude that vision transformer with multi-scale structure and efficient attention tend to hold better robustness against malicious attack. This work hopefully can serve as a general guidance on how to choose vision transformer when considering the secure factors and give deeper understanding of robustness of Vision Transformer to community.

This paper only did a simple and general survey and experiment on robustness of Vision Transformer, it could give us no convincing results but some general idea on how different structure can influence transformer. In order to get more concrete and convincing results, more experiments are required. For instance, one might do adversarial training and attack on different dataset with different scale of transformer model.

Because of different models has different architecture with countless training details and techniques, only use pre-trained model by other might not be a fair strategy. For example, in the same setting, PVT can only reach 73.00 natural accuracy over imagenette but Vit and Swin Transformer achieve over 90%. To give us a fair competition one might set model in similar pre-training criteria and similar accuracy which need to be trained from scratch.

Reference:

[1] Alexey Dosovitskiy , Lucas Beyer , Alexander Kolesnikov , Dirk Weissenborn , Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby. An image is worth 16*16 words: Transformers for Image Recognition at Scale. arXiv:2010.11929

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

[3] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, Adrian Vladu, Towards Deep Learning Models Resistant to Adversarial Attacks, ICLR 2018.

[4] Ruilin Shao. On the Adversarial Robustness of Vision Transformers. arXiv:2103.15670

[5] Sayak Paul, Pin-yu Chen. Vision transformer are robust learners. arXiv:2105.07581

[6] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, Ling Shao. Pyramid Vision Transformer: A Versatile Backbone for Dense Prediction without Convolutions. arXiv:2102.12122

[7] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, Baining Guo. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. arXiv:2103.14030

[8] Zhuofan Xia, Xuran Pan, Shiji Song, Li Erran Li, Gao Huang. Vision Transformer with Deformable Attention. arXiv:2201.00520

[9] Yichuan Mo, Dongxian Wu, Yifei Wang, Yiwen Guo, Yisen Wang, When Adversarial Training Meets Vision Transformers: Recipes from Training to Architecture, NeuIPS 2022.

[10] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, , Training data-efficient image transformers & distillation through attention, CVPR 2021.

[11]Jianwei Yang, Chunyuan Li, Pengchuan Zhang, Xiyang Dai, Bin Xiao, Lu Yuan, Jianfeng Gao. Focal Self-attention for Local-Global Interactions in Vision Transformers. arXiv:2107.00641