Understanding Adversarial Robustness of Vision Transformers via Cauchy Problem*

Zheng Wang[1 [0000−0001−7146−7503]] and Wenjie Ruan ✉[1 [0000−0002−8311−8738]]

University of Exeter, Exeter EX4 4PY, UK
{zw360, W.Ruan}@exeter.ac.uk

Abstract. Recent research on the robustness of deep learning has shown that *Vision Transformers (ViTs)* surpass the *Convolutional Neural Networks (CNNs)* under some perturbations, e.g., natural corruption, adversarial attacks, etc. Some papers argue that the superior robustness of ViT comes from the segmentation on its input images; others say that the *Multi-head Self-Attention (MSA)* is the key to preserving the robustness [30]. In this paper, we aim to introduce a principled and unified theoretical framework to investigate such argument on ViT's robustness. We first theoretically prove that, unlike *Transformers* in *Natural Language Processing*, ViTs are *Lipschitz* continuous. Then we theoretically analyze the adversarial robustness of ViTs from the perspective of *Cauchy Problem*, via which we can quantify how the robustness propagates through layers. We demonstrate that the first and last layers are the critical factors to affect the robustness of ViTs. Furthermore, based on our theory, we empirically show that unlike the claims from existing research, MSA only contributes to the adversarial robustness of ViTs under weak adversarial attacks, e.g., *FGSM*, and surprisingly, MSA actually comprises the model's adversarial robustness under stronger attacks, e.g., *PGD attacks*. We release our code via https://github.com/TrustAI/ODE4RobustViT

Keywords: Adversarial Robustness · Cauchy Problem · Vision Transformer

1 Introduction

Since *Transformers* have been transplanted from *Natural Language Processing (NLP)* to *Computer Vision (CV)*, great potential has been revealed by *Vision Transformers* for various CV tasks [19]. It is so successful that some papers even argue that CNNs are just a special case of ViTs [9]. Recently, the robustness of ViTs has been studied, for example, some research showed that ViT has superior robustness than CNNs under natural corruptions [31]. Very recently,

* W. Ruan is the corresponding author. This work is supported by Partnership Resource Fund (PRF) on Towards the Accountable and Explainable Learning-enabled Autonomous Robotic Systems from UK EPSRC project on Offshore Robotics for Certification of Assets (ORCA) [EP/R026173/1].

some researchers have also begun to investigate the robustness of ViTs against *adversarial perturbations* [26].

However, existing research on adversarial robustness for ViTs mainly focuses on *adversarial attacks*. The main idea is to adopt the attacks on CNNs to ViTs, e.g., *SAGA* [26] and *IAM-UAP* [18]. Meanwhile, some pioneering studies demonstrate that ViTs are more robust than CNNs against *adversarial patch attacks*, arguing that the *dynamic receptive field* of MSA is the key factor to its superior robustness [30]. On the other hand, some others argue that the tokenization of ViTs plays an essential role in adversarial robustness [1]. While some researchers say the patch embedding method is a critical factor to contribute the adversarial robustness of ViTs [28]. However, most existing works concerning the superior robustness of ViTs are purely based on empirical experiments in an ad-hoc manner. A principled and unified theoretical framework that can quantify the adversarial robustness of ViT is still lacking in the community.

In our paper, instead of analyzing the robustness of *Vision Transformer* purely based on empirical evidence, a theoretical framework has been proposed to examine whether MSA contributes to the robustness of ViTs. Inspired by the fact that ViTs and ResNets share a similar structure of residual additions, we show that, ViTs, under certain assumptions, can be regarded as a *Forward Euler* approximation of the underlying *Ordinary Differential Equations (ODEs)* defined as

$$\frac{d\boldsymbol{x}}{dt} = \mathcal{F}(\boldsymbol{x}, t).$$

With this approximation, each block in transformer can be modeled as the nonlinear function $\mathcal{F}(\boldsymbol{x})$. Based on the assumption that function $\mathcal{F}(\boldsymbol{x})$ is Lipschitz continuous, we then can theoretically bridge the adversarial robustness with the *Cauchy Problem* by first-order *Taylor expansion* of $\mathcal{F}(\boldsymbol{x})$. With the proposed theoretical framework, this paper is able to quantify how robustness is changing among each block in ViTs. We also observe that the first and last layers are vital for the robustness of ViTs.

Furthermore, according to our theoretical and empirical studies, different to the existing claim made by Naseer et al. [30] that MSA in ViTs strengthens the robustness of ViTs against patch attacks. We show that MSA in ViTs is *not always* improving the model's adversarial robustness. Its strength to enhance the robustness is minimal and even comprises the adversarial robustness against strong L_p norm adversarial attacks. In summary, the key contributions of this paper are listed below.

1. To our knowledge, this is the first work to formally bridge the gap between the robustness problem of ViTs and the *Cauchy problem*, which provides a principled and unified theoretical framework to quantify the robustness of transformers.
2. We theoretically prove that ViTs are Lipschitz continuous on vision tasks, which is an important requisite to building our theoretical framework.

3. Based on our proposed framework, we observe that the first and last layers in the encoder of ViTs are the most critical factors to affect the robustness of the transformers.
4. Unlike existing claims, surprisingly, we observe that MSA can only improve the robustness of ViTs under weak attacks, e.g., *FGSM attack*, and it even comprises the robustness of ViTs under strong attacks, e.g., *PGD attack*.

2 Related Work

2.1 Vision Transformers and Its Variants

To the best of our knowledge, the first work using the transformer to deal with computer vision tasks is done by Carion et al. [6], since then, it has quickly become a research hotspot, though it has to be pre-trained on a larger dataset to achieve comparable performance due to its high complexity and lack of ability to encode local information. To reduce the model complexity, *DeiT* [36] leverages the *Knowledge Distillation* [17] technique, incorporating information learned by *Resnets* [15]; *PvT* [37] and *BoTNet* [34] adopt more efficient backbones; *Swim Transformer* [24] and *DeepViT* [38] modifies the MSA. Other variants, e.g., *TNT*, *T2T-ViT*, *CvT*, *LocalViT* and *CeiT* manage to incorporate local information to the ViTs [19].

2.2 Robustness of Vision Transformer

Many researchers focus on the robustness of ViTs against *natural corruptions* [16] and empirically show that ViTs are more robust than CNNs [31]. The adversarial robustness of ViTs has also been empirically investigated. Compared with CNNs and MLP-Mixers under different attacks, it claims that for most of the white-box attacks, some black-box attacks, and Universal Adversarial Perturbations (UAPs) [29], ViTs show superior robustness [30]. However, ViTs are more vulnerable to simple FGSM attacks [5]. The robustness of variants of ViTs is also investigated and shown that the local window structure in Swim-ViT harms the robustness and argues that positional embedding and the *completeness/compactness* of heads are crucial for performance and robustness [27].

However, the reason for the superior robustness of ViTs is rarely investigated. Most of the research concentrate on frequency analysis [31]. Benz et al. argue that shift-invariance property [4] harms the robustness of CNNs. Naseer et al. say the flexible receptive field is the key to learning more shape information which strengthens the robustness of ViTs by studying the severe occlusions [30]. And yet Mao et al. argue that ViTs are still overly reliant on the texture, which could harm their robustness against out-of-distribution data [27]. Qin et al. investigate the robustness from the perspective of robust features and argue ViTs are insensitive to patch-level transformation, which is considered as non-robust features [32].

2.3 Deep Neural Network via Dynamic Point of View

The connection between differential equations and neural networks is first introduced by S. Grossberg [14] to describe a continuous additive RNN model. After ResNet had been proposed, new relations appeared that regard forward prorogation as Euler discretization of the underlying ODEs [33]. And many variants of ResNets can also be analyzed in the framework of numerical schemes for ODEs, e.g., PolyNet, FracalNet, RevNet and LMResNet [25]. Instead of regarding neural networks as discrete methods, Neural ODE has been proposed [7], replacing the ResNet with its Underlying ODEs, and the parameters are calculated by a black-box ODE solver. However, E. Dupont et al. [11] argue that neural ODEs hardly learn some representations. In addition to ODEs, PDEs and even SDEs are also involved in analyzing the Neural Network [35].

3 Preliminaries

The original ViTs are generally composed of *Patch Embedding*, *Transformer Block* and *Classification Head*. We follow the definition from [10]. Let $\boldsymbol{x} \in \mathbf{R}^{H\times W\times C}$ stands for the input image. Hence, Each image is divided equally into a sequence of $N = HW/P^2$ patches, and each one is denoted as $\boldsymbol{x}_p \in \mathbf{R}^{N\times(P^2\cdot C)}$.

$$\begin{aligned}
\boldsymbol{z}_0 &= [\boldsymbol{x}_{class}, \boldsymbol{x}_p^1\boldsymbol{E}, \boldsymbol{x}_p^2\boldsymbol{E}, ..., \boldsymbol{x}_p^N\boldsymbol{E}] + \boldsymbol{E}_{pos},\\
\boldsymbol{z}_l^{'} &= MSA(LN(\boldsymbol{z}_{l-1})) + \boldsymbol{z}_{l-1},\\
\boldsymbol{z}_l &= MLP(LN(\boldsymbol{z}_l^{'})) + \boldsymbol{z}_l^{'},\\
\boldsymbol{y} &= LN(\boldsymbol{z}_L^0),
\end{aligned}$$

where $\boldsymbol{E} \in \mathbb{R}^{P^2\cdot C\times D}, \boldsymbol{E}_{pos} \in \mathbb{R}^{(N+1)\times D}$ and $l = 1, 2, ..., L$. LN denotes *Layer Normalization*, MSA is *Multihead Self-Attention* and MLP represents *Multilayer Perceptron*. MSA is the concatenation of *Self-Attentions (SA)* before linear transformation by $W^{(O)} \in \mathbb{R}^{D\times D}$ defined by

$$MHA := \left(SA_1\ SA_2\ ...\ SA_H\right) W^{(O)},$$

where H is the number of heads and SA is defined by

$$SA := P\boldsymbol{z}W^{(V)} = softmax\left(\boldsymbol{z}W^{(Q)}W^{(K)T}\boldsymbol{z}^T\right)W^{(V)},$$

where $W^{(Q)}, W^{(K)}, W^{(V)} \in \mathbb{R}^{D\times(D/H)}$, and $\boldsymbol{z} \in \mathbb{R}^{N\times D}$ defines the inputs of transformers.

4 Theoretically Analysis

4.1 Vision Transformers are Lipschitz

To model the adversarial robustness to Cauchy Problem, we first prove that ViTs are Lipschitz functions. Unlike the conclusion drawn by Kim et al. [21]

that Dot-product self-attention is not Lipschitz, it can be proved that Vision Transformers are Lipchitz continuous since inputs are bounded between $[0, 1]$. We follow the same definition from [21] that a function $f : \mathcal{X} \to \mathcal{Y}$ is called Lipschitz continuous if $\exists K \geq 0$ such that $\forall \boldsymbol{x} \in \mathcal{X}, \boldsymbol{y} \in \mathcal{Y}$ we have

$$d_{\mathcal{Y}}(f(\boldsymbol{x}), f(\boldsymbol{x}_0)) \leq K d_{\mathcal{X}}(\boldsymbol{x}, \boldsymbol{x}_0), \tag{1}$$

where $(\mathcal{X}, d_{\mathcal{X}}), (\mathcal{Y}, d_{\mathcal{Y}})$ are given metric spaces, and given p-norm distance, the Lipschitz constant K is given by

$$Lip_p(f) = \sup_{\boldsymbol{x} \neq \boldsymbol{x}_0} \frac{\|f(\boldsymbol{x}) - f(\boldsymbol{x}_0)\|_p}{\|\boldsymbol{x} - \boldsymbol{x}_0\|_p}. \tag{2}$$

Similar to the analysis by Kim et al. [20], since Linear transformation by $W^{(V)}$ is Lipchitz and does not impact our analysis, we will drop it and focus on the non-linear part of $P\boldsymbol{z}$.

Since Patch embeddings are conducted by convolutional operations and the classification heads are fully connected layers, they are Lipchitz continuous [20]. Therefore as long as the transformer blocks are Lipschitz continuous, ViTs are Lipschitz continuous because the composite Lipchitz functions, i.e., $f \circ g$, are also Lipschitz continuous [12]. To this end, we have the Theorem (1).

Theorem 1. *(Transformer Blocks in ViTs are Lipschitz continuous) Given vision transformer block with trained parameters $\boldsymbol{w}$ and convex bounded domain $\mathcal{Z}_{l-1} \subseteq \mathbb{R}^{N \times D}$, we show that the transformer block $\mathcal{F}_l : \mathcal{Z}_{l-1} \to \mathbb{R}^{N \times D}$ mapping from z_{l-1} to z_l is Lipchitz function for all $l = 1, 2, ..., L$.*

Proof. For simplicity, we only prove the case that the number of heads H and the dimension of patch embedding D are all equal to 1. The general case can be found in Appendix.

Because the composition of the transformer block includes an MLP layer that is Lipchitz continuous, as argued by Kim et al. [20], it is the non-linear part of MSA that need to be proved Lipchitz continuous. We formulated the non-linear part as mapping $f : \mathcal{Z} \to \mathbb{R}^{N \times 1}$ shown in Equation (3)

$$f(\boldsymbol{z}) = softmax(a\boldsymbol{z}\boldsymbol{z}^T)\boldsymbol{z} = P\boldsymbol{z} = \begin{pmatrix} p_1(\boldsymbol{u}_1) & \cdots & p_N(\boldsymbol{u}_1) \\ \vdots & \ddots & \vdots \\ p_1(\boldsymbol{u}_N) & \cdots & p_N(\boldsymbol{u}_N) \end{pmatrix} \boldsymbol{z} \tag{3}$$

where $a = W^{(Q)}W^{(K)T} \in \mathbb{R}$, $\boldsymbol{z} \in \mathcal{Z}$ which is a bounded convex set and belongs to $\mathbb{R}^{N \times 1}$, P is defined by *softmax* operator. Each row in P defines a discrete probability distribution. Therefore P can be regarded as the *transition matrix* for a finite discrete *Markov Chain* with $z_1, ..., z_n$ as observed value for random valuables. Since f has continuous first deviates, *Mean Value Inequality* can be used to find Lipchitz constant. Let $\boldsymbol{z}, \boldsymbol{z}_0 \in \mathcal{Z}$ and $\|\cdot\|_p$ denote the p-norm distance

for vectors and *induced norm* for matrices. Specifically, when $p = 2$, the *induce norm* coincides with *spectral norm*, then we have

$$\|f(\boldsymbol{z}) - f(\boldsymbol{z}_0)\|_2 \leq \|J_f(\boldsymbol{\xi})\|_2 \|(\boldsymbol{z} - \boldsymbol{z}_0)\|_2, \tag{4}$$

where $\xi \in \mathcal{Z}$ is on the line through $\boldsymbol{x}$ and $\boldsymbol{x}_0$, and $J_f(\cdot)$ denotes the *Jacobian* of f. As long as the Jacobian J_f is bounded for $\mathcal{Z}$, f is Lipschitz continuous. The Jacobian J_f is shown in Equation (5) (see detail in Appendix).

$$J_f(\boldsymbol{z}) = a\big\{diag(\boldsymbol{z})Pdiag(\boldsymbol{z}) - Pdiag(\boldsymbol{z})diag(\boldsymbol{\mu}) + diag(\boldsymbol{\sigma}^2)\big\} + P, \tag{5}$$

where $\boldsymbol{\mu} = P\boldsymbol{z}$ define the mean vector for the Finite Markov Chain and the *variance* are defined by

$$\boldsymbol{\sigma}^2 = \begin{pmatrix} \sum_{k=1}^{N} p_k(\boldsymbol{u}_1)x_k^2 - \big(\sum_{k=1}^{N} p_k(\boldsymbol{u}_1)x_k\big)^2 \\ \vdots \\ \sum_{k=1}^{N} p_k(\boldsymbol{u}_N)x_k^2 - \big(\sum_{k=1}^{N} p_k(\boldsymbol{u}_N)x_k\big)^2 \end{pmatrix} = \begin{pmatrix} \sigma_1^2 \\ \vdots \\ \sigma_N^2 \end{pmatrix}, \tag{6}$$

Since every component on the right-hand-side in (5) is bounded since $\boldsymbol{z}$ is bounded. We conclude that $J_f(\boldsymbol{z})$ is also bounded, therefore the Lipchitz continuous.

Remark 1. The use of the Mean Value Theorem requires the domain $\mathcal{Z}$ to be convex, however as long as $\mathcal{Z}$ is bounded, we can always find a larger convex set $\mathcal{Z}' \supseteq \mathcal{Z}$.

Remark 2. Different from the conclusion drawn by Kim et al. [20] that the transformer is not Lipschitz continuous, ViTs are Lipschitz continuous due to the bounded input.

4.2 Model Adversarial Robustness as Cauchy Problem

Since there exists the *Residual Structure* in the *Transformer Encoder*, just like *ResNet*, which can be formulated as *Euler Method* [25], the *forward propagation* through Transformer Encoder can also be regarded as a *Forward Euler Method* to approximate the underlying *Ordinary Differential Equations (ODEs)*.

Let $f : \mathcal{X} \to \mathcal{Y}$ denote the ViTs, where $\mathcal{X} \subseteq \mathbb{R}^n$ denotes the input space and $\mathcal{Y} = \{1, 2, ..., C\}$ refers to the labels, and $\mathcal{F}_i, i = 1, ..., L$ denote the basic blocks. Notice that for simplicity, let $\mathcal{F}_1(\boldsymbol{x}_0; \boldsymbol{w}_0) + \boldsymbol{x}_0$ refer to the patch embedding and $\mathcal{F}_L(\boldsymbol{x}_{L-1}; \boldsymbol{w}_{L-1}) + x_{L-1}$ be the classification head, the rest are transformer blocks. Hence, the forward propagation can be described in Equation (7).

$$\begin{cases} \boldsymbol{x}_k & = \mathcal{F}_k(\boldsymbol{x}_{k-1}; \boldsymbol{w}_{k-1}) + \boldsymbol{x}_{k-1}, k = 1, ..., L \\ \boldsymbol{y}_{logit} & = softmax(LP(\boldsymbol{x}_L)) \\ \boldsymbol{y} & = \arg\max_{\mathcal{Y}} \boldsymbol{y}_{logit}, \end{cases} \tag{7}$$

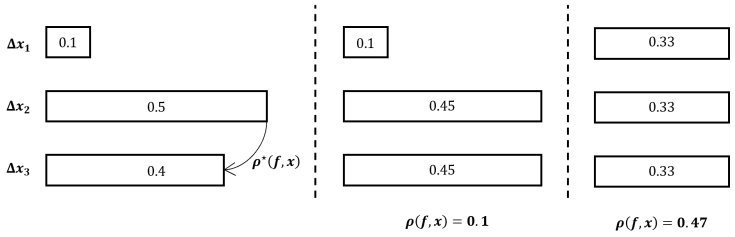

Fig. 1. Illustration of $\rho^\star(f, x)$. For better illustration L_1-norm is taken while calculating the $\rho(f, x)$.

where $\boldsymbol{x}_0 \in \mathcal{X}$, $LP(\cdot)$ stands for *Linear Projection*, $\boldsymbol{y}_{logit}$ shows the likelihood for each class and $\boldsymbol{y} \in \mathcal{Y}$ denote the classification result. As argued by Liao, et al. [23], the Transformer blocks in Equation (7) can be regarded as *Forward Euler* approximation of the underlying ODE shown below.

$$\frac{d}{dt}\boldsymbol{x}(t) = \mathcal{F}(\boldsymbol{x}, t), t \in [t_0, T] \tag{8}$$

where $\mathcal{F}(\cdot)$ corresponds to the basic blocks in ViTs and $t \in [t_0, T]$ refers to the continuous indexing of those blocks.

The backward-propagation of Equation (8) can be regarded as an estimation problem for parameters $\boldsymbol{w}$ of given boundary conditions defined by $\mathcal{X}$ and $\mathcal{Y}$, which leads to Neural ODEs [7].

Before the main theorem that models the adversarial robustness as Cauchy problem, we first define the adversarial robustness metrics. Given neural network f, and the fixed input $\boldsymbol{x} \in \mathcal{X}$, the *local Adversarial Robustness* proposed by Bastani et al. [3] is defined as

$$\rho(f, \boldsymbol{x}) \stackrel{def}{=} \inf\{\epsilon > 0 | \exists \hat{\boldsymbol{x}} : \|\hat{\boldsymbol{x}} - \boldsymbol{x}\| \leq \epsilon, f(\hat{\boldsymbol{x}}) \neq f(\boldsymbol{x})\},$$

where $\|\cdot\|$ defines the general L_p norm. Usually, p is taken as $1, 2$ and ∞. The adversarial robustness is defined as the minimum radius that the classifier can be perturbed from their original corrected result. As illustrated in figure (1), considering the the fact that even in the final larger $\Delta\boldsymbol{x}(T)_1 < \Delta\boldsymbol{x}(T)_2$, it is still possible that $softmax(LP(\boldsymbol{x}(T) + \Delta\boldsymbol{x}(T)_1))$ has been perturbed but $softmax(LP(\boldsymbol{x}(T) + \Delta\boldsymbol{x}(T)_2))$ is not, we use the minimal distortion to define the robustness as

$$\rho^\star(f, \boldsymbol{x}) \stackrel{def}{=} \inf_{\|\hat{\boldsymbol{x}}(T) - \boldsymbol{x}(T)\|} \rho(f, \boldsymbol{x}), \tag{9}$$

where $\hat{\boldsymbol{x}}(T) - \boldsymbol{x}(T) = \Delta\boldsymbol{x}(T)$.

Lemma 1. *(Existence and Uniqueness for the Solution of Underlying ODE) Since the continuous mapping $\mathcal{F}$ defined in ODE* (8) *satisfies the Lipschitz condition on $\boldsymbol{z} \in \mathcal{Z}$ for $t \in [t_0, T]$ as claimed in Theorem* (1)*, where $\mathcal{Z}$ is a bounded*

closed convex set. There exists and only exists one solution for the underlying ODE defined in (8).

Lemma 2. *(Error Bound for Forward Euler approximation) Given Forward Euler approximation shown in Equation* (7) *and its underlying ODE in Equation* (8)*. Let* $K > 0$ *denotes the Lipschitz constant for the underlying ODE, and* $\|\hat{\mathcal{F}}(x,t) - \mathcal{F}(x,t)\| \leq \delta$*, hence the error of solution is given by*

$$\|\Delta \boldsymbol{x}\| \leq \frac{\delta}{K}(e^{K|t-t_0|} - 1)$$

Since $\mathcal{F}(x,t)$ is continuous, δ can be arbitrary small as long as step for Euler approximation is small enough, namely, neural network is deep enough. The proof of lemma (1) and (2) can be found in [8].

Theorem 2. *Let* f *and* g *be two neural networks defined in Equation* (7)*, which have the underlying ODEs as shown in Equation* (8)*, and denote the basic blocks of* g *as* $\mathcal{G}_k, k = 1, ..., L^{'}$ *with its corresponding ODE defined as* $\mathcal{G}$ *to show the difference. Given point* $\boldsymbol{x} \in \mathcal{X}$ *and robustness metric* $\rho^{\star}(\cdot)$ *defined in* (9)*, classifier* f *is more robust than* g*, such that*

$$\rho^{\star}(f, \boldsymbol{x}) \leq \rho^{\star}(g, \boldsymbol{x}), \tag{10}$$

if $\forall t \in [t_0, T]$

$$\sigma_{max}(J_{\mathcal{F}}(t)) \leq \sigma_{max}(J_{\mathcal{G}}(t)) \tag{11}$$

where $J_f(t)$ *and* $J_g(t)$ *refers to the Jacobian of the basic blocks* $\mathcal{F}$ *and* $\mathcal{G}$ *w.r.t.* $\boldsymbol{x}$ *and* $\sigma_{max}(\cdot)$ *denotes the largest singular value .*

Proof. Consider 2 solutions $\boldsymbol{x}(t), \hat{\boldsymbol{x}}(t)$ of ODE defined in (8) such that for $\epsilon > 0$

$$\|\hat{\boldsymbol{x}}(t_0) - \boldsymbol{x}(t_0)\|_2 \leq \epsilon$$

and let $\Delta \boldsymbol{x}(t) = \hat{\boldsymbol{x}}(t) - \boldsymbol{x}(t), t \in [t_0, T]$ hence

$$\frac{d}{dt}\Delta \boldsymbol{x} = \mathcal{F}(\hat{\boldsymbol{x}}, t) - \mathcal{F}(\boldsymbol{x}, t) = J_{\mathcal{F}}(t)\Delta \boldsymbol{x} + \boldsymbol{r}_{\mathcal{F}}(\Delta \boldsymbol{x}), \tag{12}$$

where $\boldsymbol{r}_{\mathcal{F}}(\Delta \boldsymbol{x})$ is the residual of *Taylor Expansion* of $\mathcal{F}$ w.r.t. $\boldsymbol{x}$, such that $\|\boldsymbol{r}_{\mathcal{F}}(\Delta \boldsymbol{x})\| = \mathcal{O}(\|\Delta \boldsymbol{x}\|^2)$ [2]. Instead of $\Delta \boldsymbol{x}$, $\|\Delta \boldsymbol{x}\|_2$ is more of our interest, hence

$$\frac{d}{dt}\|\Delta \boldsymbol{x}\|_2 \leq \|\frac{d}{dt}\Delta \boldsymbol{x}\|_2 \leq \|J_{\mathcal{F}}(t)\|_2\|\Delta \boldsymbol{x}\|_2 + \mathcal{O}(\|\Delta \boldsymbol{x}\|_2^2), \tag{13}$$

since $\|\Delta \boldsymbol{x}(t_0)\|_2 = 0$ is trivial, we assume $\|\Delta \boldsymbol{x}(t_0)\|_2 > 0$. And because there exist unique solution for the ODE system, we have $\|\Delta \boldsymbol{x}(t)\|_2 > 0, t \in [t_0, T]$ therefore Equation (13) becomes

$$\frac{1}{\|\Delta \boldsymbol{x}\|_2}\frac{d}{dt}\|\Delta \boldsymbol{x}\|_2 \leq \|J_{\mathcal{F}}(t)\|_2 + \mathcal{O}(\|\Delta \boldsymbol{x}\|_2). \tag{14}$$

After integral of the both sides from t_0 to T we have

$$\int_{t_0}^{T} \frac{1}{\|\Delta \boldsymbol{x}\|_2} d\|\Delta \boldsymbol{x}\|_2 \leq \int_{t_0}^{T} \|J_{\mathcal{F}}(t)\|_2 + M\epsilon dt,$$

where $M > 0$ is a given large number. The integral for $[t_0, T]$ is given by

$$\|\Delta \boldsymbol{x}(T)\|_2 \leq \epsilon e^{\int_{t_0}^{T} \|J_{\mathcal{F}}(t)\|_2 dt + (T-t_0)M\epsilon}. \tag{15}$$

It is obvious that the perturbed output of neural network $\Delta \boldsymbol{x}(T)$ is actually bounded by the right-hand-side of Equation (15) which is determined by the $\|J_{\mathcal{F}}(t)\|_2, t \in [t_0, T]$, namely the largest *singular value* of $J_{\mathcal{F}}(t)$, denoted as $\sigma_{max}(J_{\mathcal{F}}(t))$. The rest of the proof is simple, since if $\forall t \in [t_0, T]$ (11) holds and $(T - t_0)M\epsilon$ is negligible, we have

$$\|\Delta \boldsymbol{x}_{\mathcal{F}}(T)\|_2 \leq \epsilon e^{\int_{t_0}^{T} \|J_{\mathcal{F}}(t)\|_2 dt} \leq \epsilon e^{\int_{t_0}^{T} \|J_{\mathcal{G}}(t)\|_2 dt},$$

therefore for any $\|\Delta \boldsymbol{x}_{\mathcal{F}}(t_0)\|_2 \leq \rho^{\star}(g, \boldsymbol{x})$ the classification result will also not change for f, hence the Equation (10).

Remark 3. Theorem (2) is particularly useful for adversarial perturbation since the approximation in Equation (15) relies on the narrowness of ϵ. If it is too large, the first-order approximation may fail.

Remark 4. Theorem (2) assumes that the approximation error induced in lemma (2) is small enough to neglect. For very shallow models, e.g., ViT-S_1, ViT-S_2, the relation is violated, as is shown in Table (2).

5 Empirical Study

In order to verify the proposed theorem and find out whether self-attention indeed contributes to the adversarial robustness of ViTs, we replace the self-attention with a 1-D convolutional layer, as shown in figure (2). And we name the modified model *CoViT*, which stands for *Convolutional Vision Transformer*. We use Average Pooling instead of the classification token since the classification token can only learn the nearest few features rather than the whole feature maps for CoViT.

Both ViTs and CoViTs are trained from sketches without any pertaining to ensure that they are comparable. *Sharpness-Aware Minimization (SAM)* [13] optimizer is used throughout the experiments to ensure adequate clear accurate.

5.1 Configuration and Training Result

The configurations for Both ViTs and CoViTs have an input resolution of 224 and embedding sizes of 128 and 512. The use of a smaller embedding size of 128 is to calculate the maximum singular value exactly. An upper bound is calculated

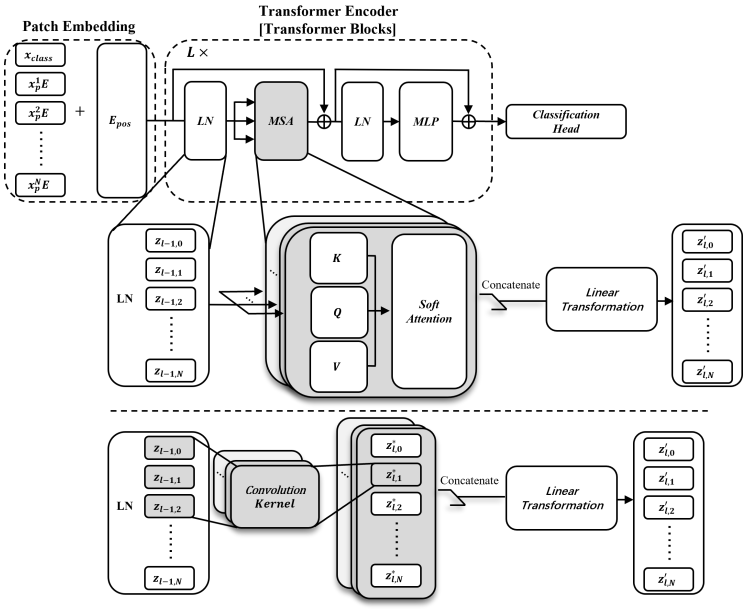

Fig. 2. Illustration of ViT and CoViTs. After *Patch Embedding*, the *Transformer Encoder* is composed of $L\times$ *Transformer Blocks*, of which in each K,Q and V stands for *Key, Query* and *Value* are computed as linear projection from former tokens z_{l-1}, hence *Self-Attention* is calculated as $softmax(\frac{QK^T}{\sqrt{D}})V$. In order to better understand whether self-attention indeed contributes to adversarial robustness, it is replaced by *1-D convolution layers* where different kernel are used and the intermediates, denoted by z_l^*, are generated before concatenation and linearly projecting to $z_l^{'}$. The kernel size can be different for each convolutional projection.

instead for models with a larger embedding size since the exact calculation is intractable. We change the number of heads for ViTs, the kernels for CoViTs, the depth, and the patch size for the experiment. All the models are divided into four groups: S, M, L, T, standing for *Small*, *Medium*, *Large*, and *Tiny* of parameter size. The tiny model uses an embedding size of 128. The detailed configuration is shown in Appendix.

All the models are trained on CIFAR10, and the base optimizer for SAM is SGD with the One-cycle learning scheduler of maximum learning rate equals to 0.1. In order to have a better performance, augmentations, including *Horizontal Flipping, Random Corp* and *Color Jitter*, are involved during training. We resize the image size to the resolution of 224×224. The model with an embedding size of 512 is trained by 150 epochs, and the tiny model with an embedding size of 128 is trained by 300 to achieve adequate performance. The performance of models within the same group is similar, and the shallow networks, e.g., ViT-S_1, ViT-S_2, CoViT-S_1, CoViT-S_2, are harder to train and may need extra training to be converged. This may be due to the optimizer used, i.e., SAM, since SAM

will try to find shallow-wide optima instead of a deep-narrow one, which requires a stronger model capacity.

The experiments are conducted on Nvidia RTX3090 with python 3.9.7, and realized by PyTorch 1.9.1. *Torchattacks* [20] is used for adversarial attacks.

5.2 Study for Small Scale Models

In order to find out whether MSA contributes to the adversarial robustness of ViTs and verify Theorem (2), tiny models with an embedding size of 128 are employed and attacked by L_2-norm PGD-20 and CW. The threshed of successful attacks for CW is set to 260. The corresponding average and standard deviation of the exact maximum singular value for the Jacobian is calculated over layers and images to indicate the overall magnitude of $\sigma_{max}(t)$ over the interval $[t_0, T]$. In other words, we calculate the mean value of $\int_{t_0}^{T} \|J_{\mathcal{F}}(t)\|_2 dt$ for 500 images to indicate the global robustness of the classifier. The PGD-20(L_2) and CW share the same setting with large-scale experiments in Table (2), except that the total iteration for PGD is 20 instead of 7.

Verification of Theorem The result, as shown in Table (1), generally matches our theoretic analysis since the most robust model has the lowest average maximum singular value. It is worth mentioning that the smaller value of $\bar{\sigma}_{max}$ cannot guarantee stronger robustness for ViTs in Table (1), since the standard deviations of σ_{max} are much larger, e.g., 11.66, than that of CoViTs.

Contribution of MSA Another observation is that CoViTs are generally more robust than ViTs. In other words, without enough embedding capacity, Self-Attention could even hurt both the robustness and generalization power. In addition, increasing the models' depth will enhance both generalization power and robustness.

Distribution of Maximum Singular Value in Each Layer In order to know which layer contributes most to the non-robustness, the distribution of σ_{max} is calculated. The layer that has the highest value of σ_{max} may dominate the robustness of the network. As is shown in figure (3), maximum singular values for the CoViTs are much concentrated around the means, reflecting more stable results for classification. And the maximum singular values for the first and last layer of all tiny models are significantly higher than that for in-between layers, indicating that the first and last layers in the transformer encoder are crucial for adversarial robustness.

5.3 Contribution of MSA to Robustness for Large Scale Models

We attack both ViTs and CoViTs with FGSM, PGD, and CW for large-scale models and compare the robust accuracy. And since it is intractable to compute exact maximum singular value for the matrix of size $(128 \cdot 512) \times (128 \cdot 512)$, an upper bound of maximum singular value is calculated as

$$\|J\|_2 \leq \left(\|J\|_1 \|J\|_\infty \right)^{\frac{1}{2}}, \tag{16}$$

Table 1. Attack result and the average maximum singular value for tiny model. All the models have embedding size of 128 with different depth and head or kernels. The cleaning accuracy and the robust accuracy for PGD-20 and CW attack are shown in the Table. The mean of exact maximum singular value σ_{max} for over layers and 500 input images are calculated with standard deviation shown in square. The highest accuracy and lowest maximum singular value are marked in bold.

Net-name	Depth	#Head/ Kernel	Clear Acc.	PGD-20(L_2)	CW(L_2)	$\bar{\sigma}_{max}$
ViT-T_1	4	1	0.819	0.397	0.0305	10.45(4.125)
ViT-T_2	4	4	0.820	0.407	0.039	18.18(11.66)
CoViT-T_1	4	$K3$	0.849	0.492	**0.083**	9.25(0.822)
CoViT-T_2	4	$4 \times K3$	**0.852**	**0.492**	0.036	**9.186(1.088)**
ViT-T_3	8	1	0.836	0.442	0.040	7.17(1.52)
ViT-T_4	8	4	0.834	0.463	0.048	10.03(2.73)
CoViT-T_3	8	$K3$	**0.860**	0.514	**0.076**	**6.413(0.562)**
CoViT-T_4	8	$4 \times K3$	0.860	**0.515**	0.065	6.662(0.386)

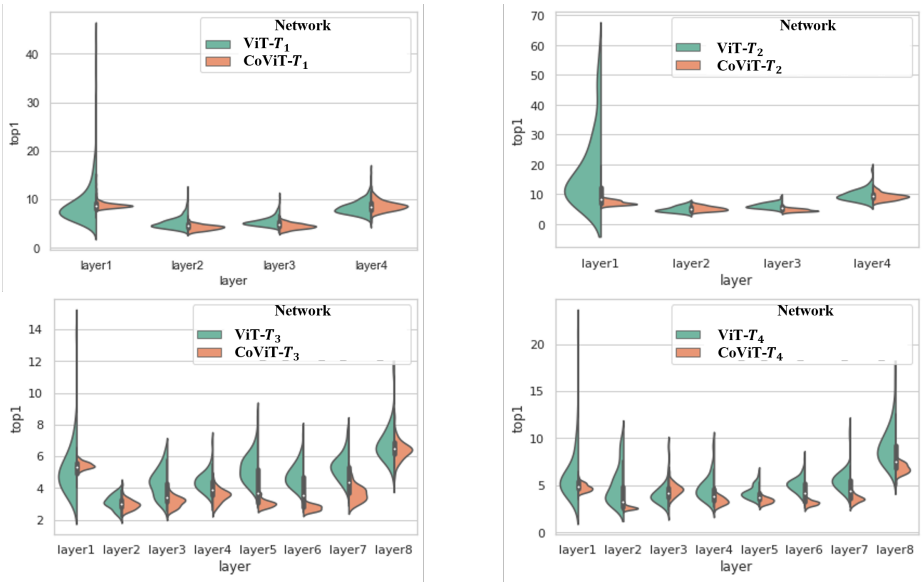

Fig. 3. Violin plot for maximum singular value for each layer of the ViTs/CoViTs. The y-axis shows the maximum singular value.

which is used as an approximation to the maximum singular value of the Jacobian. $\|J\|_1$ and $\|J\|_\infty$ denotes L_1 and L_∞ induced norm for the Jacobian. The mean value for 50 images is taken.

Table 2. Summary of attacking results and corresponding estimated largest singular value. The attacks are employed for both ViTs and CoViTs with FGSM, PGD-7 and CW, and the robust accuracy are shown for each attack. The models with patch size 32×32 are marked with $*$. $\|J\|_1$ and $\|J\|_\infty$ are the L_1 and L_∞ norm respectively. The highest accuracy and lowest estimated maximum singular value are marked in bold.

	Clean Acc.	FGSM	PGD-7(L_∞)	PGD-7(L_2)	CW(L_2)	$(\|J\|_1\|J\|_\infty)^{\frac{1}{2}}$
ViT-S_1	0.676	0.213	0.135	0.267	0.059	812.69
ViT-S_2	0.739	**0.273**	0.162	**0.348**	0.067	1003.20
CoViT-S_1	0.734	0.254	**0.173**	0.341	**0.144**	242.78
CoViT-S_2	0.737	0.244	0.163	0.328	0.143	**206.78**
ViT-S_3	0.847	0.369	0.221	0.444	0.053	296.48
ViT-S_4	0.863	**0.392**	**0.240**	**0.448**	0.065	462.12
CoViT-S_3	0.882	0.320	0.179	0.413	**0.104**	**146.48**
CoViT-S_4	0.876	0.306	0.170	0.401	0.088	150.02
CoViT-S_5	0.868	0.341	0.192	0.424	0.082	163.50
ViT-M_1	0.877	0.415	0.267	0.467	0.049	236.21
ViT-M_2	0.861	0.415	0.260	0.461	0.053	294.06
*ViT-M_3	0.853	**0.478**	**0.356**	**0.519**	0.103	139.23
CoViT-M_1	0.881	0.336	0.185	0.422	0.051	93.21
CoViT-M_2	0.882	0.337	0.197	0.417	0.086	109.79
CoViT-M_3	0.870	0.337	0.194	0.424	0.072	131.94
CoViT-M_4	0.875	0.357	0.208	0.427	0.093	99.57
*CoViT-M_5	0.861	0.416	0.303	0.480	**0.152**	**78.70**
*ViT-L	0.848	0.461	0.347	0.499	0.094	111.38
*CoViT-L_1	0.867	0.443	0.333	0.505	**0.140**	**59.54**
*CoViT-L_2	0.853	**0.466**	**0.357**	**0.528**	0.096	37.26

The *Robust Accuracy* for both ViTs and CoViTs attacked by FGSM, PGD-7 and CW is shown in Table (2). For better comparison, we set $\epsilon = 2/225$ for both FGSM and PDG attack with L_∞ norm. The step size for L_∞ PGD attack is set to $\alpha = 2/255$ and it is iterated only for 7 times to represent the weak attack. The L_2 norm PGD-7 is parameterized by $\epsilon = 2, \alpha = 0.2$. The parameters set for stronger CW attack is that $c = 1$, adversarial confidence level $kappa = 0$, learning rate for Adam [22] optimizer in CW is set to 0.01 and the total iteration number is set to 100.

As is shown in Table (2), for weak attacks, i.e., FGSM and PGD-7, ViTs are generally exhibiting higher robust accuracy within the same group of similar parameter sizes with only a few exceptions. Also, both for ViTs and CoViTs, the robustness is strengthened as the model becomes deeper with more parameters.

For a stronger CW attack, the result is almost reversed, CoViT model shows significantly better robustness and agrees with the approximation of the maximum singular value for the Jacobian. The ability of Self-Attention to avoid per-

turbed pixels is compromised as the attacking becomes stronger. And it seems that the *translation invariance* of CNNs has more defensive power against strong attacks. In addition, a larger patch size always induces better adversarial robustness for both ViTs and CoViTs.

6 Conclusion

This paper first proves that ViTs are Lipschitz continuous for vision tasks, then we formally bridge up the local robustness of transformers with the Cauchy problem. We theoretically proved that the maximum singular value determines local robustness for the Jacobian of each block. Both small-scale and large-scale experiments have been conducted to verify our theories. With the proposed framework, we open the black box of ViTs and study how robustness changes among layers. We found that the first and last layers impede the robustness of ViTs. In addition, unlike existing research that argues MSA could boost robustness, we found that the defensive power of MSA in ViT only works for the large model under weak adversarial attacks. MSA even compromises the adversarial robustness under strong attacks.

7 Discussion and Limitation

The major limitations in this paper are embodied by the several approximations involved. The first one is the approximation of the underlying ODEs to the forward propagation of neural networks with a residual addition structure. As is shown in Lemma (2), the approximation is accurate only when the neural networks are deep enough, and it is hard to know what depth is enough, given the required error bound. One possible way to make it accurate is to consider the *Difference Equation*, which is a discrete parallel theory to ODEs. The second one is the approximation of the second-order term in Equation (14). For small-size inputs, we can say that the L_2-norm of perturbations of adversarial examples is smaller enough so that the second term is negligible. However, the larger inputs may inflate the L_2-norm of perturbations since simply up sampling could result in a larger L_2-norm. Therefore, including the second term or choosing a better norm should be considered. The third approximation is shown in Equation (16). Since the size of the Jacobian depends on the size of the input image, making it impossible to directly calculate the singular value of the Jacobian for larger images, hence, we use an upper bound instead, which inevitably compromises the validation of the experiment. Moreover, since the adversarial attack can only get the upper bound of the minimal perturbations, it is also an approximation of the local robustness, as shown in Table (2).

In the experimental part, we only take into account for the small to moderate size models because it is necessary to rule out the influence of pre-training, and we have to admit that the calculation for the singular value of Jacobian w.r.t. inputs of too large size is hardly implemented.

References

1. Aldahdooh, A., Hamidouche, W., Deforges, O.: Reveal of vision transformers robustness against adversarial attacks. arXiv preprint arXiv:2106.03734 (2021)
2. Ascher, U.M., Mattheij, R.M., Russell, R.D.: Numerical solution of boundary value problems for ordinary differential equations. SIAM (1995)
3. Bastani, O., Ioannou, Y., Lampropoulos, L., Vytiniotis, D., Nori, A., Criminisi, A.: Measuring neural net robustness with constraints. Advances in neural information processing systems 29, 2613–2621 (2016)
4. Benz, P., Ham, S., Zhang, C., Karjauv, A., Kweon, I.S.: Adversarial robustness comparison of vision transformer and mlp-mixer to cnns. arXiv preprint arXiv:2110.02797 (2021)
5. Bhojanapalli, S., Chakrabarti, A., Glasner, D., Li, D., Unterthiner, T., Veit, A.: Understanding robustness of transformers for image classification. In: CVF International Conference on Computer Vision, ICCV. vol. 9 (2021)
6. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European Conference on Computer Vision. pp. 213–229. Springer (2020)
7. Chen, R.T., Rubanova, Y., Bettencourt, J., Duvenaud, D.K.: Neural ordinary differential equations. Advances in neural information processing systems 31 (2018)
8. Coddington, E.A., Levinson, N.: Theory of ordinary differential equations. Tata McGraw-Hill Education (1955)
9. Cordonnier, J.B., Loukas, A., Jaggi, M.: On the relationship between self-attention and convolutional layers. In: International Conference on Learning Representations (2020)
10. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2021)
11. Dupont, E., Doucet, A., Teh, Y.W.: Augmented neural odes. Advances in Neural Information Processing Systems 32 (2019)
12. Federer, H.: Geometric measure theory. Springer (2014)
13. Foret, P., Kleiner, A., Mobahi, H., Neyshabur, B.: Sharpness-aware minimization for efficiently improving generalization. arXiv preprint arXiv:2010.01412 (2020)
14. Grossberg, S.: Recurrent neural networks. Scholarpedia 8(2), 1888 (2013)
15. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. corr abs/1512.03385 (2015) (2015)
16. Hendrycks, D., Zhao, K., Basart, S., Steinhardt, J., Song, D.: Natural adversarial examples. CVPR (2021)
17. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015)
18. Hu, H., Lu, X., Zhang, X., Zhang, T., Sun, G.: Inheritance attention matrix-based universal adversarial perturbations on vision transformers. IEEE Signal Processing Letters 28, 1923–1927 (2021)
19. Khan, S., Naseer, M., Hayat, M., Zamir, S.W., Khan, F.S., Shah, M.: Transformers in vision: A survey. ACM Computing Surveys (CSUR) (2021)
20. Kim, H.: Torchattacks: A pytorch repository for adversarial attacks. arXiv preprint arXiv:2010.01950 (2020)
21. Kim, H., Papamakarios, G., Mnih, A.: The lipschitz constant of self-attention. In: International Conference on Machine Learning. pp. 5562–5571. PMLR (2021)

22. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
23. Liao, Q., Poggio, T.: Bridging the gaps between residual learning, recurrent neural networks and visual cortex. arXiv preprint arXiv:1604.03640 (2016)
24. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10012–10022 (2021)
25. Lu, Y., Zhong, A., Li, Q., Dong, B.: Beyond finite layer neural networks: Bridging deep architectures and numerical differential equations. In: International Conference on Machine Learning. pp. 3276–3285. PMLR (2018)
26. Mahmood, K., Mahmood, R., Van Dijk, M.: On the robustness of vision transformers to adversarial examples. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7838–7847 (2021)
27. Mao, C., Jiang, L., Dehghani, M., Vondrick, C., Sukthankar, R., Essa, I.: Discrete representations strengthen vision transformer robustness. arXiv preprint arXiv:2111.10493 (2021)
28. Mao, X., Qi, G., Chen, Y., Li, X., Duan, R., Ye, S., He, Y., Xue, H.: Towards robust vision transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12042–12051 (2022)
29. Moosavi-Dezfooli, S.M., Fawzi, A., Fawzi, O., Frossard, P.: Universal adversarial perturbations. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1765–1773 (2017)
30. Naseer, M.M., Ranasinghe, K., Khan, S.H., Hayat, M., Shahbaz Khan, F., Yang, M.H.: Intriguing properties of vision transformers. Advances in Neural Information Processing Systems 34, 23296–23308 (2021)
31. Paul, S., Chen, P.Y.: Vision transformers are robust learners. arXiv preprint arXiv:2105.07581 (2021)
32. Qin, Y., Zhang, C., Chen, T., Lakshminarayanan, B., Beutel, A., Wang, X.: Understanding and improving robustness of vision transformers through patch-based negative augmentation. arXiv preprint arXiv:2110.07858 (2021)
33. Ruseckas, J.: Differential equations as models of deep neural networks. arXiv preprint arXiv:1909.03767 (2019)
34. Srinivas, A., Lin, T.Y., Parmar, N., Shlens, J., Abbeel, P., Vaswani, A.: Bottleneck transformers for visual recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16519–16529 (2021)
35. Sun, Q., Tao, Y., Du, Q.: Stochastic training of residual networks: a differential equation viewpoint. arXiv preprint arXiv:1812.00174 (2018)
36. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention. In: International Conference on Machine Learning. pp. 10347–10357. PMLR (2021)
37. Wang, W., Xie, E., Li, X., Fan, D.P., Song, K., Liang, D., Lu, T., Luo, P., Shao, L.: Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 568–578 (2021)
38. Zhou, D., Kang, B., Jin, X., Yang, L., Lian, X., Jiang, Z., Hou, Q., Feng, J.: Deepvit: Towards deeper vision transformer. arXiv preprint arXiv:2103.11886 (2021)