SpecFormer: Guarding Vision Transformer Robustness via Maximum Singular Value Penalization

Xixu Hu[1,2], Runkai Zheng[3], Jindong Wang[2,4](✉), Cheuk Hang Leung[1], Qi Wu[1](✉), Xing Xie[2]

[1] City University of Hong Kong, Kowloon, Hong Kong SAR
[2] Microsoft Research Asia
[3] The Chinese University of Hong Kong (Shenzhen)
[4] William & Mary
qi.wu@cityu.edu.hk, jindong.wang@microsoft.com

Abstract. Vision Transformers (ViTs) are increasingly used in computer vision due to their high performance, but their vulnerability to adversarial attacks is a concern. Existing methods lack a solid theoretical basis, focusing mainly on empirical training adjustments. This study introduces **SpecFormer**, tailored to fortify ViTs against adversarial attacks, with theoretical underpinnings. We establish local Lipschitz bounds for the self-attention layer and propose the **Maximum Singular Value Penalization (MSVP)** to precisely manage these bounds By incorporating MSVP into ViTs' attention layers, we enhance the model's robustness without compromising training efficiency. SpecFormer, the resulting model, outperforms other state-of-the-art models in defending against adversarial attacks, as proven by experiments on CIFAR and ImageNet datasets. Code is released at `https://github.com/microsoft/robustlearn`.

Keywords: Vision Transformer · Adversarial Robustness · Lipschitz Continuity

1 Introduction

Vision Transformer (ViT) [17] has gained increasing popularity in computer vision. Owing to its superior performance, ViT has been widely applied to image classification [57], object detection [8], semantic segmentation [45], and video understanding [2]. More recently, the popular vision foundation models [15, 42] also adopt ViT as the basic module. Unlike CNNs [44], a classical vision backbone, ViT initially divides images into non-overlapping patches and leverages the self-attention mechanism [52] for feature extraction.

Despite its popularity, security concerns related to ViT have recently surfaced as a critical issue. Studies have demonstrated that ViT is vulnerable to malicious attacks [20, 30], which compromises its performance and system security. Adversarial examples, which are created by adding trainable perturbations

to the original inputs to produce incorrect outputs, are one of the major threats in machine learning security. Different attacks, such as FGSM [21], PGD [32], and CW attack [9], have significantly impeded neural networks including both CNNs [22, 44, 46] and ViT [17, 18, 49].

This work aims at improving the adversarial robustness of ViT. Prior arts [4, 6, 35] have empirically investigated the intrinsic robustness of ViT compared with CNNs. [6] found out that when pre-trained with a sufficient amount of data, ViTs are at least as robust as the CNNs [44] on a broad range of perturbations. [4] stated that CNNs can be as robust as ViT against adversarial attacks if they properly adopt Transformers' training recipes. Furthermore, [40, 43] used frequency filters to discover that the success of ViT' robustness lies in its lower sensitivity to high-frequency perturbations. But later works [20, 30, 55] suggest that adding an attention-aware loss to ViTs can manipulate its output, resulting in lower robustness than CNNs.

Other than the robustness analysis of ViT, there are some empirical and theoretical efforts in designing algorithms to enhance its robustness. Empirically, [14] proposed a modified adversarial training recipe of more robust ViT by omitting the heavy data augmentations used in standard training. [33] found that masking gradients from attention blocks or masking perturbations on some patches during adversarial training can greatly improve the robustness of ViT. While prior arts have mainly focused on improvements from an empirical perspective, they do not pay attention to the underlying theoretical principles governing self-attention and model robustness.

Some recent research [41, 54, 62] have investigated the theoretical properties of the self-attention mechanisms in ViTs. However, they focused on studying the stability during training [48, 54] without establishing an explicit link to robustness. While some scattered understandings about the robustness of attention mechanisms do exist, their applicability and reliability remain unclear. Therefore, it is imperative to conduct a comprehensive exploration on the robustness of ViT from *both* the theoretical and empirical perspectives. By doing so, we can possess a more thorough understanding of what influences the robustness of ViT, which is essential for the continued development of powerful and reliable deep learning models, especially large foundation models.

In this work, we propose SpecFormer, a simple yet effective approach to enhance the adversarial robustness of ViT. Concretely, we first provide a rigorous theoretical analysis of model robustness from the perspective of Lipschitz continuity [34]. Our analysis shows that we can easily control the Lipschitz continuity of self-attention by adding additional penalization. SpecFormer seamlessly integrate our proposed Maximum Singular Value Penaliztaion (MSVP) algorithm into each attention layer to help improve model stability. We further adopt the power iteration algorithm [7] to accelerate optimization. Our approach is evaluated on four public datasets, namely, CIFAR-10/100 [27], ImageNet [16] and Imagenette [24] across different ViT variants. Our SpecFormer achieves superior performance in *both* clean and robust accuracy. Under standard training, our approach improves robust accuracy by **8.96**%, **3.49**%, and **2.36**% against

FGSM [21], PGD [32], and AutoAttack [12], respectively. Under adversarial training, SpecFormer outperforms the best counterparts by **1.69**%, **1.26**%, and **0.50**% on CW [9], PGD [32], and AutoAttack [12]. The clean accuracy is also improved by **2.20**%, and **3.06**% on average in both settings.

2 Related Work

2.1 Adversarial Robustness

The early work of adversarial robustness [47] discovered that although deep networks are highly expressive, their learned input-output mappings are fairly discontinuous. Therefore, people can apply an imperceptible perturbation that maximizes the network's prediction error to cause misclassification. Since then, a vast amount of research [21, 36, 38] has been done in adversarial attack [9, 10], defense [39, 56], robustness [61], and theoretical understanding [37, 59].

The problem of overall adversarial robustness can be formulated as a min-max optimization, $\min_{\boldsymbol{\theta}} \max_{\boldsymbol{\delta}} \mathcal{L}(f(\mathbf{x}+\boldsymbol{\delta}), y)$. The solution to the inner maximization problem w.r.t the input perturbations $\boldsymbol{\delta}$ corresponds to generating adversarial samples [21, 32], while the solution to the outer minimization problem w.r.t the model parameters $\boldsymbol{\theta}$ corresponds to adversarial training [3, 51], which is an irreplaceable method for improving adversarial robustness and plays a crucial role in defending against adversarial attacks.

Existing attack approaches focus on deriving more and more challenging perturbations $\boldsymbol{\delta}$ to explore the limits of neural network adversarial robustness. For instance, Fast Gradient Sign Method (FGSM) [21] directly take the sign of the derivative of the training loss w.r.t the input perturbation as $\boldsymbol{\delta} \leftarrow \boldsymbol{\delta} + \alpha \operatorname{sign}\left(\nabla_{\boldsymbol{\delta}} \mathcal{L}(f_\theta(\mathbf{x}+\boldsymbol{\delta}), y)\right)$. Projected Gradient Descent (PGD) method [32] updates the perturbations by taking the sign of the loss function derivative w.r.t the perturbation, projecting the updated perturbations onto the admissible space, and repeating multiple times to generate more powerful attacks:

$$\boldsymbol{\delta} \leftarrow \mathit{\Pi}_\epsilon \left(\boldsymbol{\delta} + \alpha \cdot \operatorname{sign}\left(\nabla_{\boldsymbol{\delta}} \mathcal{L}\left(f_{\boldsymbol{\theta}}(\mathbf{x}+\boldsymbol{\delta}), y\right)\right)\right).$$

2.2 Robustness of Vision Transformer

Regarding the robustness of Vision Transformers, the first question people are curious about is how they compare to the classic CNN structure in terms of robustness. To address this question, numerous papers [3, 6, 43, 47] have conducted thorough evaluations of both visual model structures, analyzed possible causes for differences in robustness, and proposed solutions to mitigate the observed gaps. Building upon these understandings, [62] developed fully attentional networks (FANs) to enhance the robustness of self-attention mechanisms. They evaluate the efficacy of these models in terms of corruption robustness for semantic segmentation and object detection, achieving state-of-the-art results. Additionally, [14] compared the DeiT [49], CaiT [50], and XCiT [1] ViT variants in adversarial training and discovered that XCiT was the most effective.

This discovery sheds light on the idea that Cross-Covariance Attention could be another viable option for improving the adversarial robustness of ViTs.

In terms of the theoretical understanding of the relationship between Transformers and robustness, [62] attribute the emergence of robustness in Vision Transformers to the connection of self-attention mechanisms with information bottleneck theory. This suggests that the stacking of attention layers can be broadly regarded as an iterative repeat of solving an information-theoretic optimization problem, which promotes grouping and noise filtering. [13] analyzed that the norm of the derivative of attention models is directly related to the uniformity of the softmax probabilities. If all attention heads have uniform probabilities, the norm will reach its minimum; if the whole mass of the probabilities is on one element, the norm will reach its maximum. Considering a linear approximation of the attention mechanism, the smaller the derivative norm, the tighter the changes when perturbations are introduced, thus enhancing the robustness of the Transformer. These findings have been supported by another paper [30] that proposes an attention-aware loss to deceive the predictions of Vision Transformers. Their attack strategy exactly misguides the attention of all queries towards a single key token under the control of an adversarial patch, corresponding to the maximum derivative norm case described above.

3 Preliminaries

Denote a clean dataset as $\mathcal{D} = \{\mathbf{x}_i, y_i\}_{i=1}^N$, where $\mathbf{x}$ and y are the input and ground-truth label respectively, and denote the loss function as $\mathcal{L}$. We aim to investigate the robustness of the Vision Transformer (ViT) under two different training paradigms: standard supervised training and adversarial training (AT) [32]. In standard training, we aim to learn a classification function $f_{\boldsymbol{\theta}}$ parameterized by $\boldsymbol{\theta}$, where the optimal parameter is $\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} \mathbb{E}_{(\mathbf{x},y)\in\mathcal{D}}\ \mathcal{L}(f_{\boldsymbol{\theta}}(\mathbf{x}), y)$.

While standard training is shown to be less resilient to adversarial attacks, adversarial training (AT) is a major and effective paradigm for enhancing adversarial robustness. AT aims to improve a model's ability to resist malicious attacks by solving a min-max optimization problem. For the inner optimization, it generates adversarial examples that are as powerful as possible to deceive the model predictions, i.e., $f_{\boldsymbol{\theta}}(\mathbf{x}+\boldsymbol{\delta}) \neq y$. For the outer optimization, it adjusts the model parameters to minimize the classification error caused by the perturbed examples, resulting in enhanced robustness:

$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} \mathbb{E}_{(\mathbf{x},y)\in\mathcal{D}} \max_{\|\boldsymbol{\delta}\|_2 \leq \delta_0} \mathcal{L}\left(f_{\boldsymbol{\theta}}(\mathbf{x}+\boldsymbol{\delta}), y\right), \tag{1}$$

where δ_0 bounds the magnitude of $\boldsymbol{\delta}$ to prevent unintended semantic changes caused by the perturbation.

This work aims to improve the adversarial robustness of ViT in *both* AT and non-AT scenarios. We show that by adding a slight penalization, its robustness can be greatly enhanced in both settings.

4 Theoretical Analysis

Bounding the *global* Lipschitz constant of a neural network is a commonly used method to provide robustness guarantees [11,28]. However, the global Lipschitz bound can be loose because it needs to hold for all points in the input domain, including inputs that are far apart. This can greatly reduce clean accuracy in empirical comparisons [25,32]. Conversely, a local Lipschitz constant bounds the norm of output perturbations for inputs within a small region, typically selected as a neighborhood around each data point. This aligns perfectly with the scenario of adversarial robustness, as discussed in Section 2.1, where perturbations attempt to affect the model's output within a budget constraint. Local Lipschitz bounds are superior because they produce tighter bounds by considering the geometry in a local region, often leading to much better robustness [23, 60].

The core of our study is to bridge the gap between local Lipschitz continuity and adversarial robustness in ViTs. Primary distinctions of ViT lie in the LayerNorm and self-attention layers. The seminal work [26] proved the dot-product is not globally Lipschitz continuous and the LayerNorm is Lipschitz continuous [26]. Therefore, we only need to modify that concept and prove that the dot-product self-attention is local Lipschitz continuous. By utilizing the local Lipschitz continuity, the adversarial robustness can be strengthened.

Definition 1 (Local Lipschitz Continuity). *Suppose $\mathcal{X} \subseteq \mathbb{R}^d$ is open. A function, denoted as f, is considered locally Lipschitz continuous with respect to the p-norm, denoted as $\|\cdot\|_p$, if, for any given point $\mathbf{x}_0$, there exists a positive constant C and a positive value δ_0 such that whenever $\|\mathbf{x} - \mathbf{x}_0\|_p < \delta_0$, the following condition holds:*

$$\|f(\mathbf{x}) - f(\mathbf{x}_0)\|_p \leq C \|\mathbf{x} - \mathbf{x}_0\|_p . \tag{2}$$

The smallest value of C that satisfies the condition is called the local Lipschitz constant of f. From Eq. (2), we observe that a classifier exhibiting local Lipschitz continuity with a small C experiences less impact on output predictions when subjected to budget-constrained perturbations.

In this study, we harness the concept of local Lipschitz continuity to safeguard ViT against malicious attacks. Our primary focus is on ensuring that the attention layer maintains Lipschitz continuity in the vicinity of each input, and we employ optimization objectives to strengthen this property. This strategic approach enables us to bolster the output stability of ViT when facing adversarial attacks. We provide the formal definition of the local Lipschitz constant and the method for its calculation below.

Definition 2 (Local Lipschitz Constant). *The p-local Lipschitz constant of a network $f(\mathbf{x})$ over an open set $\mathcal{X} \subseteq \mathbb{R}^d$ is defined as:*

$$\mathrm{Lip}_p(f, \mathcal{X}) = \sup_{\substack{\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X} \\ \mathbf{x}_1 \neq \mathbf{x}_2}} \frac{\|f(\mathbf{x}_1) - f(\mathbf{x}_2)\|_p}{\|\mathbf{x}_1 - \mathbf{x}_2\|_p}. \tag{3}$$

If f is smooth and p-local Lipschitz continuous over $\mathcal{X}$, the Lipschitz constant can be computed by upper bounding the norm of Jacobian.

Theorem 1 (Calculation of Local Lipschitz Constant [19]). *Let $f : \mathcal{X} \to \mathbb{R}^m$ be differentiable and locally Lipschitz continuous under a choice of p-norm $\|\cdot\|_p$. Let $\mathbf{J}_f(x)$ denote its total derivative (Jacobian) at $\mathbf{x}$. Then,*

$$\mathrm{Lip}_p(f, \mathcal{X}) = \sup_{\mathbf{x} \in \mathcal{X}} \|\mathbf{J}_f(\mathbf{x})\|_p, \tag{4}$$

where $\|\mathbf{J}_f(\mathbf{x})\|_p$ is the induced operator norm on $\mathbf{J}_f(\mathbf{x})$.

The *global* Lipschitz constant, which takes into account the supremum over $\mathcal{X} = \mathbb{R}^d$, must ensure Eq. (2) even for distant $\mathbf{x}$ abd $\mathbf{x}_0$, which can render it imprecise and lacking significance when examining the local behavior of a network around a single input. We concentrate on 2-local Lipschitz constants, where $\mathcal{X} = B_2(\mathbf{x}_0, \delta_0) \coloneqq \{\mathbf{x} : \|\mathbf{x} - \mathbf{x}_0\|_2 \leq \delta_0\}$ represents a small ℓ_2-ball with a radius of δ_0 centered around $\mathbf{x}_0$. The choice of the 2-norm for our investigation is motivated by two key reasons. First, the 2-norm is the most commonly used norm in Euclidean space. Second, and delving further into the rationale, as revealed in [58], the sensitivity of a model to input perturbations is intricately connected to the 2-norm of the weight matrices, which in turn is closely linked to the principal direction of variation and the maximum scale change.

Proposition 1 (Model Sensitivity and Maximum Singular Value in Linear Models [58]). *In a small neighbourhood of $\mathbf{x}_0$, we can regard $f_{\boldsymbol{\theta}}$ as a linear function: $\mathbf{x} \mapsto \mathbf{W}_{\boldsymbol{\theta},\mathbf{x}_0}\mathbf{x} + \boldsymbol{b}_{\boldsymbol{\theta},\mathbf{x}_0}$, whose weights and biases depend on $\boldsymbol{\theta}$ and $\mathbf{x}_0$. For a small perturbation δ, we have*

$$\frac{\|f_{\boldsymbol{\theta}}(\mathbf{x}_0+\boldsymbol{\delta}) - f(\mathbf{x}_0)\|_2}{\|\boldsymbol{\delta}\|_2} = \frac{\left\|\left(\mathbf{W}_{\boldsymbol{\theta},\mathbf{x}_0}(\mathbf{x}+\boldsymbol{\delta}) + \boldsymbol{b}_{\boldsymbol{\theta},\mathbf{x}_0}\right) - \left(\mathbf{W}_{\boldsymbol{\theta},\mathbf{x}_0}\mathbf{x} + \boldsymbol{b}_{\boldsymbol{\theta},\mathbf{x}_0}\right)\right\|_2}{\|\boldsymbol{\delta}\|_2} = \frac{\left\|\mathbf{W}_{\boldsymbol{\theta},\mathbf{x}_0}\boldsymbol{\delta}\right\|_2}{\|\boldsymbol{\delta}\|_2} \leq \sigma_{max}\left(\mathbf{W}_{\boldsymbol{\theta},\mathbf{x}_0}\right). \tag{5}$$

The proposition above suggests that when the maximum singular value of the weight matrices is small, the function $f_{\boldsymbol{\theta}}$ becomes less sensitive to input perturbations, which can significantly enhance adversarial robustness. Inspired by this, we can re-examine the self-attention mechanism as a product of linear mappings and apply the same spirit to enhance the robustness of Vision Transformers.

Re-examining Self-Attention as a Product of Linear Mapping Operations We propose to reconsider self-attention, the fundamental module in ViT as a multiplication of three *linear mappings*, on which the aforementioned proposition on model robustness can be applied. Given an input $\mathbf{X} \in \mathbb{R}^{N \times d}$, the self-attention module in ViT is typically represented as:

$$\mathrm{Attn}\left(\mathbf{X}, \mathbf{W}^Q, \mathbf{W}^K, \mathbf{W}^V\right) = \mathrm{softmax}\left(\frac{\mathbf{X}\mathbf{W}^Q\left(\mathbf{X}\mathbf{W}^K\right)^\top}{\sqrt{D}}\right)\mathbf{X}\mathbf{W}^V, \tag{6}$$

where $\mathbf{W}^Q, \mathbf{W}^K, \mathbf{W}^V \in \mathbb{R}^{d \times D}$ are projection weight matrices corresponding to query, key, and value. N, d, and D stand for the number of tokens, data

dimension, and the hidden dimension of self-attention, respectively. By virtue of the intrinsic nature of self-attention, we conceptualized it as a three-way linear transformation, where the matrices are multiplied together:

$$\text{Attn}\left(\mathbf{X}, \mathbf{W}^Q, \mathbf{W}^K, \mathbf{W}^V\right) = \text{softmax}\left(\frac{\mathbf{X}\mathbf{W}^Q\left(\mathbf{X}\mathbf{W}^K\right)^\top}{\sqrt{D}}\right)\mathbf{X}\mathbf{W}^V = \text{softmax}\left(\frac{h_1(\mathbf{X})h_2(\mathbf{X})^\top}{\sqrt{D}}\right)h_3(\mathbf{X}), \tag{7}$$

where these linear mapping operations are formulated as:

$$h_1(\mathbf{X}) = \mathbf{X}\mathbf{W}^Q,\ h_2(\mathbf{X}) = \mathbf{X}\mathbf{W}^K,\ h_3(\mathbf{X}) = \mathbf{X}\mathbf{W}^V. \tag{8}$$

After reinterpreting the self-attention mechanism as a product of three linear mappings, we are inspired by Prop. 1 to readily discern that we can independently control the maximum singular values of these three weight matrices to imbue the model with adversarial robustness.

Theorem 2. *The self-attention layer is local Lipschitz continuous in* $B_2\left(\mathbf{X}_0, \delta_0\right)$

$$\text{Lip}_{local}(\text{Att}, \mathbf{X}_0) \leq N(N+1)\left(\|X_0\|_F + \delta_0\right)^2\left[\left\|\mathbf{W}^V\right\|_2\left\|\mathbf{W}^Q\right\|_2\left\|\mathbf{W}^{K,\top}\right\|_2 + \left\|\mathbf{W}^V\right\|_2\right], \tag{9}$$

If we make a more stringent assumption that all inputs to the self-attention layer are bounded, i.e., $\mathcal{X} \in \mathbb{R}^{N\times d}$ *represents a bounded open set, then we can determine the maximum of the input* $\mathbf{X}$ *as* $B = \max_{\mathbf{X}\in\mathcal{X}}\|\mathbf{X}\|_F$*. This allows us to establish a significantly stronger conclusion:*

$$\text{Lip}_{local}(\text{Att}, \mathbf{X}_0) \leq N(N+1)\left(B + \delta_0\right)^2\left[\left\|\mathbf{W}^V\right\|_2\left\|\mathbf{W}^Q\right\|_2\left\|\mathbf{W}^{K,\top}\right\|_2 + \left\|\mathbf{W}^V\right\|_2\right]. \tag{10}$$

This framework provides us with theoretical support, allowing us to manage the local Lipschitz constant of the attention layer by controlling the maximum singular values of $\mathbf{W}^Q, \mathbf{W}^K, \mathbf{W}^V$.

Comparison with Existing Bounds The existing literature on introducing Lipschitz continuity into Transformer models comprises three notable articles [13, 26, 41], where we abbreviate their proposed modified Transformers as L2Former [26], LNFormer [13], and LipsFormer [41]. We compare these four models from three perspectives: whether they introduce Lipschitz continuity in Transformers, whether they address the robustness problem, and whether the analysis they provide on the attention mechanism is simplified or not. A detailed summary is presented in Table 1. We can see that our analysis is the most comprehensive. More detailed comparisons can be found in Appendix D.

5 SpecFormer

Inspired by the above analysis, we propose **SpecFormer**, a more robust ViT against adversarial attacks, as illustrated in Fig. 1. SpecFormer employs **Maximum Singular Value Penalization (MSVP)** with an approximation algorithm named power iteration to reduce the computational costs.

Table 1: Comprehensive Comparison of the Baseline Models.

Model	Proposed Mechanism	Lipschitz Continuity	Robustness	Not Simplified
Transformer [17]	$\text{softmax}\left(\frac{\mathbf{x}\mathbf{W}^Q(\mathbf{x}\mathbf{W}^K)^\top}{\sqrt{D}}\right)\mathbf{X}\mathbf{W}^V$	×	×	–
L2Former [26]	$\exp\left(-\frac{\|\mathbf{x}_i^\top W^Q - \mathbf{x}_j^\top W^K\|_2^2}{\sqrt{D/H}}\right)$	✓	×	×
LNFormer [13]	$\frac{Q^\top K}{\max\{uv,uw,vw\}}$	✓	×	×
LipsFormer [41]	$\mathbf{q}_i = \frac{(\mathbf{x}_i^\top \mathbf{W}^Q)^\top}{\sqrt{\|\mathbf{x}_i^\top \mathbf{W}^Q\|^2+\epsilon}}$	✓	×	✓
SpecFormer (Ours)	$\mathcal{L}_{cls} + \lambda \cdot \sigma_{\max}^2(\mathbf{W})$	✓	✓	✓

5.1 Maximum Singular Value Penalization

MSVP aims to enhance the robustness of ViT by adding penalization to the self-attention layers. Concretely speaking, MSVP restricts the Lipschitz constant of self-attention layers by penalizing the maximum singular values of the linear transformation matrices $\mathbf{W}^Q, \mathbf{W}^K$, and $\mathbf{W}^V$. Denote the classification loss as $\mathcal{L}_{cls}$ such as cross-entropy, the overall training objective with MSVP is:

$$\mathcal{J} = \mathcal{L}_{cls} + \mathcal{L}_{msvp} = \mathcal{L}_{cls} + \lambda \cdot \left[\sigma_{\max}^2(\mathbf{W}^Q) + \sigma_{\max}^2(\mathbf{W}^K) + \sigma_{\max}^2(\mathbf{W}^V)\right], \quad (11)$$

where λ is the trade-off hyperparameter. The original Transformer [52] adopts multi-head self-attention to jointly attend to information from different subspaces at different positions. In line with that spirit, MSVP can also be added in a multi-head manner. Incorporating the summation over all the heads and layers, the overall training objective with multi-head MSVP is:

$$\mathcal{J} = \mathcal{L}_{cls} + \mathcal{L}_{msvp} = \mathcal{L}_{cls} + \lambda \sum_{l=1}^{L}\sum_{h=1}^{H}\left[\sigma_{\max}^2(\mathbf{W}_l^{Q,h}) + \sigma_{\max}^2(\mathbf{W}_l^{K,h}) + \sigma_{\max}^2(\mathbf{W}_l^{V,h})\right], \quad (12)$$

where $\mathbf{W}_l^{*,h}$ denotes the h^{th} head in the l^{th} attention layer, and $*$ could be Q, K or V. By constraining the maximum singular value of the weight matrices in the mapping, we can regulate the extent of output variations when subjected to attacks, thereby enhancing the model's robustness. The computation of maximum singular values can be seamlessly integrated into the forward process.

In the next section, we show how to perform efficient computation of maximum singular values using a convergence-guaranteed power iteration algorithm.

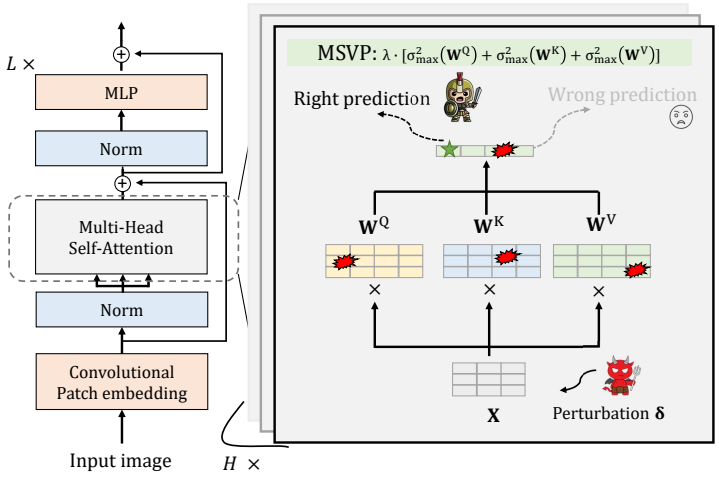


Fig. 1: SpecFormer with MSVP.

Algorithm 1 Detailed algorithm for SpecFormer

▷ Initialization:
for layer $\ell = 1$ to L **do**
  $\boldsymbol{u}_q^\ell, \boldsymbol{v}_q^\ell, \boldsymbol{u}_k^\ell, \boldsymbol{v}_k^\ell, \boldsymbol{u}_v^\ell, \boldsymbol{v}_v^\ell \leftarrow$ initialize approximation vectors for $\mathbf{W}^Q, \mathbf{W}^K, \mathbf{W}^V$.
end for
▷ Forward:
Consider a minibatch $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_k, y_k)\}$ from training data.
for layer $\ell = 1$ to L **do**
  Pass the minibatch through the layers, at the same time
  for head $h = 1$ to H **do**
    for a sufficient number of times **do** ▷ One iteration is adequate
      $\boldsymbol{v}^\ell \leftarrow (W^\ell)^\top \boldsymbol{u}^\ell$, $\boldsymbol{u}^\ell \leftarrow \mathbf{W}^\ell \boldsymbol{v}^\ell$, $\sigma^\ell \leftarrow (\boldsymbol{u}^\ell)^\top \mathbf{W}^\ell \boldsymbol{v}^\ell$
      Add $\lambda(\sigma^\ell)^2$ to maximum singular value loss $\mathcal{L}_{msvp}$.
    end for
  end for
end for
▷ Backward:
Update parameters θ by backward propagation: $\mathcal{L}_{cls} + \mathcal{L}_{msvp}$.

5.2 Power Iteration

Singular value decomposition (SVD) is the most direct way to calculate the maximum singular value. For any real matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$, there exists a singular value decomposition of the form $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$, where $\mathbf{U}$ is an $m \times m$ orthogonal matrix, $\mathbf{\Sigma}$ is an $m \times n$ diagonal matrix and $\mathbf{V}$ is an $n \times n$ orthogonal matrix. However, directly using SVD in MSVP adds $\mathcal{O}(m^2 n + n^3)$ time complexity,

creating a significant computational burden due to ViT's high hidden dimensions and numerous attention layers.

In this section, we propose to adopt the power iteration algorithm as an alternative approach to efficiently calculate the maximum singular values for the $\mathbf{W}^Q, \mathbf{W}^K, \mathbf{W}^V$ weight matrices. The power iteration method [7] is a commonly used approach to approximate the maximum singular value with each iteration taking $\mathcal{O}(mn)$ time. By selecting an initial approximation vector $\boldsymbol{u}$ and $\boldsymbol{v}$, and then performing left and right matrix multiplication with the target matrix, we can obtain a reliable and accurate underestimate of the maximum singular value in a constant number of iterations. Moreover, as the algorithm is iterative in nature, we can incorporate the updates into the model's update process, permitting us to efficiently estimate the maximum singular values with minimal cost. The convergence guarantee and proof for this algorithm can be found in Appendix C. The complete training pipeline of SpecFormer is shown in Alg. 1.

6 Experiments

6.1 Setup

Datasets. We adopt four popular benchmark datasets: CIFAR10,CIFAR100 [27], ImageNet [16], and Imagenette [24]. These datasets are commonly adopted in studies involving the robustness of vision models [33,37]. CIFAR-10 and CIFAR-100 datasets each comprise $60,000$ images, categorized into 10 classes and 100 classes, respectively. ImageNet encompasses over 1.2 million training images and $50,000$ test images, distributed across $1,000$ classes. Imagenette is a subset consisting of 10 classes that are easy to classify, selected from the ImageNet. It is often employed as a suitable proxy [33, 53] for evaluating the performance of models on the more extensive ImageNet.

Baselines. We compare SpecFormer with several notable baselines. Specifically, we evaluate SpecFormer against LipsFormer [41], which introduces Lipschitz continuity into ViT through a novel scaled cosine similarity attention (SCSA) mechanism and by replacing other unstable Transformer components with Lipschitz continuous equivalents. Additionally, we incorporate two other baseline models that implement Lipschitz continuity through different mechanisms: L2 multi-head attention [26] and pre-softmax Lipschitz normalization [13]. However, it is worth noting that the L2 multi-head attention approach requires $\mathbf{W}^{Q,h} = \mathbf{W}^{K,h}$, which is overly restrictive, potentially compromising the model's representation power. We denote the L2 attention and Lipschitz normalization methods as L2Former and LNFormer, respectively.

Implementation details. In line with prior work [33], our evaluation spans across four distinct ViT backbones: the vanilla ViT [17], DeiT [49], ConViT [18] and Swin [29]. This approach enables us to perform a thorough validation and assess the general effectiveness of the MSVP algorithm. To ensure a fair comparison, we adopt the training strategy outlined in [33]. Specifically, we use the SGD optimizer with a weight decay of 0.0001 and a learning rate of 0.1 for 40 epochs.

By default, we apply both CutMix and Mixup data augmentation. We employ FGSM [21] and PGD-2 [32] attacks for standard training, with an attack radius of 2/255. For adversarial training, we use CW-20 [9] and PGD-20 [32] attacks, with an attack radius of 8/255. We also evaluate performance using AutoAttack. In the case of the baseline models (Lipsformer, L2former, and LNFormer), we strictly adhere to the training protocols described in their respective papers.

6.2 Main Results

SpecFormer effectively enhances the adversarial robustness of ViTs with minor modifications. From Table 2, 4, 3, 5 and 6, we can see that our proposed SpecFormer achieves the best results among all the competitors. Specifically, under standard training, our SpecFormer improves over state-of-the-art robust ViT approaches by **8.96**% and **3.49**% in terms of robust accuracy on FGSM and PGD attacks, respectively, on average. Additionally, we enhance accuracy under AutoAttack for larger models by **2.36**%. Furthermore, we improve standard accuracy by **2.20**%.

Using adversarial training, our SpecFormer outperforms the best counterparts by **1.69**% and **1.26**% on CW and PGD attacks, respectively. For AutoAttack, we achieve an improvement of **0.50**% across all datasets and ViT variants, demonstrating stable robustness improvements. Our method also improves the clean accuracy of the ViT model by **3.06**%, significantly enhancing both the model's robustness and its original performance.

We have more observations from the results. 1) Some baseline methods exhibit inferior performance when compared to vanilla ViTs, potentially attributable to the imposition of excessively stringent Lipschitz continuity constraints, which limit the model's expressive capacity. 2) AutoAttack is too powerful for small models under standard training, resulting in an accuracy of 0 in most cases. Therefore, we did not include the results for AA in Table 2. 3) In contrast to other methods that ensure Lipschitz continuity in ViTs through self-attention mechanism modifications, our SpecFormer achieves this by adding a straightforward penalty term to the attention layers, without altering the original self-attention mechanism. SpecFormer's simplicity and versatility allow seamless integration into various ViT architectures, preserving their flexibility.

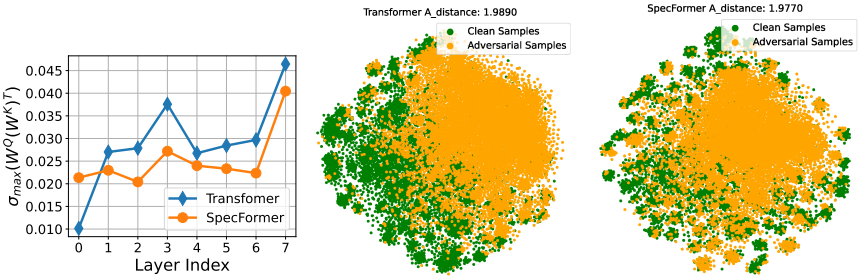
The results on ImageNet in Table 6 following the setup in [33] demonstrate that SpecFormer significantly outperforms its counterparts in both standard and robust accuracy using ViT-B as the backbone. Hyperparameter analysis is provided in Appendix E.1.

6.3 Analyzing the efficacy of MSVP

Analyzing the maximum singular values. We analyze the effectiveness of the proposed MSVP algorithm, which is the core of SpecFormer. Fig. 2a shows the maximum singular values of both ViT and our SpecFormer. It indicates that with minimal additional cost, our SpecFormer effectively limits the maximum singular values across all layers, which are smaller than those of vanilla ViTs.

Table 2: Performance (%) of SpecFormer with different ViT variants on benchmark datasets under *standard* training (using ImageNet-1k pre-trained weights). The best results are in **bold**.

Model	Method	CIFAR-10			CIFAR-100			Imagenette		
		Standard	FGSM	PGD-2	Standard	FGSM	PGD-2	Standard	FGSM	PGD-2
ViT-S	LipsFormer [41]	71.13	31.48	4.17	40.05	9.92	1.36	86.80	36.60	30.20
	L2Former [26]	79.65	39.98	13.39	53.20	15.35	5.92	92.80	58.00	37.80
	LNFormer [13]	75.82	33.72	7.75	48.81	13.04	7.27	92.00	49.00	41.80
	TransFormer [17]	87.09	45.56	22.35	63.52	19.82	7.01	94.20	72.60	48.00
	SpecFormer (Ours)	**88.52**	**50.58**	**29.53**	**69.78**	**23.92**	**9.67**	**97.20**	**84.20**	**61.60**
DeiT-Ti	LipsFormer [41]	72.54	36.14	3.46	39.72	8.66	0.96	79.40	30.60	8.00
	L2Former [26]	78.09	36.64	5.05	49.02	12.63	2.56	82.40	51.00	6.00
	LNFormer [13]	77.16	34.78	3.81	52.50	14.40	**2.83**	80.20	44.80	9.20
	TransFormer [17]	86.40	**46.10**	14.46	62.79	19.89	2.35	90.00	64.20	13.00
	SpecFormer (Ours)	**87.42**	45.71	**18.10**	**64.14**	**20.93**	1.61	**92.20**	**70.20**	**26.20**
ConViT-Ti	LipsFormer [41]	79.71	38.47	7.09	48.08	10.84	1.57	90.60	44.40	24.60
	L2Former [26]	81.33	40.17	12.76	49.86	13.86	2.03	93.40	62.20	30.80
	LNFormer [13]	75.48	28.67	3.56	51.13	11.62	2.41	84.20	36.20	15.00
	TransFormer [17]	**87.78**	**48.89**	20.26	64.68	**22.94**	**4.96**	**93.20**	68.80	**40.80**
	SpecFormer (Ours)	87.49	47.64	**20.53**	**65.57**	21.78	4.10	92.60	**69.00**	28.60
Swin-Ti	LipsFormer [41]	84.77	0.05	35.94	60.60	0.68	14.06	89.20	12.40	30.60
	L2Former [26]	97.86	44.33	**54.69**	88.24	12.76	**29.69**	99.40	66.00	95.20
	LNFormer [13]	96.90	30.96	47.66	84.54	6.42	21.88	99.20	37.40	68.80
	TransFormer [17]	97.51	42.91	53.91	87.89	14.16	25.00	**99.60**	61.40	95.31
	SpecFormer (Ours)	**98.25**	**58.15**	53.64	**89.07**	**16.24**	28.74	99.40	**73.00**	**96.80**



(a) Maximum singular values **(b)** Vanilla ViTs **(c)** SpecFormer

Fig. 2: The analysis of MSVP. (a) Maximum singular value comparison between MSVP and the vanilla Transformer. (b)&(c) tSNE [31] feature visualization.

Hence, our proposed MSVP algorithm can control the maximum singular value of the attention layers, ensuring stability even under extreme perturbations.

Feature visualization. We further illustrate feature visualizations of both vanilla ViT and our SpecFormer in Figures 2b and 2c, employing the t-SNE technique in [31]. Adversarial attacks cause adversarial samples (orange dots) to be misclassified (evidenced by the deviation of green and orange dots within classification clusters, Figure 2b). Better alignment between clean points (green) and adversarial points (orange) within each cluster indicates a more robust model. It is evident that our proposed SpecFormer's features align better than those of

Table 3: Results under **standard training** for larger models.

		CIFAR10				CIFAR100				Imagenette			
		Standard	PGD2	FGSM	AutoAttack	Standard	PGD2	FGSM	AutoAttack	Standard	PGD2	FGSM	AutoAttack
ViT-B	LipsFormer [41]	83.32	13.83	40.52	8.26	59.05	7.97	18.11	4.00	90.80	29.80	32.00	1.20
	L2Former [26]	90.67	18.80	47.01	8.33	75.03	11.71	24.87	5.83	97.60	40.40	77.40	5.20
	LNFormer [13]	86.68	14.20	40.67	5.99	65.90	8.35	17.77	4.09	92.00	31.00	38.80	3.20
	TransFormer [17]	92.97	31.88	55.02	9.12	78.24	14.14	30.05	5.91	98.40	52.60	80.47	11.60
	SpecFormer (Ours)	**96.51**	**52.91**	**60.78**	**10.39**	**85.73**	**27.48**	**34.29**	**6.58**	**99.80**	**58.40**	**83.00**	**20.00**
DeiT-S	LipsFormer [41]	81.85	13.07	36.77	8.54	56.00	6.85	16.50	3.18	87.00	23.80	43.60	3.60
	L2Former [26]	89.81	18.95	42.44	9.47	74.12	11.41	24.21	5.77	93.40	31.00	63.20	9.80
	LNFormer [13]	86.48	13.78	39.32	6.94	67.01	8.40	20.05	4.17	91.40	31.80	54.20	8.40
	TransFormer [17]	92.34	30.54	54.43	16.01	77.17	13.57	28.00	7.05	95.60	44.40	77.40	18.40
	SpecFormer (Ours)	**94.99**	**52.76**	**60.77**	**17.20**	**81.40**	**19.21**	**31.81**	**8.46**	**97.80**	**65.80**	**83.80**	**19.20**
ConViT-S	LipsFormer [41]	89.53	28.08	47.43	18.62	69.82	14.32	23.41	8.32	89.20	29.00	45.00	8.40
	L2Former [26]	92.63	34.88	52.85	18.04	74.03	13.58	25.44	7.50	94.20	42.20	75.20	10.00
	LNFormer [13]	87.84	17.96	42.68	9.54	67.79	8.84	19.29	4.43	89.00	16.20	36.80	6.60
	TransFormer [17]	93.54	38.63	54.38	**22.64**	78.52	19.79	30.38	**11.80**	96.20	49.08	75.40	10.40
	SpecFormer (Ours)	**95.58**	**57.74**	**59.09**	21.01	**82.10**	**24.13**	**33.13**	11.35	**99.20**	**76.20**	**86.60**	**21.40**

Table 4: Performance (%) of SpecFormer with different ViT variants on benchmark datasets under *adversarial* training (using ImageNet-1k pre-trained weights). The best results are in **bold**.

Model	Method	CIFAR-10				CIFAR-100				Imagenette			
		Standard	CW-20	PGD-20	AutoAttack	Standard	CW-20	PGD-20	AutoAttack	Standard	CW-20	PGD-20	AutoAttack
ViT-S	LipsFormer [41]	41.54	24.20	27.50	23.74	24.08	10.47	13.04	9.71	50.00	31.00	34.80	37.00
	L2Former [26]	63.22	33.55	35.78	36.35	37.20	13.69	15.65	18.78	84.80	54.60	54.60	62.20
	LNFormer [13]	56.04	29.92	32.74	30.56	26.37	9.84	11.63	14.55	81.00	46.80	48.80	45.00
	TransFormer [17]	71.76	**34.34**	**35.49**	46.32	36.45	11.97	12.89	**24.29**	89.60	62.80	62.40	60.60
	SpecFormer (Ours)	**72.73**	31.84	31.90	**47.07**	**41.46**	**12.80**	**13.54**	23.25	**91.60**	**67.00**	**67.00**	**64.60**
DeiT-Ti	LipsFormer [41]	39.72	24.05	26.77	23.07	23.09	9.82	12.03	9.54	39.00	27.60	28.40	23.80
	L2Former [26]	60.85	34.29	36.63	34.03	36.67	14.23	16.48	17.24	77.40	41.40	43.20	48.20
	LNFormer [13]	54.32	29.85	32.96	29.44	28.65	11.36	13.94	13.69	72.20	39.00	42.20	38.00
	TransFormer [17]	71.71	37.15	38.74	43.60	40.89	15.25	17.40	20.89	79.00	40.60	41.40	56.40
	SpecFormer (Ours)	**80.03**	**48.52**	**51.10**	**45.36**	**44.48**	**16.36**	**18.21**	**23.49**	**82.20**	**46.20**	**46.00**	**58.20**
ConViT-Ti	LipsFormer [41]	56.83	32.27	35.04	30.75	31.50	14.56	17.17	14.15	60.80	34.00	38.20	33.40
	L2Former [26]	39.36	23.84	26.42	22.21	16.53	8.06	9.65	9.03	10.00	10.00	10.00	10.00
	LNFormer [13]	49.03	28.68	31.68	27.20	29.12	13.00	15.65	12.46	72.20	39.00	42.20	22.40
	TransFormer [17]	53.09	30.87	33.63	**37.78**	31.54	**14.65**	**17.24**	14.47	68.00	41.40	43.60	**39.20**
	SpecFormer (Ours)	**67.05**	**38.72**	**41.58**	35.30	**37.92**	14.50	16.78	**18.23**	**86.60**	**47.20**	**46.20**	36.20
Swin-Ti	LipsFormer [41]	54.52	32.09	35.07	18.32	32.96	15.79	18.77	14.78	55.80	34.20	38.00	33.60
	L2Former [26]	79.20	44.72	46.20	43.00	53.63	22.47	24.39	20.89	94.00	69.80	69.40	68.60
	LNFormer [13]	77.39	43.33	45.13	41.68	52.90	22.48	24.70	20.76	93.00	67.60	68.00	66.80
	TransFormer [17]	**81.19**	**45.44**	**46.72**	**43.81**	**56.51**	23.25	24.56	**21.42**	95.20	70.80	71.00	69.40
	SpecFormer(Ours)	79.48	43.98	45.78	42.20	55.53	**23.32**	**25.10**	21.17	**95.40**	**73.20**	**72.80**	**71.80**

vanilla ViTs. This is further confirmed by the $\mathcal{A}$-distance [5], shown at the top of the images. A smaller distance signifies greater similarity between clean and attacked features, demonstrating enhanced robustness under perturbations. This contrast highlights the significant improvement in ViTs' adversarial robustness achieved by our method.

Computation cost. Finally, we analyze the computation cost of our proposed method. Table 7 displays the relative running time difference for each model to complete 10 training steps under both standard and adversarial conditions (with vanilla ViT set to 1). The table reveals that our proposed method exhibits the lowest additional computational costs compared to other baseline models, underscoring the computational efficiency of our approach.

Table 5: Results under **adversarial training** for larger models.

Model	Method	CIFAR10				CIFAR100				Imagenette			
		Standard	CW-20	PGD-20	AutoAttack	Standard	CW-20	PGD-20	AutoAttack	Standard	CW-20	PGD-20	AutoAttack
ViT-B	LipsFormer [41]	38.54	22.47	25.29	21.56	21.68	9.39	11.42	8.72	57.00	25.00	31.00	23.80
	L2Former [26]	68.91	41.37	43.89	39.78	46.84	21.52	24.58	19.94	70.20	34.60	39.20	33.80
	LNFormer [13]	52.99	30.19	33.36	29.13	34.75	16.03	18.86	14.67	64.00	31.20	35.60	29.20
	TransFormer [17]	82.70	50.17	**52.65**	48.27	**61.18**	**28.33**	**30.16**	**26.14**	**91.40**	**65.80**	**67.40**	**64.40**
	SpecFormer (Ours)	**87.22**	**51.32**	52.55	**48.80**	54.91	24.66	27.73	22.80	89.80	63.20	64.80	62.00
DeiT-S	LipsFormer [41]	41.90	25.00	27.79	24.03	25.17	11.30	13.54	10.40	47.60	31.00	34.40	30.80
	L2Former [26]	67.86	39.80	42.73	38.13	45.38	21.12	23.87	19.66	84.60	55.00	55.60	53.60
	LNFormer [13]	54.89	31.44	35.01	30.26	34.54	15.74	18.53	14.40	79.00	43.80	46.20	42.20
	TransFormer [17]	81.37	49.22	51.83	47.34	58.73	27.15	29.47	25.13	91.40	66.60	66.20	65.60
	SpecFormer (Ours)	**83.61**	**50.03**	**52.01**	**47.97**	**61.76**	**27.84**	**29.70**	**25.48**	**93.40**	**67.40**	**67.00**	**65.80**
ConViT-S	LipsFormer [41]	38.99	23.59	26.40	22.72	15.92	7.81	9.54	7.39	59.40	33.80	38.40	32.60
	L2Former [26]	32.50	23.15	24.42	22.58	23.61	10.82	13.29	10.22	10.00	10.00	10.00	10.00
	LNFormer [13]	51.03	30.18	33.54	29.27	30.60	14.52	17.31	13.45	49.60	25.20	28.00	23.80
	TransFormer [17]	63.84	36.63	39.75	35.21	29.02	13.21	15.73	12.34	89.20	62.80	63.40	60.20
	SpecFormer (Ours)	**67.71**	**38.50**	**41.49**	**36.85**	**32.95**	**15.26**	**17.94**	**14.16**	**92.40**	**67.20**	**66.80**	**65.40**

Table 6: Performance (%) of SpecFormer with different ViT variants on ImageNet datasets under standard training (using ImageNet-22k pre-trained weights). The best results are in **bold**.

Method	Standard Training			Adversarial Training			
	Standard	FGSM	PGD-2	Standard	CW-20	PGD-20	PGD-100
LipsFormer [41]	65.76	20.87	3.77	45.04	18.91	21.13	20.83
L2Former [26]	77.40	37.12	6.89	51.24	24.85	27.04	26.95
LNFormer [13]	50.84	25.78	0.54	30.93	11.53	14.08	14.04
TransFormer [17]	79.11	41.45	10.79	60.81	30.92	32.58	32.35
SpecFormer (Ours)	**80.04**	**43.51**	**11.59**	**62.30**	**31.87**	**32.82**	**32.56**

Table 7: Relative running time for 10-step training (vanilla ViT=1). The lowest additional computation costs are in **bold**.

Relative	TransFormer	LipsFormer	L2Former	LNFormer	SpecFormer
Std. training	1	5.5	2.5	4.5	**1.5**
Adv. training	1	12.3	4.3	11.7	**2.5**

7 Conclusion

This paper presents a theoretical analysis of the self-attention mechanism in ViTs through the lens of adversarial robustness and Lipschitz continuity theory. We develop the local Lipschitz constant bound based on the investigation and further introduce the Maximum Singular Value Penalization (MSVP) to enhance the robustness of ViTs. Our theoretical bounds demonstrate that the Lipschitz continuity of ViTs in the vicinity of the input can be bounded by the maximum singular values of the attention weight matrices. Empirical validation across four ViT variants on four datasets, under both standard and adversarial training and various attacks (FGSM, CW, PGD, AutoAttack), demonstrates the superiority of our proposed model. The comprehensive evaluation underscores the practical applicability and robustness of our approach, making it a valuable contribution to the fields of adversarial robustness and Vision Transformer safety.

Acknowledgements

Qi WU acknowledges the support from The CityU-JD Digits Joint Laboratory in Financial Technology and Engineering and The Hong Kong Research Grants Council [General Research Fund 11219420/9043008]. The work described in this paper was partially supported by the InnoHK initiative, the Government of the HKSAR, and the Laboratory for AI-Powered Financial Technologies.

References

1. Ali, A., Touvron, H., Caron, M., Bojanowski, P., Douze, M., Joulin, A., Laptev, I., Neverova, N., Synnaeve, G., Verbeek, J., et al.: Xcit: Cross-covariance image transformers. Advances in neural information processing systems **34**, 20014–20027 (2021)
2. Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., Schmid, C.: Vivit: A video vision transformer. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 6836–6846 (2021)
3. Bai, T., Luo, J., Zhao, J., Wen, B., Wang, Q.: Recent advances in adversarial training for adversarial robustness. In: IJCAI survey track (2021)
4. Bai, Y., Mei, J., Yuille, A.L., Xie, C.: Are transformers more robust than cnns? Advances in Neural Information Processing Systems **34**, 26831–26843 (2021)
5. Ben-David, S., Blitzer, J., Crammer, K., Pereira, F.: Analysis of representations for domain adaptation. Advances in neural information processing systems **19** (2006)
6. Bhojanapalli, S., Chakrabarti, A., Glasner, D., Li, D., Unterthiner, T., Veit, A.: Understanding robustness of transformers for image classification. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10231–10241 (2021)
7. Burden, R.L., Faires, J.D., Burden, A.M.: Numerical analysis. Cengage learning (2015)
8. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16. pp. 213–229. Springer (2020)
9. Carlini, N., Wagner, D.: Towards evaluating the robustness of neural networks. In: 2017 ieee symposium on security and privacy (sp). pp. 39–57. Ieee (2017)
10. Chen, P.Y., Sharma, Y., Zhang, H., Yi, J., Hsieh, C.J.: Ead: elastic-net attacks to deep neural networks via adversarial examples. In: Proceedings of the AAAI conference on artificial intelligence. vol. 32 (2018)
11. Cisse, M., Bojanowski, P., Grave, E., Dauphin, Y., Usunier, N.: Parseval networks: Improving robustness to adversarial examples. In: International conference on machine learning. pp. 854–863. PMLR (2017)
12. Croce, F., Hein, M.: Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In: International conference on machine learning. pp. 2206–2216. PMLR (2020)
13. Dasoulas, G., Scaman, K., Virmaux, A.: Lipschitz normalization for self-attention layers with application to graph neural networks. In: International Conference on Machine Learning. pp. 2456–2466. PMLR (2021)
14. Debenedetti, E., Sehwag, V., Mittal, P.: A light recipe to train robust vision transformers. arXiv preprint arXiv:2209.07399 (2022)

15. Dehghani, M., Djolonga, J., Mustafa, B., Padlewski, P., Heek, J., Gilmer, J., Steiner, A., Caron, M., Geirhos, R., Alabdulmohsin, I., et al.: Scaling vision transformers to 22 billion parameters. arXiv preprint arXiv:2302.05442 (2023)
16. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
17. 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 (ICLR) (2020)
18. d'Ascoli, S., Touvron, H., Leavitt, M.L., Morcos, A.S., Biroli, G., Sagun, L.: Convit: Improving vision transformers with soft convolutional inductive biases. In: International Conference on Machine Learning. pp. 2286–2296. PMLR (2021)
19. Federer, H.: Geometric Measure Theory. Classics in Mathematics, Springer Berlin Heidelberg (1969)
20. Fu, Y., Zhang, S., Wu, S., Wan, C., Lin, Y.: Patch-fool: Are vision transformers always robust against adversarial perturbations? In: International conference on learning representations (ICLR) (2022)
21. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. In: International conference on learning representations (ICLR) (2015)
22. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
23. Hein, M., Andriushchenko, M.: Formal guarantees on the robustness of a classifier against adversarial manipulation. Advances in neural information processing systems **30** (2017)
24. Howard, J.: Imagenette. `https://github.com/fastai/imagenette` (2019)
25. Huster, T., Chiang, C.Y.J., Chadha, R.: Limitations of the lipschitz constant as a defense against adversarial examples. In: ECML PKDD 2018 Workshops: Nemesis 2018, UrbReas 2018, SoGood 2018, IWAISe 2018, and Green Data Mining 2018, Dublin, Ireland, September 10-14, 2018, Proceedings 18. pp. 16–29. Springer (2019)
26. Kim, H., Papamakarios, G., Mnih, A.: The lipschitz constant of self-attention. In: International Conference on Machine Learning. pp. 5562–5571. PMLR (2021)
27. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images (2009)
28. Leino, K., Wang, Z., Fredrikson, M.: Globally-robust neural networks. In: International Conference on Machine Learning. pp. 6212–6222. PMLR (2021)
29. 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)
30. Lovisotto, G., Finnie, N., Munoz, M., Mummadi, C.K., Metzen, J.H.: Give me your attention: Dot-product attention considered harmful for adversarial patch robustness. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15234–15243 (2022)
31. Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research **9**(11) (2008)
32. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. In: International conference on learning representations (ICLR) (2018)

33. Mo, Y., Wu, D., Wang, Y., Guo, Y., Wang, Y.: When adversarial training meets vision transformers: Recipes from training to architecture. In: Advances in neural information processing systems (NeurIPS) (2022)
34. Murdock, J.A.: Perturbations: Theory and Methods. Society for Industrial and Applied Mathematics (1999). https://doi.org/10.1137/1.9781611971095, https://epubs.siam.org/doi/abs/10.1137/1.9781611971095
35. 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)
36. Nguyen, A., Yosinski, J., Clune, J.: Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 427–436 (2015)
37. Pang, T., Lin, M., Yang, X., Zhu, J., Yan, S.: Robustness and accuracy could be reconcilable by (proper) definition. In: International Conference on Machine Learning. pp. 17258–17277. PMLR (2022)
38. Papernot, N., McDaniel, P., Sinha, A., Wellman, M.: Towards the science of security and privacy in machine learning. arXiv preprint arXiv:1611.03814 (2016)
39. Papernot, N., McDaniel, P., Wu, X., Jha, S., Swami, A.: Distillation as a defense to adversarial perturbations against deep neural networks. In: 2016 IEEE symposium on security and privacy (SP). pp. 582–597. IEEE (2016)
40. Paul, S., Chen, P.Y.: Vision transformers are robust learners. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 2071–2081. No. 2 (2022)
41. Qi, X., Wang, J., Chen, Y., Shi, Y., Zhang, L.: Lipsformer: Introducing lipschitz continuity to vision transformers. In: International conference on Learning Representations (ICLR) (2023)
42. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
43. Shao, R., Shi, Z., Yi, J., Chen, P.Y., Hsieh, C.J.: On the adversarial robustness of vision transformers. arXiv preprint arXiv:2103.15670 (2021)
44. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
45. Strudel, R., Garcia, R., Laptev, I., Schmid, C.: Segmenter: Transformer for semantic segmentation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 7262–7272 (2021)
46. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1–9 (2015)
47. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R.: Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199 (2013)
48. Takase, S., Kiyono, S., Kobayashi, S., Suzuki, J.: On layer normalizations and residual connections in transformers. arXiv preprint arXiv:2206.00330 (2022)
49. 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)
50. Touvron, H., Cord, M., Sablayrolles, A., Synnaeve, G., Jégou, H.: Going deeper with image transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 32–42 (2021)

51. Tramèr, F., Kurakin, A., Papernot, N., Goodfellow, I., Boneh, D., McDaniel, P.: Ensemble adversarial training: Attacks and defenses. In: Ensemble adversarial training: Attacks and defenses (2018)
52. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
53. Wang, H., Deng, Y., Yoo, S., Lin, Y.: Exploring robust features for improving adversarial robustness. arXiv preprint arXiv:2309.04650 (2023)
54. Wang, H., Ma, S., Dong, L., Huang, S., Zhang, D., Wei, F.: Deepnet: Scaling transformers to 1,000 layers. arXiv preprint arXiv:2203.00555 (2022)
55. Wang, Z., Bai, Y., Zhou, Y., Xie, C.: Can cnns be more robust than transformers? arXiv preprint arXiv:2206.03452 (2022)
56. Xie, C., Wang, J., Zhang, Z., Ren, Z., Yuille, A.: Mitigating adversarial effects through randomization. In: International conference on learning representations (ICLR) (2018)
57. Xu, T., Chen, W., Wang, P., Wang, F., Li, H., Jin, R.: Cdtrans: Cross-domain transformer for unsupervised domain adaptation. In: International conference on learning representations (ICLR) (2022)
58. Yoshida, Y., Miyato, T.: Spectral norm regularization for improving the generalizability of deep learning. arXiv preprint arXiv:1705.10941 (2017)
59. Zhang, H., Yu, Y., Jiao, J., Xing, E., El Ghaoui, L., Jordan, M.: Theoretically principled trade-off between robustness and accuracy. In: International conference on machine learning. pp. 7472–7482. PMLR (2019)
60. Zhang, H., Zhang, P., Hsieh, C.J.: Recurjac: An efficient recursive algorithm for bounding jacobian matrix of neural networks and its applications. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 5757–5764 (2019)
61. Zheng, S., Song, Y., Leung, T., Goodfellow, I.: Improving the robustness of deep neural networks via stability training. In: Proceedings of the ieee conference on computer vision and pattern recognition. pp. 4480–4488 (2016)
62. Zhou, D., Yu, Z., Xie, E., Xiao, C., Anandkumar, A., Feng, J., Alvarez, J.M.: Understanding the robustness in vision transformers. In: International Conference on Machine Learning. pp. 27378–27394. PMLR (2022)