Fine-grained Local Sensitivity Analysis of Standard Dot-Product Self-Attention

Aaron Havens [1] **Alexandre Araujo** [2] **Huan Zhang** [1] **Bin Hu** [1]

Abstract

Self-attention has been widely used in various machine learning models, such as vision transformers. The standard dot-product self-attention is arguably the most popular structure, and there is a growing interest in understanding the mathematical properties of such attention mechanisms. This paper presents a fine-grained local sensitivity analysis of the standard dot-product self-attention, leading to new non-vacuous certified robustness results for vision transformers. Despite the well-known fact that dot-product self-attention is not (globally) Lipschitz, we develop new theoretical analysis of Local Fine-grained Attention Sensitivity (LoFAST) quantifying the effect of input feature perturbations on the attention output. Our analysis reveals that the local sensitivity of dot-product self-attention to ℓ_2 perturbations can actually be controlled by several key quantities associated with the attention weight matrices and the unperturbed input. We empirically validate our theoretical findings by computing non-vacuous certified ℓ_2-robustness for vision transformers on CIFAR-10 and SVHN datasets. The code for LoFAST is available at `https://github.com/AaronHavens/LoFAST`.

1. Introduction

The self-attention mechanism (Bahdanau et al., 2014; Vaswani et al., 2017) has become a major building block in many modern deep learning-based systems, achieving state-of-the-art performance in various applications such as vision and natural language processing. In particular, dot-product self-attention (Vaswani et al., 2017) is one of the most popular architectures used by many best-performing networks such as the well-known Transformer architecture and its variants, and has enabled successful applications such as large language models (LLM) (Brown et al., 2020; Bubeck et al., 2023) and vision transformers (ViT) (Dosovitskiy et al., 2021; Radford et al., 2021). However, unlike traditional neural network building blocks such as convolutional layers, whose structures and behaviors are well understood, the self-attention mechanism has more involved mathematical properties. For example, for a simple convolutional layer, it is well known that its operator norm is bounded (Sedghi et al., 2019), and convolution is a Lipschitz operation that always produces bounded outputs given bounded inputs (Delattre et al., 2023). However, for the popular dot-product self-attention mechanism, existing work has shown that they are surprisingly, not (globally) Lipschitz (Kim et al., 2021). The lack of Lipschitzness indicates that dot-product self-attention can theoretically be very sensitive to its input, which can impede stable learning (Qi et al., 2023) and lead to poor robustness (Zhou et al., 2022; Cisse et al., 2017). Although several architectures have been proposed to amend the popular dot-product attention mechanism to achieve Lipschitzness and bounded sensitivity (Kim et al., 2021; Dasoulas et al., 2021; Qi et al., 2023), none of them are popular in large-scale networks deployed in production, and it is still an open challenge to understand why the non-Lipschitz dot-product attention mechanism can work well in practice.

In this work, instead of amending the network structure to achieve bounded sensitivity, we aim to analyze the *local* sensitivity of the *unmodified* dot-product attention mechanism directly. Despite being non-Lipschitz, local sensitivity of the unmodified self-attention mechanism is actually sufficient for inducing certified robustness (Proposition 1). Built upon this observation, we derived novel analytical bounds for the local sensitivity of dot-product self-attention using tools from optimization and matrix theory. Our key result (Theorem 1) deciphers a few key quantities associated with the sensitivity of the dot-product self-attention operation, related to the attention weight matrix and their inputs. Our theorem can be easily interpreted, and gives us insights on how to control the local sensitivity of a Transformer. In particular, we found that the local sensitivity of the self-attention layer is directly related to the norm of its input, thus theoretically explaining the necessity of using layer normalization (Ba et al., 2016) in the popular Transformer architecture (Xiong et al., 2020). In addition,

[1]ECE & CSL, University of Illinois Urbana-Champaign [2]ECE, New York University. Correspondence to: Aaron Havens <ahavens2@illinois.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

it allows us to utilize the recent progress of 1-Lipschitz feedforward neural network layers, such as orthogonal layers (Trockman & Kolter, 2021; Prach & Lampert, 2022) and the SDP-based Lipschitz Layers (Araujo et al., 2023), to control the local sensitivity of Transformers. Note that since the self-attention layer is non-Lipschitz, naively applying 1-Lipschitz layers could not provide any guarantees without our new local results.

We confirm our theoretical findings on a few practical vision transformers by quantifying their local sensitivity and certified ℓ_2-robustness. Our experiments show that our derived local sensitivity bounds are practical for vision transformers and significantly improve against a naive approach for sensitivity analysis. In addition, we also use gradient ascent to find the maximum sensitivity empirically, and demonstrate that our theoretical bounds and empirical measurements are well-aligned. By varying the design parameters of the vision transformers (*e.g.*, number of attention heads and number of tokens), our theory predicts the observed changes in local sensitivity. As a direct application of our bounds, we also give non-vacuous (certified) adversarial robustness guarantees for vision transformers with standard dot-product self-attention mechanisms on CIFAR and SVHN datasets. Our main contributions are summarized as follows.

- We are the first to consider a fine-grained theoretical analysis of *local sensitivity* bounds of *unmodified* dot-product self-attention mechanism, contributing to the mathematical understanding of this popular network structure. Despite the non-Lipschitzness of dot-product self-attention, our local bounds are non-trivial and can lead to *non-vacuous* certified robustness for practical transformers.
- Our results give interpretable bounds that offer practical design insights into achieving low sensitivity on dot-product self-attention-based transformers. It enables us to borrow the recently developed algebraic tricks for training globally 1-Lipschitz feedforward networks to provably improve the *local* sensitivity of Transformers.
- Our theoretical results are validated through the empirical evaluation of a large range of Transformers trained with different design parameters. In addition, our tight analytical bounds allow us to achieve fast scalable computation of non-trivial deterministic certified ℓ_2-robustness guarantees for vision transformers without modifying the dot-product self-attention mechanism.

2. Related Work

Lipschitz Aspects and Regularity of Self-Attention. Since the first Lispchitz analysis of dot-product self-attention by (Kim et al., 2021), which showed that the standard dot-product self-attention is not (globally) Lipschitz, a large number of works have tried to propose variants of the original dot-product self-attention to enforce this property (Kim et al., 2021; Qi et al., 2023; Fei et al., 2022; Dasoulas et al., 2021; Ye et al., 2023). For example, (Qi et al., 2023) proposed scaled cosine similarity attention instead of dot product attention and demonstrated the Lispchitz properties of this new layer. Other works (Vuckovic et al., 2021; Castin et al., 2023) have studied the regularity of attention under a mathematical framework that uses measure theory and integral operators to model attention. Under this new framework, they show that the attention mechanism is regular (under some specific assumptions) with respect to the 1-Wasserstein distance. While this work generalizes the work of (Kim et al., 2021), the regularity over the 1-Wasserstein distance is not commonly used in practice.

Neural Networks with Prescribed Lipschitz Constant. Recently, researchers have designed neural networks with prescribed Lipschitz constant in order to better control the stability (Miyato et al., 2018), robustness (Zhang et al., 2021; Prach & Lampert, 2022; Meunier et al., 2022; Zhang et al., 2022; Araujo et al., 2023; Wang & Manchester, 2023; Li et al., 2019; Trockman & Kolter, 2021; Singla & Feizi, 2021; Yu et al., 2022; Xu et al., 2022; Havens et al., 2023; Fazlyab et al., 2023; Barbara et al., 2024), and generalization (Bartlett et al., 2017) of the network. However, most of these techniques come with important design choices with respect to the architecture that are not common in networks with state-of-the-art performance.

Robustness of Transformer Networks. Randomized smoothing (Cohen et al., 2019) has been used to obtain probabilistic certified robustness of dot-product attention (Carlini et al., 2023; Wu et al., 2023). However, randomized smoothing suffers from high computational cost. General-purpose certification tools such as CROWN (Zhang et al., 2018; Wang et al., 2021) and zonotope abstractions have also been tailored for robustness certification of dot-product attention (Shi et al., 2020; Bonaert et al., 2021). However, these prior approaches face severe scalability issues when applied to large transformers on practical datasets such as CIFAR. In this work, our analytical local sensitivity bounds can be used to provide fast scalable computation of non-trivial (deterministic) ℓ_2 certified robust accuracy for dot-product self-attention in ViT for image classification tasks such as CIFAR-10 and SVHN. In our experiments, we provide a comparison study looking at the trade-offs in terms of tightness and scalability of CROWN, and show that our approach LoFAST can complement existing deterministic verifiers via providing enhanced scalability.

3. Preliminaries and Problem Formulation

Notation We denote the spectral norm and the Frobenius norm as $\|\cdot\|$ and $\|\cdot\|_F$, respectively. Two useful facts are $\|AB\|_F \leq \|A\|\|B\|_F$, and $\|A\| = \|A^\mathsf{T}\|$. Given two matrices A and B, their Kronecker product is denoted as

$A \otimes B$. We denote the vectorization operation as vec. Let e_i denote an n-dimensional vector whose i-th entry is 1 and all other entries are 0. The $n \times n$ identity matrix is denoted by I_n. The softmax mapping on matrices with the temperature being 1 is denoted as softmax. We know that softmax is 1-Lipschitz (Gao & Pavel, 2017), i.e. $\|\text{softmax}(A) - \text{softmax}(B)\|_F \leq \|A - B\|_F$ for any two matrices A and B that have the same dimension.

Dot-Product Self-Attention. Let $x_1, x_2, \ldots, x_n$ be a sequence of n vectors, where $x_i \in \mathbb{R}^d$. For vision tasks, each x_i is a patch. This sequence is represented as a matrix X. The dot-product multi-head self-attention maps $\mathbb{R}^{n\times d}$ to $\mathbb{R}^{n\times d}$. With h heads, the l-th head maps $\mathbb{R}^{n\times d}$ to $\mathbb{R}^{n\times d/h}$ as:

$$X = \begin{bmatrix} - & x_1^\mathsf{T} & - \\ & \vdots & \\ - & x_n^\mathsf{T} & - \end{bmatrix} \in \mathbb{R}^{n\times d} \tag{1}$$

and

$$Y_l = \text{softmax}\left(\frac{XW_l^Q(XW_l^K)^\mathsf{T}}{\sqrt{d/h}}\right) XW_l^V \tag{2}$$

where $W_l^Q, W_l^K, W_l^V \in \mathbb{R}^{d\times d/h}$ denote the weight matrices for the l-th head, and the softmax operation is applied in a row-wise manner. Finally, the outputs of all heads are concatenated as

$$f(X) = [Y_1, \ldots, Y_h]\, W^O = \sum_{l=1}^{h} Y_l W_l^O, \tag{3}$$

where $W^O = [(W_1^O)^\mathsf{T}, \ldots, (W_h^O)^\mathsf{T}]^\mathsf{T} \in \mathbb{R}^{d/h\times d}$ gives the weight for the linear combination of the outputs from all the heads. For simplicity, we introduce the notation $P_l(X)$ as

$$P_l(X) = \text{softmax}\left(\frac{XW_l^Q(XW_l^K)^\mathsf{T}}{\sqrt{d/h}}\right). \tag{4}$$

Hence the dot-product self-attention can be rewritten as:

$$f(X) = \sum_{l=1}^{h} P_l(X) X W_l^V W_l^O \tag{5}$$

Residual Structure. Dot-product self-attention is typically used in a residual form. In this case, the output is defined as $f(X) = X + \sum_{l=1}^{h} P_l(X) X W_l^V W_l^O$.

Problem Statement. It is well-known that (5) is not globally Lipschitz (Kim et al., 2021). We are interested in analyzing the local sensitivity of dot-product self-attention. We consider the following model which unifies (5) and its residual variant with $H \in \mathbb{R}^{n\times n}$:

$$F(X) = HX + \sum_{l=1}^{h} P_l(X) X W_l^V W_l^O. \tag{6}$$

If $H = 0$, then (6) recovers the standard dot-product self-attention (5). If $H = I$, then (6) reduces to the residual setting. Given a local input point X and some small positive scalar ϵ, we want to prove a bound in the following form:

$$\|F(X') - F(X)\|_F \leq \delta(X, \epsilon) \tag{7}$$

for X' satisfying $\|X' - X\|_F \leq \epsilon$ where the mapping $F(\cdot)$ is defined by (6). We denote this set of ϵ-bounded perturbations centered at X as $\Omega(X, \epsilon) := \{X' : \|X' - X\|_F \leq \epsilon\}$. In principle, the tightest choice of $\delta(X, \epsilon)$ is given by the solution to the following constrained optimization problem

$$\max_{X'\in\Omega(X,\epsilon)} \|F(X') - F(X)\|_F. \tag{8}$$

One can use the projected gradient ascent method to search solutions for (8). However, there are no polynomial-time guarantees in solving the above problem globally. In addition, the bound (8) does not bring any insights for how to control the local sensitivity via network structure design. The goal of this paper is to develop a spectrum of choices for $\delta(X, \epsilon)$ that can capture the trade-off between tightness, tractability, and interpretability.

Once we figure out an efficient way to compute $\delta(X, \epsilon)$ for the above problem, we can immediately apply the analysis in a recursive manner to solve the local sensitivity analysis of multi-layer networks consisting of various dot-product self-attention layers. Specifically, consider a N-layer network:

$$F(X) = f^N \circ f^{N-1} \circ \cdots \circ f^0(X) \tag{9}$$

where f^k is either a dot-product self-attention layer (6) or a globally 1-Lipschitz operation. Applying the local sensitivity analysis in a recursive manner, we will be able to compute $\delta(X, \epsilon)$ for bounding the end-to-end local sensitivity of (9) as described by (7). Such a bound can be used to prove the certified robustness of F on the data point X subject to adversarially chosen ℓ_2 perturbations. Specifically, the following result connects the local sensitivity bound $\delta(X, \epsilon)$ to certified ℓ_2-robustness in a rigorous manner.

Proposition 1. *Suppose F is a classifier that maps any input X to the output as defined by (9). The j-th entry of $F(X)$ is denoted as $[F(X)]_j$, which gives the logits value for the j-th label class. The predicted label for X is given by $\arg\max_j [F(X)]_j$. Given an input X with the true label y satisfying $y = \arg\max_j [F(x)]_j$, if we have*

$$\mathcal{M}_\mathbf{f}(X) := [F(X)]_y - \max_{j\neq y} [F(X)]_j > \sqrt{2}\delta(X, \epsilon),$$

then for every τ satisfying $\|\tau\|_F \leq \epsilon$, we must have $\arg\max_j [F(X + \tau)]_j = y$.

The proof for the above result is almost identical to Tsuzuku et al. (2018, Proposition 1), and hence deferred to the appendix. The above proposition provides a way to compute the certified robust accuracy of dot-product self-attention using our local sensitivity analysis.

Distinction from Local-Lipschitz Bounds. We emphasize that the local bound $\delta(X,\epsilon)$ is not the same as a local-Lipschitz bound. As a matter of fact, the local Lipschitz approach can be unnecessarily conservative. Specifically, the local Lipschitz bound applies for any two arbitrary points in the ϵ-neighborhood of the original input X. In contrast, our local sensitivity analysis is weaker in the sense that the bound can only tell us the deviation of $F(X')$ from a fixed $F(X)$. However, that is still sufficient for computing certified robustness as in Proposition 1. In Appendix A, we will show explicitly how the existing local Lipschitz analysis (Xixu, 2023) can only give vacuous certified robustness results on CIFAR-10.

4. Fine-Grained Local Sensitivity Analysis

In this section, we perform the local sensitivity analysis for the dot-product self-attention where F is defined by (6). We have $F(X) = HX + \sum_{l=1}^h P_l(X)XW_l^VW_l^O$ for either $H=0$ or $H=I$. First, the following bound based on the splitting trick is standard:

$$\begin{aligned}&\|F(X')-F(X)\|_F\\ \le&\big\|H(X'-X)+\sum_{l=1}^h P_l(X)(X'-X)W_l^VW_l^O\big\|_F\\ &+\big\|\sum_{l=1}^h(P_l(X')-P_l(X))X'W_l^VW_l^O\big\|_F\end{aligned}$$

Next, we will bound the two terms on the right side. We use the following notation

$$\Delta_1(X,X')=\big\|H(X'-X)+\sum_{l=1}^h P_l(X)(X'-X)W_l^VW_l^O\big\|_F \tag{10}$$

$$\Delta_2(X,X')=\big\|\sum_{l=1}^h(P_l(X')-P_l(X))X'W_l^VW_l^O\big\|_F \tag{11}$$

If we can derive bounds in the form of:

$$\Delta_1(X,X')\le\delta_1(X,\epsilon),\quad \Delta_2(X,X')\le\delta_2(X,\epsilon)$$

which hold for all X' satisfying $\|X-X'\|_F\le\epsilon$, then we can immediately set $\delta(X,\epsilon):=\delta_1(X,\epsilon)+\delta_2(X,\epsilon)$, and obtain the following bound for the self-attention map which can be computed given a point X and perturbation radius ϵ.

$$\max_{X'\in\Omega(X,\epsilon)}\|F(X')-F(X)\|_F\le\delta_1(X,\epsilon)+\delta_2(X,\epsilon)$$

Our fine-grained analysis addresses how to reduce the conservatism in deriving $\delta_1(X,\epsilon)$ and $\delta_2(X,\epsilon)$.

Reducing Conservatism in Deriving $\delta_1(X,\epsilon)$: Based on the property of matrix norms, one can obtain the following naive upper bound for $\max_{X'\in\Omega(X,\epsilon)}\Delta_1(X,X')$ (see the appendix for a detailed derivation):

$$\delta_1^{(\text{naive})}(X,\epsilon)=\left(\|H\|+\sum_{l=1}^h\|W_l^VW_l^O\|\|P_l(X)\|\right)\epsilon. \tag{12}$$

The above bound is informative in showing that one can potentially control $\Delta_1(X,X')$ by constraining the spectral norm of $\{W_l^V,W_l^O\}_{l=1}^h$. However, the above bound can be loose quantitatively. In contrast, the best possible bound for $\delta_1(X,\epsilon)$ can be obtained via solving the following problem

$$\underset{X'\in\Omega(X,\epsilon)}{\text{maximize}}\ \Delta_1(X,X') \tag{13}$$

It turns out that this problem actually has an analytical solution. This leads to our first result stated as follows.

Lemma 1 (Key Sensitivity Metric)**.** *The exact solution to the optimization problem* (13) *is given by*

$$\delta_1(X,\epsilon):=\max_{X'\in\Omega(X,\epsilon)}\Delta_1(X,X')=\zeta(X)\epsilon \tag{14}$$

where $\zeta(X)$ is defined as

$$\zeta(X)=\big\|H\otimes I_n+\sum_{l=1}^h(P_l(X)\otimes(W_l^VW_l^O)^\mathsf{T})\big\|. \tag{15}$$

Consequently, we have $\Delta_1(X,X')\le\delta_1(X,\epsilon)=\zeta(X)\epsilon$, for all X' satisfying $\|X'-X\|_F\le\epsilon$.

A detailed proof for Lemma 1 is presented in the appendix. The main proof idea is based on the following key identity:

$$\begin{aligned}&\text{vec}\left(\sum_{l=1}^h(P_l(X)(X'-X)W_l^VW_l^O)^\mathsf{T}\right)\\ =&\left(\sum_{l=1}^h(P_l(X)\otimes(W_l^VW_l^O)^\mathsf{T})\right)\text{vec}((X'-X)^\mathsf{T})\end{aligned}$$

which enables us to solve (13) exactly via viewing it as a largest singular value problem. The quantity $\zeta(X)$ is termed as the key sensitivity metric which quantifies the local sensitivity of the self-attention around the data point X due to the error Δ_1. The computation of this metric is reasonably scalable so that one can efficiently compute this metric for ViT image classifiers for datasets like CIFAR-10 and SVHN. Later, we will show that this term is the dominating term in the bound $\Delta_1+\Delta_2$, and hence one should calculate this term exactly when fine-grained sensitivity analysis is needed.

Reducing Conservatism in Deriving $\delta_2(X,\epsilon)$ The best possible bound for $\Delta_2(X,X')$ is the solution to the following constrained maximization problem:

$$\underset{X'\in\Omega(X,\epsilon)}{\text{maximize}}\ \Delta_2(X,X') \tag{16}$$

One can apply the projected gradient ascent method to the above problem. However, there are no guarantees that the resultant solution is global due to the form of the cost function. The solution from the gradient ascent method only provides lower bound for (16). To obtain a more tractable upper bound, it is straightforward to apply the triangle inequality to show that (16) can be bounded by the following term:

$$\sum_{l=1}^{h} \Big(\max_{X' \in \Omega(X,\epsilon)} \big\| (P_l(X') - P_l(X)) \big\|_F \cdot \max_{X' \in \Omega(X,\epsilon)} \|X' W_l^V W_l^O\| \Big), \tag{17}$$

which involve two maximization problems. Now we discuss these two problems separately.

To address the term $\max_{X' \in \Omega(X,\epsilon)} \|X' W_l^V W_l^O\|$, we can apply the triangle inequality and obtain the following tractable upper bound:

$$\max_{X' \in \Omega(X,\epsilon)} \|X' W_l^V W_l^O\| \leq \|X W_l^V W_l^O\| + \|W_l^V W_l^O\| \epsilon \tag{18}$$

The above upper bound can be efficiently calculated via power iteration, and is less conservative than the naive bound $\|W_l^V W_l^O\|(\|X\| + \epsilon)$. Later, we will show that the above upper bound is reasonable for the purpose of upper bounding $\Delta_1 + \Delta_2$, since replacing it with the lower bounds obtained by the projected gradient ascent method does not affect the final overall bound value significantly.

Next, we discuss how to address

$$\max_{X' \in \Omega(X,\epsilon)} \big\| P_l(X') - P_l(X) \big\|_F \tag{19}$$

Again, one can apply the projected gradient ascent method to search for lower bounds for the above quantity. We are more interested in obtaining less conservative upper bounds that are computationally tractable. Since $\mathrm{softmax}$ is 1-Lipschitz, we can show the following holds for any X:

$$\begin{aligned} &\|P_l(X') - P_l(X)\|_F \\ &\leq \frac{1}{\sqrt{d/h}} \|X' W_l^Q (W_l^K)^\mathsf{T} (X')^\mathsf{T} - X W_l^Q (W_l^K)^\mathsf{T} X^\mathsf{T}\|_F \end{aligned} \tag{20}$$

Denoting $\Gamma = X' - X$. If $\|X' - X\|_F \leq \epsilon$, then we have $\|\Gamma\|_F \leq \epsilon$. We immediately have

$$\begin{aligned} &\|P_l(X') - P_l(X)\|_F \\ &\leq \frac{1}{\sqrt{d/h}} \|\Gamma W_l^Q (W_l^K)^\mathsf{T} X^\mathsf{T} + X W_l^Q (W_l^K)^\mathsf{T} \Gamma^\mathsf{T} \\ &\qquad + \Gamma W_l^Q (W_l^K)^\mathsf{T} \Gamma^\mathsf{T}\|_F \end{aligned} \tag{21}$$

which leads to the following bound for (19):

$$\begin{aligned} &\frac{1}{\sqrt{d/h}} \max_{\Gamma : \|\Gamma\|_F \leq \epsilon} \|\Gamma W_l^Q (W_l^K)^\mathsf{T} X^\mathsf{T} \\ &\qquad + X W_l^Q (W_l^K)^\mathsf{T} \Gamma^\mathsf{T} + \Gamma W_l^Q (W_l^K)^\mathsf{T} \Gamma^\mathsf{T}\|_F \end{aligned} \tag{22}$$

The above problem can be searched using the projected ascent method. However, there are no polynomial-time guarantees in maximizing a fourth-order polynomial subject to a quadratic norm constraint. Fortunately, when ϵ is reasonably small, the following bound is not loose due to the negligible effects of the higher-order term. We can obtain the following bound:

$$\begin{aligned} &\max_{\Gamma : \|\Gamma\|_F \leq \epsilon} \frac{1}{\sqrt{d/h}} \|\Gamma W_l^Q (W_l^K)^\mathsf{T} X^\mathsf{T} + X W_l^Q (W_l^K)^\mathsf{T} \Gamma^\mathsf{T}\|_F \\ &\qquad + \max_{\Gamma : \|\Gamma\|_F \leq \epsilon} \frac{1}{\sqrt{d/h}} \|\Gamma W_l^Q (W_l^K)^\mathsf{T} \Gamma^\mathsf{T}\|_F \end{aligned}$$

We can easily bound the second term as

$$\max_{\Gamma : \|\Gamma\|_F \leq \epsilon} \frac{1}{\sqrt{d/h}} \|\Gamma W_l^Q (W_l^K)^\mathsf{T} \Gamma^\mathsf{T}\|_F \tag{23}$$

$$\leq \frac{\epsilon^2}{\sqrt{d/h}} \|W_l^Q (W_l^K)^\mathsf{T}\|. \tag{24}$$

In addition, the exact value of the first term can be calculated using the following lemma.

Lemma 2. *The following relation holds*

$$\begin{aligned} &\max_{\Gamma : \|\Gamma\|_F \leq \epsilon} \frac{1}{\sqrt{d/h}} \|\Gamma W_l^Q (W_l^K)^\mathsf{T} X^\mathsf{T} + X W_l^Q (W_l^K)^\mathsf{T} \Gamma^\mathsf{T}\|_F \\ &= \frac{1}{\sqrt{d/h}} \|M_l(X)\| \epsilon, \end{aligned}$$

where $M_l(X)$ is given by the following specific matrix

$$\begin{aligned} M_l(X) =& I_n \otimes \begin{bmatrix} x_1^\mathsf{T} W_l^K (W_l^Q)^\mathsf{T} \\ \vdots \\ x_n^\mathsf{T} W_l^K (W_l^Q)^\mathsf{T} \end{bmatrix} \\ &+ \sum_{i=1}^{n} (e_i \otimes I_n) \otimes (x_i^\mathsf{T} W_l^Q (W_l^K)^\mathsf{T}). \end{aligned} \tag{25}$$

The dimension of $M_l(X)$ can be quite high. A bound that can be quickly computed is given by

$$\|M_l(X)\| \leq \xi_l(X) := \Big(\|W_l^Q (W_l^K)^\mathsf{T} X^\mathsf{T}\| + \|X W_l^Q (W_l^K)^\mathsf{T}\| \Big). \tag{26}$$

With this we can now state two different bound for $\Delta_2(X, X')$. Using Equation (25), we can state the following bound

$$\begin{aligned} \delta_2^{(1)}(X, \epsilon) = \sum_{l=1}^{h} \frac{\epsilon}{\sqrt{d/h}} &\Big(\|M_l(X)\| + \epsilon \|W_l^Q (W_l^K)^\mathsf{T}\| \Big) \\ &\cdot \Big(\big\|X W_l^V W_l^O\big\| + \epsilon \big\|W_l^V W_l^O\big\| \Big). \end{aligned} \tag{27}$$

Using the relaxation of $M_l(X)$ and Equation (26), we can state a looser, but more tractable bound given by:

$$\delta_2^{(2)}(X,\epsilon) = \sum_{l=1}^{h} \frac{\epsilon}{\sqrt{d/h}} \Big(\xi_l(X) + \epsilon \big\| W_l^Q (W_l^K)^\intercal \big\| \Big) \cdot \left(\big\| X W_l^V W_l^O \big\| + \epsilon \big\| W_l^V W_l^O \big\| \right). \tag{28}$$

Obviously, we have $\delta_2^{(1)}(X,\epsilon) \leq \delta_2^{(2)}(X,\epsilon)$. Putting together all the bounds that we have obtained, we can state the following local sensitivity result, which we will refer to as **Lo**cal **F**ine-grained **A**ttention **S**ensi**T**ivity (LoFAST).

Theorem 1 (LoFAST)**.** *Consider the dot-product self-attention model (6). Suppose an input point X is given. For any X' satisfying $\|X' - X\|_F \leq \epsilon$, we have*

$$\begin{aligned} \|F(X') - F(X)\|_F &\leq \zeta(X)\epsilon + \delta_2^{(1)}(X,\epsilon) \\ &\leq \zeta(X)\epsilon + \delta_2^{(2)}(X,\epsilon), \end{aligned}$$

where $\zeta(X)$ is given by Equation (15), $\delta_2^{(1)}(X,\epsilon)$ is given by Equation (27), and $\delta_2^{(2)}(X,\epsilon)$ is given by Equation (28).

The above bounds can be used to provide a competitive method for fast scalable computation of non-trivial deterministic ℓ_2 certified robustness result of dot-product self-attention on CIFAR-10 and SVHN datasets. We will show this in the numerical result section.

Insights for Network Design. Based on the simple analytical forms of (15), (27), and (28), our bounds are highly interpretable. Our theory should not suggest that weight matrices and data with small magnitude are necessarily better for network design. The right interpretation is that our bound can be used to quantify the robustness/performance trade-off for dot-product self-attention and achieve non-vacuous certified robust accuracy. Importantly, Proposition 1 states that one needs to simultaneously make the prediction margin $\mathcal{M}_{\mathbf{f}}(X)$ large and the local sensitivity $\delta(X,\epsilon)$ small for inducing certified robustness. Based on Theorem 1, if we make $\|W_l^Q\|$, $\|W_l^K\|$, $\|W_l^V\|$, $\|W_l^O\|$, and $\|X\|$ small, then our local sensitivity bound is guaranteed to be small. However, using very small matrix norm can also make the prediction margin $M_f(X)$ small (or even vacuous) for many data points. This leads to a fundamental trade-off: we want to control the matrix norms such that the local sensitivity in Theorem 1 is not too high, while we also cannot overly reduce the matrix norms (otherwise we sacrifice clean performance and the prediction margin in Proposition 1 will become too small). From this insight, it is possible to borrow the recent advancements on how to constrain weight norms from the Lipschitz network literature to design dot-product self-attention layers with weight norm being controlled. In addition, the insight on the need of controlling $\|X\|$ further justifies the use of layer normalization in training such attention layers.

5. Experiments

In this section, we will perform numerical experiments to study the conservatism introduced in our fine-grained analysis and how to use these local bounds in a scalable manner. Furthermore, we will study how our analysis can be used to inform the design of robust self-attention blocks when applied to ViT on CIFAR-10 and SVHN datasets and explore the trade-offs between performance and robustness of our regularized ViT. For concreteness, our experiments are performed under the standard residual setting, i.e. $H = I$.

5.1. Studying Conservatism in the Local Bound

In our fine-grained local sensitivity analysis of multi-head self-attention, each derivation step used to upper-bound the output introduces conservatism. Of course, these steps are important for making the local upper-bound computationally tractable and scalable. We aim to show that the conservatism introduced by these choices does not significantly degrade the effectiveness of our approach and that our key sensitivity metric in Lemma 1 is quite informative for quantifying the robustness for small values of the ℓ_2 input perturbation level ϵ. Recall that for the multi-head self-attention block F , our analysis considers two major terms, Δ_1 and Δ_2, in upper-bounding the local perturbed output at an input X with respect to the Frobenius norm. For the first term Δ_1, we have already established a tight upper-bound $\delta_1(X,\epsilon) = \zeta(X)\epsilon$ using the key sensitivity metric, which can be readily computed by power-iterations. Now we focus on the term Δ_2.

Single-head Case: Bounding Δ_2. To upper bound Δ_2 in the single-head case, we need to compute bounds for the following two multiplicative terms.

$$\Delta_2(X,X') \leq \underbrace{\big\| P(X') - P(X) \big\|_F}_{:=\Delta_{2,1}(X,X')} \cdot \underbrace{\big\| X' W^V W^O \big\|}_{:=\Delta_{2,2}(X,X')}$$

For bounding $\Delta_{2,1}$, LoFAST (Theorem 1) offers us the following two upper bounds from our fine-grained analysis:

$$\begin{aligned} \delta_{2,1}^{(1)} &= \frac{\epsilon}{\sqrt{d/h}} \left(\|M(X)\| + \epsilon \|W^Q (W^K)^\intercal\| \right) \\ \delta_{2,1}^{(2)} &= \frac{\epsilon}{\sqrt{d/h}} \Big(\xi(X) + \epsilon \big\| W^Q (W^K)^\intercal \big\| \Big) \end{aligned}$$

where $M(X)$ is given in (25), and $\xi(X)$ is defined in (26). Clearly $\delta_{2,2}^{(1)}$ is tighter than $\delta_{2,1}^{(2)}$, but more expensive to compute. For $\Delta_{2,2}$, we compare two possible bounds for the problem defined in Equation (18); $\delta_{2,2}^{(1)}$ from LoFAST, and a more conservative naive bound $\delta_{2,2}^{(2)}$.

$$\begin{aligned} \delta_{2,2}^{(1)}(X,\epsilon) &= \big\| X W^V W^O \big\| + \epsilon \big\| W^V W^O \big\|, \\ \delta_{2,2}^{(2)}(X,\epsilon) &= \big\| W^V W^O \big\| (\|X\| + \epsilon). \end{aligned}$$

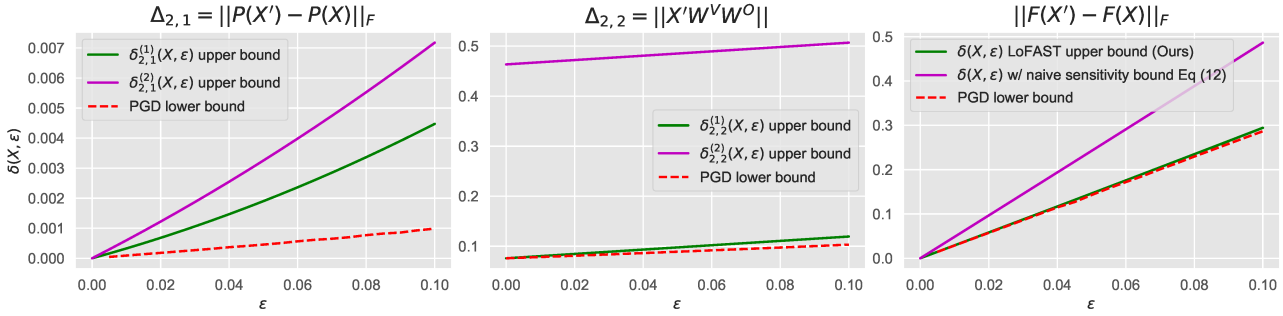

Figure 1. We compare the proposed single-head bounds for $\Delta_{2,1}$, $\Delta_{2,2}$ and the end-to-end multi-head attention bound $\|F(X') - F(X)\|_F$ across the input perturbation bound ϵ. The PGD lower bound is given by directly optimizing $\Delta_{2,1}(X, X')$, $\Delta_{2,2}(X, X')$ and $\|F(X') - F(X)\|_F$ over $X' \in \Omega(X, \epsilon)$.

To better understand the tightness of our bounds, we compare these bounds against their respective lower-bound given by directly performing PGD on $\max_{X' \in \Omega(X,\epsilon)} \Delta_{2,1}$ and $\max_{X' \in \Omega(X,\epsilon)} \Delta_{2,2}$. These results can be seen in Figure 1. It is important to note that for $\Delta_{2,1}$, although there seems to be a significant gap in the PGD lower bound and our upper bounds, the relative scale of these terms is small compared to the input perturbation for small values of ϵ (the bounds are quadratic in ϵ). For this reason, there may not be much improvement from using the tighter bound $\delta_{2,1}^{(1)}$ in most cases.

Multi-head Case: End-to-End Tightness. To further validate that Δ_1 dominates the sensitivity for controlled inputs, we compare our upper bound to the PGD lower bound of the entire multi-head attention layer with $h = 8$ heads. Through our previous study, summing over the heads, we can justify the following practical upper bound which coincides with LoFAST.

$$\begin{aligned}\|F(X') - F(X)\|_F &\leq \delta(X, \epsilon) \\ &= \zeta(X)\epsilon + \sum_{l=1}^{h} \delta_{2,1,l}^{(2)}(X, \epsilon) \cdot \delta_{2,2,l}^{(1)}(X, \epsilon).\end{aligned}$$

This multi-head bound is also evaluated in Figure 1, alongside the single-head components. To emphasize how crucial our key sensitivity metric is for tightness, we also compare the above multi-head bound with the naive sensitivity bound in Equation (12). As a lower bound, we compare against PGD which directly maximizes $\max_{X' \in \Omega(X,\epsilon)} \|F(X') - F(X)\|_F$. It becomes clear that when the spectral norm of the input X is controlled and ϵ is small, our upper bound is tight. That is because the contribution of Δ_2 is small and our bound on Δ_1 is tight. However, when the input norm of X is large, the conservative terms of $\delta_{2,1}(X, \epsilon)$, which depend on X begin to drive the estimate upwards and loosen our bound. We will use these insights to design a more robust ViT to achieve non-trivial certified accuracy on CIFAR-10 and SVHN.

5.2. Applications to Certified Robust Accuracy

Now we will apply our local analysis of the dot-product attention unit to obtain an end-to-end local sensitivity bound of ViT. With this local bound and the margin argument given in Proposition 1, we can obtain non-trivial certified robust accuracy to ℓ_2-bounded attacks. Informed by our upper bound, we can make fine-grained design choices of ViT that trade-off performance and robustness. Here we focus primarily on ViT image-classifiers, but broader data modalities such as language tasks are possible. See Appendix D for an illustrative example of certified robustness of word-embeddings for language sentiment analysis.

To achieve non-vacuous certified robustness, we need to control local sensitivity of ViT during training. Although our theory supports the standard dot-product self-attention unit commonly used in ViT, the bound needs to be propagated through other modules such as feed-forward layers and layer normalization. We use a standard ViT architecture with residual attention and feed-forward blocks and a patch-size of 16. Based on our local upper bound, we can tailor the network design to improve robustness as discussed below.

Layer Projection. Modules such as `LayerNorm` are not globally Lipschitz, but can be crucial to the performance of ViT. We instead replace `LayerNorm` units with `LayerProject` defined by:

$$\texttt{LayerProject}(x, R) = \begin{cases} \frac{x}{\|x\|_2} \cdot R & \text{if} \quad \|x\|_2 > R \\ x & \text{otherwise} \end{cases}, \tag{29}$$

where $R = \sqrt{d}$ is set to mimic the behavior of `LayerNorm`. Because projection to a closed convex set is 1-Lipschitz, we can seamlessly propagate our upper bound and maintain

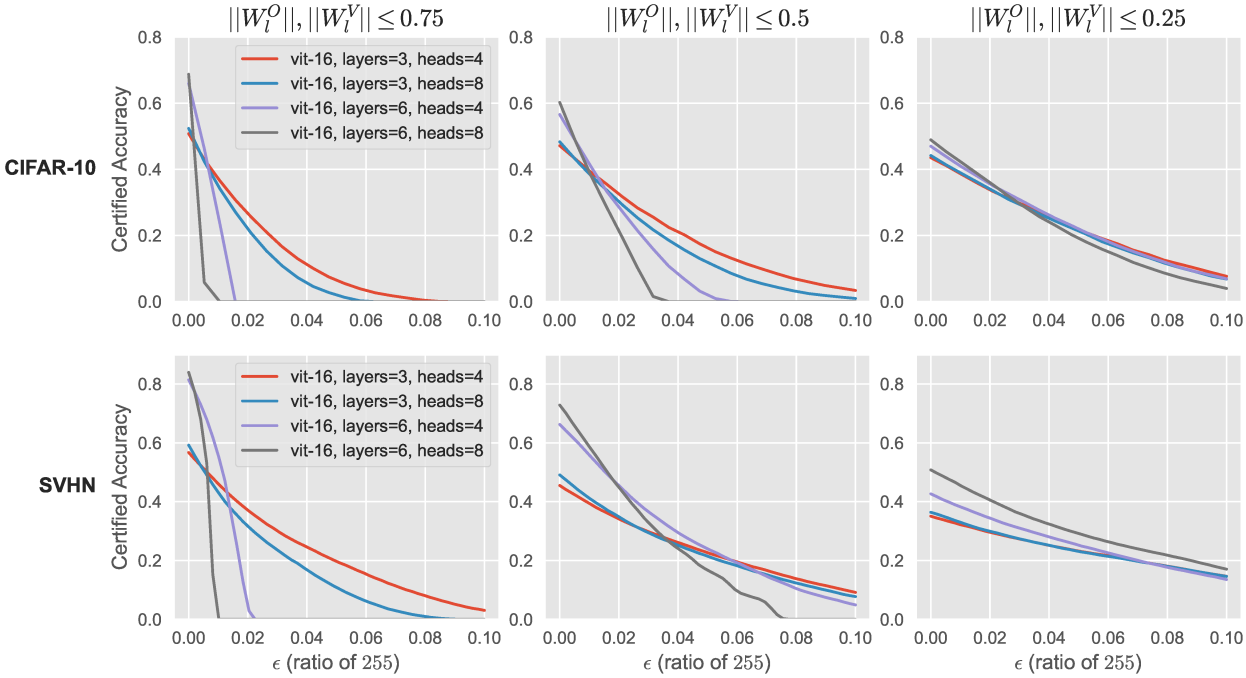

Figure 2. Certified robust accuracy on CIFAR-10 and SVHN tasks using our local sensitivity bounds under many combinations of ViT architecture parameters (number of layers, heads, and norm of weight matrices). All networks use a hidden feature dimension of 384.

the desired input scale. Additionally, `LayerProject` with $R = 1$, is applied before each attention head so that the spectral norm of the entire input $\|X\|$ is controlled. Consequently, our upper bound also remains controlled at each attention unit.

Lipschitz Constrained Layers. LoFAST will depend directly on the spectral norm of the attention weight matrices $(W_l^Q, W_l^K, W_l^V, W_l^O)$. In order to keep the expansion of our upper-bound through each layer low, we constrain the norms of these weights through SDP-based Lipschitz Layer (SLL)[1] parameterizations (Araujo et al., 2023). We also constrain all feed-forward modules and patch embedding units to be 1-Lipshitz so that we can easily propagate our upper bound with unit expansion through ViT. Such parameterizations are not uncommon. For instance, previous works on orthogonal-ViT (Fei et al., 2022) leverage orthogonal 1-Lipschitz layers in the attention unit to improve generalization on smaller data sets.

ℓ_2 Certified Robust Accuracy on CIFAR-10 and SVHN. We now apply our end-to-end local upper bound using LoFAST to obtain certified robust accuracy for ViT trained on CIFAR-10 and SVHN image datasets, both being 10-class image classification tasks. We study the effect of different ViT architecture parameters such as the number of attention heads, number of layers, and Lipschitz constant constraint of the attention weights. Our certified robustness results are presented in Figure 2. In addition to 1-Lipschitz feed-forward layers, we found it crucial to additionally constrain the Lipschitz constant of the weights W^V and W^O to be a fraction of 1. This directly affects the key sensitivity metric, which is the largest contributing factor for our local sensitivity bound on ViT. The results also show that many of these architectural factors will introduce a trade-off between clean accuracy and certified robustness. This is to be expected since, for example, the number of layers will cause our upper bound to compound, but having sufficient depth is crucial for improving clean performance.

In the extreme case of $||W_l^O||, ||W_l^V|| \leq 0.25$, our experiment makes the point that although using a smaller norm will decrease the sensitivity bound, it is not necessarily preferable since such a strict regularization will cause a loss of expressivity and sacrifice clean accuracy. It is well-known in the certified-robustness literature (Singla et al., 2022; Trockman & Kolter, 2021; Meunier et al., 2022; Prach & Lampert, 2022; Araujo et al., 2023; Wang & Manchester, 2023) that there is a trade-off between deterministic ℓ_2 certified robustness and the clean performance for standard feed-forward networks or residual networks. Therefore, it is not surprising to see a similar trade-off for ViT (notice

[1]Semidefinite programming (SDP) has been widely used to address the Lipschitz constant of deep learning models (Fazlyab et al., 2019; Pauli et al., 2021; Wang et al., 2022; Wang & Manchester, 2023; Pauli et al., 2024; Wang et al., 2024). SLL is one of most scalable Lipschitz structures derived from SDP methods.

Model	α, β-CROWN				LoFAST (Theorem 1)			
	$\epsilon = 0.02$	$\epsilon = 0.05$	$\epsilon = 0.1$	sec/sample	$\epsilon = 0.02$	$\epsilon = 0.05$	$\epsilon = 0.1$	sec/sample
ViT 3-Layer	41.2	14.1	-	6.7	29.1	16.0	3.0	0.055
ViT 4-Layer	42.9	-	-	13.3	28.9	11.9	0.8	0.077
ViT 5-Layer	44.7	-	-	20.4	27.6	8.3	0.0	0.094

Table 1: Certified robust accuracy results and average run-time for our local bound LoFAST compared to the verifier α, β-CROWN (Zhang et al., 2018; Wang et al., 2021) on CIFAR-10. In order to accommodate CROWN, we consider smaller ViTs with 128-dimensional features and only verify a subset of 1000 samples from the CIFAR-10 test set.

Dataset	Verifier	$\epsilon = 0.02$		$\epsilon = 0.05$	
		certified acc. (%)	(sec/sample)	certified acc. (%)	(sec/sample)
CIFAR-10	CROWN	41.99	4.33	14.35	6.06
	CROWN+LoFAST	**41.99**	**1.71**	**20.03**	**4.17**
SVHN	CROWN	43.73	5.68	27.45	8.60
	CROWN+LoFAST	**43.73**	**1.44**	**29.51**	**3.64**

Table 2: Certified robust accuracy results and average run-time using our local bound LoFAST as a first past and then using α, β-CROWN (Zhang et al., 2018; Wang et al., 2021) on CIFAR-10 and SVHN. We observed a significant speed up and even increased certified accuracy for $\epsilon = 0.05$, since there are some points which LoFAST can verify that CROWN can not and vice-versa. In this sense, LoFAST and CROWN can be truly complementary. In order to accommodate CROWN, we consider a smaller 3-layer ViT with 128-dimensional features and only verify a subset of 1000 samples.

that dot-product self-attention has not been covered in these previous works). We emphasize that our certified robustness results heavily rely on directly exploiting the residual structure in the key sensitivity metric of Lemma 1 (instead of the naive bound) as well as using Lipschitz controlled weights.

Comparison to General Purpose Verifier α, β-CROWN. Verifiers such as CROWN and its variants (Zhang et al., 2018; Wang et al., 2021) have been developed and integrated into the general-purpose automatic verification software AutoLiRPA (Xu et al., 2020). AutoLiRPA supports ℓ_2 perturbation models and has in the past been used for robustness certification of dot-product attention (Shi et al., 2020). However, this certification tool can be computationally expensive and has not been successfully scaled to large datasets such as CIFAR. Our analysis can serve as a complementary tool due to its enhanced scalability. To better understand how LoFAST compares to AutoLiRPA in tightness, scalability and speed, we analyze a set of smaller ViT models with 128-dim. features on a subset of the CIFAR-10 dataset (compared to 384-dim features on the entire CIFAR-10 dataset in our previous experiments). For these ViT models trained on CIFAR-10 with layers $l \in \{3, 4, 5\}$, we examine the certified robust accuracy for ℓ_2 perturbation sizes $\epsilon \in \{0.02, 0.05, 0.1\}$ and the average wall-clock time in seconds per sample on our local machine. These results are reported in Table 1. Due to memory limitations, we were not able to run AutoLiRPA on ViT with 6-layers or more. Although CROWN was able to provide tighter results for $\epsilon = 0.02$, we encountered overflow-related failures for some larger perturbation values (in Table 1, these failure entries are denoted by '-'). In addition, AutoLiRPA took considerably longer time to verify samples as the number of layers increased.

Combining LoFAST and α,β-CROWN It is possible to combine LoFAST with AutoLiRPA to achieve the best of both worlds. For example, we can always first apply LoFAST for fast verification, and then only apply AutoLiRPA to those samples that cannot be verified by LoFAST. The result of this approach for Lipschitz regularized 3-layer ViT applied to CIFAR-10 and SVHN can be found in Table 2 for perturbation sizes $\epsilon \in \{0.02, 0.05\}$. We observed a significant speed up (up to a $3\times$ speed up for smaller perturbations) and even increased certified accuracy for $\epsilon = 0.05$. This is because there are some points which LoFAST can verify that CROWN can not and vice-versa. In this sense, LoFAST and CROWN can be truly complementary.

6. Conclusion

This work has provided a fine-grained local sensitivity analysis of the standard dot-product self-attention mechanism. Our local sensitivity bound is analytical and highly interpretable, shedding light on design and sensitivity control of transformers. The theoretical results presented in this paper have been empirically validated through a comprehensive set of experiments. These findings provide a deeper understanding of the sensitivity/robustness issues of the standard dot-product self-attention models.

Acknowledgements

A. Havens and B. Hu are generously supported by the AFOSR award FA9550-23-1-0732 and the NSF award CAREER-2048168. H. Zhang was supported by the AI2050 program at Schmidt Sciences and NSF SLES award (2331967).

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning, specifically, to better understand and provably verify widely utilized transformer architectures. The pursuit of provable robustness and verification tools for neural architectures arguably allows society to minimize the harm of machine learning systems deployed in the real world.

References

Araujo, A., Havens, A. J., Delattre, B., Allauzen, A., and Hu, B. A unified algebraic perspective on Lipschitz neural networks. In *International Conference on Learning Representations*, 2023.

Ba, J. L., Kiros, J. R., and Hinton, G. E. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Bahdanau, D., Cho, K., and Bengio, Y. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

Barbara, N. H., Wang, R., and Manchester, I. R. On robust reinforcement learning with Lipschitz-bounded policy networks. *arXiv preprint arXiv:2405.11432*, 2024.

Bartlett, P. L., Foster, D. J., and Telgarsky, M. J. Spectrally-normalized margin bounds for neural networks. *Advances in neural information processing systems*, 2017.

Bonaert, G., Dimitrov, D. I., Baader, M., and Vechev, M. Fast and precise certification of transformers. In *ACM SIGPLAN International Conference on Programming Language Design and Implementation*, pp. 466–481, 2021.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *Advances in neural information processing systems*, 2020.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023.

Carlini, N., Tramer, F., Dvijotham, K. D., Rice, L., Sun, M., and Kolter, J. Z. (certified!!) adversarial robustness for free! In *International Conference on Learning Representations*, 2023.

Castin, V., Ablin, P., and Peyré, G. Understanding the regularity of self-attention with optimal transport. *arXiv preprint arXiv:2312.14820*, 2023.

Cisse, M., Bojanowski, P., Grave, E., Dauphin, Y., and Usunier, N. Parseval networks: Improving robustness to adversarial examples. In *International conference on machine learning*, pp. 854–863. PMLR, 2017.

Cohen, J., Rosenfeld, E., and Kolter, Z. Certified adversarial robustness via randomized smoothing. In *International Conference on Machine Learning*, 2019.

Dasoulas, G., Scaman, K., and 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.

Delattre, B., Barthélemy, Q., Araujo, A., and Allauzen, A. Efficient bound of Lipschitz constant for convolutional layers by Gram iteration. In *International Conference on Machine Learning*, 2023.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

Fazlyab, M., Robey, A., Hassani, H., Morari, M., and Pappas, G. Efficient and accurate estimation of Lipschitz constants for deep neural networks. *Advances in Neural Information Processing Systems*, 2019.

Fazlyab, M., Entesari, T., Roy, A., and Chellappa, R. Certified robustness via dynamic margin maximization and improved Lipschitz regularization. *Advances in Neural Information Processing Systems*, 2023.

Fei, Y., Liu, Y., Wei, X., and Chen, M. O-vit: Orthogonal vision transformer. *arXiv preprint arXiv:2201.12133*, 2022.

Gao, B. and Pavel, L. On the properties of the softmax function with application in game theory and reinforcement learning. *arXiv preprint arXiv:1704.00805*, 2017.

Havens, A., Araujo, A., Garg, S., Khorrami, F., and Hu, B. Exploiting connections between Lipschitz structures for certifiably robust deep equilibrium models. *Advances in Neural Information Processing Systems*, 2023.

Hou, B., Jia, J., Zhang, Y., Zhang, G., Zhang, Y., Liu, S., and Chang, S. Textgrad: Advancing robustness evaluation in nlp by gradient-driven optimization. In *International Conference on Learning Representations*, 2022.

Kenton, J. D. M.-W. C. and Toutanova, L. K. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, volume 1, 2019.

Kim, H., Papamakarios, G., and Mnih, A. The Lipschitz constant of self-attention. In *International Conference on Machine Learning*, pp. 5562–5571. PMLR, 2021.

Li, L. and Qiu, X. Tavat: Token-aware virtual adversarial training for language understanding. *arXiv: Computation and Language*, 2020.

Li, Q., Haque, S., Anil, C., Lucas, J., Grosse, R. B., and Jacobsen, J.-H. Preventing gradient attenuation in Lipschitz constrained convolutional networks. In *Advances in Neural Information Processing Systems*, 2019.

Meunier, L., Delattre, B. J., Araujo, A., and Allauzen, A. A dynamical system perspective for Lipschitz neural networks. In *International Conference on Machine Learning*, pp. 15484–15500. PMLR, 2022.

Miyato, T., Kataoka, T., Koyama, M., and Yoshida, Y. Spectral normalization for generative adversarial networks. In *International Conference on Learning Representations*, 2018.

Pauli, P., Koch, A., Berberich, J., Kohler, P., and Allgöwer, F. Training robust neural networks using Lipschitz bounds. *IEEE Control Systems Letters*, 6:121–126, 2021.

Pauli, P., Havens, A. J., Araujo, A., Garg, S., Khorrami, F., Allgöwer, F., and Hu, B. Novel quadratic constraints for extending lipsdp beyond slope-restricted activations. In *International Conference on Learning Representations*, 2024.

Prach, B. and Lampert, C. H. Almost-orthogonal layers for efficient general-purpose Lipschitz networks. In *European Conference on Computer Vision*. Springer, 2022.

Qi, X., Wang, J., Chen, Y., Shi, Y., and Zhang, L. Lipsformer: Introducing Lipschitz continuity to vision transformers. *arXiv preprint arXiv:2304.09856*, 2023.

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.

Sedghi, H., Gupta, V., and Long, P. M. The singular values of convolutional layers. In *International Conference on Learning Representations*, 2019.

Shi, Z., Zhang, H., Chang, K.-W., Huang, M., and Hsieh, C.-J. Robustness verification for transformers. In *International Conference on Learning Representations*, 2020.

Singla, S. and Feizi, S. Skew orthogonal convolutions. In *International Conference on Machine Learning*, 2021.

Singla, S., Singla, S., and Feizi, S. Improved deterministic l2 robustness on CIFAR-10 and CIFAR-100. In *International Conference on Learning Representations*, 2022.

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A. Y., and Potts, C. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pp. 1631–1642, 2013.

Trockman, A. and Kolter, J. Z. Orthogonalizing convolutional layers with the Cayley transform. In *International Conference on Learning Representations*, 2021.

Tsuzuku, Y., Sato, I., and Sugiyama, M. Lipschitz-margin training: Scalable certification of perturbation invariance for deep neural networks. *Advances in Neural Information Processing Systems*, 2018.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

Vuckovic, J., Baratin, A., and Combes, R. T. d. On the regularity of attention. *arXiv preprint arXiv:2102.05628*, 2021.

Wang, B., Wang, S., Cheng, Y., Gan, Z., Jia, R., Li, B., and Liu, J. Infobert: Improving robustness of language models from an information theoretic perspective. In *International Conference on Learning Representations*, 2020.

Wang, R. and Manchester, I. Direct parameterization of lipschitz-bounded deep networks. In *International Conference on Machine Learning*, pp. 36093–36110. PMLR, 2023.

Wang, S., Zhang, H., Xu, K., Lin, X., Jana, S., Hsieh, C.-J., and Kolter, J. Z. Beta-CROWN: Efficient bound propagation with per-neuron split constraints for complete and incomplete neural network verification. *arXiv preprint arXiv:2103.06624*, 2021.

Wang, Z., Prakriya, G., and Jha, S. A quantitative geometric approach to neural-network smoothness. In *Advances in Neural Information Processing Systems*, 2022.

Wang, Z., Hu, B., Havens, A. J., Araujo, A., Zheng, Y., Chen, Y., and Jha, S. On the scalability and memory efficiency of semidefinite programs for Lipschitz constant

estimation of neural networks. In *International Conference on Learning Representations*, 2024.

Wu, Q., Ye, H., Gu, Y., Zhang, H., Wang, L., and He, D. Denoising masked autoencoders help robust classification. In *International Conference on Learning Representations*, 2023.

Xiong, R., Yang, Y., He, D., Zheng, K., Zheng, S., Xing, C., Zhang, H., Lan, Y., Wang, L., and Liu, T. On layer normalization in the transformer architecture. In *International Conference on Machine Learning*. PMLR, 2020.

Xixu, H. Specformer: Guarding vision transformer robustness via maximum singular value penalization. In *Conference on Parsimony and Learning (Recent Spotlight Track)*, 2023.

Xu, K., Shi, Z., Zhang, H., Wang, Y., Chang, K.-W., Huang, M., Kailkhura, B., Lin, X., and Hsieh, C.-J. Automatic perturbation analysis for scalable certified robustness and beyond. *Advances in Neural Information Processing Systems*, 2020.

Xu, X., Li, L., and Li, B. Lot: Layer-wise orthogonal training on improving l2 certified robustness. In *Advances in Neural Information Processing Systems*, 2022.

Xu, X., Li, L., Cheng, Y., Mukherjee, S., Awadallah, A. H., and Li, B. Certifiably robust transformers with 1-Lipschitz self-attention, 2023.

Ye, W., Ma, Y., Cao, X., and Tang, K. Mitigating transformer overconfidence via Lipschitz regularization. *Conference on Uncertainty in Artificial Intelligence*, 2023.

Yu, T., Li, J., Cai, Y., and Li, P. Constructing orthogonal convolutions in an explicit manner. In *International Conference on Learning Representations*, 2022.

Zhang, B., Cai, T., Lu, Z., He, D., and Wang, L. Towards certifying l-infinity robustness using neural networks with l-inf-dist neurons. In *International Conference on Machine Learning*, pp. 12368–12379. PMLR, 2021.

Zhang, B., Jiang, D., He, D., and Wang, L. Rethinking Lipschitz neural networks and certified robustness: A boolean function perspective. *Advances in Neural Information Processing Systems*, 35:19398–19413, 2022.

Zhang, H., Weng, T.-W., Chen, P.-Y., Hsieh, C.-J., and Daniel, L. Efficient neural network robustness certification with general activation functions. *Advances in Neural Information Processing Systems*, 2018.

Zhou, D., Yu, Z., Xie, E., Xiao, C., Anandkumar, A., Feng, J., and Alvarez, J. M. Understanding the robustness in vision transformers. In *International Conference on Machine Learning*, 2022.

Zhu, C., Cheng, Y., Gan, Z., Sun, S., Goldstein, T., and Liu, J. Freelb: Enhanced adversarial training for natural language understanding. In *International Conference on Learning Representations*, 2019.

A. Local sensitivity analysis vs. local Lipschitz analysis

Local Lipschitz analysis aims at showing that for any two points (X', X'') in the ϵ-ball around X, the following bound holds

$$\|F(X') - F(X'')\|_F \leq L\|X' - X''\|_F,$$

where L is the local Lipschitz constant. This is a stronger condition than our local sensitivity analysis, and may be too strong for establishing non-trivial certified robustness results of dot-product self-attention. If one can show that the above local Lipschitz bound holds, then clearly one can choose $\delta(X, \epsilon) = L\epsilon$ to obtain a local sensitivity bound. However, given our local sensitivity bound (7), one cannot guarantee local Lipschitzness. Specifically, the local Lipschitz bound applies for any two arbitrary points in the ϵ-neighborhood of the original input X. In contrast, our local sensitivity analysis is weaker in the sense that the bound can only tell us the deviation of $F(X')$ from a fixed $F(X)$. However, that is still sufficient for computing certified robustness as stated in Proposition 1. Below we show explicitly how local-Lipscshitz-based approaches can be too conservative for computing non-trivial/practical certified robustness results.

Comparison to Existing Local-Lipschitz-based Bound in Xixu (2023). For a concrete comparison, let us examine the local Lipschitz bound from SpecFormer (Xixu, 2023), a recent work that computes the local Lipschitz bound by bounding the gradient of the self-attention unit. In Table 3, we compare the SpecFormer local-Lipschitz-based bound from (Xixu, 2023) (Theorem 4.3) and LoFAST to bound the error $\max_{X' \in \Omega(X,\epsilon)} \|F(X') - F(X)\|_F$. We can see that LoFAST is better by an order of magnitude and very close to the PGD lower-bound. As a consequence, SpecFormer is too conservative to achieve non-vacuous certified robust accuracy on CIFAR-10. Our key sensitivity metric (Lemma 1) is novel and crucial for obtaining non-vacuous certified robustness results on CIFAR-10 and SVHN.

Method	$\epsilon = 0.01$	$\epsilon = 0.03$	$\epsilon = 0.05$	$\epsilon = 0.07$	$\epsilon = 0.09$	$\epsilon = 0.10$
PGD Lower-bound	0.0286	0.0860	0.1427	0.2008	0.2582	0.2860
LoFAST Upper-bound (ours)	**0.0291**	**0.0875**	**0.1462**	**0.2052**	**0.2646**	**0.2943**
SpecFormer Upper-bound (Xixu, 2023)	16.901	52.515	90.600	131.219	174.436	197.036

Table 3: We compare our approach LoFAST against the SpecFormer method based on a local-Lipschitz bound (Xixu, 2023). We report the upper-bound $\max_{X' \in \Omega(X,\epsilon)} \|F(X') - F(X)\|_F$ for a single residual multi-head attention layer.

B. Detailed Derivations and Proofs

B.1. Proof of Proposition 1

Let X be an input and suppose that the margin of the classifier F at X satisfies $\mathcal{M}_{\mathbf{f}}(X) > \sqrt{2}\delta(X, \epsilon)$. Then for any $\|\tau\|_2 \leq \epsilon$ we have:

$$\begin{aligned}
\mathcal{M}_{\mathbf{f}}(X+\tau) =& [F(X+\tau)]_y - \max_{j \neq y} [F(X+\tau)]_j \\
=& [F(X)]_y - \max_{j \neq y} [F(X)]_j - ([F(X)]_y - [F(X+\tau)]_y) + (\max_{j \neq y} [F(X)]_j - \max_{j \neq y} [F(X+\tau)]_j) \\
=& [F(X)]_y - \max_{j \neq y} [F(X)]_k - \begin{bmatrix} 1 \\ -1 \end{bmatrix}^\top \begin{bmatrix} [F(X)]_y - [F(X+\tau)]_y \\ \max_{j \neq y}[F(X)]_j - \max_{j \neq y}[F(X+\tau)]_j \end{bmatrix} \\
\geq& [F(X)]_y - \max_{j \neq y} [F(X)]_j - \left| \begin{bmatrix} 1 \\ -1 \end{bmatrix}^\top \begin{bmatrix} [F(X)]_y - [F(X+\tau)]_y \\ \max_{j \neq y}[F(X)]_j - \max_{j \neq y}[F(X+\tau)]_j \end{bmatrix} \right| \\
\geq& [F(X)]_y - \max_{j \neq y} [F(X)]_j - \left\| \begin{bmatrix} 1 \\ -1 \end{bmatrix} \right\|_2 \|F(X) - F(X+\tau)\|_2 \\
\geq& \mathcal{M}_{\mathbf{f}}(X) - \sqrt{2}\delta(X, \epsilon) > 0
\end{aligned}$$

Therefore, $\arg\max_j [F(X+\tau)]_j = y$ for all τ such that $\|\tau\|_2 \leq \epsilon$. This completes the proof.

B.2. A Detailed Derivation of (12)

By the triangle inequality, we have

$$\begin{aligned}\Delta_1 &\leq \|H\| + \sum_{l=1}^{h} \|P_l(X)(X'-X)W_l^V W_l^O\|_F \\ &\leq \|H\| + \sum_{l=1}^{h} \|P_l(X)\| \|(X'-X)W_l^V W_l^O\|_F \\ &\leq \|H\| + \sum_{l=1}^{h} \|P_l(X)\| \|X'-X\|_F \|W_l^V W_l^O\|,\end{aligned}$$

which gives the stated bound.

B.3. Proof of Lemma 1

First, we observe that

$$\left\|H(X'-X) + \sum_{l=1}^{h} P_l(X)(X'-X)W_l^V W_l^O\right\|_F = \left\|(X'-X)^\mathsf{T} H^\mathsf{T} + \sum_{l=1}^{h} (W_l^V W_l^O)^\mathsf{T} (X'-X)^\mathsf{T} (P_l(X))^\mathsf{T}\right\|_F$$

Since $(A \otimes B)\,\mathrm{vec}(V) = \mathrm{vec}(BVA^\mathsf{T})$, we must have

$$\begin{aligned}&\mathrm{vec}\left((X'-X)^\mathsf{T} H^\mathsf{T} + \sum_{l=1}^{h} (W_l^V W_l^O)^\mathsf{T} (X'-X)^\mathsf{T} (P_l(X))^\mathsf{T}\right) \\ =& \left((H \otimes I_n) + \sum_{l=1}^{h} P_l(X) \otimes (W_l^V W_l^O)^\mathsf{T}\right) \mathrm{vec}((X'-X)^\mathsf{T}).\end{aligned}$$

Therefore, we are minimizing the ℓ_2 norm of the right side of the above equation subject to an ℓ_2 norm constraint on $\mathrm{vec}((X'-X)^\mathsf{T})$. Therefore, the maximum value is achieved by the product of the largest singular value of $\left((H \otimes I_n) + \sum_{l=1}^{h} P_l(X) \otimes (W_l^V W_l^O)^\mathsf{T}\right)$ and ϵ.

B.4. Proof of Lemma 2

To prove this lemma, we denote $\Gamma_i = x_i' - x_i \in \mathbb{R}^d$. Set $\beta_{ij} = \Gamma_i^\mathsf{T} W^Q (W^K)^\mathsf{T} x_j + x_i^\mathsf{T} W^Q (W^K)^\mathsf{T} \Gamma_j$. We can augment $\{\beta_{ij}\}$ as the following big vector:

$$\Lambda = \begin{bmatrix} \beta_{11} \\ \beta_{12} \\ \vdots \\ \beta_{1n} \\ \beta_{21} \\ \beta_{22} \\ \vdots \\ \beta_{n1} \\ \vdots \\ \beta_{nn} \end{bmatrix} = M(X) \begin{bmatrix} \Gamma_1 \\ \Gamma_2 \\ \vdots \\ \Gamma_n \end{bmatrix}$$

where $M(X)$ is given by the following specific matrix

$$
\begin{aligned}
&M(x)\\
&= \begin{bmatrix}
x_1^\top(W^K(W^Q)^\top + W^Q(W^K)^\top) & 0 & 0 & \cdots & 0\\
x_2^\top W^K(W^Q)^\top & x_1^\top W^Q(W^K)^\top & 0 & \cdots & 0\\
x_3^\top W^K(W^Q)^\top & 0 & x_1^\top W^Q(W^K)^\top & \cdots & 0\\
\vdots & \vdots & \vdots & \vdots & \vdots\\
x_n^\top W^K(W^Q)^\top & 0 & 0 & \cdots & x_1^\top W^Q(W^K)^\top\\
\vdots & \vdots & \vdots & \vdots & \vdots
\end{bmatrix}\\
&= I_n \otimes \begin{bmatrix} x_1^\top W^K(W^Q)^\top \\ \vdots \\ x_n^\top W^K(W^Q)^\top \end{bmatrix} + \sum_{i=1}^{n} (e_i \otimes I_n) \otimes (x_i^\top W^Q(W^K)^\top).
\end{aligned}
$$

Based on the above largest singular value interpretation, we can obtain the desired conclusion.

B.5. Proof of Theorem 1

We can combine Lemma 1, the bound (17), the bound (23), and Lemma 2 together, and the resultant bound is the desired one stated in this theorem.

C. Additional Experiments on Image Classification Tasks

Ablation Study of Attention Weights. In this section, we study more closely the effects of each parameter in the multi-head attention map on our bound in Theorem 1. To do this, we consider the weights (W^Q, W^K, W^V, W^O) from the first layer of a ViT trained on CIFAR-10 and a normalized input X (as the input undergoes projection prior to each attention layer in our architecture). We then perturb each element while keeping the others fixed, observing how our upper bound in Theorem 1 is affected with increasing parameter perturbation size. For the experiment, we fix $\varepsilon = 0.1$. A total of 10 samples are taken for each weight and each perturbation size. The results are presented in Figure 3. Based on this study, we can observe that the weights W^V, W^O and X account for much of the sensitivity of our bound, therefore, controlling the norm of these weights and the input is crucial to control our bound.

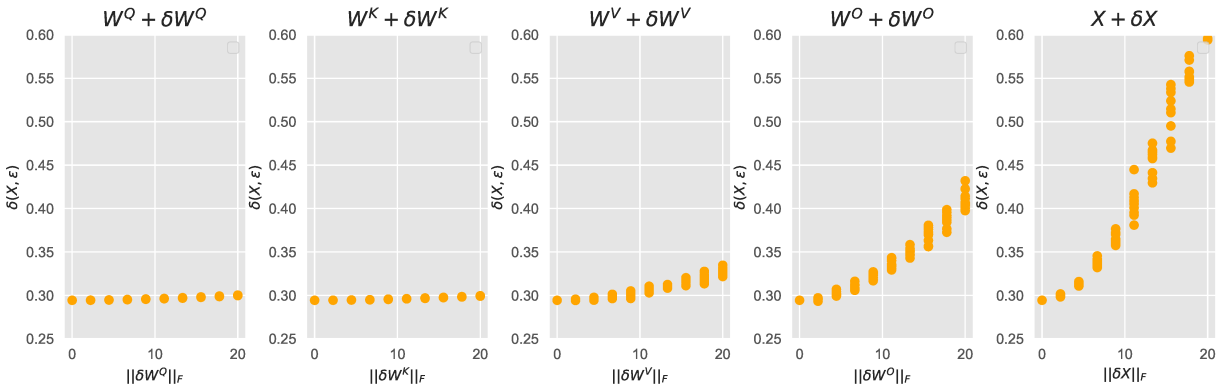

Figure 3. We perform an ablation on the weight and input and its effect on the derived local upper-bound of Theorem 1. Perturbations are applied to a trained set of self-attention weights and input (W^Q, W^K, W^V, W^O, X), perturbing one element of the tuple at a time. A total of 10 samples are taken for each parameter and each weight/input perturbation size. We set $\varepsilon = 0.1$ for all samples.

Input Norm and Tightness of Our Bounds. To study the effect of the input norm size and how it affects the tightness of our bound, we perform an extended study similar to the one in Figure 1 for several input norm scales. In this study, we are looking at the first residual self-attention layer of a pretrained ViT with 8 heads and evaluating all proposed bounds discussed in section 5.1. Our results are presented in Figure 4. We find that the bounds suggested in our paper remain tight as long as the input norm is not too large. For large inputs values, our bound eventually loses some effectiveness. This further justifies why we should perform pre-layer projection if one desires to maintain non-trivial robustness using our proposed bound.

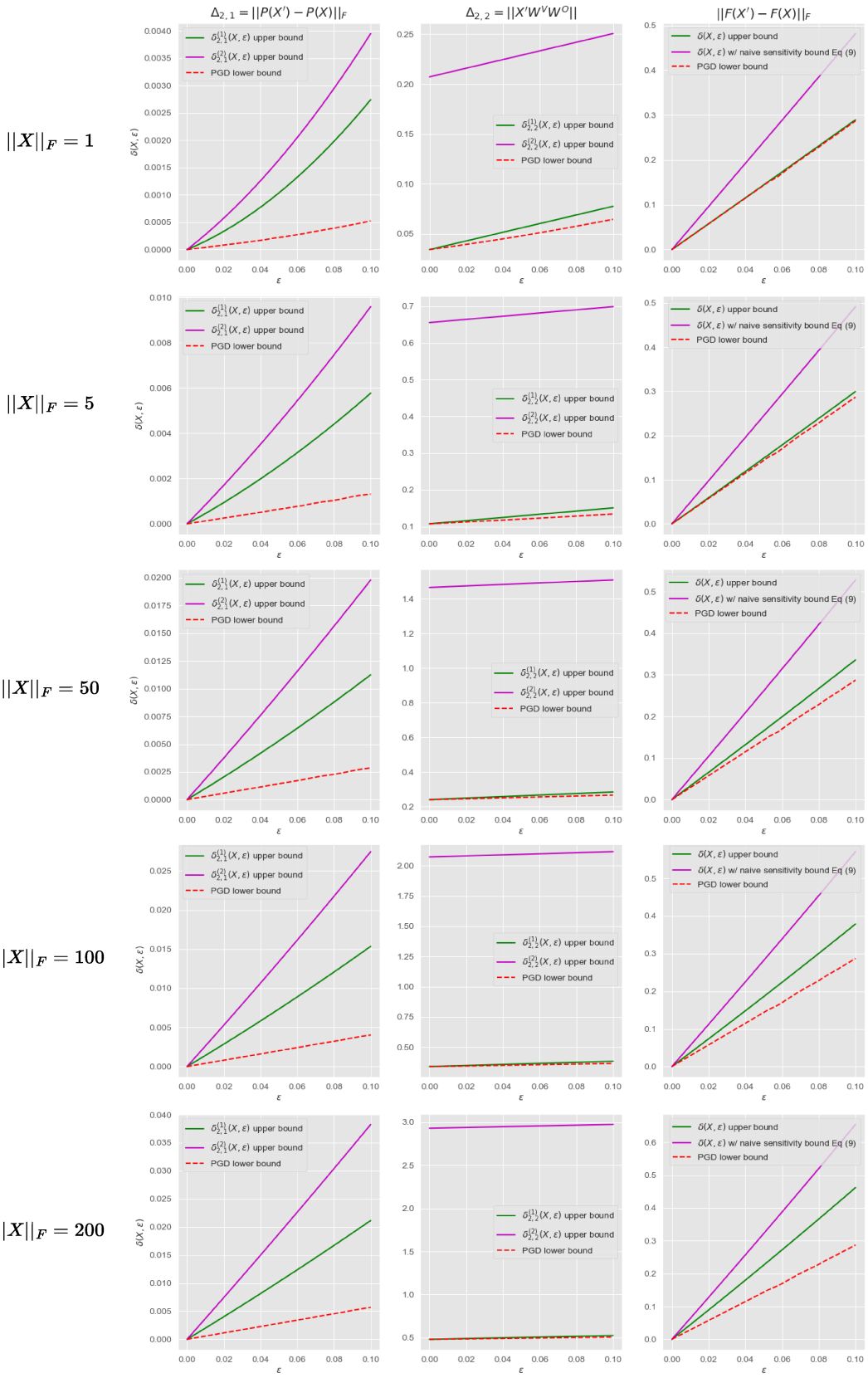

Figure 4. We repeat the conservatism study of our bound in Figure 1 with different input norms, to study the how this affects the tightness of our bound for a single multi-head attention layer.

D. Applications to Sentiment Analysis: Word Embedding Robustness

In order to broaden the application domains of our theory, we also apply our local sensitivity bounds to a sentiment analysis task, where Transformer architectures are commonly utilized (Kenton & Toutanova, 2019). In this section, we provide a robustness study on the Stanford Sentiment Tree-bank (SST) dataset (Socher et al., 2013) using transformers based on the BERT architecture (Kenton & Toutanova, 2019). We are using the version of SST that classifies sentences into two classes which indicate a positive or negative sentiment.

As in previous works (Wang et al., 2020; Zhu et al., 2019; Li & Qiu, 2020; Xu et al., 2023), we reason about ℓ_2 bounded adversarially perturbations on the word embedding space, as it is not easy to formulate perturbations on the tokens themselves using ℓ_2 perturbations. We must point out that this is a common limitation of applying sensitivity analysis to NLP benchmarks, as already noted in other works (Hou et al., 2022). The certified robustness radii we obtained measured in the ℓ_2 norm are similar to those in prior work *without* using dot-product attention (Xu et al., 2023). The results show that our sensitivity analysis bounds are indeed non-vacuous. Future study is needed to address the perturbations on the token space.

Experiment Setup for SST Sentiment Data-set Similarly to Section 5.2, we will examine the certified robust accuracy of several architectures and choices of weight norm restrictions. As mentioned before, we consider perturbations applied directly to word embeddings. In this case, ϵ describes the radius of the raw perturbation, rather a ratio of the pixel value in our vision task. The self-attention architecture designs are identical to the ones used for ViT, except we consider a embedding dimension of $d = 64$ and 32 tokens per input (i.e. $X \in \mathbb{R}^{32\times 64}$). We consider combinations of self-attention units with layers in $\{3, 6\}$ and number of heads in $\{4, 8\}$. Additionally, we train each architecture constraining the output attention weights $\{W^V, W^O\}$ to have spectral norm in $\{0.25, 0.5, 0.75\}$ using the same SLL layer. The results are presented below in Figure 5. In this case, we see that adding regularity does not necessarily decrease clean accuracy because the task is rather simple. By controlling the bound sufficiently, we can even sustain good robust accuracy while applying more layers (see the right-most panel in Figure 5).

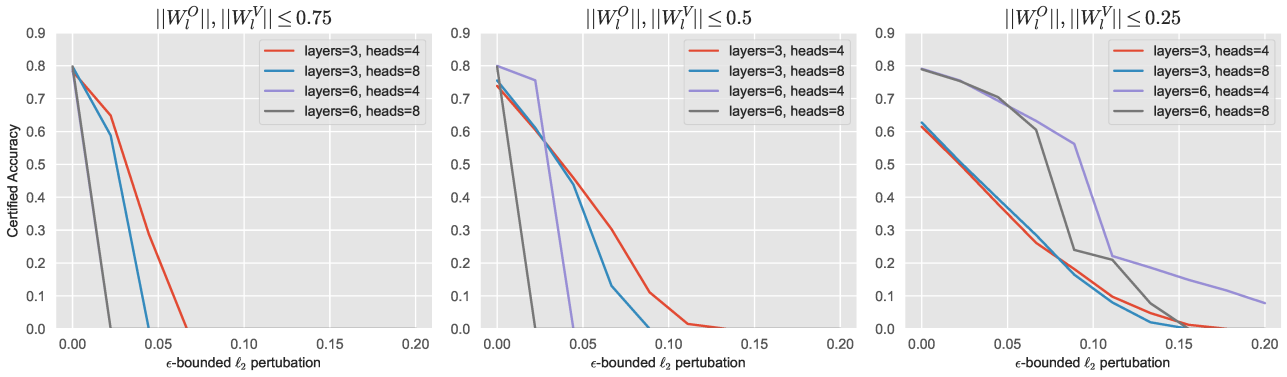

Figure 5. Certified robust accuracy on the SST using our local sensitivity bounds under many combinations of small BERT architecture parameters (number of layers, heads, and norm of weight matrices).