ROBUSTNESS VERIFICATION FOR TRANSFORMERS

Zhouxing Shi1, Huan Zhang2, Kai-Wei Chang2, Minlie Huang1, Cho-Jui Hsieh2
1Dept. of Computer Science & Technology, Tsinghua University, Beijing 10084, China
2Dept. of Computer Science, University of California, Los Angeles, CA 90095, USA
zhouxingshi@cs..tsinghua.edu.cn, huan@huan-zhang.com
kw@kwchang.net aihuang@tsinghua.edu.cn, chohsieh@cs.ucla.edu

ABSTRACT

Robustness verification that aims to formally certify the prediction behavior of neural networks has become an important tool for understanding the behavior of a given model and for obtaining safety guarantees. However, previous methods are usually limited to relatively simple neural networks. In this paper, we consider the robustness verification problem for Transformers. Transformers have complex self-attention layers that pose many challenges for verification, including cross-nonlinearity and cross-position dependency, which have not been discussed in previous work. We resolve these challenges and develop the first verification algorithm for Transformers. The certified robustness bounds computed by our method are significantly tighter than those by naive Interval Bound Propagation. These bounds also shed light on interpreting Transformers as they consistently reflect the importance of words in sentiment analysis.

1 INTRODUCTION

Deep neural networks have been successfully applied to many domains. However, a major criticism is that these black box models are difficult to analyze and their behavior is not guaranteed. Moreover, it has been shown that the predictions of deep networks become unreliable and unstable when tested in unseen situations, e.g., in the presence of small and adversarial perturbation to the input (Szegedy et al., 2013; Goodfellow et al., 2014; Lin et al., 2019). Therefore, neural network verification has become an important tool for analyzing and understanding the behavior of neural networks, with applications in safety-critical applications (Katz et al., 2017; Julian et al., 2019; Lin et al., 2019), model explanation (Shih et al., 2018) and robustness analysis (Tjeng et al., 2019; Wang et al., 2018c; Gehr et al., 2018; Wong & Kolter, 2018; Singh et al., 2018; Weng et al., 2018; Zhang et al., 2018).

Formally, a neural network verification algorithm aims to provably characterize the prediction of a network within some input space. For example, given a K-way classification model \( f : \mathbb{R}^d \rightarrow \mathbb{R}^K \), we can verify some linear specification (defined by a vector \( c \)) as below:

\[
\min_x \sum_i c_i f_i(x) \quad \text{s.t. } x \in S,
\]

where \( S \) is a predefined input space. For example, in the robustness verification problem that we are going to focus on in this paper, \( S = \{ x \mid \|x - x_0\|_p \leq \epsilon \} \) is defined as some small \( \ell_p \)-ball around the original example \( x_0 \), and setting up \( c = y_0 - 1_y \) can verify whether the logit output of class \( y_0 \) is always greater than another class \( y \) within \( S \). This is a nonconvex optimization problem which makes computing the exact solution challenging, and thus algorithms are recently proposed to find lower bounds of Eq. (1) in order to efficiently obtain a safety guarantee (Gehr et al., 2018; Weng et al., 2018; Zhang et al., 2018; Singh et al., 2019). Moreover, extension of these algorithms can be used for verifying some properties beyond robustness, such as rotation or shift invariant (Singh et al., 2019), conservation of energy (Qin et al., 2019) and model correctness (Yang & Rinard, 2019).

However, most of existing verification methods focus on relatively simple neural network architectures, such as feed-forward and recurrent neural networks, and cannot handle complex structures. In this paper, we develop the first robustness verification algorithm for Transformers (Vaswani et al., 2017) with self-attention layers. Transformers have been widely used in natural language processing (Devlin et al., 2018; Yang et al., 2019; Liu et al., 2019) and many other domains (Parmar et al., 2018).
For meaningful in practice. The importance of input words in sentiment analysis, which justifies that the computed bounds are mean-

We quantitatively and qualitatively show that the certified lower bounds consistently reflect the position dependency. Our bounds are significantly tighter than those by adapting Interval Bound Propagation (IBP) – we recursively propagate and compute linear lower bound and upper bound for each neuron with respect to the input within perturbation set $S$.

We resolve several particular challenges in verifying Transformers. **First**, Transformers with self-attention layers have a complicated architecture. Unlike simpler networks, they cannot be written as multiple layers of linear transformations or element-wise operations. Therefore, we need to propagate linear bounds differently for self-attention layers. **Second**, dot products, softmax, and weighted summation in self-attention layers involve multiplication or division of two variables under perturbation, namely cross-nonlinearity, which is not present in feed-forward networks. **Third**, neurons in each position after a self-attention layer depend on all neurons in different positions before the self-attention (namely cross-position dependency), unlike the case in recurrent neural networks where outputs depend on only the hidden features from the previous position and the current input. Previous works (Zhang et al., 2018; Weng et al., 2018; Ko et al., 2019) have to track all such dependency and thus is costly in time and memory. To tackle this, we introduce an efficient bound propagating process in a forward manner specially for self-attention layers, enabling the tighter backward bounding process for other layers to utilize bounds computed by the forward process. In this way, we avoid cross-position dependency in the backward process which is relatively slower but produces tighter bounds. Combined with the forward process, the complexity of the backward process is reduced by $O(n^2)$ for input length $n$, while the computed bounds remain comparably tight. Our contributions are summarized below:

- We propose an effective and efficient algorithm for verifying the robustness of Transformers with self-attention layers. To our best knowledge, this is the first method for verifying Transformers.
- We resolve key challenges in verifying Transformers, including cross-nonlinearity and cross-position dependency. Our bounds are significantly tighter than those by adapting Interval Bound Propagation (IBP) (Mirman et al., 2018; Gowal et al., 2018).
- We quantitatively and qualitatively show that the certified lower bounds consistently reflect the importance of input words in sentiment analysis, which justifies that the computed bounds are meaningful in practice.

## 2 Related Work

**Robustness Verification for Neural Networks.** Given an input $x_0$ and a small region $B_p(x_0, \epsilon) := \{x | \|x - x_0\|_p \leq \epsilon\}$, the goal of robustness verification is to verify whether the prediction of the neural network is unchanged within this region. This problem can be mathematically formulated as Eq. (1). If Eq. (1) can be solved optimally, then we can derive the minimum adversarial perturbation of $x$ by conducting binary search on $\epsilon$. Equivalently, we obtain the maximum $\epsilon$ such that any perturbation within $B_p(x_0, \epsilon)$ cannot change the predicted label.

Several works focus on solving Eq. (1) exactly and optimally, using mixed integer linear programming (MILP) (Jeng et al., 2019), branch and bound (BaB) (Bunel et al., 2018), and satisfiability modulo theory (SMT) (Ehlers, 2017; Katz et al., 2017). Unfortunately, due to the nonconvexity of model $f$, solving Eq. (1) is NP-hard even for a simple ReLU network (Katz et al., 2017). Therefore, we can only hope to compute a lower bound of Eq. (1) efficiently by using relaxations. Many algorithms can be seen as using convex relaxations for non-linear activation functions (Salman et al., 2019), including using duality (Wong & Kolter, 2018; Dvijotham et al., 2018), abstract domains (Gehr et al., 2018; Singh et al., 2018; Mirman et al., 2018; Singh et al., 2019), layer-by-layer reachability analysis (Wang et al., 2018b; Weng et al., 2018; Zhang et al., 2018; Gowal et al., 2018) and semi-definite relaxations (Raghunathan et al., 2018; Dvijotham et al., 2019). Additionally, robustness verification can rely on analysis on local Lipschitz constants (Hein & Andriushchenko, 2017; Zhang et al., 2019). However, existing methods are mostly limited to verifying...
networks with relatively simple architectures, such as feed-forward networks and RNNs (Wang et al., 2018a; Akintunde et al., 2019; Ko et al., 2019), while none of them are able to handle Transformers.

**Transformers and Self-Attentive Models.** Transformers (Vaswani et al., 2017) based on the self-attention mechanism, further with pre-training on large-scale corpora, such as BERT (Devlin et al., 2018), XLNet (Yang et al., 2019), RoBERTa (Li et al., 2019), achieved state-of-the-art performance on many NLP tasks. Self-attentive models are also useful beyond NLP, including VisualBERT for image generation (Parmar et al., 2018), acoustic models for speech recognition, sequential recommendation (Kang & McAuley, 2018) and graph embedding (Li et al., 2019a).

The robustness of NLP models has been studied, especially many methods have been proposed to generate adversarial examples (Papernot et al., 2016; Jia & Liang, 2017; Zhao et al., 2017; Alzantot et al., 2018; Cheng et al., 2018; Ebrahimi et al., 2018). In particular, Hsieh et al. (2019) showed that Transformers are more robust than LSTMs. However, there is not much work on robustness verification for image generation (Parmar et al., 2018), acoustic models for speech recognition, sequential recommendation and graph embedding. In this paper, we propose the first verification method for Transformers.

### 3 Methodology

We aim to verify the robustness of a Transformer whose input is a sequence of frames $X = [x^{(1)}, x^{(2)}, \ldots, x^{(n)}]$. We take binary text classification as an running example, where $x^{(i)}$ is a word embedding and the model outputs a score $y_c(X)$ for each class $c (c \in \{0, 1\})$. Nevertheless, our method for verifying Transformers is general and can also be applied in other applications.

For a clean input sequence $X_0 = [x_0^{(1)}, x_0^{(2)}, \ldots, x_0^{(n)}]$ correctly classified by the model, let $P = \{r_1, r_2, \ldots, r_t\}$ $(1 \leq r_k \leq n)$ be the set of perturbed positions, where $t$ is the number of perturbed positions. Thus the perturbed input will belong to $S_\epsilon := \{X = [x^{(1)}, x^{(2)}, \ldots, x^{(n)}] : \|x^{(r_k)} - x_0^{(r_k)}\|_p \leq \epsilon, 1 \leq k \leq t, x^{(i)} = x_0^{(i)}, \forall i \notin P\}$. Assuming that $c$ is the gold class, the goal of robustness verification is to compute

$$\min_{X \in S_\epsilon} y_c(X) - y_{1-c}(X) =: \delta_c(X).$$

If $\delta_c(X) > 0$, the output score of the correct class will always be larger than the incorrect one within $S_\epsilon$. As mentioned previously, computing the exact values of $\delta_c(X)$ is NP-hard, and thus our goal is to efficiently compute a lower bound $\delta^L_c(X) \leq \delta_c(X)$.

#### 3.1 Base Framework

We obtain $\delta^L_c(X)$ by computing the bounds of each neuron when $X$ is perturbed within $S_\epsilon$ ($\delta^L_c$ can be regarded as a final neuron). A Transformer layer can be decomposed into a number of sub-layers, where each sub-layer contains neurons after some operation. These operations can be categorized into three categories: 1) linear transformations, 2) unary nonlinear functions, and 3) operations in self-attention. Each sub-layer contains $n$ positions in the sequence and each position contains a group of neurons. We assume that the Transformer we verify has $m$ sub-layers in total, and the value of the $j$-th neuron at the $i$-th position in the $l$-th sub-layer is $\Phi^{(l,i)}_j(X)$, where $\Phi^{(l,i)}(X)$ is a vector for the specified sub-layer and position. Specially, $\Phi^{(0,i)} = x^{(i)}$ taking $l = 0$. We aim to compute a global lower bound $f_j^{(l,i),L}$ and a global upper bound $f_j^{(l,i),U}$ of $\Phi^{(l,i)}_j(X)$ for $X \in S_\epsilon$.

We compute bounds from the first sub-layer to the last sub-layer. For neurons in the $l$-th layer, we aim to represent their bounds as linear functions of neurons in a previous layer, the $l'$-th layer:

$$\sum_{k=1}^n A^{(l,i,i'),L}_{j,k} \Phi^{(l',k)}(X) + \Delta^{(l,i,i')L}_j \leq \Phi^{(l,i)}_j(X) \leq \sum_{k=1}^n A^{(l,i,i'),U}_{j,k} \Phi^{(l',k)}(X) + \Delta^{(l,i,i')U}_j, \tag{2}$$

where $A^{(l,i,i'),L}_{j,k}$, $\Delta^{(l,i,i')L}_j$ and $A^{(l,i,i'),U}_{j,k}$, $\Delta^{(l,i,i')U}_j$ are parameters of linear lower and upper bounds respectively. Using linear bounds enables us to efficiently compute bounds with a reasonable tightness. We initially have $A^{(l,i,i'),L}_{j,k} = A^{(l,i,i'),U}_{j,k} = I$ and $\Delta^{(l,i,i'),L} = \Delta^{(l,i,i'),U} = 0$. Thereby the right-hand-side of Eq. (2) equals to $\Phi^{(l,i)}_j(X)$ when $l' = l$. Generally, we use a backward
process to propagate the bounds to previous sub-layers, by substituting \( \Phi^{(t', s)} \) with linear functions of previous neurons. It can be recursively conducted until the input layer \( t' = 0 \). Since \( \Phi^{(0, k)} = x^{(k)}(\forall k \notin P) \) is constant, we can regard the bounds as linear functions of the perturbed embeddings \( \Phi^{(0, r_k)} = x^{(r_k)}(1 \leq k \leq t) \), and take the global bounds for \( x^{(r_k)} \in \mathcal{B}_p(x^{(r_k)}, \epsilon) \):

\[
\begin{align*}
\Delta_{j}^{(1, l), L} &= -\epsilon \sum_{k=1}^{t} \| \Lambda_{j,:}^{(1, i, 0, r_k), L} \|_q + \sum_{k=1}^{n} \Lambda_{j,:}^{(1, i, 0, k), L} x_0^{(k)} + \Delta_{j}^{(1, i, 0), L}, \\
\Delta_{j}^{(1, l), U} &= \epsilon \sum_{k=1}^{t} \| \Lambda_{j,:}^{(1, i, 0, r_k), U} \|_q + \sum_{k=1}^{n} \Lambda_{j,:}^{(1, i, 0, k), U} x_0^{(k)} + \Delta_{j}^{(1, i, 0), U},
\end{align*}
\]

where \( 1/p + 1/q = 1 \) with \( p, q \geq 1 \). These steps resemble to CROWN [Zhang et al., 2018] which is proposed to verify feed-forward networks. We further support verifying self-attentive Transformers which are more complex than feed-forward networks. Moreover, unlike CROWN that conducts a full backward process, we combine the backward process with a forward process (see Sec. 3.3) to reduce the computational complexity of verifying Transformers.

### 3.2 Linear Transformations and Unary Nonlinear Functions

Linear transformations and unary nonlinear functions are basic operations in neural networks. We show how bounds Eq. (2) at the \( l' \)-th sub-layer are propagated to the \( (l' - 1) \)-th layer.

**Linear Transformations** If the \( l' \)-th sub-layer is connected with the \( (l' - 1) \)-th sub-layer with a linear transformation \( \Phi^{(l', k)}(X) = W^{(l')} \Phi^{(l'-1, k)}(X) + b^{(l')} \) where \( W^{(l')}, b^{(l')} \) are parameters of the linear transformation, we propagate the bounds to the \( (l' - 1) \)-th layer by substituting \( \Phi^{(l', k)}(X) \):

\[
\Lambda^{(l, i, l'-1, k), L/U} = \Lambda^{(l, i, l', k), L/U} W^{(l')} + \sum_{k=1}^{n} \Lambda^{(l, i, l', k), L/U} b^{(l')},
\]

where “\( L/U \)” means that the equations hold for both lower bounds and upper bounds respectively.

**Unary Nonlinear Functions** If the \( l' \)-th layer is obtained from the \( (l' - 1) \)-th layer with an unary nonlinear function \( \Phi_{j}^{(l', k)}(X) = \sigma^{(l')}(\Phi_{j}^{(l'-1, k)}(X)) \), to propagate linear bounds over the nonlinear function, we first bound \( \sigma^{(l')}(\Phi_{j}^{(l'-1, k)}(X)) \) with two linear functions of \( \Phi_{j}^{(l'-1, k)}(X) \):

\[
\alpha_{j}^{(l', k), L} \Phi_{j}^{(l'-1, k)}(X) + \beta_{j}^{(l', k), L} \leq \sigma^{(l')}(\Phi_{j}^{(l'-1, k)}(X)) \leq \alpha_{j}^{(l', k), U} \Phi_{j}^{(l'-1, k)}(X) + \beta_{j}^{(l', k), U},
\]

where \( \alpha_{j}^{(l', k), L/U}, \beta_{j}^{(l', k), L/U} \) are parameters such that the inequation holds for all \( \Phi_{j}^{(l'-1, k)}(X) \) within its bounds computed previously. Such linear relaxations can be done for different functions, respectively. We provide detailed bounds for functions involved in Transformers in Appendix B.

We then back propagate the bounds:

\[
\Lambda_{j}^{(l, i, l'-1, k), L/U} = \alpha_{j}^{(l', k), L/U} \Lambda_{j}^{(l, i, l', k), L/U} + \alpha_{j}^{(l', k), U/L} \Lambda_{j}^{(l, i, l', k), L/U},
\]

\[
\Delta_{j}^{(l, i, l'-1, L/U)} = \Delta_{j}^{(l, i, l', L/U)} + \sum_{k=1}^{n} \beta_{j}^{(l', k), L/U} \Lambda_{j}^{(l, i, l', k), L/U} + \beta_{j}^{(l', k), U/L} \Lambda_{j}^{(l, i, l', k), L/U},
\]

where \( \Lambda_{j}^{(l, i, l', k), L/U} \) and \( \Lambda_{j}^{(l, i, l', k), L/U} \) mean to retain positive and negative elements in vector \( \Lambda_{j}^{(l, i, l', k), L/U} \) respectively and set other elements to 0.

### 3.3 Self-Attention Mechanism

Self-attention layers are the most challenging parts for verifying Transformers. We assume that \( \Phi^{(l-1, i)}(X) \) is the input to a self-attention layer. We describe our method for computing bounds for one attention head, and bounds for different heads of the multi-head attention in Transformers can be easily concatenated. \( \Phi^{(l-1, i)}(X) \) is first linearly projected to queries \( q^{(l, i)}(X) \), keys \( k^{(l, i)}(X) \), and values \( v^{(l, i)}(X) \) with different linear projections, and their bounds can be obtained as described in Sec. 3.2. We also keep their linear bounds that are linear functions of the perturbed embeddings.
In this way, linear bounds of \( q(k,v,L) \) are normalized into attention probabilities with a softmax, i.e. \( \tilde{\beta} \). Attention scores are computed from \( q(k,v,U) \)

### A Forward Process

We bound the output of the self-attention layer starting from \( q(l,i)(X) \), \( k(l,i)(X) \), \( v(l,i)(X) \).

#### Bounds of Multiplications and Divisions

We compute bounds with a forward process which we will show later that it can reduce the computational complexity. Attention scores are computed from \( q(l,i)(X) \) and \( k(l,i)(X) \): \( S_{l,j}^{(i)} = (q(l,i)(X))^T k(l,j)(X) = \sum_{k=1}^{d_k} q_k^{(i)}(X) k_{l,j}^{(i)}(X) \), where \( d_k \) is the dimension of \( q(l,i)(X) \) and \( k(l,j)(X) \). For each multiplication \( q_k^{(i)}(X) k_{l,j}^{(i)}(X) \), it is bounded by:

\[
\begin{align*}
q_k^{(i)}(X) k_{l,j}^{(i)}(X) & \geq \alpha_k^{(i)(L)} q_k^{(i)}(X) + \beta_k^{(i)(L)} k_{l,j}^{(i)}(X) + \gamma_k^{(i)(L)}, \\
q_k^{(i)}(X) k_{l,j}^{(i)}(X) & \leq \alpha_k^{(i)(U)} q_k^{(i)}(X) + \beta_k^{(i)(U)} k_{l,j}^{(i)}(X) + \gamma_k^{(i)(U)}.
\end{align*}
\]

We then obtain the bounds of \( S_{l,j}^{(i)} \):

\[
\begin{align*}
\Omega_{l,i}^{(i),s,L} x(r) + \Theta_{l,i}^{(i),s,L} & \leq S_{l,j}^{(i)} \leq \Omega_{l,i}^{(i),s,U} x(r) + \Theta_{l,i}^{(i),s,U}, \\
\Omega_{l,i}^{(i),s,L,U} & = \alpha_k^{(i)(L),U/L} \left( \sum_{\alpha_k^{(i)(L),U/L}} \right) + \beta_k^{(i)(L),U/L} \left( \sum_{\beta_k^{(i)(L),U/L}} \right), \\
\Omega_{l,i}^{(i),s,U,L} & = \alpha_k^{(i)(U),L/U} \left( \sum_{\alpha_k^{(i)(U),L/U}} \right) + \beta_k^{(i)(U),L/U} \left( \sum_{\beta_k^{(i)(U),L/U}} \right), \\
\Omega_{l,i}^{(i),s,U,U} & = \alpha_k^{(i)(U),U/U} \left( \sum_{\alpha_k^{(i)(U),U/U}} \right) + \beta_k^{(i)(U),U/U} \left( \sum_{\beta_k^{(i)(U),U/U}} \right)
\end{align*}
\]

In this way, linear bounds of \( q(l,i)(X) \) and \( k(l,i)(X) \) are forward propagated to \( S_{l,j}^{(i)} \). Attention scores are normalized into attention probabilities with a softmax, i.e. \( \tilde{S}_{l,j}^{(i)} = \exp(S_{l,j}^{(i)}/(\sum_{k=1}^{d_k} \exp(S_{k,j}^{(i)}))) \), where \( \tilde{S}_{l,j}^{(i)} \) is a normalized attention probability. \( \exp(S_{l,j}^{(i)}) \) is an unar way nonlinear function and can be bounded by \( \alpha_{l,j}^{(i,L)}, \beta_{l,j}^{(i,L)} \). So we forward propagate bounds of \( S_{l,j}^{(i)} \) to bound \( \exp(S_{l,j}^{(i)}) \) with \( \Omega_{l,i}^{(i),e,L} x(r) + \Theta_{l,i}^{(i),e,L} \), where:

\[
\begin{align*}
\Omega_{l,i}^{(i),e,L,U} & = \alpha_k^{(i)(L),U/L} \left( \sum_{\alpha_k^{(i)(L),U/L}} \right) + \beta_k^{(i)(L),U/L} \left( \sum_{\beta_k^{(i)(L),U/L}} \right), \\
\Omega_{l,i}^{(i),e,U,L} & = \alpha_k^{(i)(U),L/U} \left( \sum_{\alpha_k^{(i)(U),L/U}} \right) + \beta_k^{(i)(U),L/U} \left( \sum_{\beta_k^{(i)(U),L/U}} \right), \\
\Omega_{l,i}^{(i),e,U,U} & = \alpha_k^{(i)(U),U/U} \left( \sum_{\alpha_k^{(i)(U),U/U}} \right) + \beta_k^{(i)(U),U/U} \left( \sum_{\beta_k^{(i)(U),U/U}} \right)
\end{align*}
\]

By summing up bounds of each \( \exp(S_{l,j}^{(i)}) \), linear bounds can be further propagated to \( \sum_{k=1}^{d_k} \exp(S_{k,j}^{(i,k)}) \). With bounds of \( \exp(S_{l,j}^{(i)}) \) and \( \sum_{k=1}^{d_k} \exp(S_{k,j}^{(i,k)}) \) ready, we forward propagate the bounds to \( \tilde{S}_{l,j}^{(i)} \) with a division similarly to bounding \( q_k^{(i)}(X) k_{l,j}^{(i)}(X) \). The output of the self-attention \( \Phi^{(i)}(X) \) is obtained with a summation of \( \Phi^{(i)}(X) \) weighted by attention probability \( \tilde{S}_{l,j}^{(i)} \):

\[
\Phi^{(i)}(X) = \sum_{k=1}^{d_k} \tilde{S}_{l,j}^{(i)} v_{k,j}^{(i)}(X),
\]

where \( v_{k,j}^{(i)}(X) = v_{j,k}^{(i)}(X) \) whose bounds can be obtained from those of \( v_{j,k}^{(i)}(X) \).
with a transposing. Therefore, bounds of $\hat{x}_{i,j}^{(l)}$ and $\hat{y}_{i,j}^{(m)}(X)$ can be forward propagated to $\Phi^{(l,i)}(X)$ similarly to bounding $S_{i,j}^{(l)}$. In this way, we obtain the output bounds of the self-attention:

$$
\Omega_{j,:}^{(l',i),\Phi,L}(\hat{x}^{(r)}) + \Theta_{j,:}^{(l',i),\Phi,L} \hat{x}^{(r)} \leq \hat{\Phi}^{(l',i)}(X) \leq \Omega_{j,:}^{(l',i),\Phi,U}(\hat{x}^{(r)}) + \Theta_{j,:}^{(l',i),\Phi,U}.
$$

(5)

Recall that $\hat{x}^{(r)}$ is a concatenation of $x^{(r_1)}, x^{(r_2)}, \cdots, x^{(r_s)}$. We can split $\Omega_{j,:}^{(l',i),\Phi,L/U}$ into $t$ vectors with equal dimensions, $\Omega_{j,:}^{(l',i,1),\Phi,L/U}, \Omega_{j,:}^{(l',i,2),\Phi,L/U}, \cdots, \Omega_{j,:}^{(l',i,t),\Phi,L/U}$, such that Eq. (5) becomes

$$
\sum_{k=1}^{t} \Omega_{j,:}^{(l',i,k),\Phi,L} x^{(r_k)} + \Theta_{j,:}^{(l',i,k),\Phi,L} \leq \hat{\Phi}^{(l',i)}(X) \leq \sum_{k=1}^{t} \Omega_{j,:}^{(l',i,k),\Phi,U} x^{(r_k)} + \Theta_{j,:}^{(l',i),\Phi,U}.
$$

(6)

Backward Process to Self-Attention Layers  When computing bounds for a later sub-layer, the $l$-th sub-layer, using the backward process, we directly propagate the bounds at the the closest previous self-attention layer assumed to be the $l'$-th layer, to the input layer, and we skip other previous sub-layers. The bounds propagated to the $l'$-th layer are as Eq. (6). We substitute $\Phi^{(l',k)}(X)$ with linear bounds in Eq. (6):

$$
\Lambda_{j,:}^{(l,i,0,r_u),L/U} = \sum_{k'=1}^{n} \Omega_{j,:}^{(l,i,k',L/U)} + \Theta_{j,:}^{(l,i,k'),L/U} (1 \leq k \leq t),
$$

$$
\Lambda_{j,:}^{(l,i,0),L/U} = 0(\forall k \notin P),
$$

$$
\Delta_{j,:}^{(l,i,l'),L/U} = \sum_{k=1}^{n} \Omega_{j,:}^{(l,i,l',L/U)} + \Theta_{j,:}^{(l,i,l'),L/U} + \Lambda_{j,:}^{(l,i,l'),L/U}.
$$

We take global bounds as Eq. (3) and Eq. (4) to obtain the bounds of the $l$-th layer.

Advantageous of Combining the Backward Process with a Forward Process  Introducing a forward process can significantly reduce the complexity of verifying Transformers. With the backward process only, we need to compute $\Lambda^{(l,i,l',k)}$ and $\Delta^{(l,i,l')}$ ($l' \leq l$), where the major cost is on $\Lambda^{(l,i,l',k)}$ and there are $O(n^2)$ such matrices to compute. The $O(n^2)$ factor is from the dependency between all pairs of positions in the input and output respectively, which makes the algorithm inefficient especially when the input sequence is long. In contrast, the forward process represents the bounds as linear functions of the perturbed positions only instead of all positions by computing $\Omega^{(l,i,i)}$ and $\Theta^{(l,i,i)}$. Imperceptible adversarial examples may not have many perturbed positions (Gao et al., 2018; Ko et al., 2019), and thus we may assume that the number of perturbed positions, $t$, is small. The major cost is on $\Omega^{(l,i,i)}$ while there are only $O(mn)$ such matrices and the sizes of $\Lambda^{(l,i,l',k)}$ and $\Omega^{(l,i,i)}$ are relatively comparable for a small $t$. We combine the backward process and the forward process. The number of matrices $\Omega$ is $O(mn)$ in the forward process, and for the backward process, since we do not propagate bounds over self-attention layers and there is no cross-position dependency in other sub-layers, we only compute $\Lambda^{(l,i,l',k)}$ such that $i = k$, and thus the number of matrices $\Lambda$ is reduced to $O(m^2n)$. So the total number of matrices $\Lambda$ and $\Omega$ we compute is $O(m^2n)$ and is $O(n)$ times smaller than $O(m^2n^2)$ when only the backward process is used. Moreover, the backward process makes bounds tighter compared to solely the forward one, as we show in Appendix D.

4 Experiments

4.1 Datasets and Models

We conduct experiments on two sentiment analysis datasets: Yelp (Zhang et al., 2015) and SST-2 (Socher et al., 2013). Yelp consists of 560,000/38,000 examples in the training/test set and SST-2 consists of 67,349/872/1,821 examples in the training/development/test set. Each example is a sentence or a sentence segment (for the training data of SST-2 only) labeled with a sentiment polarity.

We verify the robustness of Transformers trained from scratch. For the main experiments, we consider $N$-layer models ($N \leq 3$), with 4 attention heads, hidden sizes of 256 and 512 for self-attention and feed-forward layers respectively, and we use ReLU activations for feed-forward layers. We remove the variance related terms in layer normalization, making Transformers verification bounds
tighter while the clean accuracies remain comparable (see Appendix E for discussions). Although our method can be in principal applied to Transformers with any number of layers, we do not use large-scale pre-trained models such as BERT because they are too challenging to be tightly verified for now. Current state-of-the-art verification methods for feed-forward networks either produce loose bounds for large networks (Zhang et al., 2018; Singh et al., 2019), or do not scale due to computational limits (Raghunathan et al., 2018; Dvijotham et al., 2019); Transformers contain more nonlinear operations than feed-forward networks so they are even more challenging for verification.

4.2 Certified Bounds

<table>
<thead>
<tr>
<th>Dataset</th>
<th>N</th>
<th>Acc.</th>
<th>$\ell_p$</th>
<th>Upper</th>
<th>Lower (IBP)</th>
<th>Lower (Ours)</th>
<th>Ours vs Upper</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>$\ell_1$</td>
<td>Min</td>
<td>Avg</td>
<td>Min</td>
<td>Avg</td>
</tr>
<tr>
<td>Yelp</td>
<td>1</td>
<td>91.46</td>
<td>9.332</td>
<td>3.1E-4</td>
<td>0.736</td>
<td>0.925</td>
<td>8%</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>91.54</td>
<td>11.095</td>
<td>2.2E-7</td>
<td>0.331</td>
<td>0.490</td>
<td>3%</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>91.58</td>
<td>10.745</td>
<td>8.4E-10</td>
<td>1.1E-4</td>
<td>1.2E-4</td>
<td>25%</td>
</tr>
<tr>
<td>SST</td>
<td>1</td>
<td>84.13</td>
<td>6.887</td>
<td>3.0E-4</td>
<td>2.541</td>
<td>2.747</td>
<td>37%</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>83.34</td>
<td>5.356</td>
<td>4.7E-7</td>
<td>1.661</td>
<td>1.711</td>
<td>24%</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>83.29</td>
<td>6.899</td>
<td>4.8E-10</td>
<td>1.157</td>
<td>1.186</td>
<td>17%</td>
</tr>
</tbody>
</table>

Table 1: Clean accuracies and computed bounds for 1-position perturbation. Bounds include upper bounds (obtained by an enumeration based method), certified lower bounds by IBP and our method respectively. We also report the gap between upper bounds and our lower bounds (represented as the percentage of lower bounds relative to upper bounds). We compute bounds for each possible option of perturbed positions and report the minimum (“Min”) and average (“Avg”) among them.

<table>
<thead>
<tr>
<th>N</th>
<th>Yelp</th>
<th>SST</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Lower (IBP)</td>
<td>Lower (Ours)</td>
</tr>
<tr>
<td></td>
<td>Min</td>
<td>Avg</td>
</tr>
<tr>
<td>1</td>
<td>5.3E-5</td>
<td>1.1E-4</td>
</tr>
<tr>
<td>2</td>
<td>5.8E-8</td>
<td>9.1E-8</td>
</tr>
<tr>
<td>3</td>
<td>3.1E-10</td>
<td>5.0E-10</td>
</tr>
</tbody>
</table>

Table 2: Bounds by IBP and our method for 2-position perturbation constrained by $\ell_2$-norm.

We compute certified lower bounds for different models on different datasets. We include 1-position perturbation constrained by $\ell_1/\ell_2/\ell_\infty$-norms and 2-position perturbation constrained by $\ell_2$-norm. We compare our lower bounds with those computed by the Interval Bound Propagation (IBP) (Gowal et al., 2018) baseline. For 1-position perturbation, we also compare with upper bounds computed by enumerating all the words in the vocabulary and finding the word closest to the original one such that the word substitution alters the predicted label. This method has an exponential complexity with respect to the vocabulary size and can hardly be extended to perturbations on 2 or more positions; thus we do not include upper bounds for 2-position perturbation. For each example, we enumerate possible options of perturbed positions (there are $n^2$ options), and we integrate results from different options by taking the minimum or average respectively. We report the average results on 10 correctly classified random test examples with sentence lengths no more than 32 for 1-position perturbation and 16 for 2-position perturbation. Table 1 and Table 2 present the results for 1-position and 2-position perturbation respectively. Our certified lower bounds are significantly larger and thus tighter than those by IBP. For 1-position perturbation, the lower bounds are consistently smaller than the
upper bounds, and the gap between the upper bounds and our lower bounds are reasonable compared with that in previous work on verification of feed-forward networks, e.g. in [Zhang et al., 2018], the upper bounds are in the order of 10 times larger than lower bounds. This demonstrates that our proposed method can compute robustness bounds for Transformers in a similar quality to the bounds of simpler neural networks.

### 4.3 Effectiveness of Combining the Backward Process with a Forward Process

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Acc.</th>
<th>$\ell_p$</th>
<th>Fully-Forward</th>
<th>Fully-Backward</th>
<th>Backward &amp; Forward</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Min</td>
<td>Avg</td>
<td>Time</td>
</tr>
<tr>
<td>Yelp</td>
<td>91.29</td>
<td>$\ell_1$</td>
<td>0.541</td>
<td>1.216</td>
<td>18.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\ell_2$</td>
<td>1.060</td>
<td>2.193</td>
<td>16.3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\ell_\infty$</td>
<td>0.570</td>
<td>0.976</td>
<td>16.7</td>
</tr>
<tr>
<td>SST-2</td>
<td>81.87</td>
<td>$\ell_1$</td>
<td>0.619</td>
<td>1.020</td>
<td>34.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\ell_2$</td>
<td>0.902</td>
<td>1.406</td>
<td>28.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\ell_\infty$</td>
<td>0.408</td>
<td>0.603</td>
<td>25.0</td>
</tr>
</tbody>
</table>

Table 3: Comparison of certified lower bounds and computation time (sec) by different methods.

In the following, we show the effectiveness of combining the backward process with a forward process. We compare our proposed method (Backward & Forward) to two variations: Fully-Forward propagates bounds in a forward manner for all sub-layers besides self-attention layers; Fully-Backward computes bounds for all sub-layers including self-attention layers using the backward bound propagation and without the forward process. We compare the tightness of bounds and computation time of the three methods. We use smaller models with the hidden sizes reduced by half, and we use 1-position perturbation only, to accommodate Fully-Backward with large computational cost. Experiments are conducted on an NVIDIA TITAN X GPU. Table 3 presents the results. Bounds by Fully-Forward are significantly looser while those by Fully-Backward and Backward & Forward are comparable. Meanwhile, the computation time of Backward & Forward is significantly shorter than that of Fully-Backward. This demonstrates that our method of combining the backward and forward process can compute comparably tight bounds much more efficiently.

### 4.4 Identifying Words Important to Prediction

<table>
<thead>
<tr>
<th>Method</th>
<th>Importance Score (on SST)</th>
<th>Words Identified from 10 Examples on the Yelp Dataset (split by “/”)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td><strong>Most Important Words or Symbols</strong></td>
</tr>
<tr>
<td>Grad</td>
<td>0.39</td>
<td>terrible/great/diner/best/best/food/slow/great/perfect/best</td>
</tr>
<tr>
<td>Upper</td>
<td>0.48</td>
<td>terrible/. best /and/slow/great/this/best</td>
</tr>
<tr>
<td>Ours</td>
<td>0.57</td>
<td>terrible/great/level/best/best/good/slow/great/perfect/best</td>
</tr>
</tbody>
</table>

Table 4: Average importance scores of the most/least important words identified from 100 examples respectively on SST by different methods. For the most important words identified, larger important scores are better, and vice versa. Additionally, we show most/least important words identified from 10 examples on the Yelp dataset. Boldfaced words are considered to have strong sentiment polarities, and they should appear as most important words rather than least important ones.

The certified lower bounds can reflect how sensitive a model is to the perturbation of each input token. Intuitively, if a word is more important to the prediction, the model is more sensitive to its perturbation. Therefore, the certified lower bounds can be used to identify important words. In the following, we conduct an experiment to verify whether important words can be identified by our certified lower bounds. We use a 1-layer Transformer classifier under 1-position perturbation constrained by $\ell_2$-norm. We compare our method with two baselines that also estimate local vulnerability: Upper uses upper bounds; Gradient identifies the word whose embedding has the largest $\ell_2$-norm of gradients as the most important and vice versa.

\[8\]
Quantitative Analysis on SST  SST contains sentiment labels for all phrases on parse trees, where the labels range from very negative (0) to very positive (4), and 2 for neutral. For each word, assuming its label is $x$, we take $|x - 2|$, i.e., the distance to the neutral label, as the importance score, since less neutral words tend to be more important for the sentiment polarity of the sentence. We evaluate on 100 random test input sentences and compute the average importance scores of the most or least important words identified from the examples. In Table 4, compared to the baselines (“Upper” and “Grad”), the average importance score of the most important words identified by our lower bounds are the largest, while the least important words identified by our method have the smallest average score. This demonstrates that our method identifies the most and least important words more accurately compared to baseline methods.

Qualitative Analysis on Yelp  We further analyze the results on a larger dataset, Yelp. Since Yelp does not provide per-word sentiment labels, importance scores cannot be computed as on SST. Thus, we demonstrate a qualitative analysis. We use 10 random test examples and collect the words identified as the most and least important word in each example. In Table 4, most words identified as the most important by certified lower bounds are exactly the words reflecting sentiment polarities (bold-faced words), while those identified as the least important words are mostly stopwords. Baseline methods mistakenly identify more words containing no sentiment polarity as the most important. This again demonstrates that our certified lower bounds identify word importance better than baselines and our bounds provide meaningful interpretations in practice. While gradients evaluate the sensitivity of each input word, this evaluation only holds true within a very small neighborhood (where the classifier can be approximated by a first-order Taylor expansion) around the input sentence. Our certified method gives valid lower bounds that hold true within a large neighborhood specified by a perturbation set $S$, and thus it provides more accurate results.

5 Conclusion

We propose the first robustness verification method for Transformers, and tackle key challenges in verifying Transformers, including cross-nonlinearity and cross-position dependency, for efficient and effective verification. Our method computes certified lower bounds that are significantly tighter than those by IBP. Quantitative and qualitative analyses further show that our bounds are meaningful and can reflect the importance of different words in sentiment analysis.

References


A ILLUSTRATION OF DIFFERENT BOUNDING PROCESSES

Figure 1: Illustration of three different bounding processes: Fully-Forward (a), Fully-Backward (b), and Backward&Forward (c). We show an example of a 2-layer Transformer, where operations can be divided into two kinds of blocks, “Feed-forward” and “Self-attention”. “Self-attention” contains operations in the self-attention mechanism starting from queries, keys, and values, and “Feed-forward” contains all the other operations including linear transformations and unary nonlinear functions. Arrows with solid lines indicate the propagation of linear bounds in a forward manner. Each backward arrow $A_k \rightarrow B_k$ with a dashed line for blocks $A_k, B_k$ indicates that there is a backward bound propagation to block $B_k$ when computing bounds for block $A_k$. Blocks with blue rectangles have forward processes inside the blocks, while those with green rounded rectangles have backward processes inside.

Figure 1 illustrates a comparison of the Fully-Forward, Fully-Backward and Backward & Forward processes, for a 2-layer Transformer as an example. For Fully-Forward, there are only forward processes connecting adjacent layers and blocks. For Fully-Backward, there are only backward processes, and each layer needs a backward bound propagation to all the previous layers. For our Backward & Forward algorithm, we use backward processes for the feed-forward parts and forward processes for self-attention layers, and for layers after self-attention layers, they no longer need backward bound propagation to layers prior to self-attention layers. In this way, we resolve the cross-position dependency in verifying Transformers while still keeping bounds comparably tight as those by using fully backward processes. Empirical comparison of the three frameworks are presented in Sec. 4.3.
We show in Sec. 3.2 that linear bounds can be propagated over unary nonlinear functions as long as the unary nonlinear function $\sigma(x)$, with the bounds of $x$ obtained previously as $x \in [l, u]$, we aim to derive a linear lower bound $\alpha^L x + \beta^L$ and a linear upper bound $\alpha^U x + \beta^U$, such that
\[\alpha^L x + \beta^L \leq \sigma(x) \leq \alpha^U x + \beta^U \quad (\forall x \in [l, u])\]
where parameters $\alpha^L, \beta^L, \alpha^U, \beta^U$ are dependent on $l, u$ and designed for different functions $\sigma(x)$ respectively. We introduce how the parameters are determined for different unary nonlinear functions involved in Transformers such that the linear bounds are valid and as tight as possible. Bounds of ReLU and tanh have been discussed by Zhang et al. (2018), and we further derive bounds of $e^x$, $\frac{1}{x}$, $x^2$, $\sqrt{x}$, $x^2$ and $\sqrt{\tau}$ is only used when the layer normalization is not modified for experiments to study the impact of our modification. For the following description, we define the endpoints of the function to be bounded within the range as $(l, \sigma(l))$ and $(u, \sigma(u))$. We describe how the lines corresponding to the linear bounds of different functions can be determined, and thereby parameters $\alpha^L, \beta^L, \alpha^U, \beta^U$ can be determined accordingly.

**ReLU** For ReLU activation, $\sigma(x) = \max(x, 0)$. ReLU $\sigma(x)$ is inherently linear on segments $(-\infty, 0]$ and $[0, \infty)$ respectively, so we make the linear bounds exactly as the lower bound and take a tangent line passing the two endpoints as the upper bound, and take $\sigma^L(x) = 0$ when $u < |l|$ and $\sigma^U(x) = 1$ when $u \geq |l|$ as the lower bound, to minimize the gap between the lower bound and the original function.

**Tanh** For tanh activation, $\sigma(x) = \frac{1 - e^{-2x}}{1 + e^{-2x}}$. tanh is concave for $l \geq 0$, and thus we take the line passing the two endpoints as the lower bound and take a tangent line passing $((l + u)/2, \sigma((l + u)/2))$ as the upper bound. For $u \leq 0$, tanh is convex, and thus we take the line passing the two endpoints as the upper bound and take a tangent line passing $((l + u)/2, \sigma((l + u)/2))$ as the lower bound. For $l < 0 < u$, we take a tangent line passing the right endpoint and $(d^L, \sigma(d^L))(d^L \leq 0)$ as the lower bound, and take a tangent line passing the left endpoint and $(d^U, \sigma(d^U))(d^U \leq 0)$ as the upper bound. $d^L$ and $d^U$ can be found with a binary search.

**Exp** $\sigma(x) = exp(x) = e^x$ is convex, and thus we take the line passing the two endpoints as the upper bound and take a tangent line passing $(d, \sigma(d))$ as the lower bound. Preferably, we take $d = (l + u)/2$. However, $e^x$ is always positive and used in the softmax for computing normalized attention probabilities in self-attention layers, i.e. $exp(S^{(l)})$ and $\sum_{k=1}^n exp(S^{(l)}_{i,k})$. $\sum_{k=1}^n exp(S^{(l)}_{i,k})$ appears in the denominator of the softmax, and to make reciprocal function $\frac{1}{x}$ finitely bounded the range of $x$ should not pass $0$. Therefore, we impose a constraint to force the lower bound function to be always positive, i.e. $\sigma^L(l) > 0$, since $\sigma^L(l)$ is monotonously increasing. $\sigma^L_L(x) = e^d(x - d) + e^d$ is the tangent line passing $(d, \sigma(d))$. So the constraint $\sigma^L_L(l) > 0$ yields $d < l + 1$. Hence we take $d = \min((l + u)/2, l + 1 - \Delta_d)$ where $\Delta_d$ is a small real value to ensure that $d < l + 1$ such as $\Delta_d = 10^{-2}$.

**Reciprocal** For the reciprocal function, $\sigma(x) = \frac{1}{x}$. It is used in the softmax and layer normalization and its input is limited to have $l > 0$ by the lower bounds of $exp(x)$, and $\sqrt{\tau}$. With $l > 0$, $\sigma(x)$ is convex. Therefore, we take the line passing the two endpoints as the upper bound. And we take the tangent line passing $((l + u)/2, \sigma((l + u)/2))$ as the lower bound.

**Square** For the square function, $\sigma(x) = x^2$. It is convex and we take the line passing the two endpoints as the upper bound. And we take a tangent line passing $(d, \sigma(d))(d \in [l, u])$ as the lower bound. We still prefer to take $d = (l + u)/2$. $x^2$ appears in computing the variance in layer normalization and is later passed to a square root function to compute a standard derivation. To make the input to the square root function valid, i.e. non-negative, we impose a constraint $\sigma^L_L(x) \geq 0(\forall x \in [l, u])$. $\sigma^L_L(x) = 2d(x - d) + d^2$ is the tangent line passing $(d, \sigma(d))$. For $u \leq 0$, $x^2$ is monotonously decreasing, the constraint we impose is equivalent to $\sigma^L_L(u) = 2du - d^2 \geq 0$, and with $d \leq 0$, we have $d \geq 2u$. So we take $d = \max((l + u)/2, 2u)$. For $l \geq 0$, $x^2$ is monotonously increasing, the
constraint we impose is equivalent to $\sigma^l(l) = 2dl - d^2 \geq 0$, and with $d \geq 0$, we have $d \leq 2l$. So we take $d = \max((l + u)/2, 2l)$. $\sigma^l_0(0) = -d^2$ is negative for $d \neq 0$ and is zero for $d = 0$. And for $l < 0 < u$, the constraint we impose is equivalent to $\sigma^l(0) = -d^2 \geq 0$, and thus we take $d = 0$.

**Square root** For the square root function, $\sigma(x) = \sqrt{x}$. It is used to compute a standard derivation in the layer normalization and its input is limited to be positive by the lower bounds of $x^2$ and a smoothing constant, so $l > 0$. $\sigma(x)$ is concave, and thus we take the line passing the two endpoints as the lower bound and take the tangent line passing $((l + u)/2, \sigma((l + u)/2))$ as the upper bound.

## C Linear Bounds of Multiplications and Divisions

We provide a mathematical proof of optimal parameters for linear bounds of multiplications used in Sec. 5.3. We also show that linear bounds of division can be indirectly obtained from bounds of multiplications and the reciprocal function.

For each multiplication, we aim to bound $z = xy$ with two linear bounding planes $z^L = x^L y + \beta^L y$ and $z^U = \alpha^L y + \beta^L y$, where $x$ and $y$ are both variables and $x \in [l_x, u_x], y \in [l_y, u_y]$ are concrete bounds of $x, y$ from previous layers, such that:

$$z^L = x^L y + \beta^L y + \gamma^L \leq z = xy \leq z^U = \alpha^L y + \beta^L y + \gamma^U \forall (x, y) \in [l_x, u_x] \times [l_y, u_y]$$

Our goal is to determine optimal parameters of bounding planes, i.e. $\alpha^L, \beta^L, \gamma^L, \alpha^U, \beta^U, \gamma^U$, such that the bounds are as tight as possible.

### C.1 Lower Bound of Multiplications

We define a difference function $F^L(x, y)$ which is the difference between the original function $z = xy$ and the lower bound $z^L = \alpha^L x + \beta^L y + \gamma^L$:

$$F^L(x, y) = xy - (\alpha^L x + \beta^L y + \gamma^L)$$

To make the bound as tight as possible, we aim to minimize the integral of the difference function $F^L(x, y)$ on our concerned area $(x, y) \in [l_x, u_x] \times [l_y, u_y]$, which is equivalent to maximizing

$$V^L = \int_{x \in [l_x, u_x]} \int_{y \in [l_y, u_y]} \alpha^L x + \beta^L y + \gamma^L \tag{7}$$

while $F^L(x, y) \geq 0 \forall (x, y) \in [l_x, u_x] \times [l_y, u_y])$. For an optimal bounding plane, there must exist a point $(x_0, y_0) \in [l_x, u_x] \times [l_y, u_y]$ such that $F^L(x_0, y_0) = 0$ (otherwise we can validly increase $\gamma^L$ to make $V^L$ larger). To ensure that $F^L(x, y) \geq 0$ within the concerned area, we need to ensure that the minimum value of $F^L(x, y)$ is non-negative. We show that we only need to check cases when $(x, y)$ is any of $(l_x, l_y), (l_x, u_y), (u_x, l_y), (u_x, u_y)$, i.e. points at the corner of the considered area. The partial derivatives of $F^L$ are:

$$\frac{\partial F^L}{\partial x} = y - \alpha^L$$

$$\frac{\partial F^L}{\partial y} = x - \beta^L$$

If there is $(x_1, y_1) \in \{l_x, u_x\} \times \{l_y, u_y\}$ such that $F^L(x_1, y_1) \leq F(x, y) \forall (x, y) \in [l_x, u_x] \times [l_y, u_y]$, $\frac{\partial F^L}{\partial x}(x_1, y_1) = \frac{\partial F^L}{\partial y}(x_1, y_1) = 0$ should hold. Thereby $\frac{\partial F^L}{\partial x}(x, y), \frac{\partial F^L}{\partial y}(x, y) < 0 \forall (x, y) \in [l_x, x_1] \times [l_y, y_1])$, and thus $F^L(x_1, y_1) < F^L(x_1, y_1)$ cannot be the point with the minimum value of $F^L(x, y)$. On the other hand, if there is $(x_1, y_1)(x_1 = l_x, y_1 \in (l_y, u_y))$, i.e. on one border of the concerned area but not on any corner, $\frac{\partial F^L}{\partial y}(x_1, y_1) = 0$ should hold.

Thereby $\frac{\partial F^L}{\partial y}(x, y) = \frac{\partial F^L}{\partial y}(x_1, y_1) = 0 \forall (x, y), x = x_1 = l_x$. So $F^L(x_1, y_1) = F^L(x_1, l_y) = F^L(l_x, l_y)$. This property holds for the other three borders of the concerned area. Therefore, other
points within the concerned area cannot have smaller function value \( F^L(x, y) \), so we only need to check the corners, and the constraints on \( F^L(x, y) \) become

\[
\begin{align*}
F^L(x_0, y_0) &= 0 \\
F^L(l_x, l_y) &\geq 0 \\
F^L(l_x, u_y) &\geq 0 \\
F^L(u_x, l_y) &\geq 0 \\
F^L(u_x, u_y) &\geq 0
\end{align*}
\]
equivalent to

\[
\begin{align*}
\gamma^L &= x_0 y_0 - \alpha^L x_0 - \beta^L y_0 \\
l_x l_y - \alpha^L (l_x - x_0) - \beta^L (l_y - y_0) - x_0 y_0 &\geq 0 \\
l_x u_y - \alpha^L (l_x - x_0) - \beta^L (u_y - y_0) - x_0 y_0 &\geq 0 \\
u_x l_y - \alpha^L (u_x - x_0) - \beta^L (l_y - y_0) - x_0 y_0 &\geq 0 \\
u_x u_y - \alpha^L (u_x - x_0) - \beta^L (u_y - y_0) - x_0 y_0 &\geq 0
\end{align*}
\]

We substitute \( \gamma^L \) in Eq. (7) with Eq. (8), yielding

\[
V^L = V_0[(l_x + u_x - 2x_0)\alpha^L + (l_y + u_y - 2y_0)\beta^L + 2x_0 y_0]
\]

where \( V_0 = \frac{(u_x - l_x)(u_y - l_y)}{2} \).

We have shown that the minimum function value \( F^L(x, y) \) within the concerned area cannot appear in \((l_x, u_x) \times (l_y, u_y)\), i.e. it can only appear at the border, while \((x_0, y_0)\) is a point with a minimum function value \( F^L(x_0, y_0) = 0\), \((x_0, y_0)\) can also only be chosen from the border of the concerned area. At least one of \( x_0 = l_x \) and \( x_0 = u_x \) holds.

If we take \( x_0 = l_x \):

\[
V_1^L = V_0[(u_x - l_x)\alpha^L + (l_y + u_y - 2y_0)\beta^L + 2l_x y_0]
\]

And from Eq. (8) we obtain

\[
\alpha^L \leq \frac{u_x l_y - l_x y_0 - \beta^L (l_y - y_0)}{u_x - l_x}
\]

\[
\alpha^L \leq \frac{u_x u_y - l_x y_0 - \beta^L (u_y - y_0)}{u_x - l_x}
\]

\[
l_x \leq \beta^L \leq l_x \Rightarrow \beta^L = l_x
\]

Then

\[
V_1^L = V_0[(u_x - l_x)\alpha^L + l_x(l_y + u_y)]
\]

\[
(u_x - l_x)\alpha^L \leq -l_x y_0 + \min(u_x l_y - \beta^L (l_y - y_0), u_x u_y - \beta^L (u_y - y_0))
\]

\[
= -l_x y_0 + \min(u_x l_y - l_x (l_y - y_0), u_x u_y - l_x (u_y - y_0))
\]

\[
= (u_x - l_x) \min(l_y, u_y)
\]

\[
= (u_x - l_x) l_y
\]

So

\[
\alpha^L \leq l_y
\]

To maximize \( V_1^L \), since now only \( \alpha^L \) is unknown in \( V_1^L \) and the coefficient of \( \alpha^L \), \( V_0(u_x - l_x) \geq 0 \), we take \( \alpha^L = l_y \), and then

\[
V_1^L = V_0(u_x l_y + l_x u_y)
\]
is a constant.

For the other case if we take \( x_0 = u_x \):

\[
V_2^L = V_0[(l_x - u_x)\alpha^L + (l_y + u_y - 2y_0)\beta^L + 2u_x y_0]
\]

\[
\alpha^L \geq \frac{l_x l_y - u_x y_0 - \beta^L (l_y - y_0)}{l_x - u_x}
\]

16
If we take \( x \)  

Published as a conference paper at ICLR 2020  

Then  

So \( \alpha \)  

We take \( \alpha \)  

To minimize \( V \)  

We notice that \( \alpha \)  

\[
\alpha^L \geq \frac{l_x u_y - u_x y_0 - \beta^L (u_y - y_0)}{l_x - u_x}
\]

\[
u_x \leq \beta^L \leq u_x \iff \beta^L = u_x
\]

\[
V_2^L = V_0[(l_x - u_x)\alpha^L + u_x(l_y + u_y)]
\]

\[
(l_x - u_x)\alpha^L \leq -u_x y_0 + \min(l_x l_y - \beta^L (l_y - y_0), l_x u_y - \beta^L (u_y - y_0))
\]

\[
= \min(l_x l_y - u_x l_y, l_x u_y - u_x u_y)
\]

\[
= (l_x - u_x) \max(l_y, u_y)
\]

\[
= (l_x - u_x) u_y
\]

So

\[
\alpha^L \geq u_y
\]

We take \( \alpha^L = u_y \) similarly as in the case when \( x_0 = l_x \), and then

\[
V_2^L = V_0(l_x u_y + u_x l_y)
\]

We notice that \( V_1^L = V_2^L \), so we can simply adopt the first one. We also notice that \( V_1^L, V_2^L \) are independent of \( y_0 \), so we may take any \( y_0 \) within \([l_y, u_y]\) such as \( y_0 = l_y \). Thereby, we obtain the a group of optimal parameters of the lower bounding plane:

\[
\begin{cases}
\alpha^L = l_y \\
\beta^L = l_x \\
\gamma^L = -l_x l_y
\end{cases}
\]

**C.2 Upper Bound of Multiplications**

We derive the upper bound similarly. We aim to minimize

\[
V^U = V_0[(l_x + u_x - 2x_0)\alpha^U + (l_y + u_y - 2y_0)\beta^U + 2x_0 y_0]
\]

where \( V_0 = \frac{(u_x - l_x)(u_y - l_y)}{2} \).

If we take \( x_0 = l_x \):

\[
V_1^U = V_0[(u_x - l_x)\alpha^U + (l_y + u_y - 2y_0)\beta^U + 2l_x y_0]
\]

\[
\alpha^U \geq \frac{u_x l_y - l_x y_0 - \beta^U (l_y - y_0)}{u_x - l_x}
\]

\[
\alpha^U \geq \frac{u_x u_y - l_x y_0 - \beta^U (u_y - y_0)}{u_x - l_x}
\]

\[
l_x \leq \beta^U \leq l_x \iff \beta^U = l_x
\]

Then

\[
V_1^U = V_0[(u_x - l_x)\alpha^U + l_x(l_y + u_y)]
\]

\[
(u_x - l_x)\alpha^U \geq -l_x y_0 + \max(u_x l_y - \beta^U (l_y - y_0), u_x u_y - \beta^U (u_y - y_0))
\]

\[
= \max(u_x l_y - l_x l_y, u_x u_y - l_x u_y)
\]

\[
= (u_x - l_x) \max(l_y, u_y)
\]

\[
= (u_x - l_x) u_y
\]

So

\[
\alpha^U \geq u_y
\]

To minimize \( V_1^U \), we take \( \alpha^U = u_y \), and then

\[
V_1^U = V_0(l_x l_y + u_x u_y)
\]
For the other case if we take $x_0 = u_x$:

$$V^U_2 = V_0[(l_x - u_x) \alpha^U + (l_y + u_y - 2y_0) \beta^U + 2u_x y_0]$$

$$\alpha^U \leq \frac{l_x l_y - u_x y_0 - \beta^U (l_y - y_0)}{l_x - u_x}$$

$$\alpha^U \leq \frac{l_x u_y - u_x y_0 - \beta^U (u_y - y_0)}{l_x - u_x}$$

$$u_x \leq \beta^U \leq u_x \iff \beta^U = u_x$$

Then

$$V^U_2 = V_0[(l_x - u_x) \alpha^U + u_x (l_y + u_y)]$$

$$(l_x - u_x) \alpha^U \geq -u_x y_0 + \max(\{l_x l_y - \beta^U (l_y - y_0), l_x u_y - \beta^U (u_y - y_0)\})$$

$$= \max(\{l_x l_y - u_x l_y, l_x u_y - u_x u_y\})$$

$$= (l_x - u_x) \min(l_y, u_y)$$

$$= (l_x - u_x) l_y$$

So

$$\alpha^U \leq l_y$$

To minimize $V^U_2$, we take $\alpha^U = l_y$, and then

$$V^L_2 = V_0(l_x l_y + u_x u_y)$$

Since $V^U_1 = V^U_2$, we simply adopt the first case. And $V^U_1, V^U_2$ are independent of $y_0$, so we may take any $y_0$ within $[l_y, u_y]$ such as $y_0 = l_y$. Thereby, we obtain a group of optimal parameters of the upper bounding plane:

$$\begin{align*}
\alpha^U &= u_y \\
\beta^U &= l_x \\
\gamma^U &= -l_x u_y
\end{align*}$$

C.3 LINEAR BOUNDS OF DIVISIONS

We have shown that closed-form linear bounds of multiplications can be derived. However, we find that directly bounding $z = \frac{x}{y}$ is tricky. If we try to derive a lower bound $z^L = \alpha^L x; \beta^L y + \gamma$ for $z = \frac{x}{y}$ as Appendix C.1, the difference function is

$$F^L(x, y) = \frac{x}{y} - (\alpha^L x + \beta^L y + \gamma^L)$$

The partial derivatives of $F^L$ are:

$$\frac{\partial F^L}{\partial x} = \frac{1}{y} - \alpha^L$$

$$\frac{\partial F^L}{\partial y} = -\frac{x}{y^2} - \beta^L$$

If there is $(x_1, y_1) \in (l_x, u_x) \times (l_y, u_y)$ such that $\frac{\partial F^L}{\partial x}(x, y) = 0$, unlike that the case for multiplications, none of areas $[l_x, x_1) \times [l_y, y_1)$, $[l_x, x_1) \times (y_1, u_y)$, $(x_1, u_x) \times [l_y, y_1)$ and $(x_1, u_x) \times (y_1, u_y)$ has guarantees that $\frac{\partial F^L}{\partial x}(x, y) = \frac{\partial F^L}{\partial y}(x, y) < 0$ for all $(x, y)$ with in the corresponding area. It is possible that $(x_1, y_1)$ does have a minimum function value $F^L(x_1, y_1)$ among all $(x, y) \in [l_x, u_x] \times [l_y, u_y]$, i.e. a minimum function value of $F^L(x, y)$ for $(x, y)$ within the concerned area may appear at a point other than the corners. For example, for $l_x = 0.05, u_x = 0.15, l_y = 0.05, u_y = 0.15, \alpha = 10, \beta = -20, \gamma = 2$, the minimum function value of $F^L(x, y)$ for $(x, y) \in [0.05, 0.15] \times [0.05, 0.15]$ appears at $(0.1, 0.1)$ which is not a corner of $[0.05, 0.15] \times [0.05, 0.15]$. This makes it more difficult to derive closed-form parameters such that the constraints on $F^L(x, y)$ are satisfied. Fortunately, we can bound $z = \frac{x}{y}$ indirectly by utilizing bounds of multiplications and reciprocal functions. We bound $z = \frac{x}{y}$ by first bounding a unary function $\overline{y} = \frac{1}{y}$ and then bounding the multiplication $z = x\overline{y}$.
D Tightness of Bounds by the Backward Process and Forward Process

We have discussed that combining the backward process with a forward process can reduce computational complexity, compared to the method with the backward process only. But we only use the forward process for self-attention layers and do not fully use the forward process for all sublayers, because bounds by the forward process can be looser than those by the backward process. We compare the tightness of bounds by the forward process and the backward process respectively. To illustrate the difference, for simplicity, we consider a \( m \)-layer feed-forward network \( \Phi^{(0)} = x \), \( y^{(l)} = W^{(l)}\Phi^{(l-1)}(x) + b^{(l)} \), \( \Phi^{(l)}(x) = \sigma(y^{(l)}(x)) \), \( 0 < l \leq m \), where \( x \) is the input vector, \( W^{(l)} \) and \( b^{(l)} \) are the weight matrix and the bias vector for the \( l \)-th layer, \( y^{(l)}(x) \) is the pre-activation vector of the \( l \)-th layer, \( \Phi^{(l)}(x) \) is the vector of neurons in the \( l \)-th layer, and \( \sigma(\cdot) \) is an activation function. Before taking global bounds, both the backward process and the forward process bound \( \Phi^{(l)}(x) \) with linear functions of \( x \). When taking global bounds as Eq. [3] and Eq. [4], only the norm of weight matrix is directly related to the \( \epsilon \) in the binary search for certified lower bounds. Therefore, we measure the tightness of the computed bounds using the difference between weight matrices for lower bounds and upper bounds respectively. We show how it is computed for the forward process and the backward process respectively.

D.1 The Forward Process

For the forward process, we bound each neuron \( \Phi_j^{(l)}(x) \) with linear functions:

\[
\Omega_{j;L}^{(l)} x + \Theta_{j;L}^{(l)} \leq \Phi_j^{(l)}(x) \leq \Omega_{j;U}^{(l)} x + \Theta_{j;U}^{(l)}
\]

To measure the tightness of the bounds, We are interested in \( \Omega_{j;L}^{(l)}, \Omega_{j;U}^{(l)} \), and also \( \Omega_{j;U}^{(l)} \) - \( \Omega_{j;L}^{(l)} \).

Initially,

\[
\Omega^{(0);L/U} = I, \quad \Theta^{(0);L/U} = 0, \quad \Omega^{(0);U} - \Omega^{(0);L} = 0
\]

We can forward propagate the bounds of \( \Phi^{(l-1)}(x) \) to \( y^{(l)}(x) \):

\[
\Omega_{j;L}^{(l)} y^{(l)} + \Theta_{j;L}^{(l)} \leq y^{(l)}(x) \leq \Omega_{j;U}^{(l)} y^{(l)} + \Theta_{j;U}^{(l)}
\]

where

\[
\Omega_{j;L}^{(l)} = \sum_{w_{j,i}^{(l)} > 0} W_{j,i}^{(l)} \Omega_{i;L}^{(l-1),L} + \sum_{w_{j,i}^{(l)} < 0} W_{j,i}^{(l)} \Omega_{i;L}^{(l-1),U/L}
\]

\[
\Theta_{j;L}^{(l)} = \sum_{w_{j,i}^{(l)} > 0} W_{j,i}^{(l)} \Theta_{i}^{(l-1),L} + \sum_{w_{j,i}^{(l)} < 0} W_{j,i}^{(l)} \Theta_{i}^{(l-1),U/L} + b^{(l)}
\]

With global bounds of \( y^{(l)}(x) \) which can be obtained as Eq. [3] and Eq. [4], we bound the activation function:

\[
\alpha_j^{(l)} y^{(l)} + \beta_j^{(l)} \leq \sigma(y^{(l)}(x)) \leq \alpha_j^{(l)} y^{(l)} + \beta_j^{(l)}
\]

And then bounds can be propagated from \( \Phi^{(l-1)}(x) \) to \( \Phi^{(l)}(x) \):

\[
\Omega_j^{(l)} = \begin{cases} 
\alpha_j^{(l)} & \Theta_j^{(l)} \leq 0 \\
\Omega_j^{(l)} & \Theta_j^{(l)} > 0 
\end{cases}
\]

\[
\Theta_j^{(l)} = \begin{cases} 
\alpha_j^{(l)} + \beta_j^{(l)} & \Theta_j^{(l)} \leq 0 \\
\Theta_j^{(l)} + \beta_j^{(l)} & \Theta_j^{(l)} > 0 
\end{cases}
\]

Therefore,

\[
\Omega_j^{(l)} - \Omega_j^{(l)} = (\alpha_j^{(l)} - \alpha_j^{(l)}) |W_{j}^{(l)}| (\Omega_j^{(l-1),U} - \Omega_j^{(l-1),L}) \tag{9}
\]

illustrates how the tightness of the bounds is changed from earlier layers to later layers.
D.2 THE BACKWARD PROCESS AND DISCUSSIONS

For the backward process, we bound the neurons in the $l$-th layer with linear functions of neurons in a previous layer, the $l'$-th layer:

$$
\Phi_{l'}(x) = \Lambda_{l',j} \Phi_j(x) + \Delta_{l',j} \leq \Phi_j(x) \leq \Lambda_{l',j} \Phi_j(x) + \Delta_{l',j} = \Phi_{l'}(x)
$$

We have shown in Sec. 3.1 how such bounds can be propagated to $l' = 0$, for the case when the input is sequential. For the nonsequential case we consider here, it can be regarded as a special case when the input length is 1. So we can adopt the method in Sec. 3.1 to propagate bounds for the feed-forward network we consider here. We are interested in $\Lambda_{l',j}, \Lambda_{l',j} - \Lambda_{l,j}$ and also $\Lambda_{l',j} - \Lambda_{l,j}$. Weight matrices of linear bounds before taking global bounds are $\Lambda_{l,l} - \Lambda_{l,l}$, which are obtained by propagating the bounds starting from $\Lambda_{l,l} = \Lambda_{l,l} = I$. According to bound propagation described in Sec. 3.2,

$$
\Lambda_{l,l} - \Lambda_{l,l} = \left(\alpha_{l',j} \Phi_j(x) - \lambda_{l',j} \Phi_j(x) - \lambda_{l,j} \Phi_j(x) + \lambda_{l,j} \Phi_j(x)\right) W_l
$$

illustrates how the tightness bounds can be measured during the backward bound propagation until $l' = 0$.

There is a $W_l'$ in Eq. (10) instead of $W_l'$ in Eq. (9). The norm of $(\Omega_{l,j} - \Omega_{l,j})$ in Eq. (9) can easily grow larger as $l$ increases during the forward propagation when $||W_j||$ is greater than 1, while this generally holds for neural networks to have $||W_j||$ greater than 1 in feed-forward layers. While in Eq. (10), $W_l'$ can have both positive and negative elements and tends to allow cancellations for different $W_j'$, and thus the norm of $(\Lambda_{l',j} - \Lambda_{l',j} - \Lambda_{l,j} - \Lambda_{l,j})$ tends to be smaller. Therefore, the bounds computed by the backward process tend to be tighter than those by the forward framework, which is consistent with our experiment results in Table 3.

E IMPACT OF MODIFYING THE LAYER NORMALIZATION

The original Transformers have a layer normalization after the embedding layer, and two layer normalization before and after the feed-forward part in each Transformer layer. We modify the layer normalization, $f(x) = w(x - \mu)/\sigma + b$, where $x$ is $d$-dimensional a vector to be normalized, $\mu$ and $\sigma$ are the mean and standard derivation of $\{x_i\}$ respectively, and $w$ and $b$ are gain and bias parameters respectively. $\sigma = \sqrt{(1/d) \sum_{i=1}^d (x_i - \mu)^2 + \epsilon_s}$ where $\epsilon_s$ is a smoothing constant. It involves $(x_i - \mu)^2$ whose linear lower bound is loose and exactly 0 when the range of the $x_i - \mu$ passes 0. When the $\ell_2$ norm of the perturbation is relatively larger, there can be many $x_i - \mu$ with ranges passing 0, which can cause the lower bound of $\sigma$ to be small and thereby the upper bound of $f(x)$ to be large. This can make the certified bounds loose. To resolve this, we modify the layer normalization into $f(x) = w(x - \mu) + b$ by removing the standard derivation term. We use an experiment to study the impact of this modification. We compare the clean accuracies and certified bounds of the models with modified layer normalization to models with standard layer normalization and with no layer normalization respectively. Table 3 presents the results. Certified lower bounds of models with no layer normalization or our modification are significantly tighter than those of corresponding models with the standard layer normalization. Meanwhile, the clean accuracies of models with our modification are comparable with those of the original model, and particularly higher than the 1-layer model with no layer normalization on SST. This demonstrates that it is worthwhile to modify the layer normalization in Transformers, to make the models better to be verified while keeping comparable clean accuracies.
<table>
<thead>
<tr>
<th>Dataset</th>
<th>N</th>
<th>LayerNorm</th>
<th>Acc.</th>
<th>$\ell_p$ Upper</th>
<th>Acc.</th>
<th>$\ell_p$ Lower</th>
<th>Acc.</th>
<th>Ours vs Upper</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Min Avg</td>
<td></td>
<td>Min Avg</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Standard</td>
<td>91.64</td>
<td>$\ell_1$ 188.305 197.873</td>
<td>0.012</td>
<td>0.023</td>
<td>6.6E-5 1.2E-4</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>$\ell_2$ 15.115 15.363</td>
<td>0.010</td>
<td>0.020</td>
<td>6.7E-4 1.3E-3</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>$\ell_\infty$ 2.029 2.932</td>
<td>0.002</td>
<td>0.005</td>
<td>9.5E-4 1.8E-3</td>
<td></td>
</tr>
<tr>
<td>Yelp</td>
<td>1</td>
<td>None</td>
<td>91.46</td>
<td>$\ell_1$ 8.025 12.479</td>
<td>0.764</td>
<td>0.953</td>
<td>10% 8%</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>$\ell_2$ 0.583 0.878</td>
<td>0.298</td>
<td>0.365</td>
<td>51% 42%</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>$\ell_\infty$ 0.084 0.131</td>
<td>0.030</td>
<td>0.038</td>
<td>36% 29%</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ours</td>
<td>91.46</td>
<td>$\ell_1$ 9.332 13.577</td>
<td>0.736</td>
<td>0.925</td>
<td>8% 7%</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>$\ell_2$ 0.708 0.988</td>
<td>0.286</td>
<td>0.362</td>
<td>40% 37%</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>$\ell_\infty$ 0.119 0.159</td>
<td>0.031</td>
<td>0.040</td>
<td>26% 25%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>Standard</td>
<td>91.96</td>
<td>$\ell_1$ 191.899 201.829</td>
<td>0.003</td>
<td>0.005</td>
<td>1.7E-5 2.3E-5</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>$\ell_2$ 15.281 15.543</td>
<td>0.002</td>
<td>0.003</td>
<td>1.3E-4 2.1E-4</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>$\ell_\infty$ 2.002 3.033</td>
<td>0.000</td>
<td>0.001</td>
<td>1.4E-4 1.7E-4</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>None</td>
<td>91.59</td>
<td>$\ell_1$ 8.145 15.021</td>
<td>0.603</td>
<td>0.751</td>
<td>7% 5%</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>$\ell_2$ 0.592 1.034</td>
<td>0.134</td>
<td>0.168</td>
<td>23% 16%</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>$\ell_\infty$ 0.088 0.136</td>
<td>0.011</td>
<td>0.014</td>
<td>13% 10%</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ours</td>
<td>91.54</td>
<td>$\ell_1$ 11.095 16.522</td>
<td>0.331</td>
<td>0.490</td>
<td>3% 3%</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>$\ell_2$ 0.840 1.175</td>
<td>0.105</td>
<td>0.144</td>
<td>13% 12%</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>$\ell_\infty$ 0.133 0.188</td>
<td>0.010</td>
<td>0.013</td>
<td>7% 7%</td>
<td></td>
</tr>
<tr>
<td>SST</td>
<td>1</td>
<td>Standard</td>
<td>84.06</td>
<td>$\ell_1$ 191.424 195.492</td>
<td>0.008</td>
<td>0.015</td>
<td>4.1E-5 7.4E-5</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>$\ell_2$ 15.549 15.632</td>
<td>0.006</td>
<td>0.012</td>
<td>3.8E-4 7.9E-4</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>$\ell_\infty$ 2.262 2.503</td>
<td>0.001</td>
<td>0.003</td>
<td>4.6E-4 1.2E-3</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>None</td>
<td>82.58</td>
<td>$\ell_1$ 6.887 8.859</td>
<td>2.541</td>
<td>2.747</td>
<td>37% 34%</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>$\ell_2$ 0.529 0.655</td>
<td>0.408</td>
<td>0.448</td>
<td>77% 68%</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>$\ell_\infty$ 0.088 0.110</td>
<td>0.032</td>
<td>0.035</td>
<td>36% 32%</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ours</td>
<td>84.13</td>
<td>$\ell_1$ 6.887 8.859</td>
<td>2.541</td>
<td>2.747</td>
<td>37% 34%</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>$\ell_2$ 0.529 0.655</td>
<td>0.408</td>
<td>0.448</td>
<td>77% 68%</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>$\ell_\infty$ 0.088 0.110</td>
<td>0.032</td>
<td>0.035</td>
<td>36% 32%</td>
<td></td>
</tr>
</tbody>
</table>

Table 5: Clean accuracies, upper bounds, certified lower bounds by our method of models with different layer normalization settings.