Fast and Precise Certification of Transformers

Gregory Bonaert
ETH Zurich, Switzerland
greg@gregbonaert.com

Dimitar I. Dimitrov
ETH Zurich, Switzerland
dimitar.iliev.dimitrov@inf.ethz.ch

Maximilian Baader
ETH Zurich, Switzerland
mbaader@inf.ethz.ch

Martin Vechev
ETH Zurich, Switzerland
martin.vechev@inf.ethz.ch

Abstract
We present DeepT, a novel method for certifying Transformer networks based on abstract interpretation. The key idea behind DeepT is our new Multi-norm Zonotope abstract domain, an extension of the classical Zonotope designed to handle $\ell_1$ and $\ell_2$-norm bound perturbations. We introduce all Multi-norm Zonotope abstract transformers necessary to handle these complex networks, including the challenging softmax function and dot product.

Our evaluation shows that DeepT can certify average robustness radii that are $28 \times$ larger than the state-of-the-art, while scaling favorably. Further, for the first time, we certify Transformers against synonym attacks on long sequences of words, where each word can be replaced by any synonym. DeepT achieves a high certification success rate on sequences where enumeration-based verification would take 2 to 3 orders of magnitude more time.

CCS Concepts: • Theory of computation → Program verification; Abstraction; • Computing methodologies → Neural networks.

Keywords: Abstract Interpretation, Robustness Certification, Deep Learning, Adversarial attacks, Transformer Networks

ACM Reference Format:

1 Introduction
It has been established that neural networks can be tricked into misclassifying images by small, imperceptible changes to the input, called adversarial attacks [52]. Since their discovery, the community has developed many different attacks, including $\ell_p$-norm bound perturbations [18, 36, 52] as well as geometric perturbations like rotations and translations [13, 25, 42]. The existence of adversarial examples has driven the need for methods and tools able to formally verify the robustness of neural networks against these attacks. Early works on $\ell_p$-norm bound certification included [15, 56], while later research investigated certification against geometric transformations [3, 14, 39, 44, 50]. While many works focused on certifying networks operating on continuous input spaces such as images, substantially less research exists on the setting of networks with a discrete input space, as in natural language processing (NLP) tasks, where the input is a word sequence. Recently, however, neural networks for NLP tasks were successfully attacked by exchanging words in the input sequence with synonyms [1, 5, 22, 23, 31, 34].

Transformer Networks. A successful recent neural network architecture for NLP is the Transformer network [54]. Since its introduction, these networks have delivered state-of-the-art performance across many NLP tasks [27, 58]. Further, Transformer networks were successfully applied to a
wide variety of other tasks including object detection [7], graph classification [32], speech recognition [20], and visual question answering [33]. The core part of a Transformer network is a stack of self-attention layers. For NLP tasks, the first layer of this stack is preceded by an embedding layer, mapping words to vectors in an embedding space.

**Certification of Transformer Networks.** As the set of words is discrete, to perform certification in the NLP setting, one might try enumeration in order to certify the absence of adversarial examples, by classifying every combination of synonyms. However, while this works well for short sentences consisting of words having few synonyms, enumeration is already infeasible for moderately sized sequences where a few words have many synonyms. A method to circumvent enumeration is to approximate the region of synonym embeddings using an abstract shape. This shape can then be propagated through the network (as standard in neural network verification), as illustrated in Figure 1. In the recent work of [47], a linear relaxation was used to certify Transformer networks against \( p \)-norm bound attacks applied to word embeddings for \( p \in \{1, 2, \infty\} \). However, the work of [47] does not scale to large Transformer networks, with their verifiers having either a dramatically reduced certification ability as the network becomes deeper or becoming impractically slow (superlinear verification time growth). The main challenge in Transformer network verification is creating scalable and precise methods to handle the effects of the softmax function and the dot product. This is difficult as these functions apply non-linearities on multiple terms under perturbation, simultaneously. Specifically, the challenge for the dot product are the many occurrences of multiplication, while the challenge for the softmax function is the occurrence of the exponential and reciprocal.

**This Work: Precise Certification of Transformers.** In this work, we address this challenge and present a verifier called DeepT, able to certify significantly larger Transformer networks, compared to prior work. The key idea is a new abstract domain, referred to as the Multi-norm Zonotope, which improves the precision of the standard Zonotope (that has so far been successfully used for robustness certification of images [49, 55]). Our central contribution is a set of carefully designed abstract transformers that handle critical components of the Transformer architecture, including the challenging self-attention (composed of dot products and softmax functions), balancing scalability and precision.

We specifically focus on encoder Transformer networks and show that DeepT can certify on average significantly higher robustness radii (up to more than 28x), particularly of deep Transformer networks, compared to state-of-the-art [47], in comparable time. Further, DeepT can certify robustness against attacks where each word can be replaced by synonyms. For long sentences, DeepT outperforms enumeration-based methods by 2 to 3 orders of magnitude.

**Main Contributions.** Our main contributions are:
- The first robustness certification for synonym attacks on Transformer network classification, where each word in long word sequences can potentially be replaced by any synonym, simultaneously.
- The Multi-norm Zonotope domain together with abstract transformers for all functions occurring in Transformer networks.
- Carefully constructed abstract transformers for the dot-product and the softmax function, that strike a good balance between precision and performance.
- A verifier called DeepT based on our abstract transformers, and an extensive experimental evaluation showing DeepT is able to scale certification to significantly larger Transformer networks.

**2 Overview**

We now explain the threat model we are certifying against and how this threat model relates to other threat models. Then, we provide an informal explanation of our method for certifying the robustness of Transformer networks. A detailed explanation is provided in Section 4 and Section 5.

**Threat Models.** In this work we focus on binary sentiment classification, where each input sentence is either positive or negative, a standard NLP classification task.

In our threat model T1, the adversary can perturb an input sequence by adding \( p \)-noise \((p \in \{1, 2, \infty\})\) to the embedding of a word. The threat model T2 allows the attacker to exchange every word in the input sequence with a synonym.

The two threat models are related since a good embedding will map synonyms to points in the embedding space that are close to each other [40]. If we certify robustness for an \( p \)-norm bound region around a point in the embedding space, then all words (synonyms) that have their embedding in that region will not change the classification of that sentence. Further, if we have a word we would like to replace, we can certify an \( \ell^p \)-norm bound ball covering the embeddings of all the replacement words, in order to certify against attacks as in [1]. The process is depicted in Figure 1.

**Robustness Certification.** Inspired by the success of the Zonotope domain in the context of robustness certification for images [15, 49], we introduce the Multi-norm Zonotope in order to certify the robustness of Transformer networks performing binary sequence classification. The Multi-norm Zonotope is an extension of the classical Zonotope domain, containing new noise symbols bounded by an \( \ell^p \)-norm. This improves certification against \( \ell^p \)-norm bound attacks.

We define abstract transformers for the Multi-norm Zonotope domain for all operations in the Transformer network, including affine operations, ReLU, tanh, exponential, reciprocal, dot product and softmax. It is particularly challenging to tightly approximate the dot product and the softmax...
function, since they may apply multiplications or divisions between two variables under perturbation. Additionally, the quadratic number of interactions between noise symbols leads to a blow-up of the number of noise symbols which slows down computation or even renders it infeasible. To address these issues, we carefully construct the abstract transformers for the softmax and dot product to achieve high precision while maintaining good performance. Further, the number of noise symbols is periodically reduced to lower memory use and improve performance.

We certify the robustness of a Transformer network by constructing a Multi-norm Zonotope capturing the ℓp region around the embedding of a given input sequence (See Figure 1), which is then propagated through the whole network using our abstract transformers. Thus, we obtain a Multi-norm Zonotope representing an over-approximation of the possible outputs of the Transformer network. Finally, robustness can be certified if the lower bound of \( y_t - y_f \) is positive where \( y_t \) is the true class output and \( y_f \) is the false class output.

3 Background

In this section, we describe the background needed for the rest of the paper. We first demonstrate the Transformer architecture instantiated to binary sequence classification, after which we explain the Zonotope domain and how Zonotopes are used for neural network verification.

3.1 Transformer Networks

In this work, we apply Transformer networks to the task of binary sequence classification, like sentiment prediction. The transformer network takes as input a sequence of tokens (e.g., words) and processes them to get a binary answer (Figure 2).

The length of the sequence is variable and is denoted by \( N \).

A Transformer first embeds each token \( t_j \) into \( \mathbb{R}^E \) via a discrete mapping. Then a position embedding \( b_j \) is added, to facilitate the encoding of the location \( j \). This encoded sequence is then passed through \( M \) Transformer network layers (Figure 3), after which the encoded sequence passes through a pooling layer followed by a classifier.

**Self-Attention.** The key component of a Transformer network layer (Figure 3) is its self-attention mechanism, which processes all sequence components in conjunction, and thus is able to extract relationships between them.

The self-attention receives \( N \) inputs \( \bar{x}_1, \ldots, \bar{x}_N \in \mathbb{R}^E \). We transpose and stack them into the matrix \( X \in \mathbb{R}^{N \times E} \) to simplify notation. \( X \) is then multiplied by 3 separate matrices \( W_Q, W_K, W_V \in \mathbb{R}^{E \times d_k} \) and \( W_T \in \mathbb{R}^{E \times d_v} \) to obtain the queries \( Q := X W_Q \), the keys \( K := X W_K \) and the values \( V := X W_V \) of the embeddings \( X \). The self-attention output \( Z \) is then obtained, by computing

\[
Z := \sigma \left( \frac{QK^T}{\sqrt{d_k}} \right) V = \sigma \left( \frac{XW_QW_K^{T}X^T}{\sqrt{d_k}} \right) XW_V,
\]

where \( Z \in \mathbb{R}^{N \times E} \) and the softmax function \( \sigma: \mathbb{R}^N \rightarrow \mathbb{R}^N \)

\[
\sigma_i(v_1, \ldots, v_N) = \frac{e^{v_i}}{\sum_{j=1}^{N} e^{v_j}}
\]

is applied to every element row by row. Here, \( \sigma \) indicates the \( i \)-th component of \( \sigma \). The rows \( \bar{z}_1, \ldots, \bar{z}_N \) of \( Z \) are then returned as output embeddings of the self-attention layer.

We note that the softmax is an integral part of the transformer network itself, contrary to most other neural network architectures where it is only used after the final classification layer to define the training loss.

**Multi-Head Self-Attention.** Similarly to convolutional neural networks, where multiple convolution filters can be used per layer, multiple self-attentions (called attention heads) can be combined in one layer. If we have \( A \) attention heads, each with its own \( W_K, W_Q \) and \( W_V \) matrix, we get \( A \) matrices \( Z_a \), where \( a \in \{1, \ldots, A\} \). These are horizontally stacked and multiplied from the right with a matrix \( W_0 \in \mathbb{R}^{(Ad_e) \times E} \), resulting in a matrix \( Z \in \mathbb{R}^{N \times E} \). As before, rows \( \bar{z}_1, \ldots, \bar{z}_N \) of \( Z \) are then returned.

**Residual Connections and Normalization.** The inputs \( \bar{x}_j \) of each multi-head self-attention layer are added to the corresponding outputs \( \bar{z}_j \), thus creating a residual connection [21]. This is followed by a layer normalization step [2] but without the division by the standard deviation. [47] found that not dividing by the standard deviation has no significant impact on the performance, but improves certification rates (see Section 6.6). Thus, each vector \( \bar{z}_j \) is normalized by mapping it to \( \bar{z}_j - \text{mean}(\bar{z}_j) \). This normalized value is then multiplied by a parameter after which a bias is added.

**Feed-Forward Network.** The normalization is followed by applying the same feed-forward neural network, mapping \( \mathbb{R}^{N} \) to \( \mathbb{R}^{N} \), to each of the \( N \) output embeddings. This network consists of one hidden ReLU layer of size \( H \) (hidden size). Similarly to the Multi-head self-attention, the feed-forward networks are encapsulated with a residual connection, followed by a normalization layer as described above.

**Pooling and Classification.** In order for the network to produce a final classification output, after the \( M \)-th transformer network layer, we pick its first output embedding and disregard the others (pooling). The first output embedding is then pushed through a tanh hidden layer, followed by a binary linear classifier (Figure 2).

3.2 Zonotope Certification

Our work builds upon the Zonotope abstract domain, which was used to successfully certify neural networks applied to image classification tasks [15, 49].

**Zonotopes.** A classical Zonotope [16] abstracts a set of \( N \in \mathbb{N} \) variables and associates the \( k \)-th variable with an
affine expression $x_k$ using $E \in \mathbb{N}$ noise symbols defined by

$$x_k = c_k + \sum_{i=1}^E \beta_k^i \epsilon_i = c_k + \tilde{\beta}_k \cdot \tilde{\epsilon},$$

where $c_k, \beta_k^i \in \mathbb{R}$ and $\epsilon_i \in [-1, 1]$. The value $x_k$ can deviate from its center coefficient $c_k$ through a series of noise symbols $\epsilon_i$ scaled by the coefficients $\beta_k^i$. The set of noise symbols $\tilde{\epsilon}$ is shared among different variables, thus encoding dependencies between $N$ values abstracted by the zonotope.

**Affine Abstract Transformer.** All affine arithmetic operations such as sums, subtractions or scalings can be applied directly in the abstract Zonotope domain. Since affine arithmetic is exact, the corresponding abstract transformers are optimal. Given two variables $x_1 = c_1 + \beta_1 \cdot \tilde{\epsilon}$ and $x_2 = c_2 + \beta_2 \cdot \tilde{\epsilon}$, the abstract transformer for the affine operation $z = ax_1 + bx_2 + c$ is

$$z = (ac_1 + bc_2 + c) + (a\tilde{\beta}_1 + b\tilde{\beta}_2) \cdot \tilde{\epsilon}.$$

**ReLU Abstract Transformer.** We leverage the minimal area abstract transformer for the ReLU developed in [49], which is applied element-wise. Concretely, the Zonotope abstract transformer for $\text{ReLU}(x) = \max(0, x)$ of the zonotope variable $x$ is

$$y = \begin{cases} 0, & \text{if } u < 0 \\ x, & \text{if } l > 0 \\ \lambda x + \mu + \beta_{\text{new}} e_{\text{new}} & \text{otherwise}, \end{cases}$$

where $l$ and $u$ denote the lower and upper bound of $x$ and

$$\lambda = u/(u-l), \quad \mu = 0.5 \max(-\lambda l, (1-\lambda)u), \quad \beta_{\text{new}} = 0.5 \max(-\lambda l, (1-\lambda)u).$$

**Certification.** To certify the robustness of a neural network on an input region, we express that region in zonotope form and propagate it through the network using abstract transformers. The resulting output zonotope is an over-approximation of the possible outputs of the network. Given the correct class output $y_t$ and the incorrect class output $y_f$, the robustness is proven if the lower bound of $y_t - y_f$ is positive.

### 3.3 Dual Norm

For a given vector $\tilde{x} \in \mathbb{R}^N$, the dual norm $\| \cdot \|^\ast_p$ of the $\ell^p$ norm is defined by [43] as:

$$\|\tilde{x}\|_p^\ast = \sup \{ \tilde{z} \cdot \tilde{x} \mid \tilde{x} \in \mathbb{R}^N, \|\tilde{x}\|_p \leq 1 \}. \quad (3)$$

Further, [43] shows that $\| \cdot \|^\ast_p$ is the $\ell^q$ norm with $q$ satisfying the relationship $\frac{1}{p} + \frac{1}{q} = 1$.

**Lemma 1.** Retain the above notation. The tight lower and upper bounds of $\tilde{z} \cdot \tilde{x}$ where $\tilde{x} \in \mathbb{R}^N$ s.t. $\|\tilde{x}\|_p \leq 1$ are given by

$$l_k^n = -\|\tilde{z}\|_q \quad u_k^n = \|\tilde{z}\|_q.$$

Proof. The proof follows directly from the definition of the dual norm in Equation 3 and the fact that the dual norm of the $\ell^p$ norm is the $\ell^q$ norm where $\frac{1}{p} + \frac{1}{q} = 1$. The bounds are tight because $\{ x \in \mathbb{R}^N \mid \|x\|_p \leq 1 \}$ is compact and the inner product is continuous, thus the supremum is achieved. \qed
While $\ell^\infty$ perturbations can be naturally expressed with a classical Zonotope, other perturbations such as $\ell^1$ or $\ell^2$-norm balls captured by a classical Zonotope would introduce heavy over-approximation. To address this, we introduce the Multi-norm Zonotope abstract domain, an extension of the classical Zonotope which in addition to the standard noise symbols $\epsilon_i$, contains noise symbols $\phi_j$, fulfilling the constraint $\|\phi\|_p \leq 1$, where $\hat{\phi} := (\phi_1, \ldots, \phi_{E_p})^T$ (Figure 4). If $p = \infty$, we recover the classical Zonotope. This new domain allows to easily express $\ell^p$-norm bound balls in terms of the new noise symbols $\phi$:

$$x_k = c_k + \sum_{i=1}^{E_p} a_{k,i} \phi_i + \sum_{j=1}^{E_\infty} \beta_{k,j} \epsilon_j = c_k + \tilde{a}_k \cdot \hat{\phi} + \tilde{\beta}_k \cdot \hat{\epsilon},$$

$$c_k, a_{k,i}, \beta_{k,j} \in \mathbb{R}, \quad \|\phi\|_p \leq 1, \quad \epsilon_j \in [-1, 1],$$

where $E_p$ and $E_\infty$ denote the number of $\ell^p$ and $\ell^\infty$ noise symbols respectively.

To concisely represent the $N$ zonotope output variables $x_1, \ldots, x_N$, we write $\mathbf{x} := (x_1, \ldots, x_N)^T$. Therefore, the Multi-norm Zonotope $\mathbf{x}$ can be simplified to

$$\mathbf{x} = \mathbf{c} + A \mathbf{\phi} + B \mathbf{\epsilon}$$

$$\mathbf{c} \in \mathbb{R}^N, A \in \mathbb{R}^{N \times E_p}, B \in \mathbb{R}^{N \times E_\infty} \quad \|\hat{\phi}\|_p \leq 1, \quad \epsilon_j \in [-1, 1],$$

where $A_{k,i} = a_{k,i}$ and $B_{k,j} = \beta_{k,j}$. In the rest of the section, we describe how to compute the concrete bounds of the Multi-norm zonotope and present Multi-norm Zonotope abstract transformers for all operations of the Transformer network (Figures 2 and 3). We extend the existing abstract Zonotope transformers [49] to our Multi-norm Zonotope domain and describe new and efficient abstract transformers for the dot product and the softmax function. To support softmax, we describe the abstract transformers for the exponential and the reciprocal. Further, we prove soundness of our abstract transformers and show that the ReLU, Tanh, Exponential and Reciprocal transformers are optimal with respect to the area of the corresponding input-output transformer spaces.

4.1 Computing Concrete Upper and Lower Bounds

The lower and upper interval bounds of each variable $x_k$ of a Multi-norm Zonotope $\mathbf{x}$ can be obtained by leveraging Lemma 1 to compute the bounds of $\mathbf{a}_k \cdot \hat{\phi}$:

$$-\|\mathbf{a}_k\|_q \leq \mathbf{a}_k \cdot \hat{\phi} \leq \|\mathbf{a}_k\|_q.$$

Given this, the lower and upper bounds $l_k$ and $u_k$ of $x_k$ are:

$$l_k = c_k - \|\mathbf{a}_k\|_q + \min(\|\mathbf{\beta}_k \cdot \hat{\epsilon}\|_1)$$

$$u_k = c_k + \|\mathbf{a}_k\|_q + \max(\|\mathbf{\beta}_k \cdot \hat{\epsilon}\|_1),$$

where we applied Lemma 1 on $\mathbf{\beta}_k$ for the last term.

Theorem 1. For all $k$, $l_k$ and $u_k$ are sound and tight bounds of the Multi-norm Zonotope variable $x_k$.

Proof. From the derivation above and the fact that $\mathbf{a}_k \cdot \hat{\phi}$ and $\mathbf{\beta}_k \cdot \hat{\epsilon}$ are independent follows that $l_k$ and $u_k$ are sound and tight bounds for $x_k$. □

4.2 Affine Abstract Transformer

The abstract transformer for an affine combination $z = ax_1 + bx_2 + c$ of two Multi-norm Zonotope variables $x_1 = c_1 + \tilde{a}_1 \cdot \hat{\phi} + \tilde{\beta}_1 \cdot \hat{\epsilon}$ and $x_2 = c_2 + \tilde{a}_2 \cdot \hat{\phi} + \tilde{\beta}_2 \cdot \hat{\epsilon}$, is

$$z = ax_1 + bx_2 + c$$

$$= a(c_1 + \tilde{a}_1 \cdot \hat{\phi} + \tilde{\beta}_1 \cdot \hat{\epsilon}) + b(c_2 + \tilde{a}_2 \cdot \hat{\phi} + \tilde{\beta}_2 \cdot \hat{\epsilon}) + c$$

$$= (ac_1 + bc_2 + c) + (a\tilde{a}_1 + b\tilde{a}_2) \cdot \hat{\phi} + (a\tilde{\beta}_1 + b\tilde{\beta}_2) \cdot \hat{\epsilon}.$$

Theorem 2. The affine abstract transformer for the Multi-norm Zonotope domain is exact.

Proof. Follows from the derivation above. □

4.3 ReLU Abstract Transformer

The ReLU abstract transformer defined for the classical Zonotope can be extended naturally to the multi-norm setting since it relies only on the lower and upper bounds of the variables, computed as described in Section 4.1. Concretely, the Multi-norm Zonotope abstract transformer for ReLU($x$) is $\max(0, x)$ of the zonotope variable $x$ is

$$y = \begin{cases} 0, & \text{if } u < 0 \\ x, & \text{if } l > 0 \\ \lambda x + \mu + \beta_{new} \epsilon_{new} & \text{otherwise,} \end{cases}$$

Figure 4. A Multi-norm zonotope with two variables $x = 4 + \phi_1 + \phi_2 - \epsilon_1 + 2\epsilon_2$ and $y = 3 + \phi_1 + \phi_2 + \epsilon_1 + \epsilon_2$ where $\|\hat{\phi}\|_2 \leq 1$ and $\epsilon_1, \epsilon_2 \in [-1, 1]$. The dark green region indicates the classical zonotope obtained by removing the $\hat{\phi}$ noise symbols.
where $\epsilon_{\text{new}} \in [-1, 1]$ denotes a new noise symbol, $l$ and $u$ denote the lower and upper bound of $x$ and
\[
\begin{align*}
\lambda &= u / (u - l) \\
\mu &= 0.5 \max(-\lambda l, (1 - \lambda)u) \\
\beta_{\text{new}} &= 0.5 \max(-\lambda l, (1 - \lambda)u).
\end{align*}
\]
We note that the newly introduced noise symbol $\epsilon_{\text{new}}$ is an $\ell_{\infty}$ noise symbol. This holds for all $\epsilon_{\text{new}}$ in the following.

### 4.4 Tanh Abstract Transformer

To support the tanh function present in the pooling layer (Figure 2), we extend the abstract transformer for the tanh defined in [49], to the multi-norm setting. As for ReLU, the abstract transformer is applied element-wise. The abstract transformer for the operation $y = \tanh(x)$ is
\[
y = \lambda x + \mu + \beta_{\text{new}} \epsilon_{\text{new}},
\]
where $\epsilon_{\text{new}} \in [-1, 1]$ denotes a new noise symbol and
\[
\begin{align*}
\lambda &= \min(1 - \tanh^2(l), 1 - \tanh^2(u)) \\
\mu &= \frac{1}{2} (\tanh(u) + \tanh(l) - \lambda(u + l)) \\
\beta_{\text{new}} &= \frac{1}{2} (\tanh(u) - \tanh(l) - \lambda(u - l)).
\end{align*}
\]

### 4.5 Exponential Abstract Transformer

Since the softmax function $\sigma$ requires the exponential, we rely on the framework setup in [41] to define the abstract transformer for the exponential function that minimizes the area of the output zonotope and ensures its lower bound is positive, which is required by the reciprocal abstract transformer that will be applied afterward. As described in the paper, the operation $y = e^x$ can be modeled through the element-wise abstract transformer
\[
y = \lambda x + \mu + \beta_{\text{new}} \epsilon_{\text{new}},
\]
where $\epsilon_{\text{new}} \in [-1, 1]$ denotes a new noise symbol,
\[
\begin{align*}
\lambda &= e^{\ell_{\text{opt}}} \\
\mu &= 0.5 (e^{\ell_{\text{opt}}} - \lambda l_{\text{opt}} + e^u - \lambda u) \\
\beta_{\text{new}} &= 0.5 (\lambda l_{\text{opt}} - e^{\ell_{\text{opt}}} + e^u - \lambda u)
\end{align*}
\]
and
\[
\begin{align*}
l_{\text{opt}} &= \min(t_{\text{crit}}, t_{\text{crit}, 2}) \\
t_{\text{crit}} &= \log(e^{\ell_{\text{opt}} - d} / u - l) \\
t_{\text{crit}, 2} &= l + 1 + \hat{\epsilon}.
\end{align*}
\]
Here, $\hat{\epsilon}$ is a small positive constant value, such as 0.01. The choice $l_{\text{opt}} = \min(t_{\text{crit}}, t_{\text{crit}, 2})$ ensures that $y$ is positive.

### 4.6 Reciprocal Abstract Transformer

To obtain the minimal area abstract transformer for the reciprocal function, which is required for the softmax, we instantiate the equations specified in [41] for that particular function. We note that the algorithm is applicable because the reciprocal function is convex for positive inputs and we are only required to define our transformer for positive inputs due to the softmax structure. The abstract transformer for $y = 1/x$ with $x > 0$ is given by
\[
y = \lambda x + \mu + \beta_{\text{new}} \epsilon_{\text{new}},
\]
where $\epsilon_{\text{new}} \in [-1, 1]$ denotes a new noise symbol and
\[
\begin{align*}
\lambda &= -1 / t_{\text{opt}} \\
\mu &= 0.5(1 / t_{\text{opt}} - \lambda \cdot t_{\text{opt}} + 1 / \lambda l) \\
\beta_{\text{new}} &= 0.5(\lambda \cdot t_{\text{opt}} - 1 / t_{\text{opt}} + 1 / \lambda l)
\end{align*}
\]
and
\[
\begin{align*}
l_{\text{opt}} &= \min(t_{\text{crit}}, t_{\text{crit}, 2}) \\
t_{\text{crit}} &= \sqrt{ul} \\
t_{\text{crit}, 2} &= 0.5u + \hat{\epsilon}.
\end{align*}
\]
Similarly to the exponential transformer, $\hat{\epsilon}$ is a small positive constant and $l_{\text{opt}} = \min(t_{\text{crit}}, t_{\text{crit}, 2})$ ensures that $y$ is positive.

### 4.7 Soundness and Precision of the Elementwise Abstract Transformers

**Theorem 3.** The Multi-norm Zonotope abstract transformers for ReLU, tanh, exponential and reciprocal are sound and optimal in terms of their area in input-output space.

**Proof.** As outlined in Section 4.3-4.6, all of our elementwise abstract transformers follow the functional form of their classical Zonotope counterparts. In [41, 49] the classical Zonotope transformers are applied elementwise and map a Zonotope variable $x$ to an affine combination $y = \lambda x + \mu + \beta_{\text{new}} \epsilon_{\text{new}}$, where $\epsilon_{\text{new}}$ is a new $\ell_{\infty}$ noise symbol. They prove that their choices for $\lambda, \mu$ and $\beta_{\text{new}}$ provide a sound overapproximation and are optimal in terms of the area in the input-output space. Further, $\lambda, \mu$ and $\beta_{\text{new}}$ depend only on the interval bounds $l$ and $u$ of $x$ and the underlying function. Their proofs rely only on the exactness of the affine abstract transformer and the precision and soundness of $l$ and $u$.

Our Multi-norm Zonotope transformers use the same functional form for the Multi-norm Zonotope output variable $y = \lambda x + \mu + \beta_{\text{new}} \epsilon_{\text{new}}$, where as before $\epsilon_{\text{new}}$ is a new $\ell_{\infty}$ noise symbol. Further, for our Multi-norm Zonotopes, the bounds $l$ and $u$ are also sound and precise (Theorem 1) and the affine transformer is also exact (Theorem 2). Since $\lambda, \mu$ and $\beta_{\text{new}}$ depend only on the bounds $l$ and $u$ and the underlying function, the proofs in [41, 49] transfer directly to our setting. Thus, our abstract transformers are also sound and optimal in terms of the area in input-output space. □

### 4.8 Dot Product Abstract Transformer

Next, we define the abstract transformer for the dot product between pairs of vectors of variables of a Multi-norm Zonotope. The transformer is used in the multi-head self-attention, more specifically in the matrix multiplications between $Q$ and $K$ and between the result of the softmax and $V$ (Figure
Note that the dot product transformer is required here because both operands of our matrix multiplications are represented as a part of the zonotope. This is in contrast to the matrix multiplications we have in an affine layer for example, where only one of the matrices is part of the zonotope.

The standard dot product for two vectors $\vec{v}_1, \vec{v}_2 \in \mathbb{R}^N$ is given by $y = \vec{v}_1 \cdot \vec{v}_2 = \sum_{k=1}^{N} v_{1k} v_{2k}$. Computing the dot product between two Multi-norm Zonotope vectors sharing error terms $\vec{v}_1 = \vec{c}_1 + A_1 \vec{\phi} + B_1 \vec{e}$ and $\vec{v}_2 = \vec{c}_2 + A_2 \vec{\phi} + B_2 \vec{e}$ produces the output variable $y$:

$$y = \vec{v}_1 \cdot \vec{v}_2 = (\vec{c}_1 + A_1 \vec{\phi} + B_1 \vec{e}) \cdot (\vec{c}_2 + A_2 \vec{\phi} + B_2 \vec{e})$$

$$= \vec{c}_1 \cdot \vec{c}_2 + (\vec{c}_1 + A_1 \vec{\phi} + B_1 \vec{e}) \cdot (\vec{c}_2 + A_2 \vec{\phi} + B_2 \vec{e})$$

$$+ (A_1 \vec{\phi} + B_1 \vec{e}) \cdot (A_2 \vec{\phi} + B_2 \vec{e}).$$

Next, we focus on the last term of this expression, representing interaction between noise symbols, as this term is not in the functional form of a Multi-norm Zonotope. We first expand the last term and get

$$(A_1 \vec{\phi} + B_1 \vec{e}) \cdot (A_2 \vec{\phi} + B_2 \vec{e}) = (A_1 \vec{\phi}) \cdot (A_2 \vec{\phi}) + (A_1 \vec{\phi}) \cdot (B_2 \vec{e})$$

$$+ (B_1 \vec{e}) \cdot (A_2 \vec{\phi}) + (B_1 \vec{e}) \cdot (B_2 \vec{e}).$$

Each of the 4 cases contains a different combination of noise symbols and coefficients. We calculate interval bounds for each of the combinations, e.g., $[\varepsilon, \epsilon] \times [\varepsilon, \epsilon]$ for $(A_1 \vec{\phi}) \cdot (B_2 \vec{e})$.

Then the sum of the lower and upper bounds:

$$l = l_{\phi, \phi} + l_{\phi, e} + l_{e, \phi} + l_{e, e}$$

$$u = u_{\phi, \phi} + u_{\phi, e} + u_{e, \phi} + u_{e, e},$$

binds the whole term $l \leq (A_1 \vec{\phi} + B_1 \vec{e}) \cdot (A_2 \vec{\phi} + B_2 \vec{e}) \leq u$.

Next, we show a general method to compute the interval bounds for each of the 4 cases. For the $\epsilon, \epsilon$-case, we additionally present a second argument for the computation which is slower but more precise.

**Fast Bounds $l_{\phi, \delta}, u_{\phi, \delta}$ (DeepT-Fast).** To showcase the method, we calculate interval bounds for the generic expression $(V_{\vec{\xi}_p}^T) \cdot (W_{\vec{\xi}_p})$, where $V$ and $W$ are matrices such that $V_{\vec{\xi}_p}$ and $W_{\vec{\xi}_p}$ have the same dimension and $\|V_{\vec{\xi}_p}\|_p \leq 1$ and $\|W_{\vec{\xi}_p}\|_p \leq 1$. We do this by computing an upper bound for the absolute value

$$|(V_{\vec{\xi}_p}^T) \cdot (W_{\vec{\xi}_p})| = |\xi_{p_i}^T V_{\vec{\xi}_p} \cdot W_{\vec{\xi}_p}| \leq \|\xi_{p_i}^T V\| \cdot \|W\|_{\vec{\xi}_p},$$

where $| \cdot |$ denotes either the absolute value for scalars or the elementwise absolute value for vectors or matrices. Here, we used the triangle inequality in the last step.

We use Lemma 1 to bind the elements $|\vec{w}_j \cdot \vec{\xi}_{\vec{p}}|$ of the vector $|W_{\vec{\xi}_p}|$, where $\vec{w}_j$ denotes the $j^{th}$ row of $V$ and $\vec{\xi}_{\vec{p}}$ is the dual norm of $\ell^p$. We obtain

$$|(V_{\vec{\xi}_p}^T) \cdot (W_{\vec{\xi}_p})| \leq \xi_{p_i}^T V^T |W_{\vec{\xi}_p}| \cdot \|\vec{w}_j\|_{\vec{\xi}_p}.$$
We note that DeepT-Fast and DeepT-Precise can be combined: DeepT-Fast can be used to compute the bounds for the mixed terms \((l_{ϕ, φ}, u_{ϕ, φ})\), \((l_{ϕ, e}, u_{ϕ, e})\) and \((l_{e, φ}, u_{e, φ})\), while DeepT-Precise can be used to compute \((l_{e, e}, u_{e, e})\). We leave empirical evaluation of this variant for future work.

4.9 Multiplication Abstract Transformer
The multiplication abstract transformer is a special case of the dot-product abstract transformer where the vectors have only one element.

5 Practical Robustness Certification
While our Multi-norm Zonotope transformers described so far could be directly implemented and applied in a straightforward manner, this would not scale to larger networks, and the accumulated overapproximation would reduce overall verification precision. In this section we introduce several techniques that improve the scalability and precision of our abstract transformers, enhancing their practical effectiveness: In Section 5.1, we improve performance and lower memory usage by periodically reducing the number of noise symbols, in Section 5.2, we reduce the loss of precision that occurs during softmax by using an alternative but equivalent softmax definition whose abstract transformer is more precise, and in Section 5.3 we improve precision by introducing the constraint that softmax output values need to sum to 1, ensuring the output variables form a normalized distribution. Finally, in Section 5.4, we discuss the differences between DeepT and the related work.

5.1 Noise Symbol Reduction
Through the repeated application of abstract transformers during verification, the number of \(ℓ^∞\) noise symbol grows, leading to slower verification and higher memory usage. In fact, every abstract transformer we use, except the one for affine transformations, can yield new noise symbols \(ε_{\text{new}}\). To address this, we periodically reduce the number of \(ℓ^∞\) noise symbols to ensure an upper bound on the memory usage independent of the network depth, thus creating a tunable tradeoff between precision and speed.

DecorrelateMin \[ 38 \]. A heuristic method that reduces the number of \(ℓ^∞\) noise symbols in a classical Zonotope to \(k\) is DecorrelateMin\(_k\). We follow and apply that method to our Multi-norm Zonotope setting by calculating a heuristic score \(m_j = \sum_{i=1}^{N} |B_{i,j}|\) for every \(ℓ^∞\) noise symbol \(ε_j\) and keeping the top \(k \ell^∞\) noise symbols.

Let \(I\) denote the indices of the eliminated \(ℓ^∞\) noise symbols and \(P\) the indices of the top \(k \ell^∞\) noise symbols. Then, the new Multi-norm Zonotope is:

\[
\bar{x} = \bar{c} + A\bar{ϕ} + B_P \bar{e}_P + \begin{pmatrix} ε_{\text{new},1} \sum_{i \in I} |B_{i,1}| \\ \vdots \\ ε_{\text{new},N} \sum_{i \in I} |B_{i,N}| \end{pmatrix}, \quad ε_{\text{new}} \in [-1, 1]^N.
\]

The Transformer network has residual connections in each Transformer layer, each combining two different branches of the neural network. Noise symbol reduction is applied to the input embeddings of every Transformer layer, just before the residual connection around the multi-head self-attention, avoiding the complexities of handling separate noise reductions in two branches.

5.2 Softmax Abstract Transformer
The softmax function \(σ\) is a key operation in the self-attention (Figure 3). While the transformers for exponential, sum, reciprocal and multiplication can be used to construct a transformer for the softmax directly, we rewrite the softmax to:

\[
σ(v_1, \ldots, v_N) = \frac{e^{v_1}}{\sum_{j=1}^{N} e^{v_j}} = \frac{1}{\sum_{j=1}^{N} e^{v_j-v_1}}.
\]

It is well known that the latter formula is a more numerically stable way of computing the softmax function. We find it is also beneficial for defining our abstract transformers: First, the noise symbols in \(v_1\) can cancel with the ones in \(v_j\), which reduces the overapproximation introduced through the abstract transformer for the exponential function (it gets worse the larger the inputs are). Second, we do not need the multiplication abstract transformer. Further, this method ensures that the output is always between 0 and 1, as the denominator is always positive and greater than 1.

5.3 Softmax Sum Zonotope Refinement
By construction, the outputs \(y_1, \ldots, y_N\) of the softmax function \(σ\) when applied to inputs \(x_1, \ldots, x_N\) satisfy \(\sum_{i=1}^{N} y_i = 1\), i.e. they form a probability distribution. Thus, in the multi-head self-attention, the role of the softmax is to pick some convex combination of the values \(V\), according to the similarity between the query and the keys.

However, this property is not always satisfied for the Multi-norm Zonotope obtained for \(Z\) produced by the softmax abstract transformer (Eq. 1). By abuse of notation, we call this Zonotope \(Z\). There are many valid instantiations of the noise symbols such that the Zonotope variables do not sum to 1, causing non-convex combinations of values to be picked. To address this, we enforce the constraint that the variables must sum to 1, to enforce that a convex combination is selected and to preserve the semantics of the network in our abstract domain. This is achieved by excluding from the Multi-norm Zonotope \(Z\) all invalid instantiations of values, obtaining a refined Multi-norm Zonotope \(Z’\) with lower volume, that helps to increase verification precision.

We leverage Zonotope constraint methods \[ 17 \], which produce refined Zonotopes given some equality constraints. A three step process is used to refine all Zonotope variables \(y_1, \ldots, y_N\) by:

1. computing a refined variable \(y_1’\) by imposing the equality constraint \(y_1 = 1 - (y_2 + \cdots + y_N)\),
2. refining all other variables \(y_2, \ldots, y_N\) to \(y_2’, \ldots, y_N’\).
3. tightening the bounds of the $\epsilon_i$’s to a subset of $[-1, 1]$. Note that we arbitrarily select $y_1$ as the variable to be refined first, but any other variable could have been chosen.

We now detail these three steps that lead to a refined Zonotope $Z'$ with variables $y'_1, \ldots, y'_n$ that always sum to 1 and have tighter error bounds.

**Step 1. Refining $y_1$.** We illustrate the process of obtaining a refined Zonotope with variable $z'_1$, given the equality constraint $z_1 = z_2$ for a Zonotope with two variables $z_1$ and $z_2$. The final result can then be obtained by instantiating $z_2 = 1 - (y_2 + \ldots + y_N)$ and $z_1 = y_1$ and finally $y'_1 = z'_1$.

While we know that $z_1 = z_2$ needs to hold, not all instantiations of the noise symbols satisfy this constraint. We can compute a new Multi-norm Zonotope variable $z'_1 = c' + \tilde{a}' \cdot \tilde{\phi} + \tilde{b}' \cdot \tilde{\epsilon}$, such that for all instantiations of noise symbols of $z'_1$, we have $z_1 = z_2$ and $z_1 = z'_1$, thereby enforcing the equality constraints $z_1 = z_2$.

We have

$$z_1 := c_1 + \tilde{a}_1 \cdot \tilde{\phi} + \tilde{b}_1 \cdot \tilde{\epsilon} = c_2 + \tilde{a}_2 \cdot \tilde{\phi} + \tilde{b}_2 \cdot \tilde{\epsilon} =: z_2.$$

If we solve for $\epsilon_k$ (any $k$ s.t. $\beta_k - \tilde{\beta}_k \neq 0$ works) in the equation above and substitute it in the equation $z_2 = z'_1$, we obtain the following constraints for the coefficients of $z'_1$,

$$c' = c_2 + (c_1 - c_2) \frac{\beta_k - \tilde{\beta}_k}{\tilde{\beta}_k - \beta_k}$$  \hspace{1cm} (7)

$$\tilde{a}' = \tilde{a}_2 + (\tilde{a}_1 - \tilde{a}_2) \frac{\beta_k - \tilde{\beta}_k}{\tilde{\beta}_k - \beta_k}$$  \hspace{1cm} (8)

$$\tilde{b}' = \tilde{b}_2 + (\tilde{b}_1 - \tilde{b}_2) \frac{\beta_k - \tilde{\beta}_k}{\tilde{\beta}_k - \beta_k}$$  \hspace{1cm} (9)

where $I$ are the indices of the other $\epsilon$ terms (i.e. without $\epsilon_k$).

**Choosing a Value for $\beta_k$.** In the equations above, we have one degree of freedom, namely $\beta_k$. Any value $v$ for $\beta_k$ is valid and leads to a valid affine expression $z'_1$, with the other coefficients of $z'_1$ being deduced through the equations above.

To select $v$, we opt to minimize the absolute value of the noise symbol coefficients, which acts as a heuristic for the tightness of the zonotope variable:

$$v^* = \min_v S = \min_v \left[ ||z'||_1 + ||\tilde{b}'||_1 \right].$$

The minimization problem above can be efficiently solved with $O((E_p+E_{\infty}) \log(E_p+E_{\infty}))$ complexity, using a method from [17] described in Appendix A.1. We note that to maintain precision, we disallow solutions that lead to the elimination of one of the $p$-norm noise symbols $\phi$.

**Step 2. Refining $y_2, \ldots, y_n$.** We substitute the expression for $\epsilon_k$ computed in Step 1 in the affine expressions of the variables $y_2, \ldots, y_N$, to obtain the refined Multi-norm Zonotope variables $y'_2, \ldots, y'_N$.

**Step 3. Tightening the Bounds of $\tilde{\epsilon}$.** The refined sum constraint $S = 1 - \sum_{i} y'_i = c_S + \tilde{a}_S \cdot \tilde{\phi} + \tilde{b}_S \cdot \tilde{\epsilon} = 0$ can be further leveraged to tighten the bounds of the $p$-norm noise symbols $\tilde{\epsilon}$, with non zero coefficient, by solving for $\epsilon_m$:

$$\epsilon_m = \frac{1}{\tilde{b}_S} \left( c_S + \tilde{a}_S \cdot \tilde{\phi} + \tilde{b}_S \cdot \tilde{\epsilon} \right),$$

which implies that the range of $\epsilon_m$ is restricted to $[a_m, b_m] \cap [-1, 1]$ where

$$a_m = \frac{1}{||\tilde{b}_S||_1} \left( c_S - ||\tilde{a}_S||_p - ||\tilde{b}_S||_1 \right),$$

$$b_m = \frac{1}{||\tilde{b}_S||_1} \left( c_S + ||\tilde{a}_S||_p + ||\tilde{b}_S||_1 \right).$$

Note that because the noise symbol reduction process assumes all noise symbols $\tilde{\epsilon}$ have range $[-1, 1]$, prior to it a pre-processing step occurs where all noise symbols $\epsilon_m$ with tightened bounds $[a_m, b_m] \subset [-1, 1]$ are re-written as $\epsilon_m = \frac{a_m + b_m}{2} + \frac{b_m - a_m}{2} \epsilon_{new,m}$ with $\epsilon_{new,m} \in [-1, 1]$. It has good precision, back substitution through all multiplications is memory and computationally intensive. CROWN-BaF [47] has a lower memory footprint and faster certification times by not performing full back substitution but stopping early.

However, CROWN-BaF loses a significant amount of precision for multiplication due to the partial back substitution. As our Multi-norm Zonotope domain requires only a forward pass, but still preserves important relational information between the variables throughout the network, we introduce less overapproximation in deeper layers of a network. While the CROWN / DeepPoly domain used in CROWN-Backward is more precise but also more expensive compared to the Zonotope domain used in DeepT, we find that additional precision at the cost of runtime can be gained by introducing the DeepT-Precise variant of the dot product transformer.

Further, by periodically reducing the number of noise symbols, introducing only a small loss of precision we can increase efficiency for DeepT-Fast and DeepT-Precise in a more principled way compared to back substitution early stopping as we can explicitly tune the speed/memory vs precision trade-off per layer.

For softmax, CROWN-Backward and CROWN-BaF overapproximate $\frac{e^s}{\sum e^s}$ by concatenating the transformers for the exponentials, sum, reciprocal and finally multiplication in that order. Contrary to that, DeepT-Fast and DeepT-Precise overapproximate the mathematically equivalent but more favorable formula $\frac{1}{\sum e^s}$, as elaborated in Section 5.2. Further, in Section 5.3 we introduce constraints to guarantee...
that the overapproximative softmax outputs constitute a convex combination, which allows DeepT to gain additional precision.

6 Experimental Evaluation

In this section, we evaluate the performance and precision of our methods on a range of networks. Our goal is to prove local robustness of Transformer networks around given input sequences. We consider two threat models:

- **T1**: Just one embedded word can be perturbed by an $\ell^p$-norm bound perturbation.
- **T2**: Every word in the input sequence can be replaced by one of its synonyms independently.

The first threat model T1 is used in [47] and is therefore useful as a direct comparison. The comparison shows that our methods improve certification rates and scalability to deeper and wider Transformer networks in general.

The second threat model T2 is used to show that our methods yield certification results in the space of natural language processing that can not be obtained using enumeration within a reasonable time frame. We capture these synonym attacks by constructing an $\ell^\infty$ box around the synonym embeddings with the original word embedding, which is then pushed through the network to check if certification succeeds. T2 is particularly strong if many words in the sequence have multiple synonyms as this results in a large number of possible combinations.

We focus on the threat models T1 and in Section 6.7 we focus on T2. We do not handle other threat models like adding stop words, insertions or deletions of random words.

In Appendix A.2, we investigate the performance of our Multi-norm Zonotope against $\ell^2$ norm bound attacks on a small fully-connected network trained on MNIST [28] and compare it with [24]. Further, in Appendix A.3, we evaluate the performance of our certification methods on Transformer networks classifying MNIST images. An ablation study for the softmax refinement is given in Appendix A.5 and results for an approach combining DeepT-Fast and DeepT-Precise are given in Appendix A.6.

6.1 Experimental Setup

We evaluate the certification ability, performance and scalability of our approaches by verifying Transformer networks of different sizes against perturbations of norm $p \in \{1, 2, \infty\}$. We evaluate two variants of our verifier:

- **DeepT-Fast**, which uses the faster but more approximate dot-product abstract transformer, and
- **DeepT-Precise**, which uses the more precise but less efficient dot-product abstract transformer for $p = \infty$.

We compare our method against the CROWN approach [47], the state-of-the-art method for Transformer network verification. It relies on linear relaxations and applies back-substitution repeatedly to obtain linear bounds expressed in terms of the input values. We compare against the Backward-Forward (BaF) and Backward variants of the CROWN verifier, the latter being more precise but considerably slower. The certification radii are computed using a binary search to find the highest radius $\epsilon$ such that the region representing a $\ell^p$ perturbation of radius $\epsilon$ around the input can be verified.

**Transformer Networks.** A Transformer network gets a sentence as input and predicts the binary sentiment polarity of the sentence. Unless otherwise noted, the networks have $M \in \{3, 6, 12\}$ layers, an embedding size of 128 and 4 heads. The hidden layer size of the feed forward networks is 128. In Section 6.2, we provide additional experiments with wider networks. In Section 6.3 and Section 6.4 we use an embedding size of 64 and hidden layer size of 64 due to memory constraints of the baseline. Further, in Section 6.6 we provide experiments using the standard layer normalization, confirming the observation of [47] that division by the standard deviation harms certification rates.

**Dataset.** Unless otherwise noted, we train the transformers from scratch on the SST dataset [51] which contains 67349, 872 and 1821 examples in the training, development and test set, respectively. Additionally experiments using the larger Yelp dataset [61] are provided in Section 6.2.

**Hardware.** All our experiments ran on an Nvidia 2080 Ti GPU with 11GB of memory, except for the experiments presented in Table 4 and Table 5, which ran on an Nvidia Titan X GPU with 24GB.

**Noise Symbol Reduction.** DeepT-Fast reduces the number of $\ell^\infty$ noise symbols at every Transformer network layer to 14000, while DeepT-Precise reduces the number of $\ell^\infty$ noise symbols to 10000 (Section 5.1). However, for sentences with 22 words or more, these numbers are halved to lower memory usage and improve performance. These numbers were chosen empirically in order to balance precision, speed and memory use.

<table>
<thead>
<tr>
<th>$M$</th>
<th>$\ell^p$</th>
<th>DeepT-Fast</th>
<th>CROWN-BaF</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Min</td>
<td>Avg</td>
</tr>
<tr>
<td>3</td>
<td>$\ell_1$</td>
<td>0.036</td>
<td>1.808</td>
</tr>
<tr>
<td>6</td>
<td>$\ell_2$</td>
<td>6.4e-3</td>
<td>0.330</td>
</tr>
<tr>
<td>12</td>
<td>$\ell_\infty$</td>
<td>2.1e-3</td>
<td>0.328</td>
</tr>
<tr>
<td>12</td>
<td>$\ell_1$</td>
<td>0.089</td>
<td>1.191</td>
</tr>
<tr>
<td>6</td>
<td>$\ell_2$</td>
<td>0.015</td>
<td>0.212</td>
</tr>
<tr>
<td>12</td>
<td>$\ell_\infty$</td>
<td>1.2e-3</td>
<td>0.021</td>
</tr>
<tr>
<td>3</td>
<td>$\ell_1$</td>
<td>0.358</td>
<td>0.512</td>
</tr>
<tr>
<td>6</td>
<td>$\ell_2$</td>
<td>0.074</td>
<td>0.107</td>
</tr>
<tr>
<td>12</td>
<td>$\ell_\infty$</td>
<td>7.3e-3</td>
<td>0.011</td>
</tr>
</tbody>
</table>
Table 2. Certified radius (min and avg), time (in seconds) and ratio of the average certified radii on the Yelp dataset.

<table>
<thead>
<tr>
<th>M</th>
<th>t_p</th>
<th>DeepT-Fast</th>
<th>CROWN-BaF</th>
<th>Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Min</td>
<td>Avg</td>
<td></td>
</tr>
<tr>
<td>f_1</td>
<td>3</td>
<td>0.070</td>
<td>0.415</td>
<td>21.5</td>
</tr>
<tr>
<td>CROWN-BaF</td>
<td>0.102</td>
<td>0.573</td>
<td>3.9</td>
<td>0.7</td>
</tr>
<tr>
<td>f_2</td>
<td>6</td>
<td>0.026</td>
<td>0.091</td>
<td>47.9</td>
</tr>
<tr>
<td>CROWN-BaF</td>
<td>0.023</td>
<td>0.137</td>
<td>3.9</td>
<td>1.4</td>
</tr>
<tr>
<td>f_3</td>
<td>12</td>
<td>5.7e-3</td>
<td>2.9e-3</td>
<td>99.1</td>
</tr>
<tr>
<td>CROWN-BaF</td>
<td>2.4e-3</td>
<td>0.155</td>
<td>43.3</td>
<td>1.6</td>
</tr>
</tbody>
</table>

6.2 Fast Verifiers: DeepT-Fast vs CROWN-BaF

First, we compare our tool for fast verification DeepT-Fast against CROWN-BaF, the tool for fast verification in [47]. The metrics used include the size of the certified radius, the scaling behavior and the performance. To the best of our knowledge, this is the first time verification is applied on more realistically-sized Transformer networks, with 6 and 12 layers [9]; the previous maximum was 3 layers [47]. The accuracies for the 3, 6 and 12 layer networks are 83.3%, 83.6% and 83.4% respectively on the SST dataset (Table 1) and 91.4%, 91.5% and 91.5% respectively on the Yelp dataset (Table 2).

As proposed in [47], we evaluate 10 correctly classified random test examples with sentence lengths not exceeding 32. Given a sentence with embeddings E, for every position i we compute the maximum robustness radius $\epsilon$ around the embedding of the $i$th word such that all embeddings in it are correctly classified.

Certification Results. DeepT-Fast achieves larger certified radii than CROWN-BaF in almost every setting (Table 1, Table 2). For the 6-deep Transformer network, both verifiers are fast but DeepT-Fast can certify much larger perturbation regions: on average, the certified radii are 28× larger on SST and 250× larger on Yelp compared to those of CROWN-BaF. Thus, DeepT-Fast scales much better than CROWN-BaF: while the average certified radius decreases for DeepT-Fast by a factor of 3.5× on SST and 7.5× on Yelp as the number of layers increases from 3 to 12, for CROWN-BaF they decrease by a factor of 90× on SST and 1553× on Yelp on average.

Performance. While CROWN-BaF is faster by a factor of 2-5 for the smallest Transformer network, the size of the performance gap shrinks to a factor of 1.25 on SST and 2 on Yelp for the 12-layer network because DeepT-Fast’s computational complexity scales linearly with the number of layers while for CROWN-BaF it increases superlinearly.

Overall, these results indicate that DeepT-Fast is the suitable choice when the goal is to verify deeper and more practically sized Transformer networks.

Table 3. Certified radius (min and avg), time (in seconds) and ratio of the average certified radii for Transformer networks with embedding size of 256 and hidden size of 512. Execution failures are indicated by “-”.

<table>
<thead>
<tr>
<th>M</th>
<th>t_p</th>
<th>DeepT-Fast</th>
<th>CROWN-BaF</th>
<th>Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Min</td>
<td>Avg</td>
<td></td>
</tr>
<tr>
<td>f_1</td>
<td>3</td>
<td>0.288</td>
<td>0.700</td>
<td>37.7</td>
</tr>
<tr>
<td>CROWN-BaF</td>
<td>0.644</td>
<td>0.969</td>
<td>11.1</td>
<td>1.1</td>
</tr>
<tr>
<td>f_2</td>
<td>6</td>
<td>0.055</td>
<td>0.204</td>
<td>37.9</td>
</tr>
<tr>
<td>CROWN-BaF</td>
<td>0.114</td>
<td>0.101</td>
<td>11.6</td>
<td>1.1</td>
</tr>
<tr>
<td>f_3</td>
<td>12</td>
<td>3.7e-3</td>
<td>0.015</td>
<td>35.5</td>
</tr>
<tr>
<td>CROWN-BaF</td>
<td>8.4e-3</td>
<td>0.014</td>
<td>11.5</td>
<td>1.1</td>
</tr>
</tbody>
</table>

Large Width Transformer Network Results. To evaluate the performance on wider Transformer networks, we double the embedding size to 256 and quadruple the hidden layer size to 512. We evaluated on the SST dataset, the accuracies are 83.2%, 83.2% and 83.0% for the 3, 6 and 12 layer network respectively. The results are given in Table 3.

Due to memory issues, CROWN-BaF failed to verify anything for the wider 12 layer Transformer network. DeepT-Fast, on the other hand, succeeded due to its tunable noise symbol reduction, allowing for fine-grained tradeoffs between precision and memory usage. The results are in agreement with the previous observations.

6.3 The Precision-Performance Tradeoff

To compare the two variants of the dot product abstract transformer (Section 4.8), we evaluate the certification of DeepT-Precise and DeepT-Fast against $\ell^\infty$ attacks. Additionally, we compare DeepT-Precise with CROWN-Backward since they both operate in the regime of more precise but slower verification (Table 4). Because of the reduced speed, we only perturb a randomly chosen position per example sentence instead of evaluating perturbations for all positions. The accuracies for the 3, 6 and 12 layer networks we evaluate on are 83.5%, 83.7% and 83.4% respectively.

Memory Usage of CROWN-Backward. Since CROWN-Backward requires for larger sentences more memory than the 2080 Ti GPU can provide, even in the case of a moderately sized 3 layer Transformer network, we verify a smaller Transformer network where the embedding size is 64 and the hidden size 64. We run all verifiers on a 24 GB Nvidia Titan X GPU to provide a fair timing comparison since the memory requirements of CROWN-Backward for the downscaled 12 layer Transformer network still exceed the 2080 Ti.

Results. In the trade-off between powerful certification and performance, DeepT-Fast achieves the fastest results, DeepT-Precise achieves the highest average precision while CROWN-Backward is in an intermediate position, albeit it...
has the minimum best certified radius. For the 12-layer Transformer network, DeepT-Fast achieves 10% lower certification compared to CROWN-Backward but it is 12x faster. Compared to DeepT-Precise, it achieves 30% lower certification but it is 39x faster. DeepT-Precise achieves better precision than DeepT-Fast despite more noise symbols being eliminated during the noise symbol reduction steps, highlighting the importance of the dot-product abstract transformer.

### 6.4 Comparison for $\ell^1$ and $\ell^2$ Perturbations

We also compare DeepT-Fast with CROWN-Backward for $\ell^p$ where $p \in \{1, 2\}$, as well as with CROWN-BaF (Table 5).

**Comparison with CROWN-Backward.** DeepT-Fast is 3.7x, 6.0x and 11.2x faster than CROWN-Backward for the 3, 6 and 12 layer Transformer networks respectively while having lower memory usage, with only a 7% to 9% reduction in certified radius.

**Comparison with CROWN-BaF.** DeepT-Fast’s average certification radii are larger in all settings, especially for the 12 layer Transformer network where the radii are 3.3x larger, but certification is 2x slower. For the 6 layer Transformer network, the smallest radius computed by CROWN-BaF matches the one computed by DeepT-Fast, but DeepT-Fast provides larger radii for the 3 and 12 layer Transformer networks.

**Scalability of DeepT-Fast and CROWN-BaF.** As in Section 6.3, we rely on small networks due to the comparison against CROWN-Backward. Likely due to the reduced network size, CROWN-BaF verification loses precision less quickly compared to the first set of experiments, degrading by a factor of 5.4 (instead of 90) between the 3 layer network and the 12 layer network. This still indicates that DeepT-Fast scales better compared to CROWN-BaF in width and depth.

### 6.5 Dual Norm Order

Since the final bound obtained by the DeepT-Fast dot product transformer is not symmetric in $W$ and $V$ (Section 4.8), we evaluated both orders (Table 6) which showed that on average it is slightly advantageous to apply the dual norm first on the $\ell^\infty$-norm terms and then on the $\ell^p$-norm terms.

### 6.6 Standard Layer Normalization Results

The certification experiments presented in this work evaluate our methods on Transformer networks without the division by the standard deviation in the layer normalization, as described in [47]. They report improved certification rates without a significant impact on performance. Our experiments (Table 7) confirm the observation of [47]. The accuracies of the Transformer networks using division are 81.9%, 82.2% and 81.8% respectively for the 3, 6 and 12 layer Transformer networks. The accuracies are slightly lower than the accuracies of the Transformer networks without division by the standard deviation, but the latter has almost always the larger certification radii.
Table 8. Comparison between CROWN-BaF and DeepT-Fast for robustness certification against synonym attacks on 137 sentences. The network accuracy is 81.6%.

<table>
<thead>
<tr>
<th></th>
<th>Certified Sentences</th>
<th>Certified Percentage</th>
<th>Time (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CROWN-BaF</td>
<td>121</td>
<td>89%</td>
<td>2.60</td>
</tr>
<tr>
<td>DeepT-fast</td>
<td>120</td>
<td>88%</td>
<td>2.41</td>
</tr>
</tbody>
</table>

6.7 Certification against Synonym Attacks

We also evaluate our method against threat model T2, where every word in the input sequence can be replaced by one of its synonyms independently, making it significantly more difficult to certify compared to threat model T1. Note, that the threat model T2 covers also a threat model where only a fixed number of words can be replaced by synonyms as T2 is strictly stronger. We focus on sentences of up to 26 words where the number of combinations of synonyms is at least 32,000, though many considered sentences have significantly more combinations, because they pose a more difficult verification challenge as they are the ones for which enumeration-based methods are considerably slower than our certification methods.

When comparing the certified robustness rates of DeepT-Fast against CROWN-BaF (Table 8), we see that the performance is comparable on a 3 layer Transformer network trained for certifiability using the method provided in [57]. The accuracy is 81.7%. We note that some verified sentences have millions of combinations and testing all combinations would take at least 2 orders of magnitude more time. DeepT-Fast is competitive with CROWN-BaF, even though the training method is tailored to improve the verification success rate of the CROWN method and the small size of the network. The verifiers could not be evaluated on larger networks because the certified training method [57] failed for larger networks. We conjecture that DeepT-Fast would achieve superior verification compared to CROWN-BaF in the large-network setting. Table 9 details one of these certifiable sentences with 23 million possible combinations.

6.8 Summary

At a high level, our results indicate that if one is interested in pushing certification for deeper Transformer networks (as typically used in practice), then DeepT-Fast would be the preferred choice as it scales well while being significantly more precise than current state-of-the-art. If one is interested in the maximal attainable precision, then DeepT-Precise would be preferred as it is currently the most precise Transformer network verifier.

7 Related Work

Next, we discuss related work in both neural network certification and Transformer networks.

Table 9. Example certifiable sentence with 23 million possible combinations, along with the synonyms generated by the attack in [1] for every position. Enumeration-based verification would take 2 to 3 orders of magnitude more time compared to our verification method.

<table>
<thead>
<tr>
<th>Tokens</th>
<th>#Synonyms</th>
<th>Synonyms</th>
</tr>
</thead>
<tbody>
<tr>
<td>no</td>
<td>3</td>
<td>not, without</td>
</tr>
<tr>
<td>reason</td>
<td>1</td>
<td>reasons</td>
</tr>
<tr>
<td>for</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>anyone</td>
<td>6</td>
<td>somebody, anyone, anybody,</td>
</tr>
<tr>
<td>to</td>
<td>0</td>
<td>everyone, person, nobody</td>
</tr>
<tr>
<td>invest</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>their</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>hard-earned</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>bucks</td>
<td>1</td>
<td>money</td>
</tr>
<tr>
<td>into</td>
<td>5</td>
<td>at, towards, toward, in, for</td>
</tr>
<tr>
<td>a</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>movie</td>
<td>3</td>
<td>film, films, cinema</td>
</tr>
<tr>
<td>which</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>obviously</td>
<td>5</td>
<td>clearly, naturally, apparently, plainly, definitely</td>
</tr>
<tr>
<td>did</td>
<td>4</td>
<td>did, could, got, do, does</td>
</tr>
<tr>
<td>n’t</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>invest</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>much</td>
<td>5</td>
<td>very, many, highly, greatly, heavily</td>
</tr>
<tr>
<td>into</td>
<td>6</td>
<td>at, under, towards, in, for</td>
</tr>
<tr>
<td>itself</td>
<td>6</td>
<td>himself, themselves, ourselves,</td>
</tr>
<tr>
<td>either</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>.</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

Neural Network Certification. A wide variety of certification methods guaranteeing the local robustness of neural networks around a given input have been proposed in recent years. The certification methods can be grouped into two categories: the complete, but less scalable methods and incomplete but more scalable methods.

Among the complete ones, some are relying on SMT-solvers [6, 12, 26] or mixed-integer linear programming [53]. Incomplete methods rely on a wide range of techniques including abstract interpretation [15, 48–50], linear relaxation [46, 55, 60] and duality [11]. While these methods provide deterministic guarantees, a recent development in the space of incomplete verification is randomized smoothing [8, 29, 30, 45], which provides probabilistic certificates. Due to their scalability advantages, as they are able to handle larger networks, we focus in this paper on incomplete verification methods with standard deterministic guarantees.

While most work regarding neural network certification is about \(\ell^p\) robustness, neural networks can also be certified against semantic perturbation beyond \(\ell^p\) robustness, e.g., certification against geometric perturbations on images like rotations or translations [3, 14, 39, 42] or vector fields [44].

Transformer Networks. Transformer networks [54] are neural networks designed to process sets or sequences, such
as sentences, sound or images. The Transformer network architecture and the attention mechanisms it introduces have recently led to major advances in natural language processing [9, 27, 58]. Currently, this type of network and its refinements are the de facto standard for most natural language processing tasks, similar to convolutional networks for image processing tasks. Further, transformer networks have recently been successfully used in tasks even beyond natural language processing, including object detection [7], graph classification [32] and speech recognition [20], and have also been combined with existing architectures such as convolutional neural networks to achieve state-of-the-art results in visual question answering [33] or image synthesis [59].

Adversarial Attacks on Transformer Networks. One of the early adversarial attacks in the space of NLP was presented by [1], where the attacker model allowed to exchange each word in the input sequence by one of 8 nearest neighbors in the embedding space. To reduce the number of false synonyms (words whose embeddings are close but are not synonyms in the context of the sentence), a simple language model was used and empirically validated. Later, [22] successfully attacked Transformer networks by replacing one word in the input sequence with a synonym, and observed that transformer networks tend to be more robust than LSTMs. A gradient-ascent based attack was presented in [5] which uses a FGSM-like attack on the embeddings, after which each word in the sequence gets replaced by the word having the embedding closest to the attacked embedding for that word. [23, 31, 34] present attacks tailored to perturb few words, with the key idea being to only consider perturbing words that are important for classification.

Transformer Network Certification. While neural network certification has been successfully applied to feed-forward and convolutional networks classifying images, the certification of Transformer networks has been considered only very recently. A first inclination might be to certify these networks against synonym-based attacks by enumerating all possibilities. This however works only for short sequences of words where additionally each has only a few synonyms. As discussed already, [47] certifies robustness against $\ell^p$-norm bound changes in the embedding space for $p \in \{1, 2, \infty\}$. The idea here is to certify against changes in the embedding space, as this should cover synonyms if the certified radius is large enough. Further, [57] certifies against synonym attacks for small 1-layer transformer networks, where a small subset of words can be replaced by synonyms. Here, the attacker model is similar to [1], with the restriction that just a few words can be changed. Recently, [62] used the certified training method presented in [37] to train Transformer network models to be robust against NLP attacks performing deletions, insertions and reordering of words or letters.

While training a network to be empirically more robust is orthogonal to the robustness certification task addressed in this paper, we evaluated whether our Multi-norm Zonotope could be substituted in the training procedure presented in [62]. Our experiments showed that direct substitution is not feasible due to large memory requirements of [37] when combined with the Multi-norm Zonotope domain. However, we believe this issue could be solved by applying the COLT training method [4], which has more favorable memory characteristics. We leave this for future work.

8 Conclusion

We introduced the Multi-norm Zonotope domain, a generalization of the classical Zonotope designed to model $\ell^p$ perturbations by adding additional noise symbols bound by a $\ell^p$-norm constraint. We carefully construct a set of abstract transformers for all operations of the Transformer network in a manner that balances scalability and precision, including the challenging softmax and dot product functions (the critical operations in self-attention) which require handling non-linearities applied to multiple terms under perturbation.

Using our new Multi-norm Zonotope domain and its abstract transformers, we built DeepT, the first verifier able to certify more realistically-sized Transformer networks with high precision and fast performance. We evaluated the effectiveness, performance and scalability of DeepT for certifying Transformer networks for binary classification against $\ell^p$ attacks. Our results show that DeepT achieves fast and effective certification of Transformer networks, even for deep 12-layer networks. Our method scales well, both in terms of the number of layers and in terms of the size of the input embeddings, contrarily to state-of-the-art which either becomes impractically slow or whose certification strength reduces dramatically for deeper networks.

We also demonstrate, for the first time, successful certification for Transformer networks against synonym attacks for long word sequences having many synonyms, where each word can potentially be replaced by many substitutes. Our DeepT verifier achieves an 84% success rate when verifying a certifiably-pretrained Transformer network for sentences where enumeration-based methods would take 2 to 3 orders of magnitude more time.

References

Fast and Precise Certification of Transformers

PLDI ’21, June 20–25, 2021, Virtual, Canada


PLDI ’21, June 20–25, 2021, Virtual, Canada


A Appendix

A.1 Solving the Minimization Problem

We describe the method from [17] to solve the minimization problem below with \(O((\mathcal{E}_p + \mathcal{E}_\infty) \log(\mathcal{E}_p + \mathcal{E}_\infty)) \) complexity:

\[
\nu^* = \min_n S = \min_n \left[ \frac{\mathcal{E}_p}{n} \sum_{i=1}^n |\alpha_i^p| + \sum_{j=1}^n |\beta_i^p| \right].
\]

[17] relies on two observations. First, since all coefficients in the minimization can be written in form \( r + s\beta_k^p \) with \( r, s \in \mathbb{R} \) (see Eqs. 5-6), the expression to be minimized is of the form \( S = \sum_i |r_i + s_i \beta_k^p| \). The optimal value \( \nu^* \) for \( \beta_k^p \) will cause one of the \(|r_i + s_i \beta_k^p|\) terms to be 0 and therefore \( \nu^* \) must equal \( \frac{\alpha_i}{n} \) for some \( i \in [1, \mathcal{E}_p + \mathcal{E}_\infty] \). The values \( \frac{\alpha_i}{n} \) are the candidate solutions for the minimization problem and simply testing them all would lead to an algorithm with complexity \( O((\mathcal{E}_p + \mathcal{E}_\infty)^2) \).

To improve on this, a second observation is used: each term \(|r_i + s_i \beta_k^p|\) of \( S \) has a constant negative slope before \( \frac{\alpha_i}{n} \) and a constant positive slope after it. Therefore, as \( \beta_k^p \) increases the slope of more and more \(|r_i + s_i \beta_k^p|\) terms becomes positive, showing that the slope of \( S \) increases monotonically with \( \beta_k^p \). The minimum value of \( S \) will happen at the value of \( \beta_k^p \) where the slope of \( S \) changes from negative to positive. Since the slope of \( S \) increases monotonically, we can run a binary search on \( \beta_k^p \) to efficiently find the value at which the slope of \( S \) changes sign, obtaining an algorithm with \( O((\mathcal{E}_p + \mathcal{E}_\infty) \log(\mathcal{E}_p + \mathcal{E}_\infty)) \) complexity. To maintain precision, we exclude the candidate solutions that lead to the elimination of one of the \( \ell^p \)-norm noise symbols \( \phi \). This requires an additional linear search but doesn’t change the overall complexity of the algorithm.

A.2 Comparison with GeoCert for \( \ell_p \) Perturbations

While this paper focuses on using Multi-norm Zonotopes for Transformer networks verification, the domain is general and can be applied to other architectures. We illustrate this by certifying a small fully-connected DNN for MNIST [28] and compared it to GeoCert method [24] on \( \ell_p \) perturbations. The DNN performs binary classification on MNIST [28] images with digits 1 or 7, and is composed of a feed-forward network with ReLU nonlinearities and hidden sizes of 10, 50 and 10.

Without any adaptation of our method, we evaluate certification using Multi-norm Zonotopes and compare it to GeoCert (Table 10). While the radii are significantly smaller, the certification is significantly faster. We believe the results indicate a promising future work item.

A.3 Certifying Visual Transformers for Image Classification

While originally developed for NLP, Transformer networks have been successfully applied in computer vision tasks, such as image classification [10] and object detection [7, 35], achieving competitive [7, 10] or state-of-the-art performance [35].

To evaluate the certification capabilities of DeepT in domains other than NLP, we focus on robustness certification of image classification models. More specifically, our goal will be to prove that a Vision Transformer network [10] will correctly classify images even if they have suffered a \( \ell^p \) perturbation, a common threat model in computer vision.

Model and Dataset. The Vision Transformer network has 1 encoding layer, an embedding size of 64 and the size of the hidden layer of the feed-forward network is 128. The self-attention layer has 4 attention heads. The input image is split into patches of size 7x7, each being mapped through a linear layer to an embedding of size 64 after which a positional encoding is added. We evaluate on 100 randomly selected images of the MNIST dataset [28]. The network achieves 95.54% test accuracy.

Results. We report the size of certified radius and the performance of DeepT when certifying against \( \ell^1, \ell^2 \) and \( \ell^\infty \) perturbations in Table 11.

A.4 Complete Results for the Precision-Performance Tradeoff

Table 12 contains the full results for the experiment of Section 6.3 which additionally contains the results for the CROWN-BaF verifier. DeepT-Fast achieves larger average certified radii than CROWN-BaF in all settings, with the certification gap increasing substantially for deeper networks, but is 1.5-4.4 times slower, with the performance gap reducing for deeper networks.

A.5 Evaluating the Softmax Sum Refinement

We perform an ablation study to evaluate the effect of the Zonotope refinement using the softmax sum constraint. We
Table 12. Certified radius (min and avg) and time (in seconds) for DeepT-Fast, CROWN-BaF, DeepT-Precise and CROWN-Backward for $\ell_\infty$ perturbations.

<table>
<thead>
<tr>
<th>$M$</th>
<th>DeepT-Fast</th>
<th></th>
<th>CROWN-BaF</th>
<th></th>
<th>DeepT-Precise</th>
<th></th>
<th>CROWN-Backward</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Min</td>
<td>Avg</td>
<td>Time</td>
<td>Min</td>
<td>Avg</td>
<td>Time</td>
<td>Min</td>
<td>Avg</td>
</tr>
<tr>
<td>3</td>
<td>0.013</td>
<td>0.034</td>
<td>17.2</td>
<td>0.013</td>
<td>0.033</td>
<td><strong>3.9</strong></td>
<td>0.013</td>
<td><strong>0.038</strong></td>
</tr>
<tr>
<td>6</td>
<td>0.014</td>
<td>0.031</td>
<td>33.8</td>
<td>0.014</td>
<td>0.025</td>
<td><strong>10.6</strong></td>
<td>0.014</td>
<td><strong>0.036</strong></td>
</tr>
<tr>
<td>12</td>
<td>9.3e-3</td>
<td>0.021</td>
<td>69.0</td>
<td>1.9e-3</td>
<td>6.3e-3</td>
<td><strong>44.9</strong></td>
<td>8.8e-3</td>
<td><strong>0.024</strong></td>
</tr>
</tbody>
</table>

Table 13. Refining the Zonotope by using the softmax sum constraint in the DeepT-Fast verifier leads to higher certified radius. The results are rounded, but the improvement is computed using the full precision floating point numbers. Experiments ran on a 1080 Ti with 11GB of RAM.

| $M$ | $\ell_p$ | With Constraint | | Without Constraint | | Change | |
|---|---|---|---|---|---|---|
|   | $t_1$ | 0.037 | **1.808** | 21.8 | 0.037 | 1.807 | **+0.04%** |
| 3  | $t_2$ | 6.6e-3 | 0.330 | 22.1 | 6.37e-3 | 0.330 | **+0.22%** |
|   | $c_\infty$ | 2.18e-3 | 0.032 | 21.7 | 6.07e-4 | 0.032 | **+0.49%** |
| 6  | $t_1$ | 0.095 | **1.211** | 52.4 | 0.089 | 1.200 | **+0.94%** |
|   | $t_2$ | 0.015 | **0.215** | 52.7 | 0.015 | 0.214 | **+0.88%** |
|   | $c_\infty$ | 2.03e-3 | **0.021** | 50.6 | 1.36e-3 | 0.021 | **+1.50%** |
| 12 | $t_1$ | 0.358 | **0.514** | 105.1 | 0.360 | 0.501 | **+2.62%** |
|   | $t_2$ | 0.074 | **0.107** | 109.0 | 0.075 | 0.105 | **+2.80%** |
|   | $c_\infty$ | 7.29e-3 | **0.011** | 102.5 | 7.31e-3 | 0.010 | **+3.21%** |

Table 14. Certified radius (min and avg) and time (in seconds) for the DeepT verifier that combines DeepT-Precise and DeepT-Fast, and CROWN-Backward for $\ell_\infty$ perturbations.

<table>
<thead>
<tr>
<th>$M$</th>
<th>Combined DeepT verifier</th>
<th></th>
<th>CROWN-Backward</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Min</td>
<td>Avg</td>
<td>Time</td>
<td>Min</td>
</tr>
<tr>
<td>6</td>
<td>0.014</td>
<td>0.034</td>
<td><strong>227.0</strong></td>
<td><strong>0.015</strong></td>
</tr>
<tr>
<td>12</td>
<td>9.1e-3</td>
<td>0.023</td>
<td><strong>423.7</strong></td>
<td><strong>9.9e-3</strong></td>
</tr>
</tbody>
</table>

compare the DeepT-Fast verifier with and without the refinement. We observe that the refinement leads to the certification radius improvements that increases with the depth on the neural network (Table 13). The improvement is small (0.04%-0.49%) for the 3 layer Transformer networks but increases to 2.62%-3.21% for the 12 layer Transformer networks, at a 5-9% performance cost.

### A.6 Combining DeepT-Fast and DeepT-Precise

By using the more precise abstract transformer for the dot product only in the last layer of the Transformer network and the faster version in the remaining layers, we obtain a verifier that is both faster and more precise on average than CROWN-Backward for the 6 and 12 layer Transformer networks (Table 14). For these networks, reduction to 5000 and 7000 noise symbols is applied in the last layer, and reduction to 14000 and 20000 noise symbols in the remaining layers.