Fast and Precise Transformer Certification

Gregory Bonaert, Dimitar I. Dimitrov, Maximilian Baader, Martin Vechev
Department of Computer Science
ETH Zürich, Switzerland
Adversarial Examples for Images

$x$

“panda”
57.7% confidence

$+ 0.007 \times$

sign($\nabla_x J(\theta, x, y)$)

“nematode”
8.2% confidence

$= x + \epsilon$sign($\nabla_x J(\theta, x, y)$)

“gibbon”
99.3% confidence

Adversarial Examples for NLP

✅ Original: “Perfect performance by the actor”

❌ Adversarial: “Spotless performance by the actor”

Input sentence

Classification

Positive

Negative

Certification pipeline

Fire Bonfire Blaze →编码器层→编码器层

is →

cold chilly frigid →编码器层→编码器层

Transformer network layers →negative

Input Embeddings Abstraction Output
Threat models

1. Embeddings ($\ell^p$ ball)

- Represented as a Multi-Norm Zonotope

Goal: certify robustness against embedding attacks

2. Synonyms

<table>
<thead>
<tr>
<th>Perfect</th>
<th>performance</th>
<th>by the actor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spotless</td>
<td>acting</td>
<td>comedian</td>
</tr>
<tr>
<td>Impeccable</td>
<td></td>
<td>performer</td>
</tr>
</tbody>
</table>

Goal: certify robustness against synonym attacks
Transformer networks: architecture & embeddings
Transformer networks: encoder layer

Single attention head

Encoder Layer

\[ Z = \text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \]

\[ Q = XW_Q \]
\[ K = XW_K \]
\[ V = XW_V \]
Transformer networks: encoder layer

Multiple attention heads

Encoder Layer

Multi-Head Self-Attention

\[
Z = \left( \sigma \left( \frac{XW_Q^1 W_K^1 X^T}{\sqrt{d_k}} \right) XW_V^1 \right) W_0
\]

where \( X = \left( \begin{array}{c} x_1^T \\ \vdots \\ x_n^T \end{array} \right) \) and \( Z = \left( \begin{array}{c} z_1^T \\ \vdots \\ z_n^T \end{array} \right) \)

where \( \sigma = \text{softmax} \)

Add + Normalize

FFN

Add + Normalize

FFN

\( x_1 \rightarrow x_1 \)

\( \vdots \)

\( x_n \rightarrow x_n \)

\( z_1 \rightarrow y_1 \)

\( \vdots \)

\( z_n \rightarrow y_n \)
Challenges of Transformer network verification

1. Dot Products

Challenges:
- Both terms under perturbation (first, $Q$ and $K$, second softmax(·) and $V$)
- Quadratic number of dot products

$$Attention(Q, K, V) = \text{softmax} \left( \frac{QK^T}{\sqrt{d_k}} \right) V$$

Goal: Create a fast and precise dot product abstract transformer
Challenges of Transformer network verification

2. Softmax

Challenges:

- Exponential and Division abstract transformers cause great precision loss
- Concrete output represents a probability distribution, but this information is lost during abstraction

Goal: Create a fast and precise softmax abstract transformer

\[
\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V
\]

\[
\text{softmax}_i(\nu_1, \ldots, \nu_N) = \frac{e^{\nu_i}}{\sum_{j=1}^N e^{\nu_j}}
\]
Classical Zonotopes

Linear layer with weights \[
\begin{bmatrix}
1 & -2 \\
1 & 2 \\
\end{bmatrix}
\]

\[
\begin{align*}
x &= -\epsilon_1 + 2\epsilon_2 \\
y &= \epsilon_1 + \epsilon_2
\end{align*}
\]
\(\epsilon_1, \epsilon_2 \in [-1, 1]\)

\[
\begin{align*}
x &= -5\epsilon_1 + 3\epsilon_2 \\
y &= -\epsilon_1 + 3\epsilon_2
\end{align*}
\]
\(\epsilon_1, \epsilon_2 \in [-1, 1]\)
Classical Zonotopes

ReLU

Tanh

Reciprocal

Exponential
\( \ell^p \) ball representation in Classical Zonotopes

**Challenge:**
- Classical zonotopes precisely encode \( \ell_\infty \) balls but over-approximate other \( \ell_p \) balls, because they have \( \ell_\infty \) bounded terms

\[ x_k = c_k + \sum_{j=1}^{E_\infty} \beta_k^j \epsilon_j = c_k + \bar{\beta}_k \cdot \bar{\epsilon} \]

\( c_k, \beta_k^j \in \mathbb{R}, \quad \| \epsilon_j \|_\infty \leq 1 \)

**Goal:** Create a zonotope that precisely captures an \( \ell_p \) ball
Multi-Norm Zonotopes

**Idea:** Multi-norm Zonotopes, which encode the $\ell^p$ balls using new noise symbols under an $\ell^p$ constraint

One variable

$$x_k = c_k + \sum_{i=1}^{E_p} \alpha^i_k \phi_i + \sum_{j=1}^{E_\infty} \beta^j_k \epsilon_j = c_k + \alpha_k \cdot \vec{\phi} + \beta_k \cdot \vec{\epsilon}$$

$c_k, \alpha^i_k, \beta^j_k \in \mathbb{R}$, $\|\vec{\phi}\|_p \leq 1$, $\|\epsilon\|_\infty \leq 1$

Matrix form

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

$\vec{c} \in \mathbb{R}^N$, $A \in \mathbb{R}^{N \times E_p}$, $B \in \mathbb{R}^{N \times E_\infty}$

$\|\vec{\phi}\|_p \leq 1$, $\|\epsilon\|_\infty \leq 1$
Abstract Transformer - Dot product

\[ \vec{v}_1 = (\vec{c}_1 + A_1 \phi + B_1 \vec{e}) \]
\[ \vec{v}_2 = (\vec{c}_2 + A_2 \phi + B_2 \vec{e}) \]

Developing the equations

\[ \vec{v}_1 \cdot \vec{v}_2 = (\vec{c}_1 + A_1 \phi + B_1 \vec{e}) \cdot (\vec{c}_2 + A_2 \phi + B_2 \vec{e}) \]
\[ = \vec{c}_1 \cdot \vec{c}_2 + (\vec{c}_1^\top A_2 + \vec{c}_2^\top A_1) \phi + (\vec{c}_1^\top B_2 + \vec{c}_2^\top B_1) \vec{e} + (A_1 \phi + B_1 \vec{e}) \cdot (A_2 \phi + B_2 \vec{e}) \]

Multi-Norm Zonotope Form

Goal: find a Multi-norm Zonotope representation for the last term
Abstract Transformer - Dot product

Challenge:
- Putting bounds on the interaction between noise symbols

\[(A_1 \vec{\phi} + B_1 \vec{\epsilon}) \cdot (A_2 \vec{\phi} + B_2 \vec{\epsilon}) = (A_1 \vec{\phi}) \cdot (A_2 \vec{\phi}) + (A_1 \vec{\phi}) \cdot (B_2 \vec{\epsilon}) + (B_1 \vec{\epsilon}) \cdot (A_2 \vec{\phi}) + (B_1 \vec{\epsilon}) \cdot (B_2 \vec{\epsilon})\]
Idea: Use dual norm to concretize one term, then again to concretize the 2nd term

\[
(A_1 \vec{\phi} + B_1 \vec{\epsilon}) \cdot (A_2 \vec{\phi} + B_2 \vec{\epsilon}) = (A_1 \vec{\phi}) \cdot (A_2 \vec{\phi}) + (A_1 \vec{\phi}) \cdot (B_2 \vec{\epsilon}) + (B_1 \vec{\epsilon}) \cdot (A_2 \vec{\phi}) + (B_1 \vec{\epsilon}) \cdot (B_2 \vec{\epsilon})
\]
Abstract Transformer - Dot product (DeepT-Fast)

Idea: Use dual norm to concretize one term, then again to concretize the 2nd term

\[
(A_1 \vec{\phi}) \cdot (B_2 \vec{\epsilon})
\]

\[
A_1 \vec{\phi} = \begin{bmatrix}
2\phi_1 + 3\phi_2 \\
4\phi_1 - 5\phi_2
\end{bmatrix}
\]

\[
B_2 \vec{\epsilon} = \begin{bmatrix}
2\epsilon_1 - 3\epsilon_2 + 4\epsilon_3 \\
4\epsilon_1 + 5\epsilon_2 + 5\epsilon_3
\end{bmatrix}
\]
Abstract Transformer - Dot product (DeepT-Fast)

Idea: Use dual norm to concretize one term, then again to concretize the 2nd term

\[
\begin{bmatrix}
2\phi_1 + 3\phi_2 \\
4\phi_1 - 5\phi_2 \\
\end{bmatrix}
\cdot
\begin{bmatrix}
2\epsilon_1 - 3\epsilon_2 + 4\epsilon_3 \\
4\epsilon_1 + 5\epsilon_2 + 5\epsilon_3 \\
\end{bmatrix}
\leq
\begin{bmatrix}
2\phi_1 + 3\phi_2 \\
4\phi_1 - 5\phi_2 \\
\end{bmatrix}
\cdot
\begin{bmatrix}
2\epsilon_1 - 3\epsilon_2 + 4\epsilon_3 \\
4\epsilon_1 + 5\epsilon_2 + 5\epsilon_3 \\
\end{bmatrix}
\leq
\begin{bmatrix}
2\phi_1 + 3\phi_2 & 4\phi_1 - 5\phi_2 \\
\end{bmatrix}
\begin{bmatrix}
|2\epsilon_1 - 3\epsilon_2 + 4\epsilon_3| \\
|4\epsilon_1 + 5\epsilon_2 + 5\epsilon_3| \\
\end{bmatrix}
\leq
\begin{bmatrix}
2\phi_1 + 3\phi_2 & 4\phi_1 - 5\phi_2 \\
\end{bmatrix}
\begin{bmatrix}
2 & -3 & 4 \\
4 & 5 & 5 \\
\end{bmatrix}
\leq
\begin{bmatrix}
2\phi_1 + 3\phi_2 & 4\phi_1 - 5\phi_2 \\
\end{bmatrix}
\begin{bmatrix}
9 \\
14 \\
\end{bmatrix}
\]
Abstract Transformer - Dot product (DeepT-Fast)

\[
\begin{bmatrix}
2\phi_1 + 3\phi_2 \\
4\phi_1 - 5\phi_2
\end{bmatrix}
\begin{bmatrix}
9 \\
14
\end{bmatrix}
= \begin{bmatrix}
9 \\
14
\end{bmatrix}
\begin{bmatrix}
2\phi_1 + 3\phi_2 \\
4\phi_1 - 5\phi_2
\end{bmatrix}
\leq \begin{bmatrix}
2 \\
4
\end{bmatrix}
\begin{bmatrix}
3 \\
-5
\end{bmatrix}
\begin{bmatrix}
\phi_1 \\
\phi_2
\end{bmatrix}
= \begin{bmatrix}
60 \\
106
\end{bmatrix}
\begin{bmatrix}
\phi_1 \\
\phi_2
\end{bmatrix}
= \begin{bmatrix}
60\phi_1 + 106\phi_2
\end{bmatrix}
\leq \|60\phi_1 + 106\phi_2\|_q
= 106
Abstract Transformer - Dot product (DeepT-Fast)

Idea: Use dual norm to concretize one term, then again to concretize the 2nd term

\[ |(V\hat{\xi}_{p_1}) \cdot (W\hat{\xi}_{p_2})| \leq \left\| \begin{pmatrix} \|\hat{w}_1\|_{q_2} \\ \vdots \\ \|\hat{w}_N\|_{q_2} \end{pmatrix} \right\|_T \|V\|_{q_1} \]

\[ p_1, p_2 \in \{1, 2, \infty\} \]

Q: Which of the 2 terms should be concretized first in practice?
A: The order was chosen empirically.
Abstract Transformer - Dot product (DeepT-Precise)

Challenge:
- Putting bounds on the interaction between noise symbols

\[(A_1\vec{\phi} + B_1\vec{\epsilon}) \cdot (A_2\vec{\phi} + B_2\vec{\epsilon}) = (A_1\vec{\phi}) \cdot (A_2\vec{\phi}) + (A_1\vec{\phi}) \cdot (B_2\vec{\epsilon}) + (B_1\vec{\epsilon}) \cdot (A_2\vec{\phi}) + (B_1\vec{\epsilon}) \cdot (B_2\vec{\epsilon})\]
Abstract Transformer - Dot product (DeepT-Precise)

Idea: use standard interval analysis to bound the (\( l_{\infty} \times l_{\infty} \)) dot product

\[
\begin{bmatrix}
2\varepsilon_1 - 3\varepsilon_2 \\
-1\varepsilon_1 + \varepsilon_2
\end{bmatrix}
\begin{bmatrix}
3\varepsilon_1 - 4\varepsilon_2 \\
1\varepsilon_1 + 2\varepsilon_2
\end{bmatrix}
= (2\varepsilon_1 - 3\varepsilon_2)(3\varepsilon_1 - 4\varepsilon_2) + (-1\varepsilon_1 + \varepsilon_2)(1\varepsilon_1 + 2\varepsilon_2)
= 5\varepsilon_1^2 - 10\varepsilon_1\varepsilon_2 - 8\varepsilon_2\varepsilon_1 + 14\varepsilon_2^2
= 5\varepsilon_1^2 - 18\varepsilon_1\varepsilon_2 + 14\varepsilon_2^2
\in 5[0, 1] - 18[-1, 1] + 14[0, 1]
\in [0, 5] + [-18, 18] + [0, 14]
\in [-18, 37]

Abstract Transformer - Dot product (DeepT-Precise)

Idea: use standard interval analysis to bound the $\ell_\infty \times \ell_\infty$ dot product

$$(V \vec{e}) \cdot (W \vec{e}) \in \sum_{i=1}^{\mathcal{E}_\infty} (\vec{v}_i \cdot \vec{w}_i) [0, 1] + \sum_{i \neq j}^{\mathcal{E}_\infty} (\vec{v}_i \cdot \vec{w}_j) [-1, 1]$$

Abstract Transformer - Softmax

**Challenge:**
- Exponential and Division abstract transformers cause great precision loss

**Improvement: softmax re-formulation**

**Advantages:**
1. Noise symbol cancellation
2. No multiplication (only reciprocal)
3. Output always in \([0, 1]\)

\[
\text{softmax}_i(v_1, \ldots, v_N) = \frac{e^{v_i}}{\sum_{j=1}^{N} e^{v_j}} = \frac{1}{\sum_{j=1}^{N} e^{v_j-v_i}}
\]
Abstract Transformer - Softmax

Challenges

- Concrete output represents a probability distribution, but this information is lost during abstraction.

Improvement 2: enforcing softmax properties

- Enforce output zonotope variables to be positive (by construction of our abstract transformers)
- Enforce output zonotope variables to sum to 1 (based on previous work on linear constraints on zonotopes [1])

Task: verifying Transformer Networks performing binary classification
Baseline: CROWN-BaF and CROWN-Backward [1]
Dataset: SST/Yelp sentiment polarity datasets (output = positive/negative sentiment)

Example: “Offers a breath of the fresh air of true sophistication.” → Positive sentiment

Challenge: considerably bigger networks than previous work
- Deeper networks (up to 12 layers, previous maximum was 3 layers)
- Large embedding sizes (up to 256 dimensions)
- Large hidden size of feed-forward-networks in the encoder layer (up to 512 dimensions)

Evaluation - Embedding attack ($\ell_\infty$)

<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>
</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>
</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>
</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>
</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>
</tr>
</tbody>
</table>


Evaluation - Embedding attack ($\ell_\infty$)

<table>
<thead>
<tr>
<th>$M$</th>
<th>DeepT-Fast</th>
<th>CROWN-BaF</th>
<th>DeepT-Precise</th>
<th>CROWN-Backward</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Min</td>
<td>Avg</td>
<td>Time</td>
<td>Min</td>
</tr>
<tr>
<td>3</td>
<td>0.013</td>
<td>0.034</td>
<td>17.2</td>
<td>0.013</td>
</tr>
<tr>
<td>6</td>
<td>0.014</td>
<td>0.031</td>
<td>33.8</td>
<td>0.014</td>
</tr>
<tr>
<td>12</td>
<td>9.3e-3</td>
<td>0.021</td>
<td>69.0</td>
<td>1.9e-3</td>
</tr>
</tbody>
</table>
## Evaluation - Embedding attack (all norms)

<table>
<thead>
<tr>
<th></th>
<th>$M$</th>
<th>$\ell_p$</th>
<th>DeepT-Fast</th>
<th>CROWN-BaF</th>
<th>Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Min</td>
<td>Avg</td>
<td>Time</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>$\ell_1$</td>
<td>0.036</td>
<td>1.808</td>
<td>28.8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\ell_2$</td>
<td>6.4e-3</td>
<td>0.330</td>
<td>29.2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\ell_\infty$</td>
<td>2.1e-3</td>
<td>0.032</td>
<td>26.2</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>$\ell_1$</td>
<td>0.089</td>
<td>1.191</td>
<td>63.6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\ell_2$</td>
<td>0.015</td>
<td>0.212</td>
<td>64.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\ell_\infty$</td>
<td>1.2e-3</td>
<td>0.021</td>
<td>57.1</td>
</tr>
<tr>
<td></td>
<td>12</td>
<td>$\ell_1$</td>
<td>0.358</td>
<td>0.512</td>
<td>125.9</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\ell_2$</td>
<td>0.074</td>
<td>0.107</td>
<td>129.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\ell_\infty$</td>
<td>7.3e-3</td>
<td>0.011</td>
<td>113.4</td>
</tr>
</tbody>
</table>
Embedding attack - combining the verifiers

<table>
<thead>
<tr>
<th>$M$</th>
<th>Combined DeepT verifier</th>
<th>CROWN-Backward</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Min</td>
<td>Avg</td>
</tr>
<tr>
<td>6</td>
<td>0.014</td>
<td>0.034</td>
</tr>
<tr>
<td>12</td>
<td>9.1e-3</td>
<td>0.023</td>
</tr>
</tbody>
</table>
Synonym attack example:

<table>
<thead>
<tr>
<th>Tokens</th>
<th>#Synonyms</th>
<th>Synonyms</th>
</tr>
</thead>
<tbody>
<tr>
<td>No</td>
<td>3</td>
<td>no, 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, someone, anybody, everyone, person, nobody</td>
</tr>
<tr>
<td>to</td>
<td>0</td>
<td></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, myself, yourself, herself</td>
</tr>
<tr>
<td>either</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>.</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

~23 million possible combinations
DeepT has similar performance compared to CROWN, which is expected given that the Transformer Network:

- was certifiability pretrained for CROWN [1]
- is shallow (3 layers)

### Evaluation - Synonym attacks

<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>

**Future work:** Improve scalability of certifiable training methods for Transformer Networks.

### Comparing DeepT and CROWN

<table>
<thead>
<tr>
<th></th>
<th>DeepT</th>
<th>CROWN [1]</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Precision</strong></td>
<td>Abstract transformers tailored for the attention (softmax, dot product)</td>
<td>Abstract transformers composed of the transformers for *, /, exp.</td>
</tr>
<tr>
<td><strong>Speed/memory vs Precision Trade-off</strong></td>
<td>Precisely tunable</td>
<td>Coarsely tunable</td>
</tr>
<tr>
<td><strong>Computational Cost w.r.t. depth</strong></td>
<td>$O(d)$</td>
<td>$O(d^2)$</td>
</tr>
<tr>
<td><strong>Memory Usage w.r.t. depth</strong></td>
<td>$O(1)$</td>
<td>$O(d)$</td>
</tr>
</tbody>
</table>

Summary

- Introduced the **Multi-norm Zonotope domain** alongside its abstract transformers.
- Constructed precise and fast **dot-product** and **softmax** abstract transformers for Transformer networks.
- Implemented a verifier called **DeepT** that scales certification to significantly **deeper** Transformer networks.
- Developed the **first** robustness certifier for **synonym** attacks on Transformer networks.
Thank you!