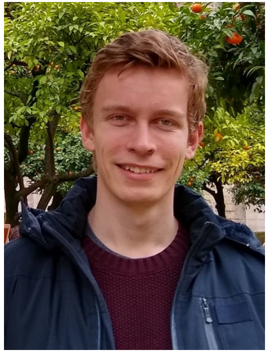
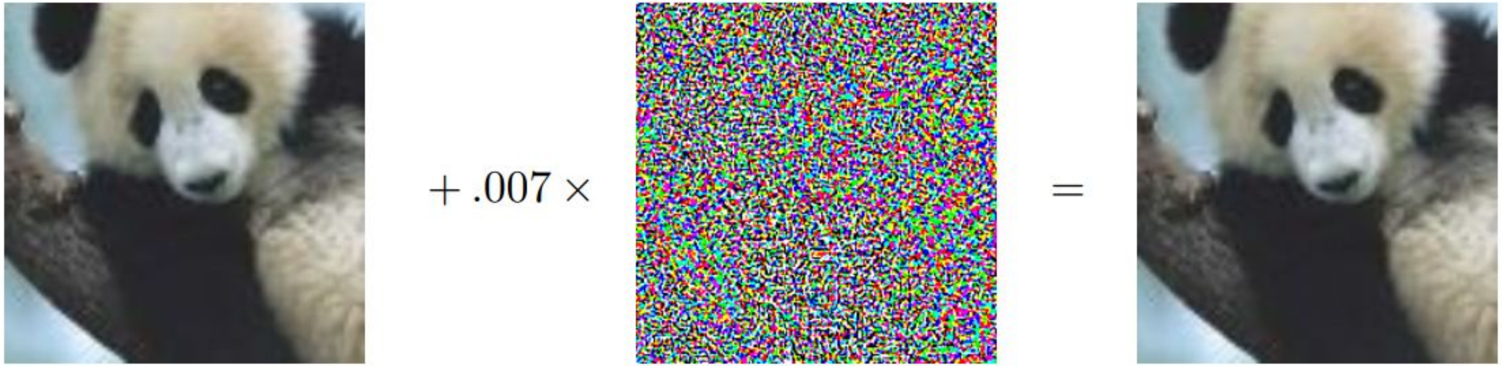
Fast and Precise Transformer Certification

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

Adversarial Examples for Images



$$\boldsymbol{x} + .007 \times \operatorname{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y)) = \boldsymbol{x} + \epsilon \operatorname{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$$

$\boldsymbol{x}$
"panda"
57.7% confidence

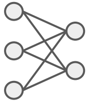
$\operatorname{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$
"nematode"
8.2% confidence

$\boldsymbol{x} + \epsilon \operatorname{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$
"gibbon"
99.3 % confidence

Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples." arXiv preprint arXiv:1412.6572 (2014).

Adversarial Examples for NLP

✔ **Original:** “**Perfect** performance by the actor” → **Positive**



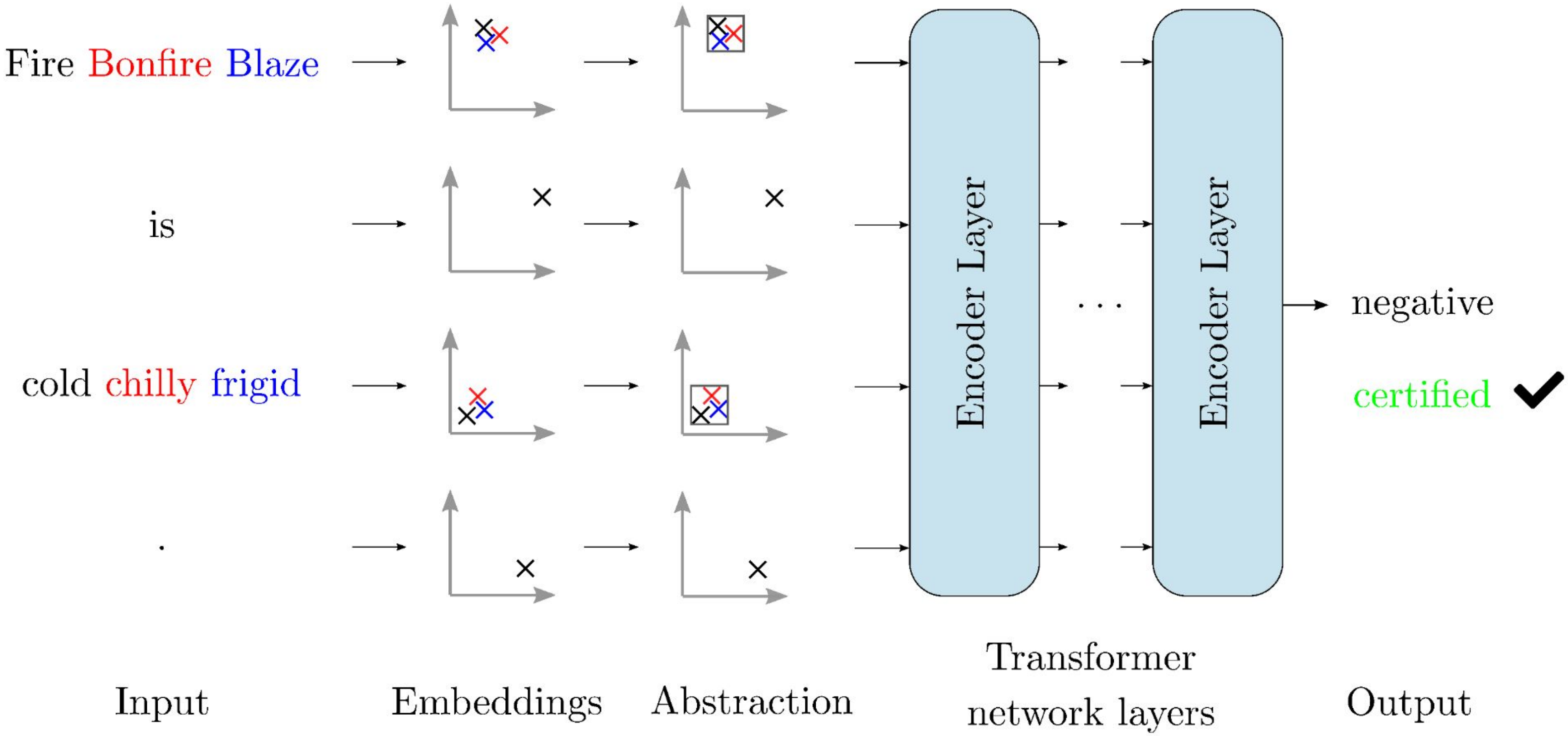
✘ **Adversarial:** “Spotless performance by the actor” → Negative

Input sentence

Classification

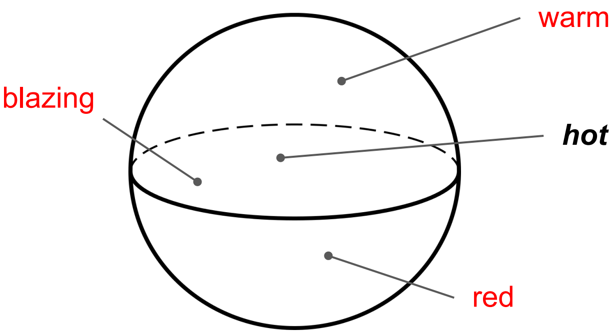
Morris, John, et al. "TextAttack: A Framework for Adversarial Attacks, Data Augmentation, and Adversarial Training in NLP." *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. 2020.

Certification pipeline



Threat models

$1.$ Embeddings (ℓ^p ball)



Represented as a Multi-Norm Zonotope

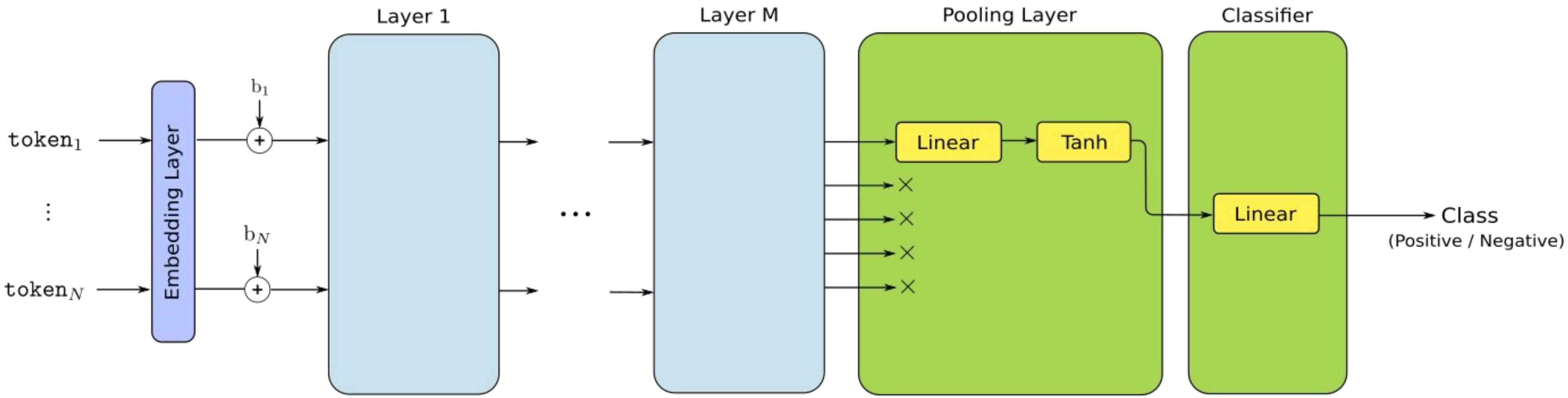
Goal: certify robustness against embedding attacks

$2.$ Synonyms

Perfect	**performance**	**by the**	**actor**
Spotless	*acting*		*comedian*
Impeccable			*performer*

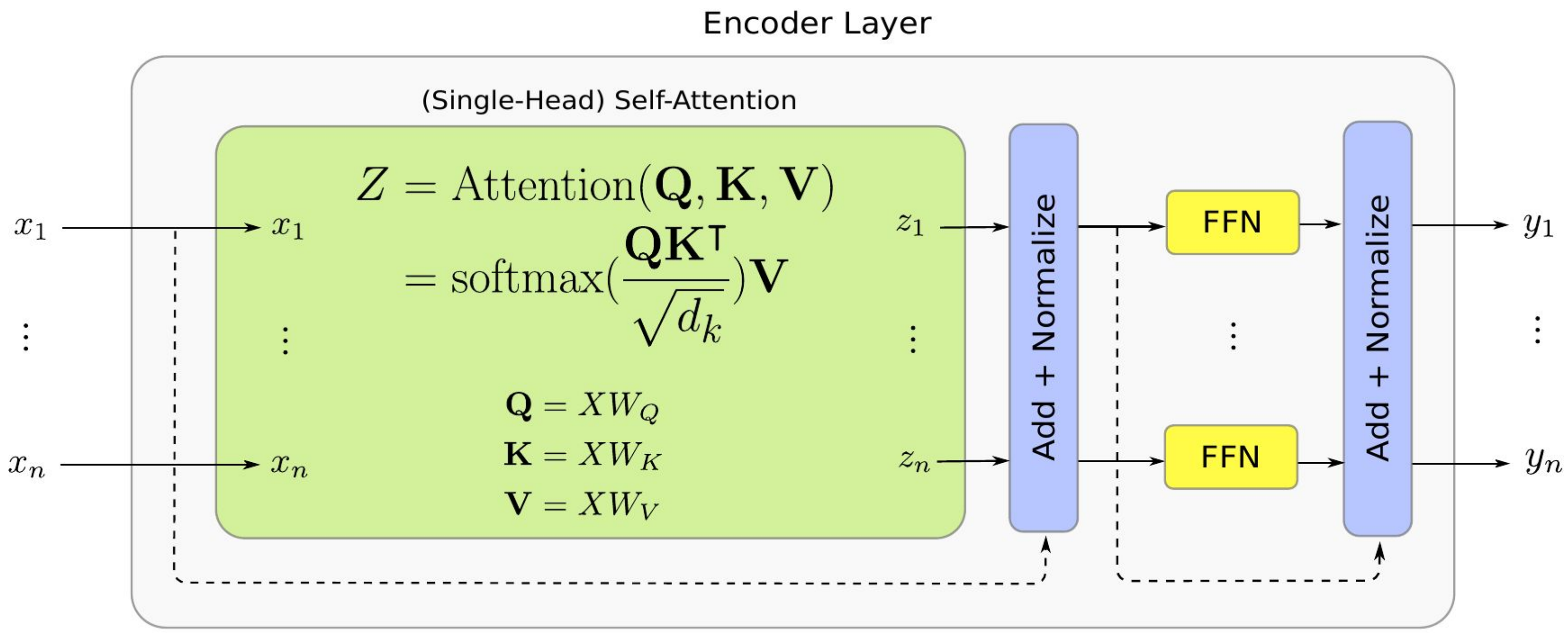
Goal: certify robustness against synonym attacks

Transformer networks: architecture & embeddings



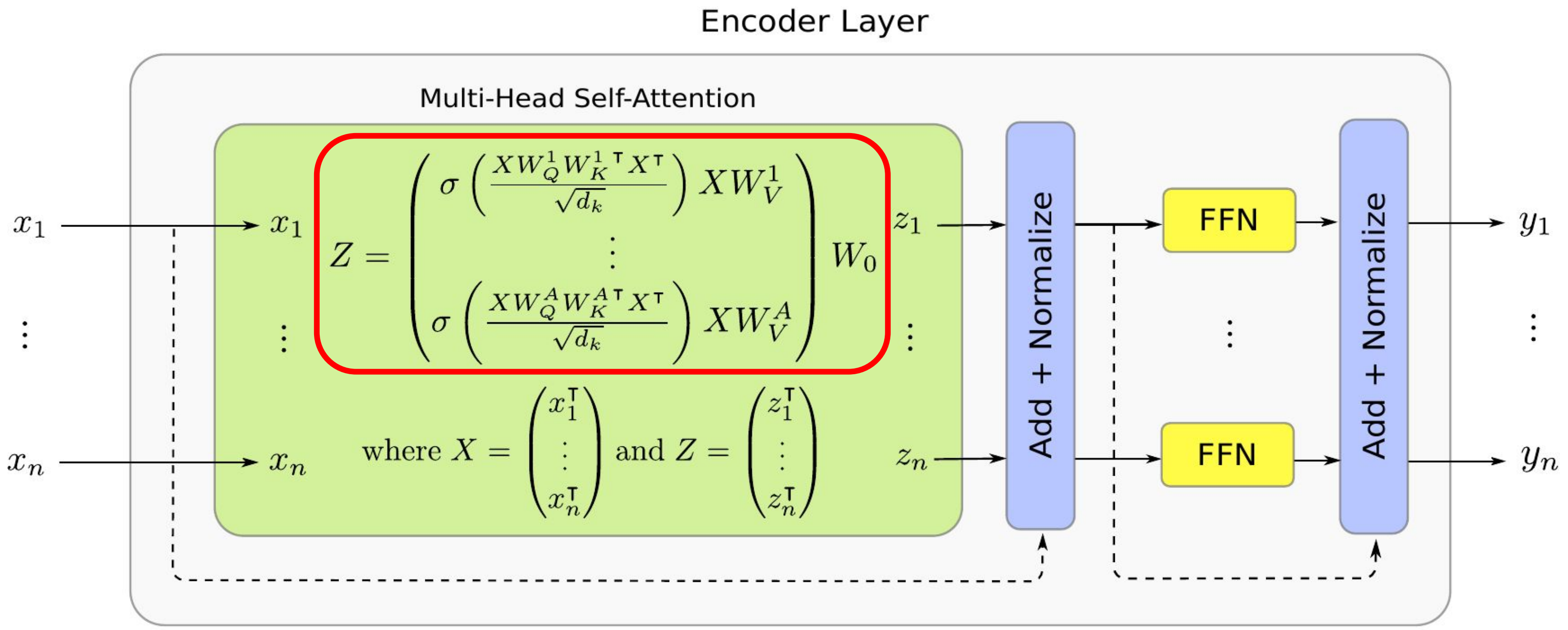
Transformer networks: encoder layer

Single attention head



Transformer networks: encoder layer

Multiple attention heads



where $\sigma = \text{softmax}$

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

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

Goal: Create a fast and precise dot product abstract transformer

Challenges of Transformer network verification

2. Softmax

Challenges:

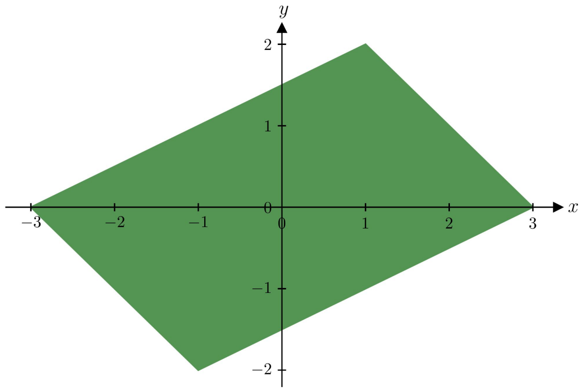
- Exponential and Division abstract transformers cause great precision loss
- Concrete output represents a probability distribution, but this information is lost during abstraction

$$\mathrm{Attention}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \boxed{\mathrm{softmax}}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right)\mathbf{V}$$

$$\mathrm{softmax}_i(\nu_1, \ldots, \nu_N) = \frac{e^{\nu_i}}{\sum_{j=1}^{N} e^{\nu_j}}$$

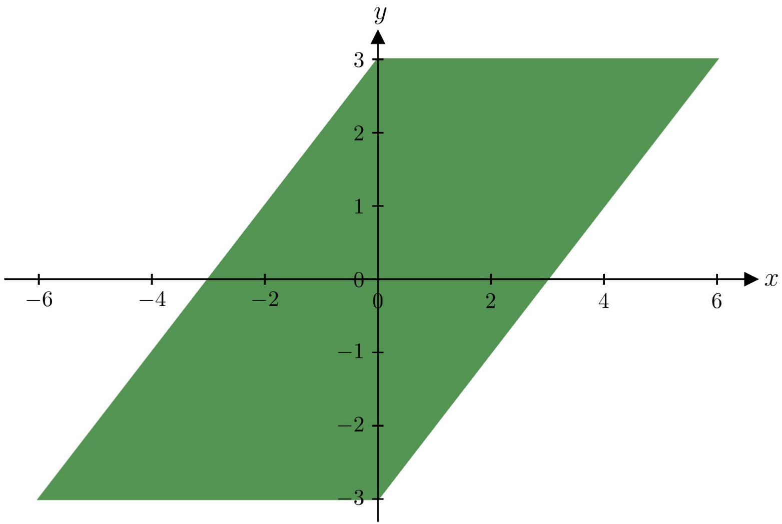
Goal: Create a fast and precise softmax abstract transformer

Classical Zonotopes



Linear layer
with weights

$$\begin{bmatrix} 1 & -2 \\ 1 & 2 \end{bmatrix}$$



$$\begin{cases} x = & -\epsilon_1 + 2\epsilon_2 \\ y = & \epsilon_1 + \epsilon_2 \end{cases}$$

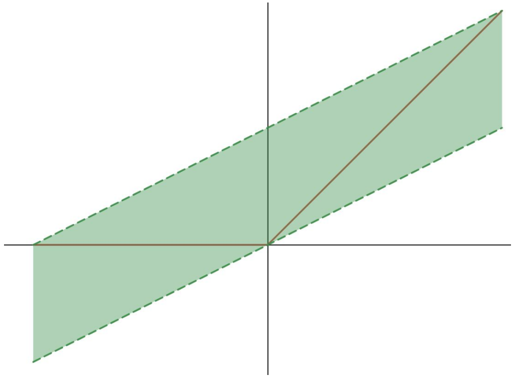
$$\epsilon_1, \epsilon_2 \in [-1,\, 1]$$

$$\begin{cases} x = & -5\epsilon_1 + 3\epsilon_2 \\ y = & -\epsilon_1 + 3\epsilon_2 \end{cases}$$

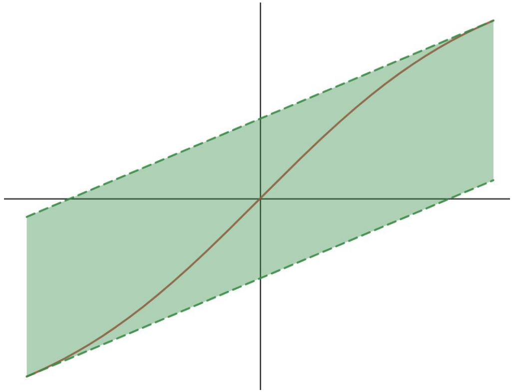
$$\epsilon_1, \epsilon_2 \in [-1,\, 1]$$

Classical Zonotopes

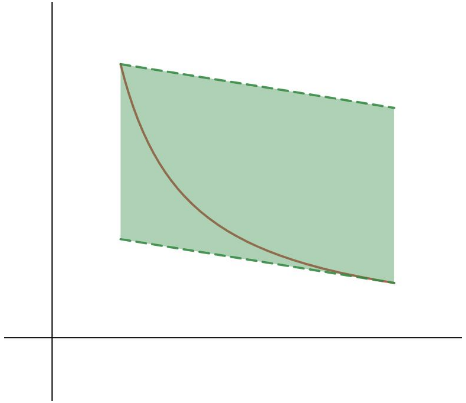
ReLU



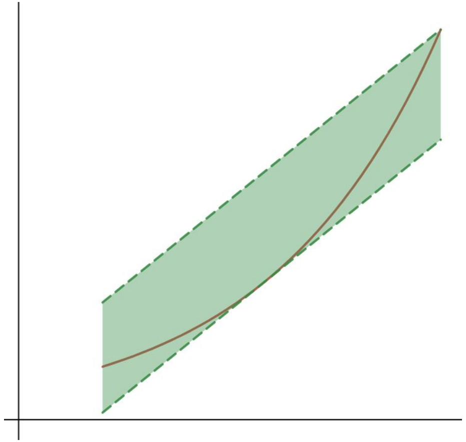
Tanh



Reciprocal



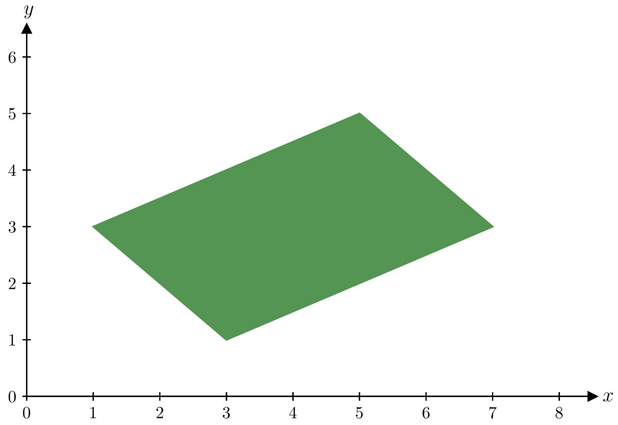
Exponential



ℓ^p ball representation in Classical Zonotopes

Challenge:

- Classical zonotopes precisely encode ℓ_∞ balls but over-approximate other ℓ_p balls, because they have ℓ_∞ bounded terms



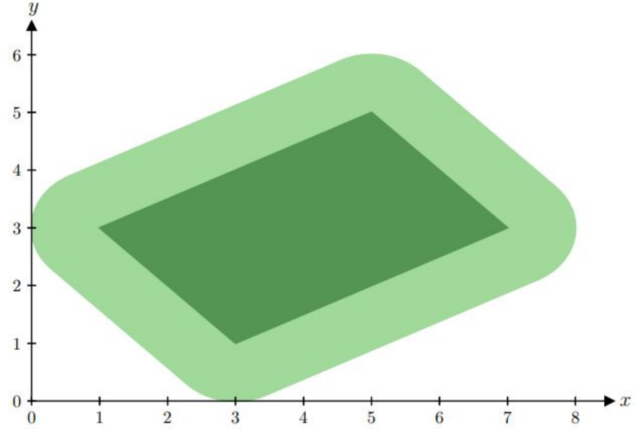
$$x_k = c_k + \sum_{j=1}^{\mathcal{E}_\infty} \beta_k^j \epsilon_j = c_k + \vec{\beta}_k \cdot \vec{\epsilon}$$

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

Goal: Create a zonotope that precisely captures an ℓ_p ball

Multi-Norm Zonotopes

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



One variable

$$x_k = c_k + \sum_{i=1}^{\mathcal{E}_p} \alpha_k^i \phi_i + \sum_{j=1}^{\mathcal{E}_\infty} \beta_k^j \epsilon_j = c_k + \vec{\alpha}_k \cdot \vec{\phi} + \vec{\beta}_k \cdot \vec{\epsilon}$$

$$c_k, \alpha_k^i, \beta_k^j \in \mathbb{R}, \quad \|\vec{\phi}\|_p \leq 1, \quad \|\epsilon_j\|_\infty \leq 1$$

Matrix form

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

$$\vec{c} \in \mathbb{R}^N, A \in \mathbb{R}^{N \times \mathcal{E}_p}, B \in \mathbb{R}^{N \times \mathcal{E}_\infty}$$

$$\|\vec{\phi}\|_p \leq 1, \quad \|\epsilon_j\|_\infty \leq 1$$

Abstract Transformer - Dot product

$$\vec{v}_1 = (\vec{c}_1 + A_1\vec{\phi} + B_1\vec{\epsilon}\,)$$

$$\vec{v}_2 = (\vec{c}_2 + A_2\vec{\phi} + B_2\vec{\epsilon}\,)$$

Developing the equations

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

$$= \underbrace{\vec{c}_1 \cdot \vec{c}_2 + (\vec{c}_1{}^\intercal A_2 + \vec{c}_2{}^\intercal A_1)\vec{\phi} + (\vec{c}_1{}^\intercal B_2 + \vec{c}_2{}^\intercal B_1)\vec{\epsilon}}_{\text{Multi-Norm Zonotope Form}} + \underbrace{(A_1\vec{\phi} + B_1\vec{\epsilon}) \cdot (A_2\vec{\phi} + B_2\vec{\epsilon})}_{\text{Needs to be developed}}$$

Goal: find a Multi-norm Zonotope representation for the last term

Abstract Transformer - Dot product

Challenge:

- Putting bounds on the interaction between noise symbols

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

Abstract Transformer - Dot product (DeepT-Fast)

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

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

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

$$\left|\begin{bmatrix}2\phi_1 + 3\phi_2 \\ 4\phi_1 - 5\phi_2\end{bmatrix} \cdot \begin{bmatrix}2\epsilon_1 - 3\epsilon_2 + 4\epsilon_3 \\ 4\epsilon_1 + 5\epsilon_2 + 5\epsilon_3\end{bmatrix}\right| \leq \left|\begin{bmatrix}2\phi_1 + 3\phi_2 \\ 4\phi_1 - 5\phi_2\end{bmatrix}\right| \cdot \left|\begin{bmatrix}2\epsilon_1 - 3\epsilon_2 + 4\epsilon_3 \\ 4\epsilon_1 + 5\epsilon_2 + 5\epsilon_3\end{bmatrix}\right|$$

$$\leq \left|\begin{bmatrix}2\phi_1 + 3\phi_2 & 4\phi_1 - 5\phi_2\end{bmatrix}\right| \begin{bmatrix}|2\epsilon_1 - 3\epsilon_2 + 4\epsilon_3| \\ |4\epsilon_1 + 5\epsilon_2 + 5\epsilon_3|\end{bmatrix}$$

$$\leq \left|\begin{bmatrix}2\phi_1 + 3\phi_2 & 4\phi_1 - 5\phi_2\end{bmatrix}\right| \begin{bmatrix}\|2 & -3 & 4\|_1 \\ \|4 & 5 & 5\|_1\end{bmatrix}$$

$$\leq \left|\begin{bmatrix}2\phi_1 + 3\phi_2 & 4\phi_1 - 5\phi_2\end{bmatrix}\right| \begin{bmatrix}9 \\ 14\end{bmatrix}$$

Abstract Transformer - Dot product (DeepT-Fast)

$$
\begin{aligned}
\left|\begin{bmatrix} 2\phi_1 + 3\phi_2 & 4\phi_1 - 5\phi_2 \end{bmatrix}\right| \begin{bmatrix} 9 \\ 14 \end{bmatrix} &= \begin{bmatrix} 9 & 14 \end{bmatrix} \left|\begin{bmatrix} 2\phi_1 + 3\phi_2 \\ 4\phi_1 - 5\phi_2 \end{bmatrix}\right| \\
&= \begin{bmatrix} 9 & 14 \end{bmatrix} \left|\begin{bmatrix} 2 & 3 \\ 4 & -5 \end{bmatrix} \begin{bmatrix} \phi_1 \\ \phi_2 \end{bmatrix}\right| \\
&\leq \begin{bmatrix} 9 & 14 \end{bmatrix} \left|\begin{bmatrix} 2 & 3 \\ 4 & -5 \end{bmatrix}\right| \left|\begin{bmatrix} \phi_1 \\ \phi_2 \end{bmatrix}\right| \\
&= \begin{bmatrix} 9 & 14 \end{bmatrix} \begin{bmatrix} 2 & 3 \\ 4 & 5 \end{bmatrix} \left|\begin{bmatrix} \phi_1 \\ \phi_2 \end{bmatrix}\right| \\
&= \begin{bmatrix} 60 & 106 \end{bmatrix} \left|\begin{bmatrix} \phi_1 \\ \phi_2 \end{bmatrix}\right| \\
&= \left|\begin{bmatrix} 60\phi_1 + 106\phi_2 \end{bmatrix}\right| \\
&\leq \|60\phi_1 + 106\phi_2\|_q \\
&= 106
\end{aligned}
$$

Abstract Transformer - Dot product (DeepT-Fast)

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

$$|(V\vec{\xi}_{p_1}) \cdot (W\vec{\xi}_{p_2})| \leq \left\| \begin{pmatrix} \|\vec{w}_1\|_{q_2} \\ \vdots \\ \|\vec{w}_N\|_{q_2} \end{pmatrix}^T |V| \right\|_{q_1}$$

$$p_1, p_2 \in \{1, 2, \infty\}$$

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

$$
\begin{aligned}
(A_1\vec{\phi} + B_1\vec{\epsilon}) \cdot (A_2\vec{\phi} + B_2\vec{\epsilon}) = {} & (A_1\vec{\phi}) \cdot (A_2\vec{\phi}) + (A_1\vec{\phi}) \cdot (B_2\vec{\epsilon}) \\
& + (B_1\vec{\epsilon}) \cdot (A_2\vec{\phi}) + \boxed{(B_1\vec{\epsilon}) \cdot (B_2\vec{\epsilon})}
\end{aligned}
$$

Abstract Transformer - Dot product (DeepT-Precise)

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

$$
\begin{aligned}
\begin{bmatrix} 2\epsilon_1 - 3\epsilon_2 \\ -1\epsilon_1 + \epsilon_2 \end{bmatrix} \begin{bmatrix} 3\epsilon_1 - 4\epsilon_2 \\ 1\epsilon_1 + 2\epsilon_2 \end{bmatrix} &= (2\epsilon_1 - 3\epsilon_2)(3\epsilon_1 - 4\epsilon_2) + (-1\epsilon_1 + \epsilon_2)(1\epsilon_1 + 2\epsilon_2) \\
&= 5\epsilon_1^2 - 10\epsilon_1\epsilon_2 - 8\epsilon_2\epsilon_1 + 14\epsilon_2^2 \\
&= 5\epsilon_1^2 - 18\epsilon_1\epsilon_2 + 14\epsilon_2^2 \\
&\in 5[0, 1] - 18[-1, 1] + 14[0, 1] \\
&\in [0, 5] + [-18, 18] + [0, 14] \\
&\in [-18, 37]
\end{aligned}
$$

Goubault, Eric, and Sylvie Putot. "Static analysis of numerical algorithms." *International Static Analysis Symposium*. Springer, Berlin, Heidelberg, 2006.

Abstract Transformer - Dot product (DeepT-Precise)

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

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

Goubault, Eric, and Sylvie Putot. "Static analysis of numerical algorithms." *International Static Analysis Symposium*. Springer, Berlin, Heidelberg, 2006.

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]

$$\mathrm{softmax}_i(v_1, \ldots, v_N) = \frac{e^{v_i}}{\sum_{j=1}^{N} e^{v_j}} = \frac{1}{\sum_{j=1}^{N} e^{v_j - v_i}}$$

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])

[1] Ghorbal, Khalil, Eric Goubault, and Sylvie Putot. "A logical product approach to zonotope intersection." *International Conference on Computer Aided Verification*. Springer, Berlin, Heidelberg, 2010.

Experimental setup

Task: verifying Transformer Networks performing binary classification

Baseline : CROWN-BaF and CROWN-Backward [1]

Dasaset: 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)

[1] Shi, Zhouxing, et al. "Robustness Verification for Transformers." In: 8th International Conference on Learning Representations (ICLR) 2020

Evaluation - Embedding attack (ℓ_∞)

M	DeepT-Fast			CROWN-BaF			DeepT-Precise			CROWN-Backward		
	Min	Avg	Time	Min	Avg	Time	Min	Avg	Time	Min	Avg	Time
3	0.013	0.034	17.2	0.013	0.033	**3.9**	0.013	**0.038**	432.2	**0.013**	0.037	65.4
6	0.014	0.031	33.8	0.014	0.025	**10.6**	0.014	**0.036**	1195.1	**0.015**	0.033	215.5
12	9.3e-3	0.021	69.0	1.9e-3	6.3e-3	**44.9**	8.8e-3	**0.024**	2676.5	**9.9e-3**	0.022	839.9

Evaluation - Embedding attack (ℓ_∞)

M	DeepT-Fast			CROWN-BaF			DeepT-Precise			CROWN-Backward		
	Min	Avg	Time	Min	Avg	Time	Min	Avg	Time	Min	Avg	Time
3	0.013	0.034	17.2	0.013	0.033	**3.9**	0.013	**0.038**	432.2	**0.013**	0.037	65.4
6	0.014	0.031	33.8	0.014	0.025	**10.6**	0.014	**0.036**	1195.1	**0.015**	0.033	215.5
12	9.3e-3	0.021	69.0	1.9e-3	6.3e-3	**44.9**	8.8e-3	**0.024**	2676.5	**9.9e-3**	0.022	839.9

Evaluation - Embedding attack (all norms)

M	ℓ_p	DeepT-Fast			CROWN-BaF			Ratio
		Min	Avg	Time	Min	Avg	Time	
3	ℓ_1	**0.036**	**1.808**	28.8	**0.036**	1.686	**9.7**	**1.07**
	ℓ_2	**6.4e-3**	**0.330**	29.2	6.2e-3	0.328	**10.6**	**1.01**
	ℓ_∞	**2.1e-3**	0.032	26.2	6.0e-4	**0.033**	**9.7**	0.99
6	ℓ_1	**0.089**	**1.191**	63.6	0.070	0.470	**29.8**	**2.53**
	ℓ_2	**0.015**	**0.212**	64.5	0.012	0.083	**29.8**	**2.56**
	ℓ_∞	**1.2e-3**	**0.021**	57.1	1.1e-3	8.0e-3	**27.0**	**2.56**
12	ℓ_1	**0.358**	**0.512**	125.9	1.3e-3	0.018	**100.7**	**28.4**
	ℓ_2	**0.074**	**0.107**	129.4	2.6e-4	3.7e-3	**106.3**	**29.3**
	ℓ_∞	**7.3e-3**	**0.011**	113.4	2.5e-5	3.5e-4	**87.4**	**29.8**

Embedding attack - combining the verifiers

M	Combined DeepT verifier			CROWN-Backward		
	Min	Avg	Time	Min	Avg	Time
6	0.014	**0.034**	**227.0**	**0.015**	0.033	289.4
12	9.1e-3	**0.023**	**423.7**	**9.9e-3**	0.022	818.8

Synonym attack example:

~23 million possible combinations

Tokens	#Synonyms	Synonyms
No	3	no, not, without
reason	1	reasons
for	0	∅
anyone	6	sombody, someone, anybody, everyone, person, nobody
to	0	∅
invest	0	∅
their	0	∅
hard-earned	0	∅
bucks	1	money
into	5	at, towards, toward, in, for
a	0	∅
movie	3	film, films, cinema
which	0	∅
obviously	5	clearly, naturally, apparently, plainly, definitely
did	4	did, could, got, do, does
n't	0	∅
invest	0	∅
much	5	very, many, highly, greatly, heavily
into	6	at, under, towards, in, for
itself	6	himself, themselves, ourselves, myself, yourself, herself
either	0	∅
.	0	∅

Evaluation - Synonym attacks

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)

	Certified Sentences	Certified Percentage	Time (s)
CROWN-BaF	121	89%	2.60
DeepT-fast	120	88%	2.41

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

[1] Xu, Kaidi, et al. "Automatic perturbation analysis for scalable certified robustness and beyond." *NeurIPS* (2020).

Comparing DeepT and CROWN

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

[1] Shi, Zhouxing, et al. "Robustness Verification for Transformers." In: 8th International Conference on Learning Representations (ICLR) 2020

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!