

Precise Verification of Transformers through ReLU-Catalyzed Abstraction Refinement

Hengjie Liu¹, Zhenya Zhang^{1,2}, Jianjun Zhao¹

¹Kyushu University

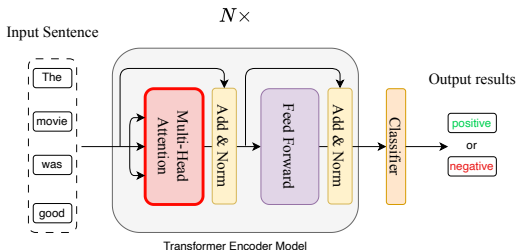
²National Institute of Informatics

Lisbon, Portugal | July 24, 2026



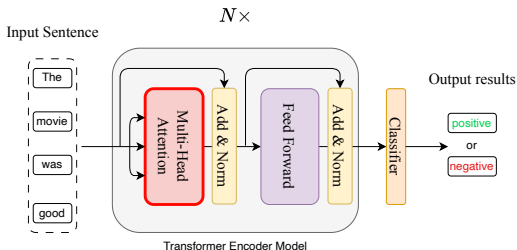
KYUSHU UNIVERSITY

Transformers Power Modern AI



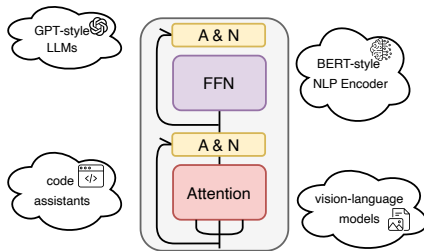
- A Transformer is a **neural architecture for sequence data** (Vaswani et al., 2017).

Transformers Power Modern AI



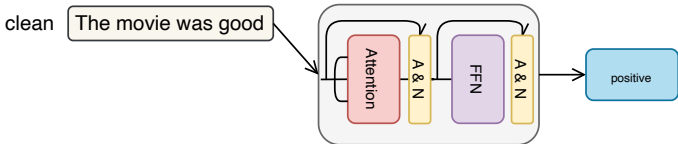
- ▶ A Transformer is a **neural architecture for sequence data** (Vaswani et al., 2017).
- ▶ Attention builds context-aware representations by connecting tokens.

Transformers Power Modern AI



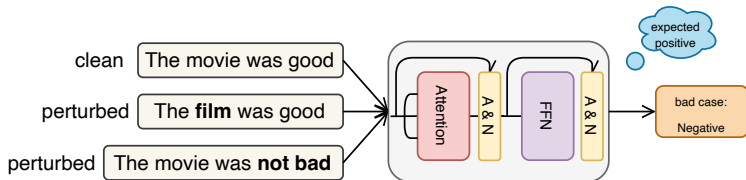
- ▶ A Transformer is a **neural architecture for sequence data** (Vaswani et al., 2017).
- ▶ Attention builds context-aware representations by connecting tokens.
- ▶ Transformers power LLMs, translation, search, code, and multimodal systems.

A Correct Answer Can Still Be Fragile



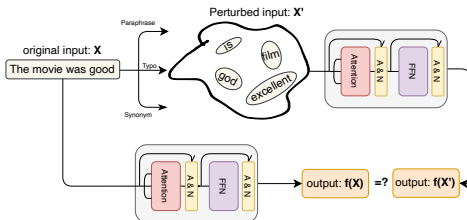
- A clean sentence can receive the expected prediction.

A Correct Answer Can Still Be Fragile



- ▶ A clean sentence can receive the expected prediction.
- ▶ Meaning-preserving perturbations can still flip that prediction.

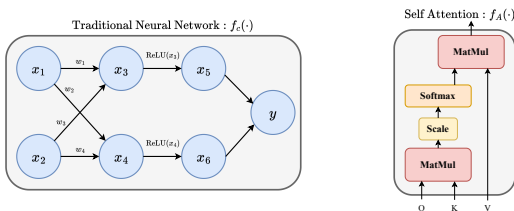
A Correct Answer Can Still Be Fragile



Specification: $\forall X' \in \text{Perturbed}(X), \quad f(X') = f(X).$

- ▶ A clean sentence can receive the expected prediction.
- ▶ Meaning-preserving perturbations can still flip that prediction.
- ▶ We want to **formally prove** that every permitted input preserves the prediction.

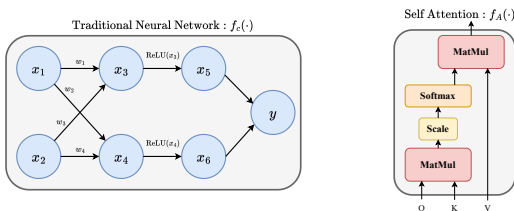
The Verification Bottleneck Hides in Attention



$$S = QK^{\top}, \quad S_{ij} = \sum_k Q_{ik}K_{jk} \quad \implies \quad z = xy.$$

- Traditional networks concentrate nonlinearity in activations such as $h = \text{ReLU}(Wx + b)$.

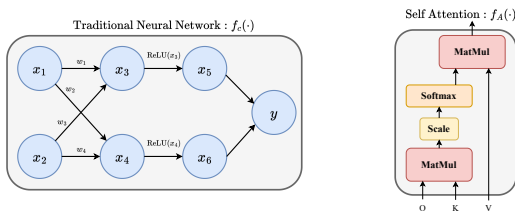
The Verification Bottleneck Hides in Attention



$$S = QK^{\top}, \quad S_{ij} = \sum_k Q_{ik}K_{jk} \quad \implies \quad z = xy.$$

- ▶ Traditional networks concentrate nonlinearity in activations such as $h = \text{ReLU}(Wx + b)$.
- ▶ Attention adds a **bilinear MatMul**: both Q and K change with the perturbed input.

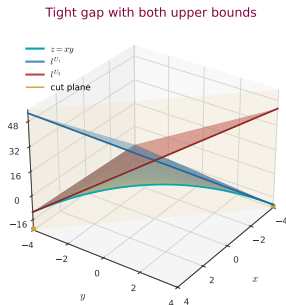
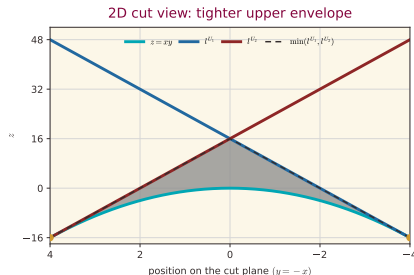
The Verification Bottleneck Hides in Attention



$$S = QK^{\top}, \quad S_{ij} = \sum_k Q_{ik}K_{jk} \quad \implies \quad z = xy.$$

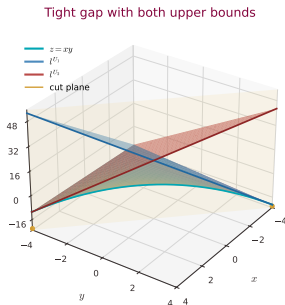
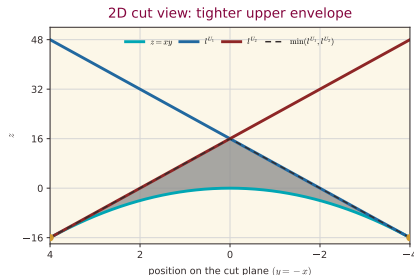
- ▶ Traditional networks concentrate nonlinearity in activations such as $h = \text{ReLU}(Wx + b)$.
- ▶ Attention adds a **bilinear MatMul**: both Q and K change with the perturbed input.
- ▶ Scalable verification therefore needs **tight linear bounds** for multiplying two numbers, $z = xy$.

BuFFeT Keeps the Better Sound Plane



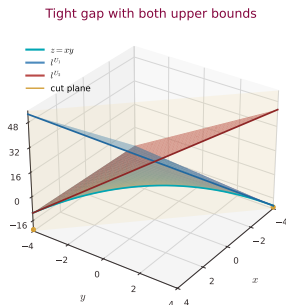
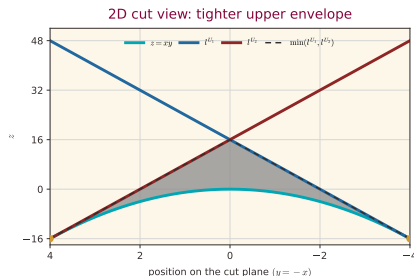
- Prior work uses one sound upper plane, l^{U_1} (Shi et al., 2020).

BuFFeT Keeps the Better Sound Plane



- Prior work uses one sound upper plane, l^{U_1} (Shi et al., 2020).
- We introduce its dual, l^{U_2} ; each plane is tighter in a different region.

BuFFeT Keeps the Better Sound Plane



- Prior work uses one sound upper plane, l^{U1} (Shi et al., 2020).
- We introduce its dual, l^{U2} ; each plane is tighter in a different region.
- Keeping the better one gives $l^* = \min(l^{U1}, l^{U2})$ without sacrificing soundness.

ReLU Turns the Minimum into One Parameter

Let $t = l^{U_1} - l^{U_2}$. Then

$$l^* = \min(l^{U_1}, l^{U_2}) = l^{U_1} - \text{ReLU}(t).$$

$$\boxed{\text{ReLU}(t) \geq \alpha t, \alpha \in [0, 1]}$$

$$\implies$$

$$\boxed{l^* \leq l_\alpha^* = (1 - \alpha)l^{U_1} + \alpha l^{U_2}}$$

- ReLU gives an exact representation of the pointwise choice.

ReLU Turns the Minimum into One Parameter

Let $t = l^{U_1} - l^{U_2}$. Then

$$l^* = \min(l^{U_1}, l^{U_2}) = l^{U_1} - \text{ReLU}(t).$$

$$\boxed{\text{ReLU}(t) \geq \alpha t, \alpha \in [0, 1]}$$

$$\implies$$

$$\boxed{l^* \leq l_{\alpha}^* = (1 - \alpha)l^{U_1} + \alpha l^{U_2}}$$

- ▶ ReLU gives an exact representation of the pointwise choice.
- ▶ Its linear relaxation turns that choice into one parameter, α .

ReLU Turns the Minimum into One Parameter

Let $t = l^{U_1} - l^{U_2}$. Then

$$l^* = \min(l^{U_1}, l^{U_2}) = l^{U_1} - \text{ReLU}(t).$$

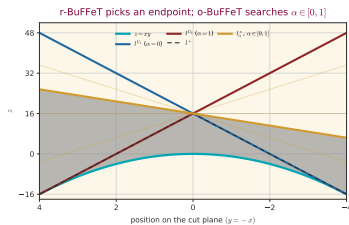
$$\boxed{\text{ReLU}(t) \geq \alpha t, \alpha \in [0, 1]}$$

$$\implies$$

$$\boxed{l^* \leq l_\alpha^* = (1 - \alpha)l^{U_1} + \alpha l^{U_2}}$$

- ▶ ReLU gives an exact representation of the pointwise choice.
- ▶ Its linear relaxation turns that choice into one parameter, α .
- ▶ $\alpha = 0$ recovers l^{U_1} ; $\alpha = 1$ recovers l^{U_2} .

Two Practical Choices: Fast or Precise



r-BuFFeT: one-pass rule (Singh et al., 2019)

$$|a| > |b| \Rightarrow \alpha = 0,$$

$$|b| > |a| \Rightarrow \alpha = 1.$$

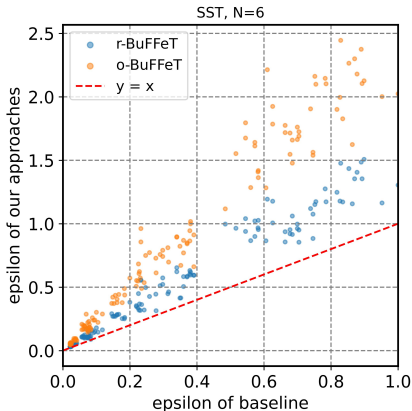
Select an endpoint from the ReLU input range.

o-BuFFeT: gradient search (Xu et al., 2021)

$$\alpha^* = \arg \min_{\alpha \in [0,1]} \mathcal{L}(M(\alpha)).$$

Optimize the verification margin like network training.

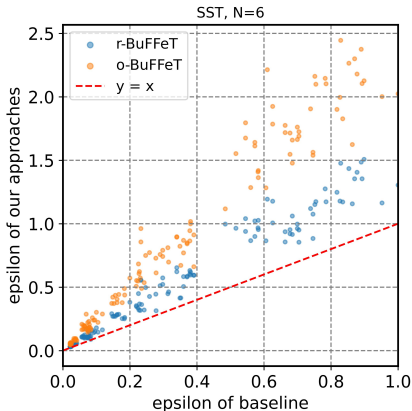
The Payoff: More Precise Verification



10 models, 1400 tasks

- SST (Socher et al., 2013) and Yelp polarity (Zhang et al., 2015); CrownBaF baseline (Shi et al., 2020).

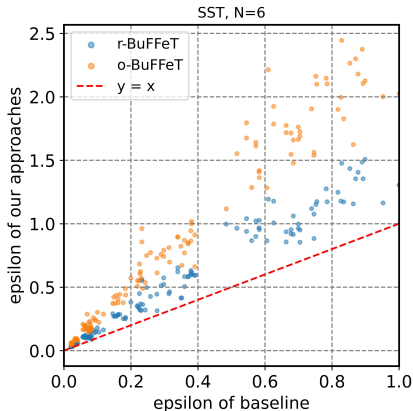
The Payoff: More Precise Verification



10 models, 1400 tasks

- SST (Socher et al., 2013) and Yelp polarity (Zhang et al., 2015); CrownBaF baseline (Shi et al., 2020).
- Most points lie above $y = x$: BuFFeT verifies larger radii.

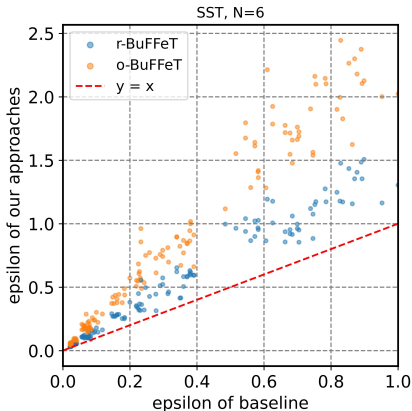
The Payoff: More Precise Verification



10 models, 1400 tasks

- SST (Socher et al., 2013) and Yelp polarity (Zhang et al., 2015); CrownBaF baseline (Shi et al., 2020).
- Most points lie above $y = x$: BuFFeT verifies larger radii.
- r-BuFFeT reaches up to $1.8\times$ with near-baseline runtime.

The Payoff: More Precise Verification



10 models, 1400 tasks

- SST (Socher et al., 2013) and Yelp polarity (Zhang et al., 2015); CrownBaF baseline (Shi et al., 2020).
- Most points lie above $y = x$: BuFFeT verifies larger radii.
- r-BuFFeT reaches up to $1.8\times$ with near-baseline runtime.
- o-BuFFeT reaches up to $3.6\times$, with higher cost.

One Idea to Remember

- Attention verification ultimately encounters a simple but difficult term: $z = xy$.

One Idea to Remember

- ▶ Attention verification ultimately encounters a simple but difficult term: $z = xy$.
- ▶ **Two sound planes + ReLU** turn a loose geometric bound into an optimizable one.

One Idea to Remember

- ▶ Attention verification ultimately encounters a simple but difficult term: $z = xy$.
- ▶ **Two sound planes + ReLU** turn a loose geometric bound into an optimizable one.
- ▶ The result is a practical speed–precision choice and up to **3.6×** the baseline radius.

Thank you

Thanks for your listening!

Appendix: Contributions

- ▶ **ReLU-catalyzed representation.** BuFFeT fuses the existing planar dot-product bound and our proposed dual through ReLU.
- ▶ **Two refinement strategies.** r-BuFFeT adapts range-based ReLU selection (Singh et al., 2019); o-BuFFeT adapts optimization-based bound search (Xu et al., 2021).
- ▶ **Evaluation and impact.** Across 10 models and 1400 tasks, BuFFeT improves precision for most tasks while bridging classic ReLU and Transformer verification.

Appendix: Runtime Data

Dataset	N	baseline	r-BuFFeT	Δ	o-BuFFeT	Δ
SST	1	0.06	0.09	1.5	5.77	96.2
	2	0.17	0.26	1.5	14.70	86.5
	3	1.05	1.76	1.7	101.72	96.9
	6	11.38	12.57	1.1	973.50	85.5
	TinyBERT	4.30	5.27	1.2	281.07	65.4
Yelp	1	0.08	0.13	1.6	2.68	33.5
	2	0.17	0.23	1.4	12.97	76.3
	3	0.92	1.26	1.4	77.61	84.4
	6	8.05	9.43	1.2	673.42	83.7
	TinyBERT	2.29	3.34	1.5	175.38	76.6

- r-BuFFeT stays in the same runtime order as the baseline.
- o-BuFFeT pays substantially more time for more precise verification.

References I

- Zhouxing Shi, Huan Zhang, Kai-Wei Chang, Minlie Huang, and Cho-Jui Hsieh. Robustness verification for transformers. In *International Conference on Learning Representations*, 2020. URL <https://openreview.net/forum?id=BJxwPJHFWs>.
- Gagandeep Singh, Timon Gehr, Markus Püschel, and Martin Vechev. An abstract domain for certifying neural networks. *Proc. ACM Program. Lang.*, 3(POPL), January 2019. doi: 10.1145/3290354.
- Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In David Yarowsky, Timothy Baldwin, Anna Korhonen, Karen Livescu, and Steven Bethard, editors, *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA, October 2013. Association for Computational Linguistics. URL <https://aclanthology.org/D13-1170/>.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.
- Kaidi Xu, Huan Zhang, Shiqi Wang, Yihan Wang, Suman Jana, Xue Lin, and Cho-Jui Hsieh. Fast and complete: Enabling complete neural network verification with rapid and massively parallel incomplete verifiers. In *International Conference on Learning Representations*, 2021. URL <https://openreview.net/forum?id=nVZtXBI6LNn>.
- Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In *Proceedings of the 29th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, page 649–657, Cambridge, MA, USA, 2015. MIT Press. URL https://proceedings.neurips.cc/paper_files/paper/2015/file/250cf8b51c773f3f8dc8b4be867a9a02-Paper.pdf.