

Certified Robustness Under Bounded Levenshtein Distance

Elias Abad Rocamora, Grigorios G. Chrysos and Volkan Cevher



ICLR

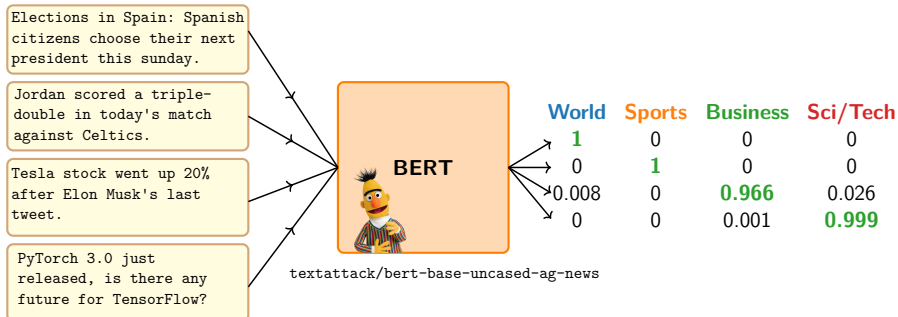
EPFL



WISCONSIN
UNIVERSITY OF WISCONSIN-MADISON

ICLR, Singapore 2025

Adversarial Attacks in Sentence Classification



Adversarial Attacks in Sentence Classification

E2lections in Spain: Spanish citizens choose their next president this sunday.

Jordan scored a triple-double in todaK's match against Celtics.

Tesla stock went up 20% after Elon Musk's last twee\t.

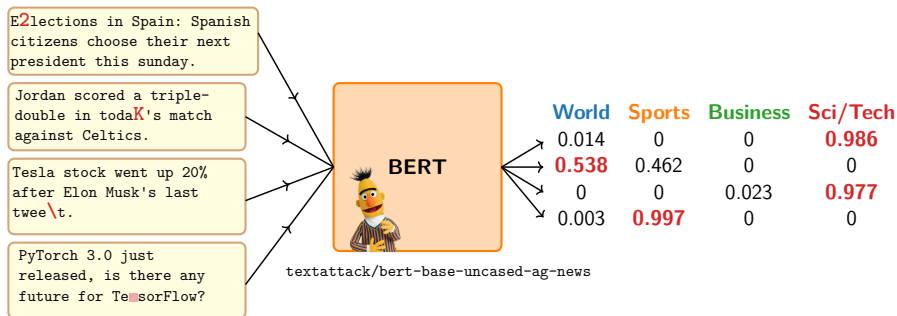
PyTorch 3.0 just released, is there any future for Te sorFlow?



textattack/bert-base-uncased-ag-news

| World | Sports | Business | Sci/Tech |
|--------------|--------------|----------|--------------|
| 0.014 | 0 | 0 | 0.986 |
| 0.538 | 0.462 | 0 | 0 |
| 0 | 0 | 0.023 | 0.977 |
| 0.003 | 0.997 | 0 | 0 |

Adversarial Attacks in Sentence Classification



How can we obtain certifiably robust models? 🤔

- The worst case margin needs to be positive:

$$g_{y,\hat{y}}(\mathbf{S}') = f_{\theta}(\mathbf{S}')_y - \max_{\hat{y} \neq y} f_{\theta}(\mathbf{S}')_{\hat{y}} > 0, \quad \forall \mathbf{S}' : d_{\text{Lev}}(\mathbf{S}, \mathbf{S}') \leq k,$$

where f_{θ} is our classifier, (\mathbf{S}, y) the original sentence and its label, d_{Lev} the Levenshtein distance and k our perturbation budget.

- The worst case margin needs to be positive:

$$g_{y,\hat{y}}(\mathbf{S}') = f_{\theta}(\mathbf{S}')_y - \max_{\hat{y} \neq y} f_{\theta}(\mathbf{S}')_{\hat{y}} > 0, \quad \forall \mathbf{S}' : d_{\text{Lev}}(\mathbf{S}, \mathbf{S}') \leq k,$$

where f_{θ} is our classifier, (\mathbf{S}, y) the original sentence and its label, d_{Lev} the Levenshtein distance and k our perturbation budget.

- **How can we check this condition?**

Brute Force: An illustrative example

\mathbf{S} = "Elections in Spain: Spanish citizens choose their next president this Sunday."

$y = 1$ ("World")

- Can we just try all of the sentences?

$$\arg \max_j f(\mathbf{S}')_j = 1 \quad \forall \mathbf{S}' \in \left\{ \begin{array}{l} \text{"Elec!tions in SpaiT: Spanish ..."} \\ \text{"3lections in Spain:4Spanish ..."} \\ \vdots \\ \text{"Elections.in Sain: Spanish ..."} \end{array} \right\}$$

Existing approaches

- The number of sentences to test grows exponentially with k 🤖

$$\left| \begin{Bmatrix} \text{"Elec!tions in SpaiT: Spanish ..."} \\ \text{"3lections in Spain:4Spanish ..."} \\ \vdots \\ \text{"Elections.in Sain: Spanish ..."} \end{Bmatrix} \right| = |\{\mathbf{S}' : d_{\text{Lev}}(\mathbf{S}, \mathbf{S}') \leq k\}| = \text{💣}$$

- Existing approaches also require a large number of forward passes:



Huang et al., Achieving verified robustness to symbol substitutions via interval bound propagation. *EMNLP 2019*.



Huang et al., Edit distance robustness certificates for sequence classifiers via randomized deletion. *NeurIPS 2023*.

- **Lipschitzness + large enough margin \Rightarrow Robustness** 🤩

$$\left. \begin{array}{l} |g_{y,\hat{y}}(\mathbf{S}) - g_{y,\hat{y}}(\mathbf{S}')| \leq G \cdot d_{\text{Lev}}(\mathbf{S}, \mathbf{S}') \\ + \\ g_{y,\hat{y}}(\mathbf{S}) > k \cdot G \end{array} \right\} \Rightarrow g_{y,\hat{y}}(\mathbf{S}') > 0 \quad \forall \mathbf{S}' : d_{\text{Lev}}(\mathbf{S}, \mathbf{S}') \leq k$$

where $g_{y,\hat{y}}(\mathbf{S}) = f(\mathbf{S})_y - f(\mathbf{S})_{\hat{y}}$.

- **This paper:** We are able to compute G for convolutional classifiers.

Some results

Table: Verified accuracy in AG-News: IBP, proposed by (Huang et al., 2019)

| p | k | Acc.(%) | Charmer | | BruteF | | IBP | | LipsLev | |
|----------|-----|---------|---------|-----------------|-----------------|-----------------|-----------------|-------|---------|---------|
| | | | Adv. | Acc.(%) Time(s) | Ver.(%) Time(s) | Ver.(%) Time(s) | Ver.(%) Time(s) | | | |
| ∞ | 1 | 65.23 | 47.90 | 5.70 | 47.87 | 16.15 | 27.77 | 16.76 | 32.33 | 0.0015 |
| | 2 | | 32.97 | 5.70 | OOT | | \times | | 11.60 | 0.0015 |
| 1 | 1 | 69.63 | 54.47 | 5.43 | 54.43 | 15.33 | 18.93 | 17.56 | 34.50 | 0.00140 |
| | 2 | | 37.77 | 5.43 | OOT | | \times | | 12.53 | 0.00140 |
| 2 | 1 | 74.80 | 62.20 | 7.32 | 62.07 | 29.12 | 29.10 | 31.54 | 38.80 | 0.00970 |
| | 2 | | 46.47 | 7.32 | OOT | | \times | | 13.93 | 0.00970 |

- Lipslev is the only method able to verify for $k > 1$.
- Lipslev is 4 orders of magnitude faster than other methods.

- **Conclusion** 🌟

- ▷ **LipsLev:** We introduce Lipschitz certification in NLP.
- ▷ First to verify Levenshtein constraints in a **single forward pass!**

- **Some interesting challenges** 🤔

- ▷ **Tokenizers:** Can we compute the Lips. constant of tokenizers?
- ▷ **Transformers:** Can we compute the Lips. constant of Self-Attention?
- ▷ **Scalability:** Can we scale the approach to larger models?

Thank you!

- **Check our paper:** More details, datasets and ablations!

Thanks for your attention!

contact: `elias.abadrocamora@epfl.ch`

