BLOCK-SPARSE ADVERSARIAL ATTACK TO FOOL TRANSFORMER-BASED TEXT CLASSIFIERS

Sahar Sadrizadeh[1], Ljiljana Dolamic[2], and Pascal Frossard[1]

[1]École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland
[2]Armasuisse S+T, Thun, Switzerland

Email: sahar.sadrizadeh@epfl.ch

EPFL

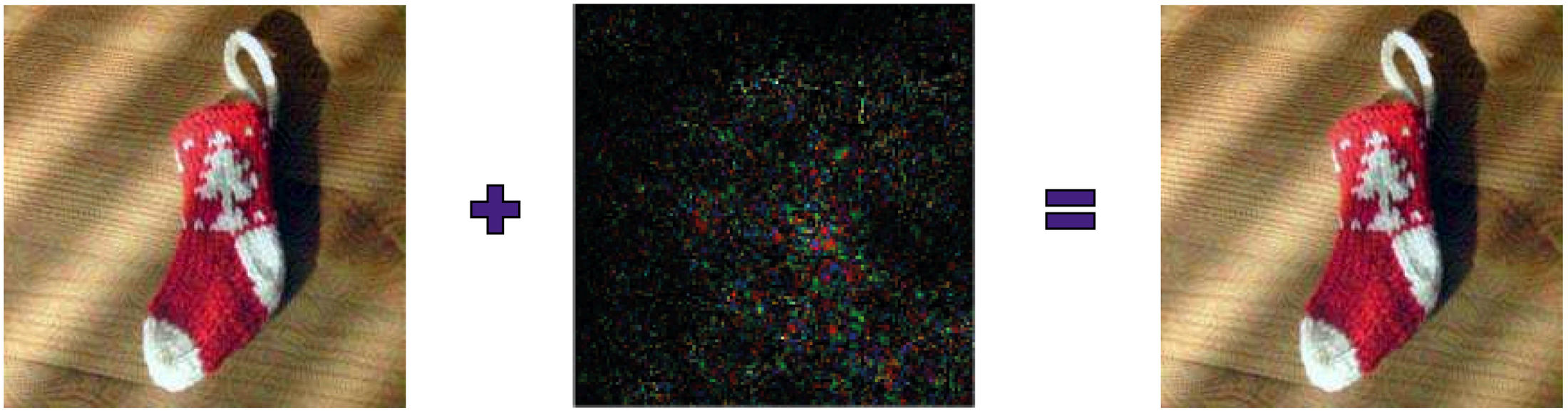
icassp 2022 Singapore

Outline

- **Introduction**
- **Proposed Method**
- **Experiments**
- **Conclusion**

Introduction

Adversarial Attacks against Images



Socks

Imperceptible Perturbation

Indian Elephant

Moosavi-Dezfooli, et al. “DeepFool: a simple and accurate method to fool deep neural network.” *CVPR*, 2016

Challenges in NLP

- Input Data is **Discrete:**
  - The system is not end-to-end differentiable (optimizing over discrete input is difficult)
  - Working on embedding space may result in invalid text
- **Perturbation** of the adversarial example from the original sentence should be **imperceptible:**
  - Semantics: Similar meaning
  - Syntax: Correct grammar (fluent sentence)

Methods for adversarial attack against images is not applicable

Word Substitution Adversarial Attacks

1. **Rank Words**
   - White-box attack: Gradients
   - Black-box attack: Masking each word and compare the output scores
2. Find possible **replacements** to maintain semantics
   - Masked Language Models
   - Embedding Space
3. **Substitute** with the word that changes the network result
   - Two checks: **POS** and similar semantics (**Universal Sentence Encoder**)

Jin, et al. "Is bert really robust? a strong baseline for natural language attack on text classification and entailment." *AAAI*, 2020.

Li, et al. "BERT-ATTACK: Adversarial Attack Against BERT Using BERT." *EMNLP*, 2020.

Problem Formulation

White-box attack against transformers

Assumption: adversarial example and input have equal lengths

Original sentence: $\mathbf{x} = x_1 x_2 ... x_n \xrightarrow{\text{embedding}} \mathbf{e_x} = [\text{emb}(x_1), \text{emb}(x_2), ..., \text{emb}(x_n)]$

$y = f(\mathbf{x})$

Adversarial example: $\mathbf{x}' = x'_1 x'_2 ... x'_n \xrightarrow{\text{embedding}} \mathbf{e}_{\mathbf{x}'} = \mathbf{e_x} + \mathbf{r_x}$

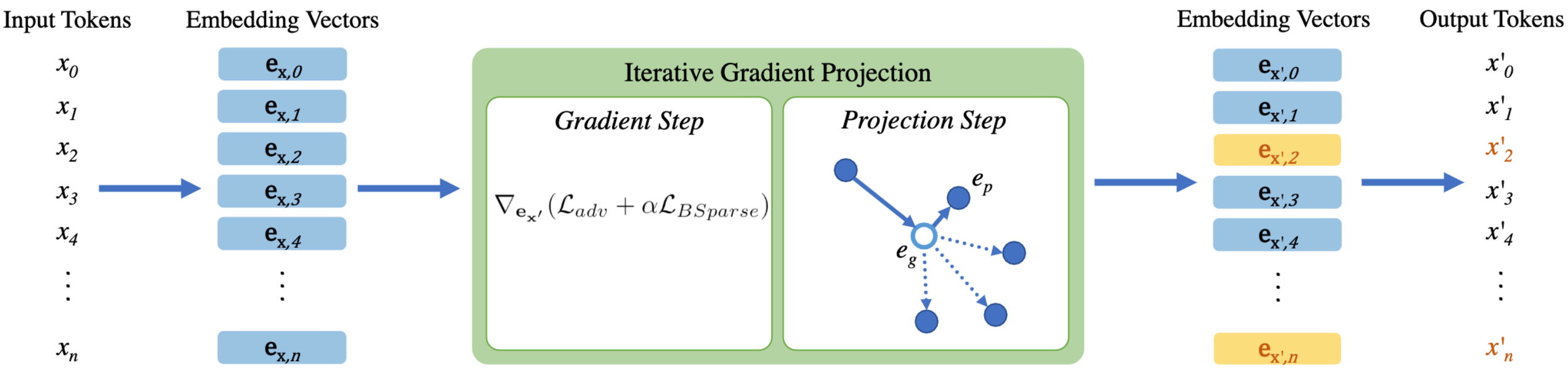
$f(\mathbf{x}') \neq y$

Perturbation: $\mathbf{r_x} = [\mathbf{r}_1, \mathbf{r}_2, ..., \mathbf{r}_n]$ $\xrightarrow{\text{Only few word changes}}$ $\mathbf{r_x}$ is block sparse

$$\mathcal{L}_{Adv} = -\mathcal{L}_f(\mathbf{e}_{\mathbf{x}'}, y) \qquad \mathcal{L}_{BSparse} = \sum_{i=1}^{n} \|\mathbf{r}_i\|_2$$

$$\hat{\mathbf{e}_{\mathbf{x}'}} = \underset{\mathbf{e}_{\mathbf{x}'} \in \mathcal{E}_{\mathcal{V}}}{\text{argmin}} \, \mathcal{L}_{Adv} + \alpha \mathcal{L}_{BSparse}$$

Block Diagram



Block diagram of the proposed method.

Algorithm

Stop if the label is wrong

Gradient descent

Projection

Algorithm 1 Block-Sparse Adversarial Attack

Input:
$f(.)$: Target classifier model, $\mathcal{V}$: Vocabulary set
$\mathbf{x}$: Tokenized input sentence, lr: Learning rate
A: Set of decreasing values for Hyper-parameter α to control the importance of the block-sparsity term
K: Maximum number of iterations
Output:
$\mathbf{x}'$: Generated adversarial example
procedure
initialization:
buffer $\leftarrow$ empty, $y \leftarrow f(\mathbf{x})$, $k \leftarrow 0$
$\forall i \in \{1, ...n\} \quad \mathbf{e_{g},}_i \leftarrow \text{emb}(x_i)$
for α in A **do**
while $f(\mathbf{e_p}) = y$ and $k \leq K$ **do**
$k \leftarrow k + 1$
Step 1: Gradient descent in the continuous embedding space:
$\mathbf{e_g} \leftarrow \mathbf{e_g} - lr.\nabla_{\mathbf{e_{x'}}}(\mathcal{L}_{adv} + \alpha\mathcal{L}_{BSparse})$
Step 2: Projection to the discrete subspace $\mathcal{E}_\mathcal{V}$ and update if the sentence is new:
for $i \in \{1, ..., n\}$ **do**
$\mathbf{e_{p},}_i \leftarrow \underset{\mathbf{e} \in \mathcal{E}_\mathcal{V}}{\text{argmin}} \frac{\mathbf{e}^\top \mathbf{e_{g},}_i}{\|\mathbf{e}\|_2 . \|\mathbf{e_{g},}_i\|_2}$
end for
if $\mathbf{e_p}$ not in **buffer then**
add $\mathbf{e_p}$ to **buffer**
$\mathbf{e_g} \leftarrow \mathbf{e_p}$
end if
end while
if $f(\mathbf{e_p}) \neq y$ **then**
break (adversarial example is found)
end if
end for
return $\mathbf{e_{x'}} \leftarrow \mathbf{e_p}$
end procedure

Experiments

Experimental Setup

- **Target model:** GPT-2 Transformer
- **3 Datasets:**
  - MNLI (natural language inference),
  - AG News (news categorization)
  - Yelp Reviews (sentiment classification).
- **Baseline:** GBDA (optimization-based white-box attack against transformers)

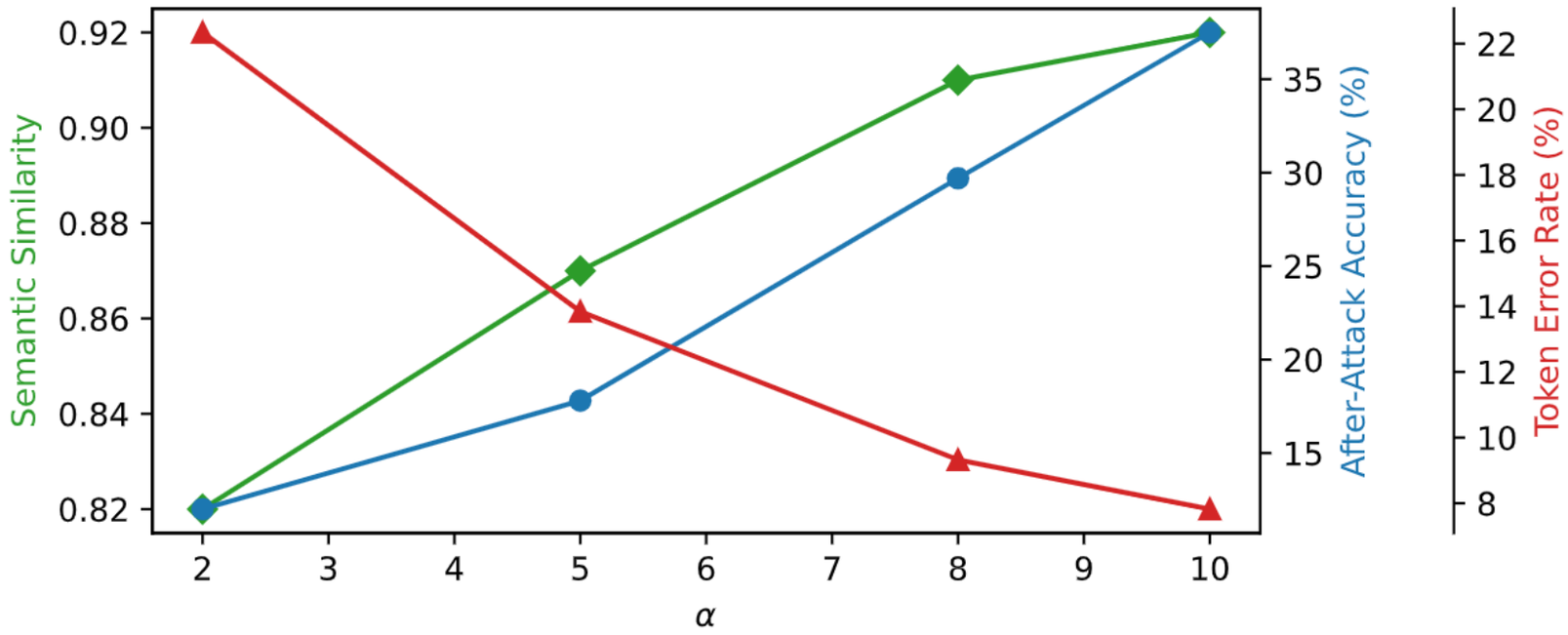
Guo, et al. "Gradient-based Adversarial Attacks against Text Transformers." *EMNLP*, 2021.

Experimental Results

Performance of white-box attack against fine-tuned GPT-2

Method	Ag News		MNLI		Yelp Reviews	
	Adv. Acc.	Sim.	Adv. Acc.	Sim.	Adv. Acc.	Sim.
Proposed	0.4	0.87	3.0 (1.3)	0.85 (0.82)	1.8	0.87
GBDA	6.6	0.90	2.8 (11.0)	0.82 (0.88)	2.9	0.94

Ablation Study



Effect of the hyper-parameter α on the performance of our attack.

Examples

Dataset	Sentence	Prediction	Text
MNLI	Original	Neutral (97.26%)	Premise: In the summer, the Sultan's Pool, a vast outdoor amphitheatre, stages rock concerts or other big-name events. Hypothesis: **Most** rock concerts take place in the Sultan's Pool amphitheatre.
	Adversarial	Entailment (99.19%)	Premise: In the summer, the Sultan's Pool, a vast outdoor amphitheatre, stages rock concerts or other big-name events. Hypothesis: **Many** rock concerts take place in the Sultan's Pool amphitheatre.
AG News	Original	Sci/Tech (99.39%)	Motorola and HP in **Linux** tie-up Motorola plans to sell mobile phone network equipment that uses **Linux**-based code, a step forward in network gear makers #39; efforts to rally around a standard.
	Adversarial	Business (83.56%)	Motorola and HP in **PC** tie-up Motorola plans to sell mobile phone network equipment that uses **PC**-based code. a step forward in network gear makers #39; efforts to rally around a standard.
Yelp	Original	Negative (99.90%)	This place holds a nostalgic appeal for people born and raised in Pittsburgh who grew up eating here. If that experience is what your looking for, please visit. If you're looking for a tasty meal, go somewhere else. 5 stars for history, **0** for food quality and flavor.
	Adversarial	Positive (96.54%)	This place holds a nostalgic appeal for people born and raised in Pittsburgh who grew up eating here. If that experience is what your looking for, please visit. If you're looking for a tasty meal, go somewhere else. 5 stars for history,**1** for food quality and flavor.

Conclusion

- Optimization problem to generate adversarial examples
  - Block sparsity constraint to ensure few tokens are modified
  - Solve by Gradient projection
- *Comparable* Performance with GBDA
  - Drops the classification accuracy to *less than 5%*
  - Semantic similarity is *more than 80%*
- **Source code:** https://github.com/sssadrizadeh/transformer-text-classifier-attack

Thanks for Your Attention

[Source Code]

Sahar Sadrizadeh

sahar.sadrizadeh@epfl.ch

Signal Processing Laboratory (LTS4), EPFL

EPFL

icassp 2022 Singapore