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
Emails: sahar.sadrizadeh@epfl.ch, ljiljana.dolamic@armasuisse.ch, pascal.frossard@epfl.ch

icassp 2022 Singapore
Paper ID: 8691

Introduction

Recently, it has been shown that, in spite of the significant performance of deep neural networks in different fields, those are vulnerable to adversarial examples. Recently, many methods have been proposed to generate adversarial examples in image data to make the systems fail, but these methods cannot be directly extended to NLP systems as:

- Input data is *discrete*.
- Definition of the *perturbation imperceptibility* is different

Problem Formulation

Consider $f: \mathcal{X} \rightarrow \mathcal{Y}$ to be the target classifier which correctly predicts the class of the input sentence $\mathbf{x}$ to be $y = f(\mathbf{x})$. Every sentence is considered to be tokenized to a sequence of tokens $\mathbf{x} = x_1 x_2 \dots x_n$. We are looking for an adversarial example $\mathbf{x}'$, which fools the target classifier and differs from the input sentence $\mathbf{x}$ in only a few tokens. We transform each token to a continuous embedding vector and represent the sentence as a sequence of embedding vectors: $\mathbf{e_x} = [\text{emb}(x_1), \text{emb}(x_2), \dots, \text{emb}(x_n)]$. Similarly, we denote the adversarial example as $\mathbf{e_{x'}} = \mathbf{e_x} + \mathbf{r_x}$ in the embedding space. In order to fool the model, we can find an adversarial example by minimizing the negative of the loss function of the classifier:

$$\mathcal{L}_{Adv} = -\mathcal{L}_f(\mathbf{e_{x'}}, y)$$

However, we want to modify only a few tokens of the input sentence. Therefore, some blocks of $\mathbf{r_x}$ that correspond to the modified tokens are non-zero, while others are zero, which means $\mathbf{r_x}$ should be block-sparse. Therefore, we need to minimize the following loss term:

$$\mathcal{L}_{BSparse} = \sum_{i=1}^{n} \|\mathbf{r}_i\|_2.$$

Therefore, our objective is to find the block-sparse perturbation that fool the target classifier by solving the following optimization problem:

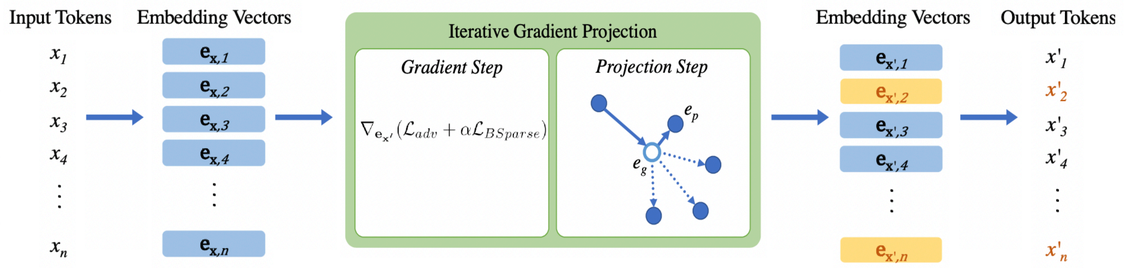
$$\hat{\mathbf{e}}_{\mathbf{x}'} = \underset{\mathbf{e}_{\mathbf{x}'} \in \mathcal{E}_{\mathcal{V}}}{\text{argmin}} \; \mathcal{L}_{Adv} + \alpha \mathcal{L}_{BSparse},$$

Proposed Method

Since we are dealing with textual data, the proposed optimization problem is discrete. In other words, the tokens of the resultant adversarial example should be in the vocabulary set $\mathcal{V}$; hence $\mathbf{e}_{\mathbf{x}'}$ should be in the discrete subspace $\mathcal{E}_{\mathcal{V}}$. We proposed to use gradient projection to solve the optimization problem as the following algorithm:

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, $y \leftarrow f(\mathbf{x})$, $k \leftarrow 0$
    $\forall i \in \{1, \dots n\}$ $\mathbf{e}_{\mathbf{g},i} \leftarrow \text{emb}(x_i)$
  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:
        $\mathbf{e_g} \leftarrow \mathbf{e_g} - lr.\nabla_{\mathbf{e}_{\mathbf{x}'}}(\mathcal{L}_{adv} + \alpha \mathcal{L}_{BSparse})$
      Step 2: Projection to the discrete subspace $\mathcal{E}_{\mathcal{V}}$ and update if the sentence is new:
        for $i \in \{1, \dots, n\}$ **do**
          $\mathbf{e_{p},i} \leftarrow \underset{\mathbf{e} \in \mathcal{E}_{\mathcal{V}}}{\text{argmin}} \frac{\mathbf{e}^\top \mathbf{e}_{\mathbf{g},i}}{\|\mathbf{e}\|_2 . \|\mathbf{e}_{\mathbf{g},i}\|_2}$
        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 $\mathbf{e}_{\mathbf{x}'} \leftarrow \mathbf{e_p}$
end procedure



Block diagram of the proposed method

Experimental Results

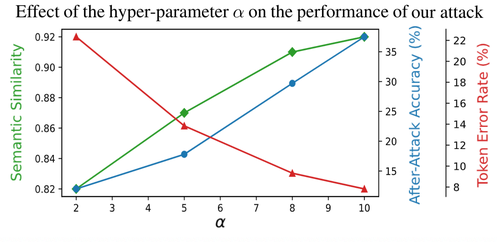
- **Target model:** GPT-2 Transformer
  - MNLI (natural language inference),
  - AG News (news categorization),
  - Yelp Reviews (sentiment classification).
- **Baseline:** GBDA [1] (optimization-based white-box attack against transformers)

[Source Code]

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

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



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.

Reference:

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