On Adversarial Examples for Text Classification 👿 By Perturbing Latent Representations

Korn Sooksatra, Bikram Khanal, and Pablo Rivas

Department of Computer Science, School of Engineering and Computer Science, Baylor University, Texas, USA.

Contributions

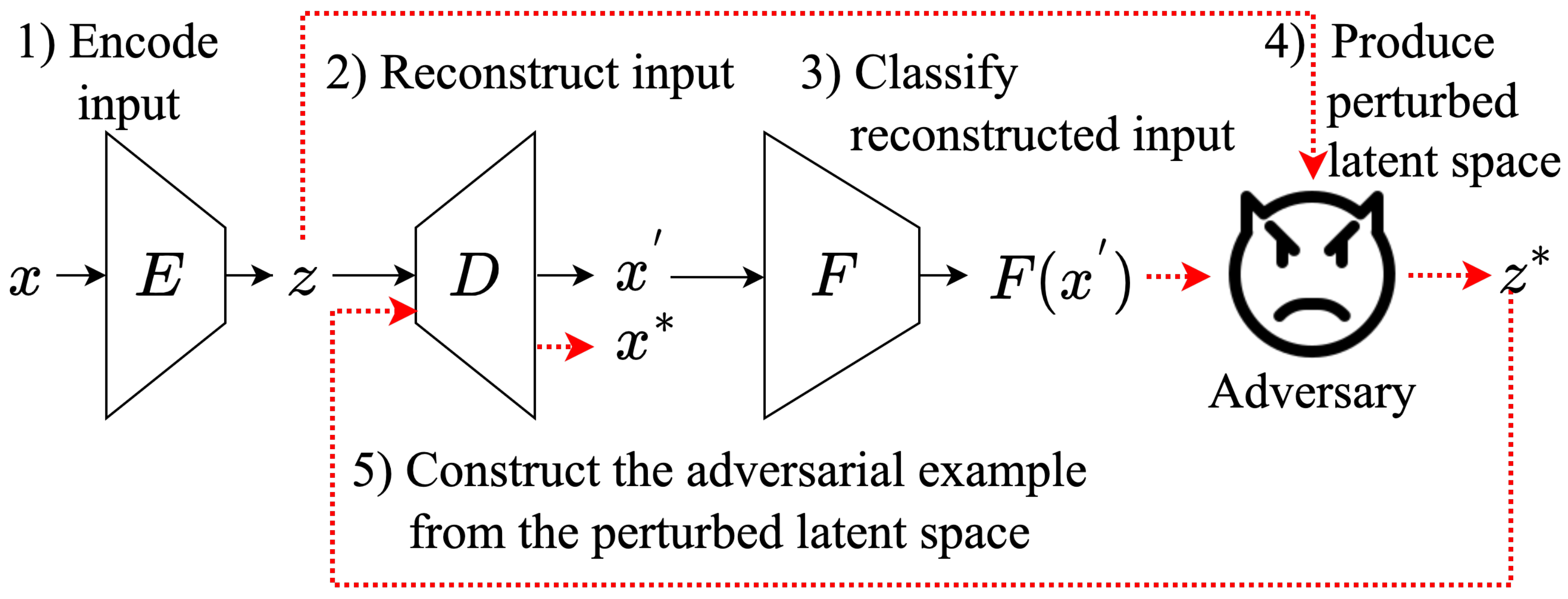
- We implement the encoder and decoder and their training scheme that can generate embedding vectors for a specific task.
- Our approach is among the first that applies a white-box adversarial attack on the embedding vectors of texts to generate adversarial examples.
- We extensively construct experiments showing that our approach can produce natural adversarial examples.

Problem Formulation

Given a text classifier C (e.g., a sentiment analyzer and a news-type classifier), our goal is to evaluate the C by finding misclassified samples (adversarial examples). We compute small perturbations δ and add them to an input of C such that the prediction is not the same as its ground-truth class. That is, given an input x and its ground-truth class y, we compute δ such that $C(x) = y$ and $C(x^*) \neq y$ where C predicts the class of x and $x^* = x + \delta$. However, the inputs are discrete in C; hence any changes to the input are obvious. Therefore, we find an embedding vector of x and compute δ instead. Then, we transform the perturbed embedding vector back to text. The next section will explain the mechanism to transform a text to an embedding vector and vice versa to guarantee that the text and its reconstruction belong to the same semantic.

Approach

Our approach consists of three main components: an encoder (i.e., E), a decoder (i.e., D) and an adversary (e.g., Fast Gradient Sign Method (FGSM) or Projected Gradient Descent (PGD)). The targeted classifier is denoted by F whose output is a vector of conference scores.


1) Encode input
2) Reconstruct input
3) Classify reconstructed input
4) Produce perturbed latent space
x
E
z
D
x'
x^*
F
$F(x')$
Adversary
z^*
5) Construct the adversarial example from the perturbed latent space

Training Encoder and Decoder

The training scheme consists of the encoder, the decoder and a small classifier (i.e., c) (not the target). The loss function for training is

$$R(X, X') + \lambda L(c(E(X)), Y_X),$$

where X is a batch of training set, Y_X is the corresponding labels, $R(\cdot,\cdot)$ is a reconstruction loss, $L(\cdot,\cdot)$ is the cross-entropy loss and λ is a balancer.

Experimental Setup

- **Dataset:** We use Ag-News dataset in this experiment. It has four classes: consists of World (W), Sport (S), Business (B) and Science/Technology (S/T).
- **Encoder and Decoder:** We choose a pretrained BERT as our encoder and two layers of LSTM as our decoder.
- **Target:** We use two layers of LSTM as our targeted Ag-news classifier.

Results

Encoder and Decoder:

Original text (Class S): us cyclists capture three medals athens , greece - tyler hamilton # 39 ; s greatest ride capped the finest olympic day for us cycling , which won three of the six medals awarded in wednesday # 39 ; s road time trials - surpassing its two total road medals won since the 1984 games in los . . .

Reconstruction (Class S): cricket : aussies crowing but india # 39 ; s grip on stump india # 39 ; s cricket board praiseds shane warne on monday as the first test against australia captain nagpur was the buttreded his team # 39 ; s chances for a test against australia .

Our Approach:

(Class S/T $\longrightarrow$ W) u . s . to share funds for more (ap) ap - the nation ' s top education department is planning to raise a new government research program in 2005 and plans to begin issuing new and ~~negative effects on the scale of the nation ' s biggest cities .~~more popular voting machines in the united states .

(Class S $\longrightarrow$ W) astros beat rockies to win nl playoff spot houston (reuters) - the houston astros have ~~picked up their first playoff berth in five years , their first big one - day winning streak in a season - clinching victory , the houston astros made the playoffs finale for their 13th straight year .~~ found a huge win over the houston astros with a huge win on their national league championship series at the houston astros .

(Class B $\longrightarrow$ S/T) google shares surge in debut on market share shares of google , the internet search engine , said its first - half profit rose 39 percent ~~, boosted by strong results in its international business .~~ as it priced its online rental market .

Conclusion and Limitation

- Our approach can produce adversarial examples from latent representations of texts.
- Although our encoder and decoder can produce a reconstructed text belonging in the same class as the input, we still need them to be visually similar with each other. We find a solution train them more efficiently.
- We do not know what a good perturbation bound in the latent space should be. Thus, it is a hyperparameter that we need to tune.