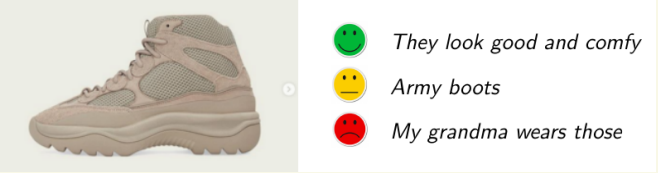
Adversarial Training for Aspect-Based Sentiment Analysis with BERT

Akbar Karimi, Leonardo Rossi, Andrea Prati
IMP Lab, University of Parma, Italy

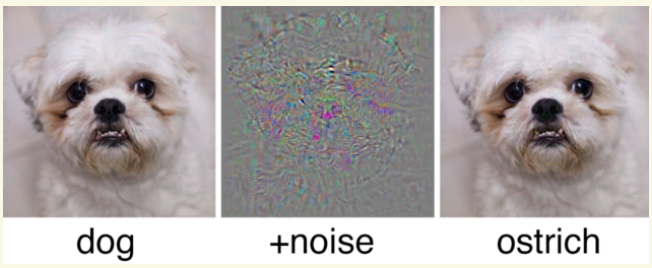
Introduction

Aspect-Based Sentiment Analysis (ABSA) deals with user opinions on a company's products or services. In this paper, we address two major problems in ABSA. The first one is Aspect Sentiment Classification (ASC). In ASC, we want to understand how a consumer feels about a particular product or service. We can classify the user's attitudes into 3 classes, namely positive, neutral, and negative. The second problem is Aspect Extraction (AE). In AE, the goal is to extract the aspects of a product or service towards which a sentiment is expressed. In order to address these two problems, we train a sentiment analysis model using BERT [1] in an adversarial manner.



Adversarial examples for images

These examples are created by adding some noise to the original images. While the result is still the same image to the human eye, strangely enough, neural networks are fooled by them and classify them as completely different entities. [3]



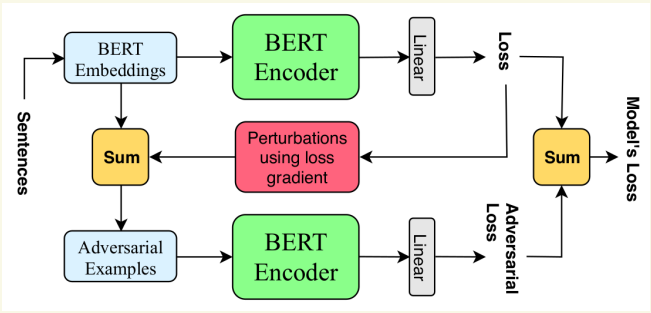
Adversarial examples for text

Creating adversarial examples are easy for images since the pixel values are continuous. But for text, which is discrete, a different approach is used. Here, we modify the word embeddings. The way this is done is that after computing the value for loss function, its gradient is calculated. Using the gradient and epsilon, which is used to specify how big the change should be, adversarial examples are created from the original ones. [2]

$$x = x - \epsilon \frac{g}{\|g\|} \tag{1}$$

x: input embedding
ϵ: size of perturbation
g: gradient of loss w.r.t. x

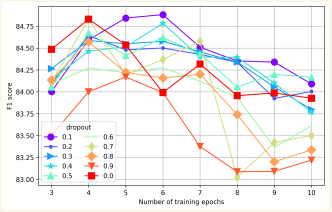
Model

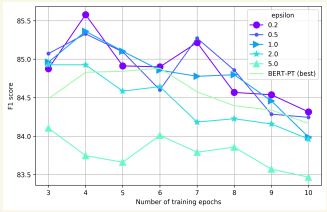


We use the BERT architecture in this work. As you can see, the model can be divided into two paths. One is the upper path where we calculate the network loss on real sentences. Here, after computing the BERT embeddings for the sentences, they go through the BERT encoder which consists of 12 transformer layers. Then a classification layer is applied and we get the loss on original examples. The gradient of this loss with respect to the real inputs is then calculated and using the Formula 1, adversarial examples are created. Then these artificial samples go through the BERT encoder again and the adversarial loss is calculated. In order to the get the total loss of the network, the two losses are summed up.

Experiments

We first carry out a hyperparamter search for BERT-PT and then perturbation search for BAT. For instance, on the left, F1 scores for different training epochs and dropout values for BERT-PT are presented. On the right, the perturbation search for BAT is reported. The figures are only for aspect extraction task on laptop data set. More experimental results can be found in the paper.





Results

Domain	Laptop	Restaurant
Methods	F1	F1
BERT-base [1]	79.28	74.1
BERT-PT [4]	84.26	77.97
BERT-PT (best)	**84.88**	**80.69**
BAT (Ours)	**85.57**	**81.50**

Table 1: Aspect Extraction (AE)

Domain	Laptop	Restaurant
Methods	Acc	Acc
BERT-base [1]	75.29	81.54
BERT-PT [4]	78.08	84.95
BERT-PT (best)	**78.89**	**85.92**
BAT (Ours)	**79.35**	**86.03**

Table 2: Aspect Sentiment Classification (ASC)

References

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[2] Takeru Miyato, Andrew M Dai, and Ian Goodfellow. Adversarial training methods for semi-supervised text classification. *arXiv preprint arXiv:1605.07725*, 2016.

[3] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna Estrach, Dumitru Erhan, Ian Goodfellow, and Robert Fergus. Intriguing properties of neural networks. In *2nd International Conference on Learning Representations, ICLR 2014*, 2014.

[4] Hu Xu, Bing Liu, Lei Shu, and Philip S Yu. Bert post-training for review reading comprehension and aspect-based sentiment analysis. *arXiv preprint arXiv:1904.02232*, 2019.

Akbar Karimi
akbar.karimi@unipr.it
University of Parma