IE643: Deep Learning : Theory and Practice **Jul-Nov 2019**

End-term Project Report : Is BERT Really Robust?

Student Name: Vamsi, Pavan, Rukmangadh *Roll No: 160050064, 160050076, 160070051*

Abstract

In this paper, we propose two new improvements to the existing baseline attacks on BERT. Prior work on adversarial attacks over BERT have relied on a Black-box setting without utilizing any specific architectural knowledge. We alleviate this through novel attention based pruning of heuristic space and thereby improving the performance over the existing work.

1 Introduction

Machine learning algorithms are often vulnerable to adversarial examples that have imperceptible alterations from the original counterparts but can fool the state-of-the-art models. These adversarial examples are helpful to evaluate the robustness of a model and to expose malicious statistical cues learnt by the model, if any. We can even improve the robustness of these models by exposing the maliciously crafted adversarial examples. This can be done so by collecting all the examples in a 'adversarial dataset' and training the model on this dataset.

2 Literature Survey

2.1 BERT: State-of-the-Art Pre-training for NLP

BERT primarily involves two steps: *pre-training* and *fine-tuning* (see 1). During pre-training, the model is trained over unlabeled corpus like Wikipedia over a set of pre-training tasks. During fine-tuning for a specific task, the model is initialized with the pre-trained parameters and fine-tuned during training possibly with the help of an additional smaller auxiliary model.

We can see that this two-step procedure goes as an analogy for transfer-learning, specifically those which exist in vision-related tasks. Also, this enables a uniform base architecture for a multitude of tasks and minimal additional parameters for training over specific tasks.

The paper analyzes the performance of two model sizes for BERT:

- BERT BASE: A smaller version of BERT comparable in size to the OpenAI Transformer for legitimate comparison.
- BERT LARGE: A huge model to achieve state of the art results as reported in the paper.

BERT is essentially a trained Transformer Encoder stack from the transformer models as described in [3]. BERT involves a number of

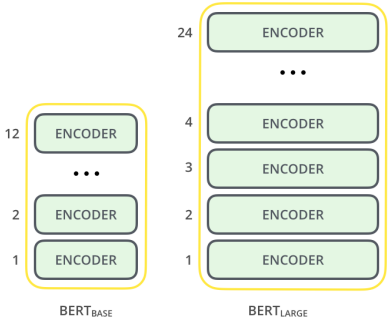

Figure 2: BERT-BASE and BERT-LARGE. [Source [1]]

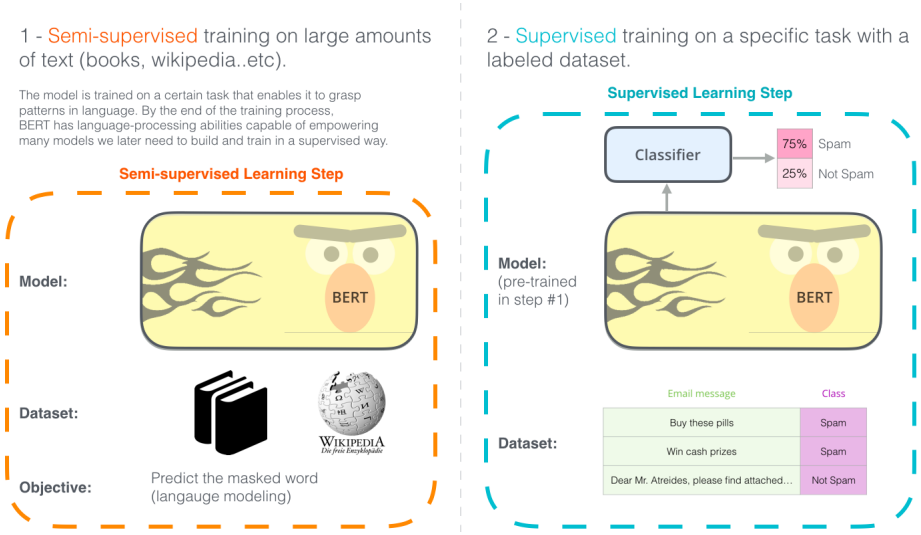

Figure 1: The two-step process involved in tuning BERT for a specific task. [Source [1]]

encoder layers (referred to as Transformer Blocks in the paper) depending on BASE/LARGE version and much larger feed-forward layers, more attention heads than those in [3].

2.1.1 Input Representations

The first token of every sequence is always a special token ([CLS]). CLS here stands for Classification. The final hidden state corresponding to this token is used as the aggregate sequence representation for classification tasks. In terms of architecture, this has been very similar to the Transformer in [3].

Also in input, sentence pairs are packed together into a single sequence and distinguished using a [SEP] token. A learned embedding called 'segment embedding' is added with denotes which sentence that token belongs to. Position embedding is added similar to the original transformer paper except that here this embedding is not sinusoidal but also learned like segment and token embeddings.

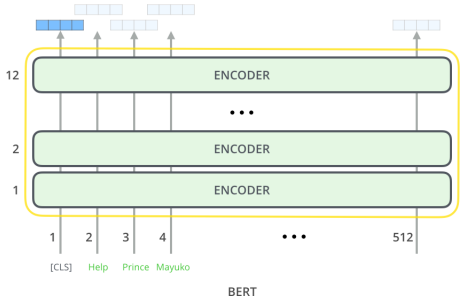

Figure 3: Input and output in BERT. [Source [1]]

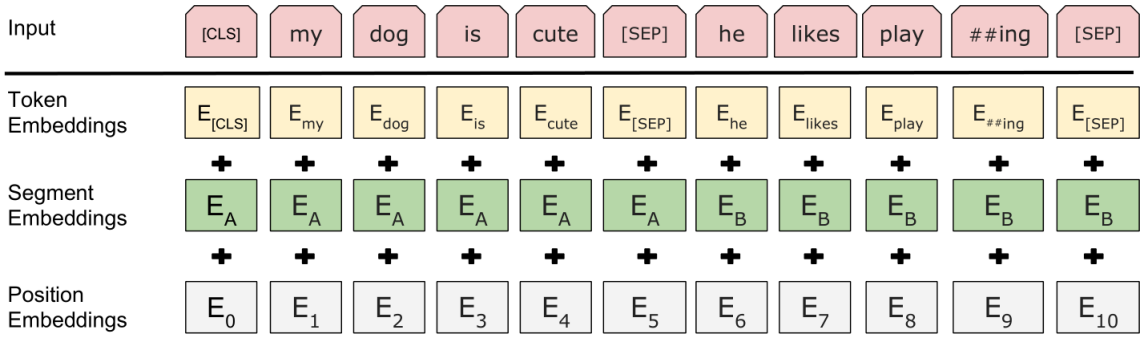

Figure 4: Input representation in BERT. [Source [2]]

2.1.2 Pre-training methods

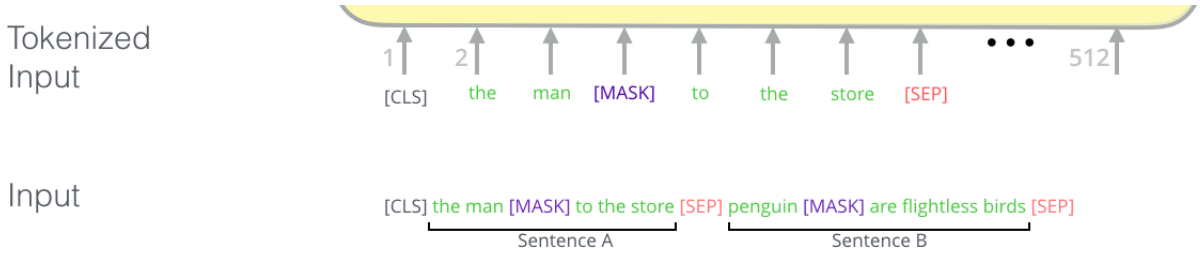

Figure 5: MLM and Two-Sentence tasks in BERT. [Source [1]]

The authors suggest that a deep bidirectional model is strictly more powerful than a uni-directional or a shallow concatenation of left-to-right and right-to-left models like in ELMo. In order to train a deep bidirectional model, BERT adopts a "masked language model" concept from earlier literature (where it's called a Cloze task). Beyond masking 15% of the input, BERT also mixes things a bit in order to improve how the model later fine-tunes. Sometimes it randomly replaces a word with another word and asks the model to predict the correct word in that position.

To make BERT better at handling relationships between multiple sentences, the pre-training process includes an additional task: Given two sentences (A and B), is B likely to be the sentence that follows A, or not? This is called the NSP (next-sentence prediction) pretraining task.

2.1.3 BERT for feature extraction

In addition to fine-tuning the model for a specific task, pretrained BERT can also be used as a feature extractor, specifically in tasks where a task-specific model architecture is unavoidable. Various ablation studies presented in the paper emperically determine the best strategy for each task.

2.2 TextFooler: Is BERT Really Robust?

This paper talks about TextFooler, a simple black-box algorithm that generates semantically meaningful adversarial examples. Also, this algorithm is computationally simple and is O(n) in the length of the text to be generated.

Black Box Setting: In a black box setting, the attacker is not aware of the model architecture, parameters, or training data. It can only query the target model with supplied inputs, getting as results the predictions and corresponding confidence scores.

2.2.1 Algorithm

This algorithm has two main steps:

- **Step 1: Word Importance Ranking**
- **Step 2: Word Transformer**

Step 1: Word Importance Ranking: It has been observed that only some key words act as influential signals for the prediction. Using this fact, this paper tries to rank the words in the sentence by their importance in contribution to the prediction. To find the importance of the word, the paper use the following metric:

$$I_{w_i} = \begin{cases} F_Y(X) - F_Y(X_{\setminus w_i}) & if\ F(X) = F(X_{\setminus w_i}) = Y \\ F_Y(X) - F_Y(X_{\setminus w_i}) & if\ F(X) = Y\ F(X_{\setminus w_i}) = \bar{Y} and\ Y \neq \bar{Y} \end{cases} \tag{1}$$

Where we represent the model by the function $F(.)$, the confidence for an output Y by $F_Y(.)$ and $X_{\setminus w_i}$ represents the example after removing word w_i from the sentence X.

Step 2: Word Transformer: This step is applied in the order of the importance of the words until either an adversarial example is generated or all the words of the sentence have been transformed. So, for a given word w_i in X the word transformation mechanism need to satisfy the following criteria:

- Have similar semantic meaning with the original one
- Fit within the surrounding context
- Force the target model to make wrong predictions

Keeping these in mind the paper proposes the following sub-steps for achieving step-2 of the algorithm:

1. *Synonym Extraction*: In this step we collect a set of closest N synonyms for the given word w_i. Lets call this set $CANDIDATES$
2. *Parts Of Speech(POS) Checking*: We perform parts of speech checking on each member of $CANDIDATES$ and remove those words whose POS is different from that of w_i. This new set is $POS(CANDIDATES)$
3. *Semantic Similarity Checking*: We replace the word w_i in X with a member from $POS(CANDIDATES)$ and get X_{adv}. We use the Universal Sentence Encoder to encode these two sentences into vectors and take their cosine similarity. This is done for each member. We threshold the similarity and get the final set of replacement words for w_i called as $FINCANDIDATES$
4. *Finalization of Adversarial Examples*: We finally compute $F_Y(X_adv)$ and compare it with $F_Y(X)$ for all members of $FINCANDIDATES$ and choose the one that causes the highest disparity.

3 Attack Environment

Task: For the purpose of our experiments we attack the task of binary classification of movie reviews as positive or negative. This is a simple enough task to perform initial testing before moving onto more complex classification tasks. We use the **IMDB reviews** dataset that contains more than around 50,000 review of various movies, tv shows, documentaries...etc.

Model: We attack the pre-trained BERT + Fine-tuned Sequence Classification Head model.

4 Methods and Experiments

InferSent is a sentence embeddings method that provides semantic representations for English sentences. InferSent trained by Facebook Inc. on a large corpus of unlabelled data is more popular and up to date than Universal Sentence Encoder (USE). InferSent uses a BiLSTM to encode a sentence to a 4096 dimensional embedding. So we replaced the USE embeddings used in the implementation with InferSent.

Model	Time	Avg. Change Rate	Success Rate
Original	4381.607 sec	6.013 %	73.600 %
+ InferSent	3102.862 sec	5.794 %	84.333 %

The way the importance of a word in the context of a sentence is found out in a different and indirect way. We are making the attack specific to BERT, so we can directly use the attention values from encoder layers of BERT to determine importance of input tokens. This makes the algorithm to avoid the need to forward every leave-one token sentence to BERT.

5 Results

Metrics:

Model	Time	Avg. Change Rate	Success Rate
Original	4381.607 sec	6.013 %	73.600 %
+ InferSent	3102.862 sec	5.794 %	84.333 %
+ Attention	2441.705 sec	7.571 %	84.333 %

Examples:

Original Text	Adversarial Text
i first seen this movie in the early 80s and we used to have it on betamax as we all know , betamax went the way of the 8 trak tape , sigh , it really had nice picture quality too anyways , i 'm glad i found this movie again , i 've been searching for it for more than 10 years ! this movie falls into the category of movies like airplane continuous jokes , oneliners , funny actions *bodylanguage* mark blankfield is absolutely hilarious his transformation from the shy dr daniel jekyll into the sex crazed partyanimal mr hyde is unforgettable , complete with goldtooth , chesthair and goldchains the part i loved best was when he hijacked the car from this poor guy and then drove to madam woo woo 's totally psychedelic experience without the drugs ! if you need laugh therapy this is the movie to do it when i first seen it , i had tears in my eyes and my belly was hurting from constantly laughing this is a movie i could watch over and over again i highly recommend it	i first seen this movie in the early 80s and we used to have it on betamax as we all know , betamax went the way of the 8 trak tape , sigh , it really had nice picture quality too anyways , i 'm happier i found this flick again , i 've been searching for it for more than 10 years ! this movie falls into the genus of movies like airplane continuous jokes , oneliners , funny actions *bodylanguage* mark blankfield is absolutely hilarious his transformation from the shy dr daniel jekyll into the sex crazed partyanimal mr hyde is unforgettable , complete with goldtooth , chesthair and goldchains the part i loved best was when he hijacked the car from this poor guy and then drove to madam woo woo 's totally psychedelic experience without the drugs ! if you need laugh therapy this is the movie to do it when i first seen it , i had tears in my eyes and my belly was hurting from constantly laughing this is a movie i could watch over and over again i highly suggest it

Original Text	Adversarial Text
when i voted my 1 for this film i noticed that 75 people voted the same out of 146 total votes that means that half the people that voted for this film feel it 's truly terrible i saw this not long ago at a film festival and i was really unimpressed by it 's poor execution the cinematography is unwatchable , the sound is bad , the story is cut and pasted from many other movies , and the acting is dreadful this movie is basically a poor rip off of three other films no wonder this was never released in the usa	when i polling my 1 for this kino i noticed that 75 people voted the same out of 146 omnibus votes that means that half the people that voted for this theatres feel it 's really frightening i enjoyed this not anymore ago at a film festival and i was vitally skeptical by it 's imperfect implemented the cinematography is unwatchable , the sound is bad , the legend is haircuts and glued from umpteen other movies , and the caretaker is scary this movie is crucially a vulnerable buzzed off of two other panorama no blitz this was never dispensed in the unidos

Original Text	Adversarial Text
i really enjoyed this movie and i usually do n't like animated pictures but i thought the cats were appealing and the story line was charming there is a good song called everybody wants to be a cat , that is a lot of fun it has some comic moments and is an interesting adventure i think it helps to be an avid cat lover to enjoy this film	i really rained this stills and i usually do n't like stimulated pictures but i thought the cats were appealing and the story line was charming there is a good song called everybody wants to be a cat , that is a lot of fun it has some comic moments and is an interesting adventure i think it helps to be an avid cat lover to enjoy this film

6 Future Work

We developed on the given heuristic model, getting better and faster results using some basic optimizations. The task of finding semantically equivalent and grammatically correct adversaries is better modelled by a neural net rather than just a heuristic search. Using a neural network can alleviate the search space and time.

The attention values from BERT can be incorporated into the original black box model in a better way to fully exploit the BERT model. Instead of selecting one word with the least confidence score, a top-k beam search kind of algorithm can be used. Another possible line of work is develop defences against the exploited bugs.

7 Conclusion

Although we did exploit BERT model to get the attention values, stronger model which attack BERT specifically can be developed. The discrete nature of input is always a difficult problem while creating models which generate adversaries for NLP based tasks.

The speed of the heuristic model is about 6 sentences per minute when done on a server with Xeon Gold 5120 (CPU) + GTX 1080 Ti (GPU). This is quite slow, considering that this is just a classification task.

We anticipate that the suggestions in future work will perhaps increase the performance of the attacks.

References

[1] Jay Alammar. The Illustrated BERT. `https://jalammar.github.io/illustrated-bert/`, 2018. [Online; accessed 06-October-2019].

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

[3] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017.