BERT for Opinion Mining and Sentiment Farming

Arti Buche\(^1\), and M. B. Chandak\(^2\)

\(^1\)Shri Ramdeobaba College of Engineering and Management, University Nagpur India
\(^2\)Shri Ramdeobaba College of Engineering and Management, University Nagpur India

**ABSTRACT**
BERT which is published by Google AI language stands for Bidirectional Encoder Representations from Transformers is a language representation model recently developed has caused a flurry in Machine learning community. The design of BERT consists of pre-training the unlabeled text contextually from both the directions left and right for all the layers, unlike the recent language representation models. For a BERT model which is pre-trained, inclusion of one output layer to the state-of-the-art models for fine tuning is done which facilitates different tasks for language understanding such as opinion mining and question answering so that the task specific architectures are not to be modified. BERT is a key technical innovation which applies bidirectional training of a popular attention model called as Transformer language modelling. The results in the paper represent a language model BERT in which training is done bi-directionally and a deeper sense of context for the language understanding is developed. In the paper, a detailed description of a novel technique of BERT in which training involves Masked Language Model (MLM) and Next Sentence Prediction (NSP) bi-directionally which was previously impossible. [Jacob Devlin, 2019]

**KEY WORDS:** BERT, BIDIRECTIONAL LANGUAGE MODEL, MASKED LANGUAGE MODEL, NEXT SENTENCE PREDICTION, OPINION MINING.

**INTRODUCTION**
Natural Language Processing (NLP) applications are ubiquitous these days as they stumble across various websites and applications in one form or another. NLP adoption has rapidly increased thanking to the concept of pre-trained models embedded with transfer learning. Pre-training of language model has shown impressive results for natural language processing tasks. People interested in developing or learning an algorithm on existing framework, pre-trained models are the best source of help for them. It is not possible to build a model from scratch always due to computational constraints or time restriction, that is why, pre-trained models came into picture. The pre-trained models can be used as a benchmark for either improvement of the existing model or testing the developed model against it.

The pre-trained language representations can be applied using two existing strategies feature based and fine tuning to downstream tasks. ELMo is the feature based approach which uses task specific architectures and which are included in pre-trained as additional features. openAI GPT which is Generative Pre-Trained Transformer is the fine-tuning approach in which the downstream tasks are used for training purpose and all pre-trained parameters for fine tuning and it uses minimal task-specific parameters. The above strategies share the same objective of using unidirectional language models during pre-training to learn general language representations. [Jacob Devlin, 2019]. The point of contend is that the current techniques curb the capacity of pre-trained models, especially for the fine-tune approaches.
The dominant limitation during pre-training is that the choice of architectures and standard language models which are unidirectional. For example, in OpenAI GPT which consists of self-attention layers, and only in these layers, the researchers have used left to right architecture which provides accessibility of every token to the previous token. The sentence level tasks which consists of such limitation are very much sub-optimal which in turn can be proved harmful during application of the approaches which are fine tuned to token level tasks like question answering task. In such cases it is crucial to incorporate understanding of language context from both the directions.

In this paper, the improvement of the approaches based on fine tuning is done with the help of BERT (Bidirectional Encoder Representations from Transformers. BERT is the key technical innovation in the field of NLP marking it as a beginning of new era. It is a momentous development as it enables to build a powerhouse as a component readily available to use. This innovation has many advantages like it saves time, energy, knowledge and the resources that require training of a language model from scratch. BERT has been built on number of clever ideas which are bubbling up recently in NLP community. The contributions of the paper are as follows:

- BERT alleviates the constraint of unidirectionality by using Masked Language Model (MLM). The demonstration of the importance of bidirectional pre-training for language models which is used by BERT and recent work in pre-training contextual representations is shown.
- Heavily engineered task specific architectures can be reduced with the help pre-trained representations is shown. Many task specific architectures outperform with the help of BERT as it is the first representation model which is fine tune based has achieved state-of-art on token level and sentence level tasks on a large suite.
- The main goal is to help researchers with fewer computational resources and encourage the community to build the gap in large amount of data which improve when trained on millions of annotated training examples and variety techniques that are used for training which is called as pre-training concept for general purpose language representation models using unannotated text in a large amount. [Jacob Devlin, 2018]

Related Work: Unsupervised feature based Approaches: An active research area for decades has been widely learning representation of words including non-neural and neural methods. An integral part of modern NLP are pre-trained word embeddings which has shown much improvement significantly over the embeddings which were learned from scratch. The objectives in pre-training word embedding s in language model included left to right training and differentiate between correct and incorrect words in both left and right context. The coarser granularities of the unsupervised feature-based approaches has been generalized into sentence embeddings or paragraph embeddings. The training objectives which were used for sentence representation includes prior work of ranking candidate next sentences, previous sentence representation with the help of left to right generation or by using auto encoder for denoising derived objectives. A step further is ELMo (Embeddings from Language Models) which trains LSTM (Long short-term memory) bi-directionally resulting into a language model which has sense of next as well as previous word.

ELMo performs generalization of traditional word embeddings form different dimension. This language model mainly represents each token based on its context i.e. concatenate representations from left to right and right to left. The existing task-specific architectures integrate these contextual word embeddings which advances ELMo for several important NLP benchmarks including opinion mining, question answering and named entity recognition. In [Oren Melamud, 2016] contextual representations were learned using LSTMs in which prediction of a single word was done through a task for both the contexts left and right. One such model is ELMo which is also feature based like LSTMs but it is not deeply bidirectional.

Unsupervised Fine-tuning Approaches: The feature-based approaches work in direction of pre-training word embeddings from unlabelled text. Recently, the contextual token produced by sentence or document encoders are pre-trained from unlabelled text and are further fine-tuned to a supervised downstream task. These approaches are proved to be an advantage as learning from scratch requires less parameters. Due to part of this advantage OPENAI GPT (Generative Pre-Trained Transformer) model achieved state-of-art results from GLUE (General Language Understanding Evaluation) benchmark on many sentence level tasks. The objectives used for pre-training by such models are auto-encoder and left to right modelling.

Transfer Learning from Supervised Data: Transfer Learning is the technique which enables the researchers and developers to use the pre-trained models developed by others by performing some modifications in it. The methods used for transfer learning in NLP are the main reason of its increased accuracy mainly in Opinion Mining. Transfer learning basically leverages the knowledge which is extracted from big datasets to increase efficiency even if limited training data is available. There are many works done in transfer learning which shows effective transfer of supervised tasks from large datasets such as machine translation and opinion mining summarization. An effective recipe of transfer learning has shown its importance in Computer Vision research also by fine tuning models which as pre-trained with ImageNet. [Lisa Torrey, 2009]

Proposed Work: BERT (Bidirectional Encoder Representations from Transformers): One of the latest milestones in Natural Language Processing marked as the beginning of a new
era of NLP is the release of BERT (Bidirectional Encoder Representations from Transformers). A new technique was open sourced in NLP for pre-training called as BERT. The core principle of BERT is to use attention modelling so that it understands relationship between different words contextually backed by Transformer. BERT’s vanilla form includes a Transformer which has two mechanisms of encoder and decoder for reading input text and outputs prediction for the task respectively. The encoder mechanism is the only requirement of BERT as the main goal is to generate language model. Transformer encoder takes input text and reads at once the entire sequence of words as opposed to the directional models (RNN and LSTM) which read the input text linearly left to right or from right to left. This technique is referred as Bidirectional more accurately non-directional as it allows the model to learn a word’s contextual information with respect to complete sentence (left as well as right of the word).

The detailed working of BERT is described as follows: The challenge in training of language models is to define a prediction of a sentence in a text. In a directional approach adopted by many models prediction of next word in a sequence is done but which lacks in context learning. So to overcome this challenge, BERT framework is divided into training steps: pre-training and fine-tuning. In pre-training step, the model undergoes training for different pre-training tasks using unlabeled text. During fine-tuning step, the pre-trained parameters of BERT model are first initialized followed by fine-tuning of all parameters using labelled data from the downstream tasks. Even if the parameters in BERT are firstly initialized with same pre-trained parameters, separate fine-tuned models are generated for each downstream task.

How does BERT work?: The attention mechanism which helps in learning the contextual relationship between words of sentence is the main concept of Transformer which BERT relies on completely. As discussed earlier BERT requires only encoder, as an encoder takes input a sequence of tokens which is followed by conversion into vectors, and further are being processed in the network. To perform this processing, the decoration of inputs is to be done with extra metadata as follows:

- **Token Embeddings**: A sentence is detected by using tokens namely [CLS] which is placed at starting of the sentence and [SEP] at the end.
- **Segment Embeddings**: To distinguish between separate sentences, a marker is used like Sentence A or Sentence B so that encoder understands.
- **Positional Embeddings**: A token is added to sentence in order to indicate the position of the words in the sentence which is called as positional embedding.

Example:

**Input** = Token Embeddings + Segment Embeddings + Position Embeddings

The transformer maintains a stack of layer in order to output a sequence as 1:1 corresponding to it input at the same index. Training in BERT makes usage of following two strategies:

**Masked Language Model (MLM)**: The main idea of masking is “simple”. It randomly masks 15% words in the input sentence i.e. replaces those words with [MASK] token. Afterwards the entire sequence is executed through attention based BERT in which the encoder predicts context based information for only the masked words provided by other non-masked words. In this naïve masking approach, the problem is that it predicts only [MASK] tokens instead of predicting correct tokens irrespective of the position of the token in the input sequence. Hence to deal with this issue, among the selected masked tokens of 15%, the tokens which are actually replaced with [MASK] token is 80%, replacement of with a random token is of 10% of time token and 10% among those remain unchanged. In the training phase of BERT, prediction of only masked tokens is done and unmasked tokens are ignored. Ultimately results for the model mingle slowly than left to right or vice versa.  

[Jacob Devlin, 2018]

**Next Sentence Prediction (NSP)**: BERT model also performs prediction of next sentences to understand the relationship between the two sentences. In next sentence prediction, the model which is under training takes input a pair sentences and learns to perform prediction of whether second sentence is related to previously sentence in the original text. As per the previously discussed model of BERT, the input consists of [SEP] token which helps in understanding the separate sentence. The input to the model is two sentences at a time such that 50% first sentence is followed by second one and 50% is a random sentence. In such situation where random sentence occurs, BERT predicts whether the second sentence among the pair of input sentences is random or not, along with the assumption that random sentences and the first sentence are disconnected with each other.

**Example**:

**Input** = [CLS][MASK]Opinion Mining is the science of using text analysis [SEP][MASK]Sentiment analysis is almost same thing.

**Label** = IsNext

Input = [CLS][MASK]Opinion Mining is the science of using text analysis [SEP] A [MASK]database is a collection of information that is organized
Label = NotNext
The basic working of BERT in prediction of whether first and second sentences are connected which each other or not, for which the input sequence needs to be passed through a transformer based model, in which transformation of the [CLS] token into a vector of shape 2x1 using a simple classification layer and with the help of softmax the NextLabel is assigned to it.

BERT model is trained by using both MLM and NSP in order to minimize the combined loss function, hence, “Together is better”. [Jacob Devlin, 2018]

BERT Architecture
BERT model architecture has four pre-trained versions of different types depending on its scale.
• BERT-Base: 12 layer, 768 hidden nodes, 12 attention heads, 110 M parameters
• BERT-Large: 21 layer, 1024 hidden nodes, 16 attention heads, 340 M parameters
• BERT is basically a multilayer bidirectional transformer encoder and requires task specific fine tuning

Flow of a word in BERT: By considering the above flow diagram, embedding layer starts with a word embedding representation. To create a new intermediate representation of the word, it performs some computation with some multi headed attention in every layer. The intermediate representations thus created are all of same size. In the above flow diagram, E1 represents the embedding representation i.e. position embeddings, intermediate representations of same token are represented by Trm and finally the output is represented by T1. In BERT model of 12 layers, a single token will have 12 intermediate representations as layers will be 12. [Matthew Peters, 2018].

Fine Tuning of BERT: A wide variety of task in general language understanding like opinion mining, natural language summarization, paraphrase detection, are outperformed the state-of-art by BERT. BERT fine-tuning for our own dataset is performed by addition of a single layer on crest of the base model. For example, one of language prediction task is question answering which involves a question as an input and the prediction of the right answer is from some corpus is to be identified. So, a context paragraph is given along with a question, the prediction includes a start and an end token of the paragraph which provides the likely answer to the given question. For such tasks BERT needs two extra vectors which help in marking the start and end of the answer as well. As in next sentence prediction training phase, input consists of two sentences similarly in this task of question answering first sentence is the given question and second sentence is the context paragraph along with two new parameters start and end vectors, created during fine-tuning. Note that in case of BERT fine tuning it is important to transform the input as per the format used during pre-training as discussed above.

Opinion Mining and Text Classification using BERT: Opinion Mining is one of the types of Natural Language Processing which helps in text classification and sentiment analysis. The main goal of Opinion Mining is to determine the polarity of the reviews or textual data as positive, negative or neutral. [Arti Buche, 2013] The steps to be followed for Opinion Mining and Text classification with the help of BERT are as follows:
• BERT model: Pre-trained BERT model is to be decided depending on the weights required for the corpus as discussed in above topic of BERT architecture.
• Preparation of Data: For using BERT model, we need to transform the input as per the format accepted by BERT. Data can be stored in a.tsv file format as given below:
  - 0 Column 0: ID of the row
  - 0 Column 1: Label of the row in integer format like 0,1,2,3...etc.
  - 0 Column 2: Same letter column for all called as throw away column expected by BERT
  - 0 Column 3: Classification column for test examples
• Training Model using Pre-trained BERT model: Transformation of the input text takes place as per requirement of the BERT with extra tokens as learned earlier. After training completes, an output is generated in which prediction results on test dataset containing probability value for the class labels can be seen.
• Fine tuning of the model: After the training is completed, plug in the task specific inputs and outputs are fine tuning of all the parameters is done from end to end. For the output, the token level tasks such as sequence tagging are fed to an output layer and [CLS] representation is being fed to an output layer for classification tasks such as opinion mining, text summarization. Fine tuning is relatively expensive as compared to pre-training of BERT.

CONCLUSION
A big milestone in the field of Natural Language Processing is BERT which is a powerful language representation model. It has great capacity to perform transfer learning in NLP. It also comes with a great promise to solve 11 different NLP tasks using the same pre-trained successfully. BERT demonstrates, due to transfer learning, rich and for many language understanding systems, unsupervised pre-training has become an integral part.
BERT can be used for prediction in various domains and one such important in which predictions also play a very important part is Stock Market Predictions using Opinion Mining which is trending nowadays.

**REFERENCES**


Hebatallah A. Mohamed Hassan, Giuseppe Sansonetti, Fabio Gasparetti, Alessandro Micarelli, Joeran Beel, “ACM RecSys 2019 Late-breaking Results”, Copenhagen, Denmark, 16th-20th September 2019


Websites:
https://gluon-nlp.mxnet.io/examples/sentence_embedding/bert.html
http://jalammar.github.io/illustrated-bert/