BERT and its Applications in Specific Text Processing Tasks

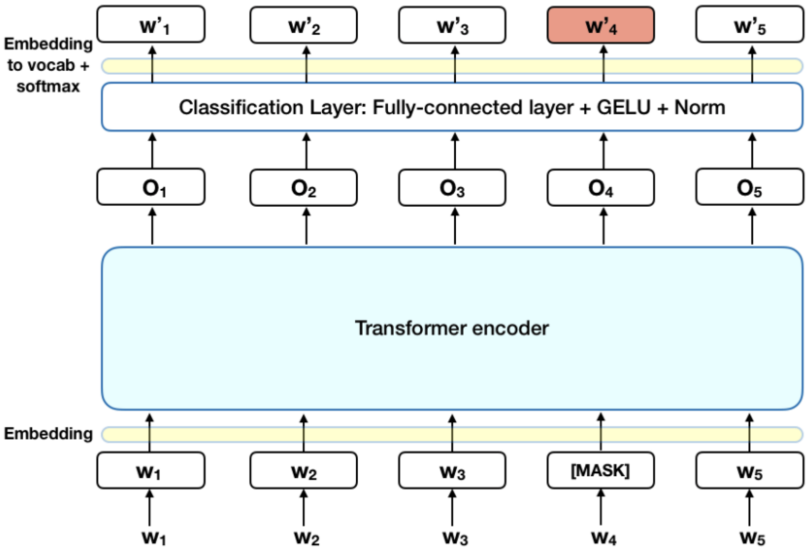
INTRODUCTION

BERT (Bidirectional Encoder Representations from Transformers) is an open source machine learning framework for Natural Language Processing (NLP). BERT is designed to help computers understand the meaning of ambiguous language in text by using surrounding text to establish context. The BERT framework was pre-trained using text from Wikipedia and can be fine-tuned with question and answer datasets.

HOW BERT WORKS

BERT makes use of Transformer, an attention mechanism that learns contextual relations between words (or sub-words) in a text. In its vanilla form, Transformer includes two separate mechanisms — an encoder that reads the text input and a decoder that produces a prediction for the task. Since BERT's goal is to generate a language model, only the encoder mechanism is necessary.

As opposed to directional models, which read the text input sequentially (left-to-right or right-to-left), the Transformer encoder reads the entire sequence of words at once. Therefore it is considered bidirectional. This characteristic allows the model to learn the context of a word based on all of its surroundings (left and right of the word). Below is a picture of its architecture.


w'1
w'2
w'3
w'4
w'5
Embedding to vocab + softmax
Classification Layer: Fully-connected layer + GELU + Norm
O1
O2
O3
O4
O5
Transformer encoder
Embedding
w1
w2
w3
[MASK]
w5
w1
w2
w3
w4
w5

APPROACHES TO APPLICATION OF BERT IN TEXT PROCESSING

When adapting BERT to specific word processing tasks, a special retraining technique is required. Such techniques can be of three types:

Further pre-training: BERT is initially trained on general-purpose texts, which may have a different data distribution than the texts of the target application area. The desire to retrain BERT on a specific text corpus can be based on an intra-task dataset (the same dataset that will be used to train the target model), an intra-subject dataset (a set of text data obtained from the same subject area) or a cross-subject dataset, in depending on the nature of the texts and the availability of data sets. Studies of the effectiveness of BERT retraining show that, first, retrained models show significantly better results compared to models without retraining, and, secondly, intra-subject learning is generally more effective than intra-task learning. Cross-subject learning does not significantly improve performance relative to the original BERT model, which is logical given that BERT is trained on a general set of texts.

Retraining Strategies: There are many ways to use BERT for a target. For example, you can use the representation provided by BERT as additional or basic characteristics in the classification model, you can use the inner layers of BERT to get information about the text. Different layers of the neural network can display different levels of syntactic and semantic information of the text. Intuitively, the earlier layers contain more general information. Accordingly, the problem arises of choosing the required layer for inclusion in a specific model. During additional training of models, the problem of so-called “catastrophic forgetting” often arises ,which is that in the process of additional training on a specific set of data, knowledge expressed in the form of model weights trained at the stage of preliminary training is quickly erased. This leads to the leveling of the use of pre-trained models. It has been shown that using low learning rates can overcome this problem.

Multitasking Learning: In the absence of pre-trained natural language models, multitasking learning is effective in leveraging shared knowledge of multiple target tasks. When researchers are faced with several word processing tasks in the same subject area, it makes sense to retrain the BERT model in these tasks simultaneously.

CONCLUSION

The model received an intense reaction from the scientific community and is now used in almost all word processing problems. BERT has shown the advantage of bidirectional contextual models of text comprehension based on the architecture of transformers with an attention mechanism. It represented a quantum leap in the field of intelligent natural language processing and consolidated the superiority of using pre-trained text representation models on huge data sets as a universal basis for building intelligent algorithms for solving specific problems. Further improvement of the neural network architecture, coupled with fine-tuning the training procedure and parameters, will inevitably lead to significant improvements in many computer NLP algorithms, from text classification and annotation to machine translation and question-answer systems.

REFERENCES

1. https://towardsdatascience.com/bert-explained-state-of-the-art-language-model-for-nlp-f8b21a9b6270
2. https://sayanchak.medium.com/practical-uses-of-bert-c384ae3a5c2a
3. Official BERT paper: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Jacob Devlin Ming-Wei Chang Kenton Lee Kristina Toutanova