# BERT Technology Review

1. Introduction

BERT is a new language representative model called Bidirectional Encoder Representations from Transformers. It is quite different from recent language models which usually pretrain deep bidirectional representations from unlabeled text by jointly conditioning on only left or right side in any layer. BERT can do this on both left and right side in all layers. It's a very power tool for natural language processing.

2. Body

BERT is a new bi-directional encoder language model based on Transformer architecture, which is the pioneer of pre-trained language models. Unlike unidirectional language models such as ELMo and GPT, BERT aims to build a bidirectional language model to better capture the contextual semantics between utterances, making it more generalizable to more tasks. The BERT model is conceptually simple but powerful, it achieved optimal results on 11 classical natural language processing tasks, including improving the GLUE dataset to 80.5% (a 7.7% improvement over the previous optimal model), the SQuAD v1.1 Q&A test dataset to 93.2 (a 1.5 point improvement), and the SQuAD v2.0 dataset to 83.1 (a 5.1 point improvement).

Current pre-training models are mainly classified into two categories: feature-based and fine-tuning, but they are mostly based on one-way language models for language learning representation, which prevents optimal training results for many sentence-level downstream tasks. Therefore, this paper proposes a bi-directional pre-training representation model called BERT, which largely alleviates the constraints brought by one-way models. At the same time, we introduce "completion" and "sentence matching" as two general tasks at word level and sentence level, respectively, to train the BERT model. The experiments show that the application of the BERT model has achieved

SOTA results for all 11 NLP tasks. Before the advent of BERT, most of the existing pre-trained language models were unidirectional model architectures. For example, the GPT model introduced by OpenAI [14] introduces the masked attention mechanism in the Transformer Decoder layer [2], which allows the model to fully learn the contextual semantics. However, the unidirectional model architecture still limits the generalization ability of the pre-trained models to NLP tasks, and many NLP tasks have difficulty learning more useful features from the unidirectional architecture, such as question and answer systems [12]. Therefore, there is a need to continue optimizing the current pre-training architecture to make it adaptable to more kinds of tasks and enhance its generalization in the NLP field.

The overall framework of BERT consists of two phases, Pre-training and Fine-tuning, In the Pre-training phase, the model is first trained on a set of generic tasks using unlabeled data. After the trained model acquires a set of initialized parameters, the model is migrated to a specific task in Fine-tuning phase, where it continues to tune the parameters using labeled data until it converges again on the specific task.

The BERT model adopts the Encoder architecture in Transformer and stacks Encoder blocks to form the final BERT architecture by introducing a multi-headed attention mechanism. To accommodate tasks of different sizes, BERT divides its architecture into two categories: base and large. The smaller base structure contains 12 Encoder units with 12 Attention blocks in each unit and a word vector dimension of 768, and the larger large structure contains 24 Encoder units with 16 Attention blocks in each unit and a word vector dimension of 1024. By using Transformer as the main framework of the model, BERT is able to capture the bidirectional relationships in the utterances more thoroughly, which greatly improves the performance of the pre-trained model in specific tasks.

The input to the BERT model consists of three components. In addition to token word vectors in the traditional sense, BERT introduces

positional word vectors and sentence word vectors. The positional word vector is the same idea as Transformer, but instead of using its formula, BERT randomly initializes it into the model and trains it together; the sentence word vector is essentially a 0-1 representation that aims to distinguish between upper and lower sentences in the input paragraph. These three word vectors with different meanings are added together to form the final word vector for the input model.

3. Conclusion

The paper showed the power of BERT. And experimental results show that deep bidirectional language models can greatly improve the performance of NLP tasks. At the same time, transfer learning of pre-trained models, which is gradually becoming an integral part of language understanding systems, can even enable some low-resource tasks to benefit from deep unidirectional architectures.