

BERT and Its Applications in Natural Language Understanding

Ziting Zhu

*College of Artificial Intelligence, China University of Petroleum-Beijing, Beijing, China
3369498653@qq.com*

Abstract. Natural Language Understanding (NLU) has long been constrained by semantic ambiguity and context dependence, whereas traditional approaches struggle to overcome the limitations of unidirectional encoding and static semantic representation. Through a literature review method, this paper systematically examines the theoretical foundations of BERT (Bidirectional Encoder Representations from Transformers) and its empirical applications in NLU. Grounded in the Transformer encoder and the Masked Language Modeling objective, BERT addresses the limitations of traditional models through bidirectional context modeling, and its fine-tuning mechanism supports efficient downstream task transfer. In the context of practical applications, BERT has improved semantic-matching accuracy in information retrieval and achieved breakthroughs in question-answering systems and the medical domain. In addition, research on the improvement of BERT is continuously advancing, including model compression, multimodal extensions, and innovative pre-training objectives, which continue to advance the model's evolution. The article concludes that BERT fills a critical gap in systematic pre-trained-model research and supplies an effective blueprint for industrial-grade NLU systems, but still has limitations in computational cost and long-document processing. Future work should therefore prioritize dynamic sparse training and unified cross-modal architectures.

Keywords: BERT, Natural language understanding, Question-answering systems, Pre-trained language models

1. Introduction

Natural Language Understanding (NLU) confronts numerous challenges such as lexical ambiguity and contextual dependency. Traditional methods have many drawbacks and limitations: unidirectional RNNs (Recurrent Neural Networks) suffer from single-direction encoding and vanishing-gradient issues, whereas shallow word embeddings provide static representations incapable of dynamically disambiguating polysemous terms according to context. However, after Vaswani et al. proposed the Transformer architecture in 2017 [1], people successfully achieved parallelized computing and hierarchical feature extraction through the self-attention mechanism, bringing about a profound transformation in this field. This major breakthrough lays the foundation for the Bidirectional Encoder Representations from Transformers (BERT), which provides an overall solution for NLU tasks through the pre-training-fine-tuning paradigm.

Since the inception of Transformer, numerous studies have made outstanding contributions to the development of BERT. Devlin et al. demonstrated that Masked Language Modeling (MLM) and Next Sentence Prediction (NSP) enable BERT to generate deep contextualized representations, achieving state-of-the-art performance on eleven NLU benchmarks [2]. Subsequent investigations—e.g., Liu et al.—have refined the architecture for domain-specific applications in medicine and law while improving computational efficiency [3]. Subsequent deployments in sentiment analysis [4] and multilingual understanding [5] have further attested to BERT’s broad generalizability.

This article will comprehensively explore the BERT model and its application in NLU through the research method of literature review, including the basic principles of the BERT model, its application in information retrieval, as well as its appropriate improvements and future development directions. The theoretical contribution of this article lies in improving and enriching the theoretical achievements, and offering suggestions for the development of the industry in practice, which has played a certain promoting role in the development of BERT.

2. Theoretical foundations of BERT

BERT (Bidirectional Encoder Representations from Transformers) is a landmark pre-trained language model in the field of natural language understanding [2]. It utilizes the Transformer encoder structure, combines innovative pre-training tasks, deeply learns bidirectional context representation, and through the fine-tuning effect of downstream tasks, significantly improves the performance of multiple NLU tasks.

2.1. Model architecture

The core of BERT is a stacked layer composed of Transformer encoders. The Transformer architecture completely abandons the RNN structure and implements sequence modeling based on the self-attention mechanism [1].

Each core layer of BERT contains two important modules: the multi-head self-attention mechanism and the feedforward neural network. By combining the two modules, this model can simultaneously focus on and weight the information of all other labels when dealing with sequence labels, perform nonlinear transformation and feature integration on each position vector of the self-attention output, and establish a global context-dependent relationship model, thereby achieving a powerful context understanding ability. The input of BERT is composed of the sum of three embeddings, mainly including Token Embedding, which is used to represent words or subwords themselves, and Segment Embedding, which is used to distinguish different sentences or text segments (such as sentence pairs separated by [SEP]), and there is a Position Embedding used to provide the absolute position of the token-encoded word in the sequence (as shown in Figure 1).

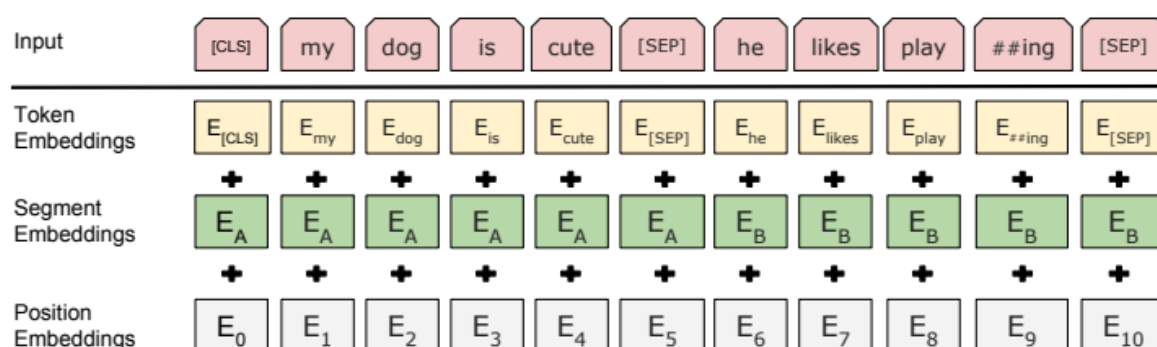


Figure 1. BERT input representation

2.2. Pre-training objectives

Traditional one-way language models (such as LSTM and GRU) can only model in one direction and cannot simultaneously utilize contextual information. However, BERT successfully breaks through the limitations of traditional one-way language models by learning deep bidirectional representations through two unsupervised tasks: the masked language model and the next step prediction.

Firstly, the MASK language model employs an innovation different from the past. It randomly masks 15% of the tokens (80% replaced with [MASK], 10% with random tokens, and 10% retained the original word), enabling the model to predict the masked word based on bidirectional context [2]. It is worth noting that when learning context-related semantics, this model has significantly enhanced its ability to discriminate polysemous words (such as "bank" in financial/geographical scenarios). Based on this, RoBERTa was further optimized to fully dynamic masking, completely eliminating the next sentence prediction task and focusing on larger corpora, longer training time, and dynamic masking mechanisms. It outperformed BERT in multiple NLU tasks [3].

The task of the next sentence prediction is to predict whether B is the true subsequent sentence of A by inputting a sentence pair (A+B) (50% positive samples + 50% random negative samples) [2]. This task enhances the model's understanding of textual coherence and is crucial for QA/NLI tasks. However, through subsequent research, it was found that the NSP task is overly simplistic. The ALBERT model replaces it with sentence order prediction (SOP), enhancing the logical reasoning ability between sentences by predicting the inversion of sentence order [6].

Overall, these training methods still have limitations in terms of adaptability to downstream tasks, consumption of computing resources, and understanding of long texts. Other methods can be introduced in the future to address the current deficiencies.

2.3. Fine-tuning mechanism

BERT performs outstandingly in multiple classic tasks, such as text classification tasks (like SST-2), natural language reasoning tasks (like MNLI), question-answering systems (like SQuAD), and sequence labeling tasks (such as named entity recognition in CoNLL-2003), all of which have achieved significant performance improvements through simple fine-tuning.

After pre-training, the fine-tuning mechanism is equally important. As shown in Figure 1, BERT ADAPTS to various NLU tasks through a unified input structure ([CLS]+ sentence A+[SEP]+ sentence B) and only requires the addition of a lightweight output layer to achieve this. Among them, the [CLS] vector is used for sentence pair classification (such as sentiment analysis), while the

Token vector is used for sequence labeling (such as named entity recognition). This enables the pre-trained backbone parameters to be fully reused, significantly reducing the migration cost. This mechanism enables the BERT model to support a wide range of downstream tasks (classification/annotation/question answering, etc.) with simple modifications, breaking through the limitation of traditional models that require custom design and perfectly integrating with pre-training tasks. BERT also supports fine-tuning strategies such as parameter sharing or freezing partial layer parameters to meet the actual demands of limited computing resources.

3. BERT in information retrieval

BERT has achieved remarkable breakthroughs in multiple tasks of natural language processing, especially in the field of information retrieval. This section will focus on the typical applications of BERT in semantic search and question-answering systems, demonstrating its advantages over traditional methods in terms of effectiveness and intelligence.

3.1. Semantic search

Traditional information retrieval systems (such as BM25, TF-IDF, etc.), mostly based on keyword matching mechanisms, achieved certain success in the early stage, but it was difficult to capture the deep semantic relationship between the query terms and the documents. For instance, "the causes of global warming" and "the factors leading to global warming" have a low degree of overlap at the keyword level, but their semantics are highly consistent. Traditional models often struggle to cope with this.

Compared with the traditional ones, BERT achieves context modeling of text through the bidirectional Transformer architecture and generates context-dependent embeddings with semantic information [2]. Based on this, a variety of semantic retrieval methods based on BERT have emerged in recent years. Among them, the double-tower structure is widely applied. A typical representative is Sentence-BERT. This model encodes queries and documents as dense vectors, respectively, and achieves extremely efficient semantic matching by calculating cosine similarity [7].

In addition, BERT has been successfully introduced in many fields to upgrade search services. In the field of large-scale industrial systems, Google announced in 2019 that it would deploy BERT to its English search system. This upgrade improved over 10% of English query results, especially enhancing performance in complex and long-tail queries. In the medical field, there is also a profound exploration of how to use fine-tuned BERT to build a semantic search engine adapted to the medical field. For example, by building a COVID-19 semantic search engine, it is possible to retrieve literature and information related to COVID-19 more accurately [8]. In the field of Chinese information processing, such as in the task of paraphrase generation, the experimental results of the model based on BERT and the column search strategy on the Chinese dataset show that compared with the baseline model, the BLEU index has increased by 10.47 and the semantic similarity index has increased by 0.105, fully demonstrating the effect of BERT in improving paraphrase generation [9]. These examples demonstrate that BERT not only excels in general semantic search but also ADAPTS to multilingual scenarios in vertical fields.

3.2. Question-answering systems

By taking advantage of BERT's bidirectional context encoding ability and self-attention mechanism, the question-answering system based on this model has also achieved a huge leap in the fields of reading comprehension and open-domain question answering compared with traditional models.

The Google AI team utilized the pioneering BERT-large model with a 24-layer transformer and 340 million parameters to achieve an F1 value as high as 92.7 on the SQuAD 1.1 dataset, successfully surpassing the human 91.6 [2], marking a brand-new breakthrough in reading comprehension for question-answering systems. By leveraging this technology, iFLYTEK made improvements on this basis in 2019, combining the ALBERT model [6] to compress parameters and launched BERT + DAE + AoA. It surpassed humans (89.452) with a new record of F1 value of 89.474 on the more challenging SQuAD 2.0 (which requires identifying unanswerable questions).

Furthermore, the emergence of the RAG model marks a new leap in open-domain question-answering technology [10]. It combines the bidirectional encoding of BERT with external retrieval, first retrieving relevant documents and then generating comprehensive answers, enabling the system to answer questions without preset text. Multi-task integration and rejection technologies (such as in-vehicle interaction) have thus been extended to open scenarios. With the continuous integration of pre-trained models and retrieval mechanisms, the application of BERT in question-answering systems will continue to expand, which also provides strong support for subsequent research on information extraction and dialogue systems.

4. Refinements and future directions of BERT

With the continuous pursuit of model performance and efficiency in the field of natural language processing, the improvement of BERT and its future development direction have become particularly important.

4.1. Model compression and acceleration

Compressing the model can effectively achieve the acceleration function, thereby enhancing the model's performance.

Knowledge distillation is one of the important technical means of BERT compression. As a typical representative, DistilBERT transfers knowledge from the large teacher model to the smaller student model by imitating the output and attention mechanism of the original BERT model, etc. This enables it to maintain approximately 95% of BERT performance while increasing the inference speed by about 60% [11]. It has perfectly achieved architecture simplification and model optimization.

Pruning technology is another important method. Evaluating the significance of weights and structures removes the parts that have a smaller impact on performance output to reduce the complexity of the model [12]. Structural pruning mainly targets the Attention Head or Feedforward layer, while non-structural pruning is based on weight sparsity.

Quantization techniques reduce the storage requirements of the model and accelerate the computation by lowering the precision of the model's weights, such as quantifying 32-bit floating-point numbers to 8-bit integers [13]. This method successfully reduced the model size by 75%, achieving efficient deployment of the model on hardware.

Future research is also exploring more efficient compression methods, such as quantized perception training (QAT) and dynamic pruning, to adapt to diverse deployment scenarios.

4.2. Multilingual and multimodal extensions

The development of cross-language models is one of the important future research directions for BERT. mBERT is based on the same pre-training task architecture as the original BERT. Through joint training with multi-language corpora, it fully understands the text and semantic similarities and differences of different languages, and performs outstandingly in tasks such as cross-language text classification and machine translation [2]. However, mBERT does not use language labels or vocabulary sharing issues to limit its performance. XLM-R, on the other hand, adopts a larger corpus, improved training methods and model architectures, and optimizes the multilingual pre-trained model, enabling it to exhibit higher accuracy than mBERT on various multilingual tasks [14].

In order for BERT to better integrate text with other types of data, multimodal expansion is another direction that requires in-depth research. Among them, ViLBERT [15] and LXMERT [16] are typical representatives of image-text combined models. ViLBERT uses two independent Transformer encoders to process text and visual information, respectively, fusing the two. By correlating images with text descriptions, it has been well applied in scenarios such as visual question answering. LXMERT has further optimized the fusion method by allowing text and visual features to interact during the encoding process, achieving a deeper modal fusion. These multimodal models can handle more complex application scenarios, such as intelligent video analysis and multimodal sentiment analysis.

With the increasing demand for modal fusion of speech, images, text, etc., unified multimodal pre-trained models will gradually become the research focus.

4.3. Innovations in pre-training objectives

Novel pre-training objectives are actively investigated. Contrastive learning markedly enhances sentence embeddings: SimCSE leverages unsupervised contrastive signals—augmenting a sentence as a positive pair—to outperform prior methods in semantic retrieval [17]. Generative pre-training, including hybrid BERT-GPT architectures [18], combines BERT's deep semantic encoding with GPT's left-to-right autoregressive generation, addressing BERT's generative shortcomings and enabling text summarization and dialogue systems. Such encoder-decoder unification heralds next-generation models capable of broader applications.

5. Conclusion

This study aims to analyze the technical principles of the BERT model and its application value in the field of NLU. The findings show that, based on Transformer and mask language modeling, BERT has broken through the limitations of traditional models. Its semantic vectorization significantly improves search accuracy and is adapted to vertical fields. Meanwhile, its compression technology and multimodal extension have promoted efficiency and cross-modal breakthroughs. By filling the gap in systematic research on pre-trained models and expanding the application boundaries of bidirectional coding theory in cross-language and generative tasks, it contributes to the existing knowledge system. It is of guiding significance for providing solutions for industrial-grade NLU systems (such as intelligent question answering and search).

However, this paper did not conduct a systematic empirical evaluation of the performance of each functional module of the BERT model. Future research will focus on quantitative analysis of BERT's performance in different task scenarios, with particular attention to its adaptability and transfer

capabilities in Chinese scenarios, low-resource languages, and multilingual corpora. Although BERT has strong language understanding and generalization capabilities, its limitations remain obvious, such as high pre-training costs, low inference efficiency, and insufficient long text processing capabilities. These issues may affect the model's universality and scalability in specific applications. Subsequent research can focus on dynamic sparse training mechanisms, hierarchical attention modeling strategies, and the construction of a unified multimodal architecture to further enhance the model's performance in multi-task and multi-domain environments. Overall, this study provides a fundamental reference and development direction for theoretical exploration and engineering implementation in the field of NLU.

References

- [1] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.
- [2] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019, June). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)* (pp. 4171-4186).
- [3] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv: 1907.11692*.
- [4] Sun, C., Qiu, X., Xu, Y., & Huang, X. (2019, October). How to fine-tune bert for text classification?. In *China national conference on Chinese computational linguistics*(pp. 194-206). Cham: Springer International Publishing.
- [5] Pires, T., Schlinger, E., & Garrette, D. (2019). How multilingual is multilingual BERT?. *arXiv preprint arXiv: 1906.01502*.
- [6] Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2019). Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv: 1909.11942*.
- [7] Reimers, N., & Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv: 1908.10084*.
- [8] Jose, A., & Harikumar, S. (2021, December). COVID-19 semantic search engine using sentence-transformer models. In *International Conference on Computational Intelligence, Cyber Security, and Computational Models* (pp. 189-200). Cham: Springer International Publishing.
- [9] Fan, Q., Kong, C., Y, L., & Yang, E. (2021). Chinese interpretation generation based on BERT and column search. *Journal of Chinese Information Processing*, 35(11), 80-90.
- [10] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33, 9459-9474.
- [11] Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv: 1910.01108*.
- [12] Han, S., Pool, J., Tran, J., & Dally, W. (2015). Learning both weights and connections for efficient neural network. *Advances in neural information processing systems*, 28.
- [13] Jacob, B., Kligys, S., Chen, B., Zhu, M., Tang, M., Howard, A., ... & Kalenichenko, D. (2018). Quantization and training of neural networks for efficient integer-arithmetic-only inference. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2704-2713).
- [14] Ruder, S., Søgaard, A., & Vulić, I. (2019, July). Unsupervised cross-lingual representation learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*(pp. 31-38).
- [15] Lu, J., Batra, D., Parikh, D., & Lee, S. (2019). Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32.
- [16] Tan, H., & Bansal, M. (2019). Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv: 1908.07490*.
- [17] Gao, T., Yao, X., & Chen, D. (2021). Simcse: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv: 2104.08821*.
- [18] Dong, L., Yang, N., Wang, W., Wei, F., Liu, X., Wang, Y., ... & Hon, H. W. (2019). Unified language model pre-training for natural language understanding and generation. *Advances in neural information processing systems*, 32.