How BERT Works and Its Strengths and Weaknesses

CS410 Text Information System Technology Review
Yang Yang (NetID: yangy19)

Introduction

BERT stands for Bidirectional Encoder Representations from Transformers. It has become one of the most popular State-of-the-Art architectures in NLP since Google first introduced and open-sourced it in 2018. The review will talk about how BERT works, what are the current research activities, especially strengths and weaknesses of BERT model and its real-world applications.

Keywords: BERT, Research, Strengths, Weaknesses, Applications

Review

How BERT Works

1. BERT's Architecture

We already knew that BERT is based on the Transformer architecture. There are basically two model sizes. One is "base" version with 12 layers, 12 heads and 110M parameters and the other one is "large" version with 24 layers, 16 heads and 340M parameter.

As we know, machines don't understand languages. Therefore, every input text has to be translated into a machine-understandable language. All the Transformer layers are exactly the "translators" – Encoder blocks, which consist of multiple self-attention "heads".

The common procedure for BERT includes two parts: pre-training and fine-tuning. Pre-training usually has two NLP tasks: Masked Language Modeling and Next Sentence Prediction.

2. Input Embedding

In order to encode the input information, every input text needs to go through text preprocessing which is input embedding including Token Embeddings, Segment Embeddings and Position Embeddings. A combination of the three embeddings represent a lot information of the input text for the model to be easily trained.

3. Pre-training on MLM and NSP

Pre-training on two tasks:

- Masked Language Modeling: Learn to understand the relationship between words

- Next Sentence Prediction: Learn to understand the relationship between sentences

1) Pre-training on MLM (Masked Language Modeling)
From the name of BERT, we have already known that the network gathers information from both directions. So it looks from the very first layer to the last one from both the left and right context, which is quite different from traditional language model's left-to-right context train sequence.
If we compare BERT with OpenAI GPT and ELMo, BERT is bidirectional, GPT is left-to-right only, and ELMO is shallowly bidirectional.

2) Pre-training on NSP (Next Sentence Prediction)
Next Sentence Prediction is required for tasks that needs an understanding between sentences, such as sequence classification. It is trying to label if the second sentence is actually the next sentence of the first one with given two.

Current Research Overview

BERT has pushed state of the art in a lot of aspects in NLP, but we have very limited knowledge on what is behind its success. Therefore, many researches started to dig into topics like what kind of information it learns and how it is represented, common modifications to its training objectives and architecture, the overparameterization issue and approaches to compression.

The first topic is localizing linguistic knowledge. In BERT model, usually lower layers seem to best capture linguistic knowledge about linear word order, middle layers about syntactic information, and final layers about task-specific questions. It appears however, that knowledge spreads across the entire model and cannot be specifically located although syntactic information appears early in the model.

The second topic is how to improve BERT training through architectural improvements, training process (e.g. larger batch sizes), and new or additional methods for pre-training (e.g. altered masking methods, new training objectives or even explicitly providing structured knowledge) and fine-tuning (e.g. amended architecture, robustness enhancements, or regularization techniques). A core result is that changing the number of layers seem to be more significant than changing the number of heads. There is even some discussion around whether pre-training is helpful on all tasks and that it should be analyzed carefully because it is so expensive.

Some papers also talk about the issue of BERT's size and offer different approaches that could help with the disadvantages that come with such a large model (e.g. reproducibility, environmental concerns, complexity). There is a detailed discussion about various methods, including pruning, disabling of attention heads, knowledge distillation, and quantification. Many promising options are discussed that result in good performance, even with reduced complexity.

Lastly, some future research directions are suggested from researches which include different and harder benchmarks, improving reasoning through explicit teaching, or building a better understanding about what happens at inference time.

1. Strengths of BERT

One obvious strength is that BERT has extremely good performance compared with its predecessor models due to its bidirectional encoding architecture. Because of this, the model makes predictions not only from the words before it but also words after it.

From some research papers, it shows that BERT did learn syntactic knowledge, semantic knowledge, and somewhat word knowledge from the context. Researches indicate that BERT representations are hierarchical rather than linear, which is similar to the syntactic tree structure. Moreover, BERT has some knowledge of semantic roles. That knowledge provides more accurate predictions compared with other approaches. This is arguably the primary value of BERT as well as the key to solving nearly all Natural Language Processing tasks.

Another strength is BERT's robustness to withstand various compression and still keeping good performance. Recent papers describe a lot of compression techniques such as head pruning, zeroing out parameters to speed up computation, and increasing batch size to something very high (32k) and have minimal performance impact. While this suggests models are not making good use of parameters, it also means that the model is flexible for customization. Users can apply any of the mentioned techniques above as needed.

2. Weaknesses of BERT

One negative aspect about BERT is that BERT has many issues with inference and reasoning of commonsense knowledge. BERT also does not learn the concept of negation and can create inference with sentences that are not grammatically correct. It is still able to perform well on many tasks through learning stereotypical answers but is not able to do "true" inference. This behavior might be explained with the fact that BERT has difficulties with learning semantic information to some extent, e.g. numbers. Without semantic information, it would also not be possible to perform inference or reasoning. Another reason for this behavior is that even though semantic information is present within BERT, it is not represented in ways in which the model can use it for inference.

One approach how this could be mitigated could be to try to teach BERT how to reason – either explicitly through basic logical reasoning techniques or implicitly through having it learn by hard tasks that teach him reasoning, or even through imitation learning.

Another weakness is overparameterization. Many parameters are trained to gain very little accuracy improvement. Training a lot of parameters can add complexity and running time, which can also slow down inference retrieval.

Some approaches have been introduced to potentially overcome this issue:

- Distillation: Removing heads and parameters that can cause a reduction in memory footprint
- Quantization: reducing the precision of parameters
- Training a student model using the original BERT as teacher model
- Training adapters over these networks to learn better task-specific stuff. This has shown better result than generalized neural networks.

Extension of BERT Applications

BERT has not only become a must-be baseline in NLP academia and industry, but also been widely expanded and incorporated its core idea and architecture in various fields.

1. Pre-trained BERT models

- BioBERT – biomedical text
- SciBERT – scientific publications
- ClinicalBERT – clinical notes
- M-BERT – task specific annotations in one language is used to fine tune model for evaluation in another language
- ERNIE – incorporates knowledge by masking entities and phrases using KG
- videoBERT – for video captioning

2. Fine-tuned BERT models

- DocBERT – document classification
- PatentBERT – patent classification

Conclusion

BERT, as one of the most popular pre-trained language representation models, obtains the State-of-the-Art results by simple fine-tuning, including text classification, masked word completion, named entity recognition and etc. Its strengths transformed the MLP landscape, but there are still weaknesses. The core architecture idea of BERT has expanded to other areas and created a variety of extension of BERT application.

References

Adhikari, A., Ram, A., Tang, R., & Lin, J. (1904). DocBERT: BERT for Document Classification. arXiv (2019). *arXiv preprint arXiv:1904.08398*.

Beltagy, I., Lo, K., & Cohan, A. (2019). SciBERT: A Pretrained Language Model for Scientific Text. *arXiv preprint arXiv:1903.10676*.

Huang, K., Altosaar, J., & Ranganath, R. (2019). Clinicalbert: Modeling Clinical Notes and Predicting Hospital Readmission. *arXiv preprint arXiv:1904.05342*.

Klein, T., & Nabi, M. (2019). Attention Is (not) All You Need for Commonsense Reasoning. *arXiv preprint arXiv:1905.13497*.

Lee, J. S., & Hsiang, J. (2019). PatentBERT: Patent Classification with Fine-Tuning a pre-trained BERT Model. *arXiv preprint arXiv:1906.02124*.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, *36*(4), 1234-1240.

Pires, T., Schlinger, E., & Garrette, D. (2019). How Multilingual is Multilingual BERT?. *arXiv preprint arXiv:1906.01502*.

Rajasekharan, A. (2019, Jun 17). *A review of BERT based models*. towards data science. https://towardsdatascience.com/a-review-of-bert-based-models-4ffdc0f15d58

Rogers, A., Kovaleva, O., & Rumshisky, A. (2020). A Primer in BERTology: What we know about how BERT works. *arXiv preprint arXiv:2002.12327*.

Sun, C., Myers, A., Vondrick, C., Murphy, K., & Schmid, C. (2019). VideoBERT: A Joint Model for Video and Language Representation Learning. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 7464-7473).

Zaki Rizvi, M.S. (2019, Sep 25). *Demystifying BERT: A Comprehensive Guide to the Groundbreaking NLP Framework*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2019/09/demystifying-bert-groundbreaking-nlp-framework/

Zhang, Z., Han, X., Liu, Z., Jiang, X., Sun, M., & Liu, Q. (2019). ERNIE: Enhanced Language Representation with Informative Entities. *arXiv preprint arXiv:1905.07129*.