Deep learning in Text Information System: BERT model and its variants

Introduction

The rise of the pre-trained nlp model represented by BERT (Devlin et al., 2018) has captured more and more sights from different industries. And many applications of this technology such as language translation and customer service robots have achieved initial success.

At the same time, the high speed development of the semiconductor industry makes it profitable to train and use these deep neural networks in industry.

It's time to think about the possibilities of applying such technology in the text information system now. In this article, we will first talk about the reasons why the pre-trained nlp model, BERT, is superior to the traditional text information algorithm. Then we will analyze the drawbacks of this technology, and the current solutions of these problems. Finally, we will give some suggestions to its applications and prospects in the text information system.

Body

What is BERT?

BERT is the state of the art natural language processing model proposed in 2018. It is designed to capture the semantic meaning of the input sentences. And since it's also a pre-trained model, BERT can be fine-tuned to many different kinds of tasks with limited computation resources.

Why is BERT more powerful than traditional text information algorithms?

Most traditional text information algorithms such as BM25 and PLSA use "bag of words" representations. Such a strategy does simplify the computation greatly, but at the same time, it hurts the performance of these algorithms a lot. The key problem brought by such representation is there are too many naive assumptions made during the computation process. The "bag of words" representation makes the algorithm ignore the relationships between word and word which could have a lot of information that is useful to the tasks in the text information system.

One the other hand, the n-gram model can alleviate this problem, but would have to occupy too much computation resources.

Deep learning models, with the help of dense word embedding and complex network structure, are able to capture such relationships with limited resource occupation. And the emergence of the BERT, makes the model able to capture not only unidirectional but bidirectional dependencies in the sentence.

What are the challenges people could meet when they apply BERT in a Text Information System? Any solutions?

If we look at the original BERT introduced in a journal article written by Jacob Devlin group in 2018, we can immediately find that the BERT model only accepts fixed length input (which is 512 tokens). That means BERT can only handle no more than a certain fixed length of the text.

This could be problematic when we try to utilize BERT in a text information system. Under some scenarios, we may need to process the long documents which can not be fitted in BERT. For example, how about we need a system that can help the researchers to search the journal articles they want?

One naive solution that we can come up with easily is that we split the long document into multiple segments and then feed these segments into our network. That's exactly what people did in the past. But this solution is defective and will hurt the performance of the network.

The information about the dependencies between segment and segment would be discarded, when we apply this naive solution. Recall that one of the reasons that we want to use the deep learning model instead of a traditional text information algorithm in our future text information system is that we want to capture the semantic meaning, the relationship between word and word. Luckily, in 2019, the Zihang Dai group proposed the Transformer-XL (Dai et al., 2019). Here, we can simply think of the Transformer (Vaswani et al., 2017) as the basic framework for the BERT. How about we build a recurrent network that can cache the last segments' output and concat it to the current input? Then we would have a network that is not only able to capture the dependencies inside each segment but also the dependencies among the segments. And that's exactly the main idea of Transformer-XL.

But as we said, Transformer is the framework of BERT and not pretraining which means it may still be too early and challenging to let the industry world to actually apply Transformer-XL. So that in the same year, XLNet (Yang et al., 2019) is proposed, it can be described as the combination of BERT and Transformer-XL network. It is pre-trained and "outperforms BERT on 20 tasks" (Yang et al., 2019, p.1) which is amazing!.Thanks to this brilliant work, the gap between theory and the commercial area is closing.

But before we can step in and touch the beautiful future of the text information system, there are still many concerns left to be resolved.

What about the efficiency problem? Although the pre-training model can help us save a lot of computation resources. Applying a deep learning model to complete the text information system task is still time consuming in most scenarios. Suppose we want to use BERT to build a search engine. And we want to find the best matching documents in one million documents given an input query. The naive way to complete this task is to pair the query with each document, concat and feed them into the network.

However, this will lead to unacceptable response time during the actual use. Imagine that we have to use a deep learning model to make one million predictions to respond to a single query. And the pre-training model can only accelerate our training progress! But this is not the end of the story, the clever researchers save our engineers again. Also in 2019, Sentence-BERT (Reimers et al., 2019) is created. In the original journal article, it said "This reduces the effort for finding the most similar pair from 65 hours with BERT / RoBERTa to about 5 seconds with SBERT, while maintaining the accuracy from BERT." (Reimers et al., 2019, p.1)

Recall that BERT is used to extract the semantic meaning of the input words. Instead of feeding query and paired documents into the network at the same time, why not just only feed one document each time instead and cache them in our database for future use!

So that we can pre-computing those one million documents to get the dense representations of their semantic meaning and store them in our database in advance. When the new query comes, we can just only run the BERT one time to get the query's semantic representation, and then we just need to apply cosine similarity or a simple siamese network to do pretty fast computation and get the matching scores between query and each candidate document for document ranking.

These are only the general idea of the Sentence-BERT, we highly recommend you to check the original paper by yourself since there are many details that we can not cover here due to the limited space.

Possible Applications

Because BERT is designed to capture the semantic meaning of input words, it can be applied to almost all the text information system tasks by fine-tuning it with other deep learning models. As we mentioned above, we can use bert to complete the Text Retrieval Task. It's easy to come up with the idea that we can use BERT to do topic mining also by simply paring the document with candidate topics. We can even apply BERT to get the semantic representations of the user-like documents and treat them as one of the features in our Recommendation system.

The future of BERT in text information system area

We are very optimistic about the future of BERT and its variants. So what's the next step then?

- Customization

User groups of different text information systems may have different information preferences. We may need to pre-train the BERT with more computer science papers if the text information system we are implementing is for computer researchers.

- Higher Efficiency

Computer science engineers shall never stop their pursuit of efficiency. Can we make BERT even faster? How about a kind of BERT variant that can allow us to input multiple documents and get the semantic representation of each document in a single time?

Conclusion

The high speed development of the semiconductor industry and recent years' theoretical breakthrough in the Natural Language processing area make it feasible and profitable to apply neural networks like BERT in the text information system area. At the same time, the ongoing modifications on the existing BERT model keep reducing the gap between the theory and the commercial area. The future of the pre-trained nlp model represented by BERT is exciting.

Reference

[1] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

[2] Dai, Z., Yang, Z., Yang, Y., Carbonell, J., Le, Q. V., & Salakhutdinov, R. (2019). Transformer-xl: Attentive language models beyond a fixed-length context. arXiv preprint arXiv:1901.02860.

[3] Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., & Le, Q. V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. Advances in neural information processing systems, 32.

[4] Reimers, N., & Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. arXiv preprint arXiv:1908.10084.

[5] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In Advances in neural information processing systems (pp. 5998-6008).