SiCheng Meng

Lan ID: meng29

CS 410: Text Information Systems

November 6, 2022

A TECH REVIEW

FROM N-GRAM TO BERT, HOW LANGUAGE MODELS EVOLVED

Intro

Language model pre-training has been proved to be effective to improve natural language processing tasks in many ways. This paper provides a overview of three language models, n-gram, RNNLM and BERT, that has been widely used in the industry. You'll learn how they have evolved based on previous models. Some comparisons between the three models are

N-Gram

The mechanism of conventional "n-gram" models reads the text by dividing sentences into n adjacent words, or sometimes n adjacent letters. MetaPy in MP 1 is an example of using this n-gram mechanism to segment texts.

The problem of n-gram is obvious – when the information in a n-word segmentation is not enough, the correct meaning cannot be interpreted. For example, if sentence "this is not organic" is divided by bi-gram, or 2-gram, we will get "this is", "is not", not organic". However, in

segmentations such as the first pair “this is”, it is not possible to read the meaning correctly due to lack of useful information, not to mention predicting the next word.

Although in MP 1, we’ve tried to remove common words (or “stop words”) such as “the”, “a”, “this”, etc., we still needed the help of a third-party lexicon – the “background words language model”. A more mainstream way to overcome the problem is to simply increase the number of words, i.e., increasing “n”. However, while the accuracy of the model is increased in this way, it also makes the model huge. RNNLM is the solution to this issue.

RNNLM

RNNLM is a model that adapts RNN (Recurrent Neural Network) to natural language processing. Compared to previous models, RNNLM is considered lighter and more accurate, and more straightforward.

It achieved higher accuracy with a simple model through neural nets. The innovative approach of RNNLM is that, instead of reading the text from a limited number of words/characters, it tries to understand the text more accurately by reading the text from the full sentence up to that point. More specifically, RNNLM reads the text by passing information in a relay format instead of reading each n-word separately. In other words, RNNLM passes the previous context as information to achieve high accuracy. The trick of this approach is through layering. More information can be found on academic websites.

One big problem of RNNLM is that, as a sentence gets longer, the accuracy becomes lower. This issue was later improved by “attention”, which is a RNNLM-based mechanism. Transformer model was built based on attention mechanisms and was first published in the paper “Attention is All You Need”.

BERT

BERT stands for “Bidirectional Encoder Representations from Transformers”. It was created and published in 2018 by researchers at Google AI Language. Since year 2020, Google has been using BERT to generate language models which were then used in almost every English-language query, as well as a few other languages.

Directional models like RNNLM read a text input sequentially, either from left-to-right or right-to-left. BERT makes use of the “encoder part” of Transformer model, which is an encoder-decoder model that uses self-attention to make sense of contextual relations between words in text. The key innovation of Transformer encoder is that it reads the whole text at once, therefore it is considered bidirectional, or more precisely, non-directional. This allows parallel processing, and the model can learn the context of a token based on both left and right sides of the it, and thus it’s much faster than any other model with the same performance.

How does BERT overcome the drawback of traditional directional approach in older models? Masked LM (MLM) and Next Sentence Prediction (NSP) are the two main training strategies being used.

Before feeding a text into BERT, a certain percentage of words (usually less than 15%) in each sequence are first replaced by a masked token, named MLM. The model then tries to predict the original values of the masked tokens, based on the 85% non-masked words in that sequence. The loss function of BERT only considers the prediction the masked tokens and ignore other words that are not masked.

NSP is the strategy used in the actual training process of BERT. The input of the model are sentences in pairs – each pair includes two sentences. Half of the input are true pairs, which means the second sentence is the actual subsequent of the first sentence in the original document, while

the other half of the input are just two random sentences from the corpus. The model will then learn to predict if the second sentence in a pair is the true subsequent sentence from the original document.

The two strategies are trained together in BERT, which minimize the combined loss functions. With no doubt, BERT is a breakthrough in the use of machine learning for NLP. The fact that it's approachable and allows fast fine-tuning will undoubtfully allow a wide range of practical applications in the future.

Conclusion

Just as RNNLM has achieved "lightweight & high accuracy" by reversing the trend of n-gram-based huge models, BERT further increased the speed and accuracy by reading the whole text in a non-directional way. Other Transformer-based models such as GPT-3 and PaLM has also been released in recent years. There is a high possibility that in the near future, people will see models that can overtake huge models in a totally new and innovative way.

References

Suetsugu, “RNNLM explained in depth and in an easy-to-understand way / Natural Language Processing (NLP)”, *www.medium.com*, 2022

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, Illia Polosukhin, “Attention is All you Need”, *Advances in Neural Information Processing Systems 30*, 2017

Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”, *Google AI Language*, 2018