Techniques and Architectures for Text Analysis in NLP

Yunfei Ouyang
University of Illinois Urbana-Champaign
yunfeio2@illinois.edu

Abstract

Utilizing machine learning techniques in Natural Language Processing for text analysis makes it one of the hottest domains in Computer Science. Recent use of large data sets, high-capacity neural net models, and supervised unsupervised learning made the text analysis performance increase by a huge margin. In this paper, we will investigate state-of-art language modelling techniques such as the **Attention** mechanism [2], as well as large language models such as **BERT** [3] and **GPT-2** [1].

Introduction

The performance of text analysis tasks is advancing at a hyper speed because of the utilization of language modelling techniques with machine learning. We will look into how different techniques are used in text analysis, especially the techniques that are built upon the sequential architecture, i.e. Recurrent Neural Networks. We also will dive into how a large language processing model works and its performance in real-world tasks.

Methodology

We will explain the following techniques, RNN, LSTM, and Attention.

RNN

RNN or so-called Recurrent Neural Networks is one of the most important neural network architectures that is heavily used in Natural Language Processing. The main difference between an RNN compared to a CNN (Convolutional Neural Network) is the ability to take variable length input, which makes it suitable for sequential data analysis such as text and speech. RNN treats each word in a sentence as a separate input within a time frame where if the current word is at time t, then the previous word is at time t-1. The utilization of sequential input makes RNN very versatile to many text analysis tasks such as classification, translation, question and answer, etc. classification tasks modify the output layer of RNN to one output; many inputs to many outputs architecture change it to Encoder Decoder composition which is suitable for machine translation.

LSTM

Long Short-Term Memory architecture utilizes the Gated Recurrent Units to accommodate the vanish gradient long-term dependencies in Recurrent Neural Networks. GRU consists of an additional memory unit as the update gate which uses the property of **tanh**, **sigmoid** and **softmax** which is used to calculate the activation value of the next unit. Similarly, LSTM adapt the GRU architecture and introduced one more unit which is so called the forget gate. Adding the forget gate gives the architecture an option to either remember or forget the old value at time t-1 or use it as an activation for the next unit, this makes LSTM can remember important long-term words dependencies as well as ignore dependencies that are not important to the sentence that is being analyzed.

Attention

In 2017, an influential paper "**Attention is all you need**" [2] was published by Google Brain and the University of Toronto. It proposed the **Transformer** architecture which relies entirely on an attention mechanism to draw global dependencies between input and output instead of requiring to work in conjunction with a recurrent network. The transformer architecture uses an Encoder and Decoder stacks for the machine translation, and it introduced Multi-Head Attention architecture which utilizes the Scaled Dot-Product Attention architecture to allow the model to jointly attend to information from different representation subspaces at different positions. The Self-Attention mechanism in the Transformer architecture yields state-of-the-art performance around its time, and it is being widely used in many NLP tasks.

Large-Scale Model

Now we have introduced cutting-edge Text analysis techniques and architectures, we are going to evaluate some recent large-scale models for the Text analysis task, especially **BERT** and **GPT-2**.

BERT [3]

BERT stands for Bidirectional Encoder Representations from Transformer architectures and is a deep learning model in which every output is connected to every input. Utilizing the attention mechanism, making it advances the state-of-the-art eleven NLP tasks. Google utilizes large data sets to train BERT making it one of the largest language models, with about 110 million parameters, BERT is able to accomplish several tasks including Sequence-to-sequence based language generation: Question and Answer, Abstract Summarization, Sentence Prediction, Conversation Generation, etc. and Natural Language Understanding Tasks such as Word sense disambiguation, Natural Language Inference, Sentiment Classifications, etc.

GPT-2[1]

GPT-2 stands for Generative Pre-trained Transformer 2 is a large model created by OpenAI in 2019. The architecture is similar to BERT where it has just the decoder blocks from the transformer architecture, whereas BERT utilizes the encoder blocks from the transformer. Note that GPT-2 applied self-attention where the decoder was only allowed to gain information from the prior words in the sentence. GPT-2 is used in many text analyses similar to BERT. Note that GPT-2 is substantially larger than BERT with 1.5 billion parameters because OpenAI trained with a dataset that consists of 8 billion web pages.

Conclusion

Recent empirical improvements in the transformer architecture demonstrated that transfer learning is an integral part of many language modelling systems, where understanding how it works fundamentally can help us optimize the model even more beyond what is capable of. The major contribution of BERT and GPT-2 allowed the same pre-trained large-scale model to work on many different NLP tasks. Further advancement can be made through different architecture, more computational power, or even larger datasets.

Reference

[1] Alec Radford, Jeffrey Wu, Rewon Child, et al. Language Models are Unsupervised Multitask Learners. *CDN OpenAI, 2019*
https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

[2] Ashish Vaswani, Noam Shazeer, Niki Parmar, et al. Attention Is All You Need. *arXiv:1706.03762, 2017*
https://arxiv.org/pdf/1706.03762.pdf

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805, 2019*
https://arxiv.org/pdf/1810.04805.pdf