StartOnAI NLP III Tutorial

Vishak Srikanth, Ayush Karupakula, Smit Ramteke

February 2021

1 Introduction

Word Embeddings and Word2Vec are explained in detail in the NLP II tutorial, but key points are provided below. Word Embeddings are a way of representing words based on vectors. Word that have similar meanings(synonyms) have similar vector representations. We use cosine similarity to measure the similarity between word vectors which is a measure of how close two words are. Cosine similarity is the value obtained by dividing the dot product between two vectors by the product of the magnitudes of two vectors. Word vectors are learned similar to a neural network. Word embeddings are needed to make human language understandable and operable for machines. Word2Vec is a specific algorithm using Word embeddings that takes in a large set of texts and produces word vectors.

2 Challenges of Word2Vec

Although Word2Vec is a great starting point for natural language processing, Word2Vec has many challenges. Word2Vec is not able to examine and understand words used in a different context, or words that it has not seen before. In addition, Word2Vec does not recognize prefixes and suffixes such as re- or -less as subwords. In addition, new word embedding matrices are required when looking at different languages, which makes coming up with a universal algorithm for multiple languages difficult.

3 Attention and Transformers

When sentences are parsed as an initial step for BERT to be used, sentences are divided into tokens.Transformers are used to parse and do composition, in order to solve the issue of the interdependence of words. Compositions group together tokens in a compound word and groups together related words, and words in phrases and clauses. Transformers consist of attention layers. Attention layers in BERT contain 12 attention mechanisms,and 12 aspects used for attention. Attention means a token can focus on another token. The attention mechanisms

of tokens can be used for composition by relating words and help to build context which is used in BERT models.

4 BERT

In the dynamic field that is NLP, novel frameworks tend to appear on a regular basis. The latest state-of-the-art library never stays that way for long because the task of Natural Language Processing is incredibly complex and will never truly become perfect. The result is a significant amount of room for improvement, which leads to a number of frameworks being developed to account for the latest framework's setbacks. One of these NLP frameworks is BERT, or "Bidirectional Encoder Representations from Transformers".

The framework, developed by Google AI Language Researchers, has been able to achieve numerous state-of-the-art results in Natural Language Processing tasks. Essentially, BERT can operate on a vast unsupervised dataset and construct a model. The trained model is able to perform various Natural Language processing tasks. Utilizing BERT, many language models have been developed such as GPT-2 by OpenAI.

The basis for BERT is its transformer architecture, as seen in the acronym "Bidirectional Encoder Representations from Transformers". In contrast to sequence-to-sequence models which struggle with long-range dependencies and prevent the more efficient process of parallelization. This is due to their sequential nature. However, the transformer model is able to deal with the same tasks that can be performed by a Seq2Seq model, but managing the long-range dependencies.

An example of one of these NLP frameworks is BERT which stands for "Bidirectional Encoder Representations from Transformers". It was published by researchers at Google AI Language. A vast number of NLP tasks have achieved state-of-the-art results by using BERT. In addition, BERT is a pre-trained model which can take a large dataset of unlabelled text and establish an understanding of the inner workings of a language.

The first part of the acronym "bidirectional" means that BERT operates, in contrast to word embeddings like Word2Vec by taking information from the left and right side of a token's context. By having bidirectionality, BERT has an increased accuracy as it is able to capture the differences between words that have the same spelling, but different meanings. For example, word2Vec would have the same word vector for words that are homonyms, therefore allowing for misinterpretation of these words in different contexts. Overall, the path taken by the field of Natural Language Processing, to eventually come to a framework that could learn language representations like BERT was lengthy and gradual, starting with word embeddings like Word2Vec, transitioning to pre-training, and further extension with GPT by OpenAI.

5 Applications and Improvements

BERT has many applications including Google Search. Google uses BERT for some Google searches and BERT helps to better understand longer search keywords by better understanding the context of the search query. Since BERT models look at the complete context of a word by considering words that come before and after it, BERT can improve the understanding of the intent or purpose behind search queries, especially for conversational queries where prepositions like "for" and "to" can significantly change the meaning. By applying BERT models to both ranking and snippets in search, BERT helped Google improve 10% of the search queries in the U.S. in English. BERT can also be used for classification of patents as it can be used to search for similar patents, or search for prior papers on a patent filing. A variant of BERT called SciBERT can be used on scientific texts and can be used to search for textual content with significant portions that include scientific vocabulary terms. Improved versions of BERT have been created, which may have been trained on a larger corpus of words, run faster, or have more data. RoBERTa is an improved version of BERT from Facebook which was trained with a larger data set and using more computing power. RoBERTa uses dynamic masking, which masks changes in tokens, and removes a task called the Next Sentence Prediction task, which is used in BERT. XLNet has larger data, and better methods for training. DistilBert is a version of BERT using less parameters and is quicker, but has only 95% of the performance of BERT.

References

1. What Are Word Embeddings for Text? by Jason Brownlee
2. Word Embeddings in NLP and its Applications
3. word2vec
4. Word Embeddings and Their Challenges
5. Understanding BERT Transformer: Attention isn't all you need
6. FAQ: All about the BERT algorithm in Google search
7. SciBERT Embeddings
8. Patent Classification by Fine-Tuning BERT Language Model
9. BERT language model
10. BERT, RoBERTa, DistilBERT, XLNet: Which one to use?
11. Understanding searches better than ever before