Finding Related ArXiv Papers with BERT

CS224U Project - Literature Review

Junshen Kevin Chen
Stanford University
jkc1@stanford.edu

1 Introduction

For this project, I propose to design a system that evaluates the level of relation between two academic papers, and suggest possible related work given a text query.

Motivation Three main factors contribute to motivating this project. First, the current state of academic research tools such as Google Scholar functioning similar to a traditional search engine, which primarily leverages keyword matching and ranking by popularity (citation counts), while having being limitedly influenced by content semantics. Second, large data dumps of academic papers such as arXiv is available to extract relationships between academic work in the form of a citation graph. Finally, the research in transfomer models such as BERT (Devlin et al., 2018) provides powerful ways to encode semantics, and trained models are readily available to be fine-tuned for this specific task.

Problem Tentatively, I formulate this project into several following problems, with each latter depending on the former:

1. Given a query abstract, and a target abstract, predict a score of related-ness of query abstract to the target abstract
2. Given a query abstract, and a target abstract, predict a citation distance, or how many edges it needs to traverse from the query to the target
3. Given a query abstract, retrieve related articles (Stretch goal involving recommender system)

In this document, I select and discuss the literature related to this project, in hopes that the outcome of this project produces some results in performing exactly this selection task. [1]

2 Select Literature Summaries

2.1 ArXiv Papers and Citation Graph

This work by (Clement et al., 2019) explores using academic papers on ArXiv as a research dataset. They build a publicly available script that extracts metadata and text from ArXiv's data dump, and a pipeline to extract a citation graph from processing the data. The raw data dump amounts to 1.4TB of pdf files.

The resultant processed dataset is 1.35 million articles at the time of publication, totalling 11 billion words. Further, the citation graph contains 6.7 million edges, and only 62% weakly connected components, making this dataset significantly larger and denser than other popular dataset that provides a citation graph. This study aims to enable further research in text segmentation, link prediction, and research trend prediction.

2.2 Document Similarity with Topics

The work by (Gong et al., 2019) attempts the problem of concept-project mapping without any deep model, producing a pipeline that maps a full-text document to a target summary document. The authors proposes a model that preprocesses both the document and the summary to generate a topic mapping, which in turn evaluates into relevance.

The authors propose a method of mapping topics into documents by projecting the vectors onto a linear space, then taking the intersection of the spaces. They experiment with topics generated by word2vec on science topics extracted from a corpus of science and discusses the result.

2.3 BERTs

Bidirectional Transformers for Language Understanding (BERT) is a powerful model developed by researchers at Google (Devlin et al., 2018). Originally developed to tackle popular NLP tasks such as

[1] To satisfy the guideline of the literature review assignment, I separate these readings into 5 topics.

GLUE and SQuAD, the model proved to be applicable to many other NLP tasks as well. RoBERTa (Liu et al., 2019) is a robustly-optimized version of BERT.2

The model is computationally expensive to train, however, there are many publicly-available trained BERT models ready to be fine-tuned for specific tasks. Huggingface (Wolf et al., 2019) provides an easy access to trained small and large BERT models. This blog post by (Keita, 2019) discusses ways to pool or use BERT's [CLS] token for encoding variable length text.

One possible limitation is that BERT is constrained to 512 tokens, and there exist papers with a longer abstract.

2.4 SBERT

In the work by (Reimers and Gurevych, 2019), the aurthors propose Sentence-BERT (SBERT), a BERT-based model fine-tuned for semantic text similarity tasks, outperfoming Google's Universal Sentence Encoder (Cer et al., 2018).

SBERT is a BERT model trained with a Siamese structure. For the variaous tasks the aurthors tackled, the authors propose three training methods: softmax classifier on the concatenated BERT output vectors, regression on the cosine similarity on the BERT output vectors, and sentence embedding trained with triplet loss.

Finally, the authors experiments with RoBERTa in comparison to BERT, and also experiment with both [CLS] token and averaging BERT embedding vectors across the token dimension.

2.5 Document Retrieval with BERT

In the paper titled Simple Applications of BERT for Ad Hoc Document Retrieval (Yang et al., 2019), researchers experiments with a simple document retrieval task of searching for related social media posts and news articles. The system is given a short query text, to produce a ranking of documents within a corpus.

The authors approaches the problem in the most obvious manner, by fine-tuning BERT to a simple sentence inference task similar to BERT's original next-sentence prediction, with [SEP] token. This project shows promising result in improving search relevance, with certain limitation in data and possibly inefficiency.

3 Compare and Contrast

Aside from the first paper on using ArXiv as a dataset, all other work tackle sub-problems of text similarity / relation. In (Gong et al., 2019), the attempt is to use existing topic embeddings learned from an external corpus, and these topic are predefiend by human. BERT lays the ground work for the deep models. In (Yang et al., 2019), the attempt is to use BERT in the most straightforward way by forwarding both the query and the candidate text, separated by the [SEP] token. Finally, in SBERT (Reimers and Gurevych, 2019), the authors are primarily concerned with the real-time performance of the deep model when given a query to search through candidates, and proposes a solution by vector similarity trained on a Siamese BERT. The results of each paper are not comparable as they do address different sub-problems and evaluates on different tasks and data.

4 Future Work

(Clement et al., 2019) provides a pipeline for the text dataset and a citation graph, this lays down the ground work in acquiring the dataset for my project. For this dataset, I propose to use the edge in the citation graph as a proxy to ”academic related-ness”, and build a model around it. In the aforementioned deep models, the problem of ”text / topic similarity” is approached as a problem of supervised learning, therefore as a first step, this project will explore how to use the citation graph to generate a good label for training.

(Yang et al., 2019) uses BERT in the most straightforward way, and I propose to extend (Reimers and Gurevych, 2019)'s work by further experimenting with various classifiers, while applying the models to the ArXiv dataset, to measure the related-ness between academic papers or abstracts, or the mixture of the two. I hypothesize that the measure of ”similarity” as used by SBERT such as one optimized by triplet loss, is not as good in suggesting papers-to-cite, as intuitively cited works are somewhat related to the literature, but not strictly similar.

Finally, as a stretch goal. I will use the result of the above to build a simple recommender system, to retrieve related articles in the database given a text query. I hypothesize that there will be some performance issues, and plan to explore how to resolve them.

References

Daniel Cer, Yinfei Yang, Sheng yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. Universal sentence encoder.

Colin B. Clement, Matthew Bierbaum, Kevin P. O'Keeffe, and Alexander A. Alemi. 2019. On the use of arxiv as a dataset.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding.

Hongyu Gong, Tarek Sakakini, Suma Bhat, and Jinjun Xiong. 2019. Document similarity for texts of varying lengths via hidden topics.

Kurita Keita. 2019. Paper dissected: "bert: Pre-training of deep bidirectional transformers for language understanding" explained.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing.

Wei Yang, Haotian Zhang, and Jimmy Lin. 2019. Simple applications of bert for ad hoc document retrieval.