Finding Related ArXiv Papers with BERT

CS224U Project - Literature Review

Junshen Kevin Chen
Stanford University
jkc1@stanford.edu

Abstract

I propose a system that evaluates the relevancy between academic papers using text encodings from BERT, using a dataset of ArXiv papers, and the citation graph extracted from it. This document contains the findings of my preliminary literature search, and some discussion among them.

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 abstract as query.

The task of "suggesting related work" appears to be under-studied based on my cursory literature search, with some preliminary experiment with the topic and the corresponding conclusions. As an on-demand document retrieval system, using deep models can introduce many problems in its inefficiency. However, not unlike product recommendation in a shopping site or a friend suggestion on a social media, which already heavily utilizes forms of deep learning to achieve the task, document recommendation can be similarly formulated. Furthermore, paper citations produeces an graph not unlike users in a social media.

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 transformer models provide powerful ways to encode semantics, and trained models are readily available to be fine-tuned for this specific task. Google has announced recently that it is using BERT to augment search result ranking, for a subset of the queries that are more semantic-specific (Nayak, 2019).

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.

2 Dataset

On the Use of ArXiv as a Dataset

Through its participation in the Open Archives Initiative, arXiv makes its data openly available. The data released includes metadata such as date, author, title, category, and most importantly, text abstract. The paper itself is also released as a data dump of pdf files, including all papers on arXiv since its founding in the 90s.

The work titled *On the Use of ArXiv as a Dataset* 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, and takes around $130 and two days to download in entirety [1].

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. The co-citation network can be used for taskes such as relationally powered classification, author attribution, segmentation, language modeling and many other NLP tasks. This research is intended to lay the ground work for further research in the field using such data.

The authors also briefly experiments with a simple logistic regressive model to predict the category of the text from simple features extracted from the citation graph, achieving 94.5% accuracy at the highest.

Future Work

I used the pipeline produced by this research to download and process the data, and construct a citation graph. I intend to use this data to train any model towards the goal of this project.

3 Transformer Models

BERT

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 GLUE and SQuAD, the model proved to be applicable to many other NLP tasks as well.

Unlike previous transformer modles such as GPT, BERT attends to all positions in the input tokens, not just ot its left. It is trained as a Masked Language Model (MLM) by masking out a portion of the input to predict the output, and in a next sentence prediction task.

One main advantage of BERT's structuer is that it can be easily fine-tuned to many other language tasks, as we can simply attach a classification layer to the end of the transformer output and update all the pre-trained parameters. The authors also discuss feature-based approaches and show that BERT is equally effective.

RoBERTa

Robustly Optimized BERT Pretraining Approach (RoBERTa) (Liu et al., 2019) is an optimized version of BERT. The authors experiments with different methods of masking the MLM, optimizers, and more training data to the original model, to achieve a more robust version of BERT. The results show that RoBERTa is able to outperform BERT in numerous NLP tasks.

DistilBERT

DistilBERT is lite version of BERT, developed by the researchers at Huggingface (Sanh et al., 2019). Aside from the reduced number of learnable parameters, the main difference of DistilBERT is that the token-type embedding and poolers are removed, while reducing the number of layers. In practice, when evaluated on the same NLP tasks, DistilBERT retains 97% of vanilla BERT's performance, while being 40% smaller and 60% faster.

Using the models

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, RoBERTa, DistilBERT models weights, and therefore we can easily apply transfer learning and adapt the pre-trained models for this particular task.

Encoding texts of variable length (as paper abstracts varies in length), and performing a prediction task, we must produce an output that is consistent in dimension. This blog post by (Keita, 2019) discusses ways to pool or use BERT's [CLS] token for encoding variable length text.

4 Text Similarity

4.1 Document Similarity with Pre-defined 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

[1] I drastically lower this cost by working directly with AWS instances on the same region the data dumps are hosted

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.

4.2 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).

As a driving goal to this research, to find the most similar pair in a 10000 sentence corpus requires 50 million forwards with BERT, taking an impossibly long time. The author approaches problem by applying transfer learning on a pre-train BERT to output embedding vectors that are close (in terms of cosine distance) when input sentences are related, and far when the are not related, such that each sentence from the copus only need to be pre-computed once.

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. The output of the model is then a variable-length embedding, which the authors experiment with both [CLS] token and averaging BERT embedding vectors, then calculating the corresponding cosine distances. Finally, the authors experiments with RoBERTa but without significant improvement. This results in a model defeating Universal Sentence Encoder not only in accuracy but also in speed.

4.3 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.

5 Compare and Contrast

On using the dataset

This dataset creates the ground work for many further research as it produces a pipeline that constructs a citation graph from the raw paper texts, which allows us to interpret academic research relatability in the form of co-citation.

However, co-citation may not entirely be accurate in predicting similarity, as paper authors tend not to cite work that overlaps with their own for obvious reasons: either because they are unaware that such work has already been done, or that they are intentionally omitted as they undermine the author's own work. Therefore, from the perspective of a search-engine like system, building from co-citation may not be best in finding similar papers.

Co-citations likely occur because the literature builds onto an existing work, or is an alternative approach, etc. Even though they may not be "similar" in the strictest sense, considering that the goal is to suggest paper that may be related to a topic (as described by a query text) using co-citations is a good choice. A hypothetical perfect system built using this information will suggest papers the author is likely to cite.

On using BERT for text similarity

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. As it is infeasible to humanly define topics to academic papers beyond a notion of "field" (as is defined by ArXiv's default categories), not only because it is non-trivial amount of work, but also due to the ever-changing nature of academia.

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, 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 re-

sults of each paper are not comparable as they do address different sub-problems and evaluates on different tasks and data. (Reimers and Gurevych, 2019) lays down the proof-of-concept in using bert-encoding as a measure of semantic similarity. One possible concern of using BERT for academic papers is that it is possible for the abstract text to exeed BERT's limit of 512 tokens.

6 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.

Pandu Nayak. 2019. Understanding searches better than ever before.

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

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter.

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.