Deep Reading of a Concept

Keerthi Kumar Kallur
Dept. of Computer Science
Stony Brook University
112608121
kkallur@cs.stonybrook.edu

Abdullah Mitkar
Dept. of Computer Science
Stony Brook University
112685069
amitkar@cs.stonybrook.edu

Vanessa Singh
Dept. of Computer Science
Stony Brook University
112657002
vassingh@cs.stonybrook.edu

Abstract

This project implements an approach in which a question answering model is trained on a large data that contains information about a particular topic. The model will use this knowledge to improve its embeddings accordingly to focus on that knowledge and answers all the questions relevant and related to the topic. Instead of training on a large dataset that discusses multiple concepts we train the model on a data that is focused on one topic to analyse if the model can understand that topic in-depth. This project will enhance the BERT model and try to answer SQuAD questions to validate deep reading of a concept.

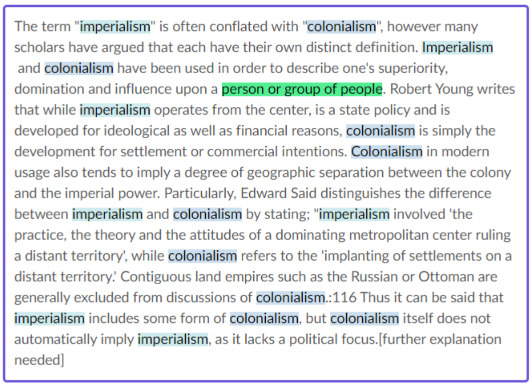
1 Introduction

The task of question answering (QA) is quite popular in recent time. There are various question-answering models that perform satisfactorily on question answering tasks spanning over various topics. However, these models come with their share of problems. They are not able to answer questions on a specific topic in depth. Our project focuses on trying to address the above stated problem. In the real world, deep reading of a concept would hugely impact systems like automated disease diagnosis, etc. Deep Reading models should be able to extract more information from each paragraph present in the data. We would like to create a model that is able to perform deep learning of a topic and is able to answer questions on a particular topic. The challenges associated with this task are: how well will the model be able to answer inference questions, will the model be able to understand what is the sub-topic being discussed in a particular paragraph.

There are multiple approaches that target the Question and Answering problem efficiently such as Stanford Question Answering Dataset[7, 8] and Explainable Multi-hop Question Answering[10]. Since this is a new topic, we didn't find any articles on deep reading of a topic. One paper that gave us some direction was BioBERT[4]. The ALBERT[1] model was also referred while implementing the project.
SQuAD[7] and Multihop QA models are trained on huge datasets which talk about a lot of topics which might confuse the model while answering deep questions about a specific topic. This is overcome by training the model on lots of data on one topic. The idea behind the project is to make the model understand one topic very well rather than understand lots of topics superficially. BioBERT[10] is similar to our idea but it discusses a model that will do very well on biomedical data whereas our aim is to build a model that can do well on any topic specific data.

In this project, we try to improve the weights of a pre-trained BERT model that can help us answer some of the topics. BERT uses a "masked language model" (MLM) pre-training objective, inspired by the Cloze task[9]. The masked language model randomly masks some of the tokens from the input, and the objective is to predict the original vocabulary id of the masked word based only on its context. Unlike left-to-right language model pre-training, the MLM objective enables the representation to fuse the left and the right context, which allows us to pre-train a deep bidirectional Transformer.

We are using two approaches to bridge this gap - using language modelling on BERT, using transfer learning on language model and QA model. The model will understand embeddings from the BERT model which will be fine tuned on the data specific to one concept. This will help the model gain a thorough understanding of the topic which will enable QA model to answer factoid questions better. It should also improve the answering of comparison, inference, definition questions.



The term "imperialism" is often conflated with "colonialism", however many scholars have argued that each have their own distinct definition. Imperialism and colonialism have been used in order to describe one's superiority, domination and influence upon a person or group of people. Robert Young writes that while imperialism operates from the center, is a state policy and is developed for ideological as well as financial reasons, colonialism is simply the development for settlement or commercial intentions. Colonialism in modern usage also tends to imply a degree of geographic separation between the colony and the imperial power. Particularly, Edward Said distinguishes the difference between imperialism and colonialism by stating; "imperialism involved 'the practice, the theory and the attitudes of a dominating metropolitan center ruling a distant territory', while colonialism refers to the 'implanting of settlements on a distant territory.' Contiguous land empires such as the Russian or Ottoman are generally excluded from discussions of colonialism.:116 Thus it can be said that imperialism includes some form of colonialism, but colonialism itself does not automatically imply imperialism, as it lacks a political focus.[further explanation needed]

Figure 1: Passage on Imperialism in SQuAD 2.0 dataset

On reading the above passage on Imperialism our model should be able to find similarity between Imperialism and Colonialism without getting confused as to what is the main subject of the article.

The results of our model were compared to the BERT model and on some chosen topics from the SQuAD dataset. We compared the results of LM on BERT followed by QA and transfer learning on language modelling and question answering.

The three outcomes of this project are:

1. We experimented on bert layers to focus on specific topics

2. Our evaluation shows that for a set of topics, our QA model's accuracy improved by 2 percent.

3. Based on our work, we can say that making BERT weights to ficus on topics has improved the performance of QA on that topic.

2 Your Task

The task is to assess the influence of additional data related to a topic on QA performance. We are using the BERT implementation of the QA task and modifying the BERT layers to learn the embeddings of a topic. The key challenges in the task are:

- Influencing the BERT layer to focus on specified topics.
- Analysis of results on SQuAD (Which has large chunk of data), difficult to analyze the effects and repercussions.

BERT [2] is currently the best performing system, which is rigorously trained on heavy GPUs over Wikipedia data. BERT has provided some pre-trained models which can be used to change the final layer and customize it to our objective.

The high level diagram of the process is below, in which the end-to-end steps planned for the task is mentioned.

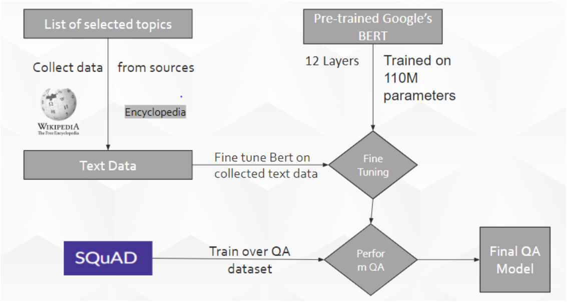


Figure 2: Flow

2.1 Baseline Model(s)

The baseline system used here is the BERT model [2], which is trained on Wikipedia data. Here at each layer, the model learns the embdedings by doing word prediction on masked data. Unlike many word embeddings, the attention is given to both sides of the context. This way the BERT takes even positional embeddings into consideration. This is proven to perform much better than older models like Word2Vec[5] and GLoVE[6]. Please refer below image for the architecture.

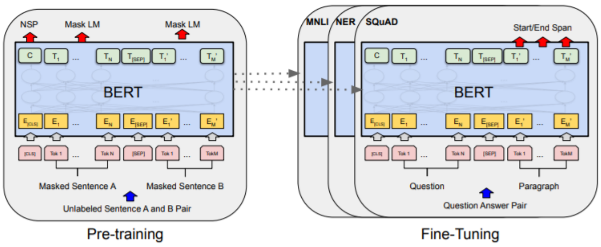


Figure 3: Bert Architecture from [2]

The left side in the above figure is the pre-trained BERT model, which allows us to fine tune it based on our requirement. It can be used to do a QA task, classification task, etc.

3 Your Approach

BERT base case embedding performs excellent on question-answering task when trained on the SQUAD 2.0 data set[8]. So we decided to take those embeddings as the weight initializer for our models. From this base model, stems the 2 ideas that we have tried.

3.1 Idea 1:

Do Language Modeling Task on BERT base embeddings for the topic in hand and then do question answering task on them.

3.2 Idea 2:

Do Language Modeling task on BERT base embeddings for the topic in hand and and do question answering task on the BERT base embedding. Next we do transfer learning where we take the 8 layers of the Question Answering model and the lower 4 layers of the language modelling model and concatenate these layers to form a final model.

3.3 Implementation Details

For the implementation of this project, we leveraged the HuggingFace github repository that contains the scripts and code for working with BERT and other models and doing the language tasks and question answering task.

3.3.1 Idea 1: LM on BERT base followed by QA

To implement this idea, we used the ıbert-base embeddings and a corpora of text 600 to 800 sentences for a topic.

- Initial model with bert-base weights.
- Do fine tuning with the masked LM task using the corpora of the topic
  - Epochs: 16
  - Base model: Bert base with 12 layers
- On this model, we ran the QA task.
  - Epochs: 2
  - Base model: Output of previous step

3.3.2 Idea 2: Transfer learning on LM model and QA model

We perform the language modeling task and the QA modeling tasks separately and then merge the 'QA component' and the 'language component' to create a hybrid model.

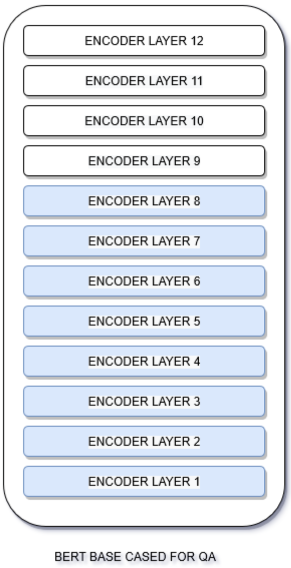


Figure 4: Encoder of the BERT QA model

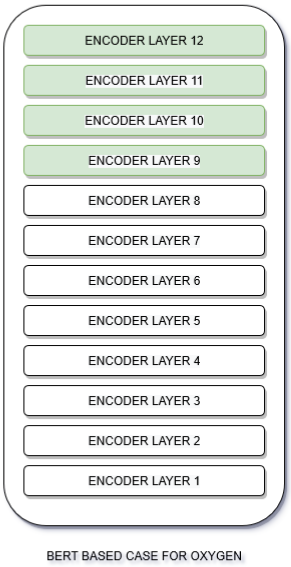


Figure 5: LM model

- Initial model A with bert-base weights.
- Do fine tuning with the masked LM task using the corpora of the topic for model A
  - Epochs: 16
  - Base model: Bert base with 12 layers
- Initial model B with bert-base weights.
- Do fine tuning with the QA task using the squad training dataset for model B
  - Epochs: 2

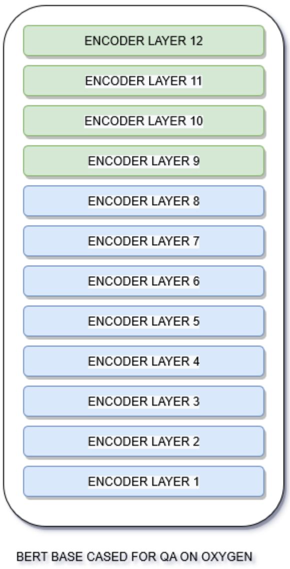


Figure 6: Hybrid model

  - Base model: Bert base with 12 layers

- Merge the top 8 layers of model A and the lower 4 layers of model B

4 Evaluation

Since this is a Question Answering task, we relied on the SQuAD Dataset for the training as well as the evaluation data. We first split the Squad evaluation dataset based on the topic at hand. We then ran the state of the art BERT base case model on this dataset and then our model. We then compared the results.

4.1 Dataset Details

For the Question Answering task, we used the SQuAD dataset for the training as well as the evaluation dataset. SQuAD picks passages from Wikipedia on which it poses some questions. These questions an be factoids, inference questions etc. These passages contain information on diverse topics.

For the Language Modeling task, based on the topics from the evaluation dataset, we extracted 600 to 800 lines from the internet sources like Wikipedia and Encylopedia and created sentences of 1 line each. In our model fine tuning has been tried for the following topics - Force, French Indian, Imperialism, Southern California, Sky, Islamism, Oxygen.

4.2 Evaluation Measures

- F1 score is used as the evaluation metric to measure accuracy of the answers.

- Exact match is used as a metric to measure the percentage of predictions that match any one of the ground truth answers exactly.

4.3 Baselines

The baseline model is a general question answering model that is trained by initialised weight from google bert-base-cased(12 layer, case sensitive) embeddings. The embedding are trained on the SQUAD training dataset for 2 epochs with a learning rate of –learning_rate of 3e-5 and a maximum seq length of 384 and doc stride of 128[Splitting the passage into 128 words].

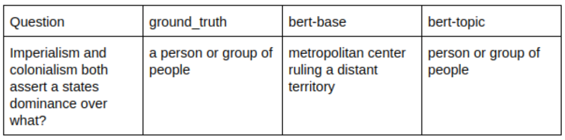
The bert-base-cased model was used as it is a light model that contains almost 97% of features of bert-large. We had to train models for several time for different topics and to save time we decided to use the bert-base-cased model.

The models were tuned for Masked LM task using the same strategy using in BERT's training this. The scripts to do this are provided in huggingface's repository mentioned above. The same for fine tuning for SQUAD.

4.4 Results

	bert		bert-topic	
Topic_Epochs	exact	f1	exact	f1
force_8	81.68	87.49	80.19	87.04
FrenchIndian_16	66.86	81.21	68.02	82.63
Imperialism_16	74.33	82.66	75.4	84.12
Imperialism_32				
Scalifornia_16	73.33	82.19	76.66	85.28
Sky_8	83.33	92.44	83.33	92.9

This result shows that our model when trained on a topic with 16 epochs is able to answer questions better than the bert-general-qa model.



Question	ground_truth	bert-base	bert-topic
Imperialism and colonialism both assert a states dominance over what?	a person or group of people	metropolitan center ruling a distant territory	person or group of people

The above example shows that our model(bert-topic) is able to answer questions on Imperialism when it is trained on data specific to Imperialism.

4.5 Analysis

A Question Answering model trained on a SQUAD data set is able to fairly answer a lot of questions on several topics. What we wanted was to create a model that is able to answer more questions on one topic. The model must read some text and be able to answer questions on the topic based on the extract. For this, we tried 2 approaches. Let us analyze the approaches separately.

4.5.1 Transfer Learning

We tried to train 2 models separately and then merge their QA components and embedding components of BERT. This model however did not work as we can see from the results.

The layers of BERT can be understood in terms its parallels with Convolutional Neural Network. The lower layers of BERT capture the semantics or the grammar of the language and as we move far towards the upper layer it tends to capture more wider information of the text. Hence, the lower 4 layers can effectively represent the embeddings.

When we have two models trained separately, they each capture the minute details(lower layers) and broader picture(upper layers) of 2 different data set. Hence, combining the broader details of one model with minute details of another models cannot essentially help in solving this problem at hand.

4.5.2 Language Modeling followed by Question Answering

This approach involves teaching the how to understand the language(MLM) using the masked language modelling task. Now that the model has fairly understood the language, we can ask this model to learn the question answering task and predict the span. Using this, the model which is initialised on the understand of a particular language and after the training is complete, it is able to learn question answering from there. Hence, after performing question answering for a little longer time, the model is able to learn question answering and perform fairly better than the bert's general question answering model.

4.6 Code

The code having the experiments and analysis performed in this project is available at this github repository.

5 Conclusions

Extracting layers from BERT base model, fine tuning it on topic specific data has a positive effect on the model and the Question Answering task is able to perform better and generate more accurate answers.

6 Future Scope

We have influenced the layers of BERT to improve our topic - specific QA model. There is scope of improvements for few tryouts and analysis in terms of performance and generalization.

- Performance of the model if it is made to focus on multiple topics at once. This might lead to the case where the weights of BERT are generalized. Bringing us back to initial stage of the experiment.

- GAN model[3] Adversarial training of the QA and the LM model can also be tried. The model for LM can be train simultaneously with Question Answering and both can effectively make each other better by reducing each other's losses. The LM task can strive to make better representation to reduce the loss of the question answering model. The QA model can in-turn make predictions on this model and the accuracy of these prediction can be used in training the LM task. In such generative setting, the LM task can create a representation of the language that performs excellent on the QA task.

References

[1] Zhenzhong Lan1 Mingda Chen, Sebastian Goodman1 Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations. 2019.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.

[3] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, pages 2672–2680, Cambridge, MA, USA, 2014. MIT Press.

[4] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. 2019.

[5] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc., 2013.

[6] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *In EMNLP*, 2014.

[7] Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for squad. 2018.

[8] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. 2016.

[9] Wilson L Taylor. "cloze procedure": A new tool for measuring readability. 1953.

[10] Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. 2018.