Transfer learning for Question Answering task

Srilakshmi Alla
srialla@utexas.edu

Abstract

Humans can answer questions across many domains. To build a human intelligent question answering (QA) system they should obtain robust performance across multiple domains. The pre-trained models liked DistilBERT showed state of art performance on SQuAD (em[1] : 77, f1[2] : 85.4) but didn't generalize well on domain specific data sets like BioASQ (em : 38.4, f1 : 53.1). This shows the current QA systems degrade across domains. In this paper, I explored pre-trained models with domain knowledge to optimize performance on both SQuAD and BioASQ.

BioBERT, a pre-trained BERT model trained on Pubmed and PMC articles achieved similar performance to DistilBERT on SQuAD and improved performance on BioASQ by 7% em and 8% f1 . The larger version of BioBERT in comparison to DistilBERT improved performance on SQuAD by 8% em and 6% f1. For BioASQ, it improved by 23% em and 21% f1 score. In this paper, I also presented qualitative analysis on predictions in addition to training and evaluating different models. I highlighted the downsides of some of the approaches and ways it can be addressed in future.

1 Introduction

We are in an era where we generate lots of unstructured text data everyday like medical records, research articles, news articles and many more. In recent years, there has been a growing interest in training machines learn and perform natural language processing (NLP) tasks like summarizing, question answering (QA) and information retrieval.

Recurrent Neural Networks (RNNs), especially long short term memory (LSTM) and gated recurrent unit (GRU) have been state of art techniques in performing NLP tasks. Architectures like encoder-decoder with attention extended the performance. This kind of design works great for a short length sequence but the complexity increases tremendously with longer inputs when using a sequential model like RNN.

With the introduction to transformers (Vaswani et al., 2017), RNNs were replaced with self-attention layers. This architecture compared to encoder-decoder with RNNs offers parallel processing, less complexity for layer and shorter path length to learn long-range dependencies in the network. There are different kinds of representation of transformers.

Bidirectional Encoder Representation for Transformers (BERT) (Devlin et al., 2018) is one representation of transformers. It is designed to pretrain bidirectional representations by taking both left and right context in all layers into consideration. As a result, by adding one additional layer to any pre-trained BERT model can be fine tuned for a wide range of tasks, such as question answering (QA), text classification, named entity recognition, relation extraction etc.

For various NLP tasks including QA, pre-trained language models achieved state of the art performance when fine tuned on a given task. However, most of the models are pre-trained on general domain corpora, which cannot be generalized to specific domain corpora. Hence, for a QA system to achieve good performance we need a model pre-trained on domain specific corpora.

Recently, Lee et al. (Lee et al., 2019) proposed BioBERT, a pre-trained BERT model trained on PubMed and PMC articles. BioBERT similar to BERT can be fine tuned to any data set. This model has proven to outperform BERT when answering biomedical related questions. Many researchers leveraged this model to even address questions related to global pandemic (Das et al., 2020). That motivated me to explore BioBERT on BioASQ data

[1] exact match
[2] f1 score

set.

In this project, I fine tuned a pre-trained BERT model DistilBERT (Sanh et al., 2019) on SQuAD, and used pre-trained BioBERT (Lee et al., 2019) on SQuAD to address question answering task for out of domain questions.

The rest of our paper is organized as follows. First, I introduced the data sets. Second, I introduced the models used in this project. Then I presented conclusions and future scope followed by the experiments, results with analysis.

2 Data sets

To promote research in area of QA, Machine Reading for Question Answering (MRQA)[3] testbed released a shared task focusing on generalization in reading comprehension (Fisch et al., 2019). This task comprised of three data sets: SQuAD (Rajpurkar et al., 2016), NewsQA (Trischler et al., 2016), BioASQ (Tsatsaronis et al., 2015). SQuAD is a Stanford Question Answering data set consisting of questions on set of Wikipedia articles. NewQA on new articles and BioASQ on biomedical text. All these data sets contains contexts, questions and span of text from context as answers.

The data sets I want to focus on in this project are SQuAD [4] and BioASQ[5]. The data sets used in this work are downloaded following instructions provided in here [6]. This copy of data, have train and dev data sets for SQuAD and only test data for BioASQ. I used SQuAD train to train the models and SQuAD dev and BioASQ test to evaluate the models.

3 Models

This section covers the architectures of the models used in the project, DistilBERT, BioBERT and BioBERT-Large.

3.1 BERT

BERT is pre-trained on BooksCorpus (800M words) and text passages from English Wikipedia (2,500M words).

$\textbf{BERT}_{\textbf{BASE}}$ has 12 layers of transformer blocks with 12 attention heads and 110 million parameters where as $\textbf{BERT}_{\textbf{LARGE}}$ has 24 layers with 16 attention heads and 340 million parameters.

[3]https://github.com/mrqa/MRQA-Shared-Task-2019

[4]https://rajpurkar.github.io/SQuAD-explorer/

[5]http://bioasq.org/

[6]https://github.com/gregdurrett/nlp-qa-finalproj

3.1.1 DistilBERT

DistilBERT [7] (Sanh et al., 2019) is a distilled version of BERT which is faster, smaller, cheaper and lighter in comparison with BERT. It is trained on same corpa as BERT. The use of larger pre-trained models for solving problems in NLP has become prevalent. As it involves lot of resources, Hugging face (Devlin et al., 2018) came up with DistilBERT which is 60% faster, 40% smaller version of BERT with sacrifice of less than 1% of performance metrics.

DistilBERT has 6 transformer layers, 12 attention heads with embedding size of 512 and hidden size of 3072. It has a vocabulary of 30k.

3.2 BioBERT

BioBERT (Lee et al., 2019) is one of first few domain specific pre-trained BERT model. It is pre-trained with PubMed Abstracts (4.5B words) and PMC Full-text articles (13.5B words).

BioBERT model used in my experiments is based on BERT-Base Architecture for QA[8]; It has 12 transformer Layers , 12 attention heads with embedding size of 512 and hidden size of 768. The size of vocabulary is 29K. This model uses word-piece vocabulary same as BERT model.

3.3 BioBERT-Large

BioBERT-Large model used in my experiments is based on BERT-Large Architecture for QA[9]; It has same architecture like BioBERT with 12 additional transformer Layers , 4 additional attention heads with embedding size of 512 and hidden size of 1024. The size of vocabulary is 30K more than BioBERT.

4 Experiments : Factoid Question Answering Task

The hypothesis behind this project is that the models pre-trained on biomedical corpa when fine tuned exceeds performance on BioASQ data set compared to any other model pre-trained on general corpa. In this project, I used RNN based architecture as general baseline, DistilBERT (Sanh et al., 2019) as a BERT baseline and BioBERT, BioBERT-

[7]https://huggingface.co/distilbert-base-uncased/blob/main/config.json

[8]https://huggingface.co/dmis-lab/biobert-base-cased-v1.1-squad/tree/main

[9]https://huggingface.co/dmis-lab/biobert-large-cased-v1.1-squad/tree/main

Model	Transformer layers	Attention heads	Embedding size	Hidden size	Vocabulary size
DistilBERT	6	12	512	3072	30k
BioBERT	12	12	512	768	29k
BioBERT-Large	24	16	512	1024	59k

Table 1: Model Parameters

Large (Lee et al., 2019) as pre-trained models on biomedical corpa.

For Factoid QA task, both SQuAD and BioASQ data sets are pre-processed to have question, context, answers and the question id. Then the context and question are fed to model's tokenizer. The tokenizer separates context and question with [SEP] token, appends [CLS] token at the beginning and [SEP] tag at the end of the tokenized output. The tokenized data is fed into the model.

The pre-trained model is fine tuned with a QA data set which met the requirements. The fine tuned model is evaluated using both SQuAD and BioASQ data sets. The output needs to be post processed to map it back to the context. The model outputs the start and end logits of the answers. It outputs n best possible predictions. The logits are scored by adding the start and end logits. To pick best indices of the answer, the scores are sorted. The indices with best score are spanned on to the context to predict the answer. The performance metrics like exact match and F1 score are used to compare different models.

The following experiments are conducted to verify my hypothesis. Table 2 shows performance metrics from the experiments.

4.1 General baseline

The general baseline model used in this project is adapted from DrQA reader (Chen et al., 2017). It uses bidirectional LSTM to encode passages and a fixed vector to encode questions. The model outputs start and end spans in the passages as answers.

I trained baseline model on SQuAD train data set and tested it using both SQuAD dev data set and BioASQ data set. For SQuAD, baseline model gave exact match of 49.03 and F1 score of 61.41. For BioASQ, it gave em of 11.5 and F1 of 20.

4.2 BERT baseline

DistilBERT is used as BERT baseline in this project. As DistilBERT is faster and smaller in comparison with BERT so I leveraged to use DistilBERT as my BERT baseline.

I fine tuned DistilBERT model on SQuAD train data set using 3 epochs with a learning rate of 2e-5 and weight decay of 0.01. I tested it using both SQuAD dev data set and BioASQ data set. For SQuAD, BERT baseline achieved em of 77 and F1 of 85.4. For BioASQ, it achieved em of 38.4 and F1 of 53.1.

DistilBERT learned context from large scale text data in Wikipedia articles and when fine tuned on large QA data sets like SQuAD, model performance boosted on both SQuAD and BioASQ data set in comparison to DrQA model. Though, there is improvement on BioASQ data set, its no where compared to performance on SQuAD data set.

As SQuAD data set contains more general questions, DistilBERT was able to perform better. As BioASQ data set contains more specific questions involving biomedical context and terminology which we don't use in general, DistilBERT seems to under perform on BioASQ data set.

4.3 BioBERT

In previous section, we observed that DistilBERT couldn't perform on BioASQ as well as it performed on SQuAD data set. The reason for this behaviour is DistilBERT is lacking the knowledge of biomedical context.

BioBERT is a BERT pre-trained model on large corpa of PubMed and PMC articles. For this experiment, I took on the shelf BioBERT fine tuned to SQuAD model and tested the model on SQuAD dev data set and BioASQ data set. For SQuAD, BioBERT achieved em of 75.9 and F1 of 84.8. For BioASQ, it achieved em of 45.7 and F1 of 61.8.

BioBERT performance on SQuAD data set is comparable to DistilBERT but it performed much better on BioASQ data set compared to DistilBERT.

4.4 BioBERT-Large-SQuAD

BioBERT-Large has similar configuration as BioBERT with 30k more vocabulary. For this ex-

Model	Train data set	Test data set	Exact Match	F1 score
Baseline	SQuAD	SQuAD	49.0	61.4
		BioASQ	11.5	20.0
DistilBERT	SQuAD	SQuAD	77.0	85.4
		BioASQ	38.4	53.1
BioBERT-SQuAD		SQuAD	75.9	84.8
		BioASQ	**45.7**	61.8
BioBERT-Large-SQuAD		SQuAD	85.4	91.6
		BioASQ	**61.8**	74.1

Table 2: Results

periment, I took on the shelf BioBERT-Large fine tuned to SQuAD model and tested the model on SQuAD dev data set and BioASQ data set. For SQuAD, BioBERT achieved em of 85.4 and F1 of 91.6. For BioASQ, it achieved em of 61.8 and F1 of 74.1.

As BioBERT-Large is a larger model with more vocabulary, this model learned and performed better than BioBERT and DistilBERT on both the data sets.

4.5 BioBERT-SQuAD fine tuned on BioASQ

In all the previous sections, all the models are fine tuned on SQuAD and we saw the models perform really well on SQuAD dev data set. If we can fine tune BioBERT models on BioASQ data set, the performance of the model can improve on BioASQ.

BioASQ is a data set of only 1500 samples. That is a small data set to fine tune any model. To mitigate the issue of small data sets, we first fine-tune all the models (DistilBERT, BioBERT and BioBERT-Large) on a large-scale extractive question answering data sets like SQuAD, and then fine-tune it on BioASQ data sets to transfer learning.

Out of 1504 questions, there are only 224 unique questions in BioASQ data set. If we randomly do train test split, we may end up with same question answer pairs in both train and test data which results in data leakage. One solution to this problem is to do strategically split data and maintain different sets of questions in train and test data.

Note: When attempted to implement this method, I ran out of computational resources so these results are not reported.

4.6 Quantitative Analysis

There is a significant boost in the exact matches and F1 scores from DistilBERT to BioBERT and BioBERT to BioBERT-Large. In this section, I want to present some analysis on answers predicted by these models.

The hypothesis behind using a model pre-trained on biomedical data is having prior knowledge of the domain can help model predict the answers correctly. Example in Table 3 supports this hypothesis. BioBERT and BioBERT-Large, having domain specific contextual knowledge picked up the right answer whereas DistilBERT picked an answer matching a word in question "Willis".

Table 4 shows the number of questions each model predicted correctly. Out of 1504 questions, all the models predicted 417 questions correctly. BioBERT and BioBERT Large predicted 202 questions correctly that are missed by DistilBERT . This proves the point that having biomedical knowledge increases exact matches. BioBERT-Large alone predicted 201 questions which other two models missed. It clearly states that having more vocabulary can help the model predict better. Surprisingly, DistilBERT and BioBERT Large predicted 109 questions correctly which BioBERT missed. I assume that the BioBERT was trained with a smaller set of documents and vocabulary.

5 Conclusions

In this project, I took an ablative approach to address some of the questions. First question is does a model performs better on adding domain knowledge. From my experiments, BioBERT which has prior knowledge of the domain outperforms DistilBERT on BioASQ. Second question is does adding more vocabulary boosts the performance. BioBERT-Large which has 30k more vocabulary than BioBERT improved exact match by 16% and F1 score by 13%.

These experiments helped me conclude that to achieve a multi-domain QA system, the model

Type	Description
Question	'Willis-Ekbom disease is also known as?'
Passage	"The article briefly summarizes the milestones leading to current knowledge and the possibility of treating one of the most widespread and perhaps least known diseases, restless legs syndrome (RLS). Until the mid-twentieth century, the syndrome first described by Willis (1685), was sporadically reported in medical literature and in most cases deemed a bizzare condition. It was only with Ekbom's detailed clinical description of the syndrome (1944) and the polygraphic recordings of Coccagna et al. (1962) that RLS became well-recognised clinical entity. Since then, almost all sleep laboratories have devoted much of their research to discovering the pathogenetic mechanisms underlying the disease and devise increasingly specific treatment. Major advances have been made in recent years, but a full understanding of RLS is still a long way off."
Ground Truth	'restless legs syndrome'
DistilBERT	'the syndrome first described by willis (1685), was sporadically reported in medical literature and in most cases deemed a bizzare condition'
BioBERT	'restless legs syndrome'
BioBERT-Large	'restless legs syndrome'

Table 3: Predictions by DistilBERT and BioBERT QA system on the BioASQ factoid data set

Distil BERT	Bio BERT	Bio BERT-Large	No. of Exact Match
1	1	1	417
1	1	0	19
1	0	1	109
1	0	0	33
0	1	1	202
0	1	0	49
0	0	1	201
0	0	0	474

Table 4: Exact Match model comparison

needs extensive knowledge of domain. With increase in availability of computational resources, we should be able to achieve state of the art performance across multiple domains.

6 Future Scope

With the extensive usage of pre-trained models, having a distilled version of BERT (DistilBERT) helped the NLP community to conduct research and answer various questions without taking too many computational resources. This is one of the reason I succeeded in fine tuning DistlBERT and failed in fine tuning BioBERT. To address this, I want to explore a distilled version of BioBERT analogous to DistilBERT. Very recently Du et al. (Du and Hu, 2020) presented the idea of distillation of BioBERT. I want to explore further more on this idea and evaluate on BioASQ.

BioBERT uses same vocabulary as BERT to maintain BERT architecture. Having a biomedical specific vocabulary can improve the performance much more. When I have enough resources to pre-train a model like BioBERT, I want to tokenize using biomedical specific vocabulary.

Though BioBERT Large, is inherited from BioBERT, there are questions which BioBERT answered correctly and BioBERT-Large failed. BioBERT-Large seems to over fit data. Hyper parameter tuning can be helped to address this issue. Adding additional drop out layers helped solve this issue in RNN architectures. I would like to explore if dropout layers bring any difference in the BioBERT-Large performance.

7 Acknowledgments

I would like to thank all the course staff for being very supportive and helping me think through my project.

References

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading wikipedia to answer open-domain questions. *CoRR*, abs/1704.00051.

Debasmita Das, Yatin Katyal, Janu Verma, Shashank

Dubey, AakashDeep Singh, Kushagra Agarwal, Sourojit Bhaduri, and RajeshKumar Ranjan. 2020. Information retrieval and extraction on covid-19 clinical articles using graph community detection and bio-bert embeddings. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*.

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

Hongbin George Du and Yanke Hu. 2020. Squeeze-biobert: Biobert distillation for healthcare natural language processing. In *International Conference on Computational Data and Social Networks*, pages 193–201. Springer.

Adam Fisch, Alon Talmor, Robin Jia, Minjoon Seo, Eunsol Choi, and Danqi Chen. 2019. MRQA 2019 shared task: Evaluating generalization in reading comprehension. *CoRR*, abs/1910.09753.

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

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100, 000+ questions for machine comprehension of text. *CoRR*, abs/1606.05250.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. 2016. Newsqa: A machine comprehension dataset. *CoRR*, abs/1611.09830.

George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, et al. 2015. An overview of the bioasq large-scale biomedical semantic indexing and question answering competition. *BMC bioinformatics*, 16(1):1–28.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.