Transfer Learning with QA Datasets

Cynthia Zhu, Fidelia Nawar, Ryan Goding
University of California, Berkeley
{cyuzhu,fnawar,rgoding}@berkeley.edu
July 30, 2022

Abstract

While pre-trained models such as BERT have shown great gains across natural language understanding tasks, whether their performance can be improved by further training the model in different domains remains unclear. In this project, we evaluated the performance of BERT models in extractive question answering tasks through multiple fine-tuning stages with different domain datasets. We obtained an F1 increase of 0.24% - 2.94% on models pre-trained on different datasets compared to models without pre-training.

1 Introduction

Reading comprehension question answering (QA) aims to answer questions given passages or documents. This QA task is difficult to accomplish because the models built need to understand the structure and semantics of a language, understand the context of the questions, and locate positions of answers and phrases. Another challenge in the NLP field is the lack of annotated datasets in all domains to allow models to perform the best in the targeted task. A trained model may not perform as well on a test set as in the training stage because the training and test sets were collected from different sources.The discrepancy across domain distributions could cause degradation in model performance. Domain adaptation of a model becomes necessary to secure a good accuracy on the target dataset.

To investigate the transfer learning of BERT models in different domains, we selected three datasets, SQUAD 2.0 [1], TriviaQA[2], and the Natural Questions[3](NQ). After training each model on the different datasets, we will see if there is a transfer learning rate improvement of the models with the different datasets in different orders. Although many models have already been built to accomplish these tasks on these datasets, we hope to see if there are performance differences in testing on different datasets in different orders and which factors seem to affect the transfer learning rate.

In our case, we adopted a supervised approach with labeled data from the target domain and target dataset. Here we applied transfer learning of DistilBERt and SpanBERT models training on SQuAD 2.0, NQ, and TriviaQA. We wanted to see the domain adaptations in dealing with the disparity between domains.

2 Background

BERT for QA

BERT, as a self-supervised approach for pre-training a deep transformer encoder before fine-tuning a specific task, optimizes two training objectives, masked language model and next sentence prediction. We chose to apply BERT for QA in this project since QA requires a comprehensive understanding of natural languages and the ability to do further reasoning and inference, and BERT is known for being able to understand long-term queries better. Since BERT's underlying model architecture is a multi-layer bidirectional transformer encoder, it relies solely on convolutional and attention mechanisms, significantly decreasing training time. In addition, its bidirectional representation is conditioned on contexts on either side of the word (left and right) simultaneously in all layers, making it an excellent model for producing

contextual representations for text. This is essential for our task since we want to be able to generate contextual representations that incorporate information from both the passage and the query.

Related Work

There are many models implemented for this task that first build contextual encodings of passage/query and then fuse them together using attention and additional layers to process the representation and predict answers. For example, the BiDAF[3] accomplishes this by proposing a bidirectional attention flow that combines both passage-to-query and query-to-passage attentions, producing a query representation of the passage. Similar to BiDAF, SAN[4], Unet[5], and QANet[6] use similar bidirectional attentions to fuse information with self-attention. Before BERT, ELMo[7] was a common pre-trained model for deep contextualized word representation, using word embeddings to improve model performance. We decided to use BERT for this project rather than the other pre-trained models mentioned above because BERT uses transformers that use self-attention, which helps the encoder look at other words in the input sentence as it encodes specific words. We believe our novel modeling approach allows us to show the effectiveness of transfer learning of large-scale datasets with supervised learning.

3 Method and Approach

3.1 Datasets

Here, we introduce the three datasets of various sizes, complexity, and information gathered that we used in the training processes of each model.

SQUAD 2.0

The SQuAD dataset is a reading comprehension dataset. It consists of questions posted by crowd workers on a set of Wikipedia articles. The answer to each question is a segment of text (span) from the corresponding reading passage, or the question might be unanswerable. This dataset combines the 100k questions from the SQuAD 1.1 with 50k unanswerable questions to look similar to the answerable ones. If a question is answerable, the answer is guaranteed to be a continuous span in the context paragraph. No significant preprocessing was required for using this dataset since Natural Questions and TriviaQA was adapted to match the format of SQuAD so that the models received input data in the same structure from all three datasets.

Natural Questions

NQ consists of anonymized and aggregated query questions from the Google search engine. Each sample contains a Wikipedia page, an annotated long answer, and one or more short answers if present on the wiki page, or marked null if no long/short answers are present. We repurposed the dataset using long answers as context and short answers as answers to evaluate models' performance on extractive QA. Doing so removes the samples without long answers from the training set. The training set contains 6.8% of samples having multiple answers, 63.4% with single answers, and 29.8% with no answers.

TriviaQA

The TriviaQA dataset complements the previous two datasets chosen in several ways. First, the TriviaQA dataset has complex, compositional questions authored by trivia enthusiasts. We only included the search-based results as context to prevent overlap with any NQ or SQuAD data. This dataset required some preprocessing to allow extractive QA. Initially, some text cleaning was lowercased to allow the usage of normalized answers. After training on TriviaQA, we obtained very poor EM/F1 results. We hypothesized that lowercasing may have negatively affected the results because we pre-trained SQuAD and NQ using cased models. Therefore, a second type of TriviaQA preprocessing was created and evaluated when computation resources were allowed.

3.2 Modeling Approach

Given the complexity of both task and datasets, we selected three models, a bidirectional GRU with attention model, DistilBERT, and SpanBERT, to study transfer learning.

Model 1: Baseline

We first adopted a bidirectional GRU and attention architecture. The model comprises a word embedding layer, which maps each word in context and question to a vector space with GloVe embeddings. Both question and context have a bidirectional GRU layer to learn the contextual cues from surrounding words to refine the embedding of the words. The outputs were fed into an attention flow layer, which couples the question and context to produce a set of query-aware feature vectors for each word. The model outputs are start position, end position, and probability of having answers, each was obtained by applying softmax on attention output. Unfortunately, we only achieved Exact Match (EM) / F1 24.48/25.85 with fine tuning on the NQ. Furthermore, we saw a large number of samples not having correct predictions. Given that the state-of-the-art approaches are often BERT based, we diverted to DistilBERT and SpanBERT to continue our transfer learning study.

Model 2: DistilBERT

Due to the long contexts in TriviaQA, we decided to use the DistilBERT base model, given that it has about half the total number of parameters of the BERT base and retains 95% BERT's performance on the language understanding benchmark GLUE.

The DistilBERT model is based on a concept of knowledge distillation, where the student is trained with a distillation loss over the soft target probabilities of the teacher. The student - DistilBERT - has the same general architecture as BERT but leveraged the training loss value from the teacher BERT masked language model. We considered using BERT loss in question answering tasks as teacher loss in the training loss function. However, past literature showed it is beneficial to use a general-purpose pretraining distillation rather than a task-specific distillation. Therefore, we kept it unchanged [10].

$$L_{ce} = \sum_i t_i * \log(s_i)$$

Model 3: SpanBERT

SpanBERT extends BERT by masking contiguous random spans rather than random tokens and training the span boundary representation to predict the entire content of the masked span without relying on the individual token reputation within it. Past work has demonstrated that SpanBERT consistently outperforms BERT on span selection tasks such as extractive question answering.

4 Experiments

First, all three datasets were reformatted to a standard format for simplicity in streamlining the modeling process. Three modeling experiments were designed, as shown in Table 1 below. The pre-trained model is fine-tuned on the three datasets in order. Each fine-tuning evaluation is performed on the current and previous datasets to calculate the F1 improvement rate [11].

Table 1: Modeling experiments

Model	Order of Fine Tuning
DistilBERT	NQ → TriviaQA → SQuAD 2.0
DistilBERT	SQuAD 2.0 → TriviaQA → NQ
SpanBERT	NQ → SQuAD 2.0 → TriviaQA

5 Results and Discussion

The metrics were chosen to evaluate model performance for question answering tasks are an exact match (EM) and F1 score. EM calculates the percentage of answers that are exactly correct as our ground truth word to word. The higher the percentage value, the better the model's performance in understanding the context and

questions and giving exact answers. Macro averaged F1 score captures the precision and recalls that words chosen as part of the answer are part of the answer.

Table 2 shows the transfer learning rate after evaluating the different models in various training orders of the datasets. The first column represents an evaluation of each model pre-trained on the respective validation dataset. The adjacent columns represent training evaluation metrics on an additional dataset and the respective scores outputted for each validation dataset. For example, in column 1 of Cross Training Experiment 2 - DistilBert QA, we trained and evaluated the model's performance on the SQuAD dataset. In column 2, we have trained the model on SQuAD and are evaluating its performance on the TriviaQA validation set. In column 3, we have trained the model on SQuAD and Trivia QA and are analyzing its performance on the TriviaQA validation set. Similarly, in column 4 of the same table, we use the SQuAD/TriviaQA trained DistilBert model on the NQ validation set. Finally, in column 5, we are training on an additional dataset, NQ, and evaluating its final performance after being trained on all three datasets.

5.1 Transfer Learning Rate Improvement

SQuAD Analysis

For DistilBert QA 1, we have trained SQuAD after training on both NQ and TriviaQA, whereas for DistilBert QA 2, we trained SQuAD before TriviaQA. We can see that training on SQuAD after TriviaQA produced much better performance than training before TriviaQA, around 50% for EM and F1 when fine-tuned on SQuAD after TriviaQA vs 1% when fine-tuned on SQuAD before. We assumed this occurred because training on larger datasets before smaller ones can produce better results since larger datasets hold more contexts and information for the model to learn from, indicating an improvement in transfer learning rate when training in this order.

Table 2: Evaluation results in cross-training experiments.

Cross Training Experiment 1 - DistilBert QA 1					
Trained on:	NQ	NQ	NQ, Trivia QA	NQ, Trivia QA	NQ, Trivia QA, SQuAD
Eval on:	NQ	Trivia QA	Trivia QA	SQuAD	SQuAD
EM	56.1	11.57	.10/ 0.48	72.83/ 50.07	59.44/ 61.07
F1	61.39	14.64	.23/ 0.48	76.23/ 50.07	62.55/ 64.00
Cross Training Experiment 2 - DistilBert QA 2					
Trained on:	SQuAD	SQuAD	SQuAD, Trivia QA	SQuAD, Trivia QA	SQuAD, Trivia QA, NQ
Eval on:	SQuAD	Trivia QA	Trivia QA	NQ	NQ
EM	60.55	17.78	.08 / 1.44	NE**/ 1.44	55.42
F1	63.55	23.71	.25 / 1.57	NE**/ 1.57	61.37
Cross Training Experiment 3 - SpanBertQA					
Trained on:	NQ	NQ	NQ, SQuAD	NQ, SQuAD	NQ, SQuAD, Trivia QA
Eval on:	NQ	SQuAD	SQuAD	Trivia QA	Trivia QA
EM	61.02	48.26	75.11	44.04 / NE**	.14 / NE**
F1	68.26	52.00	78.32	55.36 / NE**	.297 / NE**

* TriviaQA has two sets of evaluation results (<old/new>).
**Not Evaluated due to computing limitations and restrictions by Colab

We see this hold true for SpanBertQA when we trained on NQ first before fine-tuning on SQuAD, and the rate of transfer learning was improved by 10% in F1. It is also interesting to note how the performance on SQuAD was highest post-trained on all datasets, whereas this trend was not exhibited for the other two datasets.

NQ Analysis

The F1 score of DistilBERT without pre-training but post-fine-tuned on NQ is 61.39, whereas the F1 score of DistilBERT pre-trained on SQuAD followed by fine-tuning on NQ is 64.33. The rate of transfer learning was improved by 2.94% in F1. This indicates that the DistilBERT model performs better after seeing different domain data.

TriviaQA Analysis

Evaluating previously trained DistilBert models on Trivia QA validation data yielded very poor F1 results ranging from 14.64 to 23.71. After training on Trivia QA data, our F1 results plummeted to as low as 0.48. SpanBert performed better before training on Trivia QA but also saw a sharp decline after training on the large dataset. This could indicate that DistilBert or SpanBert aren't sophisticated enough to work with the Trivia QA dataset. The trivia QA dataset is more complex with its large context and answers not directly obtained by span prediction.

Figure 6.1 summarizes the transfer learning between natural questions and SQuAD. For example, the first bar shows the DistilBert model trained and then evaluated on Natural questions, and the increase in F1 with the additional training on the SQuAD dataset is shown in the second bar.

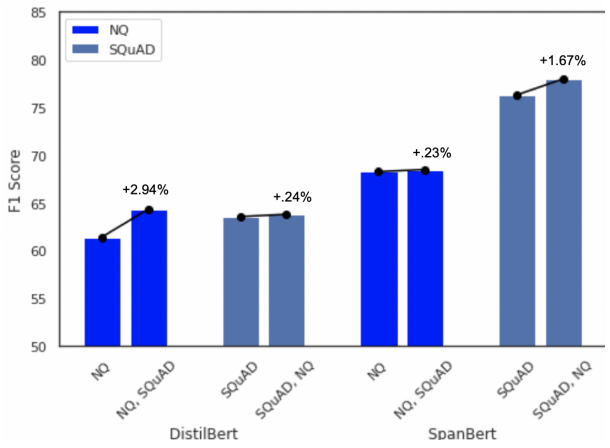

Figure 6.1: F1 Scores before and after training on datasets.

6 Conclusion

In this paper, we selected three models, a bidirectional GRU with attention model, DistilBERT, and SpanBERT, to study the transfer learning rate across three question answering datasets, SQuAD 2.0, Natural Questions, and TriviaQA. The results show that fine-tuning larger datasets and evaluating smaller ones improve the transfer learning rate. In contrast, the opposite holds true when fine-tuning on smaller datasets first and then evaluating on larger (SQuAD > TriviaQA). For future work, we hope to use larger models that can handle larger datasets like TriviaQA such as RoBERTa, which is trained on more data in larger batches. In the future, we also would like to test DistilBert QA loss as a teacher model to see if performance is improved since we simply used the DistilBert teacher model loss for this project. We believe that the training loss of Bert for QA for the loss calculation, instead of the masked language model, may provide additional relevant distillation for training.

Appendix

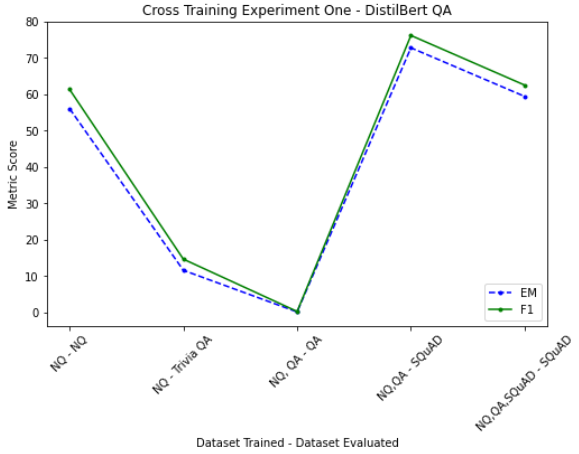

Figure A.0 - Cross Training Experiment 1 - DistilBert QA

Figure A.1 - Cross Training Experiment 2 - DistilBert QA

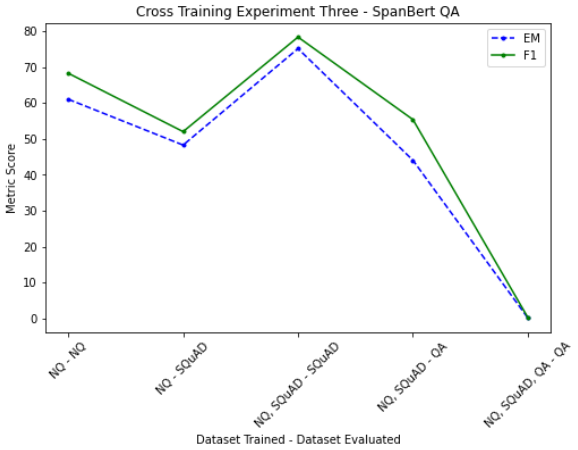

Figure A.2 - Cross Training Experiment 3 - SpanBert QA

References

1. Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know What You Don't Know: Unanswerable Questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.
2. Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.
3. Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural Questions: A Benchmark for Question Answering Research. *Transactions of the Association for Computational Linguistics*, 7:452–466.
4. [BIDAF] Seo, M., Kembhavi, A., Farhadi, A., & Hajishirzi, H. (2017). Bidirectional Attention Flow for Machine Comprehension. *ArXiv, abs/1611.01603*.
5. [SAN] Xiaodong Liu, Yelong Shen, Kevin Duh, and Jianfeng Gao. Stochastic answer networks for machine reading comprehension. CoRR, abs/1712.03556, 2017.
6. [UNet] Fu Sun, Linyang Li, Xipeng Qiu, and Yang Liu. U-net: Machine reading comprehension with unanswerable questions. CoRR, abs/1810.06638, 2018.
7. [QANet] Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V. Le. Qanet: Combining local convolution with global self-attention for reading comprehension. CoRR, abs/1804.09541, 2018.
8. [ELMo] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. CoRR, abs/1802.05365, 2018.
9. Farahani, A., Voghoei, S., Rasheed, K.M., & Arabnia, H.R. (2021). A Brief Review of Domain Adaptation. *ArXiv, abs/2010.03978*.