Reading Comprehension and Open-Domain Question Answering with BERT

Nathan Nusaputra
University of California, Berkeley
nusaputra137@berkeley.edu

Ryan Sawasaki
University of California, Berkeley
rsawasaki@berkeley.edu

Abstract

While recent advances in natural language processing models have helped solve reading comprehension tasks, progress on open-domain question answering has been challenging. In this paper, we analyze the two potential factors that make open-domain question answering a more difficult task: context length and question generation. We focus on the Stanford Question Answering Dataset (SQuAD 2.0) as our baseline reading comprehension dataset and Google Natural Questions (NQ) as our open-domain dataset, which we split into two datasets aimed at isolating the context length and question generation factors to evaluate their impact on model performance. Using a BERT-base model, we fine-tune the model on the three datasets and evaluate their performance using an F1 score metric. We conclude that the challenges designed into the NQ dataset and open-domain question answering datasets are predominantly due to longer contexts rather than the methodology used to generate questions.

1 Introduction

Advances in deep learning and accessibility to larger training datasets has led to significant progress in machine learning and artificial intelligence. One of the primary benefactors of this focused research in neural networks is the field of natural language processing, which over the last few years has seen advances in NLP tasks such as question answering, natural language inference, and text classification. Progress in the task of question answering has also been boosted by the introduction of publicly released datasets.

One of the first notable question answering datasets is the Stanford Question Answering Dataset (SQuAD), which consists of 100,000+ questions posed by crowdworkers on a set of Wikipedia passages (Rajpurkur et al., 2016). The answer is defined as a span of text from the corresponding reading passage. A couple years later, Stanford recognized that machine reading comprehension systems can often locate the correct answer within a context, but had more difficulty making guesses when the answer is not provided in the context. To address this issue, SQuAD 2.0 was released, which combined existing SQuAD data with 50,000 unanswerable questions that were written to look similar to answerable questions (Rajpurkar et al., 2018). Human performance on the SQuAD 2.0 dataset is an F1 score of 89.0.

Another notable dataset is Google Natural Questions (NQ), which is an open-domain question answering dataset consisting of 300,000+ questions on full Wikipedia pages (Kwiatkowski et al., 2019). Questions are real, aggregated queries posed in Google search. NQ provides a short answer (one or more entities) similar to SQuAD, and a long answer (paragraph). Human performance on the NQ dataset for the short answer is an F1 score of 75.7.

While SQuAD and NQ are similar in that they are both question answering datasets, there are nuances between the two that result in NQ presenting additional challenges, as evidenced by its lower human performance score. The contrast in performance scores could potentially be attributed to the differences between reading comprehension datasets (SQuAD 2.0) and open-domain datasets (NQ). In this paper we examine the impact these differences have on question answering model performance.

2 Background

On the surface, the SQuAD 2.0 and NQ datasets appear to be similar. Both datasets are organized into question-context-answer format, contain non-answerable questions, and provide context from a common source (Wikipedia). However, an in-depth review of the design and research behind the datasets reveals that there are significant differences that present unique challenges for question answering models.

SQuAD was released to address the challenge of machine reading comprehension, which is the ability to read a passage and answer questions about it. In the next evolution of question answering tasks, NQ was introduced to take on the additional challenges associated with open-domain question answering, which is the ability to produce an answer to a question with or without access to external context (open-book or closed-book). In order to replicate an open-book, open-domain question answering task, there are two fundamental differences designed into the NQ dataset.

Question Generation: NQ aggregates its data by collecting real, anonymized questions from Google search engine and the context is a Wikipedia page returned by the search given that it is one of the top 5 search results. The questions are “natural” in how they are actual queries generated from pure human curiosity. Conversely, SQuAD data was created by first selecting a passage from Wikipedia for the context and then using crowdsourced annotators for the generation of questions. In retrospect, this method of context-question collection led to notable lexical overlap between question and context because annotators often used part of the context to create questions.

Context Length: NQ attempts to replicate open-domain question answering by asserting that the question answering task begins at the point of retrieving a Wikipedia page from Google search results. NQ uses the full page as the context and retrieves a paragraph for its long answers. The short answers are a span or set of spans within the long answer paragraph. SQuAD provides a context about a paragraph long, much shorter in length in comparison to the NQ page-long context. However, the SQuAD context is similar in length to the NQ long answer.

To make a comparison between reading comprehension and open-domain datasets, a question answering model capable of handling both the SQuAD and NQ datasets is needed. We propose a BERT: Bidirectional Encoder Representations from Transformer model, which has already been shown to be fully capable of handling the SQuAD 2.0 dataset as demonstrated in the original BERT paper (Devlin et al., 2019).

3 Methods

The methods section is split into two parts, first we talk about data collection and preprocessing. Then we go into a brief description of the model and approach we took to compare the datasets.

3.1 Data Preprocessing

For this experiment, we are comparing three main datasets: SQuAD 2.0 and two variations of the Google NQ dataset. The first dataset, SQuAD 2.0 has 130,319 questions and 19,029 unique contexts (Rajpurkur et al., 2018). Contrary to the NQ dataset which has one question for each context, SQuAD has multiple questions per context. To make a comparison to the NQ datasets, we refine the SQuAD dataset by sampling one question per context and reduce the total number of examples to a size comparable to the NQ datasets. In addition, we sample questions with approximately 75% answerable and 25% non-answerable questions to closely match the answerable to non-answerable percentage split of the NQ dataset.

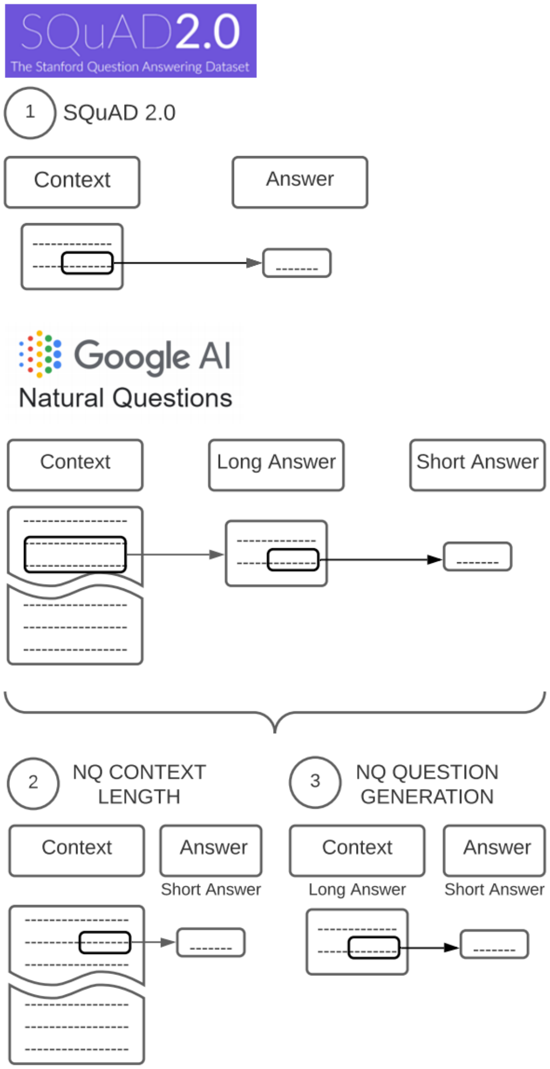

Figure 1: SQuAD and Natural Questions datasets. NQ is split into two datasets, Context Length and Question Generation

The other two datasets are from the Google NQ corpus. The original NQ dataset contains 300,000+ training examples but due to scalability and limited computing resources we took a sample of 10% of the data resulting in 30,000 examples. From this sample, we filter out questions that do not include a long answer. In addition, answers that are extracted from tables or include yes/no answers are also removed. After filtering, the NQ dataset is left with 10,608 examples that are further split into 80% train and 20% test sets. Two differences that are designed into open-domain question answering datasets are the manner in which the questions are generated and the length of context. To explore how each of these differences affect model performance, we reconstruct the NQ dataset, as shown in Figure 1. To investigate the impact of the question generation, we use the NQ long answer as the context. In doing this, the context length more closely resembles that of SQuAD 2.0 and we can observe how the model performs with the question generation factor isolated. To see the impact of context length, we will keep the NQ page-long context and implement a model to find the short answer within the context.

	SQuAD 2.0	NQ Context Length	NQ Question Generation
Number of Questions	8,563	8,486	8,486
Answerable Questions	6,468 (75.6%)	6,424 (75.7%)	6,424 (75.7%)
Non-Answerable Questions	2,095 (24.4%)	2,062 (24.3%)	2,062 (24.3%)
Question Average Tokens	11.2	9.2	9.2
Context Average Tokens	134.1	3390.1	113.9

Table 1: Dataset statistics

We conduct EDA on the datasets to see statistics on the token counts, shown in Table 1. As anticipated, we find that on average the context has a significantly larger number of tokens, while the NQ dataset that uses the long answer as the context has a comparable number of tokens to SQuAD. The average number of question tokens for the NQ and SQuAD datasets are similar.

3.2 Model

The goal of question answering is to find the span of text in each context that answers the question. We use a pre-trained BERT-base uncased model fine-tuned on our three datasets. This model was chosen for its ability to handle both the SQuAD and NQ contexts in a single model, rather than conducting a two-tier model approach.

To address the BERT token limitation, we preprocess long context tokens to fit with BERT. The preprocessing involves splitting the context by sliding a window over the full length of the context and overlapping the windows with a stride (Alberti et al., 2019). We set the window max length and stride at 384 and 128, respectively. By sliding a window over the context, we generate multiple instances of the same context.

For the fine-tuning parameters, we use a learning rate of 0.005, batch size of 16, weight decay of 0.01, and 3 epochs. These parameters are constant throughout the various experiments we run on the datasets.

MODEL BERT $_{\text{BASE-UNCASED}}$	SQuAD 2.0 F1 Test	NQ Context Length F1 Test	NQ Question Generation F1 Test
Fine-tuned on SQuAD 2.0	54.7	25.7	39.2
Fine-tuned on NQ Context Length	49.8	36.9	38.4
Fine-tuned on NQ Question Generation	43.2	43.7	56.6

Table 2: Bert-base uncased model fine-tuned on SQuAD 2.0, NQ Context Length, and NQ Question Generation

4 Results

To make a comparison between reading comprehension and open-domain question answer datasets, we conduct experiments on the SQuAD and NQ datasets comparing their F1 scores. Our baseline is the BERT model fine-tuned on SQuAD 2.0 and evaluated on a SQuAD test set. This model is compared with BERT models fine-tuned on our two created NQ datasets to observe the impact of the context length and question generation. In addition, we cross evaluate each fine-tuned model on the other test sets to determine if the models are generalizable across the reading comprehension and open-domain datasets.

4.1 BERT Fine-Tuned on SQuAD 2.0

The results of our experiments are shown in Table 2. Our baseline BERT model fine-tuned on SQuAD and evaluated on the SQuAD test set results in an F1 score of 54.7. Models typically see a decrease in performance when tested on different datasets and our results are consistent with this notion. When evaluated on the NQ Question Generation test set the F1 score drops to 39.2. The score drops even further to an F1 score of 25.7 when evaluated on the NQ Context Length test set.

4.2 BERT Fine-Tuned on NQ Context Length

Next, we evaluate the BERT model fine-tuned on the NQ Context Length dataset. The model performs significantly worse than the baseline with an F1 score of 36.9. While there may be additional factors responsible for the decrease in performance, the result suggests that the BERT model struggles to handle much longer context lengths. Unexpectedly, when the model is evaluated on the SQuAD and NQ Question Generation datasets the performance increases with an F1 score of 38.4 and 49.8, respectively.

4.3 BERT Fine-Tuned on NQ Question Generation

For our final model, we evaluate the BERT model fine-tuned on the NQ Question Generation dataset. The model performs similar to the baseline with an F1 score of 56.6. It has been speculated that the manner in which SQuAD questions were generated might have led to lexical overlap between the question and context, resulting in better model performance. However, the results from this experiment suggest that the design of the SQuAD questions and its lexical overlap does not lead to higher performance over the NQ dataset. Similar to the SQuAD fine-tuned model, when evaluated on the other test sets the results drop with an F1 score of 43.2 on the SQuAD test set and 43.7 on the NQ Context Length test set.

5 Analysis

Our experiment yielded results that warranted additional analysis. As part of our analysis, we conducted an ablation study on these components to gain a better understanding of the datasets and model performance.

5.1 SQuAD 2.0 Model Performance

The baseline model resulted in an F1 score of 54.7, which is underwhelming in comparison to previous benchmarks. The original BERT paper achieved much higher F1 scores, setting the bar with a BERT-large model F1 score of 83.1 on SQuAD 2.0 (Devlin et al., 2019). Our SQuAD dataset was reduced to make an analogous comparison to our NQ datasets. To evaluate the impact of that data reduction, we evaluate a dataset that includes all of the unique contexts with a single question, as well as the total dataset, which includes contexts with multiple questions.

The results of the study are shown in Table 3. When including all unique contexts and sampling one question per context, our dataset increases to 19,029 examples. The BERT-base uncased model trained on this dataset performs nearly 8 points higher with an F1 score of 62.2. We evaluate on the full SQuAD dataset of 130,319 training examples, which includes multiple questions per context. The BERT model trained on this dataset results in an even larger increase in performance with an F1 score of 75.6, a score more comparable to that of the BERT benchmark models. These results indicate that a combination of a larger number of examples and multiple questions per context may lead to increased model performance.

BERT $_{\text{BASE-UNCASED}}$ Fine-tuned on SQuAD 2.0	
Number of Training Examples	SQuAD 2.0 F1 Test
8,563 [a]	54.7
19,029 [a]	62.2
130,319	75.6

[a] One question per context

Table 3: Results on BERT model fine-tuned on SQuAD 2.0 with increasing size of training examples

5.2 NQ Context Length Reduction

The final study that we ran was a BERT-base uncased model trained on decreasing NQ context lengths. The full context yielded an F1 score of 36.9. With this baseline F1 score we wanted to determine if decreasing the training context size would make an impact. We found that 90% of the records for the NQ dataset had their short answer within the first 50% of the context and that 80% of the records had their

BERT $_{\text{BASE-UNCASED}}$ Fine-tuned on NQ Context Length	
Training Context Size	NQ Context Length F1 Test
Full Context	36.9
50% Context Reduction	37.7
90% Context Reduction	50.8

Table 4: Results on BERT model fine-tuned on NQ Context Length with decreasing context size

answer within the first 10% of the context. This led us to run the two reduced context tests as shown in Table 4.

We found that the F1 score increases as we decrease the context length. The results show that despite losing 20% of the records due to truncation of the context, the trade-off results in higher overall performance.

6 Conclusion

We explored two potential factors that create challenges for open-domain question answering models. Our evaluation reveals that BERT-base uncased models struggle with exceedingly long contexts, while the methods in which the questions were generated do not impact performance in comparison to reading comprehension datasets. In addition, the results suggest that fine-tuned BERT models are not generalizable across reading comprehension and open-domain datasets.

Acknowledgments

We would like to thank Mark Butler and the UC Berkeley NLP faculty for their valuable advice and mentorship on this work.

References

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, Percy Liang. 2016. *SQuAD: 100,000+ Questions for Machine Comprehension of Text* https://arxiv.org/pdf/1606.05250.pdf

Pranav Rajpurkar, Robin Jia, Percy Liang. 2018. Know What You Don’t Know: Unanswerable Questions for SQuAD https://arxiv.org/pdf/1806.03822.pdf

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, Slav Petrov. 2019. *Natural Questions: a Benchmark for Question Answering Research* https://research.google/pubs/pub47761/

Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova. 2019. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding* https://arxiv.org/pdf/1810.04805v2.pdf

Chris Alberti, Kenton Lee, Michael Collins. 2019. *A BERT baseline for the Natural Question* https://arxiv.org/pdf/1901.08634v3.pdf