Natural Language Processing (CS-388) Final Project Report

Anonymous Individual Author

Note: This project was completed by an *individual* student

Abstract

This project explores the application of pre-trained contextual embedding (PCE) models in comparison to the bidirectional attentive reader (baseline) implemented by the CS-388 staff. Particularly, I applied DistilBERT (Sanh et al., 2019) as it is 60% faster to train and retains 97% of the base BERT performance. Model comparison was conducted on the SQuAD 1.1 (Rajpurkar et al., 2016) and SQuAD adverserial datasets (Jia and Liang, 2017). The HuggingFace implementation of DistilBERT was utilized, hence I first had to ensure the SQuAD datasets used match to that provided by the class staff. Whereas the baseline model achieved 60.63% and 46.76% F1 scores on the SQuAD 1.1 and adverserial SQuAD datasets, respectively, the fine-tuned DistilBERT model achieved 85.4% and 69.4%, respectively. Analysis shows that both models struggle the most with "why" questions, and achieve lower F1 scores with increasing answer lengths.

1 Introduction

Question Answering (QA) is a rapidly growing field that aims to automatically answer natural language questions regarding one or multiple domains of knowledge (Pudaruth et al., 2016; Voorhees and Tice, 2000). QA systems target end-to-end or pipeline models that process questions and relevant documents, and correspondingly return short and precise answers. They commonly use pre-structured databases and/or a collection of documents (Ansari et al., 2016; Lende and Raghuwanshi, 2016), and consist of three major modules: 1) question processing, 2) document processing, and 3) answer processing (Soares and Parreiras, 2020). These systems are based on one or multiple algorithmic implementations, i.e. information retrieval, knowledge-based, and natural language processing (NLP) (Yao, 2014).

The application of NLP techniques in the development of QA models span a variety of algorithms (e.g. deep neural networks, naive bayes, support vector machines, etc.), techniques (tokenization, stemming, lemmatization, chunking, etc.) and frameworks (named entity recognition, position-of-speech tagging, syntactical analysis, etc.) (Soares and Parreiras, 2020). To evaluate NLP algorithms on QA tasks, multiple datasets were developed over the years, e.g. Stanford Question Ansering Dataset (SQuAD) 1.1 (Rajpurkar et al., 2016), NewsQA (Trischler et al., 2016), TriviaQA (Joshi et al., 2017), SQuAD 2.0 (Rajpurkar et al., 2018), QuAC (Choi et al., 2018), Natural Questions (Kwiatkowski et al., 2019), amongst others. Whereas various NLP algorithms were developed to tackle QA, pre-trained contextual embedding (PCE) models, e.g. BERT (Devlin et al., 2018) and ELMo (Peters et al., 2018), achieved the state-of-the-art performance.

This CS-388 project focused on the application and evaluation of DistilBERT (Sanh et al., 2019) in comparison to the course staff implementation of a bidirectional attentive reader. This report describes the referenced datasets and algorithms, performs training and evaluation, analyzes and discusses results, and finally points to potential direction for further improvement.

2 Dataset

Whereas the CS-388 final project prompt highlighted multiple datasets, this work is focused on the SQuAD 1.1 (Rajpurkar et al., 2016) and adverserial SQuAD (Jia and Liang, 2017) datasets. SQuAD 1.1 is curated for reading comprehension purposes using Wikipedia articles. Whereas it is still unclear how NLP algorithms comprehend QA text, Jia and Liang (2017) curated adverserial examples by adverserially inserting sentences to SQuAD

Dataset	CS-388	HuggingFace	Difference
SQuAD Training	86,588	87,599	1,011
SQuAD Validation	10,507	10,570	63
SQuad Adverserial	1,787	1,787	0

Table 1: SQuAD 1.1 dataset comparison as provided by CS-388 final project prompt versus HuggingFace (Wolf et al., 2020b)

paragraphs which are generated to distract NLP models without altering the label nor misleading humans.

Since this project utilizes HuggingFace tokenizers and PCE transformer implementations (Wolf et al., 2020a), it is important that these are only applied to the same SQuAD 1.1 and adverserial SQuAD dataset examples as provided by the CS-388 final project prompt [1]. As seen in Table 1, I found that the SQuAD training and validation datasets differ between the CS-388 final project prompt and HuggingFace by 1,011 and 63 examples, respectively [2]. Subsequently, the HuggingFace dataset was filtered by *id* and *question string* to exactly match that of the CS-388 final project prompt.

3 Methods

The baseline algorithm used in this project is the bidirectional attentive reader implemented by the CS-388 course staff. Preprocessing involves uncasing questions and passages and trimming them to maximum lengths of 64 and 384, respectively. Strings are then indexed and tokenized where PAD and UNK tokens are also introduced. The algorithm embeds questions and passages, computes context-to-query attention, encodes question and passage using bidirectional recurrent layers, computes query-to-context attention, and finally outputs the start and end position distributions. To determine the final answer, the start position is greedily determined based on the maximum marginal probability. The optimal span endpoint position is that which maximizes the start-end joint probability within a maximum span length size of 15 (Chen et al., 2017).

The second algorithm used in this project is DistilBERT (Sanh et al., 2019), a distilled version of BERT, which advantageously makes use of transformers. Whereas the bidirectional attentive reader is a context-free model that utilizes Word2Vec and Glove embeddings, BERT is a fully-bidirectional contextual model that represents words with respect to their context allowing for capturing contextual relationships. Since training the base BERT is computationally demanding, I decided to use DistilBERT which offers a smaller and faster version using knowledge distillation (Buciluǎ et al., 2006; Hinton et al., 2015). DistilBERT reduces 40% of the base BERT parameters while retaining 97% of its performance and boosting computational speed by 60%. In the HuggingFace tokenization step, long documents are fully preserved without truncation. Rather, each is chunked into overlapping strings/features with stride of 64 and lengths of 384. This also requires careful treatment during post-processing the final answer. After obtaining the start and end position distributions, the joint distributions are greedily computed for the top 10 start positions and top 10 end positions while ensuring end positions have greater indices and retain a maximum span length of 15 (Chen et al., 2017).

Parameter	Baseline	DistilBERT
vocab size	50,000	30,522
hidden dim	256	3,072
epochs	10	3
batch size	128	20
learning rate	1E-3	2E-5
dropout	0.0	0.1
weight decay	0.0	1E-3
gradient clip	0.5	1.0
early stopping	3	None

Table 2: Model hyperparameters.

Whereas there exist many hyperparameters, Table 2 highlights the major model and training hyperparemters used in this project. Note that all other hyperparameters are set to default values. Both were optimized using Adam. Meanwhile, model evaluation was conducted based on two QA metrics, exact match (EM) and F1 score. Given a QA example, EM is 1.0 if the predicted answer exactly

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

[2] Refer to Code at https://github.com/aljubrmj/mj-nlp-fp/blob/master/Dataset_Check.ipynb

Dataset	Baseline		DistilBERT	
	EM	F1	EM	F1
SQuAD	47.95	60.63	77.18	85.4
SQuAD Adverserial	36.43	46.76	62.17	69.4

Table 3: Performance of trained baseline and DistilBERT models.

matches the true answer; otherwise, it is 0.0, which makes it a strict metric. F1 score is a smoother metric that is defined as the harmonic average of precision and recall.

$$Precision = \frac{TruePositives}{TruePositives + FalsePositives}$$

$$Recall = \frac{TruePositives}{TruePositives + FalseNegatives}$$

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

4 Results and Discussion

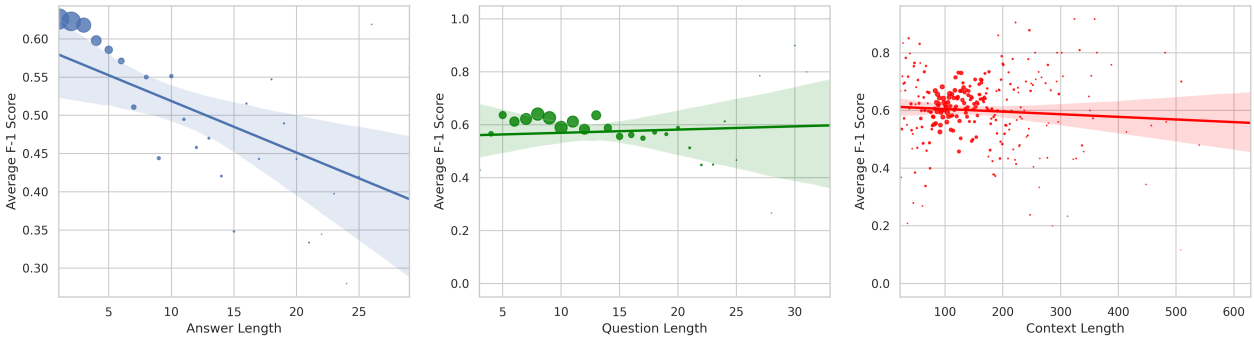
I first trained both the baseline and DistilBERT models using the aforementioned hyperparameters. Table 3 summarizes the performance of these models on both the SQuAD 1.1 and SQuAD adverserial datasets. DistilBERT clearly outperforms the baseline attentive reader by a a factor of 1.41 and 1.48 in SQuAD 1.1 and SQuAD adverserial, respectively. Note that the performance of DistilBERT on the SQuAD adverserial task (harder) is superior to that of the baseline attentive reader on the SQuAD 1.1 task (easier).

To further understand the model results, I performed sensitivity analysis where I evaluated the impact of answer, question, and context lengths individually on the F1 score [3]. Figure 1a shows the effect of these variables on the baseline model predictions of the SQuAD 1.1 task. Note that the majority of answers are shorter than 10 words. While it is expected that the baseline model would perform poorly beyond 15 words (maximum answer length based on DrQA post-processing), we can observe a clearly decreasing trend of F1 score as the answer length increases. This indicates that the model struggles with more comprehensive answers. Meanwhile, no trend was observed with respect to question and passage lengths. Figure 2a, shows the effect of question type (i.e. which, where, what, who, how, when, and why). The model performs best on "when" questions (shorter answers) and worst on "what" and "why" questions (longer answers). Evaluating the baseline model performance on SQuAD answers with and without numerical characters, I found that it achieves 72.74% and 57.69% F1 scores, respectively. This further justifies the model's particularly superior performance on "when" questions.

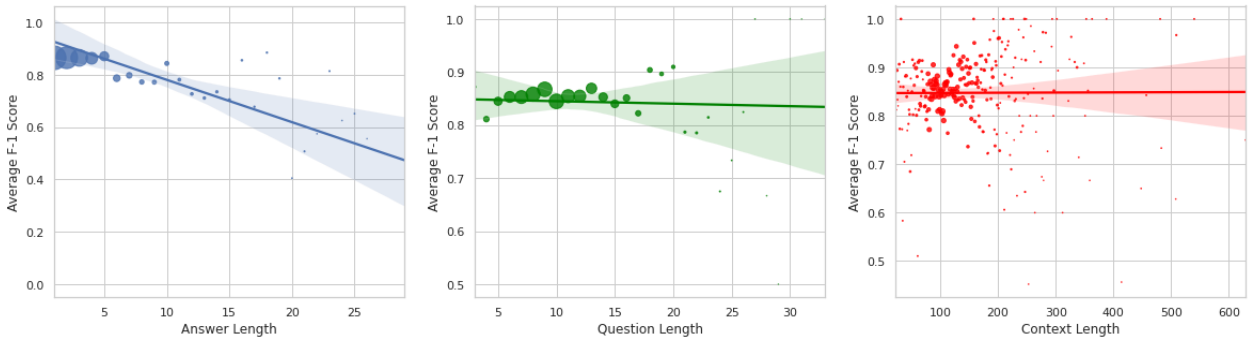
Similarly, I evaluated the DistilBERT model to see in which type of questions it particularly improved over the baseline model [4]. As seen in Figure 1b, the lengths of question and context also do not show a correlation with F1 score. Meanwhile, a decreasing trend is observed again with respect to answer lengths. Although, we can see that DistilBERT consistently achieves about 85% F1 score for in examples whose answers have lengths of less than 5 words, and the decreasing trend only starts in longer answers. This confirms that DistilBERT captured contextual information that assisted in capturing the correct answers. Figure 2b shows the analysis with respect to the question type where DistilBERT successfully achieves over 80% F1 score for all question types, excpet for "why" questions with nearly 75% F1 score. This indicates where further improvement to this model can be considered.

[3] Refer to Code at https://github.com/aljubrmj/mj-nlp-fp/blob/master/Baseline_Model_Mistakes.ipynb

[4] Refer to Code at https://github.com/aljubrmj/mj-nlp-fp/blob/master/DistilBert_Model_Analyze_Mistakes.ipynb

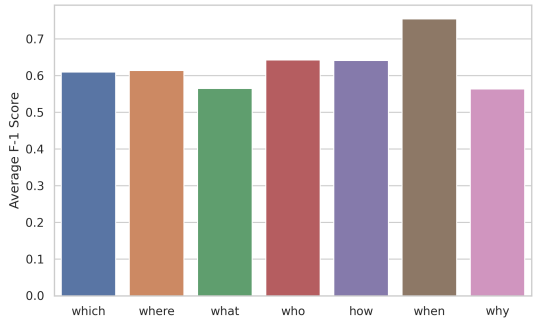


(a) Baseline

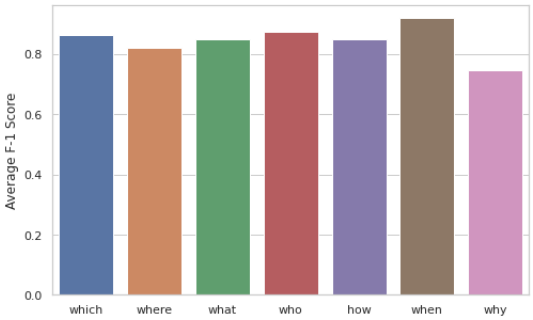


(b) DistilBERT

Figure 1: Effect of answer (left), question (middle), and passage (right) lengths on the (a) baseline and (b) DistilBERT model performance. Each marker represents the average F1 score for its corresponding length while the marker size represents the count of examples with that particular length. The line and uncertainty shaded bands are based on a simple linear regression fit.



(a) Baseline



(b) DistilBERT

Figure 2: F1 score performance of the (a) baseline and (b) DistilBERT models based on question type.

5 Conclusion

This project demonstrates the comparison of a bidirectional attentive reader as a baseline model and DistilBERT on the SQuAD 1.1 and SQuAD adverserial datasets. The results evidently indicated the superiority of DistilBERT which achieved F1 scores of 85.4 and 69.4% on the SQuAD 1.1 and adverserial SQuAD datasets, respectively. It was observed that the model performance degrades with longer answers, particularly "why" questions. Subsequently, future work can focus on improving these models in such contexts where answers are longer and require more comprehensive and contextual coverage of the provided documents.

References

Ahlam Ansari, Moonish Maknojia, and Altamash Shaikh. 2016. Intelligent question answering system based on artificial neural network. In *2016 IEEE International Conference on Engineering and Technology (ICETECH)*, pages 758–763. IEEE.

Cristian Buciluǎ, Rich Caruana, and Alexandru Niculescu-Mizil. 2006. Model compression. In *Proceedings of the 12th ACM SIGKDD international*

conference on Knowledge discovery and data mining, pages 535–541.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading wikipedia to answer open-domain questions. *arXiv preprint arXiv:1704.00051*.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wentau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. Quac: Question answering in context. *arXiv preprint arXiv:1808.07036*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. *arXiv preprint arXiv:1707.07328*.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Sweta P Lende and MM Raghuwanshi. 2016. Question answering system on education acts using nlp techniques. In *2016 world conference on futuristic trends in research and innovation for social welfare (Startup Conclave)*, pages 1–6. IEEE.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Sameerchand Pudaruth, Kajal Boodhoo, and Lushika Goolbudun. 2016. An intelligent question answering system for ict. In *2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT)*, pages 2895–2899. IEEE.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. *arXiv preprint arXiv:1806.03822*.

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

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Marco Antonio Calijorne Soares and Fernando Silva Parreiras. 2020. A literature review on question answering techniques, paradigms and systems. *Journal of King Saud University-Computer and Information Sciences*, 32(6):635–646.

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

Ellen M Voorhees and Dawn M Tice. 2000. Building a question answering test collection. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 200–207.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020a. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Thomas Wolf, Quentin Lhoest, Patrick von Platen, Yacine Jernite, Mariama Drame, Julien Plu, Julien Chaumond, Clement Delangue, Clara Ma, Abhishek Thakur, Suraj Patil, Joe Davison, Teven Le Scao, Victor Sanh, Canwen Xu, Nicolas Patry, Angie McMillan-Major, Simon Brandeis, Sylvain Gugger, François Lagunas, Lysandre Debut, Morgan Funtowicz, Anthony Moi, Sasha Rush, Philipp Schmidd, Pierric Cistac, Victor Muštar, Jeff Boudier, and Anna Tordjmann. 2020b. Datasets. *GitHub. Note: https://github.com/huggingface/datasets*, 1.

Xuchen Yao. 2014. *Feature-driven question answering with natural language alignment*. Ph.D. thesis, Johns Hopkins University.