Machine Comprehension : Question-Answering on SQUAD 2.0

Dolly Agarwal A0228490B
Lim Ernest A0201835M
Rashi Sharma A0228492X
Rohit Jain A0228500R
Simran Aggarwal A0228520M

Abstract

Deep learning approaches have been successful in achieving good performance in solving question answering (QA) task. In this project, we (i) explore all such popular existing approaches and improve their performance, (ii) propose new approach to tackle QA problem. Finally we compare performance of all implemented models and perform error analysis to understand model behaviour.

1 Introduction

The QA task is the task of determining answers to a question from a given context passage. In particular, the Stanford Question Answering Dataset (SQuAD) [1] used in this project is a reading comprehension dataset, consisting of questions on a set of Wikipedia articles. The answer to every question is a segment of text from the corresponding reading passage, or none if the answer is not present in the passage. SQuAD 2.0 is a popular dataset for machine comprehension and question answering task. As the QA task requires comprehensive understanding of natural language and ability to do inference and reasoning it can be quite challenging.

In this project, we first explore Bidirectional Attention Flow (BiDAF) model that makes use of a bi-directional attention flow mechanism to obtain a query-aware context representation without early summarization (1). Next, we have focused on improving the performance by using pre-trained Language Models(PrLM) and data augmentation. Finally, a new model architecture is proposed, which is inspired from the Retro-Reader (2)

In general, humans tend to approach the reading-comprehension problem as a two step process, where they first skim through the entire passage and try to extract the portion that is relevant to the question asked. Then, they carefully examine the extracted text to arrive at the exact answer. Inspired by how humans approach this problem, we divide the model architecture in 2 modules:

1. Relevant Context Extraction: This stage tries to extract the minimum required context to answer a given question and also an initial opinion about whether the question is answerable or not.

2. Exact Answer Extraction: This stage uses the previously extracted context to generate the final prediction.

2 Background

Transformers have demonstrated that a simple network structure based only on attention mechanisms (3) can have a good performance on machine translation tasks. This can be seen by the fact, that since 2019 transformer-based architectures have dominated the SQuAD 2.0 leaderboard. Different transformer architectures BERT, XLNet, RoBERTa, ALBERT, ELECTRA and ensemble models using transformers, have been used for QA task to achieve state of the art performances. One such example being ALBERT model trained by SRCB_DML, having an F1 score of 91.286%, which surpasses human performance. The transformer-based models uses self-attention mechanism to attain the contextualised embeddings, which is very relevant for question answering. Generally pre-trained version of these models are used as a base and fine-tuning is performed on the last few layers of the model using the SQuAD dataset to achieve high F1 scores. Different models are often ensembled together to increase performance. Li et al. (4) achieved an increase in F1 score of 1% by ensembling different ALBERT models.

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

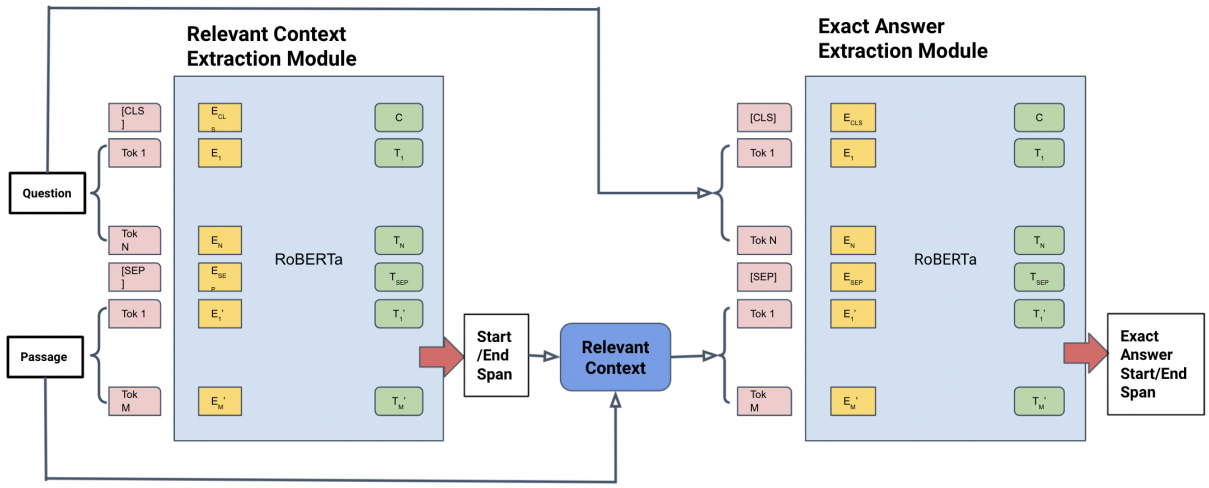

Figure 1: Proposed Model Architecture

3 Experiments

3.1 Dataset and Augmentation

We used Stanford Question Answering Dataset (SQuAD) V2.0 for our experiments (5). The input to the model is a question with a context paragraph, and the output should be the the span of text in the paragraph that answers the question. There are about half of the questions in the dataset that cannot be answered given the provided paragraph. We used 130319 examples as the training set, 11873 examples as the dev set. We experimented by augmenting the questions using the process of Back Translation. We used the library nlpaug[2] to do the same. The idea here is to help the network generalize to the syntactic variance in the question to generalize better at understanding questions and interactions between question and context.

Original Question: *What magazine rated Beyonce as the most powerful female musician in 2015?*

Question after back-translation: *Which magazine named Beyonce the most powerful female musician in 2015?*

3.2 Experimentation with existing architectures

3.2.1 Bidirectional Attention Flow Model

First model that we used to solve Q&A task was the BiDAF model. The BiDAF architecture consists of 3 basic layers: Embedding Layers, Attention and Modeling Layers, Output Layer.

In this architecture we encode the question and corresponding context paragraph using word embeddings, compute an attention matrix, and decode to find the answer to the question in the context paragraph.

3.2.2 Transformer based models

While bi-directional LSTMs are known to capture long term dependencies, their performance do not scale for larger tasks. Hence, to improve further, we use state-of-the-art transformer based model architectures. These models are much bigger and larger than traditional RNN networks.

We leverage 3 main characteristics of transformers i.e. self attention, non-sequential processing and positional embeddings. For QA task, we experimented by fine-tuning several transformer models (already pre-trained on large corpora like Wikipedia, Book Corpus etc.).

These models[3] includes case/uncased, base/large variants of BERT, XLNet, RoBERTa, ALBERT, DistilBERT. While fine-tuning we experimented with various model parameters and hyperparameters like - different tokenizers, doc stride, learning rate, number of epochs, weight decay, adam optimizer epsilon, gradient accumulation steps, maximum gradient norm, maximum answer length etc. Out of all aforementioned models, roberta-large gave the best performance as shown in Table 1

3.3 Proposed Model Architecture

Our model architecture based on human reading-comprehension is implemented as a two stage pipeline using the Roberta base model[4]: ***Extract***

[2] https://github.com/makcedward/nlpaug

[3] https://huggingface.co/transformers/pretrained_models.html

[4] Question-Answering using Transformer based models

Model	Variant	F1	EM	Params
BiDAF	-	60.8	57.4	<1M
BERT	base	80.3	77.3	108M
	large	83.9	80.8	340M
ALBERT	base	79.1	76.1	11M
	large	82.1	79.0	17M
RoBERTa	base	84.7	81.2	125M
	large	87.1	83.9	355M
	augmented	86.7	83.1	355M
XLNet	base	84.1	80.8	110M
	large	86.7	83.4	340M
DistilBERT	base	77.4	76.2	66M
Our Model	base	79.6	76.3	250M

Table 1: Comparison of model performance, BiDAF, finetuned-transformer, Proposed model

Answer Sentence Module , and ***Shrink Context Module***

3.3.1 Extract Answer Sentence Module

In this module, we first modify the training dataset to generate expanded answer spans for model training. Each original question-answer triplet $\langle C, Q, A \rangle$, where C is context, Q is question and A is the answer, is modified to $\langle C, Q, A' \rangle$ where A' captures the context relevant to the question. In order to generate A' we use spacy[5] to tokenize C into sentences and concatenate only those sentences which span the original answer A.

Using the modified train-set, we fine-tune a PrLM (specifically Roberta) to predict the relevant span from the original context. The question and context are concatenated together and passed as input to the PrLM, where as A' is used to generate the start vector and the end vector to mark the answer in the context for training.

3.3.2 Exact Answer Extraction Module

Objective of this module is to generate predictions given the relevant context extracted using the previous module. Training data is generated by modifying each original question-answer triplet $\langle C, Q, A \rangle$ to $\langle C', Q, A \rangle$ where C' is the shrunken context.

Another PrLM (Roberta) is trained using the same procedure as in Extract Answer Sentence Module to predict the actual answer given the shrunken context.

[5] https://spacy.io/api/tokenizer

4 Results and Analysis

Official metrics EM (Exact Match) and F1 score are used to evaluate the performance on the SQUAD data set. Table 1 compares the performance of each of the implemented models those are: baseline BiDAF model, fine-tuned state-of-the-art transformer based models and our proposed model pipeline. Based on the results, we make the following observations:

1. Transformer based models outperform the baseline BiDAF model.
2. Finetuned RoBERTa and XLNet produce comparable results, with highest F1 and EM score for their large variants.
3. No improvement with Data Augmentation. One possible reason might be because the questions were collected from the crowdsourced workers and might already cover syntactic variations.
4. Our pipeline model which is based on 'RoBERTa base' produces comparable results to the ALBERT model, but is not able to outperform the original RoBERTa model. The degrade in performance can be attributed to the sub par performance of the first model in extracting the relevant context from the passage. When tested independently using ground-truth dev set, the Exact Answer Extraction Module gives EM: 91.9 and F1: 95.6, which shows the potential in increasing the overall performance if the first model can be trained to correctly extract the relevant part of the context.

5 Error Analysis

Error analysis was performed using a sample of 3000 questions which had answers and predictions by Roberta Large, Distilled Bert and the newly proposed pipeline model. Of the 3000 samples used for the analysis, the predictions which did not match ground truths exactly for the models were as follows: Roberta Large: 623, Distilled Bert: 1061, Pipeline Model: 817. The results observed can be seen in Table 2.

On average questions with higher dependency distances were more likely to not match exactly in the case of all the models. Dependency distance is the distance between a key question token and the answer token. The key is computed by finding tokens that do not occur frequently in the context and is not far from the given answer.

Average Values	Pipeline Model		Roberta Large		Distilled Bert	
	Exact	Not Exact	Exact	Not Exact	Exact	Not Exact
Dependency Distance	4.2	6.36	4.37	6.34	3.86	6.52
Question Length	11.84	11.49	11.85	11.37	11.75	11.74
Answer Length	2.51	2.89	2.53	2.95	2.43	2.94

Table 2: Average values of different measures performed on predictions for error analysis.

An analysis on questions starting with 'wh' revealed that in case of all the models for the 3000 samples the number of 'why' questions for which the models got an exact match was less than for non exact matches. Among the rest of the question types, the "what" question type had the highest proportionality of predictions not matching exactly with the ground truth. This behavior commonly seen across the all models analysed, maybe due to the fact that while the question types 'which', 'when', 'who', 'whose' require much more objective answers, 'what' and 'why' type of questions can have a more general answer. In answer type analysis it was found that description type answers had a higher chance of not matching exactly. Description type answers can be much more subjective, compared to other answer types namely entity, location, number, year and human causing the models to give answers which might sometimes not be the same as ground truth.

Among the 3000 samples there were 339 questions common, among all the models analysed, for which the predictions were not exact matches to the true answers. One example of such context, question and answer was:

Context : *Following the Peterloo massacre of 1819, poet Percy Shelley wrote the political poem The Mask of Anarchy later that year, that begins with the images of what he thought to be the unjust forms of authority of his time—and then imagines the stirrings of a new form of social action. It is perhaps the first modern[vague] statement of the principle of nonviolent protest.*

Question : *The Mark of Anarchy was written to protest against what?*

Possible Answers : {*Peterloo massacre, the unjust forms of authority*}

Roberta Large, Distilled Bert and the proposed Pipeline model all failed to produce an answer for the above example. The question type in above example is "what" and the answer is description type. Although the answer is present in the first line of the passage yet the answer itself is abstract. As conjectured previously the models are not able to perform well on "what" question types and description answer types compared to other question and answer types due to the subjective nature of the answer.

6 Conclusion

In this project we tried to tackle SQuAD 2.0 challenge using BiDAF, transformers and the a new proposed pipeline model. After experimenting with multiple models, our experimental evaluations showed that fine tuning a Roberta Large model gave the best results amongst all. We also tried to analyse the predictions from our trained models in order to identify in which cases our models were failing to give correct answers.

Some limitations of our work stemmed from memory and computational limits which prevented us from training the proposed pipeline model with different hyperparameter values in order to see if we could further improve its performance. Overall, the project helped us get a good understanding on current state-of-the-art question answering models and how to build, train and evaluate transformer based models for our own tasks.

7 Future Work

The F1 score of Exact Answer Extraction Module tested independently using ground-truth dev set is 95.6. This highlights potential in the pipeline architecture if first model can be improved. For future work, we could fine-tune and improve the Relevant Context Extraction module to see overall improvements in the pipeline architecture. Also, linguistic post-processing for the "when", "Where", "Whose", "Which" questions can be experimented with. Furthermore, ensembling different transformer models is another direction to look at.

References

[1] *Bidirectional attention flow for machine comprehension*, Min Joon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi.

[2] *Retrospective Reader for Machine Reading Comprehension*, Zhuosheng Zhang and Junjie Yang and Hai Zhao, 2020

[3] *Attention is all you need*, 2017, Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez,Lukasz Kaiser, and Illia Polosukhin.

[4] *Ensemble ALBERT on SQuAD 2.0*, https://arxiv.org/pdf/2110.09665.pdf

[5] *Know What You Don't Know: Unanswerable Questions for SQuAD*, https://arxiv.org/abs/1806.03822

[6] *Ensemble BERT with Data Augmentation and Linguistic Knowledge on SQuAD 2.0*, https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1194/reports/default/15845024.pdf

[7] *BERT-A: Fine-tuning BERT with Adapters and Data Augmentation*, https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1194/reports/default/15848417.pdf

Open source code: Hugging face