# Question Answering on SQuAD 2.0 Dataset

*Yuanjun Li,*[1] *Yuzhu Zhang*[2]

[1]*Department Energy Resources Engineering, Stanford University*
[2]*Department of Electrical Engineering, Stanford University*

Stanford
CS224n NLP

## Abstract

Reading comprehension is a challenging task for machine learning, since the system must be able to model the complex interactions between the question and the context paragraph. In this study, we explore the performance difference of task-specific model BiDAF (Bi-Directional Attention Flow) and the pretrained BERT (Bidirectional Encoder Representations from Transformers) model on the QA tasks of SQuAD 2.0 (Stanford Question Answering Dataset), also proposed three self-designed model structures. The baseline BiDAF model achieves 60.5% F1 and 57.4% EM on the validation set. Our modified fine-tuned model on BERT achieves F1 and EM scores up to 76.6% and 73.6% respectively, with a promising performance on answerable questions but poor on the non-answerable ones. The proposed ensemble BERT QA + Classifier model can alleviate this problem and improve the F1 and EM scores further to 78.1% and 75.3%.

## Introduction

In 2018, large-scale pre-trained language modes such as OpenAI GPT and BERT have achieved great performance on multiple language tasks using generic model architectures. For BERT, it is pre-trained with two auxiliary tasks(the Mask Language Model task and the Next Sentence Prediction task) with large corpus to encourage the bi-directional prediction on text as well as sentence-level understanding, Since many important down-stream tasks such as Question answering (QA) and Natural Language Inference (NLI) are based on understanding the relationship between pair of sentences, BERT can perform well when fine-tuned on these downstream specific tasks without customized network architectures. On the SQuAD 2.0 leaderboard, all the top 10 model performers are developed based on BERT.

## Models

### Fine-tuned BERT-Base Model

BERT makes use of the Transformer, an attention mechanism that learns contextual relations between words in a text. For QA tasks, context-question pairs are packed together into a single sequence and separated by a special token ([SEP]). Token embeddings, segment embeddings and position embeddings are applied. Configuration of BERT-base model: it contains 12 transformer blocks, 768 hidden layers, 12 self-attention layers and 110M parameters

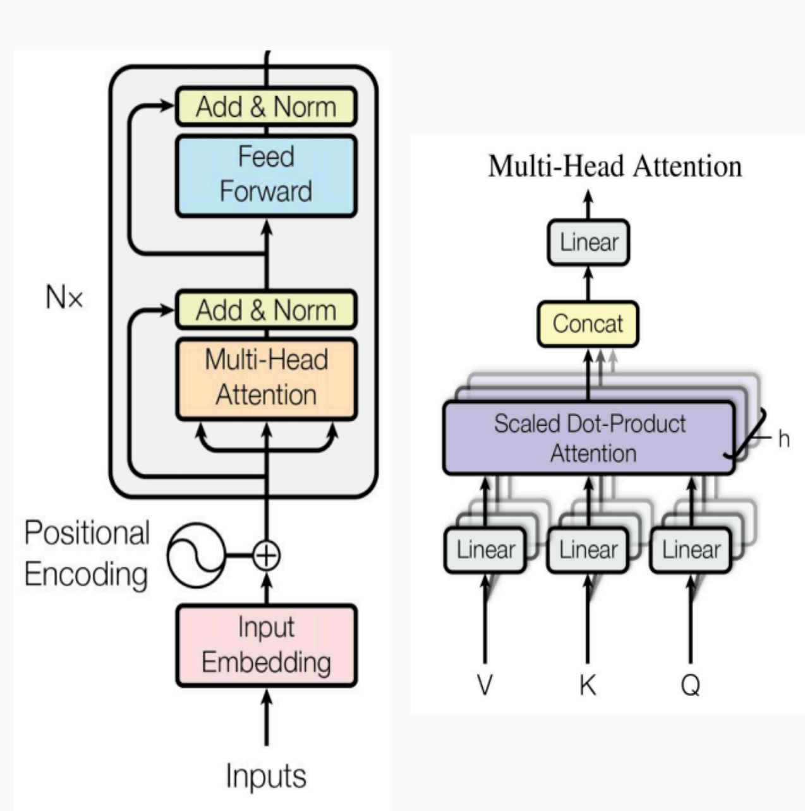

Fig.1 Transformer blocks

### Extension #1: Fine-tuned BERT QA Model with Modified Output layer

In this part, we aim to improve the model performance by altering the output layers on BERT. As shown in Fig.2, our final best-performed architecture consists of three main layers. Highway networks and bi-LSTMs are added to model the interactions on the BERT output embedding matrix.

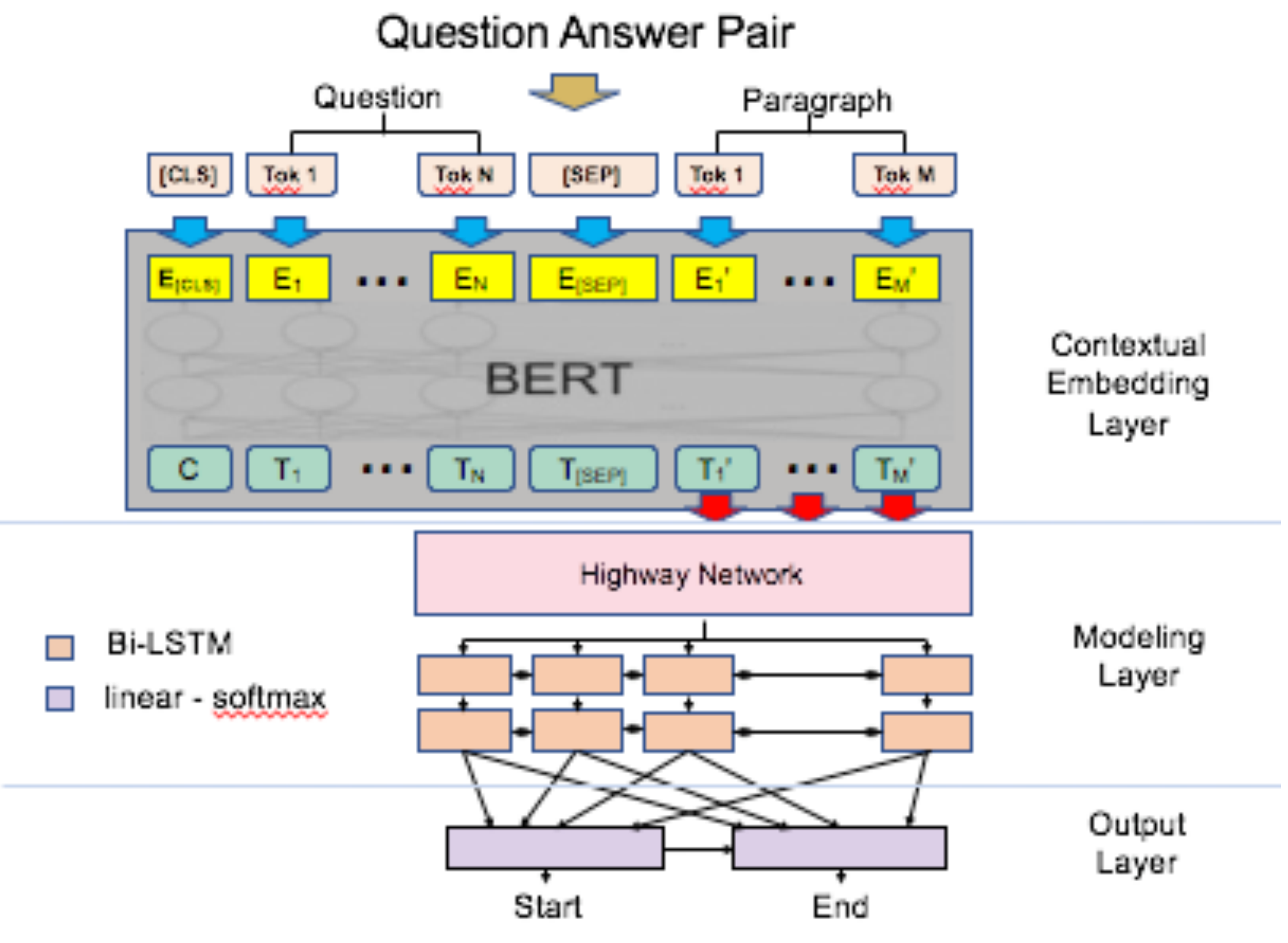

Fig.2 BERT QA Model with Modified Output layer

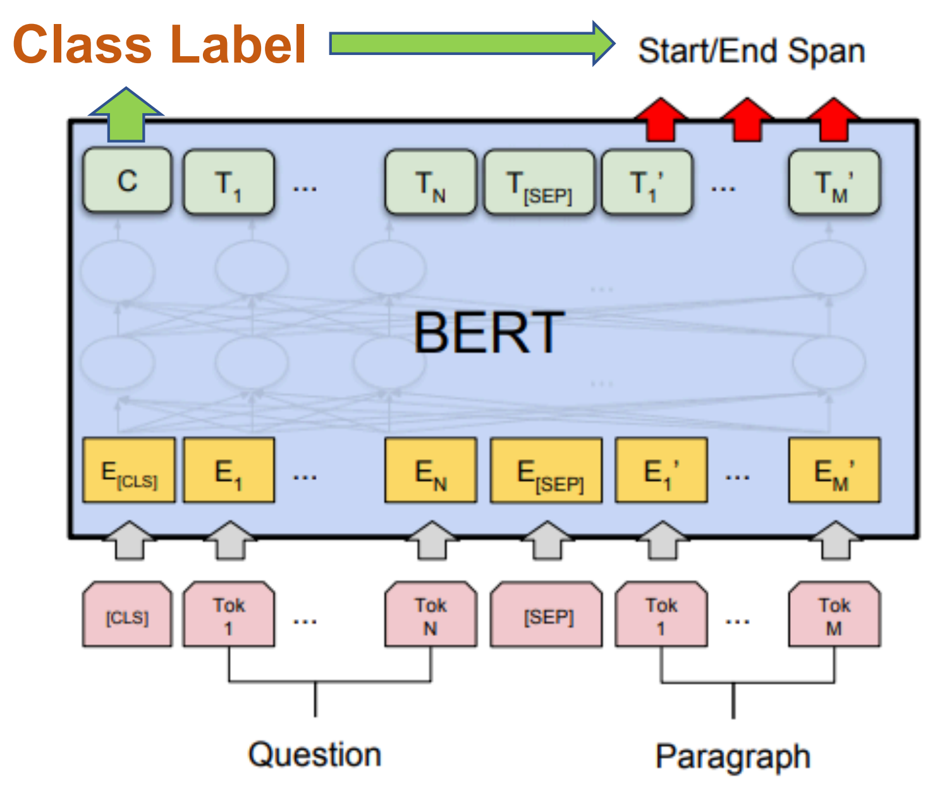

Fig.3 BERT QA + Classifier Model(Single)

### Extension #2: BERT QA + Classifier Model(Single)

We find the first token [CLS] is an indicator to emit logit for "no answer".(Fig.3) Therefore, we take the final hidden state for the first token [CLS] as extra embeddings from the QA model. The vector is denoted as $Q \in R^h$. Then the predicted class $Cp$ could be constructed with a sigmoid regression layer:

$$Cp = sigmoid(W * Q + b)$$

The modified training loss is a weighted average of the start and end position losses and also the cross entropy loss of the predicted labels $Cp$ compared to true labels $Cr$

$$Loss = [\alpha\big(CE(start_p, start_r) + CE(end_p, end_r)\big) + \beta CE(C_p, C_r)]/(2\alpha + \beta)$$

### Extension #3: BERT QA + Classifier Model(Ensemble)

Rather than combining the QA and classification tasks as a single model in extension #2. We separately fine-tuned the QA and classification models, and ensemble the two models to produce final output.(Fig.4) The model predicts no-answer as long as there is one vote from any of the two models.

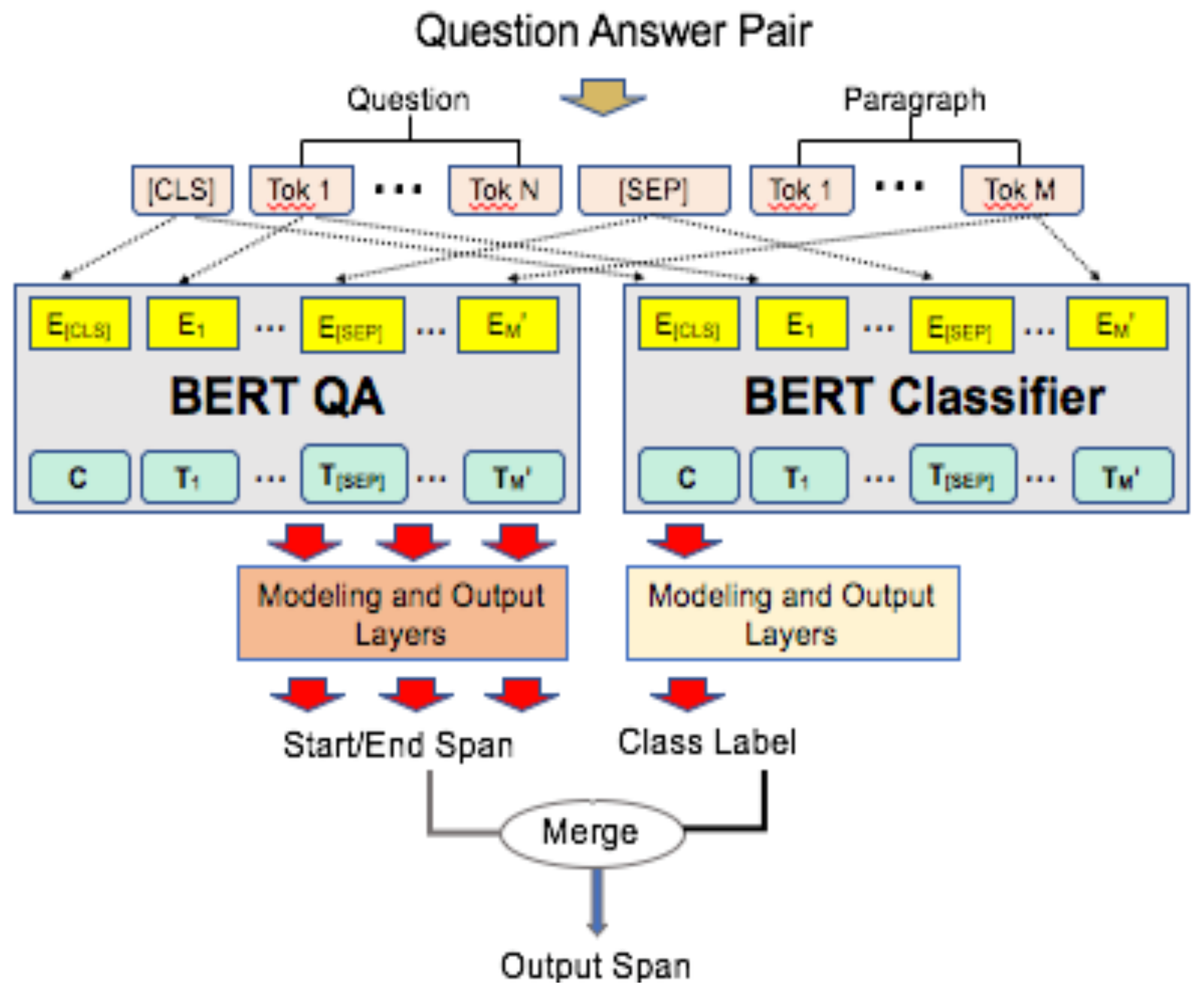

Fig.4 BERT QA + Classifier Model(Ensemble)

## Experiment Results

| Models | Total Questions | | Answerable Ques. | | Non-anwerables | |
|---|---|---|---|---|---|---|
| | F1 | EM | F1 | EM | F1 | EM |
| Baseline(BiDAF) | 60.5% | 57.4% | - | - | - | - |
| Fine-tuned BERT | 74.5% | 72.2% | 70.0% | 65.2% | 78.6% | 78.6% |
| Extension #1 | 76.6% | 73.5% | 81.3% | 74.6% | 72.4% | 72.4% |
| Extension #2 | 75.3% | 73.1% | 74.7% | 70.0% | 76.0% | 76.0% |
| Extension #3 | **78.1%** | **75.3%** | **77.1%** | **71.2%** | **79.0%** | **79.0%** |

## Conclusion

During our experiments, it is found that BERT embedding is super powerful hence it could be suitable for many NLP tasks. It is also found that, as we increase the complexity of the fine-tuned model, the performance of non-answerable questions decreases hugely. Therefore we propose two QA + Classifier models to balance the discrepancy and successfully improve the total model performance.

## Reference

[1] Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for squad. arXiv preprint arXiv:1806.03822, 2018.
[2] Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. Bidirectional attention flow for machine comprehension. arXiv preprint arXiv:1611.01603 , 2016.
[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 , 2018.
[4] Julian G. Zilly, Rupesh Kumar Srivastava, Jan Koutník, and Jürgen Schmidhuber. Recurrent highway networks. CoRR , abs/1607.03474, 2016