Who is Ernie? Just ask Bert!

Barthold Albrecht, Yanzhuo Wang, Xiaofang Zhu

Stanford Computer Science

Abstract

Question answering, like many other prominent tasks of NLP problems, has recently experienced a significant progress on established performance measures through the introduction of pretrained language model representation. At the same time, the concept of multitask learning which tries to integrate different fields of NLP has gained considerable momentum. In our work we combine both lines of research and show that adding an auxiliary tasks to a BERT-based question answering system can improve the performance for a single given task. We discuss the implications of our findings for this specific task as well as for multitask learning approaches in general.

Stanford University

Introduction

The Stanford Question Answering Dataset (SQuAD) has become the standard benchmark for assessing the reading comprehension ability of an NLP algorithm. Given that the pre-trained language model BERT immediately surpassed human level performance on SQuAD 1.1 in the F1 score while no variant of BERT yet managed to do so for SQuAD 2.0 our working hypothesis is that these models have not yet learned sufficiently well to distinguish between certain question types (answerable or not).

Goal

Our goal is to follow a multi-task learning approach using BERT as the core model and in this way improve the performance on SQuAD 2.0.

Related Work

- Ruder et al. An overview of multi-task learning in deep neural networks. arXiv preprint 1706.05098, 2017.

- Devlin et al. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint 1810. 04805, 2018.

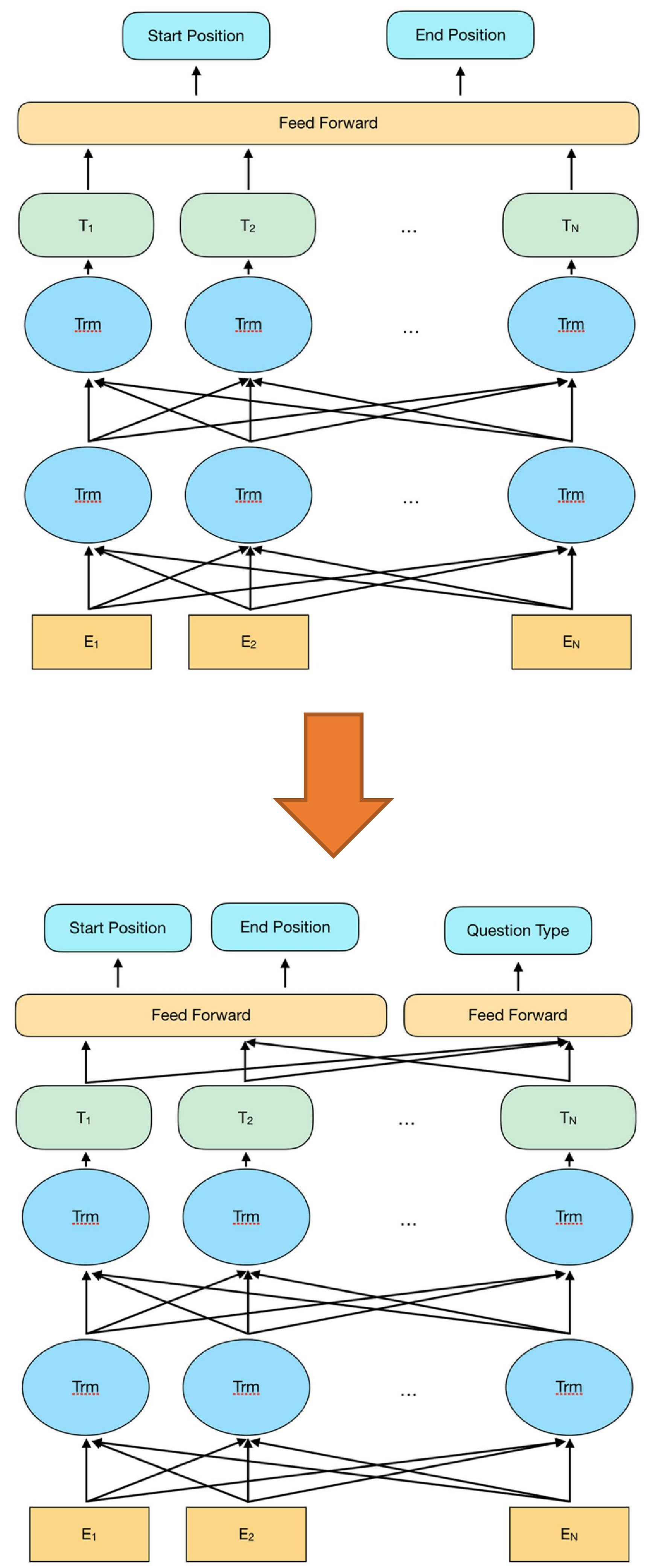
- Liu et al. Stochastic answer networks for SQuAD 2.0. arXiv preprint 1809. 09194, 2018.

Approach

- **Baseline**

BiDAF model provided to us with the starter code.

- **BERT Multitask setting**



Experiments

- **Data**

The dataset we used is the SQuAD 2.0 dataset tailored for CS224N. The train, dev and test splits are pre-defined.

Dataset	# of examples
train	129914
dev	6078
test	5921

- **Models & Parameters**

Tab. 1: Model Arch Parameters

	BERT_base	BERT_large
layer	12	24
hidden	768	1024
heads	12	16
parameters	110M	340M

Tab. 2: Model Configure Parameters

	BERT_b	BERT_mt_b	BERT_mt_l
epoch	2	2	2
b_size	6	6	4
max_seq_len	384	384	384
class_loss_factor	N/A	7	7

Results and Analysis

Tab. 3: Results

	EM	F1
BiDAF	57.491	61.097
BERT_b	73.017	76.086
BERT_mt_b	73.593	76.538
BERT_mt_l	74.038	77.412

Tab. 4: Answerable/Unanswerable Prediction

	Precision	Recall	F1
BiDAF	78.05	52.97	63.11
BERT_b	86.18	72.63	78.83
BERT_mt_b	81.01	80.52	80.76
BERT_mt_l	86.79	75.92	80.99

Conclusion

Our approach on multitask learning shows that the addition of a classification task for the answer type enhances the performance of a BERT-based model for question answering. Starting from there we will further investigate how to make additional use of the quite accurate predictions of the classifier (> 95%) and continue our respective experiments which yet failed to improve the score.