Multi-Task Deep Neural Networks for Generalized Text Understanding

George He
georgehe@stanford.edu

Introduction

Stanford Question Answering Dataset (SQuAD[3]) challenge:
Predict answers, which are a segment of a provided context/unanswerable

Evaluation Metrics
EM Score: predictions that match any one of the ground truth answers exactly
F1 Score: average overlap between the prediction and ground truth answer

Goal
Improve existing models by utilizing pretrained contextual embeddings and multi-task learning objectives to incorporate knowledge from related language understanding problems

Datasets

Sentence Classification (SC)
Corpus of Linguistic Acceptability(CoLA)
Expertly annotated for acceptability (grammar)
Multi-Genre Natural Language Inference (MNLI)
Sentence pairs - textual entailment information
Microsoft Research Paraphrase Corpus (MRPC)
Sentence pairs - paraphrase equivalence

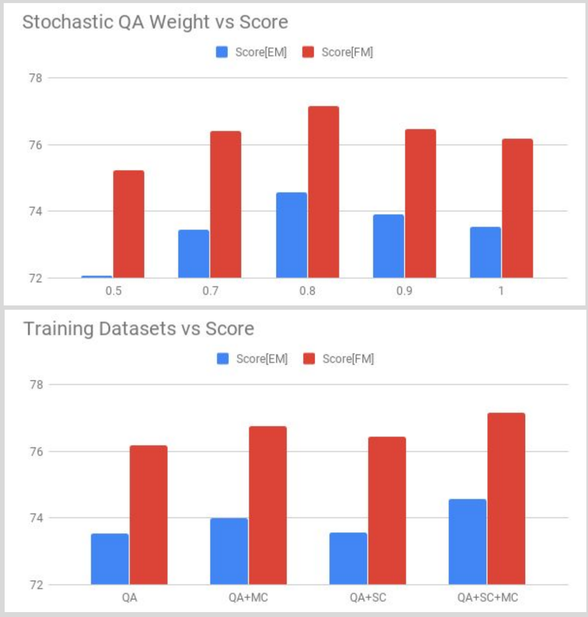
Multiple Choice (MC)
SWAG
Ground truth common-sense inference

Question Answering (QA)
Stanford Question Answering Dataset 2.0

Each task-specific dataset had its own output layer, with a shared encoding and embedding layer.

Cross entropy loss between the ground truth and the predicted outcomes was used across all datasets.

Results



Conclusion

Minor improvements from properly tuned training across different datasets

Catastrophic forgetting is an issue with mixed datasets - requires careful weighting

After experimentation, we determined an optimal weight during stochastic batch sampling for the QA dataset is 0.8.

bert-base validation EM/F1: 73.373/76.421

bert-large validation EM/F1: 77.853/81.053

Methods

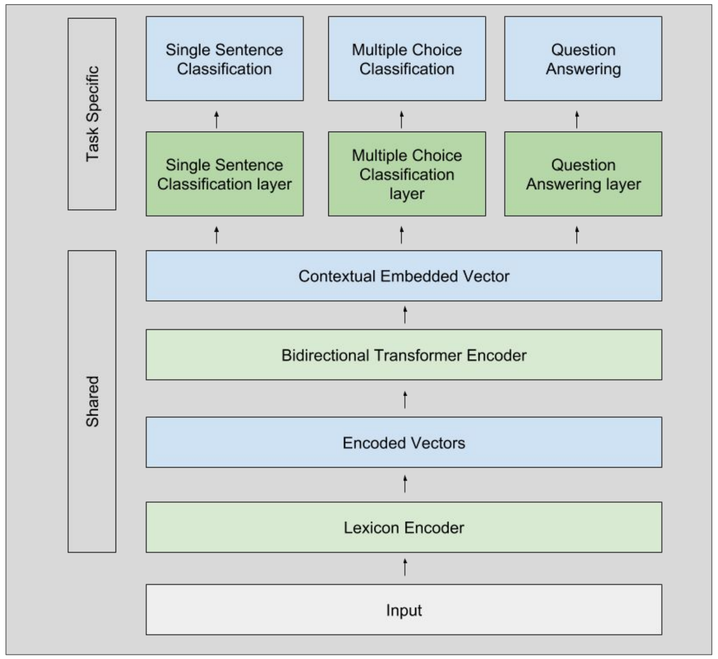
BERT[1]:
Pretrained Contextual Embeddings and Deep Bidirectional Transformers

Multi-task Learning Model[2]:
Learn from multiple datasets, applying task-specific loss to train common contextual model

Weighted Stochastic Batch Samples
Carefully incorporate new datasets to avoid catastrophic interference

We modify the task-specific BERT architecture to incorporate and train across multitask learning objectives, and apply a training routine favoring the question answering task

Architecture



Sample Answer

Context: On October 6, 1973, Syria and Egypt, with support from other Arab nations, launched a surprise attack on Israel, on Yom Kippur. This renewal of hostilities in the Arab–Israeli conflict released the underlying economic pressure on oil prices. At the time, Iran was the world\'s second-largest oil exporter and a close US ally. Weeks later, the Shah of Iran said in an interview: "Of course [the price of oil] is going to rise... Certainly! And how!... You've [Western nations] increased the price of the wheat you sell us by 300 percent, and the same for sugar and cement... You buy our crude oil and sell it back to us, refined as petrochemicals, at a hundred times the price you've paid us... It's only fair that, from now on, you should pay more for oil. Let's say ten times more."

Question: How many times more did the other nations have to pay for oil after the surprise attack?

QA+MCC+SC answer: ten times more
QA+SC answer: ten times more
QA+MCC answer: ten
QA answer: ten

References

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding.CoRR, abs/1810.04805, 2018.

[2] Weizhu Chen Jianfeng Gao Xiaodong Liu, Pengcheng He. Multi-task deep neural networks for natural language understanding. https://arxiv.org/abs/1901.11504, 2019.

[3] Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions forsquad.CoRR, abs/1806.03822, 2018.