BERT++
CS224N Default Final Project

Lea Jabbour, Sophia Barton, Juliet Okwara

Abstract

Our main goal is to produce a question answering (QA) system that performs well on SQuAD 2.0 and improves upon the BiDAF baseline, through use of the BERT model. We fine-tune BERT and add two different attention mechanisms to BERT: Attention-over-Attention (AoA) and Dynamic Coattention Network (DCN). The fine-tuned BERT model achieves the highest scores: EM score of 73.69 and F1 score of 76.98 on the test dataset.

Problem

- **Task:** create a QA system that performs well on SQuAD 2.0
    - Determine when no answer is available
    - Correctly return the span of text which answers the question when possible
- **Importance:** QA can automate reading comprehension, and extract useful information from massive amounts of text
- **Approaches:**
    - Linguistics techniques (e.g. NER, Parsing, POS)
    - Deep learning
        - Non-PCE (e.g. BiDAF)
        - PCE (e.g. ELMo, BERT)

Data

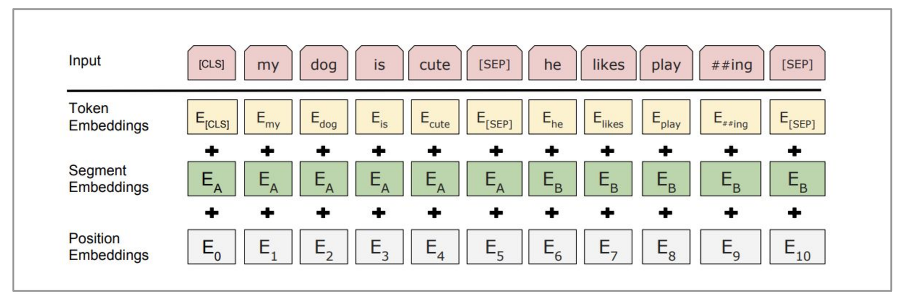
We are using the custom SQuAD dataset provided.

- train (129,941 examples): All taken from the official SQuAD 2.0 training set
- dev (6078 examples): Roughly half of the official dev set, randomly selected
- test (5921 examples): Remaining examples from the official dev set, and hand-labeled examples.

Dataset Example:
C: The kilogram-force leads to an alternate, but rarely used unit of mass: the **metric slug**...
Q: What is a very seldom used unit of mass in the metric system?
A: {"slug","answer_start":274},{"text":"metric slug", "answer_start":267}

Approach

I. **BERT**
- PCE (pretrained contextual embeddings)
- Multi-layer bidirectional Transformer encoder
- Pre-trained with two unsupervised prediction tasks: (1) Masked Language Modeling (MLM) (2) Next Sentence Prediction (NSP)
- Input representations:



II. **BERT + AoA**
- Split BERT output into queries & contexts
- Query-to-context (Q2C) attention
- Context-to-question (C2Q) attention
- Append weighted sum of Q2C to BERT output

III. **BERT + DCN** (v1, v2)
- Split BERT output into queries & contexts
- Compute attention both ways (Q2C, C2Q)
- Use C2Q to take weighted sum of Q2C, and concatenate with C2Q attention → biLSTM

Results

Table 1: Dev set scores

Model	EM	F1
BiDAF	57.32	61.1
BERT	74.48	77.7
BERT + AoA	35.36	46.9
BERT+ DCN	42.83	44.9

- BiDAF, BERT: as expected
    - BERT on test set: EM of 73.69, F1 of 76.98
- BERT + AoA, BERT + DCN: much lower than expected
    - Splitting
    - Tokens
    - Need for re-finteuning

Analysis

Table 2: NA analysis

Model	TP Rate	FP Rate	FN Rate
BERT	0.3823	0.053	0.139
BERT + AoA	0.339	0.08	0.181
BERT + DCN	0.396	0.193	0.125

I. **BERT**
- Partially correct predictions
- Struggles to locate names & locations

II. **BERT + AoA**
- Truncates predictions

III. **BERT + DCN**
- Produces single word (or partial word predictions)
- Focuses on certain parts of query, ignoring perhaps more relevant text
- Produces answer simply because query word appears close to context word

Conclusions

The BERT fine-tuned model performs best. We believe our additional attention mechanisms have potential to outperform this model, but require more hyperparameter
fine-tuning to achieve better performance.

Future Work:
- Null_score_diff_threshold tuning for NA problem
- Experiment with the bert-large-uncased model
- Ensembling.

References

1. Devlin, Chang, et. al. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*, 2018.
2. Cui, et al. *Attention-over-Attention Neural Networks for Reading Comprehension*, 2016.
3. Xiong, et al. *Dynamic Coattention Networks For Question Answering*, 2016.