SQuAD: Integrating PCE and Non-PCE approaches

TG Sido
osido@stanford.edu | Computer Science, Stanford University

Motivation

Question **A**nswering (QA) is an increasingly important NLP problem with the proliferation of chatbots and virtual assistants. In October 2018, **B**idirectional **E**ncoder **R**epresentations from **T**ransformers (**BERT**) was released and achieved state-of-the-art results on a variety of NLP tasks, including QA. We seek to extend BERT with other performant QA architectures for SQuADv2.0

Dataset

Over 150,000 examples from 23,215 Wikipedia paragraphs in the following format:

P: …Bismarck was aware that public opinion had started to demand colonies for reasons of German prestige. He was influenced by Hamburg merchants and traders, his neighbors at Friedrichsruh. The establishment of the German colonial empire proceeded smoothly, starting with German New Guinea in 1884.

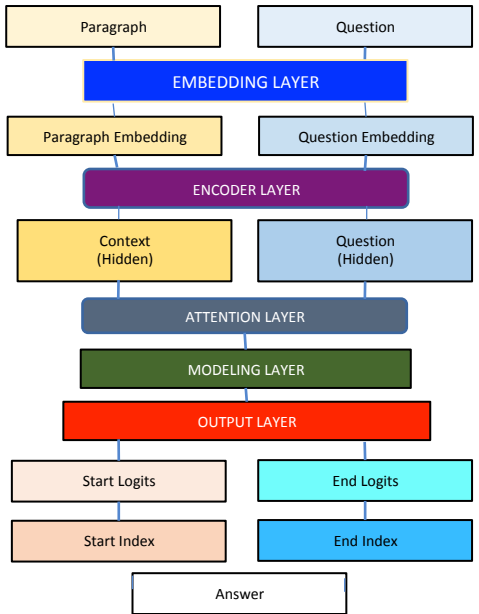
Q: Colonies were a sign of what amongst European countries?

Answerable: TRUE

A: prestige

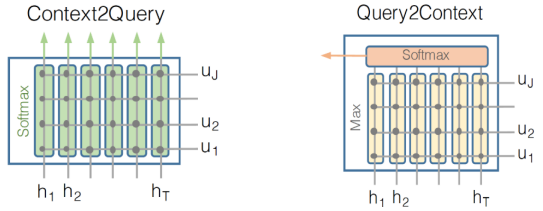
Legend: P (paragraph); **Q** (question); **A** (answer)

Solution Architecture



Bi-Directional Attention Flow (BiDAF)

- Similarity matrix S (NxM)
- C2Q: weighted sum of question states → f
- Q2C: weighted sum of context states → g
- Output: stacked combination of f and g



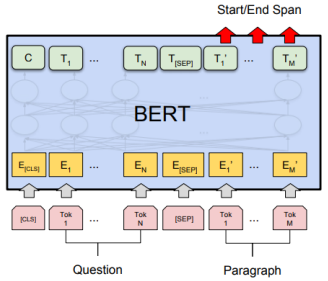
Dynamic Coattention Network (DCN)

- Affinity matrix L ((N+1)x(M + 1))
- C2Q: weighted sum of question states → a
- Q2C: weighted sum of context states → k
- 2nd level attention: weighted sum of k states → s
- Output: bi-LSTM encoding of stacked s and a

Answer-Pointer

- conditions end prediction on start prediction
- conducts two passes over modeling layer outputs with RNN/GRU
- second pass: uses final hidden state to facilitate end logits' dependence on start logits

BERT



- pre-train bidirectional representations by conditioning on both left and right context
- uses multi-layer bidirectional transformer encoder, transformer blocks, hidden states, self-attention heads, and a feed-forward filter
- uses bidirectional self-attention

References:
- Devlin, J. & Chang, M.W. (2018) BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, arXiv:1810.04805
- Seo, M. & Kembhavi, A.F. (2016) Bidirectional Attention Flow for Machine Comprehension, arXiv:1611.01603
- Xiong, C. & Zhong, V. (2016) Dynamic Coattention Networks for Question Answering., arXiv:1611.01604

BERT Integration

To integrate the non fine-tuned BERT embeddings into our BiDAF and DCN implementations, we projected the BERT embeddings down, which were originally 768-vectors, to match the GLoVE dimensions. Then, we pass in GLoVE + GLoVE ⊙ BERT, where ⊙ is the hadamard operator.

Results

Model	EM	F1
BERT-CASED	**71.9**	**75.3**
BERT-UNCASED	71.6	74.7
BERT-BiDAF	56.4	59.4
BERT-DCN	52.7	56.1
DCN	54.1	56.8
BERT-Answer-Pointer (RNN)	43.4	49.7
BERT-Linear-Answer-Pointer	67.7	71.1
BiDAF Baseline	55	58

Error Analysis*

Model	Accuracy	AvNa
BERT-CASED	**74.0**	**53.8**
BERT-UNCASED	66.0	38.5
BERT-BiDAF	68.0	38.5
BERT-DCN	52.0	38.5
DCN	48.0	7.7
BERT-Answer-Pointer (RNN)	54.0	61.5
BERT-Linear-Answer-Pointer	62.0	30.7

*selected 25 random examples to compare models

Discoveries

- BERT-Linear-Answer-Pointer often attempts to answer unanswerable questions:
  - This occurs due to magnified logits because of the lack of normalization when adding logits from linear and Answer Pointer layers
- BERT-BiDAF/DCN vs BiDAF/DCN:
  - BERT-X models converge faster to maximum EM and F1 than their non-BERT counterparts due to incorporating more contextual information via BERT's FastText approach
- BERT-Answer-Pointer suffers from early-summarization:
  - arises due to RNN, which does not have the additional memory gate that LSTMs and GRUs have, hampering the ability to understand long-range dependencies

Conclusions

- Non fine-tuned BERT embeddings can help speed up training in non-PCE implementations
- BERT-CASED is the most performant model; however, it's main issue is attempting to answer unanswerable questions
- We need to develop better mechanisms to determine answerability
- True understanding of text is still a significant challenge