Question Answering with BERT and Answer Verification

Kevin Culberg (kculberg@Stanford.edu)

Problem

The task of reading comprehension with unanswerable questions challenges a models ability to both correctly predict an answer span as well as determine if the question can even be answered. Many solutions involve adding placeholder values to the output to predict unanswerable questions. I propose combining a method of answer verification[2] with BERT[1] to improve detection of impossible questions.

Data/Task

Task: Reading Comprehension/Question Answering

- Given a question and context paragraph return a span of that paragraph as an answer to the question or return nothing if the question is impossible

Dataset: SQuAD 2.0[3] containing approximately 150,000 questions with half being impossible to answer.

- Questions are paired with context paragraphs that contain the answer as a single sub sequence if the question is answerable
- Answerable questions have multiple human answers for computing scores

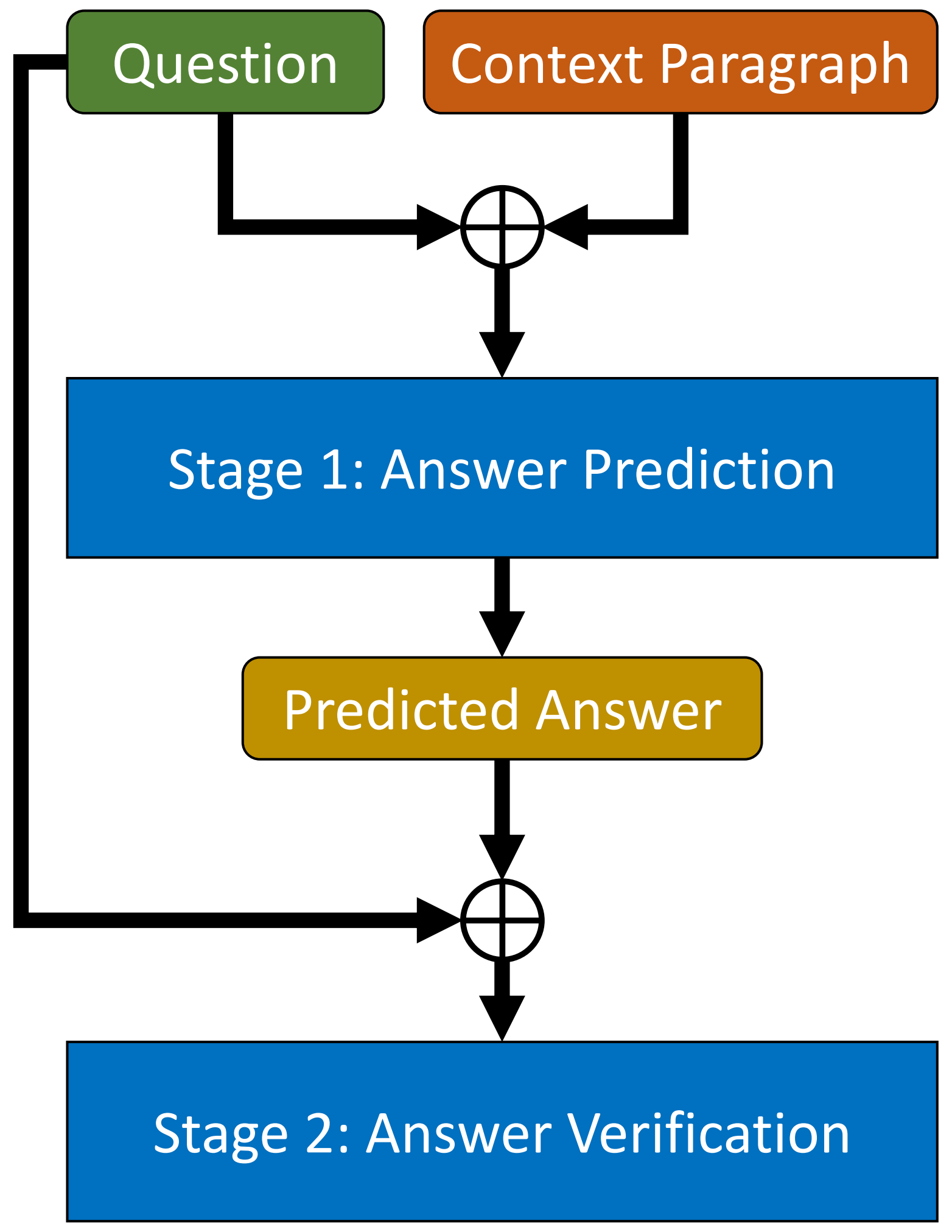
Example:

Question: *Why was Tesla returned to Gospic?*

Paragraph (answer underlined):

On 24 March 1879, Tesla was returned to Gospic under police guard for <u>not having a residence permit</u>. On 17 April 1879, Milutin Tesla died at the age of 60 after contracting an unspecified illness (although some sources say that he died of a stroke). During that year, Tesla taught a large class of students in his old school, Higher Real Gymnasium, in Gospic.

Approach



Model consists of two stages:

Stage 1:

- Input: Question and context paragraph
- Output: probability for start and end locations

Stage 2:

- Input: Question and predicted answer
- Output: probability answer is valid

Models:

- BERT-base: no stage2
- BERT+AV1: no transfer learning, uses dropout
- BERT+AV2: transfer learning from BERT-base, no dropout

Stage Architecture:

- The body for both stages is BERT with shared weights
  - 12 transformer blocks
  - Hidden size of 768
  - 12 self-attention heads
- The head for each is a linear to predict outputs

Loss:

- Combines loss for both stages
- Loss function only considers answer position loss if the question is actually answerable

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. Minghao Hu, Furu Wei, Yuxing Peng, Zhen Huang, Nan Yang, and Ming Zhou. Read + verify: Machine reading comprehension with unanswerable questions. CoRR, abs/1808.05759, 2018.
3. Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100, 000+ questions for machine comprehension of text. CoRR, abs/1606.05250, 2016.

Results

	Test Set		Dev Set			
	EM	F1	EM	F1	Precision	Recall
BERT-base	**59.679**	**63.026**	**59.230**	**62.353**	**76.23%**	57.89%
BERT+AV1	50.871	53.389	53.455	55.525	60.59%	**67.80%**
BERT+AV2	48.318	52.088	48.602	52.548	71.37%	41.32%

*Precision and recall are calculated for correctly predicting if a question is answerable.

Analysis

While the addition of answer verification decreased the number of questions incorrectly predicted to be answerable, this came at the cost of greatly increasing the number of questions falsely labeled as unanswerable. It is likely that reusing the same weights for the bodies of both stages caused the model to be pulled in two different directions simultaneously which decreased performance on both tasks. This is especially clear by examining answers predicted for answerable questions where the model with answer verification is more likely to include extra words before or after the true answer.

Conclusions

The addition of answer verification can improve the prediction of impossible questions by decreasing the number of false positives. However, to fully realize the benefit would likely require separating the weights from the two stages. This would allow both sides to be trained independently and specialize in their own tasks. Future work could also include further tuning of the weight applied to the loss function for both stages.