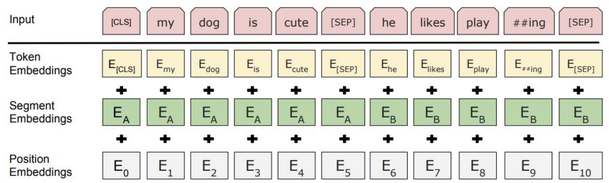
BERTNet: QANet + BERT for SQuAD 2.0 Question Answering

Hongtao Sun, Brett Szalapski, Yang Wang
{s3sunht, brettski, leonwy12} @stanford.edu

Introduction and Approach

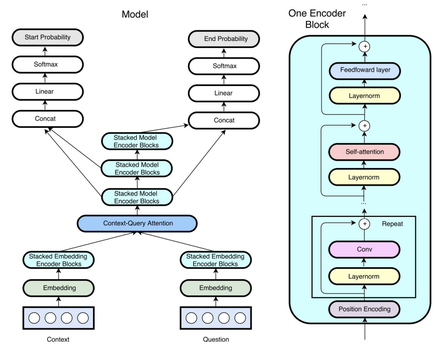
Problem Description: The SQuAD 2.0 challenge, a Question-Answering task. One of the most important Natural Language Processing challenges.

Approach:



BERT

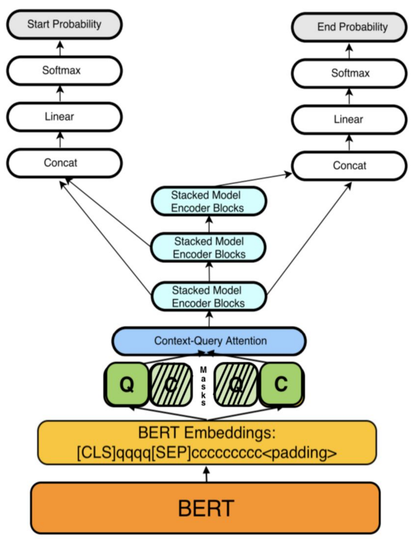
+



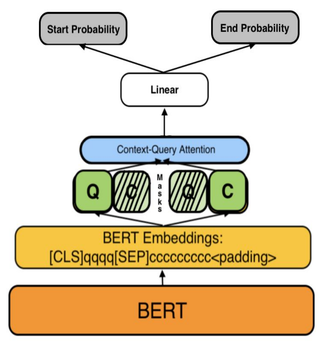
QANet

- Also trained a simpler model that used only a Context-Query attention layer after BERT to reduce overfitting.

Model Architectures



QANet + BERT



Context-Query Attention + BERT

Results

Model	Dev			Test	
	EM	F1	AVNA	EM	F1
QANet	57.44	60.97	68.22	x	x
BERT-small	72.97	76.41	80.40	x	x
QANet w/ BERT-small	**74.50**	77.34	80.49	x	x
CQ-BERT-small	74.38	**77.98**	81.67	x	x
BERT-large	**78.89**	**82.18**	85.29	**77.38**	**81.10**
CQ-BERT-large, Dropout = 0.1	77.90	80.81	83.81	77.31	80.23
QANet w/ BERT-large	77.34	80.21	83.46	73.76	76.96

References

Thomas Wolf, Victor Sanh, and Gregory Chatel et al. Pytorch pretrained bert: The big & extending repository of pretrained transformers. https://github.com/huggingface/pytorch-pretrained-BERT, 2019.

Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V. Le. Qanet: Combining local convolution with global self-attention for reading comprehension. CoRR, abs/1804.09541, 2018. URL http://arxiv.org/abs/1804.09541

Analysis

- The original BERT computes interactions between all words in the input. However, for Q&A, the interactions between context and query as separate groups may require more emphasis.
- CQ-BERT and QANet + BERT **outperform BERT-small**.
- QANet + BERT could outperform CQ-BERT with **data augmentation**.
- Failed to surpass BERT-large, but with more resources to perform further **hyperparameter tuning** it is likely that this result could be overturned.
- **Overfitting on BERT-large** variants may also contribute to the inferior performance .

Conclusion

- QANet and BERT can be combined to achieve near state-of-the-art results on the SQuAD 2.0.
- Using Context-Query Attention as the output layer for BERT may prevent overfitting and performs better, with an **F1-score of 80.23**.
- Our final model places **5th** on the CS224N 2019 Winter **leaderboard.**

- <u>Future work:</u>
  - Hyperparameter tuning on BERT-large
  - QANet Data augmentation