BERTQA: Attention on Steroids

Ankit Chadha and Rewa Sood

Introduction

Our BERTQA model tackles the question answering problem based on SQuAD 2.0 [1] and our augmented version, SQuAD 2.Q.We have tested the limits of applying attention in BERT [2] to improving the network's performance. BERT applies attention to the concatenation of the query and context vectors and thus attends these vectors in a global fashion. We propose BERTQA which adds Context-to-Query (C2Q) and Query-to-Context (Q2C) attention in addition to localized feature extraction. Our augmented dataset will be publicly available on our github [3]. After performing hyperparameter tuning, we ensembled our two best networks to get F1 and EM scores of 82.317 and 79.442 respectively. The experiments took around 300 GPU hours to train.

Model

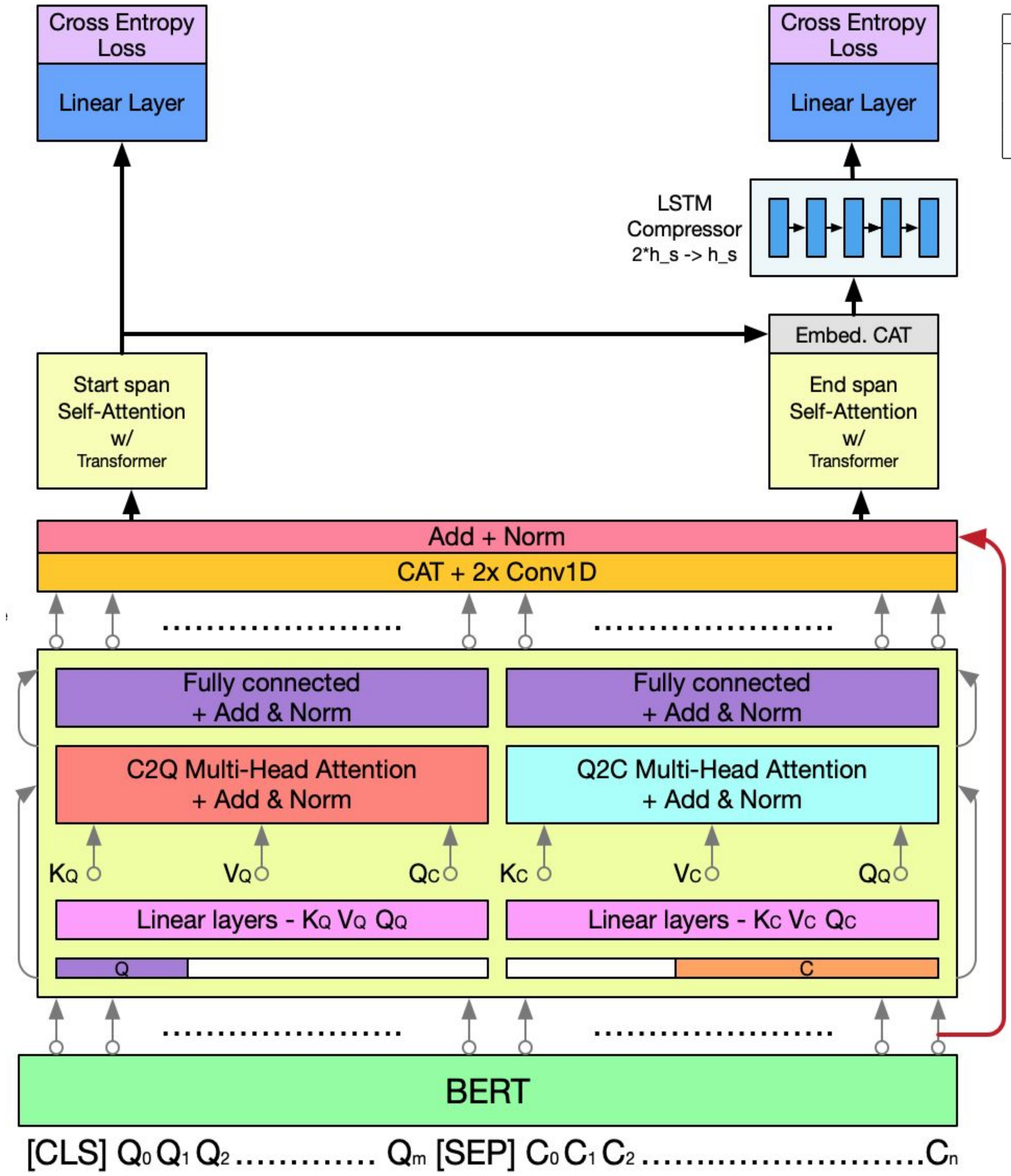
Cross Entropy Loss
Linear Layer
Cross Entropy Loss
Linear Layer
LSTM Compressor 2*h_s -> h_s
Embed. CAT
Start span Self-Attention w/ Transformer
End span Self-Attention w/ Transformer
Add + Norm
CAT + 2x Conv1D
Fully connected + Add & Norm
Fully connected + Add & Norm
C2Q Multi-Head Attention + Add & Norm
Q2C Multi-Head Attention + Add & Norm
K_Q
V_Q
Q_C
K_C
V_C
Q_Q
Linear layers - K_Q V_Q Q_Q
Linear layers - K_C V_C Q_C
Q
C
BERT
[CLS] Q_0 Q_1 Q_2............ Q_m [SEP] C_0 C_1 C_2........................C_n

Fig 1: BERTQA architecture using BERT, Co & Self Attention and LSTM

SQUAD2.Q

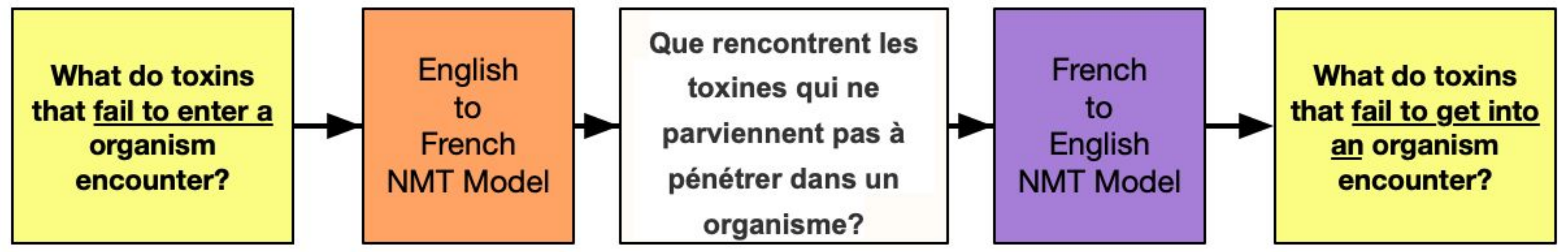
What do toxins that fail to enter a organism encounter?
English to French NMT Model
Que rencontrent les toxines qui ne parviennent pas à pénétrer dans un organisme?
French to English NMT Model
What do toxins that fail to get into an organism encounter?

Fig 2: Backtranslation example

- Question augmented version of SQuAD 2.0 using back translation
- Helps Network learn syntactic and grammatical variance in language when questions are posed in different ways based on paragraphs.
- Work can be further extended to Contexts

Results

	Base	C2Q/Q2C	Simple Skip	Transformer Skip	Inside Conv
F1	74.15	74.34	74.81	74.95	**77.03**
Has Ans F1	**80.62**	76.30	76.13	76.35	78.47
No Ans F1	68.21	72.54	73.01	73.66	**73.83**
EM	71.09	71.56	72.11	72.07	**74.37**

Table 1: Results for BERT Base compared to our additions. For the base model, the convolutional layers added inside the C2Q/Q2C coattention layers provided the best results. The transformer skip does not provide significant gains commensurate to the added overhead.

	BERT Large	Model 1	Model 2	Model 3	Ensemble
F1	80.58	82.08	81.51	81.31	**83.42**
Has Ans F1	**84.39**	84.36	83.53	84.38	81.09
No Ans F1	77.08	79.98	79.68	78.49	**85.56**
EM	77.74	78.81	78.24	78.00	**80.53**

Table 2: Results for BERT Large compared to our three best performing models and the ensembled model. The ensembled model performs the best in all categories except for the has answer F1. The significant gains in the no answer F1 can be attributed to both coattention, data augmentation and the ensembling method.

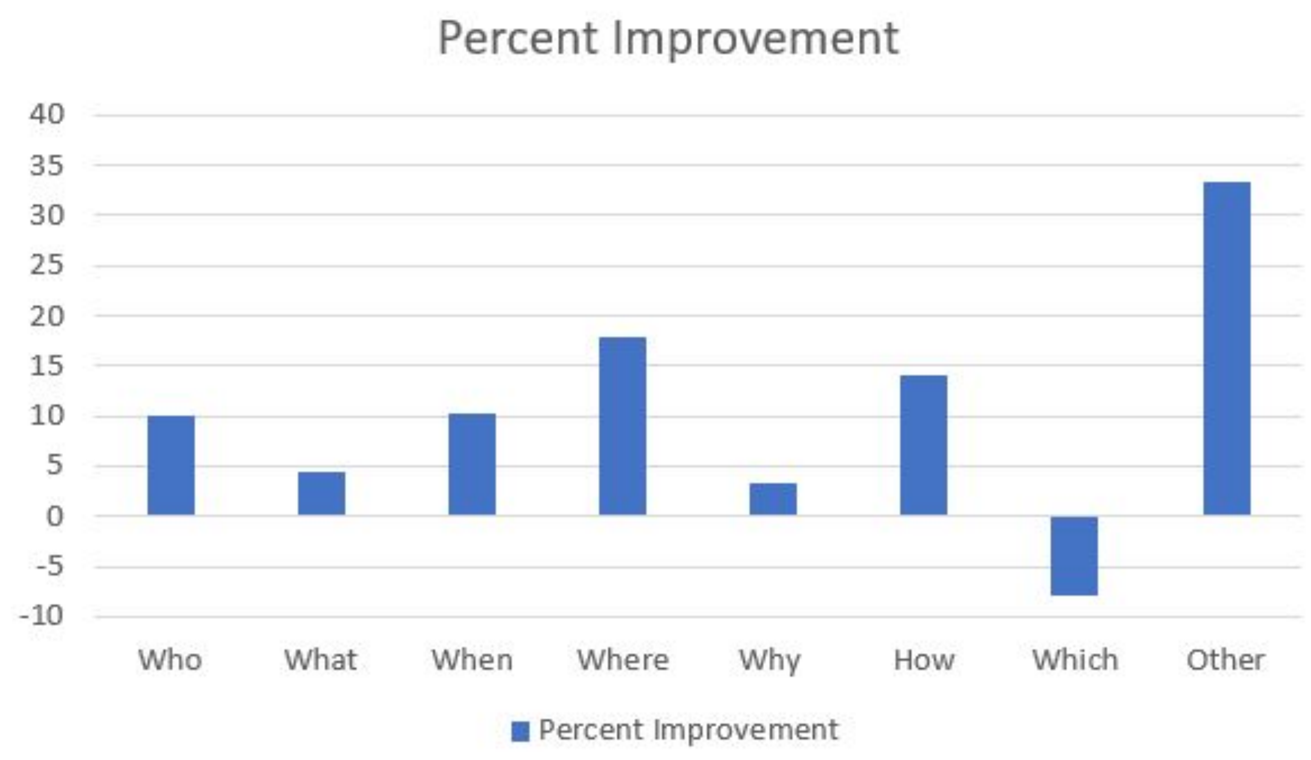

Fig 3: Performance for different question types compared to Bert large. Other refers to questions that do not fall in the other 7 categories, such as 'Is it?'.

Discussion

Context: *...The Normans were famed for their martial spirit and eventually for their Christian piety, becoming exponents of the Catholic orthodoxy into which they assimilated ...*

Question: *Who was famed for their Christian spirit?*

Baseline Bert Large Prediction: Normans
BertQA Prediction: No Answer
Golden: No Answer

Fig 4: Example cropped context and question where the BERT Large model is incorrect while our model is correct due to the coattention's ability to maintain word pair relations.

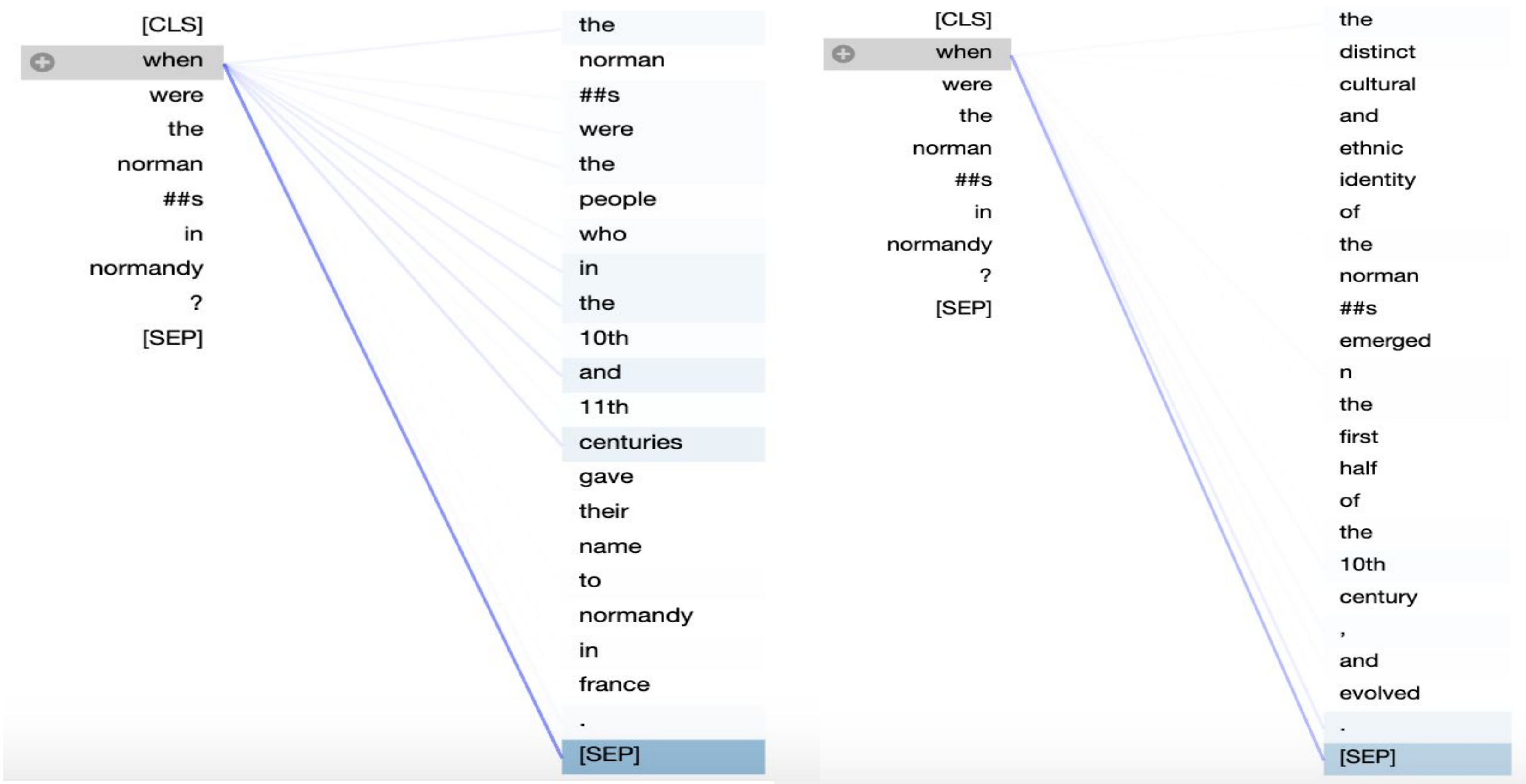

Fig 5: Attention distribution for questions with and without answers. Question with answer has attention over answer span while question without answer has attention only on separator which is outside the span of the context.

Conclusion

We present a novel architectural scheme to help the network learn directed co-attention which has improved performance over the BERT baseline. We present SQuAD 2.Q, an augmented dataset, developed using NMT backtranslation. Our ensemble model gives a ~3.5 point improvement over the Bert Large dev F1. We also learned about how different model components do or don't work together and that some architectural choices like convolutional layers that work so well in computer vision do not necessarily work as well in NLP. As future work, we would like to pre-train our directed co-attention layers and extend SQuAD 2.Q for contexts. We would like to thank the CS224n Team for all the support throughout the course and Cloud access.

References

[1] Rajpurkar et al, Know What You Don't Know: Unanswerable Questions for SQuAD, ACL 2018, https://arxiv.org/abs/1806.03822
[2] Devlin et al, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, https://arxiv.org/abs/1810.04805
[3] SQuAD2.0 Augmented Data, https://github.com/ankit-ai/SQUAD2.Q-Augmented-Dataset