BERT Squared: Read + Verify System for SQuAD

Jiayu Lou

PROBLEM

The task of Question Answering has gained prominence in the past few decades for testing the ability of machines to understand natural language, essentially a machine reading comprehension task focusing on an agent's ability to read a piece of text and subsequently answer questions about it. As one of the foundations for human interactions and communications, this particular task sees increasing attention due to the advances in computational power, brilliant algorithms, and available datasets.

RESULTS

	Model 1 Only			Model 1 + Model 2		
Metrics	**Training**	**Dev**	**Test**	**Training**	**Dev**	**Test**
EM	80.65	72.83	71.50	91.03	75.83	73.71
F1	86.35	75.14	74.28	95.16	78.69	76.41
HasAns EM	72.85	57.93	N/A	87.07	66.52	N/A
HasAns F1	81.40	62.76	N/A	93.26	72.51	N/A
NoAns EM	96.24	86.52	N/A	98.93	84.37	N/A
NoAns F1	96.24	86.52	N/A	98.93	84.37	N/A

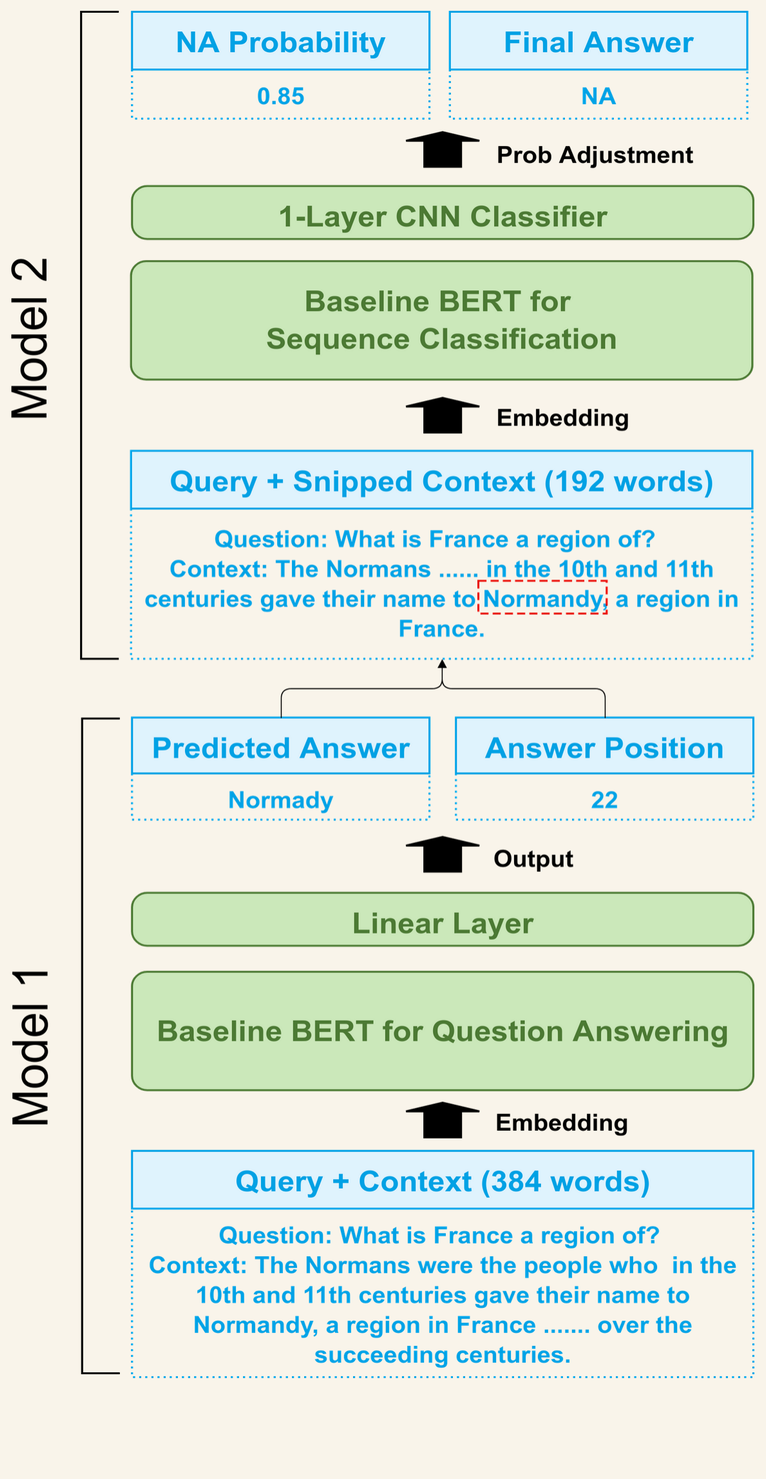
CONCLUSION

When the context is snipped, the attention distribution is "squeezed" towards several key words. The "squeezing effect" helps the attention layers to better concentrate on the relevant keywords, but sometimes could backfire and artificially inflate the confidence level for unanswerable questions and lead to mistakes.

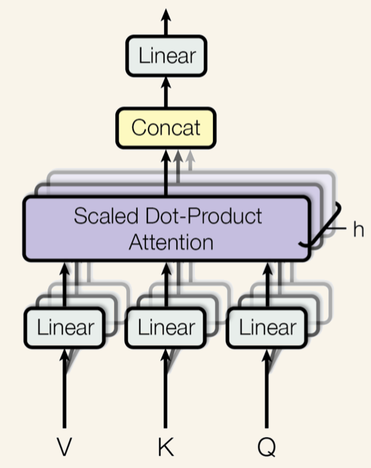
Model 2 only manages to improve Model 1 by a limited margin because it seems that Model 1 already has a good ability to "concentrate". In some effective self-attention layers the top 5 words being attended are already in the scope of snipped context and account for more than 80% of the attention distribution, therefore leaving limited space for improvement for the enforced shortening.

APPROACH

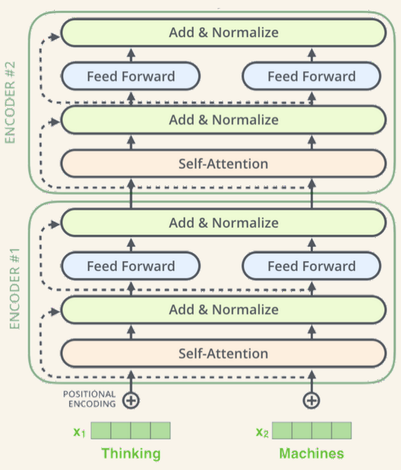
This model utilizes the baseline BERT models for preliminary prediction; after the baseline BERT predicts an answer, the model then snipped the original context to less than 192 words with the predicted answer centered in the middle, and then the snipped context is concatenated with the original query text to be fed into a second BERT which closely examines the trimmed context and predicts a binary result to determine if there is indeed an answer.



Transformers uses the Multi-Head Attention block to compute multiple attention weighted sums instead of a single attention pass.



BERT applies the bidirectional training of Transformer to language modelling and proposes two new pre-training objectives: the "masked language model" (MLM) and "next sentence prediction".



ANALYSIS

Example 1: Model 2 corrects an answerable question

Question: From which countries did the Norse originate?
Model 1 Answer: NA
Model 2 Answer: Denmark, Iceland and Norway
Correct Answer: Denmark, Iceland and Norway
Context: They were descended from Norse raiders and pirates from Denmark, Iceland and Norway who, under their leader Rollo, agreed to swear fealty to King Charles III of West Francia.

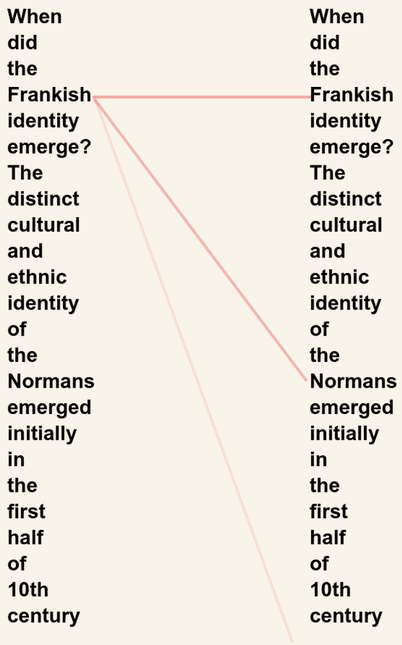
Model 1 Attention Distribution

Model 2 Attention Distribution

Example 2: Model 2 mistakes an answerable question

Question: When did the Frankish identity emerge?
Model 1 Answer: NA
Model 2 Answer: first half of the 10th century
Correct Answer: NA
Context: The distinct cultural and ethnic identity of the Normans emerged initially in the first half of the 10th century, and it continued to evolve over the succeeding centuries.

Model 1 Attention Distribution



Model 2 Attention Distribution