QaN We Pay More Attention?

Akshita Agarwal [1] Qianli Song [1] Shafat Rahman [1]

[1]Department of Computer Science, Stanford University

Introduction

Question answering (QA) is one of the hardest challenges in NLP. Unlike information retrieval tasks like named-entity recognition, QA requires a model to develop deep syntactic and semantic understanding of text as well as efficiently represent relationships between context and query. In this work, we implemented a deep learning architecture based on QANet for question answering on the SQuAD2.0 dataset. Our best model *QANet-Ensemble* achieves **F1 score of 70.47** and **EM score of 66.94**, which is an improvement of **14.9% over the baseline.**

Background

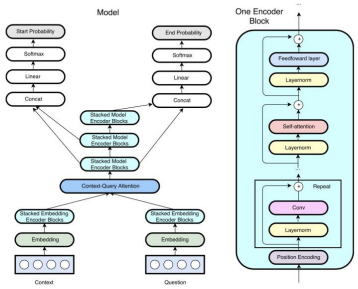

Figure 1. QANet architecture

The QANet model [1] has four major parts:

1. **Input Embedding Layer**: Similar to BiDAF. Transforms the query and context words into distributed pre-trained embeddings from GloVe combined with character-level embeddings.
2. **Encoder Block:** Main contribution of QANet. Consists of a positional encoding layer, followed by several convolutional layers, a self-attention layer with eight heads and a feed-forward layer. Although similar to transformers, it uses Depthwise separable convolutions to save memory.
3. **Context-Query Attention:** QANet borrows the context-query attention mechanism used in BiDAF [2]
4. **Output**: Computes the probability of each position in the context being start or end of answer using output of three encoder blocks.

Methods and Experiments

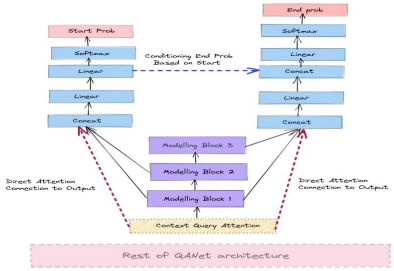

Figure 2. Proposed QANet-XtraAtt architecture

Model	Dev Loss	F1	EM	AvNA
Baseline	02.96	61.33	58.24	67.87
BiDAF-200D-CharEmbed	02.82	66.68	63.28	72.69
QANet-Lite	02.96	68.25	64.75	74.91
QANet-4-AttHeads-7-Encoders	02.70	68.52	65.05	74.79
QANet-8-AttHeads-5-Encoders	02.83	68.38	64.49	74.81
QANet-8-AttHeads-7-Encoders	02.64	67.95	64.78	74.14
QANet-Xtra-Att	02.62	68.62	65.32	74.59
QANet-CharEmbed200D-Xtra-Att	**02.79**	**69.02**	**65.35**	**75.55**
QANet-Ensemble	**02.83**	**70.47**	**66.94**	**76.49**

Table 1. Experiment Results

- **Character-level Embeddings:** We add 200 dimensional character-level embeddings to the baseline BIDAF model to aid in handling out-of-vocabulary words.
- **QANet-Lite:** Our simplest version of QANet with very few convolutional layers and 5 encoder blocks in modelling layer
- **QANet-XtraAtt:** Our proposed approach (Figure 2) modifies the output layer of QANet to add a direct connection from context-query attention layer and also condition end prediction based on start prediction
- **QANet-Ensemble:** We adopt a Majority-Voting technique to ensemble our 6 QANet models. This helps to reduce variance and achieve a significant boost in performance.

Analysis

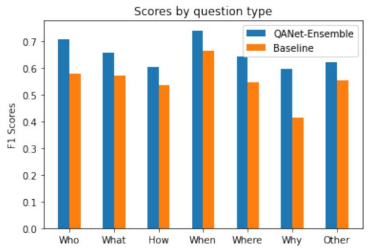

Figure 3. F1 score comparison with Baseline across various question types

Figure 3 demonstrates that among the different question types, our models performs the worst on "Why" questions. "Why" questions require logical reasoning and deeper semantic understanding of the context. This is unlike "Who" questions, which require the model to simply identify entities in the context. "Reasoning-Driven" QA techniques along with QANet might help to improve performance here.

Conclusions

1. Increasing complexity of QANet encoder blocks by increasing the number of attention heads or number of encoder blocks, does not improve performance significantly on the SQuAD2.0 dataset
2. Increasing complexity of other parts of the model, including features like character-embedding dimension and changing the output layer architecture were key factors in improved performance
3. Ensembling helped to produce effective models that attain 1.45 F1 score gain over our single best model as it reduces the variance and feature noise from the base predictors

References

[1] A. W. Y. et al., "Qanet: Combining local convolution with global self-attention for reading comprehension," *CoRR*, 2018.

[2] M. J. S. et al., "Bidirectional attention flow for machine comprehension," *CoRR*, 2016.