Applying QANet to SQuAD 2.0

Aleksander Dash, Andrew Zhang, Nolan Handali

{adash, azhang97, nolanh}@stanford.edu

Introduction

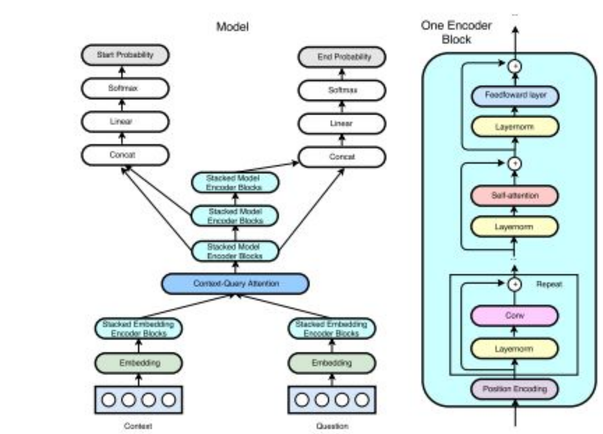
Machine question answering has been a rising area of research in NLP over the past few years, and at the heart of it is the Stanford Question Answering Dataset (SQuAD), version 2.0. In SQuAD, the model is given a question and a context paragraph, and asked to predict the answer, defined as a span from the paragraph. SQuAD 2.0 added unanswerable questions to the dataset.

In this project, we aimed to implement the QANet from Yu et al, one of the top performing models on SQuAD 1.0, and adapt it for for 2.0. QANet focuses on removing the need for sequential processing, instead using a combination of self-attention and convolution to process the input.

Data

We used the official SQuAD dataset, where the training data is taken from the official train set, and our dev and test sets are randomly selected from the official dev set. Additionally, we used pretrained GLoVe word embeddings of dimension 300.

Architecture



Methods

Input Embedding Layer:

1. Concatenate GloVe embeddings for each word with learned character embeddings.
2. Pass through a highway network, the output of which is our embedding

Embedding Encoder Layer(encoder block)

1. Add positional encoding to maintain information lost by convolutions
2. Pass this through convolutional layers with depthwise separable convolutions
3. Pass this through a self-attention layer
4. Pass through a feed-forward layer, which is then output from this block

Context-Query Attention Layer

Computes the attention score between each word in the query and context

1. Compute similarity matrix S and normalize over each row and column to get S' and S'' respectively

2. Let *C* be the matrix of the encoded context, and *Q* the matrix of the encoded query, the context-to-query attention is $A = S' * Q^T$, and the query to context attention is $B = S' * S''^T * C^T$

Model Encoder Layer:

1. This layer consists 7 encoder blocks chained together
2. Pass the input through the chained blocks 3 times to obtain M_0, M_1, M_2

Output Layer:

1. Find probability distribution of start index by passing M_0, M_1 through a softmax function, and similarly with M_0, M_2 for the end index. $p^1 = softmax(W_1[M_0; M_1]),\ p^2 = softmax(W_2[M_0; M_2]),$

To handle "No Answer", we add a special token at the start, and if the model's prediction of start and end of span is this token, output N/A

References

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.

Results

Model	Dev set		Test set	
	EM	F1	EM	F1
Baseline BiDAF	57.05	60.62	56.298	59.920
BiDAF with Char Embeddings	61.39	64.84	–	–
QANet	62.914	66.467	–	–
QANet convolutional attention	62.527	66.227	–	–
QANet multi-head attention (2 heads)	64.24	67.67	–	–
QANet multi-head attention (3 heads)	63.74	67.04	–	–
BiDAF + QANet ensemble	**66.258**	**69.432**	**63.956**	**67.442**

Analysis

Our model struggles with many no answer questions, where the query is almost exactly in the context, but there is no answer. In this example, it is likely the model did not properly associate "Triton" with being male, and so incorrectly associated Triton with the "she" in the context.

- **Question:** At what village did a Triton stop to rest on a sandy beach
- **Context:** The origin of the legendary figure is not fully known. The best-known legend, by Artur Oppman, is that long ago two of Triton's daughters set out on a journey through the depths of the oceans and seas. One of them decided to stay on the coast of Denmark and can be seen sitting at the entrance to the port of Copenhagen. The second mermaid reached the mouth of the Vistula River and plunged into its waters. She stopped to rest on a sandy beach by the village of Warszowa, where fishermen came to admire her beauty and listen to her beautiful voice. A greedy merchant also heard her songs; he followed the fishermen and captured the mermaid.
- **Answer:** N/A
- **Prediction:** Warszowa

Conclusion

Our ensemble of BiDAF with 3 different QANets achieved EM of 63.956 and F1 of 67.442 on the test set, good enough for sixth place on the leaderboard. This illustrates that QANet, originally designed with SQuAD 1.0 in mind, works well on the version 2.0 dataset, with only a few simple changes.

In the future, we would ideally have access to a GPU with more memory, as we were memory constrained while training. Additionally, we would train our best model several times with random starts, and ensemble this to test whether or not it would achieve higher performance than the result of single training sessions.