Investigating QANet's Convolution Layer

Andy Jin, Matthew Early, Jesse Doan

1Department of Computer Science, Stanford University 2Department of Linguistics, Stanford University

Problem & Background

We tackle question-answering by building QANet.
- QANet uses a self-attention layer (for global pairwise interactions) and a convolution layer in its encoder block (for local structure), in lieu of RNN.
- We implemented several QANet variations to see if this reasoning holds.

Dataset

The SQuAD dataset has 129,941 (context, question, answer) triplets for training, 6,075 for dev, 5,915 for test. Below is an example:

Question: Why was Tesla returned to Gospic?
Context paragraph: On 24 March 1879, Tesla was returned to Gospic under police guard for not having a residence permit.
Answer: not having a residence permit

We predicted (start, end) positions for the answer.

Baseline: Bidirectional Attention Flow

BIDAF uses a bidirectional RNN on the embedding output to capture temporal dependencies, and context-to-question and question-to-context attention.

Core QANet Implementation

- Input Embedding: Use 300-dimensional pretrained GloVe word vectors.
- Embedding Encoder Layer: 1 encoder block consisting of 4 convolution layers + 8-headed self-attention layer + feed-forward layer, with residual blocks and layer norms.
- Cross Attention: Compute similarity between each pair of context and query words $A = S \cdot Q^T \in \mathbb{R}^{N \times d}$.
- Model Encoder Layer: Same as embedding encoder but with 2 blocks and 7 conv layers.
- Output Layer: M0, M1, M2 = model encoder outputs $y_0 = \sigma(\text{tanh}(W_0[M0; M1; M2]))$, $y_2 = \sigma(\text{tanh}(W_2[M0; M1]; M2))$.

QANet Extensions

- Character Embeddings: Concatenate with GloVe vectors in the input embedding layer.
- Upscale and Downscale QANet: Experiment with 4, 5, 6, 7, and 9 blocks in the model encoder layer.
- Ensemble: Majority vote from QANet with 4 blocks, 5 blocks (x2), 6 blocks, 7 blocks (x2), large dropout.

Training Curves

Hyperparameters:
- Learning rate: 0.001
- Exp. moving avg. (decay rate: 0.9999)
- L2 weight decay: 3e-7
- Dropout prob: 0.1 (word embed, between layers), 0.05 (char embed)
- Stochastic depth layer dropout in encoder block.

Legend:
- Gray + Orange: 5 encoder blocks
- Blue + Red + Cyan: Baseline QANet
- Cyan: Replace convolution with global attention + feed-forward layer

Test Results (Enssemble / 5 Enc. Blocks)

- EM: 61.10 / 59.76
- F1: 63.82 / 62.83

Future Work

- Experiment with survival rates in stochastic depth dropout.
- Modify number of layers in the embedding encoder.
- Enhance the output layer to condition the end probability on the start.
- Data augmentation via back-translation.

Conclusions

- Decreasing the number of blocks in model encoder can lead to (but does not necessarily cause) increased EM/F1 scores.
- Poor results when replacing conv blocks with attention especially w.t. to self-attention suggests that conv blocks do encode local structure.
- Answers to “when” questions are much more readily captured due to few ways to reference time.

References