Investigating QANet's Convolution Layer

Andy Jin,[1] Matthew Early,[1,2] Jesse Doan[1]

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

CS 224N Winter 2022

andyjin@stanford.edu

mmearly@stanford.edu

jdoan21@stanford.edu

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; 6,078 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

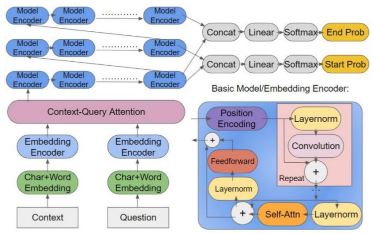
BiDAF uses a bidirectional RNN on the embedding output to capture temporal dependencies, and context-to-question and question-to-context attention.

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 $A = \overline{S} \cdot Q^T \in \mathbf{R}^{n \times d}$
- **Model Encoder Layer:** Same as embedding encoder but with 2 blocks and 7 conv layers
- **Output Layer:** M0, M1, M2 = model encoder outputs

$$p^1 = softmax(W_1[M_0; M_1]), \;\; p^2 = softmax(W_2[M_0; M_2])$$

QANet Architecture

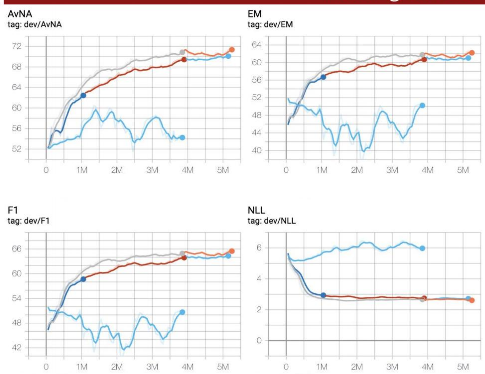


QANet Extensions

- **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
- **Global/local self attention:** Substitute conv with global self-attention and local self-attention
- **Ensemble:** Majority vote from QANet with 4 blocks, 5 blocks (x2), 6 blocks, 7 blocks (x2), large dropout

Model	# Epochs	F1 (dev)	EM (dev)
BiDAF Baseline	30	58	55
BiDAF + Char Embed	30	63.57	60.39
QANet, ensemble	**N/A**	**66.42**	**63.57**
QANet, `n_blocks=5`	40	65.67	62.34
QANet (`n_blocks=7`)	40	64.49	61.27
QANet, `n_blocks=6`	30	63.96	60.85
QANet, `n_blocks=4`	30	63.86	60.59
QANet, `dropout=0.15, n_blocks=5`	30	61.09	58.07
QANet, `survive_prob=1`	16	52.19	52.19
QANet, replace conv w/ self-attention	30	52.19	52.19
QANet, replace conv w/ "global" attention	30	51.89	51.87
QANet, no convolutions	30	50.84	50.55

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:

- **Grey + Orange**: 5 encoder blocks
- **Blue + Red + Cyan**: Baseline QANet
- **Cyan**: Replace convolution with global attention + feedforward layer

Test Results (Ensemble / 5 Enc. Blocks)

- **EM**: 61.10 / 59.763
- **F1**: 63.82 / 62.839

(dev)	**Overall**	**Who**	**What**	**When**	**Where**	**Why**	**How**	**Misc.**
Count	5951	601	2759	440	231	84	525	1311

QANet, ensemble

Pred/Truth	Answer	No Answer
Answer	2122	992
No Answer	726	2111

(dev)	**Overall**	**Who**	**What**	**When**	**Where**	**Why**	**How**	**Misc.**
EM	63.57	64.39	62.78	**70.68**	61.04	58.33	60.19	60.18
F1	66.42	66.71	66.28	**72.12**	67.03	64.11	63.67	64.46
AvNA	71.55	70.72	71.04	**76.82**	73.59	71.43	68.00	70.40

AvNA TPR = 74.51 AvNA TNR = 68.03 AvNA FPR = 31.97 AvNA FNR = 25.49

QANet

Pred/Truth	Answer	No Answer
Answer	2192	1120
No Answer	656	1983

(dev)	**Overall**	**Who**	**What**	**When**	**Where**	**Why**	**How**	**Misc.**
EM	61.27	61.06	59.70	**67.73**	60.17	52.38	60.95	58.50
F1	64.49	64.06	63.72	**69.60**	66.89	57.35	65.77	62.83
AvNA	70.16	68.72	69.52	**74.55**	73.16	70.24	70.48	70.02

AvNA TPR = 76.97 AvNA TNR = 63.91 AvNA FPR = 36.09 AvNA FNR = 23.03

QANet, ~~convs~~ → attn

Pred/Truth	Answer	No Answer
Answer	41	37
No Answer	2807	3066

(dev)	**Overall**	**Who**	**What**	**When**	**Where**	**Why**	**How**	**Misc.**
EM	51.87	52.41	53.75	43.18	49.35	47.62	**54.29**	50.11
F1	51.89	52.41	53.75	43.44	49.35	47.62	**54.29**	50.11
AvNA	52.21	52.41	53.75	48.18	49.35	47.62	**54.29**	50.19

AvNA TPR = 1.44 AvNA TNR = 98.81 AvNA FPR = 1.19 AvNA FNR = 98.56

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.r.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

- 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. In *arXiv preprint arXiv:1804.09541*, 2018.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *arXiv preprint arXiv:1706.03762*, 2017.