QANANET: Improve Question Answering By Learning Not To Answer

Zixiao Ken Wang

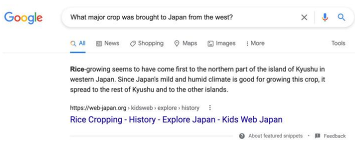
zixiaow@stanford.edu

Stanford Center for Professional Development

Stanford
CS224N Final Project

Problem

- Machine reading comprehension serves information needs at large scale
- Goal: answer question correctly by extracting span of information based on context
- Challenge
  - Text and context understanding
  - Sometimes no answer is the best answer
- Even Google search cannot fully solve this problem



Background

SQuAD (Stanford Question Answering Dataset) 2.0 [3]

- Answerable questions: 100,000
- **Unanswerable** questions: 50,000 (adversarially crowd sourced)

Example:

- Question: What major crop was brought to Japan from the west?
- Context: ...Contacts with the West also brought the introduction to China of a major food crop, sorghum, along with other foreign food products and methods of preparation.
- Baseline Prediction: sorghum
- QANANET Prediction: N/A
- Why is this example hard? The context was about crop in China instead of Japan.

Baseline: Bi-directional Attention Flow (BiDAF) [1]

- RNN + Attention
- Old SOTA for SQuAD 1.x before transformers

Methods

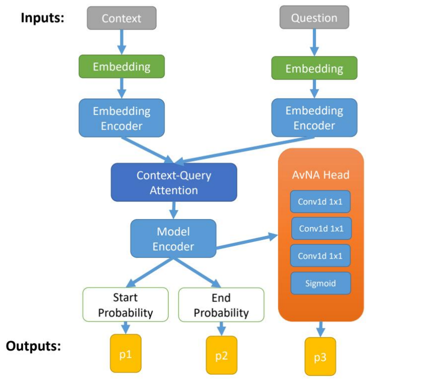
QANANET = QANET + AvNA HEAD

QANET: improve overall baseline performance [2]

- Convolution and Self-attention
- SOTA for SQuAD 1.x before BERT
- No large corpus pretraining required by BERT

AvNA Head: improve no answer predictions

- Shares major architecture as QANET
- New component: binary classification head
- New learning objective: binary cross entropy



QANET Loss Function:

$$\text{Loss} = CrossEntropy(p1, y1) + CrossEntropy(p2, y2)$$

AvNA Loss Function:

$$\text{Loss} = CrossEntropy(p3, y3)$$

y1 Golden start position

y2 Golden end position

y3 1 - No Answer, 0 - Answer

Experiments

Metrics:

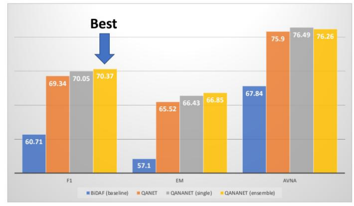
- **F1**: $2 \times \frac{presicion \times recall}{precision + recall}$ (harmonic mean)
- **E**xact **M**atch: answer has exact match with label
- **AvNA**: answer v.s. no answer binary prediction is correct

Our Best Scores

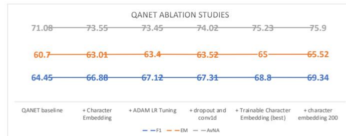
Dev: F1: 70.365 EM: 66.846

Test: F1: 66.581 EM: 62.975

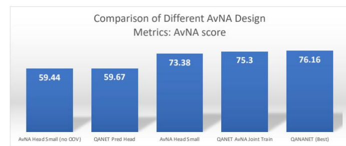
Experiment 1: model with the best dev scores



Experiment 2: QANET ablation study



Experiment 3: best AvNA design



- **This study uses character embedding 64 instead of 200.**
- **AvNA Head Small (no OOV)**: train AvNA head with 1 Conv1D layer without OOV (out of vocabulary) token
- **AvNA Head Small**: AvNA Head with 1 Conv1D layer
- **QANET Pred Head**: use original QANET no answer prediction logic for the AvNA head with AvNA loss function
- **QANET AvNA Joint Train**: train QANET and AvNA head jointly using $Loss = Loss_{QANET} + \lambda Loss_{AVNA}$ where λ=0.2 produces the best results

Analysis

AvNA Head boosts QANET

- 3 Conv1D layer AvNA Head works the best
- AvNA Head finetuning is better than train from scratch
- Single model has highest AvNA **76.49**
- Ensemble is required for best F1 score **70.37**
- AvNA Head requires manual threshold tuning (may lead to dev data overfitting)

How to make a good QANET model?

- (Large) Character Embedding (trainable) is a must
- Optimizer and learning rate are important
- Regularizations (such as layer dropout) helps

Failure Example

Question: How many Frenchmen lost Battle of Carillon?

Context: The third invasion was stopped with the improbable French victory in the Battle of Carillon, in which 3,600 Frenchmen famously and decisively defeated Abercrombie's force of 18,000 regulars…

Gold Answer: N/A

QANANET Prediction: 3,600

Explanation: our model only understands the context of number of Frenchmen in battle but does not infer the notion of lost, or losing people, especially when the "lost" keyword is not in the context.

Possible Solution: pre-training with larger English corpus using larger transformer such as BERT.

Conclusions

- **Our Best Dev Scores: F1: 70.365 EM: 66.846**
- **Our Best Test Scores: F1: 66.581 EM: 62.975**
- **QANET** outperforms baseline **BiDAF**
- **AvNA** further boots **QANET** performance
- **QANANET** cannot fully solve **SQuAD 2.0**
- Intricate context understanding may require large corpus pretraining and larger network like BERT.

Reference

- [1] Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. Bidirectional attention flow for machine comprehension. arXiv preprint arXiv:1611.01603, 2016.
- [2] Adams Wei Yu, David Dohan, Quoc Le, Thang Luong, Rui Zhao, and Kai Chen. Fast and accurate reading comprehension by combining self-attention and convolution. In International Conference on Learning Representations, 2018.
- [3] Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for SQuAD. In Association for Computational Linguistics (ACL), 2018.