QANET: Improve Question Answering By Learning Not To Answer
Zixiao Ken Wang
zixiaow@stanford.edu
Stanford Center for Professional Development

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 (Standard Question Answering Dataset) 2.0 [1]
- 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
- QANET Prediction: No Answer
- 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 1.1 before transformers

Methods

- QANET = RANET + AvNA HEAD
- QANET: improve overall baseline performance [2]
  - Convolution and Self-attention
  - SOTA for SQuAD 1.1 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

Experiment 1: model with the best dev scores

- Metrics:
  - F1: 2 * precision * recall / precision + recall (harmonic mean)
  - Exact Match: answer has exact match with label
- AvNA: answer u.s. no answer binary prediction is correct

- Our Best Scores
  - Dev: F1: 80.85 EM: 92.846
  - Test: F1: 66.581 EM: 62.975

- Experiment 3: best AvNA design

Experiment 2: QANET ablation study

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 must
- Optimizer and learning rate are important
- Regularizations (such as layer dropout) helps

Failure Example

- Question: How many Frenchmen lost Battle of Carthage? (Germ)
- Context: The third invasion was stopped by the improbable French victory in the Battle of Carthage, in which 3,000 Frenchmen famously and decisively defeated Hannibal’s force of 30,000 regulars.
- Gold Answer: 3,000
- QANET Prediction: 3,000

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
- QANET outperforms baseline BiDAF
- AvNA further boosts QANET performance
- QANET cannot fully solve SQuAD 2.0
- Intricate context understanding may require large corpus pretraining and larger network like BERT.

Reference