QANet for SQuAD 2.0

Jenny Yang,[1] Brad Nikkel[2]

[1]*Computer Science, Stanford University, jjyang1@stanford.edu*
[2]*Symbolic Systems, Stanford University, bnikkel@stanford.edu*

Stanford Computer Science

Problem

Objective: Given a query and context, QANet aims to answer the query with a selection from its corresponding context or "N/A" if no answer exists

Motivation: QANet achieved state of the art question-answer performance on SQuAD 1.1 (2018), where all queries had answers

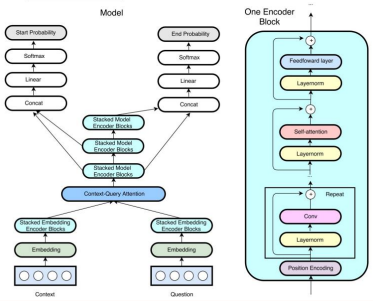
- We implemented QANet, testing its performance on question answering for SQuAD 2.0, which includes "no answers" to some questions

Background

Data: SQuAD 2.0 has ~150k questions and about half the questions have no answer half have 3 human-sourced gold answers, selections from context text.

Methods

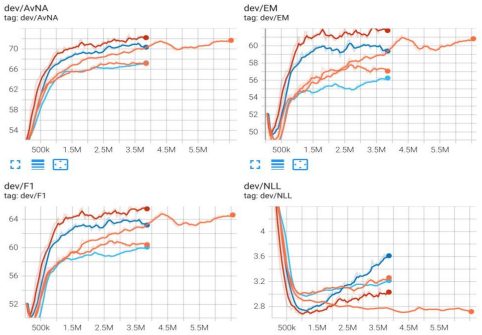
- We implement QANet as close to Yu et al.'s design as possible given memory constraints
- Tested with similar hyperparameters, except for smaller batch sizes & less heads
- **Architecture** is below:



Experiments

Evaluation Metrics:

- **F1 Score**: Harmonic mean of precision & recall. (2 x prediction x recall) / (precision + recall)
- **EM Score:** Binary exact match for ground truth answers (true/false)
- **Top Model: Ensemble** *BiDAF(200-dim)+(64-dim)+ QANet(w/ 4-heads)* EM/F1 = **63.02/66.4**



Model	Dev EM/F1	Test EM/F1
BiDAF: Baseline	57.87/60.98	-----
BiDAF: 64-dim char embeds	60.66/64.24	-----
BiDAF: 200-dim char embeds	**62.46/65.59**	**62.19/65.83**
QANet: 2 heads, 7 blocks	56.26/60.22	-----
QANet: 2 heads, 5 blocks	57.18/61.35	-----
QANet: 4 heads, 7 blocks	**61.2/65.04**	-----
QANet: 8 heads, 7 blocks	52.19/52.31	-----
Ensemble (200-dim, 4&2 head)	**63.02/66.4**	----

Short orange=BiDAF (baseline), blue=BiDAF(64-dim char embeds), red=BiDAF(200-dim char embeds), Long orange=QANet(4-heads), cyan=QANet(2-heads)

Analysis

- **Depth**: More encoding blocks did not significantly increase F1 scores but *did* significantly increase training time. If discovered earlier, we would have decrease depth and increased our batch size, allowing for more experimentation
- **Inconsistent Hyperparameters**: Due to memory limitations, we used different hyperparameters (e.g. quantities of heads & encoder blocks affected permissible batch sizes). Since we could not keep hyperparameters consistent, we cannot draw strong conclusions about their influence
- **Common Errors**: Via manual inspection, we found that our model struggles with answers containing numbers

- **Question:** How many members are on the Miasta City Council?
- **Context:** Legislative power in Warsaw is vested in a unicameral Warsaw City Council (Rada Miasta), which comprises 60 members. Council members are elected directly every four years. Like most legislative bodies, the City Council divides itself into committees which have the oversight of various functions of the city government. Bills passed by a simple majority are sent to the mayor (the President of Warsaw), who may sign them into law. If the mayor vetoes a bill, the Council has 30 days to override the veto by a two-thirds majority vote.
- **Answer:** N/A
- **Prediction:** 60

Conclusions

- **Results non-transferable to different dataset**: Original QANet needs significant modifications to perform as well on SQuAD 2.0 as it did on version 1.1 (in speed & accuracy)
- **More layers/heads are not always better:** Diminishing returns exist for layer depth/head number to performance payoff (less encoder blocks achieved nearly the same F1 scores in faster training & 8 heads was worse than 4 heads)
- **Faster?** Yu et al. found QANet faster at training and inference than RNN models. We did not find QANet faster nor more accurate than BiDAF, but results likely differ with better hardware

Future Work:

- Given we used a subset of SQuAD 2.0, training on a larger subset with 8-headed attention *and* a larger batch size of 32 would give a better comparison with Yu et al.'s results
- Several trainings with different seeds gave us significantly different results. Thus, training at different seeds and ensemble models together might be useful
- Integrate Transformer-XL to get longer range dependencies

References

Yu, Adams Wei, et al. "Qanet: Combining local convolution with global self-attention for reading comprehension." *arXiv preprint arXiv:1804.09541* (2018).