QANet on SQuAD v2.0

Alex Fan[1] and Yixuan (Sherry) Wu[1]
[1]Department of Statistics, Research Mentor: Sarthak Kanodia

CS224n: NLP with Deep Learning
Poster Presentation, Winter 2022

Introduction

Problem:
Given a context paragraph, can the machine extract the right span of words to answer a given question? And when a question is unanswerable, can it produce an empty span?

Dataset:
We will use SQuAD 2.0, which contains 141,934 crowd-sourced questions of Wikipedia articles. 129,941 of those are used as train set, about 6000 as dev set and another 6000 as test set.

Goal:
The goal is to improve the architecture of the Bi-Directional Attention Flow network model detailed in Seo et al. (2017) by:

1. Implementing the character-level embeddings
2. Adapting the QANet model introduced by Yu et al. (2018) from scratch
3. Tuning the QANet model's hyperparameters

Model Architecture

Our best performing model:
Transformer-based model adapted from QANet by Yu et al.
Design Specifics:
Hidden size: 128
# of Conv in the Contextual Embed Encoder Block: 4
# of Conv in a Stacked Encoder Block: 2
# of heads in Self-attention: 8
Used a 1d Convolution to decrease input size when necessary

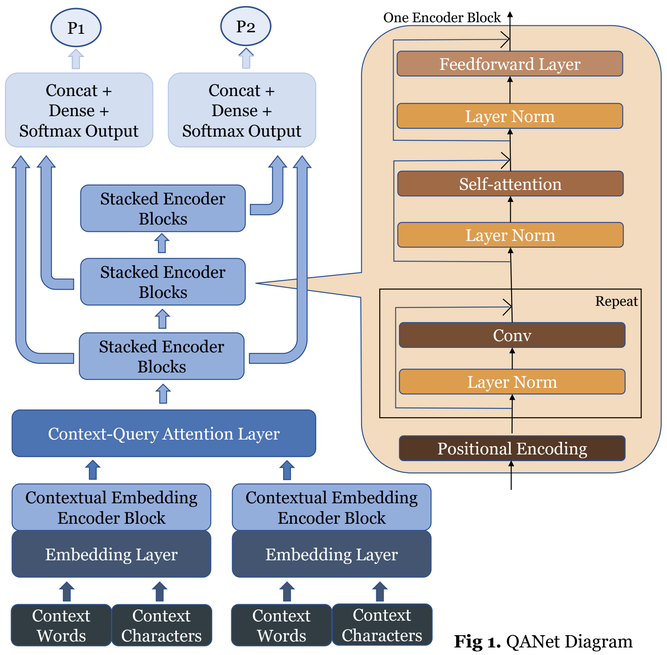

Fig 1. QANet Diagram

Hyperparameter Tuning

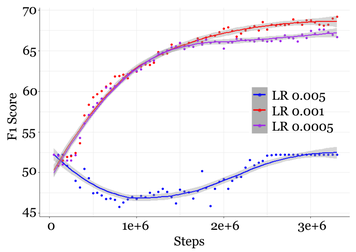

Fig 2. F1 scores with different Learning Rates

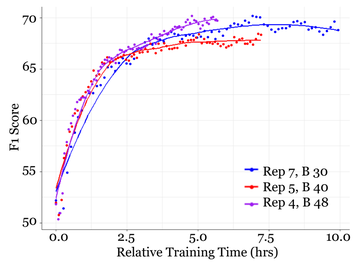

Fig 3. F1 scores vs. Training Time with different # Blocks and Batch sizes

Tuned Parameter	Value	F1	EM
Learning Rate	0.0005	67.29	64.02
	0.001*	69.17	66.82
	0.005	52.19	52.19
Dropout Prob	0.05	69.62	65.89
	0.1*	**70.17**	**66.81**
	0.2	69.17	65.82
Batch size ~ # of Stacked Encoder Blocks (see fig. 1)	#Rep = 7, B = 30*	70.17	66.81
	#Rep = 5, B = 40	68.39	64.86
	#Rep = 4, B = 48	70.17	66.83

Table 1. Parameter Tuning (*Default training param)

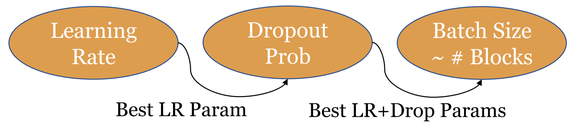

Fig 4. Tuning Sequence (due to time constraint)

Analyses

AvNA
Confusion matrix of model predictions
X: Model predictions; Y: Ground Truths
Correct: 76.5%;

	Has Answer	No Answer
Answered	2361 (39.67%)	912 (15.33%)
Not Answered	487 (8.18%)	2191 (36.82%)

Table 2. AvNA

Performance by Question Type
Table shows predictions on questions that include or start with the following words
(e.g. "Who" category contains questions with "whom", "whose", etc.)
(e.g. Questions that include for example "whatever" are not included in the "What" category)

	What	When	Where	Which	Who	Why	How	Others
F1	69.23	75.22	66.31	74.80	72.08	65.60	69.24	70.89
EM	65.79	75.21	62.10	70.83	69.22	55.81	65.00	68.10
Count	3435	456	248	216	601	86	560	348

Table 3. Performance by Q Type

Error Type Examples

1. Difficult reading comprehension questions (esp. in "Why" category)
   Q: Why was there a depreciation of the industrialized nations dollars?
   C: ... Anticipating that currency values would fluctuate unpredictably for a time, the industrialized nations increased their reserves (by expanding their money supplies) in amounts far greater than before. The result was a depreciation of the dollar and other industrialized nations' currencies...
   A: industrialized nations increased their reserves
   P: N/A
2. Confounding & Proximity of Q&A
   Q: What treaty took the place of constitutional treaty?
   C: Following the Nice Treaty, there was an attempt to reform the constitutional law of the European Union and make it more transparent; ... (40 words)... Instead, the Lisbon Treaty was enacted...
   A: the Lisbon Treaty
   P: Nice Treaty

Results

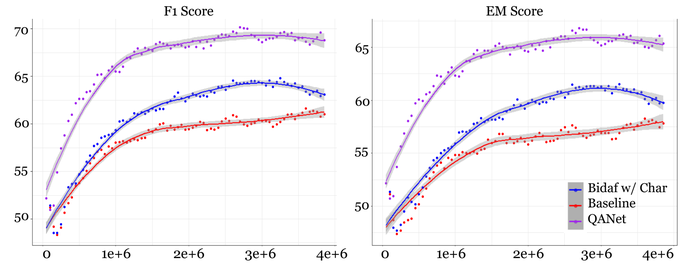

Fig 5. F1 and EM of three different models vs. Steps

The best performing model is the QANet model:
- Achieves an F1 score of 70.17 and EM score of 66.81 on SQuAD v2.0.
- +9/+8 on F1/EM over the baseline
- +6/+3 on F1/EM over the BiDAF w/ Char

Some clear best performing hyperparameters:
- LR = 0.001 does not overshoot minima
- pDrop = 0.1 much more consistent results

Some tradeoff in architecture:
- Reducing the number of stacked encoder blocks from 7->4 and increasing the batch size cut the training time from ~10 hours to ~6 hours.
- Only minor decrease in performance, suggesting that decreasing the number of blocks could be useful if faced with limited compute resources.

Future Work

Due to time and compute limitations, there is still room for improvement. These include:

1. Using a more fine-grained grid search for hyper-parameter tuning
2. Testing model architecture improvements such as including attention-fusion networks (Wang et al. 2018) within the transformer encoder blocks.
3. Using data augmentation strategies such as back-translation to boost model performance
4. Ensembling the QANet models together by averaging the output probabilities

References

[1] Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. Bidirectional attention flow for machine comprehension. In *Association for Computational Linguistics (ACL)*, 2017.
[2] 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 *ACL, 2018*.
[3] Wei Wang, Ming Yan, and Chen Wu. Multi-granularity hierarchical attention fusion networks for reading comprehension and question answering. In *Association for Computational Linguistics (ACL)*, 2018.