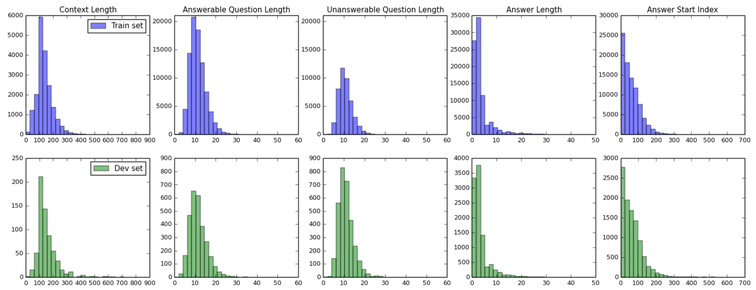
BIDAF Model Optimization and QANet Model Re-implementation on SQuAD2.0

Yulian Zhou
Department of Computer Science
zhouyl@stanford.edu

Dataset – SQuAD2.0

- A question-answering dataset based on Wikipedia articles
- Composed of contexts, related questions, and ground-truth answers
- Questions are derived from context, and answers are span-based
- Over 100,000 questions, half of which are unanswerable



Project Description

We aim to **improve the end-to-end Q&A systems for SQuAD2.0** by

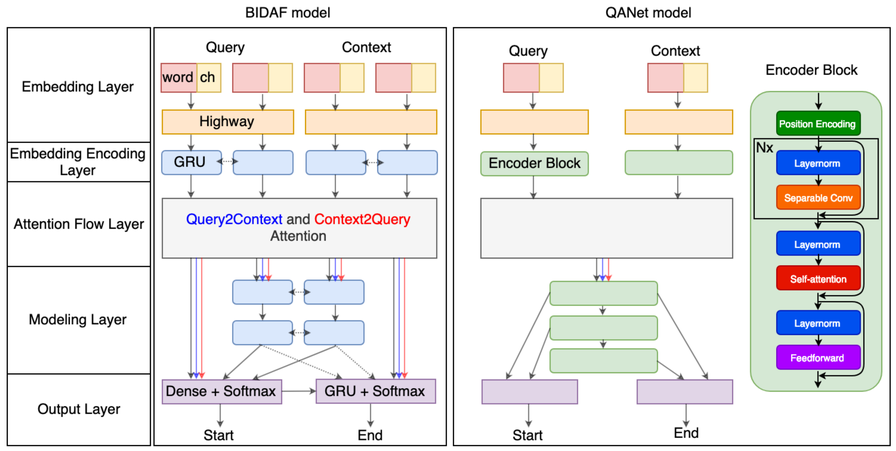
- Add building blocks to the baseline BIDAF model
- Explore attention mechanisms in BIDAF model
- Re-implement QANet

Evaluation metrics

- Exact Match (EM): whether the answer span matches exactly with the ground truth answer
- F1 scores: harmonic mean of precision and recall
- AvNA: Answer vs. No Answer prediction

Model Architecture

BIDAF: A bi-directional attention flow network[1]
QANET: A Q&A model fully based on convolution and self-attention without recurrent[2]



Model Performance

Result: The ensemble model from model 3 and 5 achieved **EM = 66.64** and **F1 score = 69.70** on the Dev. set, and **EM = 65.224** and **F1 score = 68.438** on the Test set (**None-PCE**, one submission made, first place on the Test Leaderboard as of March 18, 2019)

Table: Performance comparison of di□erent BIDAF models

Model #	Based on	Change Made	EM	F1
0	baseline	N.A.	56.49	59.88
1	0	+CharCNN	59.94	63.12
2	1	+Linear ReLU	61.60	65.17
3	2	LSTM–>GRU with increased hidden_size	**63.85**	**67.24**

Table: Performance comparison of QANet models with di□erent settings

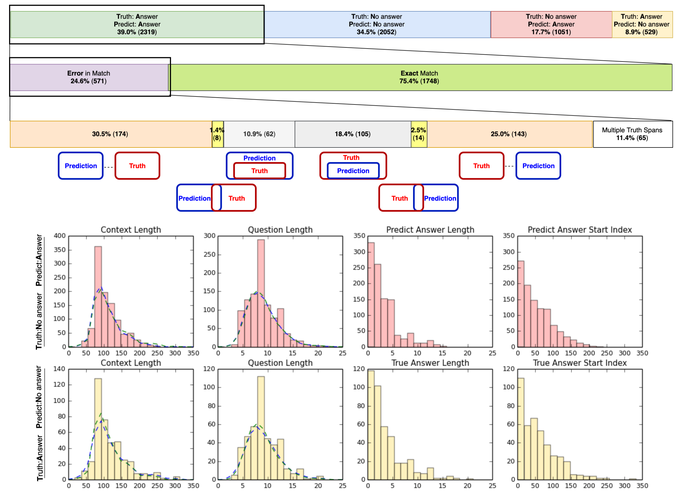
Model #	hidden_size	head_num	batch_size	EM	F1
4	96	3	24	60.78	64.36
5	96	4	20	**61.94**	**65.47**
6	128	8	8	59.62	63.47

Model Comparison

Table: Query Types and Performance Comparison of BIDAF and QANet model

		BIDAF			QANet		
Type	**#**	**EM**	**F1**	**AvNA**	**EM**	**F1**	**AvNA**
Overall	5951	67.24	63.85	73.45	65.49	61.96	72.26
What	3625	64.22	67.17	73.35	61.71	65.37	72.47
Who	688	63.52	66.24	71.37	62.50	64.65	69.62
How	569	58.88	64.90	71.35	58.88	63.61	69.60
When	451	72.06	73.55	78.71	72.06	72.78	76.72
Where	253	59.68	64.71	73.91	56.52	61.14	70.75
Which	214	69.16	72.71	78.97	67.29	71.09	78.97
Why	86	54.65	63.76	74.42	59.30	68.19	76.74
Other	65	44.62	54.38	61.54	33.85	41.94	58.46

Error Analysis



References

[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, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V Le.
Qanet: Combining local convolution with global self-attention for reading comprehension.
arXiv preprint arXiv:1804.09541, 2018.