Attending in BiDAF and QANet

Lucas Orts,[1] Yanal Qushair[1], Sophia Sanchez[1]

[1] Computer Science Department, Stanford University

Stanford | ENGINEERING
Computer Science

Introduction

MOTIVATION

The **Question Answering Task** (QA) is a salient problem in the field of NLP:

- Home assistants like Alexa and Google Home,
- Info retrieval for user-facing interfaces,
- Automated reading comprehension of online texts.

Improving QA results in not only better QA services, but better understanding of natural language semantics.

DATASET

Stanford Question Answering Dataset (**SQuAD**) 2.0.

- A set of (question, context, answer) triples based on Wikipedia text excerpts.
- For each question, the QA model attempts to return an answer to the question that is similar to the human-produced answer based on the context.
- Not all the questions can be answered from the context.

EVALUATION METRICS

The **EM** and **F1 Score** as defined in project handout.

RELATED WORK

[1] Ali Farhadi Minjoon Seo, Aniruddha Kembhavi 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:1804.09541,* 2018

[3] Caiming Xiong, Victor Zhong, and Richard Socher. Dynamic coattention networks for question answering. *arXiv:1611.01604*, 2018.

[4] Microsoft Research Asia Natural Language Computing Group. R-net: Machine reading comprehension with self-matching networks, 2017.

[5] Zihang Dai, Zhilin Yang, Yiming Yang, William W Cohen, Jaime Carbonell, Quoc V Le, and Ruslan Salakhutdinov. Transformer-xl: Language modeling with longer-term dependency. 2018.

Model Implementations

BiDAF MODEL

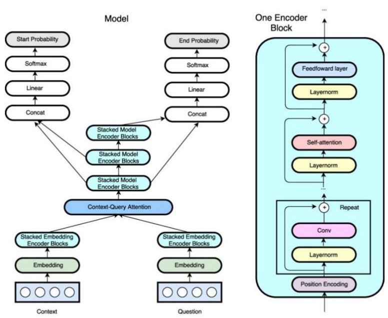
Character Embedding. We implement character-level embeddings to condition on words' internal structure to better handle out-of- vocabulary words. For each word, we concatenate an additional character-level embedding onto the GloVe vectors.

Co-Attention Layer. Based on [3], we implement two-way attention between the context and the question. This involves a second-level attention computation, which attends over representations that are themselves attention outputs.

Self-Attention Layer. Inspired by [4], we implement a self-attention layer, which directly matches the question-aware passage representation against itself using a similarity matrix similar to that of the BiDAF attention layer.

QANet

We implement a **QANet** model from scratch, based in [2]. The architecture has five layers: (1) Input Embedding Layer, (2) Embedding Encoder Layer, (3) Context-Query Attention Layer, (4) Model Encoder Layer, (5) Output Layer.



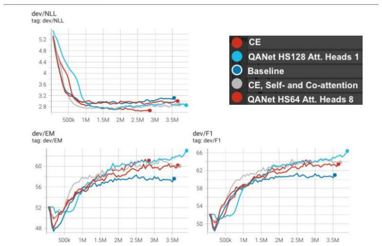
Results and Evaluation

EXPERIMENTS

Baseline. Default BiDAF implementation.

	EM	F1
Baseline	58.01	61.25
Character-Embedding (CE)	60.39	63.71
CE, Self- and Co-attention	61.25	64.65
QANet (hidden 128, att. heads 1)	53.77	57.23
QANet (hidden 64, att. heads 8)	61.27	64.32

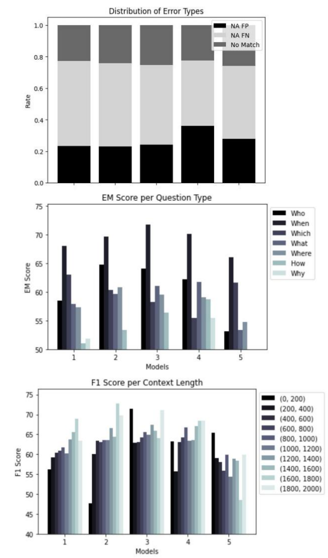
Discussion



DISCUSSION

- Best performing models are BiDAF with CE, Self- and Co-Attention with EM~61, F1~64. More attention helps for BiDAF. Hyperparameter tuning is highly impacting for QANet.
- All models, but especially the BiDAF models, tend to fail most often by giving an answer when there was none. Less obvious for QANet.
- Naturally, more abstract questions like "How" and "Why" are harder for the model. However, the best QANet does significantly better on these than BiDAF.
- Basic BiDAF models (i.e. baseline and with CE) do better the larger the context, but CE with Self- and Co-Attention does much better for smaller context windows. QANet model does not seem as impacted by context length.

Plots



Figures 1-3. From left to right: (1) baseline, (2) CE, (3) CE, Self- & Co-Attention, (4) QANet (hidden size=64, # heads=8), (5) QANet (hidden size=128, # heads=1)

CHALLENGES AND FUTURE WORK

- Limited hyperparameter tuning due to time constraints. Limited model size due to hardware constraints.
- Could ensemble BiDAF and QANet model to leverage their respective strengths.
- Could improve both models with e.g. data augmentation techniques.
- Use Transformer-XL [5] to enable learning dependencies beyond fixed length context.
- More in-depth analysis of different attention mechanisms using heat maps.