All for One or One for All: Ensemble of Diverse Augmentation for Self-Attention

Jasper McAvity
jmcavity@stanford.edu

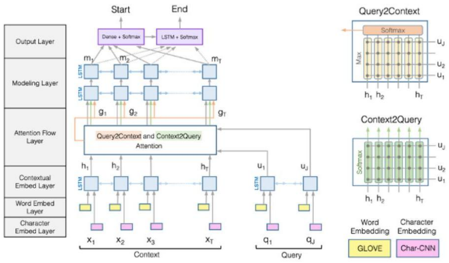
Tiffany Zhao
tiffzhao@stanford.edu

Amir Zur
amirzur@stanford.edu

Introduction

Problem: Accurate question and answering systems are crucial to Web search engines to serve information needs

Objective: Produce a model which outperforms the baseline Bidirectional Attention Flow (BiDAF) on SQuAD 2.0 introduced in (Seo et al., 2018) [1]



Data & Approach

- Official SQuAD 2.0 dataset + new SQuAD 2.0 examples produced by the teaching team:
  - ***Train*** (~130,000): official SQuAD 2.0 training set.
  - ***Dev*** (~6000): roughly half of the official dev set, randomly selected
  - ***Test*** (~6000): remaining examples from the official dev set, plus hand-labeled

Two neural network structures:

- Self-attention [2], coattention [6], and the R-NET model [3]
- QANet model [4]

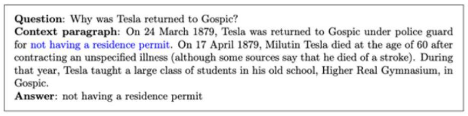
Feed QANet with:

- Data augmentation through backtranslation with Neural Machine Translation (NMT) models [2]

Finally pooling models together with:

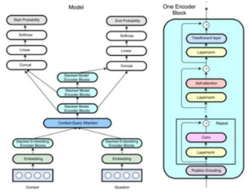
- Ensembling techniques [5]

- Example from dataset *(context, question, answer)*:



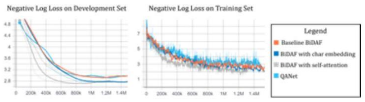
Question: Why was Tesla returned to Gospic?
Context paragraph: On 24 March 1879, Tesla was returned to Gospic under police guard for not having a residence permit. On 17 April 1879, Milutin Tesla died at the age of 60 after contracting an unspecified illness (although some sources say that he died of a stroke). During that year, Tesla taught a large class of students in his old school, Higher Real Gymnasium, in Gospic.
Answer: not having a residence permit

QANet Architecture

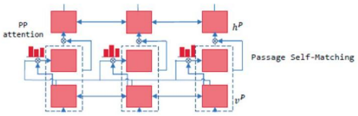


Results:

- Poorer overall performance than baseline model
- Able to learn quickly, but converges early
- Volatile loss during training



Self-Attention and R-NET



Results:

- Coattention didn’t improve upon the Baseline
- R-NET didn't improve upon the baseline
- Self-attention improved upon the baseline

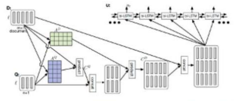
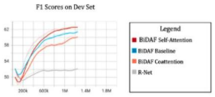
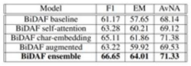


Figure 2: Coattention encoder. The affinity matrix L is not shown here. We instead directly show the normalized attention weights A^D and A^Q.



Results

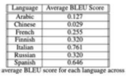
- Our best result is an ensemble of the self-attention model, the BiDAF model with character embeddings, and the BiDAF model trained on the full augmented dataset
- This model achieves higher scores on the development set than each individual model on its own
- The ensemble model achieves an **EM** score of **62.93%** and an **F1** score of **65.78%**



Model	F1	EM	AvNA
BiDAF baseline	61.17	57.65	68.14
BiDAF self-attention	63.28	60.21	69.12
BiDAF char-embedding	65.11	61.86	71.38
BiDAF augmented	63.22	59.92	69.53
BiDAF ensemble	**66.65**	**64.01**	**71.33**

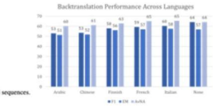
Data Augmentation

- MarianMT Tokenizer and Neural Machine Translation model to perform data augmentation
- Backtranslation of questions from training dataset for Arabic, Chinese, French, Finnish, Italian, Russian, and Spanish
- Italian and Spanish show the strongest potential for improving model performance



Language	Average BLEU Score
Arabic	0.127
Chinese	0.029
French	0.255
Finnish	0.320
Italian	0.761
Russian	0.320
Spanish	0.646

Table 2: Comparison of the average BLEU score for each language across all generated sequences.



Ensemble Methods

- **Main question:** Would a large model trained on all data perform better than an ensemble of smaller models each trained on one augmented language?
- **Large model:** QANet with 4 model encoders, and all convolutional neural layers, trained on all languages
- **Smaller model:** QANet with 3 of the 4 convolutional neural layers, and 4 model encoders, trained on a single language

Model	F1	EM	AvNA
Full augmented model	55.10	52.85	61.33
Ensemble model	55.76	55.15	57.2

Conclusions

- Best performing model is an ensemble model of fully augmented model, self-attention, and the BiDAF model
- Languages that are **more** similar to English give **better** back-translated augmented data for question-answering
- An **ensemble of small augmented models** is better than **one model trained on all data**

Future work:

- Improve QANet, find reason for poorer performance than original paper
- Implement full architecture of R-NET model, on top of self-attention layer
- Try different language families, such as Iranian or Indic languages

References:
[1] Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. Bidirectional attention flow for machine comprehension, 2018.
[2] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017. [3] Natural Language Computing Group. R-net: Machine reading comprehension with self-matching networks. May 2017.
[4] 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, 2018.
[5] Thomas G. Dietterich. Ensemble methods in machine learning, 2000.
[6] Dynamic Coattention Networks for Question Answering, 2017.

Graphic in Introduction section is from [1] and graphic in QANet Architecture section is from [4].