ORCHESTRA: An Ensemble Approach to SQuAD 2.0

Aaron Levett, Taide Ding

Department of Computer Science, Stanford University

Introduction

- Question answering and reading comprehension are crucial areas of research in Natural Language Processing, with many practical applications including digital assistants (Siri, Alexa) and web search.
- The Stanford Question Answering Dataset (SQuAD) 2.0 task tests both a system's ability to answer reading comprehension questions and to determine when a question cannot be answered given the provided passage.
- A Bidirectional Attention Flow (BiDAF) baseline model (**Figure 1**) without the original model's character embedding layer was used as our baseline.
- We made several modifications to the baseline, including augmentations to the word embedding layer, addition of a character embedding layer, and replacing the LSTMs with GRUs. We ensembled several of our models together to form our '**ORCHESTRA**' model, which achieved a max F1 score of 67.00 and EM score of 63.86 on the test set.

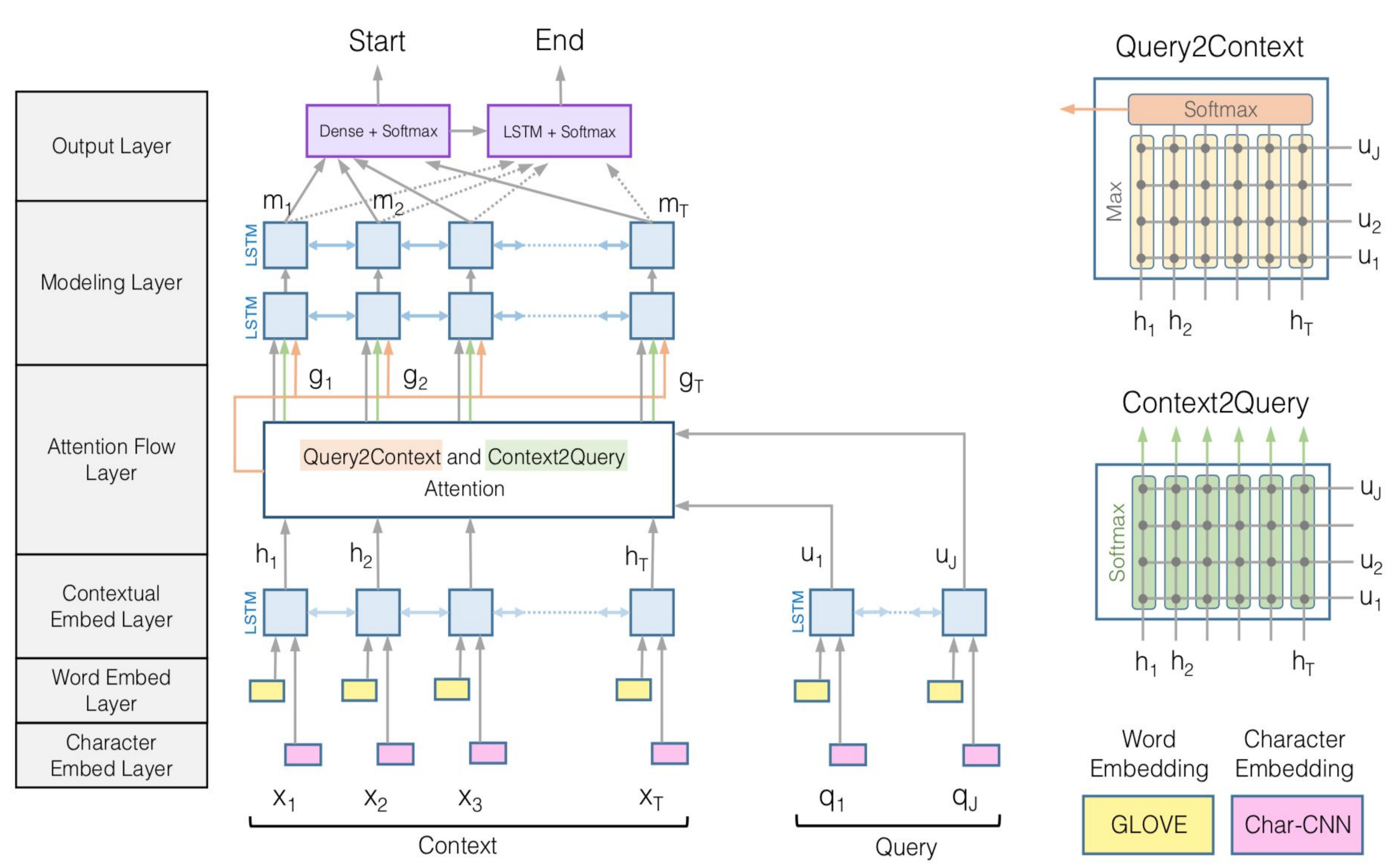

Figure 1: Bidirectional Attention Flow (BiDAF) Architecture

SQuAD 2.0 Sample Task

- **Context:** *However, some computational problems are easier to analyze in terms of more unusual resources. For example, a non-deterministic Turing machine is a computational model that is allowed to branch out to check many different possibilities at once….*
- **Question**: *What type of Turing machine can be characterized by checking multiple possibilities at the same time?*
- **Answer:** *Non-Deterministic*

Dataset

- **Training**: 129,941 labeled training examples (question/context/answer) from the official SQuAD 2.0 dataset.
- **Development**: 5,951 Examples (about half) of the official SQuAD 2.0 development dataset
- **Testing**: The remaining official SQuAD 2.0 dataset examples, with additional examples from the course teaching staff (5,915 examples)

Approach

1. Augmenting the representations for the input layer by adding a character embedding layer (**CharEmb**) and by concatenating the input GLoVe word embeddings with a token's word2vec embeddings (**w2v+GLoVe**)
2. Using GRUs in place of LSTMs in the BiDAF Model. Three such models were produced (**1GRU, 2GRU, 3GRU**), with LSTMs and GRUs mixed together (see **Table 1** for which layers used LSTM vs. GRU)
3. Ensembling the models from 1. and 2. together (see **Table 3** for permutations of models used for ensembling).

Model	Encoding	Modeling	Output
CharEmb (LSTM)	LSTM	LSTM	LSTM
CharEmb (1GRU)	LSTM	GRU	LSTM
CharEmb (2GRU)	GRU	GRU	LSTM
CharEmb (3GRU)	GRU	GRU	GRU

Table 1: Each of the Models used in our GRU vs. LSTM Experiments, specifying the architecture in each layer

Experiments and Results

- F1 and EM metrics were used for evaluation. For examples that lack an answer, F1 and EM are defined as 1 if the model correctly predicts no answer, and 0 if the model predicts there to be an answer.
- We trained all models for 30 epochs with a fixed learning rate of 0.50. Batch gradient descent with batch size of 64 was employed, and dropout probability was set at 0.2 for all experiments.

Model	F1	EM
Baseline	61.28	57.87
w2v+GLoVe	62.92	59.75
CharEmb (LSTM)	64.38	60.98
w2v+GLoVe+CharEmb (LSTM)	64.06	61.03
CharEmb (1GRU)	63.52	60.16
CharEmb (2GRU)	62.87	59.62
CharEmb (3GRU)	64.84	61.20

Table 2: F1 and EM scores for model epochs with best F1

Constituent Model	E1	E2	E3	E4	E5	E6
w2v+GLoVe	+	+		+	+	+
CharEmb (LSTM)	+		+	+	+	+
CharEmb (1GRU)						+
CharEmb (2GRU)		+	+	+	+	+
CharEmb (3GRU)						+
w2v+GLoVe+CharEmb (LSTM)					+	+
Scores	**E1**	**E2**	**E3**	**E4**	**E5**	**E6**
F1 (Dev)	65.52	65.61	66.54	66.57	67.27	67.06
F1 (Test)			64.80	65.91	67.00	
EM (Dev)	62.43	62.64	63.32	63.54	64.24	64.17
EM (Test)			61.52	62.72	63.86	

Table 3: F1 and EM Scores for Ensemble Models. E5 was submitted to the Test Non-PCE leaderboard as ORCHESTRA

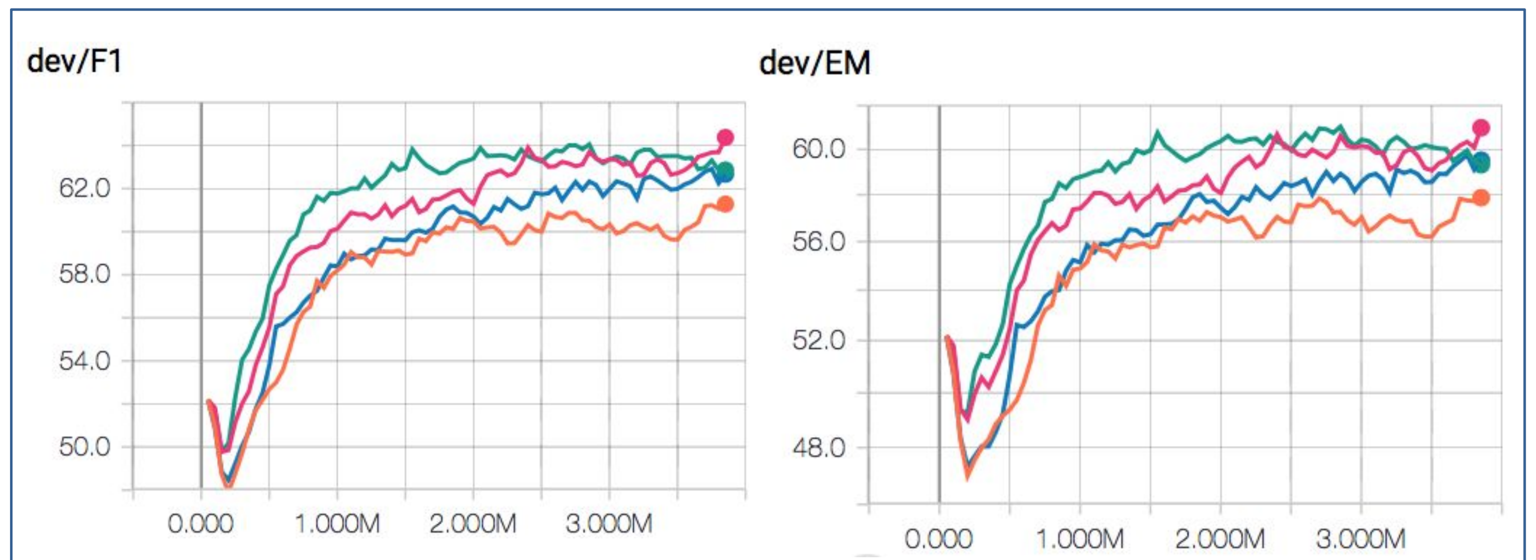

Figure 2: F1 and EM scores during training of character and word embedding experiments: **Orange** = Baseline; **Blue** = w2v+GLoVe; **Pink** = CharEmb (LSTM), **Green** = w2v+GLoVe+CharEmb (LSTM)

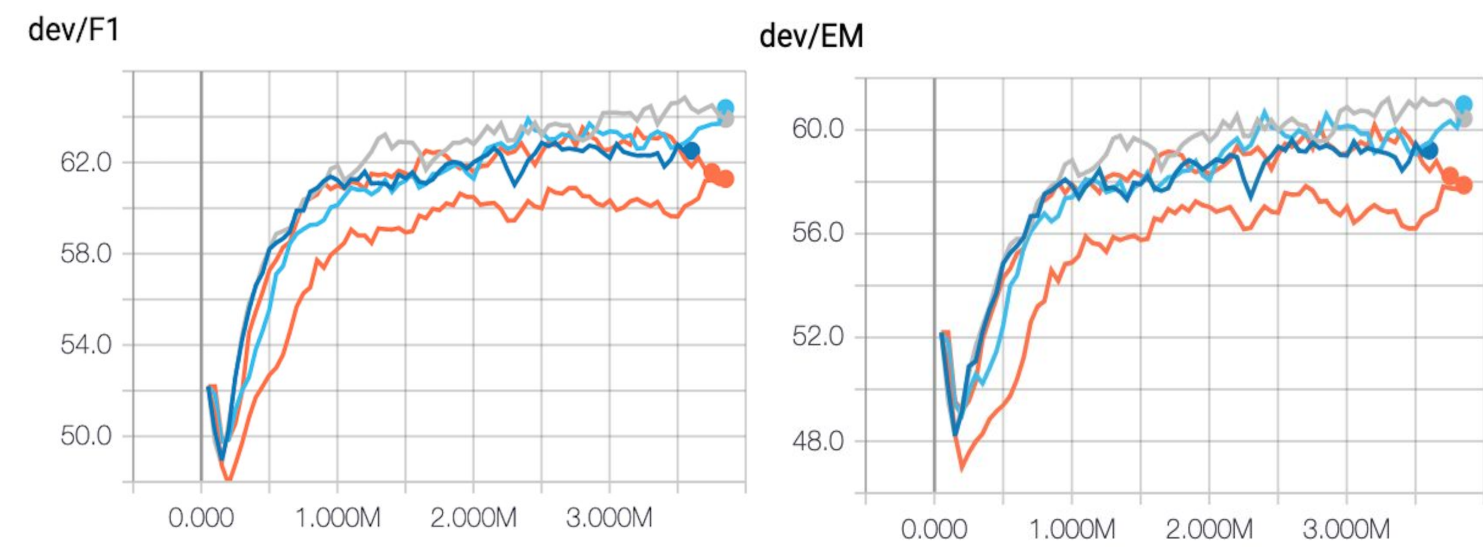

Figure 3: F1 and EM scores for GRU implementations: **Lower Orange** = Baseline; **Light Blue** = CharEmb (LSTM); **Upper Orange** = CharEmb (1GRU); **Gray** = CharEmb (2GRU); **Dark Blue** = CharEmb (3GRU)

Analysis and Discussion

- Common Types of Errors in ORCHESTRA based off of analysis of a subset of ORCHESTRA's outputs for the CS224N SQuAD 2.0 development set:
  - Answering Unanswerable 'When' Questions
  - Answer Content Correct, span does not precisely match provided answer
  - Inclusion of Adjectives Irrelevant to the Question
- Broadly, F1 and EM scores increased with the number of ensembled models

Conclusions and Next Steps

- Combining multiple word embedding models, adding a character embedding layer, and replacing LSTMs with GRUs improved the baseline's performance.
- Ensembling many different variants of the BiDAF model led to significant improvements in performance relative to any single model.
- As our best ensemble model, ORCHESTRA performed similarly on the test and dev sets (less than 0.5 difference in F1 and EM scores), indicating that significant overfitting to the dev set did not occur.
- Further work may involve tuning the hyperparameters of the models we used in our ensemble. Our experiments with decaying learning rates were inconclusive: using a cyclic learning rate that decays over time but rises once performance starts to plateau may be beneficial.

References

- [1] Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. *Bidirectional attention flow for machine comprehension.* arXiv preprint arXiv:1611.01603, 2016.
- [2] Gao Huang, Yixuan Li, and Geoff Pleiss. *Snapshot Ensembles: Train 1, Get M for Free.* International Conference on Learning Representations (ICLR), 2017.