An Exploration of Embeddings with BiDAF

Sterling Alic, Miles Zoltak
Department of Computer Science, Stanford University

Overview

- Question Answering is a well-studied topic and the Stanford Question Answering Dataset (SQuAD 2.0) is a popular dataset to evaluate model performance.
- We implemented a system to approach the task of Question Answering by extending the BiDirectional Attention-Flow (BiDAF) model with character-level embeddings.
- Our best model shows improvement over the baseline and promise for altering our existing embeddings and adding other types of embeddings.

Dataset

- We used the SQuAD 2.0 dataset
  - 150k question/context/answer triples.
  - 100k+ answerable questions and 50k unanswerable questions.
  - 500+ Wikipedia articles used to generate examples

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

Figure 1: Examples in the SQuAD 2.0 dataset are formatted as (Question, Context Paragraph, Answer) triples

Fine-tuning GloVe embeddings

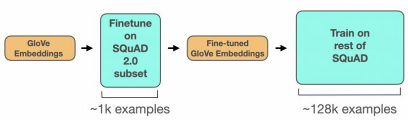

Figure 2: Our proposed method for fine-tuning GloVe embeddings. We fine-tuned the embeddings, but did not finish training.

Model Architecture: BiDAF with Character-Level Embeddings

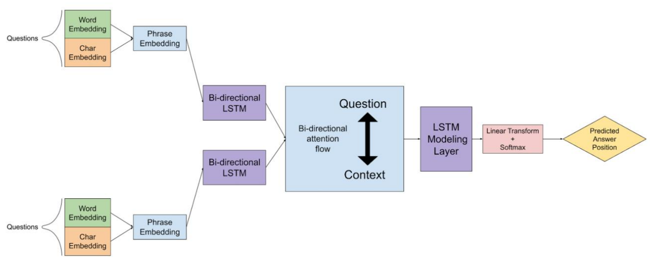

Figure 3: Our model extends the baseline BiDAF model by concatenating character-level embeddings (using a concurrent neural network with max-pooling) with the original word embeddings.

Results

Model	Dropout	Hidden Size	Batch Size	Number Epochs	EM	F1
Baseline LSTM	0.2	100	64	30	57.75	61.23
Baseline GRU + Char Embed	0.2	100	64	30	59.87	63.17
LSTM + Char Embed	0.2	100	96	30	59.30	62.81
LSTM + Char Embed	0.25	100	128	30	57.22	60.74
LSTM + Char Embed	0.15	150	64	20	60.11	63.35
LSTM + Char Embed	0.15	100	64	30	59.44	63.04
LSTM + Char Embed	0.2	100	64	30	**61.03**	**64.07**

Figure 4: Our best model is the BiDAF model including character embeddings in the input and using LSTMs for the RNN layers.

Analysis by Question-Word

Question Word	EM	F1
Who	60.428	62.261
What	60.783	63.944
Where	61.029	64.477
Which	66.667	69.957
When	68.592	69.780
Why	49.425	54.749
How	56.616	61.056
Other	38.461	42.749

Figure 5: Our best model's performance broken down by question word.

Conclusion + Future Work

- Character-level embeddings improves the model performance. Our results are still inconclusive though for GloVe embeddings fine-tuned on SQuAD.
- One limitation in our model is that our hyperparameter search did not yield an improvement in model performance, so there is potential for improvement with further search.
- Implement early stopping, such that runs in hyperparameter tuning that don't improve our evaluation metrics for a certain number of epochs terminate early.
- Given more time, we would train our model with the fine-tuned GloVe word embeddings.

References

[1] Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. Bidirectional attention flow for machine comprehension, 2018.

[2] Nicholas Dingwall and Christopher Potts. Mittens: an extension of GloVe for learning domain- specialized representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 212–217, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[3] Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for SQuAD. In *Association for Computational Linguistics (ACL)*, 2018.

[4] Matthew D. Zeiler. ADADELTA: an adaptive learning rate method. *CoRR*, abs/1212.5701, 2012.