BiDAF-inspired Preferential Multi-perspective Matching for Question Answering Task

Kexin Yu[1], Yuxing Chen[2]

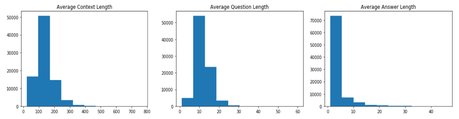
Institute for Computational and Mathematical Engineering, Stanford University[1];Symbolic Systems Program, Stanford University[2]

ABSTRACT

We combine the ideas of two high-performing SQuAD models, Bidirectional Attention Flow (BiDAF) and Bilateral Multi-Perspective Matching (BiMPM), and propose a new framework, BiDAF-inspired Preferential Multi-perspective Matching (BPMPM) for the Question Answering task. In particular, we consider multiple strategies when matching context and question embeddings in both Context-To-Question and Question-To-Context directions, apply a preferential rule when aggregating their results and aim for a more comprehensive view of the interaction between two sequences. We further refine the model using character-level embeddings in the Encoder Layer to better handle out-of-vocabulary words, Dynamic Pointing Decoder to recover from the local maximum corresponding to the initial incorrect guess in the Output layer, and smarter span selection at test time.

DATASET

The dataset is obtained from Stanford Question Answering Dataset (SQuAD). SQuAD consists of questions posed by crowdworkers on a set of Wikipedia articles, and the answer to each question is directly taken from the corresponding passage. The majority of context lengths in the dataset are below 400 words, and there are few questions with length over 25. We choose hyperparameters accordingly in our experiments.



EMBEDDINGS

We use word-level as well as character-level encoding to obtain the embedding for each word. To get the word-level embeddings, we use the pre-trained word vectors, GloVe. We've also experimented with trainable word embeddings based on GloVe. Since word-level embeddings suffer out-of-vocabulary (OOV) problem when an unseen word appears during the test time, we extract the character-level features from each word to form a distributed representation by adding character-level embedding layers. The concatenation of word-level and character-level embeddings is then passed into a three-layer highway network.

MODEL

Architecture Overview

- **Character Level Embedding Layer** generates character-level embedding by feeding each word into a Convolutional Neural Network (CNN).
- **LSTM Encoder Layer** maps each word into its pre-trained vector representation (GloVe) and concatenates it with its corresponding character-level embedding. The hybrid representation is then fed into a 1-layer bidirectional LSTM.
- **Matching Layer** (core) matches the context and question using four different strategies and aggregates their matching scores to produce a blended query-aware context representation.
- **Modeling Layer** employs another Bidirectional LSTM to further encode the query-aware context representation.
- **Output Layer** uses Dynamic Pointing Decoder and Smarter Span Selection to predict the answer.

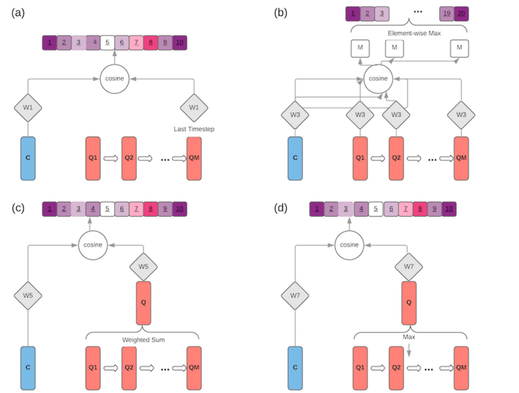
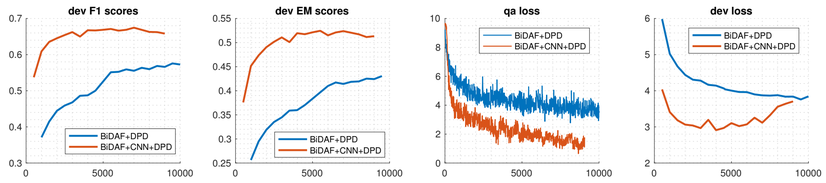


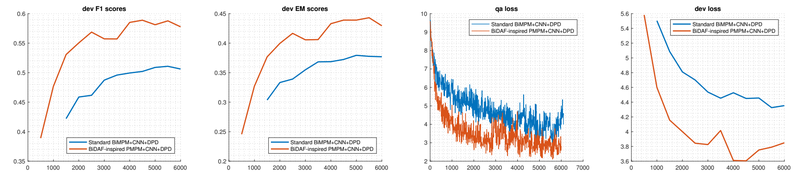
Diagram for strategies used in Preferential Multi-perspective Matching: (a) full matching (b) max-pooling matching (with the largest number of perspectives) (c) attentive matching (d) max-attentive matching.

EXPERIMENTS & RESULTS

Performance boost from character-level embeddings



Performance boost from BiDAF-inspired Preferential variant of BiMPM

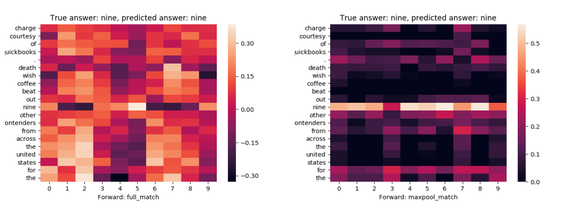


Model	Best F1	Best EM	Time of Convergence	Seconds per Batch
BiDAF + DPD	0.57567	0.43047	9000 iterations	1.5s
BiDAF + CNN + DPD	**0.73753**	**0.63217**	4000 iterations	1.7s
BiMPM + CNN +DPD	0.51079	0.37937	5000 iterations	2.2s
BPMPM + CNN + DPD	0.58778	0.44308	4000 iterations	4.7s
Ensemble	**0.75071**	**0.64238**	-	-

EVALUATION

To understand how different matching strategies work, we first implement the standard BiMPM model and visualize their matching scores in heat map. The figure shows an example where the matching score has the largest absolute value at true/predicted answer start/end with the Full Matching strategy. Moreover, each perspective provides different advice.

To further investigate which strategy makes the best decision, we compare the scores distribution of all strategies and find out that the Max-pooling Matching strategy is the most confident and reliable in making predictions. Thus, we assign higher weight to this strategy in our preferential variant of BiMPM.



CONTEXT: in early 2012, nfl commissioner roger goodell stated that the league planned to make the 50th super bowl "spectacular" and that it would be "an important game for us as a league"..
QUESTION: what *one word* did the nfl commissioner use to describe what super bowl 50 was intended to be?
TRUE ANSWER: spectacular
PREDICTED ANSWER from BiDAF: an important game for us as a league
PREDICTED ANSWER from BiMPM: spectacular

REFERENCE

[1] Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level Convolutional Networks for Text Classification. *arXiv preprint arXiv:1509.01626, 2015.*

[2] Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. Bidirectional attention flow for machine comprehension. *arXiv preprint arXiv:1611.01603, 2016.*

[3] Zhiguo Wang, Wael Hamza, and Radu Florian. Bilateral multi-perspective matching for natural language sentences. *arXiv preprint arXiv:1702.03814, 2017.*