Reading Comprehension on SQuAD: An Insight into BiDAF

Shreyash Pandey, Vivekkumar Patel

Stanford University

Introduction

- Solve Reading Comprehension using a Deep Learning approach.
- Implement BiDAF, one of the state-of-the-art approaches.
- Suggest improvements to BiDAF to improve F1 and EM scores.
- Analyze significance of each component of the model to understand it's strengths and weaknesses.

Approach

Encoder-Interaction-Output Framework

- Encoder Layer: Represent each word with word-level embeddings and (optionally) character-level embeddings.
- Interaction Layer: Capture the interaction between query and context by using attention mechanism. Generate a blended representation.
- Output Layer: Calculate probability distributions for the answer span, then choose the start and end locations of the answer.

Encoder Experiments

- Baseline: Glove embeddings for word vectors.
- CNN: CNN based character-level embeddings to augment the word embeddings.

Attention Experiments

- Baseline: Dot-product attention with the context hidden states attending to the question hidden states.
- BiDAF: Attention flows both ways - from context to question and from question to context.

Output Layer Experiments

- Full BiDAF: Two LSTM layers on top of the blended representation.
- Modified BiDAF: Augment Full BiDAF with two more LSTM layers.
- DP: Use dynamic programming to find start and end points of the answer in linear time.

Implementation

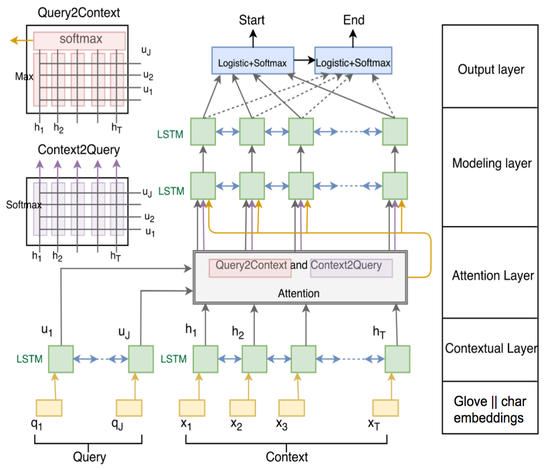

Figure 1: Model Architecture, Image borrowed from Xia et al.'s report

- Context Len: 600, Question Len: 30
- LSTM hidden size: 200
- Char embedding size: 20

Performance based on Question-type

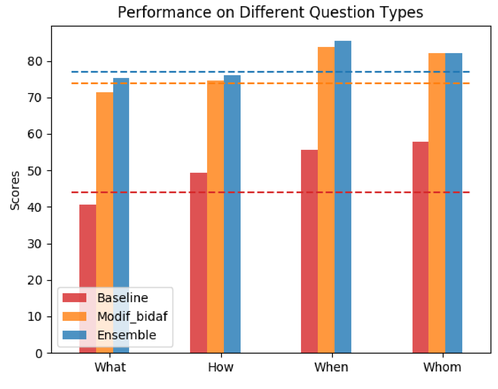

Figure 3: Highest number of questions have "What". Least number of questions have "Whom"

- LSTMs in modeling layers provide a large improvement, with perfect performance on "Whom" questions (F1 = EM).

BiDAF & Similarity Matrix

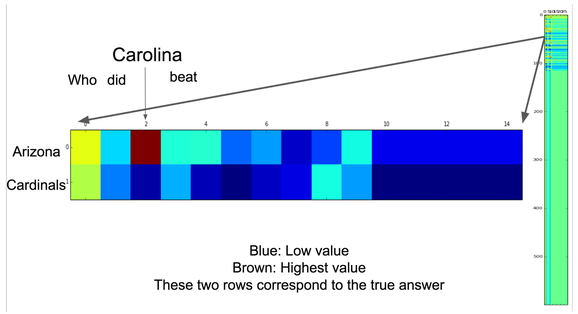

Figure 2: Visualization of the Similarity Matrix

Context: The Panthers finished the ... They defeated the Arizona Cardinals 49-15 in the NFC Championship Game ....
Question: Who did Carolina beat in the NFC Championship Game?
Answer: Arizona Cardinals

$$S_{ij} = w^T[c_i, q_j, c_i \circ q_j]$$
$$C2Q : a_i = \Sigma_{j=1}^{N} softmax(S_{i,:})_j q_j$$
$$\theta_i = \max_j S_{ij}$$
$$Q2C : c' = \Sigma_{i=1}^{N} softmax(\theta)_i c_i$$

Character Embeddings

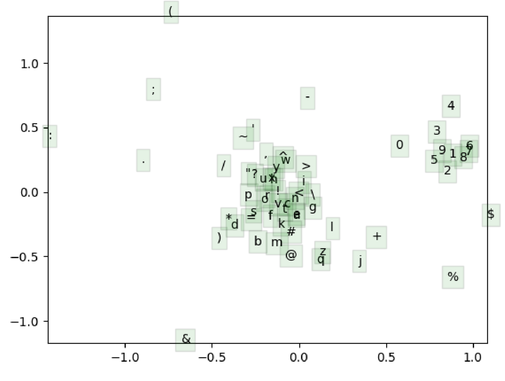

Figure 4: Notice that digits and alphabets are grouped separately

- Model using only word vectors will fail when answers are **numerical** or have special characters.
- Character embeddings solve the problem!

Performance on Dev Set

Model	F1 Score	EM Score	NA
Baseline	43.93	34.58	16.05
Bidaf_attn	49.99	39.66	14.07
Bidaf_attn+cnn	51.76	41.67	13.08
Full_bidaf	73.16	62.89	3.44
Modif_bidaf	73.72	63.83	3.45
Modif_bidaf+dp	75.13	64.02	**0.0**
Ensemble	**77.08**	**66.80**	**0.0**

Answer Length Distributions

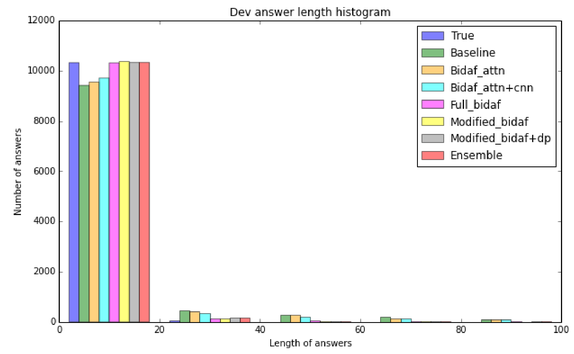

Figure 5: Distribution of answer lengths for each model

DP and Ensemble Explained

- Models predict the end before start many times.
- DP: Find end after start such that $p_1[start]p_2[end]$ is maximum, in linear time.
- Ensemble: Get start and end distributions from different models. Take weighted average based on their individual F1 Scores.

Conclusion and Future Work

- Our final model achieves a test F1 of **77.77**, EM of **68.006**.
- Strengths: Always predicts an answer, is robust to numerical answers, and captures word similarity between question and context.
- Weakness: Unable to predict the correct answer length a large number of times.
- Future work: Attention-over-Attention, N-best re-ranking.