Self-Attention and Character Embeddings for SQuAD

Gabriel Voorhis-Allen (gvoorhis@stanford.edu) and Nick Steele (nsteele@stanford.edu)
Department of Computer Science, Stanford University

Introduction

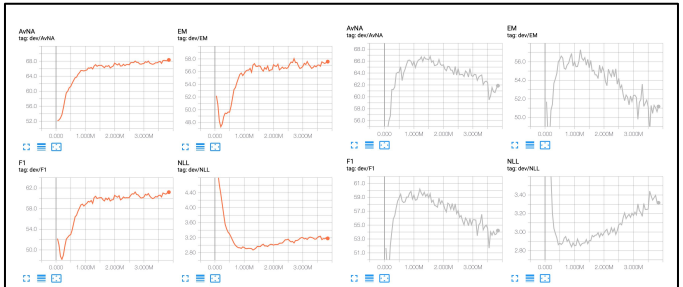
SQuAD 2.0 is a reading comprehension dataset consisting of paragraphs and questions about those paragraphs. Given a context (several sentences or paragraphs) and a related question, the goal is for the model to answer a question related to the paragraph. This challenge is significant because improved question answering capabilities have the potential to positively impact many fields including academic research, medical services, and education. SQuAD 2.0 is particularly well-suited for evaluating question answering models because of its size (about 150,000 examples of Wikipedia passages and crowdsourced questions and answers), its quality, the fact that each answer is an exact substring of the context, and the fact that it contains about 50,000 unanswerable questions. Our model combines several features that make it well-suited to question answering. It includes three primary parts:

1. A Bidirectional Attention Flow model 2. Character-level embeddings 3. A Self-Attention layer These features, in addition to hyperparameter tuning, obtain a F1 score of 61.221 on the test leaderboard.

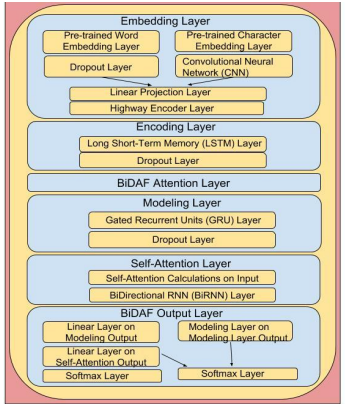
Related Work

There are a wide variety of recent papers on deep learning approaches to question answering on SQuAD. Our baseline model is based off of BiDAF by Seo et al. 2016 with no character embeddings (4). BiDAFcombines context-to question and question-to-context attention, leading to a large performance gain over previous models. Our self-attention layer is inspired by Wei et al. 2017 in RNET (9), which aggregates evidence to answer a question from the entire context to form an answer, and also created large performance gains over other models. The current top of the leaderboard (2) models use pre-trained contextual embeddings such as ELMo and Bert (13) (14). Pre-trained contextual embeddings use use word embeddings that are dependent on the context in which words appear in text, and thus post large performance gains over other models. This represents an alternative approach.

Models and Training



Approach



Results, Hyperparameter Tuning, and Architecture Search

Model	F1 Dev Score (High to Low)	EM Score
Char Embedding Only	61.13	57.67
Baseline	61.01	58.04
Char Embedding AND Self Attention (Full Model)	60.17	57.30
Char Emb. + Self Att. (Reduced Dimensionality, Default Hyperparameters): Hidden Size = 64, Modeling Output Size = 8 * Hidden Size	59.71	56.91
Self Attention Only (no Char Embeddings)	59.49	57.12
Char Emb. + Self Att. (Reduced Dimensionality): Learning Rate Decay = 0.001	57.62	54.15
Char Emb. + Self Att. (Reduced Dimensionality): Dropout = 0.5	52.21	52.21

Discussion of Results

The character embeddings improved marginally on the baseline, but self-attention (under our implementation) did not.
Lack of inclusion of a second GRU in self-attention, as well as our choice to use simpler models in the name of training efficiency, likely lead to this underperformance.
Going forward, we would like to add a second GRU to our attention function, and experiment with more complex models (more/larger layers, and more complex versions of attention such as additive attention).

Error Analysis

1. Inaccurate answer length boundaries:
Context: ...The Islamic Republic has also maintained its hold on power in Iran in spite of US economic sanctions...
Question: What Republic has maintained its control of Iran?
Answer: Islamic Prediction: Islamic Republic
Analysis: These errors are not a significant concern for us as humans likely make them quite often as well.

2. Predicting an answer when there is none:
Context: .... In 2006, Internet2 announced a partnership with Level 3 Communications to launch a brand new nationwide network, boosting its capacity from 10 Gbit/s to 100 Gbit/s…
Question: Who did Internet2 partner with to boost their capacity from 100 Gbit/s to 1000 Gbit/s?
Answer: N/A **Prediction:** Level 3 Communications
Analysis: This can be addressed by testing different values of a decreased null threshold (making it easier for the model to predict "N/A").

3. Not remembering relevant context information:
Context: ..Ogedei's grandson Kaidu refused to submit to Kublai and threatened the western frontier of Kublai's domain. ...Li Tan, the son-in-law of a powerful official, instigated a revolt against Mongol rule in 1262....
Question: Who was Kaidu's grandfather?
Answer: Ogedei **Prediction:** Li Tan
Analysis: The model is assigning more weight to words later in the context ("son in law" comes much later than "grandson"). Self-attention is likely not working properly. We can add an additional gated attention-based recurrent network in the attention layer before the attention output is passed into the biRNN (as done in RNET). We might also use a GRU instead of an RNN to better remember long-term dependencies.

4. Not interpreting complex questions correctly:
Context: ...The defeat along with economic stagnation in the defeated countries, was blamed on the secular Arab nationalism of the ruling regimes…
Question: Secular Arab nationalism was blamed for both the defeat of Arab troops as well as what type of stagnation?
Answer: economic **Prediction:** secular Arab
Analysis: The model seems to interpret the questions as "What was blamed. . . ", instead of realizing that the question was actually asking for a type of stagnation. We could increase complexity (types of attention, number of layers) or train more on these types of questions.