Incorporating Self-Attention and Character Embeddings in a Question Answering System

Zhang, Melanie

CS 224N, Computer Science, Stanford University

Abstract

In this project, I aim to improve upon a baseline contextual question-answering BIDAF SQuAD model by using Context to Context self-attention and character-level embeddings. With the addition of a self-matching attention layer that receives question-aware context representations from the existing attention layer as input, the model is able to more effectively gather evidence from the entire context to help determine the answer. The existing embedding layer is extended with character-level embeddings, which allows the model to deal with out-of-vocabulary tokens and infer more meaning from contexts and questions containing them. The model is able to improve upon the baseline EM and F1 scores substantially, achieving scores of 63.452 and 66.861 respectively.

Problem

Question answering is a problem within NLP whose solutions allow users to formulate questions using natural language and receive an informative response. This has applications in search engines, personal assistants, and more.

The SQuAD dataset consists of paragraph, question, and answers crowdsourced using Amazon Mechanical Turk. Each answer is a span of the paragraph text. This project aims to improve on a provided baseline SQuAD model, by adding a context-to-context self attention layer, character embeddings, and GRU encoding layers.

Background

The baseline BIDAF (Bidirectional Attention Flow for Machine Comprehension) model consists of:
- an embedding layer (with only word-level embeddings, although the original model uses character-level as well)
- encoder layer
- question-to-context and context-to-question attention layer
- modeling layer (another RNN encoder)
- output layer: probability distribution over the start and end positions of the answer span

This project adds self-attention to this baseline model:
- Self-attention is a paradigm that has seen great success in both RNN and Transformer models.
- A NERF incorporated Context to Context self-attention on top of a BIDAF model in 2017 landed first place on the SQuAD score leaderboard at the time of inception
- I also add character-level embeddings to the model, as done in the original BIDAF and RNet models. This allows the model to better handle out-of-vocabulary words.

Methods

Context-to-Context Self Attention
- We add a Context-to-Context self attention layer on top of the existing attention & modeling layers
- The outputs of the self attention layer are fed into 1+ RNN encoder layers

Embedding Layer
- We extend the existing embedding layer by using character-level embeddings.
- Each word is split into characters which are embedded into vectors
- Each sequence serves as input into a 2D convolutional layer
- Input channel size cn = character embedding size
- output channel size cn = hidden size
- Resulting output is concatenated with the word embedding, doubling the hidden size of the original model

Other Model Changes
- Use GRUs in place of LSTMs in each encoder layer

Experiments

- To evaluate this model's performance, I utilized the train, dev, and test sets provided.
- The inputs and outputs to the model are (question, context) pairs and the outputs are spans of the given context.
- Other configurations:
  - LR 0.5
  - Dropout probability 0.7
- Below are the Tensorboard graphs of train and dev NLL, as well as EM & F1 dev scores during training.
- The orange line is the baseline model, red is character embeddings only, and blue is the final model.

Analysis

The baseline model made occasionally made syntactic and grammatical errors, e.g. punctuation truncation (An example answer: [cell
- reproductive], which is missing a closing parenthesis). The model extended with character embeddings learns these grammatical patterns and improves upon the baseline.

The model extended with character embeddings and multi-headed self-attention improves upon question-answers that have longer contexts that require long-range dependencies within the context to answer the question. This is due to the fact that self-attention allows the model to effectively pool information across the entire context in each individual context word representation.

Conclusion

This project demonstrates the effectiveness of adding self-attention and character-level embeddings incrementally to improve a baseline SQuAD question answering model. My model is able to improve upon the baseline EM and F1 scores, achieving scores of 63.452 and 66.861.

References