HW 1: Language Models

Shreshta Bhat

February 8, 2022

A. Please list at least one correct and one incorrect prediction from your improved network, and give a proposed explanation for why the model might have gotten it wrong. Did the pooling network get these examples right?

The word 'successive' is predicted as ADJ by the improved network, which is a right classification. However, the pooling network predicts it as NOUN. This is probably because the pooling network does not consider character ngrams as the improved network does, and hence recognizing that the pattern -ive is probably an adjective. Both the improved network and pooling network categorize 'under' as NOUN instead of ADP, this is because of the class imbalance in our train set where ADP only occurs 68 times as opposed to thousands of nouns, adjectives, and verbs (2251, 9746, 4006). In fact, neither of the networks risk classifying any word as ADP.

B. Please describe the modifications you made to your LSTM and its corresponding perplexity. Include (1) a concise and precise description of the extension that you tried, (2) a motivation for why you believed this approach might improve your model, (3) a discussion of whether the extension was effective and/or an analysis of the results, and (4) a bottom-line summary of your results comparing validation perplexities of your improvement to the original LSTM. This should involve some combination of tables, learning curves, etc. and be at least half a page in length.

(1) I employed embedding dropout, by zeroing out the embeddings for a random set of words with a probability of 0.2 in the matrix. I also added another linear layer, projecting the 512 sized LSTM output to 256, before projecting it down to 128 and vocabulary size. I also tried scheduling the learning rate exponentially and dynamically changing it during training.

(2) Embedding layer is just another linear layer and adding dropout regularization to the embedding layer essentially results in ensembling, thereby reducing overfitting. Changing the hyperparameters was experimental, some combinations worked and some did not. I tried changing the learning rate during training as it is useful to slow down learning as training progresses.

(3) The extension was effective and decreased the validation perplexity, although not too significantly. Changing the hyperparameters is what made a significant difference. Adding an exponential scheduler increased the validation perplexity.

1C. What prompt did you use for the final part of this assignment? How did you choose it, and did you make any other modifications to the prediction function?

Prompt: "What a terrible movie" is negative. "That was a great film" is positive. "It was a wonderful song" is positive. "He is a disgusting man" is negative. "This is exciting to work with" is positive.

I chose simple sentences with no ambiguity and a clear positive/negative adjective. For the prediction function, I used the distribution over the last prediction to extract scores and returned 0/1 based on the maximum probability. This gave me an accuracy of 74%