CS224n: Assignment 4 Solutions

Shubham Gupta

May 5, 2021

1 NMT with RNN

1.1 1G

- Masking sentences is critical for attention to work.
- In both the encoder and the decoder, they help set the attention to zero for the padded tokens and non-zero for the actual tokens.
- In the decoder, the prevent the decoder from "peaking" into the tokens in the future. This helps ensure that the decoder focusses only on the information from the past.
- They also prevent the decoder from predicting the ¡pad¿ padding tokens that are usually present in every training batch. These tokens are not useful during prediction, since the sentence predicted generally ends with a ¡EOS¿ token instead.

1.2 1J

- **Additive Attention**
  - This was the original method introduced by Bahdanau.
  - It works well for large dimensions of data.
  - However, this is slow to compute as there it cannot be vectorized
- **Multiplicative Attention**
  - It simplifies the additive attention operation by computing $f_{att}(h_i, s_j) = h_i^T W_a s_j$
  - This is similar to additive attention in terms of complexity but it is easier to compute because of matrix operations that can be vectorized.
  - It does not work well as the number of dimensions increases.

2 Analyzing NMT Systems

-