Assignment4

April 28, 2020

1 Neural Machine Translation with RNNs

(g) The mask is used for setting padding words' weights(attention distribution score) to $-inf$, i.e. it sets the impact of padding words in source sentence to minimal for decoder. It is useful because the decoder should only pay attention to those real words rather than padding words in source sentence.

(i) Dot product attention vs Multiplicative attention
Dot product attention: Pros: Fewer parameters. Faster and space-efficient. Cons: If attention output vector is not in the same space as decoder's hidden state, the learned result may be not that optimal.

Multiplicative attention vs additive attention.
Additive attention: Pros: Captures non-linear interaction between decoder's attention output and encoder's hidden state. Cons: Computation is slower.

2 Analyzing NMT Systems