CS 224n Assignment #5

Yoshihiro Kumazawa

July 6, 2020

1. **Character-based convolutional encoder for NMT**

   (a) Convolutional architectures can operate over variable length input too since convolutional layers slide fixed-sized windows over input unlike linear layers.

   (b) The size of the padding should be 1 so that the padded vector will have size at least 5. Indeed, m_{word} could be 1 if all words in a batch happen to be some characters of length 1 like 'a', in which case we have $\mathbf{x}'_{\text{padded}} \in \mathbb{Z}^3$.

   (c) The Highway layer makes it possible to combine local features and global features. In other words, it matches our intuition that we can sometimes understand the meaning of a word by just looking at a little chunk of consecutive characters at a time but it is sometimes better to consider the whole characters in it at once. In order to simplify the network semantics in the beginning of training, I would initialize $\boldsymbol{b}_{\text{gate}}$ to be negative.

   (d) Transformers are easier to parallelize and faster to train.

   (e)

   (f)

   (g)

   (h)

   (i)

   (j)

2. **Character-based LSTM decoder for NMT**

   (a)

   (b)

   (c)

   (d)

   (e)

3. **Analyzing NMT Systems**

   (a)

   (b)   i.

        ii.

       iii.

   (c)