# CS 224n Assignment 5.

Abhishek Goswami.

February 21, 2019

1. Character-based convolutional encoder for NMT

   (a) The embedding size used for character-level embeddings is typically lower than that used for word embeddings. This is because set of characters in a language vocabulary is typically much smaller than the number of words in a language. This provides some intuition for using smaller embedding sizes for char-level embeddings.

   (b) Parameters in the word-based lookup embedding model:

   $$V_{\text{word}} * e_{\text{word}} \qquad \text{word embedding lookup parameters}$$

   Parameters in the character-based lookup embedding model:

   $$\begin{aligned} &(V_{\text{char}} * e_{\text{char}}) && \text{character embedding lookup parameters} \\ &+ ((e_{\text{word}} * e_{\text{char}} * k) + e_{\text{word}}) && \text{CNN parameters} \\ &+ ((e_{\text{word}} * e_{\text{word}}) + e_{\text{word}}) && \text{Highway parameters} \\ &+ ((e_{\text{word}} * e_{\text{word}}) + e_{\text{word}}) && \text{Highway parameters} \end{aligned}$$

   (c) When a 1D convnet computes features for a given window of the input, those features depend on the windows only. Moreover, each filter computes its output features *independent* of the other filters. This means :

      i. We can have multiple filters, each of which will produce their own activation map.

      ii. Each filter ends up capturing a different interpretation out of the same input level character embeddings.

   This can be very powerful, and something the RNN model does not capture.

   (d) i. Max Pooling
      - Advantage : max pool extracts max features. This means max pooling is highly sensitive to the existence of some pattern in the pooled region.
      - Disadvantage : A lot of data gets discarded, because we are only taking the max value of the window.

      ii. Average Pooling
      - Advantage :The entire data gets used, because we take the average of the values in the window.
      - Disadvantage : Since average pooling computes the average over the pooled region, it may end up diluting some strong signal especially if there are much smaller values in the pooled region.

   (e) (coding)

   (f) (coding)

   (g) (coding)

(h)
- Verified that input and output have expected shape
- Created a small instance of the highway network, and fed in an input of all 0's. Verified the module returned the expected output.
- Did some tests with incorrect input sizes. Verified they crashed with expected size mismatch errors.

(i)
- Verified that input and output have expected shape
- Created a small instance of the cnn network, and fed in an input of all 0's. Verified the module returned the expected output.
- Did some tests with incorrect input sizes. Verified they crashed with expected size mismatch errors.

(j) (coding)

(k) (coding)

(l) (coding)

2. Character-based LSTM decoder for NMT

(a) (coding)

(b) (coding)

(c) (coding)

(d) (coding)

(e) (coding)

(f) Corpus BLEU: 24.411367086676762

3. Analyzing NMT Systems

(a)
- traducir . present
- traduzco . not present
- traduces . not present
- traduce . present
- traduzca . not present
- traduzcas. not present

This is a bad thing for word-based NMT system. For words missing in the source language, the target word produced by a word-based NMT system would most likely be <unk >as well. A character aware NMT system could overcome this problem because whenever the word-level decoder produces <unk >, we run the CharDecoderLSTM to generate a target word one character at a time. This helps us produce a target word, instead of just printing <unk >

(b) i. Word2Vec
- financial. economic (0.463)
- neuron . nerve (0.559)
- Francisco. san (0.184)
- naturally . occurring (0.545)
- expectation . norms (0.627)

ii. CharCNN
- financial. vertical (0.301)
- neuron . Newton (0.354)
- Francisco. France (0.420)

- naturally . practically (0.302)
- expectation . exception (0.389)

iii. Word2Vec models *contextual* similarity. So words which occur nearby each other are seen next to each other e.g. Closest word to **Francisco** is **San** , because thre is a city 'San Francisco'.
CharCNN on the other hand is a submodule of a larger NMT module. CharCNN models convolution operations over character embeddings, so at best we can say it models some form of *character level* similarity. This is evident from words closest to each other . Both **financial**, and its closest word **vetical** have suffix 'al' . **naturally**, and its closest word **practically** have suffix 'ally' . Words 'expectation' and 'exception' share almost the sam characters, which shows that CharCNN models words in terms of character level similarity.

(c) Explore outputs

i.
- Source Sentence : La epifana es que la muerte es parte de la vida.
- Reference Translation : The epiphany is that death is a part of life.
- A4 Translation : <unk >is that death is part of life.
- A5 Translation : The epiphany is that death is part of life.
- Comment : The character-based decoder produced an exact translation in place of <unk >

ii.
- Source Sentence : Un amigo mo hizo eso – Richard Bollingbroke.
- Reference Translation : A friend of mine did that – Richard Bollingbroke.
- A4 Translation : A friend of mine did that – Richard <unk >
- A5 Translation : A friend of mine did that – Richard Bolla.
- Comment : The character-based decoder produced an incorrect translation in place of <unk >. It uses a different surname for Richard, which might end up referring to a completely different person.