Name: Chakrya Ros

Section 4: Machine Translation & Transfer learning (20 points)

1. Explain how the E-M algorithm works for IBM Model 2.

a) **Choose random initial values for the parameters (q, t)**
b) **E-step: Use the training data in order to estimates or guess the values of missing.**
c) **M-step: Update the values of the parameters in order to maximize the objection functions. Check whether the values have converged or not. If not, then repeat the process**.
d) **E-M algorithm is necessary because it will converge to a local maximum of the log likelihood function and it is guaranteed to never decrease to the value of its objective function on any iteration.**

2. Explain how the encoder-decoder architecture works for neural machine translation with no attention used by answering the following questions.

a) **The output at time step t is the input at time step t-1.**
b) **The output is one-time step ahead of the input. It takes the output from previous layer and outputs one hot vector representing the target word.**
c) **No idea**
d) **Loss function, like Cross-Entropy Lost, sum over the negative log likelihoods that the model gives the correct word at each position in the output sentence. The lower loss value means the model has better translation.**

3. Explain how transfer learning can be used (2 – 3 sentences) **and** what one important limitation that it faces is (2 – 5 sentences).

- Transfer learning can be used to develop model approach and pre-trained model approach. For develop model approach, we select the source task that we already learned. Then we develop the better model on this source task. This model must be better than a naïve model to ensure that some features learning has been performed. Finally, we reuse this model to be starting point for a model on the second task. For pre-trained model, we can do similar to the develop model approach. First, selecting pre-trained source model to reuse as the starting point for a model on the second task of interest. This approach is commonly used for deep learning like a word embedding and sentiment analysis.

- One important limitation for transfer learning is limited resource because in deep learning models, we need a lot of training data and data is limited for language, task and domain. For example, if we want to build language model for different languages like English and Afrikaans. In Wikipedia, English has 49M pages and Afrikaans have only 200k pages. So, it's not possible for NLP systems can trained for Afrikaans language model as they can for English. This is the problem of negative transfer. If we used transfer learning for Afrikaans as source task to reuse for second task for English. It would be end up with a bad performance or decrease accuracy of the new model.

Multiple Choice

4. How is a contextualized word embedding different than a "regular" word embedding? (choose the best single answer)
    1. contextualized word embeddings are less sparse
    2. **contextualized word embeddings capture polysemy better**
    3. contextualized word embeddings are higher dimensionality
    4. contextualized word embeddings are trained using the CBOW algorithm
5. Which of the following are common examples of a source task for transfer learning? (Choose all that apply)
    1. **copy task**
    2. edit task
    3. **language modeling**
    4. machine translation
    5. NER recognition
    6. **POS tagging**
    7. **sentiment analysis**
    8. sentence end detection
    9. summarization task

Citations:

- Post-quiz 33
- Pre-quiz 35
- Lecture 33 EM, IBM Model 2
- Lecture 35 Neural Machine Translate
- Lecture 38 Transfer learning
- https://towardsdatascience.com/neural-machine-translation-15ecf6b0b
- https://machinelearningmastery.com/transfer-learning-for-deep-learning/
- https://www.allerin.com/blog/exploring-the-limits-of-transfer-learning