Assignment 2: Sequence-to-sequence with attention 2020-2021 Spring Semester

Submission Due: 11:59pm on March 15 (Mon)

Submission: Send iPython notebook (or Python codes) and a report in PDF to ai6127.assignments@gmail.com with "AS2-[YourName]"

2. Question One **[50 marks]**
    a. Rerun the implementations of machine translation (MT) of Tutorial 6 with the multiple parallel datasets at http://www.statmt.org/wmt14/training-parallel-nc-v9.tgz as follows: **(25 marks)**
        i. Refer to http://www.statmt.org/wmt14/translation-task.html#download (News Commentary) for details
        ii. Randomly split a dataset into 5 subsets (S_1, …, S_5). Run training/testing for 5 times as follows: (5 marks)
            1. Select S_i (i=1,…,5) as test dataset and the other 4 subsets as training dataset. Train a model with the training dataset and evaluate the model against the test dataset in terms of BLEU (BLEU-1, BLEU-2, BLEU-3).
            2. Report the average of the 5 BLEU scores for each dataset.
        iii. Rerun the implementations for Question 3 **(10 marks)** and Question 6 **(10 marks)** of Tutorial 6.
            1. Do not use the filters of Q 1.b of Tutorial 6 (i.e. MAX_LENGTH, prefixes)
            2. For Question 6, if you do not find parameters that outperform the MT implementation of Question 3 for a dataset, specify which parameters you have tested and discuss why they were not effective for the dataset.
            3. Note that the file has 4 parallel datasets (CS-EN, DE-EN, FR-EN, RU-EN). Rerun the implementations for all the 4 datasets, English as target language.
    b. The Attention Decoder of Tutorial 6 is different from the attention decoder of the lecture (Sequence-to-sequence with attention). **(25 marks)**
        i. Explain the difference **(5 marks)**
        ii. Modify the Attention Decoder of Tutorial 6 to the attention decoder of Lecture 6. Evaluate it for Question 3 of Tutorial 6 against the 4 parallel datasets aforementioned. Compare its evaluation results with the results from Question 2.a of this assignment, and discuss why. **(10 marks)**
        iii. Replace the dot-product attention of Question 2.b.ii of this assignment to the following attention variants: **(10 marks)**
            1. Multiplicative attention: $\boldsymbol{e}_i = \boldsymbol{s}^T \boldsymbol{W} \boldsymbol{h}_i \in \mathbb{R}$
                a. Where $\boldsymbol{W} \in \mathbb{R}^{d_2 \times d_1}$ is a weight matrix
            2. Additive attention: $\boldsymbol{e}_i = \boldsymbol{v}^T \tanh(\boldsymbol{W}_1 \boldsymbol{h}_i + \boldsymbol{W}_2 \boldsymbol{s}) \in \mathbb{R}$
                a. Where $\boldsymbol{W}_1 \in \mathbb{R}^{d_3 \times d_1}, \boldsymbol{W}_2 \in \mathbb{R}^{d_3 \times d_2}$ are weight matrices and $\boldsymbol{v} \in \mathbb{R}^{d_3}$ is a weight vector. d_3 (the attention dimensionality) is a hyper-parameter.
            3. Refer to https://ruder.io/deep-learning-nlp-best-practices/index.html#attention for more details of the variants

4. Compare evaluation results for Question 3 of Tutorial 6 with the results from Questions 2.a and 2.b.ii of this assignment, and discuss why.