Assignment 4: Sequence to Sequence Models

Homework assignments will be done individually: each student must hand in their own answers. Use of partial or entire solutions obtained from others or online is strictly prohibited. Electronic submission on Canvas is mandatory.

Machine Translation (100 points)

A Sequence to Sequence (seq2seq) network is a model consisting of two separate RNNs called the encoder and decoder. The encoder reads an input sequence one item at a time, and outputs a vector at each step. The final output of the encoder is kept as the context vector. The decoder uses this context vector to produce a sequence of outputs one step at a time.

The attention mechanism introduced by Bahdanau et al. (2016) gives the decoder a way to "pay attention" to parts of the input, rather than relying on a single vector. For every step, the decoder can select a different part of the input sentence to consider.

First, download a pair of languages here.

- (a). (0 pts) Divide the data into train, validation, and test or use cross-validation. Report the size of each part. Use at least 20,000 parallel sentences (examples) for training your model.
- (b). (0 pts) Preprocess the text data: tokenize each sentence and build a vocabulary. Initialize an embedding vector for each word.
- (c). (50 pts) Implement a seq2seq model which includes an encoder RNN class and a decoder RNN class. For each RNN cell, try a simple RNN, LSTM, and GRU. You can also use stacked RNN layers. You can also add an attention layer to the encode-decoder RNNs.
- (d). (30 pts) Train the model and plot the training loss and validation loss every 200 iterations. Use the validation set to compute the BLEU score and report the best choice.
- (e). (10 pts) Compute the BLEU scores of the best seq2seq model you get on the testing data set.
- (f). (10 pts) Select 20 test examples. For each example, print the translation results of each model along with the ground truth. For example, if your task is translating from French to English:

```
French:  Reprise de la session
Ground-truth English:  Resumption of the session
Translation from seq2seq model:  Session resumption
Translation from seq2seq plus attention:  Repeat of the session
```

Submission Instructions You shall submit a zip file named Assignment4_LastName_FirstName.zip which contains: (Those who do not follow this naming policy will receive penalty points)

- python files (.ipynb or .py) including all the code, comments and results. You need to provide detailed comments in English.

- (optional) report(.pdf) for each task: Describe the dataset you choose and your model: size of the training set and validation set, parameters for your model, seq2seq structures, RNN cell choice, loss function, learning rate, optimizer, etc. Plot for training and validation loss. Report BLEU scores for all the options (different RNN cells, different attention scores) you have evaluated.