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.

- (a). (5 pts) Load training, validation, and test. Encode the data into token ids.
- (b). (40 pts) Implement the seq2seq model, including
  - (10 pts) an encoder,
  - (15 pts) decoder,
  - (10 pts) a seq2seq model,
  - (5 pts) and a seq2seq loss.
- (c). (50 pts) Training and test the model.
  - (15 pts) You will need to pad the batch into equal lengths,
  - (10 pts) implement a batch index sampler,
  - (15 pts) After training, you will need to translate the test data,
  - (5 pts) show 10 examples,
  - (5 pts) and compute the bleu score.

```
French:  merci.
True English:  thank you.
Translated English:  thank you.
```

- (d) (5 pts) Finally, you will need to analyze the model and translate results.

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 we choose and your model: size of the training set and validation set, parameters for your model, seq2seq structures, loss function, learning rate, optimizer, etc. Plot for training and validation loss. Report BLEU score.