CS584: Natural Language Processing Name: Due: April 23, 2024

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.

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. Name the python files by A4_firstname.py or A4_firstname.ipynb.
- (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.

Machine Translation (100 points)

A Sequence to Sequence (seq2seq) network is a model consisting of two separate RNNs called 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) Download data from IWSLT 2017. You can choose iwslt2017-en-fr (English to French); You can also choose other pairs; Load the training set, the validation set, and the test set. Encode the data into token ids.
- (b). (30 pts) Implement a seq2seq model (you can use packages for RNN or GRU modules), including
  - (5 pts) an encoder,
  - (10 pts) a decoder,
  - (10 pts) a seq2seq model,
  - (5 pts) and a seq2seq loss.
- (c). (10 pts) Implement a seq2seq model with an attention layer introduced in class.
- (d). (50 pts) Train and test both the seq2seq model and the seq2seq+Attention model.
  - (15 pts) You will need to pad the batch into equal lengths;
  - (10 pts) Implement a batch index sampler; Create a index sampler to sample data index for each batch. This is to make the sentences in each batch have similar lengths to speed up training.

    ```
    Example:
    Assume there are 7 sentences and their lengths are: [5, 2, 3, 6, 2, 3, 6].
    The batch_size is 2.
    We can make the indices in the batches as follows:
    ```

```
-- [1, 4] of length 2
-- [2, 5] of length 3
-- [0, 3] of lengths 5 and 6
-- [6] of length 6
```

  - (15 pts) Train the model; After training, you will need to translate the test data;
  - (5 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
    ```

  - (5 pts) Compute the BLEU score on the test set for both models.

- (e). (5 pts) Finally, you will need to analyze the models and their translation results.