E208 Fall 2023 Due: 11:59pm Monday October 23, 2023

Homework #7

The goal of this assignment is to train a sequence classifier that can predict if a sentence is in English or Spanish. You should use the official PyTorch documentation to build your system from scratch. You may use other online sources as well but must cite your sources and indicate clearly what portions of your code have been copied and modified from elsewhere. You may work individually or with a partner on this assignment.

Each team should submit one assignment as a single jupyter notebook on Sakai. At the top of your notebook, please indicate both team members' names and who did what. To speed up training, you may want to run your jupyter notebook in Google Colab with a GPU. Note: The datasets provided below are very large, and you don't need to train on everything! In fact, as you develop your code, I would recommend using a tiny subset of data to iterate quickly, and wait until your code is debugged to start training on larger subsets of data. It is much better to have a functioning model that is trained on 1% of the data than a non-functional model that failed to train on 100% of the data.

Part 1: Basic System with fixed-length inputs (65 points)

In the first part of the assignment you will do the following:

- Prepare the data (20 points). Get two large text files: one English file (WikiText-103, 181MB) and one Spanish file (e.g. Spanish text corpus, 155MB). Convert to lowercase and remove all punctuation except "." so the data only contains alphabet characters, whitespace, and periods. Determine a set of unique characters and map all characters to integers. Split the data into train & validation sets, and split each into chunks of fixed length.
- Train 1-layer model (20 Points). Define an LSTM model containing 1 LSTM layer followed by an output linear layer. Your model should classify a fixed-length sequence of characters as English or Spanish. Show your training & validation loss curves, along with your validation classification accuracy.
- Experimentation (20 points). Experiment with different aspects of the model: the number of LSTM layers, the number of fully connected layers, the size of the hidden layer, etc. Train the corresponding models, compare their performance, and provide plots to demonstrate the effect of at least two different hyperparameters of interest.
- Intuition (5 points). Show the output of your model for several specific sentences. Pick inputs that demonstrate the behavior of the system, and try to figure out what things the model is focusing on. Explain your intuition about what the model is doing.

Part 2: Realistic system with variable-length inputs (25 points)

In the second part of the assignment you will do the following:

- Prepare the data (10 points). Your data should be the same as in part 1, except that each sample should contain one complete sentence rather than a fixed-length sequence of characters. This means that your training & validation samples should have variable length. You will need to zero-pad your inputs.

- Train model (15 points). Use the best model architecture that you found from part 1, and train a model on your data. Be careful to handle the zero-padding correctly, since you can no longer use the same index for all batch samples. Show the training & validation loss curves and validation accuracy. Compare your results to the corresponding model in part 1.

An additional 10 points will be graded for the organization and clarity of your notebook. Your notebook should read like a tutorial and be understandable to others!