CSC 487: Deep Learning

Homework 3: Sentiment Analysis

Sentiment analysis is the task of determining sentiment (e.g., positive or negative) from text. In this homework you will experiment with sentiment analysis on a dataset of reviews from IMDB which have been annotated as having either positive or negative sentiment.

The provided starter notebook shows how to download and unpack the data, and how to set up and use the BPE tokenizer from BERT.

Instructions

1. **Bag of words model.** Create a 90/10 train/test split of the data. Create TF-IDF weighted histograms (using TfidfVectorizer) using the top 1000 words and train an MLP model (MLPClassifier) to classify them. Compute the train and test accuracy of the model (using the .score() function).
2. **RNN model.** Train a GRU to process sequences of BPE tokens output a binary sentiment prediction. (Don't forget to set the batch_first flag if needed!)

   Use an Embedding layer to map the BPE tokens to embedding vectors for input to the GRU.

   If the text is too long, take a random sub-sequence; if the text is too short, pad it using token index 0.

   I used the following hyperparameter settings:
   - Maximum sequence length: 100
   - Hidden layer size: 100
   - Num hidden layers: 3
   - Embedding size: 100
   - 10 epochs, Adam optimizer with learning rate 3e-4

Report

Your report should include the following:

- **Code explanation:**
  - Briefly explain your solution and any design choices you made that weren't specified in the instructions.

  - Clearly describe any external sources that were used (e.g. websites or AI tools) and how they were used.
- **Discussion:**
  - Compare the results of bag-of-words and RNN in terms of accuracy.

Deliverables

Python notebook and report document due Sunday, Mar. 9, 11:59 pm.