DS-GA 1011, Fall 2017

Homework Two: Global Vectors

Due October 25th

1 Introduction

In this assignment we'll be learning how to learn vector representations of words! We'll be implementing GloVe, one of the more popular, effective, and efficient approaches to learning word embeddings. Before proceeding, you should read Pennginton et al.'s (2014) paper on Global Vectors for Word Representations, this will be your primary reference.

2 Implement GloVe

Now that you've read the paper, implement it! Note that you must turn in your code in an iPython notebook. **You must also turn in a fully run notebook; we will not run your code to grade your homework.** If the notebook isn't run, expect to lose 50% of the points from this section.

2.1 Setup

Though GloVe is normally trained on massive corpora, since we want the training time to be speedy, we're using the Stanford Sentiment Treebank (SST). The code for data loading, extracting cooccurrences, and batchifying data is provided for you. For speed, we're only looking at the top 250 most frequent words.

2.2 Evaluation metric

Since we're training on such a small dataset, we're not going to test our model on standard analogy or similarity metrics. Instead, we'll use a simple scoring function which grades the model on how well it captures ten easy/simple similarity comparisons. The function returns a score between 0 and 10. Random embeddings can be expected to get a score of 5.

2.3 Implement and train (60 pts)

2.3.1 Model

Fill in the starter code provided for the `Glove()` model.

2.3.2 Training loop

Complete the `training_loop()` function.

2.3.3 Train a working model

Using the provided hyperparameters and commands, train your model. If your model works, it will converge to a score of 10.

3 Questions

Turn in a pdf, no longer than 3 pages, with answers to the following questions.

3.1 Cooccurrence Matrix (10 pts)

What do the entries on the diagonal of the 'cooccurrences' matrix represent?

3.2 Weighting Function (10 pts)

What would deleting the weighting function do and why?

3.3 Corpus Size (10 pts)

If you used 100 times more training data, would the model take 100 times as long to train? Just as long as now? Somewhere in between? Choose one and (informally) defend your answer.

3.4 Window size (10 pts)

How and why does varying window size effect the performance on syntactic and semantic tasks?

4 Logistics

You must turn in both an iPython notebook with the completed code (Sec. 2), and a pdf report with questions answered (Sec. 3). This homework is **due on October 25th, at 6:45 PM**. You can find the late homework policy in the syllabus.

References

[1] Jeffrey Pennington, Richard Socher, Christopher D. Manning, Glove: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP).*

[2] Omer Levy, Yoav Goldberg, Ido Dagan, 2014. Improving Distributional Similarity with Lessons Learned from Word Embeddings. *Transactions of the Association for Computational Linguistics (TACL).*