NLP extra credit

Vedanshi Shah and Arnav Joshi

Answer the following questions carefully (2 - 3 pages total to answer all 4 questions):

1. **Describe what either LSA or SVD is and why it appears in this paper.**
    a. *Latent Semantic Analysis (LSA)*
        i. It is used to break large matrices into smaller chunks. It is used to identify patterns in large datasets to understand the relationship between the terms and documents.
        ii. It utilizes low-rank approximations to decompose large matrices that capture statistical information about a corpus.
        iii. Matrices are of a term-document type with rows corresponding to words or terms and columns corresponding to documents in the corpus.
    b. *Singular Vector Decomposition (SVD)*
        i. It is a mathematical technique to decompose a matrix into its constituent parts. Used for reducing dimensionality of large datasets.
        ii. Breaking matrix into singular values to identify patterns and relationship between the data that is latent.
    c. *Why it appears in the paper*
        i. It appears in this paper because it is one of the primary methods for creating word embeddings, which represent words as vectors in a high-dimensional space.
        ii. The authors of the paper argue that while LSA is effective at capturing global word co-occurrence statistics, it does not perform as well as other methods, such as Word2Vec, in capturing the nuances of word relationships.
        iii. GloVe seeks to improve upon LSA by taking advantage of the information that can be gleaned from both global and local word co-occurrence statistics.
2. **What are the primary differences between skip-gram word embeddings and GloVe word embeddings?**
    a. Skip-gram and GloVe are two popular methods for generating word embeddings, which are dense vector representations of words that capture their meaning and semantic relationships.
    b. The primary difference between skip-gram and GloVe word embeddings lies in the way they capture word co-occurrence statistics. Skip-gram generates word embeddings by predicting the context words that surround a given target word, while GloVe generates word embeddings by factoring a matrix of word co-occurrence counts.
    c. Here are some specific differences between skip-gram and GloVe:

        i. Training method: Skip-gram uses a neural network with one hidden layer to train word embeddings, while GloVe uses a factorization method to train word embeddings based on co-occurrence statistics.
        ii. Context window: Skip-gram considers the context words within a fixed-size window around a target word, while GloVe considers the global co-occurrence statistics of words in a corpus.
        iii. Importance of word pairs: GloVe places more emphasis on rare word pairs that occur together, while skip-gram treats all word pairs equally.
        iv. Speed of training: Skip-gram is faster to train than GloVe, especially for large datasets.
        v. Embedding quality: The quality of the resulting word embeddings can depend on the specific task or dataset, but some studies have shown that GloVe embeddings can perform better on some tasks, while skip-gram embeddings can perform better on others.
    d. GloVe takes syntactic relationships into account as well such as the difference between “King” and “Queen” using weighted least squares regression model.
    e. Overall, skip-gram and GloVe are both effective methods for generating word embeddings, and the choice between them often depends on the specific needs of the task at hand.

3. **At the beginning of section 4.7, the authors state "A rigorous quantitative comparison of GloVe with word2vec is complicated by the existence of many parameters that have a strong effect on performance." explain this sentence using specific examples from the paper.**
    a. The authors state that a rigorous quantitative comparison of GloVe with Word2Vec is complicated by the existence of many parameters that have a strong effect on performance. This is because both GloVe and Word2Vec have several hyperparameters that can be tuned to improve performance, such as the context window size, vector dimensionality, and learning rate.
    b. Additionally, the authors note that different parameter settings may be more effective for different evaluation tasks, such as word similarity or sentiment analysis.
    c. For example, in Table 2 of the paper, the authors compare GloVe and Word2Vec embeddings trained on different corpora and with different hyperparameters, and find that the relative performance of the two methods varies depending on the evaluation task. (also annotated in the paper).
    d. Also, the authors compare the performance of GloVe with different window sizes and different weighting functions on a word similarity task. For example, they demonstrate that a larger window size tends to capture more global co-occurrence statistics, while a smaller window size tends to capture more local context information.

4. **What is your favorite figure or paragraph from this paper? Why?**
    a. I really liked the tables comparing the baselines with their GloVe model because it helps understand the performance with different metrics and changing hyperparameters. It allows us to look at the different models and how they differ.
    b. However, my favorite figure from this paper is Figure 2, which shows the geometric relationships between word embeddings for various semantic and syntactic relationships.
    c. I find this figure compelling because it illustrates how GloVe embeddings are able to capture not only semantic relationships between words (e.g. "king" and "queen"), but also syntactic relationships (e.g. "man" and "woman").
    d. Additionally, the figure demonstrates how GloVe embeddings can be used to perform analogical reasoning, such as "Paris is to France as Rome is to Italy."
    e. Overall, I think this figure provides a clear and intuitive explanation of how GloVe embeddings represent word relationships in a high-dimensional space.