Assignment-2

By : Sumit

RollNo.: 423181

Skip-gram with Negative Sampling on Wikipedia

1. Introduction

Word embeddings are dense vector representations of words that capture semantic and syntactic properties based on their distributional context. So, In this I implemented the Skip-gram model with Negative Sampling (SGNS) as described in Chapter 5 of Jurafsky and Martin. The objective mentioned in our task is to learn meaningful word vectors from a large Wikipedia corpus and evaluate them using cosine similarity, word analogy tasks, and bias detection.

2. Dataset Description

The dataset I used for training is the **enwik8** corpus, obtained from Matt Mahoney's text data repository. The enwik8 dataset is a cleaned and compressed version of the English Wikipedia dump, containing approximately 100 MB of plain text. The text is lowercased and tokenized into words. To reduce computational complexity, the vocabulary is limited to the top 30,000 most frequent words.

3. Task Description

1. Implement the Skip-gram model with Negative Sampling from scratch using PyTorch.
2. Trained the model on the Wikipedia enwik8 dataset on google collab for 2 epochs
3. Extract word embeddings learned by the model after training is complete
4. Compare the learned embeddings with pretrained Gensim Word2Vec vectors using cosine similarity.
5. Evaluate the embeddings using word analogy tasks.
6. Analyze and detect bias in the learned word embeddings as described in chapter 5

4. Skip-gram with Negative Sampling Model

The Skip-gram model aims to predict surrounding context words given a center word. Negative sampling is used to make training efficient by replacing the full softmax with a small number of randomly sampled negative words. The model learns two embedding matrices: one for center words and one for context words. Training optimizes a binary logistic loss that increases similarity between true word pairs and decreases similarity between randomly sampled word pairs.

5. Training Details

The model is trained using the Adam optimizer with an embedding dimension of 100, a context window size of 2, and 5 negative samples per positive example. Training is performed for multiple epochs over the filtered Wikipedia corpus.

6. Cosine Similarity Evaluation

Cosine similarity is used to measure the similarity between word vectors. The learned embeddings are compared with pretrained Gensim Word2Vec embeddings. A moderately high cosine similarityis observed for common words such as 'king', indicating that both models capture similar semantic relationships despite differences in training data and scale.

7. Word Analogy Task

Word analogy tasks evaluate whether vector arithmetic captures linguistic relationships. For example, the relationship 'man : king :: woman : queen' is computed using vector operations. The model successfully retrieves semantically correct answers, demonstrating that both semantic and syntactic regularities are encoded in the embedding space.

8. Bias Detection in Word Embeddings

Bias in word embeddings is analyzed by computing a gender direction vector using the difference between 'he' and 'she'. From the book I saw a example where Profession words such as 'doctor', 'nurse', and 'engineer' are projected onto this direction. The results indicate the presence of gender bias, with certain professions showing stronger associations with specific genders, consistent with observations discussed in Jurafsky and Martin.

9. Conclusion

In this experiment, Skip-gram with Negative Sampling was successfully implemented and trained on a Wikipedia corpus. The learned word embeddings demonstrate strong semantic structure, perform well on analogy tasks, and reveal inherent societal biases. This study highlights both the effectiveness and limitations of distributional word representations