Danial Hasan

Assignment 1

CSC 413
February 3, 2020

Part 1

1. Given vocbulary size V and embedding dimension d, we will have parameters $\mathbf{W} \in \mathbb{R}^{V\times d}$ and $\mathbf{b} \in \mathbb{R}^V$. In total the number of trainable parameters will be: $V\times d + V$.

2. The gradient of the loss function L w.r.t. a single parameter vector $\mathbf{w}_i$ is:

$$\frac{\partial L}{\partial \mathbf{w}_i} = 4\sum_{j=1}^{V} \left(\mathbf{w}_i^T\mathbf{w}_j + b_i + b_j - log\mathbf{X}_{ij}\right)\mathbf{w}_j \tag{1}$$

   Equivalently, if we define:

$$\mathbf{W} = \begin{bmatrix} \mathbf{w}_1^T \\ \vdots \\ \mathbf{w}_V^T \end{bmatrix} \tag{2}$$

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_V^T \end{bmatrix} \tag{3}$$

$$\frac{\partial L}{\partial \mathbf{w}_i} = 4\mathbf{W}^T\left(\mathbf{W}\mathbf{w}_i + b_i\mathbf{1} + \mathbf{b} - log\mathbf{x}_i^T\right) \tag{4}$$

   Which we can easily extend to compute the gradient for the entire W matrix.

3. By inspection of the loss plot versus embedding dimension, it seems that an embedding dimension of 10 leads to optimal validation performance. Increasing the embedding dimension increases the capacity of the model to reconstruct the original co-occurence matrix. However, increasing the capacity can lead to overfitting of the model to the co-occurence matrix of the training set due to the overparameterization. This is evident from the training loss continuing to decrease as the embedding dimension increases.

Part 2

1. Given 3 input words, vocabulary size $V = 250$, embedding dimension $d = 16$, and 128 hidden layer units, we will have $\mathbf{W}_1 \in \mathbb{R}^{250\times 16}$ for our lookup table, $\mathbf{W}_2 \in \mathbb{R}^{(3\times 16)\times 128}$, $\mathbf{b}_2 \in \mathbb{R}^{128}$, $\mathbf{W}_3 \in \mathbb{R}^{128\times 250}$, $\mathbf{b}_3 \in \mathbb{R}^{250}$. In total the number of trainable parameters will be:

$$250 \times 16 + 3 \times 16 \times 128 + 128 + 128 \times 250 + 250 = 42522 \tag{5}$$

   The greatest number of parameters is added by the hidden to output layer which has to softmax over all 250 words and has a weight matrix of dimension $\mathbb{R}^{128\times 250}$.

2. To store all possible 4-gram counts from a vocabulary of 250 words, the table would need 250^3 rows for all possible preceding 3-grams and 250 columns for all possible proceeding words. Thus the total number of entries would be 250^4, which is much larger than the ∼42000 parameters of our neural network.

Part 3

1. Output of `print_gradients()`:

```
loss_derivative[2, 5] 0.0011122317737824977
loss_derivative[2, 121] -0.9991004720395987
loss_derivative[5, 33] 0.00019032378031737032
loss_derivative[5, 31] -0.7999757709589483

param_gradient.word_embedding_weights[27, 2] -0.27199539981936854
param_gradient.word_embedding_weights[43, 3] 0.8641722267354156
param_gradient.word_embedding_weights[22, 4] -0.2546730202374645
param_gradient.word_embedding_weights[2, 5] 0.0

param_gradient.embed_to_hid_weights[10, 2] -0.6526990313918256
param_gradient.embed_to_hid_weights[15, 3] -0.13106433000472622
param_gradient.embed_to_hid_weights[30, 9] 0.11846774618169384
param_gradient.embed_to_hid_weights[35, 21] -0.10004526104604404

param_gradient.hid_bias[10] 0.25376638738156426
param_gradient.hid_bias[20] -0.03326739163635312

param_gradient.output_bias[0] -2.0627596032173052
param_gradient.output_bias[1] 0.0390200857392169
param_gradient.output_bias[2] -0.7561537928318482
param_gradient.output_bias[3] 0.21235172051123632
```

Part 4

1. Using the 3-gram 'today police are', which does not occur in the training set, the model predicts `not` to be the most likely next word, followed by '.' (a period), 'all', and 'nt' (as in *aren't*). All of these seem like plausible predictions, indicating the model has learnt a fairly good distribution to generate from.

2. Looking at one of the clusters for `tsne_plot_representation`, the words are largely related to time, e.g *days, year, days, night, time.* Comparing against the GloVe tsne visualization, there seem to be some similar clusters, such as one related to time. However, the GloVe visualization has more discrete clusters whereas the language model has more concentrated clusters. The 2D GloVe embedding is much less widespread in its distribution of words in the embedding dimension, indicating a weaker representation as compared to the higher dimensional one.

3. The words 'new' and 'york' have a distance of 3.90 as compared to the distance between 'day' and 'time' which is 2.00. This would indicate that 'new' and 'york' are not close in the learned representation. This is likely because the word 'new' is also used in many contexts other than the city, where the word 'york' is largely used in context of referring to the city.

4. 'government' and 'university' are closer than 'government' and 'political' in the representation. As both 'government' and 'university' are nouns that refer to an institution, they are more likely to be used in the same context with similar n-grams. Thus they are more likely to be interchanged within a phrase than 'government' and 'political'.