**Lab5 Report**

**Implementation**

In this project, a simple neural network based on N-Gram language model was implemented. The network comprises 3 layers including an embedding layer (which is input layer) and two hidden layers. The input layer required an matrix in which n represents the size of all unique vocabularies in the training data and m represents the dimensionality of each word. For the hidden layers, the first layer is a linear layer requiring a matrix as input which p is the value of the context size multiplying the embedding dimensionality and the activation function of this layer is Relu function. The second layer is also a linear layer requiring a matrix (n has been mentioned in previous) and the activation function of this layer is the log softmax function. Be more specific, the dimensionality of each layer would be shown in below table.

|  |  |
| --- | --- |
| Layer | Dimensionality |
| Embedding layer |  |
| First hidden layer |  |
| Second hidden layer |  |

The linear layer, in the program, utilizes the mathematical equation which is to implement the linear transform in dimensionality.

**Evaluation**

The program could run training continuously in 5 times with different hyper parameters. The final loss and the correctness of the checking and testing would be shown in the table below. The check in the program is using the sentence “The mathematician ran to the store” to check every trigram. For the test, a sentence, “The \_ solved the open problem”, will be filled with choosing given options which are “physicist” and “philosopher”.

|  |  |  |  |
| --- | --- | --- | --- |
| **Number of Test** | **Final loss** | **Correctness of checking** | **Choice of testing** |
| 1 | Epoch = 10, Learning rate = 0.001, Embedding dimensionality = 10 | | |
| 102.9999 | Incorrect | physicist |
| 2 | Epoch = 10, Learning rate = 0.03, Embedding dimensionality = 10 | | |
| 18.9814 | Correct | physicist |
| 3 | Epoch = 20, Learning rate = 0.03, Embedding dimensionality = 10 | | |
| 13.8663 | Correct | philosopher |
| 4 | Epoch = 100, Learning rate = 0.001, Embedding dimensionality = 10 | | |
| 26.5426 | Correct | physicist |
| 5 | Epoch = 20, Learning rate = 0.03, Embedding dimensionality = 20 | | |
| 13.5955 | Correct | physicist |

From the observation, when epoch keep 10 and the learning rate increase to 0.03 or the epoch comes to 100 and learning rate keep 0.001, the result could keep continuously correct. The embedding dimensionality does not impact the result so much in such a small scale data training. When the prediction starts with the context “<s> The”, the next target word is always predicted as “mathematician” rather than “physicist”. That is because the occurrence of the vocabulary combination “<s> The mathematician” has higher probability than the occurrence of the “<s> The physicist”. Obviously, the Bigram language model could be implemented in this program by changing the size of context to 1.

In the testing part, the answer is not fixed, but usually it is “physicist”. To make the choice be reasonable, the similarity among “mathematician”, “physicist” and “philosopher” have been calculated. When the answer is given “physicist”, the value of the similarity between “mathematician” and “physicist” is always larger than that between “mathematician” and “philosopher”, otherwise the answer is given “philosopher”. The similarities would be given in following table.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Number of Test | Mathematician vs  Physicist | Mathematician vs  Philosopher | Philosopher  vs  Physicist | Choice |
| 1 | 0. 6055 | 0. 4696 | 0. 3905 | physicist |
| 2 | 0. 3836 | 0. 1538 | 0. 4042 | physicist |
| 3 | 0. 3323 | 0. 4340 | 0. 3897 | philosopher |
| 4 | 0. 1614 | 0. 0600 | 0. 6083 | physicist |
| 5 | 0. 1498 | 0. 0666 | 0. 0711 | physicist |