1.Model description

Neural network language model is a kind of multilayer perceptron.

Mathematical equations are:

$$y = \varphi(\sum_{i=1}^{n} \omega_i x_i + b) = \varphi(\mathbf{w}^T \mathbf{x} + b)$$

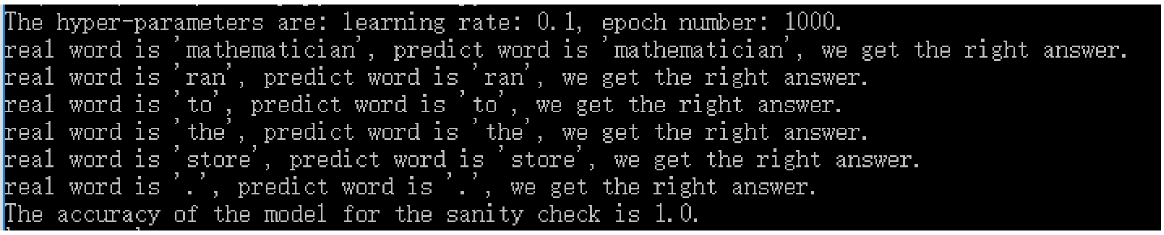
where $\mathbf{w}$ denotes the vector of weights, $\mathbf{x}$ is the vector of inputs, b is the bias and φ is the activation function. The activation function φ is often chosen to be the logistic sigmoid $\mathbf{1} / (\mathbf{1} + \boldsymbol{e}^{-x})$ or the hyperbolic tangent $\mathbf{tanh}(\boldsymbol{x})$.

A typical multi-layer perceptron (MLP) network consists of a set of source nodes forming the input layer, one or more hidden layers of the compute nodes, and an output layer of the nodes.

$$\mathbf{x} = \mathrm{f(s)} = \mathrm{B}\varphi(\mathrm{As} + \mathrm{a}) + \mathrm{b}$$

where s is a vector of inputs and $\mathbf{x}$ a vector of outputs. A is the matrix of weights of the first layer, a is the bias vector of the first layer. B and b are, respectively, the weight matrix and the bias vector of the second layer. The function φ denotes an elementwise nonlinearity.

2. Sanity check

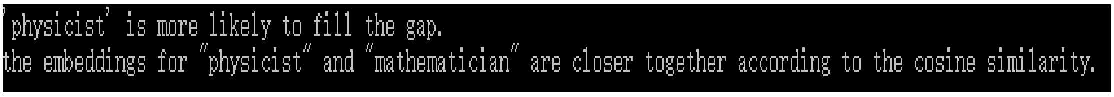


```
The hyper-parameters are: learning rate: 0.1, epoch number: 1000.
real word is ’mathematician’, predict word is ’mathematician’, we get the right answer.
real word is ’ran’, predict word is ’ran’, we get the right answer.
real word is ’to’, predict word is ’to’, we get the right answer.
real word is ’the’, predict word is ’the’, we get the right answer.
real word is ’store’, predict word is ’store’, we get the right answer.
real word is ’.’, predict word is ’.’, we get the right answer.
The accuracy of the model for the sanity check is 1.0.
```

As we can see, the model can work for every trigram.

When we set the learning rate =0.1, epoch number = 1000, the results are all correct. My model predicts for the context "START_OF_SENTENCE The" the word "mathematician" instead of "physicist". That is because the embeddings for " mathematician " and " START_OF_SENTENCE The " are closer together than the embeddings for " physicist " and " START_OF_SENTENCE The ".

3. test



```
'physicist' is more likely to fill the gap.
the embeddings for "physicist" and "mathematician" are closer together according to the cosine similarity.
```

The model can predict “The ______ solved the open problem.” Correctly. It cannot be possible with the bigram ML model. Because in training data, the times of "physicist" appears after 'The' is the same as the times of "philosopher" appears after 'The'.

As the output presents, the cosine similarity between "physicist" and "mathematician" is larger than it between "philosopher" and "mathematician", which is the right reason for the right prediction.