Neural Language Modeling Report

Ziyang Li

01/05/2018

1 Model Analysis

The N-Gram neural network language model contains two parts. The input is three one-hot-presentation words, two context words are used for the first neural and the last one is target word which is used in the last neural. The output is a vector includes probabilities for all the words in corpus.

The first part is a embedding model. In this part, the input is two one-hot-presentation

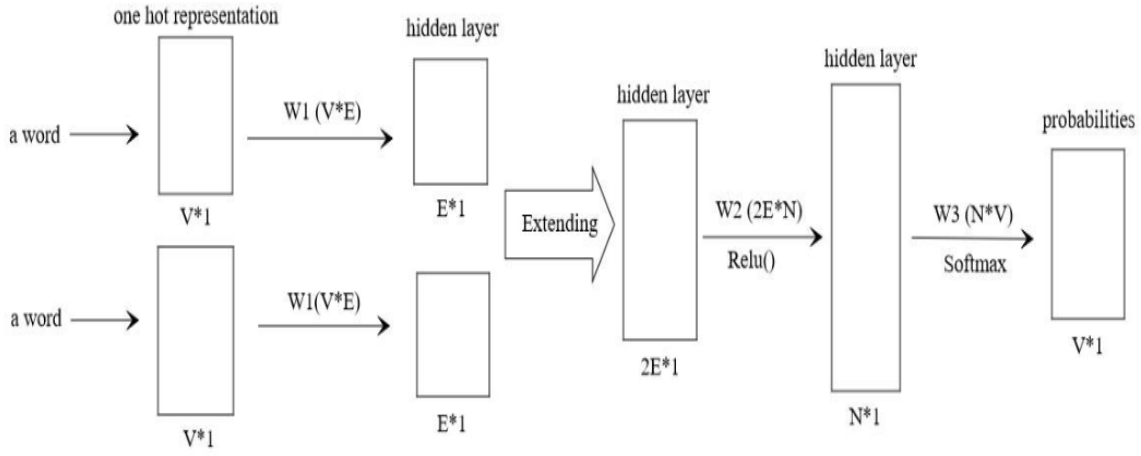

Figure 1: N-Gram network language model

words $(V * 1)$ and the output is two embeddings $(E * 1, V > E)$.

$$W_1^T * U_v = U_e \tag{1}$$

where, W_1 is a $V * E$ embedding matrix, U_v is one-hot-representation of a input word and U_e is the embedding of the input word.

The second part is a softmax model. In this part, the input is two embeddings $(E * 1)$ and the output is probabilities $(V * 1)$.

$$W_3^T * (Relu(W_2^T * [U_e^1, U_e^2])) = P \tag{2}$$

where, [ ] is a extending function which inputs two $E * 1$ vectors and outputs a $2E * 1$ vector, W_2 is a $2E * N$ matrix, Relu() is a activation function, W_3 is a $N * V$ matrix and P is a $N * 1$ vector.

Actually, W_1 is the embedding results for all the words we want. After getting probabilities P, we can use log softmax and maximise the probability of the target word to train the model.

2 Sanity check

The result of sanity check by using trigram language model is shown in figure 2. The model predicts every target word correctly when learning rate is 0.01 and epoch number is 15.

Input	output
START_OF_SENTENCE The	mathematician
The, mathematician	ran
mathematician ran	to
ran to	the
to the	store
the store	.
store .	END_OF_SENTENCE

Figure 2: Sanity check Result

The reason why predicting "mathematician" instead of predicting "physicist" is the probability of "mathematician" is higher than "physicist" when the context is "START_OF_SENTENCE The". In our training data, the number of data ("START_OF_SENTENCE The" , "mathematician") is more than data ("START_OF_SENTENCE The" , "physicist"). So, the weight for "mathematician" should be higher when input is "START_OF_SENTENCE The".

3 Test

For the Given sentence, "physicist" is more likely to fit in the gap. The cosine similarity between "physicist" with "mathematician" is 0.3317 and the cosine similarity between "philosopher" with "mathematician" is -0.6111. However, it could not work if we use the bigram ML model from lab 2. The probability of (START_OF_SENTENCE The , "physicist") is the same as ("START_OF_SENTENCE The" , "mathematician") in training data. Thus, the bigarm ML model does not work for this example. In oreder to predict the gap correctly, we need more features such as words behind the gap and the CBOW model should do better than both of N-gram language model and bigram ML model.