Word Embedding

Friday, February 23, 2024 8:08 AM

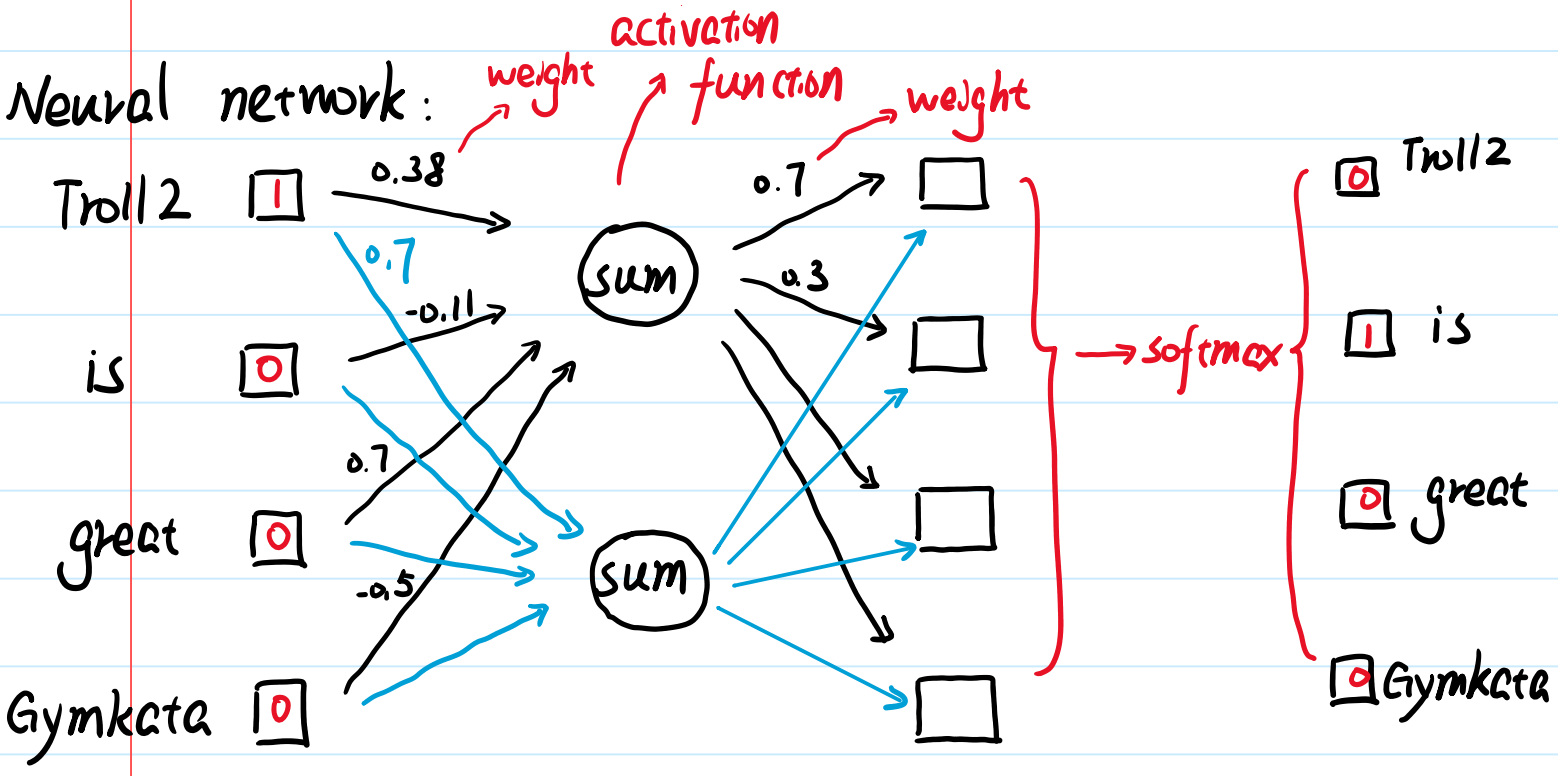
Represent words using numbers:

Example:

Troll2 is great! } training data

Gymkata is great!

Neural network:



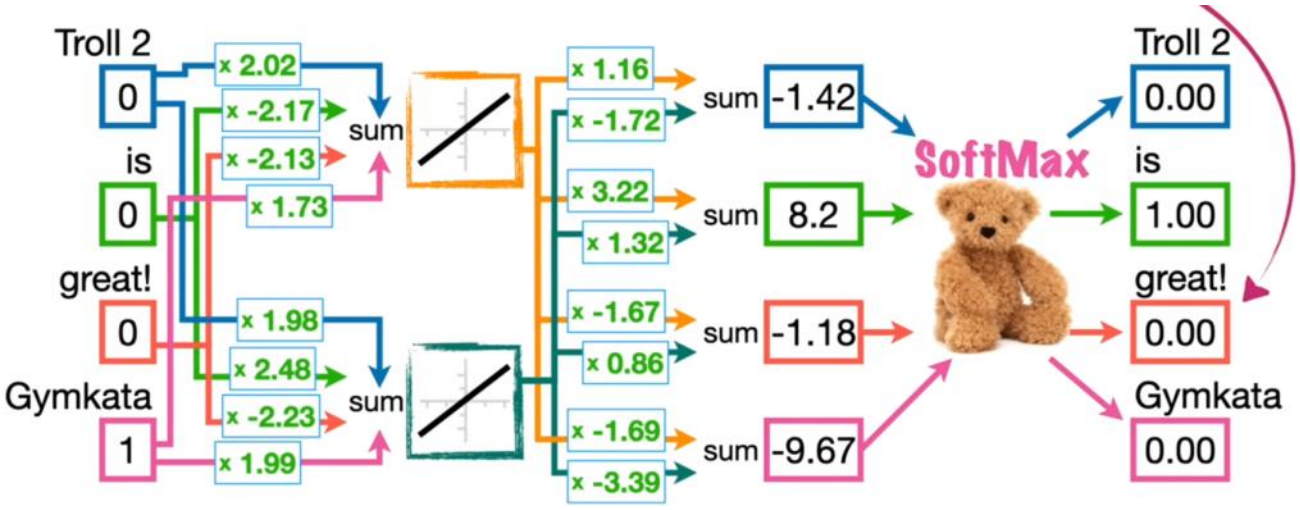
predict the next word to train weights

weight from words to the activation function is "embedding"

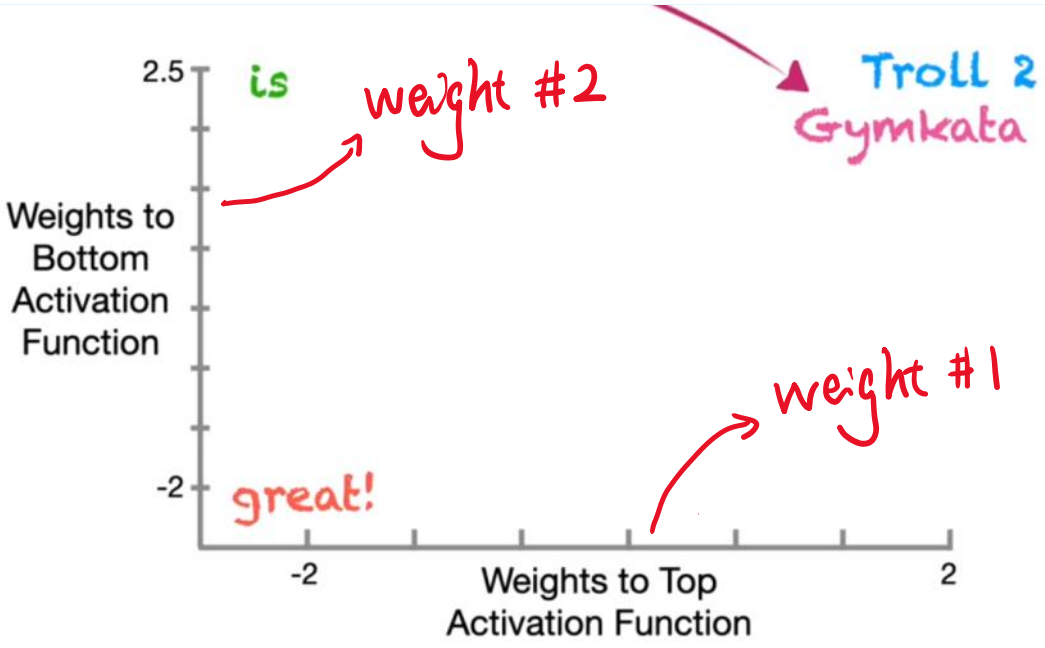
In this case, "troll2" has an embedding of [0.38, 0.7].

Word Embedding

Friday, February 23, 2024 8:08 AM



(picture from "stat Quest" on YouTube)



(picture from "stat Quest" on YouTube)

Word 2 Vec

Friday, February 23, 2024 8:08 AM

① Continuous bag of words.

use surrounding texts to predict the middle words

Troll 2 is great

predict

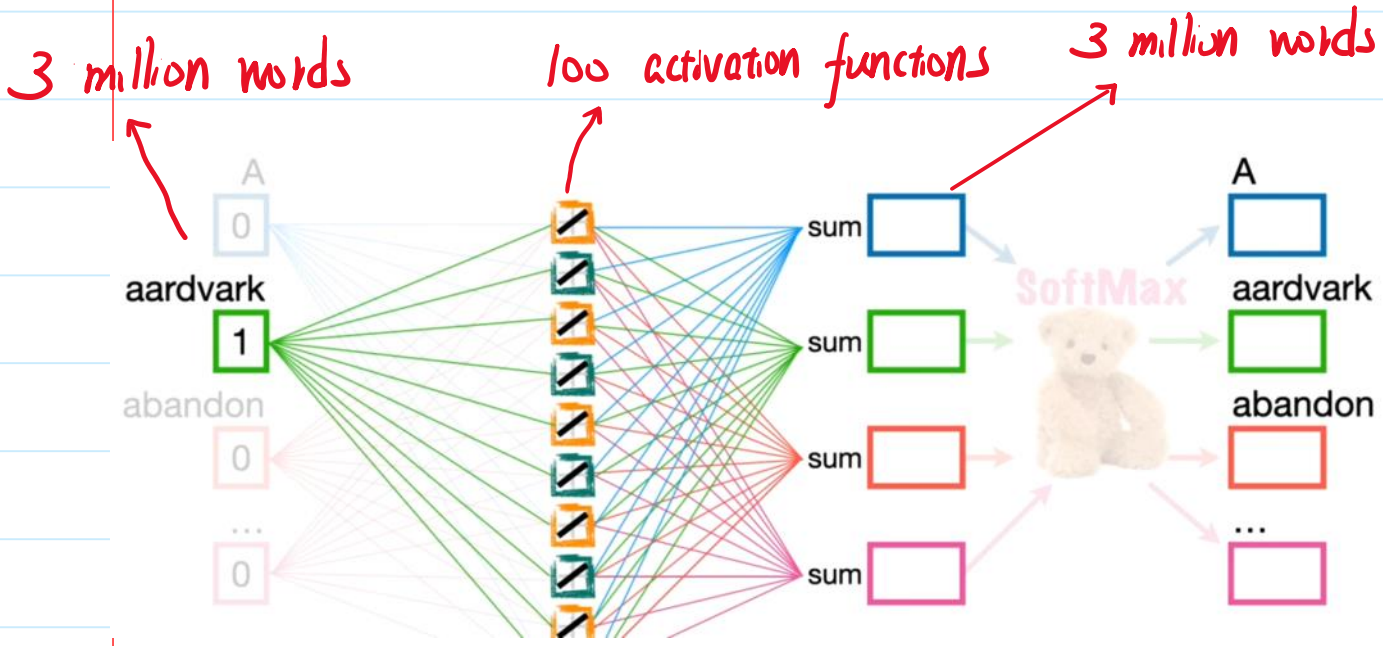
② skip gram

use middle word to predict surrounding words.

Troll 2 is great

predict

Actual word 2 vec :



(picture from "stat Quest" on YouTube)

Word2Vec

Friday, February 23, 2024 8:49 AM

The number of training parameters are about:

3 million * 100 + 100 * 3 million = 600 million

To reduce the training load, word2vec uses negative sampling.

For example, if we want to use the word "abandon" to predict its next one "is". Then we know that:

- All input words are ZEROs except for "abandon"
- In the output, only the word "is" is ONE and all other outputs are ZEROs.

This means that:

- All weights from non-"abandon" words are not important anymore because they multiply by each word (which is ZERO) is still ZERO. This means that we only need to train the weight parameters from "abandon" to the 100 activation functions, which is only 100 parameters to train.
- For the output layer, we only select one positive sample (that is "is" in our case) and one negative sample (which may be some other words). In this case, we only need to train the weights from activation functions to the positive output and negative output, which is 200 parameters in total.

By using the approach above, we reduce the total training parameters 600 million to 300.

In short, the word2vec train 300 parameters per step, NOT 600 million.

Be aware that in reality, the word2vec may select 2 to 20 negative samples in the output layer (in our case we assume it is one negative sampling for simplification).