4. Word Embeddings

Extension Note: Homework 4 due date has been extended by 1 day to **July 27 23:59UTC** .

Word Embeddings

1/1 point (graded)

Training a neural network using backpropagation and SGD moves the network weights in a direction that minimizes the loss function. If the network contains a *bottleneck*, a layer in which many inputs are reduced to only a few outputs, training will adjust the weights to maximize the useful information contained in the layer's output. In this way, a sparse input representation can be embedded in a lower-dimensional space to become a dense, *distributed* representation. Embeddings often have interesting properties like transforming semantic or visual similarity into geometric proximity. In this question, we will examine the utility of (the highly popular) word embedings.

Consider two neural networks for classifying sequences of words that differ only in their input representation: The first of which uses a sparse *one-hot* encoding of each word in which word i is identified by a vector that contains a 1 in position i and 0s elsewhere. For instance, a dictionary containing the words *word* and *embedding* might be represented as $[0\ 1]$ and $[1\ 0]$, respectively. You may assume that the dictionary used contains all words in both the training and testing sets.
The second neural network, instead, uses a pre-trained embedding of the dictionary that you may assume represents every word in the dictionary.

Assuming that both networks use $\tanh$ activations and have randomly initialized weights, and they are trained using the same training set.

Now, at test time, each network is presented with a sequence of words not seen during training. Which of the following statement(s) is/are true about the output of the network for this sequence?

training set里面没有的

☑ The first network will produce a random output ✓

□ The second network will produce a random output

☑ The second network has a fighting chance at classifying the sequence ✓

□ The first network has a fighting chance at classifying the sequence

✓

Solution:

The first network will produce a random output. As a result of the words not being seen during training, the corresponding input units always had an output of zero, which means that no gradient was backpropagated into the weights and the randomly-initialized weights remain unchanged.

The second network, however, has a fighting chance at classifying the sequence. Since the embedding process causes words that appear in similar contexts to have similar locations in vector space, the unseen words will likely fall into a region that contains words that the network **has** seen and knows how to correctly classify.

Submit You have used 1 of 2 attempts

ℹ Answers are displayed within the problem

Discussion

Show Discussion

Topic: Unit 3 Neural networks (2.5 weeks):Homework 4 / 4. Word Embeddings