Programming Assignment 2: Learning Word Representations.

22/31 points earned (70%)

You haven't passed yet. You need at least 80% to pass. Review the material and try again! You have 3 attempts every 8 hours.

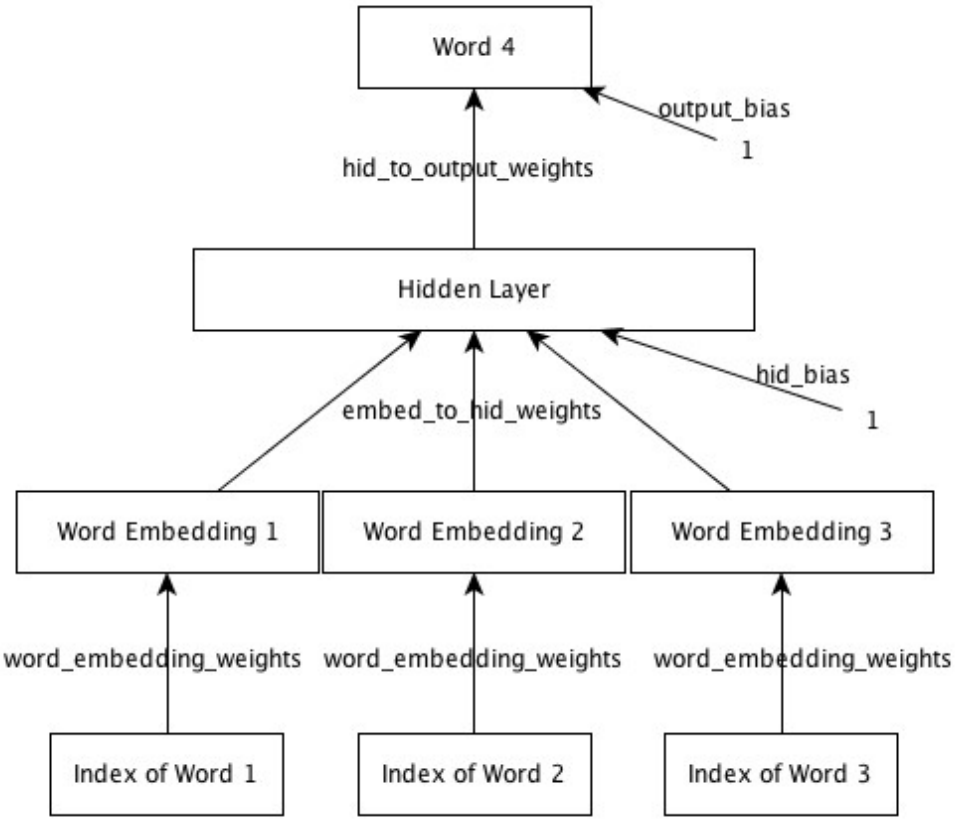
Review Related Lesson

1 / 1 points

1.

We are now ready to start using neural nets for solving real problems!

In this assignment we will design a neural net language model. The model will learn to predict the next word given the previous three words. The network looks like this:



To get started, download any one of the following archives.

assignment2.tar.gz

Or

assignment2.zip

Or each file individually:

- README.txt
- train.m
- raw_sentences.txt

0 / 4 points

2.

Train a model with 50 dimensional embedding space, 200 dimensional hidden layer and default setting of all other hyperparameters. What is average training set cross entropy as reported by the training program after 10 epochs ? Please provide a numeric answer (three decimal places). [4 points]

25.435

Incorrect Response

3 / 3 points

3.

Train a model for 10 epochs with a 50 dimensional embedding space, 200 dimensional hidden layer, a learning rate of 0.0001 and default setting of all other hyperparameters. What do you observe ? [3 points]

- [x] Cross Entropy on the training and validation set decreases very slowly.

Correct

- [ ] Cross Entropy on the training and validation set decreases very rapidly.

Un-selected is correct

- [ ] Cross Entropy on the validation set fluctuates wildly and eventually diverges.

Un-selected is correct

- [ ] Cross Entropy on the training set fluctuates wildly and eventually diverges.

Un-selected is correct

---

3 / 3 points

4.

If all weights and biases in this network were set to zero and no training is performed, what will be the average cross entropy on the training set ? Please provide a numeric answer (three decimal places). [3 points]

5.522

Correct Response

If all weights and biases are zero, the output distribution will be uniform for all inputs. The entropy will then be $\log_e(n)$ where n is the number of words in the vocabulary. In this case it will $\log_e(250)$

1 / 1 points

5.

Train three models each with 50 dimensional embedding space, 200 dimensional hidden layer.

- Model A: Learning rate = 0.001,
- Model B: Learning rate = 0.1
- Model C: Learning rate = 10.0.

Use a momentum of 0.5 and default settings for all other hyperparameters. Which model gives the lowest training set cross entropy after 1 epoch ? [3 points]

- Model A
- Model B
- Model C

Correct

---

0 / 2
points

6.
In the models trained in Question 5, which one gives the lowest training set cross entropy after 10 epochs ? [2 points]

- Model C

  This should not be selected

- Model A
- Model B

3 / 3 points

7.

Train each of following models:

- Model A: 5 dimensional embedding, 100 dimensional hidden layer
- Model B: 50 dimensional embedding, 10 dimensional hidden layer
- Model C: 50 dimensional embedding, 200 dimensional hidden layer
- Model D: 100 dimensional embedding, 5 dimensional hidden layer

Use default values for all other hyperparameters.

Which model gives the best training set cross entropy after 10 epochs of training ? [3 points]

- Model A
- Model C

Correct

- Model D
- Model B

2 / 2 points

8.
In the models trained in Question 7, which one gives the best validation set cross entropy after 10 epochs of training ? [2 points]

- [x] Model C

Correct

- [ ] Model B
- [ ] Model D
- [ ] Model A

---

0 / 3 points

9.
Train three models each with 50 dimensional embedding space, 200 dimensional hidden layer.

- Model A: Momentum = 0.0
- Model B: Momentum = 0.5
- Model C: Momentum = 0.9

Use the default settings for all other hyperparameters. Which model gives the lowest training set cross entropy after 5 epochs ? [3 points]

- [ ] Model A
- [x] Model B

This should not be selected

- [ ] Model C

---

2 / 2 points

10.

Train a model with 50 dimensional embedding layer and 200 dimensional hidden layer for 10 epochs. Use default values for all other hyperparameters.

Which words are among the 10 closest words to the word 'day'. [2 points]

- [ ] 'today'

Un-selected is correct

- [ ] 'during'

Un-selected is correct

- [ ] 'week'

Correct

- [ ] 'year'

Correct

2 / 2 points

11.
In the model trained in Question 10, why is the word 'percent' close to 'dr.' even though they have very different contexts and are not expected to be close in word embedding space? [2 points]

- Both words occur too frequently.
- We trained the model with too large a learning rate.
- The model is not capable of separating them in embedding space, even if it got a much larger training set.
- Both words occur very rarely, so their embedding weights get updated very few times and remain close to their initialization.

Correct

2 / 2 points

12.
In the model trained in Question 10, why is 'he' close to 'she' even though they refer to completely different genders? [2 points]

- Both words occur very rarely, so their embedding weights get updated very few times and remain close to their initialization.
- They differ by only one letter.
- They often occur close by in sentences.
- The model does not care about gender. It puts them close because if 'he' occurs in a 4-gram, it is very likely that substituting it by 'she' will also make a sensible 4-gram.

Correct

3 / 3
points

13.
In conclusion, what kind of words does the model put close to each other in embedding space. Choose the **most** appropriate answer. [3 points]

- [x] Words that can be substituted for one another and still make up a sensible 4-gram.

Correct

- [ ] Words that occur close in an alphabetical sort.
- [ ] Words that belong to similar topics. A topic is a semantic categorization (like 'sports', 'art', 'business', 'computers' etc).
- [ ] Words that occur close to each other (within three words to the left or right) in many sentences.