Based on the above data, answer the given subquestions.

Consider the CBOW model for learning word embeddings with embedding dimension two. The window size is one, that is, we only use the previous word as the context to predict the current word. The vocabulary is made up of the words {"one", "two", "three", "four"}. At some point during training, the word and context matrices are given below. The first column in each matrix corresponds to the embedding for "one", the second column is for "two" and so on:

$$W_{\text{word}} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ -1 & 1 & 2 & -2 \end{bmatrix}, \quad W_{\text{context}} = \begin{bmatrix} 0 & 1 & 0 & 1 \\ -1 & 0 & 1 & 1 \end{bmatrix}$$

The sample that has come up now during training is "two three". Note that "three" is to be considered as the true label here. Enter your answer correct to two places after the decimal for both sub-questions.

Find the probability of predicting "three" given "two" as context.

Answer (Numeric):

Answer

Accepted Answer : 0.2 to 0.3

Your score : 0

Using the given sample, perform one update of gradient descent using $\eta = 1$ for the word embedding corresponding to "three". If the updated word embedding of "three" is the vector (a, b), enter $b - a$ as the answer. Note that we are interested in the word embedding of "three" and not its context embedding.

Answer (Numeric):

Answer

Accepted Answer : 0.2 to 0.3

0.25

Suppose you have a vocabulary of 5,000 unique words and you want to train a CBOW model with a window size of 3 (on each side) and with an embedding dimension of 100. How many parameters (weights) will the embedding layer have?

Answer (Numeric):

Answer

Accepted Answer : 500000

Your score : 0

5,000

Assume that your CBOW model outputs a probability distribution over a vocabulary of 20,000 words for a given context. If the correct target word is word number 150, and the model's predicted probability for this word is 0.2, calculate the cross-entropy loss for this prediction. (use natural log)

$\hat{y}_t$

$-\ln(0.2)$

1.6

Answer (Numeric):

Answer

Accepted Answer : 1.57 to 1.63

1.6

Your score : 0

In a Skip-gram model with a window size of 2 (on each side), how many **unique** pairs of target and context words will be generated for the following sentence: '**Predicting the future is not magic, it is artificial intelligence**'

Note: Ignore the punctuations in your calculation

Answer (Numeric):

Answer

Accepted Answer : 32

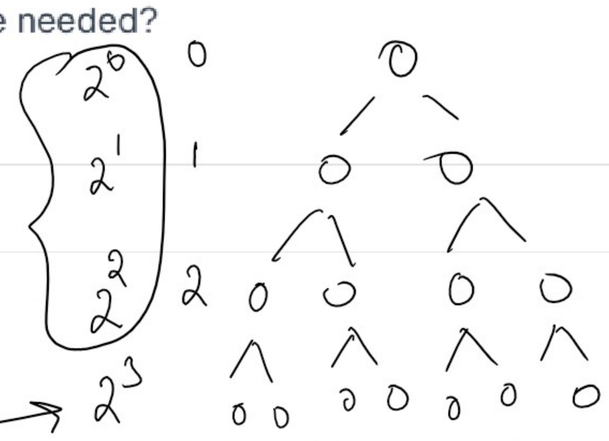
Your score : 0

If you use hierarchical softmax with a binary tree where each leaf node represents a word in the vocabulary, and the vocabulary size (V) is 16000, how many binary classifiers are needed?

Answer (Numeric):

Answer

Accepted Answer : 15998 to 16001



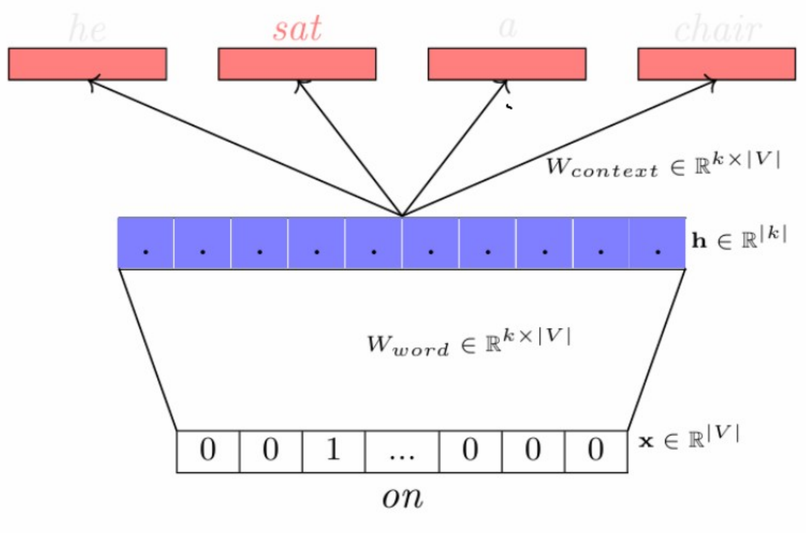
In a Skip-gram model with a vocabulary size V = 100, an embedding dimension D = 10, and a window size of 3 (on each side), using negative sampling with 5 negative samples per positive sample, what is the total number of parameters in the model?

Answer (Numeric):

Answer

Accepted Answer : 2000

Your score : 0



Given the following probabilities for a word pair (w_i, w_j):

- $p(w_i, w_j) = 0.02$
- $p(w_i) = 0.2$
- $p(w_j) = 0.3$

Calculate the PMI and PPMI for (w_i, w_j).

OPTIONS :

- -0.477, 0
- -0.301, 0
- 0.301, 0
- 0.477, 0.477
- -0.477, 0.477

Your score : 0

Given a matrix A with dimensions p × q, which of the following statements is NOT true regarding the rank-k approximation of A obtained through Singular Value Decomposition (SVD)?

OPTIONS :

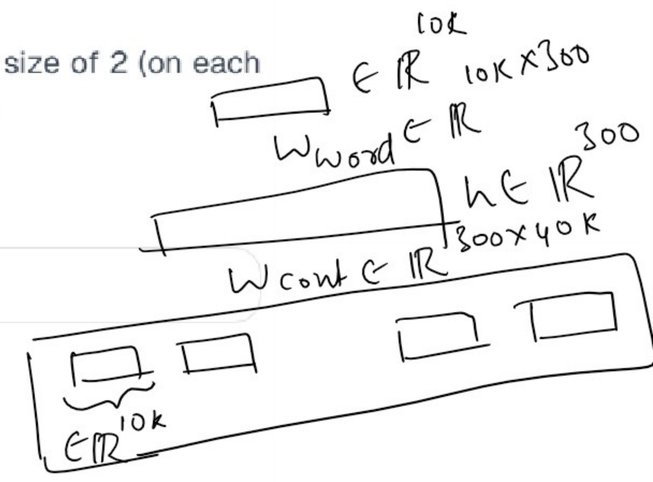
- The rank-k approximation matrix will have dimensions $p \times q$.
- The matrix U_k in the rank-k approximation has dimensions $p \times k$.
- The matrix Σ_k in the rank-k approximation has dimensions $k \times k$.
- The rank-k approximation matrix will have dimensions $k \times k$.

Suppose you have a vocabulary of 10,000 unique words and you want to train a CBOW model with a window size of 2 (on each side) and with an embedding dimension of 300. How many parameters (weights) will the embedding layer have?

Answer (Numeric):

Answer

Accepted Answer : 3000000



Assume that your CBOW model outputs a probability distribution over a vocabulary of 20,000 words for a given context. If the correct target word is word number 150, and the model's predicted probability for this word is 0.02, calculate the cross-entropy loss for this prediction.

$-\ln 0.02$

Answer (Numeric):

Answer

Accepted Answer : 3.9 to 4.0

Your score : 0

3.91

In a Skip-gram model with a window size of 2 (on each side), how many unique pairs of target and context words will be generated for the following sentence: 'The idea is to die young as late as possible'

2 3 4 4 4 4 4 3 1 2

Answer (Numeric):

Answer

Accepted Answer : 31

In a Skip-gram model with a vocabulary of 15000 words, if you choose a negative sampling rate of 10 negative samples for each positive sample, how many total samples will be generated for a sentence with 8 unique words and a window size of 3 (on each side)?

w_1 w_2 w_3 w_4 w_5 w_6 w_7 w_8

3 4 5 6 6 5 4 3

Answer (Numeric):

Answer

Accepted Answer : 396

36
36
396

$(C, w_1 w_2 \ldots w_{10})$ — 11

$(C_{36}, w_1, w_2 \ldots w_{10})$

36

10+8+18

$= 36 \times 10 + 36$

360

(Sat, on), $(Sat, ox), (Sat, magic) \ldots$ — wrong

C

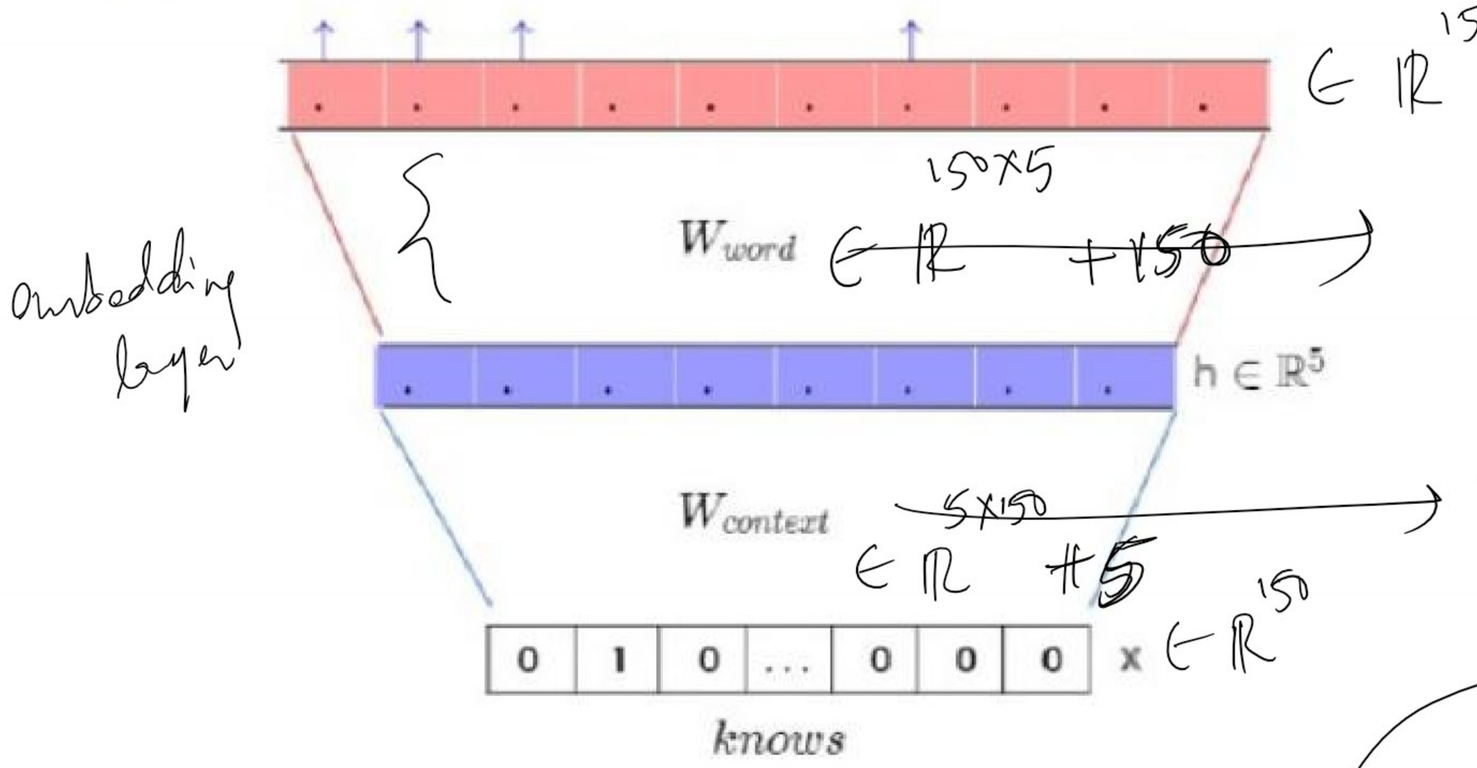
36 × 11

- $D = [(\text{sat}, \text{on}), (\text{sat}, \text{a}), (\text{sat}, \text{chair}), (\text{on}, \text{a}), (\text{on,chair}), (\text{a,chair}), (\text{on,sat}), (\text{a}, \text{sat}), (\text{chair,sat}), (\text{a}, \text{on}), (\text{chair}, \text{on}), (\text{chair}, \text{a})]$
- $D' = [(\text{sat}, \text{oxygen}), (\text{sat}, \text{magic}), (\text{chair}, \text{sad}), (\text{chair}, \text{walking})]$

$D \times (b' + 1)$

Based on the above data, answer the given subquestions.

Suppose we use the CBOW (Continuous Bag of words) model shown below to find a distributed vector representation for all the words in the vocabulary. The size of the vocabulary $|V|$ is 150 and all the words in the vocabulary are one hot encoded (as a **column vector**) and fed as input to the network. The output layer uses softmax to produce the probability score for each word given the context. Here, $\mathbf{W}_{word}$ and $\mathbf{W}_{context}$ are weight matrices.



How many number of parameters are there in the network (excluding bias)?

Answer (Numeric):

Answer

Accepted Answer : 1500

Your score : 0

The word representation for the 2-nd word in the vocabulary corresponds to

OPTIONS :

- ○ 2-nd column of $\mathbf{W}_{\mathbf{context}}$
- ○ 2-nd row of $\mathbf{W}_{\mathbf{context}}$
- ○ 2-nd column of $\mathbf{W}_{\mathbf{word}}$
- ○ 2-nd row of $\mathbf{W}_{\mathbf{word}}$

Your score : 0

Based on the above data, answer the given subquestions.

Construct a vocabulary V from the following text corpus

1 2 3 4 5 6 7

- How much wood could a woodchuck chuck
- If a woodchuck could chuck wood
- as much wood as a woodchuck could chuck

8 9

Your score : 0

What is the size of the vocabulary |V |?

Answer (Numeric):

Answer

Accepted Answer : 9

Your score : 0

Suppose we consider the three words (wood,woodchuck, much). Assume we use one-hot encoded vector representation for all these words. The statement that,"The cosine similarity between the pair (wood,woodchuck) is greater than the pair (wood,much)" is

0

equal to both 0

$\sqrt{2}$ euclidian distance

OPTIONS :

- ◯ True
- ◯ False

Your score : 0

Consider a sentence inside the quote "I may be wrong, and you may be right, and by an effort, we may get nearer to the truth"

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16

Based on the above data, answer the given subquestions.

Your score : 0

What is the size of the vocabulary, $|V|$?

Answer (Numeric):

Answer

Accepted Answer : 16

Your score : 0

Suppose all words in the vocabulary are represented using one-hot-encoded vector of size $|V|$. Then compute the ordered pair-wise (that is, Cartesian product of $V \times V$) cosine similarity between word representations and enter their sum.

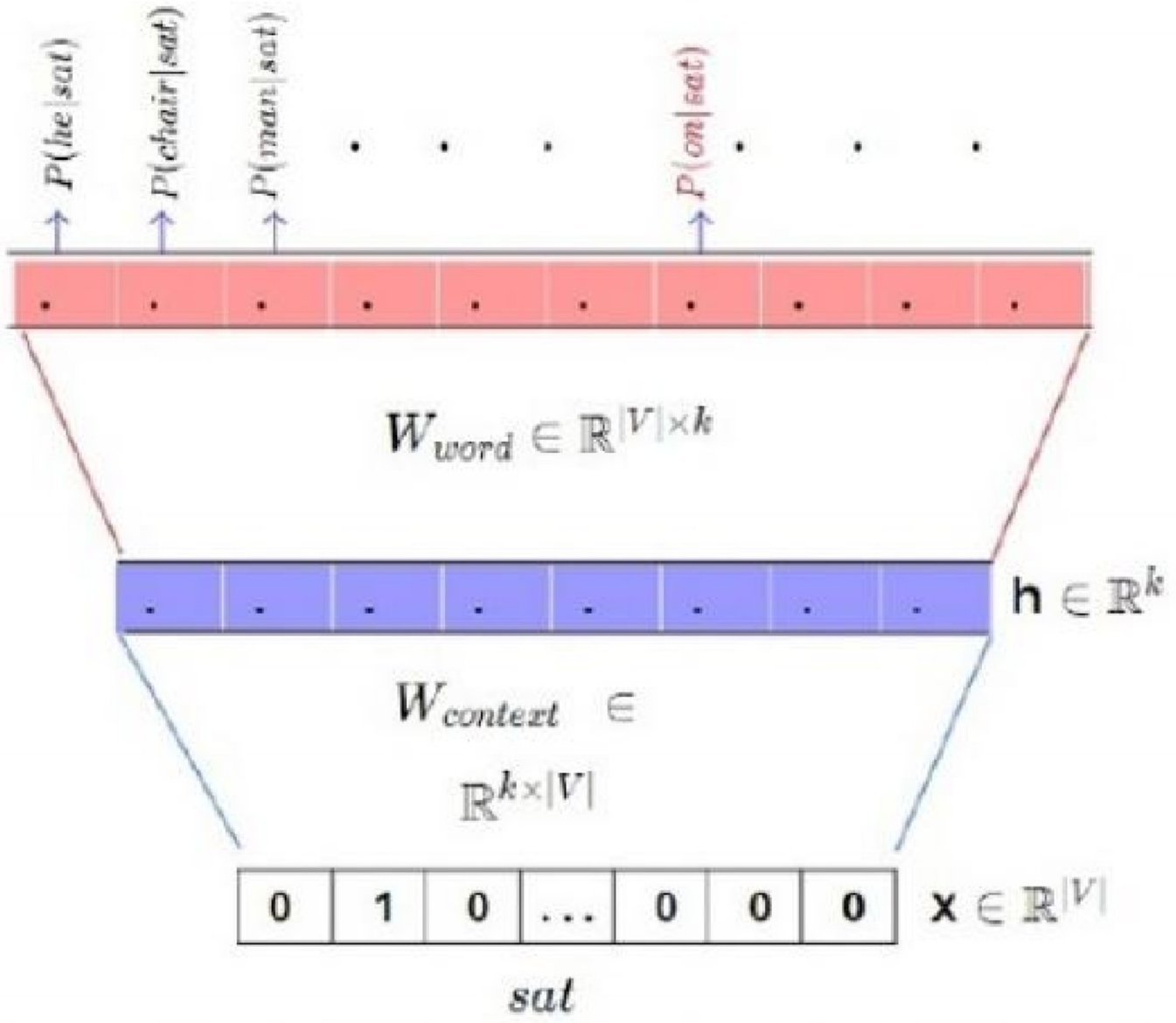
Answer (Numeric):

Answer

Accepted Answer : 16

Your score : 0

Consider a model shown below that learns the distributed vector representation of words by learning to predict the target word v_w given the context word u_c. Here, v_w and u_c are the vector representation of target word at index w of the output vocabulary and context word at index c of the input vocabulary, respectively.

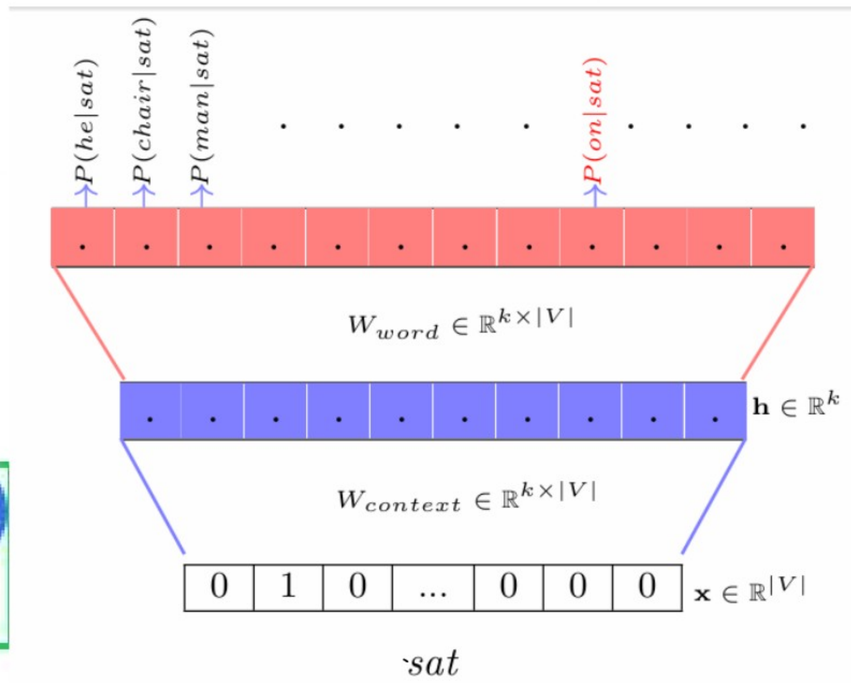


In the diagram, $|V|$ denotes the size of the vocabulary, $W_{context}$ and W_{word} are learnable parameters. The vector representation of all context words are arranged as columns of $W_{context}$ and the vector representation for all target words are arranged as row vectors in W_{word}. The parameters are initialized randomly. The input x is one-hot-representation of a word in the input vocabulary. Assume that the size of both input and output vocabulary are equal.

Suppose the input is x (one hot representation of context word) and the corresponding label is y (one hot representation of target word). The quantities $h, u_c, \hat{y}$ are computed as follows,

$$h = u_c = W_{context}\ x, \quad z = W_{word}\ u_c \quad \hat{y} = softmax(z)$$

Choose the expression that the model has to minimize using cross entropy loss (Assume natural logarithm where required).



OPTIONS :

- $-v_w u_c + \log\left(\sum_{w' \in V} exp(v_{w'} u_c)\right)$
- $-u_c v_w^T + \log\left(\sum_{w' \in V} exp(u_c v_{w'}^T)\right)$
- $-u_c v_w^T - \log\left(\sum_{w' \in V} exp(u_c v_{w'}^T)\right)$
- $-v_w^T u_c - \log\left(\sum_{w' \in V} exp(v_{w'}^T u_c)\right)$

Suppose we compute the gradients and update the representations with $\eta = 1$. Choose the correct statement(s)

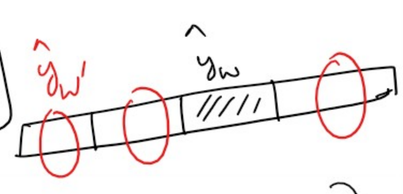
OPTIONS :

- [x] Suppose the model predicts target word v_w with probability score of 1 (that is, $\hat{y_w} = 1$). Then no elements in v_w will get modified after one iteration (i.e, parameter update).
- [ ] Suppose the model predicts target word v_w with probability score of 1 (that is, $\hat{y_w} = 1$). Then no elements in $v_{w'}, (w' \neq w)$ will get modified after one iteration (i.e, parameter update).
- [ ] Suppose the model predicts target word v_w with probability score of 0.5 (that is, $\hat{y_w} = 0.5$). Then the elements of v_w will be modified as $v_w = v_w + 0.5u_c^T$
- [ ] Suppose the model predicts target word v_w with probability score of 0.5 (that is, $\hat{y_w} = 0.5$). Then the elements of $v_{w'}$ will be modified as $v_{w'} = v_{w'} - \hat{y}_{w'} u_c^T$
- [ ] Suppose the model predicts target word v_w with probability score of 0.5 (that is, $\hat{y_w} = 0.5$). Then the elements of $v_{w'}$ will be modified as $v_{w'} = v_{w'} + \hat{y}_{w'} u_c^T$

c

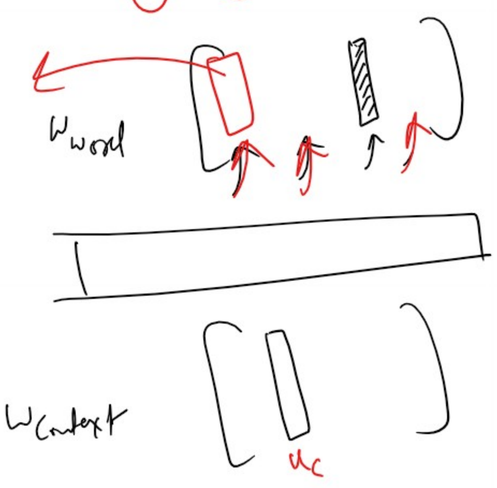
$V_{w'} = V_{w'} - \hat{y}_{w'}$

$V_w = V_w + \eta(1-\hat{y})u$



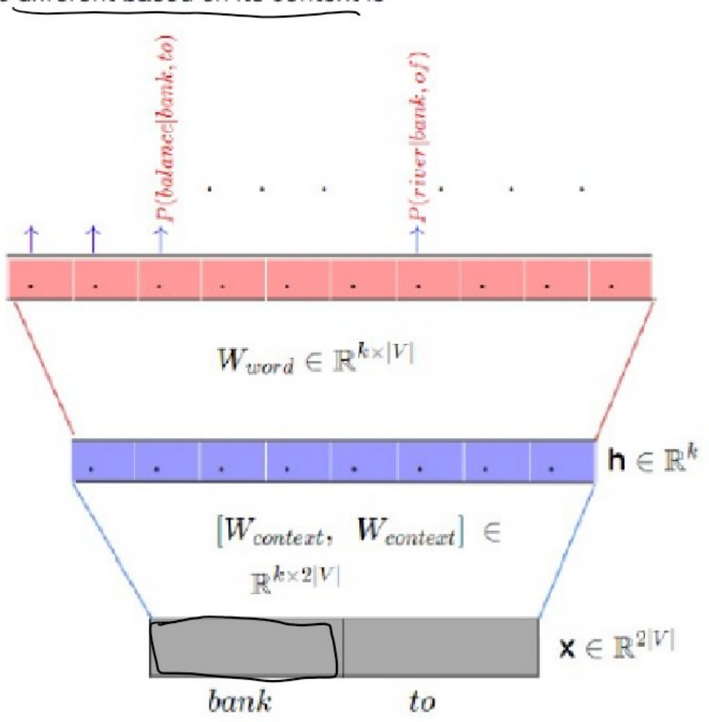
$V_{w'}$

$p("3"|"2")$



$V_{w'} = V_{w'} - \hat{y}_{w'} u_c^T$

Consider the following two sentences • A man was sitting at the bank of the river and gazing at stars in the sky • A man went to the bank to check his current balance Suppose we get the word representation for the word **bank** in both sentences using CBOW model which was trained as shown in the image below. The model was trained by building a vocabulary that contains unique words in the sentences. Then the statement that the word representation for the word **bank** will be different based on its context is

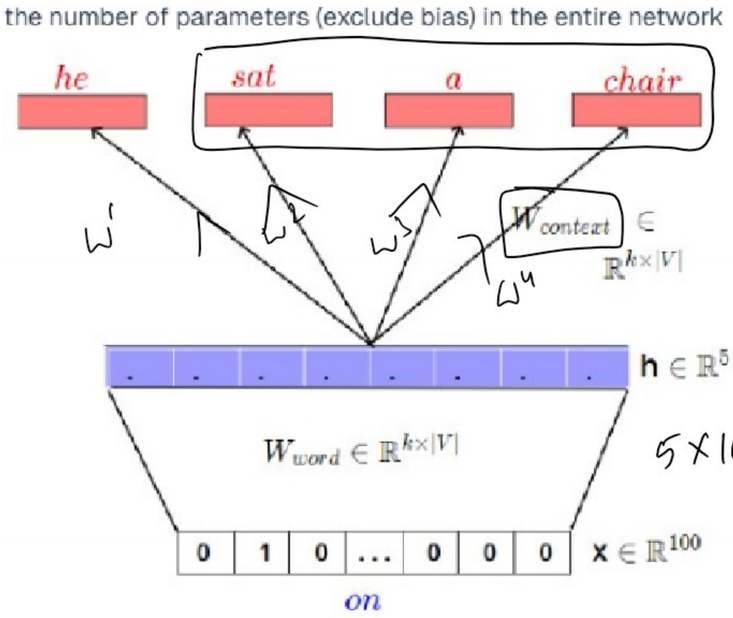


OPTIONS :

- ○ TRUE
- ✓ FALSE
- ○ Insufficient information

Your score : 0

Consider a skip-gram model shown below. Each word in the vocabulary is represented as one-hot vector of size 100×1. The embedding dimension h is 5×1. Enter the number of parameters (exclude bias) in the entire network



500 × 4 = 2000

5 × 100

500

Answer (Numeric):

Answer

Accepted Answer : 2500

A text corpus contained the following two sentences. • In science if you know what you are doing you should not be doing it • In engineering if you do not know what you are doing you should not be doing it Based on the above data, answer the given subquestions.

Your score : 0

Build a vocabulary V and enter its size $|V|$

Answer (Numeric):

Answer

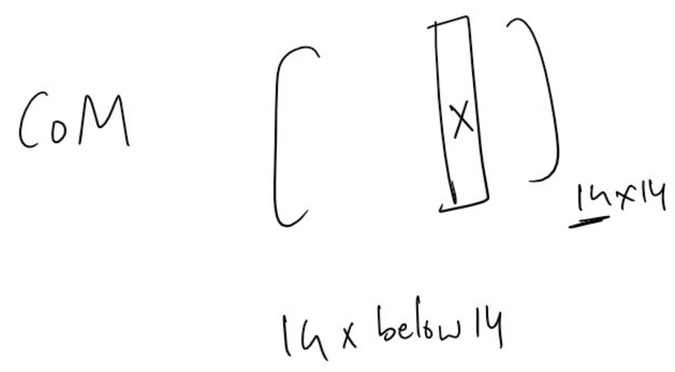
Accepted Answer : 14

Your score : 0

Suppose we build a co-occurrence matrix of size m × n, where each row corresponds to a word in the vocabulary and the columns corresponds to the context of the word. Which of the following could be a valid size of the co-occurrence matrix (select all correct answers)?

OPTIONS :

- ☐ 32 × 32
- ☐ 32 × 14
- ☐ 14 × 32
- ☑ 14 × 14
- ☑ 14 × 13
- ☐ 13 × 14
- ☑ 14 × 7
- ☐ 7 × 14



Your score : 0