Assignment 3: Word Embedding
CSPE73: Natural Language Processing

Due Date: 20/11/2022 **Max Marks: 10**

In this assignment, you will explore a classic way of generating word embeddings or representations.

- You will implement a famous model called the continuous bag of words (CBOW) model.

By completing this assignment you will:

- Train word vectors from scratch.
- Learn how to create batches of data.
- Understand how backpropagation works.
- Plot and visualize your learned word vectors.

Knowing how to train these models will give you a better understanding of word vectors, which are building blocks to many applications in natural language processing.

The Continuous bag of words model
Let's take a look at the following sentence:

'**I am happy because I am learning**'.

- In continuous bag of words (CBOW) modeling, we try to predict the center word given a few context words (the words around the center word).
- For example, if you were to choose a context half-size of say C=2, then you would try to predict the word **happy** given the context that includes 2 words before and 2 words after the center word:

C words before: [I, am]

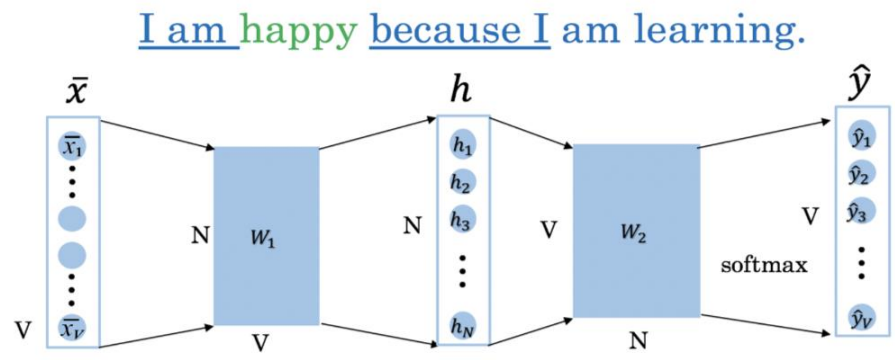
C words after: [because, I]

- In other words:

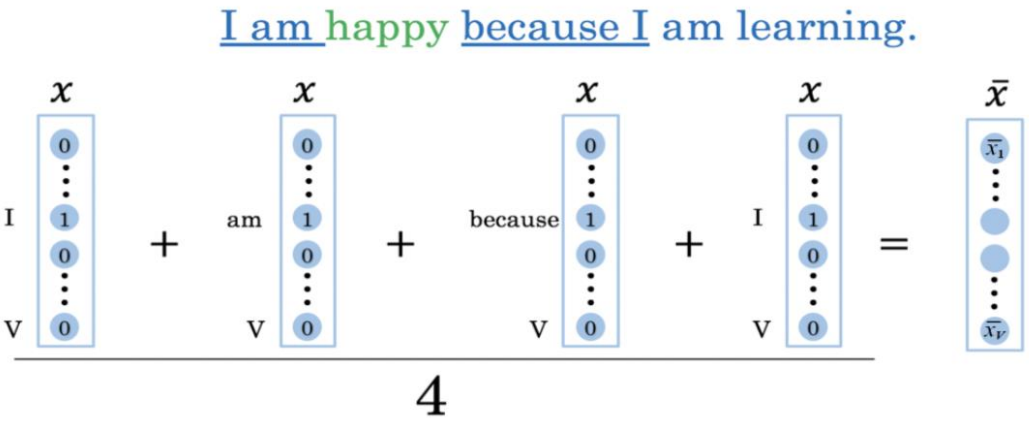
$$context = [I, am, because, I]$$

$$target = happy$$

The structure of your model will look like this:



Where $\bar{x}$ is the average of all the one hot vectors of the context words.



Once you have encoded all the context words, you can use $\bar{x}$ as the input to your model.

The architecture you will be implementing is as follows: (equation 1 to equation 4)

$$h = W_1 X + b_1$$
$$a = ReLU(h)$$
$$z = W_2 a + b_2$$
$$\hat{y} = softmax(z)$$

2 Training the Model

Exercise 01

Initializing the model:

You will now initialize two matrices and two vectors.

- The first matrix (W1) is of dimension N×V, where V is the number of words in your vocabulary and N is the dimension of your word vector.
- The second matrix (W2) is of dimension V×N.
- Vector b1 has dimensions N×1

- Vector b2 has dimensions V×1.
- b1 and b2 are the bias vectors of the linear layers from matrices W1 and W2.

2.1 Softmax

Before we can start training the model, we need to implement the softmax function as defined in equation 5:

$$\text{softmax}(z_i) = \frac{e^{z_i}}{\sum_{i=0}^{V-1} e^{z_i}} \tag{5}$$

- Array indexing in code starts at 0.
- V is the number of words in the vocabulary (which is also the number of rows of z).
- i goes from 0 to |V| - 1.

Exercise 02

Instructions: Implement the softmax function below.

- Assume that the input z to softmax is a 2D array
- Each training example is represented by a column of shape (V, 1) in this 2D array.
- There may be more than one column, in the 2D array, because you can put in a batch of examples to increase efficiency. Let's call the batch size lowercase m, so the z array has shape (V, m)
- When taking the sum from i=1⋯V−1, take the sum for each column (each example) separately.

2.2 Forward propagation

Exercise 03

Implement the forward propagation z according to equations (1) to (3).

For that, you will use as activation the Rectified Linear Unit (ReLU) given by:

$$h = W_1 X + b_1 \tag{1}$$
$$a = ReLU(h) \tag{2}$$
$$z = W_2 h + b_2 \tag{3}$$

$$f(h) = \max(0, h) \tag{6}$$

2.3 Cost function

Exercise 04

Implement the cross entropy cost function

2.4 Training the Model - Backpropagation

Exercise 05
Now that you have understood how the CBOW model works, you will train it.
You created a function for the forward propagation. Now you will implement a function that computes the gradients to backpropagate the errors.

Gradient Descent

Exercise 06
Now that you have implemented a function to compute the gradients, you will implement batch gradient descent over your training set.

3.0 Visualizing the word vectors

In this part you will visualize the word vectors trained using the function you just coded above.

The input file if from Penn Treebank Dataset

Deliverables
1. Code in one folder
2. A short report discussing the implementation of each exercise and the final result.
    a) Change context size (C=1,2,3,4)and analyze your results