NAME: KSHITIJ VINOD SALI
CLASS: BE-A ROLL NO.:
SUBJECT: LP-IV: DEEP LEARNING

LAB ASSIGNMENT - 05

Title: Implementation of the Continuous Bag of Words (CBOW) Model.

Course Outcome: CO3- Apply deep learning techniques like CNN, RNN auto encoders to solve real world problems.

Date of Completion:
Assessment Grade/Marks:
Assessor's Sign with Date:

Problem Statement: Implementation of Continuous Bag of Words (CBOW) model. Stage can be:
a) Data preperation
b) Generate training data
c) Train model
d) Output

Blooms Taxanomy Category: Create

Requirements: Python framework (keras), Python libraries (NumPy, Gensim), Dataset.

Theory:

i) What is NLP?

→ Natural Language Processing (NLP) is a field of artificial intelligence (AI) that enables computers, to understand, interpret, generate and interact with human language in a meaningful way. Its goal is to bridge the gap between human communication & computer understanding

ii) What is word embedding related to NLP?

→ Word Embedding is a ~~teq~~ technique in NLP where words/phrases from vocabulary are represented as dense numerical vectors in a multi-dimensional space. The key idea is that words with similar meanings will have similar vector representations, allowing models to capture semantic relationships.

iii) Explain Word2Vec techniques.

→ Word2Vec consists of two main neural networks architectures:

a) Continuous Bag-of-Words (CBOW) - Predicts targeted word based on surrounding context words.

b) Skip-gram - Uses targeted word to predicts its surrounding context words.

iv) Enlist applications of Word embedding in NLP.

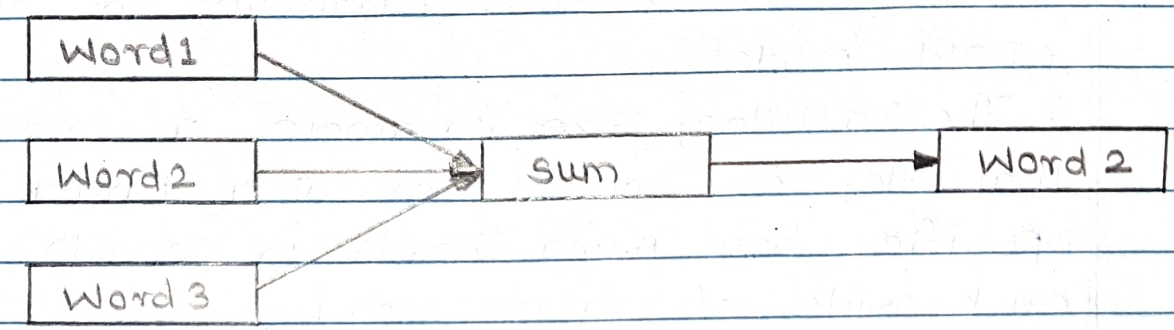
→ a) Sentiment analysis.

b) Machine translation.

c) Text classification and categorization
d) Named entity recognization
e) Question-answering systems.

v) Explain CBOW architecture.

→



The CBOW architecture is a predictive model. It takes a set of context words as input, averages their vector representations & feeds this average into neural network. The network's goal is to predict original target word from this context.

vi) What will be input to CBOW model & output to CBOW model?

→ Input - A set of context words surrounding a target word.
Output - The target word that was originally in the middle of the context.

vii) What is tokenizer?

→ A tokenizer is a tool that breaks down a block of text into smaller units called

token. These tokens can be words, characters or sub-words. Tokenization is a crucial first step in any NLP pipeline, as it converts raw text into a format that helps a model can process.

viii) Explain window size parameter in detail for CBOW model.

→ The window size parameter in CBOW model that defines "context" for given target word. It specifies how many words to consider on each side of target word.

A smaller window captures more synthetic relationships. A larger window captures more broader, semantic relationships.

ix) Explain Embedding and Lambda layer from keras.

→ Embedding Layer- This keras layer is used to create word embeddings. It's essentially a lookup table that maps integer indices to dense floating-point vectors.

Lamba Layer- This layer allows you to apply any arbiratory function or expression as a layer within your keras model.

x) What is yeild()?

→ yeild() is a generator formation keyword in Python. Instead of returning a final result & existing, a generator function pauses at the yeild statement, send back a value & waits.

When called again, it resumes exactly where it left off.

Algorithm and Steps:

i) Import following libraries gemsim & numpy set i.e., text file created. It should be preprocessed
ii) Tokenize the every word from the paragraph. You can call in-build tokenizer present in Gensim
iii) Fit the data to tokenizer.
iv) Find total number of words & total number of sentences.
v) Generate the pairs of context words and target words.
vi) Create neural network model with following parameters: Model type, Layers, Compile options
vii) Create vector file of some word for testing.
viii) Assign weights to your trained model.
ix) Use the vector created in Gemsim
x) Choose the word to get similar type of words.

Inference: In this experiment, we saw what a CBOW model is & how is work. We also implemented the model on a custom dataset & got good output. We learnt what word embeddings are & how CBOW is useful. These can be used for text recognition, speech to text conversion etc.