PAGE No.
DATE / /

* Word Embeddings:

In NLP word embedding is a term used for the representation of words for text analysis, typically in the form of real-valued vector that encodes the meaning of the word such that the words that are closer in the vector space are expected to be similar in meaning.

Types

1) Frequency Based
   i) BoW
   ii) GLOVE
   iii) TFIDF

2) Prediction Based.
   i) Word 2 Vec.

* What is Word 2 Vec ?

→

→ Word embedding technique
→ 2013 google engineers developed
→ Converts word to vectors
→ Can find Semantic meaning of words.
→ Low dimension Vectors [100-300] for words
→ Dense vector [No Overfitting]
→ Deep learning Technique.

* Word 2 Vec Creates features based on Vocabullary

→ In neural network architecture we cannot able to see and understant the created features of architecture.

features	king	Queen	Man	Woman	Monkey
Gender	1	0	1	0	1
Wealth	1	1	0.2	0.2	0
Power	1	0.2	0.2	0.2	0
Weight	0.8	0.4	0.4	0.4	0.3
Speak	1	1	1	1	0

* The underlying ~~aspect~~ assumption of word2vec is that two words sharing similar contexts also share a similar meaning and consequently a similar vector representation from the model.

* Type of Word2VeC.

→
1) CBoW
2) Skip-gram

* 1) CBOW
(Continuous Bag of Words)

Creates dummy (fake) problem → Solve → Vectors (By product)

eg.

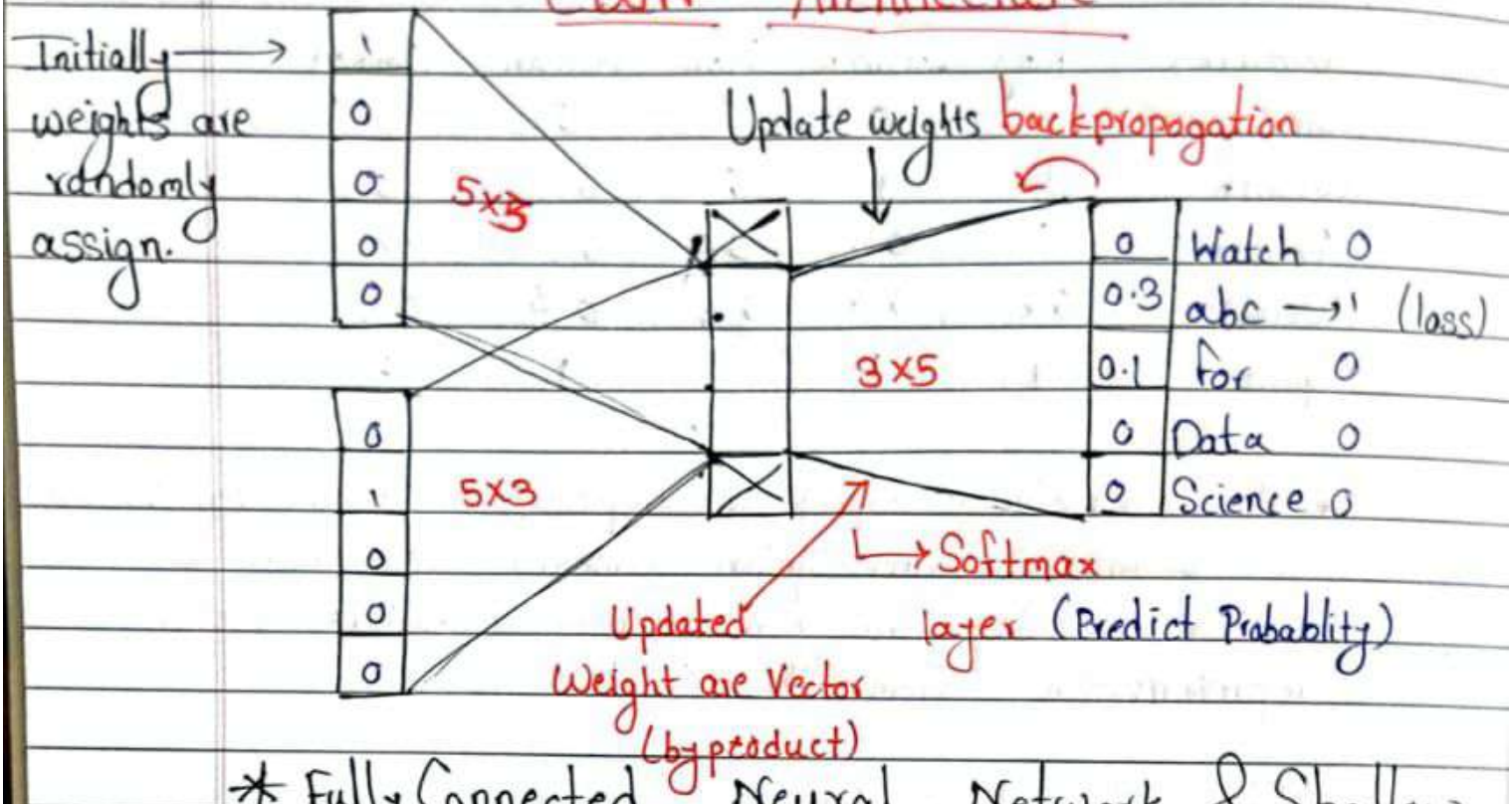
Watch	abc	for	Data	Science
↓ Context Word	↓ Target Word	↓ Context Word		

	x	y
	Watch, for	abc
	abc, Data	for
	for, Science	Data.

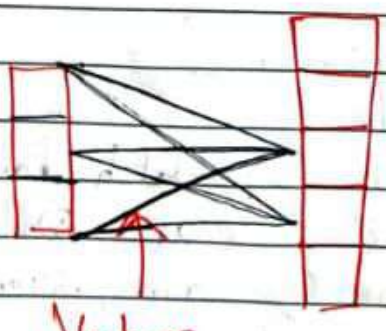
ONE

1	0	0	0	0	watch
0	1	0	0	0	abc
0	0	1	0	0	for
0	0	0	1	0	Data
0	0	0	0	1	Science

CBOW Architecture



* Fully Connected Neural Network & Shallow.
* If 1st attemp model Calculate loss between predicted and actual value.
* Based on that loss value, architecture performs back propogation and adjust weight in such a way resulting output meet the required output.
* The updated weights are the vectors for our vocabulary. (byproduct)



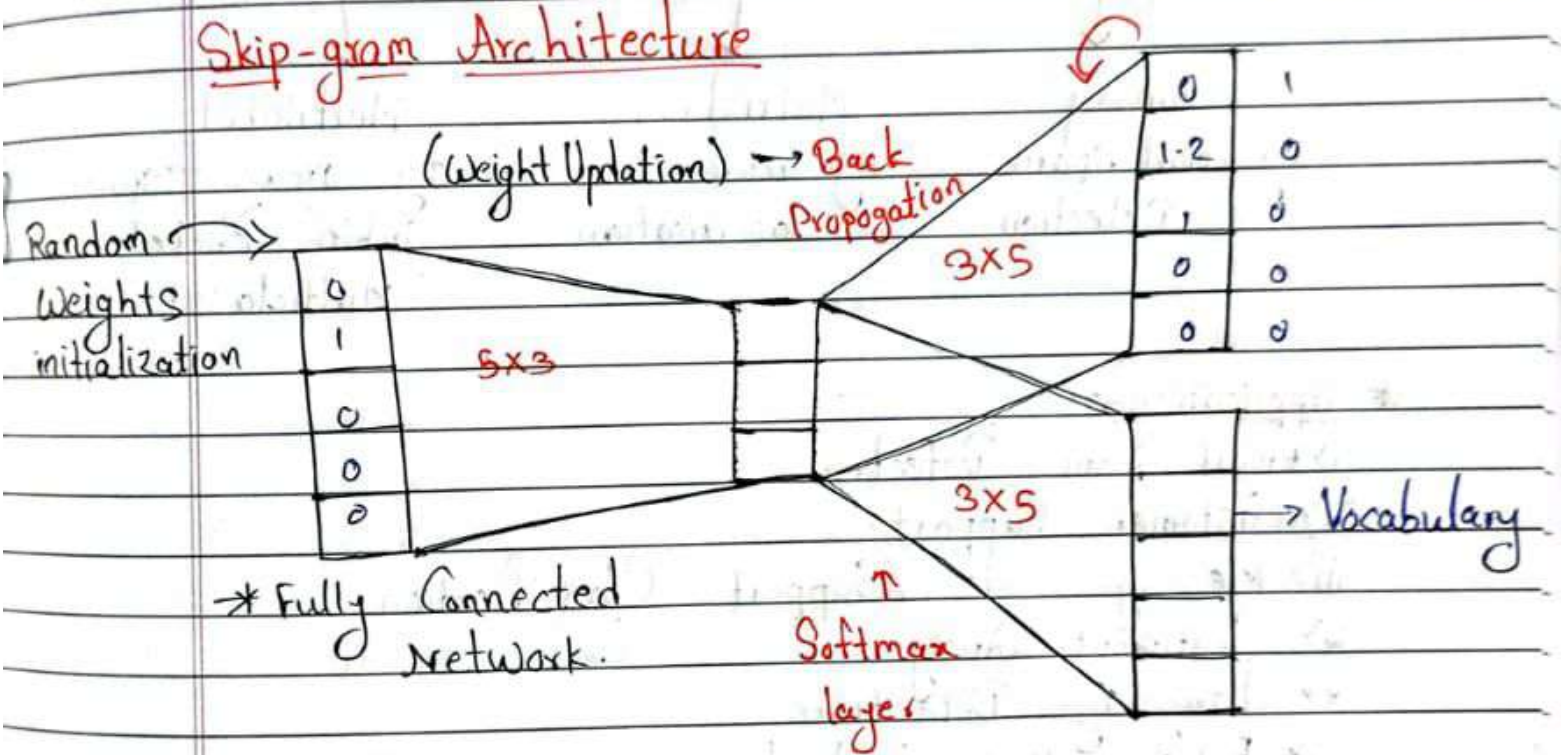
* Skip-gram

→ Works in Riverse manner of CBOW.
→ Target Value is given as input and Contextual text is output.

[Watch [bcd] for] Data Science , x y
↳ Window Size 3.

CampusX → watch for
for → CampusX, Data
Data → for, Science

Skip-gram Architecture



✓ * For Smaller Data → We use CBOW [Faster & Accurate]
✓ * For larger Data → We use Skip-gram [better results]

* To improve quality of Word2Vec Embeddings
→ Increase Training Data.
→ Increase dimension of Vectors [hidden layer nodes]
→ Increase Window Size.