TADA : Talk About Data Analytics

WORD2VEC : SKIP-GRAM
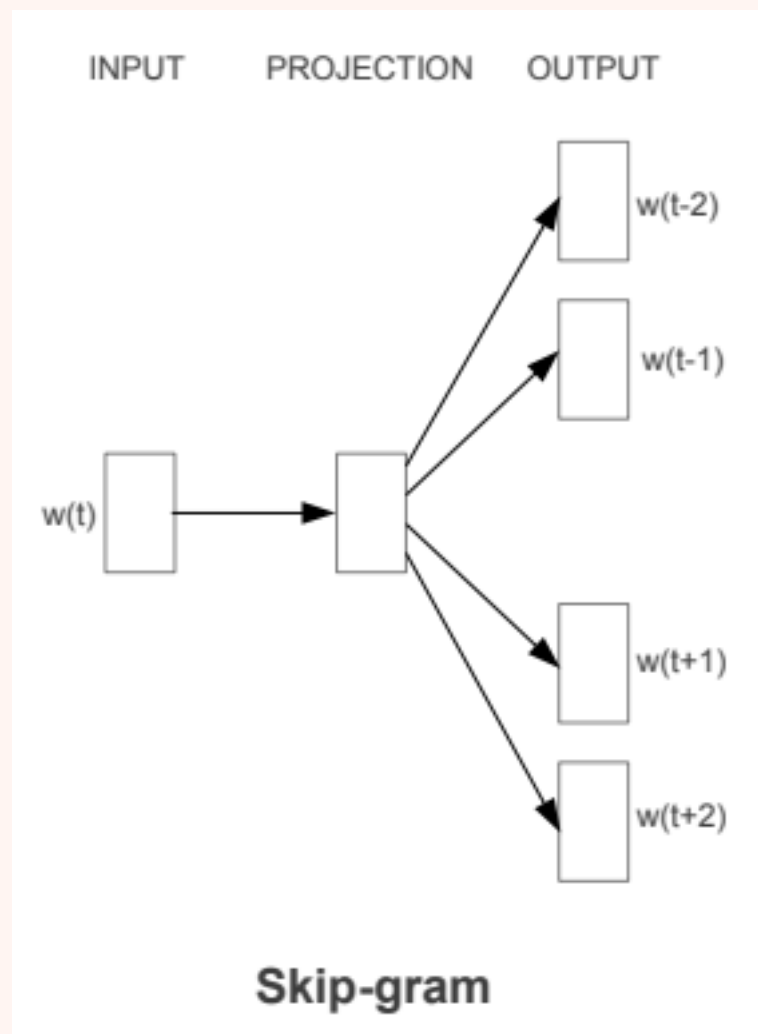
Hannah Do

Skip gram : Predicting the context word for a given target word

Given word : sat

Context words : cat, mat

ARCHITECTURE OF SKIP-GRAM

Skip-gram

> **w(t) : target word / given input**

> **One hidden layer** that performs the dot product between the weight matrix and input vector w(t)
>   ► *no activation function*

> **Output layer** that computes the dot product between the output vector of the hidden layer and the weight matrix of the output layer.
>   ► *softmax activation function* to compute probability of words appearing to be in the context of w(t)
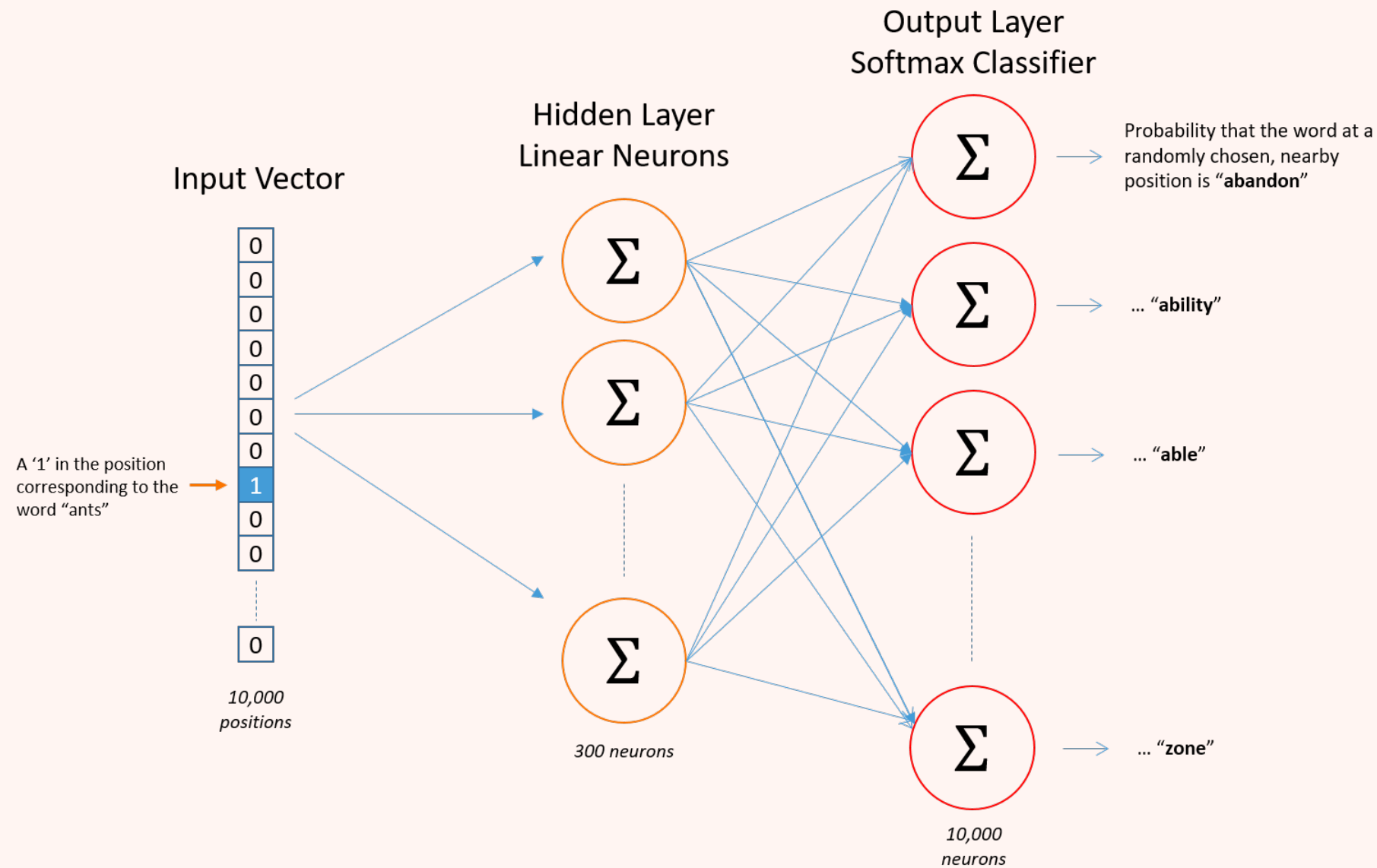
CREATING TRAINING SAMPLES

| The | quick | brown | fox jumps over the lazy dog. ⟹ | (the, quick) |
| | | | | (the, brown) |

The	quick	brown	fox	jumps over the lazy dog. ⟹	(quick, the)
					(quick, brown)
					(quick, fox)

The	quick	brown	fox	jumps	over the lazy dog. ⟹	(brown, the)
						(brown, quick)
						(brown, fox)
						(brown, jumps)

The	quick	brown	fox	jumps	over	the lazy dog. ⟹	(fox, quick)
							(fox, brown)
							(fox, jumps)
							(fox, over)

▶ The network learns the statistics from the frequency of each pairing

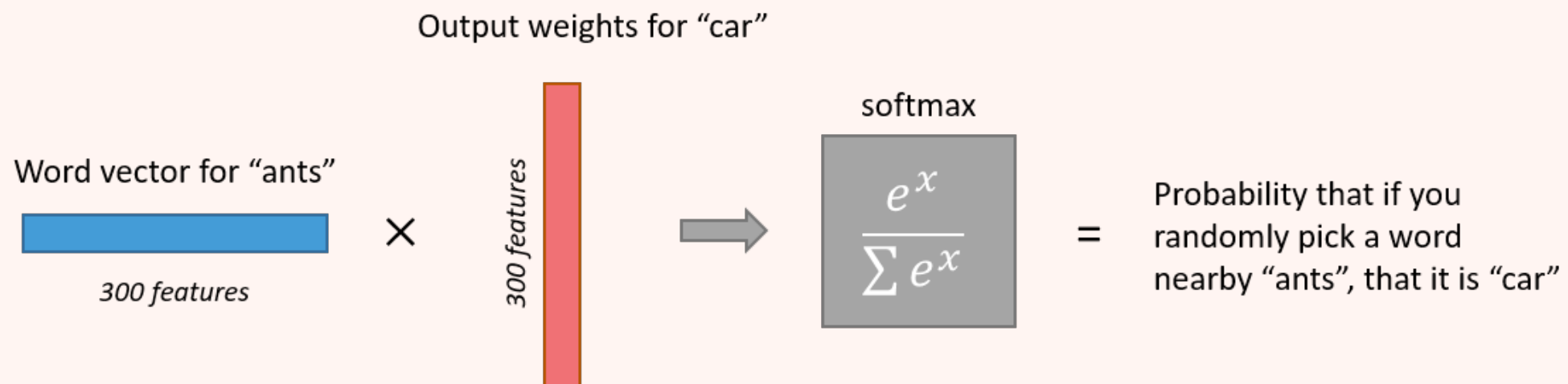▶ (*fox, jumps*) >> (*fox, crawls*)

INPUT VECTOR —> OUTPUT VECTOR

(WORDS REPRESENTED IN ONE HOT ENCODING)

(SINGLE VECTOR WITH PROBABILITY DISTRIBUTION)

Output Layer
Softmax Classifier

Hidden Layer
Linear Neurons

Input Vector

A '1' in the position corresponding to the word "ants"

Probability that the word at a randomly chosen, nearby position is "**abandon**"

... "**ability**"

... "**able**"

... "**zone**"

10,000 positions

300 neurons

10,000 neurons

WEIGHTS IN THE HIDDEN LAYER
= "WORD VECTOR" FOR THE INPUT WORD

$$[0 \quad 0 \quad 0 \quad 1 \quad 0] \times \begin{bmatrix} 17 & 24 & 1 \\ 23 & 5 & 7 \\ 4 & 6 & 13 \\ 10 & 12 & 19 \\ 11 & 18 & 25 \end{bmatrix} = [10 \quad 12 \quad 19]$$

Output weights for "car"

Word vector for "ants"

$\times$

300 features

300 features

softmax

$$\frac{e^x}{\sum e^x}$$

$=$ Probability that if you randomly pick a word nearby "ants", that it is "car"

300 features

EXAMINING OUTPUT WORD'S PROBABILITY DISTRIBUTION

SKIP GRAM VS CBOW

- Works well with small amount of data

- Represents rare words or phrases well

- Takes longer time to train

- Several times faster than the skip-gram

- Better accuracy for frequent words

REFERENCES

- **Word2Vec Tutorial - The Skip-gram Model**, Chris McCormick, 19 Apr 2016 (http://mccormickml.com/2016/04/19/word2vec-tutorial-the-skip-gram-model/)

- **Skip gram NLP context words prediction algorithm** - Towards Datascience, Sanket Doshi, 17 Mar 2019 (https://towardsdatascience.com/skip-gram-nlp-context-words-prediction-algorithm-5bbf34f84e0c#:~:text=Skip%2Dgram%20is%20one%20of,It's%20reverse%20of%20CBOW%20algorithm.&text=As%20there%20is%20more%20than,which%20makes%20this%20problem%20difficult.)