Spring25 CS598YP

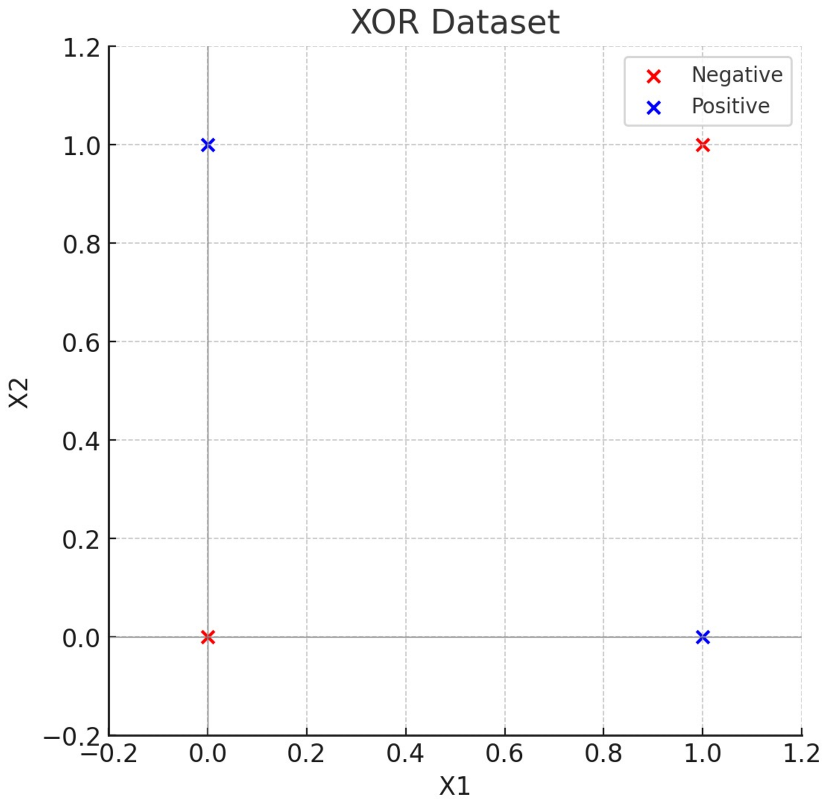
17.2: Word2Vec

Yongjoo Park

University of Illinois Urbana-Champaign

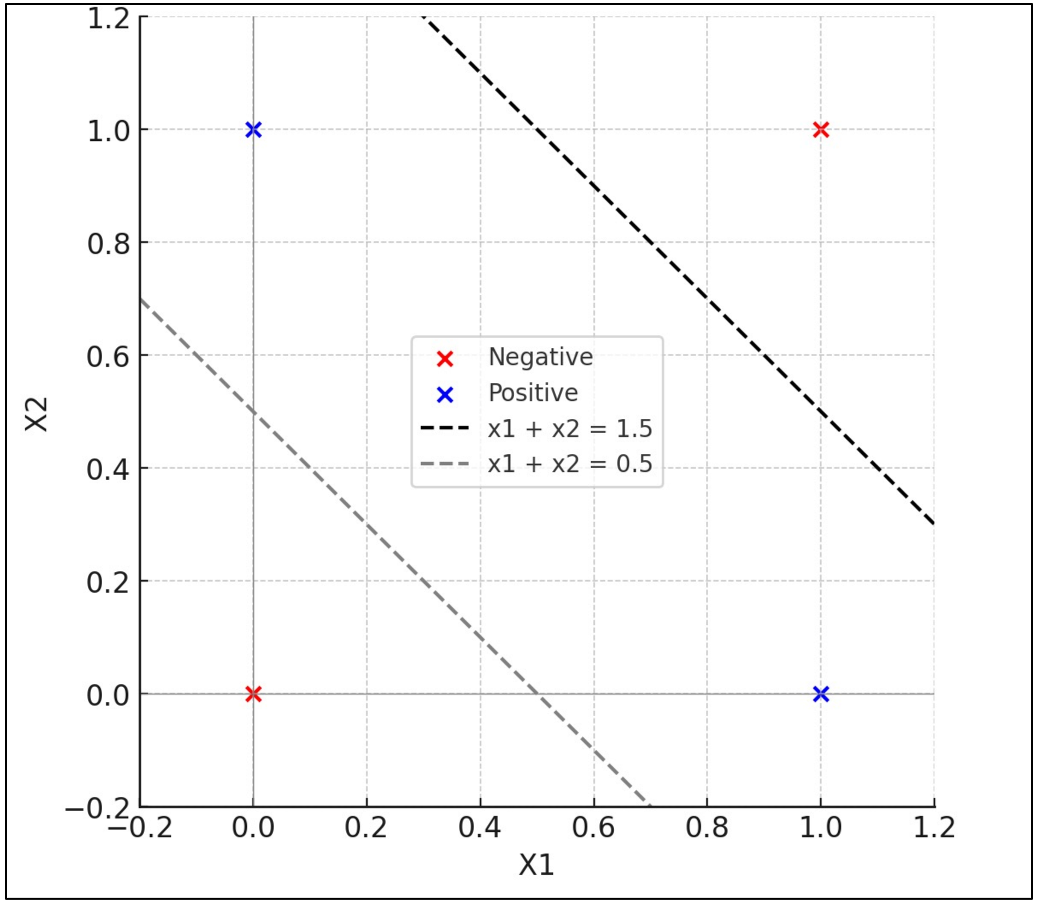
Neural Network Basics

Linear boundary cannot express XOR

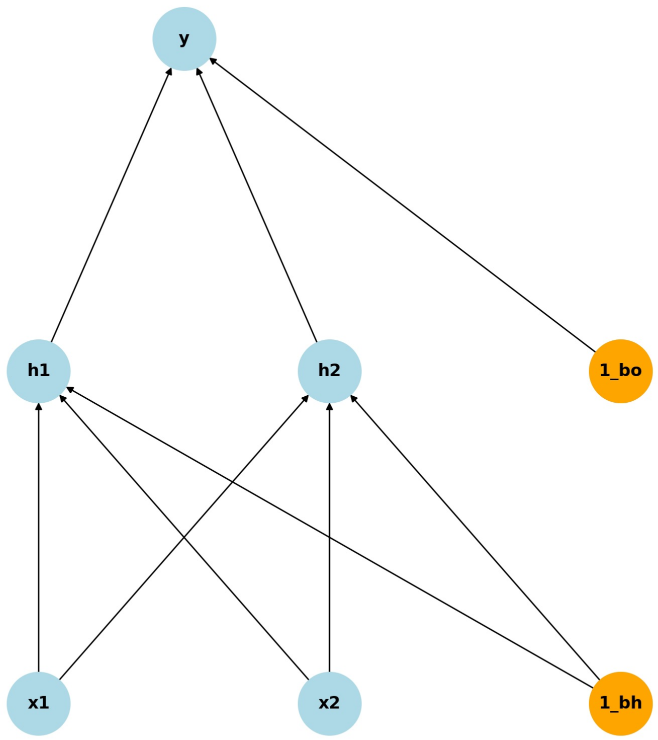


We can use non-linear mapping to express XOR

Non-linear layers can express XOR



Parameters of two-layer neural network



- $y = \sigma(w_{31}h_1 + w_{32}h_2 + b_3)$
- $h_1 = \sigma(w_{11}x_1 + w_{12}x_2 + b_1)$
- $h_2 = \sigma(w_{21}x_1 + w_{22}x_2 + b_2)$

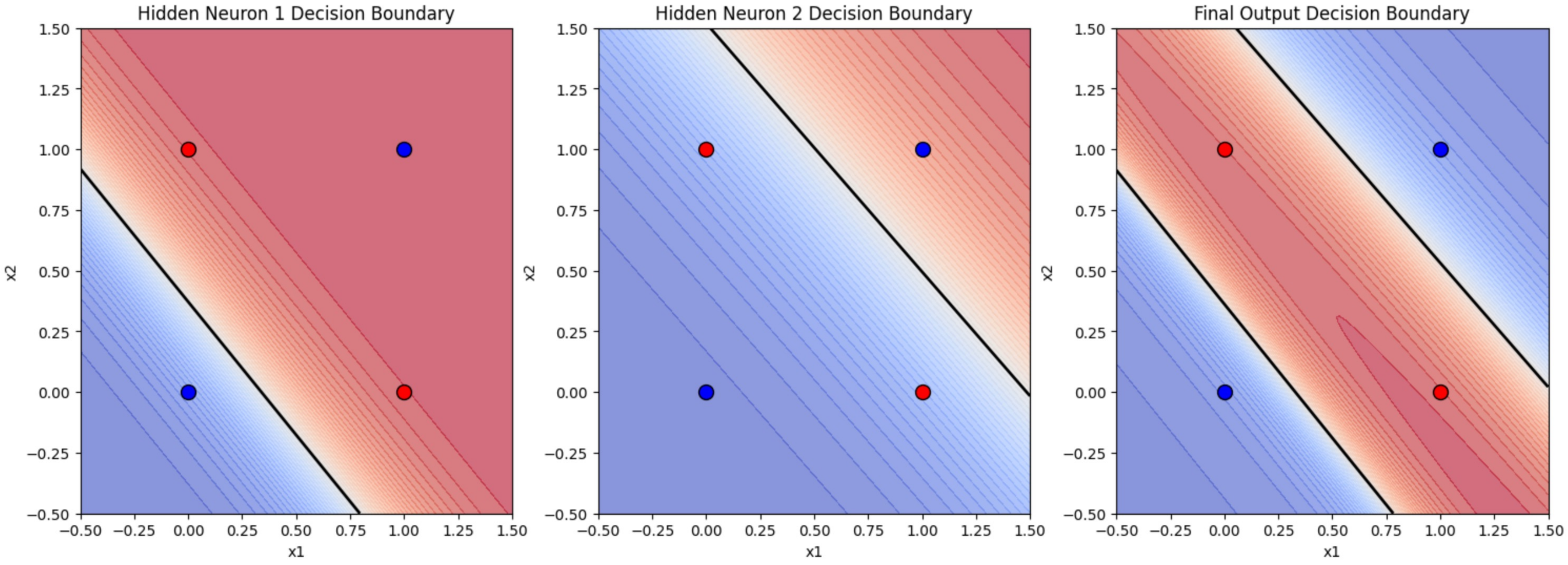
Objective function (negative log-likelihood)

- $\ell = -(t \log y + (1 - t) \log(1 - y))$

Colab example

```
Epoch 0, Loss: 0.8385518400233835
Epoch 1000, Loss: 0.6935294157721361
Epoch 2000, Loss: 0.6929689056414556
Epoch 3000, Loss: 0.6914984440909803
Epoch 4000, Loss: 0.6797520087317903
Epoch 5000, Loss: 0.6275268150325937
Epoch 6000, Loss: 0.5342532565852264
Epoch 7000, Loss: 0.4722113600512412
Epoch 8000, Loss: 0.35310833359020566
Epoch 9000, Loss: 0.1892975971815561

Final Weights and Biases:
w11: 5.683420743202077, w12: 5.182118467060412, b1: -1.9294958161165687
w21: 2.9483433333328297, w22: 2.8810666022847715, b2: -4.37592866451455
w31: 6.24617558266537, w32: -6.486694155096366, b3: -2.8210109634912017
```

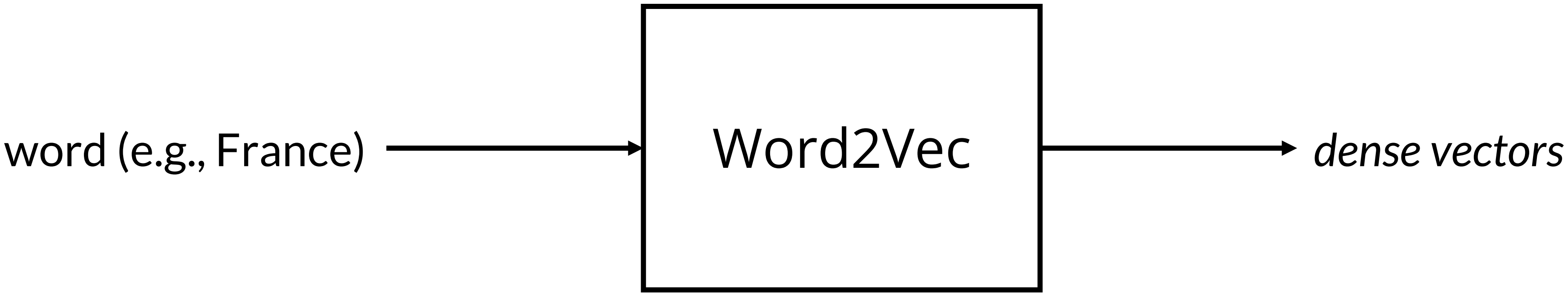


https://colab.research.google.com/drive/1LmUaoPU3xwPEhhSpOemxgAkoNHMAmCge?usp=sharing

Word2Vec: Skip-gram Model

https://mccormickml.com/2016/04/19/word2vec-tutorial-the-skip-gram-model/

Word2Vec Overview



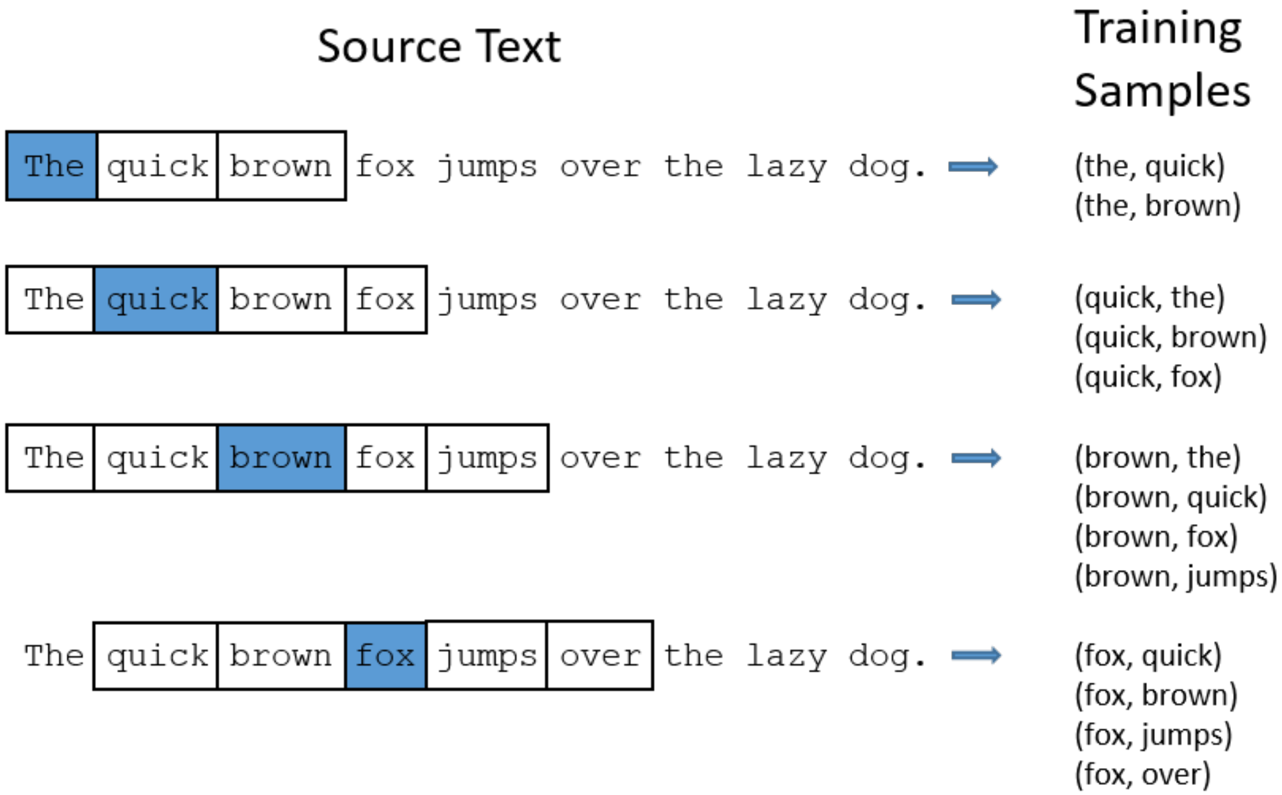
Word2Vec captures semantic relationship

Table 8: *Examples of the word pair relationships, using the best word vectors from Table 4 (Skip-gram model trained on 783M words with 300 dimensionality).*

Relationship	Example 1	Example 2	Example 3
France - Paris	Italy: Rome	Japan: Tokyo	Florida: Tallahassee
big - bigger	small: larger	cold: colder	quick: quicker
Miami - Florida	Baltimore: Maryland	Dallas: Texas	Kona: Hawaii
Einstein - scientist	Messi: midfielder	Mozart: violinist	Picasso: painter
Sarkozy - France	Berlusconi: Italy	Merkel: Germany	Koizumi: Japan
copper - Cu	zinc: Zn	gold: Au	uranium: plutonium
Berlusconi - Silvio	Sarkozy: Nicolas	Putin: Medvedev	Obama: Barack
Microsoft - Windows	Google: Android	IBM: Linux	Apple: iPhone
Microsoft - Ballmer	Google: Yahoo	IBM: McNealy	Apple: Jobs
Japan - sushi	Germany: bratwurst	France: tapas	USA: pizza

Training through *Fake Task*

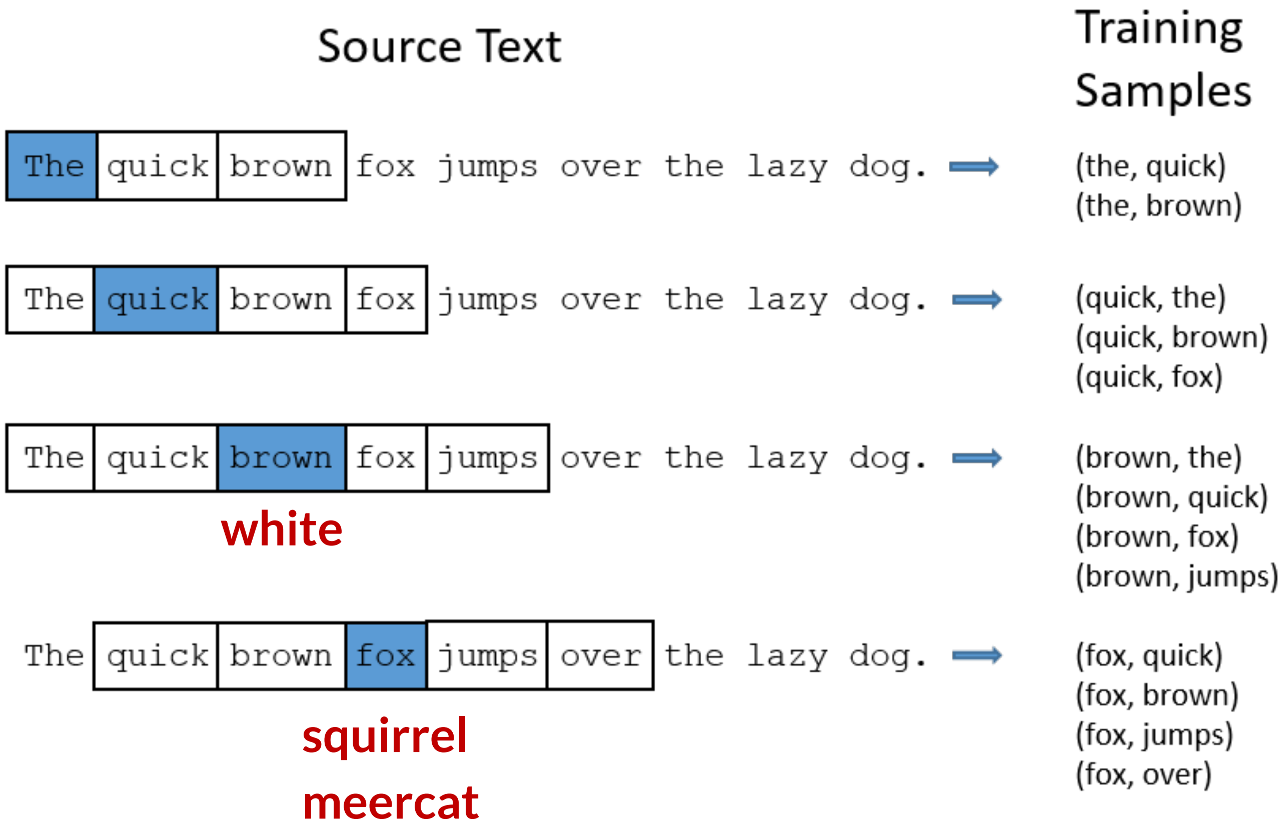
Task: Given **blue**, predict <u>other words</u> in the window



Word2vec uses C=10 past and future words

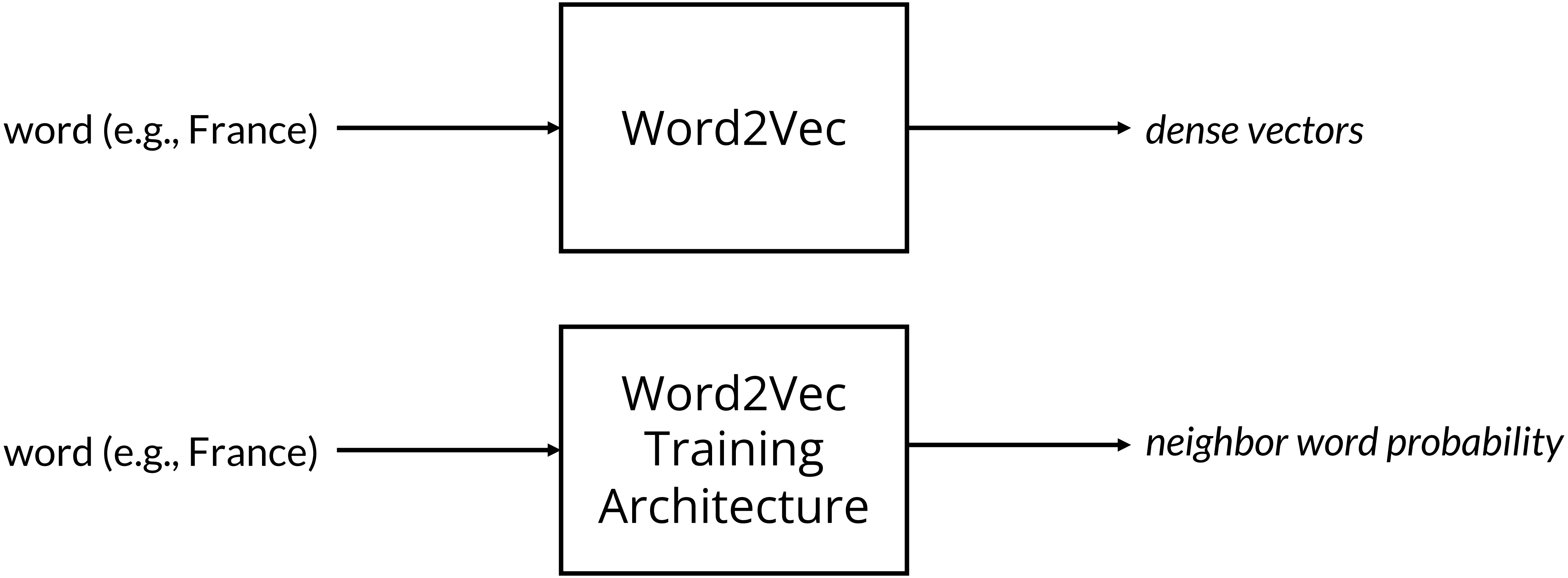
Training through *Fake Task*

Task: Given **blue**, predict <u>other words</u> in the window

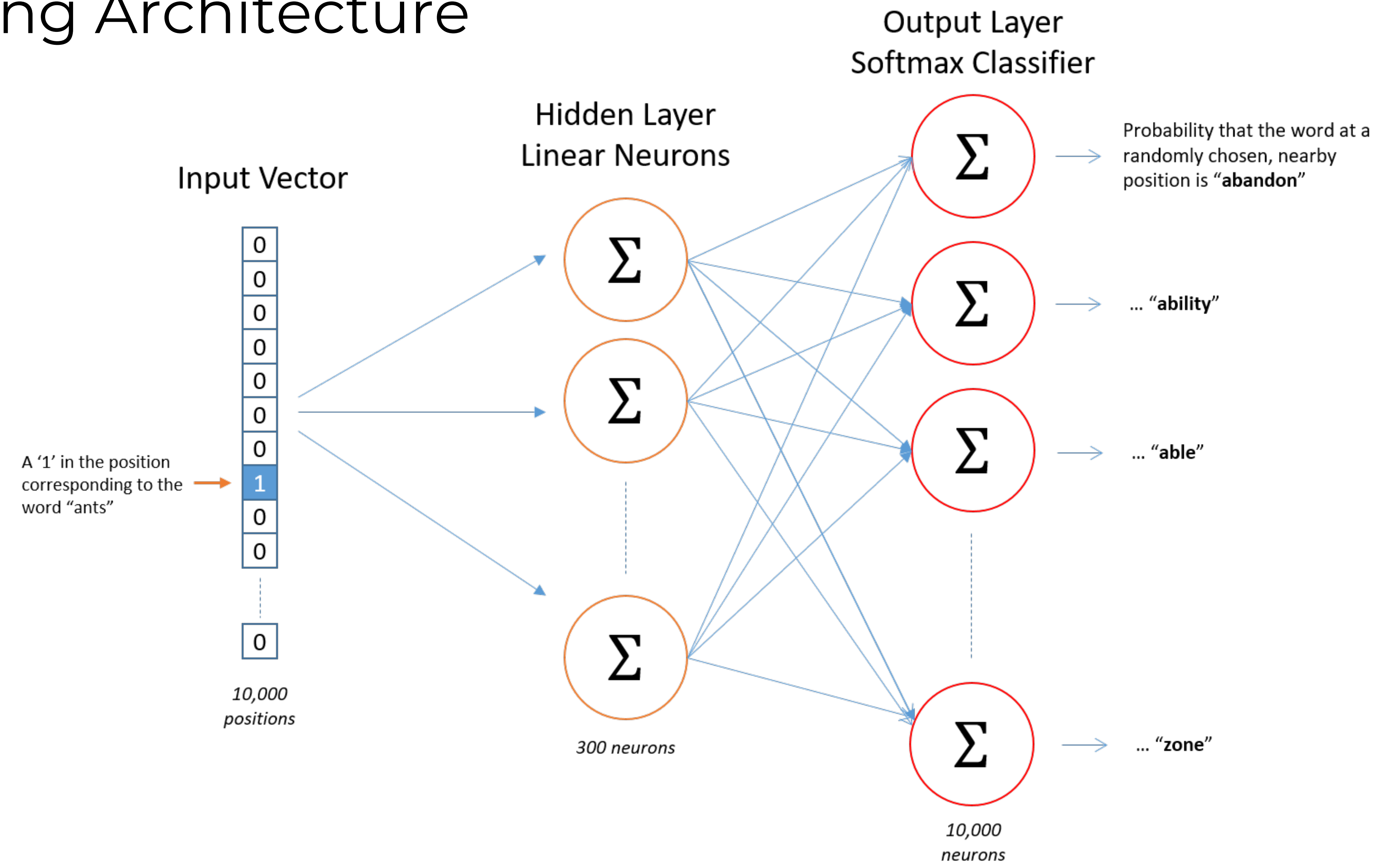


Word2vec uses C=10 past and future words

Word2Vec: Architecture for Fake Task

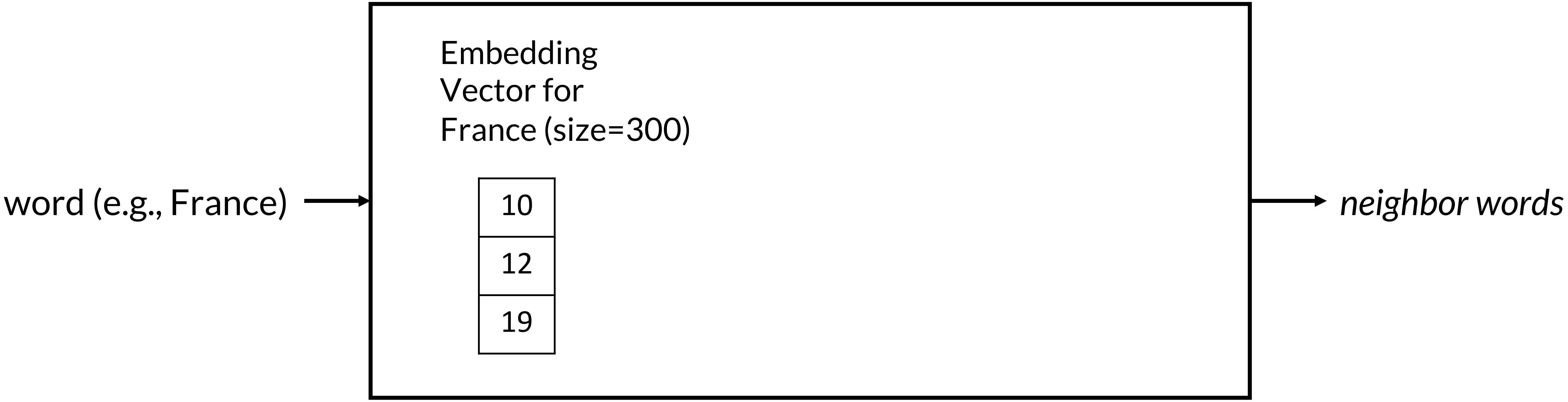


Training Architecture



Aim to learn *embeddings* through training

Training Architecture

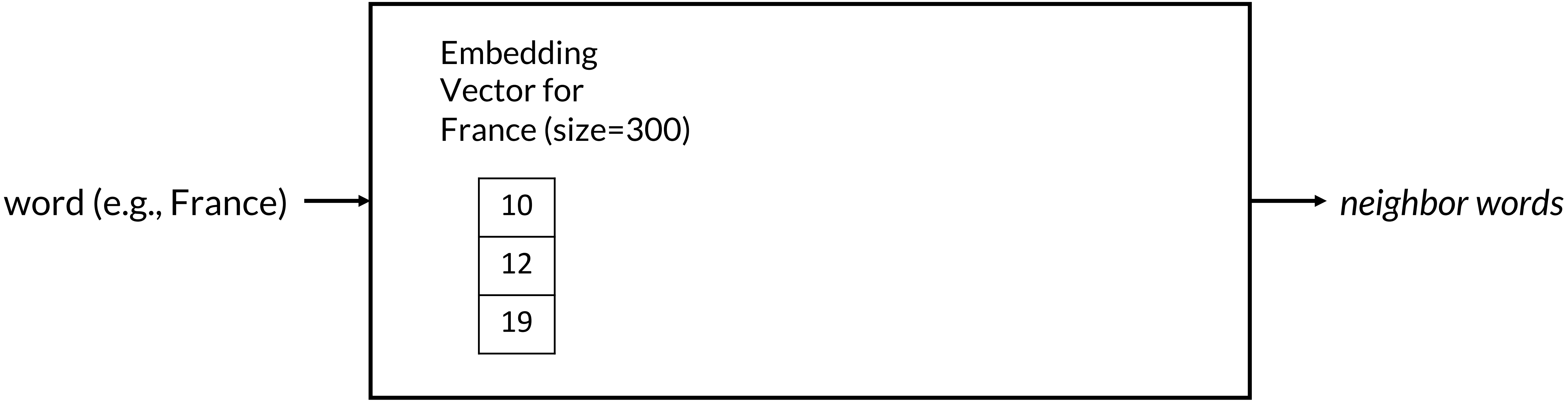


$$[0 \quad 0 \quad 0 \quad \boxed{1} \quad 0] \times \begin{bmatrix} 17 & 24 & 1 \\ 23 & 5 & 7 \\ 4 & 6 & 13 \\ \boxed{10 \quad 12 \quad 19} \\ 11 & 18 & 25 \end{bmatrix} = [10 \quad 12 \quad 19]$$

There are as many embedding vectors as the vocabulary size (e.g., 10K)

Embeddings will predict likely neighbor words

Training Architecture

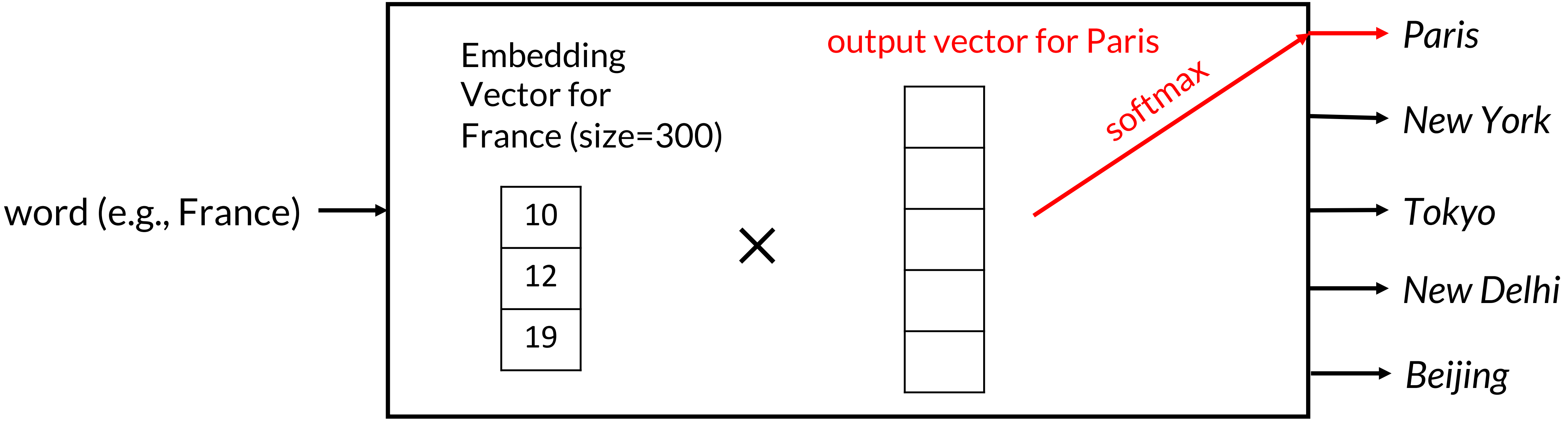


$$[0 \quad 0 \quad 0 \quad \boxed{1} \quad 0] \times \begin{bmatrix} 17 & 24 & 1 \\ 23 & 5 & 7 \\ 4 & 6 & 13 \\ \boxed{10 \quad 12 \quad 19} \\ 11 & 18 & 25 \end{bmatrix} = [10 \quad 12 \quad 19]$$

There are as many embedding vectors as the vocabulary size (e.g., 10K)

Embeddings will predict likely neighbor words

Training Architecture

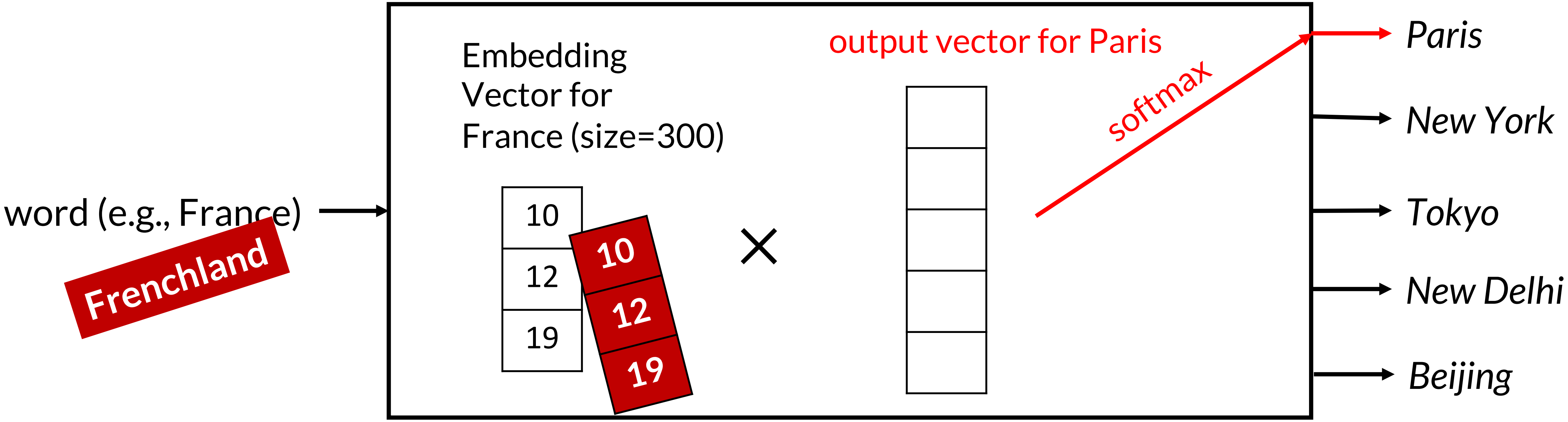


$$\begin{bmatrix}0 & 0 & 0 & 1 & 0\end{bmatrix} \times \begin{bmatrix}17 & 24 & 1 \\ 23 & 5 & 7 \\ 4 & 6 & 13 \\ 10 & 12 & 19 \\ 11 & 18 & 25\end{bmatrix} = \begin{bmatrix}10 & 12 & 19\end{bmatrix}$$

There are as many embedding vectors as the vocabulary size (e.g., 10K)

Embeddings will predict likely neighbor words

Training Architecture



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

There are as many embedding vectors as the vocabulary size (e.g., 10K)

Summary

- Can we learn a word’s meaning from its context?
- Skip-gram model: Predicts each neighbor words
- Use large corpus for training

Questions?