

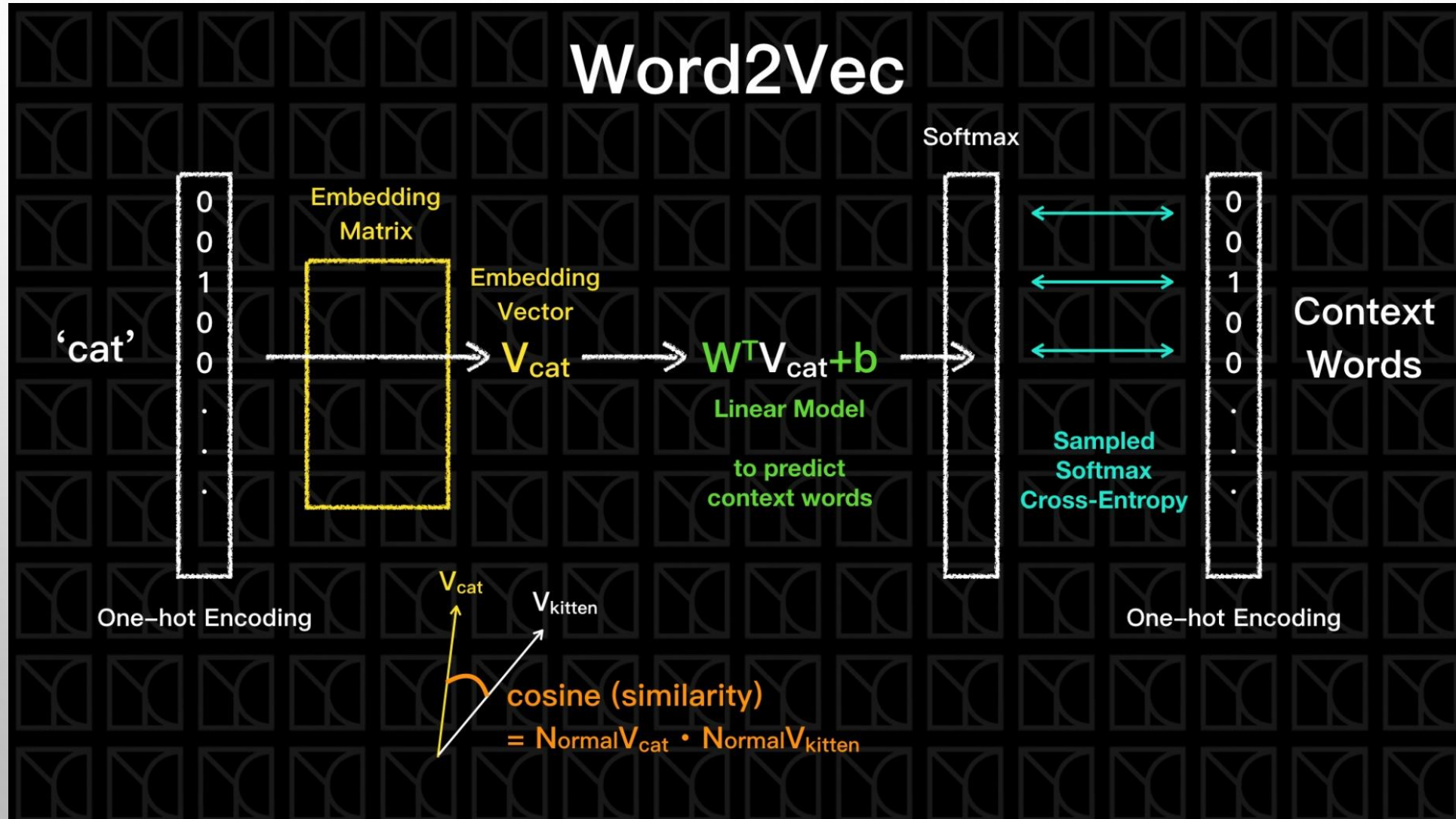
The background is a light gray gradient. It is decorated with numerous realistic water droplets of various sizes, some with highlights and shadows, scattered across the surface. In the upper center, there is a faint, circular logo or watermark that appears to contain a globe or a similar abstract design.

WORD2VEC

基礎概念

- 顧名思義是將文字轉為詞向量
- 屬於**UNSUPERVISED LEARNING**
- 最核心的概念是用前一個詞預測下一個詞
- 是一種訓練文字語意的類神經網路
- **HIDDEN LAYER**只有一層
- 可以很好地表達不同詞之間的相似和類比關係
- 使用**ONE-HOT ENCODING**
- 分為兩種模型
 - 1.**SKIP-GRAM**
 - 2.**CBOW**

流程圖



ONE-HOT ENCODING

- 也稱之為ONE OF N ENCODING
- 例: "THE QUICK BROWN FOX JUMPS OVER THE LAZY DOG"

The = [1, 0, 0, 0, 0, 0, 0, 0, 0]

quick = [0, 1, 0, 0, 0, 0, 0, 0, 0]

brown = [0, 0, 1, 0, 0, 0, 0, 0, 0]

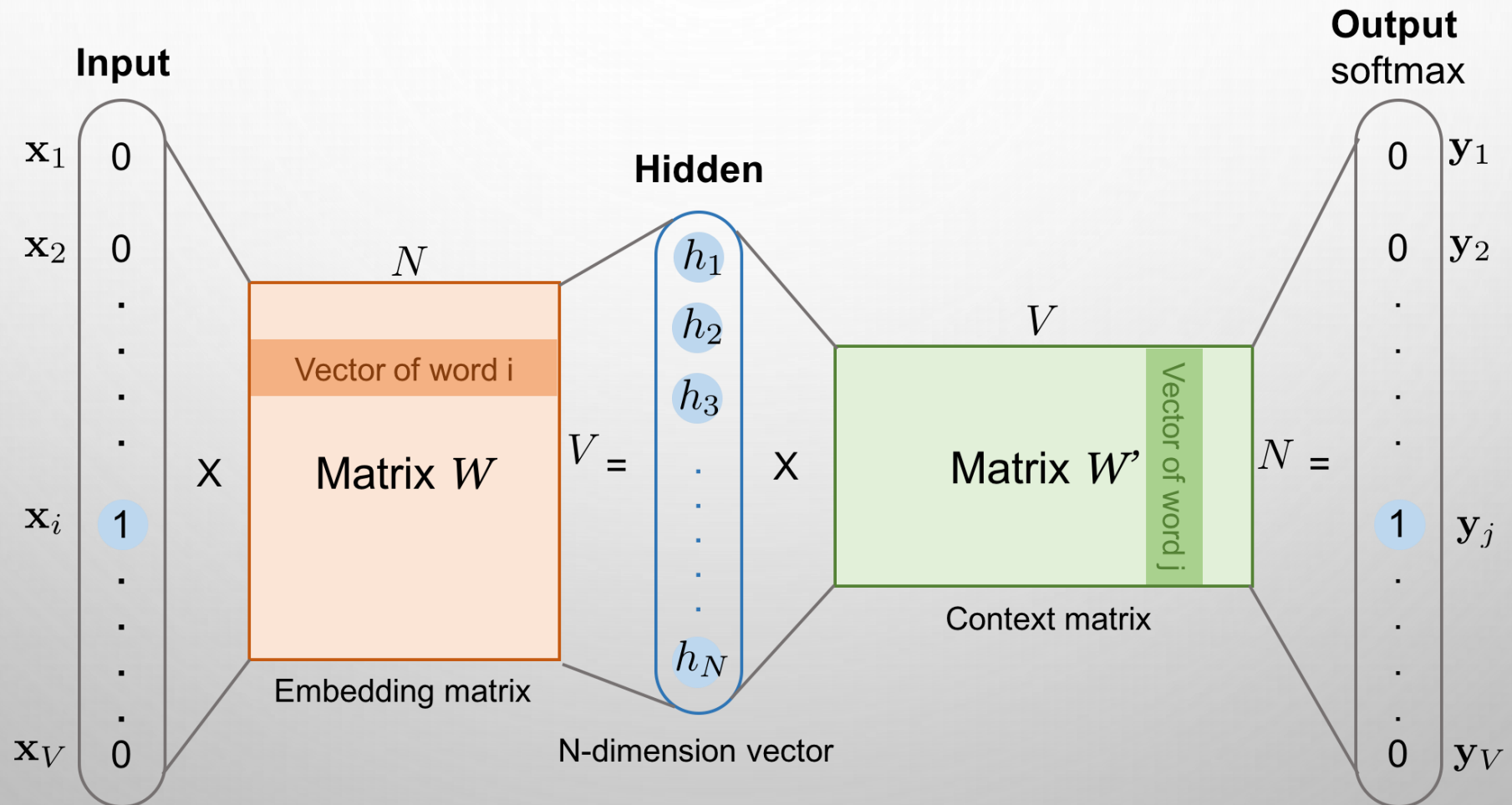
...

....

dog = [0, 0, 0, 0, 0, 0, 0, 0, 1]

WORD EMBEDDING

- 可達到降維的效果



skip-gram 和 CBOW(continuous bagging of words)

- SKIP-GRAM

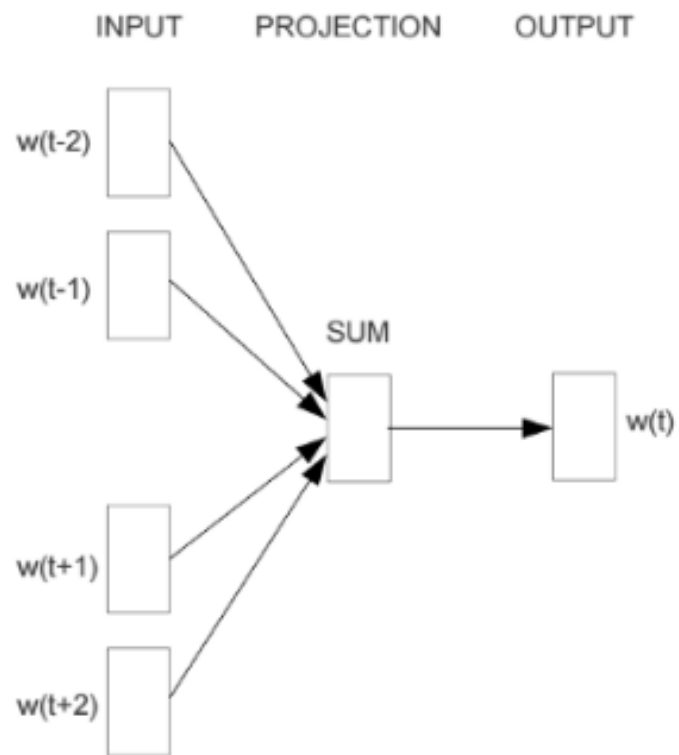
用中間詞預測上下文

- CBOW

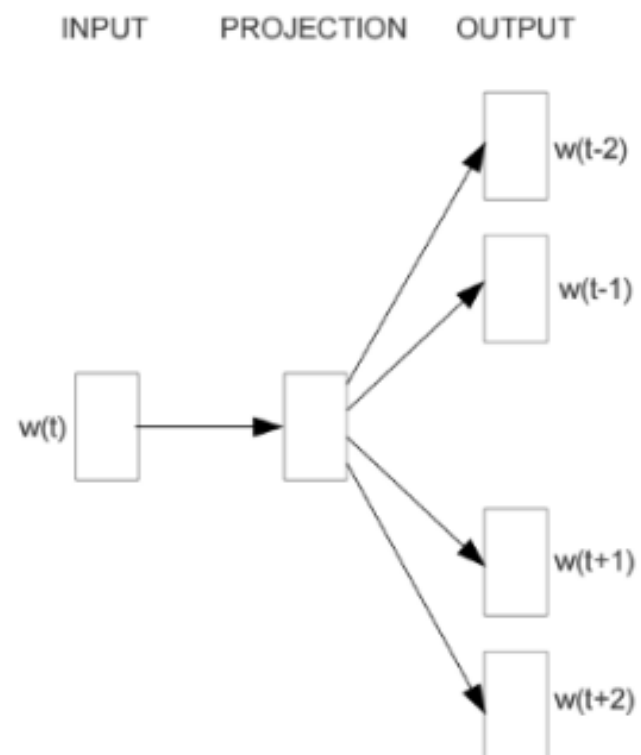
用上下文預測中間詞



兩種模型架構圖



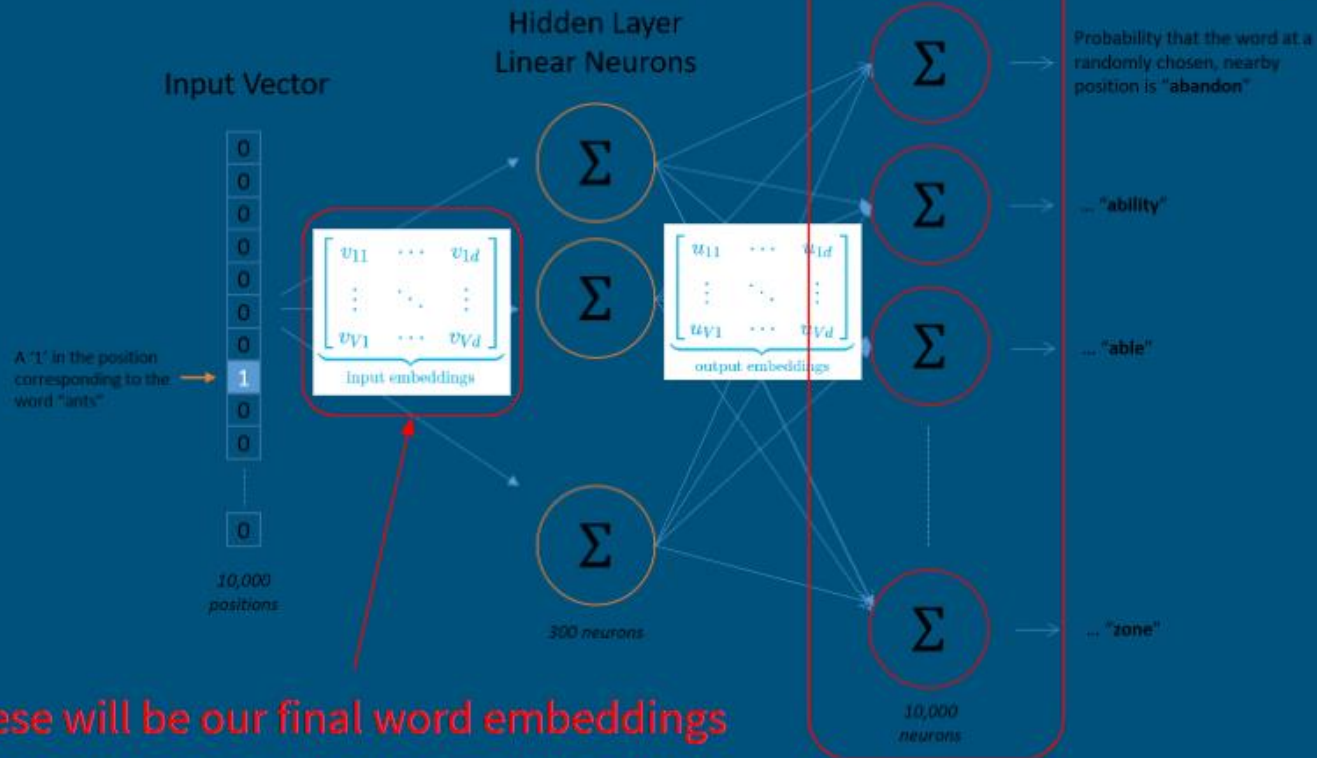
CBOW



Skip-gram

以SKIP-GRAM為例

Word2Vec Network Structure



INPUT 格式

window size=2

Source Text	Training Samples			
<table><tr><td>The</td><td>quick</td><td>brown</td></tr></table> fox jumps over the lazy dog. ➡	The	quick	brown	(the, quick) (the, brown)
The	quick	brown		
The <table><tr><td>quick</td><td>brown</td><td>fox</td></tr></table> jumps over the lazy dog. ➡	quick	brown	fox	(quick, the) (quick, brown) (quick, fox)
quick	brown	fox		
The quick <table><tr><td>brown</td><td>fox</td><td>jumps</td></tr></table> over the lazy dog. ➡	brown	fox	jumps	(brown, the) (brown, quick) (brown, fox) (brown, jumps)
brown	fox	jumps		
The quick brown <table><tr><td>fox</td><td>jumps</td><td>over</td></tr></table> the lazy dog. ➡	fox	jumps	over	(fox, quick) (fox, brown) (fox, jumps) (fox, over)
fox	jumps	over		

TRAINING

- MINIMIZE CROSS ENTROPY

OUTPUT格式

- 每個詞的機率，長度和INPUT的詞向量長度相同。
- 找出機率最大的詞。

WORD EMBEDDING 結果

Use cosine similarity

$$V(\text{China}) - V(\text{Beijing}) + V(\text{Tokyo}) = V(\text{Japan})$$

