

ĐỒ ÁN GIỮA KÌ

NHẬP MÔN XỬ LÝ

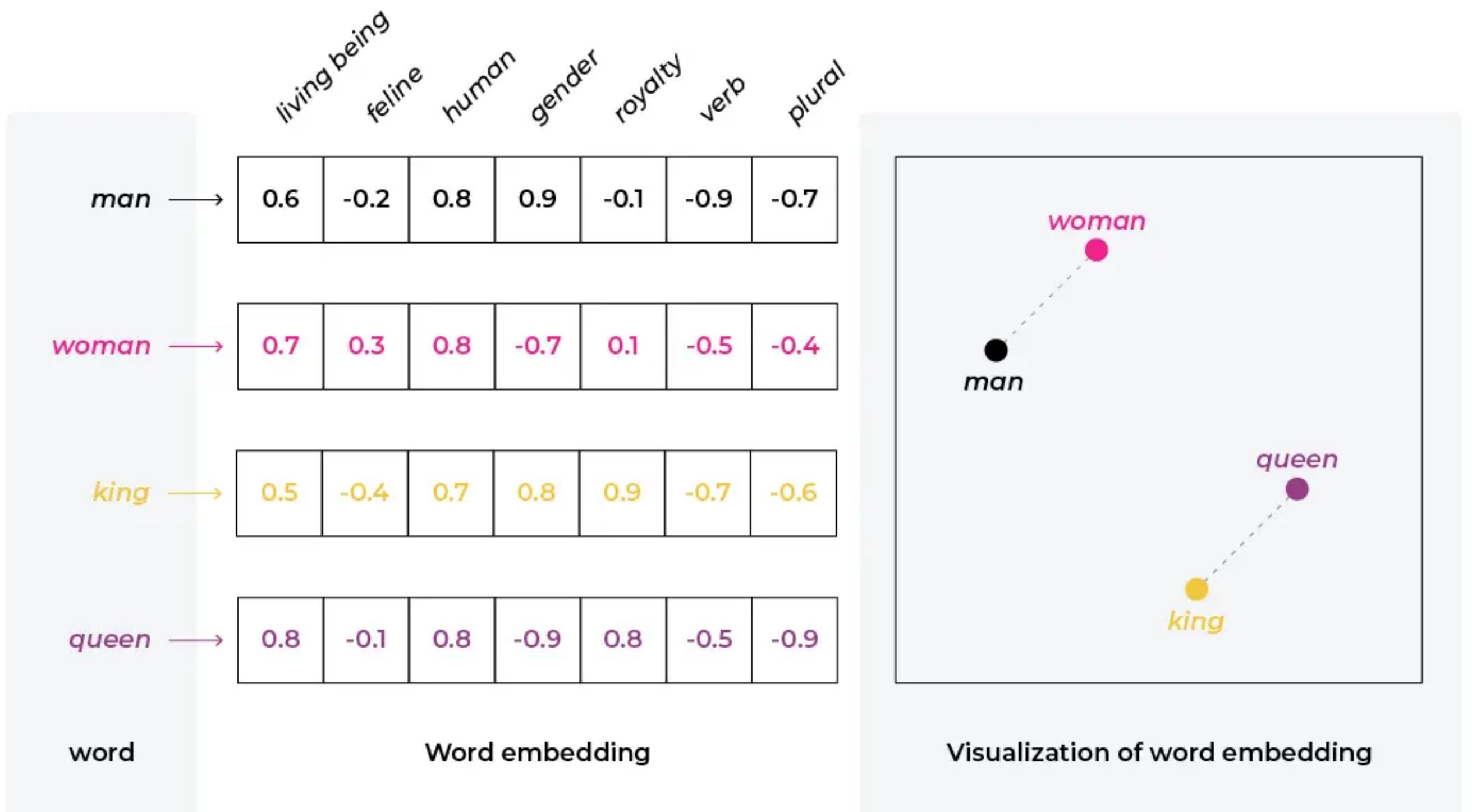
NGÔN NGỮ TỰ NHIÊN

52100322 - TRƯƠNG BÌNH THUẬN
52100878 - NGUYỄN ĐÌNH DANH
52100674 - TRẦN THỊ VẸN

NỘI DUNG THUYẾT TRÌNH

- 01** WORD EMBEDDINGS
- 02** PHƯƠNG PHÁP WORD EMBEDDING
- 03** GLOVE (GLOBAL VECTOR)
- 04** NEURAL EMBEDDING
- 05** FASTTEXT
- 06** QUARTER
- 07** NEXT PROJECT

WORD EMBEDDING



Word embedding cổ điển:

- + Bag of word
- + TF-IDF
- + Distributional Embeddings

Word embedding Neural

+
+
+

BAG OF WORD

Mỗi từ hoặc n-gram từ sẽ được mô tả là một vector có số chiều bằng đúng số từ trong bộ từ vựng. Tại vị trí tương ứng với vị trí của từ đó trong túi từ, phần tử trong vector đó sẽ được đánh dấu là 1. Những vị trí còn lại sẽ được đánh dấu là 0.



TF-IDF - TERM FREQUENCY–INVERSE DOCUMENT FREQUENCY

Tần suất- tần suất đảo nghịch từ. TF-IDF thể hiện trọng số của mỗi từ theo ngữ cảnh văn bản. TF-IDF sẽ có giá trị tăng tỷ lệ thuận với số lần xuất hiện của từ trong văn bản và số văn bản có chứa từ đó trên toàn bộ tập tài liệu.

Trong đó:

i : 1 .. D

n_i : tần số xuất hiện của từ trong văn bản i.

N_i : tổng số từ trong văn bản i.

$$tf_i = \frac{n_i}{N_i}$$

Trong đó:

D: tổng số documents trong tập dữ liệu.

d: số lượng documents có sự xuất hiện của từ.

$$idf_i = \log_2 \frac{D}{d}$$

$$tfidf_i = tf_i \times idf_i$$

DISTRIBUTIONAL EMBEDDING

Co-occurrence Matrix là một ma trận vuông đối xứng, mỗi hàng hoặc mỗi cột sẽ chính là vector biểu thị của từ tương ứng.

Số lượng từ vựng nhiều, ta thường chọn cách bỏ đi một số từ không cần thiết (ví dụ như các stopwords) hoặc sử dụng phân tích SVD (Singular Value Decomposition) để giảm kích thước của vector từ nhằm giúp cho biểu diễn của từ được rõ ràng hơn đồng thời tiết kiệm bộ nhớ dùng để lưu trữ Co-occurrence Matrix (do các Co-occurrence Matrix có kích thước rất lớn).

	John	is	not	fat	thin
John	0	2	0	1	0
is	2	0	1	0	1
not	0	1	0	1	0
fat	1	0	1	0	0
thin	0	1	0	0	0

GLOVE (GLOBAL VECTOR)

GloVe (Global Vector) là một trong những phương pháp mới để xây dựng vec-tơ từ (được giới thiệu vào năm 2014), nó thực chất được xây dựng dựa trên Co-occurrence Matrix. GloVe có bản chất là xác suất, ý tưởng xây dựng phương pháp này đến từ tỉ số sau:

$$\frac{P(k|i)}{P(k|j)} \quad P(k|i) = \frac{x_{ik}}{x_i} = \frac{x_{ik}}{\sum_m x_{im}}$$

$P(k|i)$ là xác suất xuất hiện của từ k trong ngữ cảnh của từ i , tương tự với $P(k|j)$.

x_{ik} : số lần xuất hiện của từ k trong ngữ cảnh của từ i (hoặc ngược lại).

x_i : số lần xuất hiện của từ i trong ngữ cảnh của toàn bộ các từ còn lại ngoại trừ i.

Các giá trị này chính là các mục nhập của Co-occurrence Matrix

WORD2VEC

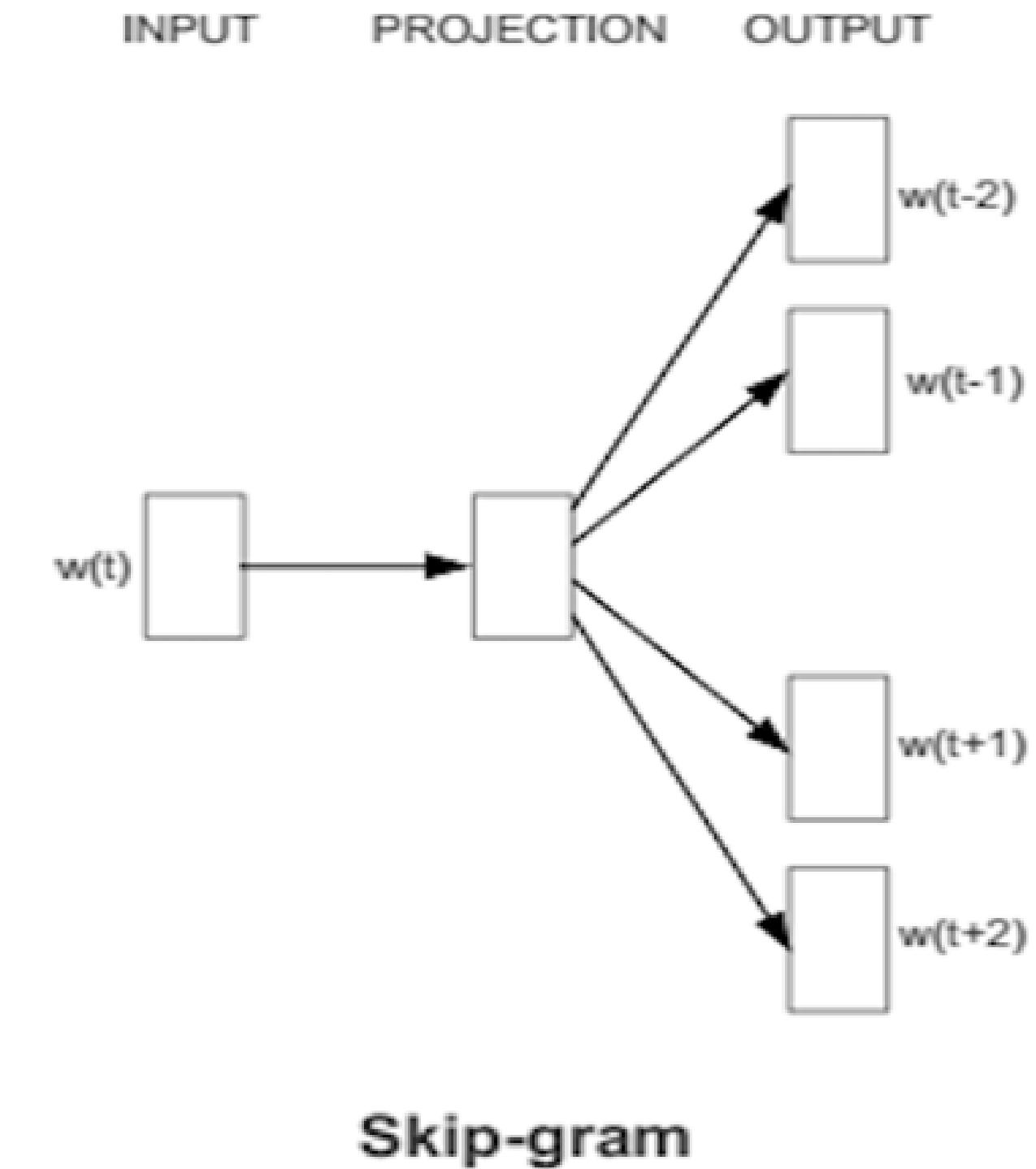
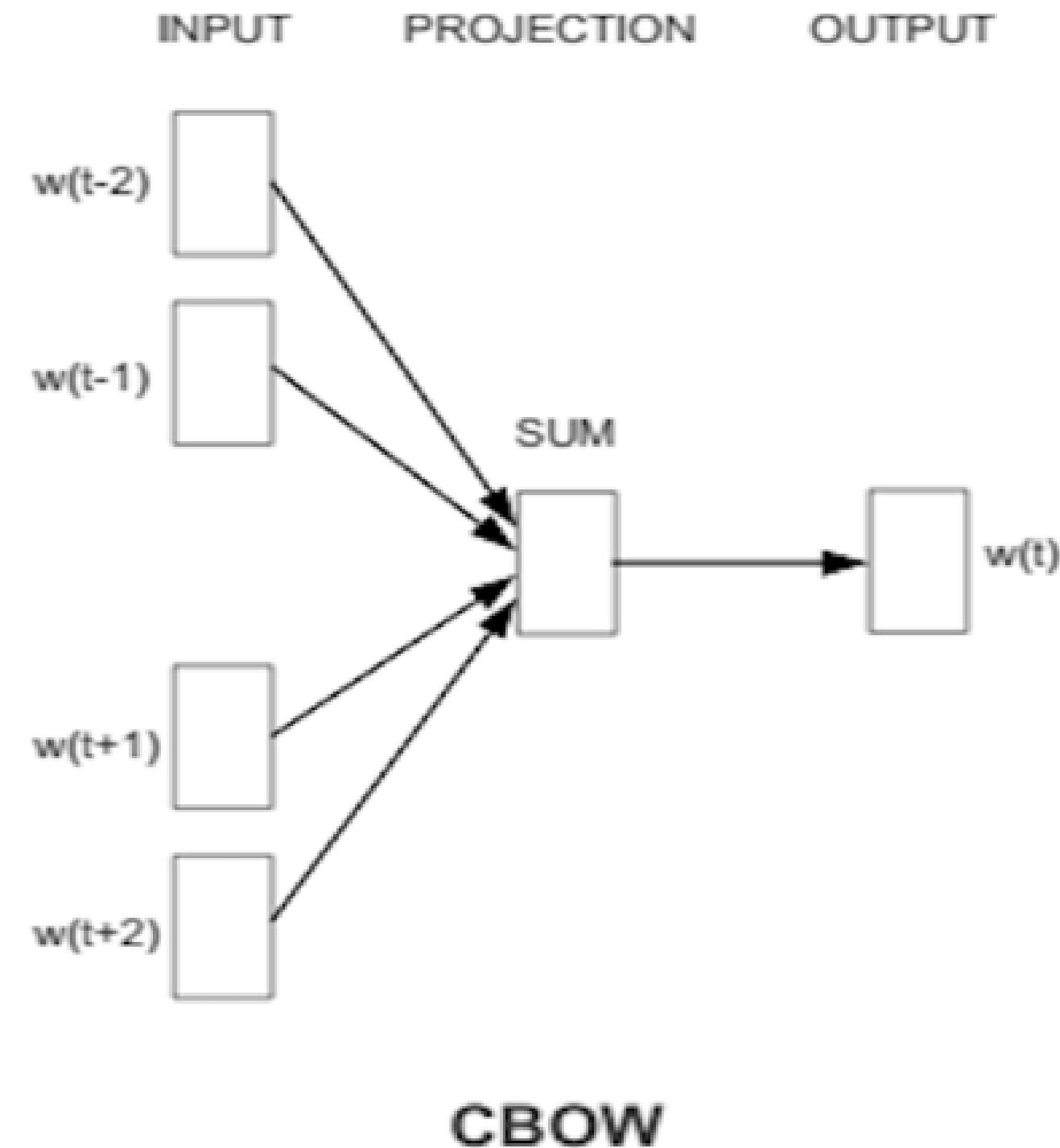
Word2vec học trực tiếp word vector có số chiều thấp trong quá trình dự đoán các từ xung quanh mỗi từ. Đặc điểm của phương pháp này là nhanh hơn và có thể dễ dàng kết hợp một câu một văn bản mới hoặc thêm vào từ vựng.

Word2vec là một mạng neural 2 lớp với duy nhất 1 tầng ẩn, lấy đầu vào là một corpus lớn và sinh ra không gian vector(với số chiều khoảng vài trăm), với mỗi từ duy nhất trong corpus được gắn với một vector tương ứng trong không gian.

WORD2VEC

	Vua	Hoàng hậu	Phụ nữ	Công chúa
Hoàng gia	0.99	0.99	0.02	0.98
Nam tính	0.99	0.05	0.01	0.02
Nữ tính	0.05	0.93	0.999	0.94
Tuổi	0.7	0.6	0.5	0.1

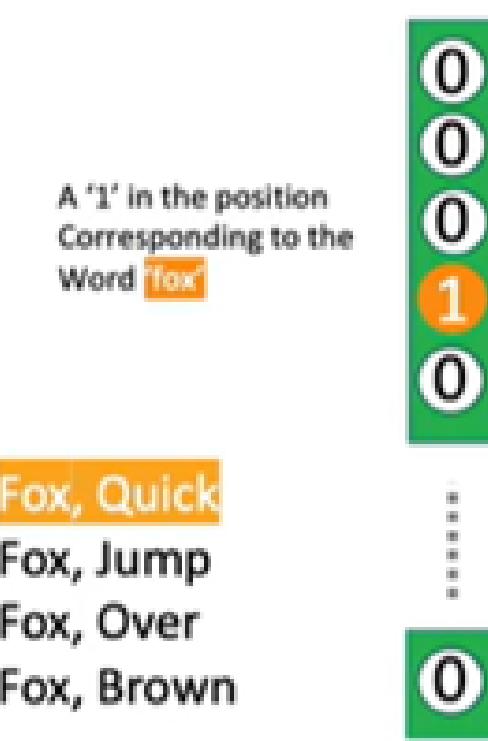
WORD2VEC



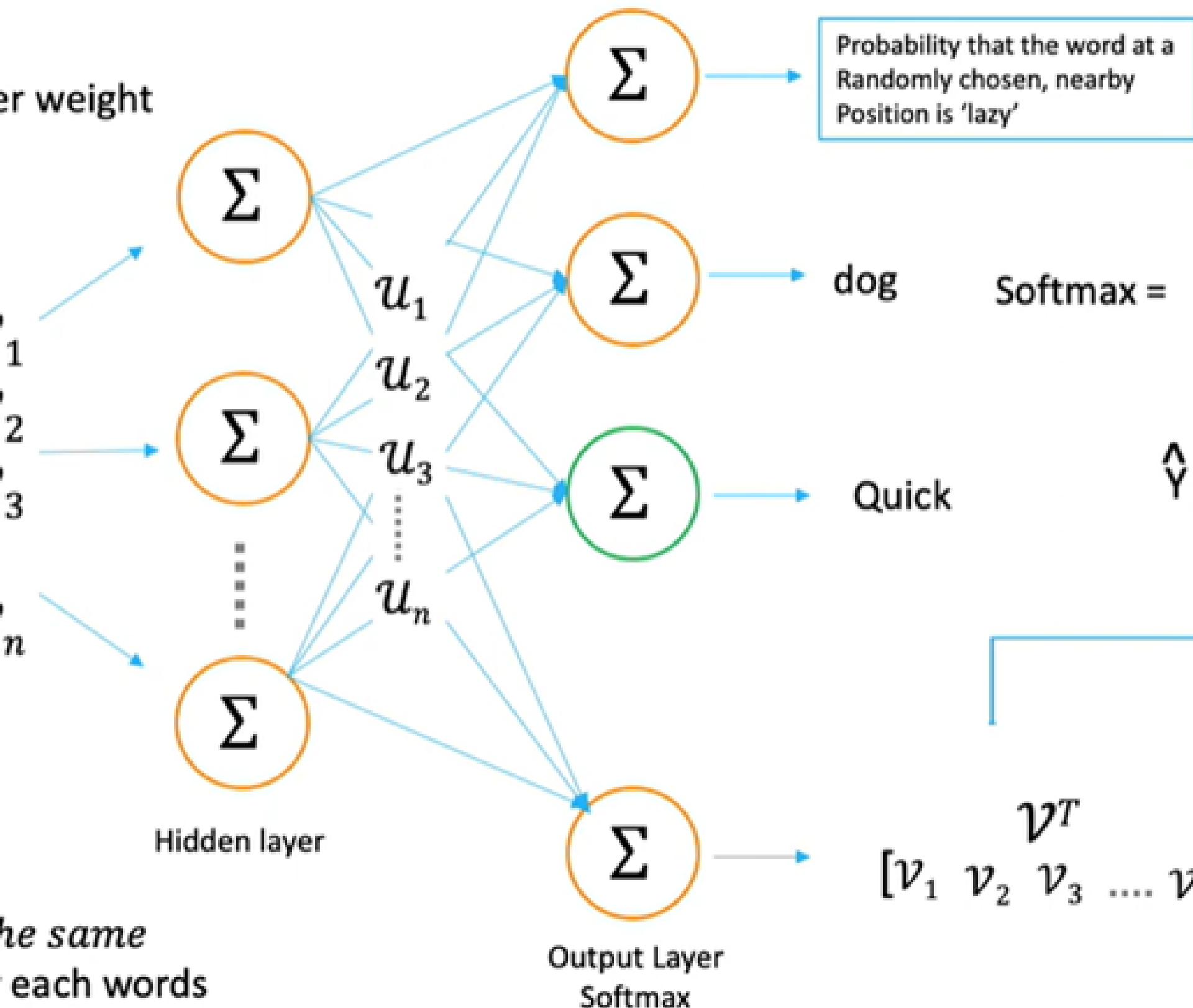
Skip-Gram

- Hoạt động tốt với lượng nhỏ dữ liệu đào tạo.(training data)
- Khi từ và cụm từ rất hiếm.

Goal: Learn hidden layer weight



V remains the same
 U changes for each words

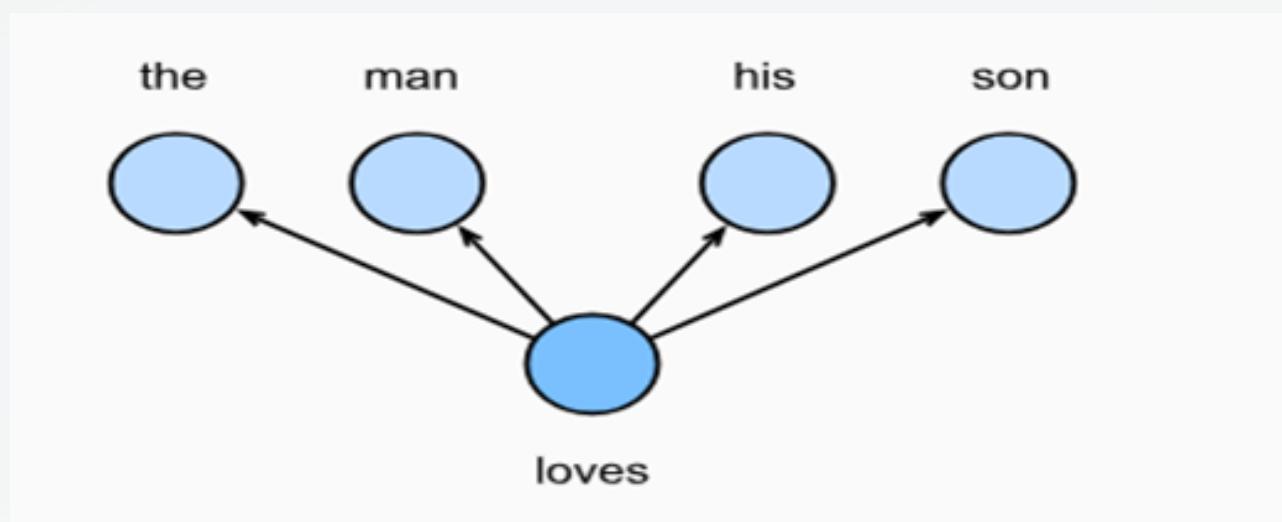


SKIP GRAM

Model details:

- Xây dựng bộ từ vựng
- Biểu diễn mỗi từ thành các one-hot-vector
- Đầu ra là một vector duy nhất, có kích thước bằng kích thước của bộ từ vựng, thể hiện xác suất của mỗi từ được là lân cận của từ đầu vào.
- Không có hàm kích hoạt trên tầng ẩn
- Hàm kích hoạt trên tầng output là softmax
- Trong quá trình training, input là 1 one-hotvector, ouput cũng là 1 one-hot-vector
- Trong quá trình đánh giá sau khi training, đầu ra phải là 1 phân bố xác suất

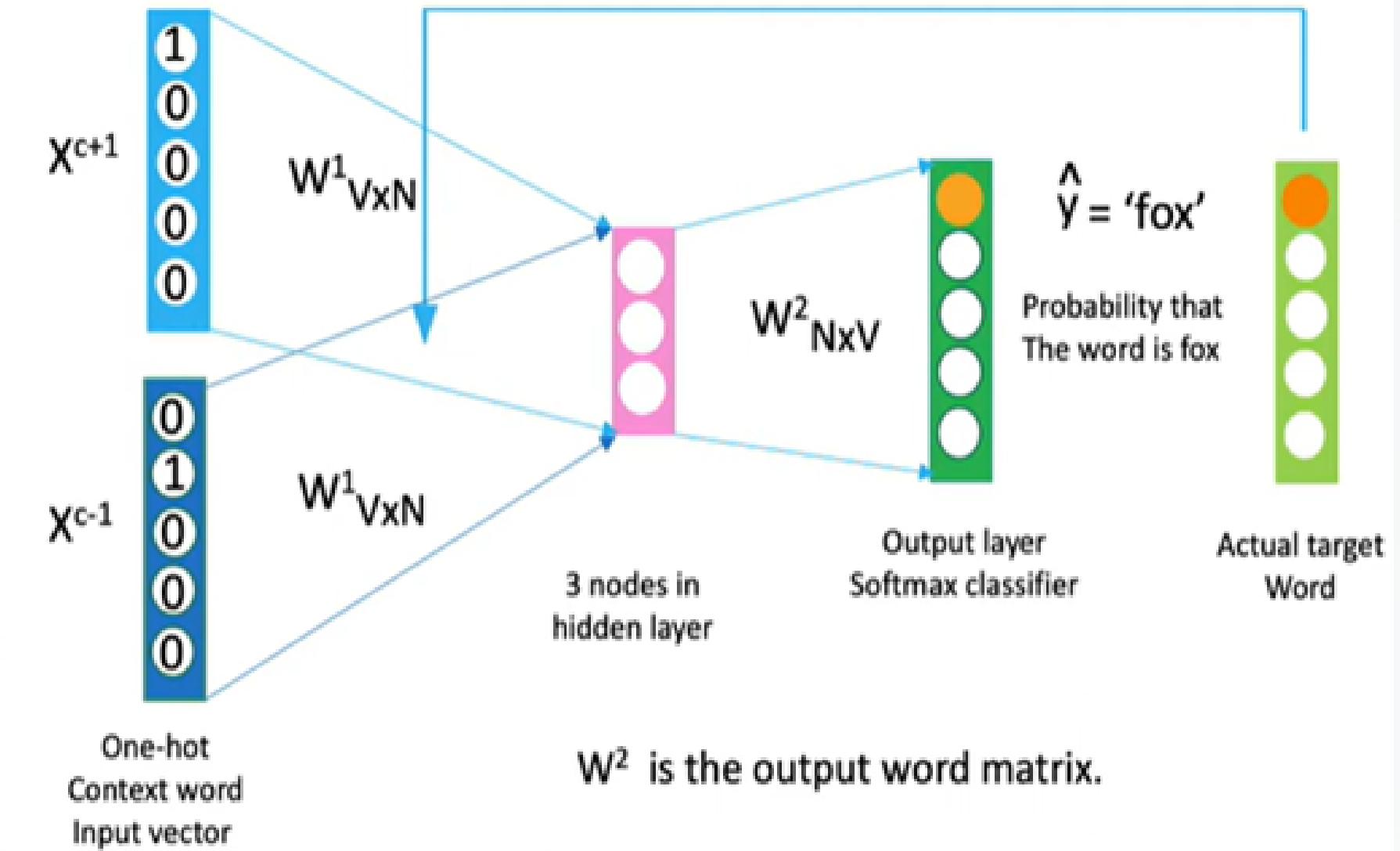
Word2vec cải tiến



MÔ HÌNH TÚI TỪ LIÊN TỤC (CBOW)

- Generate one hot word vectors
Window size = m
- Input one hot vectors or context = X^c
 $(X^{(c-m)}, \dots, X^{(c-1)}, X^{(c+1)}, \dots, X^{(c+m)})$
- Get embedded word vector for context
 $V^{c+1} = W^1 * X^{c+1}$
 $V^{c-1} = W^1 * X^{c-1}$
 \dots
 \dots
 $V^{c-m} = W^1 * X^{c-m}$

Softmax –
produce an
output between
0 and 1
(normalized prob)



W^2 is the output word matrix.

➤ Average vectors to $\hat{V} = \frac{V^{(c-m)} + V^{(c-m+1)} + \dots + V^{(c+m-1)} + V^{(c+m)}}{2m}$

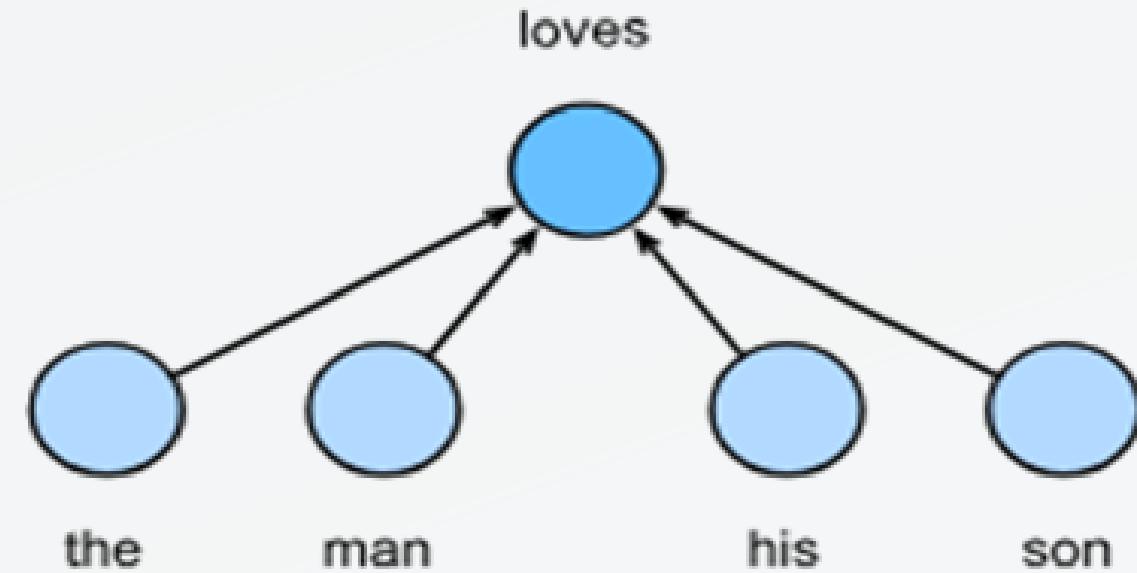
➤ Generate a score vector $z = W^2 * \hat{V}$

➤ Turn the scores into probability $\hat{y} = \text{softmax}(z)$

Prob of word fox is to be maximum

$$\text{Softmax} = \frac{e^x}{\sum_j e^x} = \frac{e^{V^T w_i u_{wo}}}{\sum_j e^{V^T w_i u_{wo}}}$$

CBOW



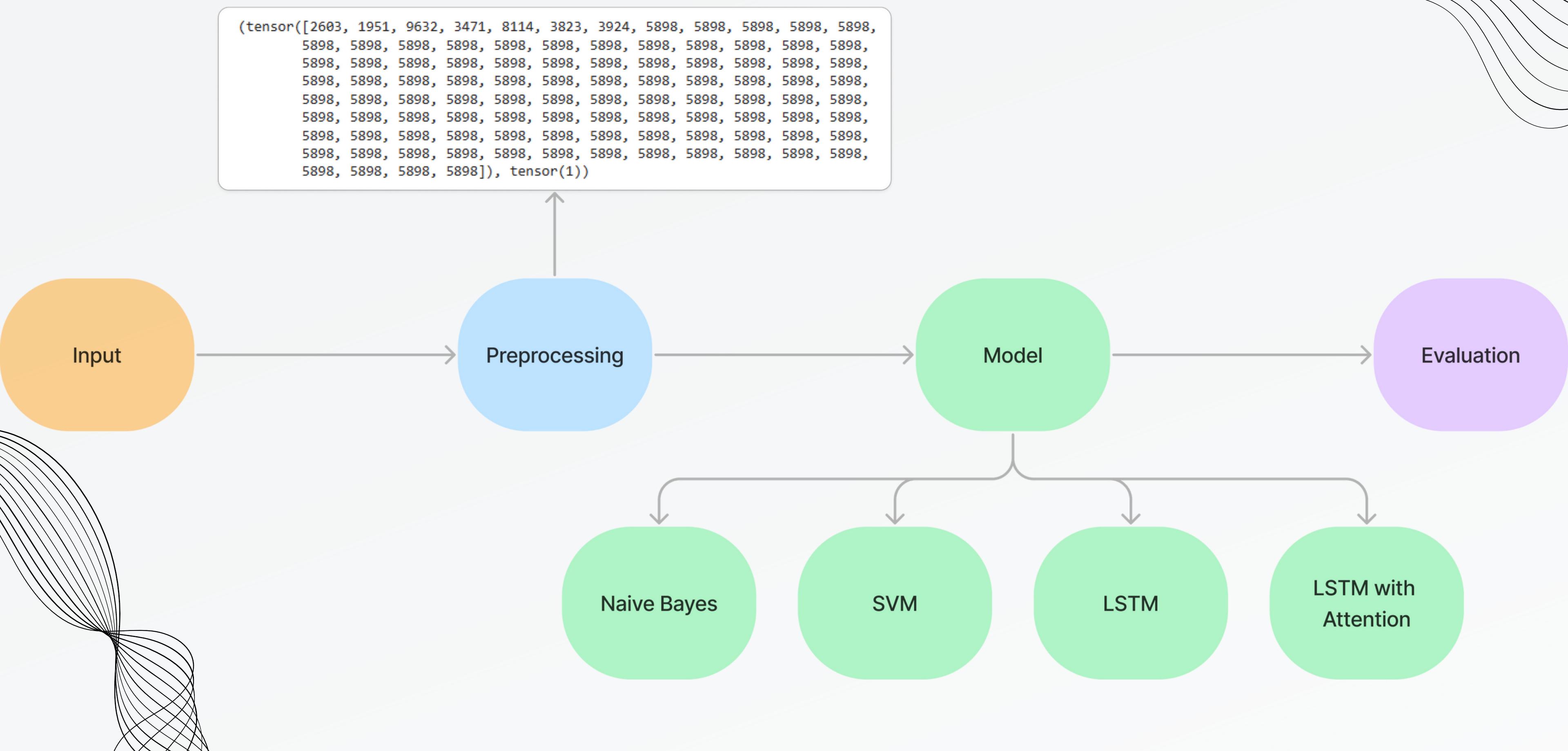
Mô hình túi từ liên tục (Continuous bag of words - CBOW) tương tự như mô hình skip-gram. Khác biệt lớn nhất là mô hình CBOW giả định rằng từ đích trung tâm được tạo ra dựa trên các từ ngữ cảnh phía trước và sau nó trong một chuỗi văn bản.

- Train nhanh hơn nhiều so với Skip-gram
- Bộ nhớ thấp. CBOW không cần phải có yêu cầu RAM lớn.
- Độ chính xác tốt hơn một chút cho frequent words

FASTTEXT VÀ WORD2VEC

Khía cạnh	Word2Vec	fastText
Xử lý từ nǎm ngoài từ vựng	Gặp khó khăn	Xử lý hiệu quả bằng cách chia các từ thành phụ từ
Biểu diễn của từ	Chỉ dựa trên từ	Xem xét thông tin phụ từ để biểu diễn phong phú hơn
Hiệu suất huấn luyện	Trung bình	Xuất sắc, đặc biệt với dữ liệu lớn
Các trường hợp sử dụng	Tìm từ tương tự, mối quan hệ giữa các từ, tương tự ngữ nghĩa	Xử lý từ nǎm ngoài từ vựng, phân tích tình cảm, xác định ngôn ngữ, hiểu biết hình thái học

CLASSIFICATION OF SEXTUALLY EXPLICIT



PREPROCESSING



**THANK'S FOR
WATCHING**

