Nghiên cứu thuật toán Glove embedding áp dụng vào bài toán Sentiment Analysis cho website bán hàng

18130038 - Lê Công Diễn
18130243 - Trịnh Quang Tiến
GVHD: TS. Nguyễn Văn Dũ

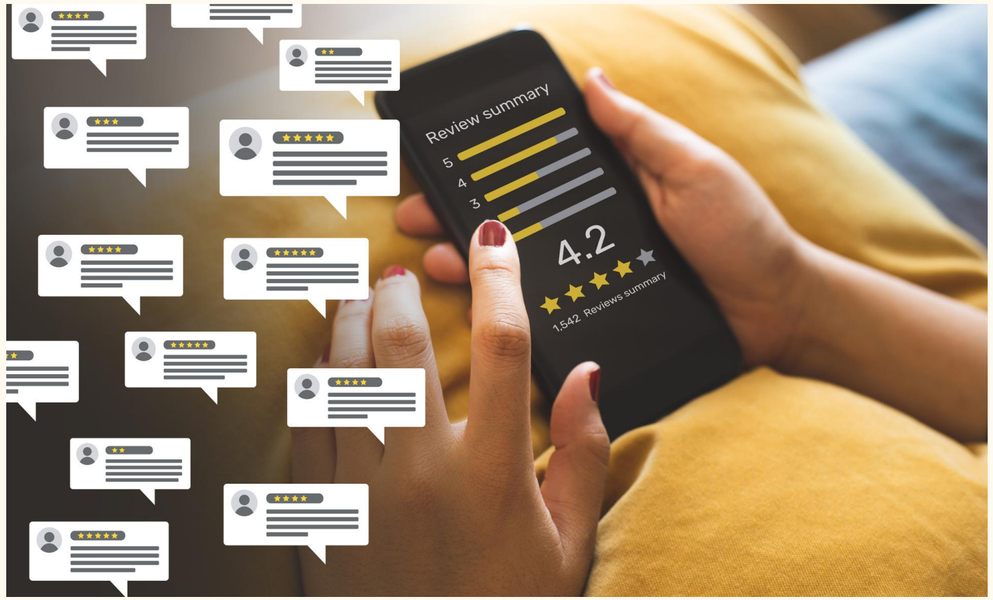
Nội dung chính

1. Đặt vấn đề
2. Giới thiệu bài toán Sentiment Analysis
3. Glove Embedding
4. Xây dựng mô hình
5. Triển khai ứng dụng
6. Demo

1

Đặt vấn đề

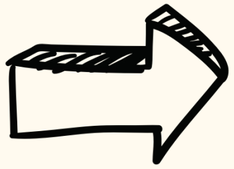
1 Đặt vấn đề



Thương mại điện tử phát triển

Dữ liệu đánh giá khách hàng nhiều

1 Đặt vấn đề

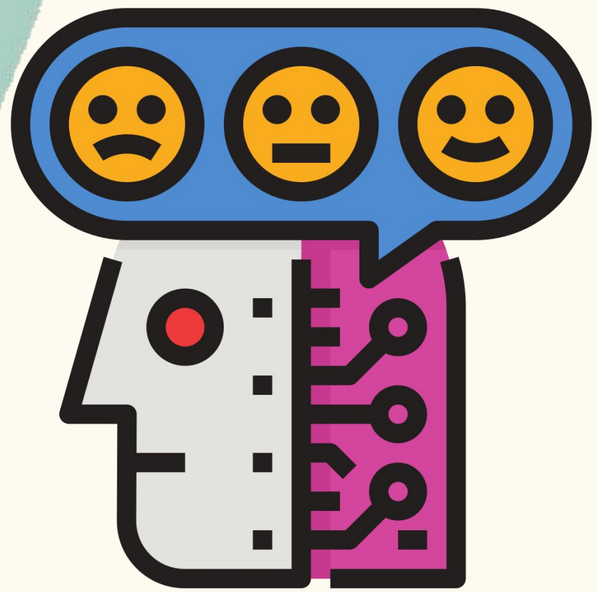


Nhu cầu phân tích lượng lớn dữ liệu đánh giá của khách hàng

Phân tích cảm xúc (Sentiment Analysis)

2

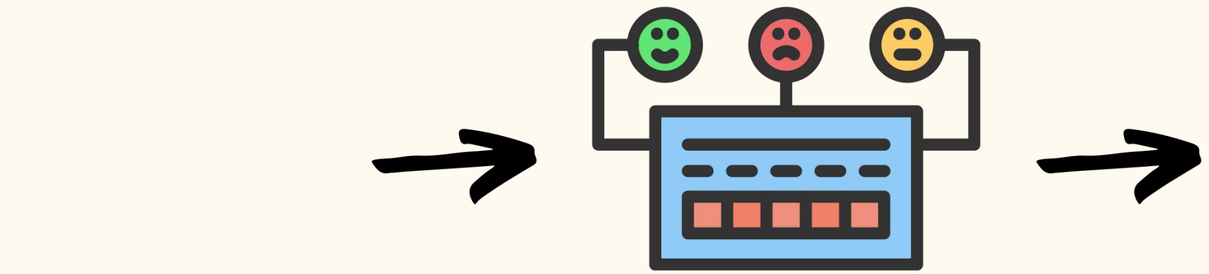
Giới thiệu bài toán Sentiment Analysis



2 Giới thiệu bài toán Sentiment Analysis

Reviews **Sentiment Analysis** **Result**

"I love this product"



"This product is bad"

2 Giới thiệu bài toán Sentiment Analysis

Ứng dụng

Mạng xã hội

Chăm sóc khách hàng

2 Giới thiệu bài toán Sentiment Analysis

Bài báo liên quan

“Comparison of Deep Learning and Rule-based Method for the Sentiment Analysis Task”

→ *Sentiment Analysis (Machine Learning based) > Sentiment Analysis (Rule Based)*

“The Importance of Neutral Examples for Learning Sentiment”

→ *Đánh giá cao việc có thêm lớp Neutral hơn trong bài toán Sentiment*

Lựa chọn nghiên cứu

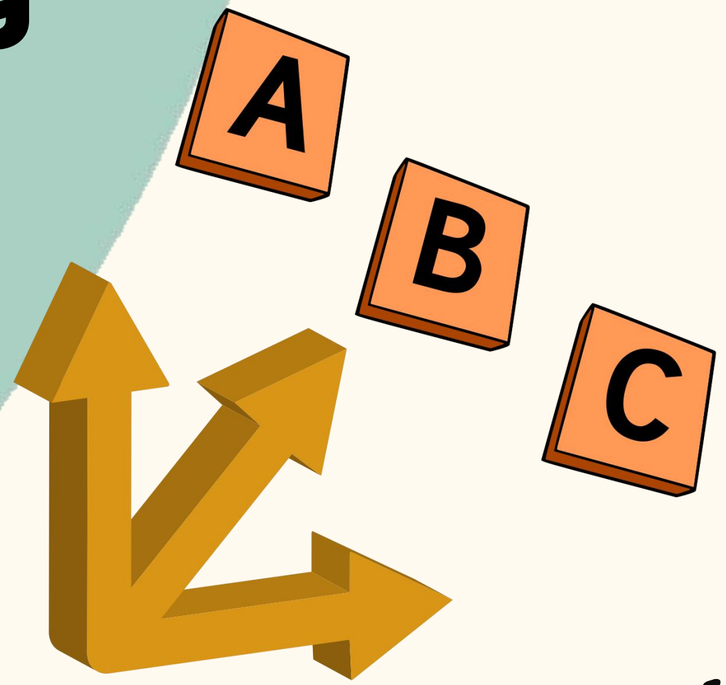
Loại Sentiment Analysis: Ba lớp (Positive/Neutral/Negative)

Loại mô hình: LSTM

Áp dụng: Website bán điện thoại di động

3

Glove Embedding

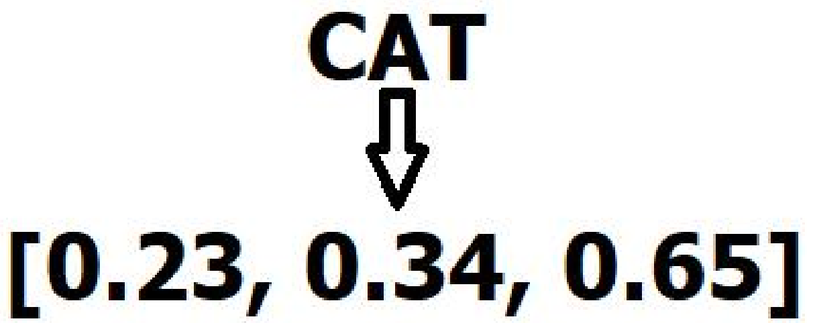


3 Glove Embedding

Glove: Glove Vector

Lịch sử: 2014 - các nhà nghiên cứu ĐH Stanford: Jeffrey Pennington, Richard Socher, Christopher D. Manning

Là mô hình vector hóa chữ thành vector, trong đó các vector có thể biểu diễn được mối quan hệ gần nhau dựa trên sự tương đồng về ngữ cảnh.



3 Glove Embedding

Tại sao dùng Glove?

Tại sao vector từ? Vì máy tính không hiểu được các chữ như con người hiểu. Ví dụ từ man, woman

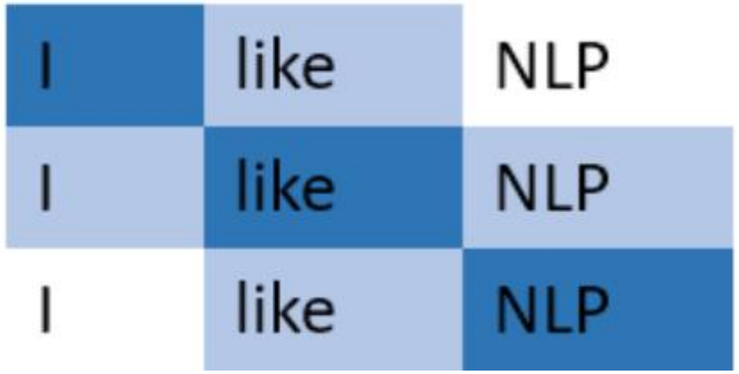
Tại sao dùng Glove? Vì mô hình phân tích cảm xúc, hoạt động hiệu quả khi các vector biểu diễn các từ có thể biểu diễn được mối quan hệ gần nhau về ngữ nghĩa giữa các từ.

Tại sao Glove làm được điều này? Glove dựa vào các từ ngữ cảnh gần với một từ để hiểu về từ đó.

3 Glove Embedding

Ma trận đồng xuất hiện (Co-occurrence matrix)

Cho ngữ liệu: “I like NLP
I like deep learning
I enjoy flying”

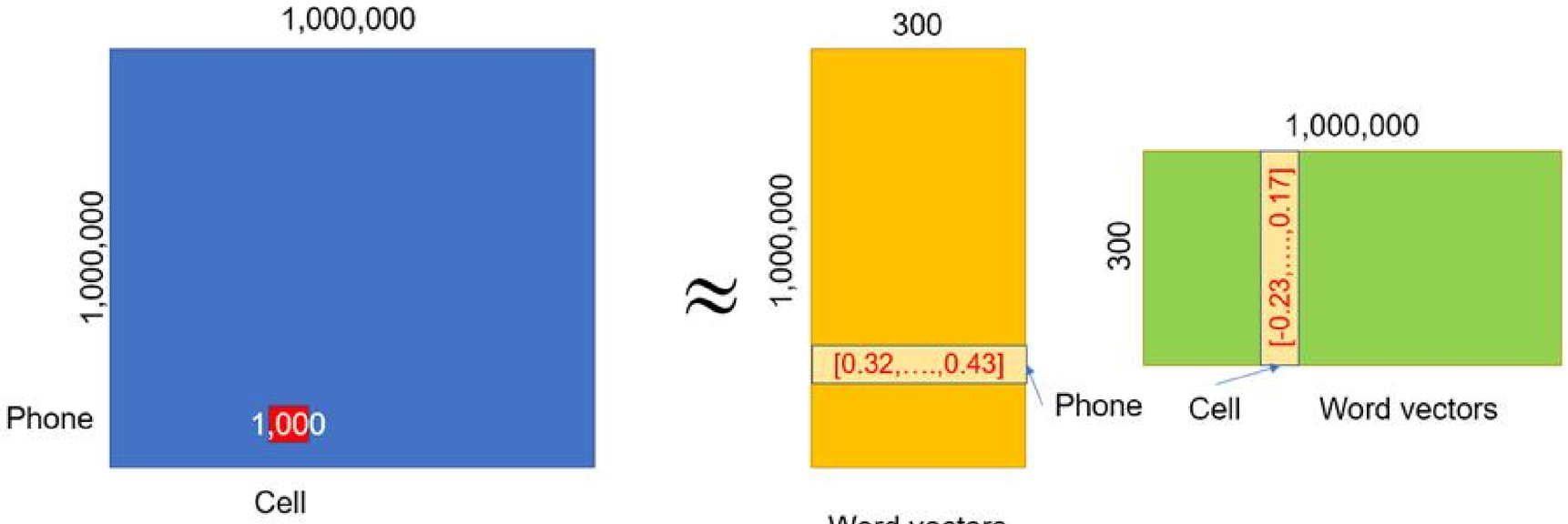


count	I	like	NLP	deep	learning	enjoy	flying
I	0	1	0	0	0	0	0
like	1	0	1	0	0	0	0
NLP	0	1	0	0	0	0	0
deep	0	0	0	0	0	0	0
learning	0	0	0	0	0	0	0
enjoy	0	0	0	0	0	0	0
flying	0	0	0	0	0	0	0

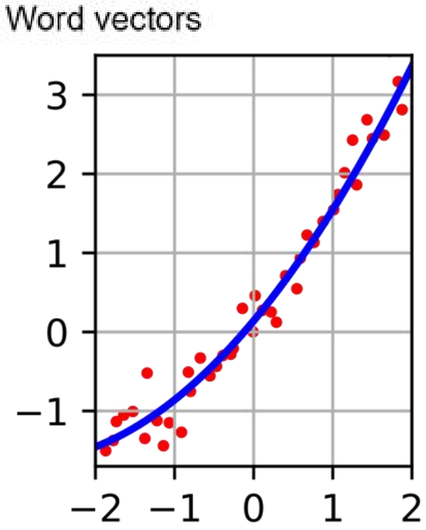
count	I	like	Nlp	deep	learning	enjoy	flying
I	0	2	0	0	0	1	0
like	2	0	1	1	0	0	0
Nlp	0	1	0	0	0	0	0
deep	0	1	0	0	1	0	0
learning	0	0	0	1	0	0	0
enjoy	1	0	0	0	0	0	1
flying	0	0	0	0	0	1	0

3 Glove Embedding

Mô hình Glove



$$J(w_i, \bar{w}_j) = \sum_{i,j=1}^{|V|} f(X_{ij}) \left(w_i^T \bar{w}_j + b_i + \bar{b}_j - \log(X_{ij}) \right)^2$$



$$w_i^{\text{final}} = \frac{w_i + \bar{w}_i}{2}$$

3 Glove Embedding

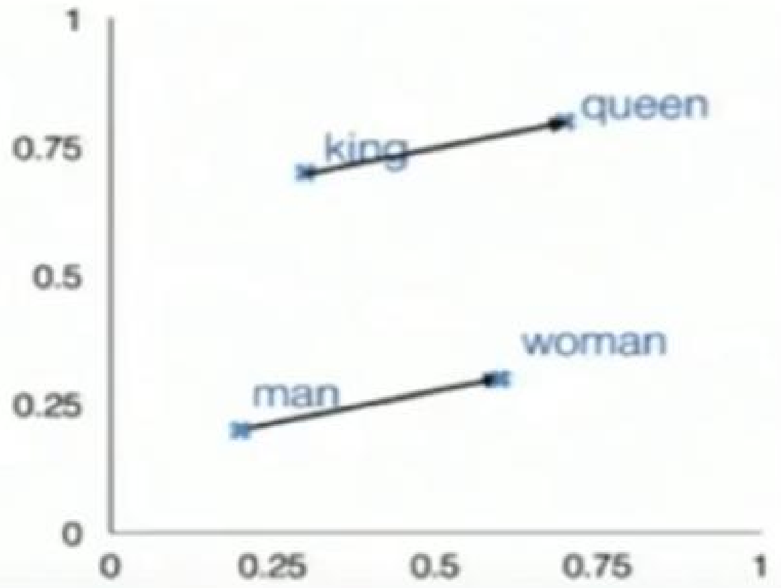
Biểu diễn Glove

3 Glove Embedding

Biểu diễn Glove

king - man + woman = queen

+	king	[0.30 0.70]
-	man	[0.20 0.20]
+	woman	[0.60 0.30]
	queen	[0.70 0.80]



3 Glove Embedding

Xử lý dữ liệu Glove: Nguồn dữ liệu

Glove

Glove Data:

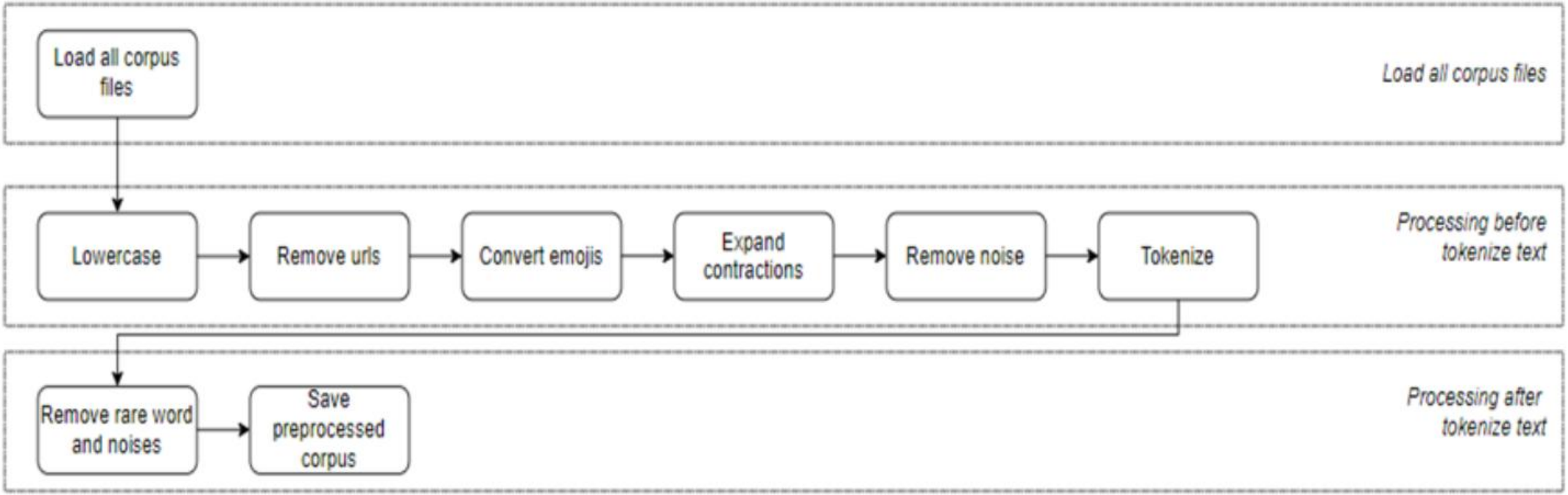
- **Nguồn dữ liệu:** Website Wortschatz Leipzig
- **Lượng dữ liệu:** 11 file - 10,3 triệu sentences
- **Người chia sẻ:** Dữ liệu được tổng hợp từ Internet
- **Loại dữ liệu:** News, News-typical, Newscrawl, Newscrawl-public, Web, Web-public, Wikipedia.
- **Thời gian:** 2007 - 2021

Cấu trúc:

- **id:** Mã định danh
- **sentence:** Văn bản

3 Glove Embedding

Xử lý dữ liệu Glove: Làm sạch dữ liệu



3 Glove Embedding

Xử lý dữ liệu Glove: Làm sạch dữ liệu

Sau khi tiền xử lý dữ liệu:

Vocabulary: 296772 vocabulary

Tokens: 164834340 tokens

File size: 1,36 GB.

```
['1,000', 'different', 'runners', 'from', 'over', '100', 'running', 'clubs', 'have', 'already', 'enjoyed',
['£10.00', 'per', 'person', 'which', 'you', 'can', 'pay', 'upon', 'arrival', 'at', 'the', 'gate']
['£10.00', 'per', 'tonne', 'going', 'to', 'local', 'worthy', 'charitie']
['£10.00', 'you', 'will', 'only', 'be', 'charged', 'this', 'if', 'your', 'application', 'is', 'approved']
['£100', 'billion', 'in', 'fact']
['£100', 'deposit', 'required', 'at', 'booking']
['10-0', 'does', 'have', 'a', 'nice', 'ring', 'to', 'it', 'it', 'is', 'double', 'figures', 'he', 'smiled']
['100', 'in', '100', 'is', 'supported', 'by', 'a', 'number', 'of', 'key', 'industry', 'stakeholders', 'inc
['£100', 'million', 'to', 'be', 'precise']
['£100', 'million', 'to', 'extend', 'the', 'biomedical', 'catalyst', 'fund', 'to', 'stimulate', 'the', 'tr
```

3 Glove Embedding

Kết quả sau khi train Glove

```
glove_model.word_vectors[glove_model.dictionary['phone']][:10]
```

```
array([ 0.31227752,  0.26768746, -0.08209904,  0.12328828,  0.17155725,
        0.18397549, -0.02893085, -0.16730548,  0.0406741 , -0.05937972])
```

Thư viện glove-python-binary

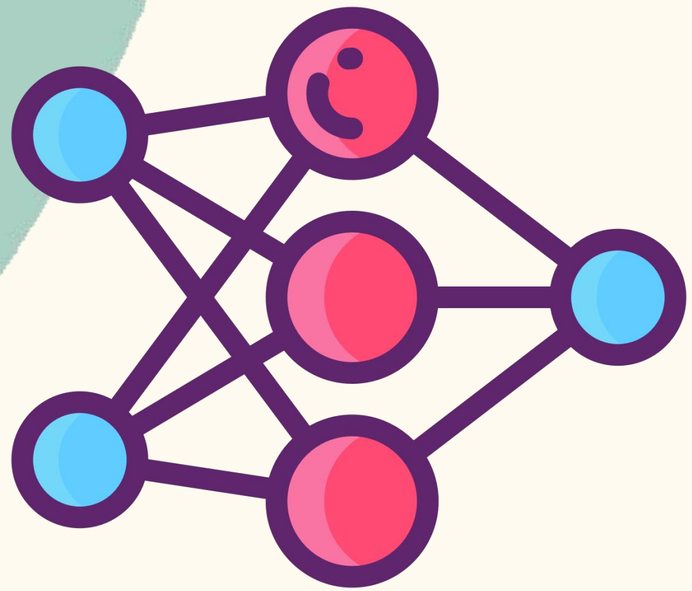
File pre-trained GloVe có 296772 từ được chuyển thành vector, file có kích thước 1,73 GB.

```
glove_model.most_similar('phone',number=10)
```

```
[('telephone', 0.602848154413826),
 ('phones', 0.602614000778562),
 ('calls', 0.602055564965323),
 ('mobile', 0.565655415136773),
 ('call', 0.542009275427078),
 ('cell', 0.492824547367746),
 ('tablet', 0.481108512176878),
 ('email', 0.450871757307091),
 ('number', 0.442318662260572)]
```

4

Xây dựng mô hình



4 Xây dựng mô hình

papers

Corpus → Clean corpus → Training → *glove-python-binary*

CAT ⇩ [0.23, 0.34, 0.65]

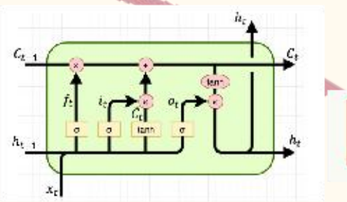
Glove model

Amazon

Rating data → Clean data → Labelled data (Negative, Neutral, Positive)

Training → LSTM (Keras)

Testing → Confusion Matrix



4 Xây dựng mô hình

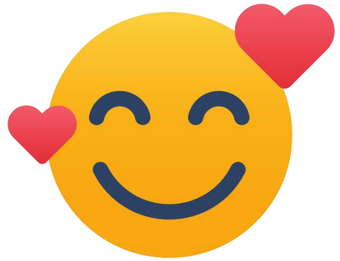
LSTM: Là mạng neuron RNN cải tiến

Cả RNN và LSTM đều giải quyết bài toán tính tuần tự

Ví dụ:

"I hate you, I don't love you"

"I love you, I don't hate you"

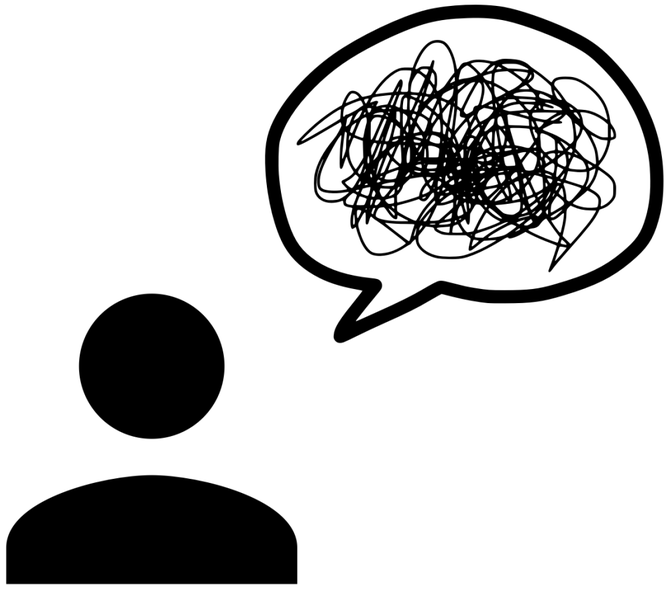


Thứ tự khác → Nghĩa khác

4 Xây dựng mô hình

Ưu điểm LSTM?

LSTM nhớ lâu hơn (long term) so với RNN



RNN

LSTM

4 Xây dựng mô hình

Xử lý dữ liệu LSTM: Nguồn dữ liệu

LSTM

amazon

LSTM Data 1: Amazon Reviews 400 000

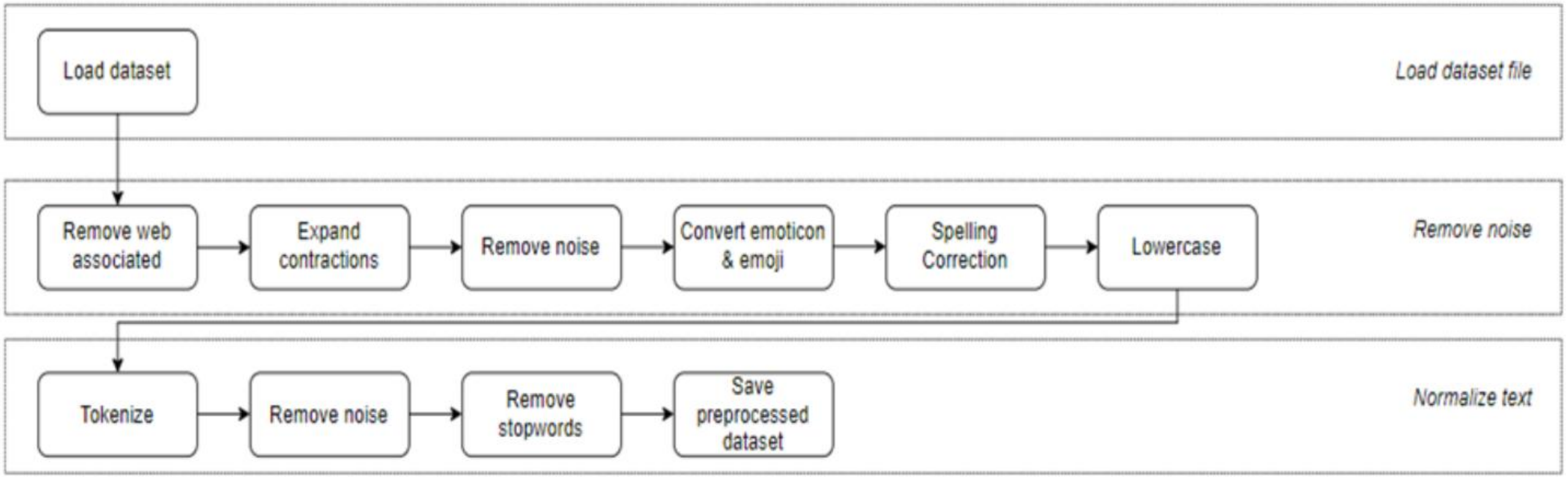
- **Tên dữ liệu:** “Amazon Reviews: Unlocked Mobile Phones”
- **Người chia sẻ:** PROMPTCLOUD
- **Tổng số dòng:** 400 000 dòng
- **Thời gian:** 9/2019

LSTM Data 2: Amazon Reviews 80 000

- **Tên dữ liệu:** “Amazon Cell Phone Review NLP ”
- **Người chia sẻ:** Wenling Yao
- **Tổng số dòng:** 80 000 dòng
- **Thời gian:** 12/2016

4 Xây dựng mô hình

Xử lý dữ liệu LSTM: Làm sạch dữ liệu



4 Xây dựng mô hình

Gán nhãn dữ liệu (LSTM)

POSITIVE

NEUTRAL

NEGATIVE

4 Xây dựng mô hình

Gán nhãn dữ liệu (LSTM)

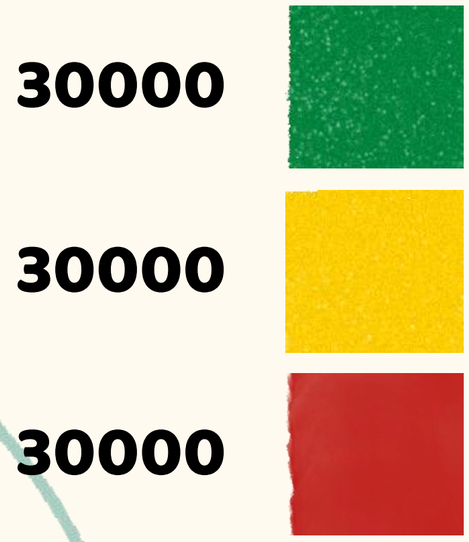
NEGATIVE + NEUTRAL

POSITIVE + NEUTRAL

4 Xây dựng mô hình

Gán nhãn dữ liệu (LSTM) - Cắt

Trước khi cắt



Sau khi cắt

4 Xây dựng mô hình

Dữ liệu cuối cùng

Trước xử lý

"The camera is very slow :("

"I was happy I got a case and a charger that ..."

"I bought this phone for my wife and I wanted ..."

Sau xử lý

"camera slow"

"happy got case charger ..."

"bought phone wife wanted..."

4 Xây dựng mô hình

Tham số

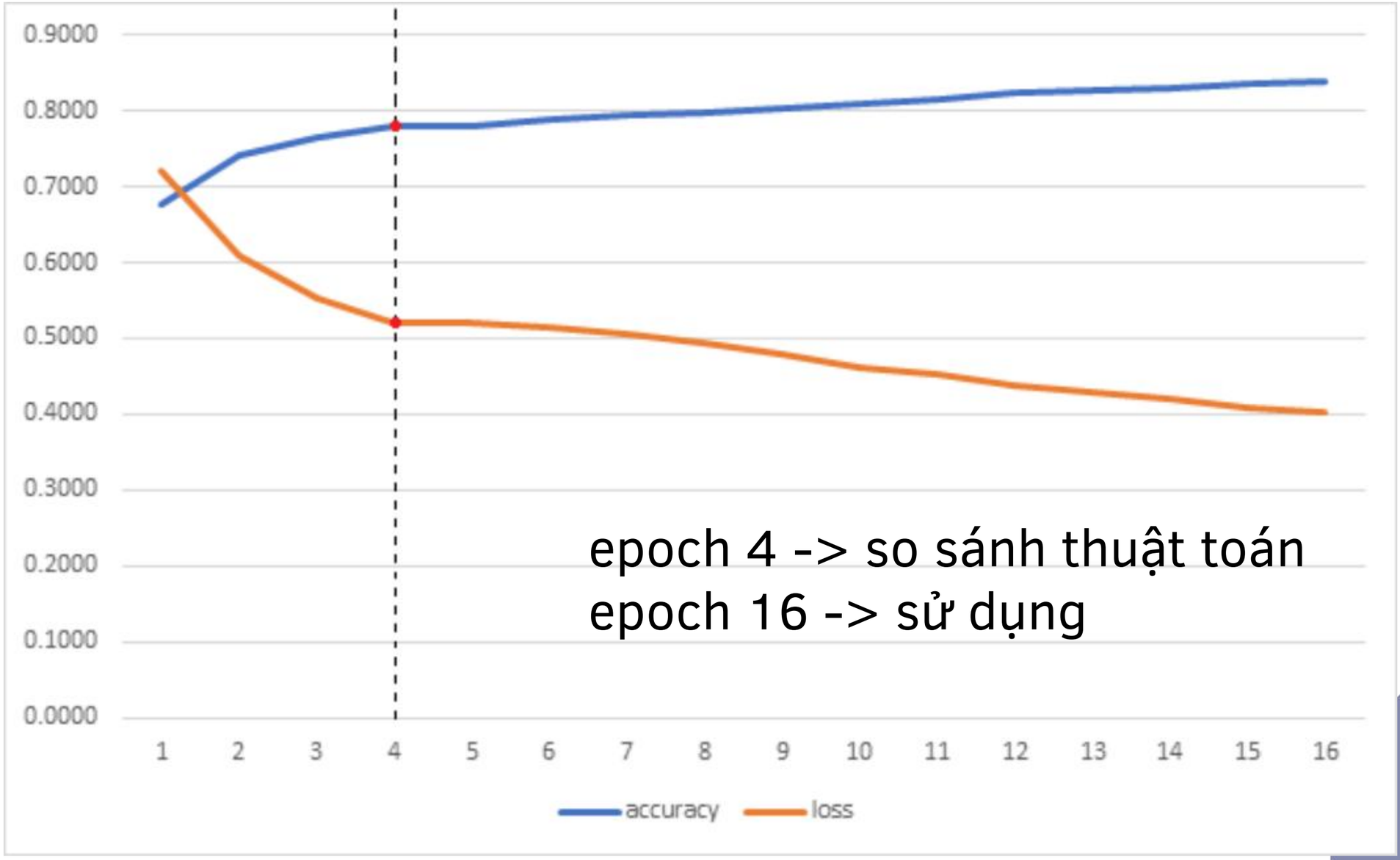
Hidden_nodes = 150
Dropout = 0.2
Recurrent_dropout = 0.2
Activation_function = softmax
epochs = 16

Kết quả

Accuracy: 0.8263, **Time:** 77628s, **Average time:** 4851.75s / epoch

		Negative	Neutral	Positive
Confusion matrix	**Negative**	5049	938	122
	Neutral	783	4470	653
	Positive	110	520	5355

4 Xây dựng mô hình



epoch 4 -> so sánh thuật toán
epoch 16 -> sử dụng

4 Xây dựng mô hình

So sánh Glove với BoW, Word2Vec

BoW (Bag Of Words)

- Vector hóa đoạn văn
- Vector hóa dựa trên tần suất
- Không quan tâm thứ tự

A = [“I love NLU”]

B = [“I love to be a student of NLU NLU”]

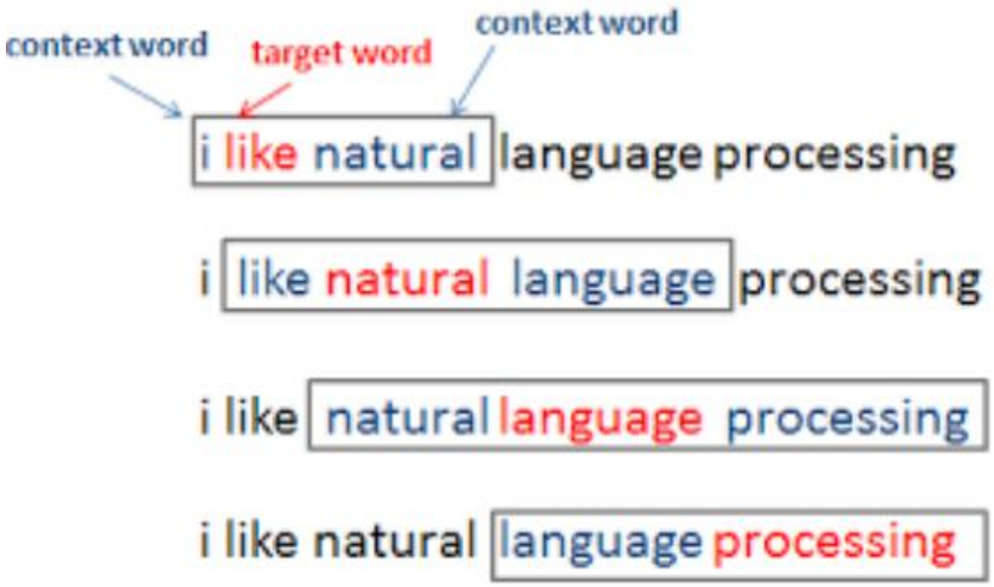
	nlu	i	love	to	be	a	student	of
A	1	1	1	0	0	0	0	0
B	2	1	1	1	1	1	1	1

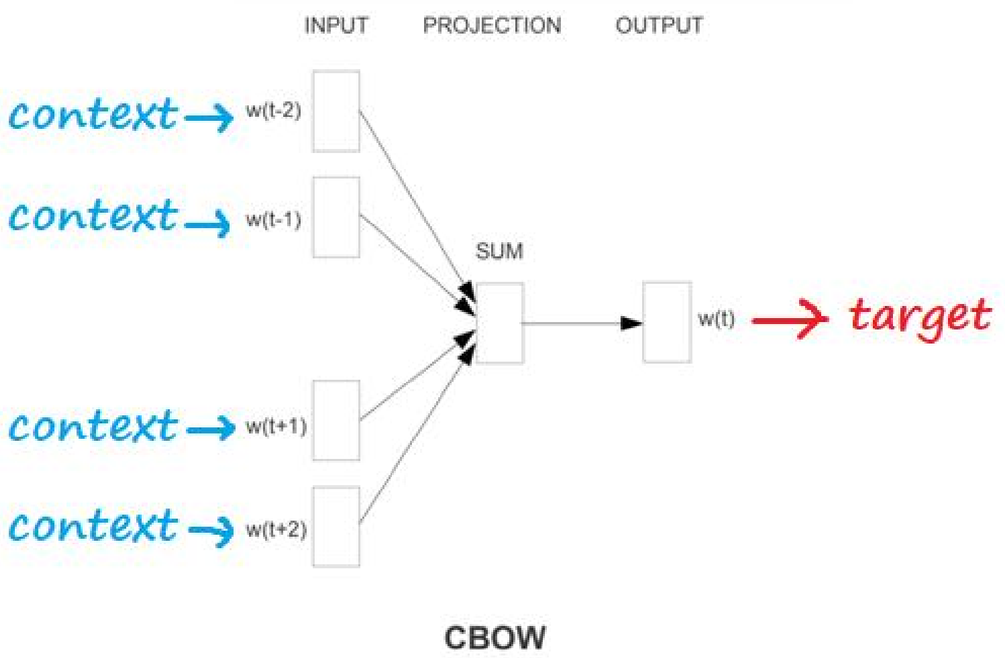
4 Xây dựng mô hình

So sánh Glove với BoW, Word2Vec

Word2Vec

- Vector hóa từ
- Tương tự Glove





4 Xây dựng mô hình

So sánh Glove với BoW, Word2Vec

	Glove (300D4E)	Word2Vec (300D4E)	BoW
Accuracy	0.7955	0.7888	0.3281
Thời gian train	Chậm	Trung bình	Nhanh
Loại mô hình	Count-based	Predictive	Count-based
Loại vector hóa	Từ (word)	Từ (word)	Văn bản (Doc)
Sử dụng mạng neuron	Không	Có	Không

4 Xây dựng mô hình

So sánh Glove với Pretrained-Glove

Dữ liệu

Glove	Pretrained Glove
- Tự xây dựng	- Đã có sẵn
- 164 triệu tokens	- 6 tỷ tokens
- 296 nghìn từ vựng	- 400 nghìn từ vựng
- Nguồn từ các bài báo + Wikipedia	- Nguồn Wikipedia + Gigaword

4 Xây dựng mô hình

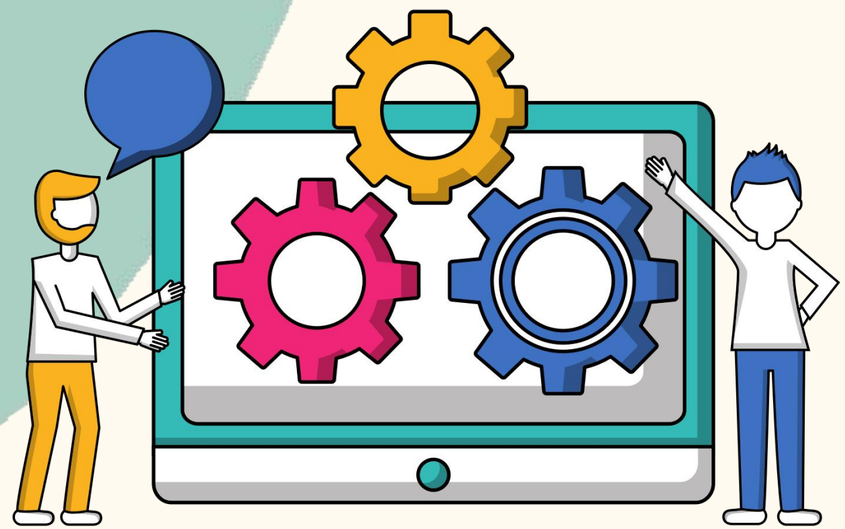
So sánh Glove với Pretrained-Glove

Kết quả so sánh (Thông qua LSTM)

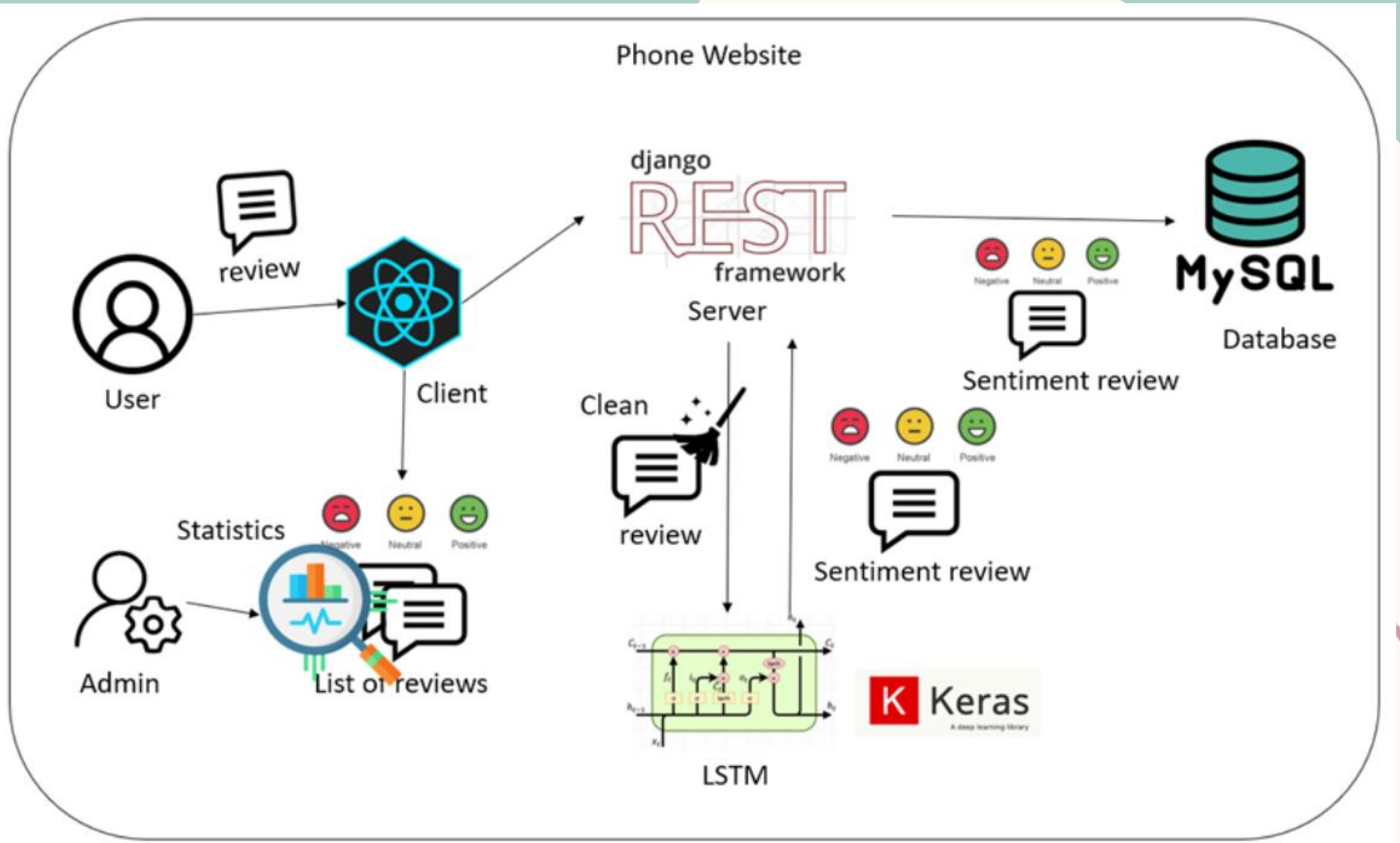
	Accuracy	
Dimension	**Trained glove (4E)**	**Pretrained glove (4E)**
50D	0.7083	0.7326
100D	0.7252	0.7652
200D	0.7764	0.7733
300D	0.7955	0.8058

5

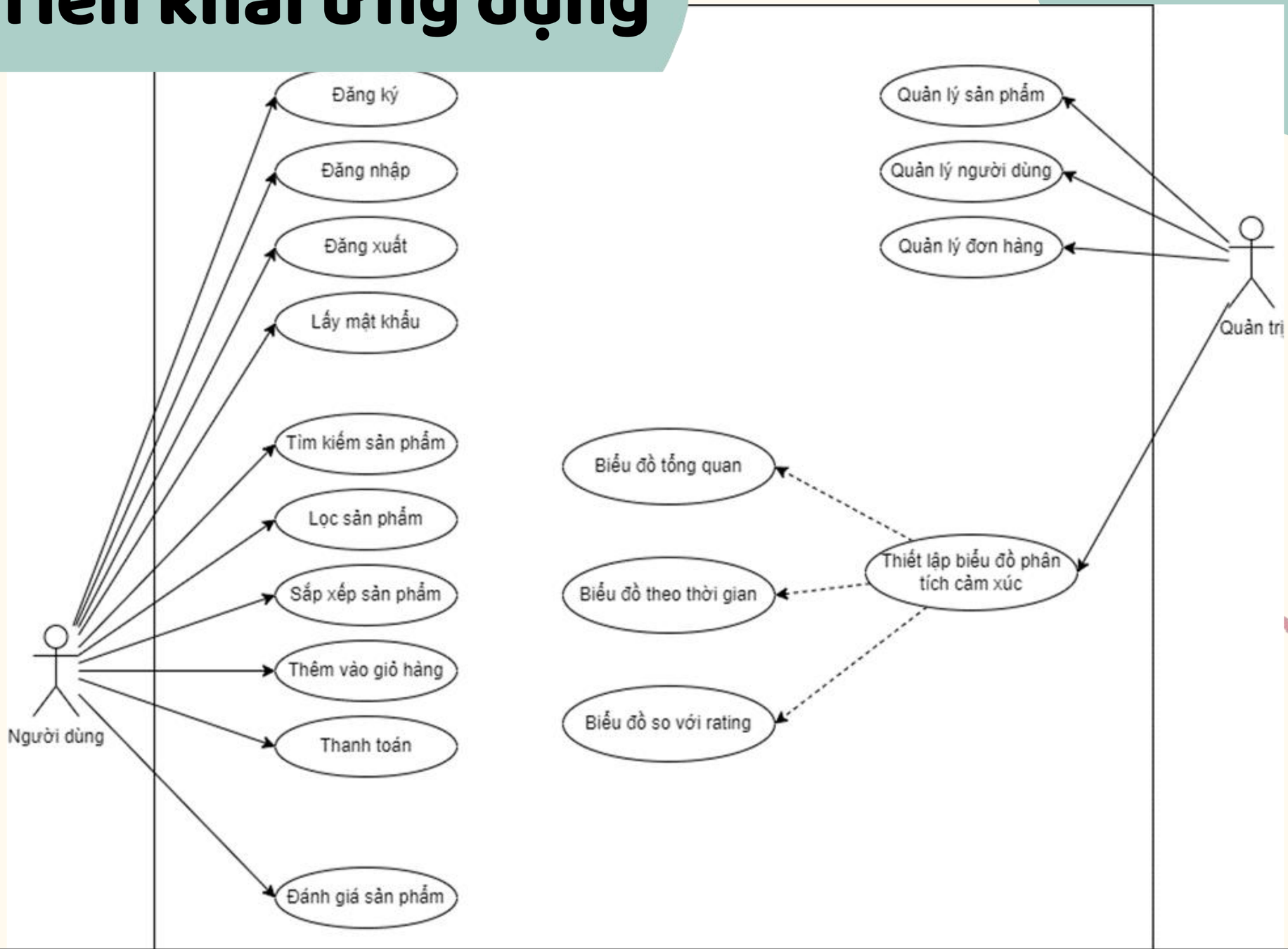
Triển khai ứng dụng



5 Triển khai ứng dụng



5 Triển khai ứng dụng



5 Triển khai ứng dụng

PHONE SHOP

Search

QUANG TIEN

VIVO T1 44W (STARRY SKY, 128 GB)

4 reviews

Price: $20.00 Status: In Stock

Desc: 4 GB RAM | 128 GB ROM | Expandable Upto 1 TB16.36 cm (6.44 inch) Full HD+ AMOLED Display50MP + 2MP + 2MP | 16MP Front Camera5000 mAh Lithium BatteryQualcomm Snapdragon 680 Processor1 Year Handset and 6 Months Accessories

Quantity 1

ADD TO CART

REVIEWS

Quang Tien

2021-03-04

I bought this A13 LTE as a replacement for my wife whose Samsung J2 that was starting to have a failure making and receiving calls (a still essential feature for us in a phone (smart or otherwise)). My wife said to me when considering a replacement, "I just want a phone that works". Pretty easy criteria for any phone, but useful to information to consider when factoring in my rating and review.She uses her smartphone primarily for calls, texting, Internet browsing, social media apps and the occasional lightweight game (like Solitaire). She snaps pictures here and there with her phone but doesn't expect or require great quality from her phone's camera. No surprise, this phone works great for all of those uses and she has no complaints about its performance with any of those functions - even when she ends up with five or more applications running in the background that she forgot to close.I consider this phone a great value because of the relatively low price, her satisfaction, and a

POSITIVE (93.92%)

Demo

Cảm ơn thầy cô đã lắng nghe!

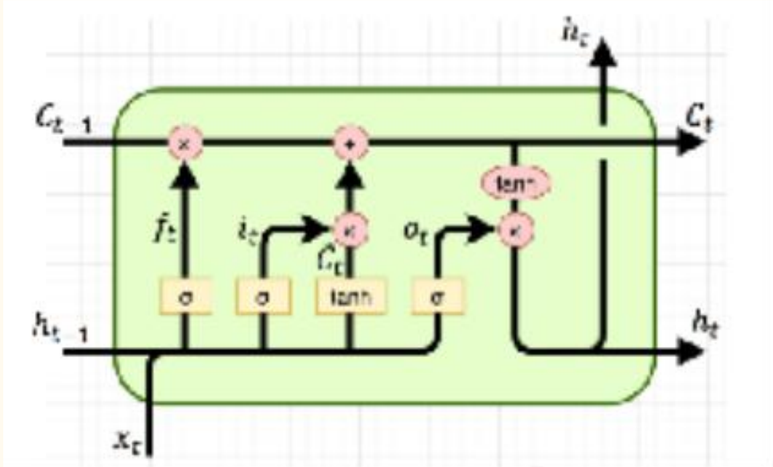
Phụ lục

Nội dung chính

CAT

⇩

[0.23, 0.34, 0.65]



Glove

LSTM

Phone Website Server

Glove embedding

Glove (Global Vector)

2014 - các nhà nghiên cứu ĐH Stanford: Jeffrey Pennington, Richard Socher, Christopher D. Manning.

Glove tối ưu vì kết hợp cả 2 phương pháp:

- Global matrix factorization
- Local context window

Mang tính toàn cục (**global**) và cục bộ (**local**)

Co-occurrence matrix

Corpus

I enjoy flying

I like NLP

I like deep learning

window size = 1

I : enjoy(1 time), like(2 times)
enjoy : I (1 time), flying(2 times)
flying : enjoy(1 time)
like : I(2 times), NLP(1 time), deep(1 time)
NLP : like(1 time)
deep : like(1 time), learning(1 time)
learning : deep(1 time)

$$X = \begin{array}{c} \\ I \\ like \\ enjoy \\ deep \\ learning \\ NLP \\ flying \\ . \end{array} \begin{array}{c} \begin{array}{ccccccc} I & like & enjoy & deep & learning & NLP & flying \end{array} \\ \begin{bmatrix} 0 & 2 & 1 & 0 & 0 & 0 & 0 \\ 2 & 0 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ \end{bmatrix} \end{array}$$

Co-occurrence matrix

like: [2,0,0,1,0,1,0]

enjoy: [1,0,0,0,0,0,1]

cos(like, enjoy) = 0.57

-> Nắm bắt được mối quan hệ gần nhau của vector

$$X = \begin{array}{c} \\ I \\ like \\ enjoy \\ deep \\ learning \\ NLP \\ flying \\ . \end{array} \begin{array}{c} \begin{array}{ccccccc} I & like & enjoy & deep & learning & NLP & flying \end{array} \\ \begin{bmatrix} 0 & 2 & 1 & 0 & 0 & 0 & 0 \\ 2 & 0 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ \end{bmatrix} \end{array}$$

Số chiều của các vectors lớn

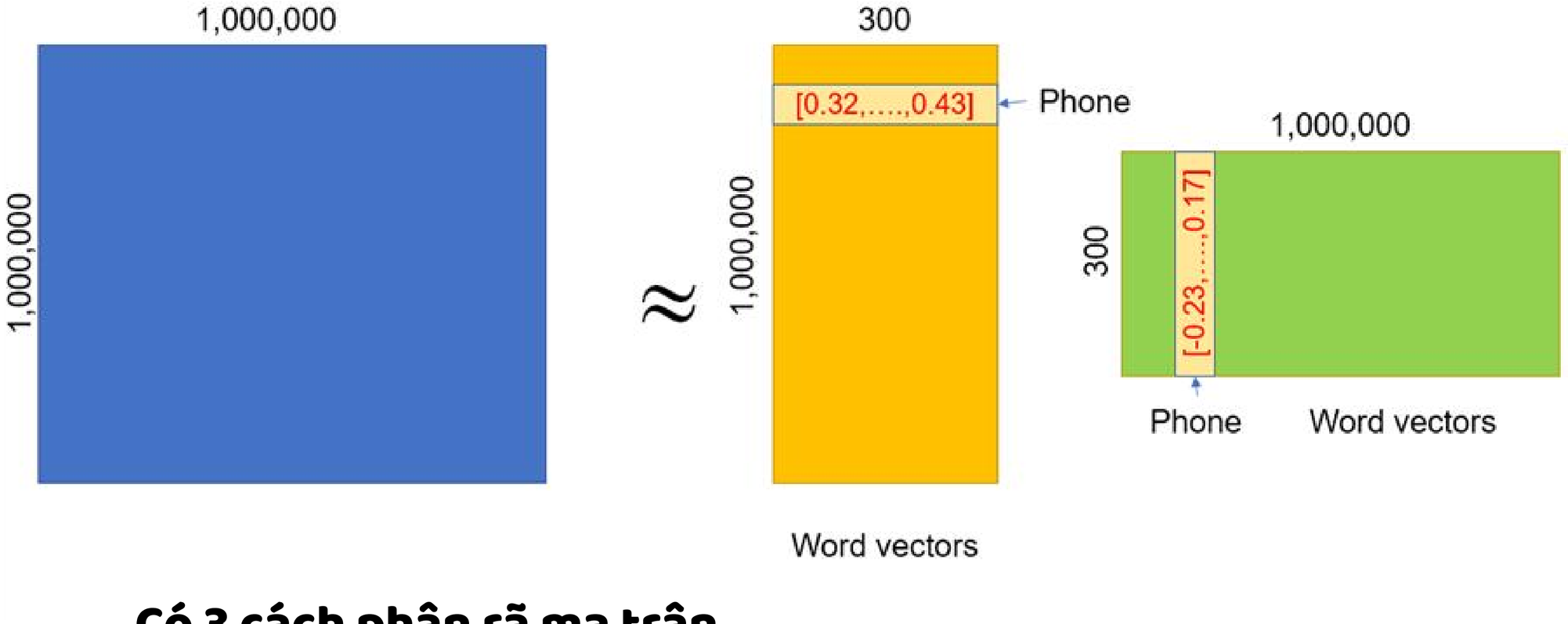
-> Tốn bộ nhớ

Ma trận còn chứa nhiều số 0

-> Giảm số chiều ma trận

-> Phân rã ma trận

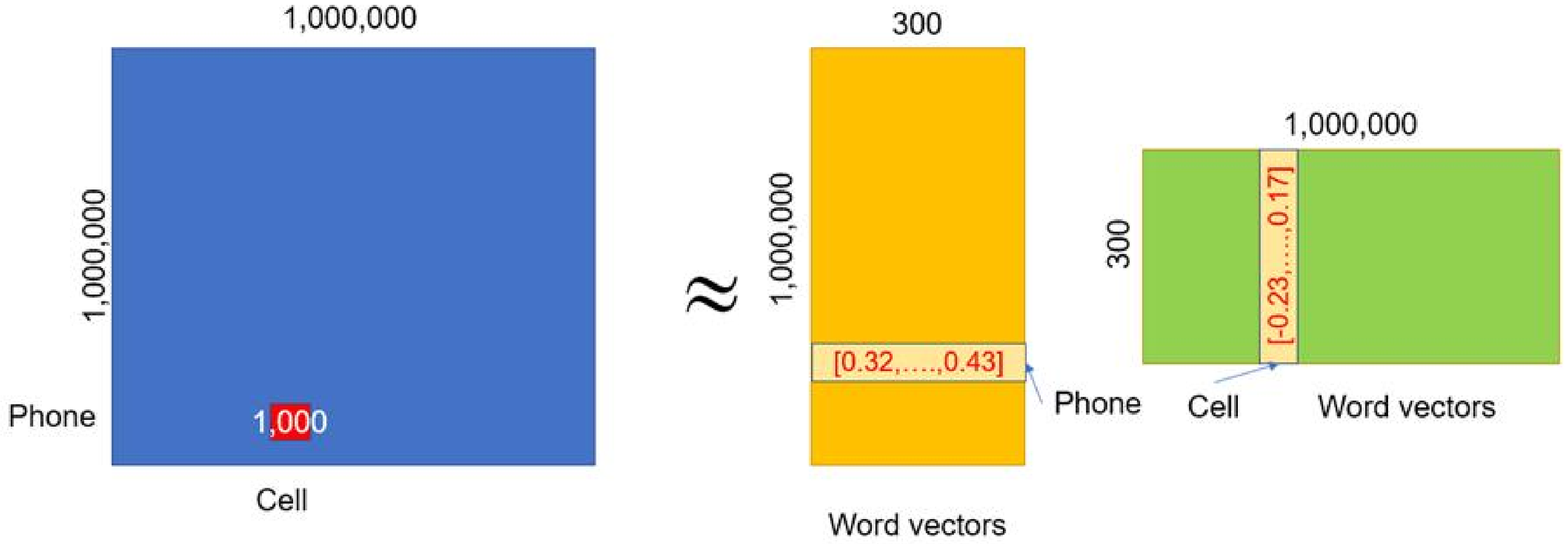
Phân rã ma trận (matrix factorization)



Có 3 cách phân rã ma trận

- **Singular Value Decomposition (SVD)**
- **Neural Language Model**
- **Brown clustering**

Phân rã ma trận (matrix factorization)



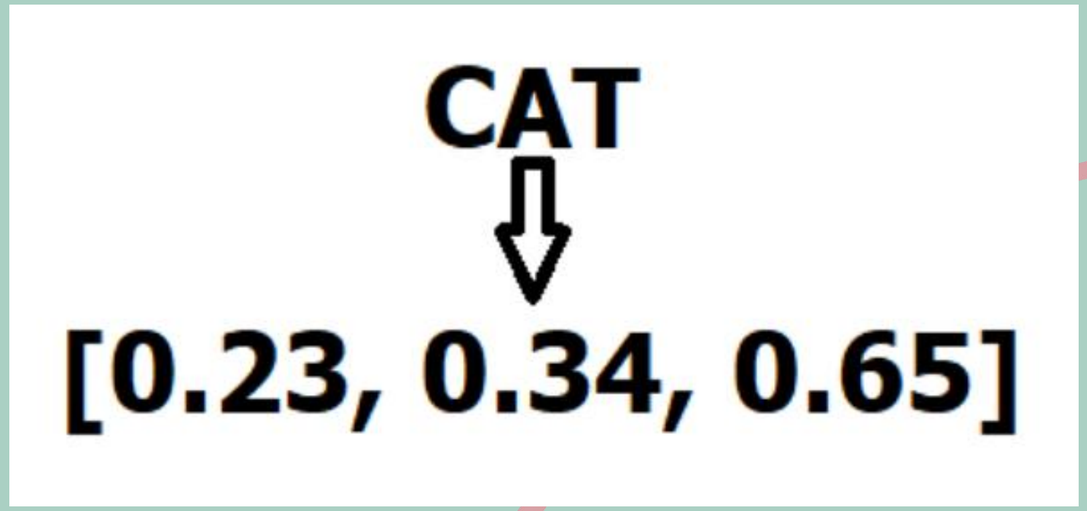
Phân rã số 1000 đó thành tích vô hướng (dot production) của hai vector phone và cell

Phương pháp này được dùng trong nhiều thuật toán word embedding: LSA, HAL,...

So sánh

Global matrix factorization (LSA, HAL,..)	Local context window (Word2vec)
Thời gian huấn luyện nhanh (nếu ngữ liệu không quá lớn)	Tăng lên theo độ lớn ngữ liệu
Sử dụng hiệu quả số liệu thống kê	Sử dụng kém hiệu quả số liệu thống kê
Chủ yếu được sử dụng để nắm bắt sự giống nhau của từ	Nắm bắt được các mối quan hệ khác giữa các từ ngoài sự giống nhau
Bị nhiễu đối với số lần xuất hiện lớn (các từ the, a, an,…)	Cải thiện chung hiệu suất trên các nhiệm vụ khác

Glove embedding



glove-python-binary

Co-occurrence matrix

genius is one percent inspiration ninety nine percent perspiration

Target Word and Corresponding Window	Elememt		Count
genius is one percent inspiration ninety nine percent perspiration	$X_{genius,is}$,	$X_{is,genius}$	+1
	$X_{genius,one}$,	$X_{one,genius}$	+1/2
	$X_{genius,percent}$,	$X_{percent,genius}$	+1/3
	$X_{genius,inspiration}$,	$X_{inspiration,genius}$	+1/4

Window size = 4 (local)

Count	genius	is	one	percent	inspiration	ninety	nine	perspiration
genius	0	1	½	1/3	¼	0	0	0
is	1	0	0	0	0	0	0	0
one	½	0	0	0	0	0	0	0
percent	1/3	0	0	0	0	0	0	0
inspiration	¼	0	0	0	0	0	0	0
ninety	0	0	0	0	0	0	0	0
nine	0	0	0	0	0	0	0	0
perspiration	0	0	0	0	0	0	0	0

Co-occurrence matrix

genius is one percent inspiration ninety nine percent perspiration

Window size = 4 (local)

Target Word and Corresponding Window	Elememt		Count
genius **is** one percent inspiration ninety nine percent perspiration	$X_{genius,is}$,	$X_{is,genius}$	+1
	$X_{genius,one}$,	$X_{one,genius}$	+1/2
	$X_{genius,percent}$,	$X_{percent,genius}$	+1/3
	$X_{genius,inspiration}$,	$X_{inspiration,genius}$	+1/4
genius **is** **one** **percent** **inspiration** **ninety** nine percent perspiration	$X_{is,one}$,	$X_{one,is}$	+1
	$X_{is,percent}$,	$X_{percent,is}$	+1/2
	$X_{is,inspiration}$,	$X_{inspiration,is}$	+1/3
	$X_{is,ninety}$,	$X_{ninety,is}$	+1/4
genius is one percent inspiration ninety nine **percent** **perspiration**	$X_{percent,perspiration}$,	$X_{perspiration,percent}$	+1
genius is one percent inspiration ninety nine percent **perspiration**	Done		

Co-occurrence matrix

genius is one percent inspiration ninety nine percent perspiration

Co-occurrence matrix (global)

Count	genius	is	one	percent	inspiration	ninety	nine	perspiration
genius	0	1	½	1/3	¼	0	0	0
is	1	0	1	½	1/3	¼	0	0
one	½	1	0	1	½	1/3	¼	0
percent	1/3	½	1	½	4/3	1	4/3	1
inspiration	¼	1/3	½	4/3	0	1	½	¼
ninety	0	¼	1/3	1	1	0	1	1/3
nine	0	0	¼	4/3	½	1	0	½
perspiration	0	0	0	1	¼	1/3	½	0

Công thức

genius is one percent inspiration ninety nine percent perspiration

$$X_{\text{genius, percent}} = 1/3$$

$$X_{\text{genius}} = 0+1+ ½ + 1/3 + ¼ +0+0+0 = 25/12$$

$$P(\text{percent}|\text{genius}) = \frac{X_{genius,percent}}{X_{genius}} = \frac{1/3}{25/12} = \frac{4}{25} \approx 0.16$$

Count	genius	is	one	percent	inspiration	ninety	nine	perspiration
genius	0	1	½	1/3	¼	0	0	0
is	1	0	1	½	1/3	¼	0	0
one	½	1	0	1	½	1/3	¼	0
percent	1/3	½	1	½	4/3	1	4/3	1
inspiration	¼	1/3	½	4/3	0	1	½	¼
ninety	0	¼	1/3	1	1	0	1	1/3
nine	0	0	¼	4/3	½	1	0	½
perspiration	0	0	0	1	¼	1/3	½	0

Công thức

$$F(w_i, w_k) = P(k|i)$$ ✘

Probability and Ratio	$k = solid$	$k = gas$	$k = water$	$k = fashion$
$P(k\|ice)$	1.9×10^{-4}	6.6×10^{-5}	3.0×10^{-3}	1.7×10^{-5}
$P(k\|steam)$	2.2×10^{-5}	7.8×10^{-4}	2.2×10^{-3}	1.8×10^{-5}
$P(k\|ice)/P(k\|steam)$	8.9	8.5×10^{-2}	1.36	0.96

Quá nhỏ x

Quá nhỏ x

Ổn v

P(solid|ice)/P(solid|steam) = 8.9 >1 → solid liên với ice hơn

P(gas|ice)/P(gas|steam) = $8.5.10^{-2} < 1$ →gas liên với steam hơn

P(water|ice)/P(water|steam) = 1.36 ≈ 1 →water đều liên quan với ice, steam

P(fashion|ice)/P(fashion|steam) = 0.96 ≈ 1 →fashion đều không liên quan đến ice, steam

Nên dùng tỷ lệ xác suất

Công thức

$$F(w_i, w_j, \bar{w}_k) = \frac{P(k|i)}{P(k|j)}$$

Công thức

$$w^T_i \bar{w}_k + b_i + \bar{b}_k = \log(X_{ik})$$

Trong đó:

i: từ đang xét

k: từ ngữ cảnh

X: ma trận đồng xuất hiện

w: vectors từ

$\bar{w}$: vectors của từ ngữ cảnh

b: bias

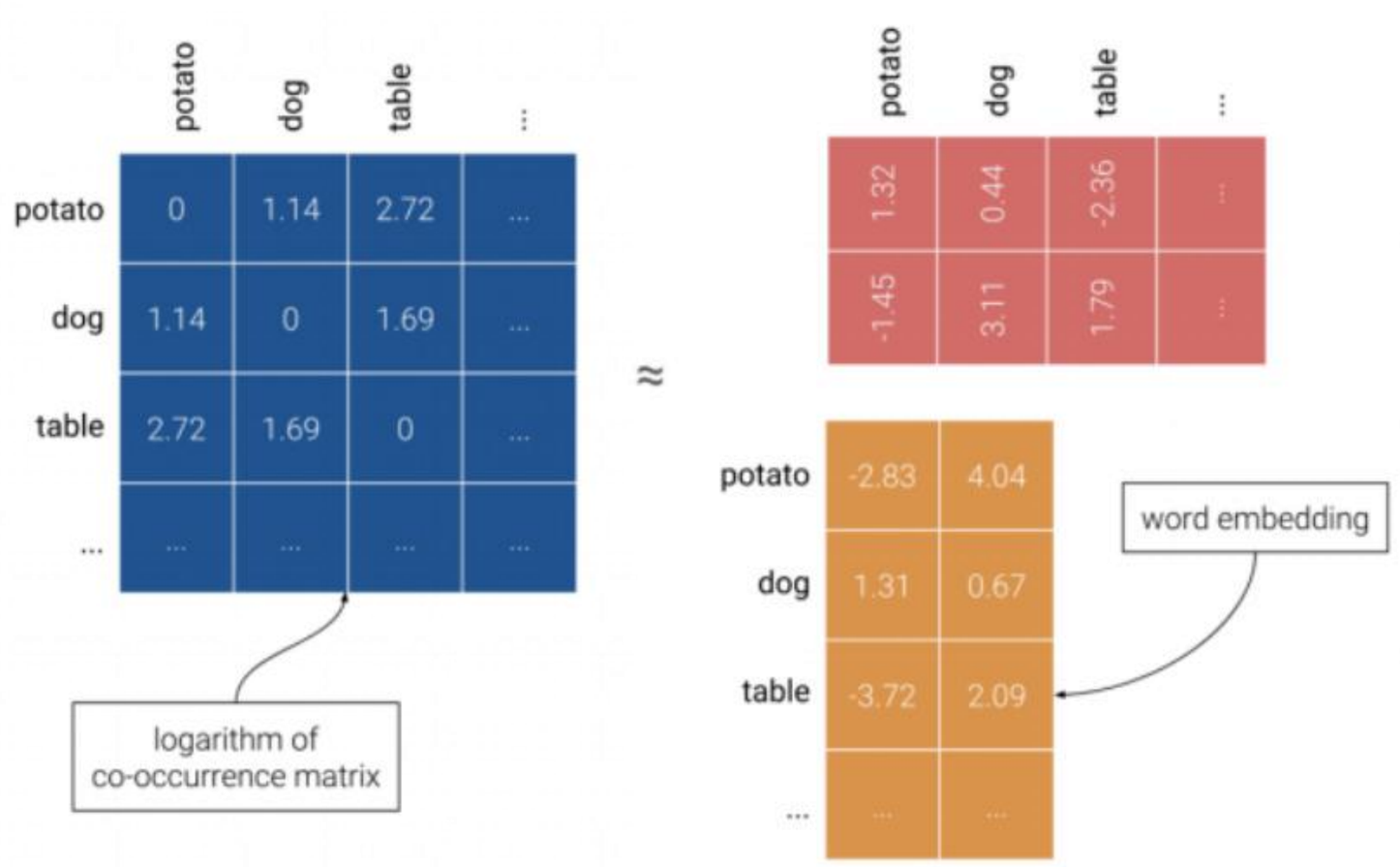
$\bar{b}$: bias

Global log-bilinear regression

$$w_i^{\text{final}} = \frac{w_i + \bar{w}_i}{2}$$

Công thức

$$w^T_i \bar{w}_k + b_i + \bar{b}_k = \log(X_{ik})$$



Global matrix factorization

Công thức

$$w^{T}_{i}\bar{w}_{k} + b_{i} + \bar{b}_{k} = \log(X_{ik})$$

Hàm Loss:

$$J(w_i, \bar{w}_j) = \sum_{i,j=1}^{n} f(X_{ij}) \left(w^{T}\bar{w}_j + b_i + \bar{b}_j - \log(X_{ij})\right)^2$$

$$w_i^{\text{final}} = \frac{w_i + \bar{w}_i}{2}$$

weighted least squares

Trong đó:

$$f(X_{ij}) = \begin{cases} \left(\frac{X_{ij}}{x_{max}}\right)^{\alpha} & \text{if } X_{ij} < X_{max} \\ 1 & \text{otherwise} \end{cases}$$

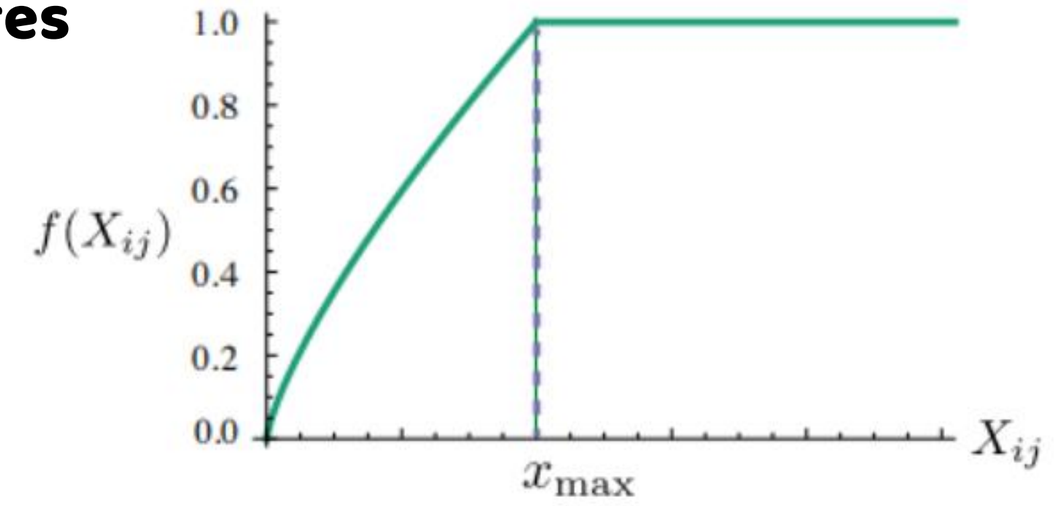
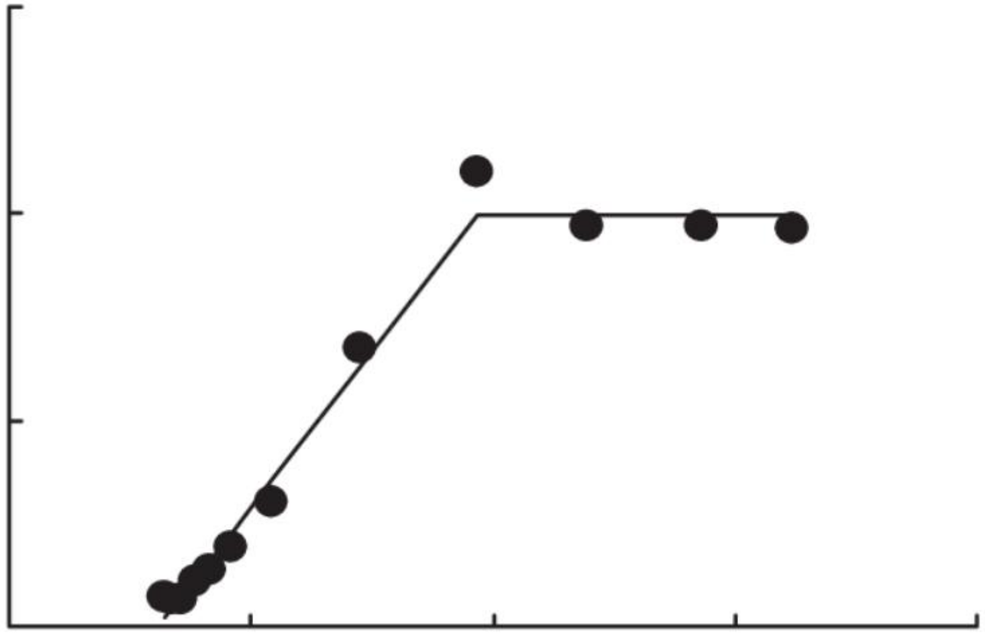
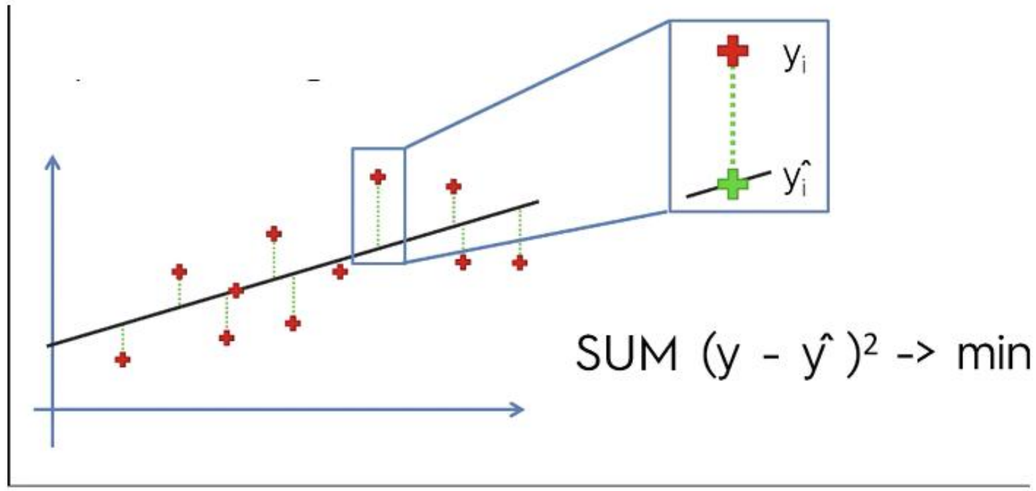


Figure 1: Weighting function f with $\alpha = 3/4$.

Công thức



bilinear regression



weighted least squares

Glove (Global Vector)

Glove dùng **global log-bilinear regression model** để tạo ra cách biểu diễn vector mang ý nghĩa tổng quát

Glove sử dụng **weighted least squares method** để huấn luyện mô hình

Kết quả

File pre-trained GloVe có 296772 từ được chuyển thành vector, file có kích thước 1,73 GB.

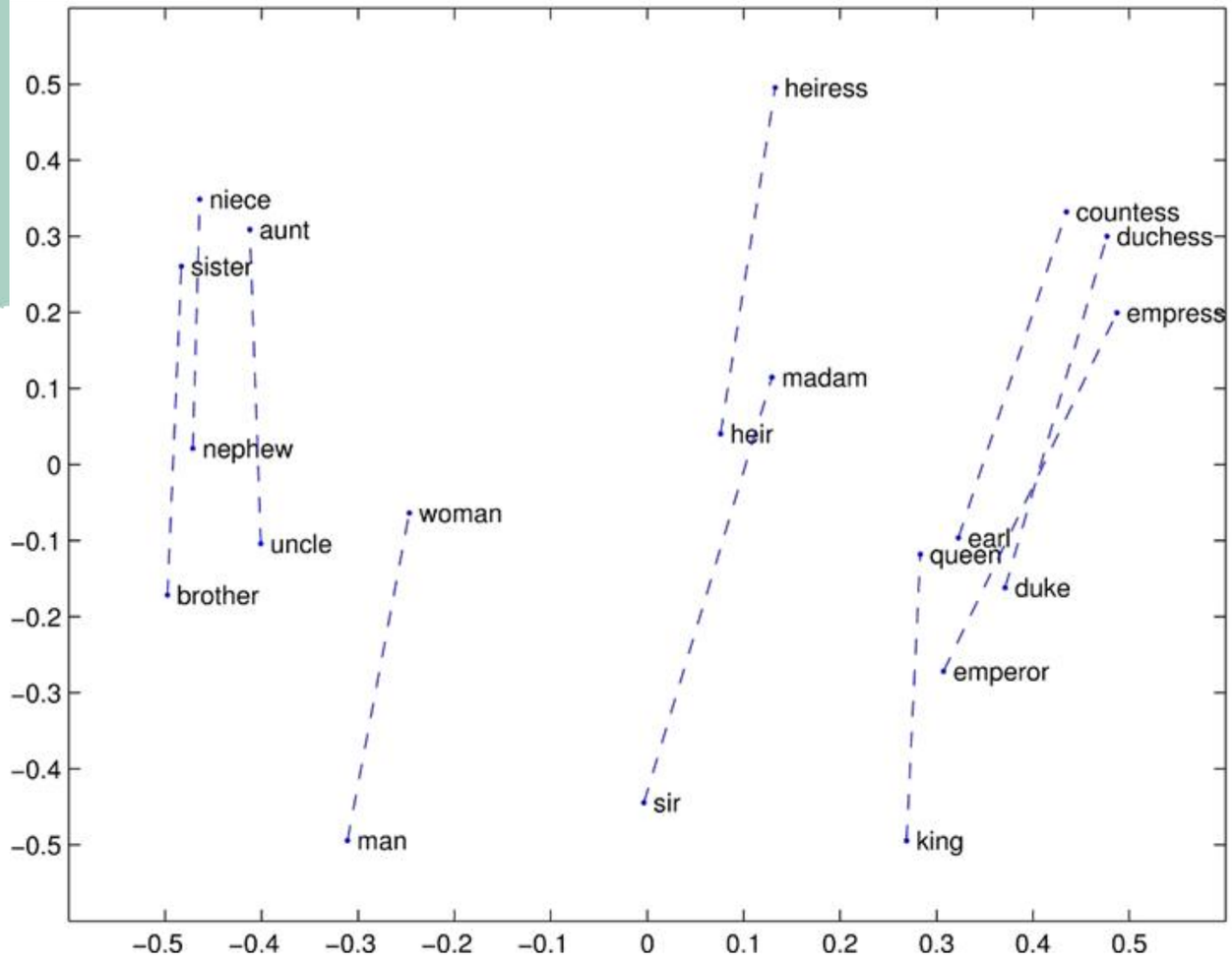
```
glove_model.word_vectors[glove_model.dictionary['phone']][:10]
```

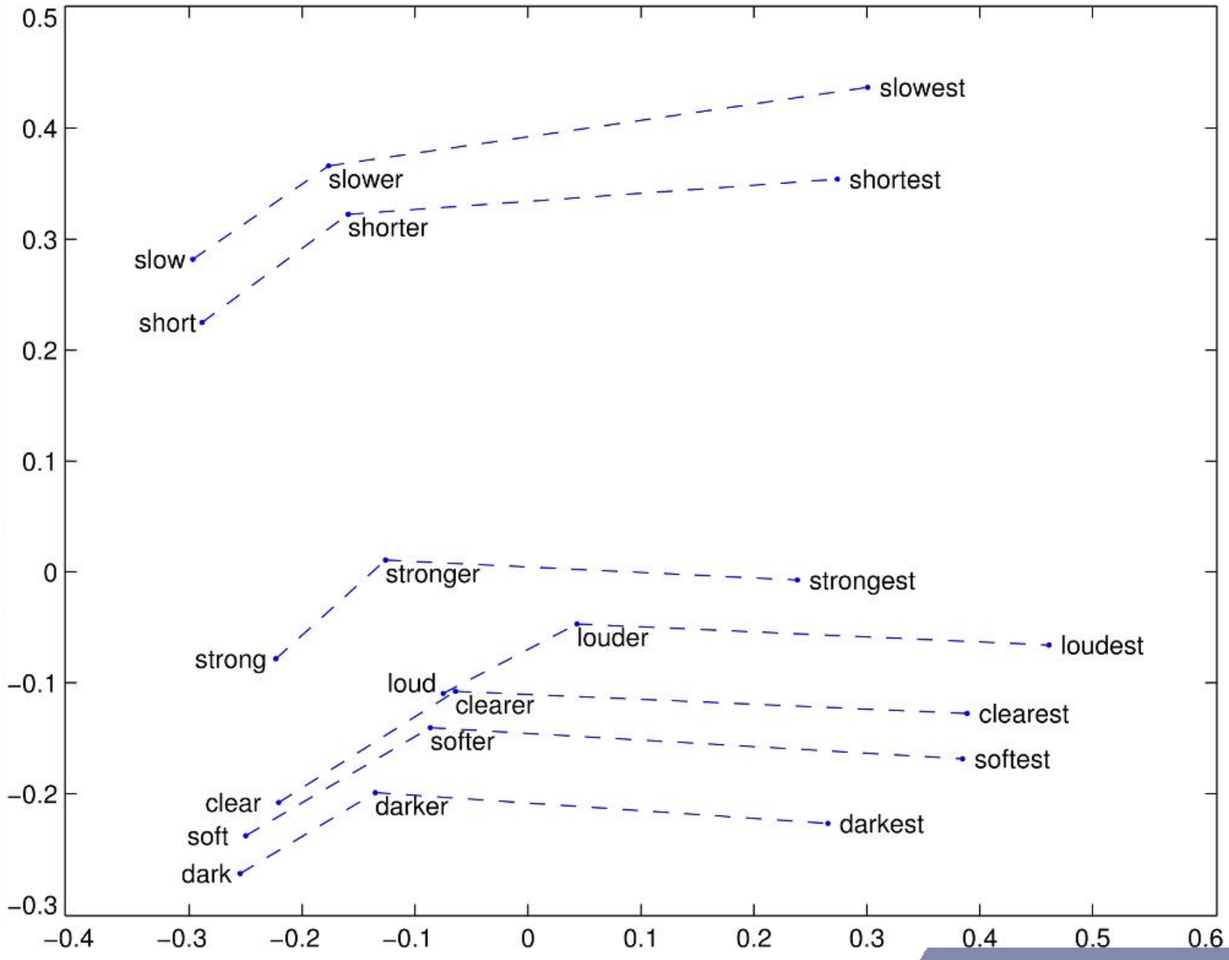
```
array([ 0.31227752,  0.26768746, -0.08209904,  0.12328828,  0.17155725,
        0.18397549, -0.02893085, -0.16730548,  0.0406741 , -0.05937972])
```

```
glove_model.most_similar('phone',number=10)
```

```
[('telephone', 0.6028481544138266),
 ('phones', 0.6026140007785626),
 ('calls', 0.602055649653235),
 ('mobile', 0.5656554151367738),
 ('call', 0.542009275427078),
 ('cell', 0.492824547367746 4),
 ('tablet', 0.4811085121768788),
 ('email', 0.4508717573070919 5),
 ('number', 0.442318662260572 3)]
```

Mối quan hệ giữa các từ

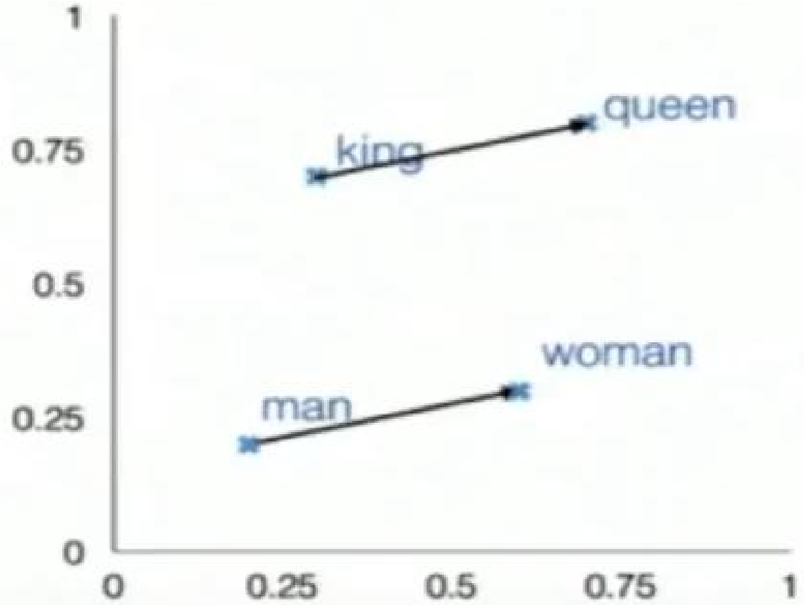




Biểu diễn toán học

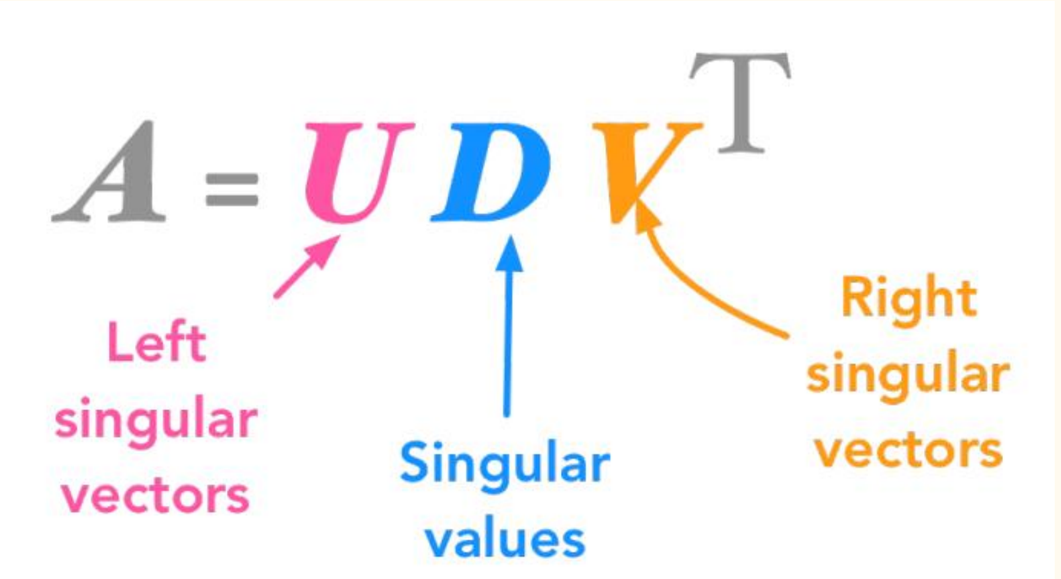
King - man + woman

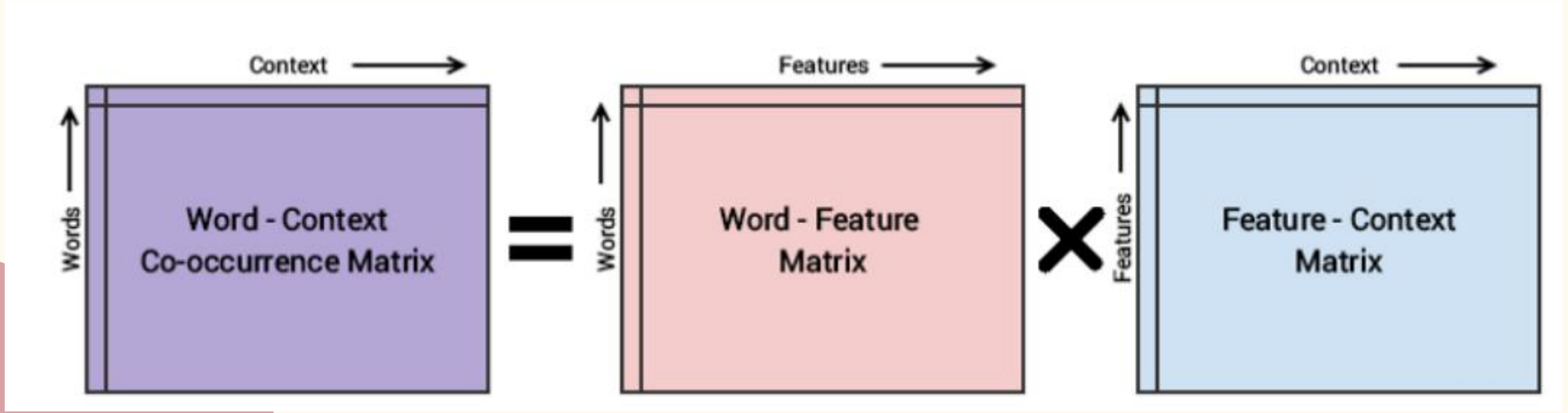
+	king	[0.30 0.70]
-	man	[0.20 0.20]
+	woman	[0.60 0.30]
	queen	[0.70 0.80]



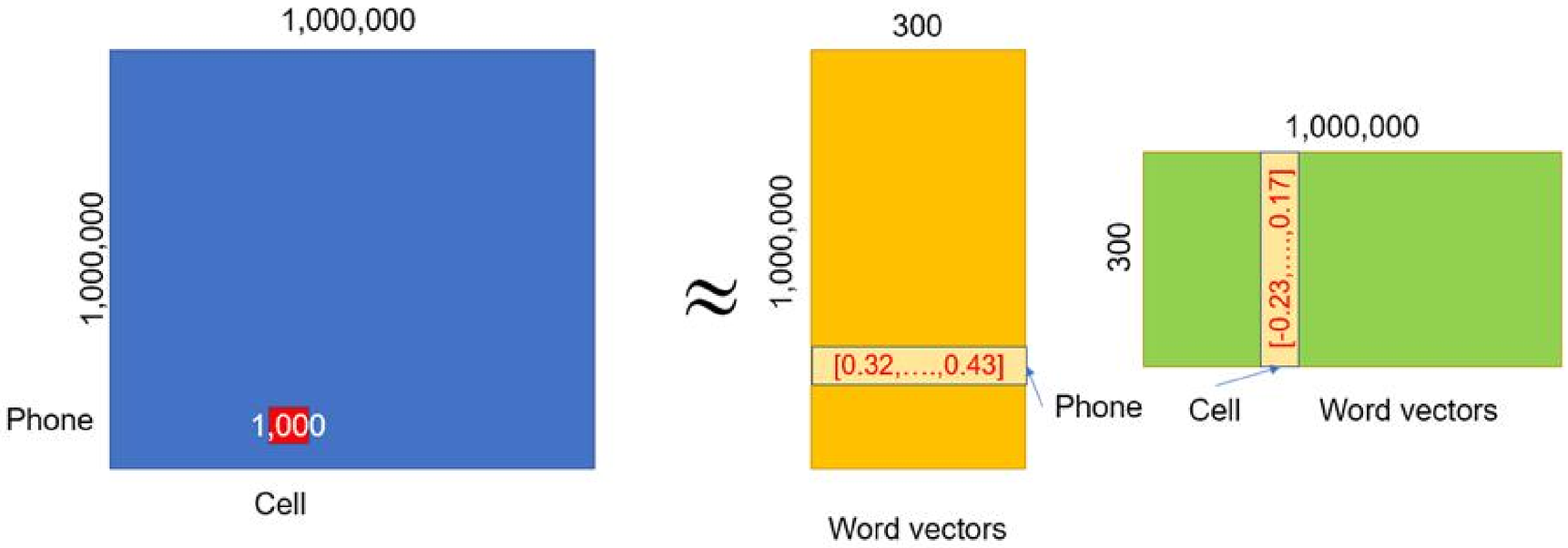
Matrix Factorization Methods

Reduce dimensionality: SVD





3 Ma trận đồng xuất hiện (Co-occurrence matrix)

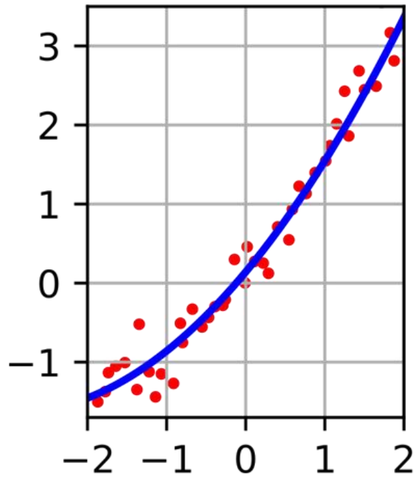


$$w_i^T \bar{w}_j = X_{ij}$$

w_i là vector của từ phone

$\bar{w}_j$ là vector của từ cell

X_{ij} là số lần phone cell cùng xuất hiện



$$w_i^{\text{final}} = \frac{w_i + \bar{w}_i}{2}$$

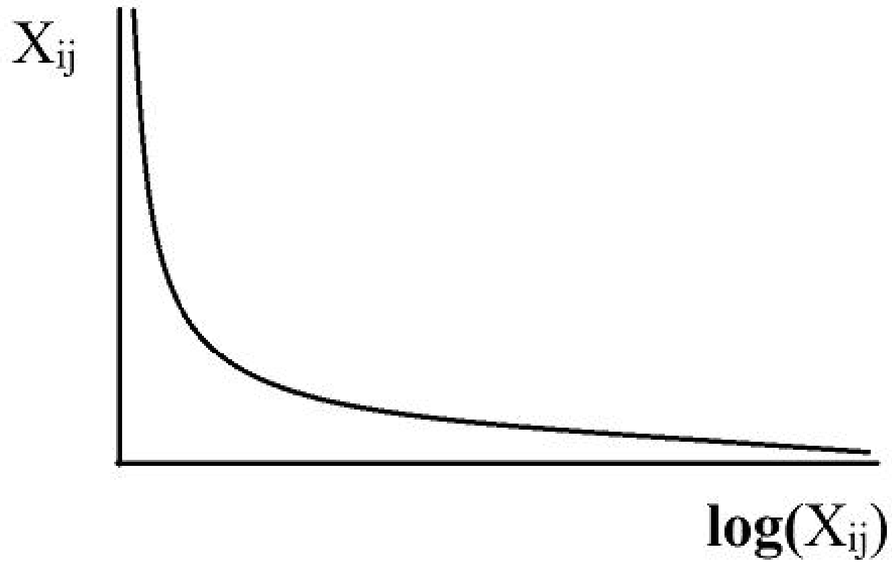
3 Mô hình Glove

$$w_i^T \bar{w}_j = X_{ij}$$

X_{ij} — quá lớn / quá nhỏ $\rightarrow \log(X_{ij})$

$$w_i^T \bar{w}_j = \log(X_{ij})$$

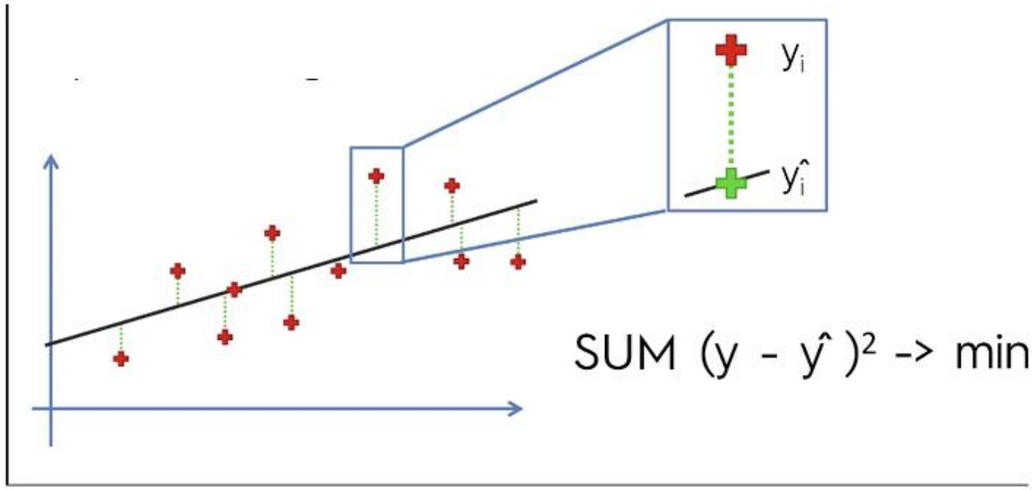
$$w_i^T \bar{w}_j \; b_i + \bar{b}_k = \log(X_{ij})$$



3 Mô hình Glove

$$\sum_{i,j=1}^{|V|} f(X_{ij}) \left(w^T \bar{w}_j + b_i + \bar{b}_j - \log(X_{ij}) \right)^2$$

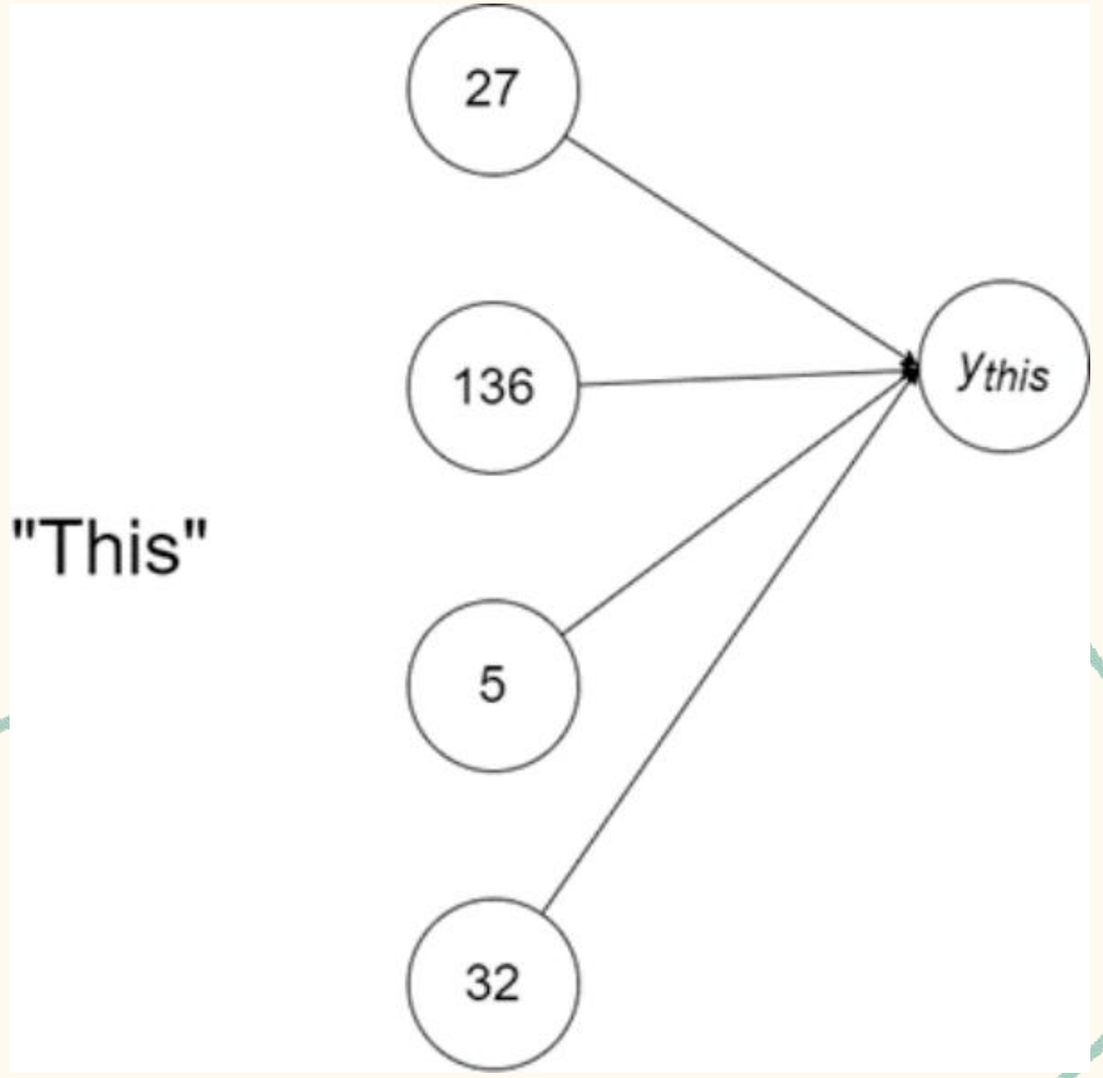
$$f(X_{ij}) = \begin{cases} \left(\frac{X_{ij}}{x_{max}} \right)^{\alpha} & \text{if } X_{ij} < X_{max} \\ 1 & \text{otherwise} \end{cases}$$



LSTM

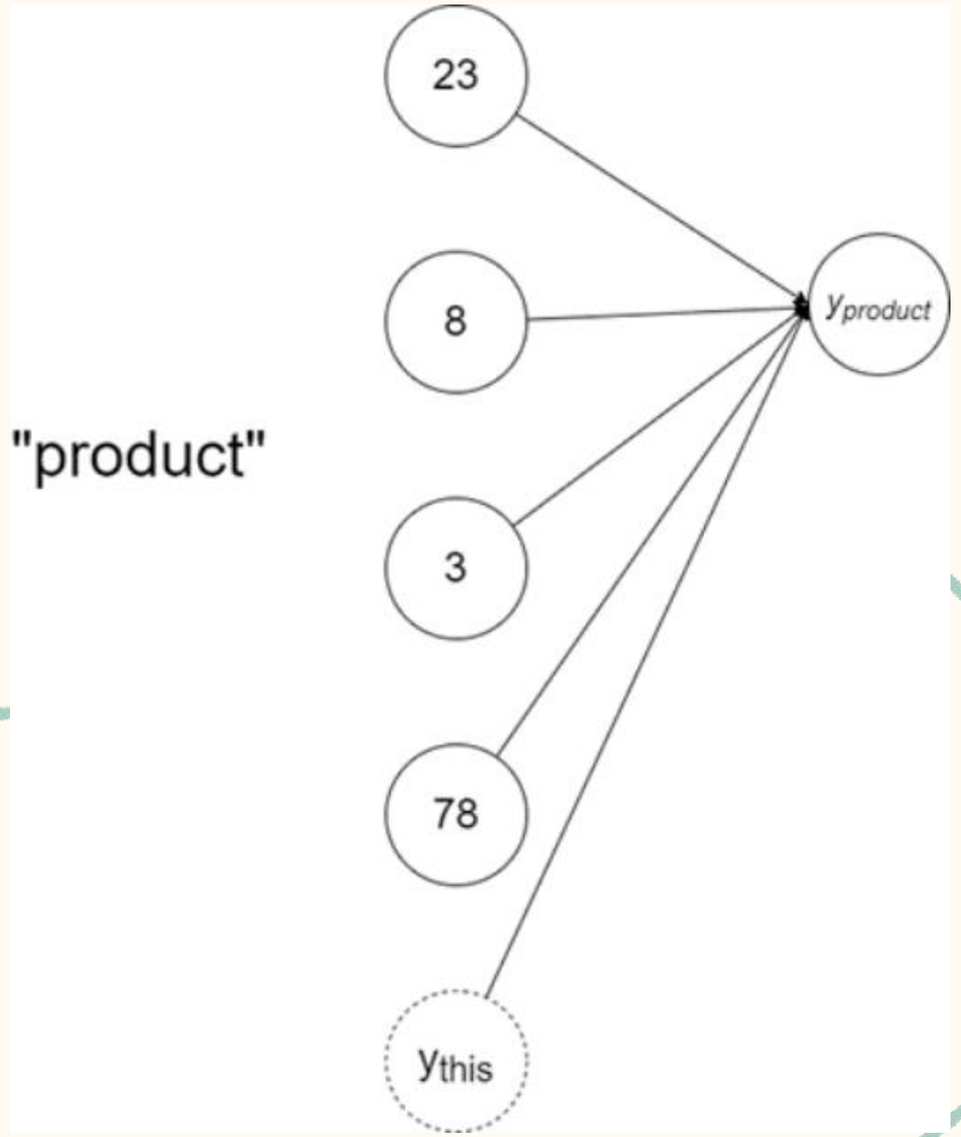
RNN

X = “This product is good”



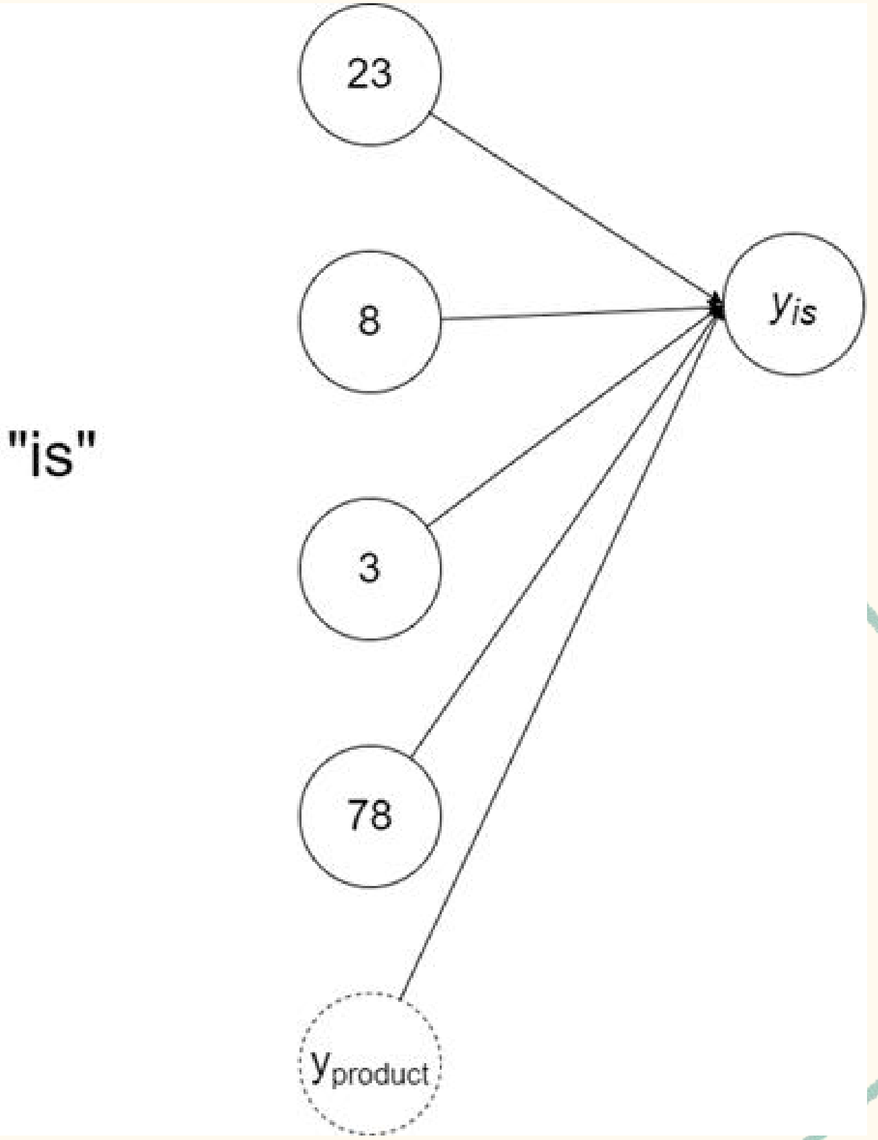
RNN

X = “This product is good”



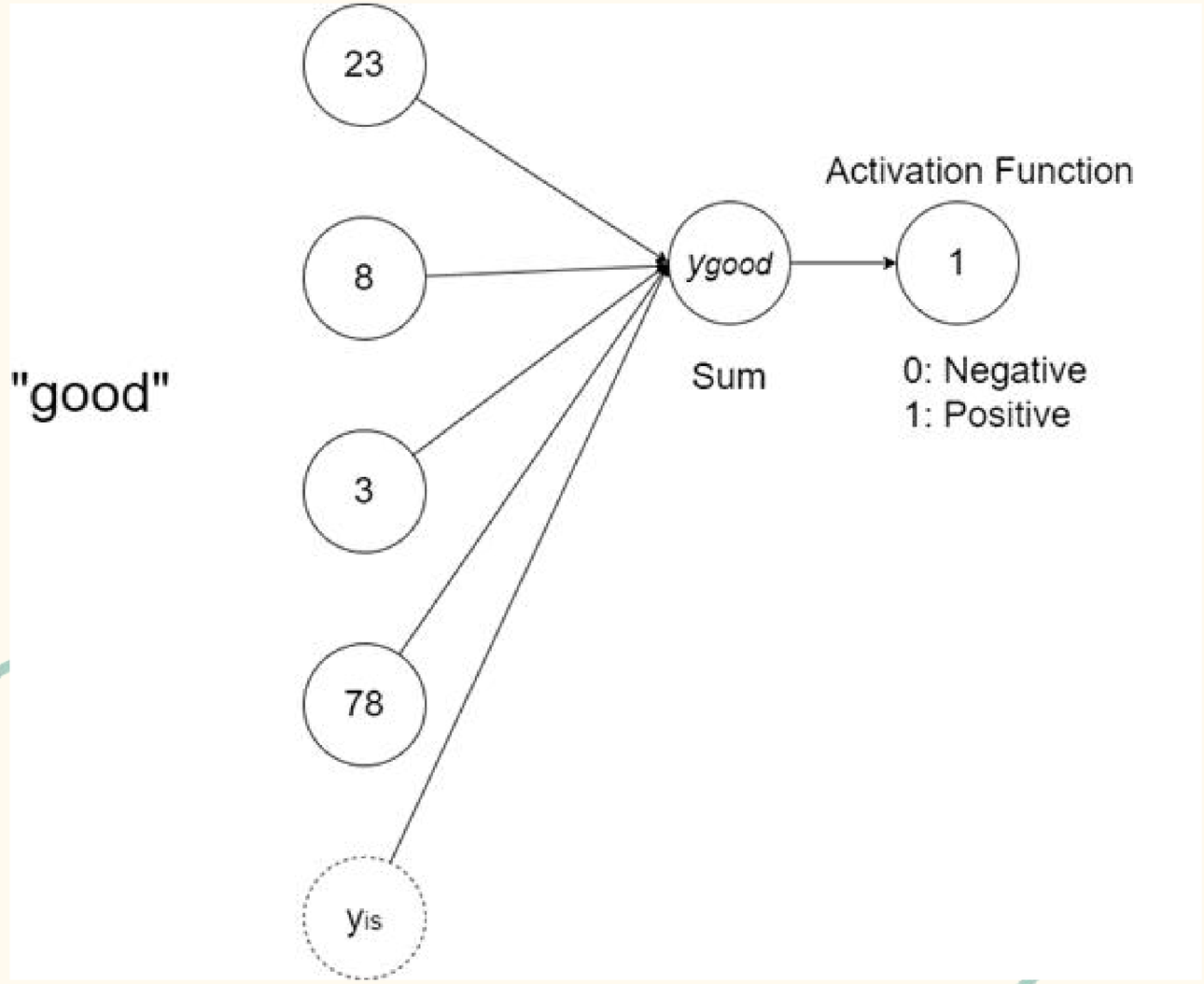
RNN

X = “This product is good”



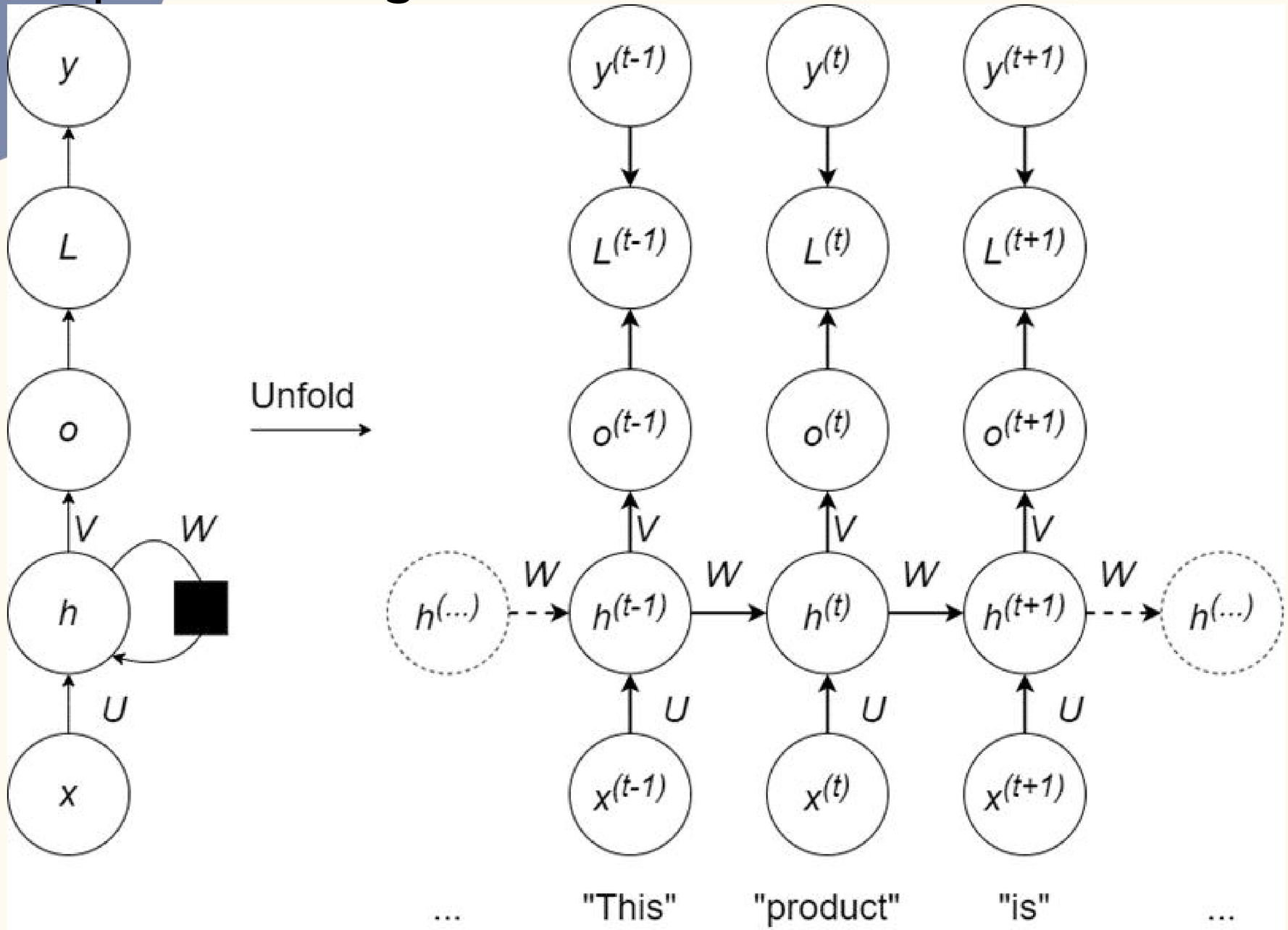
RNN

X = “This product is good”



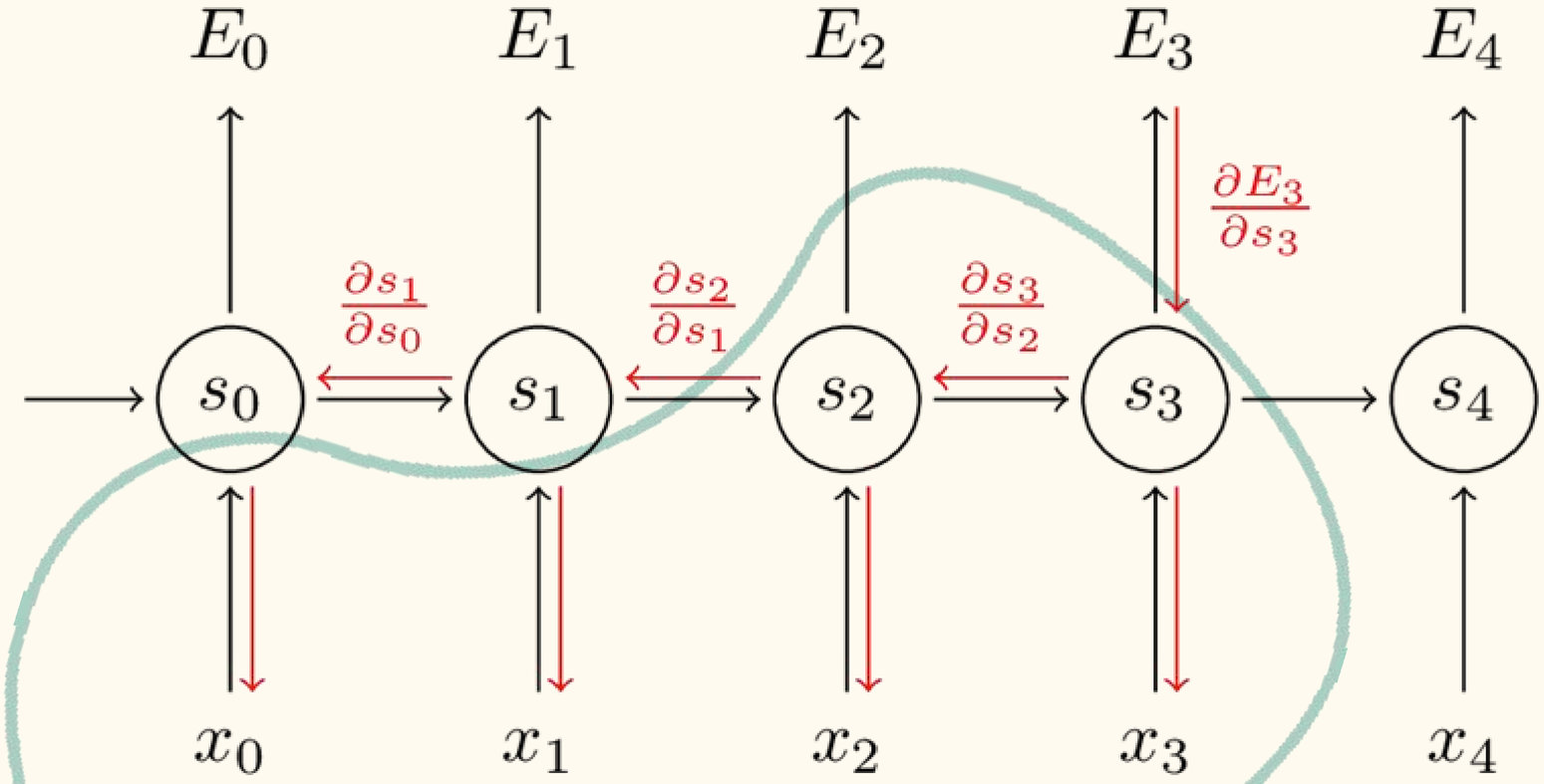
X = “This product is good”

RNN

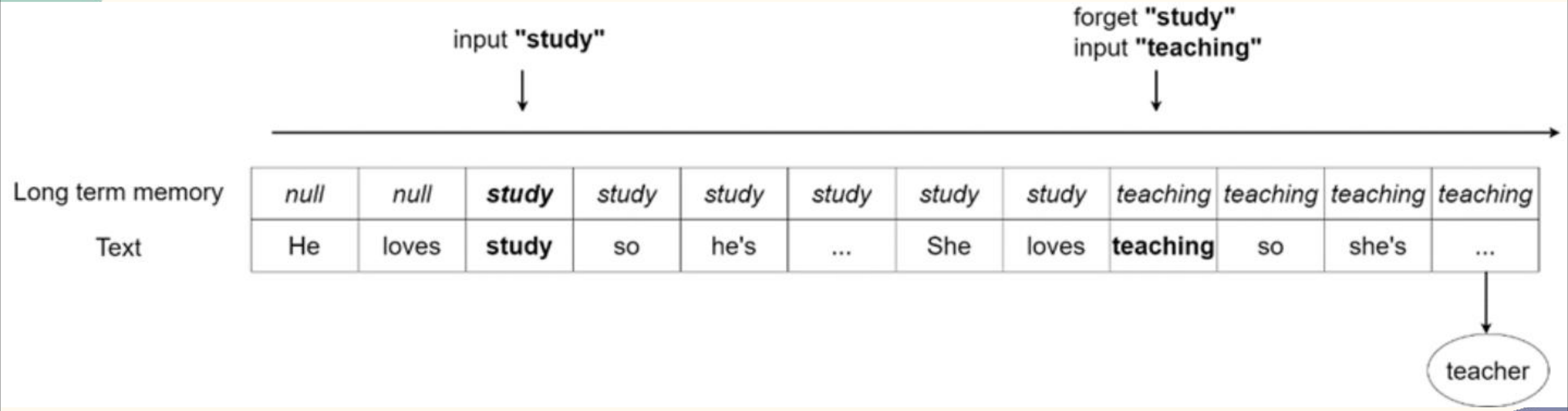


RNN
Vanishing Gradient

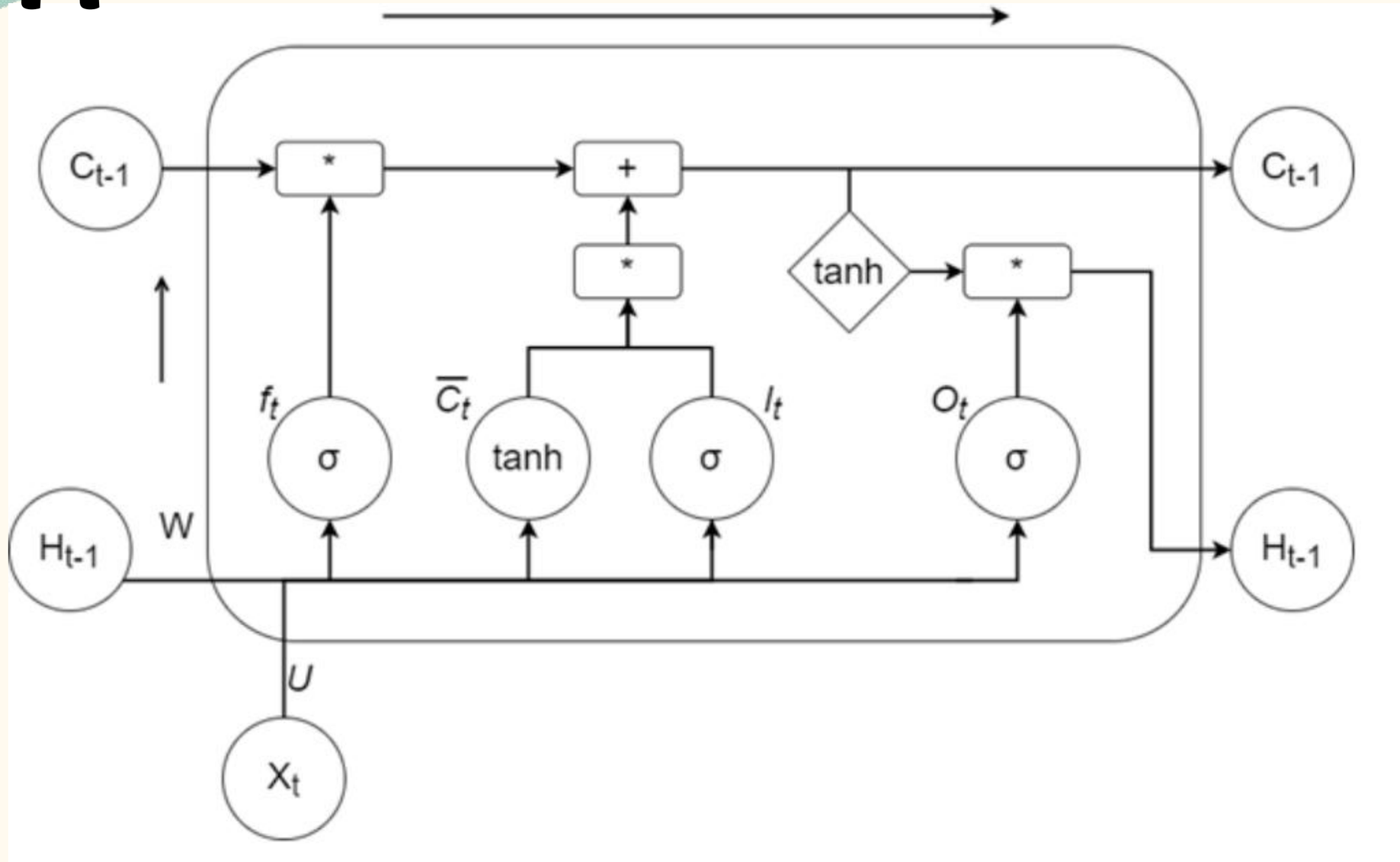
=> RNN ghi nhớ ngắn hạn



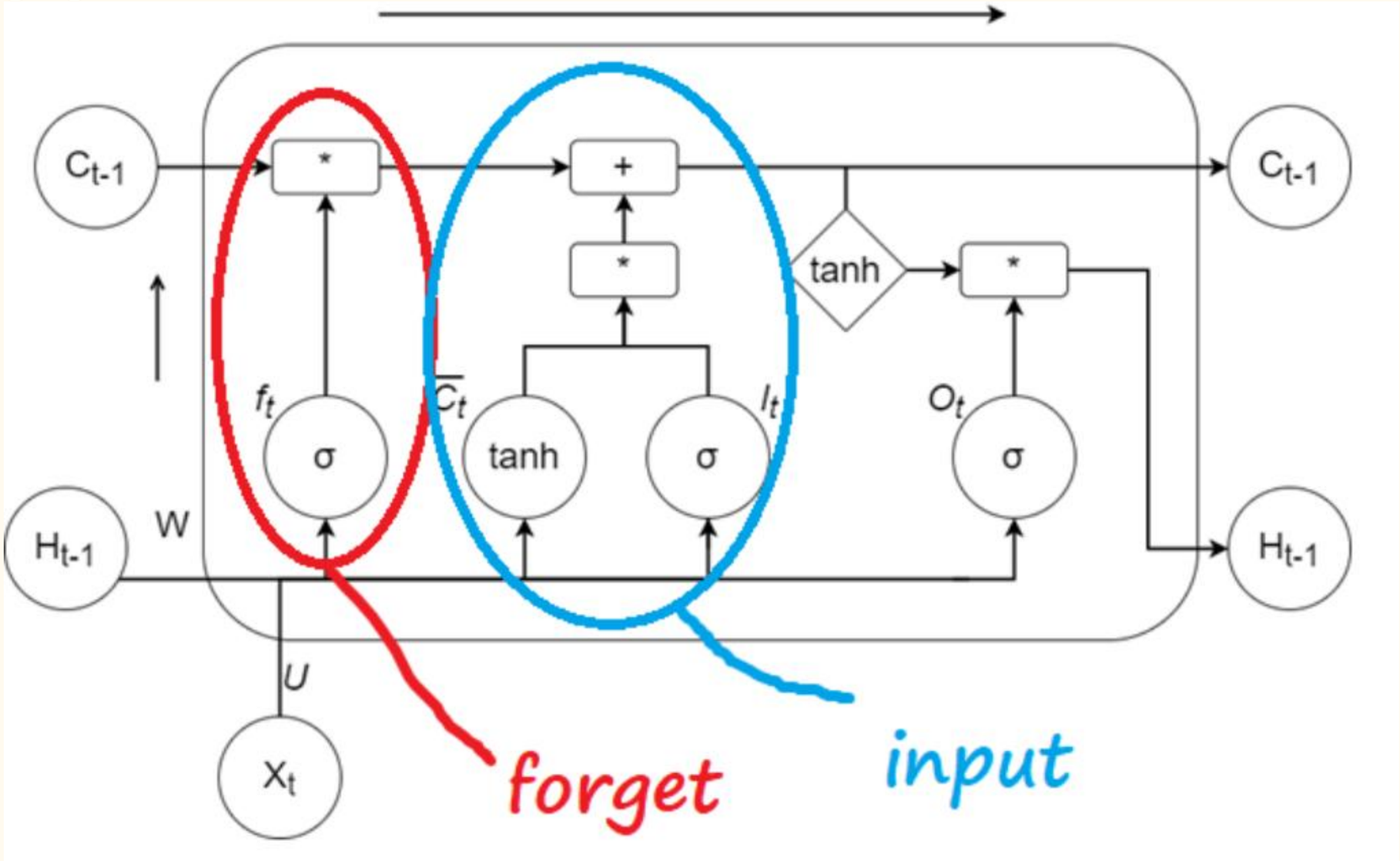
LSTM ghi nhớ dài hạn



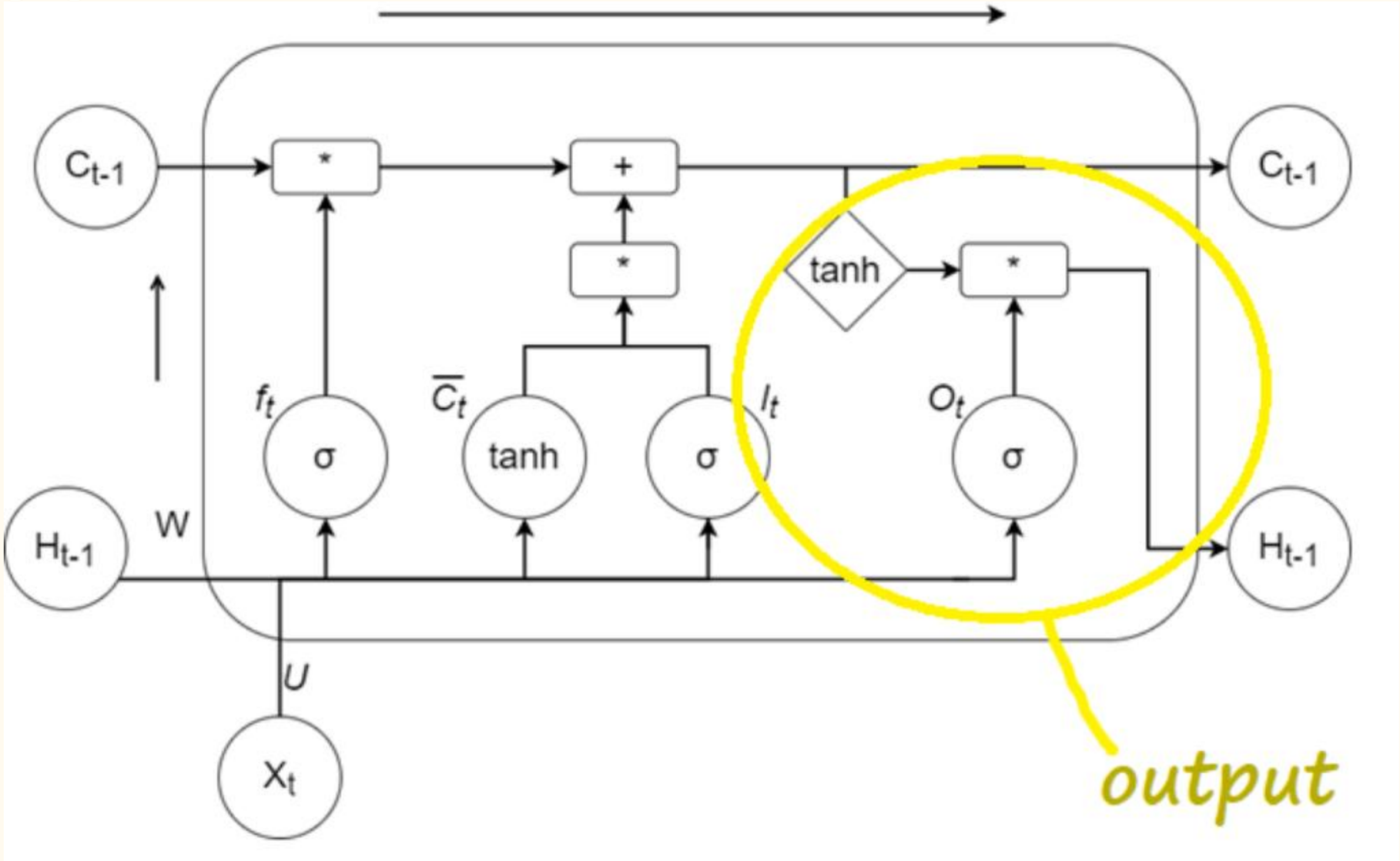
LSTM



LSTM


C_{t-1}
H_{t-1}
X_t
W
U
f_t
$\overline{C}_t$
I_t
O_t
σ
tanh
σ
σ
tanh
C_{t-1}
H_{t-1}
forget
input

LSTM



Word2Vec, BoW

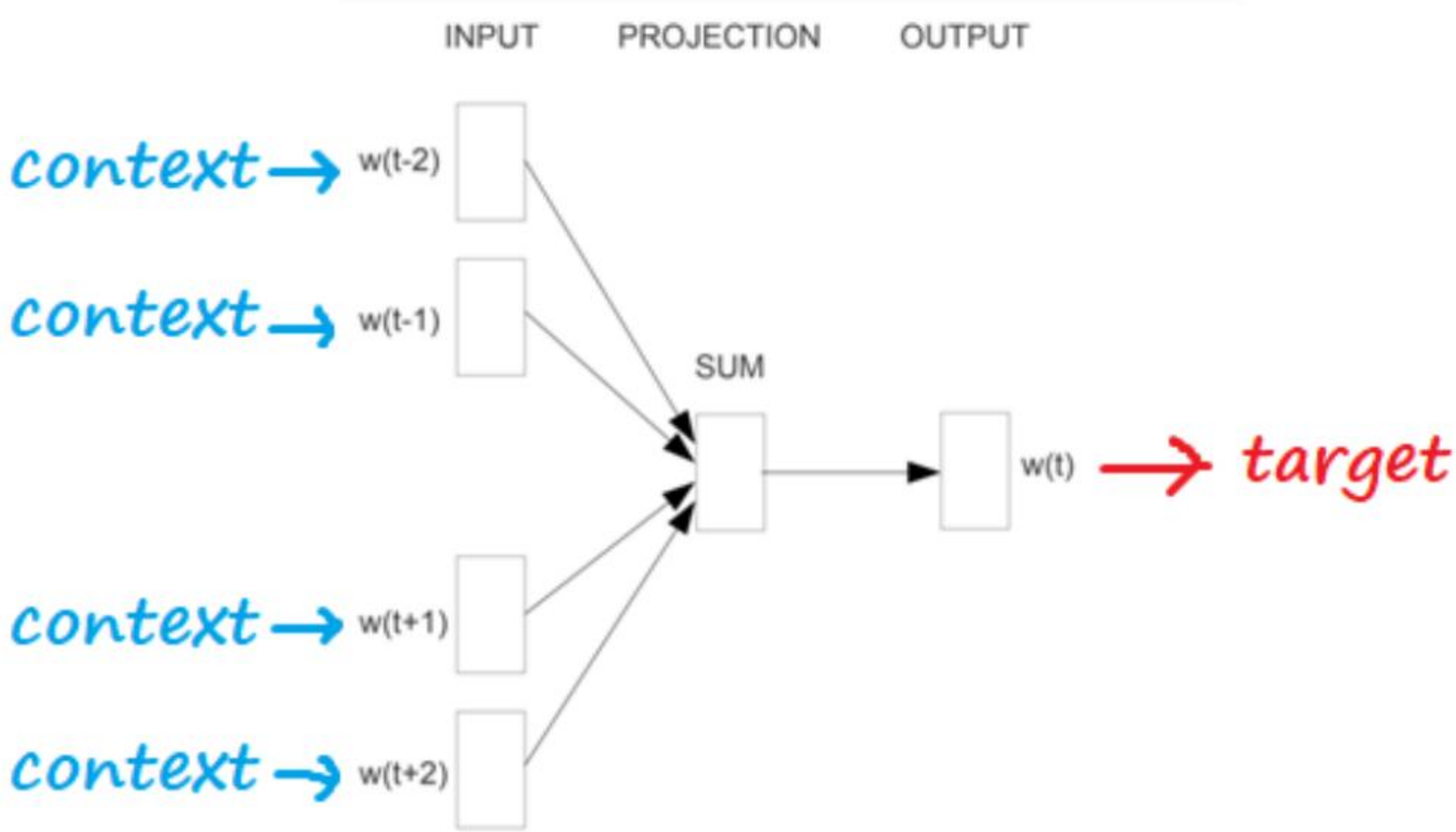
Word2Vec

(5)The future *king is* **the** *prince*

(**the**, king),
(**the**, is),
(**the**, prince),

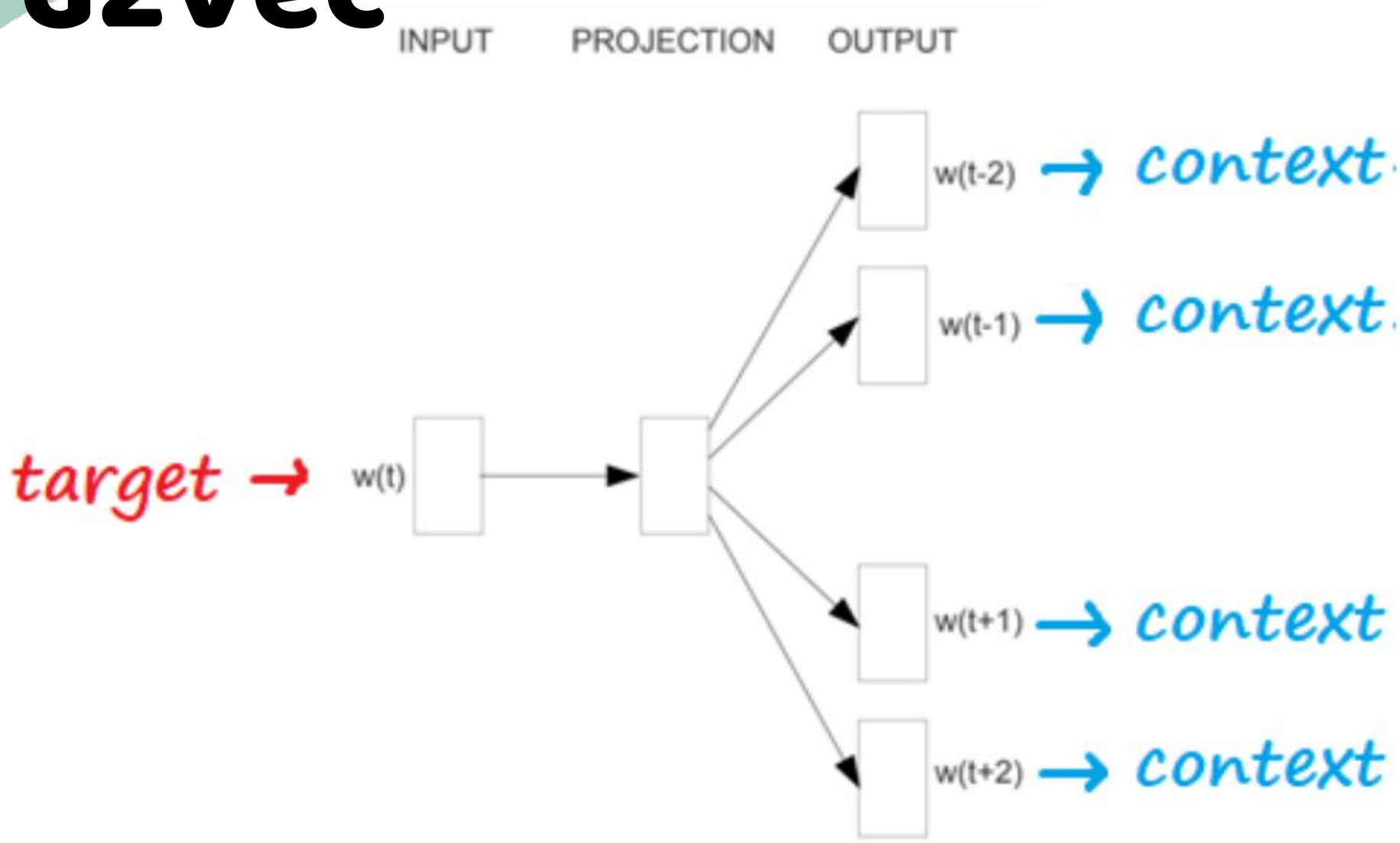
Window size

Word2Vec



CBOW

Word2Vec



Skip-gram

BoW (Bag Of Words)

- Vector hóa đoạn văn
- Vector hóa dựa trên tần suất
- Không quan tâm thứ tự

A = [“I love NLU”]
B = [“I love to be a student of NLU NLU”]

	nlu	i	love	to	be	a	student	of
A	1	1	1	0	0	0	0	0
B	2	1	1	1	1	1	1	1