Text Generation with Language Model

NHÓM 6

TRẦN TRUNG ANH – 18520473

NGUYỄN ANH KHOA – 18520923

VÕ QUỐC AN – 18520440

Giới thiệu

- Text generation: dự đoán từ tiếp theo dựa trên những từ đã có trước.
- Mục tiêu: giống nhất có thể với ngôn ngữ tự nhiên.

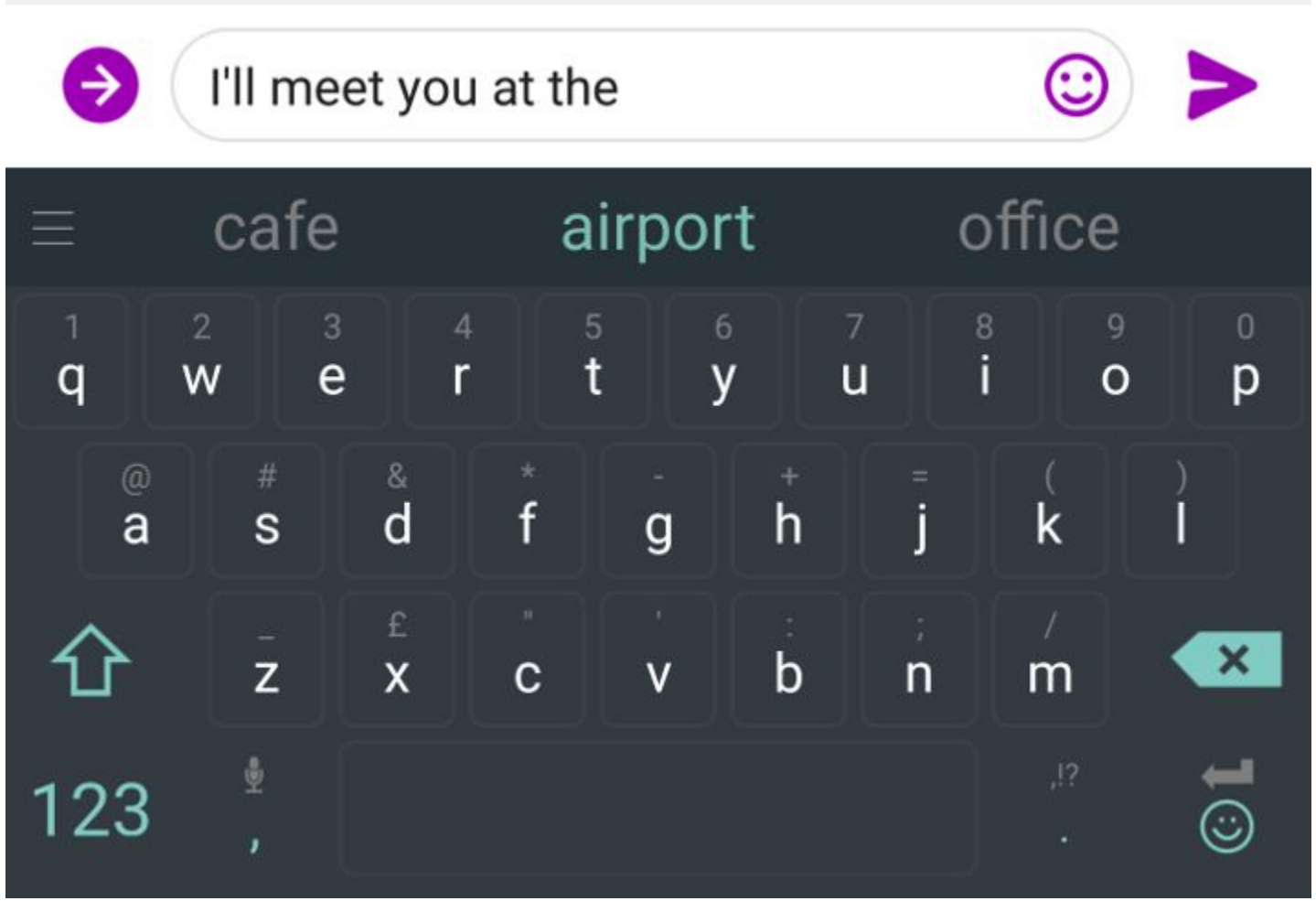
Ứng dụng

- Auto-complete.
- Hệ thống hỏi - đáp.
- Chatbot.

Ứng dụng


Your search query
Google
Options to "Remove" searches based on your browser history
skateboards for
skateboards for sale
skateboards for cheap
skateboards for kids
skateboards for beginners
skateboards for girls
skateboards for toddlers
skateboards for sale near me
skateboards for adults
skateboards for 8 year olds
skateboards for dogs
Remove
Remove
Autocomplete suggestions
Google Search
I'm Feeling Lucky
Report inappropriate predictions

Ứng dụng


I'll meet you at the
cafe
airport
office
1 q 2 w 3 e 4 r 5 t 6 y 7 u 8 i 9 o 0 p
@ a # s & d * f - g + h = j (k) l
_ z £ x " c ' v : b ; n / m
123
,
.!?

Phương pháp xây dựng

- Chuẩn bị dữ liệu.
- Xây dựng mô hình ngôn ngữ.
- Sử dụng mô hình để dự đoán từ tiếp theo.

Phương pháp xây dựng

Dữ liệu từ Kaggle
New York Times Articles

Số câu sử dụng: 6011

A Quote Disproved
Hot Stuff Turns Cold
At the Top
Years of Bizarre Behavior
Will the Court Stand Up to Mr. Trump?
Hope in Arizona
How Human Rights Groups Failed on Economic Equality
An American Tragedy in Nashville
We DonÃ¢â‚¬â„¢t Need No Education
Paul Relents
Making Change
Long Overlooked

Phương pháp xây dựng

Tiền xử lý dữ liệu:

Chuyển viết thường
Loại bỏ dấu câu
Loại bỏ kí tự đặc biệt
…

```
['a quote disproved',
 'hot stuff turns cold',
 'at the top',
 'years of bizarre behavior',
 'will the court stand up to mr trump',
 'hope in arizona',
 'how human rights groups failed on economic equality',
 'an american tragedy in nashville',
 'we dont need no education',
 'paul relents',
 'making change',
 'long overlooked']
```

Mô hình N-gram

the students opened their _____

N-gram là một chuỗi các từ liên tiếp nhau

- unigrams: “the”, “students”, “opened”, ”their”
- bigrams: “the students”, “students opened”, “opened their”
- trigrams: “the students opened”, “students opened their”
- 4-grams: “the students opened their”

Mô hình N-gram

~~as the proctor started the clock, the~~ *students opened their* _____

discard

condition on this

$$P(\boldsymbol{w}|\text{students opened their}) = \frac{\text{count}(\text{students opened their } \boldsymbol{w})}{\text{count}(\text{students opened their})}$$

Mô hình N-gram

Smoothing

- Laplace: $$P(\boldsymbol{w}|\text{students opened their}) = \frac{\text{count}(\text{students opened their } \boldsymbol{w}) + 1}{\text{count}(\text{students opened their}) + |V|}$$

- Backoff: nếu không thể tính *n-gram* thì thay bằng *n-1, n-2, ... gram*

Mô hình N-gram

Dự đoán từ tiếp theo

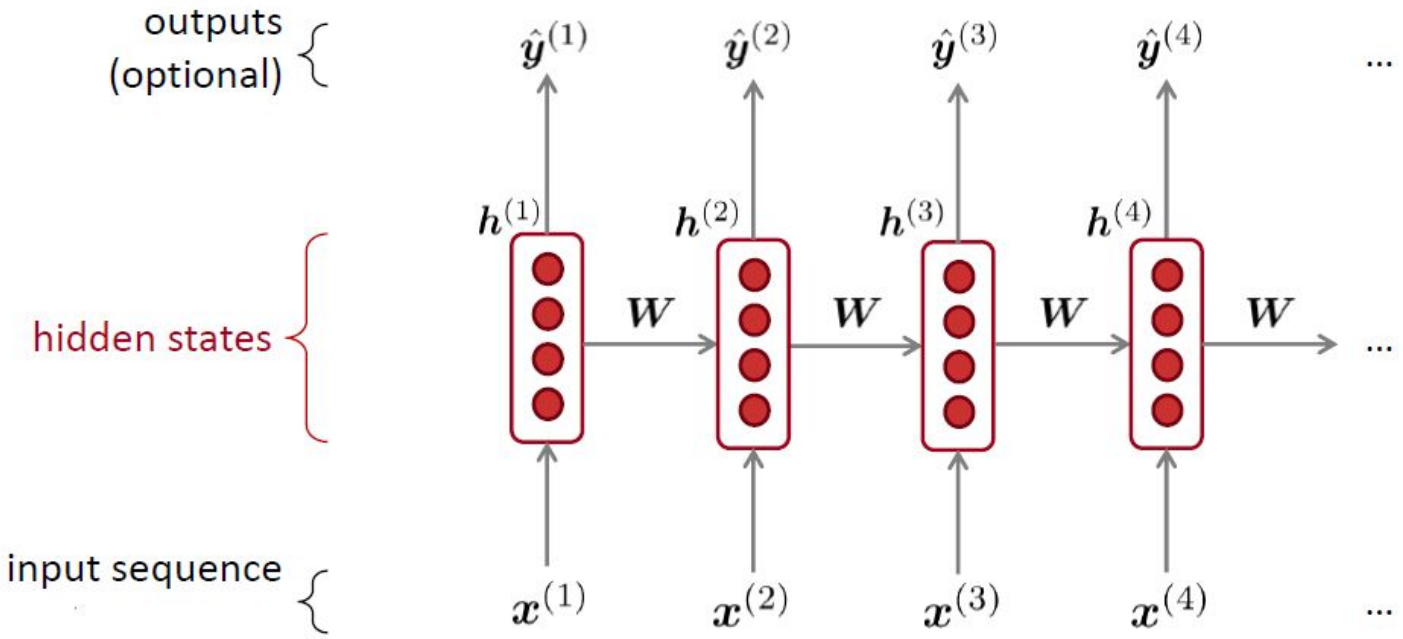
In the class, the students open their _____

books: 0.153
bags: 0.056
hands: 0.0032
brains: 0.00001

Recurrent Neural Networks (RNN)

$$\boldsymbol{h}^{(t)} = g_h(W_{\mathrm{I}}\boldsymbol{x}^{(t)} + W_{\mathrm{R}}\boldsymbol{h}^{(t-1)} + \boldsymbol{b}_h)$$

$$\boldsymbol{y}^{(t)} = g_y(W_y\boldsymbol{h}^{(t)} + \boldsymbol{b}_y)$$



Mô hình ngôn ngữ RNN

output distribution

$$\hat{\boldsymbol{y}}^{(t)} = \text{softmax}\left(\boldsymbol{U}\boldsymbol{h}^{(t)} + \boldsymbol{b}_2\right) \in \mathbb{R}^{|V|}$$

hidden states

$$\boldsymbol{h}^{(t)} = \sigma\left(\boldsymbol{W}_h\boldsymbol{h}^{(t-1)} + \boldsymbol{W}_e\boldsymbol{e}^{(t)} + \boldsymbol{b}_1\right)$$

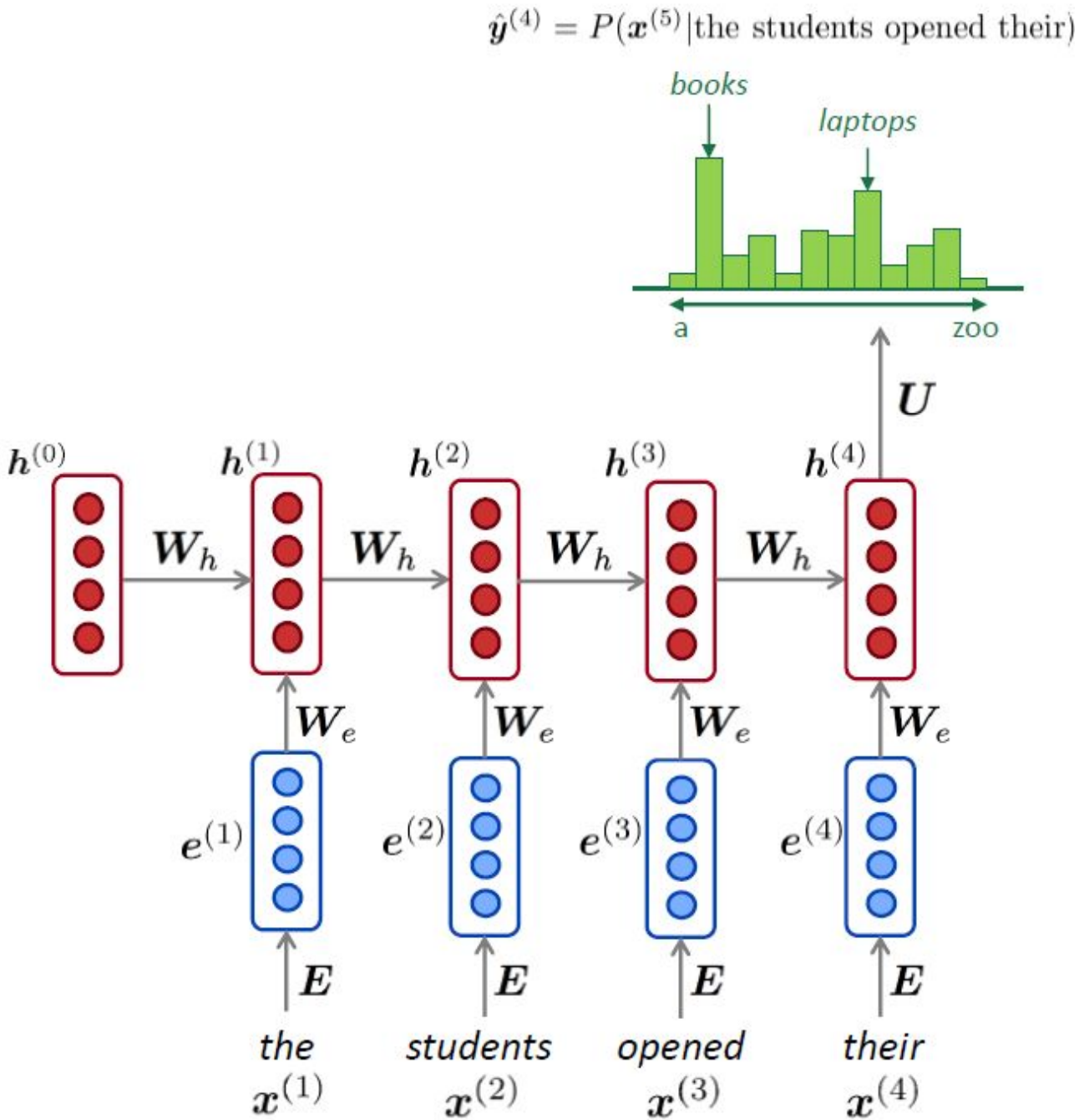
$\boldsymbol{h}^{(0)}$ is the initial hidden state

word embeddings

$$\boldsymbol{e}^{(t)} = \boldsymbol{E}\boldsymbol{x}^{(t)}$$

words / one-hot vectors

$$\boldsymbol{x}^{(t)} \in \mathbb{R}^{|V|}$$



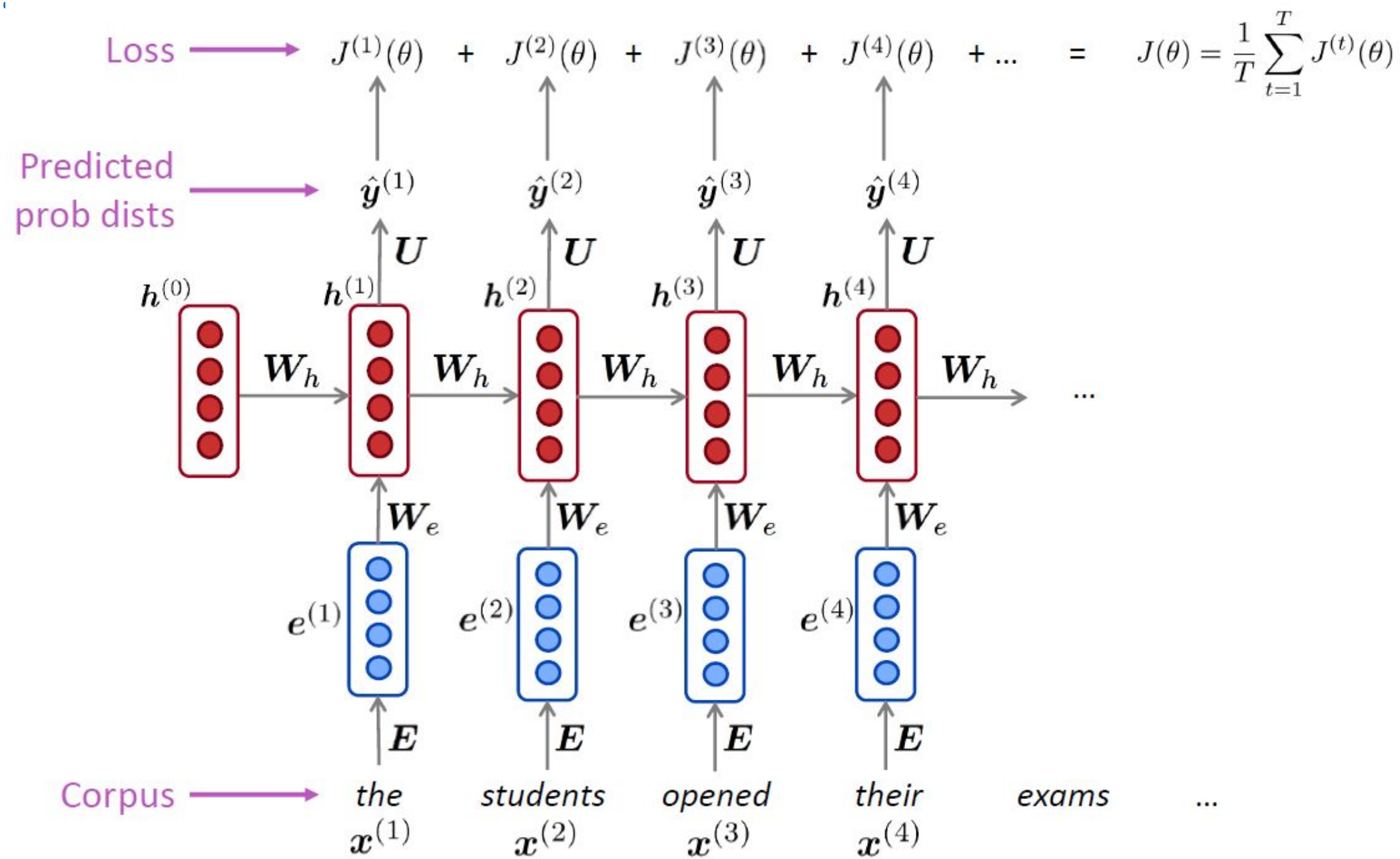
Huấn luyện RNN

- Input: one-hot của mỗi từ trong bộ dữ liệu
- Tính phân phối xác suất của tất cả các từ trong dữ liệu
- Loss function: cross-entropy

$$J^{(t)}(\theta) = CE(\boldsymbol{y}^{(t)}, \hat{\boldsymbol{y}}^{(t)}) = -\sum_{w \in V} \boldsymbol{y}_w^{(t)} \log \hat{\boldsymbol{y}}_w^{(t)} = -\log \hat{y}^t_{w_{t+1}}$$

- Tính trung bình cho toàn tập dữ liệu

$$J(\theta) = \frac{1}{T}\sum_{t=1}^{T} J^{(t)}(\theta) = \frac{1}{T}\sum_{t=1}^{T} -\log \hat{y}^t_{w_{t+1}}$$


Loss
$J^{(1)}(\theta)$ + $J^{(2)}(\theta)$ + $J^{(3)}(\theta)$ + $J^{(4)}(\theta)$ + ... = $J(\theta) = \frac{1}{T}\sum_{t=1}^{T} J^{(t)}(\theta)$
Predicted prob dists
$\hat{y}^{(1)}$ $\hat{y}^{(2)}$ $\hat{y}^{(3)}$ $\hat{y}^{(4)}$
U
$h^{(0)}$ $h^{(1)}$ $h^{(2)}$ $h^{(3)}$ $h^{(4)}$
W_h
...
W_e
$e^{(1)}$ $e^{(2)}$ $e^{(3)}$ $e^{(4)}$
E
Corpus
the $x^{(1)}$
students $x^{(2)}$
opened $x^{(3)}$
their $x^{(4)}$
exams
...

Vấn đề:

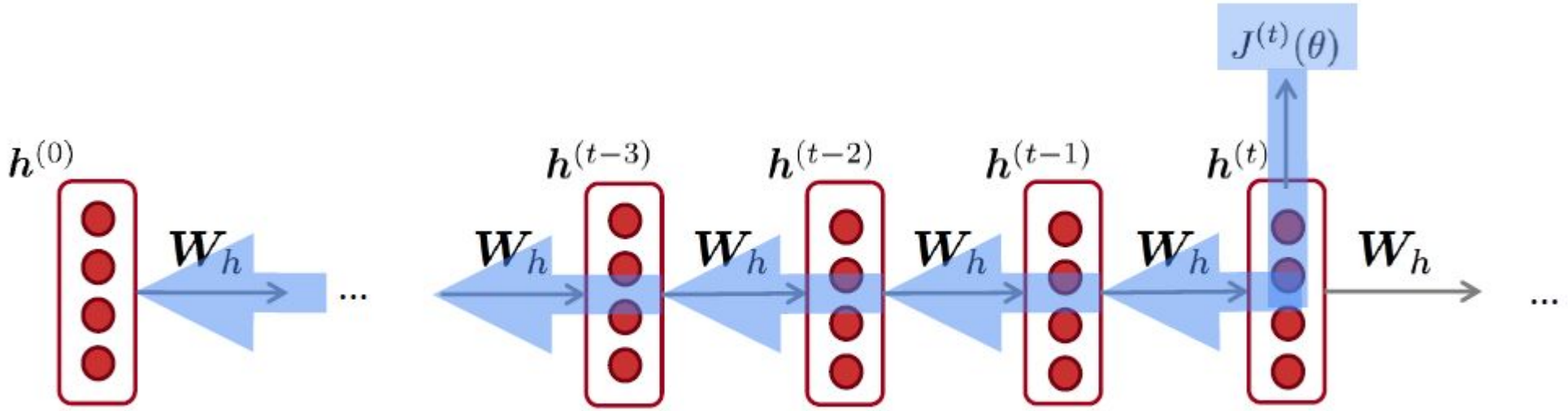
- Chi phí lớn để tính hàm mất mát và đạo hàm của toàn bộ tập dữ liệu

$$J(\theta) = \frac{1}{T}\sum_{t=1}^{T} J^{(t)}(\theta) = \frac{1}{T}\sum_{t=1}^{T} -\log \hat{y}^{t}_{w_{t+1}}$$

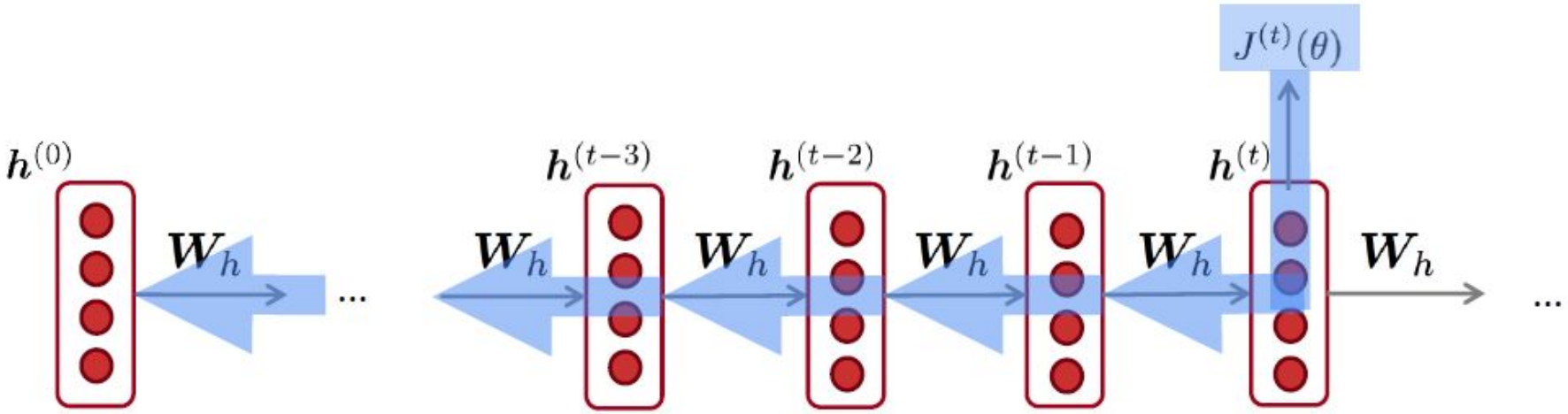
Giải pháp:

- Stochastic Gradient Descent

Backpropagation



Backpropagation through time



Tính đạo hàm theo từng timestep i= t,....0

⇒ Backpropagation through time

$$\frac{\partial J^{(t)}}{\partial \boldsymbol{W_h}} = \sum_{i=1}^{t} \frac{\partial J^{(t)}}{\partial \boldsymbol{W_h}}\bigg|_{(i)}$$

THANKS FOR LISTENING