

Dịch máy  
tiếng Việt → tiếng Anh



# THÀNH VIÊN NHÓM



Nguyễn Quang Minh



Dương Văn Dụ



Nguyễn Viết Quang  
Trưởng nhóm



Phùng Minh Hiếu



Phùng Văn Thịnh

# Khái niệm dịch máy

**Kết quả của dịch máy khác gì với dịch 1-1 bằng từ điển thông thường?**

**Ví dụ câu: ' có con mèo ở trên sàn'**

**Dịch theo từ điển**

‘Have a cat on the floor’  
‘Exist a cat on the floor’



**Dịch máy theo ngữ cảnh**

‘The cat is on the floor’  
‘There is a cat on the floor’

# Khái niệm dịch máy

## Mô hình seq2seq khác gì mô hình thông thường ?

**Sequence to Sequence  
(seq2seq ) nói chung hay GRU  
nói riêng**

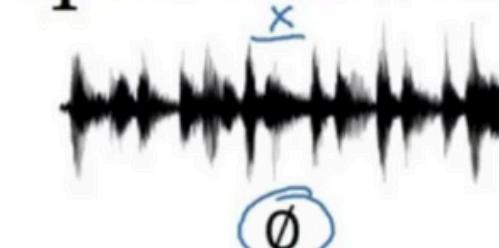
Dữ liệu đầu vào hoặc đầu ra là chuỗi.

**Mô hình thông thường**

Đầu vào và đầu ra là một số, biến phân loại,...

### Examples of sequence data

Speech recognition



"The quick brown fox jumped over the lazy dog."

Music generation



Sentiment classification

"There is nothing to like  
in this movie."



DNA sequence analysis

AGCCCCCTGTGAGGAACCTAG



AGCCC<sub>CTGTGAGGAAC</sub>CTAG

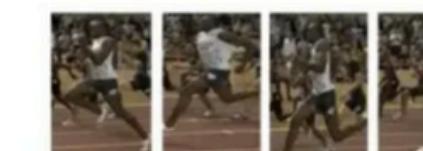
Machine translation

Voulez-vous chanter avec  
moi?



Do you want to sing with  
me?

Video activity recognition



Running

Name entity recognition

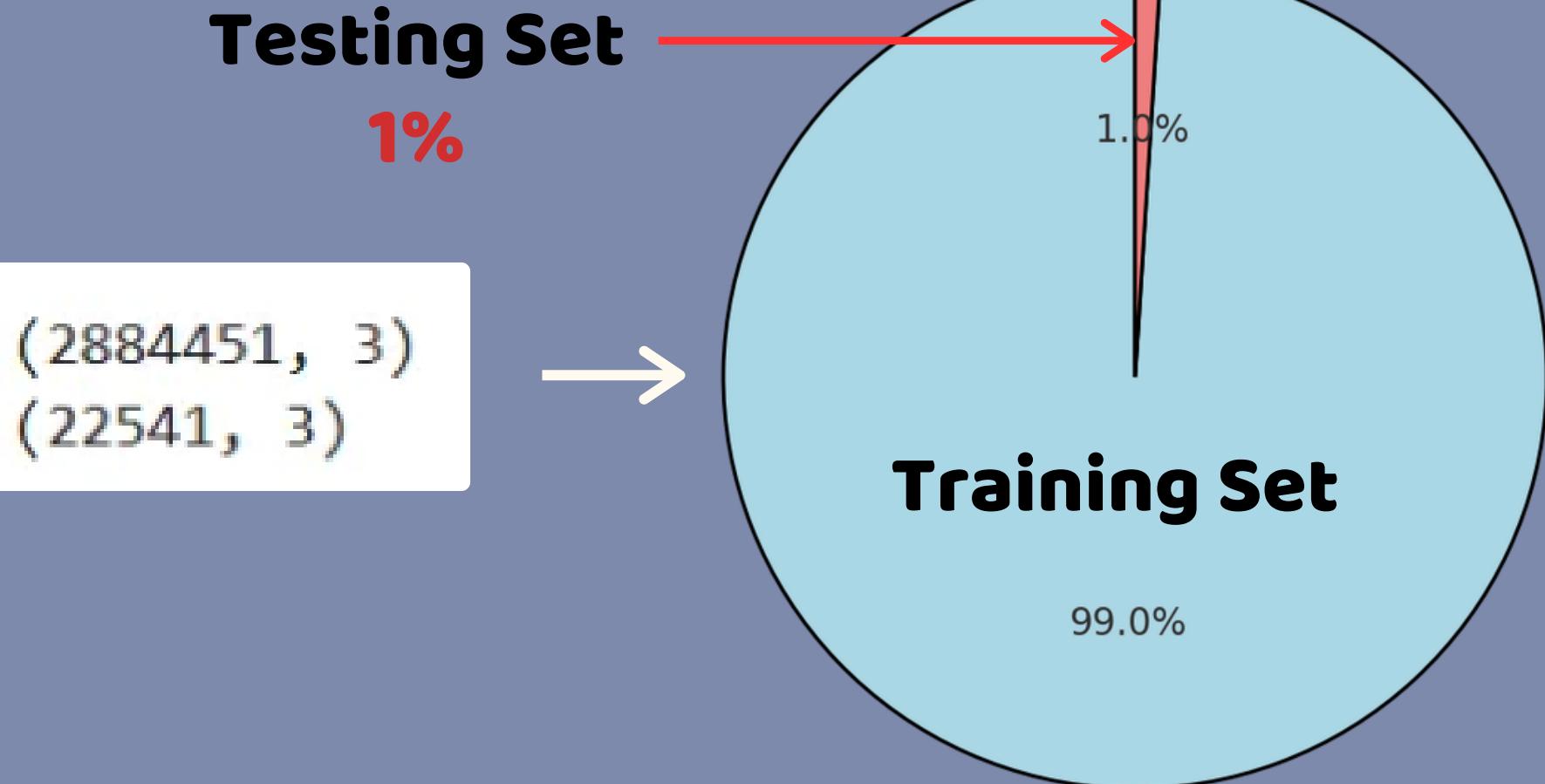
Yesterday, Harry Potter  
met Hermione Granger.



Yesterday, Harry Potter  
met **Hermione Granger**.

# DATASET

```
print(data.shape)  
print(data_test.shape)
```



```
data = pd.read_csv("/kaggle/input/my-data/new_train_ds.csv")  
data = data.dropna()  
data.head(5)
```

**Training Set**

	en	vi	source
0	- Sorry, that question's not on here.	- Xin lỗi, nhưng mà ở đây không có câu hỏi đấy.	OpenSubtitles v2018
1	He wants you to come with him immediately.	Ông ấy muốn bố đi với ông ấy ngay lập tức	OpenSubtitles v2018
2	I thought we could use some company.	Tôi nghĩ chúng ta có thể muốn vài người bạn đồn...	OpenSubtitles v2018
3	It was founded in 2008 by this anonymous progr...	Nó được sáng lập vào năm 2008 bởi một lập trìn...	TED2020 v1
4	With both of these methods, no two prints are ...	Với cả hai phương pháp, không có hai bản in nà...	TED2020 v1

```
dt_test = pd.read_csv("/kaggle/input/my-data/new_test_ds.csv")  
dt_test = dt_test.dropna()  
dt_test.head(5)
```

**Testing Set**

	en	vi	source
0	And what I think the world needs now is more c...	Và tôi nghĩ điều thế giới đang cần bây giờ là ...	TED2020 v1
1	The group is named after Bangkok, the capital ...	Nhóm được đặt theo tên của Bangkok, thủ đô của...	wikimedia v20210402
2	It is surrounded by rivers (Simpson and Coyhai...	Nó được bao quanh bởi các con sông (Simpson và...	WikiMatrix v1
3	Four years, and you never talked to her.	Bốn năm, và cậu chưa nói gì với cô ấy.	OpenSubtitles v2018
4	The man that's giving the test has serious dou...	Người cung cấp bài kiểm tra đó đang nghi ngờ v...	OpenSubtitles v2018



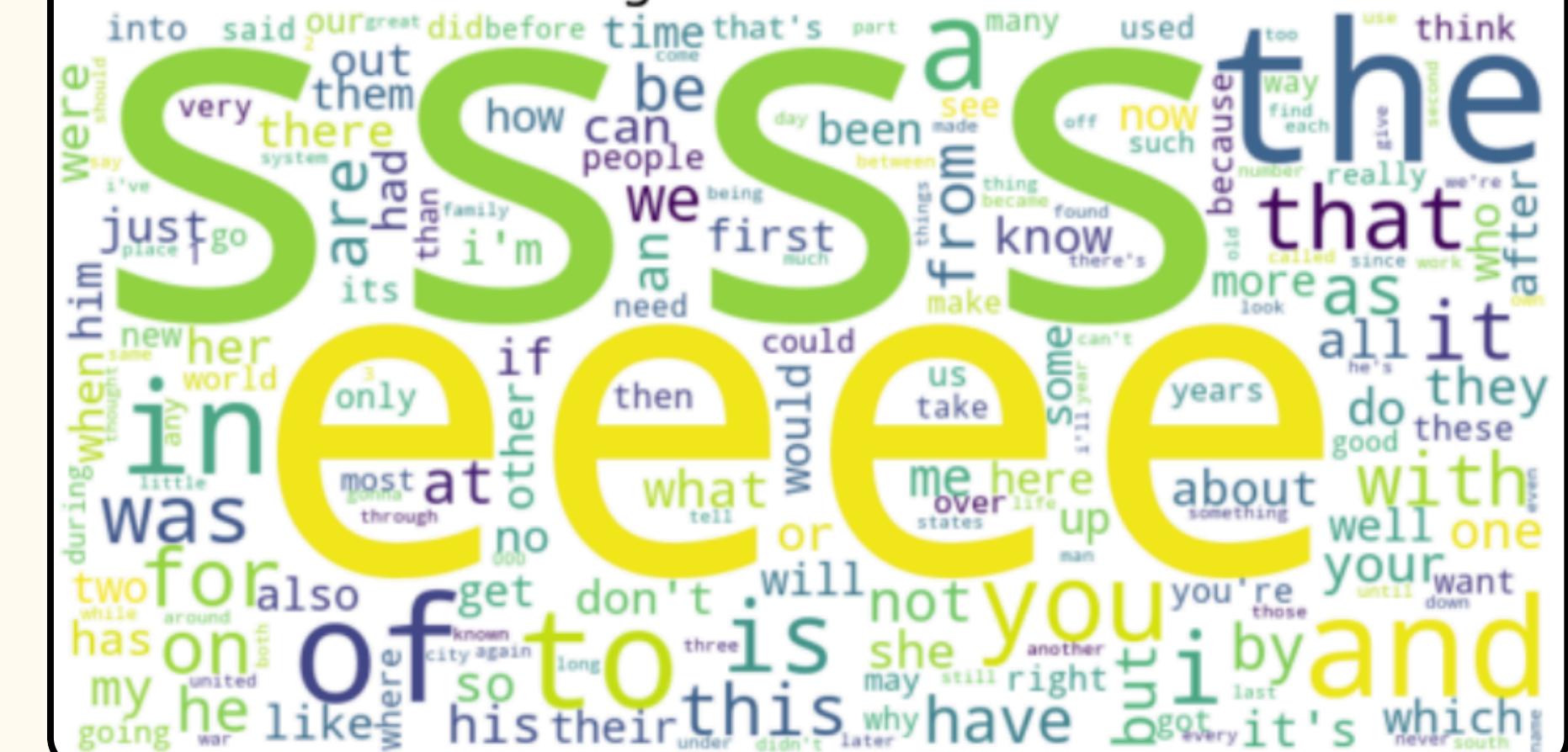
# Exploratory Data Analysis

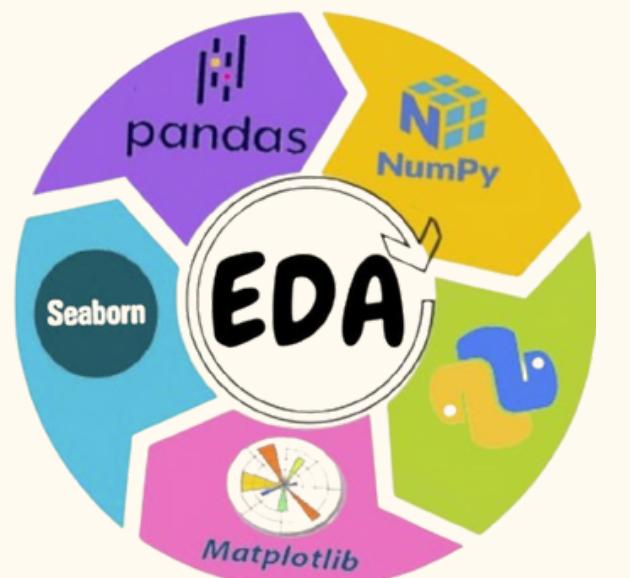
# Word cloud

Vietnamese WordCloud



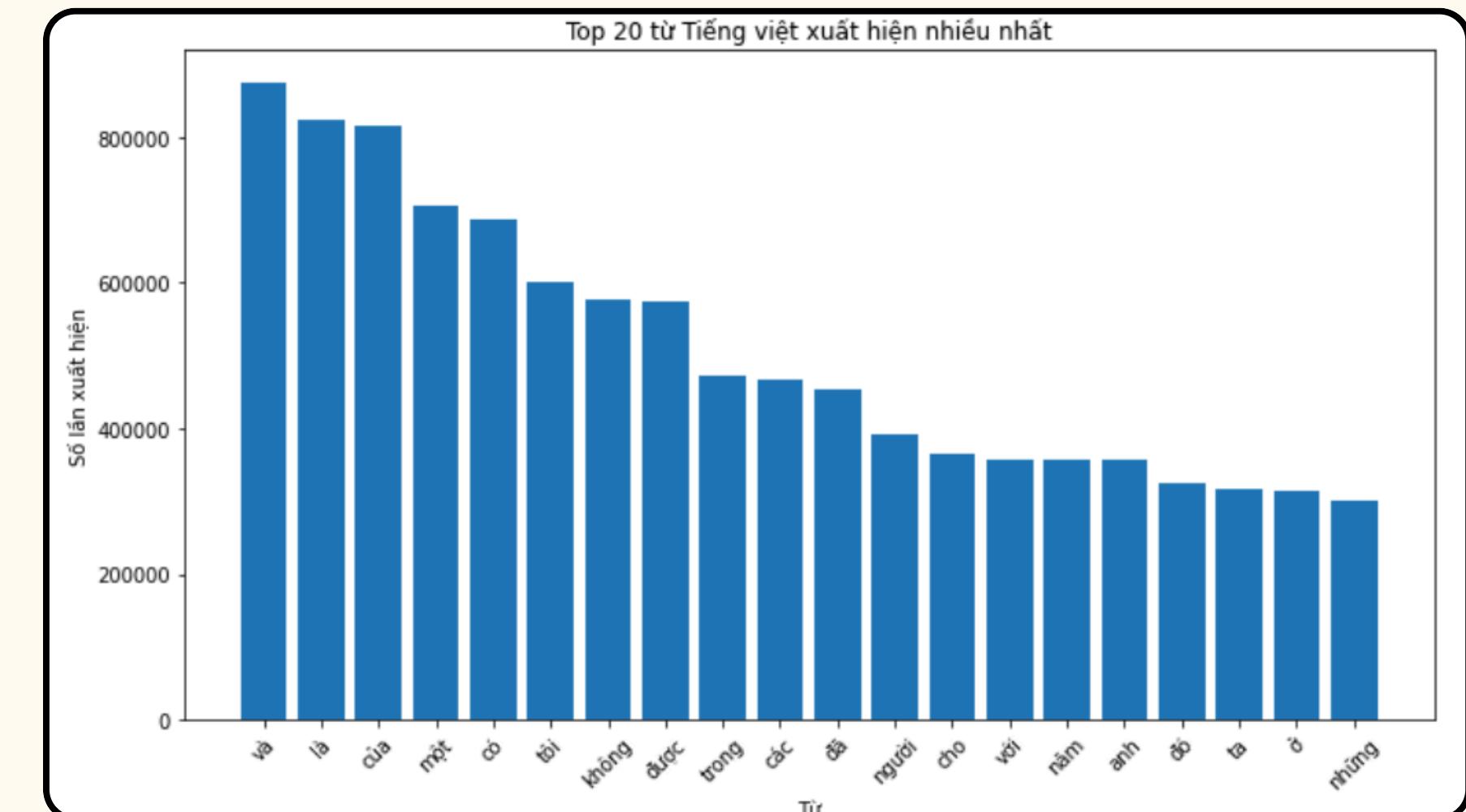
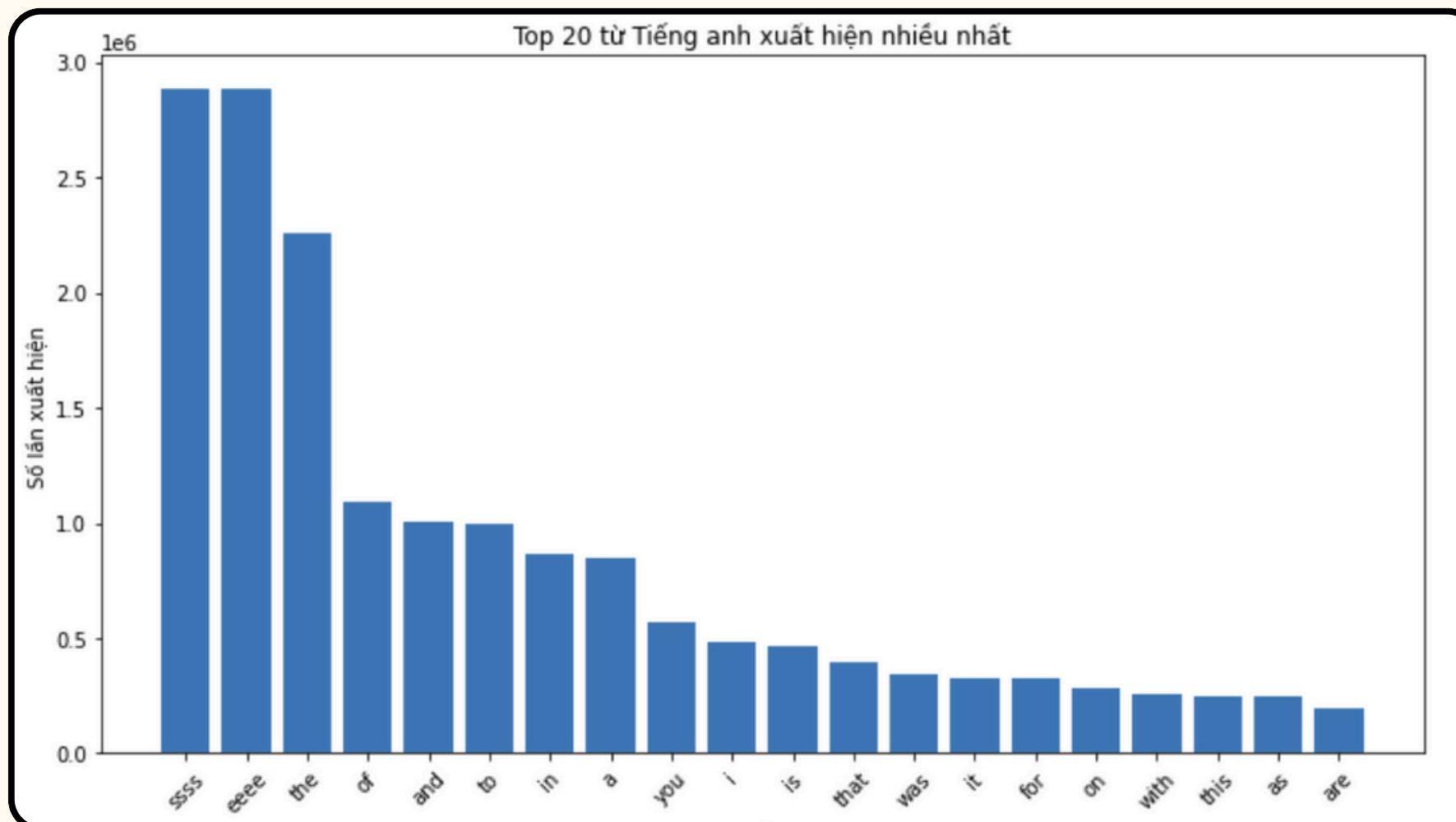
English WordCloud





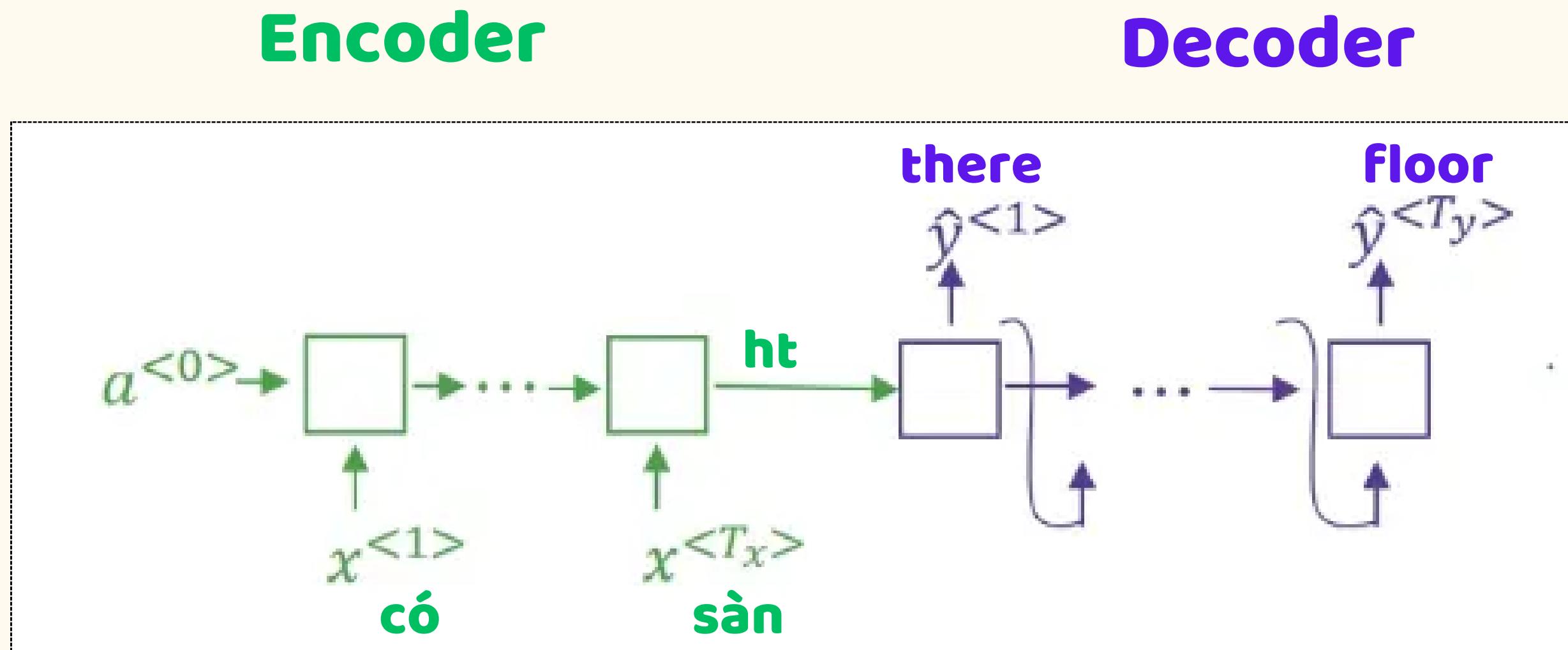
# Exploratory Data Analysis

## Biểu đồ thanh (Bar chart)

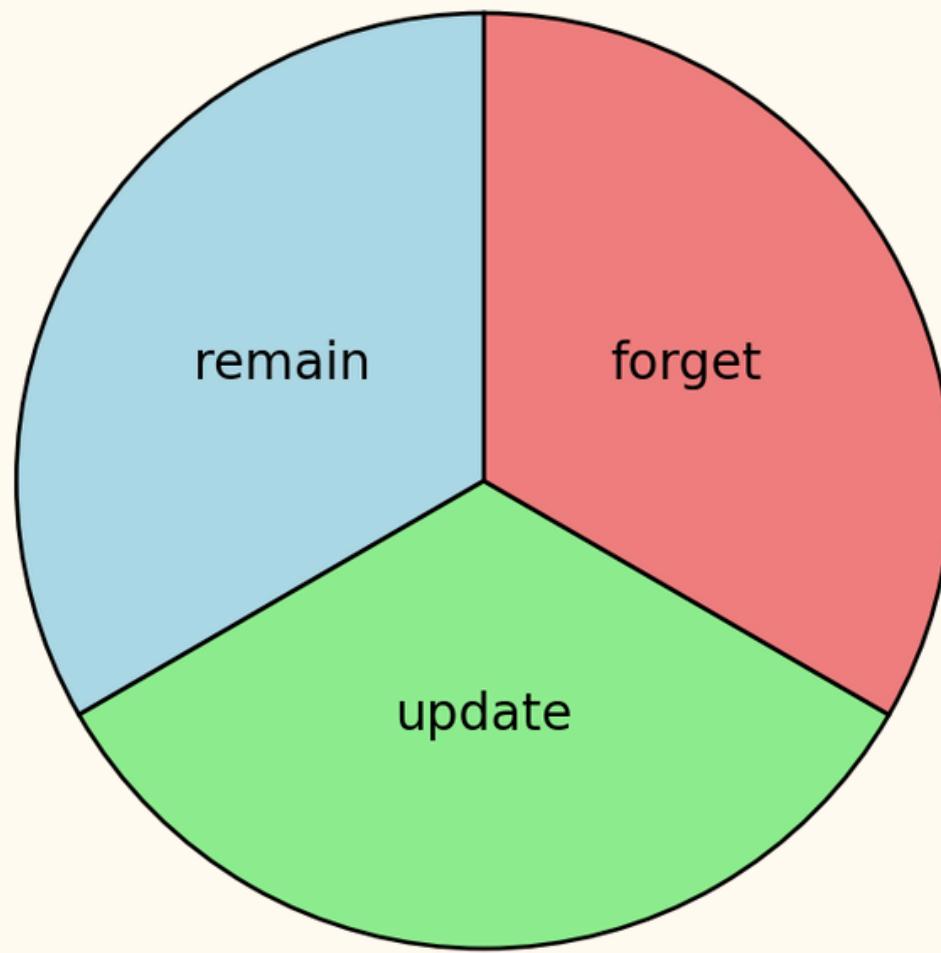


# GRU - Gated Recurrent Unit

“có con mèo ở trên sàn” → “there is a cat on the floor”

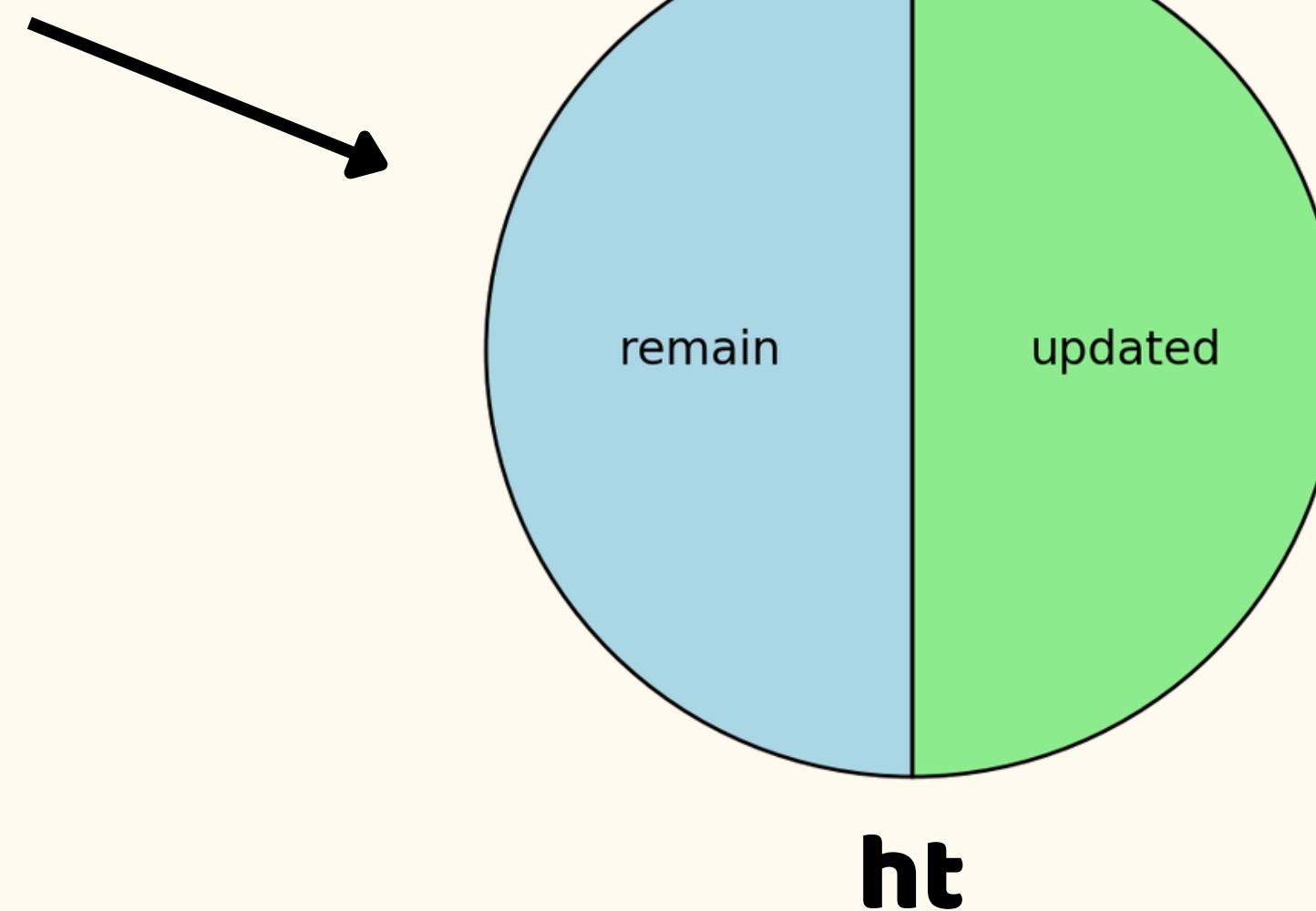


# GRU

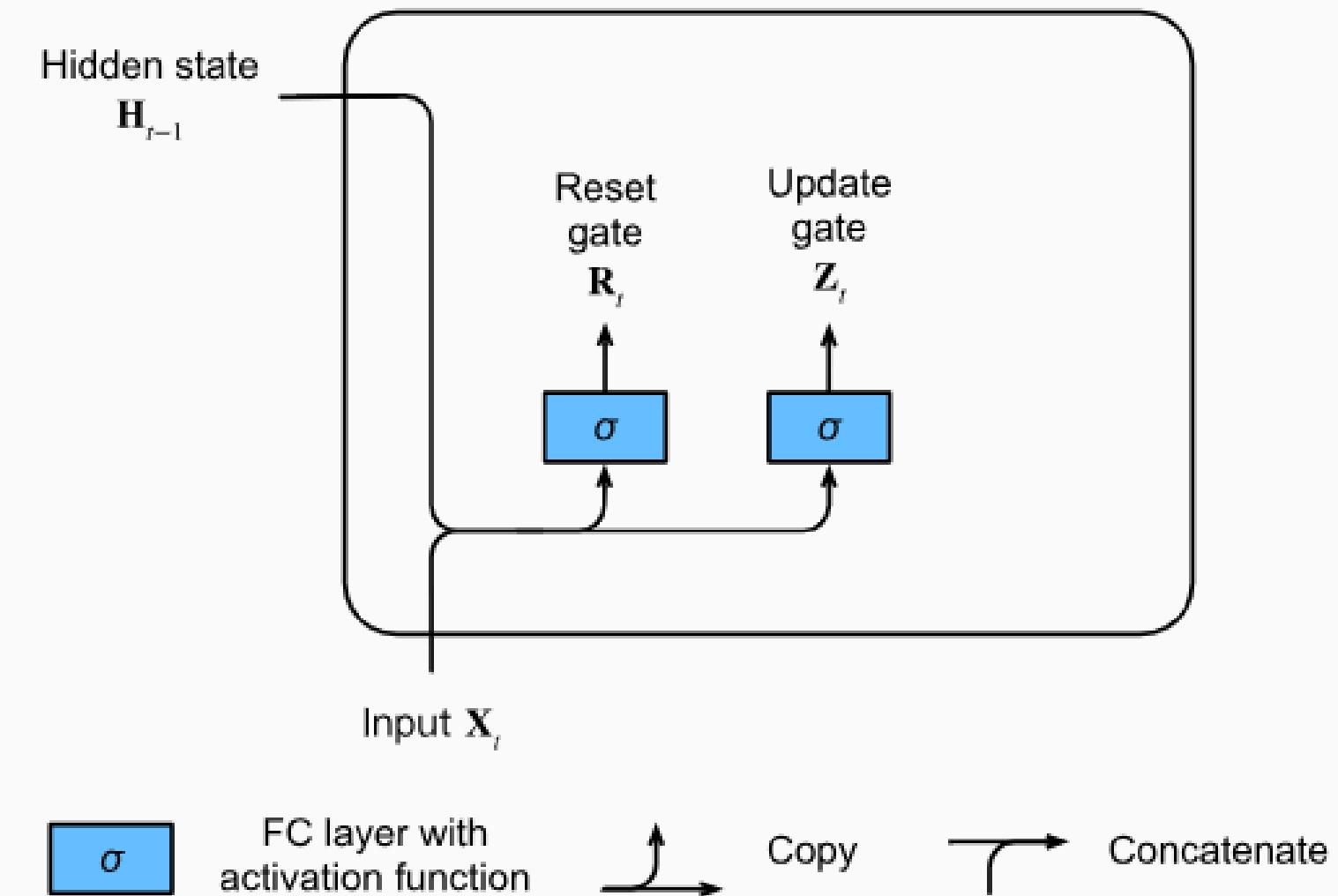


$h(t - 1) \& x_1$

Có con mèo ở trên sàn  
There is the ??

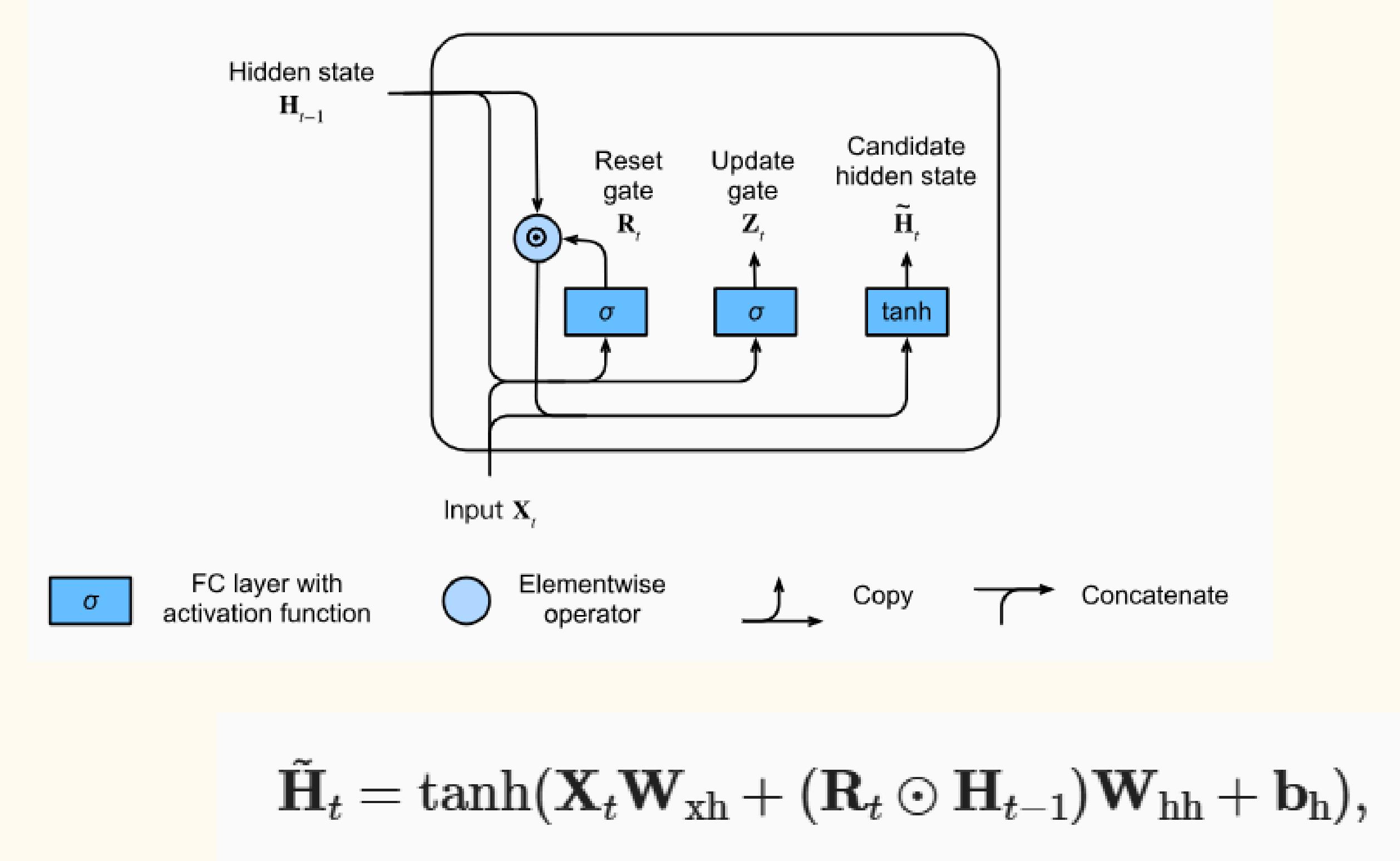


# GRU

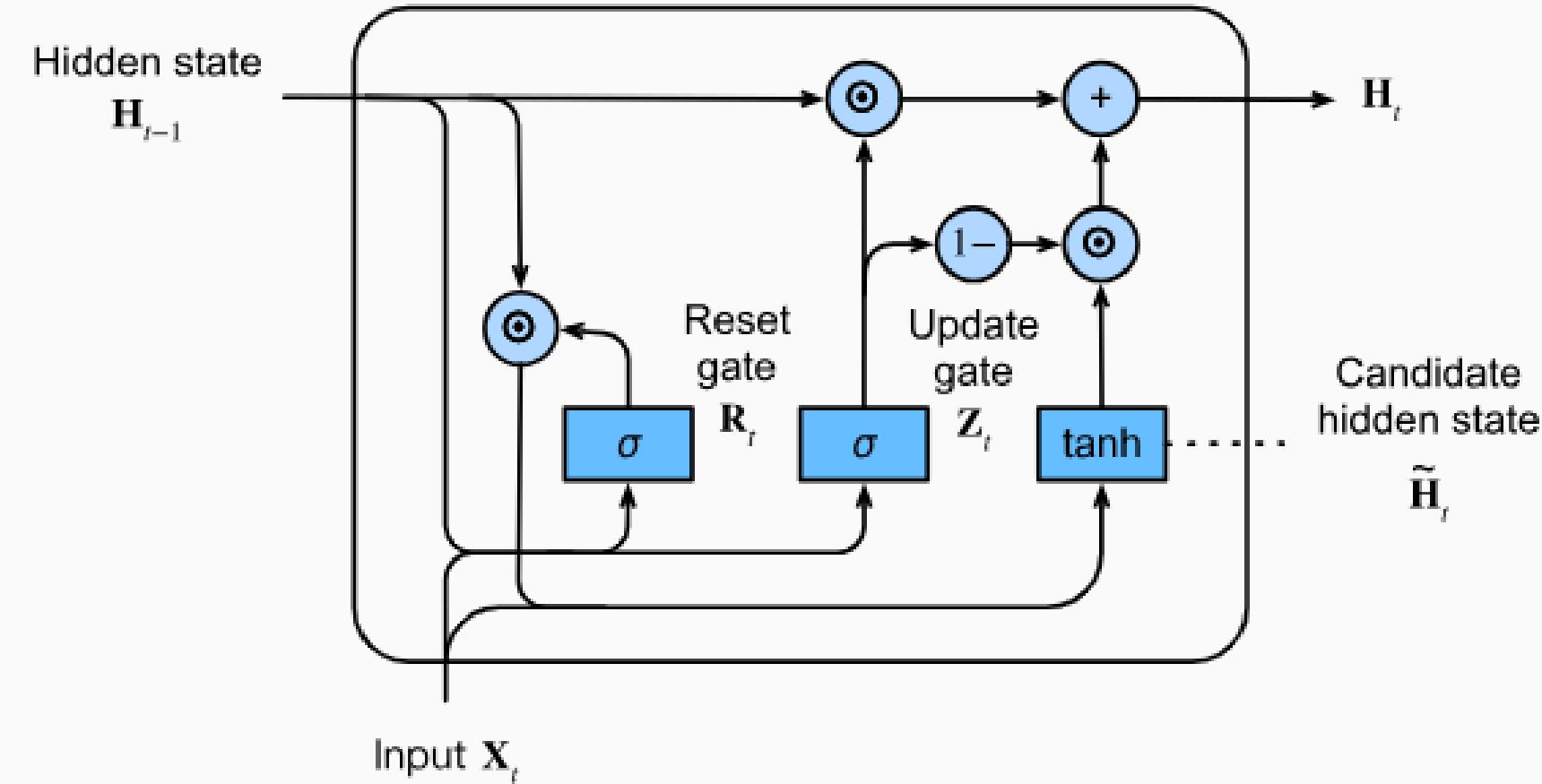


$$\begin{aligned}\mathbf{R}_t &= \sigma(\mathbf{X}_t \mathbf{W}_{\text{xr}} + \mathbf{H}_{t-1} \mathbf{W}_{\text{hr}} + \mathbf{b}_r), \\ \mathbf{Z}_t &= \sigma(\mathbf{X}_t \mathbf{W}_{\text{xz}} + \mathbf{H}_{t-1} \mathbf{W}_{\text{hz}} + \mathbf{b}_z),\end{aligned}$$

# GRU



# GRU



$\sigma$

FC layer with  
activation function

$\odot$

Elementwise  
operator

$\uparrow$

Copy

$\overline{\wedge}$

Concatenate

$$\mathbf{H}_t = \mathbf{Z}_t \odot \mathbf{H}_{t-1} + (1 - \mathbf{Z}_t) \odot \tilde{\mathbf{H}}_t.$$



# Training

```
with strategy.scope():
    model_train.fit(
        x=x_data,
        y=y_data,
        batch_size=512,
        epochs=10,
        callbacks=callbacks
    )
```

Bảng Mô Tả Epoch	
	Thông tin
1	Tổng số epoch
2	Mỗi đợt
3	Thời gian mỗi epoch



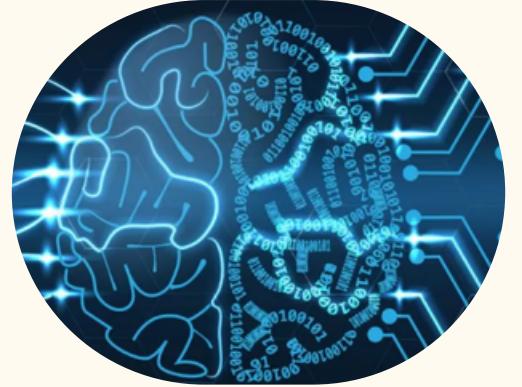
```
Epoch 1/10
5632/5632 [=====] - 1408s 248ms/step - loss: 1.8607 - val_loss: 1.4558

Epoch 00001: val_loss improved from inf to 1.45581, saving model to data_3M_train_checkpoint.keras
Epoch 2/10
5632/5632 [=====] - 1391s 247ms/step - loss: 1.3962 - val_loss: 1.2894

Epoch 00002: val_loss improved from 1.45581 to 1.28944, saving model to data_3M_train_checkpoint.keras
Epoch 3/10
5632/5632 [=====] - 1392s 247ms/step - loss: 1.2735 - val_loss: 1.2302

Epoch 00003: val_loss improved from 1.28944 to 1.23019, saving model to data_3M_train_checkpoint.keras
Epoch 4/10
5632/5632 [=====] - 1392s 247ms/step - loss: 1.2091 - val_loss: 1.1888

Epoch 00004: val_loss improved from 1.23019 to 1.18879, saving model to data_3M_train_checkpoint.keras
Epoch 5/10
5632/5632 [=====] - 1389s 247ms/step - loss: 1.1683 - val_loss: 1.1665
```



# Training

```
print(f"BLEU score: {bleu_score}")
```



```
BLEU score: 0.11498937615183447
```

```
results_df[15:20]
```

	vi	en	predict
15	Tôi và Chekaren gặp khó khăn khi chui ra khỏi ...	Chekaren and I had some difficulty getting out...	ssss i and get rough when you get out of the t...
16	Lý thuyết tiến hóa trò chơi bao gồm cả sinh họ...	Evolutionary game theory includes both biologi...	ssss the theory of evolution is as well as cul...
17	Thực tế này cho thấy rằng Baekje có kiến thức ...	This fact indicates that the Baekje had superi...	ssss this fact suggests that there is a high l...
18	Làm sao anh biết họ ở đây?	how did you know they were here?	ssss how do you know they're here eeee
19	Đã giết rất nhiều dân Nga vô tội.	Killed a lot of innocent Russians doing it.	ssss to kill many innocent russian innocent pe...



**DEMO -->**

Original sentence: Từ nơi đồng xanh thơm hương lúa, về nơi nhà cao xe giăng p  
hố

Translated: ssss from the green rice fields to the of the eeee

Original sentence: Bạn khỏe không

Translated: ssss how are you doing eeee

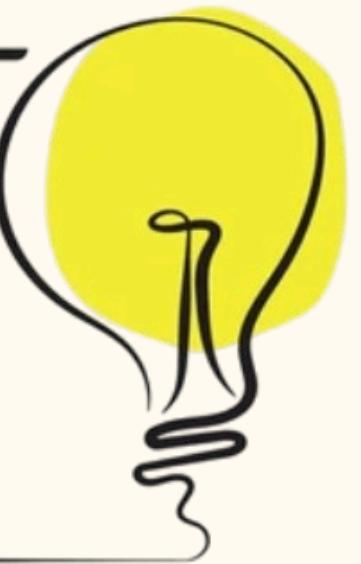
Original sentence: bạn thấy thế nào

Translated: ssss how are you feeling eeee

Original sentence: Học phát triển hệ thống thông minh

Translated: ssss learning to develop intelligent systems eeee

# Conclusion



## ^Kết luận

### KẾT QUẢ ĐẠT ĐƯỢC

phát triển được từ đầu một mô hình dịch máy từ tiếng Việt sang tiếng Anh với điểm BLEU ~ 0.115

### PHÁT TRIỂN

cải thiện hiệu suất mô hình trong tương lai với nhiều epoch hơn bằng cách train lâu hơn

# **Xin cảm ơn!**

**Hy vọng buổi trình bày này mang lại  
giá trị cho bạn.**

```

data = pd.read_csv("/kaggle/input/my-data/new_train_ds.csv")
data = data.dropna()
data.head(5)

```

## Training Set

	en	vi	source
0	- Sorry, that question's not on here.	- Xin lỗi, nhưng mà ở đây không có câu hỏi đấy.	OpenSubtitles v2018
1	He wants you to come with him immediately.	Ông ấy muốn bố đi với ông ấy ngay lập tức	OpenSubtitles v2018
2	I thought we could use some company.	Tôi nghĩ chúng ta có thể muốn vài người bạn đồn...	OpenSubtitles v2018
3	It was founded in 2008 by this anonymous progr...	Nó được sáng lập vào năm 2008 bởi một lập trìn...	TED2020 v1
4	With both of these methods, no two prints are ...	Với cả hai phương pháp, không có hai bản in nà...	TED2020 v1

## Top 20 từ Tiếng anh xuất hiện nhiều nhất

