

By Group 2

FINAL REPORT

RNN for Language Translation



Overview

- Introduction **01**
- Aims **02**
- Methodology **03**
- Experiment **04**
- Result **05**
- Analysis **06**
- Conclusion **07**



Introduction

Tầm quan trọng của dịch ngôn ngữ trong thời đại toàn cầu hóa:

- Giáo dục: Chia sẻ tài liệu học thuật và thúc đẩy trao đổi sinh viên.
- Văn hóa: Tiếp cận phim ảnh, sách báo, và âm nhạc từ các nền văn hóa khác nhau.
- Chính trị: Hỗ trợ đàm phán quốc tế và truyền tải thông điệp quốc gia.

--> ***Ứng dụng AI cho lĩnh vực sẽ mang lại giá trị lớn***

Aims

Trong dự án nhỏ này, nhóm sẽ hướng tới:

- 1 Nghiên cứu kiến trúc và cơ chế hoạt động của RNN
- 2 Xây dựng mô hình RNN cho bài toán dịch ngôn ngữ
- 3 Mở rộng bộ dataset và đánh giá hiệu quả của mô hình

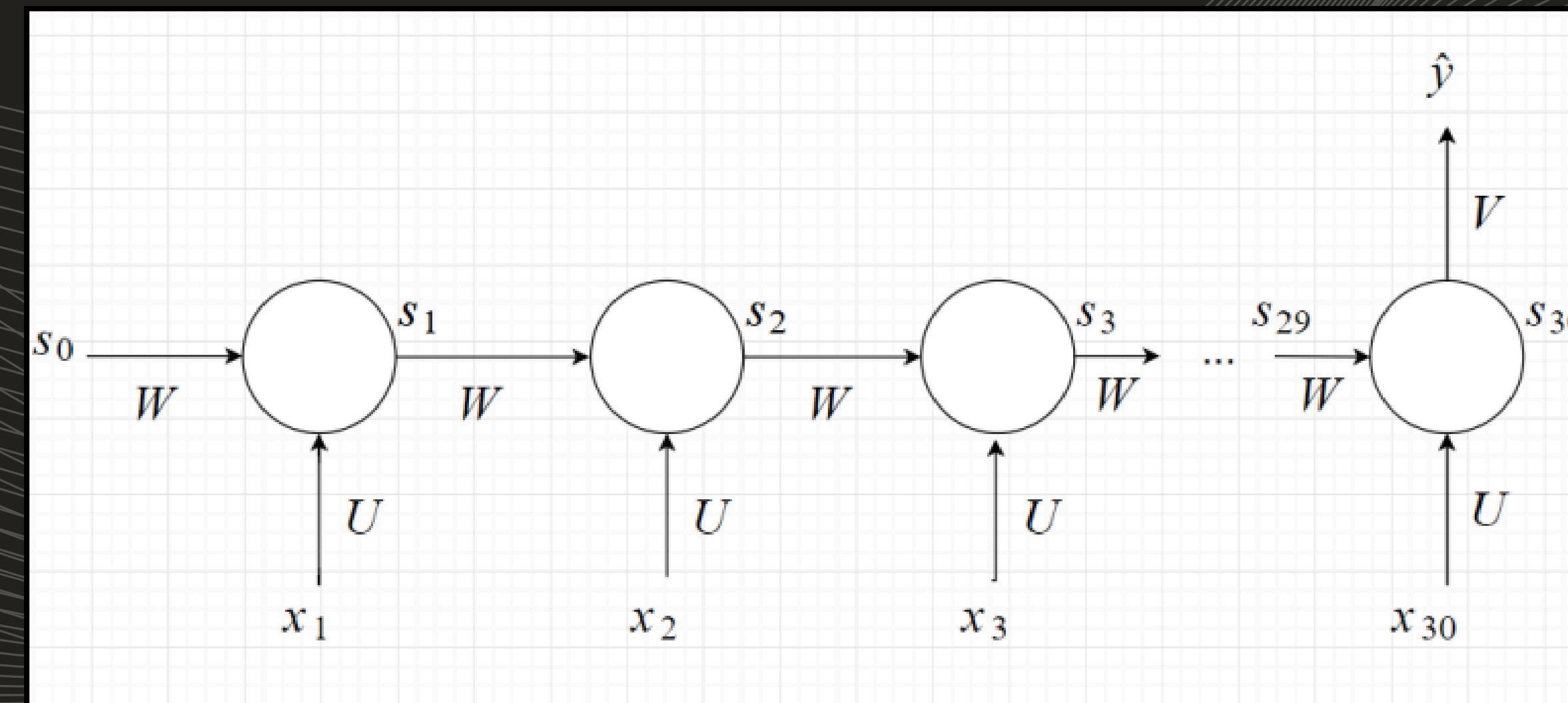
The background image shows a panoramic aerial view of a city skyline, likely Dubai, featuring a high density of modern skyscrapers, some with distinctive EMAAR branding. In the foreground, there's a large body of water with several boats and yachts docked along a waterfront promenade lined with palm trees and modern architecture.

METHODOLOGY

Recurrent Neural Network

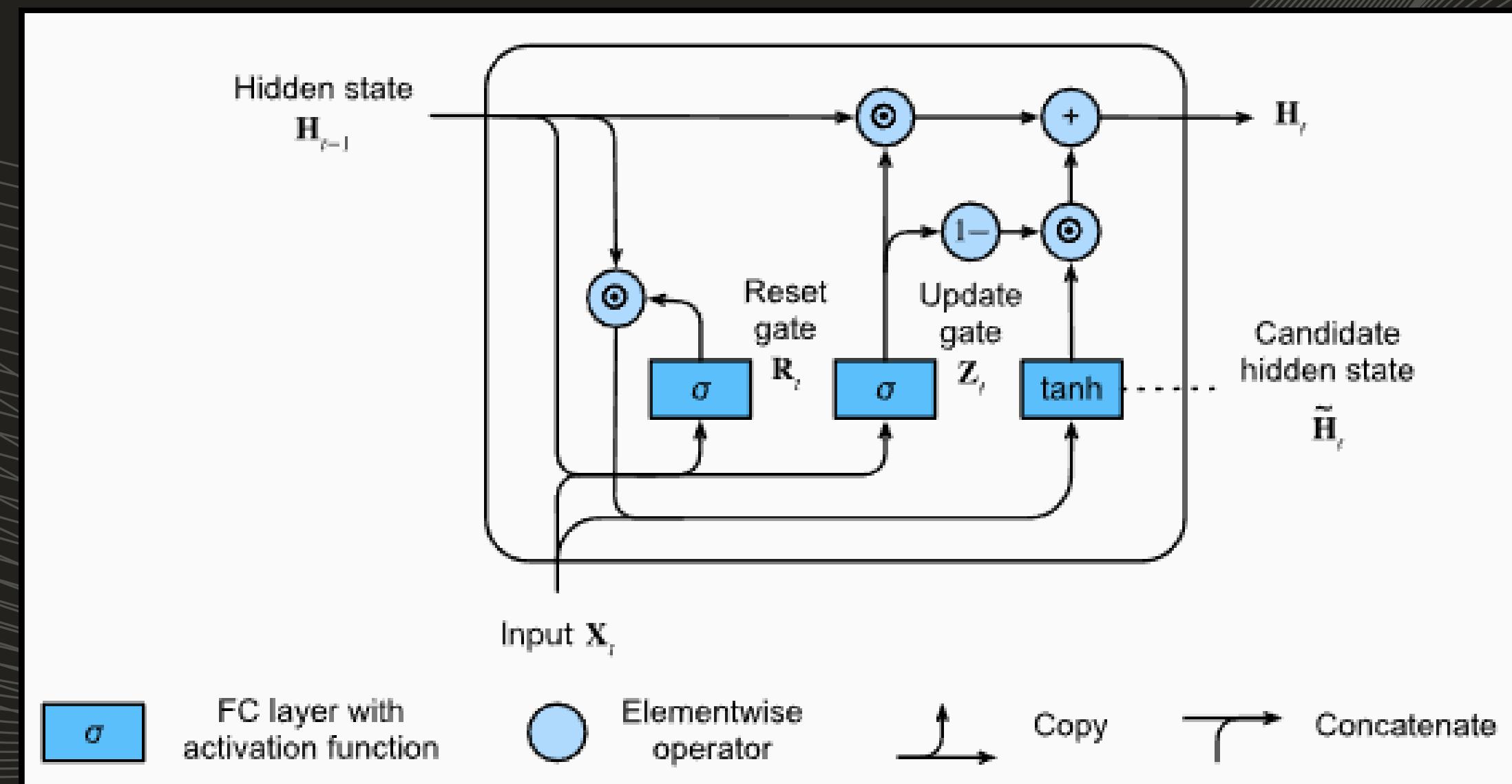
Định nghĩa:

Mạng thần kinh tái phát (RNN) là một loại mạng nơ ron nhân tạo được thiết kế cho dữ liệu tuần tự. Không giống như các mạng nơ ron truyền thẳng truyền thống, RNN có các kết nối tạo thành các chu kỳ có định hướng, cho phép thông tin có thể truyền giữ.



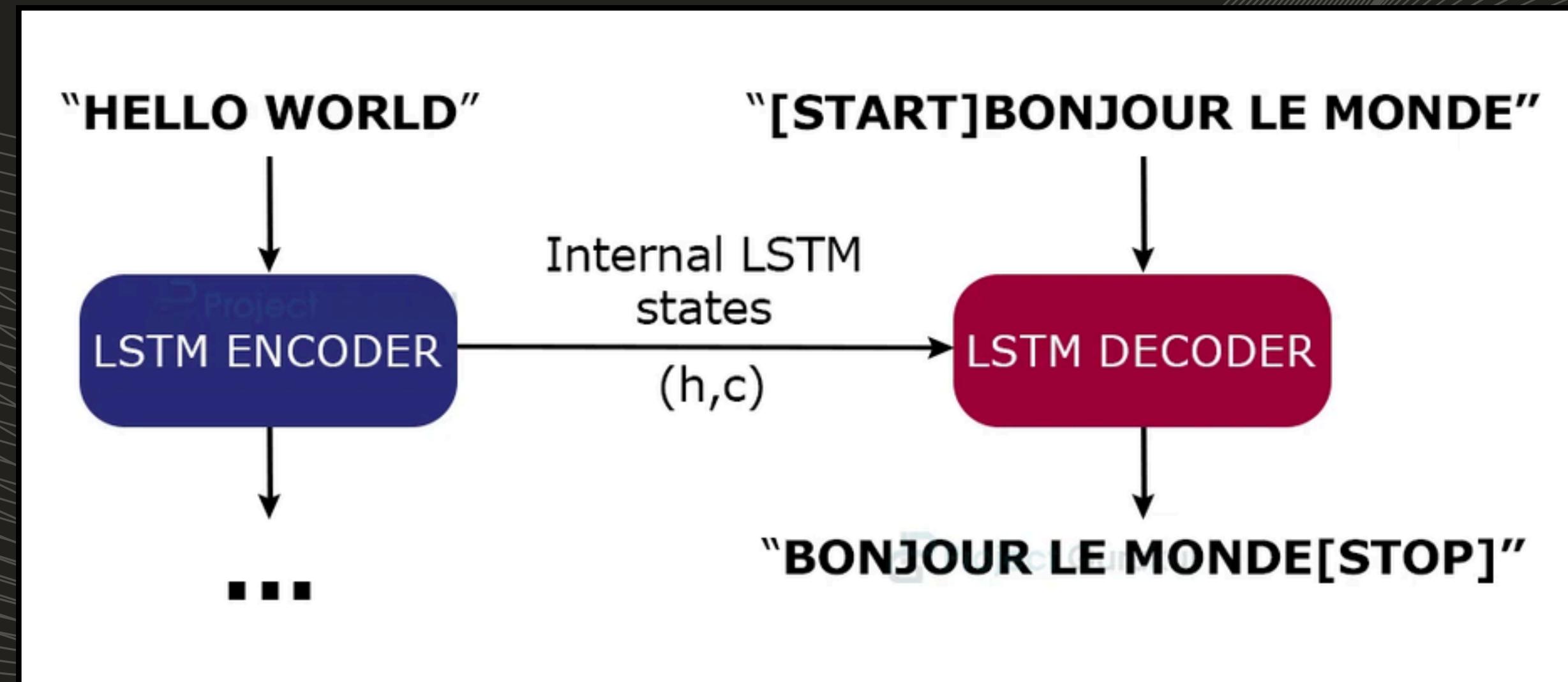
Gated Recurrent Network

Khác biệt chính giữa RNN thông thường và GRU là GRU cho phép điều khiển trạng thái ẩn, tức là ta có các cơ chế học để xem khi nào nên cập nhật và khi nào nên xóa trạng thái ẩn. Ví dụ như với các quan sát quan trọng, mô hình sẽ học để giữ nguyên trạng thái ẩn của quan sát đó. Với những quan sát không liên quan, mô hình sẽ xóa bỏ qua các trạng thái ẩn đó khi cần thiết.



Encoder - Decoder

Kiến trúc gồm có 2 phần: bộ mã hoá và bộ giải mã. Bộ mã hoá đóng vai trò mã hoá đầu vào thành trạng thái chứa vài tensor. Tiếp đó, trạng thái được truyền vào bộ giải mã để sinh ra. Trong dịch máy, bộ mã hoá biến đổi một câu nguồn thành trạng thái, chẳng hạn là một vector chứa thông tin ngữ nghĩa của câu đó. Sau đó bộ giải mã sử dụng trạng thái này để dịch câu sang ngôn ngữ đích.



The background image shows a dense urban landscape with numerous skyscrapers of varying heights. In the foreground, there's a large body of water with several boats and yachts docked along a waterfront promenade lined with palm trees. The sky is clear and blue.

EXPERIMENT

Dataset Overview

vie.txt

mn-eng-vie-translation > Text > txt > vie.txt

TieuNhatThanh2508, 2 weeks ago | 1 author (TieuNhatThanh2508)

Index	Vietnamese Phrase	Attribution	Author	
1	Run!	Chạy!	CC-BY 2.0 (France)	tatoeba.org #91
2	Run!	Chạy đi!	CC-BY 2.0 (France)	Attribution: tatoeba.org
3	Help!	Giúp tôi với!	CC-BY 2.0 (France)	Attribution: tatoeba.org
4	Help!	Cứu tôi!	CC-BY 2.0 (France)	Attribution: tatoeba.org
5	Jump!	Nhảy đi!	CC-BY 2.0 (France)	Attribution: tatoeba.org
6	Stop!	Dừng lại!	CC-BY 2.0 (France)	Attribution: tatoeba.org
7	Wait!	Đợi đã!	CC-BY 2.0 (France)	Attribution: tatoeba.org #1
8	Go on.	Tiếp tục đi.	CC-BY 2.0 (France)	Attribution: tatoeba.org
9	Hello!	Chào bạn.	CC-BY 2.0 (France)	Attribution: tatoeba.org

9429

vie.txt

A

Index	Vietnamese Word	English Translation
1	Tiếng_Việt	Tiếng_Anh
2	xin vui lòng đặt người quét rác trong tủ chổi	Please put
3	im lặng một lát	Be quiet fo
4	đọc này	Read this
5	tom thuyết phục người quản lý cửa hàng trả lại tiền cho anh ta.	Tom persu
6	tình bạn bao gồm sự hiểu biết lẫn nhau	Friendship
7	ngày mai bạn có đến không	Are you go
8	nhìn thấy vấn đề này ngay lập tức, bạn sẽ?	See to this
9	tôi đã cho bạn bè của tôi xem những tấm bưu thiếp hình ảnh.	I showed r
10	mary là em út trong ba chị em	Mary is the
11	anh ấy có hai người dì ở bên mẹ.	He has two
12	đó là những gì tôi muốn biết	That is wh

254090

Book1.xlsx

Dataset Overview

TieuNhatThanh2508, 19 hours ago | 1 author (TieuNhatThanh2508)

xin vui lòng đặt người quét rác trong tủ chổi Please put the dustpan in the broom closet
im lặng một lát Be quiet for a moment.

TieuNhatThanh2508, 19 hours

đọc này Read this

tom thuyết phục người quản lý cửa hàng trả lại tiền cho anh ta. Tom persuaded the store manager to give him back his money.

tình bạn bao gồm sự hiểu biết lẫn nhau Friendship consists of mutual understanding

ngày mai bạn có đến không Are you going to come tomorrow?

nhìn thấy vấn đề này ngay lập tức, bạn sẽ? See to this matter right away, will you?

tôi đã cho bạn bè của tôi xem những tấm bưu thiếp hình ảnh. I showed my friends these picture postcards.

mary là em út trong ba chị em Mary is the youngest of the three sisters

anh ấy có hai người dì ở bên mẹ. He has two aunts on his mother's side.

đó là những gì tôi muốn biết That is what I want to know

hành tây có thể được sử dụng trong nhiều món ăn. Onions can be used in many dishes.

Row: 263482

combined_data.txt

System Specification

- Hệ thống của máy tính x64 (64-bit).
- Processor: Intel(R) Core(TM) i7-1065G7 CPU @ 1.30GHz, 1498 Mhz, 4 Core(s), 8 Logical Processor(s).
- Installed Physical Memory (RAM): 16.0 GB.
- Total Physical Memory: Dung lượng RAM khả dụng sau khi trừ đi phần dành cho phần cứng khác là 15.8 GB.
- Available Physical Memory: Dung lượng RAM khả dụng cho các ứng dụng và hệ thống hiện tại là 5.42 GB.

Hyperparameters

- batch_size = 64
- epochs = 10
- latent_dim = 256
- num_samples = 40000
- Number of unique input Tokens: 167
- Number of unique output Tokens: 94
- Max sequence length for inputs: 193
- Max sequence length for outputs: 192

Vấn đề phát sinh

1. Lỗi bộ nhớ “Unable to allocate 14.2 GiB for an array with shape (263518, 263, 110) and data type float32”

- Nguyên nhân:
 - i. Kích thước dữ liệu lớn: Mảng được tạo có kích thước quá lớn ($263518 \times 263 \times 110$) và sử dụng kiểu dữ liệu float32.
 - ii. Hạn chế phần cứng: Bộ nhớ RAM của máy tính không đủ lớn để chứa mảng này.
- Giải pháp:
 - Xóa bớt data, giảm từ 263,518 samples xuống còn 227,546 samples.
 - Sử dụng kiểu dữ liệu nhỏ hơn: sử dụng kiểu dữ liệu dùng ít bộ nhớ hơn, ví dụ như float16 thay vì float32.



Vấn đề phát sinh

2. Máy bị đơ khi chạy đoạn mã chuyển đổi các chuỗi văn bản (câu) thành các mảng nhị phân (one-hot encoding) để mô hình tuân tự như seq2seq (sequence-to-sequence) có thể sử dụng.

- Nguyên nhân:

- Máy cố gắng tạo và thao tác trên các mảng lớn trong bộ nhớ, dẫn đến việc tiêu thụ quá nhiều tài nguyên và vượt quá khả năng xử lý của hệ thống.

- Giải pháp:

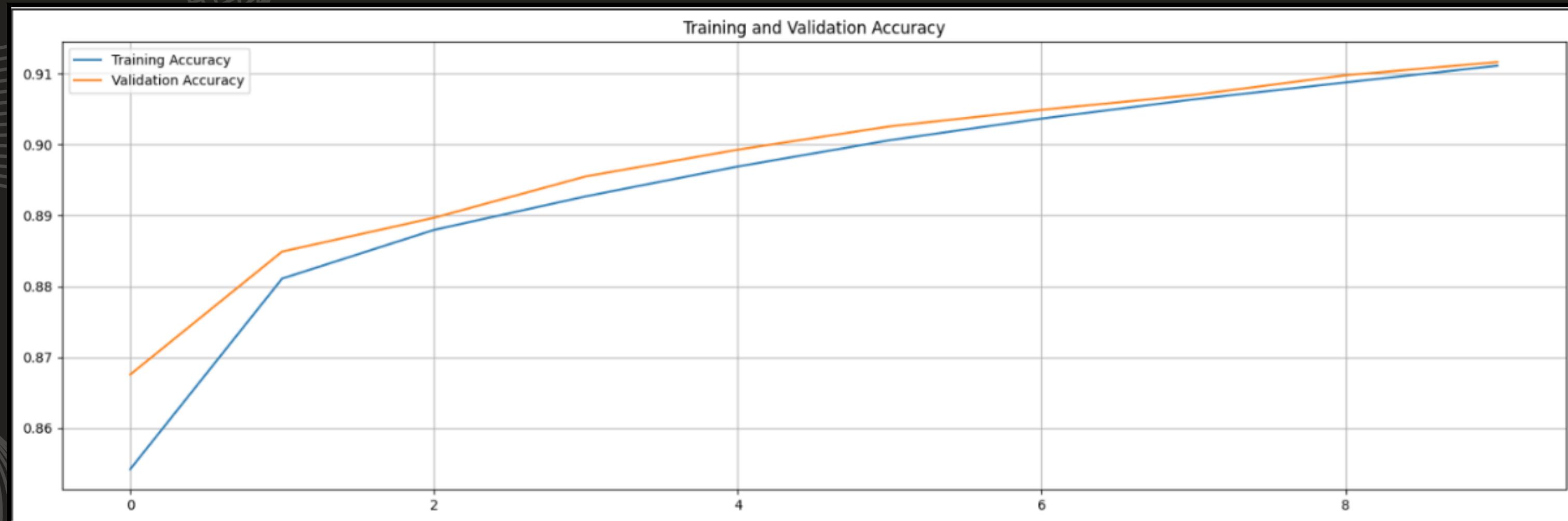
- Giới hạn lượng samples “num_samples = 40,000” . Lấy 40,000 samples đầu tiên trong tổng cộng 227,546 samples ban đầu sẽ được sử dụng trong quá trình huấn luyện.



The background image shows a dense urban landscape with numerous skyscrapers, some featuring the "EMaar" logo. In the foreground, there's a large body of water with several boats and yachts docked along a waterfront promenade lined with palm trees and modern buildings.

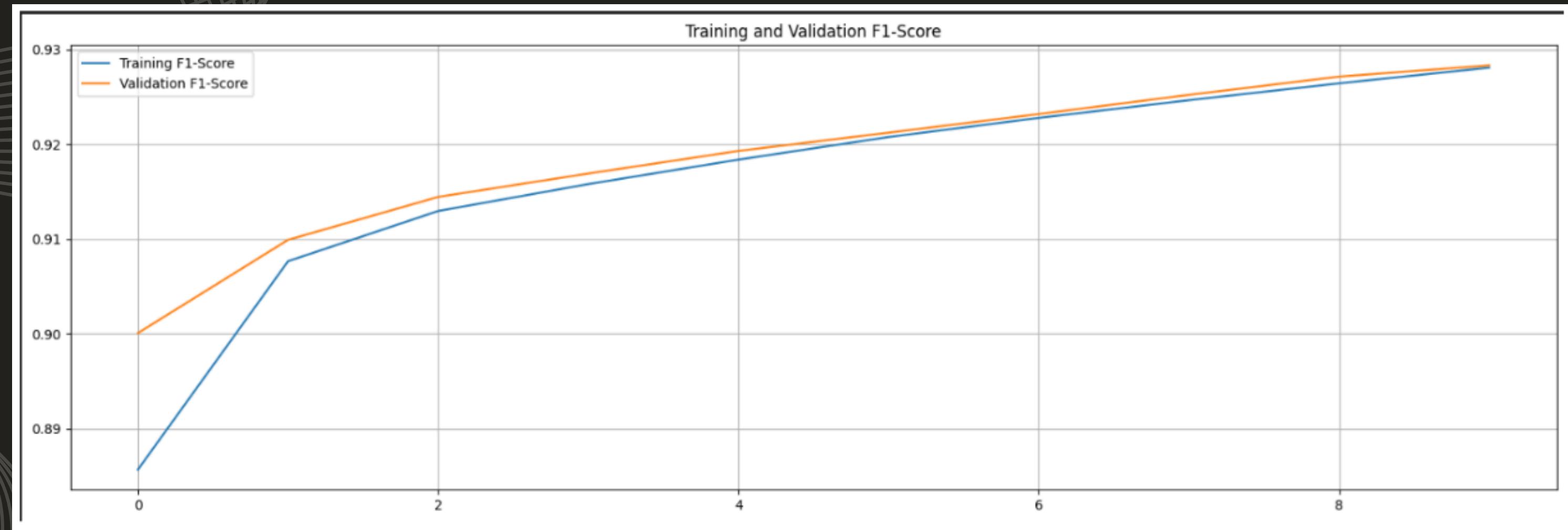
RESULT

Accuracy



Accuracy: 91.06%

F1-Score



F1_score: 0.9277

Statistics

128m37s

Training time

Accuracy

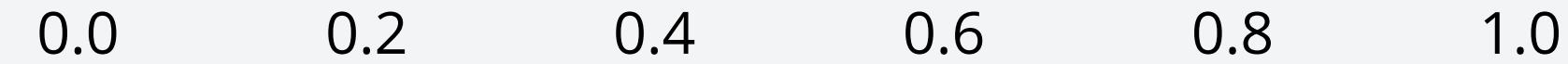
Val_acc

F1-Score

Val_F1

Loss

Val_loss



Error Analysis

- Kết luận:

- Qua 10 epoch, mô hình đã cải thiện đáng kể về độ chính xác, F1_score, và hàm mất mát. Kết quả cuối cùng cho thấy mô hình có hiệu quả cao trong việc phân loại và dự đoán.
- Độ chính xác cao và gần như tương đương giữa tập huấn luyện và tập kiểm tra.
- Điểm F1 cao, cho thấy khả năng cân bằng giữa precision và recall.
- Hàm mất mát giảm dần và đạt mức thấp, chứng tỏ mô hình học tốt và giảm thiểu sai số.
- Việc tăng samples cho dataset chỉ cải thiện một phần nhỏ độ hiệu quả so với lần huấn luyện trước

eng-vie
Loss
Accuracy
F1 score
Training time
Validation loss
Avalidation accuracy
Validation F1 score
Number of tokens

Previous Model

Conclusion

- Nhóm đã khám phá và áp dụng RNN vào dịch ngôn ngữ, từ việc thu thập dữ liệu đến xây dựng và huấn luyện mô hình.
- Mô hình RNN đã cho thấy khả năng dịch ngôn ngữ tốt với các chỉ số đánh giá tích cực, cụ thể là trong phạm trù dịch ngôn ngữ Anh - Việt.
- Hướng phát triển tương lai: Cần nghiên cứu và cải tiến thêm, chẳng hạn như tích hợp các kỹ thuật tiên tiến hơn như Transformer hoặc sử dụng RNN kết hợp với các mô hình khác để nâng cao chất lượng dịch.



Thank You

Nhóm 2

Nguyễn Đỗ Nhất Huy

Tiêu Nhật Thành

Nguyễn Thành Nhân

Gerente General