ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH

TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

BÁO CÁO ĐỒ ÁN 1

Môn học: Khai thác dữ liệu văn bản và dữ liệu

Giáo viên hướng dẫn: Nguyễn Trần Duy Minh

Lê Thanh Tùng

Thành viên:

21127680 – Trần Văn Quyết

19127336 – La Gia Bảo

21127577 – Trịnh Hoàng An

21127056 – Lâm Thiều Huy

Mục lục

Bảng phân công việc **2**
Mức độ hoàn thành **3**
I. Exploratory Data Analysis **4**
1. Độ dài của câu hỏi: 4
2. Độ dài của đoạn văn: 4
3. Phân bố của nhãn: 5
4. Hiệu suất mô hình: 5
5. Các từ phổ biến trong câu hỏi: 6
6. Các từ phổ biến trong đoạn văn: 6
II. BARTPho, CNN + FCL **6**
1. Code Structure: 6
2. Suggestion: 8
3. Kết quả model: 8
III. Thử Nghiệm Trên Mô Hình Đã Có Sẵn **8**
1. Mô Hình : T5 (Text-to-Text Transfer Transformer) 8
2. Điểm chính và cách hoạt động 8
3. Thử nghiệm mô hình với bộ dataset 10

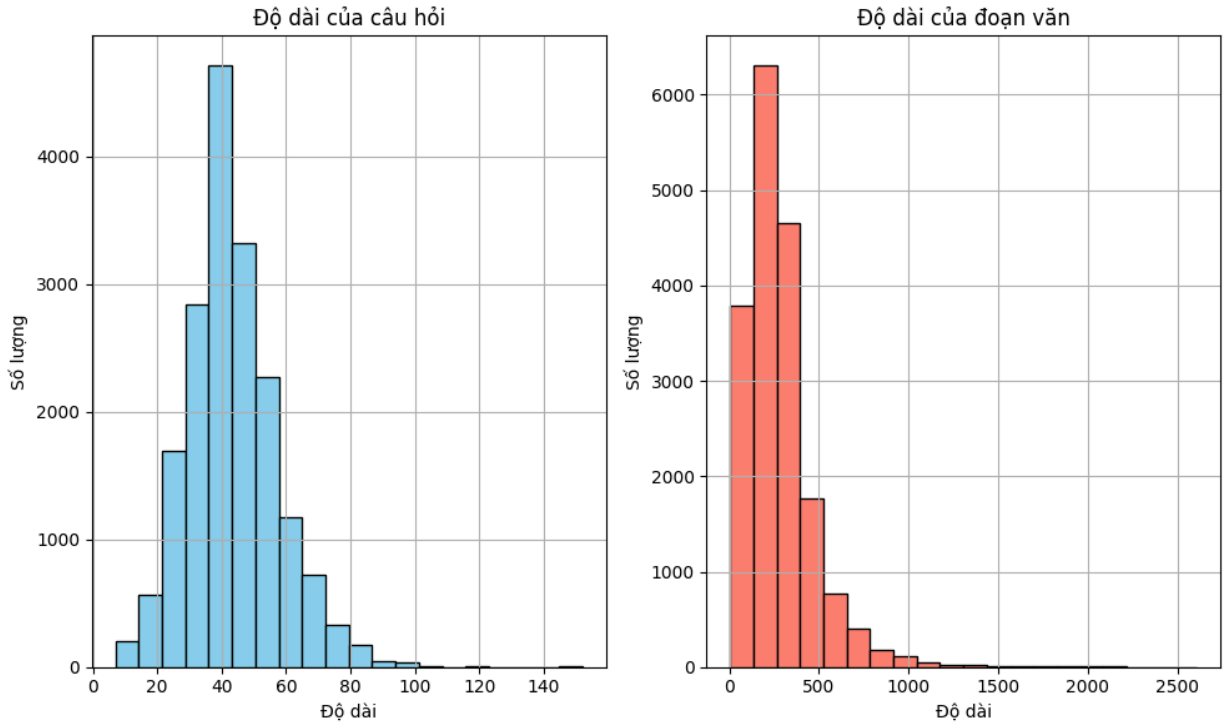
Bảng phân công việc

Tên thành viên	Nhiệm vụ được phân công
Trần Văn Quyết	- Thử nghiệm trên mô hình tự xây dựng (2.2)
La Gia Bảo	- Thử nghiệm trên mô hình đã có sẵn (III) (2.3)
Trịnh Hoàng An	- Thử nghiệm trên mô hình đã có sẵn (III) (2.3)
Lâm Thiều Huy	- Exploratory Data Analysis (I) (2.1) - Report

Mức độ hoàn thành

- Trần Văn Quyết: 100%.Train model sử dụng BARTPho embedding, CNN và Fully Connected Layer cho việc classifier. Đánh giá metric accuracy của model, validation trên từng epoch. Chưa hyper parameter và đánh giá metric khác nhau.
- La Gia Bảo: 50%.
- Trịnh Hoàng An: 100%. Tìm hiểu, training, đánh giá mô hình bằng bộ dữ liệu được giao: Đánh giá thời gian chạy, epoch, learning rate, gradient norm, loss. Tìm hiểu được chức năng, cách hoạt động của mô hình. Đánh giá tổng thể mô hình.
- Lâm Thiều Huy: 100%. Phân tích và thống kê dữ liệu: xác định độ dài của câu hỏi và đoạn văn, tính toán độ dài trung bình, tối đa và tối thiểu của chúng, phân tích phân bố của nhãn trong tập dữ liệu để hiểu sự cân bằng giữa các lớp, xác định các từ phổ biến nhất trong câu hỏi và đoạn văn để hiểu cấu trúc và ngữ cảnh của dữ liệu. Trực quan hóa dữ liệu: sử dụng biểu đồ histogram để thể hiện phân bố của độ dài câu hỏi và đoạn văn, vẽ biểu đồ cột để thể hiện phân bố của nhãn trong tập dữ liệu, sử dụng biểu đồ cột hoặc bảng để thể hiện các từ phổ biến nhất trong câu hỏi và đoạn văn. Nhận xét và phân tích kết quả: Dựa trên thông tin từ phân tích, đưa ra nhận xét và suy luận về dữ liệu, bao gồm cấu trúc, tính đa dạng, sự cân bằng và ngữ cảnh của câu hỏi và đoạn văn, phân tích những thách thức và cơ hội có thể phát sinh khi xử lý dữ liệu hoặc phát triển mô hình học máy.

I. Exploratory Data Analysis

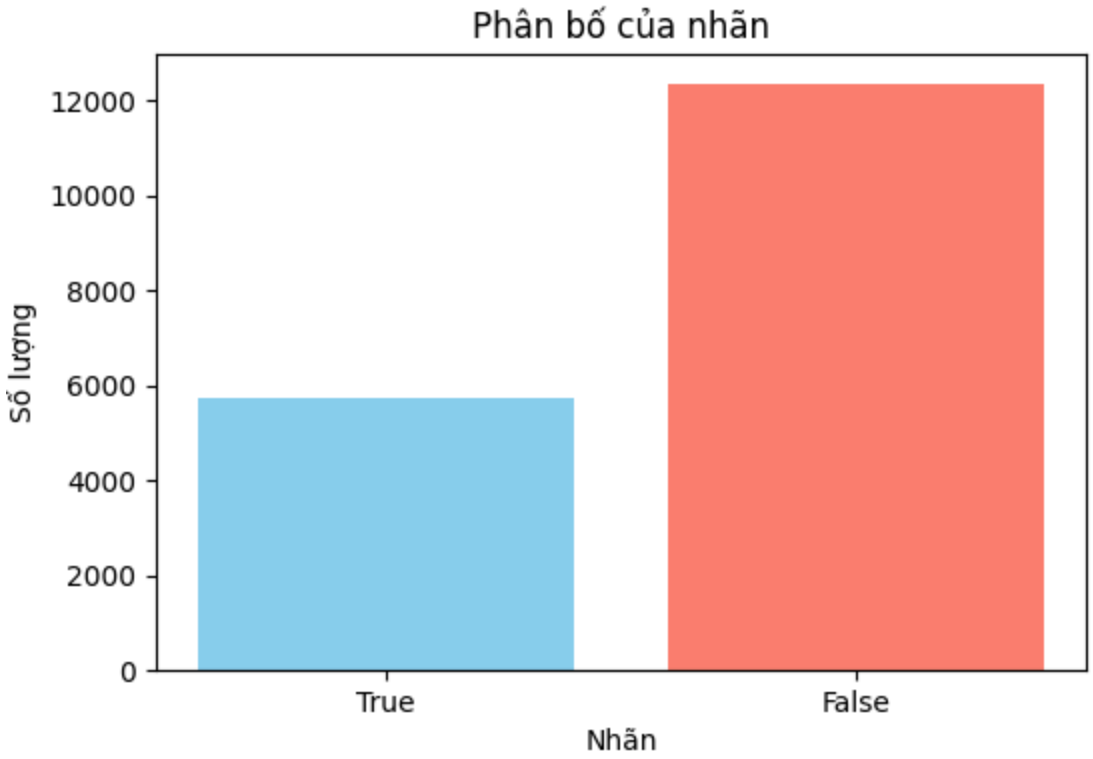


1. Độ dài của câu hỏi:

- **Độ dài trung bình** của câu hỏi là khoảng 43 ký tự, cho thấy các câu hỏi trong tập dữ liệu thường ngắn và trực tiếp. Điều này có thể cho thấy dữ liệu có thể tập trung vào các câu hỏi đơn giản hoặc cụ thể.
- **Độ dài tối đa** của câu hỏi là 152 ký tự, và **độ dài tối thiểu** là 7 ký tự. Sự đa dạng về độ dài này cho thấy có sự biến động lớn trong mức độ chi tiết và phức tạp của các câu hỏi.

2. Độ dài của đoạn văn:

- **Độ dài trung bình** của đoạn văn là khoảng 279 từ. Điều này cho thấy đoạn văn trong tập dữ liệu thường có kích thước trung bình, với đủ thông tin để cung cấp câu trả lời cho câu hỏi.
- **Độ dài tối đa** của đoạn văn là 2609 từ, và **độ dài tối thiểu** là 4 từ. Một độ biến động lớn như vậy trong độ dài của các đoạn văn có thể gợi ý đến sự đa dạng của kiểu dữ liệu, từ thông tin tổng quát đến thông tin chi tiết hoặc ngắn gọn.

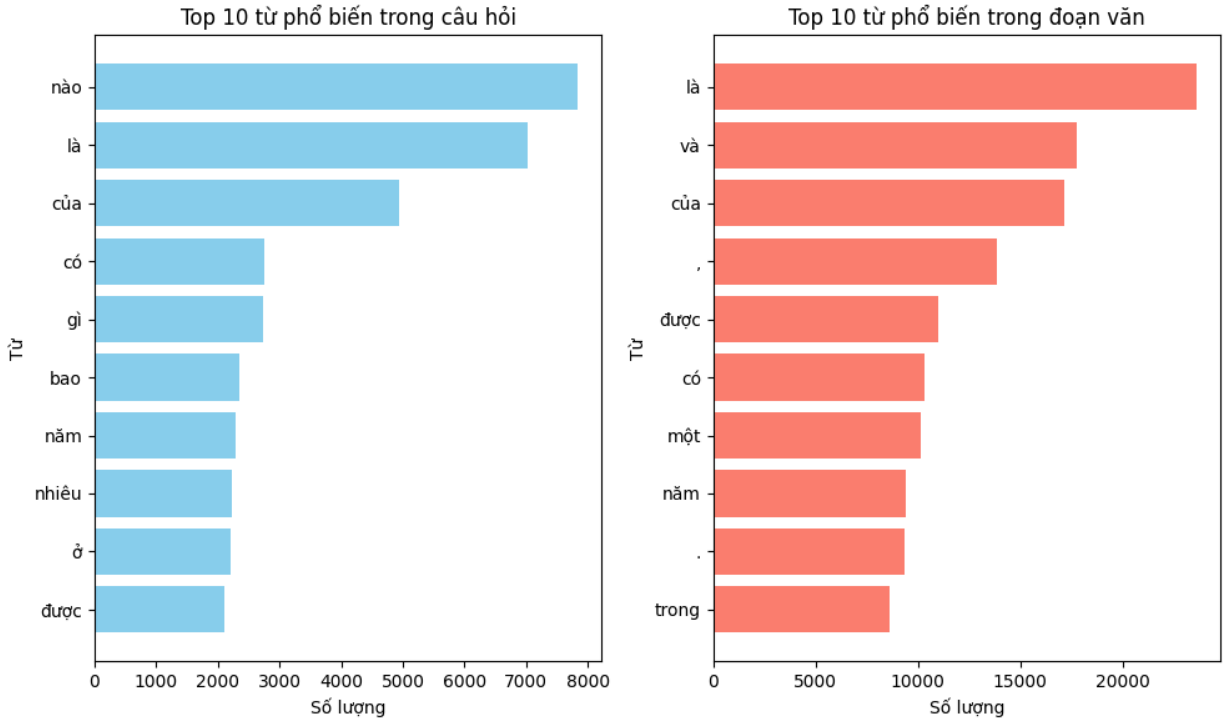


3. Phân bố của nhãn:

- Số lượng nhãn True là 5738 và số lượng nhãn False là 12370.
- Tỷ lệ của nhãn True so với nhãn False là khoảng 31.7% , và tỷ lệ của nhãn False so với nhãn True là khoảng 68.3%.
- Sự chênh lệch lớn giữa số lượng các nhãn True và False có thể ảnh hưởng đến hiệu suất của mô hình, đặc biệt trong các bài toán phân loại không cân bằng.

4. Hiệu suất mô hình:

- Do sự không cân bằng trong số lượng các nhãn, mô hình có thể có xu hướng dự đoán nhãn nhiều hơn cho lớp có số lượng lớn hơn, điều này có thể dẫn đến việc mô hình không nhạy với các trường hợp của lớp thiểu số.



5. Các từ phổ biến trong câu hỏi:

- Trong top 10 từ phổ biến nhất trong câu hỏi, chúng ta thấy các từ như "nào", "là", "của", "có", "gì", "bao", "năm", "nhiêu", "ở", "được" đều là các từ phổ biến trong câu hỏi tiếng Việt thông thường.
- Các từ này thường xuất hiện trong các câu hỏi để hỏi về thông tin cụ thể hoặc yêu cầu giải thích về một sự kiện, một số liệu, hoặc một thực thể nào đó.

6. Các từ phổ biến trong đoạn văn:

- Trong top 10 từ phổ biến nhất trong đoạn văn, chúng ta thấy các từ như "là", "và", "của", ",", "được", "có", "một", "năm", ".", "trong" thường là các từ nền được sử dụng trong văn bản tiếng Việt thông thường.
- Các từ này thường xuất hiện trong các đoạn văn để mô tả, diễn đạt ý kiến hoặc cung cấp thông tin.

II. BARTPho, CNN + FCL

1. Code Structure:

1. BARTPhoEmbedding:

- Sử dụng thư viện transformers của Hugging Face để tải một mô hình BARTPho và tokenizer được pretrained.
- Nhận một văn bản đầu vào và tạo các nhúng BARTPho bằng mô hình được pretrained.

2. TextClassifier:

- Một bộ phân loại văn bản dựa trên CNN đơn giản nhận đầu vào là các nhúng BARTPho.
- Các lớp tích chập được theo sau bởi max-pooling và một lớp kết nối đầy đủ với dropout.
- Kích thước đầu ra được điều chỉnh dựa trên số lớp.

3. VNQADataset:

- Lớp tập dữ liệu tùy chỉnh để xử lý dữ liệu đầu vào.
- Tokenize và embedding câu hỏi và văn bản bằng tokenizer BARTPho.
- Cung cấp các tensor đầu vào và nhãn cho việc train và đánh giá.

4. Train_and_evaluate function:

- Train và đánh giá mô hình trong một số lượng epoch cụ thể.
- Sử dụng hàm mất mát cross-entropy và tối ưu hóa Adam cho quá trình training.
- In độ chính xác trên tập kiểm thử cho mỗi epoch.

5. Tiền Xử Lý Dữ Liệu:

- Tải dữ liệu từ một tệp JSON và chia thành tập train và test.
- Khởi tạo tokenizer BARTPho và tạo các phiên bản của lớp tập dữ liệu tùy chỉnh.
- Tạo các bộ tải dữ liệu cho quá trình train và kiểm thử.

6. Khởi Tạo Mô Hình:

- Khởi tạo mô hình embedding BARTPho và mô hình phân loại văn bản.
- Xác định hàm mất mát (cross-entropy) và bộ tối ưu hóa (Adam).

7. Train và Đánh Giá:

- Gọi hàm train_and_evaluate để đào tạo mô hình và đánh giá hiệu suất trên tập test
- In độ chính xác trên tập kiểm thử cho mỗi epoch.

2. Suggestion:

- Model hiện tại có vẻ đơn giản. Hãy thử nghiệm với các kiến trúc khác nhau, đặc biệt là trong các lớp tích chập, để cải thiện hiệu suất.
- Hyper parameter tuning: với GridSearch sklearn để tìm và bộ siêu tham số tốt nhất cho model.

3. Kết quả model:

- Epoch 1:Test Accuracy: 69.03%
- Epoch 2: Test Accuracy: 69.03%

III. Thử Nghiệm Trên Mô Hình Đã Có Sẵn

1. Mô Hình : T5 (Text-to-Text Transfer Transformer)

Mô hình T5, hay "Text-to-Text Transfer Transformer", là một kiến trúc dựa trên Transformer được thiết kế để xử lý các nhiệm vụ xử lý ngôn ngữ tự nhiên (NLP) theo cách mới mẻ: chuyển đổi tất cả các nhiệm vụ về dạng "text-to-text". Mô hình này được giới thiệu bởi các nhà nghiên cứu tại Google Research trong bài báo "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer" vào tháng 6 năm 2020.

2. Điểm chính và cách hoạt động

a. Kiến trúc Chuyển đổi Text-to-Text:

- Tất cả các nhiệm vụ NLP, từ dịch máy, tóm tắt văn bản, đến phân loại câu và hỏi đáp, đều được mô hình T5 xử lý bằng cách chuyển đổi từ dạng văn bản này sang dạng văn bản khác.
- Đầu vào và đầu ra của mô hình đều là văn bản. Ví dụ: trong nhiệm vụ phân loại cảm xúc, đầu vào có thể là "nhiệm vụ: phân loại cảm xúc, văn bản: Tôi thấy rất vui hôm nay" và đầu ra sẽ là "tích cực".

b. Pre-training và Fine-tuning:

- Mô hình T5 được tiền huấn luyện trên một tập hợp dữ liệu lớn bằng cách sử dụng các kỹ thuật như "masked language modeling", nơi một phần của văn bản đầu vào được ẩn đi và mô hình cố gắng dự đoán nó.

- Sau đó, mô hình được tinh chỉnh cho các nhiệm vụ cụ thể bằng cách sử dụng tập dữ liệu nhỏ hơn chứa các ví dụ cụ thể từ nhiệm vụ đó.

c. Các Biến thể của Mô hình:

T5 có sẵn trong nhiều biến thể với số lượng tham số khác nhau, từ T5-Small đến T5-3B và thậm chí là T5-11B, cho phép nó được sử dụng trong các tình huống với yêu cầu tài nguyên khác nhau.

d. Linh Hoạt và Mạnh Mẽ:

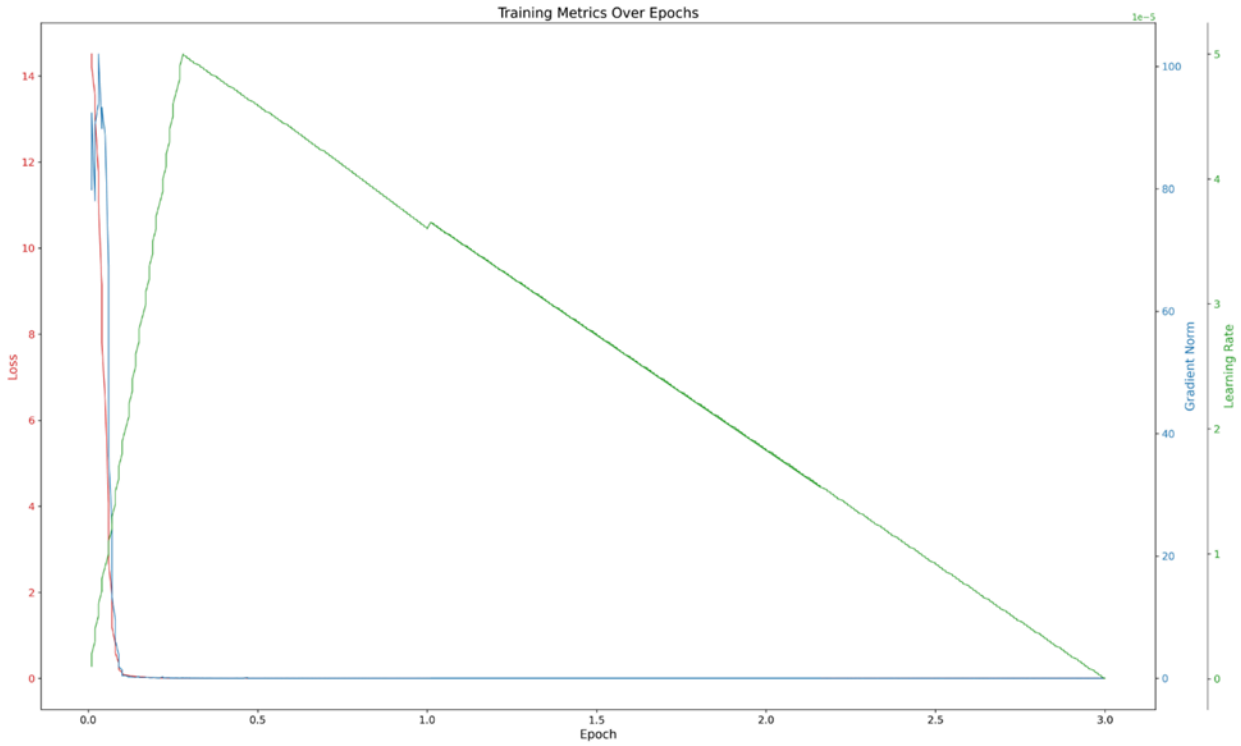
- Do thiết kế linh hoạt của nó, T5 có thể được áp dụng cho hầu như bất kỳ nhiệm vụ NLP nào mà không cần thay đổi cơ bản về kiến trúc mô hình.
- Mô hình đã thiết lập các kỷ lục mới trên nhiều bảng xếp hạng benchmark NLP khi nó được giới thiệu, chứng minh sự hiệu quả của nó trong việc xử lý các nhiệm vụ khác nhau.

e. Đào tạo và Triển khai:

- Việc đào tạo T5 yêu cầu tài nguyên máy tính lớn, thường là các GPU hoặc TPU chuyên dụng, do kích thước lớn của mô hình và tập dữ liệu tiền huấn luyện.
- Mô hình có thể được tối ưu hóa và triển khai trong các ứng dụng thực tế, từ các hệ thống chatbot đến công cụ tìm kiếm và hệ thống phân tích cảm xúc.

T5 đã góp phần làm thay đổi cách tiếp cận các nhiệm vụ NLP, chuyển từ các mô hình chuyên biệt cho mỗi nhiệm vụ sang một kiến trúc thống nhất có thể xử lý nhiều loại nhiệm vụ khác nhau chỉ bằng cách thay đổi dữ liệu đầu vào và đầu ra.

3. Thử nghiệm mô hình với bộ dataset



3.1: Biểu đồ "Training Metrics Over Epochs"

Dựa trên biểu đồ "Training Metrics Over Epochs" bạn cung cấp, ta có thể đánh giá quá trình huấn luyện mô hình dựa trên các chỉ số sau:

- I. Loss (Mất mát): Đường màu đỏ biểu diễn giá trị loss của mô hình qua từng epoch. Ban đầu, loss rất cao nhưng nhanh chóng giảm mạnh và ổn định sau khoảng 0.5 epoch. Sự giảm mạnh của loss trong giai đoạn đầu chỉ ra rằng mô hình đang học hiệu quả từ dữ liệu. Tuy nhiên, sau khoảng thời gian đầu, tốc độ giảm của loss trở nên chậm lại, dần tiệm cận với một giá trị nhỏ, điều này cho thấy mô hình bắt đầu ổn định và đã đạt được khả năng tổng quát hóa tốt trên dữ liệu huấn luyện.
- II. Gradient Norm (Độ lớn của Gradient): Đường màu xanh dương thể hiện độ lớn của gradient. Mặc dù có những biến động nhất định, nhưng nói chung gradient norm giảm theo thời gian, điều này chỉ ra rằng mô hình đang dần tiến tới một điểm tối ưu. Sự biến động không lớn của gradient norm cũng cho thấy quá trình huấn luyện ổn định và không gặp phải vấn đề về việc gradient quá lớn (gradient explosion) hay quá nhỏ (gradient vanishing).
- III. Learning Rate (Tốc độ học): Đường màu xanh lá cây biểu thị cho learning rate, có vẻ như learning rate giảm dần theo thời gian, điều này phù hợp với các

chiến lược giảm learning rate theo thời gian để giúp mô hình tinh chỉnh các tham số một cách tốt hơn khi tiến gần tới giải pháp tối ưu. Sự giảm này giúp tránh việc mô hình bị "bỏ lỡ" các giải pháp tốt do tốc độ học quá cao.

Đánh giá tổng thể:

- Mô hình đã học tốt từ dữ liệu: điều này được chứng minh bằng việc loss giảm mạnh và ổn định sau một số lượng epoch nhất định.
- Quá trình huấn luyện ổn định: không có dấu hiệu của việc gradient bùng nổ hoặc biến mất, như được thể hiện bằng sự ổn định của độ lớn gradient.
- Tốc độ học được điều chỉnh phù hợp: sự giảm dần của learning rate giúp mô hình tinh chỉnh các tham số hiệu quả và tránh việc bị overfitting.