TRƯỜNG ĐẠI HỌC THỦY LỢI



Báo CÁO BÀI TẬP LỚN MÔN XỬ LÝ NGÔN NGỮ TỰ NHIÊN

Đề tài: Phân tích cảm xúc bình luận

*NGƯỜI HƯỚNG DẪN: PGS.TS.Nguyễn Quang Hoan*

|  |  |
| --- | --- |
| Họ và tên | MSV |
| Lương Tuấn Anh | 2251262571 |

Hà Nội, Năm 2025

MỤC LỤC

[CHƯƠNG 1: GIỚI THIỆU ĐỀ TÀI 3](#_Toc201092611)

[1.1. Tóm tắt đề tài 3](#_Toc201092612)

[1.2. Mục tiêu chính 3](#_Toc201092613)

[1.3. Kết quả dự kiến 3](#_Toc201092614)

[CHƯƠNG 2: PHƯƠNG PHÁP THỰC HIỆN 4](#_Toc201092615)

[2.1. Mô tả dữ liệu 4](#_Toc201092616)

[2.2. Thu thập dữ liệu 4](#_Toc201092617)

[2.3. Gán nhãn 5](#_Toc201092618)

[2.4. Tiền xử lý dữ liệu 6](#_Toc201092619)

[2.5. Trực quan hóa 6](#_Toc201092620)

[CHƯƠNG 3: XÂY DỰNG MÔ HÌNH 8](#_Toc201092621)

[3.1. Mô hình ResNet-50 8](#_Toc201092622)

[3.1.1. Giới thiệu 8](#_Toc201092623)

[3.1.2. Cách hoạt động 8](#_Toc201092624)

[3.1.3. ResNet-50 9](#_Toc201092625)

[3.1.4. Áp dụng mô hình Resnet-50 9](#_Toc201092626)

[3.2. Mô hình EffictientNet-B0 12](#_Toc201092627)

[3.2.1. Giới thiệu 12](#_Toc201092628)

[3.2.2. Cách hoạt động 12](#_Toc201092629)

[3.2.3. Áp dụng EfficientNet-B0 12](#_Toc201092630)

[3.3. Mô hình VGG19 15](#_Toc201092631)

[3.3.1. Khái niệm 15](#_Toc201092632)

[3.3.2. Cách hoạt động 15](#_Toc201092633)

[3.3.3. Áp dụng VGG19 16](#_Toc201092634)

[3.4. So sánh giữa các mô hình 20](#_Toc201092635)

[TÀI LIỆU THAM KHẢO 21](#_Toc201092636)

# 

# DANH MỤC THUẬT NGỮ, KÝ HIỆU

|  |  |  |
| --- | --- | --- |
| **STT** | **Ký hiệu viết tắt** | **Chữ viết đầy đủ** |
| 1 | NLP | Natural Language Processing |
| 2 | TF-IDF | Term Frequency-Inverse Document Frequency |
| 3 | LR | Logistic Regression |
| 4 | NLTK | Natural Language Toolkit |
| 5 | IMDB | Internet Movie Database |
| 6 | AI | Artificial Intelligence |
| 7 | HTML | HyperText Markup Language |

# CHƯƠNG 1: GIỚI THIỆU ĐỀ TÀI

## Tóm tắt đề tài

Đề tài nghiên cứu về phân tích cảm xúc trên tập dữ liệu IMDB bằng phương pháp học máy truyền thống, trong đó sử dụng quy trình xử lý ngôn ngữ tự nhiên để chuẩn hóa dữ liệu review, trích xuất các đặc trưng bằng TF-IDF và áp dụng với mô hình Logistic Regression để phân loại Review thành tích cực hoặc tiêu cực. Kết quả nghiên cứu giúp đánh giá hiệu quả của việc kết hợp NLP cùng với Logistic Regression trong bài toán phân tích cảm xúc, hướng đến việc ứng dụng thực tế trong việc tự động hóa xử lý phản hồi của khách hàng trên các website phim ảnh.

## Lý do chọn đề tài

Việc đánh giá cảm xúc trong review phim đang ngày càng trở nên cần thiết trong thời đại số hiện nay, khi lượng phản hồi của người xem trên internet ngày một tăng. Bộ dữ liệu IMDB được sử dụng rộng rãi trong nghiên cứu NLP nhờ tính đa dạng về ngữ nghĩa và độ lớn phù hợp. Do đó, lựa chọn đề tài ‘Phân tích cảm xúc bình luận’ nhằm tạo ra một công cụ tự động xác định nhanh trạng thái tích cực hay tiêu cực của các reviewer, góp phần nâng cao trải nghiệm của người dùng và hiệu quả trong việc quản lý các nội dung trên nền tảng trực tuyến.

## Mục tiêu và phạm vi nghiên cứu

Mục tiêu chính của nghiên cứu lần này là để xây dựng lên một hệ thống có khả năng tự động nhận diện cảm xúc trong review phim trên tập dữ liệu IMDB. Cụ thể, nghiên cứu hướng tới việc đánh giá hiệu quả của quy trình xử lý NLP kết hợp trích xuất đặc trưng của TF-IDF cùng với thuật toán Logistic Regression trong việc phân tích cảm xúc của các bình luận.

Nghiên cứu được giới hạn trong việc sử dụng tập dữ liệu review phim IMDB, gồm khoảng 50,000 review được gán nhãn ‘positive’ và ‘negative’. Phạm vi bao gồm toàn bộ quy trình tiền xử lý dữ liệu, trích xuất đặc trưng, huấn luyện, đánh giá độ hiệu quả của mô hình và thử nghiệm trên các review mới.

## Phương pháp nghiên cứu và ý nghĩa khoa học, thực tiễn

Nghiên cứu được sử dụng kết hợp phương pháp thực nghiệm trên dữ liệu thực tế cùng với kỹ thuật xử lý ngôn ngữ tự nhiên. Cụ thể, quy trình bao gồm: làm sạch các review thô bằng cách chuẩn hóa, tokenization, loại bỏ stopwords, lemmatization, trích xuất các đặc trưng bằng TF-IDF rồi sử dụng thuật toán Logistic Regression để huấn luyện và đánh giá kết quả trên tập test bằng các thước đo như accuracy, precision, recall, F1-score.

Kết quả nghiên cứu góp phần khẳng định tính hiệu quả của việc kết hợp NLP với Logistic Regression trong bài toán phân tích cảm xúc bình luận. Về mặt khoa học, đề tài này giúp tổng kết kiến thức về xử lý ngôn ngữ tự nhiên trên tập dữ liệu review quy mô lớn. Về mặt thực tiễn, nghiên cứu tạo ra nền tảng giúp tự động hóa việc đánh giá review, phục vụ nhu cầu trong ngành truyền thông, phim ảnh và thương mại điện tử

# CƠ SỞ LÝ THUYẾT

## Tổng quan về xử lý ngôn ngữ tự nhiên (NLP)

### Khái niệm NLP

Xử lý ngôn ngữ tự nhiên (NLP) là một công nghệ máy học, cung cấp cho máy tính khả năng diễn giải, tương tác và hiểu được ngôn ngữ của con người. Các tổ chức ngày nay có khối lượng lớn dữ liệu thoại và văn bản từ nhiều kênh liên lạc khác nhau như email, tin nhắn văn bản, bảng tin trên mạng xã hội, tệp video, tệp âm thanh và nhiều hơn nữa. Họ sử dụng phần mềm NLP để tự động xử lý dữ liệu này, phân tích ý định hoặc cảm xúc trong tin nhắn và phản hồi bằng người thật theo thời gian thực. [1]

NLP là một lĩnh vực của trí tuệ nhân tạo (AI) tập trung vào việc nghiên cứu và phân tích ngôn ngữ tự nhiên, với mục tiêu hiểu và xử lý thông tin trong các ngôn ngữ mà con người sử dụng hàng ngày. NLP sử dụng một số phương pháp và công nghệ để cho phép máy tính hiểu, tạo ra và tương tác với ngôn ngữ tự nhiên một cách tự động. [2]

### Ứng dụng của NLP trong thực tế

NLP đã có những ứng dụng rất đa dạng trong cuộc sống hàng ngày của chúng ta. Dưới đây là một số ứng dụng của NLP:

* **Tự động hóa dịch thuật**: Công nghệ NLP giúp tự động dịch thuật giữa các ngôn ngữ khác nhau, giúp người dùng truy cập thông tin từ trang web, tài liệu hoặc tin nhắn trong ngôn ngữ mà họ không thạo
* **Tự động trả lời email**: NLP có thể được sử dụng để phân tích và trả lời các email theo cách tự động, giúp tiết kiệm thời gian và năng lực làm việc.
* **Phân loại tin tức**: NLP có thể giúp tổ chức và phân loại tin tức từ các nguồn thông tin khác nhau, giúp người dùng có thể dễ dàng theo dõi và tiếp cận thông tin theo quan tâm cá nhân
* **Phân tích cảm xúc xã hội**: NLP có thể được sử dụng để phân tích cảm xúc từ các bài đăng trên mạng xã hội, giúp các công ty và tổ chức hiểu được ý kiến và phản hồi của khách hàng.

## Phân tích cảm xúc

Phân tích cảm xúc là một bài toán trong lĩnh vực xử lý ngôn ngữ tự nhiên nhằm xác định trạng thái cảm xúc hoặc thái độ được diễn đạt trong đoạn văn bản, giúp phân loại ý kiến thành tích cực, tiêu cực hoặc trung lập.

Trong nghiên cứu lần này, bài toán được giới hạn trong việc phân tích cảm xúc nhị phân, xác định rõ review phim thuộc nhóm tích cực (positive) hay tiêu cực (negative). Phân tích cảm xúc có thể thực hiện ở nhiều mức độ khác nhau: ở mức **Từ** để xác định ý nghĩa cảm xúc của từng từ ngữ riêng lẻ, ở mức **Câu** để đánh giá cảm xúc chung của cả câu, và ở mức **Đoạn** để xác định cảm xúc tổng thể của cả một đoạn văn. Việc lựa chọn mức phân tích phù hợp sé giúp cải thiện độ chính xác và ý nghĩa của kết quả dự đoán.

## Tiền xử lý dữ liệu trong NLP

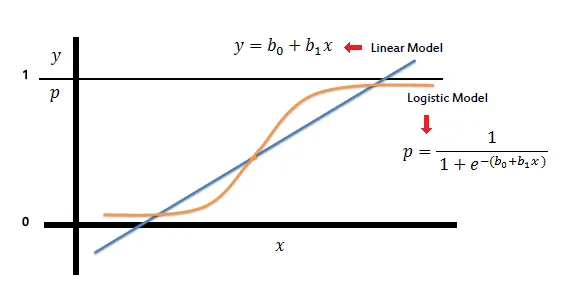
Tiền xử lý dữ liệu trong NLP là một bước quan trọng giúp chuẩn hóa đoạn văn bản thô thành dạng phù hợp để trích xuất đặc trưng và huấn luyện mô hình. Trước hết, toàn bộ đoạn review được chuẩn hóa về chữ thường (**Lowercasing**) nhằm tạo sự nhất quán trong bộ dữ liệu, giúp mô hình không bị nhầm lẫn giữa các từ giống nhau nhưng khác về kiểu chữ. Tiếp đó là loại bỏ các thẻ HTML, dấu câu, các ký tự đặc biệt để giảm thiệu độ nhiễu trong review, chỉ giữ lại các phần có ý nghĩa. Sau đó, đoạn review được tạo thành các token (từ) bằng phương pháp **Tokenization**, giúp biểu diễn đoạn văn dưới dạng dãy từ phục vụ cho việc trích xuất đặc trưng. Các từ dừng (**Stopwords**) ví dụ như: ‘and’, ‘is’, ‘the’, … sẽ được lọc ra nhằm giữ lại những thành phần có đóng góp tích cực trong việc xác định trạng thái của cảm xúc. Cuối cùng, kỹ thuật **Lemmatization** được sử dụng để chuẩn hóa các từ về dạng gốc, giúp giảm thiểu số lượng từ trong từ điển, tăng hiệu quả tổng quát hóa của mô hình trong bài toán phân tích cảm xúc.

## Biến đổi đặc trưng bằng TF-IDF

Sau khi đã hoàn thành bước tiền xử lý ở trên, đoạn review sẽ cần được biểu diễn dưới dạng số để phục vụ cho các thuật toán học máy. Phương pháp TF-IDF được sử dụng nhằm xác định mức độ quan trọng của mỗi từ trong toàn bộ tập dữ liệu. Cách tính toán TF-IDF kết hợp tần suất xuất hiện của từ đó trong tài liệu (TF) cùng mức độ hiếm của từ trong toàn bộ tập tài liệu (IDF). Kỹ thuật này giúp nhấn mạnh các từ có ý nghĩa đặc trưng trong đoạn review, đồng thời hạn chế tác động của các từ phổ biến nhưng lại ít ý nghĩa, từ đó sẽ giúp cải thiện đáng kể hiệu quả trong bài toán phân tích cảm xúc

## Thuật toán Logistic Regression

Logistic Regression là một mô hình thống kê được sử dụng để phân loại nhị phân, tức dự đoán một đối tượng thuộc vào một trong hai nhóm. Hồi quy Logistics làm việc dựa trên nguyên tắc của hàm sigmoid – một hàm phi tuyến tự chuyển đầu vào của nó thành một xác suất thuộc về một trong hai lớp nhị phân.



Logistic Regression sử dụng hàm phi tuyến để xác định xác suất của hai lớp 0 và 1

Hàm sigmoid:

Hàm mất mát Cross-Entropy:

Trong đó:

* n: số lượng mẫu dữ liệu trong tập huấn luyện.
* : giá trị thực tế của đầu ra thứ i.
* : xác suất dự đoán thuộc lớp 1 của mô hình cho đầu vào thứ i.

Hàm Cross-Entropy đo lường khoảng cách giữa hai phân phối xác suất và . Khi mô hình dự đoán chính xác, tức là nếu thì càng gần 1 và nếu thì càng gần 0, sau đó hàm mất mát sẽ tiến gần về 0.

Trong quá trình huấn luyện, chúng ta sẽ tìm cách cập nhật bộ trọng số ‘w’ sao cho giá trị hàm mất mát Cross-Entropy đạt giá trị nhỏ nhất, dẫn đến một mô hình dự báo tốt nhất. [3]

## Các phương pháp đánh giá kết quả

Để đánh giá hiệu quả của mô hình trong bài toán phân tích cảm xúc bình luận, có nhiều thước đo được sử dụng. Accuracy là tỷ lệ phần trăm số dự đoán đúng trên toàn bộ tập dữ liệu. Precision đo tỷ lệ dự đoán đúng trong số tất cả các kết quả được dự báo là tích cực (positive). Recall xác định tỷ lệ phát hiện đúng trong tổng số kết quả thực tế tích cực. F1-score là thước đo kết hợp giữa Precision và Recall, giúp đánh giá toàn diện hiệu suất mô hình. Ngoài ra, Confusion Matrix được sử dụng để tổng kết số lượng dự đoán đúng, sai trên cả hai nhãn, giúp đánh giá chi tiết mức độ nhầm lẫn của mô hình trên từng lớp.

# BỘ DỮ LIỆU & TIỀN XỬ LÝ

## Mô tả bộ dữ liệu IMDB

### Nguồn gốc bộ dữ liệu

BẮT ĐẦU VIẾT

### Cách hoạt động

Cấu trúc của ResNet được xây dựng dựa trên việc chia nhỏ mạng thành các khối residual (Residual Blocks) trong đó mỗi khối bao gồm một số lớp tích chập cùng với một kết nối tắt trực tiếp từ đầu vào đến đầu ra của khối. Ý tưởng chính là việc học một hàm residual F (x) thay vì học trực tiếp một hàm mục tiêu H (x). Điều này giúp duy trì tính ổn định và độ chính xác khi mạng có nhiều tầng.

Nguyên lý hoạt động của ResNet dựa trên việc thay vì học trực tiếp hàm H (x), mô hình sẽ học 1 hàm residual F (x) = H (x) – x. Sau đó đầu ra của mô hình sẽ là H (x) = F(x) + x.

x

Weight layer

F (x) relu x indentity

Weight layer

F (x) + x

### ResNet-50

ResNet-50 bao gồm 50 tầng với cấu trúc cụ thể:

Conv1: Tầng tích chập đầu tiên với kernel size 7x7, stride 2 và 64 filters

MaxPool: Lớp MaxPooling với kernel sixe 3x3 và stride 2

Conv2\_x: Gồm 3 khối residual với 64 filters mỗi khối

Conv3\_x: Gồm 4 khối residual với 128 filters mỗi khối

Conv4\_x: Gồm 6 khối residual với 256 filters mỗi khối

Conv5\_x: Gồm 3 khối residual với 512 filters mỗi khối

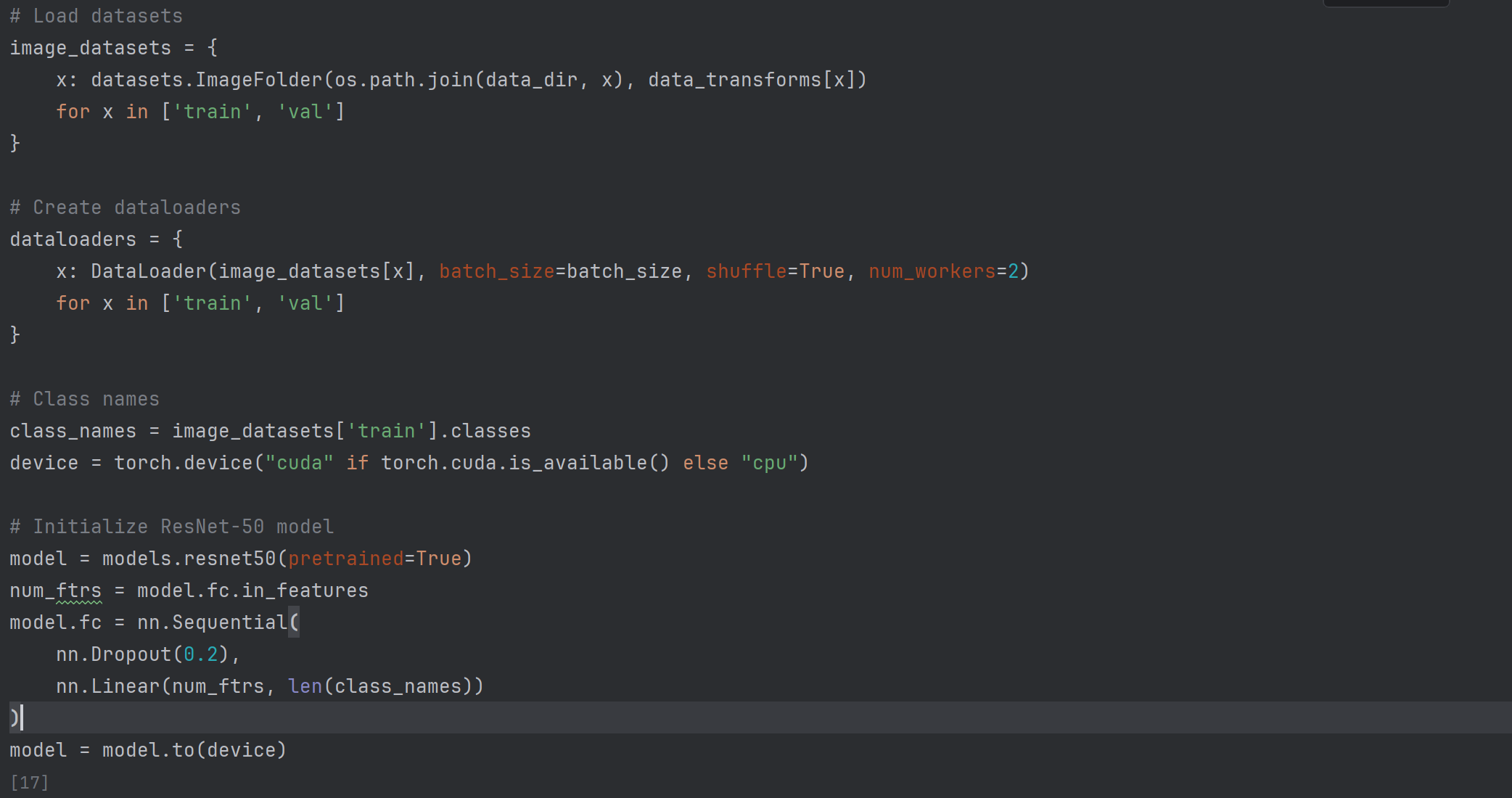
Average Pooling: lớp pooling trung bình

Fully Connected Layer: Lớp kết nối đầy đủ với số lượng neurons bằng số lớp đầu ra

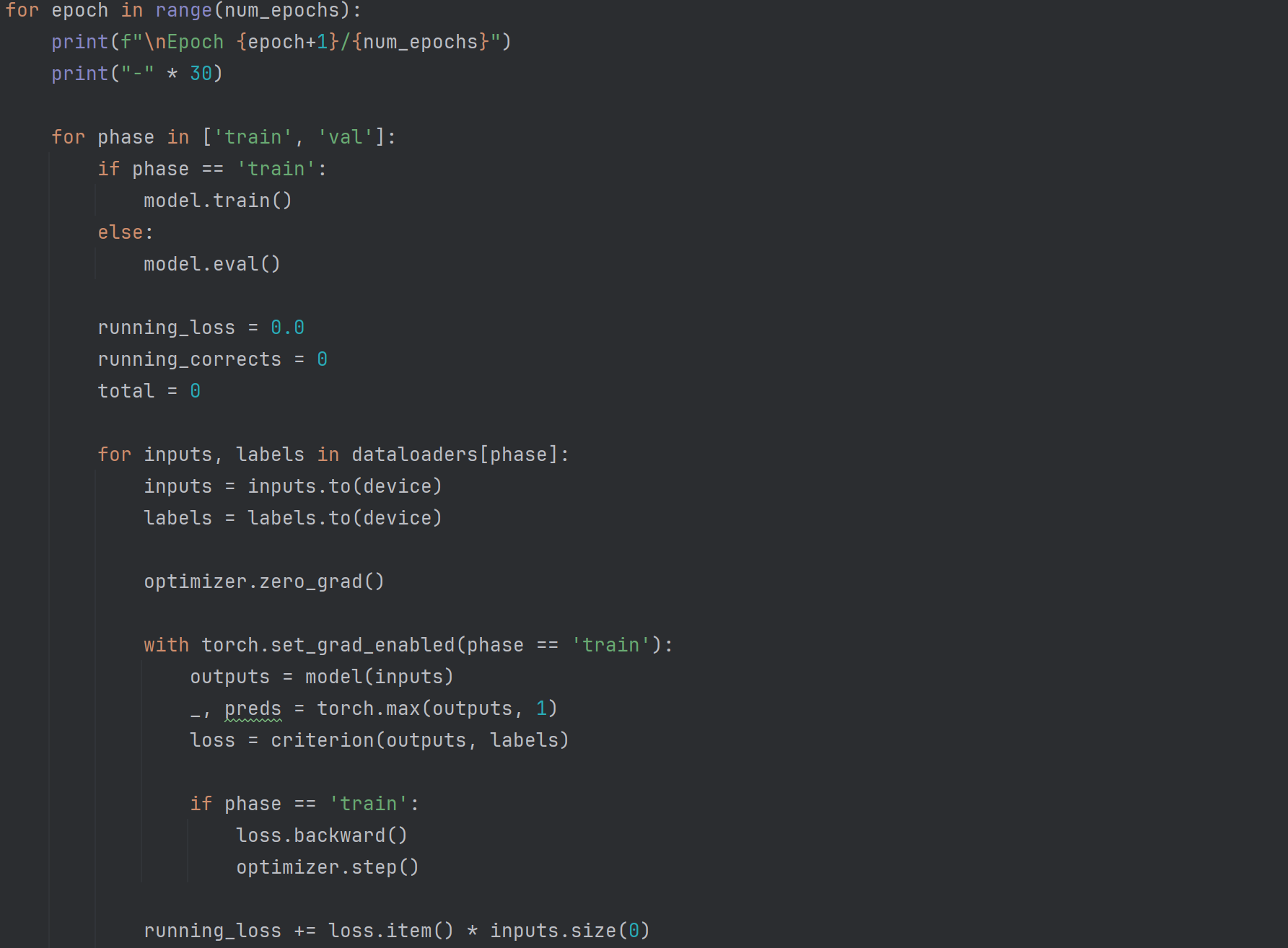
### Áp dụng mô hình Resnet-50

Load mô hình ResNet-50 với trọng số pretrained

Thay thế tầng FC cho phù hợp số lớp của bạn



Huấn luyện và đánh giá mô hình bằng cách lặp qua các epoch



Các độ đo đánh giá của mô hình:

Độ chính xác 93,52%

A screenshot of a computer screen

AI-generated content may be incorrect.

Ma trận nhầm lẫn

A graph of a number of activities

AI-generated content may be incorrect.

* Mô hình hoạt động tốt, đa phần các giá trị tập trung ở đường chéo chính
* Các nhầm lẫn nhỏ có thể xuất phát từ các việc hành vi giống nhau về hình ảnh hoặc ánh sáng, góc nhìn chưa tốt
* Cần cải thiện lớp other\_activities, vì lớp này dễ nhầm lẫn và có precison thấp nhất

## Mô hình EffictientNet-B0

### Giới thiệu

EfficientNet-B0 là một kiến trúc mạng nơ ron tích chập (CNN) hiện đại được thiết kế bởi Google AI, nổi bật vì đạt được hiệu suất cao với số lượng tham số và tính toán thấp hơn nhiều mô hình CNN truyền thống. [4]

### Cách hoạt động

Bước 1: Nhận đầu vào hình ảnh: Kích thước chuẩn 224x224x3 (ảnh màu RGB)

Bước 2: Trích xuất đặc trưng với các khối MBConv

Mỗi khối MBConv gồm: Depthwise Convolution; Pointwise Convolution, Skip connection (nối tắt); Squeeze and Excitation (SE) giúp học được mối liên hệ giữa các kênh

Bước 3: Phóng to mô hình theo 3 hướng:

Chiều sâu (depth): số lớp

Chiều rộng (widtg): số kênh

Độ phân giải (resolution): kích thước ảnh

Bước 4: Phân loại đầu ra: Sau khi trích xuất đặc trưng các đặc trưng được đưa vào lớp fully connected để phân loại

### Áp dụng EfficientNet-B0

Load mô hình efficientNet-B0

A computer screen shot of a black screen

AI-generated content may be incorrect.

Huấn luyện và đánh giá mô hình

A screenshot of a computer program

AI-generated content may be incorrect.

Các chỉ số đo của mô hình:

Độ chính xác: 96,41 %

Macro Average F1-score = 0.96: Trung bình đều nhau giữa các lớp mô hình không bị thiên lệch về 1 lớp cụ thể nào

A screenshot of a computer screen

AI-generated content may be incorrect.

Ma trận nhầm lẫn :

Hầu hết các lớp dự đoán chính xác cao

Có 1 số nhầm lẫn nhẹ : lớp texting\_phone bị nhầm trong other\_activities (5 mẫu); safe\_driving (4 mẫu); lớp other\_activities có thể chứa nhiều hành vi không dễ dàng, dễ bị chồng lấn các hành vi khác.

A diagram of a graph

AI-generated content may be incorrect.

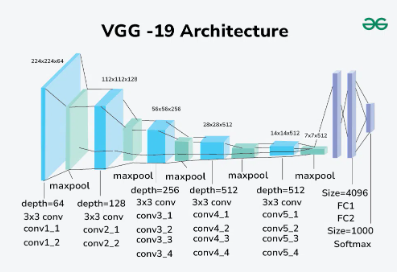
* Mô hình cho thấy khả năng phân loại hiệu quả, tuy nhiên cần được cải thiện thêm ở những lớp dễ nhầm lẫn bằng cách: tăng cường dữ liệu cho lớp này, áp dụng các kỹ thuật tiền xử lý.

## Mô hình VGG19

### Khái niệm

Các mô hình Visual Geometry Group (VGG), đặc biệt là VGG-16 và VGG-19, đã có ảnh hưởng đáng kể đến lĩnh vực thị giác máy tính kể từ khi chúng ra đời. Các mô hình này, được Visual Geometry Group từ Đại học Oxford giới thiệu, nổi bật trong Thử thách nhận dạng hình ảnh quy mô lớn ImageNet (ILSVRC) năm 2014 vì các mạng nơ-ron tích chập sâu (CNN) với kiến ​​trúc đồng nhất. VGG-19, biến thể sâu hơn của các mô hình VGG, đã thu hút được sự chú ý đáng kể do tính đơn giản và hiệu quả của nó. [4]

### ***Cách hoạt động***



VGG19 là một mạng nơron tích chập với 19 lớp trọng số, bao gồm 16 convolutional layers và 3 fully connected layers. Kiến trúc tuân theo một mô hình đơn giản và lặp lại, giúp dễ hiểu và triển khai hơn.

Các thành phần chính của kiến ​​trúc VGG-19 là:

Convolutional Layers: Bộ lọc 3x3 với bước nhảy là 1 và đệm là 1 để bảo toàn độ phân giải không gian.

Activation Function: ReLU (Đơn vị tuyến tính chỉnh lưu) được áp dụng sau mỗi lớp tích chập để đưa vào tính phi tuyến tính.

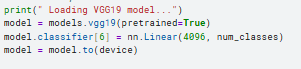
Pooling Layers: Gộp tối đa với bộ lọc 2x2 và bước nhảy là 2 để giảm kích thước không gian.

Fully Connected Layers: Ba lớp được kết nối đầy đủ ở cuối mạng để phân loại.

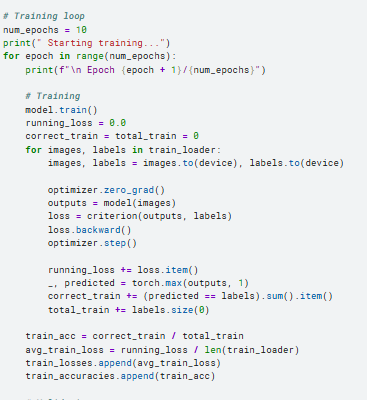
Softmax Layer: Lớp cuối cùng để đưa ra xác suất lớp

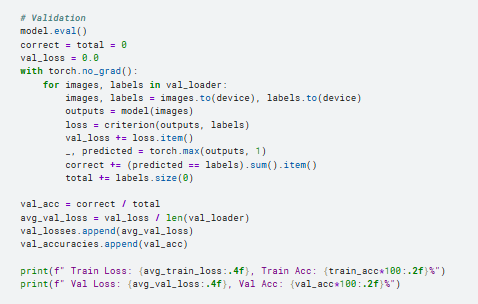
### Áp dụng VGG19

Load mô hình

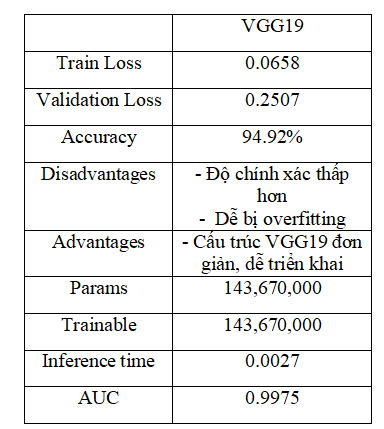


Huấn luyện và đánh giá mô hình

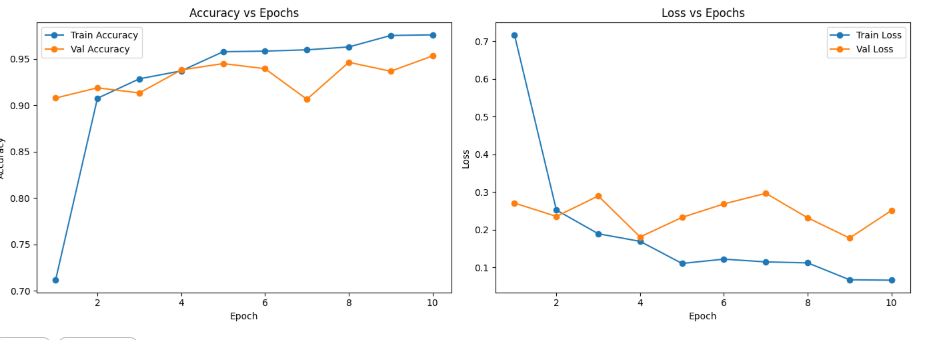




Các chỉ số đo của mô hình:



Train Loss và Validation Loss:



Train Accuracy tăng đều từ khoảng 0.75 đến 9.75 khá là ổn định

Validation Accuracy cũng tăng từ 0.89 đến 0.94 nhưng có dao động nhẹ sau epoch thứ 5

Từ đây có thể suy ra được mô hình học tốt trên cả tập Train và tập Validation, Validation Accuracy cao và sát Train Accuracy cho thấy mô hình không bị ovefitting nghiêm trọng.

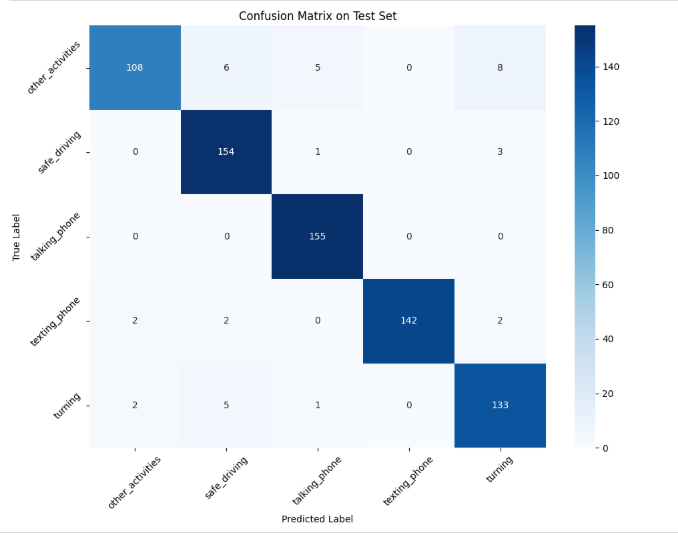
Train Loss giảm nhanh từ 0.7 đến 0.1 rất ổn định

Validation Loss giảm ban đầu nhưng giao động lên xuống từ epoch 3 đến epoch 9

Từ epoch 4 trở đi, Validation Loss giao động trong khi Train Loss vẫn giảm suy mô hình bắt đầu bị overfitting

Tuy vậy, Validation Accuracy vẫn cao cho thấy mức độ overfitting không quá nghiêm trọng

Ma trận nhầm lẫn:



Mô hình phân loại nhìn chung hoạt động **rất tốt** đặc biệt với các lớp như :

other\_activities : 108 ảnh được phân lớp đúng trên tổng số 127 ảnh

Nhìn chung mô hình hoạt động tốt đặc biệt với các nhãn talking\_phone, safe\_driving và texting\_phone

Một số ảnh còn bị phân lớp sai ở 2 lớp other\_activities và turning, điều này có thể do hành vi thiếu phân biệt hoặc bị chống lẫn bởi đặc trưng hành vi

## So sánh giữa các mô hình

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Categories | EfficientNet-B0 | | ResNet-50 | | VGG19 | |
| Correctly classified images | Misclassified images | Correctly classified images | Misclassified images | Correctly classified images | Misclassified images |
| other\_activities | 111 | 7 | 115 | 3 | 108 | 19 |
| safe\_driving | 161 | 6 | 155 | 12 | 154 | 4 |
| talking\_phone | 150 | 1 | 145 | 6 | 155 | 0 |
| texting\_phone | 147 | 9 | 143 | 13 | 142 | 6 |
| turning | 130 | 3 | 120 | 13 | 133 | 8 |

* Nên chọn mô hình EfficientNet-B0 nhờ hiệu suất toàn diện và ổn định hơn giữa các lớp; F1-score đồng đều hơn giữa các lớp và mức độ nhầm lẫn thấp.