

ĐẠI HỌC QUỐC GIA TP. HCM  
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN



**BÁO CÁO TỔNG KẾT**  
**ĐỀ TÀI KHOA HỌC VÀ CÔNG NGHỆ SINH VIÊN NĂM 2023.**

**Tên đề tài tiếng Việt:** PHÂN LOẠI VIDEO ĐỘC HẠI TIẾNG VIỆT

**Tên đề tài tiếng Anh:** CLASSIFICATION OF VIETNAMESE HARMFUL VIDEOS

**Khoa/ Bộ môn:** Khoa Khoa học và Kỹ thuật Thông tin

**Thời gian thực hiện:** 01/2024 - 06/2024

**Cán bộ hướng dẫn:** TS. Đỗ Trọng Hợp

**Tham gia thực hiện:**

TT	Họ và tên, MSSV	Chịu trách nhiệm	Điện thoại	Email
1.	Nguyễn Ngọc Hà My	Chủ nhiệm	0828924893	20521623@gm.uit.edu.vn
2.	Nguyễn Thanh Thanh Trúc	Tham gia	0949454721	20520829@gm.uit.edu.vn

**Thành phố Hồ Chí Minh – Tháng 6 /2024**

# THÔNG TIN KẾT QUẢ NGHIÊN CỨU

## 1. Thông tin chung:

- Tên đề tài: Phân loại video độc hại tiếng Việt
- Chủ nhiệm: Nguyễn Ngọc Hà My
- Thành viên tham gia: Nguyễn Thanh Thanh Trúc 20520829 KH&KTTT
- Cơ quan chủ trì: Trường Đại học Công nghệ Thông tin.
- Thời gian thực hiện: 01/2024 – 06/2024

## 2. Mục tiêu:

Trong phạm vi nghiên cứu, mục tiêu chính là:

- Xây dựng một bộ dữ liệu phục vụ nghiên cứu về tác vụ phân loại video độc hại đối với tiếng Việt.
- Đề xuất và thử nghiệm các phương pháp trích xuất đặc trưng đối với dữ liệu là audio tiếng Việt.
- Xây dựng các mô hình Deep Learning với các kiến trúc 2D CNN, 3D CNN, và các mô hình ngôn ngữ đối với tác vụ phân loại video độc hại tiếng Việt.
- Phân tích ảnh hưởng của các phương pháp trích xuất đặc trưng dựa trên kết quả dự đoán các mô hình Deep Learning.
- Đánh giá hiệu suất của các mô hình dự đoán.

## 3. Tính mới và sáng tạo:

- Hiện nay vẫn chưa có bộ dữ liệu nào về phân loại video có sử dụng yếu tố ngôn ngữ là tiếng Việt cũng như không có một nghiên cứu nào thực hiện tác vụ phân loại video độc hại tiếng Việt.
- Cùng với sự phát triển mạnh mẽ của các nền tảng mạng xã hội sử dụng các video ngắn làm yếu tố thu hút người dùng như TikTok, Youtube Shorts, Facebook Reels, ... các video độc hại có thể dễ dàng tràn lan trên mạng mà không có sự kiểm duyệt về nội dung. Điều này có thể giúp các đối tượng xấu lan truyền các video mại dâm, xu hướng bạo lực, phản động, tin giả hoặc các video có nội dung gây ám ảnh,... cho người dùng,

đặc biệt là giới trẻ. Vì vậy việc quản lý các video có thể được đăng tải trở thành một nghiệm vụ rất quan trọng.

#### **4. Tóm tắt kết quả nghiên cứu:**

- Báo cáo về các phương pháp trích xuất đặc trưng đối với tác vụ phân loại video độc hại tiếng Việt.
- Báo cáo về bộ dữ liệu và kết quả chạy thực nghiệm các mô hình Deep Learning đối với từng đặc trưng.
- Kết luận về ảnh hưởng của các đặc trưng và hiệu suất các mô hình được sử dụng.

#### **5. Tên sản phẩm: HarmfulVideosVN2023**

#### **6. Hiệu quả, phương thức chuyển giao kết quả nghiên cứu và khả năng áp dụng:**

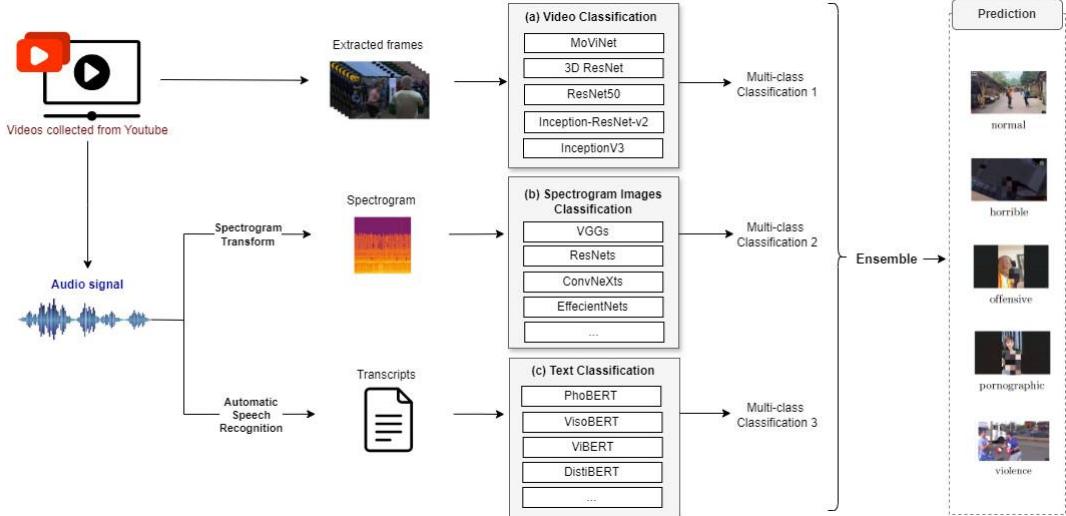
##### **a) Hiệu quả:**

- Kết quả thu được bộ dữ liệu HarmfulVideosVN2023 có thể sử dụng để nghiên cứu hiệu quả các phương pháp trích xuất đặc trưng và hiệu suất của các mô hình Deep Learning.
- Qua các thí nghiệm cho ra được các mô hình có hiệu suất cao và đáng tin cậy trong việc phân loại video độc hại tiếng Việt

##### **b) Phương thức chuyển giao kết quả nghiên cứu và khả năng áp dụng:**

- Các kết luận về ảnh hưởng của các đặc trưng đối với tác vụ phân loại video độc hại tiếng Việt có thể sử dụng để tham khảo khi lựa chọn đặc trưng đưa vào mô hình dự đoán video.
- Các nhận xét về hiệu suất các mô hình có thể sử dụng để định hướng nghiên cứu phát triển mô hình dự đoán video tiếp theo.

#### **7. Hình ảnh, sơ đồ minh họa chính:**



## 8. Tài liệu tham khảo

- [1]. Tran, Du, et al. "A closer look at spatiotemporal convolutions for action recognition." *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 2018.
- [2]. Karpathy, Andrej, et al. "Large-scale video classification with convolutional neural networks." *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 2014.
- [3]. Kondratyuk, Dan, et al. "Movinets: Mobile video networks for efficient video recognition." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021.
- [4]. Rehman, Atiq, and Samir Brahim Belhaouari. "Deep learning for video classification: A review." (2021).
- [5]. Zhang, Zhongping, et al. "Movie genre classification by language augmentation and shot sampling." *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2024.
- [6]. Solovyev, Roman, Alexandr A. Kalinin, and Tatiana Gabruseva. "3D convolutional neural networks for stalled brain capillary detection." *Computers in biology and medicine* 141 (2022): 105089.
- [7]. Kondratyuk, Dan, et al. "Movinets: Mobile video networks for efficient video recognition." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021.

## LỜI CẢM ƠN

Để hoàn thành đề tài nghiên cứu khoa học này, ngoài nỗ lực của bản thân, không thể không kể đến các cộng sự và các quý thầy cô của Trường Đại học Công nghệ thông tin ĐHQG TPHCM.

Đầu tiên tôi xin được cảm ơn TS. Đỗ Trọng Hợp, người đã đồng hành cùng chúng tôi ngay từ những ngày đầu tiên, nếu không có sự hỗ trợ của thầy thì chúng tôi đã không thể hoàn thành đề tài này. Thầy đã tận tình chỉ dẫn, tin tưởng và luôn hỗ trợ chúng tôi trong suốt quá trình thực hiện đề tài.

Bên cạnh đó, tôi cũng xin gửi lời cảm ơn bạn Thanh Trúc –người đã luôn tin tưởng và hỗ trợ hết mình cho tôi trong suốt dự án để tôi có thể có được kết quả nghiên cứu ngày hôm nay.

Chúng tôi cũng xin gửi lời tri ân sâu sắc đến cán bộ, giảng viên Trường Đại học Công nghệ Thông tin đã tận tình chỉ bảo, truyền đạt tri thức để chúng tôi có thể có những hành trang vững chắc cho mình trong quá trình nghiên cứu, và cũng tạo ra một môi trường giúp đưa sinh viên đến gần với nghiên cứu khoa học như vậy.

Trong quá trình thực hiện nghiên cứu chúng tôi đã gặp không ít khó khăn về vấn đề kỹ thuật, tài nguyên và tri thức, mặc dù đã cố gắng trong quá trình tìm hiểu, nghiên cứu và thực nghiệm để có thể đưa ra được những kết quả đáng khích lệ nhưng do kiến thức, nguồn lực và kinh nghiệm còn nhiều hạn chế nên không thể tránh khỏi được những thiếu sót. Chúng tôi hy vọng nhận được sự góp ý của thầy cô để chỉnh sửa, bổ sung.

Chân thành cảm ơn.  
Nguyễn Ngọc Hà My.

# MỤC LỤC

CHƯƠNG 1: TỔNG QUAN ĐỀ TÀI .....	12
1.1 Video độc hại .....	13
1.1.1 Định nghĩa video độc hại .....	13
1.1.2 Tình trạng video độc hại hiện nay .....	13
1.2 Phát biểu bài toán .....	14
1.3 Các thách thức .....	14
1.4 Các vấn đề cần giải quyết .....	16
CHƯƠNG 2: CÁC CÔNG TRÌNH NGHIÊN CỨU LIÊN QUAN .....	16
2.1 Các bộ dữ liệu: .....	17
2.2 Các phương pháp biểu diễn đặc trưng của video .....	18
2.3 Các mô hình phân loại video .....	19
2.4 Các phương pháp nghiên cứu xử lý âm thanh .....	21
CHƯƠNG 3: NGHIÊN CỨU THỰC NGHIỆM VÀ LÝ THUYẾT .....	22
3.1 Bộ dữ liệu .....	22
3.1.1 Thu thập dữ liệu và gán nhãn .....	23
3.1.2 Trích xuất đặc trưng .....	26
3.1.2.1 Trích xuất khung hình ảnh .....	27
3.1.2.2 Trích xuất văn bản từ âm thanh của video .....	27
3.1.2.3 Trích xuất spectrogram của âm thanh video .....	29
3.2 Các mô hình phân loại .....	30
3.2.1 Mô hình phân loại video .....	31
3.2.1.1 3D CNN với ResNet .....	32
3.2.1.2 3D CNN với ResNet50, Inception-ResNet-v2 và InceptionV3 đã được đào tạo trước .....	33
3.2.1.3 Học chuyển tiếp với MoViNet .....	34
3.2.2 Mô hình phân loại văn bản .....	34
3.2.2.1 BiLSTM (Bidirectional Long Short-Term Memory) .....	34
3.2.2.2 BERT (Bidirectional Encoder Representations from Transformers) ..	35
3.2.2.3 PhoBERT .....	36
3.2.2.4 RoBERTa (Robustly optimized BERT approach) .....	36
3.2.2.5 XLM-RoBERTa .....	37
3.2.2.6 CafeBERT .....	37
3.2.2.7 ViSoBERT .....	38
3.2.2.8 viBERT .....	38
3.2.2.9 DistilBERT .....	39
3.2.2.10 vELECTRA .....	39
3.2.3 Mô hình phân loại ảnh spectrogram .....	39
3.2.3.1 VGG .....	40
3.2.3.2 DenseNet .....	41
3.2.3.3 ResNet .....	42
3.2.3.4 Inception .....	44
3.2.3.5 Xception .....	45
3.2.3.6 NASNetMobile .....	46
3.2.3.7 ConvNeXt .....	47
3.2.3.8 EfficientNet .....	48
3.2.4 Kết hợp các mô hình sử dụng phương pháp ensemble .....	49
CHƯƠNG 4: THỰC NGHIỆM VÀ ĐÁNH GIÁ .....	51

4.1. Thiết kế thí nghiệm .....	51
4.2. Kết quả .....	53
4.2.2 Kết quả của các mô hình dự đoán với các ảnh spectrogram trích xuất từ âm thanh của video .....	59
4.2.3 Kết quả các mô hình dự đoán đối với văn bản trích xuất từ âm thanh video .....	78
4.2.4 Kết quả sau khi kết hợp các kết quả dự đoán từ các đặc trưng .....	78
CHƯƠNG 5: KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN .....	79
TÀI LIỆU THAM KHẢO .....	80

## DANH MỤC HÌNH

Hình 1 Minh họa bài toán phân loại video độc hại Tiếng Việt.....	1
Hình 2 Video minh họa các nhãn của bộ dữ liệu.....	1
Hình 3 Phân bố nhãn trong bộ dữ liệu HarmfulVideosVN2023 .....	1
Hình 4 Minh họa quy trình xử lý dữ liệu và các mô hình phân loại .....	1
Hình 5 Các khung hình ảnh từ video "bình thường" .....	1
Hình 6 Các khung hình ảnh từ video "kinh dị" .....	1
Hình 7 Các khung hình ảnh từ video "bạo lực" .....	1
Hình 8 Minh họa quá trình trích xuất văn bản từ âm thanh của video .....	1
Hình 9 Dữ liệu văn bản trích xuất từ âm thanh của các video. ....	1
Hình 10 Ânh spectrogram trích xuất từ âm thanh của một video .....	1
Hình 11 Minh họa quá trình trích xuất spectrogram từ âm thanh của video .....	1
Hình 12 Cấu trúc mạng (2+1)D tích chập .....	1
Hình 13 Kiến trúc của BiLSTM [7] .....	1
Hình 14 Quy trình phân loại ảnh Spectrogram .....	1
Hình 15 Kiến trúc DenseNet121 [8] .....	1
Hình 16 Kiến trúc ResNet [3] .....	1
Hình 17 Minh họa kiến trúc InceptionV3 .....	1
Hình 18 Ma trận nhầm lẫn của mô hình ResNet đối với tập train .....	1
Hình 19 Ma trận nhầm lẫn của mô hình pre-trained ResNet50 đối với tập train .....	1
Hình 20 Ma trận nhầm lẫn của mô hình ResNet đối với tập test .....	1
Hình 21 Ma trận nhầm lẫn của mô hình pre-trained ResNet50 đối với tập test .....	1
Hình 22 Ma trận nhầm lẫn của mô hình pre-trained Inception-ResNet-v2 đối với tập train .....	1
Hình 23 Ma trận nhầm lẫn của mô hình pre-trained Inception-ResNet-v2 đối với tập test .....	1
Hình 24 Ma trận nhầm lẫn của mô hình pre-trained InceptionV3 đối với tập train .....	1
Hình 25 Ma trận nhầm lẫn của mô hình pre-trained InceptionV3 đối với tập test .....	1
Hình 26 Đồ thị Loss và Accuracy của mô hình ResNet .....	1
Hình 27 Đồ thị Loss và Accuracy của mô hình pre-trained ResNet50 .....	1
Hình 28 Đồ thị Loss và Accuracy của mô hình pre-trained Inception-ResNet-v2 .....	1
Hình 29 Đồ thị Loss và Accuracy của mô hình pre-trained InceptionV3 .....	1
Hình 30 Ma trận nhầm lẫn của mô hình VGG16 .....	1
Hình 31 Ma trận nhầm lẫn của mô hình DenseNet121 .....	1
Hình 32 Ma trận nhầm lẫn của mô hình ResNet50 .....	1
Hình 33 Ma trận nhầm lẫn của mô hình DenseNet169 .....	1
Hình 34 Ma trận nhầm lẫn của mô hình DenseNet201 .....	1
Hình 35 Ma trận nhầm lẫn của mô hình EfficientNetB7 .....	1
Hình 36 Ma trận nhầm lẫn của mô hình NASNetMobile .....	1
Hình 37 Ma trận nhầm lẫn của mô hình InceptionResNetV2 .....	1
Hình 38 Ma trận nhầm lẫn của mô hình ConvNeXtTiny .....	1
Hình 39 Ma trận nhầm lẫn của mô hình EfficientNetB3 .....	1
Hình 40 Ma trận nhầm lẫn của mô hình EfficientNetB0 .....	1
Hình 41 Ma trận nhầm lẫn của mô hình EfficientNetB1 .....	1
Hình 42 Ma trận nhầm lẫn của mô hình ConvNeXtSmall .....	1
Hình 43 Ma trận nhầm lẫn của mô .....	1
Hình 44 Ma trận nhầm lẫn của mô hình InceptionVGG19 .....	1

Hình 45 Ma trận nhầm lẫn của mô hình Xception .....	1
Hình 46 Ma trận nhầm lẫn của mô hình ResNet50V2 .....	1
Hình 47 Ma trận nhầm lẫn của mô hình ConvNeXtBase .....	1
Hình 48 Ma trận nhầm lẫn của mô hình EfficientNetV2M .....	1
Hình 49 Ma trận nhầm lẫn của mô hình EfficientNetV2L .....	1
Hình 50 Ma trận nhầm lẫn của mô hình EfficientNetB2 .....	1
Hình 51 Ma trận nhầm lẫn của mô hình EfficientNetB4 .....	1
Hình 52 Ma trận nhầm lẫn của mô hình EfficientNetB7 .....	1
Hình 53 Đồ thị học với Accuracy và Loss của mô hình ResNet50 .....	1
Hình 54 Đồ thị học với Accuracy và Loss của mô hình ConvNeXtSmall .....	1
Hình 55 Đồ thị học với Accuracy và Loss của mô hình EfficientNetB0 .....	1
Hình 56 Đồ thị học với Accuracy và Loss của mô hình EfficientNetB1 .....	1
Hình 57 Đồ thị học với Accuracy và Loss của mô hình EfficientNetB2 .....	1
Hình 58 Đồ thị học với Accuracy và Loss của mô hình EfficientNetB3 .....	1
Hình 59 Đồ thị học với Accuracy và Loss của mô hình EfficientNetB4 .....	1
Hình 60 Đồ thị học với Accuracy và Loss của mô hình EfficientNetB7 .....	1
Hình 61 Đồ thị học với Accuracy và Loss của mô hình EfficientNetV2M .....	1
Hình 62 Đồ thị học với Accuracy và Loss của mô hình EfficientNetV2L .....	1
Hình 63 Đồ thị học với Accuracy và Loss của mô hình ConvNeXtTiny .....	1
Hình 64 Đồ thị học với Accuracy và Loss của mô hình ConvNeXtBase .....	1
Hình 65 Đồ thị học với Accuracy và Loss của mô hình InceptionResNetV2 .....	1
Hình 66 Đồ thị học với Accuracy và Loss của mô hình ResNet50V2 .....	1
Hình 67 Đồ thị học với Accuracy và Loss của mô hình NASNetMobile .....	1
Hình 68 Đồ thị học với Accuracy và Loss của mô hình DenseNet201 .....	1
Hình 69 Đồ thị học với Accuracy và Loss của mô hình DenseNet169 .....	1
Hình 70 Đồ thị học với Accuracy và Loss của mô hình VGG16 .....	1
Hình 71 Đồ thị học với Accuracy và Loss của mô hình DenseNet121 .....	1
Hình 72 Đồ thị học với Accuracy và Loss của mô hình InceptionV3 .....	1
Hình 73 Đồ thị học với Accuracy và Loss của mô hình VGG19.....	1
Hình 74 Đồ thị học với Accuracy và Loss của mô hình Xception. ....	1

## **DANH MỤC BẢNG**

Bảng 1 Thống kê chi tiết số lượng video trong bộ dữ liệu .....	1
Bảng 2 Thống kê số lượng văn bản trích xuất được .....	1
Bảng 3 Kết quả các mô hình phân loại video .....	1
Bảng 4 Kết quả các mô hình phân loại video dựa trên phô âm thanh .....	1
Bảng 5 Bảng kết quả các mô hình phân loại văn bản .....	1
Bảng 6 Kết quả sau khi kết hợp các kết quả dự đoán từ các đặc trưng .....	1

## **DANH MỤC TỪ VIẾT TẮT**

<b>STT</b>	<b>Thuật ngữ</b>	<b>Mô tả</b>
1	STT	Speech-to-text
2	ASR	Automatic Speech Recognition
3	CNN	Convolutional Neural Network
4	FSTCN	Factorized Spatial-Temporal Convolutional Network
5	RNN	Recurrent Neural Network
6	LSTM	Long-Short Term Memory

## **Chương 1: TỔNG QUAN ĐỀ TÀI**

*Nội dung chương này bao gồm định nghĩa video độc hại, tình trạng video độc hại hiện nay, phát biểu bào toàn, các thách thức còn tồn tại và chỉ ra những vấn đề mà đề tài cần tập trung, nghiên cứu giải quyết.*

### **1.1 Video độc hại**

#### **1.1.1 Định nghĩa video độc hại**

“Video độc hại” là thuật ngữ thường được dùng để mô tả những video chứa nội dung gây hại đến người xem, đặc biệt là trẻ em và thanh thiếu niên. Nội dung độc hại trong video có thể bao gồm những yếu tố bạo lực, kích động, khiêu dâm, xúc phạm, tự tử, kinh dị... Các video độc hại thường chứa những cảnh bạo lực đánh nhau, cảnh khiêu dâm đòi trụy không phù hợp với một số lượng lớn người xem, cảnh tự tử hay tự gây thương tích làm ảnh hưởng tiêu cực đến tâm lý của người xem, lời nói hoặc hành động xúc phạm quốc tịch, tôn giáo, giới tính của cá nhân khác, hay dùng lời nói thiếu văn minh để chửi nhau.

Nội dung độc hại trong video có thể gây ra nhiều hậu quả tiêu cực đến xã hội. Video có nội dung bạo lực, tự tử, khiêu dâm, kích động có thể gây ảnh hưởng mạnh như gây căng thẳng, lo lắng, hoặc gây tổn thương tâm lý đến trạng thái tâm lý và sức khỏe tinh thần của người xem, đặc biệt là trẻ em và thanh thiếu niên. Ngoài ra, video độc hại chứa nội dung bạo lực, kích động căm ghét hoặc khuyến khích hành vi tiêu cực có thể làm tăng cường những hành vi tương tự trong xã hội. Video độc hại có thể lan truyền thông tin không chính xác, tin tức giả mạo hoặc thông điệp sai lệch, gây hiểu lầm, xuyên tạc thông tin và đánh mất niềm tin vào các nguồn tin tức đáng tin cậy.

#### **1.1.2 Tình trạng video độc hại hiện nay**

Hiện nay, tình trạng video độc hại vẫn là một vấn đề nghiêm trọng trên Internet và các nền tảng trực tuyến. Video chứa nội dung bạo lực, kích động căm ghét và khủng bố vẫn tồn tại trên mạng gây ảnh hưởng tiêu cực đến tâm lý người xem và thúc đẩy hành vi bạo lực trong xã hội. Việc lan truyền và tiếp cận với những video khiêu dâm và nội dung liên quan đến lạm dụng trẻ em có

thể gây hại nghiêm trọng đến tâm lý của trẻ em. Video chứa thông tin giả mạo hoặc sai lệch gây rối loạn thông tin và ảnh hưởng đến sự tin tưởng và sự hiểu biết của người xem.

Mạng xã hội và các nền tảng trực tuyến khác tạo điều kiện thuận lợi cho các video độc hại lan truyền nhanh chóng, khiến việc kiểm soát và ngăn chặn nội dung độc hại trở nên khó khăn. Để đối phó với tình trạng tràn lan video chứa nội dung không phù hợp như trên, các nền tảng trực tuyến cần áp dụng các biện pháp hạn chế các nội dung này như lọc nội dung tự động, kiểm duyệt nội dung hay khuyến khích người dùng báo cáo nội dung độc hại.

## 1.2 Phát biểu bài toán

Trong nghiên cứu này, chúng tôi tập trung phân loại video độc hại tiếng Việt dựa trên các mô hình học sâu. Bài toán được xác định như sau:

- Đầu vào: video.
- Đầu ra: nhãn của video đó, là bình thường hay thuộc loại video độc hại nào (kinh dị, xúc phạm, phản cảm hay bạo lực).



Hình 1 Minh họa bài toán phân loại video độc hại Tiếng Việt

Bài toán phân loại video độc hại là một trong những bài toán quan trọng hiện nay nhằm giảm thiểu các ảnh hưởng tiêu cực đến con người và xã hội. Việc phân loại được những video có nội dung không lành mạnh này góp phần đưa ra biện pháp giám sát các video độc hại trên mạng xã hội, đồng thời hoàn thiện hệ thống kiểm duyệt các video trước khi đưa tới người dùng và từ đó tạo ra những nền tảng mạng xã hội lành mạnh hơn, đặc biệt là cho trẻ em.

## 1.3 Các thách thức

Hiện tại, việc áp dụng các mô hình speech-to-text (STT) hoặc automatic speech recognition (ASR) cho tiếng Việt đã có sự phát triển, nhưng vẫn có một số thách thức khi áp dụng cho loại dữ liệu như âm thanh của video “độc hại” với âm thanh không rõ ràng và nhiều từ ngữ không phù hợp. Để xây dựng mô hình STT cho tiếng Việt, cần có một bộ dữ liệu đủ lớn và đa dạng, chứa các mẫu âm thanh tiếng Việt độc hại. Tuy nhiên, việc thu thập dữ liệu âm thanh độc hại có thể gặp khó khăn, đặc biệt là khi nó liên quan đến các từ ngữ không phù hợp. Các mô hình STT hiện tại cho tiếng Việt đang có hiệu suất tốt đối với âm thanh rõ ràng và từ vựng thông thường, tuy nhiên, khi đối mặt với âm thanh không rõ ràng và từ ngữ không phù hợp, hiệu suất nhận dạng có thể giảm, các từ ngữ không phù hợp có thể được nhận dạng sai hoặc bị bỏ qua. Việc nhận dạng và xử lý tiếng nói độc hại là một thách thức phức tạp, các từ ngữ không phù hợp trong âm thanh độc hại có thể gây ra những khó khăn trong quá trình nhận dạng và có thể yêu cầu các bước xử lý bổ sung như lọc từ ngữ không phù hợp hoặc xử lý ngữ cảnh.

Việc thực hiện demo với mô hình phân loại audio độc hại tiếng Việt dựa trên thời gian thực có thể thực hiện, tuy nhiên demo có thể phụ thuộc vào một số yếu tố và đòi hỏi kiến thức và kỹ năng về xử lý âm thanh, huấn luyện mô hình và triển khai hệ thống:

- Bộ dữ liệu đủ lớn và đa dạng chứa các mẫu âm thanh độc hại và không độc hại. Việc thu thập, gán nhãn và xử lý dữ liệu âm thanh có thể đòi hỏi thời gian và công sức.
- Một mô hình phân loại audio độc hại hiệu quả cần được xây dựng và huấn luyện, đòi hỏi kiến thức và kỹ năng về xử lý âm thanh, trích xuất đặc trưng và mô hình hóa. Việc huấn luyện mô hình có thể mất thời gian và yêu cầu tài nguyên tính toán đáng kể.
- Thực hiện phân loại audio độc hại trong thời gian thực đòi hỏi xử lý nhanh chóng và hiệu quả của mô hình. Nếu mô hình phức tạp hoặc yêu cầu nhiều tài nguyên tính toán, việc thực hiện thời gian thực có thể gặp khó khăn. Cần xem xét đủ tài nguyên tính toán và môi trường chạy để đảm bảo mô hình có thể xử lý âm thanh trong thời gian thực.

- Để thực hiện demo trong thời gian thực, cần xây dựng một hệ thống hoặc ứng dụng có khả năng xử lý âm thanh đầu vào và trả về kết quả phân loại ngay lập tức. Điều này có thể yêu cầu kiến thức về triển khai mô hình, xử lý dữ liệu âm thanh và tích hợp hệ thống.

#### **1.4 Các vấn đề cần giải quyết**

Các vấn đề mà đề tài tập trung nghiên cứu và giải quyết bao gồm:

- Thu thập và phát triển một bộ dữ liệu có khả năng sử dụng cho việc nghiên cứu mô hình phân loại video tiếng Việt.
- Trích xuất các đặc trưng từ video và phân tích ảnh hưởng các đặc trưng đó với nhiệm vụ phân loại video độc hại tiếng Việt.
- Phân loại video độc hại tiếng Việt dựa trên các đặc trưng đã trích xuất được áp dụng những mô hình học sâu hiện đại cho văn bản tiếng Việt, ảnh và video.
- Nhận xét về hiệu quả của các đặc trưng và các mô hình hiện đại đối với nhiệm vụ phân loại video độc hại dựa trên bộ dữ liệu đã thu thập.

## CHƯƠNG 2: CÁC CÔNG TRÌNH NGHIÊN CỨU LIÊN QUAN

Chương này sẽ trình bày các nghiên cứu liên quan đến các bộ dữ liệu là video ở trong và ngoài nước, các phương pháp biểu diễn đặc trưng của video, các mô hình phân loại video và các nghiên cứu liên quan đến xử lý âm thanh trong tác vụ phân loại video.

### 2.1 Các bộ dữ liệu:

Hiện nay đã có một số bộ dữ liệu chuẩn được sử dụng cho việc nghiên cứu tác vụ phân loại video, một số bộ dữ liệu đáng chú ý gần đây có thể kể đến như là bộ Something-something [14] năm 2017 và JHMDB [19] năm 2011 chứa các video ngắn mô tả các tương tác giữa con người và vật thể. Tập dữ liệu này tập trung vào việc ghi lại các hành động tinh tế và nhiều sắc thái thay vì các hành động có cấu trúc cao hoặc được xác định trước. Năm 2018, nhóm nghiên cứu của Abu-El-Haija và những người cộng sự công bố bộ dữ liệu Youtube-8M [9] được chú thích bởi 4800 thực thể trực quan và mô hình pre-train bằng bộ dữ liệu này có khả năng cải thiện hiệu quả học trên những bộ dữ liệu khác. Ngoài ra còn có bộ dữ liệu chứa hình ảnh các hoạt động của con người được trích xuất từ các video trên YouTube như MPII Human Pose [28] hay bộ Sports-1M [20] bao gồm các video được phân loại dựa trên các môn thể thao. Phần lớn các bộ dữ liệu nói trên phụ vụ tác vụ nhận diện hành động hoặc vật thể sửa dụng thông tin âm thanh và hình ảnh từ video và đạt được những thành tựu đáng kể khi áp dụng các mô hình học sâu. Một vài chủ đề phân loại video dựa trên nội dung như phân loại video của các thể loại game, thể loại phim, hoặc các thể loại video (thể thao, hoạt hình, âm nhạc, dự báo thời tiết,...) [41, 42, 43]. Tuy nhiên, các nghiên cứu này thường bỏ qua yếu tố ngôn ngữ được thể hiện trong video. Vào năm 2024, Zhongping Zhang và cộng sự [44] đã công bố một phương pháp mới gọi là Movie-CLIP để cập nhập những thiếu sót trong các mô hình phân loại thể loại video hiện nay bằng cách kết hợp yếu tố ngôn ngữ và lấy mẫu dữ liệu hiệu quả từ video. Kết quả thu được cải thiện 6-9% trên độ đo mAP trên bộ dữ liệu MovieNet và Condensed Movies.

Đa số các bộ dữ liệu lớn kể trên phục vụ tác vụ nhận diện hành động/ vật thể và các nghiên cứu mô hình học máy hiện đại cũng tập trung giải quyết tác vụ trên và đã đạt được những thành tựu đáng kể. Ở Việt Nam, các bộ dữ liệu video được công bố là UIT-Anomaly của Dung T.T Vo và cộng sự, bộ dữ liệu bao gồm tổng cộng 224 video từ 6 nhãn khác nhau để phát hiện bất thường từ video [1]. Cùng chủ đề và ý tưởng với UIT-Anomaly, VNAnomaly do nhóm nghiên cứu của Tu N. Vu và cộng sự [4] là bộ dữ liệu phát hiện bất thường khác với dữ liệu được thu thập từ camera giám sát. Năm 2016, DH Thuan và cộng sự [5] đã tiến hành trích xuất đặc trưng của sóng và bờ biển từ video của camera giám sát nhằm nghiên cứu về sự phục hồi của biển Nha Trang sau các cơn bão qua thời gian. Năm 2021, một nhóm sinh viên Trường ĐH Công nghiệp Hà Nội đã nghiên cứu thực hiện đề tài “phát hiện hành vi bất thường trong video dựa trên các dạng đặc trưng khác nhau, ứng dụng cho bài toán phát hiện gian lận thi cử” sử dụng các mô hình, thuật toán để trích xuất và phân tích ảnh hưởng của các đặc trưng về khuôn mặt, cử động cơ thể đối với tác vụ phát hiện hành vi bất thường trong kiểm tra.

Tuy có rất nhiều bộ dữ liệu video được thu thập và gán nhãn ở trong và ngoài nước, nhưng hiện vẫn chưa ghi nhận bộ dữ liệu video nào làm về phân loại nội dung của video có sử dụng yếu tố ngôn ngữ là tiếng Việt. Điều này vừa là cơ hội, vừa là thách thức với chúng tôi khi xây dựng bộ dữ liệu và lựa chọn mô hình sao cho phù hợp với tác vụ. Nhóm chúng tôi hy vọng rằng bộ dữ liệu chúng tôi thu thập sẽ được ghi nhận và phục vụ nghiên cứu trong tương lai.

## 2.2 Các phương pháp biểu diễn đặc trưng của video

Các phương pháp biểu diễn đặc trưng cho tác vụ phân loại video ban đầu là những phương pháp thủ công truyền thống dựa trên các giải thuật học sâu hiện đại, một trong những phương pháp phổ biến là điểm quan tâm không gian thời gian (spatiotemporal interest points viết tắt là STIPs) [23], còn được gọi là điểm chính về không gian thời gian hoặc điểm nổi bật về không gian thời gian. Chúng là các vị trí cụ thể trong video thể hiện những thay đổi đáng kể hoặc các mẫu đặc biệt theo cả không gian và thời gian, nhằm mục đích nắm bắt cả đặc điểm không gian và thời gian trong dữ liệu video. Quỹ đạo dày đặc được

cải tiến (improved Dense Trajectories viết tắt là iDT) [15] là phương pháp trích xuất đặc trưng không gian thời gian được xây dựng dựa trên phương pháp Quỹ đạo dày đặc (Dense Trajectories) ban đầu do Wang và cộng sự đề xuất vào năm 2011, giới thiệu một số cải tiến để nâng cao tính chính xác và mạnh mẽ của việc trích xuất đặc trưng. Ngoài ra còn có SIFT-3D [32] và HOG3D [21], ...

### 2.3 Các mô hình phân loại video

Các mô hình dựa trên hình ảnh (2D CNN) như yolo, VGG16, Mobilenet, DenseNet, ... có thể được sử dụng thông tin không gian của khung hình để phân loại video dựa trên các khung hình độc lập. Mặc dù các mô hình dựa trên hình ảnh này vẫn có khả năng giải quyết nhiệm vụ phân loại video theo nội dung, nhận dạng hành động hay phát hiện sự kiện, tuy nhiên chúng vẫn còn nhiều hạn chế vì không thể nắm bắt hết được các tính chất phức tạp của video. Vì vậy đòi hỏi sự tinh chỉnh các mạng lưới có khả năng giúp mô hình học được nhiều loại thông tin như hình ảnh và cả tiếng nói trong không gian và thời gian. Một cải tiến từ mạng tích chập nhằm giải quyết vấn đề này là mô hình 3D CNN (3D Convolutional Neural Network), là một dạng mạng nơ-ron tích chập được sử dụng để xử lý dữ liệu không gian và thời gian trong video. Nó mở rộng kiến trúc của Convolutional Neural Networks (CNNs) bằng cách áp dụng các lớp convolution theo không gian và thời gian trên các khối dữ liệu không gian 3D.

Mô hình 3D CNN có thể vừa ứng dụng trong cả hình ảnh 3D vừa ứng dụng trong dữ liệu là video. Đầu vào của mô hình 3D CNN trong phân loại video là một chuỗi các khung hình liên tiếp từ video. Thông thường, một video được chia thành các khung hình (frames) có kích thước nhất định. Mỗi khung hình đại diện cho một hình ảnh trong video. Để sử dụng mô hình 3D CNN, các khung hình này được tổ chức thành một chuỗi tuần tự (sequence) theo thứ tự thời gian. Chuỗi này thể hiện sự liên kết không chỉ trong không gian mà còn trong thời gian, cho phép mô hình học được thông tin đặc trưng không gian và thông tin đặc trưng thời gian từ video. Một số mô hình 3D CNN phổ biến bao gồm C3D, I3D (Inflated 3D CNN), và R(2+1)D. Các mô hình này có kiến trúc

phức tạp hơn so với CNN 2D truyền thông thường, cho phép học được cả đặc trưng không gian và đặc trưng thời gian từ các khung hình liên tiếp trong video. Mô hình 3D CNN đã chứng minh hiệu quả trong nhiều nhiệm vụ phân loại video như nhận dạng hành động, phân loại video theo nội dung, nhận dạng đối tượng di động, và phát hiện sự kiện trong video.

Factorized Spatial-Temporal Convolutional Network (FSTCN) là một mô hình dựa trên hình ảnh trong phân loại video. Mô hình này tách riêng thông tin không gian và thông tin thời gian bằng cách sử dụng lớp Convolutional dạng factorized. Thông thường, mô hình Convolutional Neural Network (CNN) truyền thống sử dụng các lớp Convolutional 2D để xử lý cả đặc trưng không gian và thời gian trong video. Tuy nhiên, trong FSTCN, các lớp Convolutional 2D được thay thế bằng sự kết hợp của lớp Convolutional 1D và Convolutional 2D, dẫn đến việc tách biệt thông tin không gian và thông tin thời gian. Việc tách biệt này giúp giảm số lượng trọng số cần học và làm giảm độ phức tạp tính toán và số lượng trọng số cần học, đồng thời giúp mô hình tập trung vào việc học các đặc trưng của không gian và thời gian một cách riêng biệt.

Không chỉ các mô hình CNN, các mô hình RNN cũng có thể sử dụng để phân loại video. Có hai mô hình RNN phổ biến trong phân loại video là "classical LSTM" (LSTM cổ điển) và "encoder-decoder LSTM" (LSTM mã hóa-giải mã). Mô hình LSTM cổ điển trong phân loại video thường sử dụng một lớp LSTM duy nhất để mô hình hóa thông tin thời gian trong chuỗi các khung hình. Mỗi khung hình của video được truyền qua mạng LSTM theo thứ tự thời gian. Đầu ra của LSTM sau cùng (thường là ở khung hình cuối cùng) được sử dụng để phân loại video thành các lớp đã cho. Để trích xuất đặc trưng không gian từ các khung hình, ta có thể sử dụng mạng tích chập 2D (CNN) trước đó, sau đó đưa các đặc trưng đã trích xuất vào mô hình LSTM để mô hình hóa thông tin thời gian. Mặc khác, mô hình LSTM mã hóa-giải mã (Encoder-Decoder LSTM) trong phân loại video thường được sử dụng để xử lý các tác vụ như dự đoán chuỗi hoặc tạo mô tả video. Mô hình LSTM mã hóa-giải mã bao gồm hai phần: phần mã hóa (encoder) và phần giải mã (decoder).

Phần mã hóa có nhiệm vụ mô hình hóa thông tin thời gian trong chuỗi các khung hình để tạo ra một biểu diễn ngữ cảnh (context representation). Biểu diễn này sau đó được đưa vào phần giải mã để sinh ra đầu ra dự đoán hoặc mô tả. Cả hai mô hình trên đều có thể được kết hợp với các lớp mạng nơ-ron khác như CNN để trích xuất đặc trưng không gian từ các khung hình của video. Cách kết hợp và cấu trúc của mô hình sẽ phụ thuộc vào mục tiêu và yêu cầu cụ thể của bài toán phân loại video.

Việc huấn luyện các mô hình phân loại video đòi hỏi sự sẵn có của bộ dữ liệu đã được gán nhãn. Việc thu thập và gán nhãn dữ liệu cho video đòi hỏi đến sự đầu tư lớn về thời gian và nỗ lực. Tính đến thời điểm hiện tại, chưa có bộ dữ liệu nào phục vụ cho nghiên cứu phân loại video dựa trên yếu tố âm thanh với ngôn ngữ là tiếng Việt. Bộ dữ liệu mà chúng tôi đã tiến hành thu thập, gán nhãn và nghiên cứu sẽ là bộ dữ liệu đầu tiên về phân loại video dựa trên nội dung có sử dụng dụng yếu tố ngôn ngữ Việt.

## 2.4 Các phương pháp nghiên cứu xử lý âm thanh

Trích xuất đặc trưng âm thanh từ tín hiệu âm thanh gốc là bước quan trọng đầu tiên trong xử lý âm thanh. Hiện nay, có nhiều phương pháp phổ biến để rút trích đặc trưng âm thanh, trong đó Mel-frequency cepstral coefficients (MFCCs) [12] là một phương pháp phổ biến. Phương pháp này sử dụng biến đổi Fourier để chuyển đổi tín hiệu âm thanh thành miền tần số Mel, sau đó trích xuất các hệ số cepstral để biểu diễn đặc trưng của âm thanh. Một phương pháp khác là Spectrogram [29], nó biểu thị phổ tần số của tín hiệu âm thanh theo thời gian. Spectrogram được tạo ra bằng cách chia tín hiệu âm thanh thành các khung nhỏ và tính biến đổi Fourier cho mỗi khung. Spectrogram hiển thị biên độ và phân bố tần số của âm thanh. Ngoài ra, còn phương pháp Mel-scaled spectrogram [34], một biến thể của spectrogram, chuyển đổi trực tiếp tần số sang miền tần số Mel. Điều này giúp phương pháp này tương thích hơn với cách mà người nghe cảm nhận âm thanh, tạo ra một biểu đồ phổ có thang đo tần số Mel.

Phương pháp Fourier (FFT) [29] được sử dụng để biến đổi tín hiệu âm thanh từ miền thời gian sang miền tần số, cho phép phân tích phổ tần số của âm thanh. Ngoài ra, phương pháp biến đổi wavelet [26] cũng được áp dụng để phân tích tín hiệu âm thanh cả ở miền thời gian và miền tần số, mang lại thông tin về cả biên độ và thời gian của tín hiệu.

Các phương pháp nhận dạng giọng nói, như Hidden Markov Models (HMMs) [30] và các mạng nơ-ron sâu (deep neural networks) [13], là những phương tiện phổ biến trong lĩnh vực nhận dạng giọng nói. HMMs sử dụng mô hình xác suất để mô phỏng các trạng thái của giọng nói. Trong khi đó, mạng nơ-ron sâu đã đạt được sự thành công lớn trong việc nhận dạng giọng nói. Các kiến trúc như Convolutional Neural Networks (CNNs) và Recurrent Neural Networks (RNNs) thường được áp dụng để trích xuất đặc trưng và phân loại giọng nói.

## CHƯƠNG 3: NGHIÊN CỨU THỰC NGHIỆM VÀ LÍ THUYẾT

Chương này trình bày về bộ dữ liệu chúng tôi thu thập và các phương pháp nghiên cứu được sử dụng.

### 3.1 Bộ dữ liệu

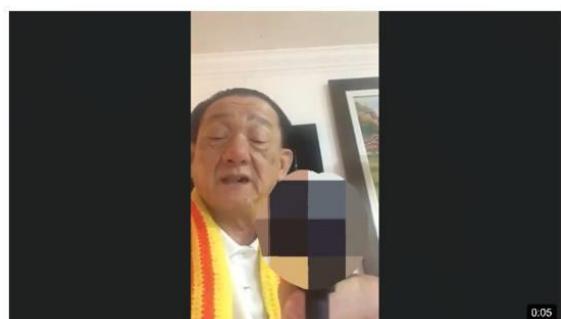
Nghiên cứu được thực nghiệm trên bộ dữ liệu HarmfulVideosVN2023, một bộ dữ liệu về các video độc hại mà nhóm thu thập trên nền tảng Youtube. Bộ dữ liệu bao gồm 1.589 video được gán 5 nhãn “bình thường”, “kinh dị”, “xúc phạm”, “nhạy cảm” và “bạo lực” với độ đồng thuận 96%. Hình 2 minh họa các video với nhãn tương ứng.



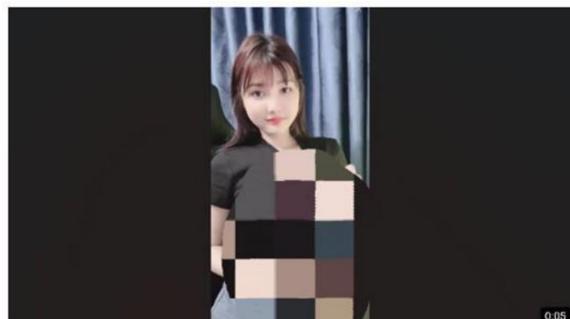
Bình thường



Kinh dị



Xúc phạm



Phản cảm



## Bạo lực

Hình 2 Video minh họa các nhãn của bộ dữ liệu

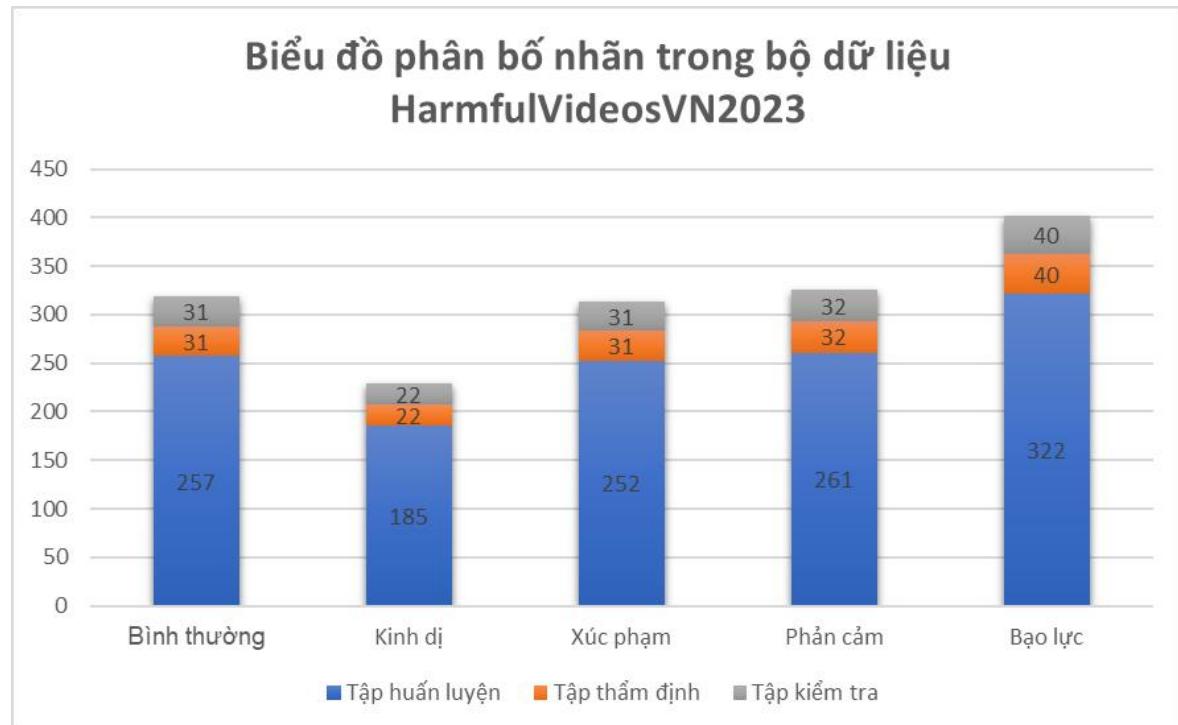
### 3.1.1 Thu thập dữ liệu và gán nhãn

Bộ dữ liệu video được thu thập từ Youtube bằng cách sử dụng YouTube Data API của Google, từ khóa để tìm kiếm và tải video dựa trên định nghĩa của mỗi nhãn. Các nhãn của bộ dữ liệu được định nghĩa như sau:

- Bình thường: là các video có nội dung bình thường, không vi phạm tiêu chuẩn cộng đồng.
- Kinh dị: là các video có nội dung kích động, kích thích một cách không phù hợp với trẻ em, bao gồm hình ảnh đáng sợ, kinh dị, hoặc gây sợ hãi.
- Xúc phạm: là các video chứa nội dung có tính xúc phạm, công kích, hoặc nhạo báng đối với cá nhân hoặc nhóm người khác. Video có thể bao gồm những hành vi, lời nói hoặc hành động không tôn trọng, phân biệt chủng tộc, kích động bạo lực, xúc phạm về giới tính, tôn giáo, quốc gia, hoặc những đặc điểm cá nhân khác.
- Phản cảm: là các video có thể chứa các tình huống và hành động nhạy cảm, như hành vi tình dục giữa người lớn, quảng cáo hoặc trình diễn nội dung khiêu dâm, hoặc hiển thị cơ thể một cách không phù hợp hoặc mở cửa.
- Bạo lực: các video chứa hình ảnh hoặc nội dung có hành vi bạo lực hoặc bạo lực vật lý đối với người hoặc động vật. Video có thể bao gồm những hình ảnh nhạy cảm, bạo lực, đau đớn, gây tổn thương hoặc lạm dụng. Ví dụ hình ảnh về cuộc tấn công vũ trang, hành động bạo lực hoặc đánh nhau, video về động vật bị hành hạ hoặc bị tấn công, hình ảnh chứng kiến vụ tai nạn, thảm kịch hoặc tình huống nguy hiểm gây tổn thương cho con người.

Sau khi thu thập video, nhóm tiến hành cắt video với độ dài trung bình là 5 giây tập trung vào nội dung của nhãn và gán nhãn theo video thay vì từng khung frames của video. Điều này sẽ gây khó khăn cho mô hình phân loại vì mô hình không thể biết được rằng những video được gán nhãn độc hại dựa trên frame nào vì có những frame ảnh tĩnh trong video vẫn được coi là bình thường, điều này sẽ vừa là thách thức với các mô hình, vừa là cơ hội để các mô hình

hiện đại hơn ra đời có khả năng học từ những dữ liệu mơ hồ. Bộ dữ liệu bao gồm 1.589 video định dạng mp4. Trong đó, tập huấn luyện (Train) có 1277 video, tập thẩm định (Validation) có 156 video và tập kiểm tra (Test) có 156 video. Hình 4 thể hiện phân bố nhãn trong bộ dữ liệu. Số lượng video trong mỗi nhãn được thống kê trong Bảng 1.

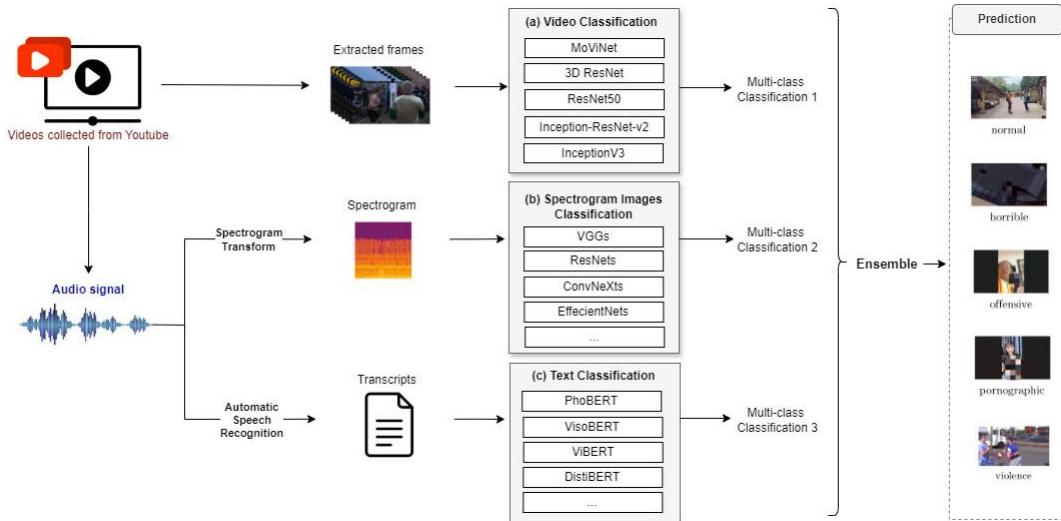


Hình 3 Phân bố nhãn trong bộ dữ liệu HarmfulVideosVN2023

Label	Train	Val	Test	Total
Bình thường	257	31	31	319
Kinh dị	185	22	22	229
Xúc phạm	252	31	31	314
Nhạy cảm	261	32	32	325
Bạo lực	322	40	40	402

Bảng 1 Thống kê chi tiết số lượng video trong bộ dữ liệu

Sau khi xây dựng bộ dữ liệu HarmfulVideosVN2023, nhóm rút trích đặc trưng từ video, sau đó huấn luyện mô hình. Quy trình thực hiện được minh họa trong Hình 5:



Hình 4 Minh họa quy trình xử lý dữ liệu và các mô hình phân loại

Các đặc trưng của video được phân chia thành hai phần chính là âm thanh và khung ảnh. Đối với phần âm thanh của video, nhóm nghiên cứu thực hiện quá trình trích xuất đặc trưng, bao gồm việc thu được văn bản và tạo ảnh quang phổ spectrogram từ tín hiệu âm thanh. Văn bản trích xuất từ âm thanh được đưa vào các mô hình phân loại văn bản, trong khi ảnh spectrogram trích xuất từ âm thanh được đưa vào các mô hình phân loại ảnh. Các frame của video được đưa vào các mô hình phân loại video để thực hiện quá trình phân loại dựa trên nội dung hình ảnh.

Tổng cộng, quá trình này giúp kết hợp thông tin từ cả hai thành phần chính của video, âm thanh và hình ảnh, để tạo ra một hệ thống phân loại video đa chiều và đa dạng, sử dụng đa dạng mô hình phân loại cho mỗi loại đặc trưng. Điều này có thể cung cấp một cái nhìn toàn diện và độ chính xác cao về nội dung của video trong quá trình phân loại và phân tích.

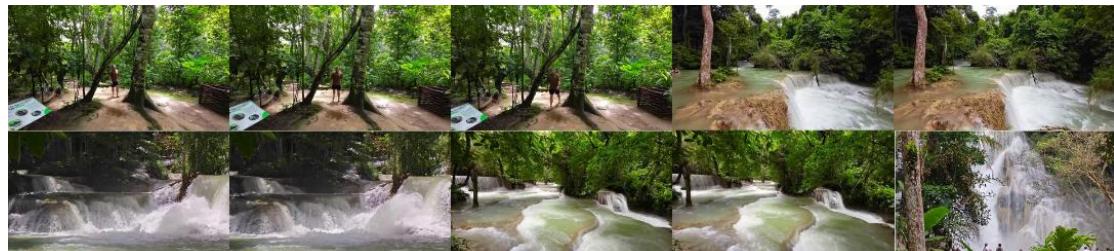
### 3.1.2 Trích xuất đặc trưng

Vì tác vụ phân loại video độc hại tiếng Việt này yêu cầu cả thông tin hình ảnh (cho nhãn “Bình Thường”, “Kinh Dị”, “Nhạy Cảm” và “Bạo Lực”) và thông tin âm thanh (cho nhãn “Bình Thường” và “Xúc Phạm”) nên chúng tôi đã tiến hành các phương pháp trích xuất đặc trưng để phân tích ảnh hưởng của chúng. Quá trình rút trích đặc trưng của chúng tôi bao gồm: trích xuất các khung hình

ảnh từ video, trích xuất văn bản từ âm thanh của video và trích xuất hình ảnh spectrogram từ âm thanh của video.

### 3.1.2.1 Trích xuất khung hình ảnh

Từ dữ liệu thô của video có độ dài vài giây, chúng tôi sử dụng 2 khung hình ảnh mỗi giây để đưa vào mô hình.



Hình 5 Các khung hình ảnh từ video "bình thường"



Hình 6 Các khung hình ảnh từ video "kinh dị"



Hình 7 Các khung hình ảnh từ video "bạo lực"

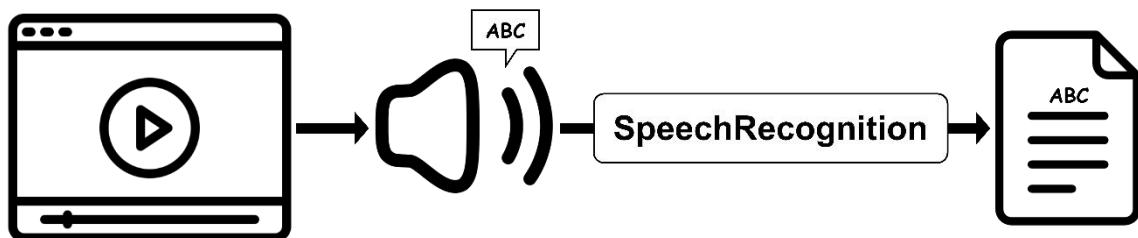
### 3.1.2.2 Trích xuất văn bản từ âm thanh của video

Trích xuất văn bản từ âm thanh là quá trình chuyển đổi âm thanh từ một tệp có chứa âm thanh thành văn bản tương ứng. Trong nghiên cứu này, chúng tôi sử dụng thư viện SpeechRecognition để trích văn bản từ âm thanh của video. SpeechRecognition là một thư viện Python mạnh mẽ cho phép nhận dạng và trích xuất văn bản từ dữ liệu âm thanh hoặc giọng nói.

Trong nghiên cứu này, thư viện SpeechRecognition là phù hợp để dùng cho việc trích xuất văn bản từ âm thanh. Tuy rằng hiệu suất và độ chính xác có

thể không cao như các thư viện khác nhưng SpeechRecognition hỗ trợ nhiều ngôn ngữ, đồng thời chi phí và thời gian tính toán thấp. Ngoài ra SpeechRecognition còn có một vài ưu điểm so với các thư viện khác thích hợp để trích xuất văn bản từ âm thanh trong nghiên cứu này.

	ƯU ĐIỂM	NHƯỢC ĐIỂM
SpeechRecognition	<ul style="list-style-type: none"> <li>- Dễ sử dụng và tích hợp với các dịch vụ nhận dạng giọng nói.</li> <li>- Hỗ trợ nhiều ngôn ngữ.</li> <li>- Nhiều tài liệu hướng dẫn.</li> </ul>	<ul style="list-style-type: none"> <li>- Hiệu suất, độ chính xác có thể không cao bằng các giải pháp phức tạp.</li> </ul>
Pocketsphinx	<ul style="list-style-type: none"> <li>- Nhỏ và phù hợp cho ứng dụng trên điện thoại di động.</li> <li>- Hoạt động offline.</li> </ul>	<ul style="list-style-type: none"> <li>- Hỗ trợ ngôn ngữ có thể hạn chế.</li> </ul>
Vosk	<ul style="list-style-type: none"> <li>- Độ chính xác tương đối cao.</li> <li>- Hỗ trợ nhiều ngôn ngữ.</li> <li>- Hoạt động offline.</li> </ul>	<ul style="list-style-type: none"> <li>- Đòi hỏi kiến thức về xử lý giọng nói.</li> </ul>
Deepspeech (mozilla)	<ul style="list-style-type: none"> <li>- Sử dụng mô hình học sâu để cải thiện độ chính xác.</li> <li>- Hỗ trợ nhiều ngôn ngữ.</li> </ul>	<ul style="list-style-type: none"> <li>- Kích thước mô hình lớn, yêu cầu tài nguyên tính toán cao.</li> </ul>



Hình 8 Minh họa quá trình trích xuất văn bản từ âm thanh của video

index	name	label	text
412	0156.mp4	horrible	bọn họ Sau một hồi cổ gắng đã không thể được con quý
421	0165.mp4	horrible	một game
922	0229.mp4	pornographic	nếu mà nó
212	0213.mp4	normal	không có chỉ là tức
444	0003.mp4	offensive	Địt mẹ chúng mày thằng ranh con này chứ phải tôi tôi đầm cho mày nha
686	0245.mp4	offensive	mày chửi tao vậy đó hả
1189	0235.mp4	violent	Trước tiên là con đi khám cái quần của con
1070	0116.mp4	violent	ba đối di con
115	0116.mp4	normal	chỉ còn một cái hang đá duy nhất ở Đông Hoàng đó là hang đá mặt trăng thi minh chưa tới thời
521	0080.mp4	offensive	Đập mẹ mày mày Bùa nay mày
526	0085.mp4	offensive	Đu má mày dám chửi tao hả mày ngoan lắm Mày học theo hướng nào
563	0122.mp4	offensive	cái đít cái con mèo thế Mày chán sống được con ạ với mày đợi bố
301	0045.mp4	horrible	anh ta di xúc động mạnh
619	0178.mp4	offensive	máy bơm xăng cho loài ốc Tiên sư bố cái thẳng trời đánh thánh vật đi tắm bổ mày đúng không cái loại môn Sinh học bác sĩ tôi có i
770	0077.mp4	pornographic	những câu khiến họ bị thương cũng không nhẹ khi họ đang chưa lành vết thương tại một ngôi chùa cổ
235	0236.mp4	normal	người đang đứng đợi mua và ăn ở đây hấp dẫn quá Chắc là khoai sê kêu một cái
660	0219.mp4	offensive	Địt mẹ nhà mày Hay mày thích bố mày
221	0222.mp4	normal	đồn biên phòng Cửa khẩu quốc tế Lao Bảo thuộc Bộ đội Biên phòng tỉnh Quảng Trị cho biết đơn
382	0126.mp4	horrible	rời bỏ những vết khâu trên người liên tục xuất hiện Lisa không suy nghĩ liền dùng kéo

Hình 9 Dữ liệu văn bản trích xuất từ âm thanh của các video.

Nội dung các văn bản trích xuất được có nhãn offensive (“xúc phạm”) là có thể phân biệt được khi chứa những từ ngữ mang tính tiêu cực, các nhãn khác như violent (bạo lực), normal (bình thường), horrible (kinh dị) và pornographic (phản cảm) đều mang sắc thái văn bản tích cực hoặc trung tính nên khó có thể phân biệt với nhau. Vì vậy không thể sử dụng đặc trưng văn bản này một cách riêng lẻ để dự đoán nhãn của các video mà cần kết hợp nó với các đặc trưng khác.

Do các video thu thập được có những video không sử dụng tiếng nói hoặc âm thanh thu được không rõ ràng để mô hình có thể trích xuất được văn bản từ tiếng nói nên số lượng dữ liệu sau khi trích xuất âm thanh giảm so với số lượng video. Bảng thống kê số lượng nhãn với dữ liệu văn bản sau khi đã loại bỏ đi các giá trị NaN và biểu đồ trực quan được thể hiện dưới đây.

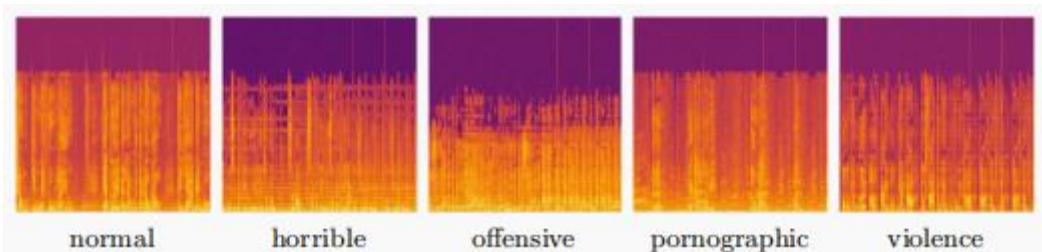
Label	Train	Val	Test	Total
Bình thường	193	23	22	238
Kinh dị	121	14	10	145
Xúc phạm	128	17	13	158
Nhạy cảm	47	2	7	56
Bạo lực	163	29	26	218

Bảng 2 Thống kê số lượng văn bản trích xuất được

### 3.1.2.3 Trích xuất spectrogram của âm thanh video

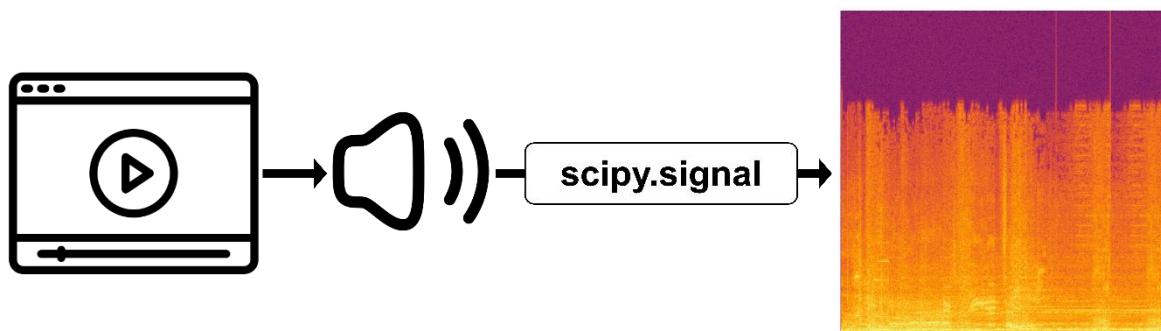
Spectrogram là biểu đồ thường được sử dụng để biểu diễn biên độ của tín hiệu âm thanh theo thời gian. Việc chuyển đổi dữ liệu âm thanh thành biểu

đồ spectrogram là một quá trình quan trọng trong việc xử lý và nhận dạng âm thanh. Trích xuất spectrogram của âm thanh từ video là quá trình chuyển đổi tín hiệu âm thanh thành hình ảnh mà mỗi điểm ảnh trong đó thể hiện mức độ năng lượng của âm thanh ứng với một khoảng thời gian cụ thể.



Hình 10 Ảnh spectrogram trích xuất từ âm thanh của một video

Trích xuất spectrogram giúp biểu diễn âm thanh dưới dạng dữ liệu số và tối ưu hóa các mô hình học máy cũng như thuật toán xử lý âm thanh. Quang phổ log mel có kích thước nhỏ hơn tín hiệu âm thanh gốc, giúp giảm chi phí tính toán và tăng tốc độ xử lý dữ liệu. Chúng tôi sử dụng thư viện SciPy trong Python để xử lý tín hiệu âm thanh của video và trích xuất sang ảnh Spectrogram.



Hình 11 Minh họa quá trình trích xuất spectrogram từ âm thanh của video

### 3.2 Các mô hình phân loại

Chúng tôi đã áp dụng các mô hình hiện đại để phân loại video dựa trên từng đặc trưng như: mô hình 3D CNN với kiến trúc ResNet được đào tạo lại từ đầu; các kiến trúc CNN nổi tiếng khác như ResNet50, Inception-ResNet-v2 và InceptionV3 với các trọng số đào tạo trước từ bộ dữ liệu ImageNet với kiến trúc 2D của mô hình cùng tên; học chuyển tiếp với MoViNet được đào tạo trước trên bộ dữ liệu Kinetics. Các mô hình 3D CNN này sẽ sử dụng đặc trưng

là các khung hình ảnh từ video làm input. Tiếp theo, chúng tôi sử dụng các mô hình 2D CNN được đào tạo trước trên bộ dữ liệu ImageNet như các mô hình ResNet, các mô hình VGG, các mô hình DenseNet, ... cho phân loại hình ảnh spectrogram. Cuối cùng, đối với văn bản trích xuất từ lời nói trong video, chúng tôi áp dụng BiLSTM và các mô hình phân loại ngôn ngữ hiện đại như BERT, PhoBERT, viBERT, ViSoBERT,... và nhận được các kết quả đầy hứa hẹn. Cuối cùng, chúng tôi thử kết hợp các đặc trưng trên với nhau bằng phương pháp Ensemble, chúng tôi chọn ra các mô hình có kết quả dự đoán tốt nhất và dùng chúng để đưa ra kết quả dự đoán cuối cùng.

### 3.2.1 Mô hình phân loại video

Các mô hình hiện đại ngày nay tập trung nghiên cứu, xử lý các tác vụ nhận diện hành động, phát hiện, trích xuất vật thể. Các nghiên cứu được tiến hành trên nhiều bộ dữ liệu lớn và đạt được hiệu quả cao. Tuy nhiên các kiến trúc này lại không đạt hiệu quả như vậy đối với tác vụ phân loại video dựa trên nội dung (bình thường, xúc phạm, khiêu dâm, kinh dị hay bạo lực) của video trên bộ dữ liệu HarmfulVideosVN2023 mà chúng tôi đã thu thập. Các mô hình mà chúng tôi sử dụng trong đề tài này là 3D CNN với kiến trúc mạng ResNet, các mô hình 3D CNN được huấn luyện trước trên bộ dữ liệu ImageNet và học chuyển tiếp với MoViNet.

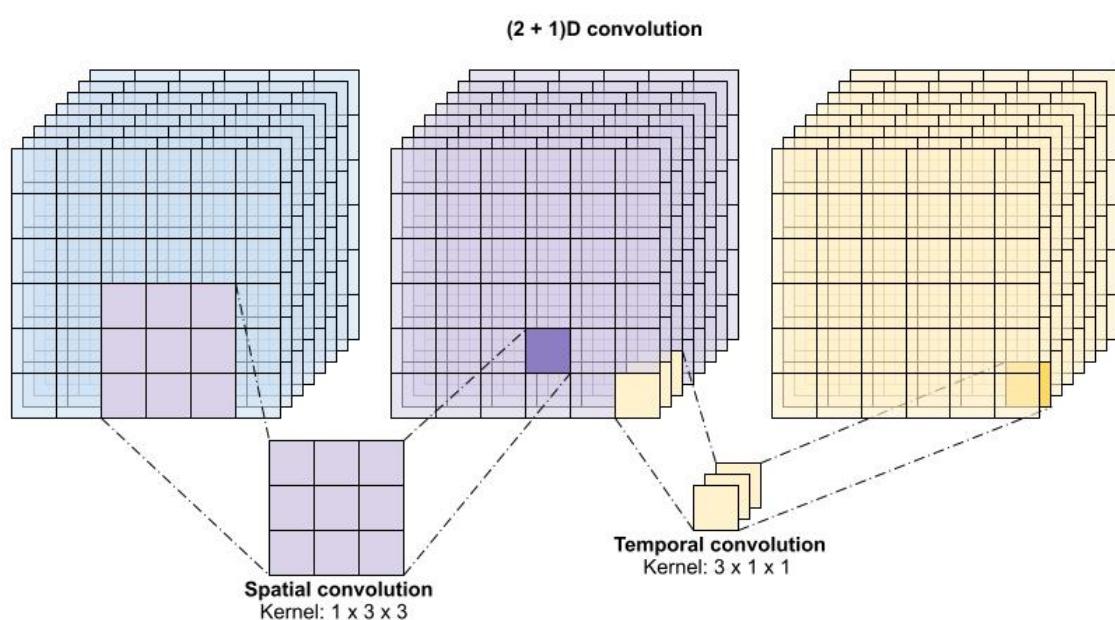
Mô hình 3D CNN có đặc trưng so với mô hình 2D CNN ở số chiều không gian của bộ lọc thực hiện phép tích tích chập. Trong khi ở 3D CNN, kernel có thể trượt theo ba hướng (chiều dài, chiều rộng, chiều sâu) thì 2D CNN có kernel trượt theo hai chiều (chiều dài và chiều rộng). Chính điểm khác biệt này đã giúp cho các mô hình 3D CNN có khả năng xử lý được đặc tính không - thời gian của video mà các mô hình 2D CNN không thể làm được. Tuy nhiên, các mô hình 3D CNN lại không có khả năng trong việc dự đoán dữ liệu streaming vì các mô hình này cần nhìn toàn bộ các frames ảnh để đưa ra dự đoán, trong khi đó các mô hình 2D CNN lại có khả năng đưa ra dự đoán với dữ liệu hiện có và cập nhập kết quả theo thời gian. Các nghiên cứu trước đó đã chỉ ra rằng các mô hình 2D CNNs dựa trên hình ảnh vẫn có khả năng đạt được hiệu

quả cao gần với các phương pháp hiện đại trên dữ liệu video với tác vụ nhận diện hành động, chúng tôi vẫn sử dụng mô hình 3D CNN trong đề tài này thay vì 2D CNN bởi vì đối với bộ dữ liệu của chúng tôi, trong cùng một video được gán nhãn là độc hại vẫn có những frame ảnh không độc hại. Hơn nữa chúng tôi không gán nhãn cho từng frame ảnh để các mô hình 2D CNN thông thường có thể học được.

### 3.2.1.1 3D CNN với ResNet

Chúng tôi sử dụng mạng tích chập (2+1)D CNN để phân loại video trong dự án này dựa trên bài báo A Closer Look at Spatiotemporal Convolutions for Action Recognition bởi nhóm tác giả D. Tran (2017) [6]. Gọi nó là mạng (2+1)D bởi nó chia tích chập 3D thành hai giai đoạn hoạt động riêng biệt liên tiếp nhau, trong đó có mạng tích chập không gian (Spatial convolution) 2D và tích chập thời gian (Temporal convolution) 1D được biểu diễn như hình dưới đây. Việc chia tích chập 3D thành 2D và 1D như vậy giúp mô hình có khả năng biểu diễn các hàm phức tạp hơn bởi thay vì học cùng lúc đặc trưng không gian thời gian, (2+1)D phân giải quá trình này thành hai giai đoạn riêng biệt thông qua phân tách chiều không gian và thời gian.

Type equation here.



Hình 12 Cấu trúc mạng (2+1)D tích chập

Trong đề tài này, chúng tôi sử dụng kiến trúc mạng ResNet dưới dạng các lớp (2+1)D. Các thay đổi kích thước trong mô hình có tác dụng nhận diện các mẫu đối với các tác vụ nhận diện hành động và vật thể. Ngoài ra, đối với các tác vụ khác như phân loại video dựa trên nội dung thì việc thay đổi kích thước giúp loại bỏ các thông tin dư thừa, giảm kích thước dữ liệu, do đó giảm chi phí tính toán và giúp mô hình học nhanh hơn.

### **3.2.1.2 3D CNN với ResNet50, Inception-ResNet-v2 và InceptionV3 đã được đào tạo trước**

Việc sử dụng các kiến trúc 3D CNN như ResNet50, Inception-ResNet-v2 và InceptionV3 đã được đào tạo trước [46] trên bộ dữ liệu ImageNet [39] là một phương pháp hiệu quả trong quá trình phân loại video. ImageNet, một bộ dữ liệu quy mô lớn và phổ biến trong lĩnh vực thị giác máy tính, chứa hình ảnh và nhãn cho hàng triệu hình ảnh từ hàng ngàn đối tượng khác nhau. Đối tượng trong ImageNet bao gồm động vật, vật nuôi, đồ vật, cảnh quan, con người và nhiều loại hình ảnh khác. Bộ dữ liệu ImageNet đã được sử dụng rộng rãi trong quá trình huấn luyện và đánh giá các mô hình thị giác máy tính, đặc biệt là các mô hình nơ-ron tích chập (CNN). Nó đã trở thành tiêu chuẩn để đánh giá và so sánh hiệu suất của các mô hình trong nhiều bài toán, như phân loại hình ảnh, nhận dạng đối tượng, phân loại video và nhiều tác vụ khác liên quan đến thị giác máy tính. Sự sử dụng các kiến trúc đã được đào tạo trước này giúp tận dụng tri thức được học từ ImageNet để cải thiện hiệu suất của mô hình trong nhiệm vụ phân loại video của chúng tôi.

Các trọng số đào tạo trước của các mô hình 3D CNN với kiến trúc mạng ResNet50, Inception-ResNet-v2 và InceptionV3 mà chúng tôi sử dụng được chuyển từ các trọng số đã được huấn luyện từ bộ dữ liệu ImageNet của các mô hình 2D CNN cùng tên. Điều này sẽ giúp tiếp kiệm thời gian và chi phí tính toán so với đào tạo lại từ đầu. Các kiến trúc mạng ResNet50, Inception-ResNet-v2 và InceptionV3 sẽ được chúng tôi trình bày kỹ hơn trong phần 3.2.3 cùng với các kiến trúc mạng VGG16 và DenseNet121 mà chúng tôi sử dụng để phân loại ảnh quang phổ spectrogram.

### **3.2.1.3 Học chuyển tiếp với MoViNet**

Học chuyển tiếp (transfer learning) là một kỹ thuật mạnh mẽ trong học máy, nơi một mô hình đã được huấn luyện trước trên một tập dữ liệu lớn được tinh chỉnh (fine-tuned) trên một tập dữ liệu cụ thể nhỏ hơn. Kỹ thuật này đặc biệt hữu ích khi dữ liệu mới không đủ lớn để huấn luyện một mô hình từ đầu (from scratch). Trong phân loại video, học chuyển tiếp với MoViNet (Mobile Video Networks) [47] đã chứng tỏ hiệu quả cao, giúp tận dụng sức mạnh của mô hình đã được huấn luyện trên các tập dữ liệu video lớn và đa dạng.

MoViNet là một họ mô hình được thiết kế bởi Google để tối ưu hóa việc phân loại video trên các thiết bị di động. Các mô hình MoViNet có cấu trúc gọn nhẹ nhưng mạnh mẽ, có khả năng xử lý video theo thời gian thực với độ chính xác cao. Được huấn luyện trước trên các tập dữ liệu lớn như Kinetics-600, MoViNet có thể nhận diện một loạt các hành động và hoạt động trong video. Trong nghiên cứu này, chúng tôi sẽ thử nghiệm khả năng của MoViNet trong việc phân loại video dựa vào nội dung.

## **3.2.2 Mô hình phân loại văn bản**

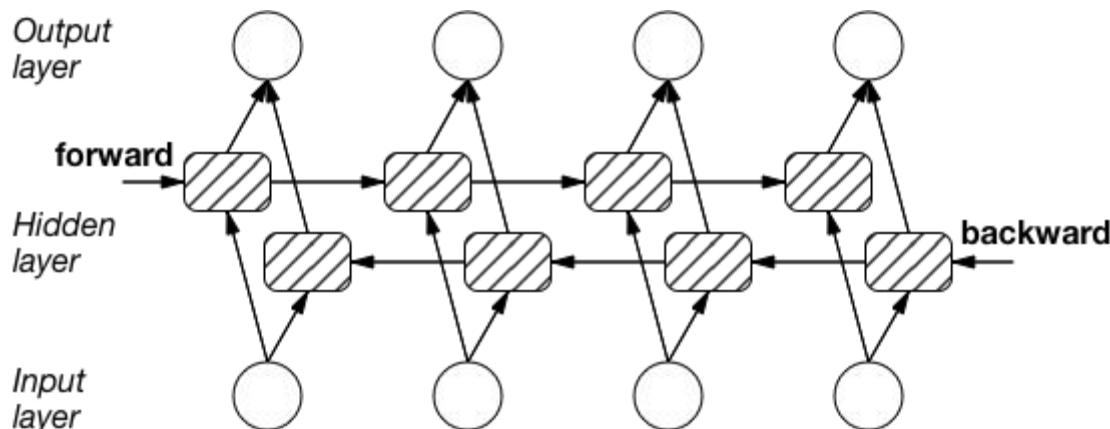
Để phân loại văn bản trích xuất từ âm thanh của video, chúng tôi sử dụng mô hình BiLSTM cổ điển và các mô hình ngôn ngữ hiện đại bao gồm: các mô hình ngôn ngữ được huấn luyện trên dữ liệu là tiếng Anh như BERT, RoBERTa, XLM-RoBERTa, DistilBERT và các mô hình ngôn ngữ được huấn luyện trên dữ liệu tiếng việt PhoBERT, vELECTRA, CafeBERT, ViSoBERT, viBERT.

### **3.2.2.1 BiLSTM (Bidirectional Long Short-Term Memory)**

Trong nghiên cứu này, chúng tôi sử dụng mô hình BiLSTM để phân loại văn bản trích từ âm thanh của video. BiLSTM là một loại mạng nơ ron thần kinh sử dụng phổ biến trong lĩnh vực xử lý ngôn ngữ tự nhiên (NLP). BiLSTM được mở rộng từ mô hình LSTM [17], một loại kiến trúc nơ ron thần kinh được thiết kế để xử lý vấn đề lưu giữ thông tin trong một thời gian dài và quyết định khi nào nên quên thông tin không quan trọng. BiLSTM được mở rộng bằng cách thêm lớp LSTM vào phía sau một lớp LSTM. Mô hình BiLSTM sử dụng

hai lớp LSTM, một chạy theo chiều thuận và một chạy theo chiều ngược (Bidirectional – hướng đảo ngược) của chuỗi đầu vào, giúp mô hình có thể học thông tin từ cả hai hướng, đồng thời cải thiện khả năng hiểu chuỗi và mối quan hệ trong chuỗi.

Mô hình BiLSTM nhận chuỗi đầu vào và chia thành các phần tử như từ, ký tự, mỗi phần tử được biểu diễn bằng một vector. Chuỗi này sẽ được đưa vào lớp LSTM thuận chiều (Forward LSTM). Lớp LSTM thuận chiều xử lý chuỗi theo chiều từ trái sang phải. Các trạng thái ẩn và các cổng của LSTM được cập nhật dựa trên thông tin từ phần tử hiện tại và trạng thái trước đó. Lớp LSTM ngược chiều (Backward LSTM) xử lý chuỗi theo chiều từ phải sang trái. Cũng giống như lớp LSTM thuận chiều, nó cập nhật trạng thái ẩn và các cổng dựa trên thông tin từ phần tử hiện tại và trạng thái trước đó. Kết quả đầu ra từ cả hai lớp LSTM thông thường được cộng dồn hoặc lấy trung bình để tạo ra đầu ra cuối cùng của mô hình.



Hình 13 Kiến trúc của BiLSTM [7]

Nhờ khả năng học được từ cả hai phía, mô hình BiLSTM có khả năng nắm bắt được ngữ cảnh phức tạp hơn so với các mô hình chỉ học được thông tin một chiều.

### 3.2.2.2 BERT (Bidirectional Encoder Representations from Transformers)

BERT (Bidirectional Encoder Representations from Transformers) là một mô hình ngôn ngữ tiên tiến trong lĩnh vực xử lý ngôn ngữ tự nhiên (NLP), được giới thiệu bởi Jacob Devlin và đồng nghiệp vào năm 2018 [27]. BERT đã đạt

được những thành tựu ấn tượng trong nhiều ứng dụng của xử lý ngôn ngữ tự nhiên, như phân loại văn bản, dịch ngôn ngữ và hiểu ngữ cảnh.

Điểm độc đáo của BERT nằm ở khả năng hiểu ngữ cảnh của từng từ trong câu bằng cách xem xét cả hai phía của mỗi từ. Truyền thống, các mô hình chỉ tập trung vào từ trước hoặc từ sau, trong khi BERT "nhìn" cả hai hướng. Điều này giúp BERT có khả năng hiểu mối quan hệ phức tạp giữa các từ trong câu. Một ưu điểm quan trọng của BERT là khả năng áp dụng và tinh chỉnh những kiến thức học được từ một tập dữ liệu lớn cho các nhiệm vụ cụ thể khác mà không cần lượng dữ liệu lớn. Điều này làm cho BERT trở thành một công cụ mạnh mẽ và linh hoạt trong việc xử lý và hiểu ngữ cảnh trong các nhiệm vụ ngôn ngữ tự nhiên.

### 3.2.2.3 PhoBERT

PhoBERT [48] là một mô hình ngôn ngữ tự nhiên đặc biệt được phát triển dựa trên kiến trúc BERT, nhưng được đào tạo trước trên dữ liệu tiếng Việt. Tên “PhoBERT” xuất phát từ việc kết hợp “Phở” - một món ăn nổi tiếng của Việt Nam, và “BERT” - mô hình nền tảng nổi tiếng trong xử lý ngôn ngữ tự nhiên.

Mục tiêu chính của PhoBERT là cung cấp khả năng hiểu ngôn ngữ tự nhiên đối với văn bản tiếng Việt. Nó giữ lại những đặc điểm quan trọng của BERT, như khả năng biểu diễn ngữ cảnh của từng từ trong câu thông qua quá trình “nhìn” cả hai chiều.

PhoBERT giúp nâng cao khả năng xử lý ngôn ngữ tự nhiên cho các ứng dụng tiếng Việt, bao gồm phân loại văn bản, tạo ra các ứng dụng dịch máy và xử lý ngôn ngữ tự nhiên. PhoBERT đã đóng góp đáng kể vào sự phát triển của xử lý ngôn ngữ tự nhiên trong cộng đồng ngôn ngữ tự nhiên Việt Nam và cung cấp một công cụ quan trọng cho nghiên cứu trong lĩnh vực này.

### 3.2.2.4 RoBERTa (Robustly optimized BERT approach)

RoBERTa [49] là một mô hình ngôn ngữ tự nhiên được phát triển dựa trên kiến trúc BERT, nhưng đã trải qua các cải tiến và tối ưu hóa đặc biệt. Mục tiêu của RoBERTa là nâng cao hiệu suất so với BERT thông qua các cải tiến trong quá trình đào tạo và kiến trúc mô hình. Mô hình này đã đạt được nhiều

thành công đáng kể trong nhiều tác vụ xử lý ngôn ngữ tự nhiên, trở thành một trong những mô hình tiêu biểu trong lĩnh vực này.

Trong quá trình đào tạo, RoBERTa sử dụng phương pháp đặt dấu chấm hỏi động, tức là mỗi mẫu dữ liệu sẽ sử dụng một tập hợp ngẫu nhiên các từ để đào tạo. Điều này giúp mô hình học được các biểu diễn tốt hơn cho ngữ cảnh xung quanh từ. RoBERTa cũng áp dụng kỹ thuật Dropout trong quá trình đào tạo, loại bỏ một số lượng ngẫu nhiên các đơn vị, nhằm cải thiện khả năng tổng quát hóa của mô hình.

### 3.2.2.5 XLM-RoBERTa

XLM-RoBERTa [45] (Cross-lingual Robustly Optimized BERT pretraining Approach) là một mô hình ngôn ngữ được phát triển bởi Facebook AI, dựa trên kiến trúc Transformer của BERT nhưng được cải tiến để hỗ trợ đa ngôn ngữ. Điểm nổi bật của XLM-RoBERTa là khả năng xử lý 100 ngôn ngữ khác nhau nhờ vào việc sử dụng một bộ dữ liệu huấn luyện không lò từ CommonCrawl, đảm bảo mô hình có khả năng hiểu và tạo ra văn bản một cách chính xác trong nhiều ngữ cảnh ngôn ngữ.

XLM-RoBERTa được huấn luyện mà không sử dụng bất kỳ thông tin cụ thể về ngôn ngữ nào, điều này giúp nó đạt được hiệu suất vượt trội trong các nhiệm vụ xử lý ngôn ngữ tự nhiên đa ngôn ngữ như phân loại văn bản, dịch ngôn ngữ và trả lời câu hỏi. Các thử nghiệm cho thấy XLM-RoBERTa vượt trội hơn các mô hình trước đó như mBERT và XLM, đặc biệt là trong các bài kiểm tra chuẩn như GLUE, XNLI và MLQA. Điều này làm cho XLM-RoBERTa trở thành một công cụ quan trọng trong nghiên cứu và ứng dụng xử lý ngôn ngữ tự nhiên đa ngôn ngữ.

### 3.2.2.6 CafeBERT

CafeBERT [50] là một mô hình ngôn ngữ tiên tiến được phát triển dựa trên kiến trúc BERT (Bidirectional Encoder Representations from Transformers), tối ưu hóa cho tiếng Việt. CafeBERT được đào tạo trên một tập dữ liệu lớn và đa dạng, giúp cải thiện hiệu suất trong các tác vụ xử lý ngôn ngữ tự nhiên như phân loại văn bản, phân tích cảm xúc, và trả lời câu hỏi. Mô hình này không chỉ cải thiện độ chính xác mà còn giúp giảm thiểu thiên lệch và lỗi

dịch ngữ nghĩa trong ngữ cảnh tiếng Việt, tạo điều kiện thuận lợi cho các ứng dụng AI trong nhiều lĩnh vực khác nhau như truyền thông, giáo dục và thương mại.

### 3.2.2.7 ViSoBERT

ViSoBERT [51] (Vietnamese Social Media BERT) là một mô hình ngôn ngữ dựa trên kiến trúc Transformer, được thiết kế đặc biệt cho tiếng Việt, với mục tiêu cải thiện hiệu suất xử lý ngôn ngữ tự nhiên trên các nền tảng mạng xã hội. Mô hình này được huấn luyện trên tập dữ liệu lớn bao gồm các bài viết và bình luận từ các trang mạng xã hội phổ biến tại Việt Nam, nhằm nắm bắt được các đặc điểm ngôn ngữ, phong cách viết, và từ vựng đặc trưng của người dùng mạng xã hội.

ViSoBERT sử dụng kiến trúc BERT (Bidirectional Encoder Representations from Transformers), một kỹ thuật học sâu cho phép mô hình học được ngữ cảnh của từ dựa trên cả hai phía (trái và phải) của từ đó trong câu. Điều này giúp ViSoBERT có khả năng hiểu ngữ cảnh tốt hơn và xử lý hiệu quả các nhiệm vụ như phân loại văn bản, nhận diện thực thể có tên (NER), và phân tích cảm xúc.

Việc sử dụng ViSoBERT đã chứng minh hiệu quả vượt trội so với các mô hình trước đây khi áp dụng vào các bài toán thực tế liên quan đến tiếng Việt, đặc biệt là trong môi trường ngôn ngữ phong phú và đa dạng của mạng xã hội. Mô hình này mở ra nhiều tiềm năng cho các ứng dụng về phân tích dữ liệu xã hội, quảng cáo, và quản lý cộng đồng trên các nền tảng mạng xã hội.

### 3.2.2.8 viBERT

viBERT [52] (Vietnamese Bidirectional Encoder Representations from Transformers) là một mô hình ngôn ngữ dựa trên kiến trúc BERT, được điều chỉnh riêng cho tiếng Việt. viBERT được huấn luyện trên một lượng lớn dữ liệu văn bản tiếng Việt, bao gồm cả nguồn dữ liệu không có cấu trúc và có cấu trúc như sách, báo và các trang web. Mô hình này đã cải thiện đáng kể hiệu suất trong nhiều tác vụ xử lý ngôn ngữ tự nhiên (NLP) tiếng Việt như phân loại văn bản, nhận diện thực thể có tên, và phân tích cảm xúc.

Công trình này giúp lấp đầy khoảng trống trong nghiên cứu NLP cho tiếng Việt, cung cấp một công cụ mạnh mẽ cho các nhà nghiên cứu và phát triển ứng dụng tiếng Việt. Kết quả thực nghiệm cho thấy viBERT đạt hiệu suất cao, tương đương hoặc vượt trội so với các mô hình trước đó trên nhiều bộ dữ liệu chuẩn, chứng minh khả năng ứng dụng rộng rãi của nó trong thực tiễn.

### 3.2.2.9 DistilBERT

DistilBERT [53] là một mô hình ngôn ngữ tự nhiên được phát triển dựa trên BERT (Bidirectional Encoder Representations from Transformers), nhưng nhẹ hơn và nhanh hơn. Được giới thiệu bởi nhóm nghiên cứu của Hugging Face, DistilBERT sử dụng kỹ thuật "knowledge distillation" để giảm số lượng tham số xuống khoảng 40%, giúp tăng tốc độ xử lý mà vẫn duy trì được 97% hiệu suất của BERT trên các bài kiểm tra ngôn ngữ tự nhiên tiêu chuẩn. Mô hình này phù hợp cho các ứng dụng yêu cầu xử lý văn bản nhanh và hiệu quả trong các môi trường có tài nguyên tính toán hạn chế.

### 3.2.2.10 vELECTRA

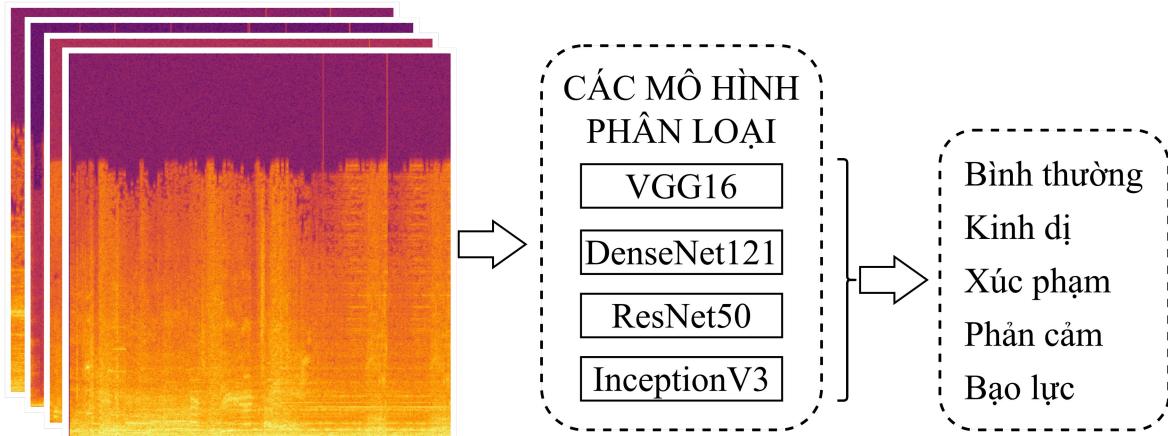
vELECTRA [52] là một mô hình xử lý ngôn ngữ tự nhiên tiên tiến, được phát triển nhằm tối ưu hóa hiệu suất và hiệu quả so với các mô hình trước đó. Dựa trên kiến trúc ELECTRA, vELECTRA sử dụng một cơ chế độc đáo gọi là "discriminator-generator", trong đó mô hình generator tạo ra các từ giả và mô hình discriminator học để phân biệt giữa từ giả và từ thật. Quá trình huấn luyện này giúp vELECTRA trở nên mạnh mẽ hơn trong việc hiểu ngữ cảnh và tạo ra văn bản tự nhiên.

Điểm nổi bật của vELECTRA là khả năng học hỏi hiệu quả từ dữ liệu, giảm thiểu yêu cầu về tài nguyên tính toán so với các mô hình lớn khác, đồng thời đạt được độ chính xác cao trong nhiều tác vụ ngôn ngữ. Điều này làm cho vELECTRA trở thành một công cụ mạnh mẽ và tiết kiệm trong nghiên cứu và ứng dụng trí tuệ nhân tạo.

## 3.2.3 Mô hình phân loại ảnh spectrogram

Phổ âm thanh của các video được biểu diễn dưới dạng các ảnh spectrogram. Các ảnh này tuy nhìn bằng mắt thường khó có thể nhận ra điểm

khác biệt, tuy nhiên lại có khả năng đóng vai trò là đặc trưng để phân loại các âm thanh khác nhau. Chúng tôi đã thử nghiệm khả năng sử dụng đặc trưng là các ảnh biểu diễn phổ âm thanh này để phân loại nội dung của video. Các mô hình mà chúng tôi sử dụng là những mô hình đã được đào tạo trước trên tập dữ liệu lớn hơn. Kết quả chúng tôi thu được khá bất ngờ là những mô hình này lại có khả năng trong việc phân loại video chỉ bằng những hình ảnh phổ âm.



Hình 14 Quy trình phân loại ảnh Spectrogram

### 3.2.3.1 VGG

Các mô hình VGG được đào tạo trước trên bộ dữ liệu ImageNet chúng tôi sử dụng trong nghiên cứu này là VGG16 [33] và VGG19. Đây là các mạng nơ-ron tích chập sâu (CNN) được phát triển bởi nhóm nghiên cứu tại Đại học Oxford. Những mô hình này nổi bật nhờ vào kiến trúc đơn giản nhưng hiệu quả, đạt được kết quả tốt trong các bài toán phân loại ảnh.

#### VGG16:

- Cấu trúc: VGG16 gồm 16 lớp bao gồm 13 lớp tích chập (convolutional) và 3 lớp kết nối đầy đủ (fully connected).
- Chi tiết: Mỗi khối tích chập được sau bởi một lớp gộp tối đa (max pooling) để giảm chiều kích. Các lớp tích chập đều sử dụng bộ lọc kích thước 3x3, giúp tăng độ sâu mà không làm tăng quá nhiều tham số.
- Ưu điểm: Đơn giản, dễ hiểu và triển khai, hiệu suất cao trong nhiều bài toán phân loại ảnh.
- Nhược điểm: Số lượng tham số lớn (138 triệu), yêu cầu tài nguyên tính toán cao.

## VGG19:

- Cấu trúc: VGG19 mở rộng từ VGG16, gồm 19 lớp với 16 lớp tích chập và 3 lớp kết nối dày đặc.
- Chi tiết: VGG19 có thêm ba lớp tích chập so với VGG16, giúp tăng độ sâu và khả năng biểu diễn của mô hình.
- Ưu điểm: Cải thiện khả năng học tập và nhận dạng đối với các chi tiết phức tạp hơn.
- Nhược điểm: Tăng thêm số lượng tham số và tài nguyên tính toán so với VGG16.

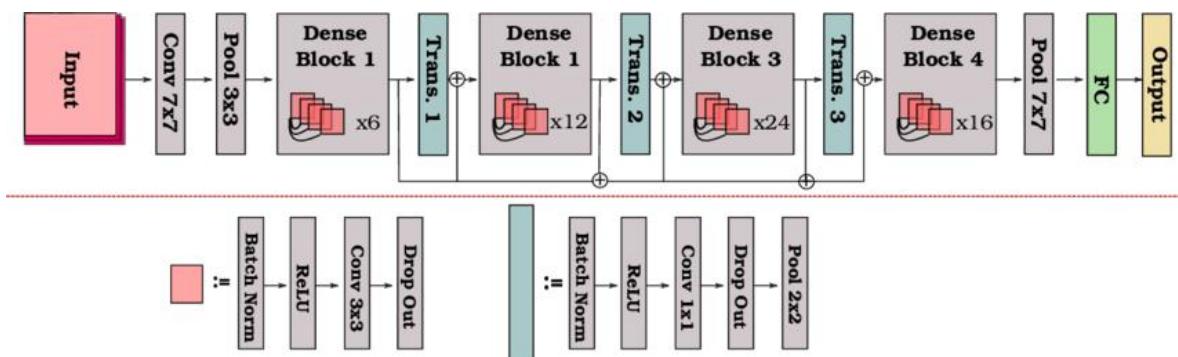
### 3.2.3.2 DenseNet

Các mô hình DenseNet được đào tạo trước trên bộ dữ liệu ImageNet chúng tôi sử dụng trong nghiên cứu này là DenseNet121, DenseNet169 và DenseNet201

**DenseNet121** [18] là một kiến trúc mạng nơ-ron tích chập kết nối từng lớp với mọi lớp khác theo kiểu chuyển tiếp. Nó có nhiều ưu điểm khác nhau, bao gồm giải quyết vấn đề biến mất độ dốc, cải thiện việc truyền bá tính năng, khuyến khích tái sử dụng tính năng và giảm đáng kể số lượng của các tham số.

$$x \rightarrow [x, f_1(x), f_2(x, f_1(x)), f_3(x, f_1(x), f_2(x, f_1(x))), \dots]$$

Công thức này mô tả hoạt động của kết nối dày đặc trong kiến trúc DenseNet. Trong DenseNet, đầu ra của mỗi lớp được kết nối với đầu vào của tất cả các lớp tiếp theo. Điều này có nghĩa là đầu ra của một lớp không chỉ được truyền cho lớp tiếp theo mà còn cho tất cả các lớp tiếp theo.



Hình 15 Kiến trúc DenseNet121 [8]

**DenseNet169** là một biến thể khác trong dòng mô hình DenseNet, với tổng cộng 169 lớp. Giống như DenseNet121, nó cũng sử dụng các khối kết nối dày đặc để đảm bảo luồng thông tin mượt mà giữa các lớp. Với số lượng lớp nhiều hơn, DenseNet169 có khả năng học được các đặc trưng phức tạp hơn từ dữ liệu đầu vào, điều này đặc biệt hữu ích trong các tác vụ yêu cầu độ chính xác cao như phân loại hình ảnh chi tiết và nhận dạng đối tượng trong các bộ dữ liệu lớn. Tuy nhiên, việc tăng số lớp cũng đồng nghĩa với việc yêu cầu tài nguyên tính toán và bộ nhớ cao hơn.

**DenseNet201** DenseNet201 là phiên bản sâu hơn trong các mô hình DenseNet, với 201 lớp. Mô hình này tiếp tục kế thừa kiến trúc kết nối dày đặc đặc trưng của DenseNet, giúp tối ưu hóa quá trình học sâu và duy trì sự nhất quán trong việc truyền tải gradient. DenseNet201 có khả năng biểu diễn các đặc trưng phức tạp từ dữ liệu, làm cho nó trở thành lựa chọn lý tưởng cho các ứng dụng yêu cầu mức độ chi tiết và độ chính xác cực cao, chẳng hạn như phân loại hình ảnh trong y tế hoặc nhận diện khuôn mặt. Mặc dù mạnh mẽ, nhưng DenseNet201 cũng đòi hỏi tài nguyên tính toán lớn và thời gian huấn luyện lâu hơn so với các mô hình DenseNet có số lớp ít hơn.

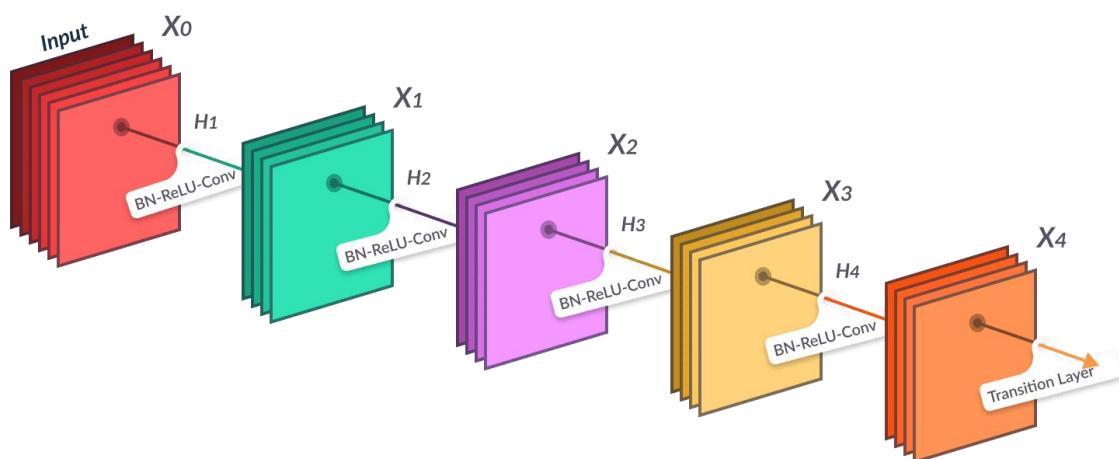
### 3.2.3.3 ResNet

Các mô hình ResNet được đào tạo trước trên bộ dữ liệu ImageNet chúng tôi sử dụng trong nghiên cứu này là ResNet50 và ResNet50V2

**ResNet50** là một mô hình mạng nơ-ron sâu nổi bật trong lĩnh vực nhận dạng hình ảnh và học sâu. Được giới thiệu bởi Kaiming He và cộng sự vào năm 2015 [16], ResNet50 bao gồm 50 lớp, sử dụng kiến trúc Residual Network (ResNet) để giải quyết vấn đề tiêu biến gradient khi độ sâu mạng tăng lên. Điểm nổi bật của ResNet50 là các khối residual, cho phép các lớp học hỏi sự khác biệt giữa đầu vào và đầu ra của chúng thay vì học trực tiếp các biểu diễn đầu ra. Điều này giúp duy trì hiệu quả huấn luyện ngay cả với các mạng rất sâu. ResNet50 đã đạt được nhiều kết quả ấn tượng trong các thử thách nhận dạng hình ảnh, đặc biệt là trong cuộc thi ImageNet Large Scale Visual Recognition Challenge (ILSVRC)

**ResNet50V2** là phiên bản cải tiến của ResNet50, được giới thiệu với mục tiêu tối ưu hóa hơn nữa khả năng huấn luyện và hiệu suất của mạng. Trong ResNet50V2, các khối residual được tái thiết kế để hoán đổi thứ tự của phép cộng và lớp chuẩn hóa (batch normalization), giúp cải thiện tính ổn định trong quá trình huấn luyện. Cấu trúc này cũng giảm thiểu hiện tượng tiêu biến gradient, nhờ đó mô hình có thể đạt được độ chính xác cao hơn với cùng số lớp. ResNet50V2 thường được áp dụng trong các ứng dụng nhận dạng và phân loại hình ảnh yêu cầu độ chính xác cao và hiệu quả huấn luyện mạnh mẽ, đồng thời nó đã chứng minh hiệu quả vượt trội trong nhiều bài kiểm tra và thách thức khác nhau. Trong các mạng thần kinh truyền thống, mỗi lớp có gắng tìm hiểu ánh xạ cơ bản của dữ liệu đầu vào, việc này có thể ngày càng trở nên khó khăn khi mạng ngày càng sâu hơn. Ngược lại, ResNet học ánh xạ dư bằng cách tìm hiểu sự khác biệt giữa đầu vào và đầu ra của mỗi lớp. Điều này giúp mạng dễ dàng ước tính ánh xạ mong muốn hơn và cho phép đào tạo thành công các mạng rất sâu.

Kiến trúc ResNet thường bao gồm một số khối dư, trong đó mỗi khối bao gồm nhiều lớp tích chập, sau đó là các hàm chuẩn hóa hàng loạt và kích hoạt phi tuyến tính như ReLU. Việc bỏ qua các kết nối trong ResNet cho phép các gradient di chuyển dễ dàng hơn trong quá trình đào tạo, cho phép mạng học hiệu quả hơn và đạt được độ chính xác cao hơn.



Hình 16 Kiến trúc ResNet [3]

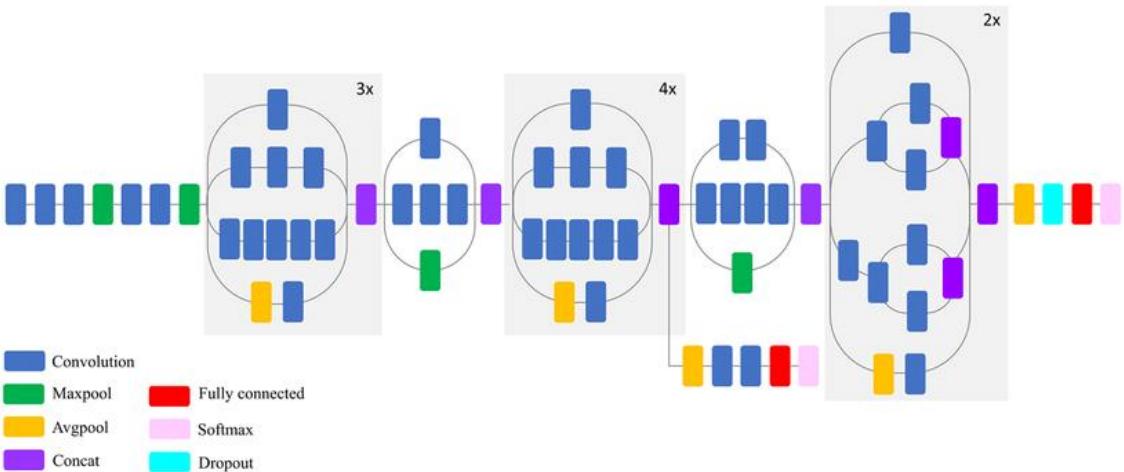
### 3.2.3.4 Inception

Các mô hình Inception được đào tạo trước trên bộ dữ liệu ImageNet chúng tôi sử dụng trong nghiên cứu này là InceptionV3 và InceptionResNetV2

**Mô hình InceptionV3** [35] được cập nhật từ mô hình InceptionV1. Nó tối ưu hóa mạng bằng nhiều phương pháp khác nhau nhằm nâng cao khả năng thích ứng của mô hình. InceptionV3 có hiệu năng tốt trên nhiều tác vụ thị giác máy tính, bao gồm nhận diện đối tượng, phân loại ảnh, và nhiều ứng dụng khác trong lĩnh vực thị giác máy tính.

InceptionV3 với kiến trúc “Inception”, sử dụng các Inception module. Mỗi module kết hợp các loại nhân tích chập có kích thước khác nhau để hiệu quả hóa quá trình học đặc trưng. Kiến trúc InceptionV3 được minh họa trong hình 3-8. Các thành phần chính của mô hình bao gồm:

- Inception Module gồm các lớp tích chập với các nhân có kích thước khác nhau để trích xuất đặc trưng ở các tỷ lệ không gian khác nhau và các lớp pooling (max pooling và average pooling) để giảm kích thước không gian và giữ lại thông tin quan trọng.
- Inception Blocks được tạo thành bởi các Inception module được xếp chồng lên nhau. Các block này giúp mô hình học được đặc trưng ở các mức độ phức tạp và ảnh hưởng đến nhiều tỷ lệ không gian.
- Reduction Blocks được sử dụng để giảm kích thước không gian của dữ liệu giữa các Inception blocks, giúp tăng tốc quá trình huấn luyện và giảm độ phức tạp của mô hình.



Hình 17 Minh họa kiến trúc InceptionV3

**Inception-ResNet-v2** đạt được hiệu suất tốt trên nhiều tập dữ liệu và thường được sử dụng trong các cuộc thi và ứng dụng thực tế về thị giác máy tính. Nó được đề xuất bởi Christian Szegedy và cộng sự trong bài báo "Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning" vào năm 2016 [40].

Inception-ResNet-v2 là một kiến trúc mạng nơ-ron sâu (deep neural network) được thiết kế để phân loại ảnh. Inception-ResNet-v2 là sự kết hợp giữa hai kiến trúc Inception và ResNet nhằm tận dụng ưu điểm của cả hai kiến trúc này.

Kiến trúc của Inception-ResNet-v2 sử dụng mô-đun Inception (hoặc Inception-v4), nó sử dụng các convolutional layer với các kích thước kernel khác nhau cùng một lúc, giúp mô hình học được các đặc trưng ở nhiều mức độ phức tạp. Inception-ResNet-v2 còn tích hợp các residual connections từ ResNet, các residual connections giúp giảm vấn đề biến mất đạo hàm và làm cho việc huấn luyện mô hình trở nên dễ dàng hơn, đặc biệt là với các mô hình có số lượng layer lớn. Sự kết hợp của những đặc điểm này giúp Inception-ResNet-v2 đạt được hiệu suất tốt trên nhiều tác vụ thị giác máy tính, đặc biệt là trong việc phân loại hình ảnh trên các tập dữ liệu lớn.

### 3.2.3.5 Xception

Xception (Extreme Inception) là một mô hình mạng nơ-ron sâu được giới thiệu bởi François Chollet vào năm 2016 [54]. Mô hình này là sự mở rộng

của kiến trúc Inception, với cải tiến chính là việc thay thế các khối Inception bằng các khối Separable Convolution (tách biệt và chập). Xception được xây dựng dựa trên giả thuyết rằng sự phân tách không gian và kênh trong phép chập có thể thực hiện tốt hơn và hiệu quả hơn so với các phép chập thông thường.

Cụ thể, Xception sử dụng các khối Depthwise Separable Convolution, nơi các phép chập không gian và chập kênh được thực hiện riêng biệt. Điều này giúp giảm số lượng tham số và tính toán cần thiết, đồng thời tăng cường khả năng biểu diễn của mô hình. Kiến trúc Xception bao gồm 36 khối convolutional chính, theo sau bởi các lớp fully-connected, và đã đạt được nhiều thành công trong các bài toán nhận dạng hình ảnh lớn, bao gồm cả trên bộ dữ liệu ImageNet. Xception thể hiện hiệu suất vượt trội so với nhiều mô hình tiền nhiệm, đặc biệt là trong việc cân bằng giữa độ chính xác và hiệu quả tính toán.

### 3.2.3.6 NASNetMobile

NASNetMobile [55] là một mô hình mạng nơ-ron sâu được thiết kế bởi Google AI thông qua quá trình Neural Architecture Search (NAS), một phương pháp tự động hóa việc tìm kiếm kiến trúc mạng tối ưu. Được giới thiệu vào năm 2017, NASNetMobile nhằm cung cấp hiệu suất cao trong nhận dạng hình ảnh với yêu cầu tài nguyên tính toán thấp, phù hợp cho các thiết bị di động và nhúng.

Kiến trúc NASNetMobile được tối ưu hóa bằng cách sử dụng các khối cell (cell blocks) tự tìm kiếm và điều chỉnh trong quá trình huấn luyện, giúp tạo ra một mạng lưới hiệu quả và mạnh mẽ. Các khối cell này được chia thành hai loại: Normal Cell và Reduction Cell, với chức năng khác nhau trong việc duy trì và giảm kích thước đặc trưng của đầu vào. NASNetMobile sử dụng các cell này theo một cấu trúc phân cấp, từ đó tạo ra một mạng nơ-ron sâu với khả năng tổng quát hóa tốt và độ chính xác cao.

Một trong những ưu điểm nổi bật của NASNetMobile là khả năng cân bằng giữa độ chính xác và hiệu quả tính toán. Mô hình này đã đạt được kết quả ấn tượng trên nhiều bài toán nhận dạng hình ảnh, đặc biệt là trên bộ dữ liệu ImageNet, đồng thời tiêu tốn ít tài nguyên hơn so với nhiều mô hình truyền thống khác. Điều này khiến NASNetMobile trở thành lựa chọn lý tưởng cho

các ứng dụng trên thiết bị di động và các hệ thống nhúng có hạn chế về tài nguyên.

### 3.2.3.7 ConvNeXt

ConvNeXt [56] là một dòng mô hình mạng nơ-ron tích chập (CNN) tiên tiến được phát triển nhằm cạnh tranh với các mô hình Transformers trong các nhiệm vụ nhận dạng hình ảnh. Các mô hình ConvNeXt được đào tạo trước trên bộ dữ liệu ImageNet chúng tôi sử dụng trong nghiên cứu này là ConvNeXtTiny, ConvNeXtSmall, ConvNeXtBase và ConvNeXtLarge

**ConvNeXtTiny** là phiên bản nhỏ nhất trong dòng mô hình ConvNeXt. Với một cấu trúc gọn nhẹ, ConvNeXtTiny bao gồm số lượng lớp ít hơn và ít tham số hơn so với các phiên bản lớn hơn, giúp nó trở thành một lựa chọn phù hợp cho các ứng dụng đòi hỏi hiệu quả tính toán cao và chi phí tài nguyên thấp. Mặc dù kích thước nhỏ, ConvNeXtTiny vẫn duy trì được hiệu suất nhận dạng hình ảnh ấn tượng nhờ vào việc sử dụng các kỹ thuật tiên tiến trong kiến trúc mạng.

**ConvNeXtSmall** tăng cường thêm một số lớp và tham số so với ConvNeXtTiny, cho phép nó xử lý tốt hơn các nhiệm vụ phức tạp hơn mà vẫn giữ được hiệu suất tính toán tương đối cao. Sự cân bằng giữa kích thước mô hình và hiệu suất giúp ConvNeXtSmall trở thành một lựa chọn lý tưởng cho các ứng dụng yêu cầu mức độ chính xác cao hơn mà không đòi hỏi quá nhiều tài nguyên phần cứng.

**ConvNeXtBase** là phiên bản trung bình trong dòng mô hình ConvNeXt, được thiết kế để cung cấp hiệu suất cao cho các nhiệm vụ nhận dạng hình ảnh phức tạp. Với số lượng lớp và tham số tăng đáng kể so với ConvNeXtSmall, ConvNeXtBase có khả năng học hỏi và nhận diện các đặc điểm phức tạp hơn trong dữ liệu hình ảnh, giúp nó đạt được kết quả tốt hơn trong các bài toán khó.

**ConvNeXtLarge** là phiên bản lớn nhất và mạnh mẽ nhất trong dòng mô hình ConvNeXt. Với số lượng lớp và tham số lớn nhất, ConvNeXtLarge được thiết kế để tối đa hóa hiệu suất trên các nhiệm vụ nhận dạng hình ảnh phức tạp nhất. Mô hình này phù hợp với các ứng dụng yêu cầu độ chính xác cao nhất và

có khả năng sử dụng nhiều tài nguyên tính toán, chẳng hạn như trong các hệ thống nhận dạng hình ảnh công nghiệp và nghiên cứu khoa học chuyên sâu.

Tất cả các phiên bản ConvNeXt đều được thiết kế để tận dụng những ưu điểm của CNN truyền thống và cải tiến chúng để đạt được hiệu suất cạnh tranh với các mô hình học sâu hiện đại như Transformers, đồng thời giữ vững các lợi thế về tính hiệu quả và khả năng mở rộng của CNN.

### 3.2.3.8 EfficientNet

Các mô hình EfficientNet được đào tạo trước trên bộ dữ liệu ImageNet chúng tôi sử dụng trong nghiên cứu này là EfficientNetB0, EfficientNetB1, EfficientNetB2, EfficientNetB3, EfficientNetB4, EfficientNetB7, EfficientNetV2M và EfficientNetV2L

EfficientNet là một họ các mô hình học sâu được phát triển bởi Google, nổi bật với khả năng mở rộng hiệu quả. Dưới đây là tóm tắt về các phiên bản chính của EfficientNet và EfficientNetV2:

- **EfficientNetB0:**

- EfficientNetB0 là phiên bản cơ sở của họ EfficientNet, được phát triển sử dụng kỹ thuật Neural Architecture Search (NAS).
  - Nó sử dụng kiến trúc MBConv và một cách tiếp cận mới gọi là "compound scaling", kết hợp tăng cường đồng thời chiều sâu, chiều rộng và độ phân giải của mạng.

- **EfficientNetB1, EfficientNetB2, EfficientNetB3, EfficientNetB4, EfficientNetB5, EfficientNetB6, EfficientNetB7:**

- Các phiên bản này là các mở rộng của EfficientNetB0 với các tham số mở rộng khác nhau, từ B1 đến B7.

- Mỗi phiên bản này tăng dần về kích thước và số lượng tham số, nhằm cân bằng giữa hiệu suất và độ phức tạp tính toán.

- EfficientNetB7 là phiên bản lớn nhất và phức tạp nhất trong nhóm này, với hiệu suất cao nhất trên các tác vụ nhận dạng hình ảnh.

- **EfficientNetV2M và EfficientNetV2L:**

- EfficientNetV2 là phiên bản cải tiến của EfficientNet với mục tiêu tối ưu hóa tốc độ huấn luyện và hiệu suất mô hình.

- EfficientNetV2M và EfficientNetV2L là các phiên bản mở rộng của EfficientNetV2, với M (Medium) và L (Large) biểu thị kích thước và độ phức tạp tăng dần.
- Các phiên bản V2 này cải tiến thuật toán MBConv và giới thiệu Fused-MBConv, giúp cải thiện tốc độ huấn luyện và độ chính xác trên các tập dữ liệu lớn.

EfficientNet và EfficientNetV2 đều thể hiện khả năng tối ưu hóa tài nguyên tính toán và hiệu suất cao trên nhiều tác vụ thị giác máy tính, làm cho chúng trở thành lựa chọn phổ biến trong các ứng dụng học sâu hiện đại.

### **3.2.4 Kết hợp các mô hình sử dụng phương pháp ensemble**

Kết hợp các mô hình học máy bằng phương pháp ensemble là một kỹ thuật nhằm tận dụng sự đa dạng và sự khác biệt giữa các mô hình để cải thiện hiệu suất và độ chính xác của dự đoán. Thay vì sử dụng một mô hình duy nhất, ensemble sử dụng nhiều mô hình và kết hợp kết quả từ chúng để đưa ra dự đoán cuối cùng. Có nhiều phương pháp kết hợp các mô hình trong ensemble, nhưng hai phương pháp phổ biến nhất là Voting và Bagging:

- Voting: Phương pháp Voting kết hợp các dự đoán từ các mô hình thành viên bằng cách sử dụng đa số phiếu. Có ba loại voting phổ biến là Majority Voting, Weighted Voting và Soft Voting. Majority Voting đưa ra dự đoán cuối cùng dựa trên dự đoán được chọn nhiều nhất từ các mô hình. Weighted Voting gán trọng số cho mỗi mô hình và tính toán dự đoán cuối cùng dựa trên trọng số. Soft Voting lấy trung bình xác suất dự đoán từ mỗi mô hình và dự đoán cuối cùng dựa trên các xác suất đó.
- Bagging: Phương pháp Bagging tạo ra nhiều mô hình độc lập bằng cách huấn luyện trên các tập dữ liệu con được rút ra từ tập dữ liệu huấn luyện gốc. Các dự đoán từ các mô hình thành viên được kết hợp để đưa ra dự đoán cuối cùng. Phương pháp Bagging thường được sử dụng với các mô hình quyết định (decision trees) và có thể tăng khả năng tổng quát của ensemble.

Ứng dụng ensemble rộng rãi và hiệu quả trong nhiều trường hợp vì:

- **Đa dạng hóa:** Ensemble tận dụng sự đa dạng giữa các mô hình thành viên. Các mô hình trong ensemble có thể học các khía cạnh khác nhau của dữ liệu và có những kiến thức riêng. Khi kết hợp lại, ensemble có khả năng tổng quát hóa tốt hơn và giảm thiểu tác động của nhiễu và overfitting.
- **Cải thiện hiệu suất:** Ensemble thường cho kết quả tốt hơn so với một mô hình đơn lẻ. Khi các mô hình độc lập được kết hợp lại, các sai sót của mỗi mô hình có thể được bù đắp và dự đoán cuối cùng trở nên chính xác hơn.
- **Ôn định:** Ensemble cũng có khả năng ổn định hơn. Với việc sử dụng nhiều mô hình, việc một mô hình không hoạt động tốt trên một số trường hợp cụ thể không ảnh hưởng nhiều đến kết quả tổng thể.

Tuy nhiên, việc xây dựng và huấn luyện ensemble có thể tốn tài nguyên tính toán và thời gian hơn so với một mô hình đơn lẻ. Cần lựa chọn các mô hình thành viên đa dạng và tối ưu các phương pháp kết hợp để đạt được kết quả tốt nhất từ ensemble.

## CHƯƠNG 4: THỰC NGHIỆM VÀ ĐÁNH GIÁ

Chương này trình bày quá trình cài đặt thực nghiệm với bộ dữ liệu thu thập và đánh giá chi tiết kết quả đạt được trên mô hình được sử dụng.

### 4.1. Thiết kế thí nghiệm

Chúng tôi đã thực hiện tổng cộng 38 thực nghiệm trên 3 đặc trưng là các frame ảnh của video, các hình ảnh phổ âm thanh spectrogram và văn bản trích xuất từ giọng nói của video. Trong đó có 5 mô hình phân loại video, 23 mô hình huấn luyện trước để phân loại ảnh spectrogram, và 10 mô hình phân loại văn bản. Chúng tôi sử dụng hàm tối ưu Adam với learning\_rate = 0.0001, kích thước ảnh là 224x224 và 3 kênh (RGB). Với các mô hình phân loại video, chúng tôi chọn số lượng khung hình cho mỗi video 5 giây là 10, huấn luyện với 50 epochs có sử dụng phương pháp dừng sớm với patience = 3. Hàm mất mát chúng tôi sử dụng là Sparse Categorical Cross Entropy (Số học phân loại chéo thưa thớt) là một hàm mất mát thường được sử dụng trong các bài toán phân loại đa lớp. Hàm mất mát này thích hợp khi số lượng lớp (classes) là lớn và các nhãn (labels) được biểu diễn dưới dạng chỉ mục (index) của lớp thay vì dạng mã hóa one-hot đầy đủ. Điều này có ý nghĩa là đầu ra dự đoán và nhãn thực tế của mỗi mẫu chỉ là một số nguyên thay vì một vectơ one-hot. Công thức tính toán hàm mất mát Sparse Categorical Cross Entropy cho một mẫu dữ liệu cụ thể là:

$$L_{CE} = - \sum_{i=1}^n t_i \log(p_i), \text{cho } n \text{ nhãn}$$

Trong đó:

- $t_i$  là nhãn đúng và  $p_i$  là xác suất Softmax cho nhãn thứ i
- Hàm softmax thường được sử dụng trong mạng nơ-ron để chuyển đổi đầu ra của mạng thành một phân phối xác suất trên các lớp hoặc các nhãn. Cho một vectơ đầu vào z có các phần tử  $z_1, z_2, \dots, z_k$ , hàm softmax tính toán giá trị softmax cho mỗi phần tử  $z_i$  bằng cách sử dụng công thức:

$$\text{Softmax}(z_i) = \frac{e^{z_i}}{\sum_{j=1}^k e^{z_j}}$$

Hàm mất mát Sparse Categorical Cross Entropy tính toán sự sai khác giữa log xác suất của lớp dự đoán và log xác suất của lớp thực tế. Mục tiêu là tối thiểu hóa sự sai khác này để mô hình học cách dự đoán chính xác lớp thực tế của các mẫu dữ liệu.

Ngoài ra, theo các nghiên cứu trước đó về phân loại video, phân loại văn bản và phân loại hình ảnh, chúng tôi sử dụng 4 độ đo chính để đánh giá hiệu suất của mô hình, bao gồm:

- Accuracy (Độ chính xác): Độ chính xác là một độ đo đơn giản nhưng quan trọng để đánh giá khả năng phân loại của mô hình. Nó đo lường tỷ lệ dự đoán chính xác so với tổng số mẫu dữ liệu.

$$\text{Accuracy} = (\text{Số lượng dự đoán chính xác}) / (\text{Tổng số mẫu dữ liệu})$$

- F1-Score: F1-Score là một độ đo kết hợp của precision và recall, thường được sử dụng khi cả precision và recall đều quan trọng. Nó là trung bình điều hòa của precision và recall, mang giá trị từ 0 đến 1, với 1 là giá trị tốt nhất. Công thức tính F1-Score là:

$$\text{F1-Score} = 2 * (\text{Precision} * \text{Recall}) / (\text{Precision} + \text{Recall})$$

- Precision (Độ chính xác dương tính): Precision đo lường tỷ lệ các dự đoán dương tính chính xác so với tổng số các dự đoán dương tính. Precision là một chỉ số về độ chính xác của các dự đoán dương tính và đo lường khả năng của mô hình phân loại xác định đúng các mẫu dương tính.

$$\text{Precision} = (\text{Số lượng dự đoán dương tính chính xác}) / (\text{Tổng số dự đoán dương tính})$$

- Recall (Độ bao phủ hoặc độ phục hồi): Recall đo lường tỷ lệ các dự đoán dương tính chính xác so với tổng số các mẫu thực sự là dương tính. Recall là một chỉ số về khả năng phát hiện các mẫu thực sự dương tính và đo lường khả năng của mô hình phân loại bao phủ đúng các mẫu dương tính.

$$\text{Recall} = (\text{Số lượng dự đoán dương tính chính xác}) / (\text{Tổng số mẫu dương tính})$$

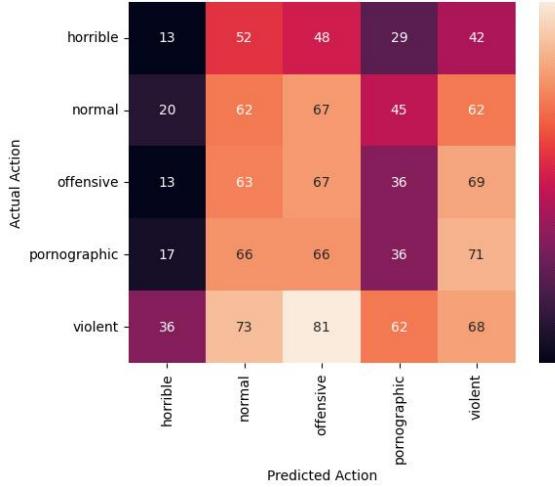
Các độ đo Accuracy, F1-Score, Precision và Recall đóng vai trò quan trọng trong đánh giá hiệu suất của các mô hình phân loại. Accuracy đo lường tỷ lệ dự đoán chính xác, trong khi F1-Score cung cấp một phép đo kết hợp của cả Precision và Recall. Precision tập trung vào độ chính xác của các dự đoán dương tính, trong khi Recall tập trung vào khả năng phát hiện các mẫu thực sự dương tính. Các kết quả chúng tôi thu được bao gồm bảng kết quả sử dụng 4 độ đo, ma trận nhầm lẫn và đồ thị học với độ chính xác (Accuracy) và hàm mất mát (Loss) được trình bày trong phần 4.2 dưới đây.

## 4.2. Kết quả

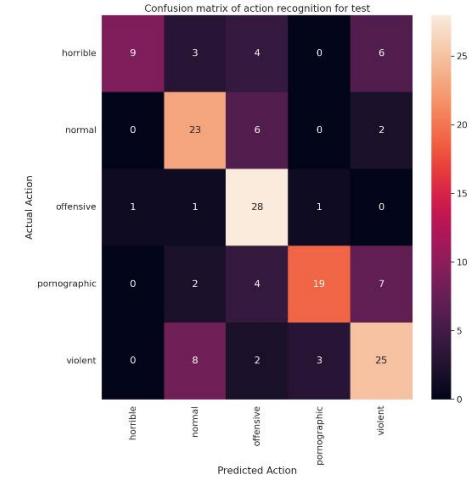
### 4.2.1 Kết quả của các mô hình dự đoán với các frames ảnh của video

Dựa vào các ma trận nhầm lẫn ta thấy rằng các mô hình 3D CNN với các trọng số được đào tạo trước trên bộ dữ liệu ImageNet không có khả năng phân loại các nhãn của video dựa trên các khung ảnh (các kết quả dự đoán chỉ dự đoán ra một nhãn duy nhất). Bộ dữ liệu ImageNet chứa hình ảnh và các nhãn cho hàng triệu hình ảnh từ hàng ngàn đối tượng khác nhau bao gồm động vật, đồ vật, cảnh quan, con người,... nên các trọng số được huấn luyện từ bộ dữ liệu này không phù hợp để dự đoán nhãn của video dựa trên nội dung có sử dụng ngôn ngữ tiếng Việt. Mặc khác, mô hình 3D CNN được train từ đầu đã cho thấy khả năng của nó trong việc nắm bắt đặc trưng không gian và thời gian của các khung hình, từ đó đưa ra kết quả dự đoán chính xác hơn, độ chính xác chúng tôi thu được là 65,58% trong khi độ chính xác trong bài báo gốc là 68,9%). Tuy nhiên, để đào tạo kiến trúc mô hình 3D CNN từ đầu đòi hỏi rất nhiều thời gian huấn luyện cũng như chi phí tính toán so với mô hình 3D CNN được đào tạo trước.

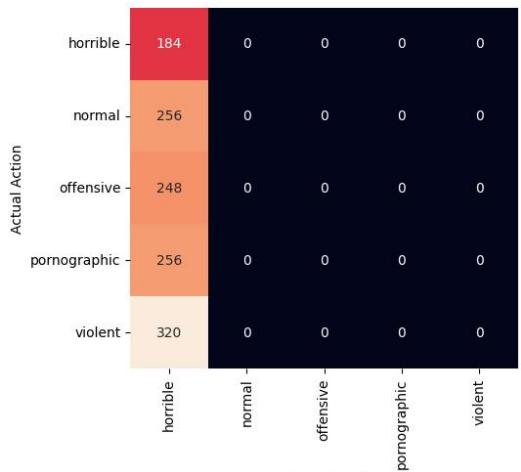
Học chuyên tiếp với MoViNet trên bộ dữ liệu Kinetics thu được kết quả dự đoán chính xác (độ chính xác 80,77%) sau khi đào tạo với 2 epochs. Cho thấy rằng các mô hình được đào tạo trước trên bộ dữ liệu Kinetics này có khả



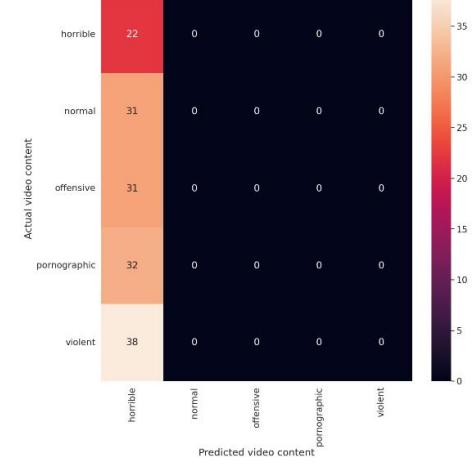
Hình 18 Ma trận nhầm lẫn của mô hình  
ResNet đối với tập train



Hình 20 Ma trận nhầm lẫn của mô hình  
ResNet đối với tập test

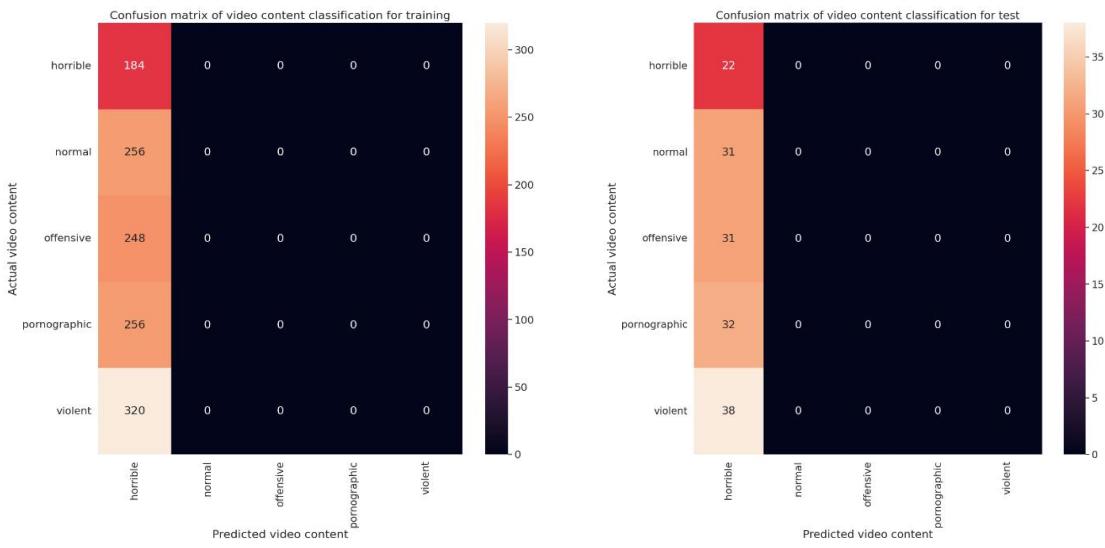


Hình 19 Ma trận nhầm lẫn của mô hình  
pre-trained ResNet50 đối với tập train



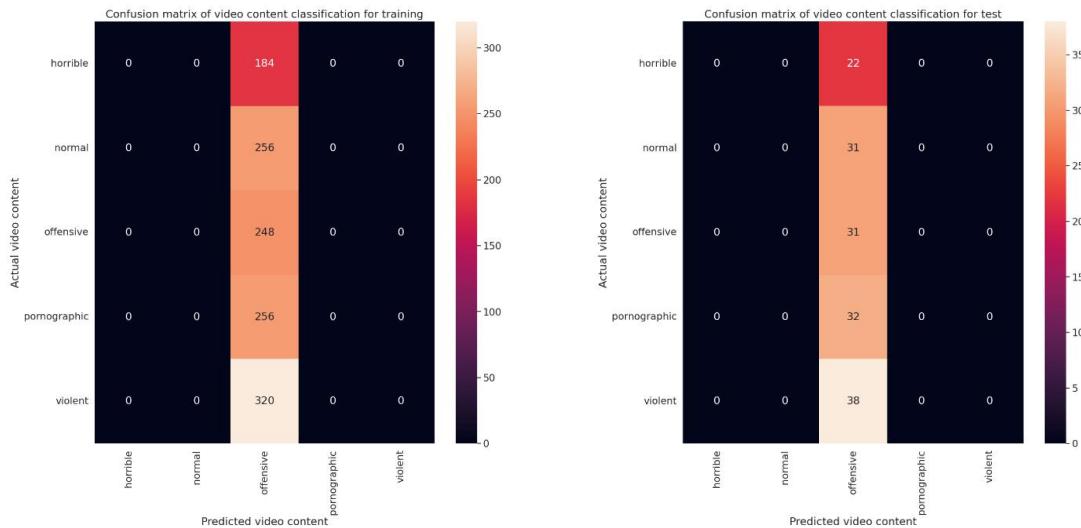
Hình 21 Ma trận nhầm lẫn của mô hình  
pre-trained ResNet50 đối với tập test

năng ứng dụng cao trong việc phân loại video một cách đáng tin cậy, nhanh chóng và tiết kiệm chi phí.



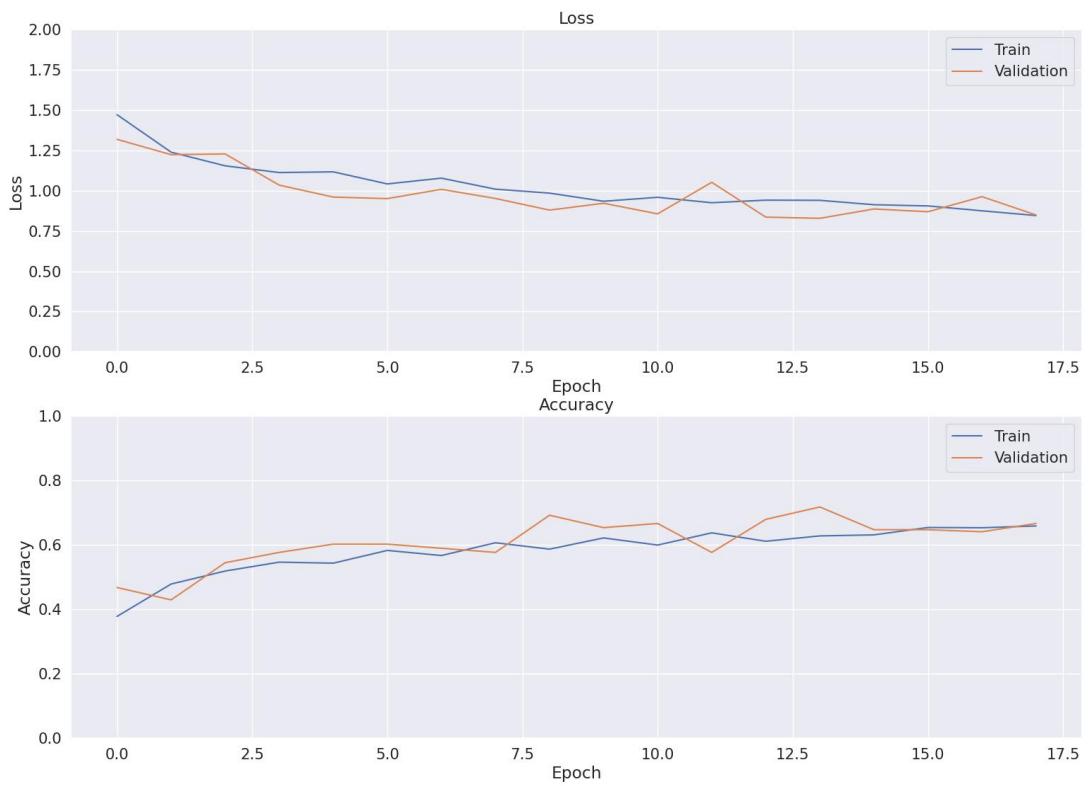
Hình 22 Ma trận nhầm lẫn của mô hình pre-trained Inception-ResNet-v2 đối với tập train

Hình 23 Ma trận nhầm lẫn của mô hình pre-trained Inception-ResNet-v2 đối với tập test

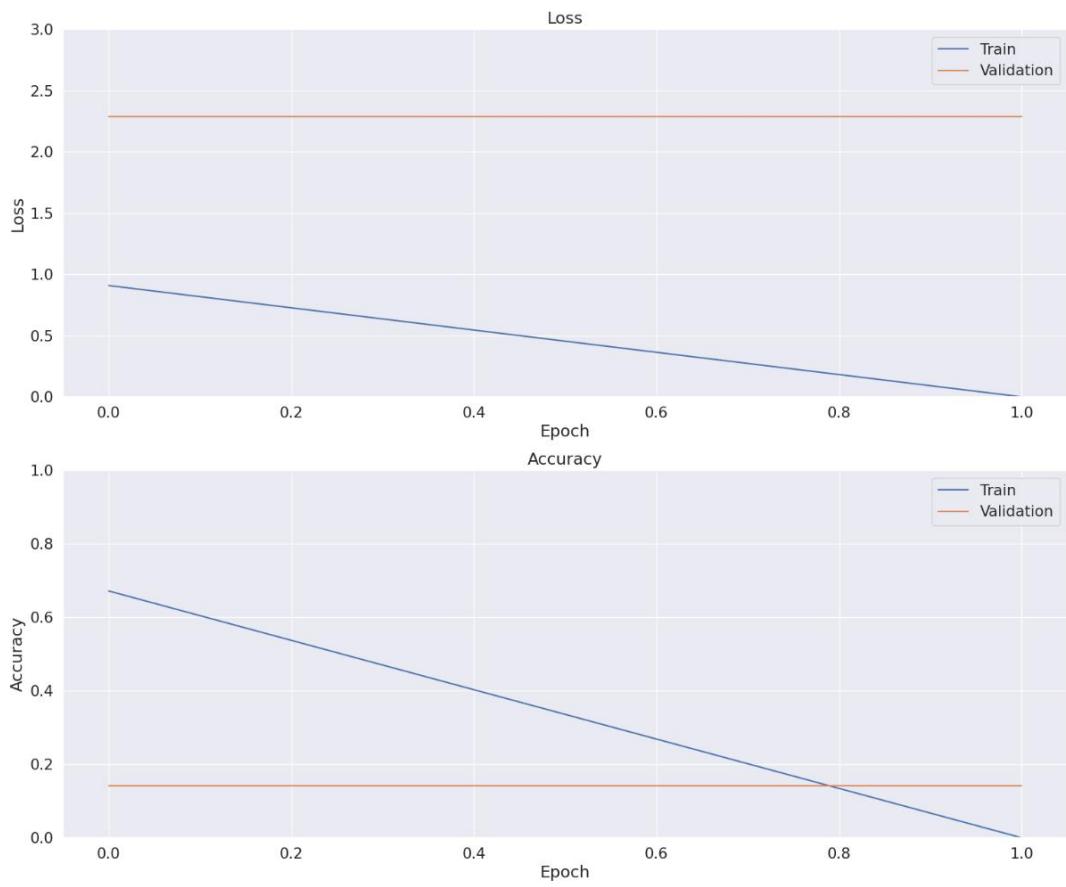


Hình 24 Ma trận nhầm lẫn của mô hình pre-trained InceptionV3 đối với tập train

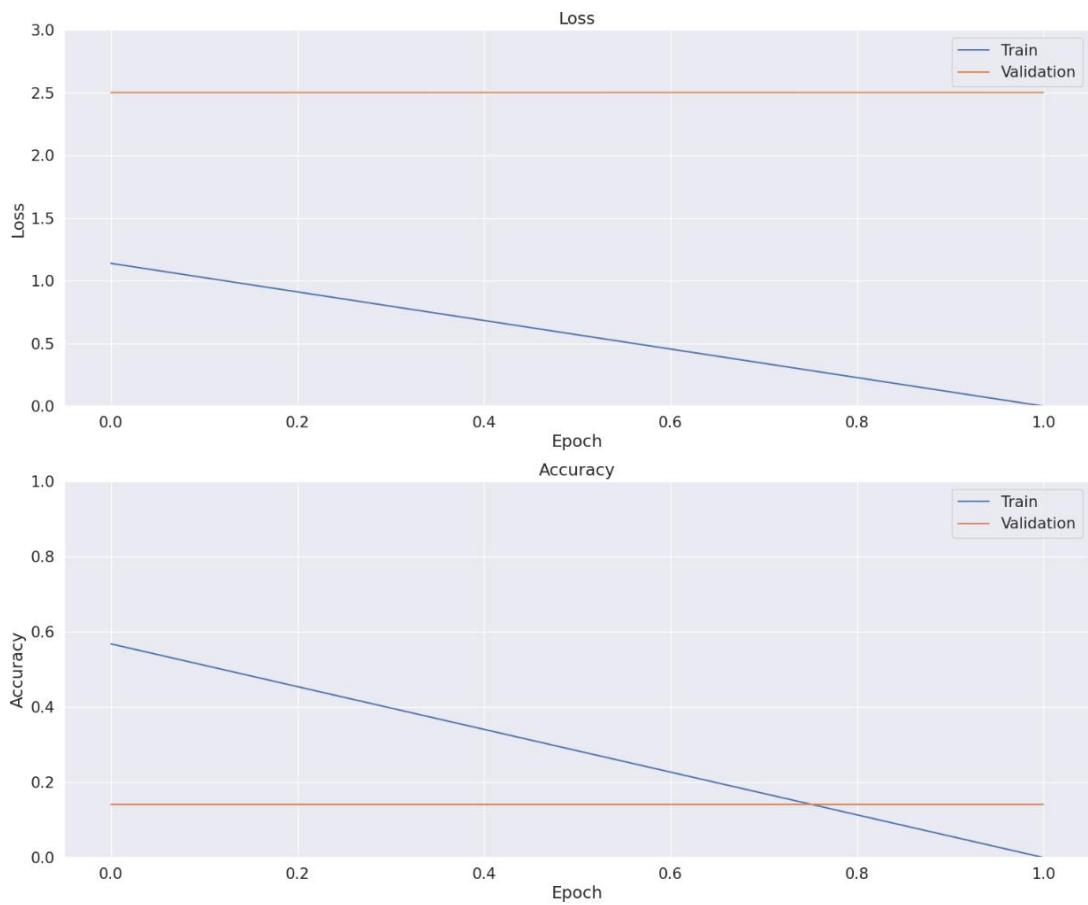
Hình 25 Ma trận nhầm lẫn của mô hình pre-trained InceptionV3 đối với tập test



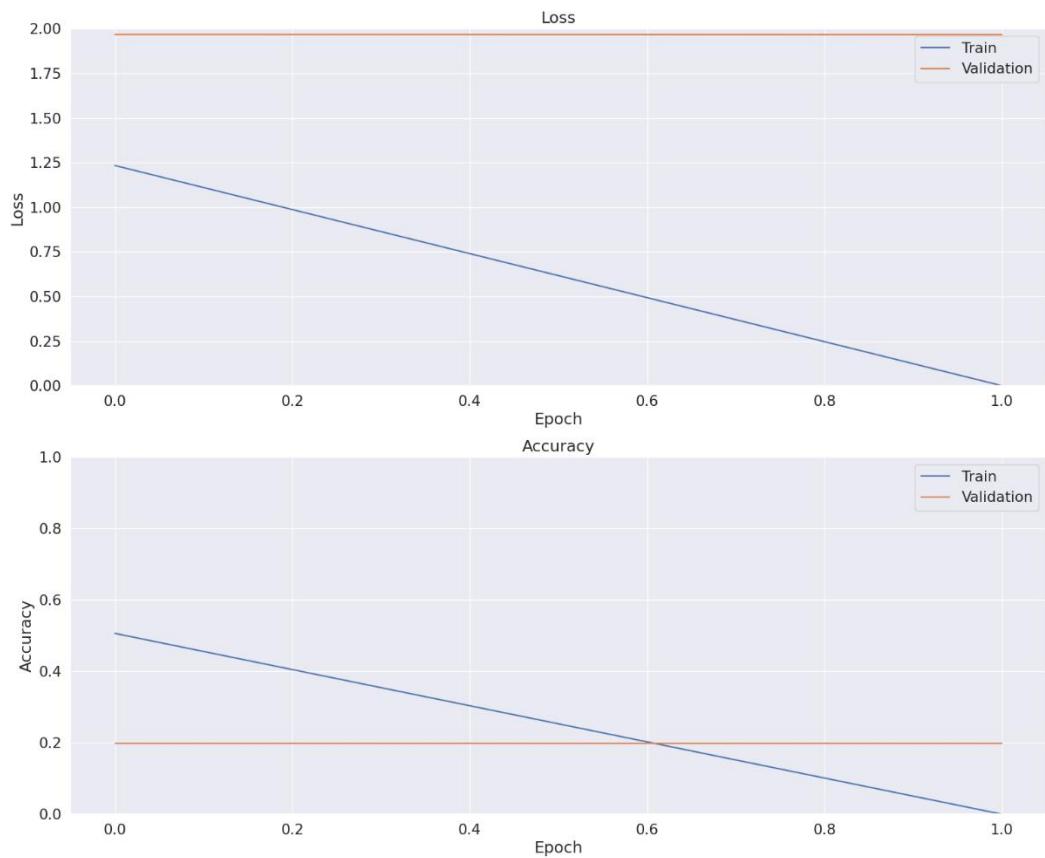
Hình 26 Đồ thị Loss và Accuracy của mô hình ResNet



Hình 27 Đồ thị Loss và Accuracy của mô hình pre-trained ResNet50



Hình 28 Đồ thị Loss và Accuracy của mô hình pre-trained Inception-ResNet-v2



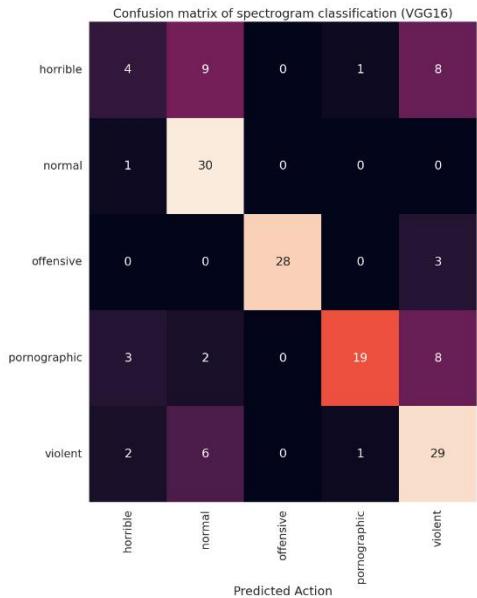
Hình 29 Đồ thị Loss và Accuracy của mô hình pre-trained InceptionV3

Các mô hình 3D CNN được đào tạo trước trên bộ dữ liệu ImageNet nhanh chóng đạt giá trị hàm Loss = 0 chỉ sau 2 epochs huấn luyện tuy nhiên độ chính xác đạt được không cao trong khi đó, mô hình 3D CNN đào tạo từ đầu cần nhiều thời gian và chi phí tính toán hơn. Kết quả thực nghiệm các mô hình được trình bày trong Bảng 3 dưới đây.

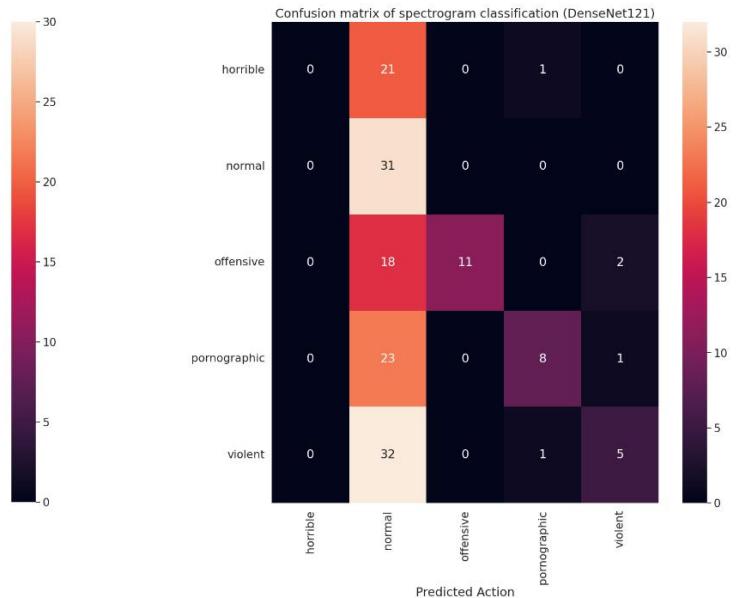
Mô hình	Accuracy	Precision	Recall	F1
ResNet	0.6558	<b>0.7218</b>	0.6611	0.6901
Pre-trained ResNet50	0.1428	0.1429	1.0	0.2501
Pre-trained Inception-ResNet-v2	0.1429	1.429	1.0	0.2501
Pre-trained InceptionV3	0.2013	0.2013	1.0	0.3353
MoViNet	<b>0.8077</b>	0.7184	<b>0.7551</b>	<b>0.7363</b>

Bảng 3 Kết quả các mô hình phân loại video

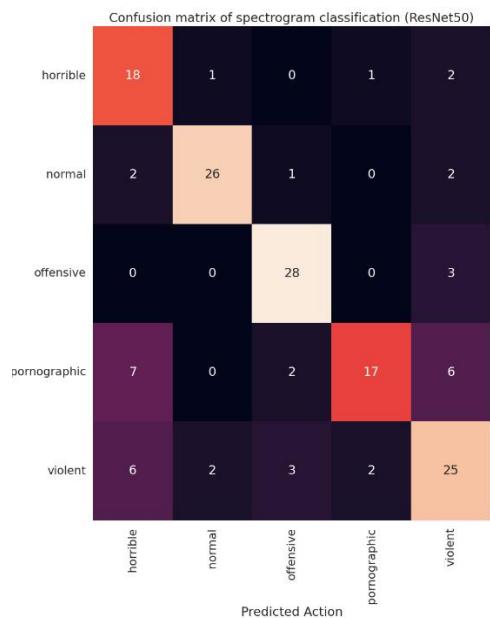
#### 4.2.2 Kết quả của các mô hình dự đoán với các ảnh spectrogram trích xuất từ âm thanh của video



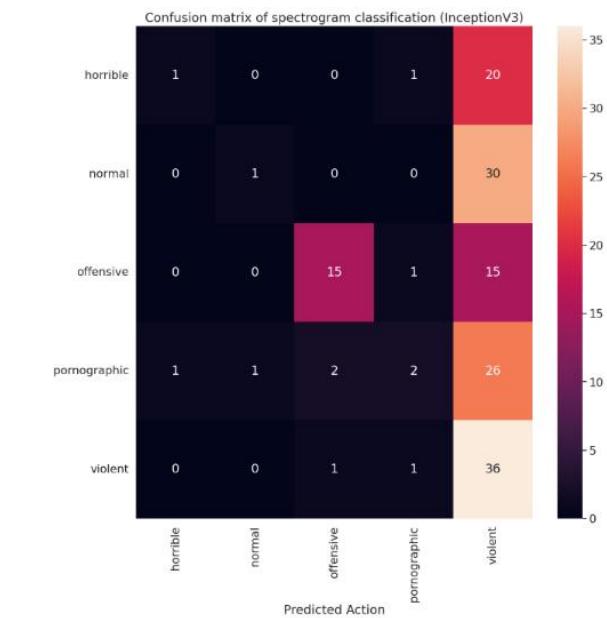
Hình 30 Ma trận nhầm lẫn của mô hình VGG16



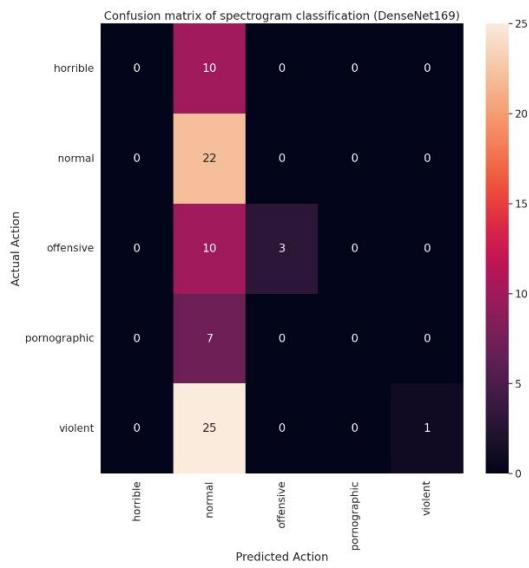
Hình 31 Ma trận nhầm lẫn của mô hình DenseNet121



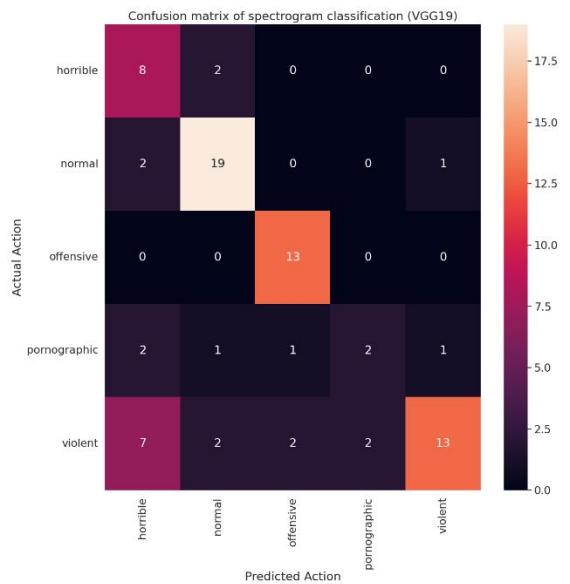
Hình 32 Ma trận nhầm lẫn của mô hình ResNet50



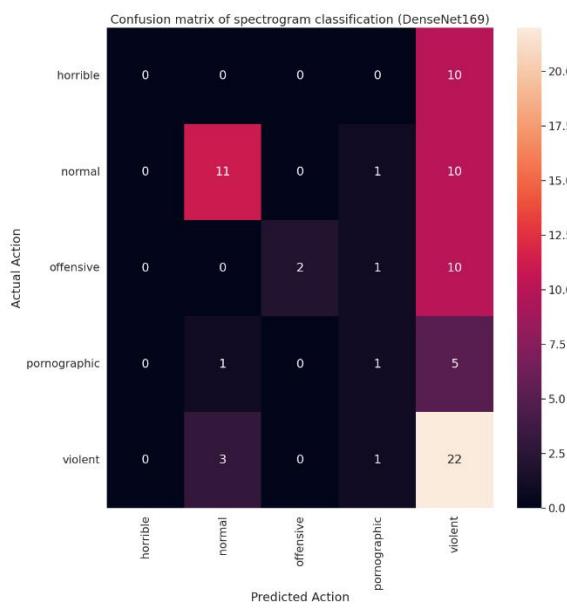
Hình 43 Ma trận nhầm lẫn của mô hình InceptionV3



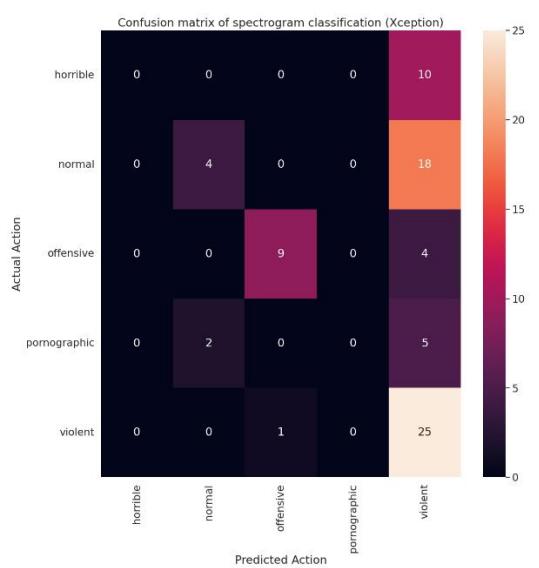
Hình 33 Ma trận nhầm lẫn của mô hình DenseNet169



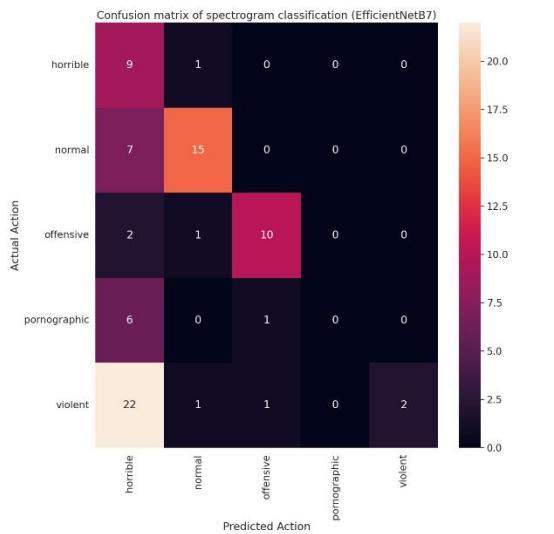
Hình 44 Ma trận nhầm lẫn của mô hình InceptionVGG19



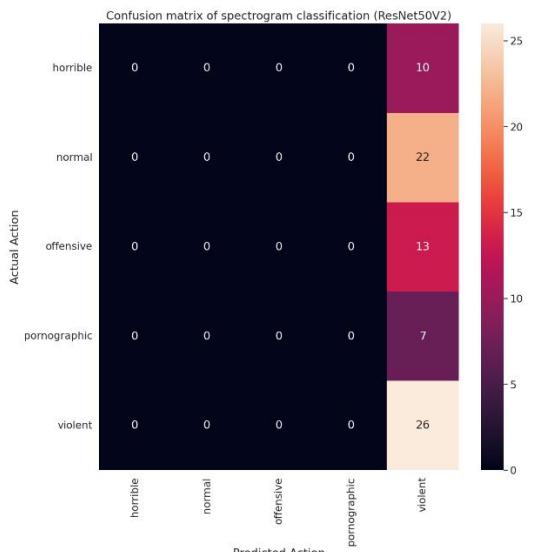
Hình 34 Ma trận nhầm lẫn của mô hình DenseNet201



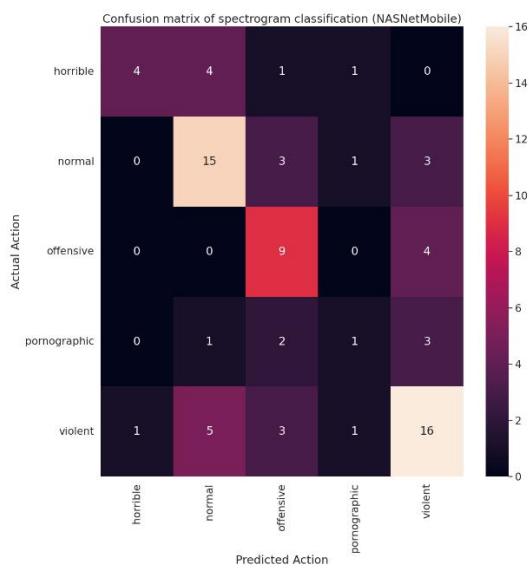
Hình 45 Ma trận nhầm lẫn của mô hình Xception



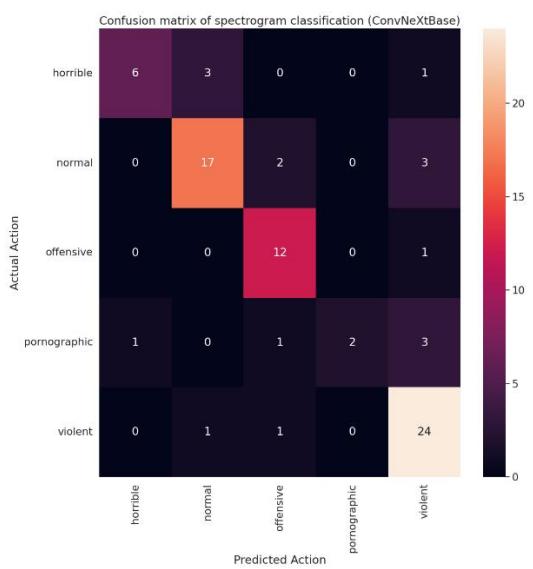
Hình 35 Ma trận nhầm lẫn của mô hình EfficientNetB7



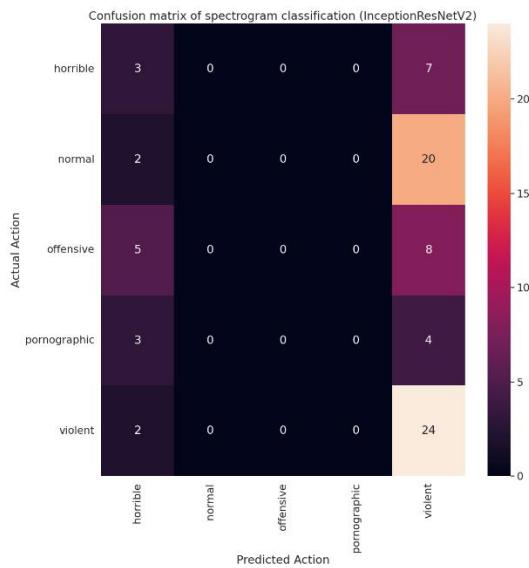
Hình 46 Ma trận nhầm lẫn của mô hình ResNet50V2



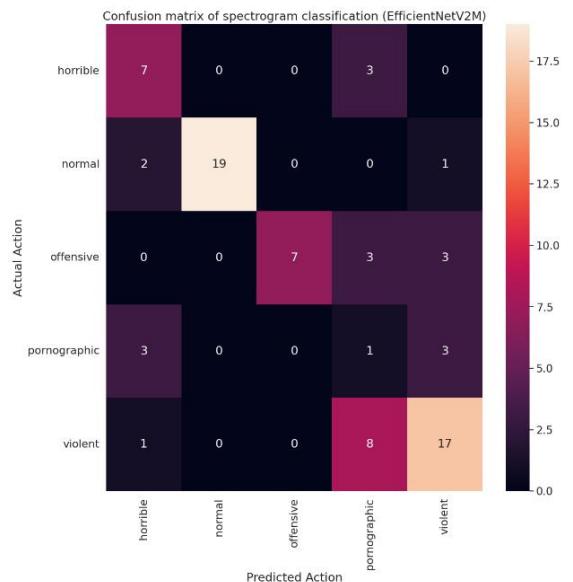
Hình 36 Ma trận nhầm lẫn của mô hình NASNetMobile



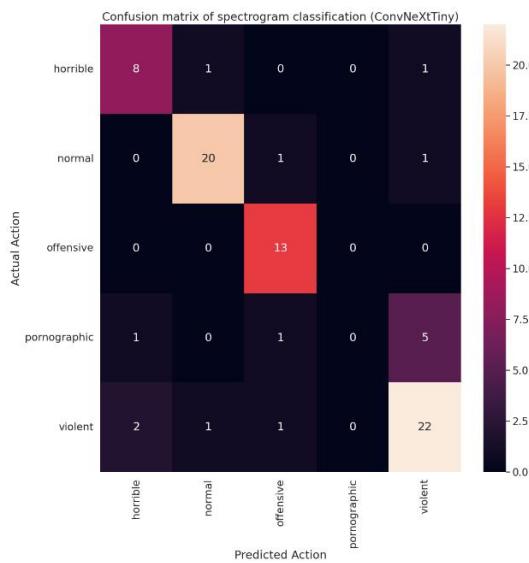
Hình 47 Ma trận nhầm lẫn của mô hình ConvNeXtBase



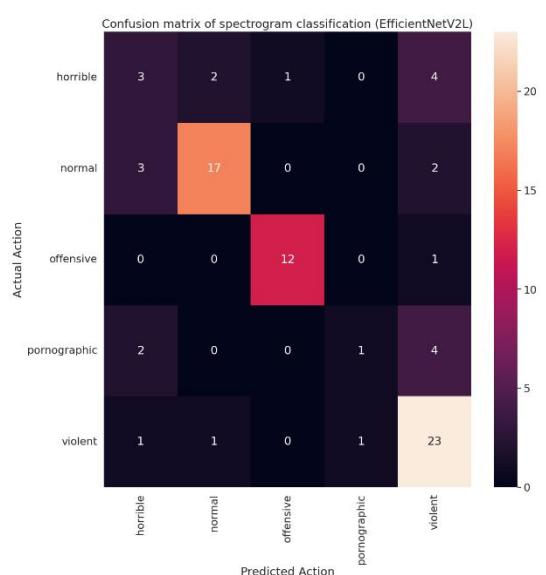
Hình 37 Ma trận nhầm lẫn của mô hình  
InceptionResNetV2



Hình 48 Ma trận nhầm lẫn của mô hình  
EfficientNetV2M

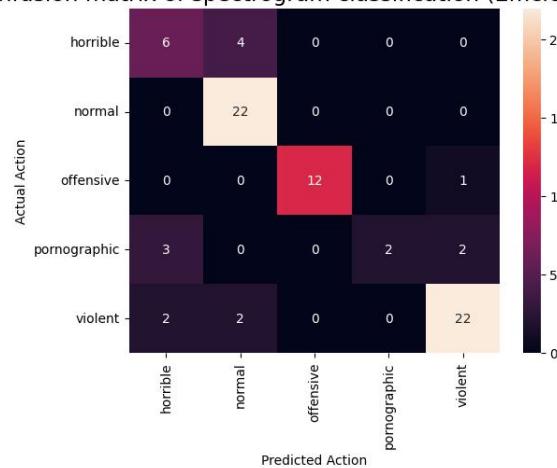


Hình 38 Ma trận nhầm lẫn của mô hình  
ConvNeXTiny



Hình 49 Ma trận nhầm lẫn của mô hình  
EfficientNetV2L

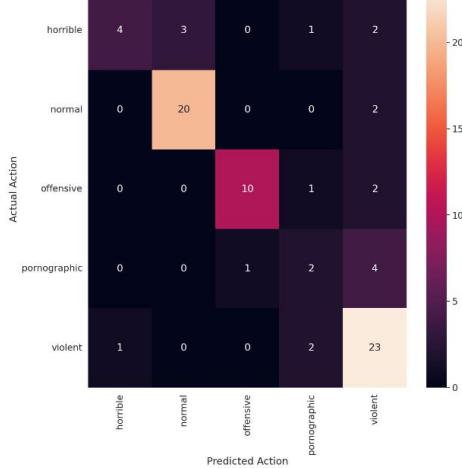
Confusion matrix of spectrogram classification (EfficientNetB3)



Hình 39 Ma trận nhầm lẫn của mô hình

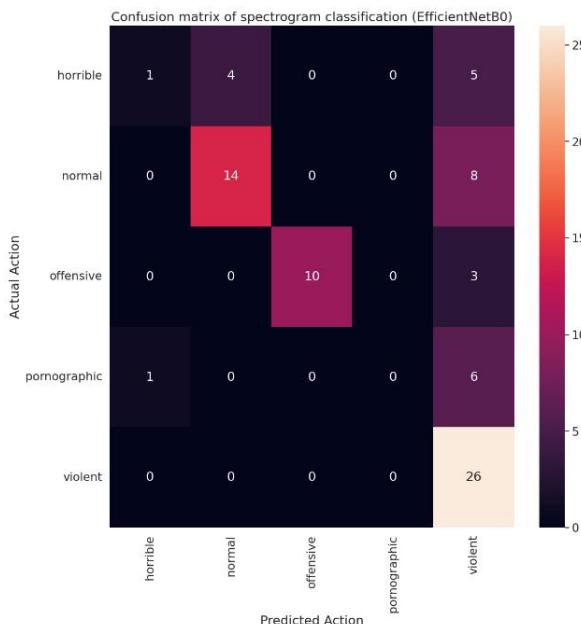
EfficientNetB3

Confusion matrix of spectrogram classification (EfficientNetB2)



Hình 50 Ma trận nhầm lẫn của mô hình

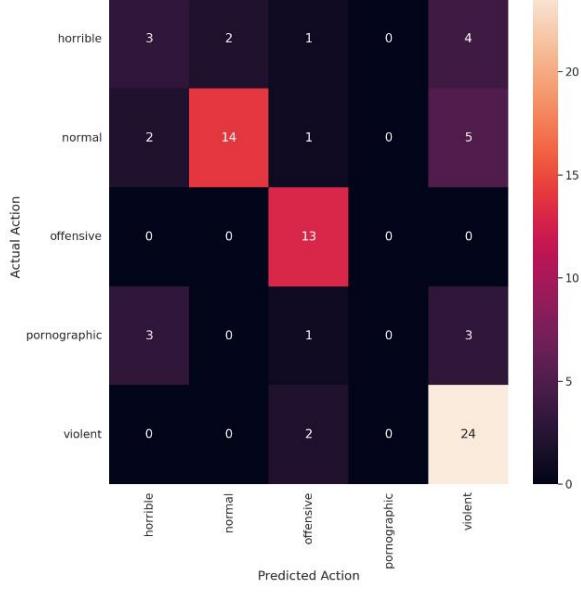
EfficientNetB2



Hình 40 Ma trận nhầm lẫn của mô hình

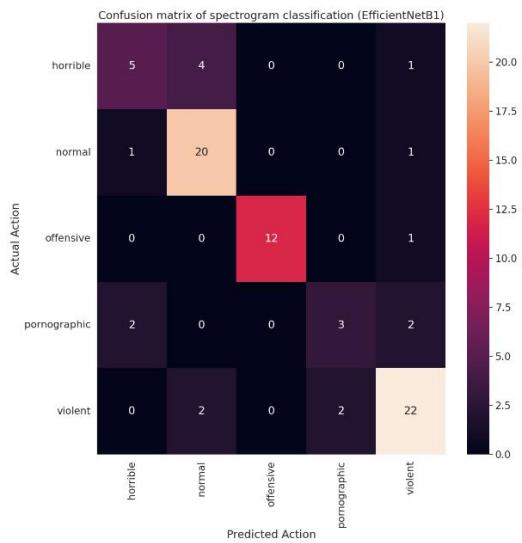
EfficientNetB0

Confusion matrix of spectrogram classification (EfficientNetB4)

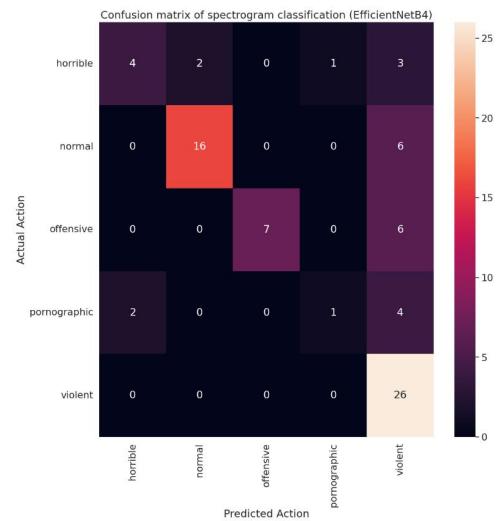


Hình 51 Ma trận nhầm lẫn của mô hình

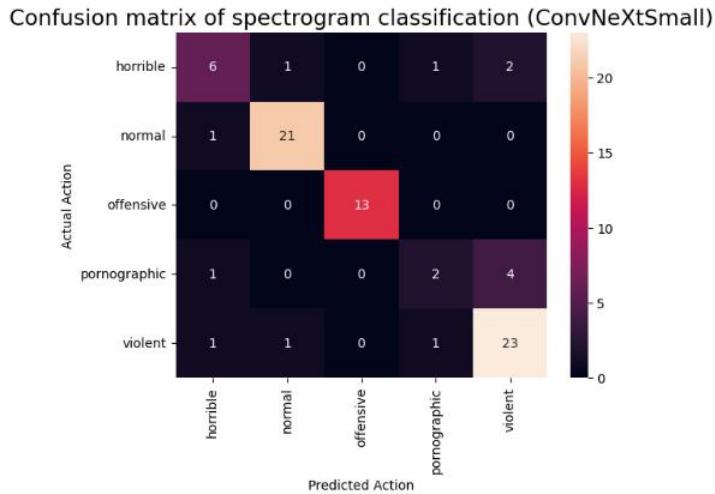
EfficientNetB4



Hình 41 Ma trận nhầm lẫn của mô hình EfficientNetB1



Hình 52 Ma trận nhầm lẫn của mô hình EfficientNetB7



Hình 42 Ma trận nhầm lẫn của mô hình ConvNeXtSmall

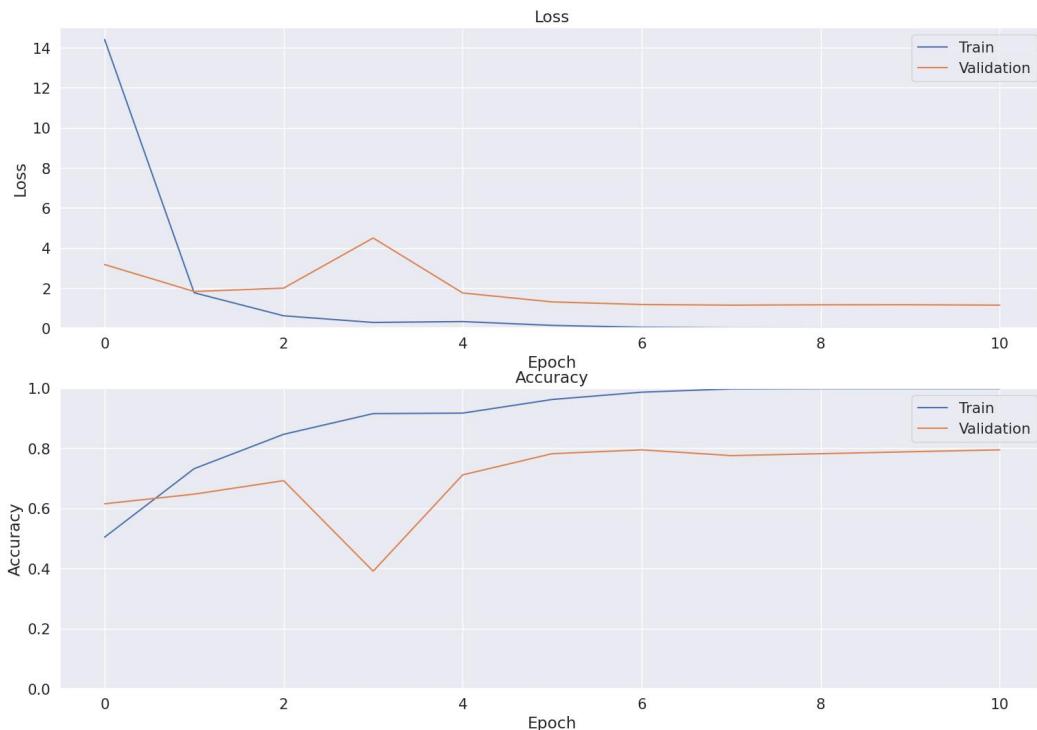
Các mô hình 2D CNN được đào tạo trước trên bộ dữ liệu ImageNet như ConvNeXts và EfficientNets, sử dụng đặc trưng là hình ảnh phô âm thanh, cho ra kết quả dự đoán chính xác (độ chính xác của ConvNeXtSmall là 83,34% và EfficientNetB3 là 82,05%) với tốc độ đào tạo nhanh hơn rất nhiều so với các mô hình 3D CNN đào tạo trên các khung hình ảnh từ video. Trong khi đó, các mô hình như DenseNet và ResNet50V2 có hiệu suất kém với độ chính xác và điểm F1 thấp. EfficientNetB3 và EfficientNetB1 cũng đạt kết quả tốt, đặc biệt

là về recall, cho thấy khả năng phát hiện cao các trường hợp dương tính. Tuy nhiên, mô hình ResNet50 nổi bật với độ chính xác cao nhất trong số các mô hình truyền thống, cho thấy sự ưu việt trong việc phân loại video dựa trên phổ âm thanh. Các mô hình 2D CNN này có khả năng nhận diện các nhãn yêu cầu yếu tố hình ảnh như “nhạy cảm”, “kinh dị”, “bình thường” và “bạo lực”. Điều này là vì các video khác nhau thuộc cùng một nhãn có thể có những đặc trưng giống nhau. Ví dụ như các video “kinh dị” sẽ có những âm thanh rùng rợn, u ám còn các video “xúc phạm” sẽ chứa giọng nói,...

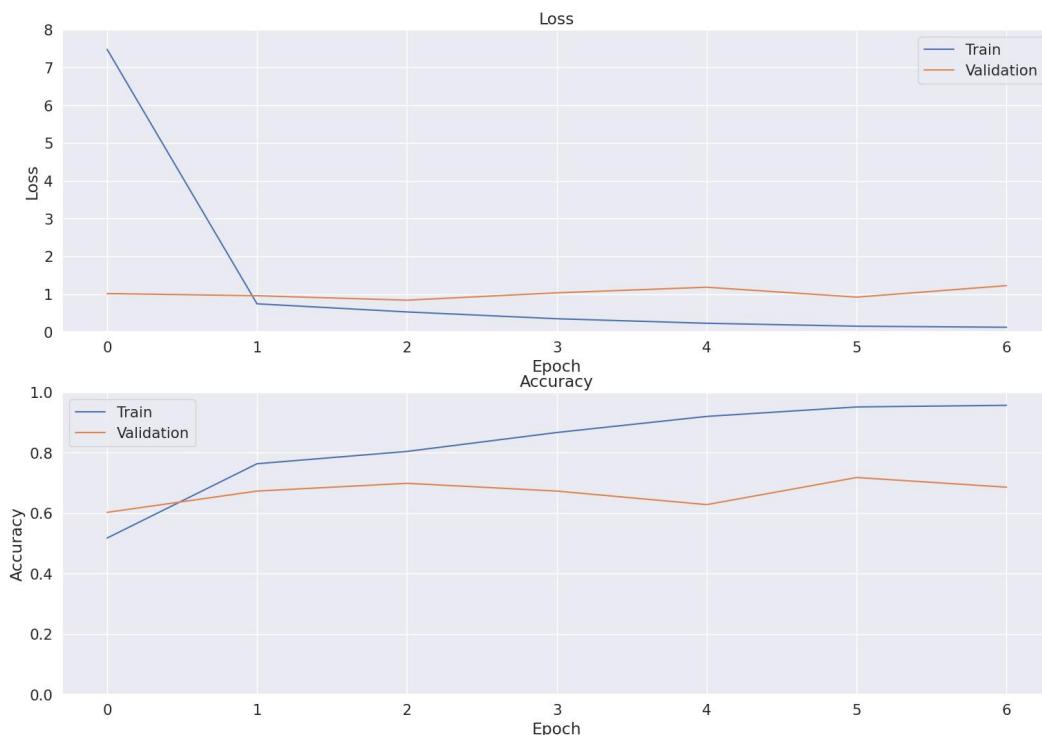
Mô hình	Accuracy	Precision	Recall	F1
VGG16	0.7143	0.6819	0.7094	0.672
VGG19	0.7051	0.6898	0.6784	0.6544
DenseNet121	0.3571	0.3472	0.5346	0.3039
DenseNet169	0.3333	0.2538	0.4595	0.1815
DenseNet201	0.4615	0.3286	0.4739	0.3146
ResNet50	0.7403	0.7499	0.7547	0.7389
ResNet50V2	0.3333	0.2000	0.0667	0.0100
Xception	0.4872	0.3671	0.3941	0.3273
InceptionV3	0.3571	0.3143	0.5034	0.2601
InceptionResNetV2	0.3462	0.2446	0.1162	0.1559
NASNetMobile	0.5769	0.5065	0.5531	0.5099
ConvNeXtTiny	0.8077	0.7110	0.6415	0.6735
ConvNeXtSmall	<b>0.8334</b>	<b>0.7450</b>	0.7746	<b>0.7530</b>
ConvNeXtLarge	0.8205	0.7124	0.6764	0.6921
ConvNeXtBase	0.7821	0.7009	0.8333	0.7192
EfficientNetB0	0.6538	0.5011	0.5639	0.4878
EfficientNetB1	0.7949	0.7214	0.7618	0.7358
EfficientNetB2	0.7561	0.6497	0.7217	0.6686
EfficientNetB3	0.8205	0.7310	<b>0.8422</b>	0.7437
EfficientNetB4	0.6923	0.5719	0.5278	0.5266
EfficientNetB7	0.4615	0.4856	0.5725	0.4029

EfficientNetV2M	0.6538	0.5798	0.6627	0.6013
EfficientNetV2L	0.7179	0.6047	0.6075	0.6075

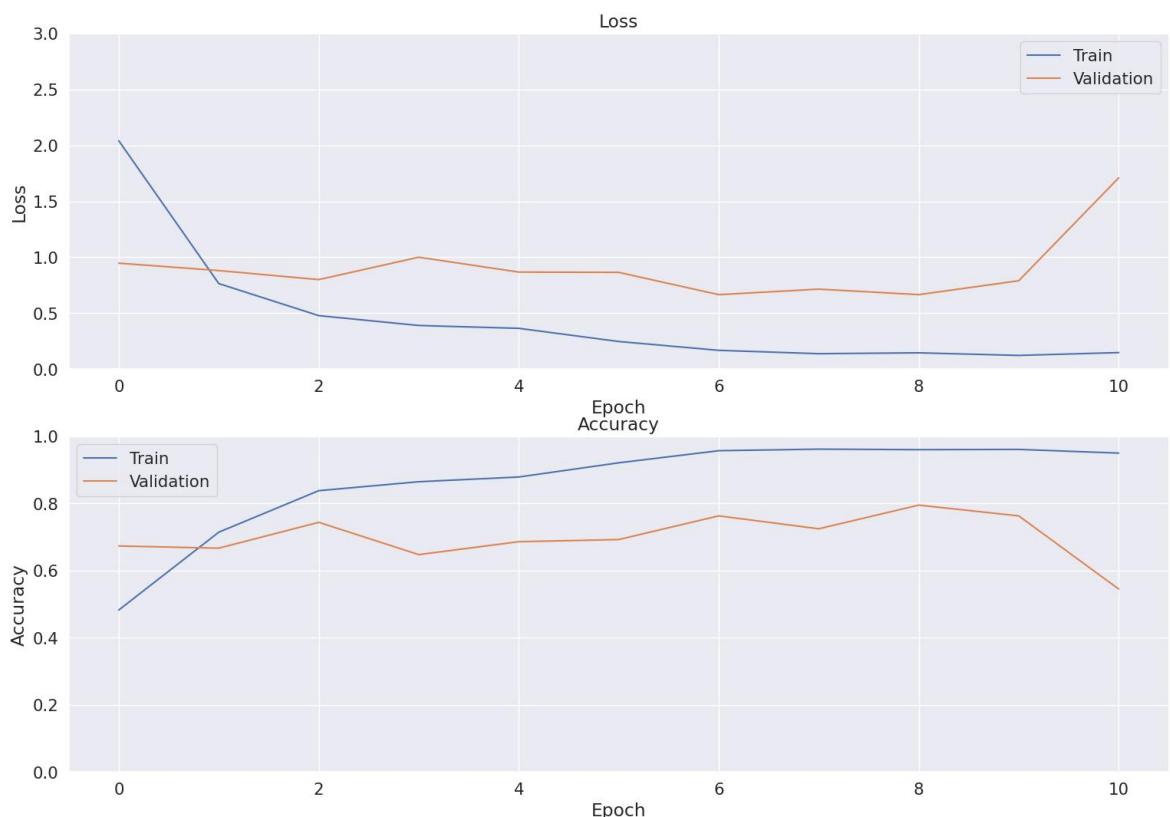
Bảng 4 Kết quả các mô hình phân loại video dựa trên phô âm thanh



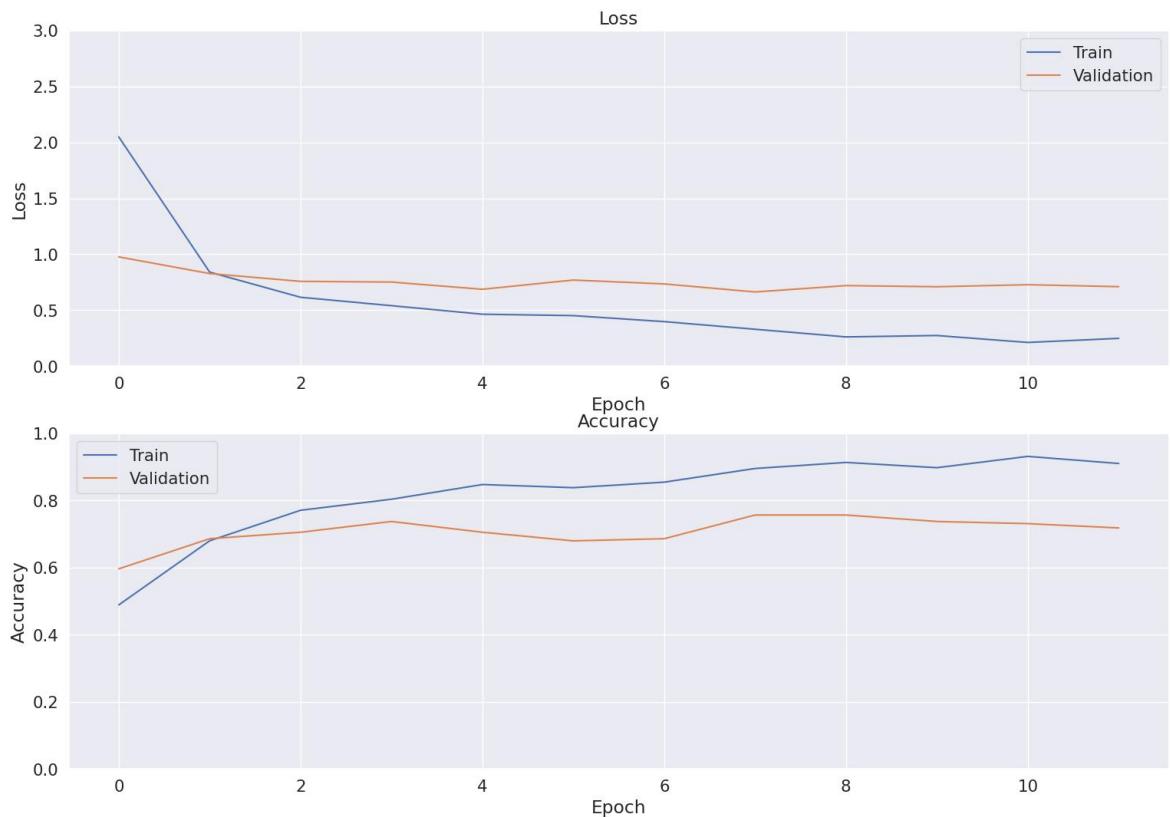
Hình 53 Đồ thị học với Accuracy và Loss của mô hình ResNet50



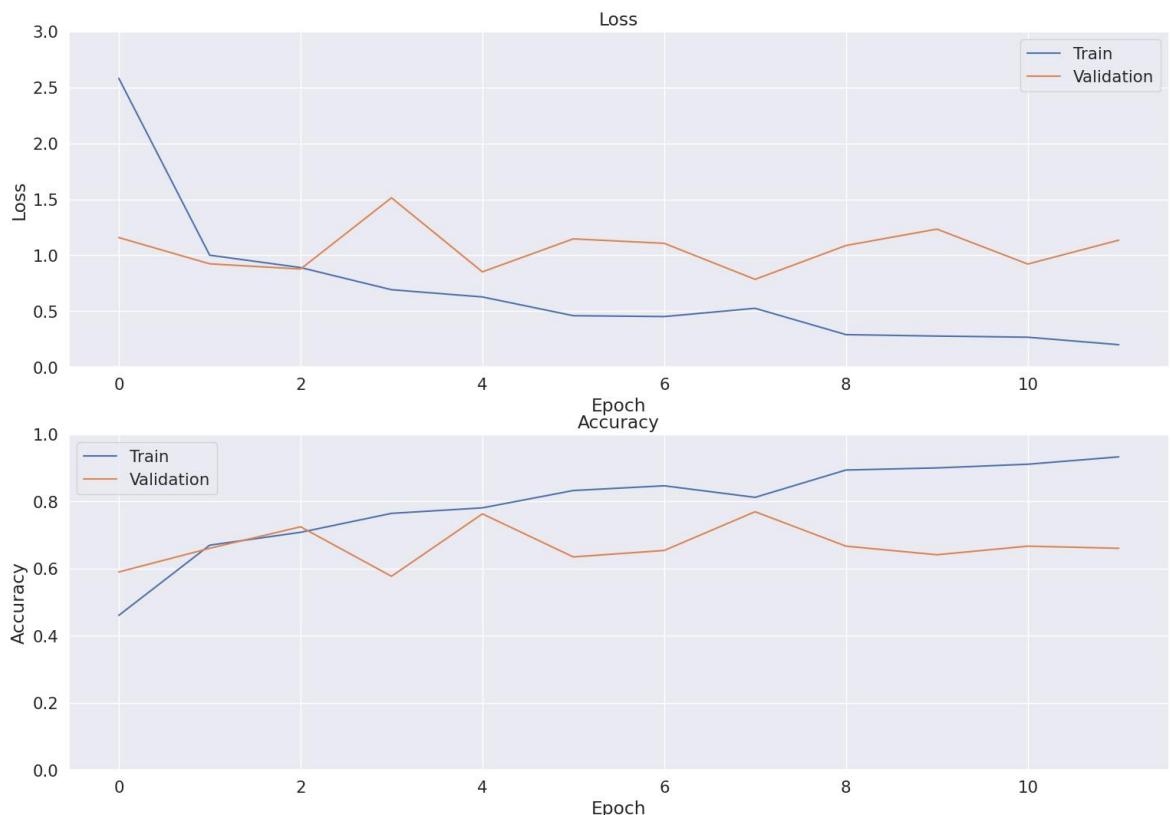
Hình 54 Đồ thị học với Accuracy và Loss của mô hình ConvNeXtSmall



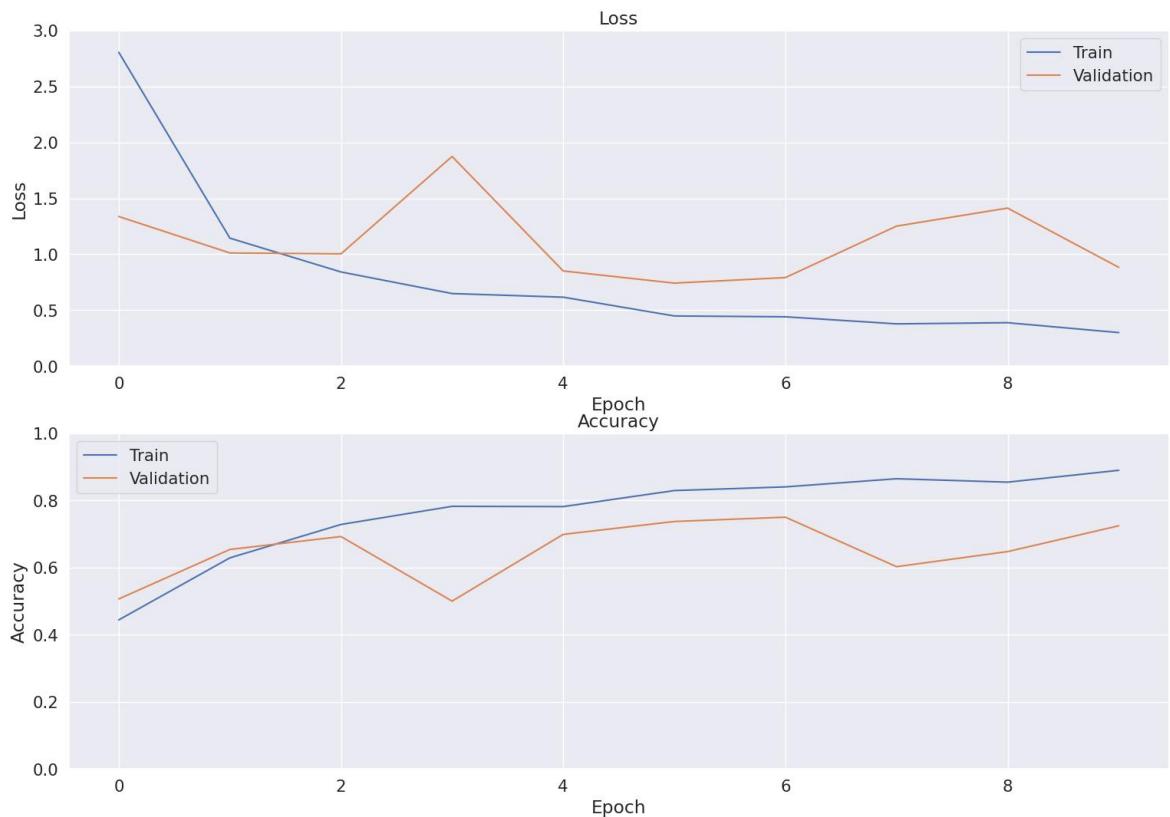
Hình 55 Đồ thị học với Accuracy và Loss của mô hình EfficientNetB0



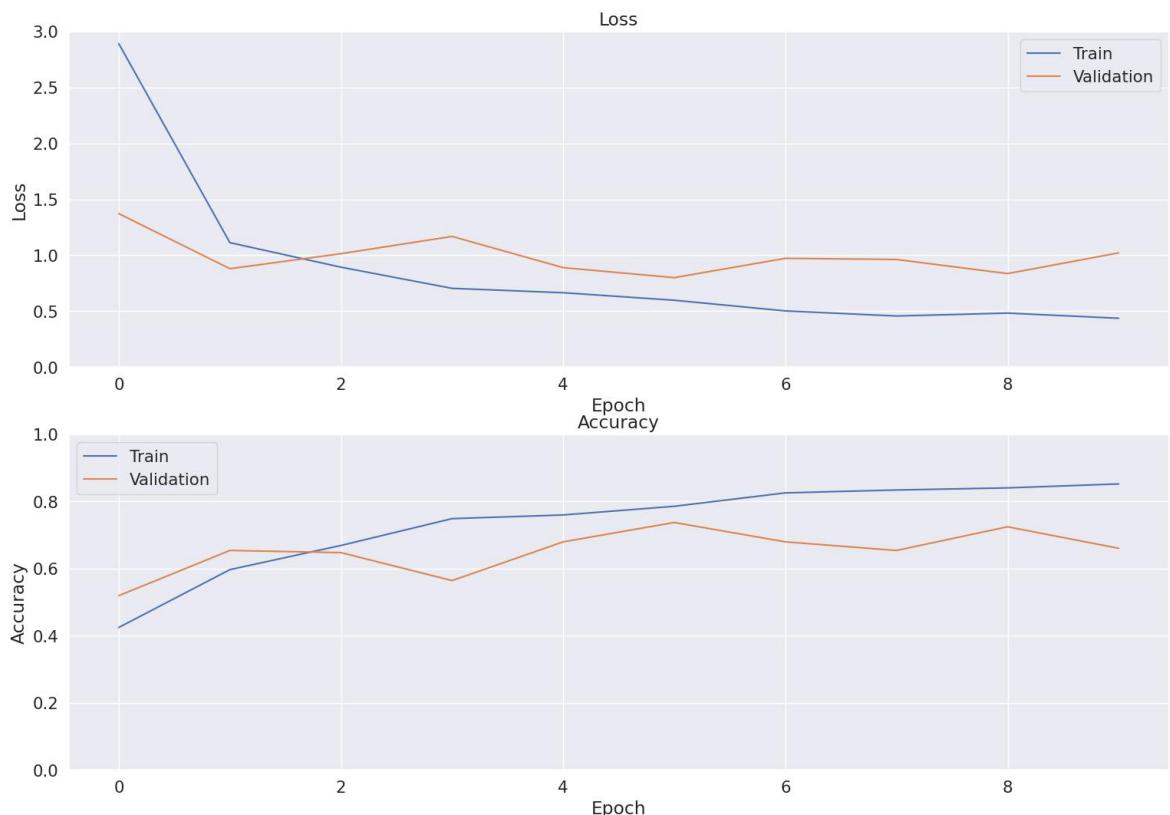
Hình 56 Đồ thị học với Accuracy và Loss của mô hình EfficientNetB1



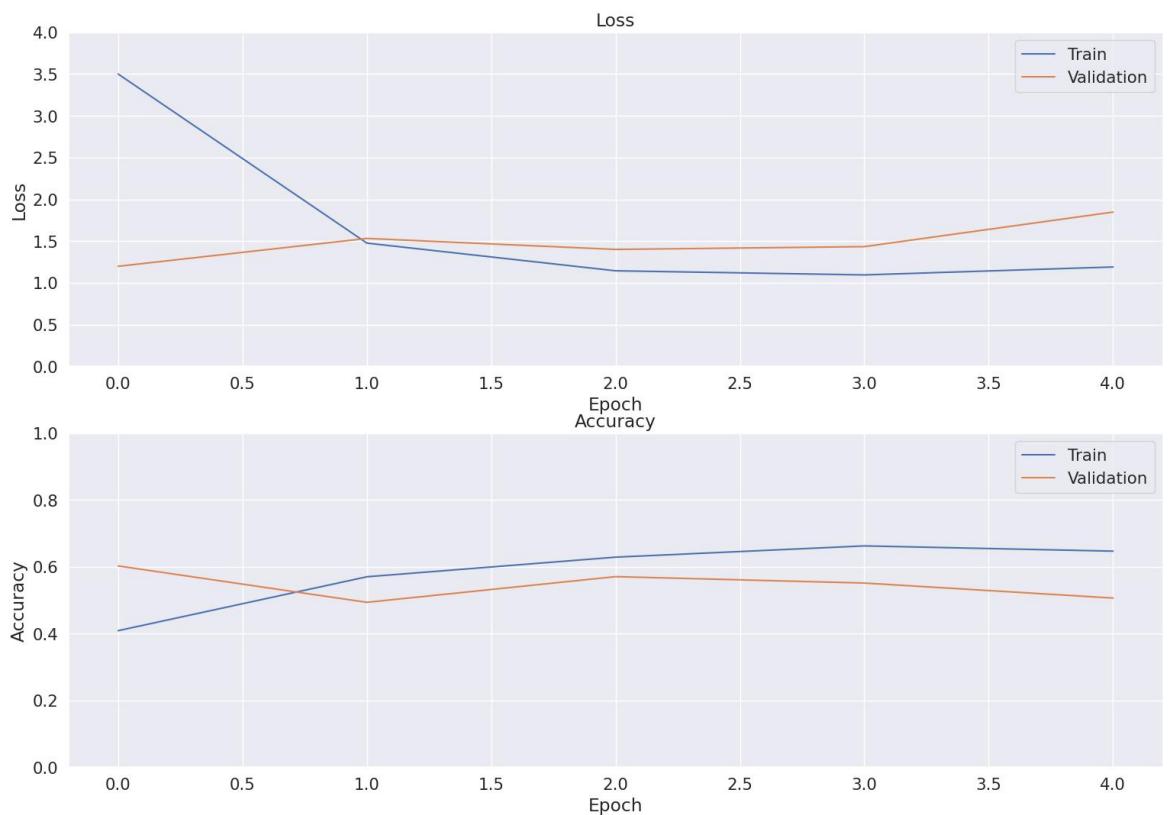
Hình 57 Đồ thị học với Accuracy và Loss của mô hình EfficientNetB2



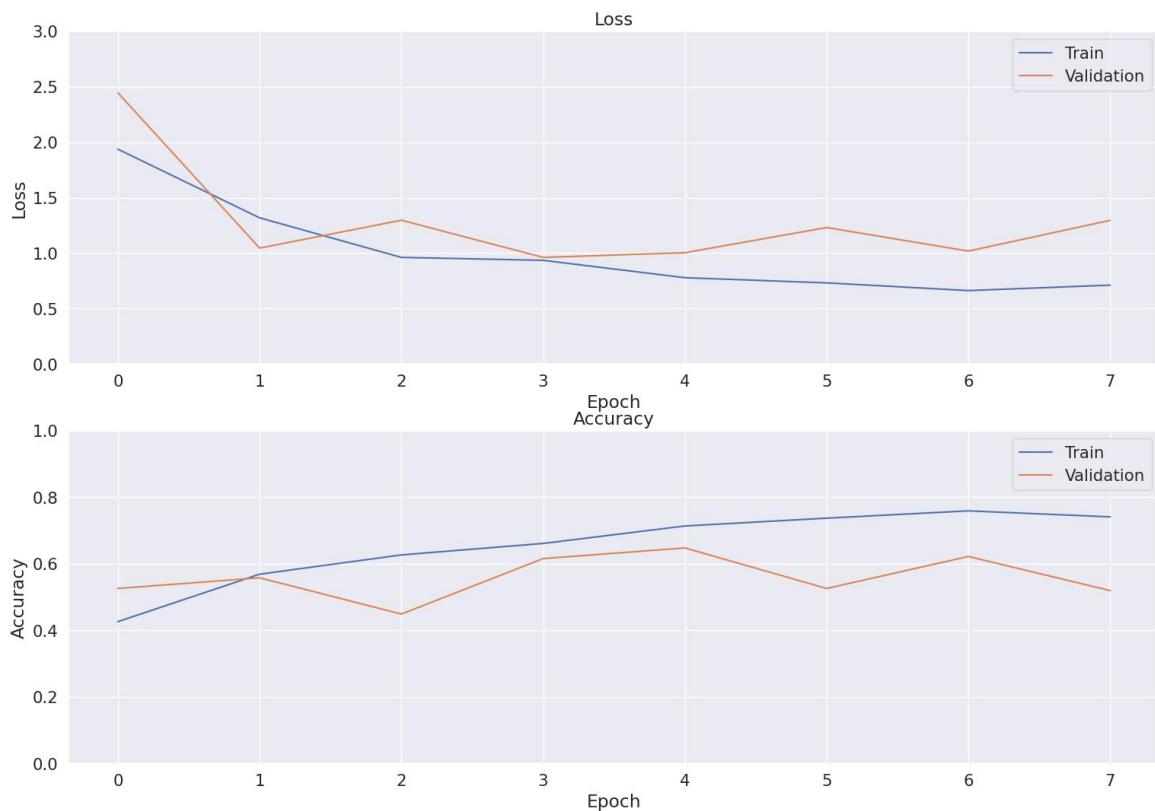
Hình 58 Đồ thị học với Accuracy và Loss của mô hình EfficientNetB3



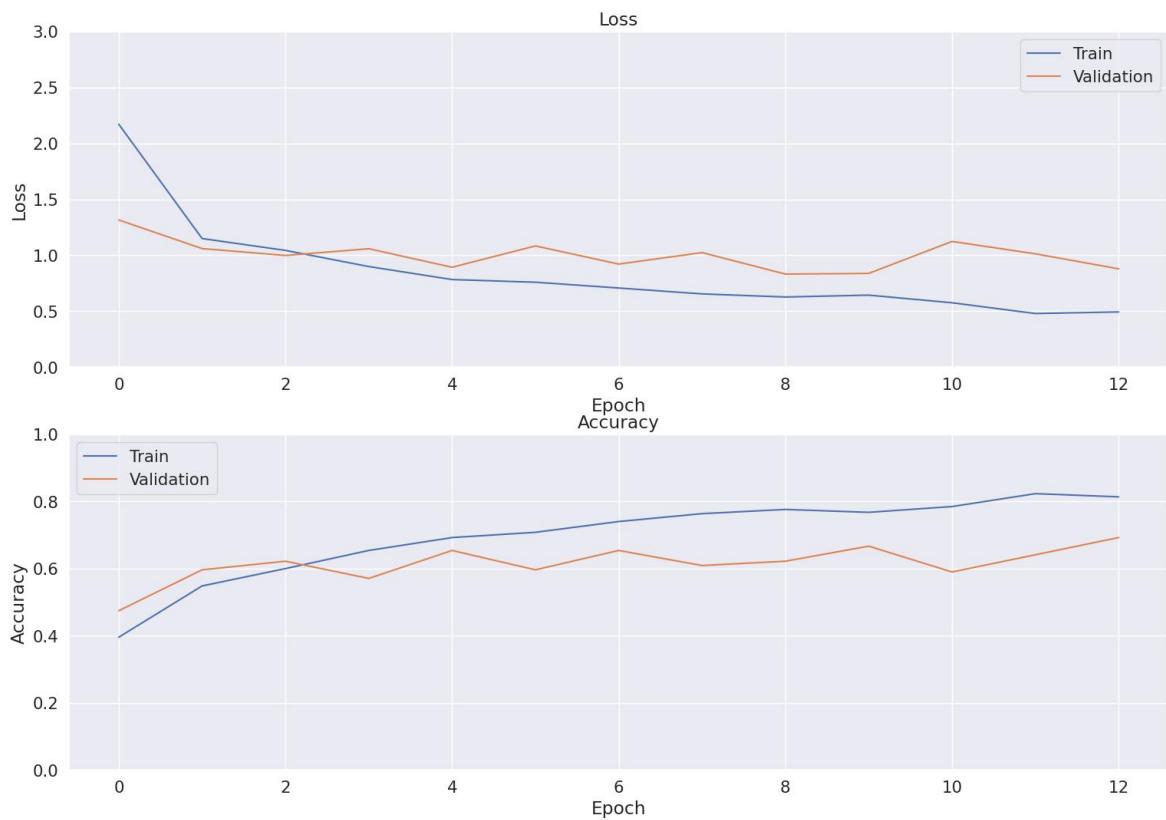
Hình 59 Đồ thị học với Accuracy và Loss của mô hình EfficientNetB4



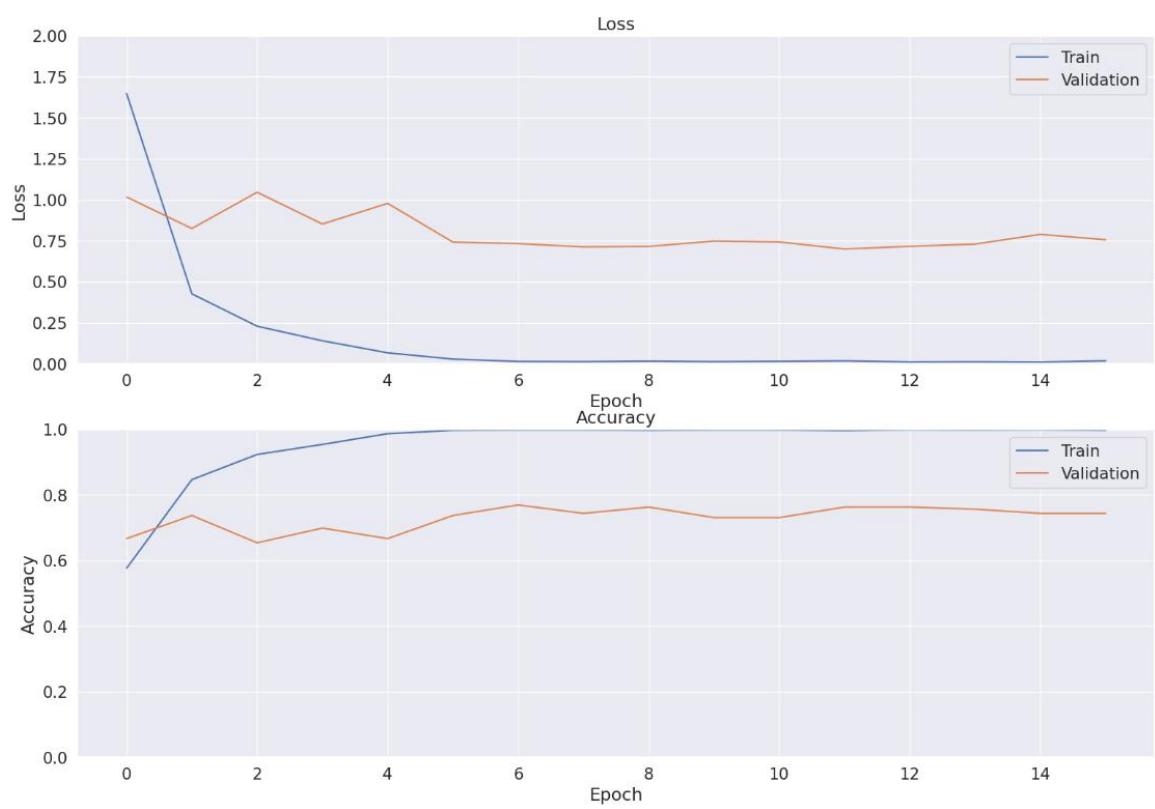
Hình 60 Đồ thị học với Accuracy và Loss của mô hình EfficientNetB7



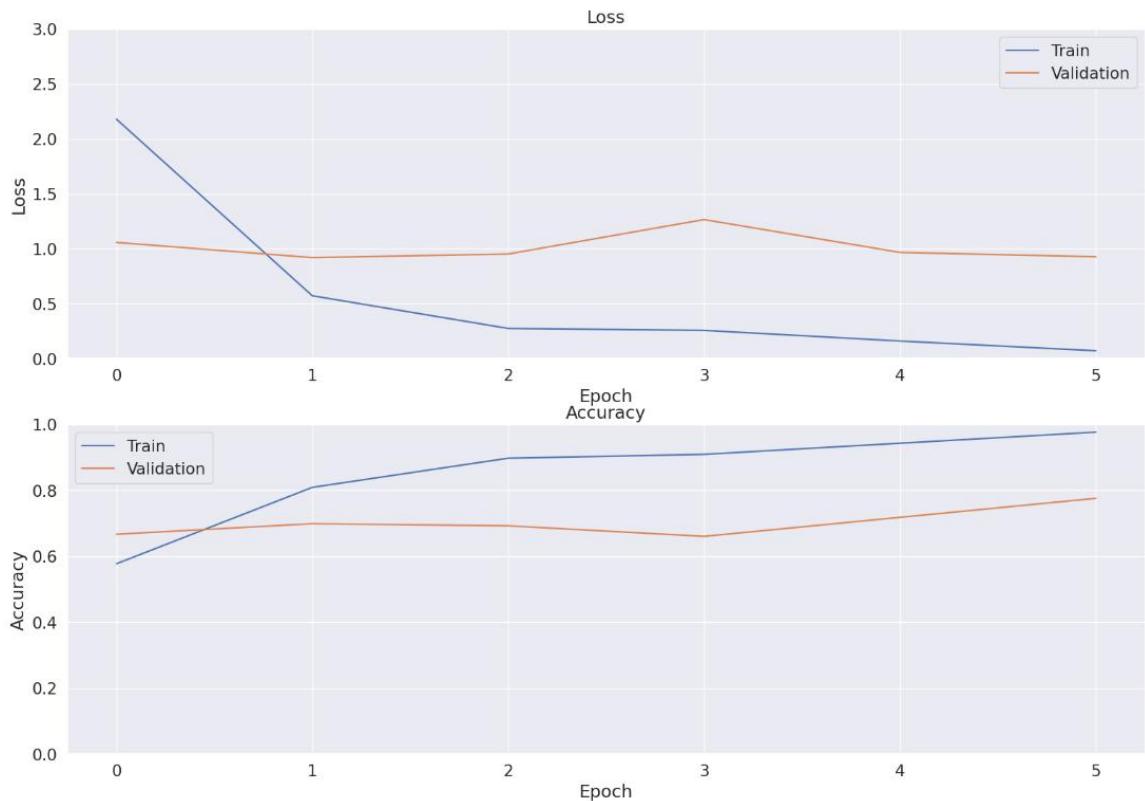
Hình 61 Đồ thị học với Accuracy và Loss của mô hình EfficientNetV2M



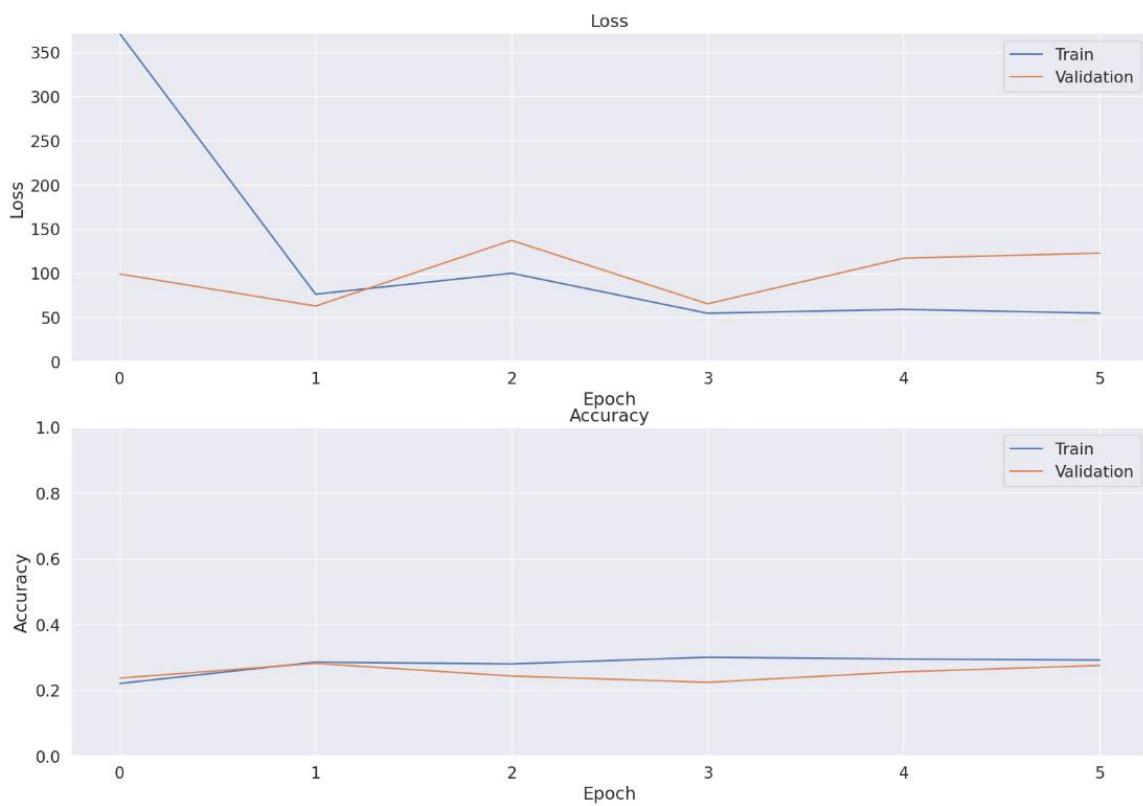
Hình 62 Đồ thị học với Accuracy và Loss của mô hình EfficientNetV2L



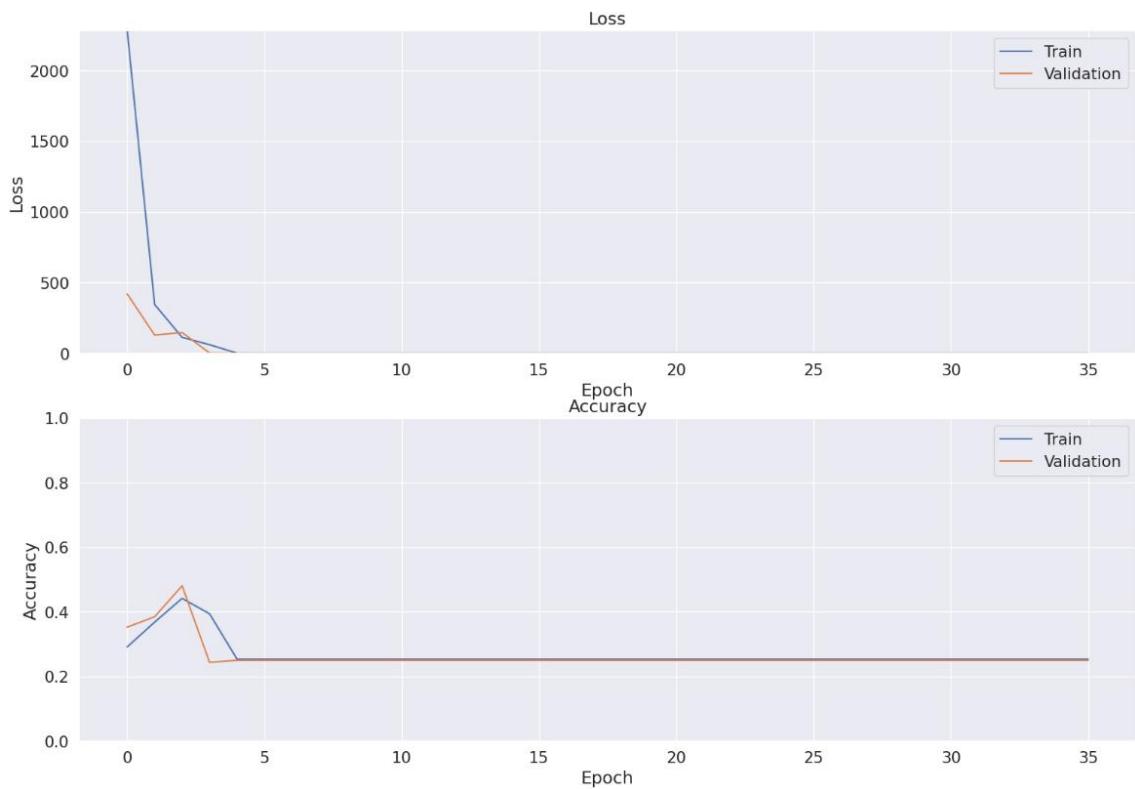
Hình 63 Đồ thị học với Accuracy và Loss của mô hình ConvNeXtTiny



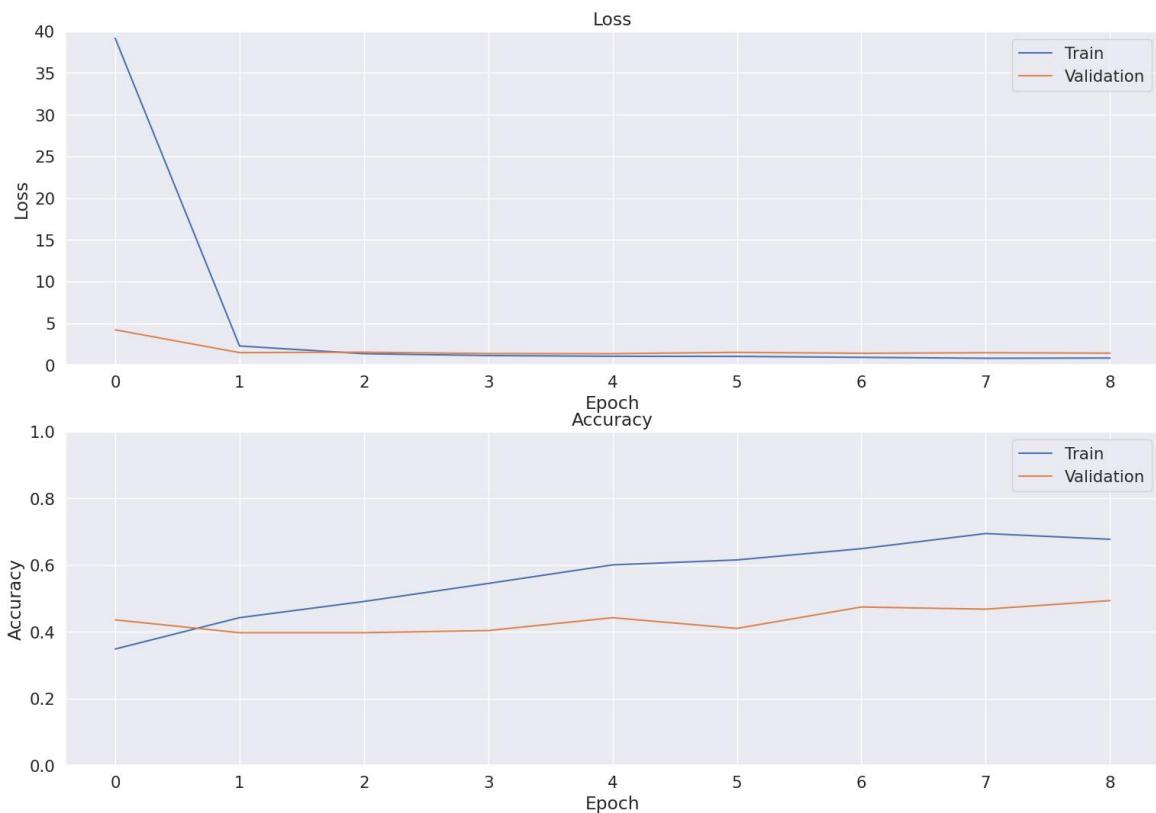
Hình 64 Đồ thị học với Accuracy và Loss của mô hình ConvNeXtBase



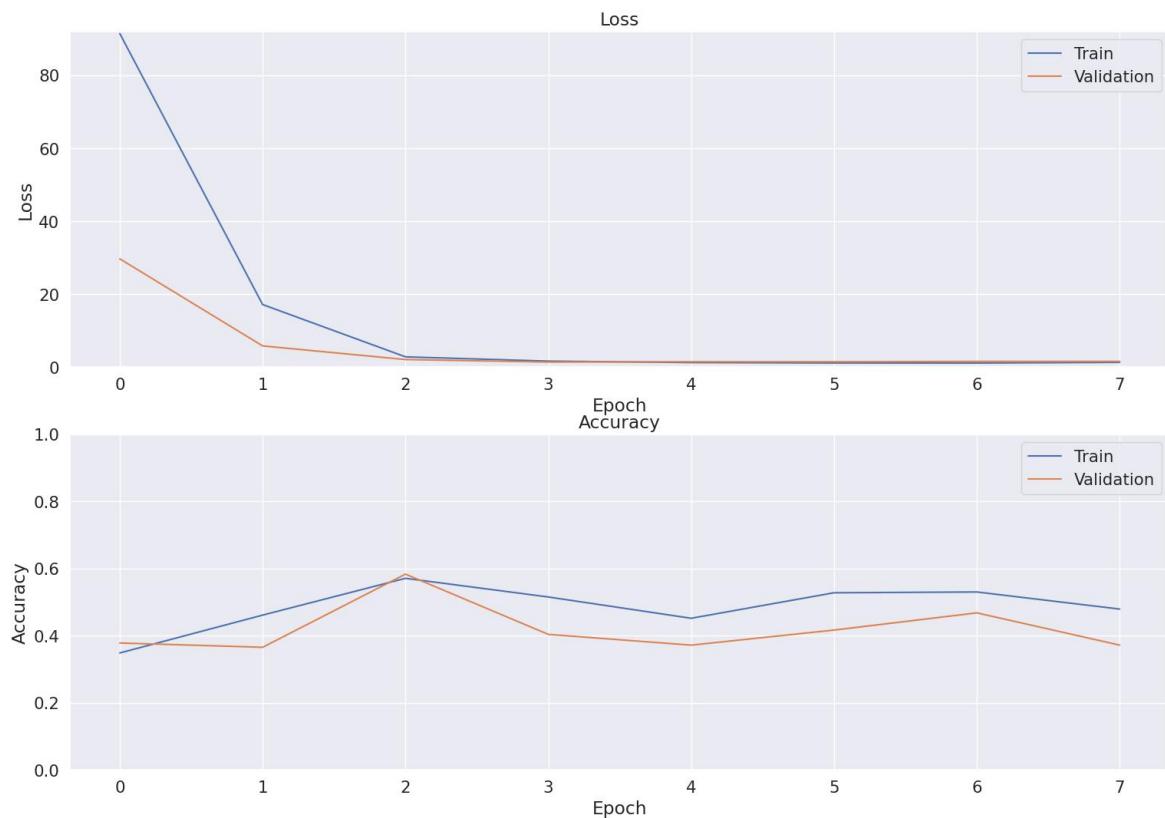
Hình 65 Đồ thị học với Accuracy và Loss của mô hình InceptionResNetV2



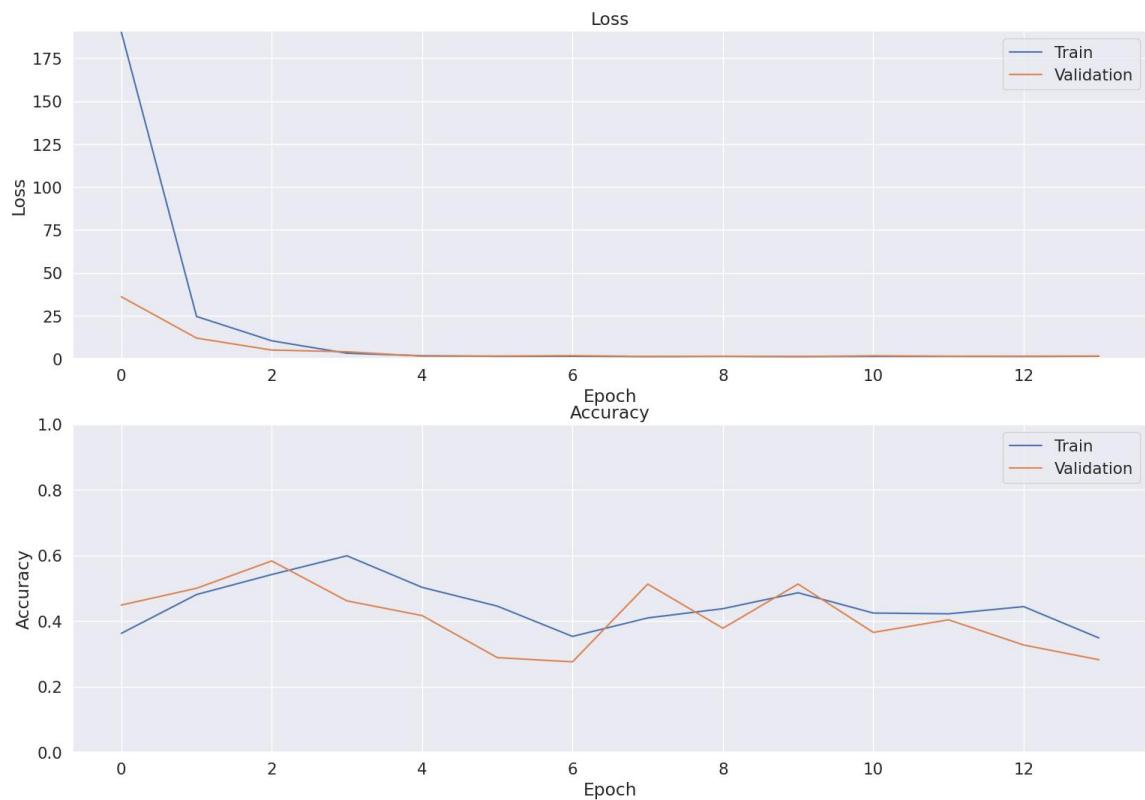
Hình 66 Đồ thị học với Accuracy và Loss của mô hình ResNet50V2



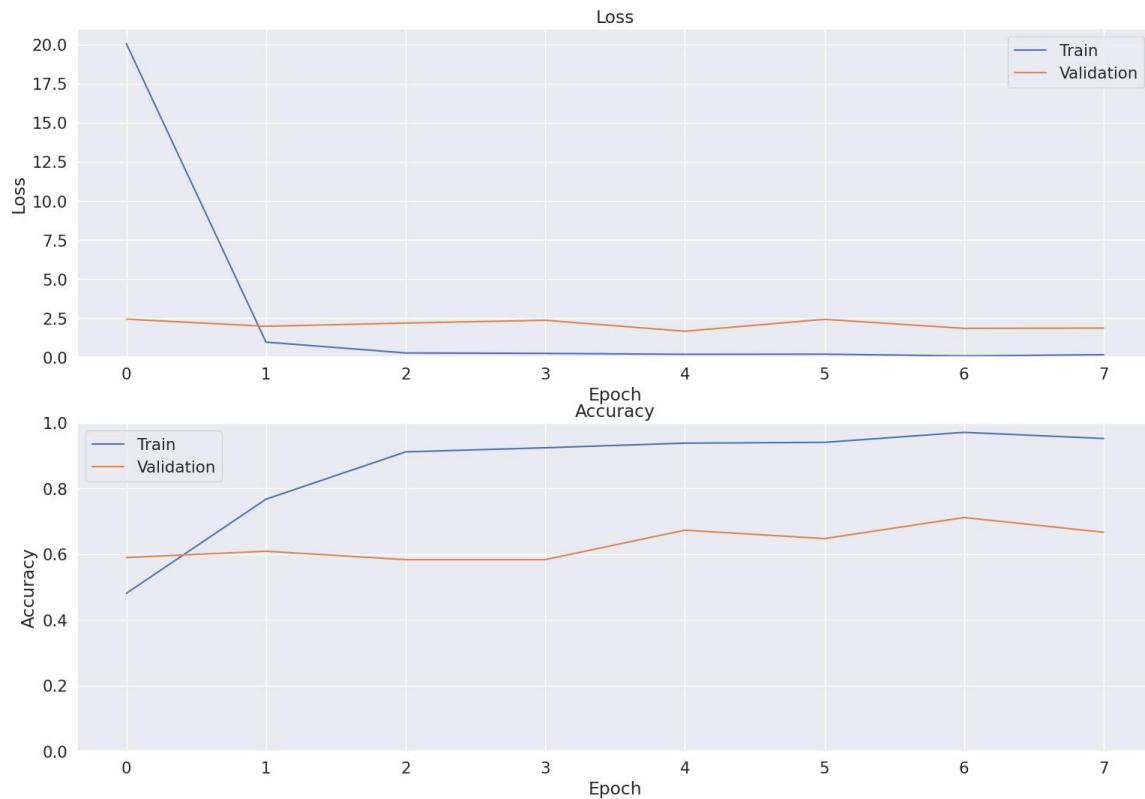
Hình 67 Đồ thị học với Accuracy và Loss của mô hình NASNetMobile



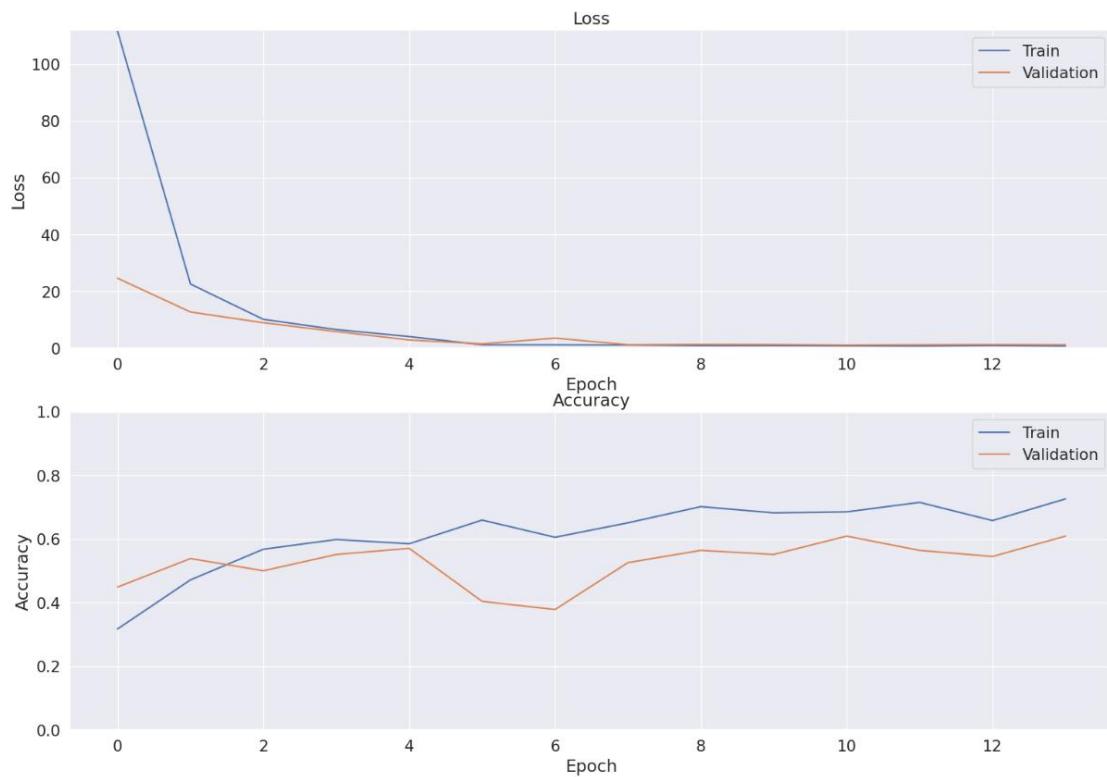
Hình 68 Đồ thị học với Accuracy và Loss của mô hình DenseNet201



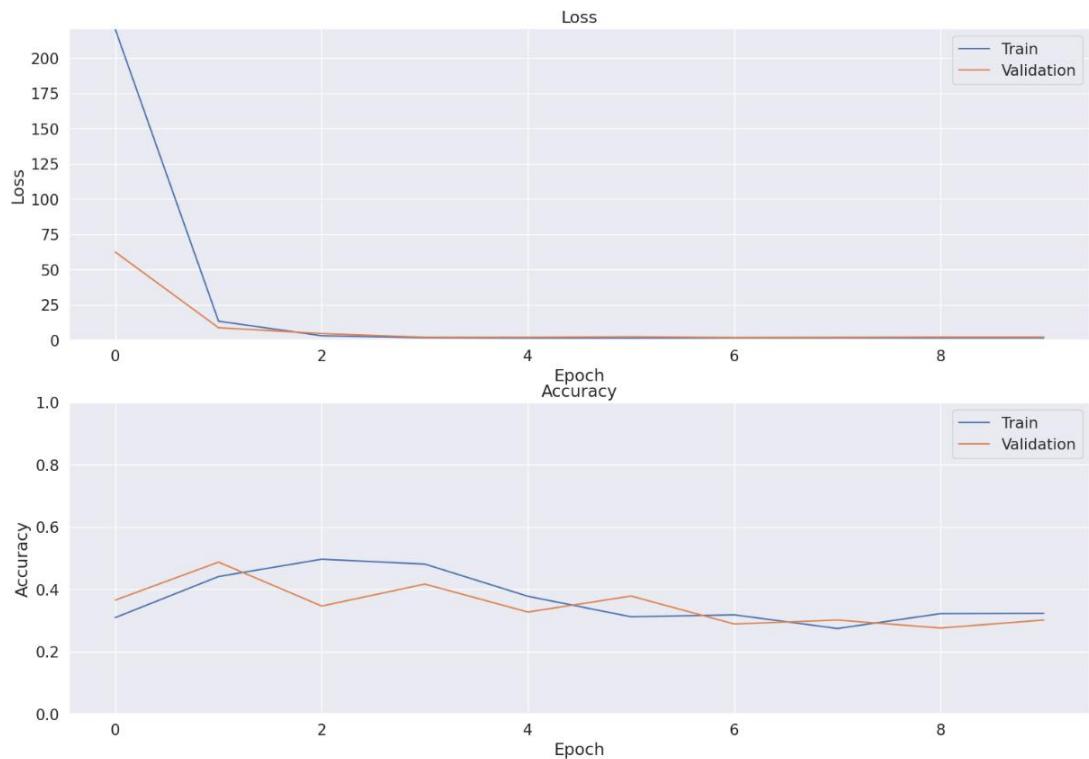
Hình 69 Đồ thị học với Accuracy và Loss của mô hình DenseNet169



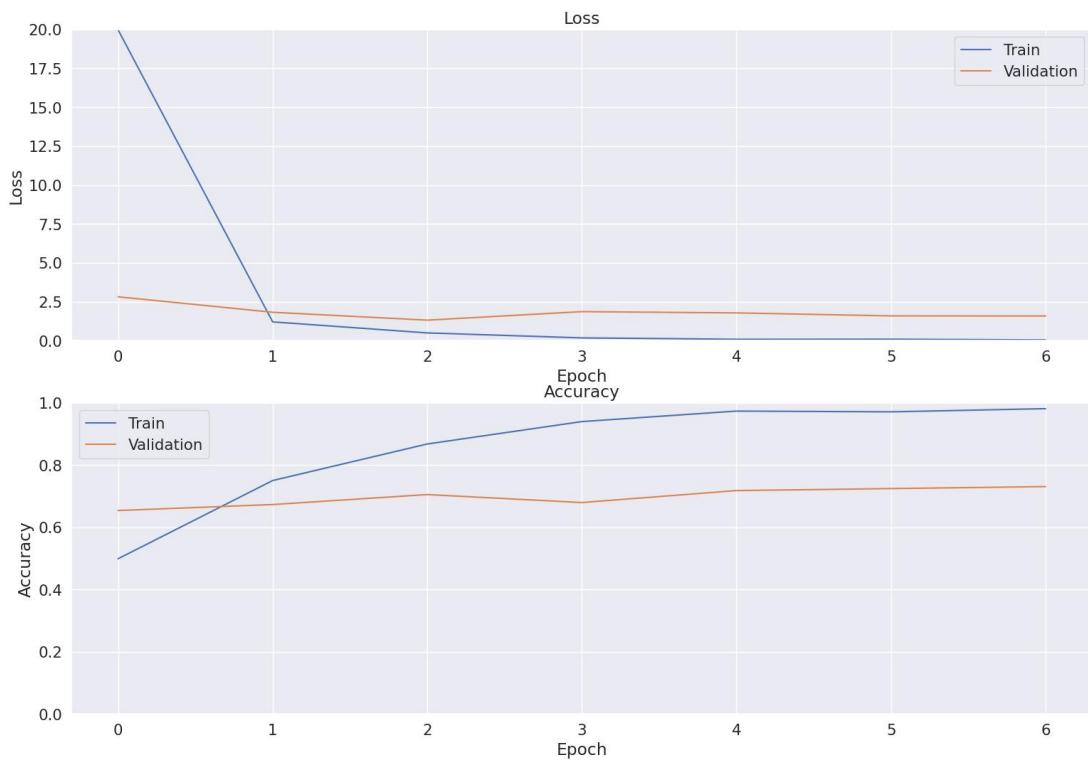
Hình 70 Đồ thị học với Accuracy và Loss của mô hình VGG16



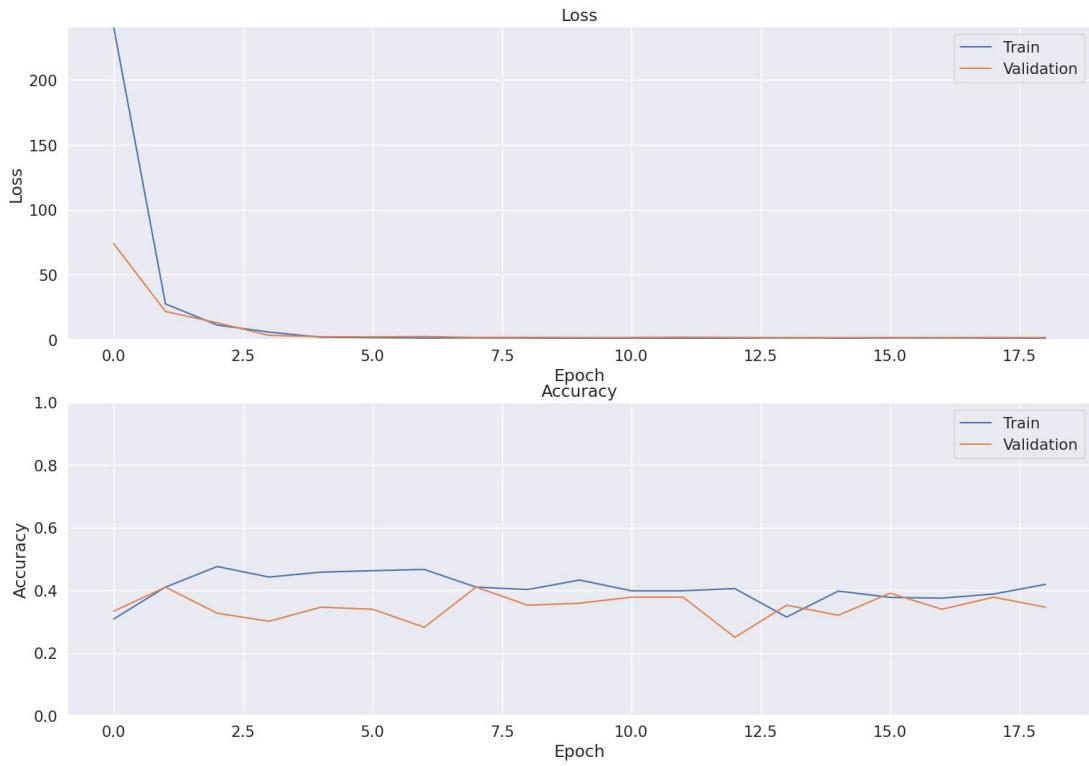
Hình 71 Đồ thị học với Accuracy và Loss của mô hình DenseNet121



Hình 72 Đồ thị học với Accuracy và Loss của mô hình InceptionV3



Hình 73 Đồ thị học với Accuracy và Loss của mô hình VGG19.



Hình 74 Đồ thị học với Accuracy và Loss của mô hình Xception.

### 4.2.3 Kết quả các mô hình dự đoán đối với văn bản trích xuất từ âm thanh video

Bảng kết quả cho thấy hiệu suất của các mô hình phân loại văn bản thông qua các chỉ số chính như accuracy, precision, recall và F1-score. Trong đó, PhoBERT nổi bật với tất cả các chỉ số rất cao, đặc biệt là precision đạt mức tối đa 1.0 và F1-score đạt 0.9315, cho thấy mô hình này có khả năng phân loại chính xác và đồng nhất cao. ViSoBERT và viBERT cũng có hiệu suất khá tốt với accuracy lần lượt là 0.6923 và 0.7436, cùng với F1-score tương đối cao là 0.5983 và 0.6414. Ngược lại, các mô hình như BERT, RoBERTa, XLM-RoBERTa, DistilBERT và CafeBERT có kết quả thấp, đặc biệt là các chỉ số precision và F1-score đều dưới 0.2, vì chỉ cho ra kết quả dự đoán một nhãn duy nhất là “bình thường”. BiLSTM và vELECTRA có kết quả trung bình với các chỉ số không quá cao nhưng cũng không quá thấp. Nhìn chung, PhoBERT là mô hình có hiệu suất tốt nhất trong số các mô hình được so sánh.

Mô hình	accuracy	precision	recall	F1-score
BiLSTM	0.5812	0.4736	0.5359	0.4827
PhoBERT	<b>0.8718</b>	<b>1.0</b>	<b>0.8718</b>	<b>0.9315</b>
BERT	0.2949	0.1917	0.2949	0.1921
RoBERTa	0.2821	0.0564	0.2000	0.0880
XLM-RoBERTa	0.2821	0.0564	0.2000	0.0880
DistilBERT	0.2821	0.0564	0.2000	0.0880
CafeBERT	0.2821	0.0564	0.2000	0.0880
ViSoBERT	0.6923	0.6357	0.6031	0.5983
viBERT	0.7436	0.7671	0.6493	0.6414
vELECTRA	0.5256	0.4758	0.4612	0.4536

Bảng 5 Bảng kết quả các mô hình phân loại văn bản

Mô hình	accuracy	precision	recall	F1-score
3DresNet+ResNet50+PhoBERT	0.7692	0.7029	0.7692	0.7308

Bảng 6 Kết quả sau khi kết hợp các kết quả dự đoán từ các đặc trưng

## CHƯƠNG 5: KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

Trong đề tài này, chúng tôi tập trung nghiên cứu về phân loại video độc hại tiếng Việt tràn lan trên mạng xã hội hiện nay và đã thành công xây dựng bộ dữ liệu HarmfulVideosVN2023, đây là bộ dữ liệu đầu tiên về phân loại video độc hại có sử dụng yếu tố ngôn ngữ là tiếng Việt. Chúng tôi còn đề xuất các phương pháp trích xuất đặc trưng hình ảnh, âm thanh, và văn bản đối với bộ dữ liệu trên bao gồm trích xuất phổ âm thanh và trích xuất văn bản tiếng Việt từ video và tiến hành phân tích ảnh hưởng các đặc trưng đối với tác vụ phân loại video độc hại tiếng Việt bằng cách áp dụng các mô hình dự đoán đối với từng đặc trưng trên.

Bằng các thực nghiệm, chúng tôi nhận thấy khả năng của bộ dữ liệu HarmfulVideosVN2023 với vai trò phục vụ nghiên cứu phân loại video độc hại tiếng Việt của mình. Đồng thời qua các thí nghiệm áp dụng các mô hình phân loại hiện đại với từng đặc trưng riêng lẻ, chúng tôi rút ra được hiệu suất và chi phí tính toán khi sử dụng các đặc trưng là các khung hình ảnh từ video, hình ảnh phổ âm thanh và văn bản giọng nói đối để phân loại video độc hại tiếng Việt. Ngoài ra, việc sử dụng nhiều loại mô hình khác nhau còn giúp chúng tôi tìm ra những mô hình nào có khả năng dự đoán chính xác cho tác vụ của mình.

Các mô hình dự đoán sử dụng đặc trưng là các khung hình ảnh từ video tồn tại nhiều tài nguyên và chi phí tính toán, tuy nhiên việc sử dụng các mô hình được đào tạo trước trên bộ dữ liệu Kinetics hoặc các bộ dữ liệu về video tương tự có thể giải quyết tốt bất cập này. Đối với đặc trưng là hình ảnh phổ âm thanh, các mô hình được đào tạo trước trên bộ dữ liệu ImageNet với kiến trúc ConvNeXt giúp chúng tôi thu được các kết quả dự đoán đáng tin cậy và chi phí tính toán thấp nhất. Hơn nữa, các mô hình transformer được đào tạo trên bộ dữ liệu tiếng Việt cho ra kết quả dự đoán khá chính xác khi học với dữ liệu và văn bản giọng nói trích từ video, nổi trội nhất là PhoBERT cho ra kết quả dự đoán chính xác nhất trong tất cả các mô hình của tất cả các đặc trưng mà chúng tôi trích xuất được với độ chính xác là 87,18%.

## TÀI LIỆU THAM KHẢO

- [1]. D. T. T. Vo, T. M. Tran, N. D. Vo and K. Nguyen, "UIT-Anomaly: A Modern Vietnamese Video Dataset for Anomaly Detection," 2021 8th NAFOSTED Conference on Information and Computer Science (NICS), Hanoi, Vietnam, 2021, pp. 352-357, doi: 10.1109/NICS54270.2021.9701556.
- [2]. Luong, Hieu-Thi, and Hai-Quan Vu. "A non-expert Kaldi recipe for Vietnamese speech recognition system." Proceedings of the Third International Workshop on Worldwide Language Service Infrastructure and Second Workshop on Open Infrastructures and Analysis Frameworks for Human Language Technologies (WLSI/OIAF4HLT2016). 2016.
- [3]. <https://trituenhantao.io/kien-thuc/resnet-mang-hoc-sau-dung-nghia/>
- [4]. T. N. Vu, T. T. Dinh, N. D. Vo, T. M. Tran and K. Nguyen, "VNAomaly: A novel Vietnam surveillance video dataset for anomaly detection," 2021 8th NAFOSTED Conference on Information and Computer Science (NICS), Hanoi, Vietnam, 2021, pp. 266-271, doi: 10.1109/NICS54270.2021.9701540.
- [5]. Thuan, Duong H., et al. "Typhoon impact and recovery from continuous video monitoring: A case study from Nha Trang Beach, Vietnam." Journal of Coastal Research 75 (2016): 263-267.
- [6]. Tran, Du, et al. "A closer look at spatiotemporal convolutions for action recognition." Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. 2018.
- [7]. [zalo.careers/blog/nhan-dien-ten-rieng-ner-voi-bidirectional-long-short-term-memory-va-conditional-random-field-CJO](https://zalo.careers/blog/nhan-dien-ten-rieng-ner-voi-bidirectional-long-short-term-memory-va-conditional-random-field-CJO)
- [8]. Radwan, Noha. Leveraging sparse and dense features for reliable state estimation in urban environments. Diss. University of Freiburg, Freiburg im Breisgau, Germany, 2019.
- [9]. Abu-El-Haija, Sami, et al. "Youtube-8m: A large-scale video classification benchmark." arXiv preprint arXiv:1609.08675 (2016).
- [10]. Alsina-Pagès, Rosa Ma, et al. "homesound: Real-time audio event detection based on high performance computing for behaviour and surveillance remote monitoring." Sensors 17.4 (2017): 854.
- [11]. Cao, Wenming, et al. "A comprehensive survey on geometric deep learning." IEEE Access 8 (2020): 35929-35949.
- [12]. Davis, Steven, and Paul Mermelstein. "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences." IEEE transactions on acoustics, speech, and signal processing 28.4 (1980): 357-366.
- [13]. Deng, Li, and Dong Yu. "Deep learning: methods and applications." Foundations and trends® in signal processing 7.3–4 (2014): 197-387.
- [14]. Goyal, Raghav, et al. "The" something something" video database for learning and evaluating visual common sense." Proceedings of the IEEE international conference on computer vision. 2017.

- [15]. H. Wang and C. Schmid, "Action Recognition with Improved Trajectories," 2013 IEEE International Conference on Computer Vision, Sydney, NSW, Australia, 2013, pp. 3551-3558, doi: 10.1109/ICCV.2013.441.
- [16]. He, Kaiming, et al. "Deep residual learning for image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.
- [17]. Hochreiter, Sepp, and Jürgen Schmidhuber. "Long short-term memory." Neural computation 9.8 (1997): 1735-1780.
- [18]. Huang, Gao, et al. "Densely connected convolutional networks." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.
- [19]. JHMDB Dataset: <http://jhmdb.is.tue.mpg.de/dataset>
- [20]. Karpathy, Andrej, et al. "Large-scale video classification with convolutional neural networks." Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. 2014.
- [21]. Klaser, Alexander, Marcin Marszałek, and Cordelia Schmid. "A spatio-temporal descriptor based on 3d-gradients." BMVC 2008-19th British Machine Vision Conference. British Machine Vision Association, 2008.
- [22]. Kondratyuk, Dan, et al. "Movinets: Mobile video networks for efficient video recognition." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021.
- [23]. Laptev and Lindeberg, "Space-time interest points," Proceedings Ninth IEEE International Conference on Computer Vision, Nice, France, 2003, pp. 432-439 vol.1, doi: 10.1109/ICCV.2003.1238378.
- [24]. Lee, Hogyun, Seungmin Lee, and Taekyong Nam. "Implementation of high-performance objectionable video classification system." 2006 8th International Conference Advanced Communication Technology. Vol. 2. IEEE, 2006.
- [25]. Li, Shasha, et al. "Adversarial perturbations against real-time video classification systems." arXiv preprint arXiv:1807.00458 (2018).
- [26]. Mallat, Stéphane. A wavelet tour of signal processing. Elsevier, 1999.
- [27]. Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018).
- [28]. MPII Human Pose Dataset: <http://human-pose.mpi-inf.mpg.de/>
- [29]. Oppenheim, Alan V. Discrete-time signal processing. Pearson Education India, 1999.
- [30]. Rahiner, L., and B. Juang. "Fundamentals of Speech Recognition." (1993).
- [31]. Rehman, Atiq, and Samir Brahim Belhaouari. "Deep learning for video classification: A review." (2021).
- [32]. Scovanner, Paul, Saad Ali, and Mubarak Shah. "A 3-dimensional sift descriptor and its application to action recognition." Proceedings of the 15th ACM international conference on Multimedia. 2007.
- [33]. Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." arXiv preprint arXiv:1409.1556 (2014).

- [34]. Slaney, M. "Auditory Toolbox Ver. 2." technical report# 1998-010 (1998).
- [35]. Szegedy, Christian, et al. "Rethinking the inception architecture for computer vision." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.
- [36]. [https://www.tensorflow.org/tutorials/video/video\\_classification](https://www.tensorflow.org/tutorials/video/video_classification)
- [37]. Wyse, L. (2017). Audio spectrogram representations for processing with convolutional neural networks. arXiv preprint arXiv:1706.09559.
- [38]. Wyse, Lonce. "Audio spectrogram representations for processing with convolutional neural networks." arXiv preprint arXiv:1706.09559 (2017).
- [39]. Deng, Jia, et al. "Imagenet: A large-scale hierarchical image database." *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009.
- [40]. Szegedy, Christian, et al. "Inception-v4, inception-resnet and the impact of residual connections on learning." Proceedings of the AAAI conference on artificial intelligence. Vol. 31. No. 1. 2017.
- [41]. Yadav, Ashima, and Dinesh Kumar Vishwakarma. "A unified framework of deep networks for genre classification using movie trailer." *Applied Soft Computing* 96 (2020): 106624.
- [42]. Choi, Eunhye, et al. "Commercial video games and cognitive functions: video game genres and modulating factors of cognitive enhancement." *Behavioral and Brain Functions* 16 (2020): 1-14.
- [43]. Göring, Steve, et al. "Automated genre classification for gaming videos." *2020 IEEE 22nd International Workshop on Multimedia Signal Processing (MMSP)*. IEEE, 2020.
- [44]. Zhang, Zhongping, et al. "Movie genre classification by language augmentation and shot sampling." *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2024.
- [45]. Conneau, Alexis, et al. "Unsupervised cross-lingual representation learning at scale." *arXiv preprint arXiv:1911.02116* (2019).
- [46]. Solovyev, Roman, Alexandr A. Kalinin, and Tatiana Gabruseva. "3D convolutional neural networks for stalled brain capillary detection." *Computers in biology and medicine* 141 (2022): 105089.
- [47]. Kondratyuk, Dan, et al. "Movinets: Mobile video networks for efficient video recognition." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021.
- [48]. Nguyen, Dat Quoc, and Anh Tuan Nguyen. "PhoBERT: Pre-trained language models for Vietnamese." *arXiv preprint arXiv:2003.00744* (2020).
- [49]. Liu, Yinhan, et al. "Roberta: A robustly optimized bert pretraining approach." *arXiv preprint arXiv:1907.11692* (2019).
- [50]. Nguyen-Thuan Do, Phong, et al. "VLUE: A New Benchmark and Multi-task Knowledge Transfer Learning for Vietnamese Natural Language Understanding." *arXiv e-prints* (2024): arXiv-2403.
- [51]. Nguyen, Quoc-Nam, et al. "ViSoBERT: A Pre-Trained Language Model for Vietnamese Social Media Text Processing." *arXiv preprint arXiv:2310.11166* (2023).

- [52]. Tran, Thi Oanh, and Phuong Le Hong. "Improving sequence tagging for Vietnamese text using transformer-based neural models." *Proceedings of the 34th Pacific Asia conference on language, information and computation*. 2020.
- [53]. Sanh, Victor, et al. "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter." *arXiv preprint arXiv:1910.01108* (2019).
- [54]. Chollet, François. "Xception: Deep learning with depthwise separable convolutions." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.
- [55]. Zoph, Barret, et al. "Learning transferable architectures for scalable image recognition." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018.
- [56]. Liu, Zhuang, et al. "A convnet for the 2020s." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022.

**Cơ quan chủ trì**

(ký, họ tên và đóng dấu)

**Chủ nhiệm đề tài**

(ký, họ và tên)