**Đề cương**

**Ứng dụng CNN-LSTM trong phát hiện Deepfake**

Giảng viên:  TS Nguyễn Ngọc Tự

Thành viên:

Đặng Hải Thịnh – 240202014

                    Nguyễn Đức Chí Đạt – 240202005

                    Trương Gia Thạch – 240202012

Mục lục

[**1.** **Giới thiệu chung** 2](#_Toc195804243)

[**2.** **Tóm tắt** 2](#_Toc195804244)

[**3.** **Phương pháp luận** 2](#_Toc195804245)

[**3.1.** **Thách thức trong phát hiện Deepfake trên mạng xã hội** 2](#_Toc195804246)

[**3.2.** **CNN và EfficientNet-B0** 3](#_Toc195804247)

[**3.3.** **Mạng LSTM** 4](#_Toc195804248)

[**3.4.** **Cơ chế tập trung (Attention Mechanism)** 5](#_Toc195804249)

[**3.5.** **Lý do không dùng các kiến trúc thay thế** 5](#_Toc195804250)

[**3.6.** **Tóm tắt** 5](#_Toc195804251)

[**4.** **Kiến trúc** 6](#_Toc195804252)

[**4.1.** **Tổng quan** 6](#_Toc195804253)

[**4.2.** **Tầng 1: Trích xuất dữ liệu** 6](#_Toc195804254)

[**4.3.** **Tầng 2: Tăng cường dữ liệu** 8](#_Toc195804255)

[**4.4.** **Tầng 3: Trích xuất đặc trưng không gian với EfficientNet-B0** 10](#_Toc195804256)

[**4.5.** **Tầng 4: Mô hình hóa thời gian với LSTM và Attention** 11](#_Toc195804257)

[**4.6.** **Tầng 5: Phân loại** 11](#_Toc195804258)

[**5.** **Huấn luyện và phân tích** 11](#_Toc195804259)

[**6.** **Kết luận và định hướng phát triển** 11](#_Toc195804260)

## **Giới thiệu chung**

Trong những năm gần đây, sự phát triển nhanh chóng của công nghệ trí tuệ nhân tạo, đặc biệt là kỹ thuật học sâu (Deep Learning), đã mở ra nhiều khả năng mới trong việc tạo ra các nội dung giả mạo, trong đó nổi bật là công nghệ deepfake. Deepfake cho phép tạo ra các video có khuôn mặt, giọng nói hoặc hành động được làm giả một cách tinh vi đến mức khó có thể phân biệt bằng mắt thường.

Trên các nền tảng mạng xã hội như Facebook, TikTok, YouTube, sự lan truyền của các nội dung deepfake không những ảnh hưởng đến danh tiếng cá nhân mà còn có thể bị lợi dụng để phát tán thông tin sai lệch, gây rối loạn xã hội, tấn công an ninh quốc gia hoặc can thiệp vào chính trị. Trong bối cảnh đó, việc xây dựng một hệ thống phát hiện deepfake chính xác, hiệu quả, có khả năng hoạt động tốt trên các video mạng xã hội là một nhu cầu cấp thiết và mang ý nghĩa thực tiễn cao.

## **Tóm tắt**

Mục tiêu chính của nghiên cứu này là xây dựng một kiến trúc học sâu kết hợp giữa mạng nơ-ron tích chập (CNN) và mạng nơ-ron hồi tiếp (LSTM), cụ thể sử dụng EfficientNet-B0 làm bộ trích xuất đặc trưng không gian, kết hợp với LSTM để học đặc trưng theo chuỗi thời gian, và áp dụng cơ chế chú ý (Attention Mechanism) nhằm nâng cao hiệu quả phát hiện deepfake từ video. Mô hình hướng đến việc xử lý hiệu quả các video ngắn, chất lượng thấp, chứa nhiều biến dạng – đặc trưng phổ biến của nội dung trên mạng xã hội.

## **Phương pháp luận**

Việc phát hiện nội dung deepfake trên mạng xã hội là một trong những thách thức quan trọng và cấp thiết trong lĩnh vực an toàn thông tin số. Phần này tập trung làm rõ cơ sở lý luận, cách tiếp cận kỹ thuật, cũng như giải thích chi tiết về lý do lựa chọn từng thành phần kiến trúc mô hình học sâu bao gồm CNN (với EfficientNet-B0), LSTM, và Attention Mechanism nhằm xây dựng một giải pháp hiệu quả, khả thi và có thể triển khai thực tế.

## **Thách thức trong phát hiện Deepfake trên mạng xã hội**

Mạng xã hội là môi trường phân phối video cực kỳ phổ biến, đồng thời cũng là nơi xuất hiện nhiều nội dung deepfake tinh vi và lan truyền nhanh chóng. Tuy nhiên, việc phát hiện deepfake trên mạng xã hội không hề đơn giản, vì vấp phải nhiều thách thức đặc thù:

* **Chất lượng video thấp, bị nén nhiều lần:** Video được tải lên mạng xã hội thường bị giảm độ phân giải, cắt xén, nén theo chuẩn riêng (như H.264, VP9), làm mất đi các đặc trưng hình ảnh quan trọng.
* **Nội dung ngắn, chuyển cảnh nhanh:** Deepfake trên mạng xã hội thường chỉ kéo dài vài giây, có thể xen lẫn hiệu ứng, filter, cắt ghép nhanh – khiến việc phân tích toàn cục và thời gian gặp khó khăn.
* **Biểu hiện giả mạo rất tinh vi:** Các video giả hiện đại thường không có lỗi rõ rệt như trước (bóp méo mặt, nhấp nháy mắt thiếu tự nhiên), mà sử dụng mô hình học sâu để làm giả biểu cảm, chuyển động, thậm chí đồng bộ môi và giọng.
* **Dữ liệu thật - giả rất giống nhau:** Trong một số trường hợp, ngay cả chuyên gia cũng khó phân biệt video thật và giả nếu không có dấu hiệu trực quan.

Do đó, việc thiết kế mô hình phát hiện deepfake cần đạt được hai mục tiêu chính:

1. Hiệu quả đặc trưng hóa nội dung hình ảnh (không gian)
2. Nắm bắt được đặc trưng thay đổi theo thời gian (chuỗi động học)

## **CNN và EfficientNet-B0**

Convolutional Neural Networks (CNN) là nền tảng không thể thiếu trong các hệ thống phân tích ảnh và video do khả năng học đặc trưng không gian (spatial features) hiệu quả. Trong nghiên cứu này, chúng tôi chọn sử dụng **EfficientNet-B0** làm mạng trích xuất đặc trưng đầu vào nhờ những ưu điểm sau:

**a. Cân bằng giữa độ chính xác và hiệu suất tính toán**

* EfficientNet-B0 được xây dựng dựa trên chiến lược **compound scaling**, tức là mở rộng mô hình theo ba chiều: độ sâu (depth), độ rộng (width) và độ phân giải ảnh (resolution) một cách cân bằng.
* Với chỉ ~5.3 triệu tham số, mô hình này vẫn đạt được độ chính xác trên ImageNet gần tương đương với ResNet-50 nhưng yêu cầu ít tài nguyên hơn, rất phù hợp cho hệ thống xử lý nhiều video song song.

**b. Khả năng tổng quát hóa cao**

* EfficientNet được huấn luyện trên tập ImageNet lớn, do đó có khả năng học được nhiều đặc trưng thị giác quan trọng, giúp mô hình có năng lực phân biệt cao kể cả trong bối cảnh dữ liệu có nhiễu như mạng xã hội.

**c. Thích nghi tốt với ảnh khuôn mặt**

* Kích thước ảnh đầu vào của EfficientNet-B0 (224x224) phù hợp với kích thước trung bình của ảnh mặt được crop từ video.
* Với các khối MBConv, mô hình có khả năng học các đặc trưng ở nhiều cấp độ (từ cạnh biên, kết cấu da, đến biểu cảm toàn cục).

**d. Khả năng tích hợp với pipeline thời gian thực**

* Do số lượng tham số nhỏ và tốc độ xử lý cao, EfficientNet-B0 có thể chạy ổn định trên GPU trung bình (như RTX 2060 hoặc T4), từ đó dễ dàng tích hợp vào các hệ thống giám sát mạng xã hội trực tiếp.

## **Mạng LSTM**

Deepfake không chỉ là vấn đề phân tích từng frame riêng lẻ, mà còn nằm ở sự thay đổi động theo thời gian – ví dụ:

* Chuyển động mắt không đồng nhất.
* Cử động môi không khớp với lời nói.
* Nhịp chớp mắt bất thường.
* Sự lệch pha ánh sáng khi quay đầu.

Do đó, sử dụng mạng học chuỗi thời gian là cần thiết. Chúng tôi lựa chọn LSTM (Long Short-Term Memory) vì các lý do sau:

**a. Khả năng ghi nhớ trạng thái dài hạn**

* LSTM khắc phục hiện tượng mất thông tin (vanishing gradient) trong chuỗi dài mà các RNN truyền thống gặp phải.
* Nhờ đó, mô hình có thể ghi nhớ biểu hiện chuyển động ở nhiều frame liên tiếp.

**b. Giữ được thứ tự thời gian**

* Thứ tự của các biểu cảm rất quan trọng: ví dụ, nếu mô hình thấy một mắt nhắm mà không thấy khớp với frame trước, nó có thể đánh giá đây là bất thường.
* LSTM không cần xáo trộn frame như trong một số mô hình CNN 3D, do đó duy trì được ngữ cảnh động học.

**c. Cấu trúc đơn giản, dễ huấn luyện**

* So với Transformer-based video models, LSTM có cấu trúc gọn nhẹ, phù hợp với các tập dữ liệu trung bình và môi trường có tài nguyên giới hạn.

**d. Khả năng kết hợp hiệu quả với CNN**

* Output từ EfficientNet-B0 trên mỗi frame là vector đặc trưng (feature vector) có thể được đưa trực tiếp vào chuỗi LSTM.
* Do đó, kiến trúc kết hợp CNN-LSTM trở thành lựa chọn tự nhiên và hiệu quả.

## **Cơ chế tập trung (Attention Mechanism)**

Cơ chế tập trung (attention) giúp mô hình xác định các khung hình có **trọng số thông tin cao**, nhờ đó tập trung vào những phần có khả năng chứa dấu hiệu deepfake rõ nhất. Trong video ngắn trên mạng xã hội, không phải mọi khung hình đều quan trọng như nhau:

* Một số frame chỉ chứa cảnh tĩnh, hoặc không rõ mặt.
* Một số frame lại hiển thị rõ ràng méo mó vùng mắt, mồm hoặc các lỗi deepfake.

Mô hình Attention sẽ học trọng số cho mỗi bước thời gian , từ đó tính tổng có trọng số cho toàn chuỗi:

trong đó ​ là vector đặc trưng tại thời điểm , và ​ được học thông qua mạng phụ (attention layer).

Lợi ích:

* **Loại bỏ frame nhiễu:** Giảm ảnh hưởng của frame không chứa mặt hoặc bị biến dạng.
* **Tăng độ chính xác phân loại:** Hướng trọng số vào những biểu hiện đáng nghi trong chuỗi thời gian.
* **Giải thích được mô hình:** Có thể trực quan hóa attention map để hiểu vì sao mô hình đưa ra dự đoán.

## **Lý do không dùng các kiến trúc thay thế**

Một số phương án thay thế như CNN 3D (C3D, I3D), hoặc transformer-based (ViViT, TimeSformer) đã được cân nhắc, nhưng chưa phù hợp trong giai đoạn đầu do:

* **C3D/I3D**: Yêu cầu nhiều GPU và tài nguyên RAM; khó mở rộng nếu frame input không cố định.
* **Transformers**: Hiệu quả cao nhưng huấn luyện chậm, yêu cầu batch size lớn, không hiệu quả với dữ liệu nhỏ hoặc phân bố lệch.
* **GRU**: Nhẹ hơn LSTM nhưng hiệu suất thường thấp hơn trong bài toán chuỗi phức tạp như deepfake.

## **Tóm tắt**

Kiến trúc đề xuất (EfficientNet-B0 + LSTM + Attention) có những đặc điểm giúp nó phù hợp với môi trường mạng xã hội:

| **Yêu cầu** | **Đặc điểm mô hình đề xuất** |
| --- | --- |
| Xử lý video ngắn, phân giải thấp | EfficientNet trích đặc trưng tốt ở ảnh nhỏ |
| Nhận diện biểu cảm động | LSTM học được sự biến thiên thời gian |
| Phân biệt frame quan trọng | Attention chọn lọc frame chứa tín hiệu mạnh |
| Hiệu quả tính toán | Mô hình nhỏ gọn, có thể triển khai nhanh |
| Khả năng mở rộng | Có thể nâng cấp lên EfficientNet-B2, multi-head attention nếu cần |

## **Kiến trúc**

## **Tổng quan**

## **Tầng 1: Trích xuất dữ liệu**

* + 1. **Mục tiêu giai đoạn trích xuất dữ liệu**

Trong bối cảnh dữ liệu video trên mạng xã hội có đặc điểm không đồng nhất về chất lượng, độ phân giải, và cách thức giả mạo, việc trích xuất dữ liệu đóng vai trò then chốt trong việc chuẩn hóa đầu vào cho mô hình học sâu. Mục tiêu chính của giai đoạn này là:

* **Trích xuất khung hình tiêu biểu từ video** với mật độ vừa đủ.
* **Xác định và cắt khuôn mặt** có chất lượng cao nhất từ mỗi khung hình.
* **Đánh giá và lọc khuôn mặt** dựa trên các tiêu chí: độ rõ nét (blur), độ chính diện (frontal), kích thước, và sự phù hợp đặc trưng hình thái (feature).
* **Ghi nhãn (label)** dữ liệu khuôn mặt cho tập huấn luyện.
  + 1. **Kiến trúc trích xuất và pipeline thực thi**

Pipeline trích xuất dữ liệu được xây dựng theo hướng song song hóa và tối ưu tài nguyên GPU. Các bước chính bao gồm:

**Bước 1: Khởi tạo thư mục và mô hình**

* Tạo các thư mục để lưu trữ khung hình trích xuất từ video real và fake.
* Khởi tạo bộ phát hiện khuôn mặt chất lượng cao InsightFace (buffalo\_l) với hỗ trợ CUDAExecutionProvider.
* Khởi tạo mô hình DNN (OpenCV) sử dụng Caffe model (proto + weights) để phát hiện nhanh các khung hình có khả năng chứa khuôn mặt.

**Bước 2: Quét tìm các khung hình có mặt người**

* Dựa trên tổng số khung hình, lựa chọn các khung hình cách đều để giảm thiểu trùng lặp và tối ưu tài nguyên.
* Với mỗi batch khung hình nhỏ (300x300), áp dụng DNN để phát hiện nhanh các vị trí có khả năng chứa khuôn mặt.
* Lưu trữ khung hình ứng viên (đầy đủ độ phân giải) có độ tin cậy cao (dựa trên confidence của DNN).

**Bước 3: Phát hiện khuôn mặt chất lượng cao**

* Với mỗi khung hình có mặt người, chuyển sang không gian RGB và áp dụng InsightFace để phát hiện tất cả các khuôn mặt.
* Với mỗi khuôn mặt phát hiện, đánh giá chất lượng dựa trên ba chỉ số:

| **Tiêu chí** | **Cách tính** |
| --- | --- |
| frontal\_score | = -(yaw + pitch), càng nhỏ càng tốt (gần chính diện) |
| size\_score | = diện tích khuôn mặt (width × height) |
| feature\_score | = điểm hình thái học dựa trên vị trí tương đối của mắt, mũi, miệng |

* Chuẩn hóa (normalize) điểm số và tính norm\_score tổng hợp để lựa chọn khuôn mặt tốt nhất.

**Bước 4: Kiểm tra độ mờ và cắt khuôn mặt**

* Với khuôn mặt tốt nhất, thực hiện đánh giá chỉ số blur\_score (độ sắc nét) bằng phương pháp Laplacian variance.
* Nếu đạt ngưỡng sắc nét tối thiểu (ví dụ: > 8), tiến hành cắt khuôn mặt từ ảnh gốc, resize về kích thước chuẩn (ví dụ: 224x224 hoặc 160x160), và lưu trữ với tên định danh duy nhất.

**Bước 5: Lưu nhãn và xử lý bất đồng bộ**

* Ghi lại thông tin nhãn (filename, label) cho mỗi ảnh khuôn mặt trích xuất.
* Sử dụng AsyncWriter để thực hiện lưu ảnh bất đồng bộ nhằm giảm tắc nghẽn I/O.
* Cập nhật tiến trình vào các file checkpoint nhằm tránh trích xuất lại khi có gián đoạn.
  + 1. **Ưu điểm nổi bật của thiết kế**
* **Tối ưu hóa GPU + CPU:** InsightFace sử dụng GPU để đảm bảo chất lượng phát hiện khuôn mặt, trong khi DNN (OpenCV) thực hiện chọn lọc sơ bộ trên CPU nhằm tiết kiệm tài nguyên.
* **Chiến lược chọn khung hình dựa trên confidence:** Giảm đáng kể thời gian xử lý trên video dài hoặc nhiễu.
* **Chấm điểm khuôn mặt đa chiều:** Không chỉ dựa vào kích thước hay độ chính diện, mà còn xem xét cả hình thái học và độ sắc nét.
* **Cắt ghép an toàn:** safe\_crop() đảm bảo không bị lỗi out-of-bound khi crop hình từ ảnh gốc.
* **Bất đồng bộ lưu ảnh:** Ghi ảnh và nhãn song song với xử lý chính, tăng hiệu năng và giảm thời gian trễ.
  + 1. **Minh họa pipeline xử lý**

[Video] → [Quét tìm khung hình có mặt người bằng DNN]

→ [Phát hiện mặt bằng InsightFace]

→ [Chấm điểm mặt] → [Chọn mặt tốt nhất] → [Kiểm tra độ mờ]

→ [Crop & Resize] → [Lưu ảnh & nhãn] → [CSV Label]

## **Tầng 2: Tăng cường dữ liệu**

* + 1. **Mục tiêu giai đoạn tăng cường dữ liệu**

Dữ liệu huấn luyện cho bài toán phát hiện video giả mạo (deepfake) thường có sự không cân bằng giữa các lớp (real và fake) và khác biệt về số lượng khung hình trích xuất giữa các video. Để cải thiện tính tổng quát và độ ổn định của mô hình học sâu, hệ thống áp dụng kỹ thuật tăng cường dữ liệu (augmentation) với hai mục tiêu chính:

* Bổ sung khung hình còn thiếu cho từng video nhằm chuẩn hóa đầu vào về số lượng.
* Cân bằng tỷ lệ giữa video thật (real) và video giả (fake) bằng cách tạo thêm dữ liệu tổng hợp từ lớp thiểu số.
  + 1. **Chiến lược và kỹ thuật tăng cường dữ liệu**

a. Tăng cường ảnh ngẫu nhiên (Random Augmentation)

Một tập hợp các phép biến đổi được áp dụng ngẫu nhiên cho ảnh khuôn mặt nhằm tạo ra sự đa dạng trong không gian đặc trưng:

| **Phép biến đổi** | **Mô tả** |
| --- | --- |
| Lật ngang (flip) | Xác suất 50% |
| Thay đổi độ sáng/độ tương phản | Hệ số alpha ∈ [0.7, 1.3], độ lệch sáng β ∈ [-25, 25] |
| Xoay ảnh (rotation) | Góc xoay ngẫu nhiên trong khoảng [-15°, +15°] |
| Phóng to/thu nhỏ (scaling) và tịnh tiến (translation) | Phép affine: scale ∈ [0.9, 1.1], dịch chuyển ±5% kích thước ảnh |

**b. Tiền xử lý và nhóm ảnh theo video**

* Mỗi ảnh khuôn mặt được ánh xạ về một group tương ứng với video gốc (dựa vào tên ảnh).
* Các ảnh trong cùng một group sẽ được xử lý tập thể, đảm bảo rằng nếu một video thiếu khung hình, hệ thống sẽ tự động bổ sung số lượng ảnh còn thiếu bằng augmentation từ các khung hình có sẵn.

**c. Cân bằng lớp (Real vs Fake)**

Sau bước bổ sung, hệ thống kiểm tra tỷ lệ giữa tổng số ảnh real và fake:

* Nếu số ảnh real < fake, hệ thống sẽ sinh thêm ảnh real bằng cách chọn ngẫu nhiên ảnh real hiện có và áp dụng augmentation.
* Các ảnh cân bằng được lưu trữ riêng trong thư mục balanced để thuận tiện trong quản lý.
  + 1. **Ưu điểm nổi bật của thiết kế**
* Đảm bảo nhất quán đầu vào: Mỗi video đầu vào sẽ cung cấp đúng số lượng khung hình đầu vào (FRAME\_COUNT) cho mạng LSTM sau này.
* Tăng đa dạng hóa mẫu huấn luyện: Giúp mô hình học sâu học được các đặc trưng biến thể về hình thái, ánh sáng và tư thế khuôn mặt.
* Tự động cân bằng dữ liệu: Hệ thống không cần thủ công kiểm tra tỷ lệ nhãn trước huấn luyện.
  + 1. **Minh họa pipeline xử lý**

[Ảnh từ CSV gốc] → [Nhóm ảnh theo video] → [Phát hiện group thiếu ảnh]

→ [Tăng cường ảnh còn thiếu] → [Cân bằng lớp real/fake nếu cần]

→ [Lưu ảnh mới + cập nhật nhãn vào CSV]

## **Tầng 3: Trích xuất đặc trưng không gian với EfficientNet-B0**

* + 1. **Mục tiêu giai đoạn trích xuất đặc trưng**

EfficientNet-B0 được sử dụng như một bộ **feature extractor** – trích xuất các đặc trưng không gian từ từng khung hình chứa khuôn mặt đã được phát hiện và cắt sẵn. Các đặc trưng này sẽ được truyền vào khối LSTM để khai thác quan hệ thời gian giữa các khung hình liên tiếp.

* + 1. **Cấu trúc chi tiết**

EfficientNet-B0 có tổng cộng 9 khối chính (stages), mỗi khối gồm nhiều lớp convolution kết hợp với Mobile Inverted Bottleneck Convolution (MBConv) – kỹ thuật giúp giảm số lượng tham số trong khi vẫn duy trì chất lượng biểu diễn đặc trưng.

| **Tầng (Stage)** | **Cấu trúc** | **Kích thước đầu vào (HxW)** | **Số lượng output channels** | **Đặc điểm** |
| --- | --- | --- | --- | --- |
| Stem | Conv3x3, stride 2 | 128×128 | 32 | Trích xuất đặc trưng cơ bản |
| Stage 1 | MBConv1, k=3 | 64×64 | 16 | Giai đoạn nhẹ, duy trì thông tin biên |
| Stage 2 | MBConv6, k=3 | 32×32 | 24 | Bắt đầu tăng kênh, giảm không gian |
| Stage 3 | MBConv6, k=5 | 16×16 | 40 | Giai đoạn trung gian, trích xuất họa tiết phức tạp |
| Stage 4 | MBConv6, k=3 | 16×16 | 80 | Độ sâu tăng lên, học biểu diễn trừu tượng hơn |
| Stage 5 | MBConv6, k=5 | 8×8 | 112 | Cân bằng giữa chi tiết và tổng quát |
| Stage 6 | MBConv6, k=5 | 8×8 | 192 | Giai đoạn sâu, học đặc trưng ngữ nghĩa |
| Stage 7 | MBConv6, k=3 | 4×4 | 320 | Giai đoạn cuối, tóm tắt đặc trưng không gian |
| Head | Conv1x1 + Pooling | 4×4 | 1280 (flatten) | Đầu ra vector đặc trưng cuối cùng |

* + 1. **Các kỹ thuật tối ưu hóa**

 **Swish Activation**: Hàm kích hoạt Swish (x \* sigmoid(x)) được sử dụng thay cho ReLU nhằm duy trì gradient tốt hơn trong vùng âm, giúp huấn luyện ổn định.

 **Depthwise Separable Convolution**: Tách convolution thành 2 bước – convolution theo kênh và sau đó là kết hợp – giúp giảm đáng kể số lượng tham số.

 **Squeeze-and-Excitation (SE) Block**: Cơ chế chú ý theo kênh, giúp mô hình tập trung vào những kênh đặc trưng quan trọng. SE block xuất hiện trong hầu hết các MBConv block.

* + 1. **Ưu điểm nổi bật**

| **Tiêu chí** | **Lý do lựa chọn EfficientNet-B0** |
| --- | --- |
| Hiệu suất cao / tham số thấp | Giảm thời gian huấn luyện và bộ nhớ khi xử lý video |
| Kích thước đặc trưng phù hợp | Vector 1280 chiều là đủ biểu diễn mà không quá dư thừa |
| Đầu vào ảnh kích thước nhỏ | Phù hợp với ảnh khuôn mặt (128×128) |
| Dễ kết hợp với LSTM | Có thể trích xuất đặc trưng tuần tự từng khung hình |
| Tương thích attention | Các attention maps có thể trực quan hóa dễ dàng |

## **Tầng 4: Mô hình hóa thời gian với LSTM và Attention**

## **Tầng 5: Phân loại**

## **Huấn luyện và phân tích**

## **Kết luận và định hướng phát triển**