

Phát hiện hành vi bạo lực (Detecting violent behavior)

Thành viên:

Nguyễn Thanh Tùng Lê Văn Vượng Nguyễn Quang Thịnh

Dai Nam University

Dai Nam University

Dai Nam University

MSV: 1771029729

MSV: 1771020771

MSV: 1771020650

Ngày 11 tháng 3 năm 2025

Tóm tắt

Phát hiện hành vi bạo lực là một trong những bài toán quan trọng trong thị giác máy tính và an ninh. Bằng cách ứng dụng mô hình mạng nơ-ron tích chập (CNN), chúng tôi nhằm phát triển mô hình tự động nhận diện các hành vi bạo lực từ dữ liệu video. Trong nghiên cứu này, chúng tôi áp dụng hai loại mô hình CNN: CNN 2D (xử lý khung hình) và CNN 3D (xử lý chuỗi video) sử dụng TensorFlow. Kết quả thử nghiệm cho thấy CNN 3D đạt độ chính xác cao hơn với 79.23%, trong khi CNN 2D đạt 72.45

Mục tiêu chính của nghiên cứu là hiểu rõ hơn về mạng nơ-ron tích chập và tối ưu hóa mô hình nhận diện hành vi bạo lực. Chúng tôi tinh chỉnh các siêu tham số như tốc độ học, dropout, batch normalization và pooling. Cuối cùng, chúng tôi phân tích để tìm mối quan hệ giữa số lớp và độ chính xác. Kết quả nghiên cứu cung cấp thông tin quan trọng cho việc xây dựng hệ thống giám sát tự động và an ninh.

1. Giới thiệu

Bạo lực là vấn đề nghiêm trọng ở nhiều quốc gia, gây ảnh hưởng tiêu cực đến xã hội, an ninh và chất lượng cuộc sống. Hành vi bạo lực có thể xảy ra ở nhiều môi trường khác nhau như trường học, nơi làm việc, đường phố, và các sự kiện thể thao. Việc phát triển hệ thống nhận diện hành vi bạo lực trong video không chỉ hỗ trợ giám sát an ninh mà còn có tiềm năng ứng dụng trong nhiều lĩnh vực như bảo vệ công cộng, kiểm soát giao

thông, và hỗ trợ pháp lý.

Các nghiên cứu trước đây chủ yếu dựa vào các phương pháp trích xuất đặc trưng truyền thống như HOG (Histogram of Oriented Gradients), optical flow, và các mô hình học máy truyền thống như SVM (Support Vector Machine). Tuy nhiên, những phương pháp này có hạn chế về độ chính xác do khó khăn trong việc tổng quát hóa và khả năng nhận diện các hành vi phức tạp.

Gần đây, các mô hình học sâu, đặc biệt là CNN, đã chứng minh hiệu quả vượt trội trong các bài toán nhận dạng hình ảnh và video. CNN có khả năng học các đặc trưng không gian từ dữ liệu, giúp cải thiện độ chính xác trong nhận diện hành vi bạo lực. Trong nghiên cứu này, chúng tôi sử dụng CNN 2D để phân tích khung hình tĩnh và CNN 3D để khai thác thông tin theo thời gian, giúp hệ thống có khả năng nhận diện chính xác hơn.

Tập dữ liệu được sử dụng trong nghiên cứu là Hockey Fight Videos, một tập dữ liệu chứa các video ghi lại cảnh đánh nhau trong các trận đấu khúc côn cầu trên băng. Tập dữ liệu này bao gồm hai nhóm: video có hành vi bạo lực (fights) và video không có hành vi bạo lực (non-fights). Đây là một tập dữ liệu phổ biến trong nghiên cứu phát hiện bạo lực, giúp kiểm tra khả năng tổng quát hóa của mô hình.

Trong bài báo này, chúng tôi đánh giá hiệu suất của CNN 2D và CNN 3D trên tập dữ liệu Hockey Fight Videos, thực hiện thử nghiệm trên nhiều kiến trúc khác nhau để tìm ra mô hình tối ưu nhất. Bên cạnh đó, chúng tôi cũng phân tích tác động của các siêu tham số quan trọng đến kết

quả dự đoán và đề xuất hướng cải tiến cho các nghiên cứu trong tương lai.

2. Công việc liên quan

2.1 Nhận diện bạo lực trong video

Nhận diện bạo lực trong video là một chủ đề quan trọng trong lĩnh vực thị giác máy tính, đặc biệt trong các ứng dụng giám sát an ninh và bảo vệ an toàn công cộng. Các phương pháp truyền thống trước đây chủ yếu dựa vào trích xuất đặc trưng thủ công kết hợp với các thuật toán máy học để phát hiện hành vi bạo lực. Một số phương pháp phổ biến bao gồm:

+ SVM (Support Vector Machine): Một số nghiên cứu đã áp dụng thuật toán SVM kết hợp với đặc trưng HOG (Histogram of Oriented Gradients) để nhận diện bạo lực trong video. Phương pháp này có độ chính xác tốt trên các bộ dữ liệu nhỏ nhưng gặp khó khăn khi làm việc với dữ liệu phức tạp hoặc có nhiều yếu tố gây nhiễu.

+ Optical Flow: Kỹ thuật này phân tích trường vận tốc của pixel để phát hiện chuyển động bất thường trong video. Optical Flow thường được sử dụng trong các hệ thống giám sát nhằm phát hiện các hành vi đáng ngờ, tuy nhiên phương pháp này nhạy cảm với nhiễu như thay đổi ánh sáng, góc quay và độ phân giải của video.

Mặc dù các phương pháp trên đã đạt được một số thành công nhất định, chúng vẫn tồn tại nhiều hạn chế như:

+ Cần trích xuất đặc trưng thủ công: Việc xác định đặc trưng quan trọng trong video đòi hỏi nhiều công sức và phụ thuộc vào chuyên môn của người thiết kế.

+ Không hiệu quả trên dữ liệu lớn: Các phương pháp này hoạt động kém trên các video có nhiều yếu tố gây nhiễu, đặc biệt khi dữ liệu có sự thay đổi mạnh về góc quay, điều kiện ánh sáng hoặc nội dung phức tạp.

2.2 Sử dụng Deep Learning để phát hiện bạo lực

Những năm gần đây, Deep Learning đã trở thành một phương pháp tiếp cận hiệu quả trong nhận diện bạo lực nhờ khả năng học đặc trưng tự động từ dữ liệu mà không cần trích xuất thủ công. Một số mô hình phổ biến bao gồm:

+ CNN 2D (Convolutional Neural Network 2D): Các mô hình CNN 2D như VGG16, ResNet50 được áp dụng để phân loại hành vi bạo lực dựa trên từng khung hình riêng lẻ. Tuy nhiên, nhược điểm chính của phương pháp này là không xem xét mối quan hệ giữa các khung hình theo thời gian, dẫn đến mất đi thông tin về chuyển động.

+ CNN 3D (Convolutional Neural Network 3D): Các mô hình như C3D (Convolutional 3D) hoặc I3D (Inflated 3D ConvNet) sử dụng tích chập 3D để học đặc trưng không gian-thời gian trong video. Điều này giúp phân biệt chính xác hơn các hành vi bạo lực dựa trên chuỗi chuyển động liên tiếp thay vì chỉ dựa vào một khung hình đơn lẻ.

So với các phương pháp truyền thống, Deep Learning có nhiều ưu điểm nổi bật:

+ Tự động học đặc trưng: Không cần trích xuất thủ công, giúp giảm bớt sai sót do con người.

+ Khả năng tổng quát tốt hơn: Hoạt động hiệu quả trên các bộ dữ liệu lớn, có khả năng nhận diện chính xác cả khi video có sự thay đổi về góc quay, độ phân giải hoặc ánh sáng.

2.3 Phương pháp tiếp cận của nghiên cứu này

Trước hết ta cần hiểu CNN là gì ?

- CNN, viết tắt của Convolutional Neural Network (Mạng Nơ-ron Tích chập), là một loại mạng nơ-ron nhân tạo được thiết kế đặc biệt để xử lý và phân tích dữ liệu có cấu trúc lưới, chẳng hạn như hình ảnh, video và dữ liệu âm thanh. CNN đã đạt được những thành công đáng kể trong nhiều lĩnh vực, đặc biệt là trong thị giác máy tính, nhờ khả năng tự động học các đặc trưng phức tạp từ dữ liệu đầu vào.

Được biểu diễn Công thức:

$$S(i, j) = (I * K)(i, j) = \sum_m \sum_n I(i + m, j + n)K(m, n)$$

Trong nghiên cứu này, chúng tôi tập trung vào việc so sánh hiệu suất của CNN 2D và CNN 3D trên cùng một bộ dữ liệu Hockey Fight Videos để đánh giá khả năng phát hiện hành vi bạo lực trong video thể thao.

Cụ thể, các bước thực hiện như sau:

- **Áp dụng CNN 2D (VGG16)** để nhận diện hành vi bạo lực dựa trên từng khung hình riêng lẻ.
- **Áp dụng CNN 3D (C3D)** để học đặc trưng không gian-thời gian, từ đó nhận diện bạo lực dựa trên chuỗi chuyển động trong video.

- **Sử dụng TensorFlow/Keras** để xây dựng, huấn luyện và đánh giá mô hình.
- **Đánh giá kết quả** bằng các chỉ số Accuracy, Precision, Recall và F1-score để xác định mô hình nào phù hợp hơn cho bài toán phát hiện bạo lực.

Nghiên cứu này không chỉ giúp đánh giá hiệu suất của các mô hình Deep Learning trong phát hiện bạo lực mà còn cung cấp hướng đi cho việc triển khai các hệ thống giám sát tự động trong thực tế.

3. Dataset

3.1 Giới thiệu bộ dữ liệu Hockey Fight

Bộ dữ liệu Hockey Fight Videos là một trong những tập dữ liệu phổ biến trong nghiên cứu nhận diện hành vi bạo lực trong video. Bộ dữ liệu này được thiết kế để phân loại video thành hai nhóm chính:

+ Fight (Bạo lực): Bao gồm các video có cảnh đánh nhau giữa các vận động viên trong các trận đấu khúc côn cầu.

+ Non-Fight (Không bạo lực): Bao gồm các video diễn biến bình thường của trận đấu mà không có hành vi bạo lực.



Hình 1



Hình 2



Hình 3



Hình 4

Bộ dữ liệu này chứa tổng cộng 1.000 video, với số lượng gần như cân bằng giữa hai lớp Fight và Non-Fight. Mỗi video có độ dài trung bình từ 3 đến 10 giây, với độ phân giải đa dạng.

Bộ dữ liệu này phù hợp với nghiên cứu nhận diện bạo lực do:

+ Tính thực tế cao: Các video được thu thập từ các trận đấu thể thao thực tế.

+ Mức độ phức tạp: Hành vi bạo lực xảy ra trong môi trường động, có nhiều yếu tố gây nhiễu như khán giả, chuyển động nhanh của cầu thủ và thay đổi góc quay camera.

+ Tính cân bằng: Số lượng video giữa hai lớp gần như bằng nhau, giúp giảm thiểu bias trong quá trình huấn luyện mô hình.

3.1.1 Bộ dữ liệu tổng hợp các hành động bạo lực thực tế đã thu thập được

Ngoài bộ dữ liệu Hockey Fight, chúng tôi còn tự thu thập và gán nhãn một tập dữ liệu riêng gồm các hành động bạo lực từ nhiều nguồn khác nhau. Bộ dữ liệu này bao gồm:

+ Các cảnh bạo lực từ các sự kiện thể thao khác ngoài khúc côn cầu, chẳng hạn như bóng đá, bóng rổ, quyền anh, MMA.

+ Video từ các hệ thống giám sát an ninh ghi lại các hành vi xô xát, đánh nhau ở nơi công cộng.

+ Dữ liệu từ các bộ phim, chương trình truyền hình có cảnh bạo lực thực tế.

+ Video trích xuất từ mạng xã hội, các nền tảng phát trực tuyến có nội dung liên quan đến hành vi bạo lực.

Quá trình thu thập dữ liệu này được thực hiện theo một số tiêu chí sau:

1. Mỗi video đều được xem xét kỹ lưỡng để đảm bảo phân loại chính xác.
2. Gán nhãn dữ liệu bằng cách phân loại thành hai nhóm chính: Bạo lực (Fight) và Không bạo lực (Non-Fight).
3. Lọc bỏ các video có chất lượng kém, không rõ nội dung để đảm bảo độ chính xác của mô hình.
4. Đảm bảo tính đa dạng của dữ liệu, bao gồm nhiều góc quay, điều kiện ánh sáng, và đối tượng khác nhau.

Việc kết hợp bộ dữ liệu Hockey Fight với tập dữ liệu tự thu thập giúp mở rộng phạm vi ứng dụng của mô hình, tăng tính tổng quát và khả năng nhận diện hành vi bạo lực trong nhiều ngữ cảnh thực tế hơn.

Việc kết hợp bộ dữ liệu Hockey Fight với tập dữ liệu tự thu thập giúp mở rộng phạm vi ứng dụng của mô hình, tăng tính tổng quát và khả năng nhận diện hành vi bạo lực trong nhiều ngữ cảnh thực tế hơn.

3.2 Tiền xử lý dữ liệu

Trước khi đưa dữ liệu vào mô hình, cần thực hiện một số bước tiền xử lý nhằm tối ưu hiệu suất huấn luyện và giảm thiểu nhiễu.

3.2.1 Trích xuất khung hình từ video

Do CNN hoạt động chủ yếu trên ảnh tĩnh, ta cần trích xuất khung hình từ video. Tuy nhiên, không phải tất cả các khung hình đều quan trọng như nhau, do đó ta trích xuất theo tần số FPS = 20 (20 khung hình mỗi giây) đây là số lượng khung hình vừa đủ để vừa tối ưu hiệu suất vừa tránh tình trạng học quá nhiều của mô hình.

3.2.2 Chuẩn hóa kích thước ảnh

Các video trong bộ dữ liệu có độ phân giải khác nhau, do đó các khung hình trích xuất được cần được resize về kích thước 224x224 pixels nhằm đồng bộ đầu vào cho mô hình. Đây là kích thước chuẩn thường được sử dụng trong các mô hình CNN như ResNet, VGG, MobileNet.

3.2.3 Chuẩn hóa dữ liệu

Chuyển đổi ảnh về dạng số: Các khung hình được chuyển thành mảng số (numpy array) với các giá trị pixel trong khoảng [0, 255].

Chuẩn hóa dữ liệu: Mỗi pixel được chia cho 255 để đưa về khoảng [0, 1], giúp tăng hiệu quả huấn luyện và tránh hiện tượng gradient vanishing/exploding.

3.2.4 Chia tập dữ liệu

Dữ liệu gồm 2 phần chính:

Tập dữ liệu	Tỉ lệ	Số lượng video
Training	80	800
Testing	20	200

+ Tập huấn luyện (Training set): Sử dụng để mô hình học các đặc trưng quan trọng của hành vi bạo lực.

+ Tập kiểm tra (Testing set): Dùng để đánh giá hiệu suất mô hình trên dữ liệu mới.

4. Phương thức

Trong nghiên cứu này, chúng tôi sử dụng hai phương pháp chính để phát hiện hành vi bạo lực từ video: CNN 2D và CNN 3D. Mô hình CNN 2D sử dụng mạng VGG16 để xử lý từng khung hình riêng lẻ, trong khi đó CNN 3D học các đặc trưng không gian-thời gian từ chuỗi khung hình liên tiếp, có nghĩa là CNN 3D có thể trích xuất thay vì từng khung hình riêng lẻ nó có thể trích xuất một chuỗi khung hình liên tiếp và xử lý luôn chuỗi đó.

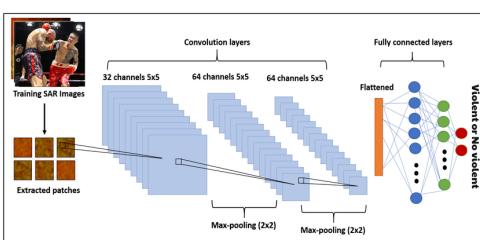
4.1 Mô hình CNN 2D

4.1.1 Đầu vào của mô hình

Dữ liệu đầu vào: Ảnh đơn có kích thước 224x224x3 (RGB). Tiền xử lý: Ảnh được trích xuất từ video với FPS = 20, sau đó được resize về 224x224 pixel và chuẩn hóa.

4.1.2 Cấu trúc mô hình CNN 2D

Mô hình CNN 2D được sử dụng trong nghiên cứu này dựa trên kiến trúc VGG16, một mô hình mạnh mẽ đã được huấn luyện trước trên tập dữ liệu ImageNet. VGG16 được lựa chọn vì khả năng trích xuất đặc trưng hiệu quả từ từng khung hình của video.



Cấu trúc mô hình bao gồm các thành phần sau:

1. Lớp Tích chập (Convolutional Layers):

- Các lớp tích chập (ví dụ: block1_conv1, block2_conv1, ...) có nhiệm vụ học các đặc trưng cơ bản từ hình ảnh đầu vào, như đường nét, góc cạnh và cấu trúc hình học.
- Công thức tích chập được biểu diễn như sau:

$$S(i, j) = (I * K)(i, j) = \sum_m \sum_n I(i+m, j+n)K(m, n)$$

- Trong đó:

- $S(i, j)$ là giá trị của bản đồ đặc trưng đầu ra tại vị trí (i, j) .
- I là ma trận đầu vào (khung hình video).
- K là bộ lọc (kernel) hoặc ma trận trọng số.
- $*$ là toán tử tích chập.
- m và n là các chỉ số chạy qua kích thước của bộ lọc K .

2. Lớp Gộp (Pooling Layers):

- Các lớp gộp (ví dụ: block1_pool, block2_pool, ...) giúp giảm kích thước dữ liệu và tập trung vào các đặc trưng quan trọng nhất.
- Gộp cực đại (Max Pooling) là phương pháp gộp phổ biến được sử dụng trong VGG16.

3. Lớp Kết Nối Đầy Đủ (Fully Connected Layers):

- Sau khi trải qua các lớp tích chập và gộp, các đặc trưng được đưa vào các lớp kết nối đầy đủ (ví dụ: fc1, fc2).
- Lớp fc2 trong VGG16 là một lớp kết nối đầy đủ, nơi mô hình học các đặc trưng phức tạp hơn, tổng hợp thông tin từ các lớp trước đó.
- Đây là nơi mô hình học các mối quan hệ phức tạp giữa các đặc trưng, như chuyển động của người, hành động tấn công, xô đẩy, ...

Link mô hình: <https://colab.research.google.com/drive/1NGqUrKYbvkkY06AT3bQ7sRaDahw8kf38?usp=sharing>

4.2 Mô hình CNN 3D

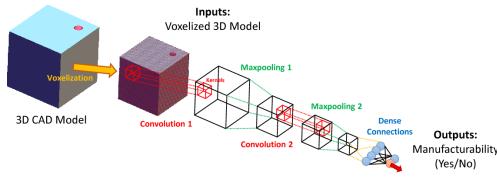
4.2.1 Đầu vào của mô hình

Dữ liệu đầu vào: Chuỗi 16 khung hình liên tiếp, mỗi khung có kích thước 224x224x3 (RGB).

Tiền xử lý:

- Trích xuất khung hình từ video.
- Chọn 16 khung hình liên tiếp để tạo một chuỗi không gian-thời gian.
- Resize về 224x224 pixels và chuẩn hóa.

4.2.2 Cấu trúc mô hình



Mô hình CNN 3D được xây dựng để phân loại video thành bạo lực hoặc không bạo lực, sử dụng các lớp tích chập 3D để học đặc trưng không gian và thời gian từ chuỗi khung hình.

Cấu trúc mô hình thực tế cũng tương tự như VGG16, tuy nhiên điểm khác biệt là thay vì các lớp tích chập và gộp 2D, chúng ta sử dụng các lớp tích chập và gộp 3D.

Các lớp tích chập 3D giúp mô hình học các đặc trưng 3D từ dữ liệu đầu vào. Công thức tích chập 3D được biểu diễn như sau:

$$S(x, y, z) = (I * K)(x, y, z) = \sum_m \sum_n \sum_p I(x+m, y+n, z+p)K(m, n, p)$$

Trong đó:

- $S(x, y, z)$ là giá trị của bản đồ đặc trưng đầu ra tại vị trí (x, y, z) .
- I là khối dữ liệu đầu vào 3D.
- K là bộ lọc (kernel) 3D.
- $*$ là toán tử tích chập 3D.
- m, n, p là các chỉ số chạy qua kích thước của bộ lọc 3D.

Các lớp gộp 3D giúp giảm kích thước dữ liệu và tập trung vào các đặc trưng quan trọng nhất. Gộp cực đại 3D (3D Max Pooling) là phương pháp gộp phổ biến được sử dụng.

Cài đặt chi tiết mô hình CNN 3D có thể được tìm thấy tại: <https://colab.research.google.com/drive/1NGqUrKYbvkkY06AT3bQ7sRaDahw8kf38?usp=sharing>

4.3 Huấn luyện mô hình

Sau khi xây dựng mô hình CNN 2D và CNN 3D, chúng tôi tiến hành huấn luyện mô hình với các tham số sau:

- **Bộ tối ưu (Optimizer):** Adam - phù hợp với dữ liệu lớn và tối ưu tốt.
- **Hàm mất mát (Loss function):** Categorical Cross-Entropy - do đây là bài toán phân loại hai lớp.
- **Số epoch:** 15 - đủ để mô hình học đặc trưng nhưng tránh overfitting.
- **Batch size:** 30 (VGG16), 16 (3D CNN) - giúp mô hình học hiệu quả mà không tiêu tốn quá nhiều RAM.
- **Early Stopping:** Dừng huấn luyện nếu accuracy trên tập validation không cải thiện sau 5 epoch.

Cải thiện hiệu suất mô hình CNN 3D Đối với mô hình CNN 3D, chúng tôi sử dụng thêm kỹ thuật Batch Normalization để cải thiện hiệu suất. Batch Normalization là kỹ thuật chuẩn hóa dữ liệu, giúp tăng tốc quá trình huấn luyện và giảm thiểu tình trạng overfitting. Nhờ đó, chúng tôi có thể tăng số lượng khung hình đầu vào lên gấp 3 lần so với ban đầu (từ 20 khung hình lên 60 khung hình).

Điều này mang lại kết quả đáng kể hơn so với mô hình VGG16 (CNN 2D).

4.4 Đánh giá mô hình

Sau khi huấn luyện, mô hình sẽ được kiểm tra trên tập test và đánh giá dựa trên các chỉ số:

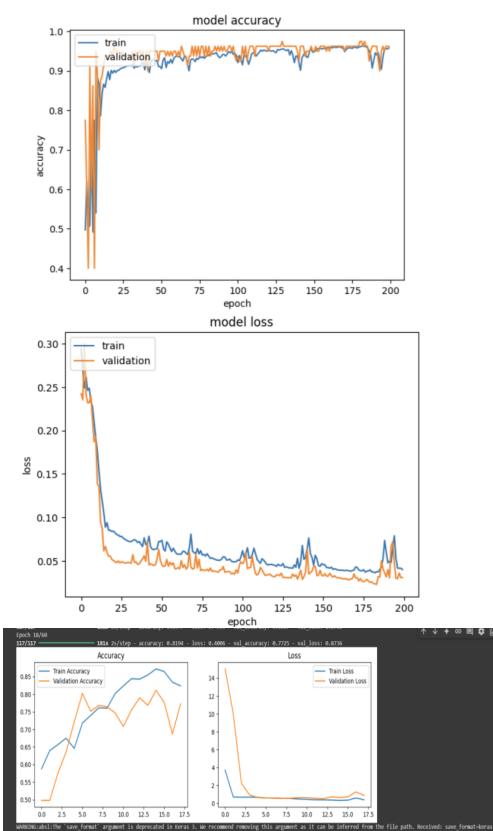
- + Accuracy: Độ chính xác của mô hình khi dự đoán.
- + Model Loss: Kiểm tra lượng dữ liệu bị thất thoát.

5. Kết quả

Sau khi xây dựng và huấn luyện mô hình CNN 2D và CNN 3D trên tập dữ liệu Hockey Fight Videos và cả bộ dữ liệu chúng tôi tự thu thập và đã tiến hành đánh giá hiệu suất của từng mô hình dựa trên các tiêu chí: độ chính xác (Accuracy), thời gian huấn luyện và khả năng làm "roi rót" dữ liệu.

5.1 Kết quả huấn luyện và kiểm tra

Bảng sau tổng hợp kết quả đánh giá hai mô hình:



Mô hình VGG16 có tỉ lệ Accuracy (Train Test) là xấp xỉ 100%, đồng thời tỉ lệ mất mát dữ liệu là xấp xỉ 5%

+ CNN 3D đã hoàn toàn cho thấy sự hiệu quả của chúng về độ mất mát dữ liệu cả tập dữ liệu huấn luyện và test song với đó khi chúng tôi thử sử dụng cả hai mô hình cũng cho thấy rằng mô hình có khả năng nắm bắt thông tin hiệu quả hơn là 3D CNN.

+ CNN 2D có thời gian huấn luyện ngắn hơn đáng kể, do chỉ xử lý từng khung hình riêng lẻ thay vì phân tích chuỗi thời gian như CNN 3D.

+ CNN 3D đòi hỏi nhiều tài nguyên hơn và cần GPU mạnh để có thể huấn luyện nhanh chóng.

5.2 So sánh hiệu suất trên dữ liệu thực tế

Để phân tích sâu hơn, chúng tôi sử dụng các video thực tế để kiểm tra tính chính xác của mô hình:

CNN 2D (VGG16): - VGG16:



+ Có thể thấy tỉ lệ dự đoán của mô hình VGG16 là khá cao mặc dù đây chỉ là khung hình về giao thông bình thường.



+ Còn đây là kết quả của 3D CNN, có thể thấy cùng một khung hình tuy nhiên rõ ràng về khả năng dự đoán đúng là thuộc về phần 3D CNN

+ Tuy nhiên, mô hình vẫn có một số lỗi khi phát hiện các cảnh va chạm không phải bạo lực (ví dụ: vận động viên ngã do mất thăng bằng).

5.3 Phân tích lỗi Hạn chế

Mặc dù đạt độ chính xác cao, cả hai mô hình vẫn gặp một số vấn đề:

Lỗi của CNN 2D

+ Chỉ xử lý từng khung hình đơn lẻ, mặc dù nhanh nhưng có thể các đặc trưng quan trọng sẽ bị thất thoát.

+ Một số hành vi bạo lực diễn ra trong thời gian ngắn bị bỏ lỡ.

+ Khi có nhiều yếu tố gây nhiễu (ánh sáng kém, nhiều người trong khung hình), mô hình dễ nhầm lẫn.

Lỗi của CNN 3D

+ Yêu cầu nhiều tài nguyên tính toán hơn, khó triển khai trên thiết bị hạn chế GPU.

+ Nếu số lượng khung hình đầu vào không đủ, mô hình có thể bỏ sót đặc trưng quan trọng.

+ Khi có các tình huống vận động mạnh nhưng không phải bạo lực (ví dụ: tranh bóng), mô hình có thể nhận diện sai.

6. Tổng kết

6.1 Kết luận

Từ các kết quả thực nghiệm, chúng tôi rút ra những kết luận chính sau:

+ CNN 3D đạt hiệu suất tốt hơn CNN 2D trong nhận diện bạo lực nhờ khả năng học đặc trưng không gian-thời gian.

+ CNN 2D nhanh hơn và tốn ít tài nguyên hơn, nhưng kém hiệu quả trong việc nhận diện hành vi có yếu tố chuyển động.

+ Cả hai mô hình đều có một số lỗi nhất định, đặc biệt là khi dữ liệu có quá nhiều yếu tố gây nhiễu (góc quay, ánh sáng, số lượng người tham gia...).

6.2 Hướng phát triển

Để cải thiện khả năng phát hiện hành vi bạo lực trong video, chúng tôi đề xuất một số hướng nghiên cứu trong tương lai:

+ Kết hợp CNN 3D với giá tốc để phát hiện chuyển động và hành vi của con người rõ hơn

+ CNN 3D có khả năng học đặc trưng không gian-thời gian, nhưng không có trí nhớ dài hạn. Chúng tôi đề xuất kết hợp CNN 3D với LSTM (Long Short-Term Memory) để cải thiện việc nhận diện hành vi phức tạp.

+ Yêu cầu nhiều tài nguyên về phần cứng hơn bởi vậy tốn nhiều chi phí hơn và nặng hơn về dung lượng so với VGG16.

Lợi ích:

Cải thiện khả năng phân biệt hành vi bạo lực với chuyển động thông thường. Giảm sai số khi phát hiện các hành vi có ngữ cảnh phức tạp.

2. Mở rộng và cải thiện bộ dữ liệu

Hiện tại, bộ dữ liệu Hockey Fight và bộ dữ liệu do chúng tôi tự thu thập chỉ chứa bạo lực trong một bối cảnh nhất định. Để tăng tính tổng quát của mô hình, chúng tôi đề xuất:

+ Thu thập dữ liệu từ nhiều nguồn khác nhau (camera giám sát, video thực tế...).

+ Bổ sung các trường hợp va chạm không phải bạo lực để mô hình học cách phân biệt chính xác hơn.

+ Tăng cường dữ liệu (Data Augmentation) bằng cách xoay ảnh, thay đổi độ sáng, làm mờ ảnh để giúp mô hình

hoạt động tốt hơn trong điều kiện thực tế.

Lợi ích:

+ Giúp mô hình tổng quát hóa tốt hơn trên các tình huống thực tế.

+ Giảm lỗi nhận diện sai khi điều kiện môi trường thay đổi.

3. Tối ưu hóa mô hình để chạy trên thiết bị di động

Để ứng dụng trong thực tế, mô hình cần chạy hiệu quả trên các thiết bị có tài nguyên hạn chế như điện thoại hoặc hệ thống giám sát an ninh.

+ Sử dụng TensorFlow Lite hoặc ONNX để tối ưu hóa mô hình.

+ Áp dụng kỹ thuật pruning và quantization để giảm kích thước mô hình mà không làm mất nhiều độ chính xác.

+ Chuyển đổi mô hình thành định dạng nhẹ để có thể triển khai trên hệ thống camera giám sát thời gian thực.

Lợi ích:

+ Giúp mô hình có thể chạy trực tiếp trên các hệ thống an ninh hoặc ứng dụng di động.

+ Giảm yêu cầu về phần cứng, mở rộng phạm vi ứng dụng trong thực tế.

6.3 Tổng kết

CNN 3D có hiệu suất vượt trội hơn CNN 2D trong phát hiện bạo lực. Tuy nhiên, mô hình còn gặp một số hạn chế về tài nguyên tính toán và tính tổng quát hóa.

Hướng phát triển tiềm năng bao gồm kết hợp CNN 3D với LSTM và gia tốc, mở rộng bộ dữ liệu, và tối ưu hóa mô hình để triển khai thực tế.

Mục tiêu dài hạn là xây dựng hệ thống phát hiện bạo lực tự động, chính xác và có thể hoạt động trong thời gian thực trên nhiều loại thiết bị khác nhau.

7.Mã nguồn và Tài liệu tham khảo

7.1.Mã nguồn

- Github: https://github.com/nguyenthanhtung2k4/BTL_Violent_Behavior
- Colab: <https://colab.research.google.com/drive/1NGqUrKYbvkkY06AT3bQ7sRaDahw8kf38?usp=sharing>

7.2.Tài liệu tham khảo

1. Krizhevsky, A., Sutskever, I., Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. *Advances in neural information processing systems*, *25*.
2. ResNet (Residual Networks): <https://arxiv.org/abs/1512.03385>
3. Inception Networks: <https://arxiv.org/abs/1512.00567>
4. DenseNet: <https://arxiv.org/abs/1608.06993>
5. Nguồn internet: <https://www.crcv.ucf.edu/projects/real-world/>