


THÔNG TIN CHUNG CỦA BÁO CÁO

- Link YouTube video của báo cáo (tối đa 5 phút):
<https://youtu.be/MaOGLZxBIBI>
- Link slides (dạng .pdf đặt trên Github):
https://github.com/binhpt310/CS2205.APR2023/blob/main/Slide_NCKH.pdf
- Mỗi thành viên của nhóm điền thông tin vào một dòng theo mẫu bên dưới
- Sau đó điền vào Đề cương nghiên cứu (tối đa 5 trang), rồi chọn Turn in

<ul style="list-style-type: none">● Họ và Tên: Phạm Thanh Bình● MSSV: 230201003 	<ul style="list-style-type: none">● Lớp: CS2205.APR2023● Tự đánh giá (điểm tổng kết môn): 9/10● Số buổi vắng: 0● Link Github: https://github.com/binhpt310/CS2205.APR2023
---	--

ĐỀ CƯƠNG NGHIÊN CỨU

TÊN ĐỀ TÀI

PHÁT HIỆN BẤT THƯỜNG TRONG CÁC VIDEO THU LẠI ĐƯỢC TỪ
CAMERA GIÁM SÁT

TÊN ĐỀ TÀI TIẾNG ANH (IN HOA)

VIDEO ANOMALY DETECTION IN SURVEILLANCE CAMERAS

TÓM TẮT (Tối đa 400 từ)

Phát hiện bất thường trong video là một đề tài khá khó thực hiện. Tuy đã có nhiều mô hình tiên tiến được phát triển và phát hiện được các bất thường tương đối chính xác, nhưng với lượng dữ liệu đầu vào khổng lồ là các video giám sát hàng ngày thì những mô hình này vẫn chưa đáp ứng đủ về độ chính xác. Do đó chúng tôi nghiên cứu mô hình phát hiện bất thường trong video sử dụng mạng **Convolutional LSTM**

Autoencoder.

Mô hình này được huấn luyện bằng cách sử dụng hai bộ dataset chính: “Avenue Dataset for Abnormal Event Detection” và “UCF-Crime”, cùng với các video từ các bộ dataset khác có chứa video từ camera giám sát. Mạng mà chúng tôi nghiên cứu kết hợp khả năng trích xuất đặc trưng không gian của mạng nơ ron tích chập (CNNs) với khả năng trích xuất đặc trưng thời gian của mạng Long Short-Term Memory (LSTM) trong kiến trúc của mạng Autoencoder.

GIỚI THIỆU (Tối đa 1 trang A4)

Phát hiện bất thường trong video là một chủ đề thiết thực trong cuộc sống của chúng ta, vì nó có thể giúp giám sát và bảo vệ các không gian công cộng và riêng tư khác nhau khỏi các mối đe dọa hoặc trong trường hợp khẩn cấp. Tuy nhiên, việc xem xét thủ công lượng lớn dữ liệu video được tạo bởi camera giám sát là không thực tế và không hiệu quả.

Đã có nhiều thuật toán liên quan đến phát hiện bất thường trong video giám sát từng được nghiên cứu và phát triển, một vài ví dụ phổ biến có thể kể đến như:

Autoencoder, Isolation Forest, One-Class SVM, GMM, K-means,... Tuy nhiên, những thuật toán này không thể đáp ứng việc phát hiện mọi bất thường ở mọi ngữ cảnh trong các video. Ngoài ra, bất thường có thể được định nghĩa theo nhiều cách khác nhau tùy thuộc vào ngữ cảnh và ứng dụng cụ thể. Ví dụ, trong video giám sát, hành vi bất thường có thể là một chiếc xe đi ngược chiều hoặc một người đi bộ băng qua đường một cách bất cẩn, còn trong video y tế, hành vi bất thường có thể là một cơn co giật hoặc một bất thường về nhịp tim. Ngoài ra, dữ liệu trong video có thể bị nhiễu bởi nhiều yếu tố, chẳng hạn như tiếng ồn hình ảnh, tiếng ồn âm thanh và thay đổi về ánh sáng.

Để có thể đáp ứng được việc phát hiện các bất thường trong video một cách hiệu quả và tối ưu nhất, chúng tôi đã nghiên cứu mô hình mạng **Convolutional LSTM Autoencoder**. Mạng này kết hợp các lớp Conv3D, ConvLSTM2D và Conv3DTranspose với nhau, được huấn luyện với trình tối ưu hoá Adam (Adam optimizer) và hàm mất mát Mean Square Error (MSE).

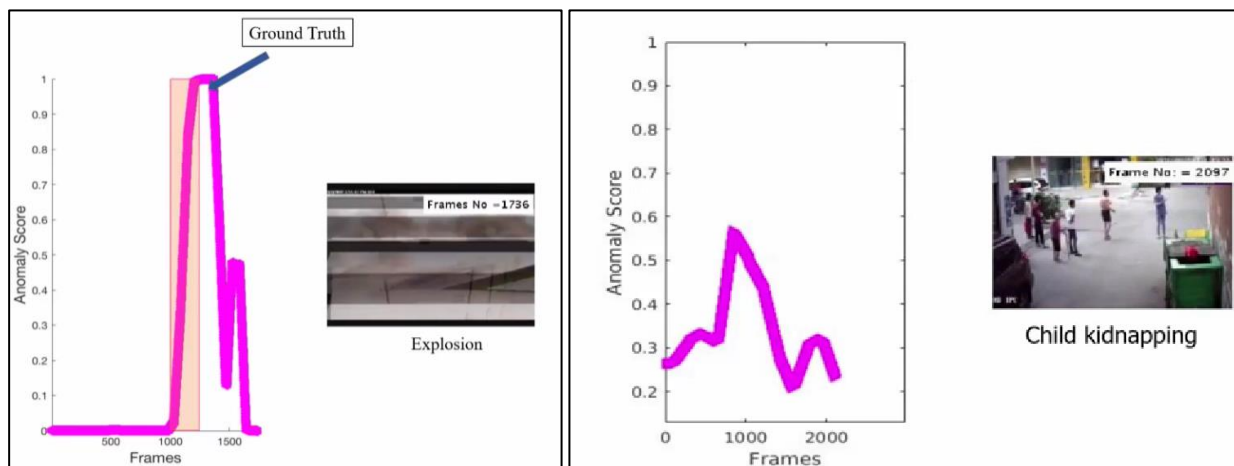
MỤC TIÊU

(Viết trong vòng 3 mục tiêu, lưu ý về tính khả thi và có thể đánh giá được)

- Phát triển thành công mô hình phát hiện bất thường trong video với độ chính xác đạt từ 80% trở lên.
- Cải thiện quá trình training của mô hình sao cho không bị tốn quá nhiều tài nguyên và thời gian, vì mỗi bộ dataset thường gồm rất nhiều video và dung lượng của chúng có thể lên tới hàng trăm GB.
- Nếu hoàn thành 2 mục tiêu trên, chúng tôi sẽ nghiên cứu mô hình có thể phát hiện bất thường với input là video trong thời gian thực thay vì là video có sẵn.

Input: Video được thu từ camera giám sát.

Output: Hiển thị cảnh báo sự kiện bất thường mỗi khi chiếu tới frame có điểm số bất thường cao trong video.



Hình 1: Biểu đồ hiển thị điểm bất thường mỗi khi duyệt qua từng frame

NỘI DUNG VÀ PHƯƠNG PHÁP

(Viết nội dung và phương pháp thực hiện để đạt được các mục tiêu đã nêu)

1. Nội dung

- Về phần dataset, chúng tôi sử dụng bộ dataset “UCF-Crime” gồm 1900 video, chia thành 14 lớp chứa 13 loại bất thường và 1 loại là hoạt động thông thường thu được từ camera giám sát trên toàn thế giới.
- Bộ dataset “Avenue Dataset for Abnormal Event Detection” thì có tổng cộng 37 video được chọn lọc kĩ càng và có sẵn các video đã được dán nhãn kèm đánh dấu bất thường ở các frame.
- Đối với các video đầu vào, chúng tôi chọn lọc các video từ các dataset sao cho hạn chế các video bị nhiễu, nhòe, độ phân giải quá thấp,... Ngoài ra chúng tôi cũng sẽ bỏ bớt khỏi bộ dataset huấn luyện các video có những bất thường khá mơ hồ cũng như không được thể hiện rõ ràng.
- Ví dụ: ánh sáng trong video nhấp nháy nhưng video không có sự kiện nào xảy ra hay có một vật thể nào di chuyển; người đi bộ tăng tốc đột ngột và biến mất khỏi tầm nhìn của camera.
- Xây dựng mô hình huấn luyện và dùng đầu ra của mô hình để triển khai một chương trình để tính điểm số bất thường trong từng frame của video.
- Mô hình này cũng có thể được lưu lại để những lần khác có thể huấn luyện tiếp

tục.

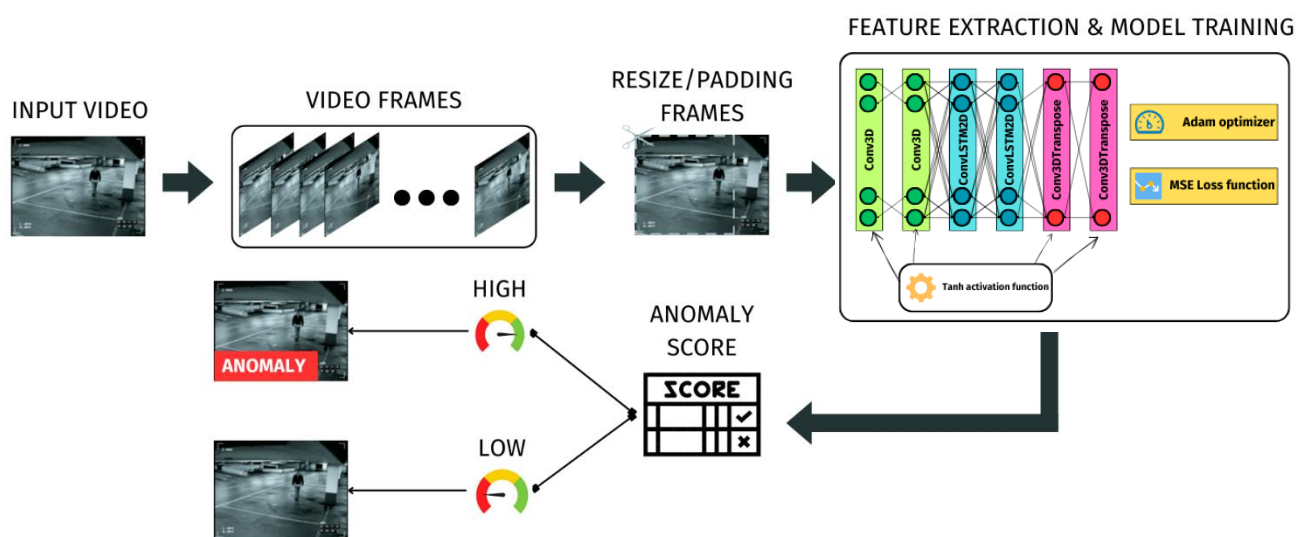
- Để có thể trực quan hoá cho người dùng, chúng tôi sẽ cho chương trình hiển thị cảnh báo bất thường mỗi khi tới frame có điểm số bất thường cao.

2. Phương pháp

- Đầu tiên, video đầu vào ở cả quá trình huấn luyện và thử nghiệm đều có độ phân giải khác nhau, nên chúng tôi sẽ padding đối với các video có độ phân giải nhỏ và resize các video có độ phân giải lớn lại sao cho bằng với chuẩn của mô hình.
- Tiếp theo, chúng tôi sẽ xây dựng một mô hình Tuần tự (Sequential) từ thư viện Keras của Python gồm các thành phần như sau:
 - **Convolutional Layers (Conv3D):** là lớp đầu tiên của mô hình, các lớp 3 chiều này sẽ học đặc trưng từ từng frame của video.
 - **Convolutional LSTM Layers (ConvLSTM2D):** các lớp này là một lớp kiểu của Mạng nơ ron tái phát triển (RNN), chúng sẽ học các dependencies tạm thời trong dữ liệu.
 - **Convolutional Transpose Layers (Conv3DTranspose):** các lớp này còn được gọi là lớp giải mã, các lớp này sẽ thực hiện nghịch đảo của thao tác tích chập từ lớp 3D trước đó. Chúng còn được sử dụng để lấy mẫu lại từ bản đồ đặc trưng (feature map).
 - **Hàm kích hoạt:** mô hình sẽ sử dụng hàm “tanh”, hàm này sẽ đảm bảo rằng các giá trị đầu ra trong phạm vi mà ta mong muốn, đây là một điều cần thiết đối với các tác vụ như tạo ảnh hoặc khử nhiễu hình ảnh.
 - **Hàm mất mát và thuật toán tối ưu:** để tối ưu đầu ra của mô hình chúng tôi sử dụng hàm mất mát MSE và thuật toán tối ưu Adam để đầu ra của mô hình có thể chính xác hơn.
- Sau khi hoàn chỉnh cấu trúc của mô hình, chúng tôi sẽ điều chỉnh các tham số như batch size, epoch, learning rate sao cho phù hợp với cấu hình phần cứng dùng để huấn luyện mô hình.
- Tiếp theo, mô hình sẽ duyệt từng frame của video đầu vào và sửa đổi frame về

định dạng grayscale, thay đổi kích thước các frame lại sao cho phù hợp với mô hình.

- Cuối cùng, mô hình sẽ thông qua mô hình để tính điểm trên từng frame của video dựa vào hàm mất mát MSE. Nếu điểm của frame nào vượt ngưỡng thì sẽ được tính là bất thường.



Hình 1: Sơ đồ hoạt động của mô hình

KẾT QUẢ MONG ĐỢI

(Viết kết quả phù hợp với mục tiêu đặt ra, trên cơ sở nội dung nghiên cứu ở trên)

- Các frame có bất thường được hiển thị chính xác giống như khi chúng ta nhận thấy bằng mắt thường trong video.
- Mô hình có độ chính xác cao hơn các mô hình phát hiện bất thường trong video đã từng được triển khai trong các bài báo.

TÀI LIỆU THAM KHẢO (Định dạng DBLP)

- [1]. Sultani, Waqas, Chen Chen, and Mubarak Shah. Real-world Anomaly Detection in Surveillance Videos. In CVPR, 2018.
- [2]. Jing Ren, Feng Xia, Yemeng Liu, and Ivan Lee. Deep Video Anomaly Detection: Opportunities and Challenges. In ICDMW, 2021.