THÔNG TIN CHUNG CỦA NHÓM

- Link YouTube video của báo cáo (tối đa 5 phút): *https://youtu.be/5ho-zSubVTA*
- Link slides (dạng .pdf đặt trên Github của nhóm): *https://github.com/NLeVinh/CS2205.FEB2025/blob/main/VinhNguyenLe_CS2205.FEB2025.DeCuong.FinalReport.Template.Slide.pdf*

- Họ và Tên: Nguyễn Lê Vinh
- MSSV: 240101084

- Lớp: CS2205.FEB2025
- Tự đánh giá (điểm tổng kết môn): 9/10
- Số buổi vắng: 0
- Số câu hỏi QT cá nhân: 4
- Link Github: https://github.com/NLeVinh/CS2205.FEB2025

ĐỀ CƯƠNG NGHIÊN CỨU

TÊN ĐỀ TÀI (IN HOA)

PHÁT HIỆN GIAN LẬN TRONG THI TRỰC TUYẾN THÔNG QUA PHÂN TÍCH VIDEO WEBCAM VÀ ÂM THANH SỬ DỤNG MÔ HÌNH TRANSFORMER ĐA MODAL

TÊN ĐỀ TÀI TIẾNG ANH (IN HOA)

MULTIMODAL ONLINE EXAM CHEATING DETECTION THROUGH VIDEO AND AUDIO ANALYSIS USING TRANSFORMER MODEL

TÓM TẮT *(Tối đa 400 từ)*

Gian lận trong thi cử là một vấn đề đáng lo ngại, đặc biệt trong bối cảnh các kỳ thi ngày càng được tổ chức theo hình thức trực tuyến nhằm đáp ứng nhu cầu học tập linh hoạt sau đại dịch COVID-19. Việc phát hiện hành vi gian lận một cách chính xác và kịp thời là yếu tố quan trọng để đảm bảo tính công bằng và minh bạch trong đánh giá năng lực. Tuy nhiên, môi trường thi từ xa đặt ra nhiều thách thức mới cho công tác giám sát, nhất là khi các hành vi gian lận ngày càng trở nên tinh vi và khó nhận diện.

Nhiều hệ thống hiện tại chỉ xử lý dữ liệu đơn kênh như hình ảnh webcam hoặc âm thanh, dẫn đến việc bỏ sót các hành vi gian lận phức tạp như thì thầm, ra hiệu, hoặc thậm chí là nhận chỉ dẫn từ người khác. Mỗi hành vi có thể thể hiện qua nhiều tín hiệu như ánh mắt, tư thế đầu và âm thanh môi trường, đòi hỏi một hệ thống giám sát có khả năng phân tích tổng hợp đa nguồn dữ liệu. Bên cạnh đó, việc giám sát thủ công không còn phù hợp trong khi dữ liệu phát sinh nhanh và liên tục trong suốt thời gian thi.

Một số nghiên cứu đã bước đầu áp dụng mô hình CNN-LSTM cho dữ liệu đa modal và cho kết quả khả quan. Tuy nhiên, LSTM còn hạn chế trong việc học quan hệ dài hạn và không tận dụng tốt tương quan phức tạp giữa các modal. Để khắc phục điều này, đề tài đề xuất mô hình Transformer đa modal, với hai nhánh encoder cho video và âm thanh được tích hợp qua cơ chế cross-attention nhằm học đồng thời đặc trưng

không gian, thời gian và mối liên kết giữa các kênh dữ liệu.

Đề tài sử dụng bộ dữ liệu thực tế MSU Online Exam Proctoring Dataset[1], bao gồm video và âm thanh được ghi lại từ các tình huống thi trực tuyến với nhãn hành vi được gán thủ công. Mô hình được huấn luyện và đánh giá dựa trên các độ đo Accuracy, Precision, Recall, F1-score đồng thời được so sánh với các mô hình đơn lẻ Vision Transformer, Audio Transformer và CNN-LSTM baseline để làm rõ tiềm năng của kiến trúc Transformer đa modal trong phát hiện gian lận trong thi trực tuyến.

GIỚI THIỆU *(Tối đa 1 trang A4)*

Gian lận trong thi cử trực tuyến đang trở thành một vấn đề nổi bật hiện nay, nhất là trong bối cảnh giáo dục chuyển đổi mạnh mẽ sang hình thức học tập và kiểm tra từ xa. Việc không phát hiện và xử lý kịp thời các hành vi gian lận có thể gây ảnh hưởng nghiêm trọng đến tính công bằng, chất lượng đào tạo và uy tín của hệ thống giáo dục. Do đó, một hệ thống giám sát tự động ứng dụng các mô hình học sâu là điều cần thiết. Video từ webcam và âm thanh từ môi trường xung quanh là hai nguồn tín hiệu phổ biến nhất được sử dụng trong các hệ thống giám sát hiện nay. Tuy nhiên, những dữ liệu này có đặc điểm đa dạng, phức tạp và biến đổi theo thời gian, đặc biệt khi các hành vi gian lận thường diễn ra ngắn, tinh vi và không dễ nhận biết. Mỗi hành vi có thể biểu hiện thông qua biểu cảm khuôn mặt, ánh mắt, tư thế đầu hoặc âm thanh phát ra xung quanh, do đó cần một phương pháp kết hợp đa luồng dữ liệu đồng thời theo thời gian thực.

Một số nghiên cứu gần đây [1, 2, 3] đã ứng dụng các mô hình học sâu như CNN hoặc LSTM để phân tích từng loại dữ liệu riêng biệt như video hoặc âm thanh nhằm phát hiện hành vi gian lận. Gần đây, nghiên cứu của Lamba và Sharma [4] đã bước đầu kết hợp hai modal bằng mô hình CNN-LSTM, cho thấy hiệu quả cao hơn so với mô hình đơn lẻ trong phát hiện gian lận. Tuy nhiên, kiến trúc tuần tự của LSTM còn hạn chế trong việc học các mối quan hệ dài hạn và chưa khai thác được mối liên hệ đặc trưng

[1] Sumit Saboo, S. Lamba, Nidhi Sharma: MSU Online Exam Proctoring Dataset. Mendeley Data, V1 (2021). https://www.kaggle.com/datasets/sumitsaboo/msu-online-exam-proctoring-dataset

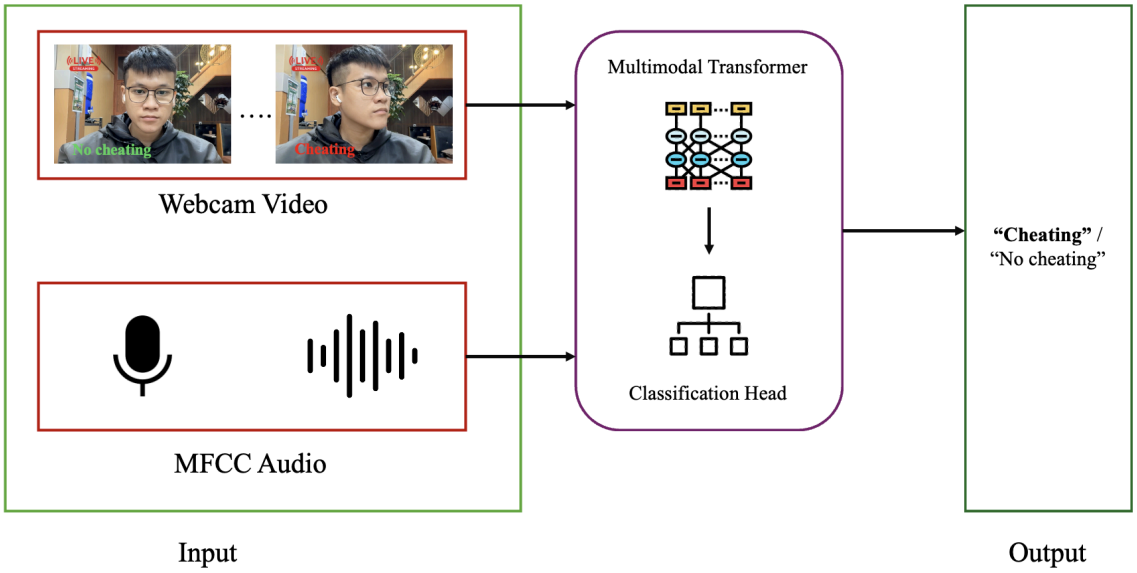
giữa hình ảnh và âm thanh.

Trên cơ sở đó, đề tài này đề xuất một mô hình giám sát thi trực tuyến ứng dụng Transformer đa modal bao gồm:

+ Hai nhánh Encoder riêng biệt cho video (Vision Transformer) và âm thanh (Transformer trên chuỗi MFCC)
+ Cơ chế cross-attention để tích hợp đồng thời đặc trưng không gian và thời gian, khai thác mối liên hệ giữa hai modal.
+ Kết quả đầu ra của mô hình là nhãn hành vi "Cheating" hoặc "No cheating".

Để đảm bảo tính thực tiễn, nghiên cứu sử dụng MSU Online Exam Proctoring Dataset, một bộ dữ liệu công khai gồm video và âm thanh được gán nhãn hành vi gian lận và không gian lận. Mô hình sẽ được đánh giá trên bộ dữ liệu này sử dụng các độ đo Accuracy, F1-score, Precision và Recall, đồng thời so sánh với các mô hình đơn modal và CNN-LSTM baseline. Từ đó, đề tài làm rõ tiềm năng của kiến trúc Transformer đa modal trong việc nâng cao độ chính xác và độ tin cậy của hệ thống giám sát thi cử trực tuyến.

- Input: Chuỗi khung hình (video frames) và dữ liệu âm thanh thu từ webcam và micro.
- Output: Nhãn phân loại hành vi: "Cheating" hoặc "No cheating"



Hình 1: Sơ đồ mô hình phát hiện gian lận thi trực tuyến từ video webcam và âm thanh

MỤC TIÊU *(Viết trong vòng 3 mục tiêu)*

- Nghiên cứu, xây dựng mô hình Transformer đa modal có khả năng kết hợp dữ liệu video webcam và âm thanh môi trường để phát hiện gian lận trong thi trực tuyến.
- Thực nghiệm và đánh giá hiệu quả của mô hình trên các độ đo Accuracy, Precision, Recall, F1-score; so sánh với các mô hình đơn modal Vision Transformer, Audio Transformer và mô hình CNN-LSTM.
- Xây dựng ứng dụng prototype mô phỏng hệ thống giám sát thi trực tuyến tích hợp mô hình đã huấn luyện để thực nghiệm minh hoạ.

NỘI DUNG VÀ PHƯƠNG PHÁP

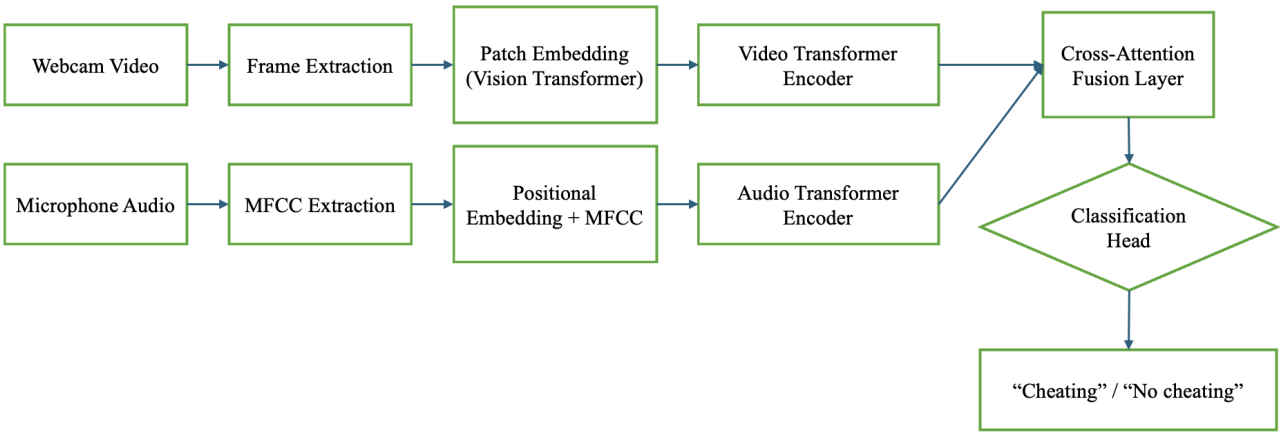
1. Nội dung

- Tìm hiểu tổng quan về vấn đề gian lận trong thi cử trực tuyến, các dạng hành vi thường gặp như nói nhỏ, ra hiệu, liếc nhìn tài liệu, sử dụng tai nghe.
- Nghiên cứu các mô hình học sâu đơn modal như Vision Transformer (ViT) cho video [5] và Transformer encoder cho chuỗi đặc trưng âm thanh trích xuất bằng MFCC [6].
- Tìm hiểu kiến trúc và cơ chế tích hợp của mô hình Transformer đa modal, đặc biệt là cách sử dụng cross-attention [7] để kết hợp thông tin và học mối quan hệ tương quan giữa các nguồn dữ liệu.
- Sử dụng MSU Online Exam Proctoring Dataset - một tập dữ liệu công khai gồm video webcam và âm thanh môi trường được ghi lại đồng thời trong các phiên thi trực tuyến. Dữ liệu được thu thập từ 24 thí sinh với nhiều hình thức gian lận khác nhau và được gán nhãn rõ ràng theo hành vi gian lận hoặc không gian lận.
- Tìm hiểu các phương pháp đánh giá mô hình phân lớp hành vi trong chuỗi video-âm thanh như Accuracy, Precision, Recall, F1-score [8].
- Huấn luyện, kiểm tra và so sánh các mô hình trên cùng tập dữ liệu gồm:
    + Các mô hình đơn modal (Vision Transformer, Audio Transformer).
    + Mô hình baseline CNN-LSTM tích hợp.
    + Mô hình Transformer đa modal được đề xuất.

- Xây dựng ứng dụng prototype mô phỏng hệ thống giám sát trực tuyến: Ứng dụng cho phép người tải lên video và âm thanh cuộc thi để phân tích hành vi của thí sinh. Nếu phát hiện dấu hiệu bất thường, hệ thống sẽ đưa ra cảnh báo. Trong trường hợp không phát hiện gian lận, hệ thống sẽ xác nhận hợp lệ. Giao diện ứng dụng sẽ trực quan hoá các đoạn video hoặc âm thanh đáng ngờ giúp hỗ trợ giám thị đánh giá và xem xét thêm.

2. Phương pháp nghiên cứu

- Tìm hiểu cơ bản về các hành vi gian lận thi cử trực tuyến, thống kê các dạng hành vi thường gặp qua báo cáo và tài liệu giáo dục.
- Nghiên cứu chi tiết kiến trúc Transformer đơn modal (ViT, Audio Transformer) và cách kết hợp chúng thành mô hình đa modal thông qua cross-attention. Đồng thời tham khảo các công trình nghiên cứu liên quan về multimodal learning.
- Tiền xử lý dữ liệu từ MSU Online Exam Proctoring Dataset, bao gồm tách khung hình từ video bằng OpenCV, trích xuất đặc trưng âm thanh MFCC bằng Librosa và đồng bộ hoá hai dòng dữ liệu theo nhãn hành vi đã được phân loại.
- Huấn luyện và thực nghiệm với các mô hình đã chọn, điều chỉnh siêu tham số và ghi nhận kết quả qua nhiều lần thử nghiệm. So sánh các mô hình dựa trên các độ đo đánh giá Accuracy, Precision, Recall, F1-score.
- Phát triển ứng dụng minh hoạ trên nền web, cho phép tải video và âm thanh, chạy mô hình phát hiện gian lận và trả kết quả trên giao diện.



Hình 2: Pipeline mô hình Transformer đa modal phát hiện gian lận thi trực tuyến.

KẾT QUẢ MONG ĐỢI

- Báo cáo tổng quan các phương pháp phát hiện gian lận thi trực tuyến đã được nghiên cứu, bao gồm mô hình đơn modal và CNN-LSTM baseline, mô hình đề xuất Transformer đa modal.
- Đưa ra kết quả thực nghiệm, so sánh, đánh giá độ hiệu quả giữa các mô hình thông qua các độ đo Accuracy, Precision, Recall, F1-score.
- Xây dựng ứng dụng trên nền tảng web cho phép tải video - âm thanh để phát hiện hành vi gian lận, hiển thị cảnh báo và hỗ trợ giám thị đánh giá trực quan.
- Thử nghiệm ứng dụng trong môi trường lớp học trực tuyến mô phỏng, tiếp tục thu thập dữ liệu để cải thiện mô hình và mở rộng bộ dữ liệu huấn luyện.

TÀI LIỆU THAM KHẢO *(Định dạng DBLP)*

[1]. Mahesh Navale, Aryan Jadhav, Mahesh Kadam, Shalmali Karandikar, Siddhi Kate: "A Computer Vision-Based Solution for Exam Cheating Detection." International Journal of Innovative Research in Information Security, Vol. 3(5), pp. 20–25 (2021).

[2]. Kamran Shaukat, Shuai Luo, Vijay Varadharajan, Imran A. Hameed, Sheng Chen, Dake Liu, Jinhai Li: "Student Cheating Detection in Higher Education by Implementing Machine Learning Techniques." Sensors, Vol. 23(8), Article 4149 (2023).

[3]. Zhiwei Hu, Yifan Jing, Guoping Wu, Hongyi Wang: "Multi-Perspective Adaptive Paperless Examination Cheating Detection System Based on Image Recognition." Applied Sciences, Vol. 14(10), Article 4048 (2024).

[4]. Sumit Lamba, Nidhi Sharma: "Deep Learning-Based Multimodal Cheating Detection in Online Proctored Exams." Journal of Electrical Systems, Vol. 20(3), pp. 7375–7383 (2024).

[5]. Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, Cordelia Schmid: "ViViT: A Video Vision Transformer." CoRR, abs/2103.15691 (2021).

[6]. C. S. Sonali, Chinmayi B S, Ahana Balasubramanian: "Transformer-based Sequence Labeling for Audio Classification based on MFCCs." CoRR, abs/2305.00417 (2023).

[7]. Yun Liu, Zheng Li, Tianyang Xu, Yang Chen, Yi Zhang: "Multimodal Transformer Using Cross-Channel Attention for Object Detection." CoRR, abs/2310.13876 (2023).

[8]. David M. W. Powers: "Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation." Journal of Machine Learning Technologies, Vol. 2(1), pp. 37–63 (2011).