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

Nguyễn Lê Vinh

University of Information Technology
HCMC, Vietnam

What ?

We proposed a multimodal Transformer-based framework for detecting cheating in online exams, in which we have:

- Designed a model combining webcam video (Vision Transformer) and MFCC audio (Audio Transformer) using cross-attention fusion.
- Compared with single-modal models (ViT, Audio Transformer) and CNN-LSTM baselines.
- Built a prototype application for automated exam monitoring.

Why ?

- Cheating in online exams is ***a growing threat to fairness and credibility*** in remote assessments.
- Existing methods often use ***single-modal input*** , making it ***hard to detect subtle cheating*** behaviors like whispering or signaling.
- A ***multimodal*** approach ***combining video and audio improves accuracy*** in real-time cheating detection

Overview

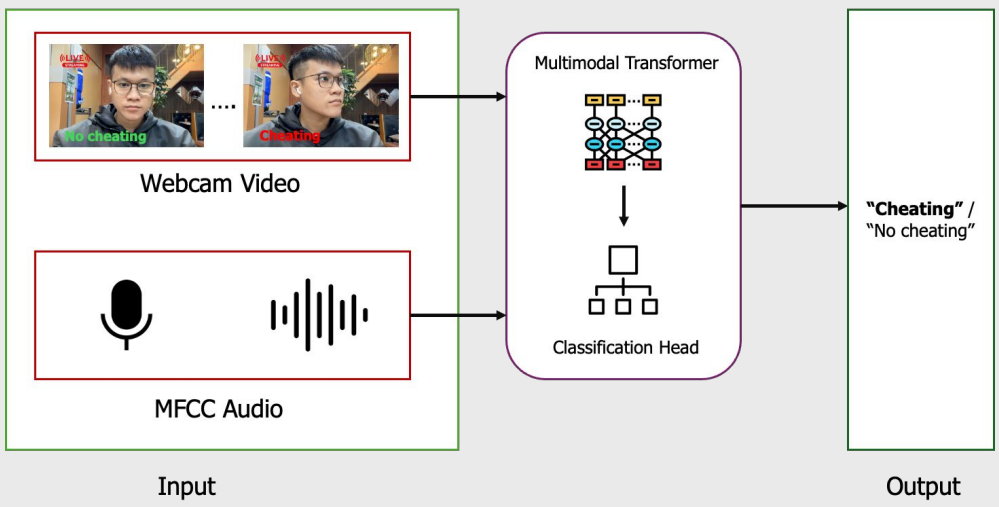

Figure 1. Diagram of the online exam cheating detection model using webcam video and audio

Description

1. CONTENT

- Research on common cheating behaviors in online exams (whispering, signaling,...)
- Study of Vision Transformer (ViT) and Audio Transformer models for video and audio analysis.
- Research on the integration of multimodal features using cross-attention.
- Utilization and preprocessing of the MSU Online Exam Proctoring Dataset, including synchronized video, audio and labeled cheating behaviors.
- Training and evaluation of single-modal models, the CNN–LSTM baseline, and proposed multimodal Transformer architecture.
- Development of a prototype application for automatic cheating detection and suspicious behavior visualization.

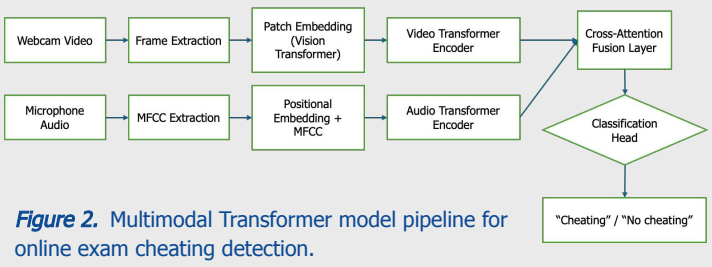

Figure 2. Multimodal Transformer model pipeline for online exam cheating detection.

2. METHOD RESEARCH

- Study **common online exam cheating behaviors** through **educational reports, articles and real-world cases.**
- Explore **ViT and Audio Transformer architectures** and their integration via **cross-attention** in multimodal Transformer models, **based on related deep learning literature.**
- Preprocess the **MSU Online Exam Proctoring Dataset** by **extracting video frames (OpenCV)**, **MFCC audio features (Librosa)**, and synchronizing modalities with behavior labels.
- **Train and evaluate** single-modal models, CNN-LSTM baseline, and the proposed multimodal Transformer using metrics: Accuracy, Precision, Recall and F1-score.
- **Develop a web-based prototype application** that enables users to upload synchronized video and audio exam data and automatically analyzes behavioral patterns using the trained model.

3. EXPECTED RESULTS

- Report on **experimental results** comparing model performance using Accuracy, Precision, Recall and F1-score.
- Develop **a web-based application** that detects and visualizes suspicious behavior from uploaded exam video-audio data.
- Deploy and test the system in simulated online exam settings to **enrich training data and improve model robustness.**

NII

Nguyễn Lê Vinh – Trường Đại học Công nghệ Thông tin. ĐHQG TP.HCM
TEL : 0837152541 Email : vinhnl.19@grad.uit.edu.vn