BÁO CÁO ĐỒ ÁN CUỐI KỲ

Môn học

CS519 - PHƯƠNG PHÁP NGHIÊN CỨU KHOA HỌC

Lớp học

CS2205.NOV2024

Giảng viên

PGS.TS. LÊ ĐÌNH DUY

THÔNG TIN CHUNG CỦA NHÓM

- Link YouTube video của báo cáo (tối đa 5 phút):
  https://www.youtube.com/watch?v=O_iAzIlb-yA
- Link slides (dạng .pdf đặt trên Github của nhóm):
  (ví dụ: https://github.com/mynameuit/CS2205.xxx/TenDeTai.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*
- *Lớp Cao học, mỗi nhóm một thành viên*

- Họ và Tên: Nhuyễn Văn Tùng
- MSSV: 240202017

- Lớp: CS2205.PHƯƠNG PHÁP NCKH
- Tự đánh giá (điểm tổng kết môn): 9/10
- Số buổi vắng: 0
- Link Github: https://github.com/mynameuit/CS2205.xxx/

ĐỀ CƯƠNG NGHIÊN CỨU

TÊN ĐỀ TÀI (IN HOA)

PHÁT HIỆN DEEPFAKE SỬ DỤNG MÔ HÌNH HỌC SÂU

TÊN ĐỀ TÀI TIẾNG ANH (IN HOA)

DEEPFAKE DETECTION USING DEEP LEARNING MODELS

TÓM TẮT *(Tối đa 400 từ)*

Deepfake, sử dụng **Generative Adversarial Networks (GANs)**, đặt ra thách thức lớn trong **an ninh mạng** và **xác thực danh tính**. Các phương pháp phát hiện truyền thống ngày càng kém hiệu quả trước những kỹ thuật giả mạo tiên tiến.

Nghiên cứu này đề xuất **mô hình học sâu kết hợp** gồm **CNN, ResNet50 và Vision Transformer (ViT)** nhằm cải thiện khả năng phát hiện Deepfake. CNN trích xuất đặc trưng cục bộ, ResNet50 học đặc trưng sâu qua các lớp residual, trong khi ViT mô hình hóa mối quan hệ không gian, giúp nhận diện dấu hiệu giả mạo chính xác hơn.

Ngoài ra, nghiên cứu áp dụng **huấn luyện đối kháng** để tăng khả năng chống **tấn công đối kháng**, đồng thời tích hợp **Blockchain** nhằm đảm bảo tính toàn vẹn dữ liệu. Mô hình được huấn luyện trên tập dữ liệu **DFDC, FaceForensics++, Real vs. Fake**, sử dụng **Fine-tuning và Transfer Learning** với **TensorFlow và PyTorch**. Hiệu suất đánh giá dựa trên **Precision, Recall, F1-score**.

Kết quả mong đợi là nâng cao **độ chính xác** trong phát hiện Deepfake, đóng góp vào **bảo mật thông tin, truyền thông số và phòng chống giả mạo dữ liệu**.

Từ khóa: *Deepfake Detection, CNN, ResNet50, Vision Transformer, Adversarial Training, Blockchain, Machine Learning, Cybersecurity.*

GIỚI THIỆU *(Tối đa 1 trang A4)*

Deepfake là một công nghệ AI sử dụng Generative Adversarial Networks (GANs) để tạo ra video và hình ảnh giả mạo, đặt ra nhiều thách thức về an ninh mạng, bảo vệ danh tính và truyền thông số. Các phương pháp phát hiện Deepfake truyền thống dựa trên phân tích đặc trưng hình ảnh thủ công hoặc mô hình học máy đơn giản đang dần trở nên kém hiệu quả trước những kỹ thuật làm giả ngày càng tinh vi. Do đó, việc sử dụng các mô hình học sâu như CNN, ResNet50 và ViT để phát hiện Deepfake trở thành một hướng nghiên cứu tiềm năng.

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

Nghiên cứu này hướng đến các mục tiêu sau:

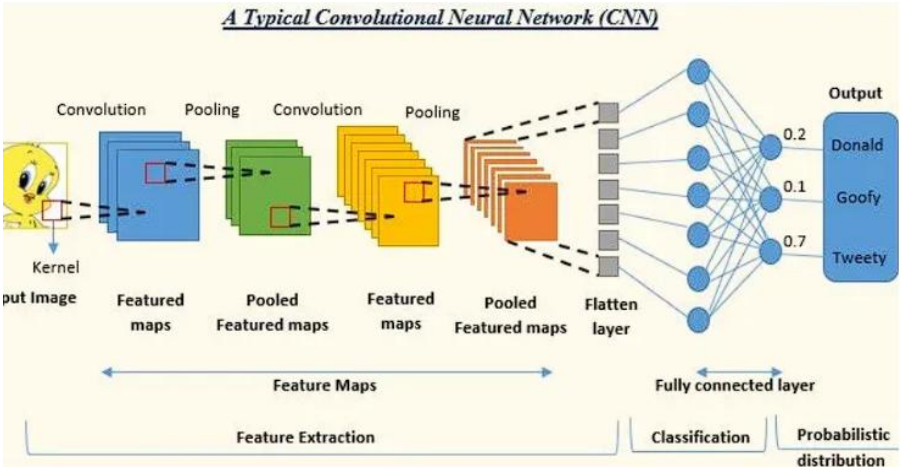
- Phân tích các kỹ thuật tạo Deepfake và những rủi ro an ninh liên quan.
- So sánh hiệu suất của các mô hình học sâu hiện có trong việc phát hiện Deepfake.

- Đề xuất mô hình kết hợp CNN, ResNet50 và ViT nhằm nâng cao độ chính xác.

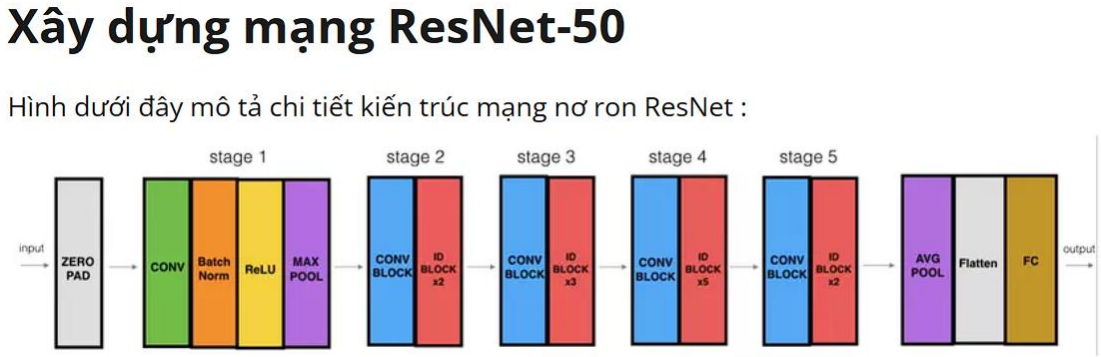
NỘI DUNG VÀ PHƯƠNG PHÁP

Mô hình hiện có:

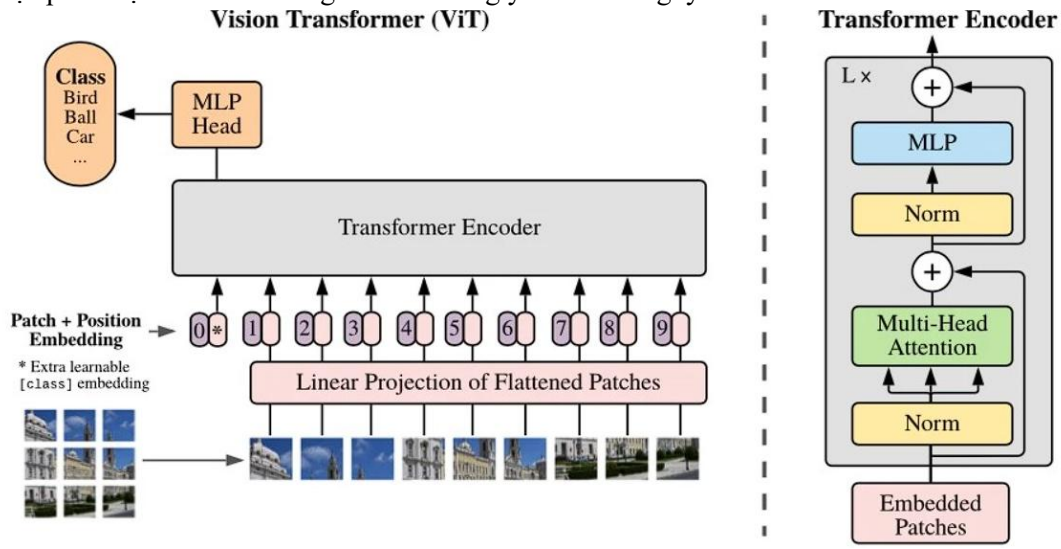
- CNN: Khai thác các đặc trưng cục bộ từ hình ảnh, phù hợp với việc nhận diện giả mạo nhưng có hạn chế về khả năng tổng quát hóa.



- ResNet50: Mô hình sâu hơn với các lớp residual giúp cải thiện khả năng học đặc trưng nhưng vẫn gặp khó khăn trong việc phát hiện các mẫu Deepfake phức tạp.



- ViT: Áp dụng transformer để mô hình hóa mối quan hệ không gian giữa các điểm ảnh, hiệu quả trong việc phát hiện các bất thường tinh vi nhưng yêu cầu tài nguyên tính toán cao.



Giải pháp đề xuất:

Mô hình kết hợp **CNN + ResNet50 + ViT** nhằm tận dụng ưu điểm của từng phương pháp:

- CNN trích xuất đặc trưng cục bộ ban đầu.
- ResNet50 học các đặc trưng sâu để tối ưu khả năng nhận diện.
- ViT tăng cường khả năng phân tích mối quan hệ không gian giúp cải thiện độ chính xác tổng thể.
- Bảo vệ mô hình: Áp dụng huấn luyện đối kháng nhằm nâng cao khả năng chống tấn công và sử dụng Blockchain để đảm bảo tính xác thực của dữ liệu.

Phương pháp thực hiện:

- Dữ liệu: Sử dụng các tập dữ liệu DFDC, FaceForensics++, Real vs. Fake để huấn luyện mô hình.
- Huấn luyện: Dùng TensorFlow và PyTorch, áp dụng Fine-tuning và Transfer Learning.
- Đánh giá: Dựa trên Precision, Recall, F1-score và kiểm tra khả năng chống tấn công đối kháng.

Dùng hàm mất mát Binary Cross-Entropy (BCE) cho phân loại Deepfake:

- Hàm BCE được sử dụng cho bài toán phân loại nhị phân (thật hoặc giả).
- Nếu nhãn thực tế (Deepfake), chỉ số hạng đầu tiên sẽ có giá trị cao nếu dự đoán gần 1.
- Nếu nhãn thực tế (hình ảnh thật), chỉ số hạng thứ hai sẽ có giá trị cao nếu dự đoán gần 0.
- Giúp tối ưu mô hình bằng cách giảm khoảng cách giữa dự đoán và thực tế.

Hàm kích hoạt Softmax:

$$\sigma(z)_i = \frac{e^{z_i}}{\sum_{j=1}^{n} e^{z_j}}$$

- Softmax biến đổi đầu ra của mô hình thành phân phối xác suất.
- Giá trị lớn nhất trong đầu ra sẽ có xác suất cao nhất, giúp phân loại hình ảnh dễ dàng hơn

Hàm chuẩn hóa Layer Normalization trong ViT: dùng để chuẩn hóa đầu vào của mỗi lớp giúp giảm hiện tượng biến đổi dữ liệu quá mức.

$$\hat{x}_i = \frac{x_i - \mu}{\sigma + \epsilon}\gamma + \beta$$

Hàm Residual Connection trong ResNet50: Giúp duy trì thông tin qua các lớp học sâu bằng cách cộng thêm đầu vào gốc x vào đầu ra $F(x)$

$$y = F(x) + x$$

KẾT QUẢ MONG ĐỢI

Bảng so sánh hiệu suất

Mô hình	Precision	Recall	F1-score	Khả năng chống tấn công
CNN	85%	80%	82%	Thấp
Resnet50	88%	83%	85%	Trung bình
ViT	90%	85%	87%	Khá
Mô hình đề xuất	93%	89%	91%	Cao

TÀI LIỆU THAM KHẢO *(Định dạng DBLP)*

- *Goodfellow et al. (2014). Generative Adversarial Networks.*
- *Rossler et al. (2019). FaceForensics++: Learning to Detect Manipulated Facial Images.*
- *Dosovitskiy et al. (2020). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale.*
- *Chollet (2017). Xception: Deep Learning with Depthwise Separable Convolutions.*
- *Wang et al. (2020). Detecting Deepfake Videos with Temporal Artifacts.*