Phát Hiện Deepfake Sử Dụng Hình Học Sâu

Nguyễn Văn Tùng

Trường ĐH Công Nghệ Thông Tin - Đại Học Quốc Gia Thành Phố Hồ Chí Minh

What ?

Nghiên cứu này tập trung vào **phát hiện Deepfake bằng mô hình học sâu:**

- 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
- Giải pháp đề xuất **kết hợp CNN, ResNet50 và ViT** để cải thiện độ chính xác trong phát hiện Deepfake.

Why ?

- Deepfake ngày càng tinh vi, dễ qua mặt con người và hệ thống cũ, gây rủi ro nghiêm trọng trong bảo mật danh tính, tin tức giả mạo và lừa đảo trực tuyến.
- Phương pháp truyền thống không đủ mạnh, khó phát hiện Deepfake thế hệ mới, cần học sâu (CNN, ResNet50, ViT) để cải thiện độ chính xác và khả năng nhận diện.
- Mô hình kết hợp **CNN + ResNet50 + ViT** giúp tận dụng dụng ưu điểm của từng phương pháp.

Overview

Deepfake là công nghệ sử dụng **Generative Adversarial Networks (GANs)** để tạo ra hình ảnh, video giả mạo, đặt ra nhiều thách thức về **an ninh mạng, bảo vệ danh tính và truyền thông số**. Các phương pháp phát hiện truyền thống, dựa trên **phân tích đặc trưng hình ảnh thủ công hoặc mô hình học máy đơn giản**, ngày càng kém hiệu quả trước các kỹ thuật giả mạo tinh vi.

Nghiên cứu này đề xuất **mô hình kết hợp CNN, ResNet50 và ViT** nhằm **nâng cao độ chính xác** trong phát hiện Deepfake. Mô hình này tận dụng **khả năng trích xuất đặc trưng cục bộ của CNN, học đặc trưng sâu của ResNet50 và phân tích mối quan hệ không gian của ViT** để phát hiện bất thường.

Mô hình được huấn luyện trên các tập dữ liệu **DFDC, FaceForensics++, Real vs. Fake**, sử dụng **TensorFlow và PyTorch**, và đánh giá dựa trên các chỉ số **Precision, Recall, F1-score**. Nghiên cứu này hướng đến **cải thiện khả năng phát hiện Deepfake**, giúp bảo vệ thông tin số và giảm thiểu rủi ro từ nội dung giả mạo.

Description

Nghiên cứu này tập trung vào việc **phát hiện Deepfake** – công nghệ sử dụng **Generative Adversarial Networks (GANs)** để tạo ra hình ảnh, video giả mạo, gây thách thức lớn về **an ninh mạng, bảo vệ danh tính và truyền thông số**.

Giải pháp đề xuất

Mô hình kết hợp CNN, ResNet50 và ViT để khai thác tối đa ưu điểm của từng phương pháp:

CNN: Trích xuất đặc trưng cục bộ từ hình ảnh.

ResNet50: Học các đặc trưng sâu để tối ưu khả năng nhận diện.

ViT: Mô hình hóa mối quan hệ không gian giữa các điểm ảnh, giúp phát hiện bất thường tinh vi.

Phương pháp thực hiện

- **Chọn bộ dữ liệu**: Sử dụng các tập **DFDC, FaceForensics++, Real vs. Fake** để huấn luyện.
- **Huấn luyện**: Dùng **TensorFlow, PyTorch**, áp dụng **Fine-tuning & Transfer Learning**.
- **Tiêu chí đánh giá**: Dựa trên các chỉ số **Precision, Recall, F1-score**, kiểm tra khả năng chống tấn công đối kháng.

Kết quả dự kiến

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**

NII

Nguyễn Văn Tùng – Trường Đại học Công Nghệ Thông Tin

TEL : 0972452727 Email : tungnv.19@grad.uit.edu.vn