PHÁT HIỆN DEEPFAKE SỬ DỤNG MÔ HÌNH HỌC SÂU

Nguyễn Văn Tùng - 240202017

Tóm tắt

- Lớp: CS2205.PHƯƠNG PHÁP NCKH
- Link Github của nhóm:
- Link YouTube video: https://www.youtube.com/watch?v=O_iAzIlb-yA
- Nguyễn Văn Tùng

Giới thiệu

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

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.

Nội dung và Phương pháp

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.

Nội dung và Phương pháp

Phương pháp thực hiện:

- **Chọn bộ 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.
- **Tiêu chí đánh giá:** Dựa trên Precision, Recall, F1-score và kiểm tra khả năng chống tấn công đối kháng.

Nội dung và Phương pháp

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

$$L = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)\right]$$

- 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.

Nội dung và Phương pháp

Hàm kích hoạt Softmax cho lớp đầu ra: $\sigma(z)_i = \frac{e^{z_i}}{\sum_{j=1}^{n} e^{z_j}}$

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

Hàm Residual Connection trong ResNet50: $y = F(x) + x$

Kết quả dự kiến

Bảng so sánh hiệu suất và mô hình đề xuất với kết quả mong đợi

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

- 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.*