THÔNG TIN CHUNG CỦA BÁO CÁO

- Link YouTube video của báo cáo: *https://www.youtube.com/watch?v=hh9pKNBizOg*
- Link slides (dạng .pdf đặt trên Github): *https://github.com/tranblb/CS2205.APR2023/blob/main/220201018_FinalReport_Slide.pdf*

• Họ và Tên: Bùi Lê Bảo Trân
• MSSV: 220201018
 
	• Lớp: CS2205.APR2023
	• Tự đánh giá (điểm tổng kết môn): 9/10
	• Số buổi vắng: 0
	• Link Github: *https://github.com/tranblb/CS2205.APR2023*
	• Mô tả công việc và đóng góp của cá nhân cho kết quả của nhóm:
	  ○ Lên ý tưởng đề tài
	  ○ Viết đề cương
	  ○ Làm slide báo cáo
	  ○ Trình bày đề tài

ĐỀ CƯƠNG NGHIÊN CỨU

TÊN ĐỀ TÀI (IN HOA)

PHÁT HIỆN HÌNH ẢNH KHUÔN MẶT NGƯỜI DEEPFAKE SỬ DỤNG MẠNG NƠ - RON TÍCH CHẬP

TÊN ĐỀ TÀI TIẾNG ANH (IN HOA)

DEEPFAKE DETECTION FOR HUMAN FACE IMAGES USING CONVOLUTION NEURAL NETWORKS

TÓM TẮT

Hình ảnh deepfake khuôn mặt người là kết quả của việc sử dụng trí tuệ nhân tạo và học sâu để tạo ra những hình ảnh giả mạo có độ chân thực cao của khuôn mặt người. Kỹ thuật này có thể bị lạm dụng để lan truyền thông tin sai lệch gây nên những ảnh hưởng tiêu cực đến cá nhân và xã hội. Trong nghiên cứu này, chúng tôi sử dụng mạng nơ-ron tích chập (Convolutional Neural Network - CNN) và bộ dữ liệu công khai để huấn luyện mô hình nhằm phân loại chính xác hình ảnh khuôn mặt thật và deepfake. Dữ liệu được chia thành tập huấn luyện (80%) và tập kiểm tra (20%). Độ tin cậy và hiệu suất của mô hình được đánh giá trên tập kiểm tra và so sánh với các phương pháp phát hiện deepfake khác. Mục đích cuối cùng nghiên cứu là để ngăn chặn những hậu quả tiêu cực như hủy hoại danh tiếng, lừa đảo, gian lận, ... đồng thời bảo vệ quyền riêng tư, đảm bảo thông tin đáng tin cậy và tạo môi trường trực tuyến an toàn

GIỚI THIỆU

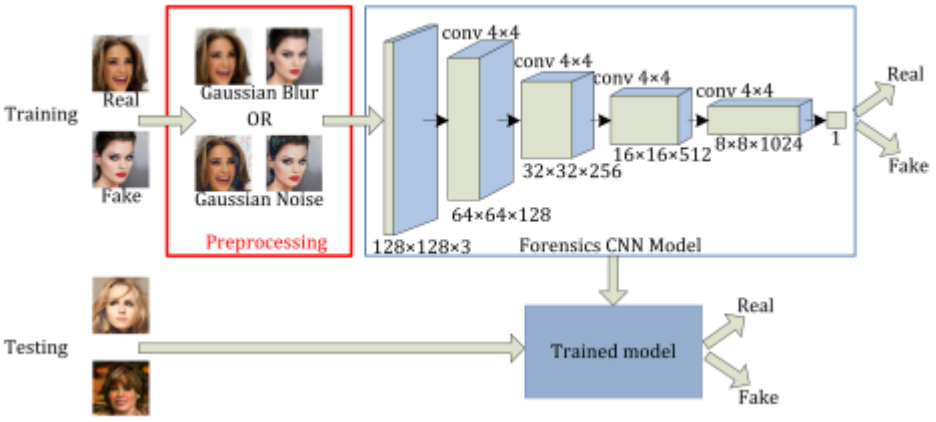
Deepfake là một công nghệ đột phá trong việc tạo ra những video hoặc hình ảnh giả mạo có tính chân thực cao, đặc biệt là trong lĩnh vực khuôn mặt người. Bằng cách kết hợp giữa các thuật toán học sâu và trí tuệ nhân tạo, deepfake cho phép tái tạo khuôn mặt người: không chỉ đơn thuần là việc ghép khuôn mặt một cách thô sơ, mà còn điều chỉnh các chi tiết nhỏ như di chuyển môi, biểu cảm và ngôn ngữ cơ thể để tạo ra một hình ảnh mà người xem dễ dàng tin rằng đó là người thật

Ví dụ về ảnh giả do Style-Gan tạo ra

Những hình ảnh deepfake có thể được sử dụng để đưa thông tin sai lệch, giả mạo, vi phạm quyền riêng tư và có khả năng gây hại cho các cá nhân, tổ chức. Để đối phó với vấn đề này và giảm thiểu hậu quả tiêu cực, phát hiện hình ảnh deepfake khuôn mặt người đã trở thành một lĩnh vực nghiên cứu quan trọng.

Ở nghiên cứu này, chúng tôi sử dụng Mạng nơ-ron tích chập (Convolutional Neural Network - CNN) và các bộ dữ liệu có sẵn công khai để huấn luyện mô hình phát hiện hình ảnh giả mạo.

- MICC-F220: bộ dữ liệu này bao gồm 220 hình ảnh; 110 bị giả mạo và 110 bản gốc.
- MICC-F2000: bộ dữ liệu này bao gồm 2000 hình ảnh; 700 bị giả mạo và 1300 bản gốc.
- MICC-F600: bộ dữ liệu này bao gồm 440 ảnh gốc, 160 ảnh giả mạo và 160 ảnh thật



Input: Hình ảnh chứa khuôn mặt bao gồm ảnh thật và ảnh deepfake
Output: Nhãn xác định "deepfake" hoặc "real" cho từng hình tương ứng

MỤC TIÊU

- Xây dựng mô hình CNN chuyên dụng cho việc phát hiện deepfake có khả năng phân loại chính xác giữa hình ảnh khuôn mặt gốc và deepfake.
- Đánh giá độ tin cậy và hiệu suất của mô hình CNN phát hiện deepfake thông qua việc so sánh với các phương pháp phát hiện deepfake khác
- Việc phát hiện các hình ảnh deepfake nhanh chóng, dễ dàng giúp ngăn ngừa các tác động tiêu cực: hủy hoại danh tiếng và uy tín, lừa đảo và gian lận,... đồng thời bảo vệ quyền riêng tư, đảm bảo thông tin đáng tin cậy và tạo môi trường trực tuyến an toàn

NỘI DUNG VÀ PHƯƠNG PHÁP

1. Nội dung:

- Thu thập và tiền xử lý các bộ dữ liệu: Các hình ảnh được thu thập từ nhiều nguồn khác nhau có thể bị nhiễu, trùng lặp, không đồng nhất,... việc tiền xử lý dữ liệu giúp chuẩn hóa dữ liệu sao cho phù hợp với mô hình huấn luyện. Bộ dữ liệu được chia nhỏ thành các tập huấn luyện và tập kiểm tra nhằm đảm bảo tính khách quan và xác định hiệu suất thực tế của mô hình
- Xây dựng một mô hình CNN có khả năng phân biệt giữa hình ảnh thật và hình ảnh Deepfake. Huấn luyện mô hình trên tập dataset đã qua xử lý giúp đảm bảo tính nhất quán và đáng tin cậy của quá trình huấn luyện
- Đánh giá và kiểm tra mô hình trên tập dữ liệu kiểm tra. Điều này để đo lường khả năng phân loại chính xác hình ảnh deepfake và hình ảnh thật của mô hình. Đồng thời, so sánh kết quả với một mô hình đã được công nhận là hiệu quả có thể xác định xem mô hình có đạt được hiệu suất tương đương hay không
- Nếu mô hình chưa đạt hiệu suất mong đợi thì tiến hành tinh chỉnh và cải thiện mô hình

2. Phương pháp:

- Thu thập dữ liệu và tiền xử lý
    - Đảm bảo rằng tập dữ liệu có đủ đa dạng để mô hình có thể học các đặc trưng phân biệt.
    - Xác định các nhãn cho dữ liệu, ví dụ như "Real" và "Deepfake". Cần đảm bảo nhãn là chính xác và đúng với từng hình ảnh.
    - Thực hiện các phép biến đổi dữ liệu để mở rộng tập dữ liệu huấn luyện và tránh overfitting, chẳng hạn như xoay, phóng to, thu nhỏ hoặc làm nhiễu hình ảnh.
    - Chia tập dữ liệu thành tập huấn luyện (80%) và tập kiểm tra (20%). Tập huấn luyện được sử dụng để huấn luyện mô hình, tập kiểm tra để đánh giá hiệu suất

- Huấn luyện mô hình:
    - Tiến hành huấn luyện mô hình trên tập dữ liệu huấn luyện.

- Huấn luyện mô hình trong nhiều epoch (vòng lặp) cho đến khi hiệu suất của mô hình không cải thiện thêm trên tập kiểm tra.

- Đánh giá và kiểm tra mô hình:
  - Đánh giá mô hình trên tập kiểm tra bằng cách tính toán các độ đo như độ chính xác (accuracy), độ nhạy (recall) và độ đặc hiệu (precision).
  - So sánh kết quả với một mô hình đã được công nhận là hiệu quả có thể xác định xem mô hình có đạt được hiệu suất tương đương hay không.

- Tinh chỉnh và cải thiện:
  - Nếu mô hình không đạt hiệu suất mong đợi thì thì tiến hành tinh chỉnh và cải thiện: thử các biến thể khác của kiến trúc mô hình, thay đổi siêu tham số hoặc thêm các kỹ thuật tiền xử lý dữ liệu mới để cải thiện hiệu suất,...

KẾT QUẢ MONG ĐỢI

- Mô hình CNN được kỳ vọng sẽ có khả năng phát hiện và phân loại chính xác giữa hình ảnh gốc và deepfake.
- So với các phương pháp khác, mô hình có độ tin cậy cao hơn, giảm thiểu số lượng false positives (nhầm lẫn hình ảnh gốc) và false negatives (không phát hiện deepfake)
- Góp phần phát hiện các thông tin sai lệch, hành vi lừa đảo cũng như ngăn ngừa tác động tiêu cực của công nghệ này

TÀI LIỆU THAM KHẢO

[1]. Luca Guarnera, Oliver Giudice, Francesco Guarnera, Alessandro Ortis, Giovanni Puglisi, Antonino Paratore, Linh M. Q. Bui,Marco Fontani, Davide Alessandro Coccomini, Roberto Caldelli, Fabrizio Falchi,Claudio Gennaro, Nicola Messina, Giuseppe Amato, Gianpaolo Perelli, Sara Concas, Carlo Cuccu, Giulia Orrù, Gian Luca Marcialis, Sebastiano Battiato : The Face Deepfake Detection Challenge. ICIAP 2021
[2]. Asad Malik, Minoru Kuribayashi, Sani M. Abdullahi, Ahmad Neyaz Khan: DeepFake Detection for Human Face Images and Videos: A Survey. IEEE 2022
[3]. Luca Guarnera, Oliver Giudice, Sebastiano Battiato:
Fighting Deepfake by Exposing the Convolutional Traces on Images. IEEE 2020
[4]. Yogesh Patel, Sudeep Tanwar, Pronaya Bhattacharya, Rajesh Gupta, Turki Alsuwian, Innocent E. A. Davidson: An Improved Dense CNN Architecture for Deepfake Image Detection. IEEE 2023
[5]. Norah M. Alnaim, Zaynab M. Almutairi, Manal S. Alsuwat, Hana H. Alalawi, Aljowhra Alshobaili. DFFMD: A Deepfake Face Mask Dataset for Infectious Disease Era With Deepfake Detection Algorithms. IEEE 2023
[6]. TechVidvan: DeepFake Detection using Convolutional Neural Networks.
url: *https://techvidvan.com/tutorials/deepfake-detection-using-cnn/*