HỌC TỔ HỢP SỬ DỤNG TRANSFORMER VÀ MẠNG NƠ-RON TÍCH CHẬP CHO NHẬN DẠNG KHUÔN MẶT KHI ĐEO KHẨU TRANG

Đinh Trọng Hậu - 250101018

Tóm tắt

- Lớp: CS2205.CH201
- Link Github của nhóm: https://github.com/tronghau112/CS2205.CH201
- Link YouTube video: https://www.youtube.com/watch?v=QC4JuZk_Xoc
- Họ tên: Đinh Trọng Hậu

Giới thiệu

Bối cảnh vấn đề

Đại dịch COVID-19 khiến việc đeo khẩu trang trở thành bắt buộc, làm giảm nghiêm trọng hiệu quả của các hệ thống nhận diện khuôn mặt truyền thống.

- Mất đi các đặc điểm nhận dạng quan trọng: mũi, miệng, cằm (khoảng 50% khuôn mặt).
- **Theo báo cáo NIST (2020):** Tỷ lệ lỗi của các thuật toán tăng mạnh từ **5% đến 50%** khi xử lý ảnh có khẩu trang.
- Nhu cầu cấp thiết về hệ thống nhận diện "không chạm" nhưng độ chính xác cao.

Mục tiêu

Mục tiêu chính

Xây dựng hệ thống nhận diện khuôn mặt mạnh mẽ (robust), có khả năng hoạt động chính xác ngay cả khi đối tượng đeo khẩu trang che kín phần lớn khuôn mặt.

Giải pháp

Sử dụng kỹ thuật **Học kết hợp (Ensemble Learning)**, tận dụng ưu điểm của cả mạng CNN truyền thống và kiến trúc Transformer hiện đại.

Nội dung: Các mô hình đề xuất

Fine-tuned CNNs

Sử dụng 3 mạng CNN phổ biến được tinh chỉnh :

- **VGG16:** Mô hình kinh điển.
- **EfficientNet:** Hiệu suất cao, ít tham số.
- **FaceNet:** Được huấn luyện chuyên biệt trên dữ liệu khuôn mặt.

Transformer

Áp dụng **Vision Transformer**:

- Chia ảnh thành các mảnh (patches).
- Học mối quan hệ ngữ cảnh toàn cục (Global Context) bằng cơ chế Self-Attention.

Ensemble Learning

Kết hợp sức mạnh của cả hai:

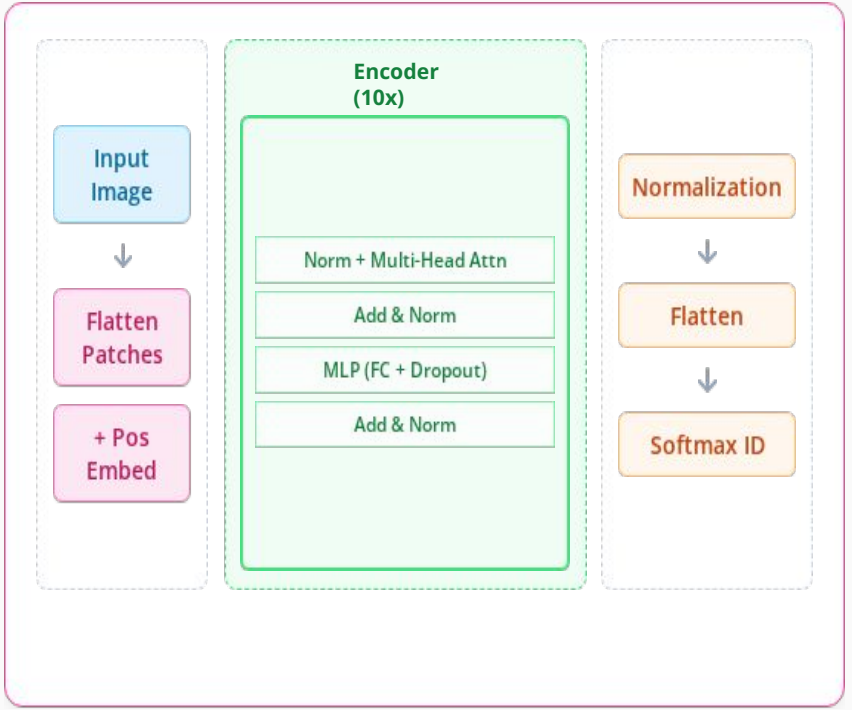
- Kết hợp 2 mô hình CNN.
- Kết hợp 2 mô hình Transformer.
- Tăng độ tin cậy và giảm phương sai.

Phương pháp: Kiến trúc Transformer

Cơ chế hoạt động Vision Transformer

Mô hình xử lý ảnh đầu vào thông qua chuỗi các bước:

- **Patch Partition:** Chia ảnh thành các mảnh nhỏ
- **Linear Embedding:** Chuyển đổi các mảnh thành vector đặc trưng.
- **Transformer Encoder:** Khối quan trọng nhất, dùng Self-Attention để học quan hệ toàn cục.
- **MLP Head:** Phân lớp danh tính.

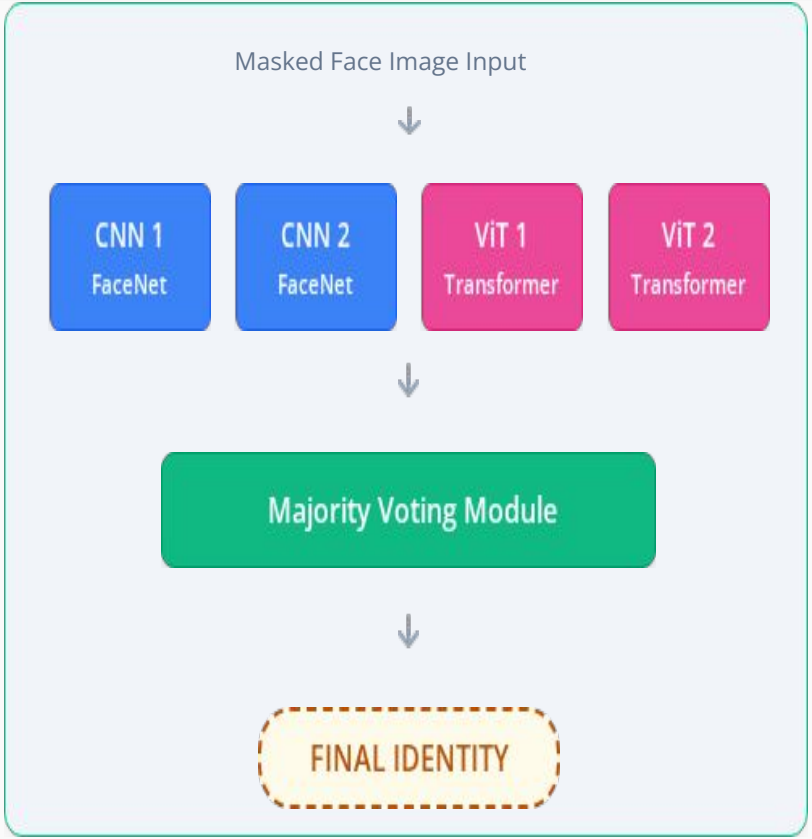


Phương pháp: Kiến trúc Transformer

Quy trình Kết hợp

Mô hình đề xuất sử dụng 4 mạng nơ-ron hoạt động song song:

- **2 x CNN (FaceNet):** Tinh chỉnh (Fine-tuned) để trích xuất đặc trưng cục bộ.
- **2 x Transformer:** Trích xuất ngữ cảnh toàn cục.
- **Majority Voting:** Tổng hợp kết quả dự đoán từ 4 mô hình để đưa ra quyết định cuối cùng.



Phương pháp: Dữ liệu tổng hợp

Synthetic Masked LFW

Do thiếu dữ liệu thực, bộ dữ liệu tổng hợp được tạo ra từ LFW:

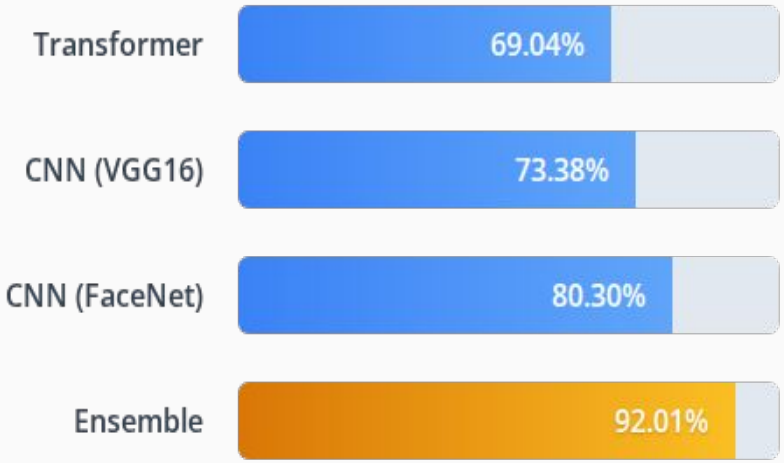
Tổng số ảnh **26,466**	Số đối tượng (Subjects) **5,749**
Tỷ lệ Train/Test **95% / 5%**	Số ảnh gốc (LFW) **13,233**

Data Augmentation: Để tăng cường khả năng tổng quát hóa, các kỹ thuật sau được áp dụng:

- Horizontal Flip
- Rescaling
- Zooming

Kết quả dự kiến

So sánh Độ chính xác



Mô hình Ensemble vượt trội hoàn toàn so với các mô hình đơn lẻ.

Đánh giá & Định hướng

- **Hiệu quả vượt trội:** Sự kết hợp giúp tăng độ chính xác đáng kể **(>10%)** so với mô hình tốt nhất là FaceNet.
- **Sự bổ sung hoàn hảo:** Tận dụng đặc trưng cục bộ của CNN và ngữ cảnh toàn cục của Transformer để "nhìn xuyên" vật cản.
- **Hướng phát triển:** Cải thiện khả năng nhận diện với các vật cản phức tạp khác và mở rộng sang bài toán xác thực.

Tài liệu tham khảo

"Ensemble Learning using Transformers and Convolutional Networks for Masked Face Recognition"

Mohammed R. Al-Sinan, Aseel F. Haneef, and Hamzah Luqman

Information and Computer Science Department, King Fahd University of Petroleum and Minerals

arXiv:2210.04816v1 [cs.CV] (2022)