

Nhận dạng Nhóm Tuổi từ Ảnh Khuôn Mặt

So sánh Hiệu quả của EfficientNet, ResNet và Vision Transformer

Nguyễn Tân Tài - 23521376
Giảng viên hướng dẫn: Mai Tiến Dũng

Ngày 15 tháng 12 năm 2025

Nội dung trình bày

- 1 1.1 Giới thiệu đề tài
- 2 2. Mô hình so sánh

- 3 3. Tập dữ liệu và tiền xử lý
- 4 6. Kết luận và đề xuất

Giới thiệu đề tài

Bối cảnh:

- Sự phát triển mạnh mẽ của **Thị giác máy tính (Computer Vision)**
- Nhu cầu ngày càng cao về **hiểu đặc điểm nhân khẩu học** từ hình ảnh
- Khuôn mặt chứa nhiều thông tin quan trọng: *tuổi, giới tính, cảm xúc*

Vấn đề đặt ra:

Làm thế nào để **tự động nhận dạng nhóm tuổi** từ ảnh khuôn mặt với
độ chính xác cao và chi phí tính toán hợp lý?

Ý nghĩa và ứng dụng thực tiễn

Ứng dụng thực tế:

- Quảng cáo theo độ tuổi
- Kiểm soát nội dung
(Age-restricted)
- Hệ thống giám sát thông minh
- Phân tích hành vi người dùng

Ý nghĩa nghiên cứu:

- So sánh các kiến trúc học sâu phổ biến
- Đánh giá **hiệu suất vs chi phí**
- Gợi ý mô hình phù hợp cho từng kịch bản

Mục tiêu nghiên cứu

Mục tiêu chính:

- ① Xây dựng hệ thống phân loại nhóm tuổi từ ảnh khuôn mặt
- ② So sánh hiệu quả của:
 - EfficientNet
 - ResNet
 - Vision Transformer
- ③ Đánh giá toàn diện về:
 - Độ chính xác
 - Khả năng tổng quát hóa
 - Chi phí tính toán

Câu hỏi nghiên cứu trung tâm:

"Kiến trúc nào mang lại sự cân bằng tốt nhất giữa hiệu suất và tài nguyên tính toán cho bài toán phân loại tuổi?"

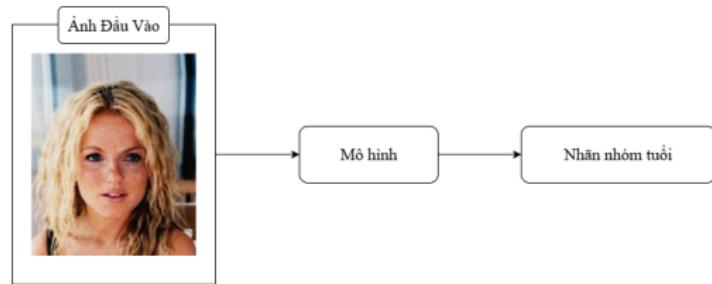
1.1. Bài toán và ứng dụng thực tế

Bài toán chính

Phân loại nhóm tuổi từ ảnh khuôn mặt thành 5 nhóm

5 nhóm tuổi:

- Trẻ em (0–12)
- Thiếu niên (13–19)
- Thanh niên (20–39)
- Trung niên (40–59)
- Người cao tuổi (60+)



Hình 1: Pipeline tổng quát của bài toán phân loại nhóm tuổi từ ảnh khuôn mặt.

1.2. Thách thức nghiên cứu

Khó khăn:

- Đa dạng sinh học
- Ranh giới mờ giữa nhóm
- Dữ liệu không cân bằng
- Điều kiện chụp khác nhau

Câu hỏi nghiên cứu:

“Kiến trúc nào (*EfficientNet*, *ResNet*, hay *Vision Transformer*) cho kết quả tốt nhất với tài nguyên hợp lý?”

2.1. Các kiến trúc được đánh giá

Họ mô hình	Variant	Params (M)	Đặc điểm
ResNet	18, 34	11.7, 21.8	CNN cổ điển
EfficientNet	B0, B3	5.3, 12.0	Hiệu quả cao
Vision Transformer	ViT-B/16	86.0	Transformer cho ảnh

Mục tiêu so sánh:

- Hiệu suất (Accuracy, F1-Score)
- Hiệu quả tính toán (Params, FLOPs)
- Thời gian inference

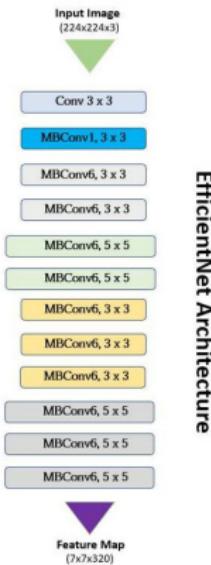
2.2. Kiến trúc EfficientNet

Ý tưởng chính:

- **Compound Scaling:** Cân bằng
 - Depth (độ sâu)
 - Width (độ rộng)
 - Resolution (độ phân giải)
- **MBConv blocks:** Tối ưu bộ nhớ

Lợi thế:

- Hiệu suất cao
- Tham số ít
- Tốc độ nhanh



EfficientNet Architecture

Hình: 2.Kiến trúc EfficientNet

2.3. Kiến trúc ResNet (18 / 34)

Ý tưởng chính:

- **Residual Learning:**

- Học phần dư

$$F(x) = H(x) - x$$

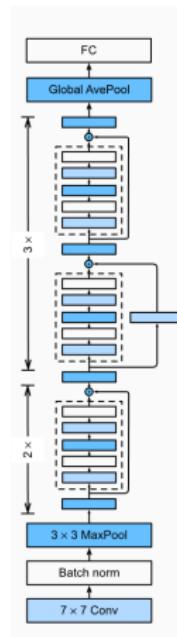
- Skip connection giúp truyền gradient tốt

- **Basic Block:**

- Conv 3×3 + BatchNorm + ReLU
- Phù hợp mạng vừa và nhỏ

Lợi thế:

- Huấn luyện ổn định
- Dễ fine-tune
- Baseline mạnh cho dữ liệu



Hình: 3.Residual Block trong ResNet

2.4. Kiến trúc Vision Transformer (ViT-B/16)

Ý tưởng chính:

- **Patch Embedding:**

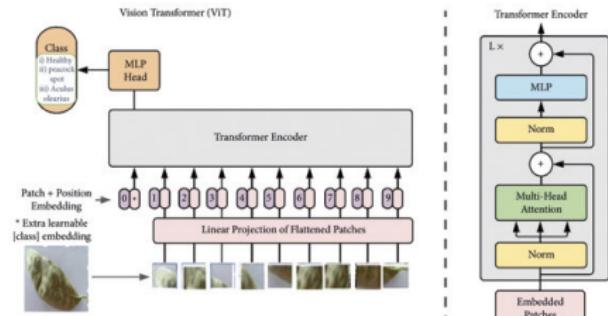
- Chia ảnh thành các patch 16×16
- Mỗi patch được ánh xạ thành vector

- **Self-Attention:**

- Học quan hệ toàn cục giữa các vùng ảnh
- Không dùng convolution

Đặc điểm:

- Khả năng biểu diễn mạnh
- Phụ thuộc nhiều vào dữ liệu
- Cần pretrain trên tập lớn



Hình: 4.Kiến trúc Vision Transformer (ViT).

So sánh nhanh các kiến trúc

- ResNet: ổn định, baseline mạnh
- EfficientNet: hiệu quả tham số tốt nhất
- ViT: mạnh về lý thuyết nhưng cần nhiều dữ liệu

3.1. Tập dữ liệu UTKFace

Thông tin:

- 23,625 ảnh khuôn mặt
- 200×200 pixels
- Nhãn: Tuổi, Giới tính, Chủng tộc

Tiền xử lý:

- ① Phát hiện và căn chỉnh mặt
- ② Chuẩn hóa pixel
- ③ Phân nhóm 5 lớp
- ④ Tăng cường dữ liệu

UTKFace Validation Dataset Samples



Hình: 5. Mẫu ảnh từ validation set

3.2. Phân chia dữ liệu

Tỷ lệ: 70%-15%-15%

Nhóm tuổi	Train	Val	Test	Tổng
Trẻ em	2,382	510	512	3,404
Thiếu niên	825	176	178	1,179
Thanh niên	8,316	1,782	1,782	11,880
Trung niên	3,180	681	683	4,544
Cao tuổi	1,811	403	404	2,617
Tổng	16,514	3,552	3,559	23,625

Vấn đề:

- Mất cân bằng nghiêm trọng
- Thanh niên: 8,316 mẫu
- Thiếu niên: 825 mẫu (chỉ 10%)

Giải pháp:

- Class weighting
- Oversampling
- Augmentation tập trung

Tăng cường Dữ liệu (Data Augmentation)

Mục tiêu:

- Cân bằng phân bố dữ liệu giữa các nhóm tuổi
- Giảm overfitting, tăng khả năng tổng quát hóa
- Mô phỏng biến thiên thực tế khi chụp ảnh khuôn mặt

Thách thức:

- Nhóm Thiếu niên (G2) chỉ chiếm **3.5%** dữ liệu
- Đặc trưng tuổi (*nếp nhăn, cấu trúc mặt*) dễ bị phá vỡ
- Augmentation quá mạnh có thể gây **nhiễu nhăn**

Các Phép Biến đổi được Áp dụng

Biến đổi hình học:

- **Lật ngang** ($p = 0.5$)
- **Xoay nhẹ** ($\pm 10^\circ$)

Mục tiêu: mô phỏng góc chụp tự nhiên
nhưng vẫn bảo toàn đặc trưng tuổi.

Biến đổi màu sắc:

- Điều chỉnh **độ sáng / tương phản** (nhẹ)

Chiến lược áp dụng:

- **Train:** có augmentation
- **Val/Test:** không augmentation

Lý do Thiết kế Augmentation Nhẹ

Đặc thù bài toán phân loại tuổi:

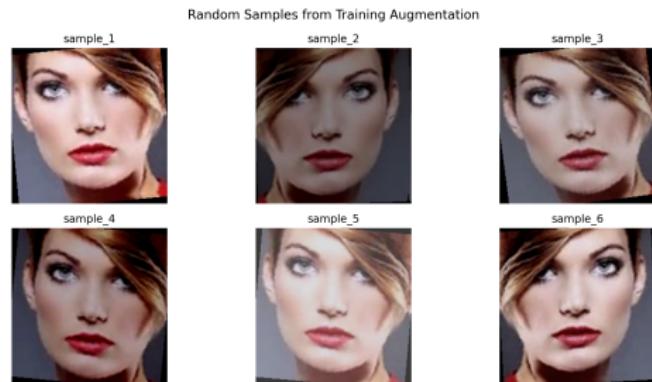
- Tuổi được biểu hiện qua **chi tiết tinh tế**
- Nhạy cảm với biến đổi hình học lớn

Cần bảo toàn:

- Nếp nhăn, độ đàn hồi da
- Cấu trúc xương mặt
- Tỷ lệ các bộ phận

Chủ động tránh:

- Xoay lớn ($>15^\circ$)
- Thay đổi màu mạnh
- Cắt xén sâu



Hình 6: Augmentation nhẹ giúp giữ đặc trưng tuổi.

Hạn chế và Hướng Phát triển

Hạn chế hiện tại:

- Augmentation còn thủ công
- Chưa tối ưu riêng cho từng nhóm tuổi
- Nhóm dữ liệu nhỏ vẫn khó học

Hướng phát triển:

● Theo nhóm tuổi:

- Trẻ em: tăng đa dạng
- Cao tuổi: hạn chế biến đổi

● Kỹ thuật nâng cao:

- MixUp / CutMix
- AutoAugment
- Augmentation học được

4.1. Cấu hình huấn luyện

Tham số	Giá trị
Framework	PyTorch
Optimizer	AdamW
Batch size	64
Epochs	50 + Early Stopping
Loss	Cross-Entropy + Class Weight
Input size	224×224
Scheduler	ReduceLROnPlateau

Chiến lược Learning Rate:

- **Classifier / FC / Head:** 1×10^{-3}
- **Backbone:** 1×10^{-4}

Chiến lược fine-tuning:

- ❶ Khởi tạo mô hình với trọng số huấn luyện trước (pre-trained)
- ❷ Huấn luyện classifier mới
- ❸ Fine-tune toàn bộ mạng (ưu tiên các layer cuối)
- ❹ Giảm learning rate dần

4.2. Đánh giá hiệu suất

Chỉ số chính:

- **Accuracy:** Độ chính xác tổng
- **Macro-F1:** Quan trọng với dữ liệu không cân bằng
- **Precision/Recall:** Cho từng lớp

Phân tích lỗi:

- Confusion Matrix
- Visual analysis

5.1. So sánh tổng quan các mô hình

Model	Params (M)	Acc (%)	F1 (%)	Infer (ms)
ResNet18	11.7	78.2	78.0	8.2
ResNet34	21.8	78.4	78.3	12.5
EfficientNet-B0	5.3	79.7	80.0	6.8
EfficientNet-B3	12.0	79.8	80.3	15.3
ViT-B/16	86.0	75.6	75.1	22.1

Nhận xét:

- **EfficientNet-B3**: Accuracy và F1 cao nhất
- **EfficientNet-B0**: Hiệu quả tham số tốt nhất (5.3M, Acc ≈ 80%)
- **ResNet18/34**: Ổn định, chênh lệch không lớn
- **ViT**: Hiệu suất thấp do dữ liệu hạn chế

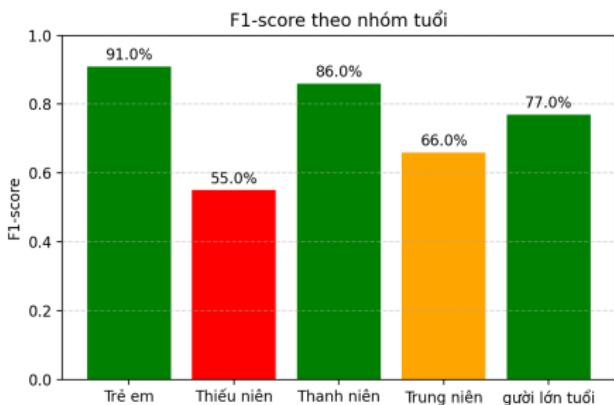
5.2. Phân tích hiệu suất theo nhóm tuổi

F1-score theo nhóm tuổi:

- **Trẻ em (G1): 93%**
- **Thanh niên (G3): 86%**
- **Trung niên (G4): 66%**
- **Người lớn tuổi (G5): 77%**
- **Thiếu niên (G2): 55%**

Nhận xét nhanh:

- Nhóm biên tuổi (G2, G4) khó phân biệt
- Nhóm có đặc trưng rõ đạt kết quả cao



Hình 7: F1-Score theo nhóm tuổi

5.3. Phân tích lỗi và khó khăn

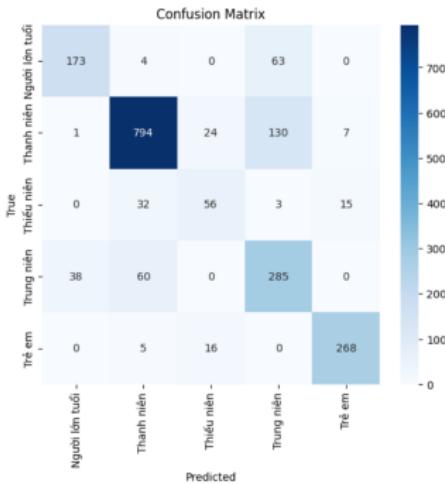
Loại lỗi chính:

- Nhầm lẫn nhóm liền kề

- Thiếu niên ↔ Thanh niên
- Trung niên ↔ Người cao tuổi

- Mẫu khó

- Tuổi sinh học khác tuổi thật
- Ánh sáng / chất lượng ảnh kém
- Biểu cảm, make-up



Hình 8: Confusion Matrix của EfficientNet-B3.

6.1. Kết luận chính

- ① **EfficientNet** cho hiệu quả tổng thể tốt nhất
- ② **EfficientNet-B3** đạt Acc cao nhất ($\sim 80\%$)
- ③ **EfficientNet-B0** tối ưu cho triển khai nhẹ
- ④ **ResNet** ổn định, dễ huấn luyện
- ⑤ **ViT** chưa phù hợp với quy mô dữ liệu hiện tại

Đề xuất theo kịch bản sử dụng:

Scenario	Mô hình đề xuất
Độ chính xác cao nhất	EfficientNet-B3
Thiết bị giới hạn tài nguyên	EfficientNet-B0
Huấn luyện nhanh, ổn định	ResNet18/34
Nghiên cứu nâng cao	Vision Transformer

6.2. Hạn chế và hướng phát triển

Hạn chế hiện tại:

- Phụ thuộc chất lượng phát hiện mặt
- Hiệu suất thấp trên nhóm thiểu số
- Chưa tối ưu cho real-time
- Dataset chủ yếu phương Tây

Hướng phát triển:

• Ngắn hạn:

- Ensemble models
- Knowledge distillation
- Data augmentation mạnh

• Dài hạn:

- Thu thập data người Việt
- Multi-task learning
- Real-time deployment

DEMO