| **BỘ GIÁO DỤC VÀ ĐÀO TẠO**  **TRƯỜNG ĐẠI HỌC SÀI GÒN**  **KHOA CÔNG NGHỆ THÔNG TIN** |
| --- |
|  |
|  |
| **LAB 6: PHÂN TÍCH LUẬN VĂN TỐT NGHIỆP** |
| Giảng viên hướng dẫn: Ths. Đỗ Như Tài |
| 3122410193 - Nguyễn Phan Tuấn Kiệt (trưởng nhóm) |
| 3122410194 – Nguyễn Thế Kiên |
| 3122410200 – Phạm Văn Kiệt |
|  |
| **Thành phố Hồ Chí Minh, tháng 04/2025** |

# **Giới thiệu**

Luận văn tốt nghiệp này của sinh viên Đại học Bách Khoa TP.HCM tập trung vào lĩnh vực nhận diện cảm xúc mặt người, ứng dụng mạng học sâu có cơ chế chú ý. Phần tóm tắt sau đây sẽ đi vào chi tiết các khía cạnh chính của công trình nghiên cứu này theo yêu cầu.

## Tên đồ án Nhận diện cảm xúc mặt người sử dụng mạng học sâu có chú ý

## Tác giả Phạm Quí Luận

## Nơi công bố Luận văn tốt nghiệp đại học, Khoa Khoa học - Kỹ thuật Máy tính, Trường Đại học Bách Khoa, Đại học Quốc gia Thành phố Hồ Chí Minh

## Năm công bố Thành phố Hồ Chí Minh, tháng 12 năm 2019

# **Mục tiêu nghiên cứu**

* Luận văn tập trung vào việc cải thiện độ chính xác của việc nhận diện cảm xúc trên các tập dữ liệu thực tế phức tạp ("in-the-wild") bằng cách sử dụng mạng học sâu.
* Áp dụng và cải tiến các mô hình học sâu hiện đại, đặc biệt là mạng Residual Masking Network tích hợp cơ chế chú ý
* Xây dựng một bộ dữ liệu mới (VEMO) chứa hình ảnh người Việt Nam để thúc đẩy nghiên cứu và ứng dụng trong bối cảnh Việt Nam, đồng thời giải quyết tranh luận về tính phổ quát của biểu cảm.
* Thực nghiệm, so sánh và đánh giá hiệu suất của mô hình đề xuất với các kiến trúc mạng học sâu hiện đại khác dưới cùng một điều kiện huấn luyện công bằng.

# **Cơ sở lý thuyết**

* **Cảm xúc con người thông qua biểu thị nét mặt**
  + Biểu thị cảm xúc trên khuôn mặt: Phân tích cách cảm xúc được thể hiện qua các cơ mặt, vai trò của nét mặt trong giao tiếp, và sự liên kết giữa hoạt động cơ mặt và trạng thái cảm xúc (3.1.1)
  + Hệ thống mã hóa cơ mặt (FACS - Facial Action Coding System): Sử dụng FACS để mô tả khách quan các chuyển động cơ mặt (Action Units - AUs) và tổ hợp chúng để xác định các biểu cảm cơ bản (3.1.2)
* **Mạng nơ-ron tích chập *(CNN)* trong bài toán nhận dạng ảnh**
  + Tổng quan về CNN: Giới thiệu cấu trúc cơ bản, cách các nơ-ron được sắp xếp theo 3 chiều (rộng, cao, sâu), và ưu điểm so với mạng nơ-ron thông thường cho dữ liệu ảnh (3.2.1)
  + Các lớp cơ bản trong CNN: Mô tả chức năng và hoạt động của
    - Lớp Tích chập (Convolutional Layer - CONV) với các khái niệm như trường tiếp nhận, chia sẻ trọng số, stride, padding (3.2.3)
    - Lớp Gộp (Pooling Layer - POOL) để giảm kích thước không gian và kiểm soát overfitting (3.2.4)
* **Kiến trúc mã hóa - giải mã và Cơ chế chú ý**
  + Kiến trúc mã hóa - giải mã: Giới thiệu khái niệm, vai trò của phần mã hóa (tạo thể hiện tiềm ẩn) và giải mã (tạo đầu ra), lấy ví dụ từ kiến trúc U-Net (3.3.1)
  + Cơ chế chú ý (Attention Mechanism): Giải thích ý tưởng cơ chế chú ý giúp mô hình tập trung vào các vùng thông tin quan trọng, nguồn gốc từ dịch máy và ứng dụng trong thị giác máy tính, đặc biệt là trong Residual Attention Network và cách nó có thể áp dụng cho bài toán nhận diện cảm xúc (3.3.2)
* **Phương pháp nhận diện khuôn mặt**  
  Đề cập đến việc sử dụng các phương pháp có sẵn (như trong OpenCV) để thực hiện bước phát hiện khuôn mặt, là tiền đề cho mô-đun phân loại cảm xúc (3.4)

# **Phương pháp thu thập và xử lý thông tin**

* **Phương pháp thu thập dữ liệu**  
  *Sử dụng hai bộ dữ liệu chính*
  + **FER2013:** Bộ dữ liệu công khai từ cuộc thi Kaggle năm 2013, gồm ảnh xám 48x48, chứa dữ liệu thực tế phức tạp với 7 lớp cảm xúc, nhưng bị mất cân bằng và có nhiễu (1.3 - 4.2)
  + **EMO:** Bộ dữ liệu do tác giả tự xây dựng, tập trung vào hình ảnh người Việt Nam. Thu thập từ nguồn ảnh công cộng (Google Images, Flickr) và trích xuất từ video phim, Youtube. Gắn nhãn thủ công kết hợp với ảnh từ AffectNet. Tổng cộng ~36,470 ảnh màu đa dạng, chia theo tỉ lệ 0.7 (huấn luyện) - 0.15 (kiểm thử) - 0.15 (kiểm tra) (1.2 - 4.3)
* **Xử lý thông tin *(dữ liệu)***
  + **Phát hiện khuôn mặt:** Sử dụng phương pháp có sẵn trong OpenCV 3.4 cho những ảnh chưa được xử lý (6.2)
  + **Làm giàu dữ liệu (Data Augmentation):** Áp dụng các phép biến đổi như lật ảnh theo chiều ngang và xoay ảnh ngẫu nhiên (từ -30 đến 30 độ) để tăng tính đa dạng và chống overfitting [source: 281, 862, 863, 869]. Sử dụng thư viện imgaug (6.2)
  + **Lớp định nghĩa dữ liệu (Dataset Class):** Xây dựng lớp kế thừa từ torch.utils.data.Dataset để tải và chuẩn bị dữ liệu. Ảnh được sửa kích cỡ về 224x224, chuyển thành 3 kênh (nếu cần), áp dụng làm giàu dữ liệu (trong giai đoạn huấn luyện), và chuyển đổi thành Tensor (6.2)
  + **Tiền xử lý khác:** Chuẩn hóa kích thước ảnh đầu vào (2.2.1), chuẩn hóa giá trị pixel (như trong nghiên cứu của Y. Tang (2.1.1)

# 

# **Kết quả đạt được**

* **Về dữ liệu**Xây dựng thành công bộ dữ liệu VEMO, là bộ dữ liệu đầu tiên về biểu cảm mặt người tập trung vào hình ảnh người Việt Nam, có độ khó cao và sát thực tế, sẽ được công khai (7.1)
* **Về kiến trúc đề xuất *(Residual Masking Network)*** (7.1)
  + Đề xuất kiến trúc mới tích hợp cơ chế chú ý (lấy cảm hứng từ kiến trúc mã hóa-giải mã) vào các khối trích xuất đặc trưng (Residual Unit), huấn luyện end-to-end
  + Đạt độ chính xác 74.14% trên tập FER2013, cao hơn nhiều mạng hiện đại khác được huấn luyện lại trong cùng điều kiện
  + Đạt độ chính xác 65.949% trên tập VEMO, cao hơn các mạng ResNet và ResAttNet được huấn luyện lại
  + Kết quả trên cả hai tập dữ liệu cho thấy mô hình hoạt động tốt với dữ liệu thực tế phức tạp và mất cân bằng
  + Trực quan hóa bằng GradCAM và gộp trung bình kênh cho thấy mô hình tập trung vào các vùng mặt quan trọng liên quan đến cảm xúc
* **Về phương pháp thực nghiệm và kết quả báo cáo** (7.1)
  + Thiết lập khung thực nghiệm công bằng, có khả năng tái lặp kết quả
  + Kết quả huấn luyện lại nhiều mạng hiện đại cung cấp cơ sở so sánh khách quan
  + Áp dụng phương pháp học kết hợp (ensemble learning) đơn giản (trung bình tổng xác suất) của mô hình đề xuất và 6 CNN khác, đạt độ chính xác 76.82% trên FER2013, thuộc nhóm kết quả hàng đầu được công bố
* **Về bài toán**  
  Đóng góp một phương pháp hiệu quả cho bài toán nhận diện cảm xúc từ ảnh tĩnh trong điều kiện thực tế, có tiềm năng ứng dụng (7.1)

# **Hạn chế của công trình**

# Luận văn không có mục **"Hạn chế"** riêng biệt, tuy nhiên, dựa trên nội dung ở chương **"Kết luận"** (6.8) và **"Tổng kết"** (7.1, 7.2), có thể suy luận ra một số hạn chế và điểm cần cải thiện sau:

* **Độ phức tạp và số lượng tham số:** Mạng Residual Masking Network có số lượng tham số lớn (142.9 triệu), điều này có thể là một điểm trừ về mặt hiệu quả tính toán và triển khai trên các thiết bị hạn chế tài nguyên [Trang 74, 84].
* **Thiếu thực nghiệm sâu rộng về siêu tham số:** Quá trình thực nghiệm chưa khám phá hết các cấu hình siêu tham số khác nhau của Residual Masking Network (ví dụ: độ sâu của Masking Block ở các giai đoạn khác nhau) [Trang 84, 85].
* **Phạm vi kiểm chứng:** Mạng được thiết kế cho bài toán phân lớp nhưng chủ yếu chỉ được kiểm chứng trên bài toán nhận diện cảm xúc. Hiệu quả trên các bộ dữ liệu phân lớp ảnh khác chưa được chứng minh [Trang 84].
* **Phụ thuộc vào bước phát hiện khuôn mặt:** Hiệu quả của toàn hệ thống vẫn phụ thuộc vào chất lượng của bước phát hiện và cắt khuôn mặt ban đầu. Các phương pháp phát hiện điểm mốc truyền thống gặp khó khăn với ảnh nhỏ hoặc góc mặt nghiêng [Trang 85], dù mạng đề xuất không trực tiếp dùng điểm mốc nhưng chất lượng ảnh đầu vào vẫn quan trọng.
* **Học kết hợp còn đơn giản:** Phương pháp học kết hợp được sử dụng (trung bình tổng xác suất) là tương đối đơn giản và thực hiện ngoại tuyến. Các kỹ thuật ensemble phức tạp hơn có thể cải thiện kết quả hơn nữa [Trang 65].
* **Chưa xây dựng hệ thống hoàn chỉnh:** Luận văn tập trung vào mô-đun phân loại cốt lõi, nhưng chưa xây dựng và thí điểm một hệ thống nhận diện cảm xúc hoàn chỉnh để tích hợp và kiểm chứng trong ứng dụng thực tế [Trang 85].