**TRƯỜNG ĐẠI HỌC SÀI GÒN**

KHOA CÔNG NGHỆ THÔNG TIN



BÁO CÁO BÀI TẬP

**Phân tích luận văn tốt nghiệp**

**HỌC PHẦN:** PHƯƠNG PHÁP NGHIÊN CỨU

**THUỘC NHÓM NGÀNH:** CÔNG NGHỆ THÔNG TIN

**GIẢNG VIÊN:** ĐỖ NHƯ TÀI

THÀNH VIÊN NHÓM 77:

3123410024 - NGÔ THƯỢNG BẢO

3123410242 - NGUYỄN TRỌNG NHÂN

3123410174 - NGUYỄN PHẠM TUẤN KHÔI

3123410387 - NGUYỄN HỮU TRI

**Mục lục**

[I. Tên đồ án 1](#_Toc195901207)

[II. Tác giả 1](#_Toc195901208)

[III. Nơi công bố 2](#_Toc195901209)

[IV. Năm công bố 2](#_Toc195901210)

[V. Mục tiêu nghiên cứu 2](#_Toc195901211)

[VI. Cơ sở lý thuyết 3](#_Toc195901212)

[VII. Phương pháp thu thập và xử lý thông tin 5](#_Toc195901213)

[VIII. Kết quả đạt được 6](#_Toc195901214)

[IX. Hạn chế của công trình 8](#_Toc195901215)

# Tên đồ án

"Nhận diện cảm xúc mặt người sử dụng mạng học sâu có chú ý" do sinh viên Phạm Quí Luận thực hiện tại Trường Đại học Bách Khoa TP. Hồ Chí Minh, tháng 12/20191.

# Tác giả

Tác giả: Phạm Quý Luân  
Là tác giả viết luận văn đề tài tốt nghiệp về nhận diện cảm xúc mặt người sử dụng mạng học sâu có chú ý, theo năm công bố được ghi trên đồ án, tác giả từng là sinh viên Khoa Khoa học – kỹ thuật máy tính của trường Đại học Bách Khoa thuộc Đại học Quốc gia Thành phố Hồ Chí Minh. Phần tác giả được anh đặt ngắn gọn ở trang bìa và không có thông tin nào thêm trong phần luận văn để giành chỗ cho để tài của mình.

# Nơi công bố

Nơi công bố: Khoa Khoa học - Kỹ thuật Máy tính Trường Đại học Bách Khoa - Đại học Quốc gia TP. Hồ Chí Minh

# Năm công bố

Luận văn được công bố vào tháng 12 năm 2019. Đây là luận văn tốt nghiệp đại học được hoàn thành tại Trường Đại học Bách Khoa – Đại học Quốc gia TP. Hồ Chí Minh, Khoa Khoa học – Kỹ thuật Máy tính, dưới sự hướng dẫn của TS. Trần Tuấn Anh. Luận văn tuân thủ đúng cấu trúc một công trình khoa học bậc đại học với đầy đủ phần lời cảm ơn, lời cam đoan, mục lục, nội dung chi tiết và phụ lục. Công trình cũng đã được phản biện bởi giảng viên chuyên môn là TS. Nguyễn Hồ Mẫn Rạng.

# Mục tiêu nghiên cứu

* Phát hiện và phân tích cảm xúc từ các chuyển động trên khuôn mặt người trong môi trường phức tạp.
* Sử dụng mạng học sâu có cơ chế chú ý (Residual Masking Network) để phân lớp cảm xúc dựa trên ảnh đầu vào.
* Áp dụng phương pháp học kết hợp nhiều mô hình hiện đại nhằm tăng cường độ chính xác nhận diện cảm xúc.
* Xây dựng và phát triển bộ dữ liệu khuôn mặt người Việt Nam nhằm góp phần phát triển bài toán nhận diện cảm xúc tại Việt Nam.
* Đạt độ chính xác cao trên tập dữ liệu chuẩn FER2013, với kết quả tốt nhất đạt 76.82%.

# Cơ sở lý thuyết

Gồm 4 trọng tâm chính:

* **Cảm xúc con người thông qua biểu thị nét mặt**

Tác giả cung cấp cái nhìn toàn diện hơn về các biểu cảm cảm xúc trên khuôn mặt người cũng như khẳng định rằng khuôn mặt con người là một trong những phần quang trọng nhất trong việc phân tích biểu cảm của con người.

Bao gồm 2 ý chính:

* **Biểu thị cảm xúc trên khuôn mặt người**

Bàn về cảm xúc trên khuôn mặt con người, cách mà cảm xúc được sinh ra từ các tín hiệu của não bộ truyền đến cơ thông qua các dây thần kinh, cũng như nói về sự phổ quát về cảm xúc trên khuôn mặt người

* **Hệ thống mã hóa cơ mặt:**

Là hệ thống mô tả đầy đủ những hành vi trên gương mặt con người bằng một tập hợp các đơn vị hoạt động được định nghĩa (Action Unit - AUs). Hệ thống làm việc dựa trên việc phát hiện một hoặc nhiều chuyển động cơ mặt gọi là AUs, sau đó tham khảo đến một mục lục để chỉ ra loại cảm xúc nào đang được biểu hiện. Có 46 đơn vị hoạt động khác nhau được mô tả trong [14]. Từ đó, Ekman và Rosenberg mô tả các biểu cảm khuôn mặt từ tổ hợp của các đơn vị hoạt động ấy.

* **Mạng nơ ron tích chập trong bài toán nhận ảnh**

Nói về mạng nơ ron tích chập, khác biệt của nó so với mạng nơ ron thông thường cũng như các thông tin tổng quan về mạng, các lớp cơ bản và lớp tích chập của nó, cách mà mạng hoạt động.

* **Kiến trúc mã hóa - giải mã và cơ chế chú ý đối với bài toán nhận ảnh**

Gồm 2 phần như trong tiêu đề

* **Kiến trúc mã hóa – giải mã**

Kiến trúc mã hóa - giải mã theo tên gọi của nó, bao gồm 2 phần riêng biệt: phần mã hóa và phần giải mã. Trong bài toán nhận ảnh, ảnh đầu vào sẽ được đi qua phần mã hóa trước để tạo ra thể hiện tiềm ẩn sau đó phần thể hiện sẽ tiếp tục đi qua phần giải mã để cho ra đầu ra mong muốn.

* **Cơ chế chú ý**

Là cơ chế giúp mô hình máy tính tập trung vào các điểm đáng chú ý để đưa ra kết quả như mong muốn, cơ chế này là đầu tiên được giới thiệu lần đầu trong bài báo Neural Machine Translation by Jointly Learning to Align and Translate được viết bởi Dzmitry Bahdanau, Kyunghyun Cho, và Yoshua Bengio

* **Phương pháp nhận diện khuôn mặt**

Phần này tác giả tuyệt đối tham khảo và sử dụng lại chức năng của OpenCV nên không được trình bày thêm. Nhưng để làm rõ hơn thì OpenCV là phần mềm máy tính dành cho thị giác máy tính được phục vụ cho các việc như nhận diện khuôn mặt hay giám sát dữ liệu,…

# Phương pháp thu thập và xử lý thông tin

* **Phương pháp thu nhập thông tin:**

1. Tập dữ liệu sử dụng:

FER2013: Một bộ dữ liệu phổ biến, chứa các ảnh khuôn mặt được gán nhãn cảm xúc, thu thập từ trong thực tế (in-the-wild).

VEMO: Bộ dữ liệu do tác giả tự xây dựng, bao gồm ảnh các khuôn mặt của người Việt Nam với nhiều biểu cảm khác nhau (giận dữ, ghê tởm, sợ hãi, hạnh phúc, buồn bã, bất ngờ và bình thường). Bộ dữ liệu này được tạo ra để hỗ trợ nghiên cứu nhận diện cảm xúc phù hợp với đặc điểm người Việt.

1. Cách thu thập dữ liệu VEMO:

Không có mô tả chi tiết, nhưng tác giả đã đề cập rõ rằng tác giả đã tự thu thập ảnh khuôn mặt người Việt Nam trong thực tế để tạo ra VEMO, góp phần phát triển bài toán tại Việt Nam (chi tiết ở trong Chương 4).

**- Phương pháp xử lý thông tin:**

1. Tiền xử lý dữ liệu:

+ Ảnh được chuẩn hóa, ảnh đầu vào được cắt, căn chỉnh khuôn mặt, và tăng cường dữ liệu như xoay ảnh, lật ảnh, nhiễu Gaussian.

1. Xây dựng mô hình:

+ Residual Masking Network (RMN): Mạng nơ-ron tích chập kết hợp cơ chế chú ý để trích xuất đặc trưng biểu cảm khuôn mặt tốt hơn.

+ Kết hợp học sâu: Sử dụng các mô hình học kết hợp (ensemble) để nâng cao độ chính xác.

1. Đánh giá mô hình:

+ Thực nghiệm trên tập FER2013 và VEMO

+ Tiêu chí đánh giá: độ chính xác (accuracy), ma trận nhầm lẫn (confusion matrix), và trực quan hóa bằng Grad-CAM.

# Kết quả đạt được

1. Trên tập dữ liệu FER2013:  
- Mô hình Residual Masking Network (RMN) kết hợp học ensemble đạt độ chính xác 76.82%, cao hơn tất cả các mô hình hiện đại khác được huấn luyện lại cùng cấu hình.  
- Các kết quả so sánh:  
 • VGG19: 71.16%  
 • ResNet18: 72.08%  
 • Inception V3: 73.24%  
 • CBAM-ResNet50: 74.35%  
- Mô hình RMN của tác giả còn vượt qua cả mô hình từng chiến thắng cuộc thi FER2013 (Yichuan Tang, 2013) với độ chính xác 71.2%.

2. Trên bộ dữ liệu VEMO:  
- RMN đạt độ chính xác cao hơn các mô hình hiện đại khác.  
- Đây là một trong những mô hình đầu tiên được thử nghiệm và đánh giá trên tập dữ liệu chứa khuôn mặt người Việt.

3. Kết quả trực quan hóa (GradCAM):  
- Các biểu đồ GradCAM chứng minh mạng RMN tập trung vào các vùng mặt quan trọng như mắt, miệng, chân mày.  
- Điều này thể hiện khả năng học đặc trưng tốt hơn các mô hình baseline.

4. Đóng góp chính:  
- Đề xuất kiến trúc Residual Masking Network (RMN) cho bài toán nhận diện cảm xúc.  
- Phát triển bộ dữ liệu VEMO dành riêng cho người Việt.  
- Chứng minh hiệu quả của việc áp dụng attention và học kết hợp trong xử lý ảnh khuôn mặt.

**Tổng kết: Đánh giá kết quả đạt được**

Trong suốt quá trình thực hiện luận văn, tác giả đã nghiên cứu, thử nghiệm và cải tiến một mạng học sâu - Residual Masking Network cho bài toán nhận diện cảm xúc. Ngoài ra còn thực hiện huấn luyện lại các mạng hiện đại hiện nay và kết quả đạt được cũng nằm trong nhóm đầu của các công bố khoa học.  
• Về dữ liệu: Tập dữ liệu VEMO là bộ dữ liệu đầu tiên về biểu cảm mặt người có hình ảnh người Việt Nam, được gán nhãn bởi người Việt, có độ khó cao và sát thực tế. Dữ liệu sẽ được công khai, hứa hẹn đóng góp cho nghiên cứu và công nghiệp trong nước.  
• Về kiến trúc: Mạng Residual Masking Network tích hợp cơ chế chú ý vào quá trình trích xuất đặc trưng. Luận văn cũng đề xuất một loại kết nối mới giữa các khối đặc trưng và mô hình được huấn luyện end-to-end không cần đặc trưng trung gian. Kết quả thử nghiệm cho thấy hiệu quả nổi bật với bài toán phân lớp.  
• Về thực nghiệm: Tất cả các mô hình được huấn luyện trong cùng môi trường, ngôn ngữ và cấu hình nhằm đảm bảo tính công bằng và khả năng tái lập. Kết quả và cách cài đặt được báo cáo chi tiết, có thể dùng làm benchmark cho các nghiên cứu sau này.  
• Về kết quả: Mô hình đạt độ chính xác cao, các ma trận nhầm lẫn trên hai tập dữ liệu VEMO và FER2013 đều cho thấy hiệu quả cao. Đặc biệt, trên tập FER2013, kết quả nằm trong nhóm đầu so với các công bố khoa học, thậm chí vượt qua một số mô hình sử dụng kỹ thuật ensemble.  
Tổng thể, luận văn đã ứng dụng hiệu quả kiến thức về mạng học sâu để giải quyết bài toán nhận diện cảm xúc mặt người, mang lại giá trị cả trong nghiên cứu và ứng dụng thực tế.

# Hạn chế của công trình

* Mặc dù phương pháp đề xuất có kết quả tốt hơn một số phương pháp hiện đại, độ chính xác vẫn chưa đạt mức hoàn hảo, còn khoảng 23% sai số trên tập dữ liệu chuẩn.
* Bài toán nhận diện cảm xúc mặt người vẫn còn nhiều thách thức do sự đa dạng về biểu cảm, độ sáng, góc độ, cường độ biểu cảm và sự xuất hiện trong môi trường thực tế.
* Bộ dữ liệu Việt Nam mới được xây dựng và phát triển, có thể chưa đa dạng và phong phú như các bộ dữ liệu quốc tế, ảnh hưởng đến khả năng tổng quát của mô hình.
* Các phương pháp học sâu thường yêu cầu tài nguyên tính toán lớn và thời gian huấn luyện dài, có thể là một hạn chế trong ứng dụng thực tế.
* Một số trường hợp dự đoán sai vẫn tồn tại, cần tiếp tục nghiên cứu để cải thiện độ chính xác và khả năng nhận diện trong điều kiện thực tế đa dạng hơn.

Tóm lại, đồ án tập trung vào phát triển mô hình nhận diện cảm xúc mặt người sử dụng mạng học sâu có chú ý, đồng thời xây dựng bộ dữ liệu khuôn mặt người Việt Nam, đạt được kết quả khả quan nhưng vẫn còn hạn chế về độ chính xác và khả năng ứng dụng trong thực tế phức tạp.