TRƯỜNG ĐẠI HỌC SÀI GÒN

`

KHOA CÔNG NGHỆ THÔNG TIN

Bài Tập 6: Phân Tích Luận Văn Của Phạm Quí Luận

Môn Học: Phương Pháp Nghiên Cứu Khoa Học

Giảng viên: Đỗ Như Tài

Nhóm

Phạm Tấn Khương – 3122410191

Hoàng Vũ - 3122560089

Huỳnh Thanh Bình - 3122410033

Nguyễn Minh Tú - 3120411167

Thành phố HCM, ngày 18 tháng 4 năm 2025

|  |  |  |
| --- | --- | --- |
| Hoàng Vũ |  | 100% |
| Huỳnh Thanh Bình |  | 100% |
| Phạm Tấn Khương |  | 100% |
| Nguyễn Minh Tú |  | 100% |

Mục lục

[**Chương I : Tổng quan về tiểu luận** 4](#_Toc195880358)

[**Chương II : Mục tiêu nghiên cứu và cơ sở lý thuyết** 4](#_Toc195880359)

[**1. Phạm vi nghiên cứu** 4](#_Toc195880360)

[**2. Mục tiêu nghiên cứu** 4](#_Toc195880361)

[**3. Cơ sở lý thuyết** 4](#_Toc195880362)

[**3.1 Cảm xúc con người thông qua biểu thị nét mặt** 4](#_Toc195880363)

[**3.2 Mạng nơ-ron tích chập trong nhận diện ảnh** 5](#_Toc195880364)

[**3.2.4 Lớp Gộp (Pooling Layer – POOL)** 6](#_Toc195880365)

[**3.3 Kiến trúc mã hóa - giải mã & Cơ chế chú ý** 7](#_Toc195880366)

[**3.3.1 Kiến trúc mã hóa - giải mã** 7](#_Toc195880367)

[**3.3.2 Cơ chế chú ý (Attention Mechanism)** 8](#_Toc195880368)

[**3.4 Phương pháp nhận diện khuôn mặt** 9](#_Toc195880369)

[**Chương III. Phương pháp thu nhập và xử lý thông tin** 10](#_Toc195880370)

[**1. Phương pháp thu thập thông tin** 10](#_Toc195880371)

[**1.1. Nguồn dữ liệu** 10](#_Toc195880372)

[**1.2. Hình thức thu thập** 10](#_Toc195880373)

[**2. Phương pháp xử lý thông tin** 10](#_Toc195880374)

[**2.1. Phát hiện và cân chỉnh khuôn mặt** 10](#_Toc195880375)

[**2.2. Làm giàu dữ liệu (Data Augmentation**) 10](#_Toc195880376)

[**2.3. Chuẩn hóa dữ liệu** 10](#_Toc195880377)

[Chương IV. Kết luận 11](#_Toc195880378)

[1.Kết quả đạt được 11](#_Toc195880379)

[Dựa trên nội dung luận văn, các kết quả chính đạt được bao gồm: 11](#_Toc195880380)

[**2. Hạn chế của công trình** 12](#_Toc195880381)

# **Chương I : Tổng quan về tiểu luận**

Tên đề tài: Nhận diện cảm xúc mặt người sử dụng mạng học sâu có chú ý

Tác giả : Phạm Quí Luận

Nơi công bố : Đại học Bách Khoa TPHCM

Năm công bố : 12/2019

# **Chương II : Mục tiêu nghiên cứu và cơ sở lý thuyết**

## **1. Phạm vi nghiên cứu**

* Do tính đa dạng và chưa có chuẩn hóa chung trong lĩnh vực điện toán cảm xúc, tác giả giới hạn phạm vi nghiên cứu như sau:
* Về phương pháp: Tập trung vào khảo sát, thử nghiệm và cải tiến các mô hình học sâu hiện đại, đặc biệt là các kiến trúc mạng nơ-ron có cơ chế chú ý.
* Về dữ liệu: Chỉ sử dụng dữ liệu ảnh tĩnh trong môi trường thực tế phức tạp (in-the-wild), không phải dữ liệu lab. Hai bộ dữ liệu chính là FER2013 và VEMO (dữ liệu người Việt tự xây dựng).
* Về đánh giá: Để đảm bảo công bằng khi so sánh, tác giả tự huấn luyện lại các mô hình chuẩn dưới cùng điều kiện, thay vì chỉ so sánh với kết quả được công bố trước đó.

## **2. Mục tiêu nghiên cứu**

* Cải thiện độ chính xác của mô hình nhận diện cảm xúc trên ảnh thực tế phức tạp.
* Xây dựng một bộ dữ liệu khuôn mặt người Việt nhằm nâng cao tính phù hợp và đóng góp cho cộng đồng nghiên cứu trong nước.
* Chấp nhận và khai thác giả thuyết biểu cảm phổ quát (Universal Hypothesis) – biểu cảm được chia thành 6 loại cơ bản và bài toán được xử lý theo hướng phân lớp (classification).
* Thực hiện đối chiếu kết quả với cả các mô hình hiện đại và các báo cáo khoa học có liên quan để đánh giá tính hiệu quả của mô hình đề xuất.

## **3. Cơ sở lý thuyết**

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

#### **3.1.1 Biểu thị cảm xúc**

* Cảm xúc được thể hiện rõ qua biểu cảm khuôn mặt.
* Ekman & Friesen xác định 6 cảm xúc cơ bản là: Giận dữ, Ghê tởm, Sợ hãi, Hạnh phúc, Buồn bã và Ngạc nhiên (sau này thêm cả “Bình thường” và “Khinh bỉ”).
* Mặc dù có tranh cãi về tính phổ quát (universal hypothesis), các nghiên cứu và bộ dữ liệu hiện nay vẫn dựa vào giả thuyết này để xây dựng hệ thống phân lớp cảm xúc.

#### **3.1.2 Hệ thống mã hóa cơ mặt – FACS (Facial Action Coding System)**

* Phân tích biểu cảm dựa trên các đơn vị hành động cơ mặt (Action Units - AUs) như nhíu mày, nâng khóe miệng, v.v.
* FACS giúp định lượng chuyển động cơ học khuôn mặt để suy ra cảm xúc.

### **3.2 Mạng nơ-ron tích chập trong nhận diện ảnh**

CNN (Convolutional Neural Networks) là mạng nơ-ron được thiết kế riêng cho dữ liệu hình ảnh, giúp giảm số lượng tham số nhờ kết nối cục bộ và chia sẻ trọng số. Mỗi lớp trong CNN biến đổi một khối đầu vào 3 chiều thành một khối đầu ra 3 chiều thông qua các phép biến đổi khả vi.

#### **3.2.1 Tổng quan**

* Mỗi lớp của CNN có 3 chiều: rộng, cao, sâu.
* CNN gồm nhiều lớp cơ bản như: Convolution (CONV), ReLU, Pooling (POOL) và Fully Connected (FC).
* Kiến trúc mẫu: [INPUT - CONV - RELU - POOL - FC].

Ảnh có chứa văn bản, ảnh chụp màn hình, thiết kế

Nội dung do AI tạo ra có thể không chính xác.

#### **3.2.2 Các lớp cơ bản**

* CONV: tính toán đặc trưng cục bộ bằng các bộ lọc.
* RELU: áp dụng hàm kích hoạt.
* POOL: giảm kích thước không gian.
* FC: tính điểm lớp cuối.

#### **3.2.3 Lớp tích chập (CONV)**

* CONV dùng các bộ lọc (filter) trượt trên ảnh đầu vào để tạo ra feature maps.
* Kết nối cục bộ: mỗi nơ-ron chỉ kết nối với một vùng nhỏ đầu vào.
* Siêu tham số chính:
  + Depth: số lượng bộ lọc.
  + Stride: bước trượt của bộ lọc.
  + Zero-padding: đệm thêm 0 để giữ kích thước đầu ra.
  + Công thức tính kích thước đầu ra:

Ảnh có chứa Phông chữ, số, văn bản, ảnh chụp màn hình

Nội dung do AI tạo ra có thể không chính xác.

* Chia sẻ trọng số giúp giảm số tham số và tăng tính khái quát của mô hình.

Ảnh có chứa văn bản, biểu đồ, ảnh chụp màn hình, thiết kế

Nội dung do AI tạo ra có thể không chính xác.

### **3.2.4 Lớp Gộp (Pooling Layer – POOL)**

Lớp gộp giúp giảm kích thước không gian của đặc trưng đầu vào, từ đó giảm số tham số và chi phí tính toán, đồng thời hạn chế overfitting. Hai phương pháp phổ biến là:

* Max Pooling: lấy giá trị lớn nhất trong vùng.
* Average Pooling: lấy giá trị trung bình trong vùng.

Pooling không học tham số và hoạt động độc lập trên từng kênh đặc trưng. Nó giúp mạng ổn định hơn với dịch chuyển nhỏ và thường đặt sau lớp ReLU. Ví dụ, Max Pooling với vùng 2x2 và stride 2 sẽ giảm kích thước chiều rộng và chiều cao đi một nửa.

### **3.3 Kiến trúc mã hóa - giải mã & Cơ chế chú ý**

### **3.3.1 Kiến trúc mã hóa - giải mã**

#### **3.3.1.1 Giới thiệu khái quát:**

Kiến trúc mã hóa - giải mã gồm hai phần chính: phần mã hóa (encoder) và phần giải mã (decoder).

Ảnh đầu vào sẽ được qua phần mã hóa để tạo ra thể hiện tiềm ẩn, sau đó được giải mã để cho ra đầu ra mong muốn.

#### **3.3.1.2 Ứng dụng trong nhận diện ảnh:**

* Mạng U-Net là một trong những kiến trúc nổi tiếng trong lĩnh vực phân đoạn ảnh, đặc biệt là trong phân đoạn ảnh y tế.
* Mạng U-Net bao gồm các lớp convolutional để trích xuất đặc trưng, kết hợp giữa phần mã hóa và giải mã để tăng dần độ phân giải của ảnh.
* Lớp convolution 1x1 cuối cùng giúp phân lớp, từ đó cho ra kết quả phân đoạn chính xác.

#### **3.3.1.3 Cơ chế hoạt động:**

* Mã hóa: Trích xuất đặc trưng từ ảnh đầu vào, giảm kích thước ảnh nhưng vẫn giữ lại thông tin quan trọng.
* Giải mã: Tăng độ phân giải của đặc trưng, kết hợp với thông tin từ phần mã hóa để đưa ra kết quả chính xác.
* Mạng U-Net giúp cải thiện phân đoạn ảnh, đặc biệt trong các bài toán y tế khi có ít dữ liệu huấn luyện.

Ảnh có chứa văn bản, ảnh chụp màn hình, biểu đồ, hàng

Nội dung do AI tạo ra có thể không chính xác.

### **3.3.2 Cơ chế chú ý (Attention Mechanism)**

#### **3.3.2.1 Giới thiệu cơ chế chú ý:**

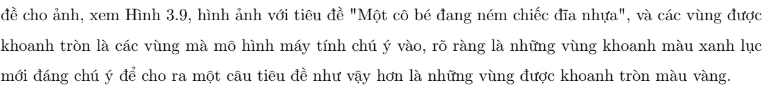
* Cơ chế chú ý (Attention Mechanism) giúp mô hình học sâu tập trung vào các phần quan trọng của dữ liệu, cải thiện hiệu suất và giảm nhiễu.
* Ban đầu, cơ chế chú ý được phát triển trong dịch máy, nhưng sau đó được mở rộng sang nhiều bài toán khác như phân loại ảnh và phân đoạn ảnh.

#### **3.3.2.2 Ứng dụng trong nhận diện ảnh:**

* Trong bài toán nhận diện cảm xúc khuôn mặt, các vùng như mắt và miệng đóng góp nhiều hơn so với các vùng khác trên khuôn mặt như cằm hay tóc.
* Mô hình sử dụng điểm mốc khuôn mặt để xác định vùng cần chú ý, từ đó giúp tăng độ chính xác trong phân loại cảm xúc.

Ảnh có chứa cỏ, ngoài trời, trang phục, người

Nội dung do AI tạo ra có thể không chính xác.



### **3.4 Phương pháp nhận diện khuôn mặt**

Phát hiện mặt người là bài toán cơ bản được xây dựng từ nhiều năm nay, có nhiều phương pháp được đưa ra như sử dụng các phương thức truyền thống tới các mạng học sâu hiện đại. Vào tháng 8, 2017, với bản OpenCV 3.3 được phát hành, một phiên bản phát hiện khuôn mặt bằng mạng Single Shot Detector được cài đặt bằng Caffe đã được sẵn sàng sử dụng ở dưới mô-đun dnn . VÌ phần này tôi tuyệt đối tham khảo và sử dụng lại chức năng của OpenCV nên không trình bày thêm ở đây.

# **Chương III. Phương pháp thu nhập và xử lý thông tin**

## **1. Phương pháp thu thập thông tin**

### **1.1. Nguồn dữ liệu**

Nghiên cứu sử dụng hai bộ dữ liệu chính phục vụ cho bài toán nhận diện cảm xúc khuôn mặt:  
- FER2013: Là một tập dữ liệu công khai trên nền tảng Kaggle, gồm 35.887 ảnh đen trắng kích thước 48×48 pixel, phân loại theo 7 biểu cảm cảm xúc cơ bản.  
- VEMO: Là bộ dữ liệu do chính tác giả xây dựng, bao gồm ảnh khuôn mặt người Việt Nam với đa dạng biểu cảm trong điều kiện thực tế.

### **1.2. Hình thức thu thập**

- Dữ liệu FER2013 được thu thập thông qua tải trực tiếp từ nguồn công khai.  
- Bộ dữ liệu VEMO được thu thập từ nhiều nguồn thực tế, sau đó được gán nhãn theo cảm xúc bởi người thực hiện.

## **2. Phương pháp xử lý thông tin**

### **2.1. Phát hiện và cân chỉnh khuôn mặt**

- Sử dụng thư viện Dlib để phát hiện 68 điểm đặc trưng trên khuôn mặt.  
- Áp dụng biến đổi affine để căn chỉnh khuôn mặt về một dạng chuẩn, phục vụ cho quá trình huấn luyện.

### **2.2. Làm giàu dữ liệu (Data Augmentation**)

Nhằm tăng độ đa dạng và khắc phục hiện tượng overfitting, dữ liệu được làm giàu bằng các kỹ thuật sau:

- Lật ảnh theo chiều ngang.

- Xoay ảnh trong khoảng ±30 độ.

- Thêm nhiễu Gaussian.

- Cắt ảnh ngẫu nhiên.

Kết quả: Tăng số lượng mẫu huấn luyện lên gấp 15 lần so với ban đầu.

### **2.3. Chuẩn hóa dữ liệu**

- Toàn bộ ảnh được resize về kích thước 224×224 pixel.  
- Chuyển ảnh từ 1 kênh (grayscale) sang 3 kênh màu (RGB) để tương thích với các mô hình học sâu.

- Áp dụng chuẩn hóa (mean subtraction và normalization) để đảm bảo dữ liệu đầu vào đồng đều.

## Chương IV. Kết luận

## 1.Kết quả đạt được

# Dựa trên nội dung luận văn, các kết quả chính đạt được bao gồm:

1. Hiệu suất của Residual Masking Network:

Mạng Residual Masking Network (ResMaskingNet) đạt độ chính xác 74.14% trên tập dữ liệu FER2013, vượt qua nhiều mạng phân loại hiện đại như VGG19 (70.8%), ResNet18 (72.9%), và DenseNet121 (73.16%).

Khi kết hợp ResMaskingNet với 6 mô hình CNN (phương pháp Ensemble), độ chính xác đạt 76.82% trên tập FER2013, cao nhất so với các công bố khoa học được so sánh (ví dụ: Ensemble 8 CNNs đạt 75.2%).

Trên tập dữ liệu VEMO (dữ liệu khuôn mặt người Việt Nam), ResMaskingNet đạt độ chính xác 65.949%, vượt qua các mạng như ResNet18 (63.94%) và ResNet34 (64.84%).

1. Xây dựng tập dữ liệu VEMO:

Tác giả xây dựng và phát triển tập dữ liệu VEMO chứa hình ảnh khuôn mặt người Việt Nam, góp phần thúc đẩy nghiên cứu bài toán nhận diện cảm xúc tại Việt Nam.

1. Ứng dụng cơ chế chú ý (Attention Mechanism):

ResMaskingNet tích hợp cơ chế chú ý vào phần trích xuất đặc trưng, cải thiện hiệu quả phân loại cảm xúc. Kết quả trực quan hóa bằng GradCAM chứng minh khả năng giải thích quyết định của mô hình.

Hiệu quả trên dữ liệu thực tế:

Mô hình được thử nghiệm trên các hình ảnh thực tế (ví dụ: cảnh trong phim "Mắt Biếc", "Harry Potter"), cho thấy khả năng nhận diện cảm xúc tốt trong các tình huống thực tế.

1. Thực nghiệm công bằng:

Các thí nghiệm được tiến hành trong môi trường công bằng, sử dụng cùng cấu hình và phương pháp, đảm bảo kết quả đáng tin cậy và có thể tái hiện.

1. Khả năng xử lý dữ liệu mất cân bằng:

Ma trận bối rối cho thấy mô hình xử lý tốt dữ liệu mất cân bằng, nhận diện hiệu quả 7 lớp cảm xúc cơ bản (giận dữ, ghê tởm, sợ hãi, hạnh phúc, buồn bã, ngạc nhiên, bình thường).

1. Tiềm năng ứng dụng:

Kết quả không chỉ có giá trị trong nghiên cứu mà còn có thể áp dụng thực tế, ví dụ trong các hệ thống giám sát hoặc phân tích cảm xúc.

## **2. Hạn chế của công trình**

Luận văn cũng chỉ ra một số hạn chế như sau:

1. **Thời gian nghiên cứu hạn chế**:

Do thời gian có hạn, tác giả chưa thể thử nghiệm ResMaskingNet trên các kích cỡ mạng khác nhau hoặc tối ưu hóa thêm các siêu tham số để đạt hiệu suất tốt hơn.

1. **Số lượng tham số lớn**:

ResMaskingNet có số lượng tham số lớn (**142.9 triệu**), cao hơn nhiều so với các mạng như ResNet18 (11.2 triệu) hay DenseNet121 (6.9 triệu), dẫn đến chi phí tính toán cao và có thể không phù hợp với các ứng dụng yêu cầu tài nguyên thấp.

1. **Chưa thử nghiệm trên các bài toán khác**:

Mô hình chủ yếu được thiết kế và thử nghiệm cho bài toán phân loại cảm xúc. Tác giả chưa kiểm chứng hiệu quả của ResMaskingNet trên các bài toán phân loại khác hoặc các tập dữ liệu lớn hơn.

1. **Hạn chế về tập dữ liệu**:

Mặc dù tập dữ liệu VEMO là một đóng góp quan trọng, luận văn không đề cập chi tiết về quy mô hoặc độ đa dạng của tập dữ liệu này, có thể ảnh hưởng đến khả năng khái quát hóa của mô hình.

1. **Chưa triển khai hệ thống hoàn chỉnh**:

Công trình tập trung vào nghiên cứu và thử nghiệm, chưa xây dựng một hệ thống hoàn chỉnh để tích hợp phương pháp vào thực tế hoặc triển khai thí điểm.