**ỦY BAN NHÂN DÂN THÀNH PHỐ HỒ CHÍ MINH**

**TRƯỜNG ĐẠI HỌC SÀI GÒN**

**KHOA CÔNG NGHỆ THÔNG TIN**

**Phân tích luận văn tốt nghiệp**

# NHẬN DIỆN CẢM XÚC MẶT NGƯỜI SỬ DỤNG MẠNG HỌC SÂU CÓ CHÚ Ý

Sinh viên thực hiện:

**Văn Hoàng Như Ý - 3122410493**

**Nguyễn Hoàng Thanh Phương - 3122410329**

**Lê Hồng Sơn - 3121410423**

**Đỗ Hữu Lộc - 3123410201**

Giảng viên: **Đỗ Như Tài**

Thành phố Hồ Chí Minh, ngày 15 tháng 4 năm 2025

# Mục lục

## 1. Thông tin chung của bài báo

| Thông tin | Chi tiết |
|---|---|
| Tác giả | Phạm Quí Luận |
| Giảng viên hướng dẫn | TS. Trần Tuấn Anh |
| Phản biện | TS. Nguyễn Hồ Mẫn Rạng |
| Trường | Trường Đại học Bách Khoa – ĐHQG TP.HCM |
| Thời gian thực hiện | Năm 2019 (cụ thể tháng 12/2019) |

## 2. Mục tiêu nghiên cứu

### a. Bối cảnh xác định mục tiêu

Tác giả đã xây dựng mục tiêu nghiên cứu dựa trên các nhận định quan trọng:
Bài toán nhận diện cảm xúc có nhiều ứng dụng thực tế: tương tác người - máy, giám sát cảm xúc tài xế, học tập thông minh, thương mại, quản lý đám đông,…

Sự khác biệt giữa dữ liệu lab-controlled và dữ liệu in-the-wild dẫn đến hiệu quả mô hình khác nhau:

- Mô hình đạt >90% trên dữ liệu trong phòng thí nghiệm (như CK+, MMI)
- Nhưng <80% khi áp dụng vào dữ liệu ảnh thực tế phức tạp, như FER2013

Do đó, cần mô hình tốt hơn, linh hoạt hơn để áp dụng cho môi trường thực tế.

b. Phạm vi giới hạn nghiên cứu

Tác giả đã chủ động giới hạn phạm vi nghiên cứu rõ ràng để tránh lan man và nâng cao tính khả thi:

- Tập trung vào dữ liệu ảnh tĩnh, không xử lý video hay chuỗi thời gian
- Dữ liệu được xử lý là loại in-the-wild (ảnh trong môi trường thực tế, không qua chuẩn hóa khắt khe)
- Chấp nhận giả thuyết có 6 cảm xúc cơ bản (Ekman & Friesen) để xử lý bài toán như một bài toán phân lớp
- Không đi sâu vào EEG hay dữ liệu cảm xúc phi hình ảnh như âm thanh, nhịp tim…

c. Mục tiêu cụ thể

Từ phạm vi trên, tác giả đặt ra các mục tiêu rõ ràng như sau:

- Xây dựng mô hình học sâu có cơ chế chú ý
  - Cụ thể là mô hình Residual Masking Network (RMN) – cho phép mô hình "chú ý" vào vùng ảnh quan trọng trên khuôn mặt.
  - Kết hợp với nhiều kiến trúc CNN hiện đại (VGG, ResNet, Inception) để tăng độ chính xác.
- Tự thu thập và xây dựng tập dữ liệu người Việt Nam (VEMO)
  - Lấy cảm hứng từ nghiên cứu của Guoying Zhao – người đã xây dựng tập dữ liệu khuôn mặt người Trung Quốc.
  - Mục tiêu là để phát triển mô hình có khả năng áp dụng tại Việt Nam.
- Huấn luyện và đánh giá trên dữ liệu thực tế
  - Dùng tập FER2013 và VEMO
  - So sánh RMN với các mô hình chuẩn đã có
  - Dùng ma trận nhầm lẫn và độ chính xác để đánh giá mô hình
- Phản biện học thuật
  - So sánh kết quả tự huấn luyện lại các mô hình hiện đại với RMN
  - Đồng thời đối chiếu với các kết quả đã được công bố trong các bài báo khoa học

## 3. Cơ sở lý thuyết

Cơ sở lý thuyết của luận văn là sự kết hợp giữa lý thuyết tâm lý học về cảm xúc, kỹ thuật xử lý ảnh số, và kiến trúc mạng học sâu trong lĩnh vực thị giác máy tính (computer vision). Tác giả trình bày và sử dụng ba nhóm kiến thức nền quan trọng sau:

### a. Cơ sở lý thuyết về cảm xúc khuôn mặt người

Nhận diện cảm xúc khuôn mặt là một bài toán xuất phát từ tâm lý học và sinh học thần kinh. Nghiên cứu này dựa vào khái niệm về sáu cảm xúc cơ bản được đề xuất bởi Ekman & Friesen, gồm: giận dữ (anger), sợ hãi (fear), ghê tởm (disgust), buồn bã (sadness), vui vẻ (happiness), và ngạc nhiên (surprise).

Những cảm xúc này có thể được biểu hiện rõ ràng thông qua các thay đổi vi mô trên khuôn mặt như chuyển động cơ ở vùng lông mày, mắt, miệng... Như vậy, việc nhận diện cảm xúc thực chất là phân tích và phân loại biểu hiện khuôn mặt thành một trong sáu nhóm cảm xúc trên.

Trong bối cảnh nghiên cứu của tác giả, cảm xúc khuôn mặt được xử lý như một bài toán phân loại ảnh với sáu lớp tương ứng.

## b. Cơ sở lý thuyết về mạng học sâu (Deep Learning)

Trong lĩnh vực thị giác máy tính, mạng nơ-ron tích chập (Convolutional Neural Network – CNN) là một trong những mô hình mạnh mẽ nhất để xử lý ảnh.

Tác giả trình bày kỹ các thành phần cơ bản trong CNN:

- Lớp tích chập (Convolutional Layer): giúp trích xuất đặc trưng cục bộ trong ảnh như đường viền, góc cạnh, kết cấu...
- Lớp phi tuyến (Activation Layer): thường dùng hàm ReLU để tạo tính phi tuyến, giúp mô hình học được các mối quan hệ phức tạp.
- Lớp gộp (Pooling Layer): giảm chiều dữ liệu, tăng tính khái quát và giảm overfitting.
- Lớp kết nối đầy đủ (Fully Connected Layer): dùng cho phần phân loại đầu ra.

Tác giả cũng trình bày một số kiến trúc CNN phổ biến dùng để so sánh trong luận văn như VGGNet, ResNet, và Inception. Mỗi kiến trúc đều có ưu nhược điểm riêng, trong đó ResNet nổi bật nhờ khả năng huấn luyện mạng rất sâu thông qua kết nối residual (skip connection).

## c. Cơ sở lý thuyết về cơ chế chú ý (Attention)

Cơ chế chú ý là thành phần quan trọng được tích hợp vào mô hình đề xuất của tác giả: Residual Masking Network. Attention được lấy cảm hứng từ cơ chế tập trung của con người –

tức là không xử lý toàn bộ ảnh một cách đồng đều, mà chọn ra những vùng quan trọng nhất để phân tích sâu hơn.

Trong bối cảnh nhận diện cảm xúc khuôn mặt, những vùng mang nhiều thông tin cảm xúc (như mắt, lông mày, miệng) nên được mạng chú ý nhiều hơn so với các vùng như trán, má hoặc nền ảnh.

Tác giả sử dụng một kiến trúc gọi là Residual Masking Block, kết hợp giữa:

- Kết nối tắt (Residual Connection)
- Mặt nạ không gian (Masking Attention)
- Cơ chế học tự động vùng quan trọng trong ảnh

Mục tiêu của cơ chế này là tự động định vị và khuếch đại những vùng chứa đặc trưng cảm xúc, trong khi làm mờ đi những vùng ít liên quan. Điều này giúp mô hình hoạt động hiệu quả hơn trong môi trường ảnh thực tế, nhiều nhiễu.

# 3. Phương pháp thu thập và xử lý thông tin

## a. Tập dữ liệu và phương pháp thu thập

Trong nghiên cứu, tác giả sử dụng hai bộ dữ liệu chính:

- Bộ dữ liệu FER2013 (Facial Expression Recognition 2013)
- Bộ dữ liệu VEMO (Vietnamese Emotion)

## b. Xử lý dữ liệu đầu vào

### 1. Tiền xử lý ảnh

- Tác giả tiến hành chuẩn hóa kích thước ảnh về cùng độ phân giải 48x48 pixel để phù hợp với kiến trúc mô hình.
- Dữ liệu được chuyển sang ảnh xám (grayscale) nếu cần, nhằm giảm số chiều và tăng tốc quá trình học.
- Một số kỹ thuật tăng cường dữ liệu (data augmentation) được áp dụng như:
    - Lật ảnh ngang (horizontal flip)
    - Dịch ảnh nhẹ
    - Xoay góc nhỏ
- Mục tiêu là tăng khả năng tổng quát hóa của mô hình và tránh hiện tượng overfitting.

### 2. Trích xuất vùng khuôn mặt

- Trong một số trường hợp, tác giả sử dụng kỹ thuật phát hiện khuôn mặt (face detection) để tách riêng vùng khuôn mặt khỏi nền ảnh.
- Việc trích xuất này giúp mô hình chỉ học những đặc trưng liên quan đến cảm xúc khuôn mặt, không bị nhiễu bởi hậu cảnh hoặc các chi tiết không liên quan.

## c. Kiến trúc mô hình huấn luyện

### 1. Mô hình Residual Masking Network

- Mô hình được đề xuất có cấu trúc gồm:
    - Các lớp CNN để trích xuất đặc trưng
    - Các khối Residual Masking Block để học cơ chế chú ý
    - Lớp fully connected để phân loại cảm xúc

- Mỗi khối masking có khả năng học "vùng mặt quan trọng" và làm nổi bật chúng trong toàn bộ biểu diễn đặc trưng.

### 2. So sánh với các mô hình hiện đại khác

- Tác giả huấn luyện lại các mô hình phổ biến như: VGGNet, ResNet, Inception và so sánh kết quả trên cùng dữ liệu.

- Các mô hình này được thiết lập với cùng cấu hình huấn luyện để đảm bảo tính công bằng khi so sánh.

3. **Cài đặt huấn luyện**

- Ngôn ngữ lập trình: Python
- Framework: TensorFlow hoặc Keras
- Bộ chia dữ liệu: training (80%), validation (10%), test (10%)
- Loss function: categorical crossentropy
- Optimizer: Adam
- Số epoch huấn luyện: được điều chỉnh linh hoạt dựa trên kết quả validation

4. **Đánh giá mô hình**

- Sử dụng độ chính xác (accuracy) và ma trận nhầm lẫn (confusion matrix) để đánh giá chất lượng mô hình.
  Ngoài ra, tác giả còn kiểm tra khả năng phân biệt giữa các cảm xúc dễ nhầm như: sợ hãi và ngạc nhiên, buồn và giận…

# 4. Kết Quả đạt được

## 1. Độ chính xác mô hình Residual Masking Network (RMN)

- Mô hình RMN do tác giả đề xuất đã được huấn luyện và đánh giá trên bộ dữ liệu FER2013.
- Độ chính xác trên tập kiểm tra (test set) đạt:
  76.82%, cao hơn so với các mô hình phổ biến khác như VGGNet, ResNet, Inception khi huấn luyện lại trong cùng điều kiện.
- Đây là một kết quả đáng chú ý vì:
  - FER2013 là bộ dữ liệu khó, chứa ảnh "in-the-wild" (ảnh thực tế, nhiều nhiễu, biểu cảm đa dạng, chất lượng ảnh không cao).
  - Nhiều mô hình hiện đại thường chỉ đạt 70–75% accuracy khi không sử dụng dữ liệu bổ sung hoặc huấn luyện đặc biệt.

## 2. Hiệu quả trên dữ liệu người Việt (VEMO)

- Mô hình được fine-tuned và kiểm thử trên bộ dữ liệu người Việt (VEMO) do chính tác giả thu thập.
- Kết quả cho thấy:
  - Khả năng tổng quát hóa tốt khi áp dụng mô hình từ dữ liệu quốc tế (FER2013) sang người Việt.

- Mô hình vẫn giữ được độ chính xác ổn định, đặc biệt trong việc phân biệt các cảm xúc phổ biến như "vui", "buồn", "ngạc nhiên".

## 3. So sánh với các mô hình khác

- Các mô hình được huấn luyện lại trong cùng điều kiện để đảm bảo tính công bằng:
  - VGGNet
  - ResNet
  - Inception
- RMN vượt trội hơn nhờ:
  - Tăng khả năng tập trung vào các vùng mặt mang thông tin cảm xúc
  - Cấu trúc residual giúp học sâu mà không gây gradient vanishing

HẠN CHẾ CỦA NGHIÊN CỨU

## 1. Giới hạn của dữ liệu

**a) Dữ liệu huấn luyện chưa đủ đa dạng**

- Bộ dữ liệu chính được sử dụng là FER2013, tuy phổ biến nhưng vẫn có nhiều điểm yếu:
  - Độ phân giải ảnh thấp (48x48), không đủ chi tiết để mô hình khai thác các tín hiệu
  - Dữ liệu được thu thập qua Google Images, nên có thể tồn tại sự không đồng nhất trong cách thể hiện cảm xúc giữa các nền văn hóa.

**b) Bộ dữ liệu người Việt (VEMO) còn nhỏ**

- Mặc dù việc xây dựng bộ dữ liệu VEMO là một đóng góp giá trị, nhưng:
  - Quy mô bộ dữ liệu chưa đủ lớn để mô hình học sâu có thể khai thác toàn diện.
    Dữ liệu chưa được gán nhãn bởi các chuyên gia hoặc hệ thống đánh giá nhất quán.
  - Còn thiếu đa dạng về độ tuổi, giới tính, bối cảnh và biểu hiện cảm xúc tự nhiên.

---

## 2. Giới hạn về phương pháp

**a) Chỉ xử lý ảnh tĩnh (static images)**

- Nghiên cứu chỉ tập trung vào ảnh đơn (frame-based), không xử lý chuỗi ảnh liên tục (video).
- Do đó, không tận dụng được thông tin theo thời gian (temporal cues) như tốc độ thay đổi biểu cảm, một yếu tố quan trọng trong nhận diện cảm xúc thực tế.

**b) Giả định cảm xúc là một trong 6 nhãn cơ bản (theo Ekman)**

- Việc sử dụng 6 nhãn cảm xúc cơ bản giúp đơn giản hóa bài toán phân loại, nhưng đồng thời:
    - Bỏ qua các cảm xúc phức hợp hoặc trung gian như lo lắng, xấu hổ, mỉa mai...
    - Không tính đến yếu tố văn hóa và bối cảnh, vốn ảnh hưởng mạnh đến cách thể hiện cảm xúc.

## 3. Giới hạn về tích hợp và ứng dụng

- Luận văn chưa triển khai hệ thống nhận diện cảm xúc theo thời gian thực (real-time).
- Mô hình chỉ mới đánh giá trên tập dữ liệu có sẵn, chưa kết nối với ứng dụng thực tế như:
    - Nhận diện cảm xúc trong lớp học online
    - Hỗ trợ tư vấn tâm lý
      Điều chỉnh phản hồi trong hệ thống chăm sóc khách hàng

## 4. Chưa khai thác dữ liệu cảm xúc đa mô thức

- Công trình chỉ khai thác biểu hiện khuôn mặt, trong khi cảm xúc có thể được phản ánh qua: Giọng nói
    - Dữ liệu sinh học (EEG, nhịp tim)
      Văn bản (text sentiment)
- Việc kết hợp các nguồn tín hiệu này (gọi là multi-modal emotion recognition) sẽ giúp nâng cao độ chính xác và khả năng tổng quát hóa, nhưng chưa được đề cập trong phạm vi nghiên cứu này.