**Bộ Giáo dục và Đào tạo**

**Trường Đại học Sài Gòn**

Khoa Công nghệ thông tin

Nhóm 3:

Thành viên:

1. Lê Khánh Hoàng
2. Nguyễn Tiến Trung
3. Phan Hoàng Vũ
4. Nguyễn Minh Thuận

BÀI TẬP 6

Tìm hiểu và tóm tắt lại  
các nội dung trong bài luận văn: Tên đồ án; Tác giả, Nơi công bố, Năm công bố; Mục tiêu nghiên  
cứu; Cơ sở lý thuyết; Phương pháp thu thập và xử lý thông tin; Kết quả đạt được;  
Hạn chế của công trình.

Tóm tắt

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

**Tác giả**: Phạm Quí Luận

**Giảng viên hướng dẫn**: TS. Trần Tuấn Anh

**Nơi công bố**: Trường Đại học Bách Khoa – Đại học Quốc gia TP.HCM

**Năm công bố**: 12/2019

**Mục tiêu nghiên cứu**:

**1. Phát triển mô hình nhận diện cảm xúc khuôn mặt từ ảnh tĩnh trong điều kiện thực tế phức tạp (in-the-wild).**

* **Mục tiêu cốt lõi:** Tạo ra một hệ thống có khả năng "đọc vị" cảm xúc trên khuôn mặt người chỉ từ một bức ảnh duy nhất (ảnh tĩnh).
* **Điểm khó khăn và thú vị:** Yếu tố "trong điều kiện thực tế phức tạp" (in-the-wild) là điểm mấu chốt. Nó chỉ ra rằng tác giả không chỉ muốn xây dựng một mô hình hoạt động tốt trong phòng thí nghiệm với những bức ảnh khuôn mặt hoàn hảo (ánh sáng tốt, góc chụp chính diện, không bị che khuất). Thay vào đó, mục tiêu là đối phó với những thách thức từ ảnh chụp thực tế, ví dụ như:
  + **Ánh sáng không đều:** Ảnh có thể quá sáng, quá tối, hoặc có bóng đổ phức tạp.
  + **Góc chụp đa dạng:** Khuôn mặt có thể không nhìn thẳng vào camera mà hơi nghiêng, cúi xuống, hoặc ngước lên.
  + **Biểu cảm tự nhiên và tinh tế:** Cảm xúc trong đời thực có thể không rõ ràng như trong các bộ dữ liệu được tạo dựng sẵn.
  + **Sự che chắn:** Khuôn mặt có thể bị che một phần bởi tóc, kính, khẩu trang, hoặc tay.
  + **Nền ảnh phức tạp:** Các yếu tố xung quanh khuôn mặt có thể gây nhiễu cho việc nhận diện.
  + **Sự khác biệt về chủng tộc và giới tính:** Mô hình cần hoạt động tốt trên nhiều đối tượng khác nhau.
* **Ý nghĩa:** Một mô hình như vậy sẽ có tính ứng dụng cao hơn trong thực tế, ví dụ như phân tích phản hồi của người dùng trong video, nhận diện cảm xúc trong các bức ảnh trên mạng xã hội, hoặc ứng dụng trong các hệ thống tương tác người-máy tự nhiên.

**2. Đề xuất và thử nghiệm mạng Residual Masking Network có cơ chế chú ý.**

* **Mục tiêu sáng tạo:** Tác giả không chỉ sử dụng một mô hình có sẵn mà còn muốn tự mình phát triển một kiến trúc mạng nơ-ron mới, được đặt tên là "Residual Masking Network."
* **Điểm đặc biệt của mô hình:**
  + **Mạng Residual (Residual Network):** Đây là một kiến trúc mạng sâu phổ biến, giúp giải quyết vấn đề "vanishing gradient" (mất mát thông tin khi mạng quá sâu), cho phép huấn luyện các mạng rất lớn và phức tạp, từ đó có khả năng học được các đặc trưng phức tạp của khuôn mặt.
  + **Cơ chế chú ý (Attention Mechanism):** Đây là một thành phần quan trọng giúp mô hình tập trung vào những vùng quan trọng nhất của khuôn mặt khi đưa ra dự đoán về cảm xúc. Ví dụ, khi nhận diện sự "buồn bã," mô hình có thể tập trung nhiều hơn vào vùng mắt và miệng đang trễ xuống. Cơ chế này giúp mô hình bỏ qua những vùng không liên quan hoặc bị nhiễu.
  + **"Masking Network" (Mạng che phủ):** Phần này có thể liên quan đến việc mô hình tự động xác định và tập trung vào các vùng khuôn mặt có chứa thông tin cảm xúc quan trọng, đồng thời giảm sự ảnh hưởng của các vùng bị che khuất hoặc ít liên quan.
* **Mục tiêu thử nghiệm:** Sau khi đề xuất kiến trúc mạng mới, tác giả sẽ tiến hành các thí nghiệm để đánh giá hiệu quả của nó trong việc nhận diện cảm xúc, đặc biệt là trong điều kiện "in-the-wild."

**3. So sánh hiệu quả của mô hình đề xuất với các mô hình hiện đại.**

* **Mục tiêu đánh giá khách quan:** Để chứng minh rằng mô hình "Residual Masking Network" mà họ đề xuất là tốt và có những ưu điểm nhất định, tác giả sẽ so sánh hiệu suất của nó với các mô hình nhận diện cảm xúc khuôn mặt tiên tiến nhất hiện có (state-of-the-art models).
* **Các tiêu chí so sánh:** Thông thường, việc so sánh sẽ dựa trên các chỉ số đánh giá hiệu suất như độ chính xác (accuracy), độ nhạy (sensitivity), độ đặc hiệu (specificity), và F1-score trên các bộ dữ liệu kiểm thử tiêu chuẩn.
* **Ý nghĩa:** Việc so sánh này sẽ giúp cộng đồng khoa học thấy được những đóng góp mới của mô hình được đề xuất so với những gì đã có, đồng thời chỉ ra những điểm mạnh và điểm yếu của nó.

**4. Xây dựng bộ dữ liệu biểu cảm người Việt Nam (VEMO) nhằm góp phần phát triển lĩnh vực này tại Việt Nam.**

* **Mục tiêu đóng góp cho cộng đồng:** Tác giả nhận thấy sự thiếu hụt các bộ dữ liệu biểu cảm khuôn mặt đặc trưng cho người Việt Nam. Hầu hết các bộ dữ liệu công khai hiện có được thu thập từ các nền văn hóa khác, và có thể không hoàn toàn phù hợp để huấn luyện các mô hình nhận diện cảm xúc cho người Việt do sự khác biệt về cách biểu lộ cảm xúc giữa các nền văn hóa.
* **Ý nghĩa của bộ dữ liệu VEMO:**
  + **Nguồn tài nguyên quý giá:** VEMO sẽ là một nguồn dữ liệu quan trọng cho các nhà nghiên cứu và phát triển trong lĩnh vực nhận diện cảm xúc tại Việt Nam.
  + **Nâng cao độ chính xác:** Việc huấn luyện mô hình trên dữ liệu biểu cảm của người Việt có thể giúp cải thiện đáng kể độ chính xác của các hệ thống nhận diện cảm xúc khi áp dụng cho người Việt.
  + **Thúc đẩy nghiên cứu:** VEMO sẽ tạo điều kiện cho các nghiên cứu sâu hơn về cách người Việt Nam biểu lộ cảm xúc và phát triển các ứng dụng phù hợp với văn hóa địa phương.
* **Tính tiên phong:** Việc xây dựng một bộ dữ liệu như VEMO cho thấy sự chủ động và mong muốn đóng góp vào sự phát triển của lĩnh vực trí tuệ nhân tạo nói chung và nhận diện cảm xúc nói riêng tại Việt Nam.

Tóm lại, nghiên cứu này có một lộ trình rõ ràng và đầy tham vọng: giải quyết một bài toán khó (nhận diện cảm xúc trong điều kiện thực tế), đề xuất một giải pháp mới (mạng Residual Masking với cơ chế chú ý), chứng minh hiệu quả của giải pháp đó thông qua so sánh, và tạo ra một nguồn tài nguyên có giá trị cho cộng đồng nghiên cứu tại Việt Nam (bộ dữ liệu VEMO). Đây là một hướng nghiên cứu có ý nghĩa khoa học và tiềm năng ứng dụng cao.

Cơ sở lý thuyết:

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

Phần này tập trung vào việc thiết lập **tầm quan trọng của khuôn mặt như một kênh biểu đạt cảm xúc** và giới thiệu các khái niệm cơ bản liên quan.

* **3.1.1 Biểu thị cảm xúc trên khuôn mặt người:**
  + **Luận điểm trung tâm:** Khuôn mặt là một công cụ phức tạp và mạnh mẽ để thể hiện cảm xúc, vượt xa các yếu tố tinh thần và tương tác xã hội. Tác giả nhấn mạnh rằng trong phạm vi nghiên cứu này, họ **chỉ tập trung vào khía cạnh biểu lộ cảm xúc** của khuôn mặt.
  + **Dẫn chứng và lập luận:**
    - Trích dẫn nghiên cứu của **Giáo sư Albert Mehrabian**: Tác giả đưa ra kết quả nghiên cứu nổi tiếng về ba yếu tố trong giao tiếp, trong đó **hành vi phi ngôn ngữ (bao gồm nét mặt) chiếm đến 55% hiệu quả giao tiếp**. Điều này củng cố tầm quan trọng của việc nghiên cứu biểu cảm khuôn mặt.
    - **Đặc điểm cấu trúc cơ mặt:** Tác giả giải thích sự phức tạp của hệ thống cơ mặt (hơn 40 cơ tự trị), cách chúng gắn kết với xương và mô, và khả năng hoạt động độc lập của từng cơ. Điều này làm nổi bật tiềm năng biểu đạt cảm xúc phong phú của khuôn mặt.
    - **Vai trò của hệ thần kinh:** Tác giả mô tả cách các dây thần kinh mặt kết nối các cơ mặt với não, bao gồm cả thân não (kiểm soát biểu cảm vô thức) và vỏ não (kiểm soát biểu cảm có ý thức). Sự phân biệt này giải thích tại sao biểu cảm thật và giả lại khác nhau.
    - **Liên kết giữa biểu cảm và xử lý cảm xúc:** Tác giả trình bày kết quả nghiên cứu MRI cho thấy **hạch hạnh nhân (amygdala)** trong thân não đóng vai trò quan trọng trong cả việc xử lý cảm xúc (sợ hãi, vui sướng) và điều khiển các phản ứng thể chất liên quan đến cảm xúc, bao gồm cả biểu cảm khuôn mặt.
    - **Giả thuyết phổ quát (Universality Hypothesis):** Tác giả giới thiệu giả thuyết cho rằng biểu cảm và nhận thức cảm xúc trên khuôn mặt là **giống nhau trên toàn cầu, bất kể văn hóa**. Họ trích dẫn công trình tiên phong của **Charles Darwin** và nghiên cứu thực nghiệm của **Paul Ekman** (bao gồm cả nghiên cứu với bộ tộc ở New Guinea) để ủng hộ quan điểm này ở một mức độ nhất định. Ekman kết luận rằng có một số **cảm xúc cơ bản** được biểu đạt bằng các kiểu nét mặt tương tự nhau trên toàn thế giới.
* **3.1.2 Hệ thống mã hóa cơ mặt (Facial Action Coding System - FACS):**
  + **Giới thiệu FACS:** Tác giả giới thiệu FACS do Ekman và Friesen phát triển như một hệ thống **mô tả chi tiết các hành vi trên khuôn mặt** thông qua tập hợp các **Đơn vị Hoạt động (Action Units - AUs)** được định nghĩa.
  + **Cơ chế hoạt động:** FACS hoạt động bằng cách phát hiện một hoặc nhiều AUs và sau đó đối chiếu với một bảng mục lục để xác định cảm xúc đang được biểu hiện. Tác giả đề cập đến 46 AUs khác nhau.
  + **Bảng tổ hợp AUs theo cảm xúc:** Tác giả cung cấp một bảng tóm tắt (Bảng 3.1) về cách các AUs cụ thể thường kết hợp với nhau để biểu thị các cảm xúc cơ bản như hạnh phúc, buồn bã, bất ngờ, sợ hãi, giận dữ và khinh bỉ. Điều này cho thấy một cách tiếp cận **dựa trên cơ học** để hiểu và phân tích biểu cảm khuôn mặt.
  + **Vi cảm xúc (Microexpressions):** Tác giả giới thiệu khái niệm vi cảm xúc – những biểu cảm khuôn mặt **diễn ra rất nhanh và không có chủ ý**, thường xuất hiện trong tình huống căng thẳng hoặc khi cố gắng che giấu cảm xúc thật. Tác giả nhấn mạnh tiềm năng của vi cảm xúc trong việc **phát hiện sự dối trá**.
  + **Tóm tắt phần:** Tác giả kết luận phần này bằng cách khẳng định rằng cảm xúc có nguồn gốc thần kinh và khuôn mặt là một bộ phận quan trọng để phân tích biểu cảm.

**3.2 Mạng nơ-ron tích chập trong bài toán nhận ảnh.**

Phần này chuyển sang giới thiệu **công cụ chính** mà tác giả sẽ sử dụng để xây dựng mô hình nhận diện cảm xúc: **mạng nơ-ron tích chập (Convolutional Neural Networks - CNNs/ConvNets)**.

* **3.2.1 Tổng quan về mạng nơ-ron tích chập:**
  + **So sánh với mạng nơ-ron thông thường:** Tác giả giải thích sự tương đồng cơ bản giữa CNNs và mạng nơ-ron thông thường (cấu tạo từ các nơ-ron có trọng số và độ chệch, có thể học được, biểu diễn một hàm khả vi, sử dụng hàm mất mát và các kỹ thuật học tương tự).
  + **Điểm khác biệt cốt lõi:** Sự khác biệt nằm ở **kiến trúc**, trong đó CNNs đưa ra **giả định rằng đầu vào là hình ảnh**. Điều này cho phép mã hóa các thuộc tính nhất định vào kiến trúc, làm cho mô hình hiệu quả hơn và giảm số lượng tham số.
  + **Cấu trúc lớp trong CNN:** Tác giả nhấn mạnh rằng các lớp trong CNN được sắp xếp theo **ba chiều: chiều rộng, chiều cao và chiều sâu (khối đặc trưng)**, khác với mạng nơ-ron thông thường chỉ có các lớp kết nối đầy đủ.
* **3.2.2 Các lớp cơ bản trong Mạng Nơ-ron Tích chập:**
  + **Ba loại lớp chính:** Tác giả giới thiệu ba loại lớp cơ bản để xây dựng CNNs: **Lớp Tích chập (Convolutional Layer - CONV), Lớp Gộp (Pooling Layer - POOL) và Lớp Kết nối Đầy đủ (Fully-Connected Layer - FC)**.
  + **Kiến trúc mẫu:** Tác giả đưa ra một kiến trúc mẫu đơn giản [INPUT - CONV - RELU - POOL - FC] và giải thích chức năng của từng lớp trong việc xử lý một ảnh đầu vào ví dụ (kích thước 32x32x3).
    - **INPUT:** Chứa giá trị pixel thô của ảnh.
    - **CONV:** Tính toán đầu ra của các nơ-ron kết nối với các vùng cục bộ trong đầu vào bằng phép tích vô hướng với bộ lọc (kernel).
    - **RELU (Rectified Linear Unit):** Áp dụng hàm kích hoạt phi tuyến tính theo từng phần tử.
    - **POOL:** Thực hiện thao tác giảm mẫu (downsampling) theo chiều không gian.
    - **FC:** Tính toán điểm số của các lớp phân loại, kết nối với tất cả các nơ-ron ở lớp trước.
  + **Quá trình biến đổi và học:** Tác giả giải thích cách CNNs biến đổi ảnh đầu vào qua các lớp để cuối cùng đưa ra dự đoán. Các lớp CONV và FC chứa các **tham số (trọng số và độ chệch)** được học thông qua **Gradient Descent** để giảm thiểu hàm mất mát. Các lớp RELU và POOL thực hiện các hàm cố định.
  + **Hình 3.2 (Tiny VGG Net):** Minh họa một kiến trúc CNN mẫu, cho thấy sự biến đổi của khối đặc trưng qua các lớp.
* **3.2.3 Lớp tích chập (Convolutional Layer - CONV):**
  + **Tầm quan trọng:** Tác giả nhấn mạnh vai trò trung tâm của lớp tích chập, đến mức CNN được đặt tên theo nó.
  + **Bộ lọc có thể học:** Lớp CONV được xem như một tập hợp các **bộ lọc (filter/kernel) có thể học được**, có kích thước nhỏ theo chiều không gian nhưng kéo dài theo toàn bộ chiều sâu của khối đầu vào.
  + **Tích chập và bản đồ đặc trưng (feature map):** Tác giả mô tả quá trình **trượt (tích chập)** từng bộ lọc trên khối đầu vào để tạo ra **bản đồ đặc trưng**, phản ánh phản ứng của bộ lọc tại mọi vị trí không gian.
  + **Kết nối cục bộ (Local Connectivity) và trường tiếp nhận (Receptive Field):** Tác giả giải thích rằng mỗi nơ-ron trong lớp CONV chỉ kết nối với một vùng cục bộ của lớp trước (trường tiếp nhận), giúp giảm số lượng tham số.
  + **Ví dụ minh họa:** Hai ví dụ được đưa ra để làm rõ cách tính toán số lượng kết nối dựa trên kích thước đầu vào và trường tiếp nhận.
  + **Các siêu tham số: depth, stride và zero-padding:** Tác giả giới thiệu ba siêu tham số quan trọng điều khiển kích thước khối đầu ra và cách kết nối:
    - **Depth (Số bộ lọc K):** Xác định số lượng bản đồ đặc trưng đầu ra.
    - **Stride (S):** Xác định bước nhảy của bộ lọc khi trượt trên đầu vào.
    - **Zero-padding (P):** Thêm các phần tử 0 xung quanh biên của đầu vào để kiểm soát kích thước đầu ra.
  + **Công thức tính kích thước đầu ra:** Tác giả cung cấp công thức toán học để tính kích thước không gian của khối đầu ra dựa trên kích thước đầu vào (W1), kích thước bộ lọc (F), stride (S) và padding (P).
  + **Hình 3.4 (Mô phỏng sắp xếp không gian):** Minh họa ảnh hưởng của stride đến kích thước đầu ra trong một chiều không gian.
  + **Sử dụng zero-padding:** Tác giả giải thích cách zero-padding có thể được sử dụng để duy trì kích thước không gian của đầu vào và đầu ra khi stride là 1.
  + **Ràng buộc của stride:** Tác giả lưu ý rằng các siêu tham số cần được chọn sao cho bộ lọc có thể "trượt" vừa vặn trên đầu vào.
  + **Ví dụ thực tế (Krizhevsky et al., 2012 - AlexNet):** Tác giả đưa ra ví dụ về lớp tích chập đầu tiên trong kiến trúc AlexNet để minh họa việc sử dụng các siêu tham số và tính toán kích thước đầu ra.
  + **Chia sẻ trọng số (Weight Sharing):** Đây là một đặc điểm quan trọng giúp giảm đáng kể số lượng tham số trong lớp CONV. Các nơ-ron trong cùng một **lát cắt độ sâu (depth slice)** sử dụng cùng một bộ trọng số và độ chệch.
  + **Hình 3.5 (Bộ lọc học được trong AlexNet):** Minh họa các bộ lọc được học ở lớp đầu tiên, cho thấy chúng phát hiện các đặc trưng cơ bản như cạnh và màu sắc.
  + **Tóm tắt lớp tích chập:** Tác giả tổng kết các đặc điểm chính của lớp tích chập, bao gồm đầu vào, các siêu tham số, đầu ra, số lượng tham số và cách tính toán đầu ra.
* **3.2.4 Lớp Gộp (Pooling Layer - POOL):**
  + **Chức năng:** Giảm kích thước không gian của khối đặc trưng, giảm số lượng tham số và tính toán, và kiểm soát overfitting.
  + **Hoạt động độc lập theo chiều sâu:** Lớp POOL hoạt động trên từng lát cắt độ sâu của khối đặc trưng một cách độc lập.
  + **Hàm gộp:** Sử dụng một hàm (thường là max) để tính toán giá trị đầu ra cho mỗi vùng không gian.
  + **Max-Pooling phổ biến:** Tác giả đề cập đến hình thức phổ biến nhất là **Max-Pooling** với bộ lọc 2x2 và stride 2, giúp giảm 75% thông tin không gian.
  + **Đặc tả lớp gộp:** Tác giả tóm tắt các đặc điểm của lớp gộp, bao gồm đầu vào, các siêu tham số (kích thước gộp F và stride S), đầu ra và việc không có tham số học.
  + **Hình 3.6 (Lớp gộp giảm kích thước):** Minh họa cách lớp gộp giảm kích thước không gian của khối đầu vào và hoạt động của Max-Pooling.

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

Phần này giới thiệu hai khái niệm kiến trúc quan trọng mà tác giả sẽ kết hợp trong mô hình đề xuất của mình: **kiến trúc mã hóa - giải mã** và **cơ chế chú ý**.

* **3.3.1 Kiến trúc mã hóa - giải mã (Encoder-Decoder Architecture):**
  + **Khái niệm cơ bản:** Kiến trúc này bao gồm hai phần riêng biệt: **phần mã hóa (encoder)** để trích xuất một biểu diễn tiềm ẩn (latent representation) từ ảnh đầu vào, và **phần giải mã (decoder)** để tái tạo hoặc tạo ra đầu ra mong muốn từ biểu diễn tiềm ẩn này.
  + **Biểu diễn toán học:** Tác giả mô tả quá trình này bằng các hàm toán học E (encoder) và D (decoder), với đầu ra cuối cùng là O = D(E(I)).
  + **Mạng U-net:** Tác giả đề cập đến **mạng U-net** như một kiến trúc mã hóa - giải mã nổi tiếng, ban đầu được phát triển cho phân đoạn ảnh y tế (Hình 3.7). Tác giả ngụ ý rằng kiến trúc của họ có thể lấy cảm hứng từ U-net.
  + **Cấu trúc U-net:** Tác giả mô tả U-net có phần mã hóa (giảm kích thước đặc trưng) và phần giải mã (tăng kích thước đặc trưng), kết hợp với các lớp convolutional liên tiếp và một lớp convolutional 1x1 ở cuối để thực hiện phân loại (trong trường hợp phân đoạn).
* **3.3.2 Cơ chế chú ý (Attention Mechanism):**
  + **Bối cảnh phát triển:** Tác giả đặt cơ chế chú ý trong bối cảnh sự phát triển của mạng học sâu với CNNs, từ LeNet đến ResNet và GoogLeNet, nhấn mạnh xu hướng xây dựng các mạng sâu và rộng hơn. Họ chỉ ra rằng luận văn của họ tập trung vào khía cạnh **tính chú ý** của mô hình.
  + **Nguồn gốc:** Cơ chế chú ý được giới thiệu lần đầu trong bài báo về **dịch máy nơ-ron (Neural Machine Translation)** của Bahdanau et al., để cải thiện hiệu suất dịch cho các câu dài bằng cách cho phép mô hình **tập trung vào các phần liên quan của câu nguồn** khi dịch từng từ.
  + **Ứng dụng rộng rãi:** Sự thành công trong dịch máy đã dẫn đến việc nghiên cứu và ứng dụng cơ chế chú ý trong nhiều bài toán khác như đặt tiêu đề ảnh, phân loại và phân đoạn ảnh.
  + **Ví dụ đặt tiêu đề ảnh (Hình 3.9):** Tác giả minh họa cách cơ chế chú ý giúp mô hình tập trung vào các vùng quan trọng của ảnh (ví dụ: cô bé, chiếc đĩa) khi tạo ra chú thích.
  + **Ứng dụng trong nhận diện cảm xúc khuôn mặt:** Tác giả liên hệ cơ chế chú ý với bài toán của mình bằng cách chỉ ra rằng **các vùng khác nhau trên khuôn mặt đóng góp khác nhau vào việc dự đoán cảm xúc** (ví dụ: mắt và miệng quan trọng hơn cằm hoặc tóc). Họ đề cập đến các nghiên cứu sử dụng **điểm mốc khuôn mặt** để trích xuất các vùng quan tâm và đưa vào mạng sâu.
  + **Nghiên cứu về mạng học sâu với cơ chế chú ý cho phân loại ảnh (Wang et al., 2017 - Hình 3.11):** Tác giả giới thiệu một nghiên cứu sử dụng **các mô-đun chú ý xếp chồng lên nhau**. Họ mô tả cách khối đặc trưng được tách thành hai nhánh và kết hợp lại bằng **Công thức 3.1 (học chú ý thặng dư)** để tránh việc chú ý sai làm mất thông tin quan trọng.
* **Phương pháp nhận diện khuôn mặt:**
  + **Nền tảng:** Tác giả đề cập đến bài toán phát hiện khuôn mặt như một bước tiền xử lý quan trọng.
  + **Sử dụng OpenCV:** Tác giả tuyên bố sẽ **tham khảo và sử dụng lại chức năng phát hiện khuôn mặt bằng mạng Single Shot Detector (SSD) được tích hợp trong OpenCV 3.3 (mô-đun dnn)** và do đó sẽ không trình bày chi tiết về phần này.

**Tóm lại, cơ sở lý thuyết mà tác giả trình bày đã xây dựng một nền tảng vững chắc cho nghiên cứu của họ:**

* **Chứng minh tầm quan trọng của khuôn mặt và biểu cảm nét mặt trong việc thể hiện cảm xúc.**
* **Giới thiệu hệ thống FACS như một phương pháp khoa học để mô tả và phân tích biểu cảm khuôn mặt.**
* **Cung cấp một cái nhìn tổng quan về mạng nơ-ron tích chập (CNNs), kiến trúc mạng chính được sử dụng trong bài toán nhận ảnh, và giải thích chi tiết các thành phần cơ bản của CNNs.**
* **Giới thiệu kiến trúc mã hóa - giải mã và cơ chế chú ý như hai thành phần kiến trúc nâng cao có tiềm năng cải thiện hiệu suất của mô hình nhận diện cảm xúc, đặc biệt là trong điều kiện thực tế phức tạp.**
* **Nêu rõ phương pháp tiếp cận để phát hiện khuôn mặt, một bước quan trọng trong quy trình nhận diện cảm xúc từ ảnh.**

Phương pháp thu thập và xử lý thông tin:

**1. Thu thập Dữ liệu (Chương 4):**

* **Đa dạng nguồn dữ liệu (Mục 4.1):** Tác giả đã sử dụng nhiều nguồn khác nhau để thu thập dữ liệu, bao gồm cả các bộ dữ liệu công khai (CK+, JAFFE, MMI, FER-2013, AFEW 7.0, SFEW 2.0, Oulu-Casia, Emotio-Net, AffectNet) và bộ dữ liệu tự xây dựng (VEMO). Điều này cho thấy sự nỗ lực trong việc tận dụng nguồn dữ liệu sẵn có và đồng thời giải quyết vấn đề thiếu hụt dữ liệu đặc trưng cho người Việt Nam.
* **Tập dữ liệu FER2013 (Mục 4.2):**
  + Đây là bộ dữ liệu chính được sử dụng để huấn luyện, đánh giá và so sánh các mô hình.
  + Dữ liệu được thu thập từ một cuộc thi trên Kaggle, bao gồm ảnh xám kích thước 48x48 với 7 lớp cảm xúc.
  + Tác giả đã nhận diện những thách thức của bộ dữ liệu này, bao gồm sự mất cân bằng dữ liệu và các mẫu bị lỗi.
* **Tập dữ liệu VEMO (Mục 4.3):**
  + Đây là một đóng góp quan trọng của nghiên cứu, một bộ dữ liệu biểu cảm khuôn mặt người Việt Nam.
  + **Phương pháp thu thập:**
    - **Ảnh từ web:** Thu thập từ Google Image, Flickr với các từ khóa liên quan đến người Việt Nam trong nhiều ngữ cảnh khác nhau (sinh viên, trẻ em, đám cưới, show truyền hình,...). Điều này cố gắng bao phủ sự đa dạng trong điều kiện thực tế.
    - **Video từ phim và Youtube:** Trích xuất hình ảnh khuôn mặt từ 92 đoạn phim.
    - **Quy trình trích xuất ảnh từ video:** Sử dụng mô-đun phát hiện khuôn mặt V&J, trích xuất 1 frame và bỏ qua 5 frame tiếp theo để giảm trùng lặp.
  + **Gắn nhãn dữ liệu:** Được thực hiện thủ công bởi 3 người Việt Nam trẻ tuổi (18-23) để đảm bảo tính phù hợp văn hóa.
  + **Dữ liệu bổ sung:** Sử dụng thêm 30 nghìn ảnh người trên toàn thế giới được đánh nhãn bởi chuyên gia từ bộ dữ liệu AffectNet.
  + **Phân chia dữ liệu:** Tổng cộng 36470 ảnh được chia thành tập huấn luyện (70%), kiểm thử (15%), và kiểm tra (15%).
  + Tác giả cũng chỉ ra sự phân phối dữ liệu của bộ VEMO qua biểu đồ.

**2. Xử lý Dữ liệu (Chương 6):**

* **Phát hiện khuôn mặt (Mục 6.2):**
  + Đối với các ảnh chưa có sẵn vùng khuôn mặt, tác giả sử dụng phương pháp phát hiện khuôn mặt được cung cấp bởi OpenCV 3.4. Đây là bước tiền xử lý quan trọng để đảm bảo mô hình tập trung vào vùng thông tin cần thiết.
* **Các phép biến đổi làm giàu dữ liệu (Data Augmentation) (Mục 6.2):**
  + Để tăng tính đa dạng của dữ liệu huấn luyện và cải thiện khả năng tổng quát hóa của mô hình, tác giả sử dụng hai phép biến đổi cơ bản:
    - Lật ảnh theo trục dọc (Flip Left Right).
    - Xoay ảnh ngẫu nhiên trong khoảng -30 đến 30 độ.
  + Các phép biến đổi này được thực hiện bằng thư viện imgaug. Việc sử dụng các phép biến đổi đơn giản giúp đảm bảo tính công bằng khi so sánh với các mô hình khác.
* **Lớp định nghĩa dữ liệu (Dataset Class) (Mục 6.2):**
  + Tác giả xây dựng một lớp Dataset tùy chỉnh kế thừa từ torch.utils.data.Dataset của PyTorch.
  + **Quy trình lấy dữ liệu (\_\_getitem\_\_)**:
    - Đọc ảnh và nhãn từ danh sách dữ liệu.
    - Thay đổi kích thước ảnh về 224x224 (kích thước đầu vào thống nhất cho các mô hình).
    - Chuyển ảnh xám thành ảnh 3 kênh bằng cách lặp lại kênh xám 3 lần.
    - Áp dụng các phép làm giàu dữ liệu (nếu đang trong giai đoạn huấn luyện).
    - Chuyển đổi ảnh sang định dạng Tensor để có thể sử dụng trong PyTorch.

**3. Cài đặt Huấn luyện (Mục 6.3):**

* Tác giả đã thiết lập các cấu hình huấn luyện cơ bản một cách nhất quán cho tất cả các kiến trúc mạng để đảm bảo sự công bằng trong so sánh. Các cấu hình này bao gồm: kích thước ảnh đầu vào, số kênh, số lớp đầu ra, tốc độ học, momentum, weight decay, batch size, số lượng workers cho dataloader, số lượng epoch tối đa, và các cơ chế dừng huấn luyện sớm (early stopping) dựa trên validation accuracy. Việc này đảm bảo một quy trình huấn luyện có kiểm soát và có thể tái lặp.
* Việc sử dụng seed ngẫu nhiên (1234) cho tất cả các mô-đun ngẫu nhiên giúp đảm bảo tính tái sản xuất của các thí nghiệm.
* Mô hình tốt nhất được chọn dựa trên độ chính xác trên tập kiểm định.

**Tóm lại, phương pháp thu thập và xử lý thông tin của luận văn này bao gồm:**

* **Thu thập dữ liệu đa dạng:** Sử dụng cả bộ dữ liệu công khai và tự xây dựng (VEMO) để giải quyết các mục tiêu nghiên cứu. Quá trình thu thập VEMO được mô tả chi tiết, bao gồm nguồn, phương pháp trích xuất và quy trình gắn nhãn thủ công.
* **Tiền xử lý cơ bản:** Phát hiện khuôn mặt bằng OpenCV để tập trung vào vùng quan trọng.
* **Làm giàu dữ liệu có kiểm soát:** Sử dụng các phép biến đổi đơn giản để tăng tính đa dạng của dữ liệu huấn luyện.
* **Xây dựng lớp Dataset hiệu quả:** Quản lý việc tải và xử lý dữ liệu trong PyTorch.
* **Thiết lập cấu hình huấn luyện nhất quán:** Đảm bảo tính công bằng và khả năng tái lặp trong quá trình thực nghiệm.

Kết quả đạt được:

Tuyệt vời! Dựa trên phần "Đánh giá kết quả đạt được" (7.1), đây là tóm tắt ngắn gọn về kết quả và đóng góp chính của công trình này:

**Kết quả đạt được và đóng góp chính:**

* **Bộ dữ liệu VEMO:** Xây dựng thành công bộ dữ liệu đầu tiên về biểu cảm khuôn mặt người Việt Nam, được thu thập và gắn nhãn bởi người Việt, có độ khó cao và sát với thực tế. Bộ dữ liệu này sẽ được công khai, góp phần phát triển lĩnh vực nhận diện cảm xúc tại Việt Nam.
* **Mạng Residual Masking Network:** Đề xuất một kiến trúc mạng học sâu mới tích hợp cơ chế chú ý vào quá trình trích xuất đặc trưng. Mạng này được huấn luyện end-to-end và cho thấy kết quả nổi bật trong bài toán phân lớp biểu cảm.
* **Phương pháp thực nghiệm công bằng và tái sinh:** Các thí nghiệm được thực hiện trên cùng một nền tảng, khung thức và cấu hình, đảm bảo tính công bằng và khả năng tái tạo. Kết quả được báo cáo chi tiết, hữu ích cho các nghiên cứu sau này.
* **Hiệu suất mô hình:** Mạng Residual Masking Network đạt độ chính xác cao trên cả bộ dữ liệu VEMO và FER2013. Đặc biệt trên FER2013, kết quả nằm trong nhóm dẫn đầu so với các công bố khoa học và vượt trội hơn một số mạng hiện đại, thậm chí cạnh tranh với các phương pháp ensemble learning.

Hạn chế của công trình:

**Chưa thử nghiệm trên tập dữ liệu lớn:** Hiệu năng của kiến trúc Residual Masking Network có thể chưa được kiểm chứng đầy đủ trên các bộ dữ liệu quy mô lớn hơn.

**Cơ chế chú ý chưa được khai thác sâu:** Khối chú ý có thể chưa được thử nghiệm và tối ưu hóa trên nhiều kiến trúc và bài toán khác nhau.

**Siêu tham số chưa được tối ưu toàn diện:** Các siêu tham số của riêng mạng Residual Masking Network có thể còn tiềm năng để được điều chỉnh và cải thiện hơn nữa.

**Chưa xây dựng hệ thống ứng dụng hoàn chỉnh:** Công trình tập trung vào nghiên cứu mô hình và đánh giá trên bộ dữ liệu, chưa triển khai thành một hệ thống hoàn chỉnh và thử nghiệm ứng dụng thực tế.