* **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.
* **Mục tiêu nghiên cứu**

1. Cải thiện độ chính xác của mô hình nhận diện cảm xúc trên ảnh thực tế phức tạp.
2. 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.
3. 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).
4. 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.

* **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**

* **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.
* **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.
* **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

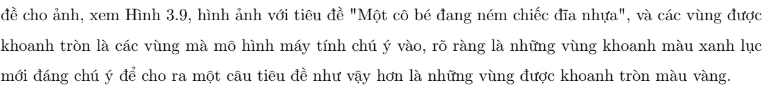
Nội dung do AI tạo ra có thể không chính xác.**

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

* **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.
* **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.