


THÔNG TIN CHUNG CỦA NHÓM

- Link YouTube video của báo cáo (tối đa 5 phút):
<https://youtu.be/1McomnfEhDU>
- Link slides (dạng .pdf đặt trên Github của nhóm):
<https://github.com/IIIDrAgOoN/CS519.O11/blob/main/slide.pdf>

<ul style="list-style-type: none">● Họ và Tên: Đoàn Nguyễn Trần Hoàn● MSSV: 21520239 	<ul style="list-style-type: none">● Lớp: CS519.O11● Tự đánh giá (điểm tổng kết môn): 8.5/10● Số buổi vắng: 0● Số câu hỏi QT cá nhân: 2● Số câu hỏi QT của cả nhóm: 15● Link Github: https://github.com/IIIDrAgOoN/CS519.O11● Mô tả công việc và đóng góp của cá nhân cho kết quả của nhóm:<ul style="list-style-type: none">○ Lên ý tưởng○ Viết phần report○ Làm Poster○ Làm video YouTube
<ul style="list-style-type: none">● Họ và Tên: Nguyễn Quốc Trường● MSSV: 21521604	<ul style="list-style-type: none">● Lớp: CS519.O11● Tự đánh giá (điểm tổng kết môn): 8.5/10● Số buổi vắng: 1● Số câu hỏi QT cá nhân: 11● Link Github: https://github.com/IIIDrAgOoN/CS519.O11● Mô tả công việc và đóng góp của cá nhân cho kết quả của nhóm:<ul style="list-style-type: none">○ Lên ý tưởng



- Viết phần Proposal
- Làm Slide
- Làm video Youtube

ĐỀ CƯƠNG NGHIÊN CỨU

TÊN ĐỀ TÀI (IN HOA)

VISION TRANSFORMER KẾT HỢP GIẢM NHIỄU CHO NHẬN DIỆN CẢM XÚC KHUÔN MẶT

TÊN ĐỀ TÀI TIẾNG ANH (IN HOA)

VISION TRANSFORMER WITH NOISE REDUCTION FOR FACIAL EXPRESSION RECOGNITION

TÓM TẮT (Tối đa 400 từ)

Nhận diện cảm xúc khuôn mặt là một lĩnh vực nghiên cứu quan trọng trong thị giác máy tính, có ứng dụng trong nhiều lĩnh vực. Tuy nhiên, nhận diện cảm xúc khuôn mặt vẫn còn là một bài toán khó, do sự đa dạng biểu cảm của con người. Vision Transformer (ViT) là một mô hình học sâu có khả năng học được các đặc trưng toàn cục giữa các patch và mối quan hệ không gian của chúng, điều này giúp ViT có khả năng nhận diện cảm xúc khuôn mặt chính xác hơn. Tuy nhiên, ViT cũng có một số nhược điểm, như khả năng học tập chậm và dễ bị quá khớp. Đồng thời, khi gặp phải các hình ảnh có các yếu tố gây nhiễu, hiệu suất của các mô hình nhận diện cũng sẽ bị ảnh hưởng. Để khắc phục những nhược điểm này, các nhà nghiên cứu đã đề xuất kết hợp ViT với các kỹ thuật giảm nhiễu. Kỹ thuật giảm nhiễu có thể giúp ViT học đặc trưng của ảnh hiệu quả hơn. Một số kỹ thuật giảm nhiễu thường được sử dụng trong nhận diện cảm xúc khuôn mặt bao gồm: Kỹ thuật giảm nhiễu dựa trên lọc và kỹ thuật giảm nhiễu dựa trên mô hình. Trong các kỹ thuật giảm nhiễu hiện có, các nhà khoa học đã đề xuất kỹ thuật giảm nhiễu dựa trên Attentive pooling. Dựa trên đó, chúng tôi xây dựng một mô hình kết hợp giữa ViT và Attentive Pooling, kỹ thuật này sử dụng một mô hình học máy để học cách tập trung vào các vùng quan trọng trong hình ảnh, giúp giảm nhiễu hiệu quả hơn.

GIỚI THIỆU (Tối đa 1 trang A4)

Nhận diện cảm xúc khuôn mặt là một lĩnh vực nghiên cứu trong trí tuệ nhân tạo liên

quan đến việc xác định cảm xúc của một người từ biểu hiện khuôn mặt của họ. Nhận diện cảm xúc khuôn mặt có ứng dụng trong nhiều lĩnh vực khác nhau, chẳng hạn như giao tiếp tự động, chăm sóc sức khỏe, và an ninh.

Có rất nhiều mô hình và phương pháp được sử dụng để giải quyết bài toán này, từ cổ điển cho đến hiện đại. Trong đó các mô hình ViT cũng đã được huấn luyện để giải quyết bài toán nhận diện cảm xúc khuôn mặt, một số tiêu biểu như là MViT, TransFER,... Tuy nhiên, hiệu quả của các mô hình này cũng không quá vượt trội so với các mô hình, phương pháp trước đó.

Để cải thiện hiệu quả của mô hình ViT trên bài toán nhận diện cảm xúc, ta có thể thực hiện nhiều phương pháp khác nhau. Ở đây, chúng tôi sử dụng phương pháp kết hợp giảm nhiễu, giúp mô hình tập trung vào phân tích các vùng ảnh quan trọng, liên quan nhiều tới cảm xúc con người.

Trong đề tài này, chúng tôi nghiên cứu cải thiện hiệu quả bằng cách kết hợp mô hình ViT hiện có với phương pháp giảm nhiễu Attentive Pooling trong bài toán nhận diện cảm xúc khuôn mặt có trong ảnh.

Input: Một ảnh của khuôn mặt người (có đầy đủ các chi tiết trên khuôn mặt) cần được nhận diện cảm xúc.

Output: Một chuỗi ký tự mô tả cảm xúc nhận diện được từ ảnh khuôn mặt người đã truyền vào.



MỤC TIÊU

(Viết trong vòng 3 mục tiêu, lưu ý về tính khả thi và có thể đánh giá được)

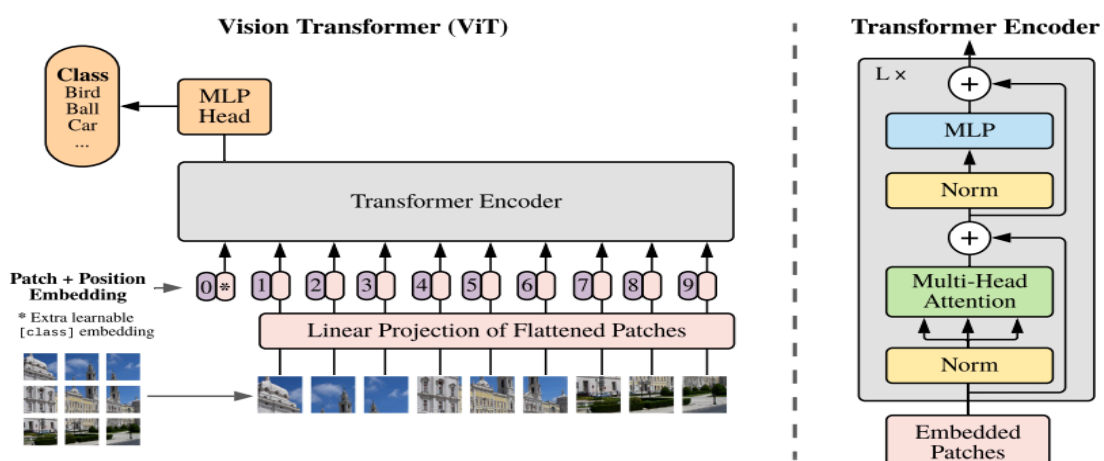
- Nghiên cứu mô hình Vision Transformer hiện có và cải thiện hiệu suất và chi phí tính toán của nó trong bài toán nhận diện cảm xúc khuôn mặt.
- Nghiên cứu phương pháp tạo ra attention map để giảm nhiễu.
- Kết hợp 2 nghiên cứu ở trên trong một mô hình và đánh giá mô hình trên các bộ dữ liệu của bài toán nhận diện cảm xúc khuôn mặt người như FER+, AffectNet, RAF-DB.

NỘI DUNG VÀ PHƯƠNG PHÁP

(Viết nội dung và phương pháp thực hiện để đạt được các mục tiêu đã nêu)

Nội dung:

- Tìm hiểu về cấu trúc mô hình ViT.



- Nghiên cứu kết hợp mô hình ViT với phương pháp giảm nhiễu Attentive Pooling.
- Chọn lọc và điều chỉnh các bộ dữ liệu FER+, RAF-DB, AffectNet cho phù hợp, để mô hình có thể học tập từ dữ liệu và có thể hoạt động được trên các ảnh ngoài bộ dữ liệu mà ta tải lên này.
- Huấn luyện mô hình ViT kết hợp Attentive Pooling (APViT) sử dụng bộ dữ liệu sau khi xử lý, sau đó đánh giá và so sánh với các phương pháp trước đó.
- Xây dựng chương trình ứng dụng minh họa.

Phương pháp:

- **Vision Transformer (ViT):**

Tìm hiểu cấu trúc mô hình ViT. Cài đặt và đánh giá thử một số mô hình ViT truyền thống trong bài toán nhận diện cảm xúc khuôn mặt, chẳng hạn như mô hình TransFER nhằm đánh giá và so sánh với phương pháp cải tiến.

Nghiên cứu và sử dụng mạng CNN (ResNet50) nhằm trích xuất đặc trưng có trong ảnh khuôn mặt người và tạo thành feature map.

Sau đó tạo ra attention map bằng cách sử dụng liên tiếp 2 lớp Conv có kích thước cửa sổ trượt là 1×1 để giảm chiều sâu feature map. Từ attention map, nghiên cứu kết hợp với các phương pháp giảm nhiễu để loại bỏ các yếu tố nhiễu ảnh hưởng tới hiệu quả hoạt động của mô hình ViT.

Nghiên cứu và cài đặt hai phương pháp giảm nhiễu là Attentive Patch Pooling và Attentive Token Pooling để kết hợp với mô hình ViT.

- **Attentive Patch Pooling (APP):**

Từ Attention map được trích xuất từ ảnh thông qua mạng ResNet50, ảnh được chia làm các vùng (Patch) với mỗi vùng có một mức độ quan trọng và cần thiết khác nhau liên quan tới bài toán. Nghiên cứu và sử dụng phương pháp APP nhằm tập trung và đánh dấu vào các vùng ảnh chứa thông tin quan trọng và cần thiết, chọn ra top k patch quan trọng nhất của ảnh. Sau đó loại bỏ các patch có chứa thông tin không cần thiết hoặc ảnh hưởng không lớn tới việc xử lý bài toán.

- **Attentive Token Pooling (ATP):**

Nghiên cứu và cài đặt ATP block (một Transformer Encoder được cải tiến bằng ATP) nhằm xác định mối quan hệ giữa các patch được chọn ra từ APP trong phạm vi toàn cục. Từ các patch thu được, sử dụng chúng để tạo nên các token embedding, là một vector dùng để biểu diễn token tương ứng, thể hiện đặc trưng của 1 vùng trong ảnh. Sau đó sử dụng attention module để tính toán trọng số cho từng token. Dựa trên trọng số của các token đã tính ra, thực hiện chọn top k token quan trọng nhất, liên quan tới bài toán nhận diện cảm xúc đang xử lý để giảm thiểu ảnh hưởng do nhiễu và tiết kiệm thời gian tính toán.

- Huấn luyện mô hình ViT kết hợp với 2 module giảm nhiễu APP và ATP trên các bộ dữ liệu có sẵn FER+, AffectNet và RAF-DB. So sánh và đánh giá kết quả dựa trên độ đo accuracy với các phương pháp trước đây.
- Xây dựng chương trình ứng dụng cho phép người dùng nhập một ảnh, và trả về kết quả là cảm xúc của từng khuôn mặt của người có trong ảnh.

KẾT QUẢ MONG ĐỢI

(Viết kết quả phù hợp với mục tiêu đặt ra, trên cơ sở nội dung nghiên cứu ở trên)

- Báo cáo các phương pháp và kỹ thuật của phương pháp Vision Transformer kết hợp với giảm nhiễu (Attentive Pooling) được sử dụng trong bài toán nhận diện cảm xúc khuôn mặt. Kết quả thực nghiệm, đánh giá và so sánh phương pháp này với các phương pháp trước đó.
- Tăng hiệu suất và giảm chi phí tính toán khi áp dụng trên các bộ dữ liệu có sẵn FER+, AffectNet và RAF-DB.
- Chương trình minh họa cho phép nhận diện cảm xúc của tất cả khuôn mặt người có trong ảnh chụp.

TÀI LIỆU THAM KHẢO *(Định dạng DBLP)*

- [1]. Dosovitskiy Alexey, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani et al:
An image is worth 16x16 words: Transformers for image recognition at scale.
arXiv preprint arXiv: 2010.11929. 2020 Oct 22.
- [2]. Xue Fanglei, Qiangchang Wang, Zichang Tan, Zhongsong Ma, Guodong Guo:
Vision transformer with attentive pooling for robust facial expression recognition.
IEEE Transactions on Affective Computing: 2022 Dec 5.
- [3]. F. Xue, Q. Wang, and G. Guo:
TransFER: Learning Relation-aware Facial Expression Representations with Transformers.
ICCV: Mar 2021.