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

Đoàn Nguyễn Trần Hoàn - 21520239
Nguyễn Quốc Trường - 21521604

Tóm tắt

- Lớp: CS519.O11
- Link Github của nhóm: https://github.com/lIlDrAgOoN/CS519.O11
- Link YouTube video:
- Ảnh + Họ và Tên của các thành viên:

Đoàn Nguyễn Trần Hoàn
21520239

Nguyễn Quốc Trường
21521604

Giới thiệu

- 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...
- Mô hình ViT đã và đang được sử dụng để giải quyết bài toán này.
- Để cải thiện hiệu quả của mô hình ViT trên bài toán nhận diện cảm xúc, ta có thể kết hợp mô hình ViT hiện có với phương pháp giảm nhiễu Attentive Pooling, iúp mô hình tập trung vào phân tích các vùng ảnh quan trọng, liên quan nhiều tới cảm xúc con người.

Input:

Output:

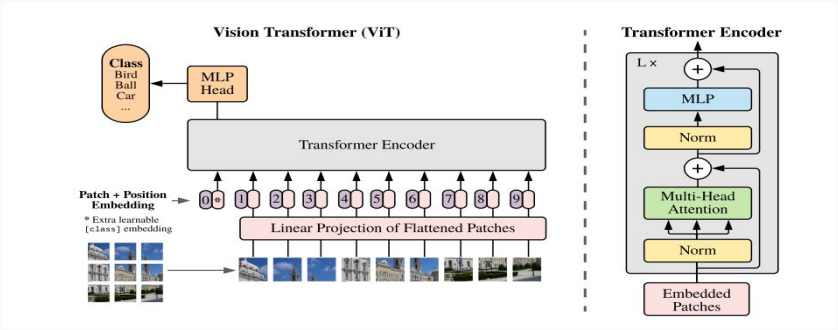
Happy

Mục tiêu

- 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

Nội dung:



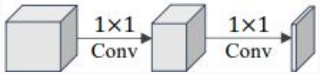
- Vision Transformer (ViT):

  Tìm hiểu và nghiên cứu 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.

- Hình thành Attention map:

  Sử dụng mạng ResNet50 thực hiện trích xuất các đặc trưng của ảnh và tạo thành một feature map gồm nhiều patch bằng nhau. Sau đó áp dụng liên tiếp 2 lớp Conv 1x1 để giảm chiều sâu feature map và tạo ra attention map

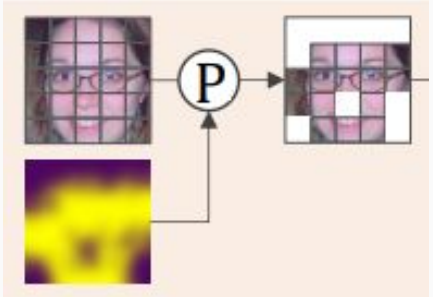
  Từ attention map, thực hiện loại bỏ các yếu tố nhiễu bằng nhiều phương pháp khác nhau.

- Phương pháp loại bỏ nhiễu:

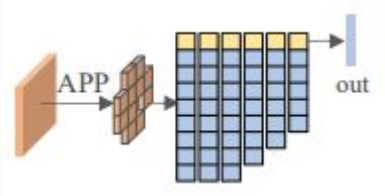
  Sử dụng phương pháp giảm nhiễu Attentive Pooling với hai module là Attentive Patch Pooling và Attentive Token Pooling.

Nội dung và Phương pháp

- Attentive Patch Pooling (APP):
  Ảnh sau khi được chia thành các patch bằng nhau.
  Chọn ra các patch quan trọng và loại bỏ các patch nhiễu.

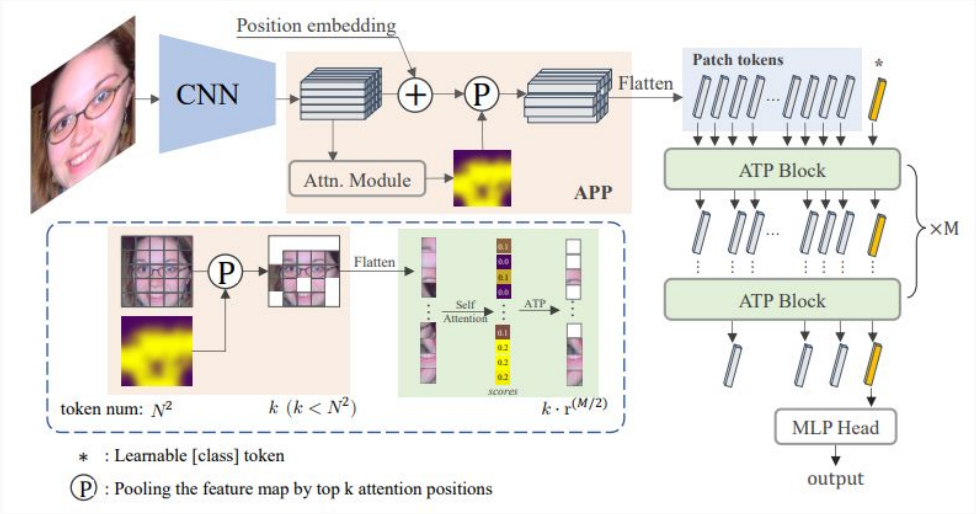


- Attentive Token Pooling (ATP):
  Sử dụng cơ chế attention được cải tiến của Transformer Encoder nhằm tập trung vào các patch token có trọng số (giá trị attention) cao nhất. Đó là các token quan trọng và có liên quan nhất tới bài toán nhận diện cảm xúc.
  Từ đó giảm thiểu ảnh hưởng từ nhiễu và tiết kiệm thời gian tính toán của mô hình.



Nội dung và Phương pháp

- Kết hợp ViT và hai module giảm nhiễu APP và ATP để hình thành mô hình Attentive Pooling modules with Vision Transformer(APViT):



- Chọn lọc và điều chỉnh các bộ dữ liệu FER+, RAF-DB, AffestNet cho phù hợp. Huấn luyện mô hình APViT trên các bộ dataset này.
- Xây dựng chương trình ứng dụng minh họa

Kết quả dự kiến

- 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

- 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.
- 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.
- F. Xue, Q. Wang, and G. Guo, “TransFER: Learning Relation-aware Facial Expression Representations with Transformers,” in ICCV, Mar. 2021.