Vision Transformer kết hợp giảm nhiễu cho nhận diện cảm xúc khuôn mặt

Đoàn Nguyễn Trần Hoàn [1,2] **Nguyễn Quốc Trường** [1,2]

[1] Trường ĐH Công Nghệ Thông Tin [2] Vietnam National University

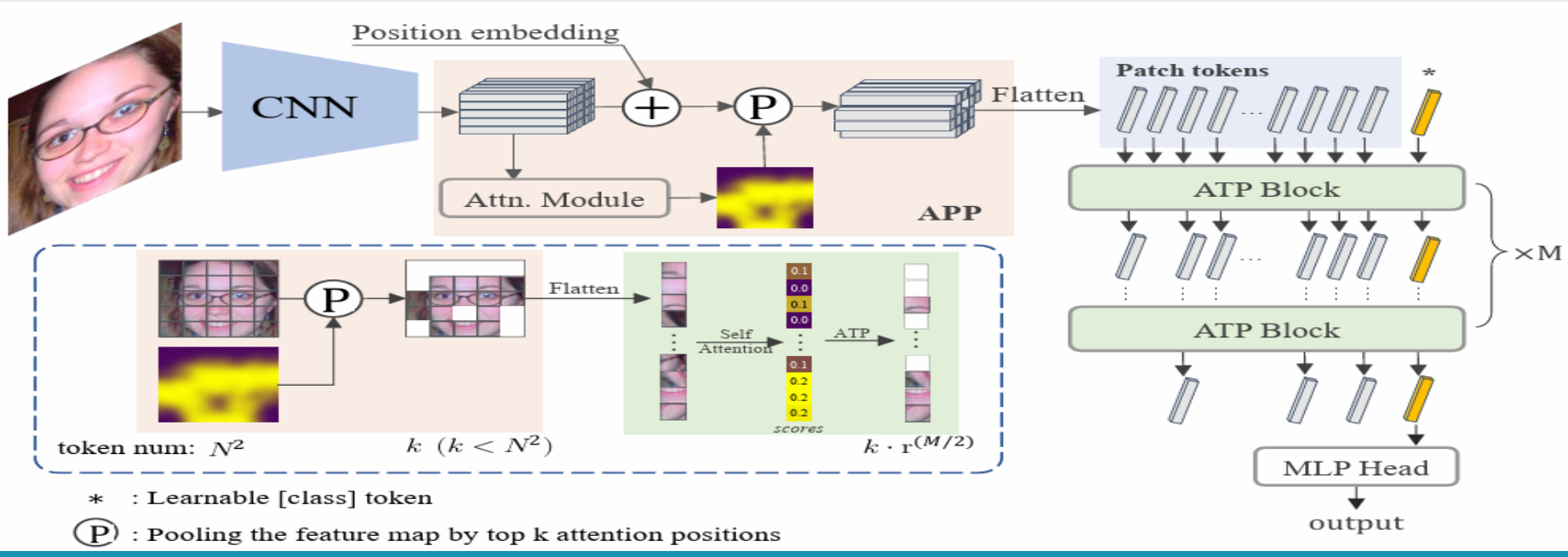
What ?

- Nghiên cứu mô hình Vision Transformer và cách cải thiện hiệu suất của nó.
- Nghiên cứu phương pháp tạo ra attention map để giảm nhiễu.
- Kết hợp 2 phương pháp trên trong một mô hình. Đánh giá và so sánh mô hình với các phương pháp nổi trội hiện nay.

Why ?

- 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.
- Vision Transformer có khả năng trích xuất đặc trưng tốt khi được huấn luyện trên dataset đủ lớn. Tuy nhiên hiện nay các dataset về cảm xúc khuôn mặt còn hạn chế vì vậy cần có phương pháp loại bỏ nhiễu tốt để ViT học tốt các đặc trưng của ảnh.

Overview

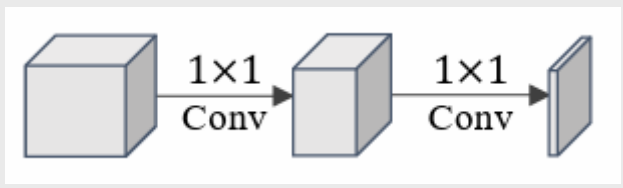


Description

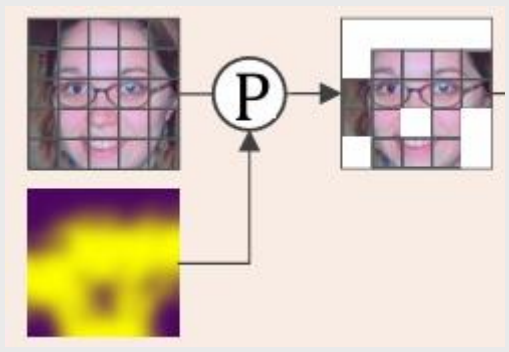
1. Attentive Patch Pooling (APP)

Phân chia ảnh thành các patch có kích thước bằng nhau. Chọn ra các patch quan trọng, loại có các patch nhiễu. Cụ thể:

- Tạo ra attention map bằng cách sử dụng liên tiếp 2 lớp Conv có kích thước cửa sổ trượt là 1x1 để giảm chiều sâu của feature map.

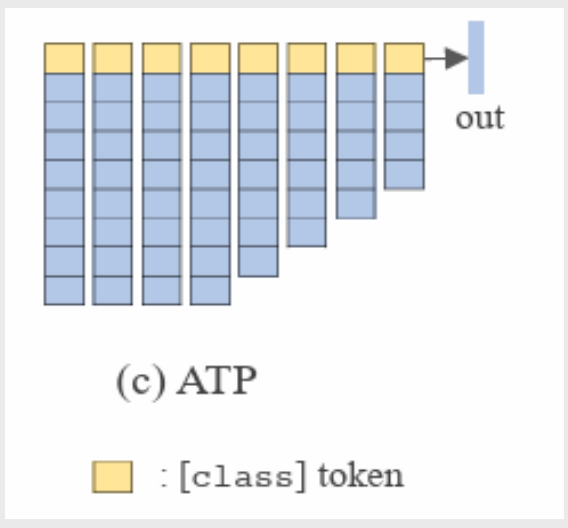


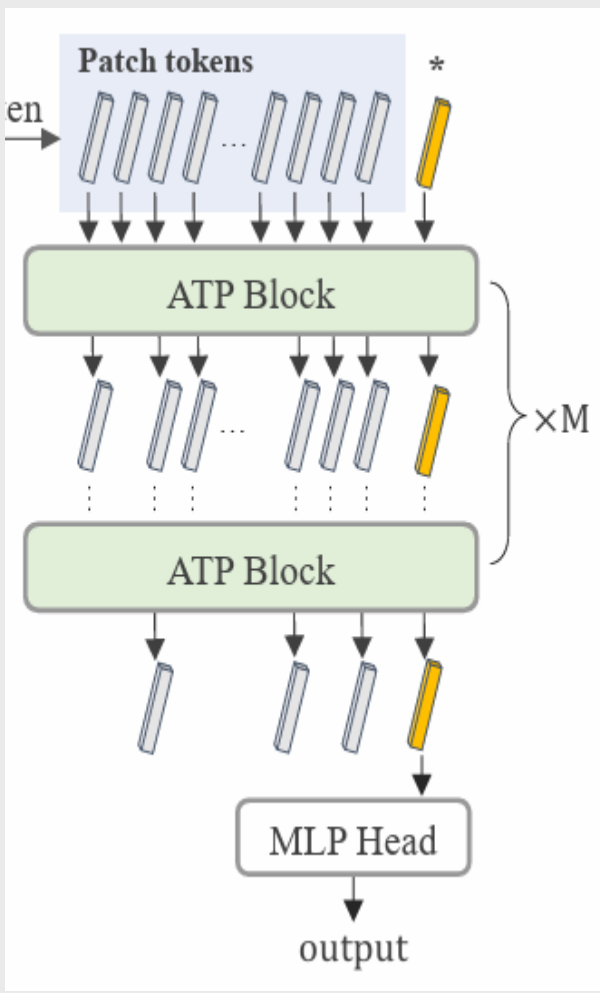
- Giữ lại top-k patch có giá trị attention cao nhất.



2. Attentive Token Pooling (ATP)

- Sử dụng cơ chế attention của Transformer để tập trung vào các patch token có giá trị attention cao nhất để giảm thiểu ảnh hưởng do nhiễu và đồng thời tiết kiệm tính toán.
- Điểm khác biệt với mô hình ViT: giảm dần số lượng token (do các nhà nghiên cứu nhận thấy trong mô hình DeepViT, các tầng sâu bị suy giảm giá trị attention)





NII

Đoàn Nguyễn Trần Hoàn – Trường Đại học Công Nghệ Thông Tin TPHCM

TEL : 0912342169 Email : 21520239@gm.uit.edu.vn