Dual-branch Cross-Patch Attention Learning for Group Affect Recognition

Hongxia Xie, PhD Candidate
Advisor: Prof. Wen-Huang Cheng
National Yang Ming Chiao Tung University, the Institute of Electronics

Abstract

Group affect refers to the subjective emotion that is evoked by an external stimulus in a group, which is an important factor that shapes group behavior and outcomes.

Most of the existing methods are proposed to detect faces and objects using pre-trained detectors and summarize the results into group emotions by specific rules. However, such affective region selection mechanisms are heuristic and susceptible to imperfect faces and objects from the pre-trained detectors.

In this work, we incorporate the psychological concept called **Most Important Person (MIP).** It represents the most noteworthy face in the crowd and has an affective semantic meaning. We propose **the Dual-branch Cross-Patch Attention Transformer (DCAT)** which uses global image and MIP together as inputs.

With parameters less than 10x, the proposed DCAT outperforms state-of-the-art methods on two datasets of group valence prediction, GAF 3.0 and GroupEmoW datasets.

Introduction

The **Most Important Person (MIP)** of an image is often the group leaders who can influence the emotion of a group.

For example, in Figure 1.(a) the expression of the MIP is clear and in accordance with group emotion, even in a large group in the right side image.

However, only considering MIP for group affect recognition is not enough since it may have facial occlusions, making it difficult to detect their emotions. Moreover, the expression of the MIP may not always correspond to the group's emotion in some cases, as shown in Figure 1.(b) .

Figure 1. The MIP in group.

Proposed Method

Based on the above observations, we present a dual-pathway vision transformer model, the Dual-branch Cross-Patch Attention Transformer (DCAT), including a global branch and a MIP branch.

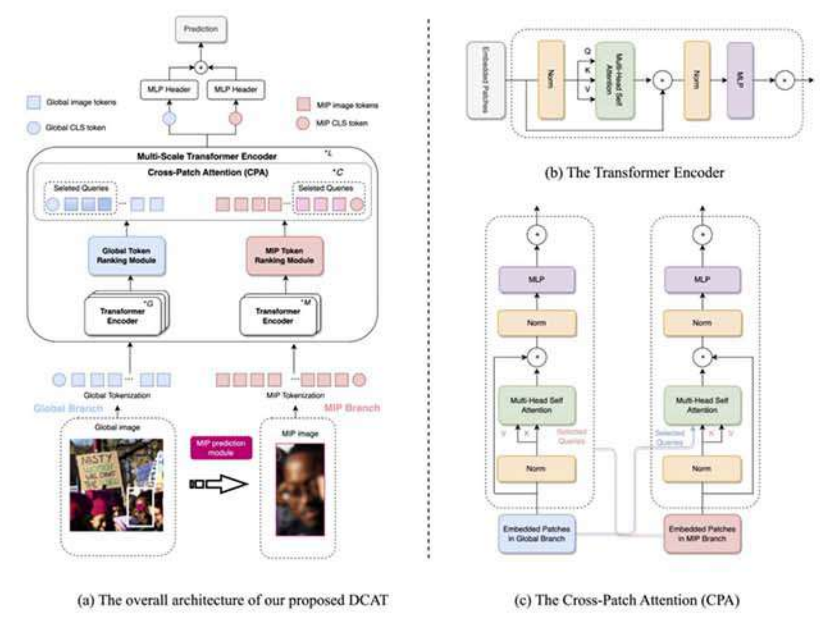

Figure 2. The overall architecture of our proposed DCAT.

Results

We conduct experiments on two challenging group valence prediction datasets: GroupEmoW and GAF 3.0.

Table 1. Overall Accuracy Comparison on the GroupEmoW dataset. "Source" refers to the types of features used in each paper.

Methods	Acc (%)	Source
Group affect methods		
ResNet34 [19]	68.13	Global, Faces
SE-ResNet-50 [21]	69.79	Global
Efficientnet-b2 [36]	72.33	Global, Faces
CAER-Net [26]	80.61	Global, Faces
GNN [17]	82.38	Global
GNN [17]	84.62	Global, Faces, Objects
ConGNN [39]	85.59	Global, Faces, Objects
WACV 21 [23]	89.36	Global, Faces
GNN [17]	89.93	Global, Faces
WACV 21 [23]	90.18	Global, Faces, Objects
Vision transformers		
ViT [8]	87.50	Global
P2T-base [40]	86.50	Global
DeepViT [47]	83.90	Global
Conformer-L [31]	83.35	Global
Focal-B [44]	87.98	Global
CrossViT [4]	88.48	Global
TransFG [18]	89.47	Global
RegionViT-B [3]	89.49	Global
CSwin-L [7]	89.90	Global
DVT [38]	87.90	Global
Evo-LeViT [42]	87.83	Global
TCFormer-L [46]	89.24	Global
DCAT-ViT	89.55	Global, MIP
DCAT-CSwin-L	**90.47**	Global, MIP

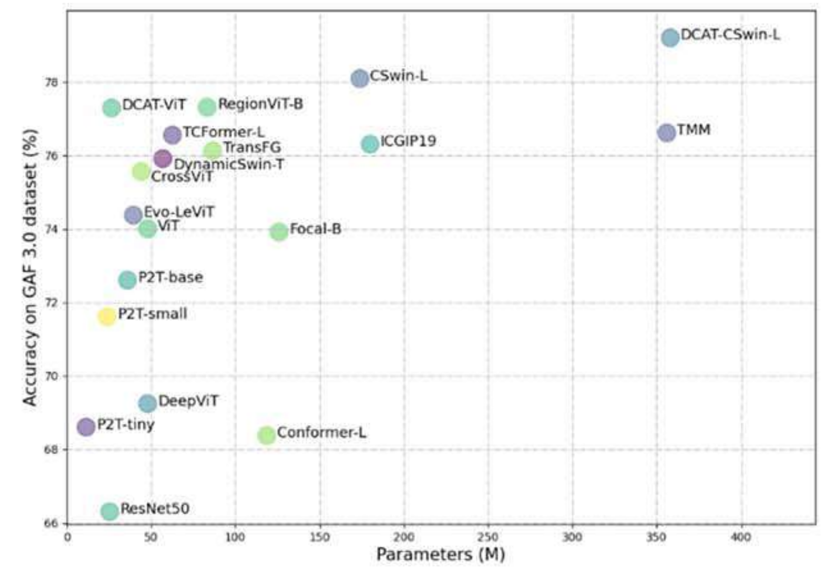

Figure 3. Model parameter and accuracy trade-off comparison on the GAF 3.0 dataset.

Feature Map Visualization

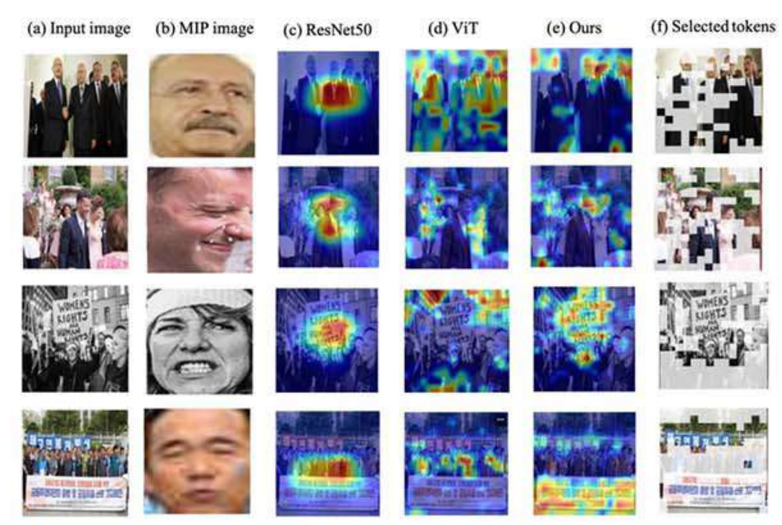

Figure 4. Comparison of feature map visualization among ResNet50, ViT and our proposed method.

Conclusions

In this work, by incorporating the MIP concept, we propose a dual-pathway vision transformer model, the Dual-branch Cross-Patch Attention Transformer (DCAT) for group affect recognition.

Moreover, we propose the CPA module so that the important tokens from both paths can be interacted with.

In terms of accuracy and parameters, the proposed DCAT outperforms both state-of-the-art group affect models and vision transformer models. It is also the first study to explore group affect recognition beyond CNNs.

Moreover, our proposed model can be utilized in other group-level affect tasks, *i.e.*, group cohesion analysis..

Contact

For more research details, please contact *hongxiaxie.ee08@nycu.edu.tw*