An Attention Model for Group-Level Emotion Recognition

Aarush Gupta[1], Dakshit Agrawal[1], Hardik Chauhan[1], Jose Dolz[2] and Marco Pedersoli[2]

Indian Institute of Technology Roorkee[1], École de Technologie Supérieure[2]

Overview

- Previous SOTA models for group -level emotion recognition are too sophisticated.
- They do not account for the fact that some faces may be misleading.

Goal: To find a simpler model for group-level emotion recognition that gives more importance to significant faces.

Introduction

- Two-branched model (Fig. 2):
  - **Global-Level CNN:** detects emotions on the basis of the image as a whole.
  - **Local-Level CNN:** detects emotions on the basis of the faces present in the image.
- The features of each face are merged into a single representation by an attention mechanism.

Datasets

- Split EmotiW validation data into VAL and EVAL.
- Images from the **EmotiC dataset** (Fig. 1) are employed for pre-training.

Figure 1: Some samples of the EmotiC Dataset

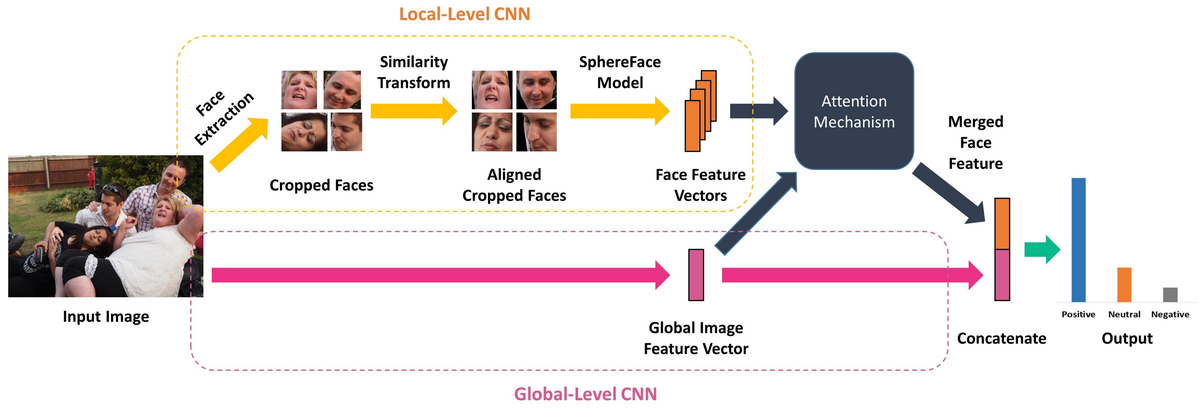

Figure 2: Overview of the proposed attention model

Global-Level CNN

- The surroundings are an important cue for recognizing emotion.
- Deploy state-of-the-art pre-trained classification network to learn global features of the whole image.

Local-Level CNN

- Use the MTCNN model for face extraction.
- Apply similarity transform using the facial landmarks.
- Pass resulting aligned image through a pre-trained SphereFace network.

Attention Mechanisms

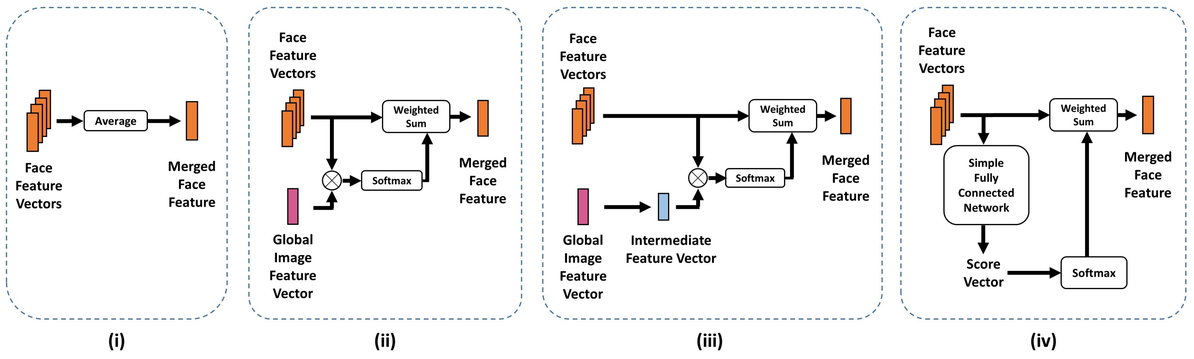

Figure 3: Attention mechanisms to merge the Face Feature Vectors: (i) Average (ii) Attention_A (iii) Attention_B (iv) Attention_C

Results

Table 1: Model Results

	Model	VAL	EVAL
	Global	69.50%	70.80%
	Local	71.18%	72.40%
Attention Models	Average	73.03%	73.90%
	Attention_A	73.18%	73.00%
	Attention_B	74.26%	75.20%
	Attention_C	73.66%	**76.20%**

Table 2: Quantitative Results

Model	EVAL	EmotiW 2018 Test Dataset			
		Positive	Neutral	Negative	Overall
Baseline	–	75.00%	50.00%	53.00%	61.00%
Single	78.20%	66.59%	57.97%	58.87%	61.84%
Ensemble	80.90%	71.33%	60.48%	59.71%	**64.83%**

Figure 4: Sample predictions of our best model

Conclusion

- Main contribution lies in using attention mechanism to merge the facial features.
- We also explored the use of a larger similar dataset for pre-training.
- Future work includes exploring cues such as pose and context.

Scan me