Group Emotion Recognition with Individual Facial Emotion CNNs and Global Image based CNNs

Lianzhi Tan, Kaipeng Zhang, Kai Wang, Xiaoxing Zeng, **Xiaojiang Peng**, Yu Qiao

Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences

{xj.peng,yu.qiao}@siat.ac.cn

November 17, 2017

Overview

Motivation

- Individual facial emotion is the most important cue for group emotion recognition
- Group emotion is relevant to the global scene, e.g. an image is likely to be positive taken from a wedding party while negative from a funeral

Figure: In the left image, there is no scene but only faces. An India wedding in the right image.

Contribution

1. An aligned facial emotion CNN
2. An non-aligned facial emotion CNN
3. An evaluation of several global image based CNNs
4. An extra dataset searched using train/val sets
5. Winner of the GReco task in the EmotiW17 challenge

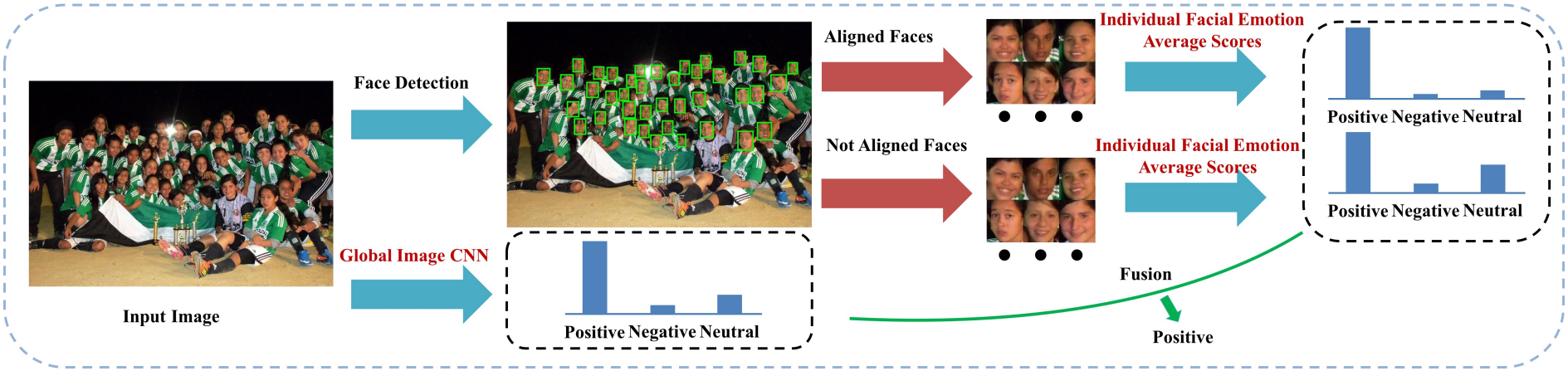
Related Work

- A similar group-level happiness estimation in the last challenge EmotiW16
- Dhall *et al*. propose the current task and introduce the baseline as follows
  - Top-down component (global): scene features like CENTRIST or GIST
  - Bottom-up component (local): BOW with low-level features like PHOG and LPQ, BOW with high-level facial action unit feature
- The last winner extracts the CENTRIST feature and ResNet18-based feature vectors for individual faces, and then feeds them into LSTM units

All use global image based features and individual facial features!

Our Approach

Global features? Local features? Go to deeper CNNs!



Our method contains two kinds of CNNs, namely the individual facial emotion CNNs and the global image based CNNs as indicated by red texts. The final prediction is made by averaging all the scores of CNNs from all faces and the image.

Individual Facial Emotion CNNs

We consider two individual facial emotion models

- Aligned model: pretrained on face recognition dataset with high resolution (i.e. 112×96)
- Non-aligned model: pretrained on facial expression dataset with low resolution (i.e. 48×48)

Face detection:

- We use MTCNN [1] to detect and align faces in the images

[1] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Processing Letters, 2016.

Aligned Facial Emotion CNNs

Our aligned model is composed of 64 convolutional layers, and two fully-connected layers with a batch normalization layer for FC1.

Type	Filter Size/Stride	Output Size
Conv1.x	$3 \times 3/2$ $\begin{bmatrix} 3 \times 3/1 \\ 3 \times 3/1 \end{bmatrix} \times 3$	48 x 56 x 64 48 x 56 x 64
Conv2.x	$3 \times 3/2$ $\begin{bmatrix} 3 \times 3/1 \\ 3 \times 3/1 \end{bmatrix} \times 8$	24 x 28 x 128 24 x 28 x 128
Conv3.x	$3 \times 3/2$ $\begin{bmatrix} 3 \times 3/1 \\ 3 \times 3/1 \end{bmatrix} \times 16$	12 x 14 x 256 12 x 14 x 256
Conv4.x	$3 \times 3/2$ $\begin{bmatrix} 3 \times 3/1 \\ 3 \times 3/1 \end{bmatrix} \times 3$	6 x 7 x 512 6 x 7 x 512
FC1_bn	512	512
FC2	3	3

Implementation

Training:

- We take the model (i.e., ResNet64) from [1] which is trained on the Webface dataset [2] using an Angular Softmax loss
- We finetune it on the ExpW facial expression dataset [3] with large-margin softmax (L-Softmax) [4]
- We finally finetune on EmtiW2017 dataset

For fintuning, the learning rate is started from 0.01, and divided by 10 at the 1.2K, 1.8K iterations. A complete training is finished at 2.2K iterations. No dropout is used.

Tools: Caffe (a version which supports L-Softmax and BN)

1 Weiyang Liu. Sphereface: Deep hypersphere embedding for face recognition. In ICCV, 2017.

2 Dong Yi. Learning face representa- tion from scratch. arXiv:1411.7923, 2014.

3 Zhanpeng Zhang. From Facial Expression Recognition to Interpersonal Relation Prediction. CoRR abs/1609.06426, 2016.

4 Weiyang Liu. Large margin softmax loss for convolutional neural networks. In ICML, 2016.

Non-aligned Facial Emotion CNNs

All detected faces are scaled to 48×48 without alignment.

Architecture:

- ResNet34 (We also test ResNet18 and ResNet101 which are inferior to ResNet34 in our observation)

Training:

- We pretrain on FERPlus expression dataset [1] from scrach
- We then finetune it on EmtiW2017

For training from scratch, lr is initilized by 0.01, and divided by 10 every 4 epochs and we stop training at the tenth epoch.
For finetuning, lr is initilized by 0.001 with same training strategy as above.
We set a dropout of 0.9 after the average pooling layer for both steps.

1 EmadBarsoum,ChaZhang,CristianCanton-Ferrer,and Zhengyou Zhang. Training Deep Networks for Facial Expression Recognition with Crowd-Sourced Label Distribution. CoRR abs/1608.01041 (2016).

Evaluation on Validation Set

In our observation, removing small faces maybe help since only large faces are most relevant to annotations.
We consider two schemes: average scores from all faces or the largest 3 faces.

	Aligned(96×112)		Non-aligned (48×48)
	Softmax	L-Softmax	Softmax
keep top 3 faces	73.8	79.7	70.73
keep all faces	74.1	80.19	69.97

Global Image based CNNs

Architecture: VGG19, BN-Inception, ResNet101
Training:

- We pretrained VGG19 on Places dataset [1]
- BN-Inception and ResNet101 are downloaded from website which are pretrained on ImageNet

▷ All BN layers are fixed if there are.
▷ For finetuning, all images are scaled to have a minimum side of 256 and then cropped as 224×224 randomly.
▷ We set a dropout of 0.9 after the average pooling layer for ResNet101 and BN-Inception, and 0.5 for both the FC6 layer and the FC7 layer of VGG19.
▷ We also test L-Softmax for VGG19 and BN-Inception.

1 Bolei Zhou,Agata Lapedriza,Jianxiong Xiao, etc. Learning deep features for scene recognition using places database.NIPS, 2014.

Evaluation on Validation Set

Extra Data Collection

With the search by image engin of Baidu and Google, we also search similar images using the train/val sets, and carefully select about 2K images for each class.

Testing: we average the scores of the center crop and its horizontal flipping.

	VGG19		BN-Inception		ResNet101	
	Softmax	L-Softmax	Softmax	L-Softmax	Softmax	L-Softmax
+extra data	-	73.2	-	67.3	**74.7**	73.2
-	67.2	70.6	60.54	65.3	71.04	-

Model Combination

- average score fusion
- weighted score fusion (guided by the individual accuracy)

Model (score weight)	Acc. on validation set (%)
Aligned	80.2
Aligned (3.0) + Global (0.5)	81.0
Aligned (3.0) + Global (0.5) + Non-aligned (0.5)	82.7
Aligned (1.0) + Global (1.0) + Non-aligned (1.0)	78.4

Our submissions

- 1: Aligned (3.0) + Global ResNet101 (0.5) + Non-aligned (0.5)
- 2: Aligned (3.0) + Global VGG19 (0.5) + Non-aligned (0.5)
- 3: Aligned (3.0, train on trainset) + Global ResNet101 (0.5) + Non-aligned (0.5)
- 4: Aligned (3.0, train on trainset) + Global VGG19 (0.5) + Non-aligned (0.5)
- 5: Aligned (3.2) + Global ResNet101 (0.8, extra data) + Non-aligned (0.6)
- 6: Aligned (3.2) + Global VGG19 (0.8, extra data) + Non-aligned (0.6)
- 7: Aligned (3.2, train on trainset) + Global ResNet101 (0.8, extra data) + Non-aligned (0.6)

Runs	Validation	Test			
	Overall	Positive	Neutral	Negative	Overall
1	82.7	82.0	57.6	76.7	79.9
2	81.7	82.3	57.0	74.7	79.1
3	-	85.2	63.6	67.6	78.9
4	-	85.2	64.84	66.6	78.8
5	83.7	83.9	58. 8	76.4	**80.9**
6	83.8	83.9	58.2	74.0	79.8
7	-	85.5	59.4	70.0	79.1

Conclusion

- With traditional Softmax, the high-resultion facial model and global image model perform similarly (about 74%)
- Facial model can be improved by a more discriminative loss function: L-Softmax
- Aligned and non-aligned facial models are complementary.

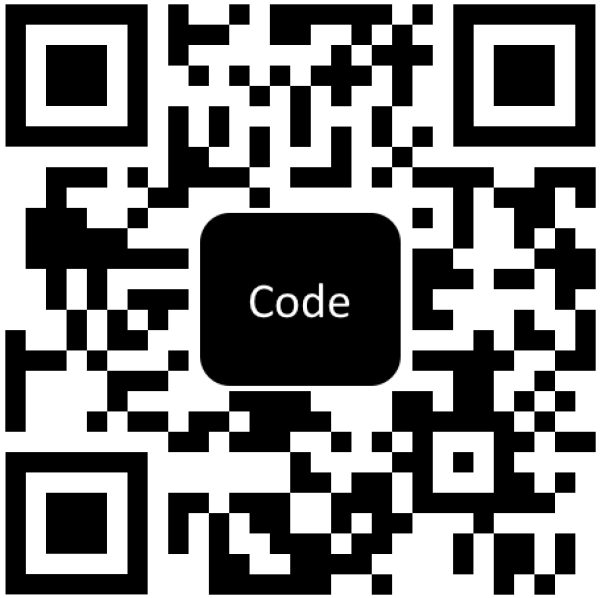
Q&A

▷ 1: Why did you use the aligned facial model which only trained on trainset for submission 3 and 4?

- L-Softmax is not stable in our observation
- We use it to make sure that the model performs best on validation set.

▷ 2: Did you try other fusion methods like feature-level fusion, multiple kernel learning, and training fusion weights?

- No. We think feature-level fusion is better but we don't have time. Training fusion weights maybe overfit training data since there are only 3 dimentionities.



Thank you for your attention.