Deep Face Emotion Recognition

Zhiqin Zhang

1 School of computer science, Wuhan Donghu University, Wuhan 430000, China

Abstract. We extend the CNN based face emotion recognition to deal with the confusion of emotion recognition. We achieve state-of-the-art results on complex environment such as low or local light and blurry face details by using multiple input features fusion and mask loss which can focus on the valid local facial features, without any further refining and weighting multiple results module. Moreover, due to DenseNet construction of the model, our approach has much less parameters. Our method was tested on the Emotion Recognition in the Wild Challenge, Static Facial Expression Recognition sub-challenge (SFEW) and shown to provide a substantial 35.38% improvement over baseline results.

1. Introduction

Convolutional Neural Networks (CNNs) are driving major advances in many computer vision tasks, such as image classification [1], object detection [2, 3] and semantic image segmentation [4]. The last few years have witnessed outstanding improvements on CNN-based models. Very deep architectures [5, 6] have shown impressive results on standard benchmarks such as ImageNet or MSCOCO [7]. State-of-the-art CNNs heavily reduce the input resolution through successive pooling layers and thus, are well suited for applications where a single prediction per input image is expected (e.g. image classification task).

Recently, a new CNN architecture, called DenseNet, was introduced. DenseNets are built from dense blocks and pooling operations, where each dense block is an iterative concatenation of previous feature maps. This architecture can be seen as an extension of ResNets [6], which performs iterative summation of previous feature maps. However, this small modification has some interesting implications: (1) parameter efficiency, DenseNets are more efficient in the parameter usage; (2) implicit deep supervision, DenseNets perform deep supervision thanks to short paths to all feature maps in the architecture (similar to Deeply Supervised Networks [8]); and (3) feature reuse, all layers can easily access their preceding layers making it easy to reuse the information from previously computed feature maps. The characteristics of DenseNets make them a very good fit for semantic segmentation as they naturally induce skip connections and multi-scale supervision.

In this paper, we extend DenseNets to work as face emotion recognition base model prior to the softmax layer. To solve the problem of blurry face details, multiple mixed input features are inputted to the net which can contain different scale and local face detail features. Because different local face details have different importance to emotion recognition, a mask loss was proposed which can focus on different face area.

2. Related Work

Facial emotion recognition has received increasing interest in the last two decades. The nature of the problems considered by previous work is reflected in the benchmarks used to measure and report
performances. For example, early relevant data-sets contained only constrained images taken in laboratory controlled conditions. More recently, unconstrained, in-the-wild photos have been considered with the release of the EmotiW challenge [9].

With the emergence of very large classification data-sets in recent years and the improvement of computation, deep CNNs have been applied in various image classification problems, including, e.g., object recognition [2], scene recognition, face verification, age and gender classification [10], and more.

We proposed the use of fusion features as input to CNN models and used mask loss to obtain 25.38% improvement over baseline results. We show that doing so boosts performance well beyond that obtained with CNN models trained on RGB alone.

3. Deep emotion recognition structure

In this section, we detail the proposed model for face emotion recognition. First, we review the recently proposed DenseNet architecture. Second, we introduce the construction of the novel mixed multiple features input and discuss its advantages. Finally, we wrap up with the details of the main architecture and mark loss.

3.1. Review of DenseNets

In order to ease the training of very deep networks, ResNets introduce a residual block that sums the identity mapping of the input to the output of a layer. The resulting output $x_\ell$ becomes

$$x_\ell = H_\ell(x_{\ell-1}) + x_{\ell-1} \tag{1}$$

Allowing for the reuse of features and permitting the gradient to flow directly to earlier layers. In this case, $H$ is defined as the repetition (2 or 3 times) of a block composed of Batch Normalization (BN), followed by ReLU and a convolution.

Pushing this idea further, DenseNets design a more sophisticated connectivity pattern that iteratively concatenates all feature outputs in a feedforward fashion. Thus, the output of the $l$th layer is defined as

$$x_\ell = H_\ell([x_{\ell-1}, x_{\ell-2}, \ldots, x_0]) \tag{2}$$

Where $[ \ldots ]$ represents the concatenation operation. In this case, $H$ is defined as BN, followed by ReLU, a convolution and dropout. Such connectivity pattern strongly encourages the reuse of features and makes all layers in the architecture receive direct supervision signal.

3.2. The proposed structure

DenseNet was used as our base model, but we reduced the total layers to 56 and the growth rate was set to 12, in our experiments we found that the proposed reduction densely can get better recognition precision than the original 121 layers DenseNet, the reason may be that the less parameters are better to fit to the emotion recognition task which reduce over-fitting. The details of the propose method are as follows.

1. Mixed features input

The net input was composed of three kinds of features including a) source face RGB image b) first half of face image c) half bottom of face image. The two half parts of the face original image are overlapped and the ratio is 1:2. In our experiments we found the mixed features can reduce 28% error rate. The reason may be that the dominant emotion features are lie on the half bottom of face. We resized all the features to the same dimension and concatenated the three kind’s features as input. The Details showed in figure 1.

2. Mask loss

Different facial areas have different importance to facial emotion recognition, so we choose to weight the facial area according to the coordinates. We added an auxiliary mask loss prior to the final classification cross entropy loss. A suitable feature map was chosen to multiply by a mask weight, and
then the weighted feature map was inputted to the next layer for auxiliary mask loss, the details showed in figure 1. The mask loss formula is as follows:

\[ l_{\text{mask}} = L_{\text{crossentropy}}(f_{\text{softmax}}(f_{\text{gp}}(f_{\text{cov}}(m(x_i) \cdot f_m))) \) \] (3)

Where \( m \) represents the mask weight, if \( x_i \) in the third feature area, \( m(x_i) = 0.7 \), if \( x_i \) in the second feature area, \( m(x_i) = 0.3 \). \( x_i \) is the coordinate of the 3d feature map (here we ignored the samples number channel). \( f_m \) represents the feature map weighted with the mask, \( f_{\text{cov}} \) represents the identity map layer, \( f_{\text{gp}} \) represents the global average pooling layer which converts the prior 3d feature map into vector map, \( f_{\text{softmax}} \) is the softmax normalization layer. So the final loss becomes:

\[ L = w_1 \cdot l + w_2 \cdot l_{\text{mask}} \] (4)

Where \( w_1 = 1.0 \), \( w_2 = 0.45 \). \( l \) represents the last standard cross entropy loss.

(3) Architecture and training details

We did not initialize our models using any pre-trained face recognition model weights and train them from scratch with SGD, with an initial learning rate of \( 1e^{-2} \) and an exponential decay of 0.995 after each epoch. All models are trained on data augmented with random crops and vertical flips. We monitor mean accuracy and use patience of 100. We regularized our models with a weight decay of \( 1e^{-4} \) and a dropout rate of 0.2. The architecture was showed in Figure 1.

Figure 1. Densenet 56 structure. RGB, first half and half bottom images are resized to the same dimension as input to the proposed net. The final loss is composed of mask loss and original densenet cross entropy loss

4. Experiments

Our tests were performed on the EmotiW 2015 benchmark [9] which includes data from version 2.0 of the Static Facial Expression in the Wild benchmark. It was assembled by selecting frames from different videos of the Acted Facial Expressions in the Wild (AFEW), and then assigning them one of the following seven emotion labels: angry, disgust, happy, sad, surprise, fear and neutral. Images from this data set are unconstrained and cover a wide range of head poses and ages, both genders, different occlusions and resolution qualities.

Table 1. Additionally provides an accuracy in each emotion category for the training and validation subsets.

|        | Anger   | Disgust | Fear    | Happy   | Neutral | Sad    | Surprise | Total  |
|--------|---------|---------|---------|---------|---------|--------|----------|--------|
| Train  | 95.3%   | 88.6%   | 93.7%   | 87.8%   | 86.7%   | 93%    | 91%      | 90.2%  |
| Val    | 89%     | 85.5%   | 89%     | 82.7%   | 81%     | 88%    | 82%      | 86%    |

We also summarized the results on the validation set for all of the different network architectures, image presentations and ablation study considered. The results showed that our DenseNet architecture can achieve 18% improvement, 10% and 7.4% for fusion features and mask loss respectively. Our results should be compared against the baseline performance for the benchmark. These were obtained using features produced from pyramids of Histogram of Gradients and Local Phase Quantization extracted from the aligned faces and classified using a fusion of separate support vector machines. Our
single model results can obtain a boosted performance substantially, by a remarkable 35.38% improvement in performance.

Finally, we analyzed a selection of correct and wrong classification results for all of the six classes. The results showed that even in local light environment, good performance can be achieved. Some bad cases showed that poor performance may be caused by wrong face alignment.

5. Conclusions
We present a substantial improvement over existing baseline results on the Emotion Recognition in the Wild Challenge (EmotiW 2015), Static Facial Expression Recognition sub-challenge (SFEW). To achieve this performance boost, we make a number of novel contributions: We propose to apply multiple feature fusion as the input to CNNs rather than RGB. In order to eliminate over-fitting and deal with clutter unconstrained environment light, a DenseNet-like structure and mask loss were proposed. Our results clearly demonstrate the advantage of looking beyond RGB as the input space for CNNs, as well as the complementary information offered by mask loss and network architectures.

Acknowledgement
This research was financially supported by Hubei Provincial Department of education scientific research project, which the number is B2017301.

References
[1] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. CoRR, abs/1409.1556, 2014.
[2] J. Redmon, S. K. Divvala, R. B. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. CoRR, abs/1506.02640, 2015.
[3] S. Ren, K. He, R. B. Girshick, and J. Sun. Faster R-CNN: towards real-time object detection with region proposal nets. CoRR, abs/1506.01497, 2015.
[4] A. Kundu, V. Vineet, and V. Koltun. Feature space optimization for semantic video segmentation. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016. pp 3168-3175.
[5] G. Huang, Z. Liu, K. Q. Weinberger, and L. van der Maaten. Densely connected convolutional networks. CoRR, abs/1608.06993, 2017.
[6] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. CoRR, abs/1512.03385, 2015.
[7] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In European Conference on Computer Vision (ECCV), 2014.
[8] C. Lee, S. Xie, P. W. Gallagher, Z. Zhang, and Z. Tu. Deeply-supervised nets. In International Conference on Artificial Intelligence and Statistics (AISTATS), 2015.
[9] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon. Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark. In Proc. Int. Conf. Comput. Vision Workshops. IEEE, 2011. pp 2106–2112.
[10] Shixing Hen, Shixing Zhang, Caojin and Dong, Ming and Le, Jialiang and Rao, Mike. Using Ranking-CNN for Age Estimation. CVPR 2017. pp 742-751.