An Improvised Deep Learning method for Improving Emotion Classification

Satyajit nayak[1], Suren Ku. Sahu[2], Rakhi Jha[3]

[1,3]Associate Professor, Department of Computer Science Engineering, Gandhi Institute For Technology (GIFT), Bhubaneswar

[2]Assistant Professor, Department of Computer Science Engineering, Gandhi Engineering College, Bhubaneswar

Publishing Date: January 27, 2017

Abstract

In this work, we show that the discriminative power of a deep neural network can be improved at three different levels: (i) inputting discriminative features, (ii) designing an optimal network architecture and (iii) changing the loss function. This work suggests that only increasing the depth or the width of a deep network may not always be the best solution while requiring computationally efficient models. Here, we show that there is scope for improving the classifier accuracy at each of the three levels. We have carried out all our experiments on the FERplus dataset and show that the facial emotion recognition accuracy can be independently improved up to 2.5% by adding better features, 1.8% by modifying the loss function and up to 3.1% by combining the two ideas. In separate experiments, we show that the computational complexity can be reduced by a factor of 24.3, while simultaneously increasing the FER by 0.95% by modifying the architecture of the model, input features and the loss function.

Index Terms : **Emotion Classification, Deep Learning, Loss function, discriminative features.**

I. INTRODUCTION AND RELATED WORK

Human emotions can be categorized into the following broad classes: neutral, anger, disgust, fear, happiness, sadness, surprise and contempt. Developing intelligent systems that are able to recognize these emotions have a lot of practical use. Such as determining patient feeling or comfort level during healthcare. Monitoring the driver's fatigue condition in a smart car and alert if the driver is feeling drowsy or sleepy to take a break. In the case of e-learning, the presentation can be adjusted based on the style of learner. ATM should automatically detect fear and should not dispense money. In the Gaming industry, automatic facial expression can give feedback if the game is successful in providing an enjoyable experience. Similarly, there are many more practical applications of automatic facial emotion recognition (FER). So, there is a need to develop computationally efficient automatic FER systems or techniques to improve the performance of such systems.

Facial action coding system (FACS) is developed by Ekman et al. [1]. Facial expression can be analyzed by mapping facial action units for each part of the face (eyes, nose, mouth corners) into codes. Another approach includes feature extraction using pyramid histogram of gradients [2]. Here, the facial edge contours are constructed using a Canny edge detector. Histograms are calculated by dividing the edge maps into different pyramid resolution levels. The histogram vectors are then concatenated to generate the final feature to be used by the SVM or AdaBoost classifier. In boosted local binary patterns (LBP) [3] for facial emotion recognition, a face is divided into small regions from which LBP histogram features are computed. For a given class, a template is generated from the histogram of facial features. Then the nearest neighbour classifier is used for classifying a given image. Deep learning has significantly improved the performance in image classification tasks, compared to the traditional machine learning techniques [4], [5], [6], [7], [8]. In [9] and [10], the authors have explored modifying the input for better reconstruction by using multiple interpolations for the task of super-resolution of natural and document images and in [11], by creating a rich, diverse training data-set capturing multiple variations in the input low-resolution images for the super-resolution of binary document images. In [12] and [13], the authors have explored modifying the objective function for improving super-resolution and denoising algorithms, by adding an edgepreserving loss to the mean square error loss function.

For the task of facial emotion recognition, the current state of the art models are by Microsoft [14], and Pandey et al. [15], [16]. Barsoum et al. [14] proposed a miniature version of VGG net, called VGG13 shown in Fig. 1, which has 8.75 million parameters. The dataset used is the FERplus dataset, which has 8 classes, adding neutral to the existing seven classes. The reported test accuracy is ≈ 84%. Levi et al. [17] convert images to a local binary pattern and maps it to a 3D metric space, which is used as an input to the CNN thus largely addressing the problem of appearance variation due to illumination. In [18], the CNN based AlexNet architecture is fine-tuned using Cohn-Kanade [19] dataset for real-time emotion detection. In [22], the

authors use landmark points to generate a grid on the face. The localized regions are used for feature extraction using LBP and normalized central moments. The extracted features are classified using SVM. In [20], the authors extract temporal appearance features and geometric features (landmark points) using two CNNs in parallel, which are combined together for classification.

This paper extends the work reported in [15] [16]. Here, we have shown how the change in the loss function, together with addition(s) to the input features affect the performance of an emotion classifier. We also show that there can be many ways to increase the discriminative power of a network and that merely increasing the depth and the width of a deep network may not always be the only solution for the same.

The following are the contributions of this work: • We show that there is good scope for modifying the input feature space to obtain better FER performance.

- We illustrate how the change in the loss function can add further performance gain to the existing architectures (see Tables I & II and Figs. 2 & 4).
- We also show that it is possible to obtain computationally efficient, deep architectures, without a major reduction in performance.
- We have shown that there are other options, such as modifications to the input features and the loss function to improve the classifier accuracy, than simply increasing the width and depth of the network (see the original FER result in Table I and Pandey et al. [16] in Table II).

II. DATASET USED FOR THE STUDY

We use the FERplus dataset, which contains approximately 35000 images divided into 8 classes: neutral, anger, disgust, fear, happiness, sadness, surprise and contempt. It improves upon the FER dataset by crowd-sourcing the tagging operation. Ten taggers are asked to choose one emotion per image and a distribution of emotions is obtained for each image. The training set contains approximately 28000 images. The remaining are divided equally into validation and test sets. The original image size is 48×48 pixels.

III. EXPERIMENTS

A. Details of the Baseline Architecture

Initially, we have recreated the baseline model VGG13 (Fig.:1) in Tensorflow. We have used the same preprocessing techniques employed in the original code. For training, we use the online data augmentation strategy as used in the paper. When selecting the images for training, testing and validation, we have used the majority voting technique as described in the paper. Every image in the dataset has more than one label, due to the fact that it was shown to 10 annotators, who had to select one of the emotions among the 8 classes. Using majority voting we have selected only the images that have more than 50% of the votes going to one emotion. Thus, only those emotions that have a clear class are used for training. The validation and test images are re-sized to 64×64 pixels. We have used softmax with cross-entropy loss for multi-class classification. We have used Adam optimizer with an initial learning rate of 0.0003, which has been multiplied by a factor of 0.97 every 2 epochs. We have obtained an overall accuracy of 83.4% as compared to the reported value of 83.9%.

B. Using the Laplacian and Gradient of the Facial Images

The Sobel operator, when applied on an image, approximates the gradient of the image and identifies the regions of high spatial frequency corresponding to the edges. Similarly, the Laplacian operator is a second order spatial derivative that identifies the regions of intensity changes around the edges. when used on facial images, these derivative operators have the ability to identify the edges along with the landmark points. In the present work, we concatenate channelwise Laplacian and/or gradient of the facial images together with the original image. The concatenated images are fed as input to the network.

C. Adding the Center Loss Function

The center loss [31] has been used in face recognition tasks to improve the discriminating power of the network by learning more differentiating features. We have used the centre loss, together with the softmax loss, as shown in equation 3 below. Intuitively, softmax places the features from different classes apart, while the center loss pulls the individual class features towards the respective class centers. The idea is to decrease intra-class variance, while at the same time increasing the inter-class variance. The total loss (TL) function has been formulated as follows:

$$L_C = \frac{1}{2}\sum_{j=1}^{M} ||x_j - c^j||_F^2 \quad (1)$$

$$L_{softmax} = -\sum^{M} log \frac{exp(W_j x_j + b_j)}{\sum_{i=1}^{C} exp(W_i x_j + b_i)} \quad (2)$$

$$L_{total} = L_{softmax} + \lambda L_C \quad (3)$$

where, x_j and c_j are a feature vector and its corresponding class center, W_j and b_j are the class weights and biases. λ is a hyperparameter used to decide the relative contribution from the center loss. Figure 3 shows how the accuracies of the baseline architecture (Fig. 1) and its proposed modifications change as a function of the value of λ. Softmax alone has been found to be insufficient for the task of face recognition, and hence other losses such as triplet [33] and contrastive loss [32] have been tried. However, the triplet and contrastive losses need a large amount of training data, as well as a careful sampling of positive and negative images, which makes the training process more complex. The center loss has been shown to be useful in this context, by requiring only a small amount of training data and also eliminating the need for the pairing of images. Based on these insights, we have employed center loss in our work with a view to handle some hard to distinguish classes such as anger and disgust. We have reported the results of our findings in Tables I and II.

Figure 4 shows the classwise accuracy of the VGG13 network (Fig. 1) trained using total loss (equation 3) and the proposed input modifications.

IV. Results and Discussion

A. Impact of the Use of Laplacian and Gradient Images

Figure 2 and Table I compare the accuracies of the model proposed in [26] and its input modified forms [16] with those after the addition of centre loss. Here O is the original VGG-13 model, SL is softmax loss (equation 2), S and L are the Sobel and Laplacian filtered images, and TL is the total loss defined in equation (3). These results suggest that the discriminative power of the network improves more with the modification of the input features than that of the loss function (see Table I).

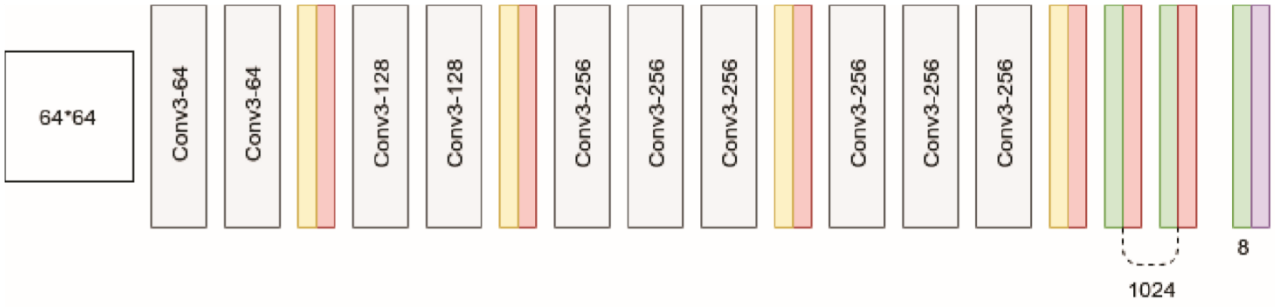

Figure 1. VGG13 network used for facial emotion classification by Barosoum et al. [14]. The colors gray, yellow, red, green and purple represent the convolution layer, max-pooling layer, dropout layer, fully connected layer and the softmax layer, respectively.

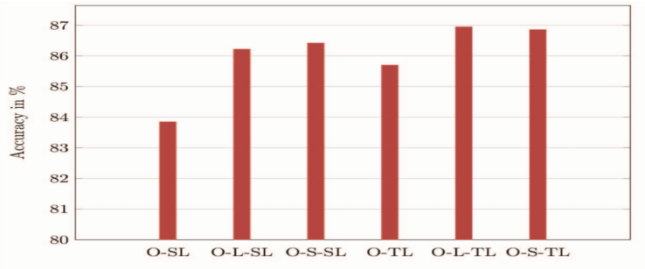

Figure 2. Accuracies of the VGG13 model [26] and its input modified forms [16] after modifying the loss function. O denotes the original model; S and L are the Sobel and Laplacian filtered images; SL is the softmax loss and TL is the total loss defined in equation (3).

Exploring the loss function, together with the input features can further add performance gain.

Table II lists the accuracies of the light weight models originally proposed by Pandey et al. [16] after the loss function is modified. We infer that in a light weight architecture, modifying the input features, rather than the loss function, may be a better option. However, the results show that modifying both may be more effective than only the "input features".

Figure 4 shows that inputting the Sobel/Laplacian filtered image along with the original image discriminates the classes better than the model trained with only the total loss given by eqn. 3 (see the accuracies of the classes "disgust" and "contempt" in Fig. 4).

Table I

Improvement of the FER accuracies of the VGG13 model with ([16]) and without ([26]) the input modifications (Sobel (S) or Laplacian(L)), on the FERPlus dataset, due to the addition of the centre loss function. SL: softmax loss; TL: total loss. The values of lambda reported are for the best results.

Models	Avg	Min	Max	Parameters	Lambda
O-SL ([26])	83.85	83.15	84.89	8.75 million	–
O-L-SL ([16])	86.22	85.94	86.56	8.75 million	–
O-S-SL ([16])	86.42	86.08	86.55	8.75 million	–
O-TL	85.70	85.24	85.97	8.75 million	0.05
O-L-TL	86.95	86.59	87.32	8.75 million	1.0
O-S-TL	86.86	86.48	87.24	8.75 million	0.1

Overall, the results reported in Tables I and II suggest that we need to search the space of models to obtain a better

Table II

Impact of modified loss function on the models proposed by Pandey et al. [16], computationally more efficient than VGG13. Results of the model on the FERPlus [26] dataset, with and without the input modifications.

Models	Accuracy %	Parameters	λ
base + SL [16]	81.95	0.36 million	-
base+L+SL [16]	84.26	0.36 million	-
base+S+SL [16]	84.47	0.36 million	-
base + TL	82.37	0.36 million	0.5
base+L+TL	83.70	0.36 million	1
base+S+TL	84.80	0.36 million	0.01

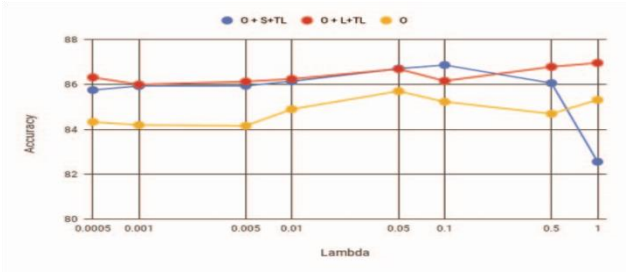

Figure 3. Variation of the accuracies of the original and the modified VGG13 networks as a function of the value of lambda. Lamda values in the x-axis are shown in log scale.

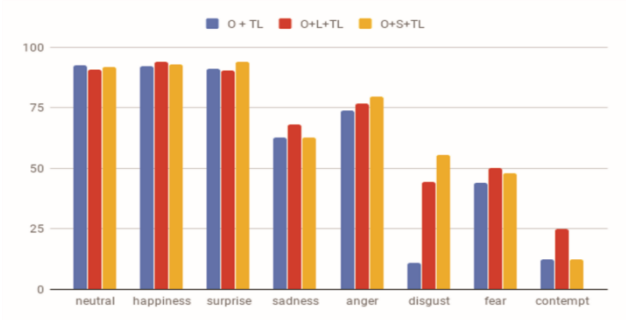

Figure 4. Classwise accuracies of the VGG13 network with the proposed modifications to the input feature and the use of total loss.

model *in terms of accuracy and/or complexity*. For example, O+SL in Table I is a more accurate classifier than base+SL in Table II, but the latter being light weight (1/24th the number of parameters), can run on low end devices easily. But once the classifier is decided, its accuracy can further be improved with the modifications proposed by us (exploring the input or the loss function or both). This suggests that to obtain the optimal classifier, we need to explore all the above possibilities.

V. Conclusion

With deep architectures, there is a large scope for feature engineering in the input image space to obtain computationally efficient models: (i) Laplacian and Sobel operators can be used to improve the discriminating power of a classifier; (ii) Search for computationally efficient, better architectures (iii) Changing the loss function can further add performance gain. We have carried out all our experiments on FERplus dataset and show that the FER accuracy can be independently improved up to 2.5% by adding better features, 1.8% by modifying the loss function and up to 3.1% by combining the two ideas. In separate experiments, we show that the computational complexity can be reduced by a factor of 24, while simultaneously increasing the FER by 0.95% by making all the above three modifications. Thus, the accuracy of any existing classifier can be improved with our proposed modifications, instead of learning a large CNN to achieve similar performance gain.

References

[1] Friesen, E. and Ekman, P., "Facial action coding system: a technique for the measurement of facial movement", Palo Alto 3, 1978.

[2] Y. Bai, L. Guo, L. Jin, Q. Huang, "A novel feature extraction method using pyramid histogram of orientation gradients for smile recognition", ICIP, 2009.

[3] C. Shan, S. Gong, and P. W. McOwan, "Facial expression recognition based on local binary patterns: a comprehensive study", Image and Vision Computing, 2009.

[4] A. Krizhevsky, I. Sutskever, GE Hinton, "ImageNet classification with deep convolutional neural networks", Advances in Neural Information Processing Systems, 1097-1105, 2012.

[5] K. Simonyan, A. Zisserman, "Very deep convolutional networks for largeScale image recognition", arXiv preprint arXiv:1409.1556, 2014.

[6] C. Szegedy, "Going deeper with convolutions", Proc. IEEE conference on computer vision and pattern recognition, 2015.

[7] K. He, X. Zhang, S.Ren, J. Sun, "IDeep residual learning for image recognition", Computer Vision and Pattern Recognition (CVPR), 2016.

[8] F. Chollet, "Xception: Deep learning with depthwise separable convolutions", arXiv preprint, 1610.02357, 2017.

[9] Pandey, Ram Krishna, and A. G. Ramakrishnan. "A hybrid approach of interpolations and CNN to obtain super-resolution." arXiv preprint arXiv:1805.09400 (2018).

[10] Pandey, Ram Krishna, Shishira R. Maiya, and A. G. Ramakrishnan. "A new approach for upscaling document images for improving their quality." 14th IEEE India Council International Conference (INDICON). IEEE, 2017.

[11] Pandey, Ram Krishna, K. Vignesh, A. G. Ramakrishnan and Chandrahasa B. "Binary document image super resolution for improved readability and OCR performance." arXiv preprint arXiv:1812.02475 (2018).

[12] Pandey, Ram Krishna, N. Saha, S. Karmakar, and A. G. Ramakrishnan, "MSCE: An edge-preserving robust loss function for improving superresolution algorithms." International Conference on Neural Information Processing. Springer, Cham, 2018.

[13] Pandey, Ram Krishna, Harpreet Singh, and A. G. Ramakrishnan. "Improvement of image denoising algorithms by preserving the edges." International Conference on Computer Analysis of Images and Patterns. Springer, Cham, 2019.

[14] E. Barsoum, C. Zhang, C. Ferrer, Z. Zhang, "Training deep detworks for facial expression recognition with crowd-sourced label distribution", Proc. 18th ACM International Conf. on Multimodal Interaction, 2016.

[15] Pandey, Ram Krishna, Souvik Karmakar, A. G. Ramakrishnan, and Nabagata Saha, "Improving facial emotion recognition systems Using gradient and Laplacian images," arXiv preprint arXiv:1902.05411, 2018.

[16] Pandey, Ram Krishna, Souvik Karmakar, A. G. Ramakrishnan, and Nabagata Saha,"Improving facial emotion recognition systems with crucial feature extractors" In 20th International Conference on Image Analysis and Processing, Trento, Italy, September, 2019.

[17] G. Levi and T. Hassner, "Emotion recognition in the wild via convolutional neural networks and mapped binary patterns", Proc. ACM International Conference on Multimodal Interaction (ICMI), Nov. 2015.

[18] S. Ouellet, "Real-time emotion recognition for gaming using deep convolutional network features", CoRR, vol. abs/1408.3750, 2014.

[19] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar and I. Matthews, "The extended Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression", IEEE Computer Society Conf. on Computer Vision and Pattern Recognition - Workshops, San Francisco, CA, 2010.

[20] H. Jung, S. Lee, J. Yim, S. Park, J. Kim, "Geometric feature-based facial expression recognition in image sequences using multi-class adaboost and support vector machines", Proc. IEEE Int. Conf. on Comput. Vis. 2015.

[21] A. Octavio, M. Valdenegro-Toro, P. Plger, "Real-time convolutional neural networks for emotion and gender classification", arXiv preprint arXiv:1710.07557, 2017.

[22] D. Ghimire, J. Lee, "Geometric feature-based facial expression recognition in image sequences using multi-class adaboost and support vector machines", Sensors, 2013.

[23] Lin, Min, Qiang Chen, and Shuicheng Yan, "Network in network". arXiv preprint arXiv:1312.4400, 2013.

[24] Jaderberg, Max, Karen Simonyan, and Andrew Zisserman, "Spatial transformer networks", Adv. in neural information process. systems, 2015.
[25] https://towardsdatascience.com/a-simple-guide-to-the-versions-of-theinception-network-7fc52b863202, last accessed: 12/12/2018.
[26] https://github.com/Microsoft/FERPlus/, last accessed: 12/12/2018.
[27] M Sandler, A Howard, M Zhu, A Zhmoginov, LC Chen, "Mobilenetv2: inverted residuals and linear bottlenecks", Proc. IEEE Conference on Computer Vision and Pattern Recognition, 4510-4520, 2018
[28] Hubel, D. H. and Wiesel, T. N, "Receptive fields, binocular interaction and functional architecture in the cat's visual cortex", J. Physiol. 160: 106-154 (1982).
[29] Lundqvist, Daniel, Anders Flykt, and Arne hman, "The Karolinska directed emotional faces (KDEF)", CD ROM from Department of Clinical Neuroscience, Psychology section, Karolinska Institutet 91: 630, 1998.
[30] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, Hartwig Adam, "Mobilenets: efficient convolutional neural networks for mobile vision applications", arXiv:1704.04861.
[31] Wen, Yandong, et al. "A discriminative feature learning approach for deep face recognition", European conference on computer vision. Springer, Cham, 2016.
[32] Hadsell, Raia, Sumit Chopra, and Yann LeCun. "Dimensionality reduction by learning an invariant mapping", IEEE Computer Society Conf. on Computer Vision and Pattern Recognition (CVPR'06). Vol. 2. IEEE, 2006.
[33] Schroff, Florian, Dmitry Kalenichenko, and James Philbin. "Facenet: A unified embedding for face recognition and clustering", Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.
[34] https://towardsdatascience.com/review-mobilenetv1-depthwiseseparable-convolution-light-weight-model-a382df364b69, last accessed: 7/6/2018