FER-Net: facial expression recognition using densely connected convolutional network

Hui Ma and Turgay Celik✉

Convolutional neural network (CNN) architectures have shown excellent image classification performance on large-scale visual recognition tasks. If a CNN architecture contains a shorter connection between layers close to the input and those close to the output, the training can be deeper, more accurate and efficient. In this Lette, the authors propose a densely connected CNN architecture for facial expression recognition (FER-Net), which connects the output of each convolution layer to the inputs of the next convolution layers in the architecture. Experiments conducted on a publicly available dataset show that FER-Net produces state-of-the-art results in facial expression recognition.

Introduction: Facial expression, as one of the most important means for humans to express emotion and connotation in the process of communication, plays an important role in the development of human–computer interaction systems. In recent years, facial expression recognition (FER) has become an important research topic due to its various applications in, such as, health care and data-driven animation [1]. The main goal of a FER system is to identify human emotional states based on a given facial image.

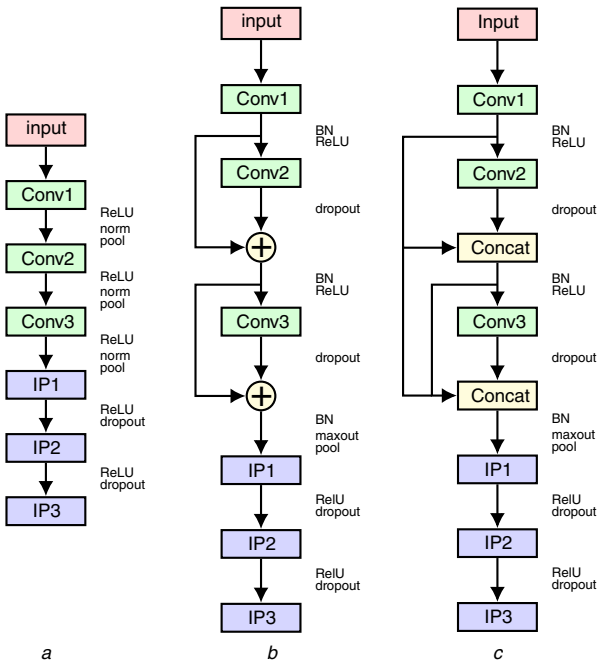
Deep convolutional neural network (DCNN) architectures have made a series of breakthroughs in image classification tasks [2, 3]. With the deepening of the network, when the input information (or the gradient from the back-propagation) goes through many layers, it can vanish and 'wash out' by the time it reaches the end (or the beginning) of the network. In order to address this problem, several DCNN architectures have been proposed. Highway Networks [4] and ResNet [5] bypass the signal from one layer to the next by identifying intermediate connections between the layers. Stochastic depth [6] shortens the ResNet by randomly dropping layers during training to get better information and gradient flow. FractalNets [7] repeatedly combines several parallel layer sequences with different numbers of convolution blocks to obtain a large nominal depth while maintaining many short paths in the network. Although these methods differ in terms of network architecture and training strategy, they all have a key characteristic: 'they create short paths from earlier layers to the later layers'. Motivated by the literature [8], we introduced densely connected DCNN for facial expression recognition (FER-Net). To support the feed-forward information flow, each convolution layer gets additional input from the all previous convolution layers and passes its own feature map to all subsequent convolution layers in the architecture.

Network architecture: The network architectures of CNN, ResNet and FER-Net are shown in Fig. 1. In order to test the performances of different methods, the networks are designed with the same architectural components as shown in Figs. 1*a*–*c*. The details of the network components are given in Fig. 1*d* where 'IP3' represents the output of the network.

The FER-Net architecture adopts the typical architectural designs of CNN and ResNet. However, in contrast to the ResNet architecture, the FER-Net architecture never combines the intermediate features from the convolution layers by summation before they are passed into the next convolution layer. Instead, it distributes features within the network by concatenating them to support the feed-forward input information flow.

As shown in Fig. 1*c*, the FER-Net architecture is designed to improve the information flow by creating direct connections from any convolution layer to all subsequent convolution layers and to the first densely connected layer. Let $\boldsymbol{x}_i \in \mathbb{R}^{h\times w\times d_i}$ and $\boldsymbol{y}_i \in \mathbb{R}^{h\times w\times d_i}$, respectively, be the input and the output feature maps at the ith layer, then the ith layer of the FER-Net receives the input from the preceding layers as follows: $\boldsymbol{x}_i = H_i(\boldsymbol{y}_{i-1}, \boldsymbol{y}_{i-2}, \ldots, \boldsymbol{y}_1)$, where $H_i(\cdot)$ is a function which simply concatenates the input feature maps $\boldsymbol{y}_{i-1}, \boldsymbol{y}_{i-2}, \ldots, \boldsymbol{y}_1$, i.e. H_i: $\mathbb{R}^{h\times w\times d_{i-1}} \times \mathbb{R}^{h\times w\times d_{i-2}} \times \cdots \times \mathbb{R}^{h\times w\times d_1} \mapsto \mathbb{R}^{h\times w\times (d_{i-1}+d_{i-2}+\cdots+d_1)}$.

The FER-Net enhances the feed-forward information flow within the network by distributing the output of each layer to the succeeding layers in the architecture. Meanwhile, the ResNet only achieves a limited information flow considering only two preceding layers of each layer, i.e. $\boldsymbol{x}_i = \boldsymbol{y}_{i-1} + \boldsymbol{y}_{i-2}$. In contrast to the ResNet and FER-Net, as shown in Fig. 1*a*, there is no feed-forward information distribution in the CNN architecture.



Layer type	filter size	stride	pad	output size
input image	48 × 48			42 × 42
Conv1	5 × 5	1	1	40 × 40 × 32
Conv2	5 × 5	1	2	40 × 40 × 32
Conv3	5 × 5	1	2	40 × 40 × 64
Pooling	3 × 3	2		20 × 20 × 128
IP1				2048
IP2				1024
IP3				7

d

Fig. 1 *DCNN architectures for facial expression recognition*
a CNN
b ResNet
c FER-Net
d Parameters of the architectures.
Legend: 'Input' (input greyscale image); 'Conv*i*' (*i*th convolution layer); 'IP*i*' (*i*th fully connected layer); 'ReLU' (rectified linear unit layer); 'Norm' (normalisation layer); 'Pool' (pooling layer); 'Dropout' (dropout layer); 'BN' (batch normalisation layer); and 'Maxout' (maximum feature map layer)

Data balancing: We utilised the FER2013 dataset [9] for FER task. The FER2013 dataset is a Kaggle challenge dataset and contains 35,887 greyscale images of size 48 × 48 collected and labelled using Google's image search API. The dataset is split into training (28,709 images), validation (3589 images) and test (3589 images) sets. There are seven classes in the dataset representing facial expressions: 'angry', 'disgust', 'fear', 'happy', 'sad', 'surprise' and 'neutral'.

the number of samples of facial expression classes before data balancing						
3995	436	4097	7215	4830	3171	4965
the number of samples of facial expression classes after data balancing						
3995	4360	4097	5215	4830	3171	4965

Fig. 2 *Sample statistics of FER2013 training dataset before and after data balancing*

The statistics of the training dataset is given in (see Fig. 2) which shows that the training dataset is highly imbalanced and learning from imbalanced dataset is a challenging task. In a typical imbalanced dataset the number of instances of some classes, called the majority classes, are significantly higher than the number of instances of the remaining classes, called the minority classes. In order to re-balance the class distribution by means of under- or over-sampling, the majority

class ‘Happy’ is randomly under-sampled to reduce its class size. Meanwhile, in this paper, the number of samples of the minority class ‘Disgust’ is increased by a linear transformation of the greyscale images in the class as follows. Considering that the greyscale range of and each image f in a minority class is in $[a, b]$, a linear transformation is applied to the pixels of f to generate a synthetic image g in range $[c, d]$, i.e. $g = (d - c)/(b - a)f + c$. Some synthetic images from the linear transformation process of a ‘Disgust’ class sample are shown in Fig. 3. The class statistics after data balancing is shown in Fig. 2.

Fig. 3 *Synthetic images (b-j) generated from a ‘Disgust’ class sample (a) using the linear transformation*

Experiments: Experiments are carried out on a desktop computer with Intel(R) Core(TM) i5-8600 K, 3.6 GHz × 6CPUs, a single GeForce GTX 1080, on Ubuntu 16.04 LTS OS. The DCNN architectures were implemented in deep learning library Caffe.

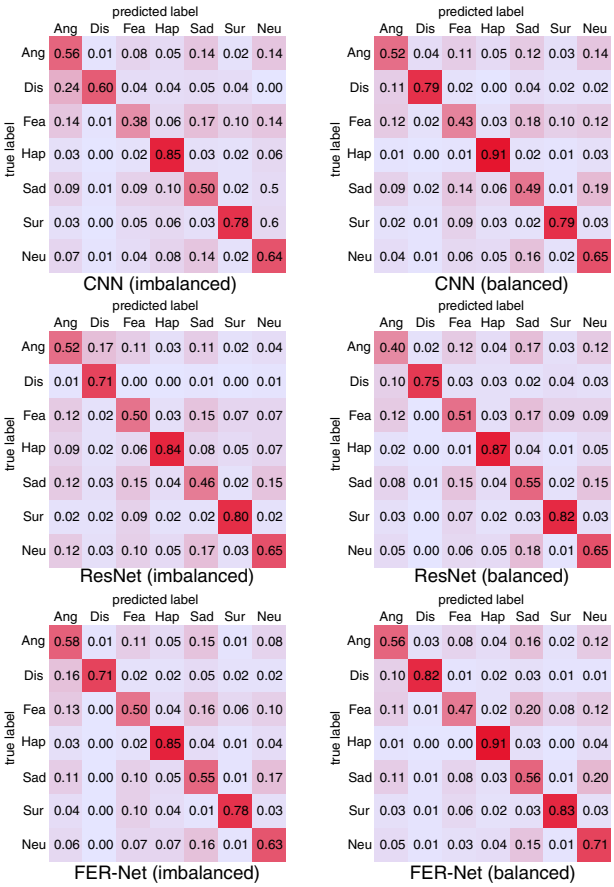


Fig. 4 *Test performances as confusion matrix of different architectures trained on original (imbalanced) and balanced training datasets. Legend: ‘Angry’ (Ang); ‘Disgust’ (Dis); ‘Fear’ (Fea); ‘Happy’ (Hap); ‘Sad’ (Sad); ‘Surprise’ (Sur); and ‘Neutral’ (Neu)*

We first compared FER-Net against to ResNet and CNN trained on the original (imbalanced) and balanced datasets. The experimental results are given in Table 1 and the corresponding confusion matrixes are shown in Fig. 4. The results show that, although CNN and ResNet have shorter average forward-backward time and smaller network size compared to FER-Net, FER-Net achieves about 3% improvement in accuracy with respect to CNN and ResNet on both imbalanced and balanced datasets. It is also clear that data balancing with a linear transformation yields about 2, 2.5 and 3% accuracy improvements for CNN, ResNet and FER-Net, respectively. The confusion matrixes show that all three networks produce the lowest or the second lowest accuracy on the class ‘Fear’ on both imbalanced and balanced datasets. The class ‘Fear’ is mostly confused with the classes ‘Angry’ and ‘Sad’ as these emotions have similar visual characteristics, which makes it difficult for the networks to discriminate between them. Meanwhile, the networks achieve highest accuracy in classifying ‘Happy’ class, which has discriminative visual features resulting from widened mouth.

Table 1: Test performances of different architectures on FER2013 dataset

Model	Train dataset	Accuracy, %	Average computing time, ms	Model size, M
CNN	imbalanced	63.50	5.89	10.8
ResNet (ReLU-Maxout)	imbalanced	64.14	25.41	61.1
FER-Net (ReLU)	imbalanced	66.54	57.59	428.4
CNN	balanced	65.68	5.61	10.8
ResNet (ReLU-Maxout)	balanced	66.51	25.68	61.1
FER-Net (ReLU)	balanced	69.51	57.59	428.4
FER-Net (Maxout)	balanced	69.01	46.29	218.4
FER-Net (Maxout-ReLU)	balanced	68.36	53.13	428.1
FER-Net (ReLU-Maxout)	balanced	69.72	50.55	428.4
Multiple Deep Network [10]	—	52.29	—	—
Net B_DAL [11]	—	58.33	—	—
Net B_DAL_MSE [11]	—	59.15	—	—
Net Net B [11]	—	60.91	—	—
Net Subnet2 [12]	—	61.58	—	—
Net Subnet1 [12]	—	61.74	—	—
Net Subnet3 [12]	—	62.44	—	—

The FER-Net employs ReLU as activation function, however, the performance of the network may change according to the type of activation function utilised. Thus, we experimented with different activation functions to test the performance of FER-Net on the balanced dataset. The experimental results are given in Table 1 which shows that the combination of ReLU-Maxout achieves the highest test accuracy. The results also show that the performance of the FER-Net does not significantly change with respect to the type of activation function used. The maximum performance change for different activation functions stays within 1% margin.

In Table 1, we also compared the performance of FER-Net against to the state-of-the-art architectures designed for facial expression recognition task on FER2013 dataset. The results show that the FER-Net outperforms all methods considered in this Letter on both imbalanced and balanced training datasets.

Conclusion: In this Letter, we introduced a FER-Net for facial expression recognition. The FER-Net densely distributes features within the network by concatenating them to support the feed-forward input information flow. By doing so it achieves the state-of-the-art accuracy in facial expression recognition task on a publicly available dataset. We also showed that data balancing using a simple linear transformation of training images results in considerable performance improvements.

© The Institution of Engineering and Technology 2019
Submitted: *27 November 2018* E-first: *21 January 2019*
doi: 10.1049/el.2018.7871
One or more of the Figures in this Letter are available in colour online.

Hui Ma and Turgay Celik (*School of Information Science and Technology, Southwest Jiaotong University, Chengdu, People’s Republic of China*)

✉ E-mail: turgay.celik@swjtu.edu.cn

References

1 Martinez, B., Valstar, M.F., Jiang, B., *et al.*: ‘Automatic analysis of facial actions: a survey’, *Trans. Affect. Comput.*, 2017, pp. 1–1, DOI: 10.1109/TAFFC.2017.2731763
2 Cai, Z., Fan, Q., Feris, R.S., *et al.*: ‘A unified multi-scale deep convolutional neural network for fast object detection’. European Conf. on Computer Vision (ECCV), Amsterdam, Netherlands, October 2016, pp. 354–370
3 Krizhevsky, A., Sutskever, I., and Hinton, G.E.: ‘ImageNet classification with deep convolutional neural networks’, *Commun. ACM*, 2017, **60**, (6), pp. 84–90

4 Srivastava, R.K., Greff, K., and Schmidhuber, J.: 'Training very deep networks'. NIPS 2015, Montreal, Canada, December 2015, pp. 2377–2385
5 He, K., Zhang, X., Ren, S, *et al*.: 'Deep residual learning for image recognition'. The IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 26 June–1 July 2016
6 Huang, G., Sun, Y., Liu, Z., *et al*.: 'Deep networks with stochastic depth'. European Conf. on Computer Vision (ECCV), Amsterdam, Netherlands, October 2016, pp. 646–661
7 Larsson, G., Maire, M., and Shakhnarovich, G.: 'Fractalnet: ultra-deep neural networks without residuals'. 5th International Conference on Learning Representations, Toulon, France, April 2017, pp. 1–11, 1605.07648
8 Huang, G., Liu, Z., Maaten, L., *et al*.: 'Densely connected convolutional networks'. The IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, July 2017
9 Goodfellow, I.J., Erhan, D., and Carrier, P.L.: 'Challenges in representation learning: a report on three machine learning contests'. Neural Information Processing Systems (NIPS), Lake Tahoe, NV, USA, December 2013, pp. 117–124
10 Yu, Z., and Zhang, C.: 'Image based static facial expression recognition with multiple deep network learning', 2015 ACM on International Conference on Multimodal Interaction, Seattle, WA, USA, November 2015, pp. 435–442
11 Zhai, Y., Liu, J., and Zeng, J.: 'Deep convolutional neural network for facial expression recognition', 9th International Conference on Image and Graphics (ICIG), Shanghai, China, September 2017, pp. 211–223
12 Liu, K., Zhang, M., and Pan, Z.: 'Facial expression recognition with CNN ensemble', 2016 International Conference on Cyberworlds (CW), Chongqing, China, September 2016, pp. 163–166