The Visual Computer
https://doi.org/10.1007/s00371-019-01630-9

ORIGINAL ARTICLE

Using CNN for facial expression recognition: a study of the effects of kernel size and number of filters on accuracy

Abhinav Agrawal[1] · **Namita Mittal**[1]

Abstract
Facial expression recognition is a challenging problem in image classification. Recently, the use of deep learning is gaining importance in image classification. This has led to increased efforts in solving the problem of facial expression recognition using convolutional neural networks (CNNs). A significant challenge in deep learning is to design a network architecture that is simple and effective. A simple architecture is fast to train and easy to implement. An effective architecture achieves good accuracy on the test data. CNN architectures are black boxes to us. VGGNet, AlexNet and Inception are well-known CNN architectures. These architectures have strongly influenced CNN model designs for new datasets. Almost all CNN models known to achieve high accuracy on facial expression recognition problem are influenced by these architectures. This work tries to overcome this limitation by using FER-2013 dataset as starting point to design new CNN models. In this work, the effect of CNN parameters namely kernel size and number of filters on the classification accuracy is investigated using FER-2013 dataset. Our major contribution is a thorough evaluation of different kernel sizes and number of filters to propose two novel CNN architectures which achieve a human-like accuracy of 65% (Goodfellow et al. in: Neural information processing, Springer, Berlin, pp 117–124, 2013) on FER-2013 dataset. These architectures can serve as a basis for standardization of the base model for the much inquired FER-2013 dataset.

Keywords Deep learning · CNN · FER-2013

1 Introduction

Convolutional networks (ConvNets) have been employed successfully for a wide range of tasks that include large-scale image classification systems, visual recognition challenges and high-dimensional shallow feature encodings. CNN-based methods are known to achieve state-of-the-art accuracy in various image classification challenges. For ICML 2013 workshop contest [1] named *Challenges in Representation Learning* later hosted on kaggle, one of the top performing entries [2] used CNN with SVM to achieve state-of-the-art accuracy of 71% on FER-2013 dataset.

VGGNet, AlexNet and Inception are well-known CNN architectures for image classification. Almost all models known to achieve high accuracy on FER-2013 dataset use VGGNet, Inception or AlexNet architectures as baseline. These architectures were originally proposed for Imagenet dataset which is a part of ILSVRC challenge. Due to its immense nature, Imagenet dataset has overshadowed the research to design new CNN models for other datasets. To the best of our knowledge, no independent study is made on the effects of parameters namely kernel size and number of filters on the accuracy of classification taking FER-2013 as dataset as the starting point.

This work makes an attempt to arrive at a CNN architecture specifically for FER-2013 by making an extensive study of variation of accuracy with CNN parameters.

In this study, a bottom-up approach is taken to design a CNN network by studying characteristics of a constituent layer (see Table 1). We have used the term constituent layer to refer to a layer which acts as a building block of our deeper network. This constituent layer is studied for various combinations of kernel size and number of filters. Combinations which show good convergence are used for constructing deeper network.

As a result of this study, we propose two network architectures referred to as Model1 and Model2 (see Tables 2, 3),

✉ Abhinav Agrawal
abhinav0653@gmail.com

[1] Department of CSE, MNIT, Jaipur 302017, India

Published online: 23 January 2019

Table 1 Constituent layer

CONV k_sz × k_sz × num_filters, BATCH NORM
CONV k_sz × k_sz × 7, RELU, STRIDE (128 × 128)
SOFTMAX

Table 2 Architecture of Model1

Input data (64 × 64) grayscale image
Data augmentation
CONV 8 × 8 × 32,BATCH NORM
CONV 8 × 8 × 32, RELU, STRIDE (2 × 2)
CONV 8 × 8 × 32, BATCH NORM
CONV 8 × 8 × 32, RELU, STRIDE (2 × 2)
CONV 8 × 8 × 32, BATCH NORM
CONV 8 × 8 × 32, RELU, STRIDE (2 × 2)
CONV 8 × 8 × 32, BATCH NORM
CONV 8 × 8 × 32, RELU, STRIDE (2 × 2)
CONV 8 × 8 × 32, BATCH NORM
CONV 8 × 8 × 32, RELU, STRIDE (2 × 2)
CONV 8 × 8 × 32, BATCH NORM
CONV 8 × 8 × 32, RELU, STRIDE (2 × 2)
CONV 8 × 8 × 32, BATCH NORM
CONV 8 × 8 × 32, RELU, STRIDE (2 × 2)
CONV 8 × 8 × 32, BATCH NORM
CONV 7 × 7 × 7, RELU, STRIDE (1 × 1)
SOFTMAX

respectively, both of which achieve human-like accuracy on FER-2013 dataset. Model2 is a reduced variant of Model1.

The architecture of Model1 is unique in the fact that it not only uses fixed kernel size, but also fixes the number of filters across the depth of network. In all the popular CNN architectures like VGGNet and AlexNet, the number of filters increases with depth. The architecture of Model2 is derived from Model1. In this architecture, the number of filters decreases with network depth. Model2 is more compact than Model1. Both architectures use kernel size of 8 which is unique. This kernel size is derived as a result of the study.

Both Model1 and Model2 have very less number of training parameters (see Table 4) as compared with VGGNet. Moreover, both models don't use dropout to improve accuracy. Comparison of proposed models with state-of-the-art (see Table 5) shows that the architecture of the proposed models is the most suitable for FER-2013 dataset.

2 Related work

Krizhevsky [3] in 2012 trained a CNN architecture, popularly known as AlexNet, on GPU to achieve good accuracy on the ILSVRC-2010 dataset. After that, there has been increasing research to improve the original architecture of Krizhevsky to achieve better accuracies on Imagenet dataset.

Table 3 Architecture of Model2

Input data (64 × 64) grayscale image
Data augmentation
CONV 8 × 8 × 32,BATCH NORM
CONV 8 × 8 × 32, RELU, STRIDE (2 × 2)
CONV 8 × 8 × 32, BATCH NORM
CONV 8 × 8 × 32, RELU, STRIDE (2 × 2)
CONV 8 × 8 × 32, BATCH NORM
CONV 8 × 8 × 32, RELU, STRIDE (2 × 2)
CONV 8 × 8 × 16, BATCH NORM
CONV 8 × 8 × 16, RELU, STRIDE (2 × 2)
CONV 8 × 8 × 16, BATCH NORM
CONV 8 × 8 × 16, RELU, STRIDE (2 × 2)
CONV 8 × 8 × 16, BATCH NORM
CONV 8 × 8 × 16, RELU, STRIDE (2 × 2)
CONV 8 × 8 × 8, BATCH NORM
CONV 8 × 8 × 8, RELU, STRIDE (2 × 2)
CONV 8 × 8 × 8, BATCH NORM
CONV 7 × 7 × 7, RELU, STRIDE (1 × 1)
SOFTMAX

The original model by Krizhevsky had convolution layers with kernel sizes 11, 5 and 3 followed by fully connected layers. This architecture was a unique combination of convolution layers followed by max-pool operations. AlexNet architecture also involved use of response normalization for the first few layers. Although the network achieved good accuracies, the choice of its parameters was empirical which left a scope for further optimization of the network.

With the success of AlexNet, many experiments were done to improve its performance [4]. Simonyan and Zisserman [5] from Oxford University did extensive experiments on the effect of network depth on classification accuracy for the Imagenet dataset. In order to determine the optimal depth, authors fixed the kernel size to 3 and increased the depth of network. Filters were increased steadily in multiples of 2 along the depth. They demonstrated the effectiveness of network with 19 weight layers out of which 16 were convolution and rest were fully connected layers. This network was called VGG19. This established use of deeper networks for achieving better accuracy. VGGNet has gained widespread acceptance due to its simplicity.

For facial expression recognition, many of the proposed CNN architectures are deeply influenced by research on Imagenet dataset. Wan et al. [6] documented their efforts to apply VGGNet and AlexNet to FER-2013 dataset. They selectively extracted layers from VGGNet to come up with their own

Table 4 Comparison of proposed models with VGGNet

Property	Model1	Model2	VGGNet
Input shape	64 × 64	64 × 64	64 × 64
Weight layers	17	17	19
Conv layers	16	16	16
Filter size	Fixed	Variable	Variable
Kernel size	Fixed to 8	Fixed to 8	Fixed to 3
Training params	931,527	464,183	20,037,575
Model size	7.6 MB	3.8 MB	160 MB

Table 5 Comparison of proposed models with state-of-the-art

Model	Acc (%)	Params	Baseline architecture
Proposed Model2	65.23	0.46M	Original
Proposed Model1	65.77	0.93M	Original
Sang et al. [15]	71	4.92M	VGGNet
Tang [2]	69.3	7.17M	AlexNet
Wan et al. [6]	65.34	14M	AlexNet+VGGNet
Liu (subnet 1) [10]	61.74	84M	VGGNet

model for FER-2013 dataset. The prominent feature of their network was the use of dropout. Using dropout and reduced architecture, they were able to achieve state-of-the-art accuracy on FER-2013.

Arriaga et al. [7] considered both VGGNet and Inception architectures in their work to derive a model for FER-2013 dataset. By using these architectures, authors were able to design a reduced model to achieve state-of-the-art accuracy on FER-2013 dataset. Their objective was to improve accuracy while keeping the number of training parameters low. They found that 90% of the training parameters for a CNN come from fully connected layers. So, they removed fully connected layers and were able to achieve good accuracy on FER-2013 dataset while keeping model size very low.

Al-Shabi et al. [8] in their work titled *Facial expression using hybrid-SIFT aggregator* used CNN for FER-2013 dataset. In this work, authors started with an aim to reduce the training samples and simultaneously achieve good accuracy. CNNs require a large number of training samples. On the other hand, handcrafted feature extraction techniques like local binary pattern feature extractor with SVM classification, HOG, Haar, SIFT, Gabor filters with fisher linear discriminant, and Local phase quantization (LPQ) and their combinations give good results with less number of training samples. This work combined the best of both approaches by using SIFT transform on images and adding its output to fully connected dense layer of CNN to achieve good results on FER-2013 dataset.

Attempts have also been made by Gogic et al. [9] to address the tradeoff between accuracy and speed of a neural network-based FER classifier on multiple facial expression datasets (CK+, MMI, JAFFE, and SFEW 2.0). In their paper titled *Fast facial expression recognition using local binary features and shallow neural networks*, authors have used shallow neural networks and feature vectors based on localized binary patterns to achieve fast and accurate network.

It is important to note that although manual feature selection-based techniques give nearly hundred percent accuracies on datasets like CK+ and JAFFE, they are not directly comparable with purely CNN-based methods. FER-2013 dataset is specifically targetted for CNN-based facial expression recognition and differs from other datasets like CK+, JAFFE and SFEW in the number of images. FER-2013 has more than thirty thousand images, while CK+, JAFFE and SFEW have less than a few thousand images each.

Liu et al. [10] tried to apply CNN to the problem of facial expression recognition in their work *Facial expression recognition using CNN ensemble*. In their work, authors have proposed three different convNet architectures, each of which is able to perform well for some particular emotion as compared with others. By using an ensemble of these convNets, they are able to achieve an accuracy of 65% on FER-2013 dataset.

Shin et al. [11] tried to get an ideal CNN architecture for facial expression recognition in the paper titled *Baseline CNN architecture for facial expression recognition*. In this work, authors try to derive baseline architecture for facial expression recognition by using the best models available till date and observing the common features in them.

A survey paper by Li and Deng [12] on deep learning-based facial expression shows that the highest test accuracies on FER-2013 dataset using single CNN network are in a range of 67–71%. Any accuracy in this range is considered to be decent, and architectures achieving it are high performers.

Fine tuning of existing CNN architectures and adapting them to FER-2013 is visible in works discussed above. This work attempts to make an independent study of the effect of network parameters on classification accuracy using FER-2013 dataset. We have made a study of variation of the accuracy of the constituent layer with changes in network parameters using FER-2013 dataset. The layers which give better convergence are selected for designing deeper network.

Next section shows the study of variation of test accuracy of the constituent layer for various combinations of kernel sizes and the number of filters.

3 Study

For this study, the following points are taken into consideration.

1. FER-2013 dataset is used for this study. It is split into training and testing samples with split ratio 80:20. Out of 35,887 images, 28,709 images are used for training and 7178 for testing.
2. Accuracy in this text refers to testing accuracy on 7178 samples which are not a part of training.
3. Input data are normalized by dividing it with 255.
4. Input shape is fixed to 64×64. FER-2013 dataset has images with resolution 48×48. Thus, each image in the dataset is upscaled to 64×64 for this study.
5. Experiments are conducted on constituent layer (see Table 1). It consists of two convolution layers stacked atop each other. The second convolution layer is a substitution for max-pool operation. It is introduced to keep network simple [13].
6. Accuracy is measured with variation in kernel size (k_sz) and the number of filters (num_filters). They are varied in multiples of two. The values considered are:
   - Kernel size: 2, 4, 8, 16, 32, 64
   - Number of filters: 2, 4, 8, 16, 32, 64, 128, 256
7. For each kernel size mentioned above, all the mentioned number of filters are analyzed. For notational convenience in graphs, a combination is represented as 32_8. It implies that the number of filters is 32 and kernel size is 8.
8. ADAM is chosen as the default optimizer for this study. For initial training of deep learning networks, ADAM is the best overall choice [14].
9. While designing deeper networks the number of filters and kernel size are kept same for all layers. This differs from VGGNet and other popular architectures where either one or both are variable.
10. Training of network is done on NVIDIA 940MX GPU using Keras and Tensorflow.

Following subsections discuss the systematic progress of the study.

3.1 Variation of constituent layer accuracy with kernel size and number of filters

Figure 1a–e shows the effect of number of filters and kernel size on the test accuracy for FER-2013 dataset using constituent layer (see Table 1). From the figures, it is clear that constituent layer saturates at an accuracy within the range of 25%–30% for all cases. It is due to the low depth of network.

It is also visible that accuracy plot is highly unstable for very low and very high kernel sizes. Very low kernel size of 2 (Fig. 1a) leads to a very unstable network. Very high kernel sizes like 32 (Fig. 1d) and 64 (Fig. 1e) do not converge at all for some combinations of network parameters. Values of kernel sizes from 8 to 16 yield good convergence. Further analyzing the results it is found that the following combinations of network parameters give the best convergence while achieving good accuracy.

- Number of filters: 16, Kernel size:4
- Number of filters: 32, Kernel size:8
- Number of filters: 4, Kernel size:16
- Number of filters: 2, Kernel size:32

3.2 Variation of network accuracy with network depth

In the last section, a subset of network parameters was found which gave good convergence and less loss. By conducting experiments on the depth of network, it was further found that the network with 32 filters and kernel size 8 showed notable improvement in accuracy for FER-2013 dataset with the increase in depth. The network constructed using this combination is shown in Table 2.

Figure 2 shows the depth analysis on a network with 32 filters and kernel size of 8. From Fig. 2, it is clear that network accuracy is improving with the increase in the depth of network. Network with 16 convolution layers gives the best accuracy. This network is chosen for further analysis in the next section where a study is made on the effect of optimizer change on this network.

3.3 Variation of network accuracy with optimizer change

Figure 2b shows the effect of using different optimizers on the network derived in the last section. It is visible that SGD

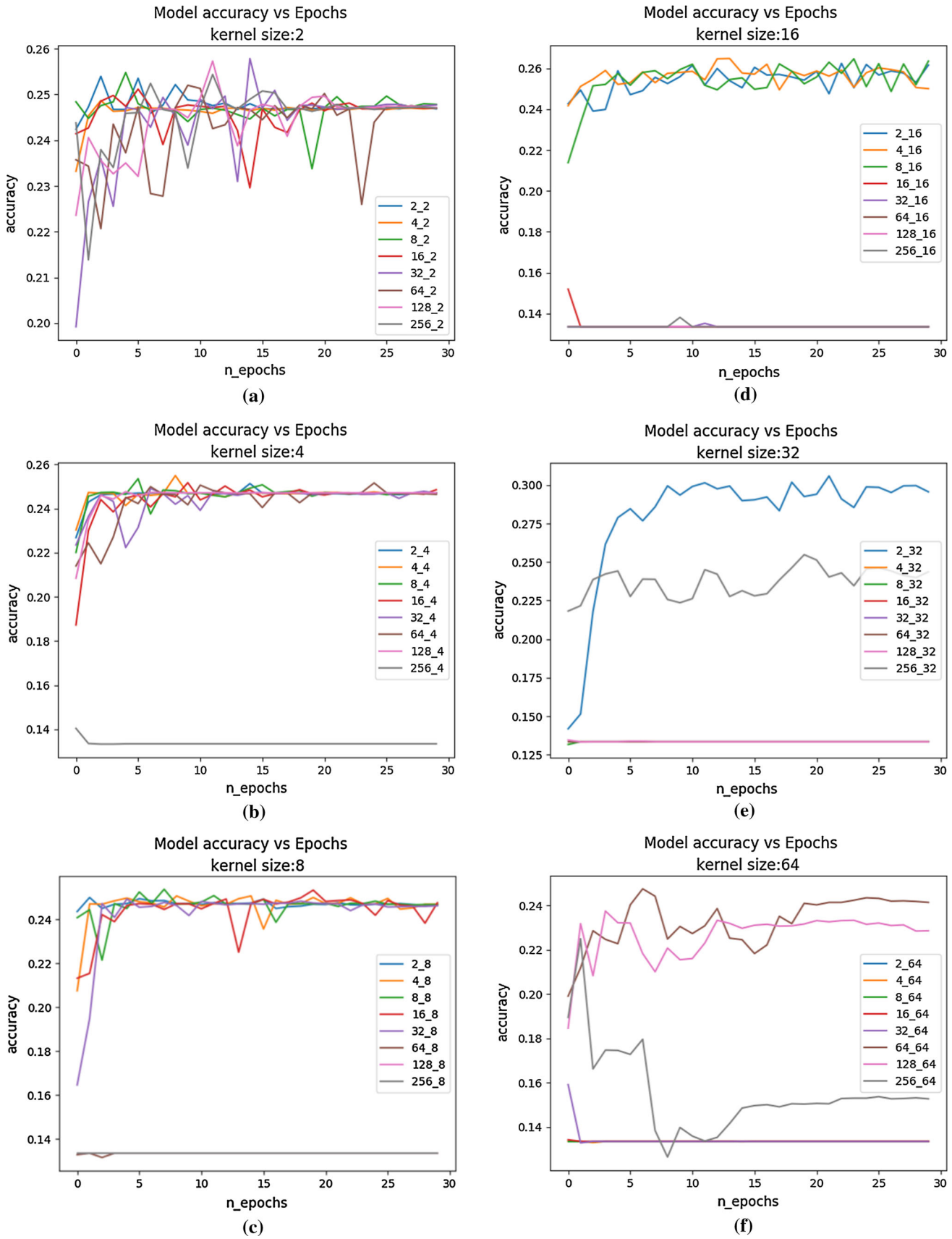
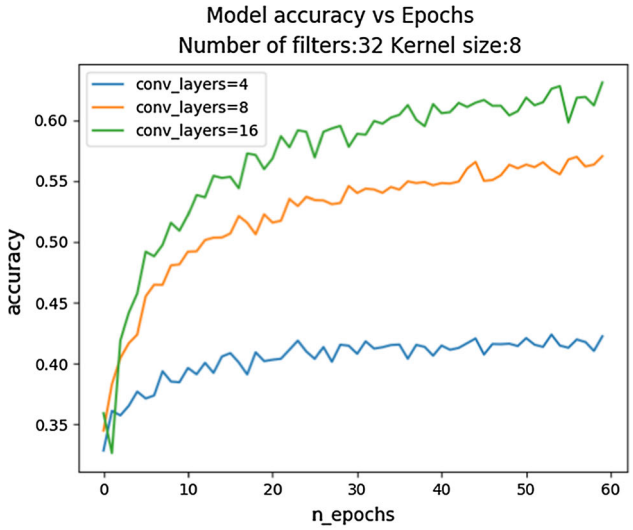
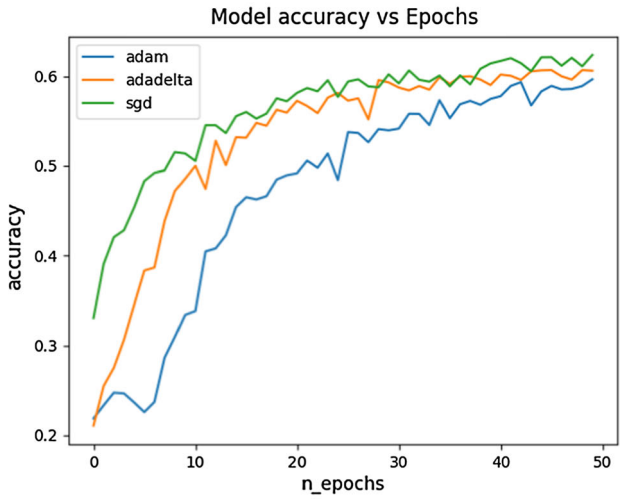


Fig. 1 Variation of model accuracy with kernel size and number of filters

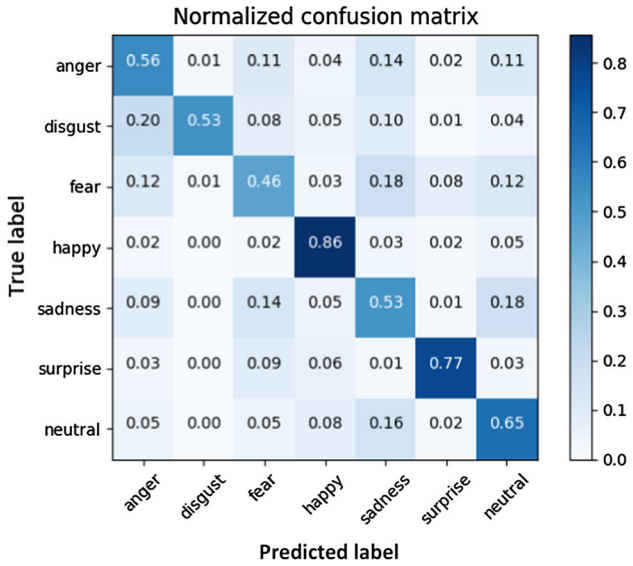


(a) Variation of accuracy with network depth

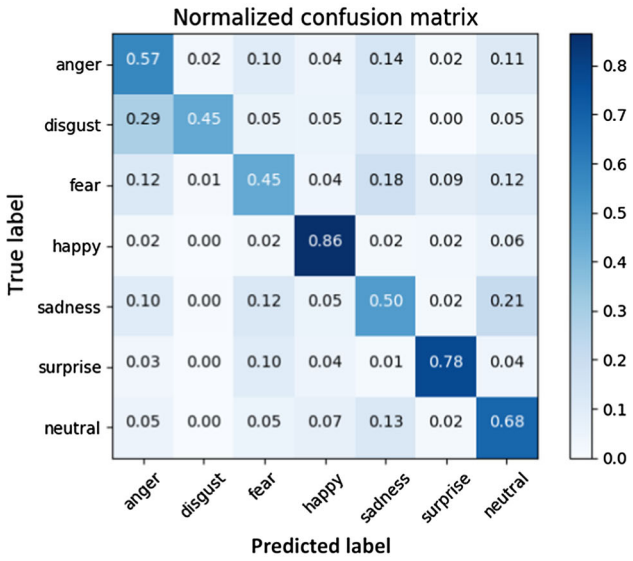


(b) Effect of optimizer on accuracy

Fig. 2 **a** Variation of accuracy with network depth, **b** Effect of optimizer on accuracy



(a) Confusion Matrix: Model1



(b) Confusion Matrix: Model2

Fig. 3 Confusion matrices for Model1 and Model2

optimizer gives a slight improvement in accuracy as compared to default optimizer ADAM that was chosen at the beginning of the study. We call the architecture obtained at this stage as Model1. It is shown in Table 2.

3.4 Reduced model

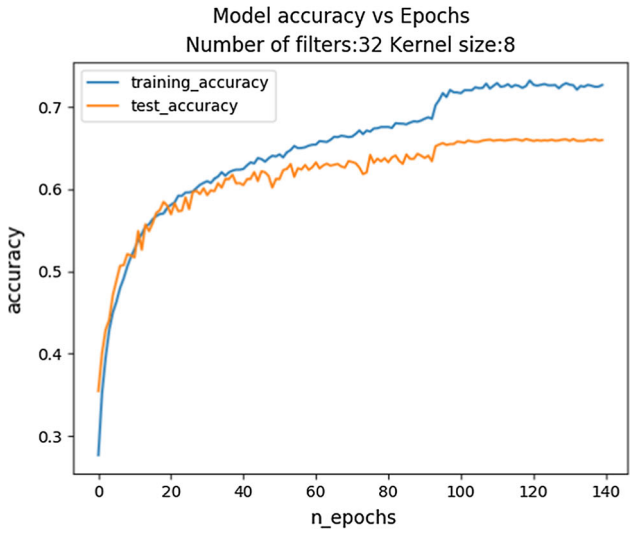
To derive Model2 from Model1, further experiments were done by us with an objective of reducing model complexity. The result was an architecture referred to as Model2. It is shown in Table 3. It is a reduced variant of Model1 because it uses less number of filters in deeper layers as a consequence of which model size and training time is reduced.

4 Proposed models

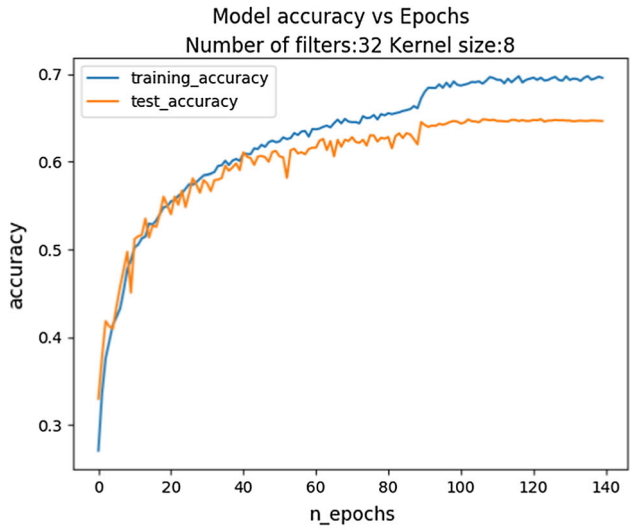
Based on our study, we propose two CNN architectures namely Model1 and Model2 which are shown in Tables 2 and 3, respectively. These architectures are specifically suitable for FER-2013 dataset because of their simplicity and less number of training parameters.

5 Results

Figure 3 a, b shows confusion matrix of normalized classification accuracies achieved by Model1 and Model2, respec-



(a) Accuracy variation: Model1



(b) Accuracy variation: Model2

Fig. 4 Accuracy variation for Model1 and Model2

tively, on FER-2013 dataset. From confusion matrices, it is observed that both models are able to classify emotions of surprise and happiness with a much higher accuracy as compared with other emotions which humans also find difficult to classify.

Figure 4a, b shows the variation of accuracy with epochs for Model1 and Model2, respectively. Both the proposed models achieve an accuracy better than 65% on FER-2013 dataset. This accuracy is at par with the human accuracy achieved on this dataset as mentioned by Goodfellow et al. [1].

Table 4 shows the comparison of proposed models with VGG19. It is visible that the proposed models have very less number of training parameters as compared with VGG19. Further, Table 5 shows the comparison of proposed models with state-of-the-art models which use regular convolution layers for FER-2013 dataset. The proposed models achieve high accuracy on FER-2013 dataset while maintaining a low model complexity by keeping the number of training parameters low. In particular, proposed Model2 has the least number of training parameters among all other models and is still able to achieve human-like accuracy (65%) on the FER-2013 dataset, making it the best suited model for the dataset.

It can also be noted that Model1 and Model2 have architectural differences (see Tables 2, 3). Model1 is uniform and has the same value for kernel size and number of filters across all layers. Model2 is non-uniform and has variation in the number of filters across depth. Model1 is especially suitable for hardware implementation due to its uniform architecture.

Proposed models are unique in their architectures. Compared with any other model proposed for FER-2013 dataset, the proposed models:-

- Do not use dropout.
- Do not have fully connected layers.
- Use same kernel size 8 throughout the network.
- Have a unique architecture which is derived as a result of the study on the constituent layer.

6 Conclusions

In this paper, we have presented two novel CNN architectures based on the study of FER-2013 dataset. These architectures are not only simple but also unique in terms of the selection of hyper-parameters across network layers. Also, by our study, we have shown that kernel size and the number of filters have a significant impact on the accuracy of the network. This study can be further extended to systematize CNN model design by developing a relationship between these parameters and accuracy of the network.

Compliance with ethical standards

Conflict of interest Both the authors declares that they have no conflict of interest.

References

1. Goodfellow, I.J., Erhan, D., Carrier, P.L., Courville, A., Mirza, M., Hamner, B., Zhou, Y.: Challenges in representation learning: a report on three machine learning contests. In: Lee, M., Hirose, A., Hou, Z. G., Kil R.M. (eds.) Neural Information Processing, ICONIP 2013. Lecture Notes in Computer Science, vol. 8228, Springer, Berlin, Heidelberg (2013)
2. Tang, Y.: Deep learning using support vector machines. CoRR, arXiv:1306.0239 (2013)

3. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: NIPS, pp. 1106–1114 (2012)
4. Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., Liu, T., Wang, X., Wang, G.: Recent advances in convolutional neural networks. arXiv preprint arXiv:1512.07108 (2015)
5. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arxiv:cs/arXiv:1409.1556 (2014)
6. Wan, W., Yang, C., Li, Y.: Facial Expression Recognition Using Convolutional Neural Network. A Case Study of the Relationship Between Dataset Characteristics and Network Performance. Stanford University Reports, Stanford (2016)
7. Arriaga, O., Valdenegro-Toro, M., Plger, P.: Real-Time Convolutional Neural Networks for Emotion and Gender Classification. Preprint. arXiv:1710.07557 (2017)
8. Al-Shabi, M., Cheah, W.P., Connie, T.: Facial expression recognition using a hybrid CNN-SIFT aggregator. arXiv preprint arXiv:1608.02833 (2016)
9. Gogic, I., Manhart, M., Pandzix, I.S., et al.: Fast facial expression recognition using local binary features and shallow neural networks. Vis Comput. https://doi.org/10.1007/s00371-018-1585-8
10. Liu, K., Zhang, M., Pan, Z.: Facial expression recognition with CNN ensemble. In: International Conference on Cyberworlds IEEE, pp. 163–166 (2016)
11. Shin, M., Kim, M., Kwon, D.-S.: Baseline CNN structure analysis for facial expression recognition. In: 2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN). IEEE (2016)
12. Li, S., Deng, W.: Deep Facial Expression Recognition: A Survey arXiv:1804.08348 (2018)
13. Springenberg, J.T., Dosovitskiy, A., Brox, T., Riedmiller, M.: Striving for Simplicity: The all Convolutional Net. arXiv preprint arXiv:1412.6806 (2014)
14. Ruder, S.: An Overview of Gradient Descent Optimization Algorithms. arXiv:1609.04747 (2016)
15. Sang, D.V., Dat, N.V., Thuan, D.P.: Facial expression recognition using deep convolutional neural networks. In: 9th International Conference on Knowledge and Systems Engineering (KSE) (2017)

Abhinav Agrawal is an M.tech student at Department of Computer Science and Engineering, Malaviya National Institute of Technology, Jaipur, India. He holds a B.tech in Electronics and Communication Engineering from Indian Institute of Technology, Dhanbad. His current research interests are deep learning, convolution neural networks, image processing and computer vision.

Dr. Namita Mittal is an Associate Professor at Department of CSE, Malaviya National Institute of Technology, Jaipur, India. She is a recipient of Career Award for Young Teachers (CAYT) by AICTE. Her current research areas are DBMS, Information Retrieval, Data mining and NLP. She has published several research papers in reputed international conferences and journals. She is also a member of review committees for Refereed journals/ conferences. She is SMIEEE, Member of ACM, CCICI, and SCRS.