Facial Emotion recognition with convolutional Neural Networks

Maryam Honari Jahromi[1] and **Alona Fyshe**[2*]

Abstract

Artificial intelligence has made significant advances toward computationally smarter systems during last decades. However, there is a need to create technologies which can track human emotions and respond to them. In this study we tried to recognize facial expressions using convolutional neural networks(CNN). We explored three neural network architectures and achieved accuracy of 68.2% on FER2013 dataset. We also found out single CNN shows a promising performance without aid of sever preprocessing of input images. We also developed an online application to predict emotions in real-time.Overall the result demonstrate feasibility of the CNNs on static and dynamic facial expression recognition tasks.

1. Introduction

Emotions play an important role in social communications among human beings. While humans use facial expression to recognize emotions without difficulty, it is not a trivial task for computers. Having machines which understand human emotions can facilitate interactions of human and computers. Another possible application of this is surveillance and behavioral analysis systems. Take fatigue detection of drivers as an example.
There are two main approaches toward answering this question. Some facial expression systems use appearance-based methods to model person's emotion. These model work based on facial shape and texture. Other methods presume emotions as a combination of movements in different regions of the face such as around eyes, nose and mouth. The entire facial movements are described in a psychological framework named Facial Action Coding System (FACS). This framework enumerates 46 atomic facial action units (FAUs) based on muscle movements and defines facial expressions as a combination of several FAUs[4]. Our aim for this study is to find out **How well a simple convolutional neural network works on facial recognition task without the aid of comprehensive image preprocessing or other complicated models?** It is a crucial question if we want to use such models in online environments where it needs to perform fast on computationally restricted devices.

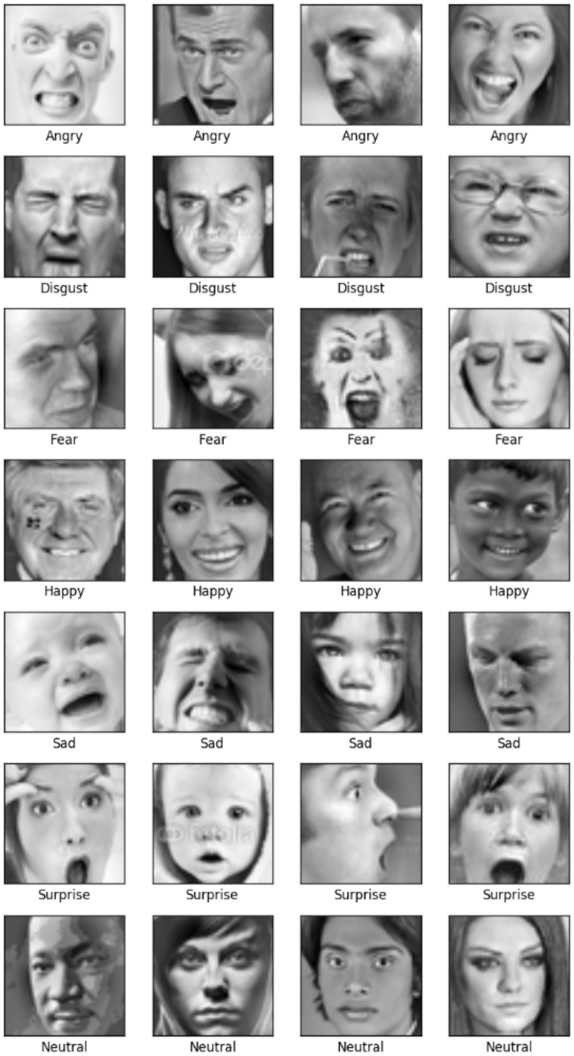


Figure 1. samples of FER2013 dataset and their corresponding emotion depicting variations of pose, illumination, occlusion and age.

*Language Learning Lab, Department of Computer Science, University of Victoria. [1]mhonari@uvic.ca [2]afyshe@uvic.ca

2. Related Work

Historically, appearance-based approaches of Facial Expression Recognition (FER) used hand-crafted methods like boosted LBP features[19],Haar features[15] and Gabor wavelets[1] to extract features from images. However, over the last couple of years, variation of deep learning networks achieved promissing results on both dynamic video-based[2] and static image-based FER[18][10].

Levi et. al [12] showed a significatn amount of improvement in static FER in the wild by applying Local Binary Pattern (LBP) to transform images to an illumination invarient, 3D space. They claim this transformation address the lack of enough labeled data for emotion classification tasks. They applied LBP codes on CASIA webface dataset [16] which was used to train an ensemble of CNN models. The models in this ensemble resembled the previous successful models in imagenet challenge like VGGnet. [14] Their final model shows a substantial accuracy up to 54.56% on SFEW dataset, an improvement of 15% compared to previous approaches.

Mollahosseini et al. [13] made an effort to fill the generalization gap of FER. They found most of existing approaches to be dependent on the conditions of a single dataset and unable to get promising results on unseen datasets. They proposed a Deep CNN with inception layers that gives competitive results on most of publicly available datasets.

Face registration is the process of finding facial landmarks(eg. eyes, nose and lip corners) and rotate image to align these landmarks to specific locations in the final image. However, in real world conditions, we may encounter non-alignable faces due to failure to detect facial landmarks. Kim et al. [10] found it helpful to do face registration only when it is possible to detect facial landmarks and keep both alignable and non-alignable faces in the training and testing dataset. They also proposed alignment-mapping networks (AMNs) to learn face alignment operation.

Also, Khorrami et al. [9] achieved state of art results on two standard expression recognition datasets (Cohn-Kanade dataset [7] : 88% and Toronto Face Dataset: 96%) using zero-bias CNN trained from scratch. They also showed that neural networks are capable of learning high-level features that highly matches to what we refer to as Facial Action units (FAU).

Recently Fan et al. proposed a novel approach to emotion recognition in videos [5]. They use both video and audio information to do classification task Enabling them to beat EmotiW 2016 competition winner and achieve state-of-the-art performance 55.3% on AFEW dataset. Their architecture is a CNN-LSTM network that utilize both spatial and temporal features of data.

3. Our Approach

In this section we describe how we divided and preprocessed our dataset. Then we explain details of three chosen networks based on, but not necessarily equal to previous researches. In section 4 we evaluate and compare obtained results of all three models.

3.1. Dataset

We trained and tested our models on Kaggle's facial expression challenge dataset[6]. This is a comparatively large dataset of low resolution images. It contains 35887, 48 by 48 grayscale images displaying 7 emotion categories: 'anger', 'disgust','fear','happiness','sadness', 'surprise' and 'neutral' as baseline. As illustrated in Figure 1 faces vary in age, pose, illumination and expression intensity which resembles wild environment. Images were collected by Google image search queries and labeled based on surrounding text. Although there might be some mislabeled images, it doesn't make classification task harder. According to authors, human accuracy on FER2013 is around 65%. Data set was divided to training, validation and test set of 28789, 3589 and 3589 samples respectively. Data distribution is shown in Figure 2

3.2. Data Preprocessing

We normalized images by removing each pixel mean and dividing its value by standard deviation across whole training dataset. Thus pixel values changed to be in the range of [-1, 1] having zero mean. We took advantage of histogram equalization to improve image contrast and illumination. This method distribute frequency of pixels uniformly across the possible range of pixel values. In order to augment data we randomly flip images horizontally or shift them by 10% of their width or height. We did data augment ion on the fly for each batch of training data. To be fair, we did the same preprocessing on test and validation set, Otherwise those sets would be easier to predict compared to training set.

3.3. Networks

We tried to compare different network architectures in terms of depth. [8].

- Our first model was a shallow CNN consists of two convolutional and two fully connected layers as shown in Figure 3. Each convolutional layer (C) uses kernels of size 3X3 with stride 1 and followed by a max-pooling (MP) layer. Batch normalization (BN) after each convolution layer and drop out (D) after first fully connected (FC) layer was applied to reduce chance of overfitting. We used Rectified Linear Unit(ReLU) as activation function and Softmax as our loss function.

- Our next experiment was based on a network similar to what Kim et.al used [10]. This network has three

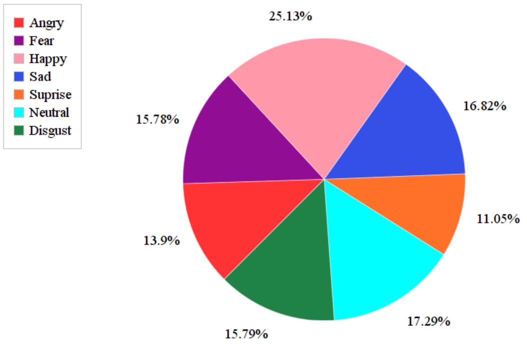


Figure 2. Distribution of emotion categories in FER 2013 dataset.

convolutional layers, one more than the previous one. The structure is as follows: 32C5X5 → MP3X3 → BN → 32C4X4 → MP3X3 → BN → 64C5X5 → MP3X3 → BN → 1024FC → D → 7FC. In Figure 7, we showed the input/output volume of each layer. Although this model is deeper compared to our shallow network but has less parameters Since the activation volume is decreased with another CMP layer before arriving to fully connected layers.

- To explore the impact of depth on accuracy of FER task, we implemented an architecture similar to VGG-B where last CC-MP block was deleted. [14] To avoid overfitting we add drop out and batch normalization after each CC-Mp block. Also drop out and L2-regularization was applied to the fully connected layers.

3.4. Training

The networks were implemented using Keras framework with tensorflow backend and trained on a GTX 860M GPU with 4 GB of video memory. We used adam optimizer with 0.01 as initial learning rate and we half the learning rate, if the validation accuracy doesn't improve after 25 epochs. All networks were trained for at least 60 epochs. Those networks which showed improving accuracy were trained for more than 60 epochs until they terminated due to early stopping. The source code and other scripts can be found on https://github.com/mahon94/empathy

4. Results and comparison

The performance of networks from previous section is given in Table 1. The evaluated accuracy on FER2013 test set shows that deep models can achieve higher performance. However, deeper doesn't imply more parameters. The depth of medium network increased compared to the shallow one, but the number of parameters decreased. This indicates that we can achieve high accuracy with reasonable number of parameters.

Table 1. Network differences in terms of depth and number of parameters. Test accuracy of networks on FER2013 dataset increases by A: data augmentation and depth increase.

Name	Depth	Parameters	Accuracy
Shallow	2	4.7 m	48.06%
Medium	3	2.4 m	56.42%
Medium + A	3	2.4 m	63.13%
Deep	11	9.4 m	58.98%
Deep + A	11	9.4 m	68.23%

Another method that improved performance about 10% is data augmentation, as demonstrated in Table 2. Data augmentation helps network to become invariant to shifts and generalize better. This observation indicates that larger datasets are needed for FER task.

Table 2. Medium Network accuracy on FER2013 dataset based on various preprocessing. A: data augmentation N: normalization HE: histogram equalization

Name	Val. Accuracy	Test Accuracy
Medium+N	55.8%	56.42%
Medium+N+A	61.0%	63.1%
Medium+N+A+HE	58.1%	60.1%

To compare accuracy on different emotions we illustrated confusion matrix of medium model in Figure 4. It is obvious that our model preforms much better on positive emotions with accuracy of 81% on happy, 73% on surprise and 72% on neutral. The most problematic class is 'Disgust' which mostly misclassified as angry. Figure 8 shows some misclassified samples for each emotion categories. Surprisingly some of the images are confusing and ambiguous for human.Perhaps this speaks for the difficulty of recognizing expressions from discrete emotions.

5. Layer visualization

We applied layer visualization on last convolutional layer of the medium network based on methods discussed by Yosinski et al. [11][17] We start optimizing on a noise image and the result Figure 5 exhibits some of best filters. Images with higher loss value on selected layer are assumed to be better looking.

Probably training on a more heterogeneous dataset with higher resolution images, gives more perceivable visualizations. It worth mentioning that we didn't used any training data visualize layers, but we can guess our network were trained on faces only based on weights which can arises security concerns.

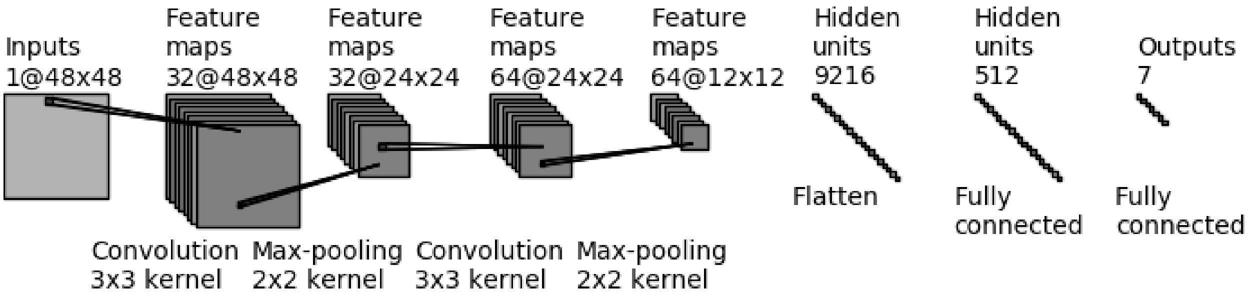
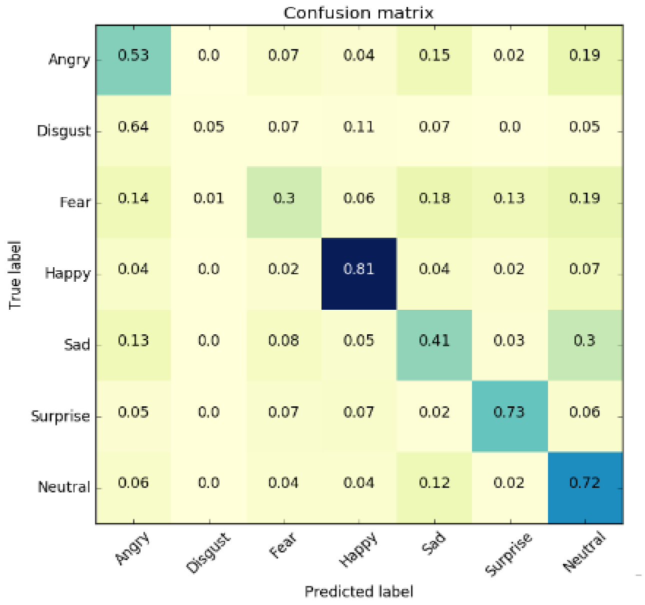


Figure 3. The architecture of shallow network: 2 convolutional layers followed by two fully connected layers



	Angry	Disgust	Fear	Happy	Sad	Surprise	Neutral
Angry	0.53	0.0	0.07	0.04	0.15	0.02	0.19
Disgust	0.64	0.05	0.07	0.11	0.07	0.0	0.05
Fear	0.14	0.01	0.3	0.06	0.18	0.13	0.19
Happy	0.04	0.0	0.02	0.81	0.04	0.02	0.07
Sad	0.13	0.0	0.08	0.05	0.41	0.03	0.3
Surprise	0.05	0.0	0.07	0.07	0.02	0.73	0.06
Neutral	0.06	0.0	0.04	0.04	0.12	0.02	0.72

Figure 4. The confusion matrix of medium model.

6. Dynamic Facial emotion recognition

As mentioned in previous sections, the main goal of FER is to enhance human computer interaction which involves dynamic interactions. To achieve this goal we need models that can preform in real-time or through video. Having datasets that demonstrate transition between different emotions leads to better models on this task. Yet, it worth to explore capabilities of our model on real-time dynamic FER.

6.1. Live Application

In order to investigate dynamic process of FER we developed an application to recognize emotions in real-time video inspired by [3] we used Haar-Cascade filter provided by OpenCV library to detect the nearest face appearing on each video frame. Then the extracted patch resized to 48x48 image and turned into a grayscale image. The result is fed into one of CNNs and the output probability of each emotion is shown on the screen. The highest probability was considered to be the most dominant emotion of the subject at the time. This illustration manner also helps to visualize probability of all other emotions including second and third highly probable ones.

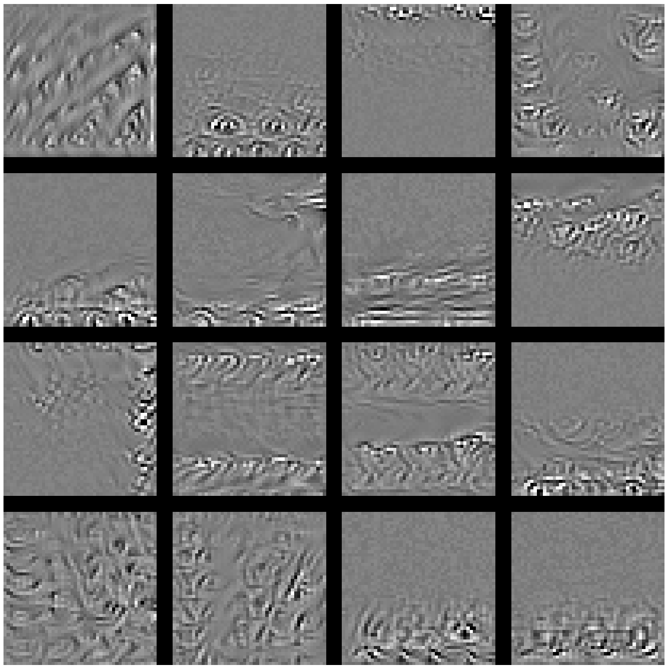


Figure 5. Layer visualization with optimization approach: Last convolutional layer in medium network after 50 update iteration.

We could run this real-time application using medium network on GPU. Finding nearest face in image takes roughly 100 ms and a forward path through network takes about 5 ms since it is done by GPU. We can state that our frame-rate is 10 frames/second. Although it is hard to assess dynamic performance of our model, it seems to be able to predict all 7 classes of emotions. However, we found it to struggling in bad illumination conditions. Though most negative emotions must be exaggerated to be detected by network, it does a good job on detecting subtle positive emotions.

We took a closer look on negative emotions to investigate

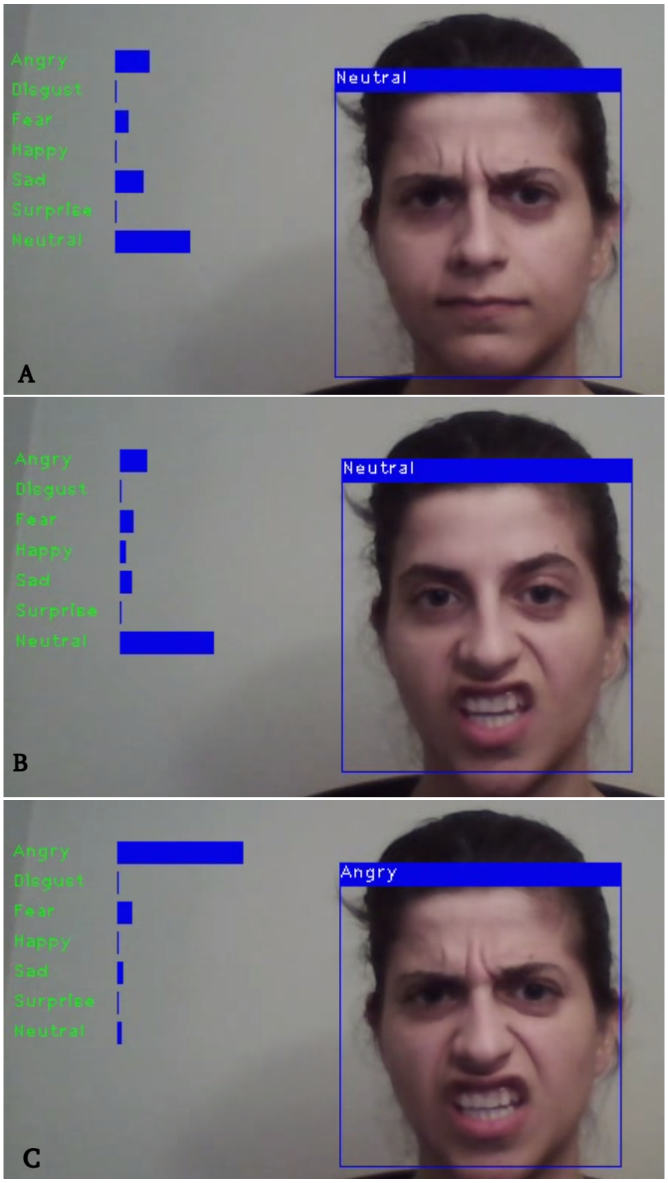


Figure 6. "The subject detected to be angry only when action unit 4 and 22 both are present. A. Action unit 4(Brrow Lowerer) preformed to show anger. The probability of anger is the second highest probability but the subject is detected as neutral. B. Action unit 22(Lip Funneler) preformed to show anger. The probability of anger is the second highest probability but the subject is detected as neutral. C. Both action units preformed and the subject detected as angry. "

detection of action units. As an example 'anger' contains action units 4 (Brrow Lowerer) and 22 (Lip Funneler).[4] Figure 6 shows that when the subject does only one of these action units, the probability of anger goes higher but the output still is 'Neutral'. The output emotion changes to 'anger' only when the subject does both of mentioned action units.

7. Conclusion

We developed various convolutional neural networks (CNNs) in terms of depth and data preprocessing and evaluated their performance. Our best model achieves 68.23% accuracy which is near to the winner of Kaggle's FER winner with accuracy of 71.16%. We applied some interesting layer visualization techniques. Also, we deployed a video application to do dynamic FER.

8. Future work

In this study we used simple CNN to do dynamic FER. In future we would like to try Recurrent neural networks on video datasets to do this task. We also like to analyze neurons of the network in terms of learned Facial action units. Recently, fascinating work has been done on face generation with generative adversarial networks (GANs). We would like to see whether it is possible to generate other emotions of a given face with the same identity.

Acknowledgement

we would like to appreciate Alona Fyshe and language learning lab members to organize this Directed studies course on deep learning. Our current understanding and intuition to this field wouldn't be possible without their help. We also thank developers of Keras and tensorflow.

References

[1] Marian Stewart Bartlett et al. "Fully automatic facial action recognition in spontaneous behavior". In: *FGR 2006: Proceedings of the 7th International Conference on Automatic Face and Gesture Recognition*. Vol. 2006. 2006, pp. 223–230. ISBN: 0769525032. DOI: 10.1109/FGR.2006.55.

[2] Youyi Cai et al. "Video based emotion recognition using CNN and BRNN". In: *Communications in Computer and Information Science*. Vol. 663. 2016, pp. 679–691. ISBN: 9789811030048. DOI: 10.1007/978-981-10-3005-5_56. arXiv: arXiv:1011.1669v3.

[3] Enrique Correa. "Emotion Recognition using Deep Convolutional Neural Networks". In: (2016).

[4] P Ekman and W Friesen. "False, felt, and miserable smiles". In: *Journal of Nonverbal Behavior* 6.4 (1982), pp. 238–258. URL: http://scholar.google.com/scholar?q=related:-8pw6P4UvJcJ:scholar.google.com/%7B%5C&%7Dhl=en%7B%5C&%7Dnum=20%7B%5C&%7Das%7B%5C_%7Dsdt=0,5.

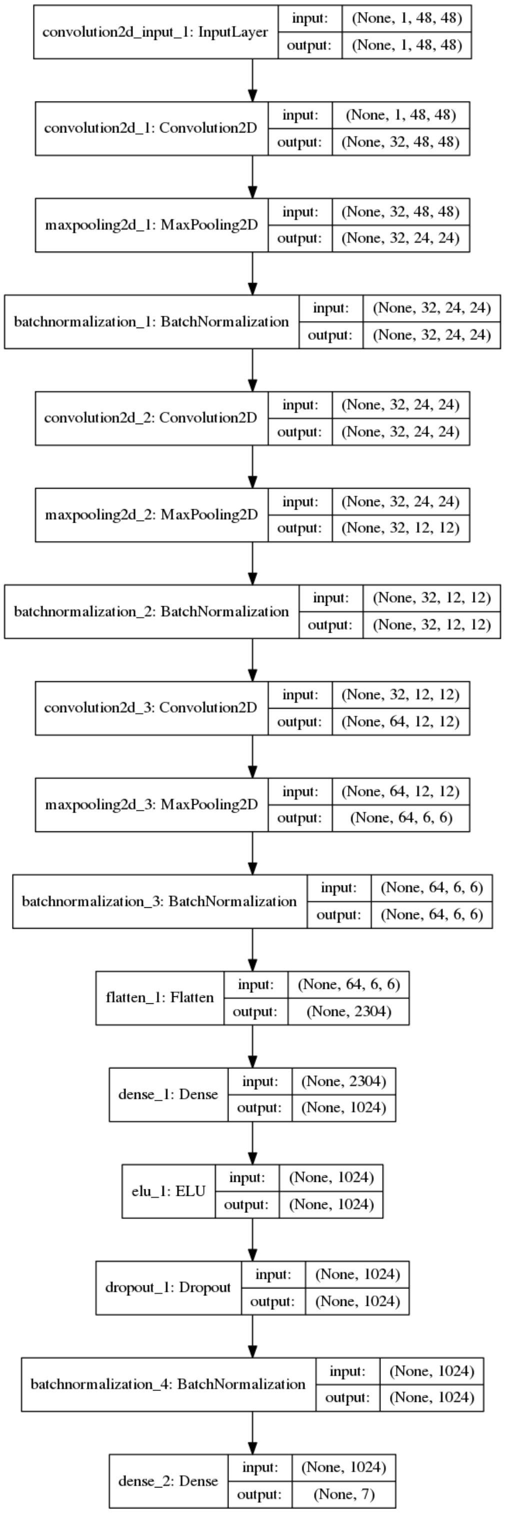

convolution2d_input_1: InputLayer | input: (None, 1, 48, 48) | output: (None, 1, 48, 48)
convolution2d_1: Convolution2D | input: (None, 1, 48, 48) | output: (None, 32, 48, 48)
maxpooling2d_1: MaxPooling2D | input: (None, 32, 48, 48) | output: (None, 32, 24, 24)
batchnormalization_1: BatchNormalization | input: (None, 32, 24, 24) | output: (None, 32, 24, 24)
convolution2d_2: Convolution2D | input: (None, 32, 24, 24) | output: (None, 32, 24, 24)
maxpooling2d_2: MaxPooling2D | input: (None, 32, 24, 24) | output: (None, 32, 12, 12)
batchnormalization_2: BatchNormalization | input: (None, 32, 12, 12) | output: (None, 32, 12, 12)
convolution2d_3: Convolution2D | input: (None, 32, 12, 12) | output: (None, 64, 12, 12)
maxpooling2d_3: MaxPooling2D | input: (None, 64, 12, 12) | output: (None, 64, 6, 6)
batchnormalization_3: BatchNormalization | input: (None, 64, 6, 6) | output: (None, 64, 6, 6)
flatten_1: Flatten | input: (None, 64, 6, 6) | output: (None, 2304)
dense_1: Dense | input: (None, 2304) | output: (None, 1024)
elu_1: ELU | input: (None, 1024) | output: (None, 1024)
dropout_1: Dropout | input: (None, 1024) | output: (None, 1024)
batchnormalization_4: BatchNormalization | input: (None, 1024) | output: (None, 1024)
dense_2: Dense | input: (None, 1024) | output: (None, 7)

Figure 7. The architecture of medium network: 3 convolutional network followed by 2 fully connected layer. Input/out volume is mentioned.

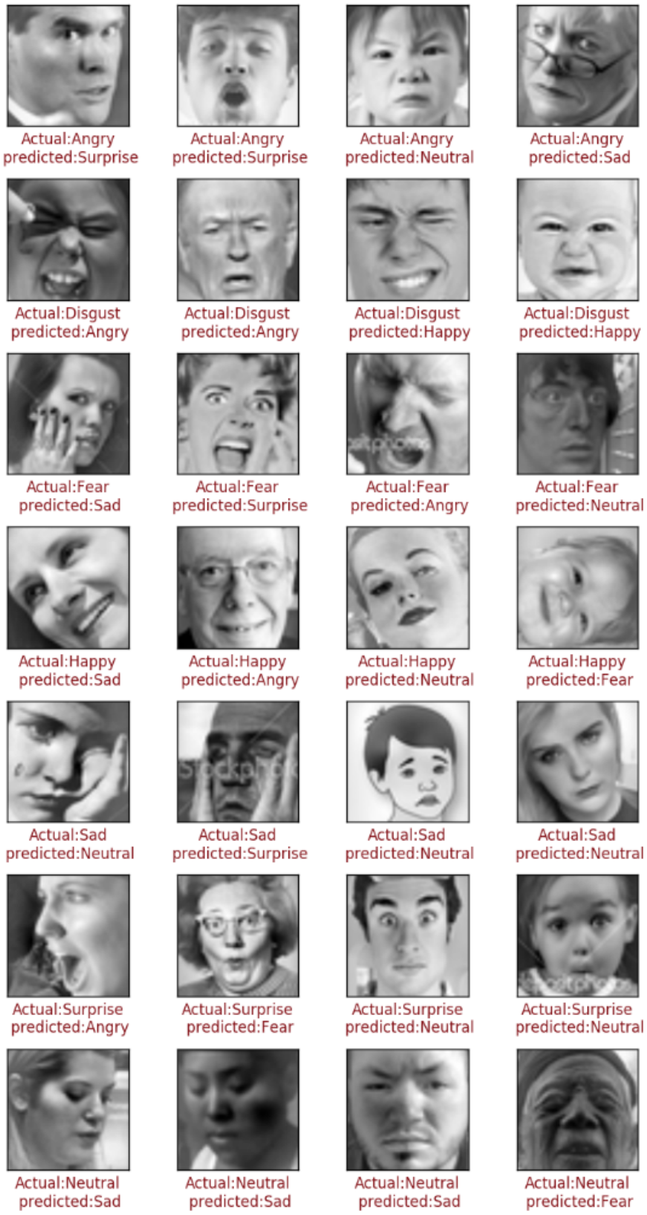

Actual:Angry predicted:Surprise
Actual:Angry predicted:Surprise
Actual:Angry predicted:Neutral
Actual:Angry predicted:Sad
Actual:Disgust predicted:Angry
Actual:Disgust predicted:Angry
Actual:Disgust predicted:Happy
Actual:Disgust predicted:Happy
Actual:Fear predicted:Sad
Actual:Fear predicted:Surprise
Actual:Fear predicted:Angry
Actual:Fear predicted:Neutral
Actual:Happy predicted:Sad
Actual:Happy predicted:Angry
Actual:Happy predicted:Neutral
Actual:Happy predicted:Fear
Actual:Sad predicted:Neutral
Actual:Sad predicted:Surprise
Actual:Sad predicted:Neutral
Actual:Sad predicted:Neutral
Actual:Surprise predicted:Angry
Actual:Surprise predicted:Fear
Actual:Surprise predicted:Neutral
Actual:Surprise predicted:Neutral
Actual:Neutral predicted:Sad
Actual:Neutral predicted:Sad
Actual:Neutral predicted:Sad
Actual:Neutral predicted:Fear

Figure 8. Samples of misclassified images.

[5] Lijie Fan and Yunjie Ke. “Spatiotemporal Networks for Video Emotion Recognition”. In: (2017), pp. 1–8. arXiv: `1704.00570`. URL: `http://arxiv.org/abs/1704.00570`.

[6] Ian Goodfellow et al. *Challenges in Representation Learning: A report on three machine learning contests*. 2013. URL: `http://arxiv.org/abs/1307.0414`.

[7] T Kanade and J.F. Cohn. “Comprehensive database for facial expression analysis”. In: *Proceedings of the 4th IEEE International Conference on Automatic*

Face and Gesture Recognition (2000), pp. 46–53. ISSN: 0-7695-0580-5. DOI: 10.1109/AFGR.2000.840611. URL: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=840611.

[8] *Keras: Deep Learning library for Theano and TensorFlow*. 2017. URL: https://keras.io/.

[9] Pooya Khorrami, Tom Le Paine, and Thomas S. Huang. "Do Deep Neural Networks Learn Facial Action Units When Doing Expression Recognition?" In: *Proceedings of the IEEE International Conference on Computer Vision*. Vol. 2016-February. 2016, pp. 19–27. ISBN: 9781467383905. DOI: 10.1109/ICCVW.2015.12. arXiv: 1510.02969.

[10] Bo-kyeong Kim et al. "Fusing Aligned and Non-Aligned Face Information for Automatic Affect Recognition in the Wild : A Deep Learning Approach". In: (2016), pp. 1499–1508. ISSN: 21607516. DOI: 10.1109/CVPRW.2016.187.

[11] *Layer visualization in Keras*. 2017. URL: https://blog.keras.io/how-convolutional-neural-networks-see-the-world.html.

[12] Gil Levi. "Emotion Recognition in the Wild via Convolutional Neural Networks and Mapped Binary Patterns". In: *Icmi*. 2015, pp. 503–510. ISBN: 9781450339834. DOI: 10.1145/2823327.2823333.

[13] Ali Mollahosseini, David Chan, and Mohammad H. Mahoor. "Going Deeper in Facial Expression Recognition using Deep Neural Networks". In: *arXiv preprint arXiv:1511.04110* (2015). DOI: 10.1109/WACV.2016.7477450. arXiv: 1511.04110. URL: http://arxiv.org/abs/1511.04110.

[14] Karen Simonyan and Andrew Zisserman. "Very Deep Convolutional Networks for Large-Scale Image Recognition". In: *ImageNet Challenge* (2014), pp. 1–10. ISSN: 09505849. DOI: 10.1016/j.infsof.2008.09.005. eprint: 1409.1556. URL: http://arxiv.org/abs/1409.1556.

[15] Jacob Whitehill and Christian W. Omlin. "Haar features for FACS AU recognition". In: *FGR 2006: Proceedings of the 7th International Conference on Automatic Face and Gesture Recognition*. Vol. 2006. 2006, pp. 97–101. ISBN: 0769525032. DOI: 10.1109/FGR.2006.61.

[16] Dong Yi et al. "Learning Face Representation from Scratch". In: *arXiv* (2014). arXiv: arXiv:1411.7923v1.

[17] Jason Yosinski et al. "Understanding Neural Networks Through Deep Visualization". In: *International Conference on Machine Learning - Deep Learning Workshop 2015* (2015), p. 12. arXiv: 1506.06579. URL: http://arxiv.org/abs/1506.06579.

[18] Zhiding Yu and Cha Zhang. "Image based Static Facial Expression Recognition with Multiple Deep Network Learning". In: *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction* (2015), pp. 435–442. DOI: 10.1145/2818346.2830595.

[19] Guoying Zhao and M Pietikainen. "Dynamic texture recognition using local binary patterns with an application to facial expressions". In: *Pattern Analysis and Machine . . .* 29 (2007), pp. 1–14. ISSN: 0162-8828. DOI: 10.1109/TPAMI.2007.1110. URL: http://www.computer.org/portal/web/csdl/doi/10.1109/TPAMI.2007.1110%7B%5C%%7D5Cnhttp://ieeexplore.ieee.org/xpls/abs%7B%5C_%7Dall.jsp?arnumber=4160945.