Facial Emotion Detection

by

Khelawan Singh

1. Abstract

As is well known, emotions in real-life scenarios have a significant impact on information processing, attitude formation, and decision making. Although several efforts on FER or facial expression recognition have been published recently, a reliable and robust FER system remains a challenge due to human facial diversity and image variability. To date, all research and work have proposed either single network or ensemble models. Ensemble models are more accurate, but are associated with many models and datasets, and some optimized datasets, to improve accuracy and increase computational complexity. Most of the research in this area has focused on improving accuracy, but this study found that human faces contain a variety of emotions, and single-label moods can be very annoying in such situations. I am using the proposed model for real-world scenarios. This paper presents a single standalone CNN model implemented in a real-time intelligent emotion detection system that validates its accuracy through transfer learning, performs tasks such as face detection, emotion classification, and provides a live list of probabilistic labels. I suggest Real-time from webcam - feed in mixed steps. Human emotions are the mental state of feelings and are spontaneous. There is no clear connection between emotions and facial expressions and there is significant variability making facial recognition a challenging research area. Features like Histogram of Oriented Gradient (HOG) and Scale Invariant Feature Transform (SIFT) have been considered for pattern recognition. These features are extracted from images according to manual predefined algorithms. In recent years, Machine Learning (ML) and Neural Networks (NNs) have been used for emotion recognition. In this report, a Convolutional Neural Network (CNN) is used to extract features from images to detect emotions.

2. Keyword

- Webcams
- Computational modeling
- Face recognition
- Probabilistic logic
- Real-time systems
- Noise measurement
- Feeds

3. Introduction

A study by psychologist Mehrabian on human communication found that 55% of information is conveyed through facial expressions, 38% through supporting language such as tone and voice, and only 7% through spoken language. Despite deep learning approaches showing efficiency in automatically recognizing facial emotions, the study of emotions was beginning to gain interest by the early 1970s. We introduced six basic emotions: sadness, surprise . The computer scientist has been interested in studying emotions since his late 1980s, culminating with published work on the development of automatic emotion recognition systems. Most conventional approaches consist of her three steps. The first step in practice is to detect the face position. Then, we extract the geometric features of the face to generate a specific vector and classify the emotion with the maximum score . These methods require a lot of technical manipulation. The sheer volume of data makes characterization very difficult. CNN showed great potential soon after its launch in the late 1990s. It has been proven in various image classification tasks and has also been used in emotion recognition. One of the main advantages of deep learning CNNs is their ability to learn features directly from data, avoiding the tedious and manual feature generation used in other supervised learning strategies. However, the use of CNNs at that time was limited by the lack of training data and processing power. Since the 2010s, as processing power and the accumulation of large datasets have increased, CNNs have become much more viable tools in image processing, pattern recognition, and feature extraction. In this article, faces are detected by hair cascade feature extraction and mood classification is performed by different layers of the CNN model. Facial sensation/emotion recognition frameworks are an important part of computational reasoning, with applications in psychometric testing, driver fatigue observation, intelligent game planning, versatile applications for naturally embedding emotions into language, medically useful frameworks, and more. , you can fill out accreditable applications in various fields. Introverts, keen insightful robots, etc.

Motivation

Emotions play an essential role in identifying the mood of a human being. There are generally six raw emotions: happy, sad, anger, fear, surprise, disgust, contempt. It is seen that research work focuses on the four main emotions named happy, sad, angry and neutral. Also, recognising the emotions plays an important role in camera surveillance to capture the suspects; for example, in the case of a feared person, a system raising the alarm can help. Emotion recognition systems can be used as a sub-module of various applications like recommending music and various camera surveillance systems.

Challenges

Defining a facial expression as representative of a certain emotion can be difficult even for humans. Studies show that different people recognize different emotions in the same facial expression. And it's even harder for AI. There's an ongoing debate as to whether existing emotion recognition solutions are accurate at all. A lot of factors make emotion recognition tricky. Emotion recognition shares a lot of challenges with detecting moving objects in video, identifying an object, continuous detection, incomplete or unpredictable actions, etc. Let's take a look at the most widespread technical challenges of implementing an ER solution and possible ways of overcoming them.

As with any machine learning and deep learning algorithms, ER solutions require a lot of training data. This data must include videos at various frame rates, from various angles, with various backgrounds, with people of different genders, nationalities, and races, etc. However, most public dataset aren't sufficient. They aren't diverse enough in terms of race and gender and contain limited sets of emotional expressions.

4. Literature work

Studies made

Many CNN variants have achieved remarkable results with a classification accuracy between 65 % and 76.82 %. Assembling multiple models has shown improved performance. For instance, in order to enhance performance, Liu et. al. trained three distinct CNNs and combined them. The greatest single-network accuracy they obtained was 62.44 percent. Minaee et. al. obtained a 70.02 percent accuracy using an attentional convolutional network in an end-to-end deep learning framework. Tang et. al. replaced the softmax layer with a support vector machine in a deep neural network and achieved a classification accuracy of 71.52 %. Shi et. al. suggested a unique amended representation module (ARM) to replace the pooling layer and obtained 71.38 percent testing accuracy. Pramerdorfer et. al. compared the performance of VGG architecture, Inception architecture, and ResNet architecture. According to their findings, VGG had the greatest accuracy of 72.7 percent, followed by ResNet at 72.4 percent and Inception at 71.6 percent. Other studies have attempted to enhance their model's performance on FER2013 by improving the data and incorporating supplementary training data, however, this is outside the scope of this study.

Datasets and their reported results

A competition named Facial Expression Recognition Challenge was held in 2013. During this competition, competitors were encouraged to build the best system for detecting which emotion is being conveyed by a picture of a human face in a facial expression identification competition. Pierre Luc Carrier and Aaron Courville created and launched the 'Facial Expression Recognition 2013' FER-2013 dataset in this competition. The dataset was produced by using the Google image search API. There are 4953 "Anger" photos, 547 "Disgust" images, 5121 "Fear" images, 8989 "Happiness" images, 6077 "Sad" images, 4002 "Surprise" images, and 6198 "Neutral" images among the 35887 images. Ian Goodfellow conducted several small-scale tests to assess human performance on this job and discovered that human accuracy on the FER-2013 was 65+5%. The top four teams utilized discriminatively trained convolutional neural networks with image transformation, as illustrated in fig 1 below. Each work done in this field after the year 2015, using FER2013 dataset has been mentioned in fig 2 below along with their reported accuracies.

Fig 1

Team	Members	Accuracy
RBM	Yichuan Tang	71.162%
Unsupervised	Yingbo Zhou, Chetan Ramaiah	69.267%
Maxim Milakov	Maxim Milakov	68.821%
Radu + Marius + Cristi	Radu Ionescu, Marius Popescu, Cristian Grozea	67.484%

S. No	Method	Dataset	Accuracy	Author/year
1	sequential fully-CNN Less no of parameters trained with ADAM optimizer	FER2013 IMDB	66%	Octavio Arriaga et al [2018] [12]
2	Image based static facial expression recognition with multiple deep network learning	FER2013	72.00%	Yu et al [2015] [13]
3	MRMR-based ensemble pruning for facial expression recognition.	FER2013	70.66%	Li et al. [2017] [14]
4	Using a Hybrid CNN–SIFT Aggregator.	FER2013	73.4%	Connie et al. [2017] [15]
5	Human Emotions Recognition for Realizing Intelligent Internet of Things	FER2013	71.91%	Hua et al. [2019] [16]
6	Local Learning with Deep and Handcrafted Features	FER2013	73.25%	Mariana-Iuliana Georgescu [2020] [17]
7	CNN OpenCV implemented in the Raspberry Pi consists of 3 main processes detection, feature extraction, emotion classification.	FER2013	65.97%	Lutfiah Zahara Purnawarman Musa [2020] [18]
8	Saptial features VGG CNN capture spatial and audio features from videos.	FER2013 EmotiW	60.03 %	Boris Knyazev Roman Shvetsov [2017] [19]
9	adopt the VGGNet architecture, fine-tune its hyperparameters, various optimization methods, saliency maps	FER2013	73.28 %	Yousif Khaireddin Zhuofa Chen [2021] [20]
10	SVM, saliency maps extract important regions for detecting different facial expressions	FER2013,CK +, FERG, JAFFE.	70.02%	Shervin Minaee, Amirali Abdolrashidi [2019] [21]
11	CNN, Random Search algorithm applied on a search space defined by discrete values of hyperparameters	FER2013	72.16%.	Adrian Vulpe-Grigoraşi Ovidiu Grigore [2021] [22]

Fig 2

5. Methodology

The final model was trained on sets which has 7 sentiments. Hence, the dataset was reduced to a total of 24282 images in the train set, 5937 images in the validation set, and 6043 images in the test set.

Data Augmentation

The first thing we did with the model was to account for the variability of facial expression recognition. We applied massive data augmentation to training images using image data preprocessing and image data generator in Keras. This expansion involves rescaling (also called normalizing) the image data to a range between 0 and 1 suitable for neural network models by dividing each pixel by 255. We rotate the image by 30 degrees, shift the image shear and zoom by 0.3x, perform horizontal and vertical shift of the image by 0.4x, and finally flip the image horizontally and set the blending mode. increase. To "closest". Each technique is applied randomly. For verification and testing, normalize the image by dividing each pixel by 255. The image is then set to grayscale in the generator and cropped to a size of 48x48 pixels. If Shuffle is set to true, a random image from the train folder directory will be sent every 20 in "categorical" class mode. That is, the labels are encoded as categorical vectors. The same goes for validation generators.

Model Description

Our proposed model is inspired by VGG-16 and is a sequential model incremented by the add(method provided by Keras since April 2020. Each layer has 4 convolution stages and 3 fully connected The layers make up the model. The convolution stage is responsible for feature extraction, dimensionality reduction, and nonlinearity, while the fully connected layers classify the input as described by the extracted features. The convolutional stage contains two convolutional blocks consisting of a convolutional layer, an ELU activation and a batch normalization layer, and the features of each convolutional block are 32 in the first block and 2 It increases to 64 for the 3rd block, 128 for the 3rd block, and 256 for the last block, each followed by a max pooling layer and a dropout layer. The above layers are suggested to improve performance. For example, the Rectified Linear Unit (ReLU) activation function was replaced with ELU (Exponential Linear Unit) as the activation function to avoid gradient scattering issues and speed up training. Max pooling is used to get the maximum score from the convolutional block and downsample the input to aid generalization. A dropout layer is used throughout to prevent overfitting, as the drawing accuracy is kept under control and lower than the test and validation accuracy. Stack normalization is used to speed up the learning process by preventing vanishing and exploding gradients.

The first fully connected layer is the dense layer passed after the data is flattened, followed by ELU activation, batch normalization, and dropout layers. The second layer repeats the same layer without smoothing the input. The third fully connected layer is a dense layer followed by "SoftMax" activations that are often used for the last layer in classification networks. The elements of the output vector are in the range (0, 1) and sum to 1.

where the flatten function transforms or flattens the aggregated feature map into a single column, the channel dimension. Then the dense layer, which is actually a fully connected network layer, feeds the output of the flattening function to all neurons, each of which produces a single output, which is reproduced by ELU. Activated and normalized. Then feed the next two dense layers again. Here, later layers are activated by the SoftMax function to predict classification. In this case, 7 emotions.

Learning rate and optimizer

One important aspect that can affect performance is the learning rate. A low learning rate can cause the model to converge much slower and end up in suboptimal local minima. Oscillations around the minimum or loss divergence can be caused by a high learning rate. A learning rate scheduler, which adjusts the learning rate during training, is a useful approach. For example, as the number of iterations increases, the learning rate decreases gradually or dramatically due to time-based decay. After a certain number of epochs, the learning rate drops by a factor of 1 due to lower levels. During training, adaptive learning rate plans attempt to automatically change the learning rate based on local gradients. For example, cosine annealing periodically resets the learning rate and reuses “good weights” during the training process. ReduceLROnPlateau is used to monitor test loss, and reduces the learning rate by 0.001 if no gain is detected after up to 3 epochs of perseverance.

The most commonly used optimizer is Stochastic Gradient Descent (SGD). This is a simple technique that updates model parameters based on the gradient of a single data point. AdaGrad adaptively scales the learning rate for each network dimension. RMSProp significantly slows down the learning rate. Adam combines the advantages of AdaGrad and his

RMSProp by adjusting the learning rate and incorporating gradient momentum, among other things, for an optimal optimizer.

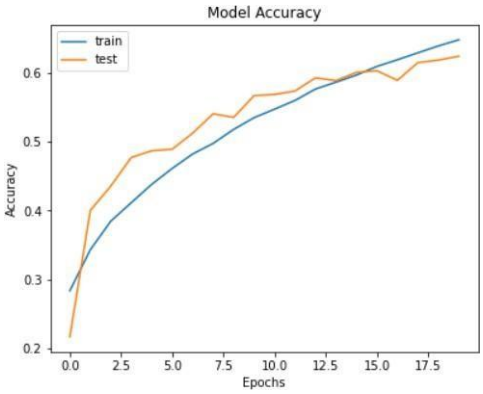
Training

We ran all layers for 100 epochs optimizing the cross-entropy loss. The model is saved every time the maximum test accuracy is achieved. Early stopping is set with the patience of 10, i.e. to stop the training after 10 epochs if there is no improvement in validation loss and restore the model weights from the epoch with the best validated accuracy. Hence, the validation loss was also kept in check.

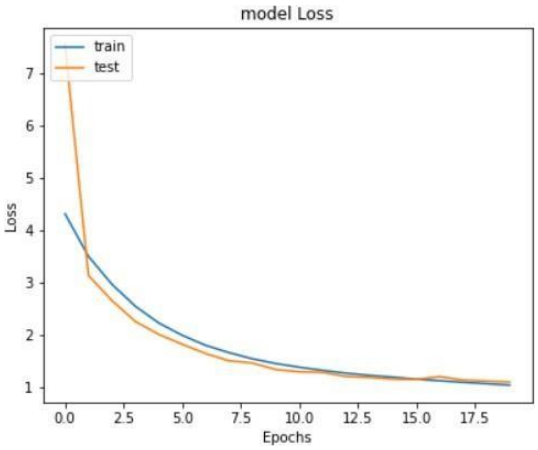
6. Result

To properly analyze the proposed model, accuracy plots showing training and testing accuracies and loss plots showing training and testing losses are plotted against the number of epochs using matplotlib. As can be seen in Figures 1 and 2, the first 5 epochs have significantly improved accuracy and loss. The model continues to improve significantly, successfully surpassing the claimed human accuracy of 55% at 10 epochs, but after 15 epochs, both loss and accuracy show little progress, reaching near It stays the same and doesn't get any better around 18 epochs. Even at very low learning rates. The negated improvement in training accuracy also indicates that the model is not overfitting and that the dropout layer helped achieve this.

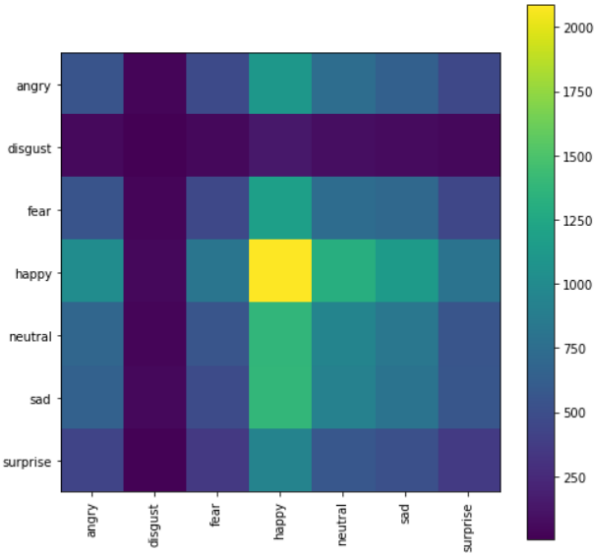
- **Model accuracy variation with number of epochs**



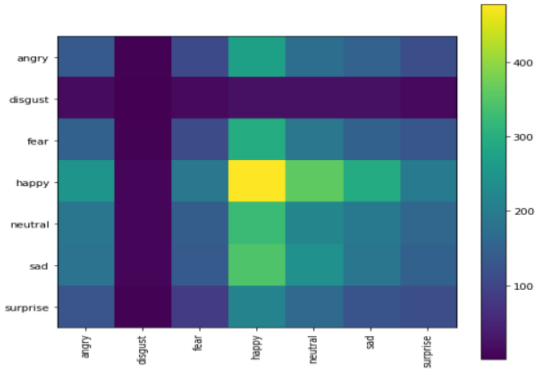
- **Model loss with number of epochs**



- **Confusion matrix of training set**



- **Confusion matrix on test set**



We can exhibit the confusion matrix together with the heatmap to better examine our model. Each row of the

matrix represents a real setiment, while each column provides a projected sentiment. Table No. 3 shows the classification report for the testing data, which allows us to go further into the performance of each sentiment. As it can be seen, this system has the greatest precision in predicting 'happy' sentiments. and the lowest precision for 'sad'. The model's overall accuracy, macro avg, and weighted avg on a scale can also be seen all lying in the range of 0 to 1, where 1 represents 100%.

- **Classification report of training set**

Classification Report

	precision	recall	f1-score	support
angry	0.14	0.14	0.14	3995
disgust	0.01	0.01	0.01	436
fear	0.14	0.11	0.13	4097
happy	0.25	0.29	0.27	7215
neutral	0.18	0.19	0.18	4965
sad	0.17	0.16	0.17	4830
surprise	0.11	0.11	0.11	3171
accuracy			0.18	28709
macro avg	0.14	0.14	0.14	28709
weighted avg	0.18	0.18	0.18	28709

- **Classification report of test set**

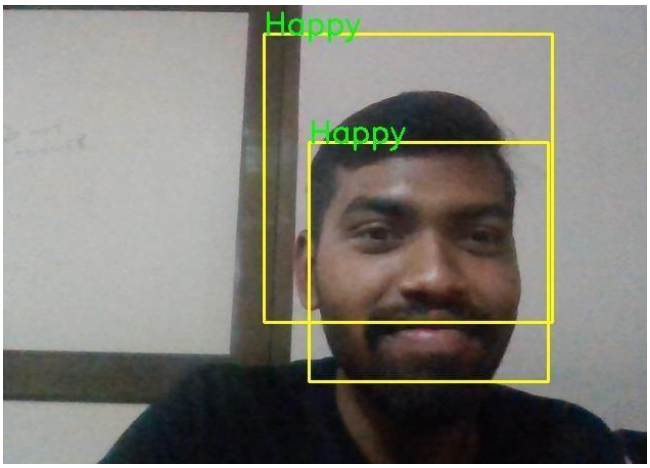
Classification Report

	precision	recall	f1-score	support
angry	0.13	0.14	0.14	958
disgust	0.03	0.01	0.01	111
fear	0.14	0.11	0.12	1024
happy	0.24	0.27	0.26	1774
neutral	0.16	0.18	0.17	1233
sad	0.17	0.15	0.16	1247
surprise	0.13	0.14	0.13	831
accuracy			0.17	7178
macro avg	0.14	0.14	0.14	7178
weighted avg	0.17	0.17	0.17	7178

6.1 Implementation

The model is showing result correct and detecting Emotion successfully.

Happy



Angry

7. Conclusion and future direction

To conclude, we closely observed and learned about various work and research done in this field. All implemented single standalone CNNs which are trained on the FER2013 dataset were studied and their achieved results were noted. We then developed, a CNN model . After this, we started developing the model from the scratch and after working on 3-4 models with different sets and combinations of layers, the final model presented in this paper was proposed, which achieved the highest accuracy of 64.5 % , outperforming all single network models we know and studied in this work. Coming to the dataset, FER2013 remains a challenging dataset as it is wild and crowdsourced,we have not done any modification in the dataset. Furthermore, we designed and implemented a real-time Intelligent System for Sentiment Recognition, which accomplishes the tasks of face detection using haar cascade, sentiment

classification using the proposed model's saved weights, and gives a live list of probabilistic labels simultaneously in real-time from a webcam feed in one blended step. In future we will implement our model to develop a web-based application using flask. And in future this emotion detection model can be used in various application like emotion based music player and emotion based movie recommendation.

REFERENCES

1). Arriaga, H. Bonn-Rhein-Sieg, & M. Valdenegro, (2019). Real time Convolutional Neural Networks for Emotion and Gender Classification. 5.

2) Z. Yu and C. Zhang. Image based static facial expression recognition with multiple deep network learning. In Proceedings of ICMI, pages 435–442. ACM, 2015.

3) D. Li and G.Wen. MRMR-based ensemble pruning for facial expression recognition. Multimedia Tools and Applications, pages 1– 22, 2017.

4) T. Connie, M. Al-Shabi, W. P. Cheah, and M. Goh. Facial Expression Recognition Using a Hybrid CNN–SIFT Aggregator. In Proceedings of MIWAI, volume 10607, pages 139–149. Springer, 2017.

5) W. Hua, F. Dai, L. Huang, J. Xiong, and G. Gui. HERO: Human Emotions Recognition for Realizing Intelligent Internet of Things. IEEE Access, 7:24321–24332, 2019.

6) M.-I. Georgescu, R. T. Ionescu, & M. Popescu(2019). Local Learning with Deep and Handcrafted Features for Facial Expression Recognition. IEEE Access, 7, 64827–64836.

7)L. Zahara , P. Musa, E. Prasetyo Wibowo, I. Karim, , & S. Bahri Musa (2020). The Facial Emotion Recognition (FER-2013) Dataset for Prediction System of Micro-Expressions Face Using the Convolutional Neural Network (CNN) Algorithm based Raspberry Pi. 2020 Fifth International Conference on Informatics and Computing (ICIC), 1–9.

8) B. Knyazev, R. Shvetsov, N. Efremova, & A. Kuharenko(2017). Convolutional neural networks pretrained on large face recognition datasets for emotion classification from video. [20] [20] Y. Khaireddin, Z. Chen, (2021). Facial Emotion Recognition: State of the Art Performance on FER2013.

9) S. Minaee, A. Abdolrashidi, (2019). Deep-Emotion: Facial Expression Recognition Using Attentional Convolutional Network. [22][22] A. Vulpe-Grigoraşi, O. Grigore (2021). Convolutional Neural Network Hyperparameters optimization for Facial Emotion Recognition. 2021 12th International Symposium on Advanced Topics in Electrical Engineering (ATEE), 1-5