AI Programming Project

Tuyen Dao Cong, Hieu Le Trung

November 2021

Abstract

This report is a compilation of the results of research and development of the facial emotional recognition problem conducted during the 10 weeks of AIP391 (AI programming project) subject. We conducted surveys of the-state-of-the-art models used for FER, identifying bottlenecks and learning lessons to overcome them for optimizing our model not only in terms of performance but also in terms of implementation time. This report is the final results that we did our best to achieve.

Contents

1 Introduction

Automatic facial expression recognition (FER) is considered as one of the most challenging tasks in computer vision. Automatic facial expression recognition (FER) system is a technology capable of identifying facial expressions (FEs) by analyzing visual cues or features that are extracted from a digital image or a video frame. FER admits a wide range of applications in human–computer interaction, behavioral psychology, and human expression synthesis like human

behavior understanding, mental disorder detection, cognition human emotions, safe driving , photo-realistic human expression synthesis, computer graphics animation and other similar tasks [8]. One interesting societal application of the FER system is to assist visually impaired persons (VIPs) in their day-to-day communication. Such a system could render a better sense of living their life [10].

Human emotions have been examined in studies with the help of acoustic and linguistic features , facial expressions, body posture, hand movement, direction of gaze , and utilization of electroencephalograms (EEGs) and electrocardiograms (ECGs) [1]. Though humans are very good at recognizing the emotional states of a person, for a computer, the task is very complicated [10]. Up to now, recognizing basic expressions under controlled conditions can now be considered a solved problem (The term basic expression refers to a set of expressions that convey universal emotions, usually **anger, disgust, fear, happiness, sadness,** and **surprise**). However, recognizing such expressions under naturalistic conditions is more challenging. This is due to variations in head pose and illumination, occlusions, and the fact that unposed expressions are often subtle. However, reliable FER under naturalistic conditions is mandatory in the aforementioned applications, therefore need to be solved effectively [9].

It is noticeable that the detailed local features like eyes and mouth corners that are exhibited by different FEs in face images and these features are essential for identifying the emotion of individual subjects. If using handcrafted features, they will limit the performance because of the need of extracting accurately all the correlated handcrafted features is very challenging [8]. Even though several Deep Learning (DL) networks exist for FER, most of them do not pan well when they are challenged with data that require a thorough understanding of the inherent features for FER [6].

2 Related work

2.1 Alexnet

The AlexNet is a deep Convolutional Neural Network (CNN) constructed as a combination of convolutional and fully connected layers. It consists of an input layer followed by five convolutional layers and three fully connected layers. Each convolutional layer consists of convolutional filters and a nonlinear activation function ReLU. The output from the last layer is passed through the normalized exponential Softmax function that maps a vector of real values into the range [0, 1] that add up to 1. These values represent the probabilities of each class (see Figure 1).

Input size is fixed due to the presence of fully connected layer, which takes 227×227 input. During the experiment, the number of classes of output needs to be adjusted to seven. With the FER2013 dataset, original images with size 48×48 are used bilinear interpolation to match the input requirements. In paper [1] the result mentioned after training was 61.0% accuracy and in [11] the accuracy is 61.1%. At the same time, data augmentation is not used. This result is also reproduced by us. However, in paper [8], the accuracy result of Alexnet is up to 77%, raises questions about the problem of reproducing the results in the study or there was a new improvement in image pre-processing

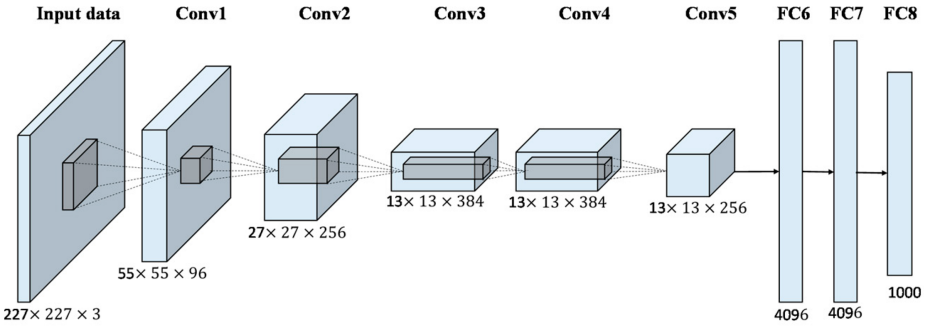


Figure 1: Structure of the AlexNet used for ImageNet showing five convolutional layers Conv1 - Conv5 and three fully connected layers FC6 - FC8. The last layer provides an output to the Softmax function (Source)

not mentioned.

One thing that should be noticed is that the network structures for recognizing facial emotions are quite simple and quite similar, even the network structures have good results (one reason is the trade-off between bias and variance in a relatively small dataset). This has raised the idea that because the neural network space is fairly uniform and the corresponding papers lack details on architecture selection, can we encode the structure of each model and use random search algorithms such as evolutionary algorithms to find an optimal network structure in the feasible space.

2.2 Fernet

In this study, a simple CNN reckoned were proposed as **FER-net** for FER. Five publicly available benchmarking datasets, namely FER2013, JAFFE, CK+, KDEF, and RAF datasets are considered. From the obtained results, the proposed model is considered as the state of the art method in almost all cases [8].

The study reviewed various methods in face detection or the use of handcraft feature and Machine Learning tools for emotion classification. The article also mentioned a large number of DL models . It also point out that it is unclear whether legacy works performing well in lab-controlled environments would provide satisfactory performance on difficult real-time datasets such as Facial Expression Recognition 2013 (FER2013).

The detailed architecture of FER-net is explained in Figure 2. It consists of four convolution layers (C1, C2, C3, and C4), four max-pooling layers (P1, P2, P3, and P4), and two fully connected layers (F1 and F2). Batch-normalization is applied to the outputs of four convolutional layers and the two fully connected layers. Further, convolved features are fed into the activation function rectified linear unit (ReLU). Finally, the output of the second fully connected layer is fed into the softmax layer. Dropout is applied to each convolution layer of 0.25 and 0.5 to fully connected layers. Categorical-cross entropy is used to measure the loss in this structure. When training, the early stopping technique is applied. Finally, adam optimizer is used for optimization and weight update. It's worth noting that most settings for this network structures are explained

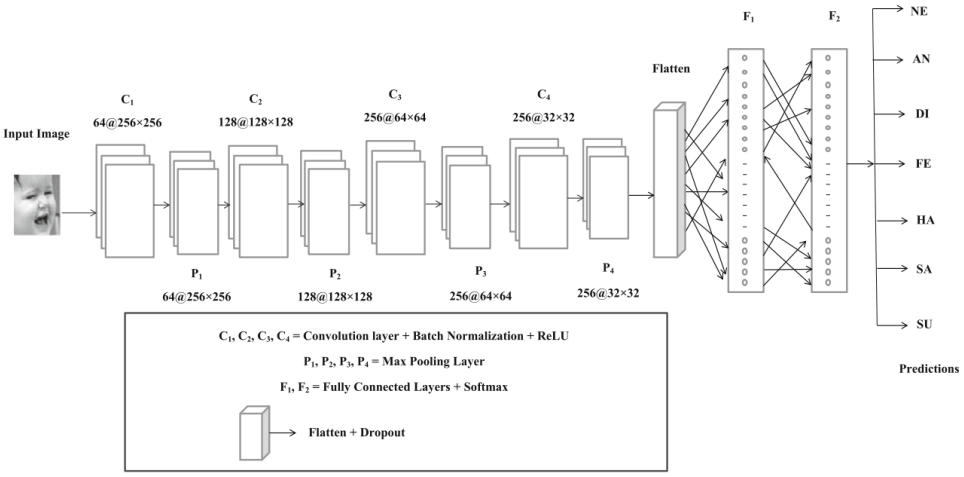


Figure 2: Schematic block diagram of the FER-net [8].

by the reasons behind. In contrast, in many other papers they have lack details on architecture selection, the reason for this discrepancy is unknown [9].

All the datasets are divided into three parts mainly for training sets, validation sets, and testing sets separately. The ratio of dividing datasets for training, validation, and testing is 80%, 10%, and 10%, respectively. Image augmentation techniques such as translation, scaling, rotation, flipping the images vertically and horizontally, adding noise to the images are also applied.

The evaluation metrics including accuracy, precision, recall, and F1-score are considered to compare the performance of the proposed FER-net over twenty-one state-of-the-art models. The result was recorded that the proposed CNN model FER-net, yields good classification accuracies along with other metrics for JAFFE (97%), CK+ (98%), and KDEF datasets (83%) in all most all the FEs. However, the performance is satisfactory for the other two datasets FER2013 (up to 79%) and RAF dataset (82%). Therefore, FER-net outperforms legacy works in almost all the cases.

The proposed model FER-net is simple as compared to state-of-the-art models and is preeminent in terms of accuracy, as well as execution time [8]. However, we have not been able to reproduce this result, one reason is the limit of resource. In our experiment, we keep the original input image size 48×48 and only use FER2013 datasets as training (while this paper use 5 datasets for training), beside that we don't use data augmentation also. The highest result we have until now is about 72% (while in this paper, the accuracy is 79%).

2.3 VGGnet

The study [6] aim to construct method for improving prediction accuracy on FER2013 using CNNs. A new single-network structure was proposed by adopting the VGG network and constructing various experiments to explore different optimization algorithms and learning rate schedulers, then thoroughly tuning the model and training hyperparameters.

Using only the FER2013 dataset for training and adhering to the official training, validation, and test sets as introduced by the ICML (International Conference on Machine Learning) . Data augmentation is also used as an important step when using FER-2013 dataset [9].

About **VGGNet**, VGGNet is a classical convolutional neural network architecture used in large-scale image processing and pattern recognition. In this case, a variant of VGGNet is considered with the structure shown in Figure 3. The network consists of 4 convolutional stages and 3 fully connected layers. Each of the convolutional stages contains two convolutional blocks and a max-pooling layer. The convolution block consists of a convolutional layer, a ReLU activation, and a batch normalization layer. The first two fully connected layers are followed by a ReLU activation. The convolutional stages are responsible for feature extraction, dimension reduction, and non-linearity. The models are evaluated using validation accuracy and tested using standard ten-crop averaging.

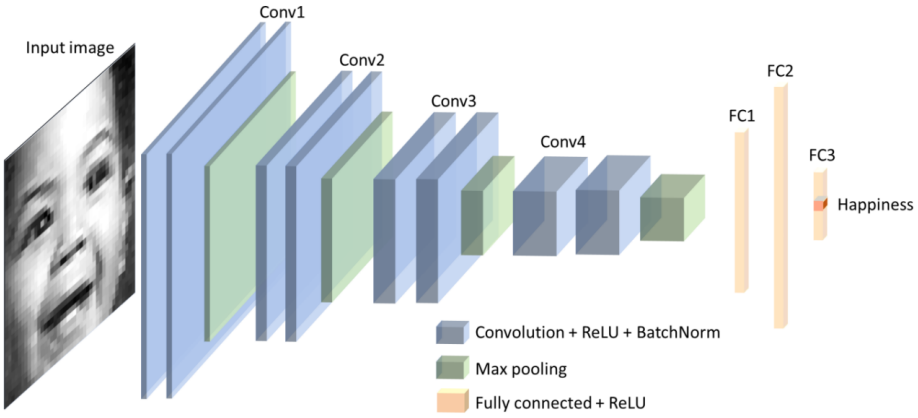


Figure 3: VGGNet architecture [6].

All experiments are trained for 300 epochs optimizing the cross-entropy loss while varying the optimizer used as well as the learning rate schedulers and maintain other parameters constant. The first experiment intends to find the best optimizer in training architecture. For this, six different algorithms of optimizers are considered: SGD, SGD with Nesterov Momentum, Average SGD, Adam, Adam with AMSGrad, Adadelta, and Adagrad. Two different variations are investigated in this experiment. In the first variation, all algorithms are run with a fixed learning rate. In the second variation, instead using decay learning rate. The validation accuracy attained by our model using the different optimizers was recorded. The model using the SGD with Nesterov momentum performs the best in both experiments attaining a validation accuracy of 73.2 % and 73.5 %.

The next experiment aimed is to find the optimal learning rate scheduler. In this case, the same architecture is retained, using the optimal optimizer decided by the previous experiment (the SGD with Nesterov momentum) with 5 different schedulers: Reduce Learning Rate on Plateau (RLRP), Cosine Annealing (Cosine), Cosine Annealing with Warm Restarts (CosineWR), One Cycle Learning

Rate (OneCycleLR), and Step Learning Rate (StepLR). All other parameters are maintained constant. The validation and testing accuracies attained by our models was recorded. The important thing to note is that Reducing Learning Rate on Plateau (RLRP) performs best. It achieves a validation accuracy of 73.59 % and a testing accuracy of 73.06 %.

To further increase our model's accuracy, the last experiment with hyper-tuning the final weights of our model after training with our optimal algorithms founded. The parameters are reloaded and train for a final 50 epochs using an initial learning rate . Cosine Annealing and Cosine Annealing with Warm Restarts are used as the learning rate schedulers. A another variation is made in which the validation set is combined into training to allow for a larger dataset set when tuning. The results shows that both models in two variation perform better after fine-tuning and perform the best training on the second variation. The final best model achieves an accuracy of 73.28 %, the highest testing accuracy attained ever.

Finally, this study achieves a new single-network state-of-the-art classification accuracy on FER2013 using a VGGNet. Tuning all hyperparameters was made towards an optimized model for facial emotion recognition. When using a saliency map for visualizing and judging by all the images, some advantages and some disadvantages of the model has been pointed out. Based on the disadvantages, it has opened up a new approach that if a model that can more effectively identify the facial features in an image and drop all useless information will perform even more better. Moreover, image pre-processing techniques need to be focused more and the method of ensemble of models can also be applied.

2.4 Resnet

The study [9] reviewed the state of the art in CNN based FER, highlighted key differences between the individual works, and compared and discussed their performance with a focus on the underling CNN architectures. On this basis, existing bottlenecks have been identified and consequently means for advancing the state of the art in this challenging research field.

Typically, with a dataset considered as small as FER2013, the use of shallow CNN structure is considered reasonable because it helps avoid the model from overfitting. At the same time, data augmentation, face registration, some form of illumination correction and using esemble of CNNs is also frequently applied associated with the use of shallow CNNs [3]. However, this study has proved that the small size of available FER datasets such as FER2013 is not the limiting factor. First, deeper networks do not necessarily have more parameters. Second, deeper networks impose a stronger prior on the structure of the learned decision function, and this prior will effectively combat overfitting. Third, modern deep CNNs still can achieve impressive results on datasets with a similar size, such as in CIFAR10 dataset. So the first bottleneck is that we have assumed CNNs do not have to be as deep for FER, that a CNN with shallow structure is already able to learn discriminative high-level features. This has limited the feasible space to search for a breakthrough solution to the FER problem and this study show that overcoming this bottleneck by employing modern deep CNNs leads to a significant improvement in FER2013 performance. (the result for FER2013 test accuracy accuracy of 75.2%). However, we still cannot reconstruct this result.

At the same time, the biggest bottleneck was pointed out that currently hinders FER performance is the fact that there is no publicly available dataset that is large enough for current deep learning standards. In contrast, FER2013, one of the largest FER image datasets available, has only 35,887 images. Compiling a large FER dataset is a laborious task due to the challenging annotation process, including assigning correct expression labels in presence of subtle expressions, partial occlusions, and pose variations is a challenging task for humans.

Finally, a demonstration that the use of an ensemble of such CNNs is also a key point outperforms state of the art methods without the use of additional training data or requiring face registration (usually used in shallow CNN structure). For the future, the analysis has left many new directions for investigating ways for overcoming these bottlenecks, such as a focus on FER specific data augmentation. Furthermore, studying the bias that affects the datasets and investigating the abibility of creating a new, more comprehensive, and publicly available FER dataset are also important for achieve new breakthroughs for FER problem [9].

3 Data Preparation

	FER2013	CK+	JAFFE	KDEF & AKDEF
ANGRY	4,953	135	30	710
DISGUST	547	177	31	710
FEAR	5,121	75	31	710
HAPPY	8,989	207	31	710
NEUTRAL	6,198	54	30	710
SAD	6,077	84	30	710
SURPRISE	4,002	249	30	710
TOTAL	35,887	981	213	4,970

Figure 4: Statistical information of four datasets

First, let's about the dataset that was used much for FER and throughout this report is FER2013. It was introduced at the International Conference on Machine Learning (ICML) in 2013 and became a benchmark in comparing model performance in emotion recognition. It is one specific emotion recognition dataset that encompasses the difficult naturalistic conditions and challenges. FER2013 is a large, publicly available dataset consisting of 35,887 face crops, it can be split into training, validation, and test sets with 28,709, 3,589, and 3,589 samples, respectively (adhering to the official training, validation, and test sets as introduced by the ICML) [6]. In this dataset, basic expression labels are provided for all samples. All images are grayscale and have a resolution of 48×48 pixels. The faces have been automatically registered so that the face is more or less centred and occupies about the same amount of space in each image (see Figure 5). The dataset is challenging as the depicted faces vary significantly

(a) (b)

Figure 5: (a) A training image for emotion happy and (b) a training image for emotion surprise in FER2013 dataset.

(a) (b)

Figure 6: (a) A training image for emotion fear and (b) a training image for emotion sad in CK+ dataset.

in terms of person age, face pose, and other factors. Human performance on this dataset is estimated to be 65.5 % [12].

If FER2013 is a data set consisting of facial emotion expressions under naturalistic conditions, the remaining 3 datasets are consisting of basic expressions under controlled conditions. And recognizing such basic expressions under controlled conditions (controlled in frontal faces and posed expressions) can now be considered a solved problem [9]. **The Extended Cohn-Kanade (CK+)** dataset contains 981 images from a total of 123 different subjects, ranging from 18 to 50 years of age with a variety of genders and heritage. These photos are frames cut from video sequences, each video shows a facial shift from the neutral expression to a targeted peak expression, a resolution of 48×48 pixels and they are all grayscale image (see Figure 6). Pictures are labelled with one of seven expression classes: anger, contempt, disgust, fear, happiness, sadness, and surprise. The CK+ database is widely regarded as the most extensively used laboratory-controlled facial expression classification database available, and is used in the majority of facial expression classification methods [7].

The JAFFE dataset consists of 213 images of different facial expressions from 10 different Japanese female subjects. Each subject was asked to do 7 facial expressions (6 basic facial expressions and neutral) and the images were annotated with average semantic ratings on each facial expression by 60 annotator. Each subject performed six basic emotions plus neutral (30 angry, 29 disgust, 33 fear, 30 happiness, 31 sad, 30 surprises and 30 neutral) in which each expression contains 3 to 4 images per subjects. All the facial images have been taken under strict controlled conditions of similar lighting and no occlusion such as hair or glasses. All the expression in frontal view and the resolution of the original image are 256×256 pixels and all in form 8-bit grayscale. They all have the original form of .tiff file [5].

(a) (b) (c) (d) (e)

Figure 7: Five different angles of a sample of fear emotion in KDEF dataset.

The Karolinska Directed Emotional Faces (KDEF) is a set of totally 4,900 pictures of human facial expressions. It contains 70 individuals, each displaying 7 different emotional expressions, each expression being photographed (twice) from 5 different angles (see Figure 7). **The Averaged KDEF (AKDEF)** is a set of averaged pictures created from the original KDEF images.

4 Methods

4.1 Proposed model

We propose a small CNN model which includes convolutional layers and Inverted Residual block in **MobileNetV2**. Real time model requires reasonable amount of parameters whereas the accuracy on the test set must still be high enough to give good result. We reduce the the number of parameters by using lightweight depthwise convolutions, combination of traditional convolutions, batch normalization and pooling helps model get richer feature for final prediction. At the

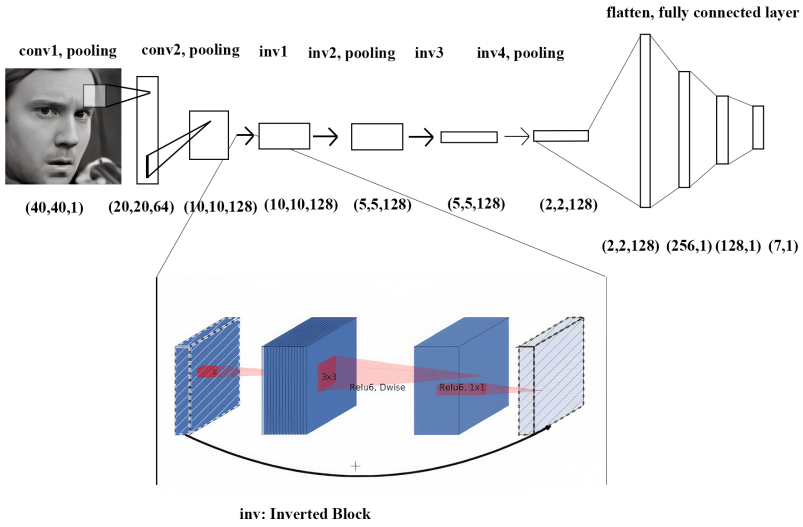


Figure 8: Our proposed model architecture.

first stage, two traditional convolution layers following by batch normalization

and activation function Relu are used to capture general features. Furthermore, the second layer uses kernel size 5×5 in order to get more enriched contextual features, the rest of model use kernel size 3×3. As mentioned before, Inverted blocks keep model can run realtime on weak devices without GPU because of the same size of block's input and output. In addition to this, max pooling layers use (2,2) size kernel and stride 2 help reduce computational cost hanks to decrease feature map size in half. Moreover, skip connection from input to output of block protects the model from the vanishing gradient problem. The authors of MobileNetV2 have argued that intermediate layers have responsibility to learn features by using none-linear transform so increasing the depth is essential. Four Inverted blocks have been used in case depth size of intermediate layers are 256, 256, 512, 512 respectively. After feature extraction part, flatten and fully connected layers are used to feed the feature vectors as classification part which has to predict labels. Using dropout does not make more difference on result cause of small model size that overfitting couldn't occur. Additionally, using double Inverted Residual blocks for each middle layer's size 256, 512 helps the model to capture completely high-level part feature, that increases the final accuracy.

5 Results

5.1 Metrics

Due to the model's confusion between fear, sad, and neutral labels can lead to poor results in reality recognition, we use Accuracy and F1-score [4] as metrics to evaluate how good of the model on private test set.

$$\text{Accuracy} = \frac{\text{total true predictions}}{\text{total predictions}} \tag{1}$$

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{2}$$

Where as:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{3}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{4}$$

When evaluating model, precision answers for the question: "How many positive predictions are really positive labels?". In model which would bring a bad result when make false positive predicting need a high precision. On the other hand, recall answer for the question: "How many reality positive labels are classified correctly?", that means the model has high recall will have very low level of missing positive label. For instance, model with high recall could be highly evaluated in Cancer prediction problem. F1-score is a balance metric between precision and recall, adjusting the model has the best fit F1-score is necessary for each type of model.

Total parameters	784,263
Estimated size	9.62 (MB)
Optimization	SGD
Criterion	Cross Entropy
Learning rate	0.01
Batch size	64

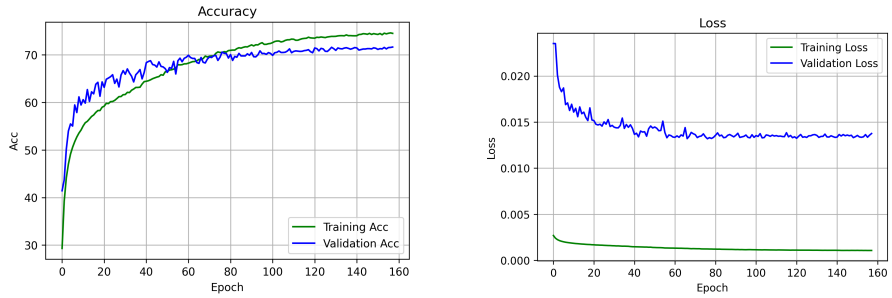
Table 1: Proposed model's details.

Besides, real-time model requires fast recognition enough as model has a high fps (frames per second). Because input of the model is face crop images, hence we need to preprocess by detecting faces and cropping before feeding into the model. Here, face detection with Haar Cascade is used to ensure lightly and fast requirement.

5.2 Experimental results

Our proposed model has a reasonable number of parameters, therefore training process on FER2013 data set was quite quickly, approximately 4 hours on Google Colaboratory environment with support of GPU.

After training 158 epoches, the model got 71.63 % accuracy on public test and got 68.52 % accuracy on private test set. The model's F1 score is 68.65 % on private test set.



(a) Accuracy on train and val set

(b) Loss on train and val set

Figure 9: Accuracy and Loss

In figure 9 shows the model's accuracy and loss after 158 epoches on train set and val set. Clearly, after epoch 100, the model shows signs of convergence, so both accuracy and loss are stable after this level and change little. We used the technique of reducing the learning rate to make the training process better. In the FER2013 dataset, the disgust label is unbalanced and less than other labels very much, leading to the model's results on this label being quite low and often misclassifying. On the other side, sad label could be confused as fear and neural class. This describes obviously in figure 10 confusion matrix. In experimental result, our model could run on system with no GPU and gained up to 10 fps with faces detection by Haar Cascade method [2].

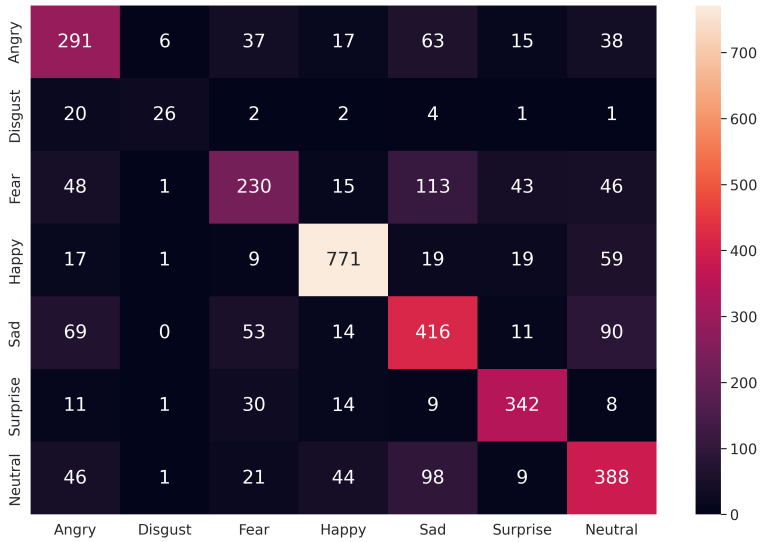


Figure 10: Confusion matrix.

6 Appendix

Source Code: Github

References

[1] Yusra Khalid Bhatti et al. "Facial Expression Recognition of Instructor Using Deep Features and Extreme Learning Machine". In: *Computational Intelligence and Neuroscience* 2021 (30th Apr. 2021). Ed. by Pasi A. Karjalainen, pp. 1–17. ISSN: 1687-5273, 1687-5265. DOI: 10.1155/2021/5570870. URL: https://www.hindawi.com/journals/cin/2021/5570870/.

[2] Li Cuimei et al. "Human face detection algorithm via Haar cascade classifier combined with three additional classifiers". In: *2017 13th IEEE International Conference on Electronic Measurement Instruments (ICEMI)*. 2017, pp. 483–487. DOI: 10.1109/ICEMI.2017.8265863.

[3] Ian J. Goodfellow et al. "Challenges in Representation Learning: A report on three machine learning contests". In: *arXiv:1307.0414 [cs, stat]* (1st July 2013). arXiv: 1307.0414. URL: http://arxiv.org/abs/1307.0414.

[4] Cyril Goutte and Eric Gaussier. "A Probabilistic Interpretation of Precision, Recall and F-Score, with Implication for Evaluation". In: vol. 3408. Apr. 2005, pp. 345–359. ISBN: 978-3-540-25295-5. DOI: 10.1007/978-3-540-31865-1_25.

[5] Miyuki Kamachi, Michael Lyons and Jiro Gyoba. "The japanese female facial expression (jaffe) database". In: *Availble: http://www. kasrl. org/jaffe. html* (Jan. 1997).

[6] Yousif Khaireddin and Zhuofa Chen. "Facial Emotion Recognition: State of the Art Performance on FER2013". In: (), p. 9.

[7] Patrick Lucey et al. "The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression". In: *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*. 2010, pp. 94–101. DOI: 10.1109/CVPRW.2010.5543262.

[8] Karnati Mohan et al. "FER-net: facial expression recognition using deep neural net". In: *Neural Computing and Applications* 33.15 (Aug. 2021), pp. 9125–9136. ISSN: 0941-0643, 1433-3058. DOI: 10.1007/s00521-020-05676-y. URL: https://link.springer.com/10.1007/s00521-020-05676-y.

[9] Christopher Pramerdorfer and Martin Kampel. "Facial Expression Recognition using Convolutional Neural Networks: State of the Art". In: *arXiv:1612.02903 [cs]* (8th Dec. 2016). arXiv: 1612.02903. URL: http://arxiv.org/abs/1612.02903.

[10] Sumeet Saurav, Ravi Saini and Sanjay Singh. "EmNet: a deep integrated convolutional neural network for facial emotion recognition in the wild". In: *Applied Intelligence* 51.8 (Aug. 2021), pp. 5543–5570. ISSN: 0924-669X, 1573-7497. DOI: 10.1007/s10489-020-02125-0. URL: https://link.springer.com/10.1007/s10489-020-02125-0.

[11] Melissa N. Stolar et al. "Real time speech emotion recognition using RGB image classification and transfer learning". In: *2017 11th International Conference on Signal Processing and Communication Systems (ICSPCS)*. 2017 11th International Conference on Signal Processing and Communication Systems (ICSPCS). Surfers Paradise, QLD: IEEE, Dec. 2017, pp. 1–8. ISBN: 978-1-5386-2887-4. DOI: 10.1109/ICSPCS.2017.8270472. URL: http://ieeexplore.ieee.org/document/8270472/.

[12] Lixu Wang et al. *Eavesdrop the Composition Proportion of Training Labels in Federated Learning*. 2019. arXiv: 1910.06044 [cs.LG].