AI Programming Project AIP391
July 2022

Real Time Facial Expression Recognition

Nguyễn Thanh Đảm; Nguyễn Phúc Nguyên Anh; Nguyễn Lê Phương Hà

Advisor: Nguyễn Quốc Trung

Abstract: Real Time Facial Expression Recognition is also an important topic in the field of Artificial Intelligence (AI). And that's exactly what we've been working on in the AI Programming Project (AIP391) for the past 10 weeks. This report is intended to document our exploration of some models to solve that problem. We did our best to make it, we hope it is useful.

Contents

1. Introduction

Facial expression recognition (FER) is a technology that can analyze signs, expressions or features extracted from images or video frames, thereby distinguishing emotions on human faces. Real time facial expression recognition (FER) is really a challenging project, because emotions are complex and they change quickly and continuously, it is difficult for even humans to recognize them and easily confuse them. However, FER is an interesting project, FER plays an important role and has many applications in many fields as well as in life. 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 [1]. Not only that, it is especially useful in other situations. 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 [2].

Although it is said that human emotions are complex and difficult to understand, it is possible for computers to recognize basic emotions (anger, disgust, fear, happiness, sadness, and surprise) under favorable conditions. Under favorable conditions such as sufficient light, clear frontal images (head pose), fine features, and so on; identification is easier. Under natural conditions, it is uncertain and can be missed. This is also a challenge to overcome to apply FER to real applications.

Nowadays, Deep Learning (DL) models are powerful tools to handle large amounts of data. One of the most popular deep neural networks is Convolutional Neural Networks (CNN). With the emergence of the convolutional neural network, many scholars tend to use the convolutional neural network to extract image features [3]. Of course, there are many other methods, but here we mainly refer to CNN.

2 Related work

2.1 AlexNet

AlexNet is a convolutional neural network (CNN) architecture, suggested by Alex Krizhevsky, Ilya Sutskever and Geoffrey Hinton in 2012. The AlexNet 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 (Rectified Linear Unit is a function with a fast computation speed thanks to the derivative having only two values {0, 1} and without the exponentiation of base e like the sigmoid function but still non-linear). 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 from which the input can be classified. (see Figure 1)

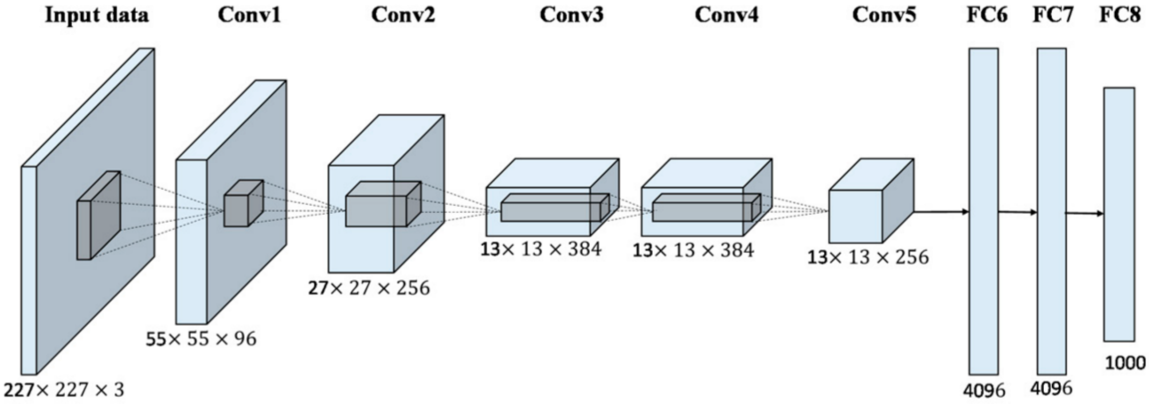


Figure 1: Structure of the AlexNet

The dataset FER-2013 (kaggle) contains original images that are normalized to 48x48 pixels in grayscale. In paper [4], with the dataset FER-2013 (kaggle), using Alexnet, the mentioned accuracy is 61.0%. In the case of FER-2013, the literature shows an overall low trend in accuracy because of high variance and occlusion conditions of the dataset. This is clearly demonstrated in the Data Preparation. In another paper [1], the accuracy is really amazing, up to 77.0%. This is really hard to believe and raises many questions: data preprocessing, data augmentation, and so on.

2.2 ResNet

Residual Neural Network (ResNet) was introduced to the public in 2015, was a commonly used architecture. Currently, there are many variations of ResNet architecture with different numbers of layers such as ResNet-18, ResNet-34, ResNet-50, ResNet-101, ResNet-152, and so on. Previous architectures often improve accuracy by increasing the depth of CNN. But experiments show that up to a certain depth threshold, the accuracy of the model will saturate and even backfire and make the model less accurate. When traversing too many layers of depth can cause original information to be lost, Microsoft researchers solved this problem on ResNet by using shortcuts. Skip connections or shortcuts are used to jump over some layers. (see Figure 2)

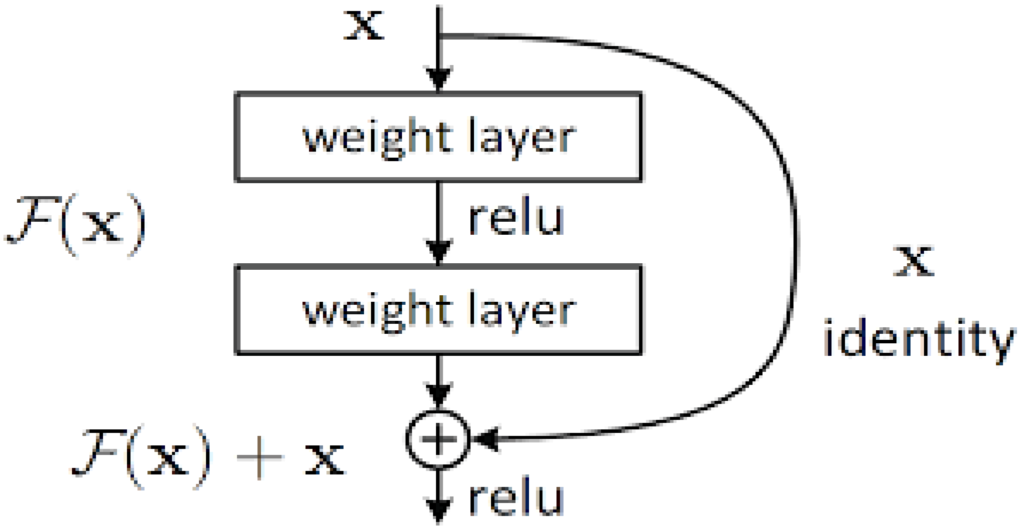

x
weight layer
$\mathcal{F}(\mathbf{x})$
relu
x
identity
weight layer
$\mathcal{F}(\mathbf{x}) + \mathbf{x}$
relu

Figure 2: Skip connections or shortcuts

In paper [3], the dataset was collected from 20 subjects of different ages, different careers and different races, including seven kinds of facial emotions pictures: happy, sadness, fear, anger, surprise, disgust, and neutral. In the final, they have 700 images in total. Although this dataset is not large, it also contributes to the richness of data for the projects. They use the current popular convolutional neural network algorithm, combined with the ResNet-50 residual network (see Figure 3), which has achieved a good effect in the multi-classification task. And the mentioned accuracy is very high, up to 95.39%. This is also understandable because the dataset is small and other factors occurring under experimental conditions.

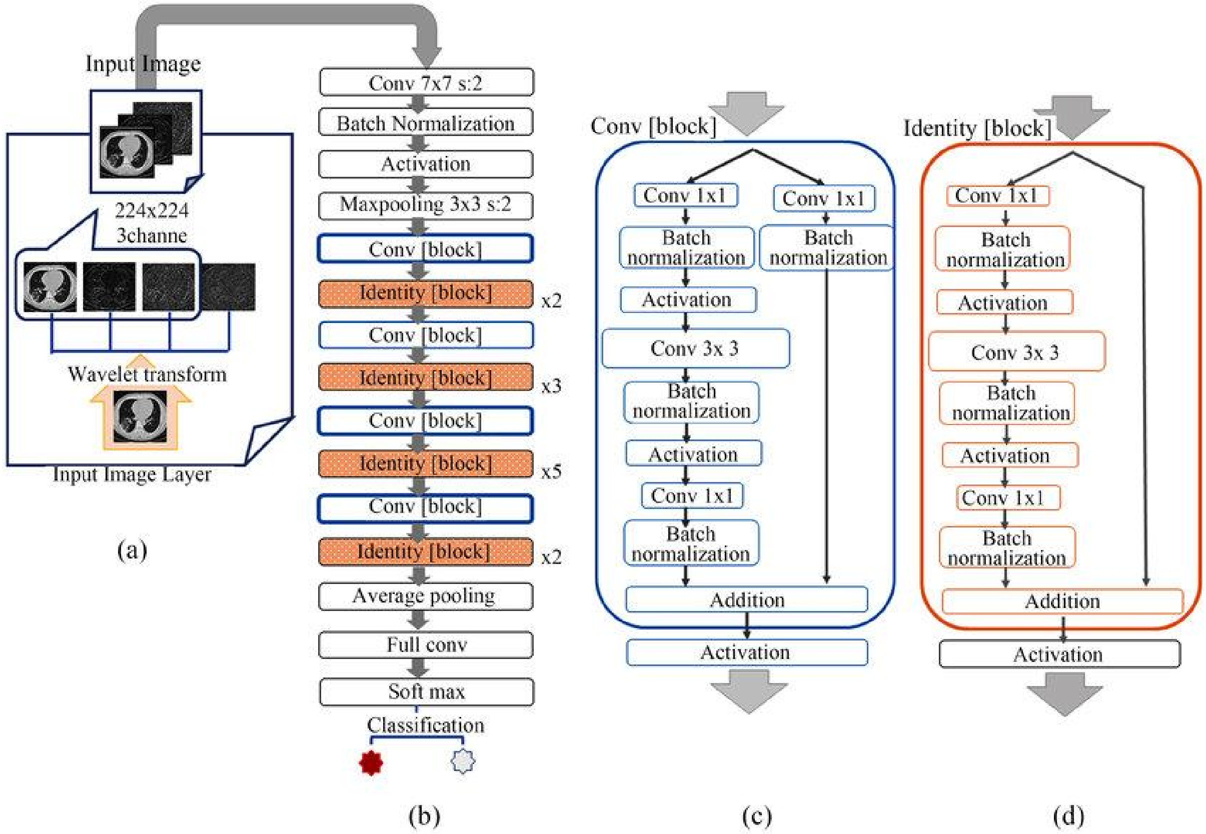


Figure 3: Structure of the ResNet-50

2.3 FerNet

Even though several 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. The proposed FER-net is specifically designed in order to learn the detailed local features like eyes and mouth corners that are exhibited by different Facial Expressions (FE) in face images. Micro-FEs play an important role in FER. Besides, traditional CNN-based methods suffer from the overfitting problem on small datasets. However, datasets with reliable expressions are relatively difficult to collect and tend to be small [1].

In paper [1], FER-net (see Figure 4) 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.

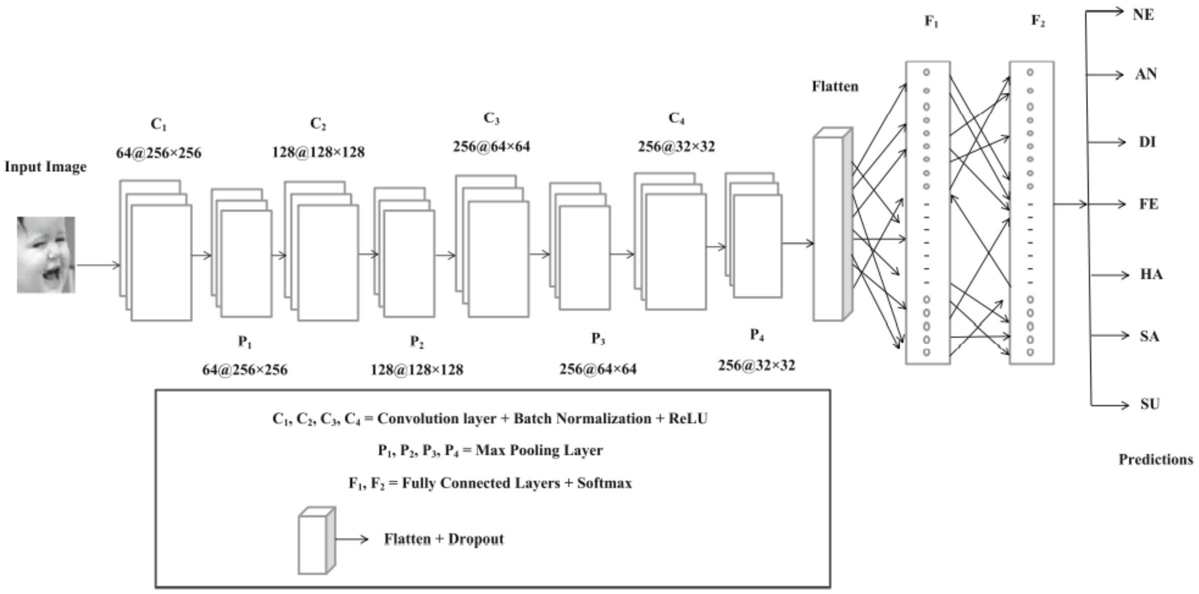


Figure 4: Structure of the FER-net

In this study, five publicly available benchmarking datasets, namely FER2013, JAFFE, CK+, and KDEF, and RAF, are considered to validate the FER-net. All the images are of 256x256 pixels except images present in FER2013 and RAF dataset. So, all the images of FER2013 are resized from 48x48 to 256x256 pixels using bilinear interpolation in the preprocessing step. Similarly, all the images of the RAF dataset are resized from 100x100 to 256x256 using the same algorithm. The accuracy of the model FER-net is respectively for JAFFE (97%), CK+ (98%), and KDEF datasets (83%). However, the performance is satisfactory for the other two datasets FER2013 (up to 79%) and RAF dataset (82%).

2.4 Mini_Xception Net

Mini-Xception model which is an enhanced model of Xception architecture using residual networks for Emotion expression and Recognition. Using the FER-2013 database gives better performance than the existing method. There are seven types of emotion expressions such as (e.g., anger, disgust, fear, happy, sad, surprise, and neutral.) we tried to recognize. In modern years, many works have introduced an end-to-end plan for emotion expression recognition, utilizing deep learning models. Although emotion recognition is a great task, it still seems emotion huge area for development. The accuracy obtained for Emotion expression and recognition using Mini_Xception is 95.60% and precision and recall rate is 93% and 90% respectively. Further the accuracy can be increased by training the Mini-Xception algorithm using the original image dimensions of 48x48. The number of Convolutional layers and the size of filter can also be increased to improve accuracy.

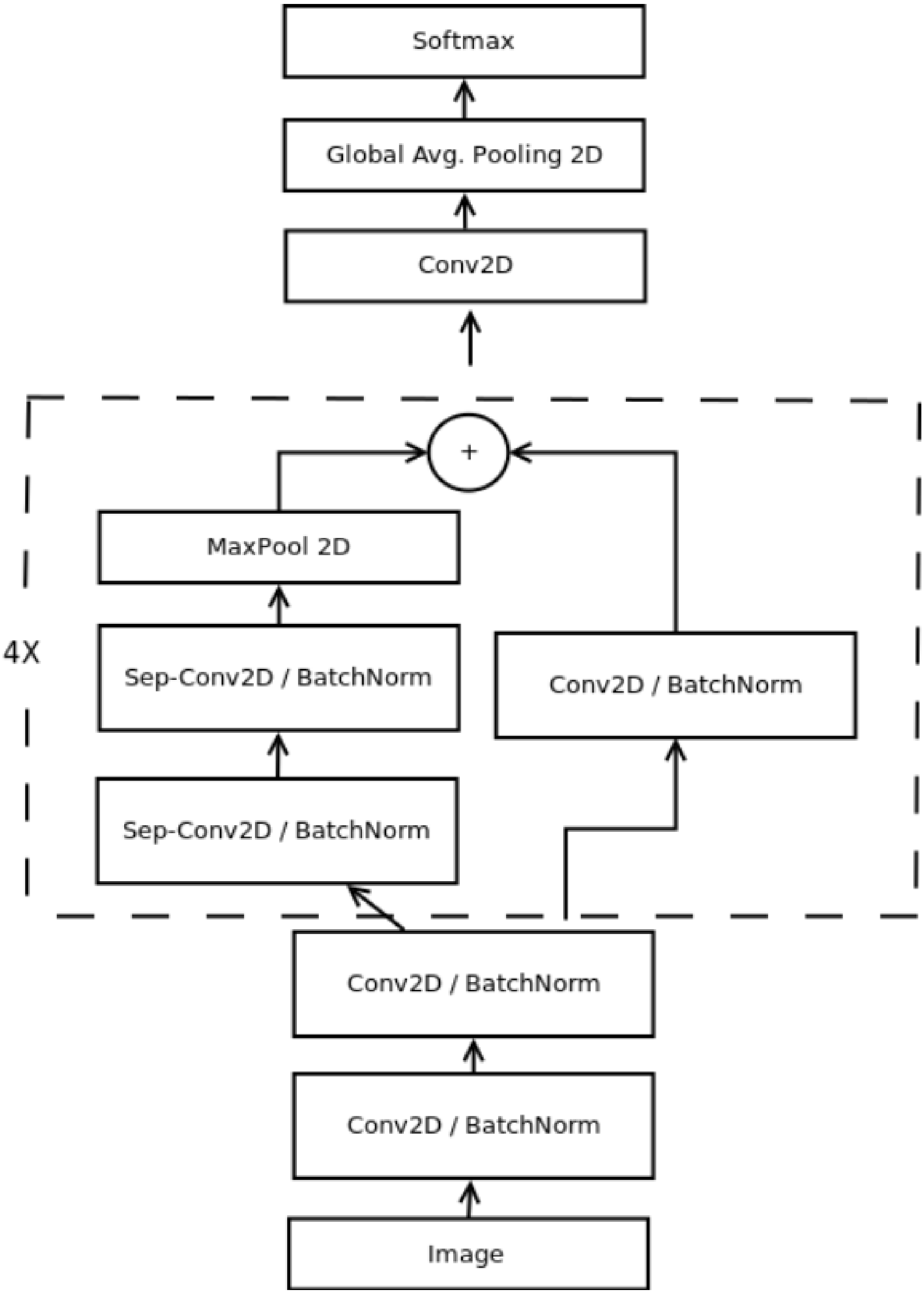
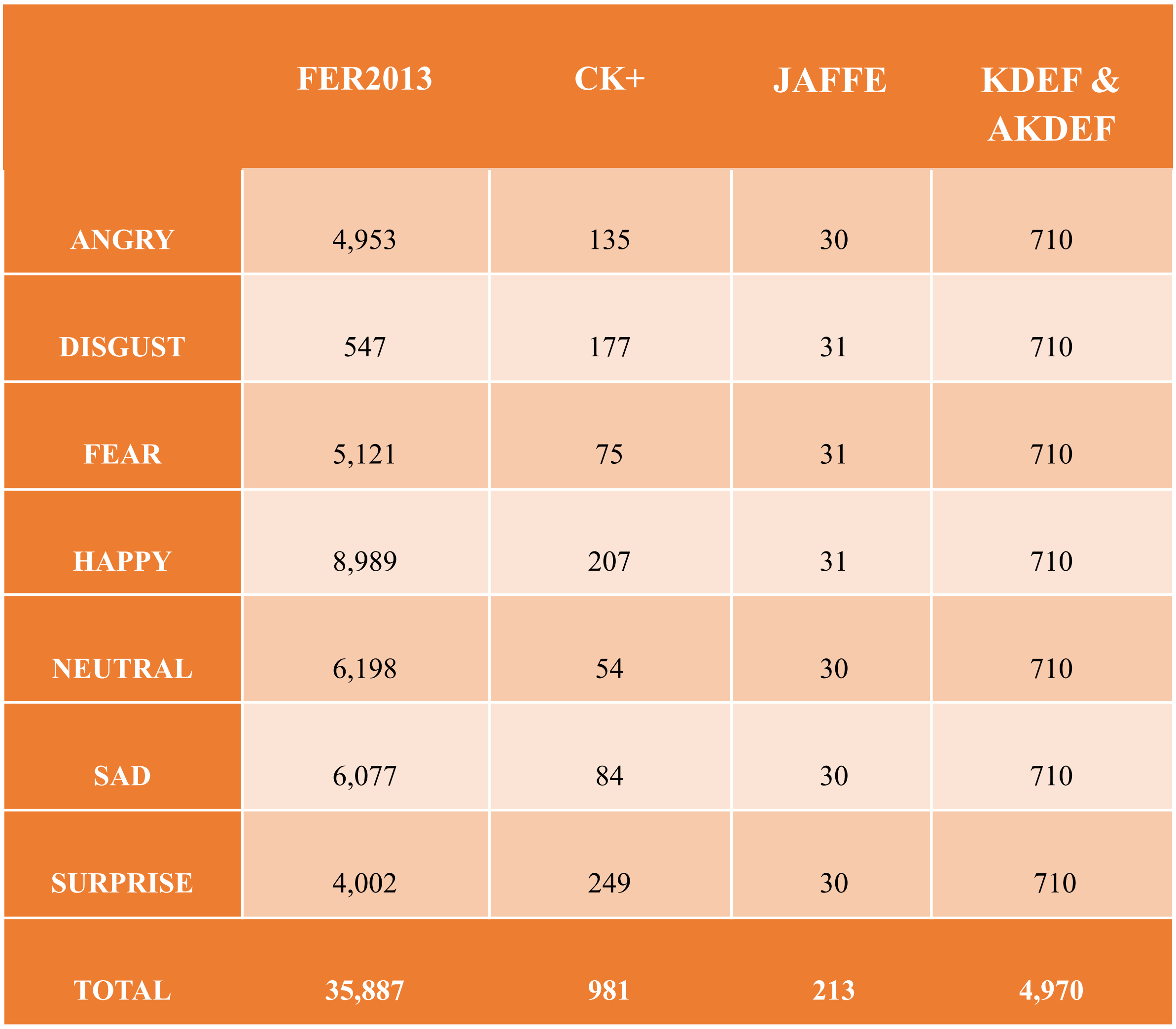


Figure 5: Our proposed model for real-time classification.

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 6: Statistical information of four datasets

FER is a well-studied field with numerous available datasets. We used FER2013 as our main dataset and drove up accuracy on its test set by using CK+, RaFD and MMI as auxiliary datasets. We also created our own web app dataset to tune our models to work better in real world scenarios.

FER2013(Facial Expression Recognition 2013 Dataset)

First, let's talk 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 well-studied dataset and has been used in ICML competitions and several research papers. It is one of the more challenging datasets with human-level accuracy only at 65.5% and the highest performing published works achieving 75.2% test accuracy. Easily downloadable on Kaggle, the dataset's 35,887 contained images are normalized to 48x48 pixels in grayscale. FER2013 is, however, not a balanced dataset, as it contains images of 7 facial expressions, with distributions of Angry (4,953), Disgust (547), Fear (5,121), Happy (8,989), Sad (6,077), Surprise (4,002), and Neutral (6,198). If FER2013 is a data set consisting of facial emotion expressions under naturalistic conditions, the remaining 3 datasets consist 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.

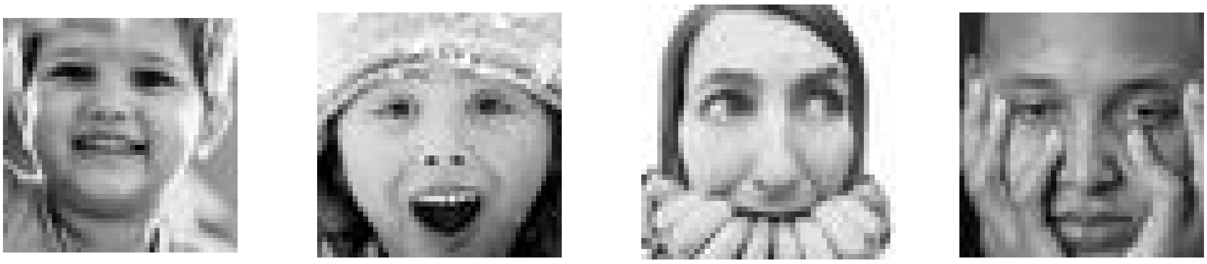


Figure 7: Happy, Surprise, Fear, Sad picture in FER2013

CK+ (Extended Cohn-Kanade dataset) dataset contains 593 video sequences from a total of 123 different subjects, ranging from 18 to 50 years of age with a variety of genders and heritage. Each video shows a facial shift from the neutral expression to a targeted peak expression, recorded at 30 frames per second (FPS) with a resolution of either 640x490 or 640x480 pixels. Out of these videos, 327 are labeled 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.

JAFFE (Database.Japanese Female Facial Expression) database was taken from publicly available data which consists of 213 facial expression images of 10 subjects of Japanese female. 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. The image has the resolution in grayscale. 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 x 256 pixels

KDEF (Database. Karolinska Directed Emotional Faces) database is another publicly available dataset consists a set of 4900 facial expression images. It contains 70 individuals, each displaying 7 different emotional expressions, each expression being photographed (twice) from 5 different angles.

4 Methods

4.1 Proposed model

VGGNet is a classical convolutional neural network architecture used in large-scale image processing and pattern recognition [5]. Our variant of VGGNet is shown in Figure 1. 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. Batch normalization is used here to speed up the learning process, reduce the internal covariance shift, and prevent gradient vanishing or explosion [6]. The first two fully connected layers are followed by a ReLU activation. The third fully connected layer is for classification. The convolutional stages are responsible for feature extraction, dimension reduction, and non-linearity. The fully connected layers are trained to classify the inputs as described by extracted features.

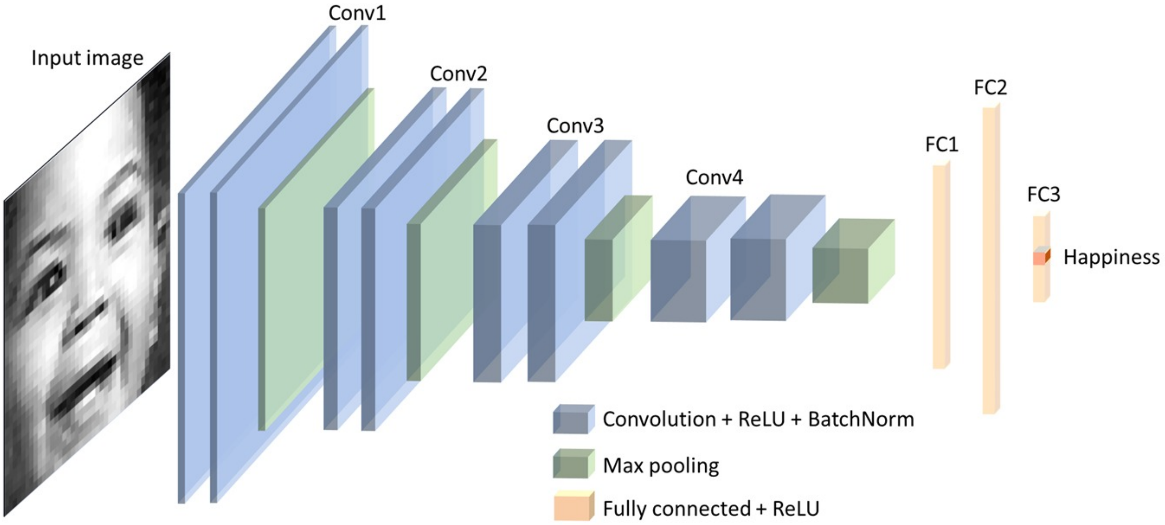


Figure 8： VGGNet architecture. A face expression image is fed into the model. The four convolutional blocks (Conv) extract high-level features of the image and the fully-connected (FC) layers classify the emotion of the image

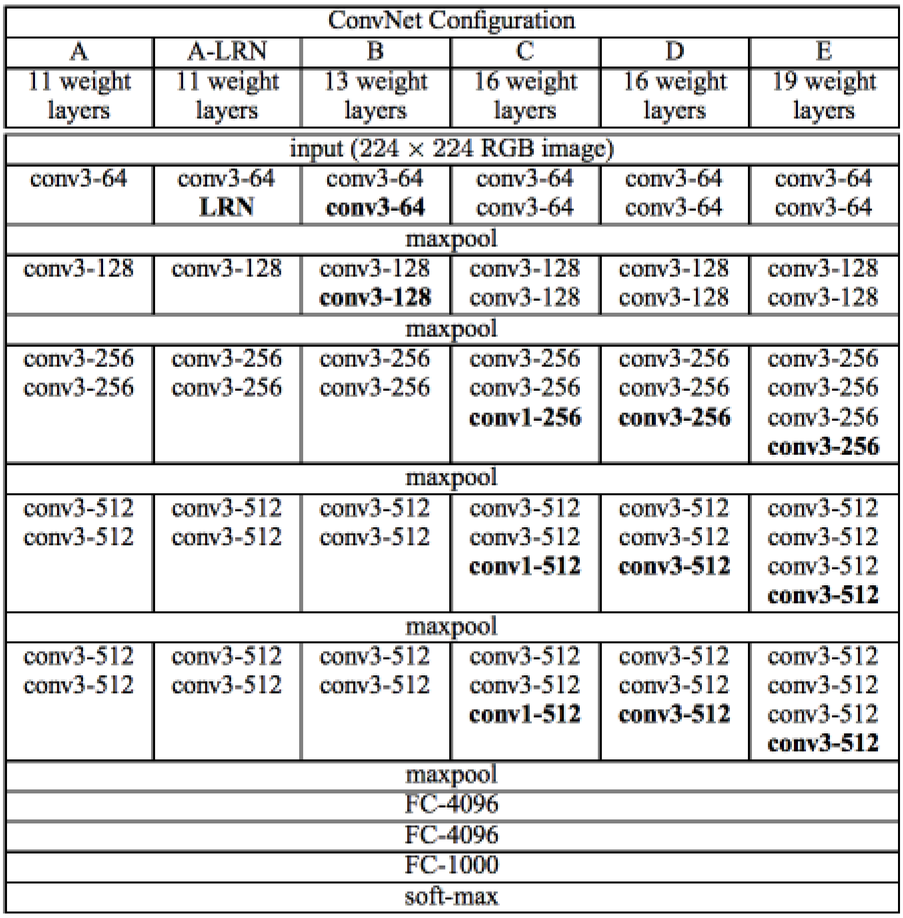
4.2 Architecture

During training, the input to VGGNet is a fixed-size 224 × 224 RGB image. The only pre-processing we do is subtracting the mean RGB value, computed on the training set, from each pixel. The image is passed through a stack of convolutional (conv.) layers, where we use filters with a very small receptive field: 3 × 3 (which is the smallest size to capture the notion of left/right, up/down, center). In one of the configurations we also utilise 1 × 1 convolution filters, which can be seen as a linear transformation of the input channels (followed by non-linearity). The convolution stride is fixed to 1 pixel; the spatial padding of conv. layer input is such that the spatial resolution is preserved after convolution, i.e. the padding is 1 pixel for 3 × 3 conv. layers. Spatial pooling is carried out by five max-pooling layers, which follow some of the conv. layers (not all the conv. layers are followed by max-pooling). Max-pooling is performed over a 2 × 2 pixel window, with stride 2 . A stack of convolutional layers (which has a different depth in different architectures) is followed by three Fully-Connected (FC) layers: the first two have 4096 channels each, the third performs 1000-way ILSVRC classification and thus contains 1000 channels (one for each class). The final layer is the soft-max layer.

4.3 Configurations

An VGGnet architecture in model similar to VGG-B in Table 1 [7] but with one CCP block less. We also use dropout after each such block (this improved the validation accuracy by around 1%). The backend consists of a single hidden layer with 1024 units

VGGnet have less parameters than any of the architectures used in the pertinent literature, despite being significantly deeper.Even the very deep ResNet has fewer parameters than most of these architectures. We did not specifically search for architectures that perform well on FER2013. Our goal is to confirm that modern deep architectures generally perform well, not to obtain the absolute best accuracies on this dataset.



ConvNet Configuration					
A	A-LRN	B	C	D	E
11 weight layers	11 weight layers	13 weight layers	16 weight layers	16 weight layers	19 weight layers
input (224 × 224 RGB image)					
conv3-64	conv3-64 **LRN**	conv3-64 **conv3-64**	conv3-64 conv3-64	conv3-64 conv3-64	conv3-64 conv3-64
maxpool					
conv3-128	conv3-128	conv3-128 **conv3-128**	conv3-128 conv3-128	conv3-128 conv3-128	conv3-128 conv3-128
maxpool					
conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256 **conv1-256**	conv3-256 conv3-256 **conv3-256**	conv3-256 conv3-256 conv3-256 **conv3-256**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
FC-4096					
FC-4096					
FC-1000					
soft-max					

Table 1: ConvNet configurations (shown in columns). The depth of the configurations increases from the left (A) to the right (E), as more layers are added (the added layers are shown in bold). The convolutional layer parameters are denoted as "conv ⟨ receptive field size ⟩ - ⟨ number of channels ⟩ ". The ReLU activation function is not shown for brevity.

Network	**A,A-LRN**	**B**	**C**	**D**	**E**
Number of parameters	133	133	134	138	144

Table 2: **Number of parameters** (in millions).

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.

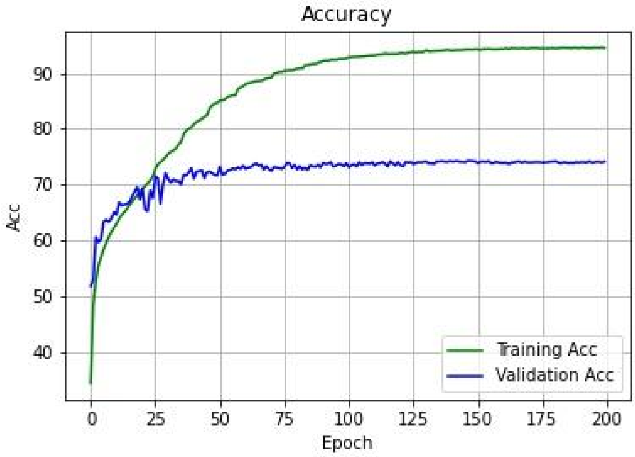
Total parameters	784,263
Estimated size	9.62 (MB)
Optimization	SGD
Criterion	Cross Entropy
Learning rate	0.01
Batch size	64

Table 3: Proposed model's details.

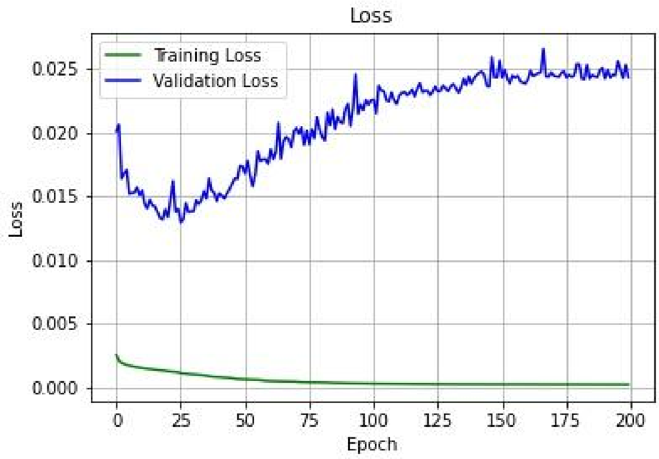
Besides, real-time model requires fast recognition enough as model has a high fps (frames per second). Because input of the model is face crop images, hence we need to preprocess by detecting faces and cropping before feeding into the model. Here, face detection with Haar Cascade is used to ensure lightly 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 200 epoches, the model got top 1 accuracy 73.28 % accuracy, top 2 accuracy 86.45% on public test and got 69.52 % accuracy on private test set. The model's F1 score is 73.27 % on private test set.



(a) Accuracy on train and val set



(b) Loss on train and val set
Figure 9: Accuracy and Loss

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 [8].

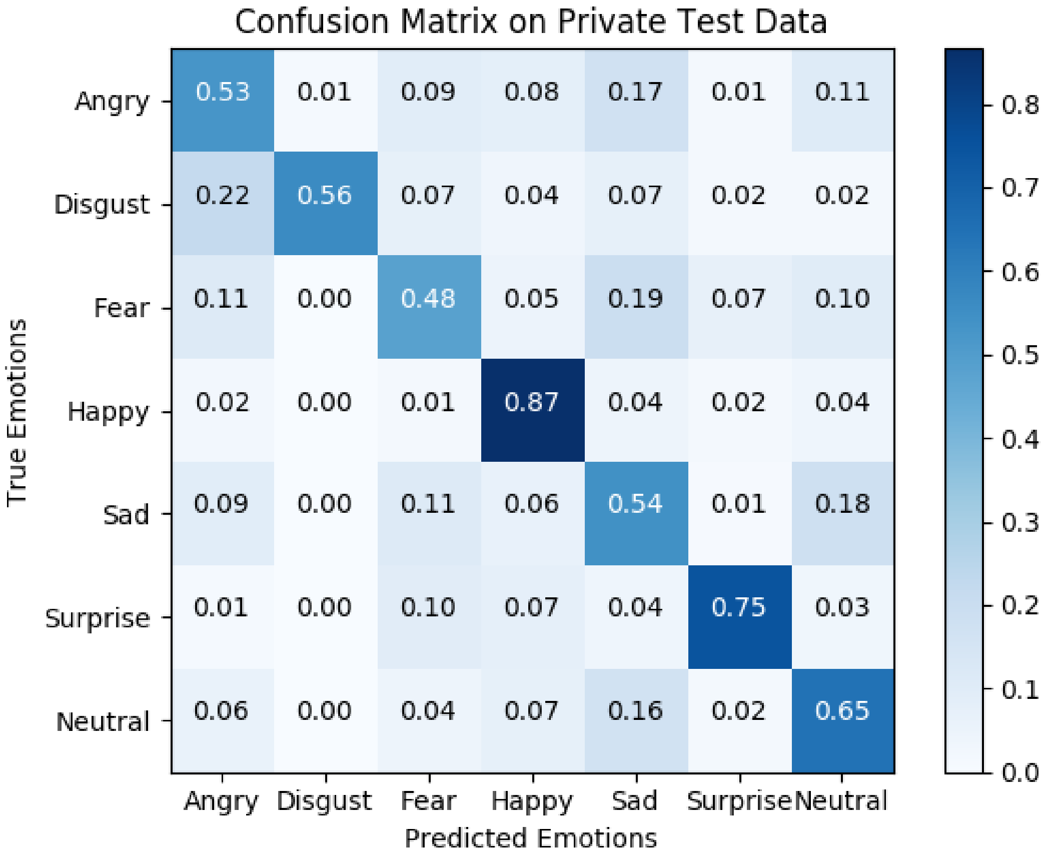


Figure 10: Confusion matrix.

6 Appendix

Item	Link
Source Code	FER project
Teamwork Plan Management	Plan Management

Contribution Rate

Member	**Research** (%)	**Attention** (%)	**Ideas** (%)	**Implement** (%)	**Attitude** (good/bad)
Nguyễn Phúc Nguyên Anh	100%	100%	100%	90%	Good
Nguyễn Thanh Đảm	100%	100%	100%	100%	Good
Nguyễn Lê Phương Hà	100%	100%	100%	90%	Good

Comment	Our teamwork is strong overall. However, we overcame the disagreements that arose during the project and the subpar equipment to clearly complete our work.

1.

References

[1] 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.

[2] 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.

[3] Li, Bin and Lima, Dimas. “Facial expression recognition via ResNet-50”. In: International Journal of Cognitive Computing in Engineering 2.0 (Feb. 2021), pp. 57-64. issn: 2666-3074. doi: 10.1016/j.ijcce.2021.02.002. url: https://www.sciencedirect.com/science/article/pii/S2666307421000073.

[4] 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. byPass A. Karjalainen, pp. 1–17. issn: 1687-5273, 1687-5265. doi: 10.1155/2021/5570870. url: https://www.hindawi.com/journals/cin/2021/5570870/.

[5] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” in 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings, 2015.

[6] E. Pranav, S. Kamal, C. Satheesh Chandran, and M. H. Supriya, “Facial Emotion Recognition Using Deep Convolutional Neural Network,” in 2020 6th International Conference on Advanced Computing and Communication Systems, ICACCS 2020, 2020, doi: 10.1109/ICACCS48705.2020.9074302.

[7] Simonyan, K. and Zisserman, A. (2015) Very Deep Convolutional Networks for Large-Scale Image Recognition. The 3rd International Conference on Learning Representations (ICLR2015). https://arxiv.org/abs/1409.1556

[8] 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.