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. [6] 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. [8]

- **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.* [8] Recently, FER has been *widely studied* and *significant progress has been made in this field.* 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**. In fact, reliable FER under naturalistic conditions is **mandatory** in the aforementioned applications, therefore need to *be solved effectively.* [7]

- It is *noticeable* that **the detailed local features** like eyes and mouth corners that *are exhibited by different FEs in face images.* These micro-expressions occur in everyone, often *unconsciously* and *in an unnoticeable manner to an interlocutor.* Such Micro-FEs *play an important role in FER.* These features are **essential** for identifying the emotion of individual subjects. When **using handcrafted features**, there are *clear shortcomings* that limit the performance in identifying FE because of *extracting accurately all the correlated handcrafted features is very challenging.* [6] 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.** [5]

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**)

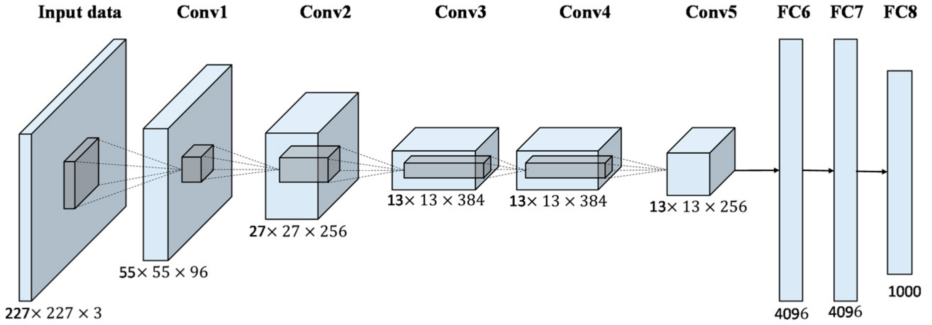


Figure 1: *Structure of the AlexNet used for ImageNet showing five convolutional layers Conv1 - Conv5 and three fully connected layers FC6 - FC8. The last layer provides an output to the Softmax function*

Input size **is fixed** due to the presence of fully connected layer, which takes 227 × 227 2-dimensional input. During the experiment, **the number of classes of output needs to be adjusted** to seven corresponding to seven basic facial emotions mentioned. With the **FER2013 dataset**, in [1], original images with size 48×48 are used **bilinear interpolation** to match the input vector size requirements. Then use the Alexnet network structure for training the parameters. In paper [1] the result received was **61.0%** accuracy and in [9] the accuracy is **61.1%** (note that the above results are only in the related work and not the main content of each paper). At the same time, data augmentation is not used when training. This result *is also reproduced* by us. However, in paper [6], the accuracy result of Alexnet is up to **77%**, raises questions about the problem of *reproducing the results in the paper* or *there was a new improvement in image pre-processing not mentioned in the paper*

One thing that should be noticed through reading the papers is that the network structures for recognizing facial emotions are **quite simple** and **quite similar**, even the network structures have good results (One reason that can be explained is that **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 probability search algorithms such as genetic 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. These datasets consist of **seven basic FEs**, namely **NE, AN, DI, FE, HA, SA,** and **SU,** which are classified by the FER-net compared with *twenty-one state-of-the-art models.* The FER-net extracts feature from face regions automatically, then these features are fed to a **softmax classifier** for identifying FEs. It is clear from the obtained results that the proposed model is **superior to the state of the art in almost all cases** except broad learning. Moreover, the proposed model is **simple as compared to state-of-the-art models** and is **preeminent in terms of accuracy, as well as execution time.** [6]

The paper **reviewed various methods in face detection** and **the use of handcraft feature and Machine Learning tools for emotion classification**, the paper also pointed out that **the difficulty of using this method** is *to extract accurately all the correlated handcrafted features due to the effect of variations caused by emotional state.* Thus, handcrafted features *have clear shortcomings that limit their performance in identifying FEs.* The article also **mentioned a large number of quite simple DL models** and **summary the main ideas of them**. It also point out that the FER2013 dataset is **a difficult set** and show that it is **unclear** from the literature *whether legacy works performing well on datasets trained in lab-controlled environments would provide satisfactory performance on real-time datasets such as Facial Expression Recognition 2013 (FER2013).* Furthermore, **traditional CNN-based methods** will *suffer from the overfitting problem on such small and difficult datasets.* However, it is **noticeable** that *datasets with reliable expressions are relatively difficult to collect and tend to be small.*

The proposed FER-net is **specifically designed in order to learn the detailed local features like facial corners** that are exhibited by different FEs in face images. It also point out that *more complex networks are usually able to learn deep features*, but they often result in **overfitting** as *the number of involved parameters is high*, so it is *reasonable* to *use a such simple CNN network to classify static expressions that perform well on small datasets.* The detailed architecture of FER-net is explained in **Figure 2**. It consists of **four convolution layers (C1, C2, C3, and**

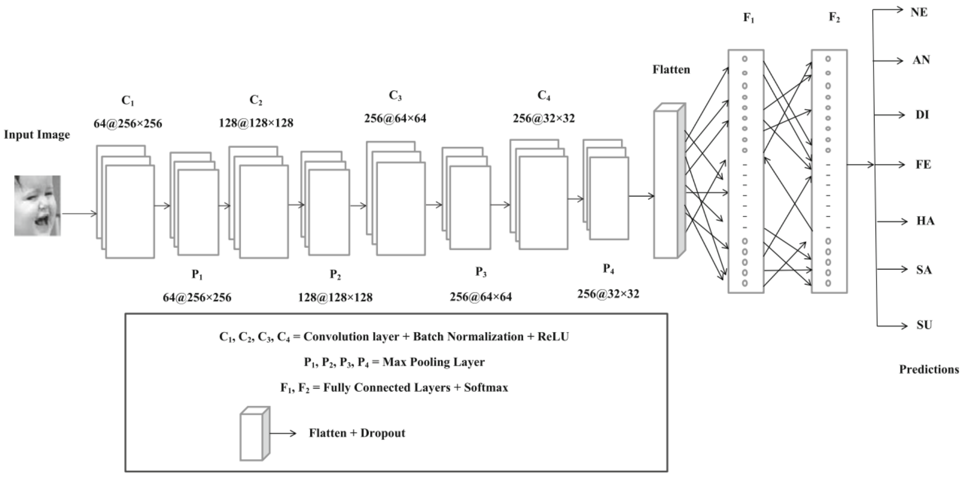
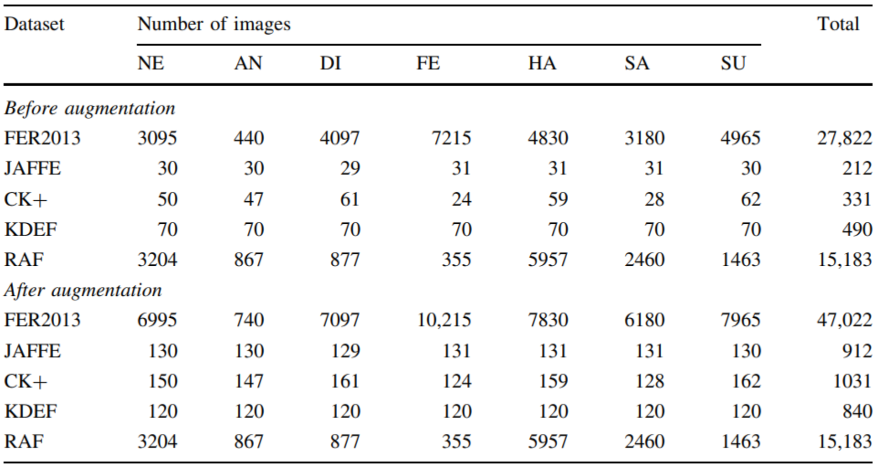


Figure 2: *Schematic block diagram of the FER-net*

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 added to *overcome the over-fitting problem* by *shutting down some of the neurons in FER-net while training.* 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 by the reasons behind.* In contrast, in many other papers, as the corresponding papers *lack details on architecture selection*, *the reason for this discrepancy is unknown.* [7]. Specifically, **Batch normalization** is applied to *normalize the output of the input layer and hidden layers* by *adjusting the mean and the scale of the activation functions* because *a high learning rate can be achieved without causing a vanishing gradient problem* by virtue of the batch-normalization, therefore gives better performance after the activation function. **Pooling operation** is applied to convolved feature maps obtained to *reduce the overfitting problem.* The pooling is also known as **subsampling**. It is able to *reduce the spatial representation of an image* by *reducing the number of parameters associated with CNN.*

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. Then training FER-net using the train set all the images of **size**

256 × 256 **pixels** (as in the structure of FER-net), if the original image in some datasets is not satisfactory in size then **using bilinear interpolation** to resize the images in the pre-processing step. Image processing techniques such as *translation, scaling, rotation, flipping the images vertically and horizontally, and adding noise to the images* are applied to *increase the size of the datasets* (**Data augmentation**). (see **Figure 3**)



Dataset	Number of images							Total
	NE	AN	DI	FE	HA	SA	SU	
Before augmentation								
FER2013	3095	440	4097	7215	4830	3180	4965	27,822
JAFFE	30	30	29	31	31	31	30	212
CK+	50	47	61	24	59	28	62	331
KDEF	70	70	70	70	70	70	70	490
RAF	3204	867	877	355	5957	2460	1463	15,183
After augmentation								
FER2013	6995	740	7097	10,215	7830	6180	7965	47,022
JAFFE	130	130	129	131	131	131	130	912
CK+	150	147	161	124	159	128	162	1031
KDEF	120	120	120	120	120	120	120	840
RAF	3204	867	877	355	5957	2460	1463	15,183

Figure 3: *Statistical information of five datasets*

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.* Finally, testing sets are fed into trained FER-net one after another in order to obtain **confusion matrices** .Then confusion matrices *help to calculate the values of metrics, which are used further to analyze the performance of FER-net.* 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 **(83%)** datasets 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** [6]. 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%**). At the same time, *compared to the res-*

ults presented in this paper when comparing performance of different models, we also *could not reproduce that results.* Is there *a new data pre-processing method that has outperformed the other*?

2.3 VGGnet

The paper [5] 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** to achieve state-of-the-art results at a testing accuracy of **73.28%**. This is **a remarkable result** and it has opened up *a new approach that adopting existing models and rigorously tuning it.* The paper also points out and concludes from other studies **the development points of neural networks so far** as well as **their importance** (including **optimizer** and **learning rate scaler**). In the other hand, **the state-of-the-art methods for FER** is also mentioned **along with a summary of the main ideas of each case study** (using **ensemble of models**, using **SVM**, using **a novel amend representation module (ARM)** to substitute the pooling layer, ...). However, **the key point** was noticed is that in order to *improve the ensemble performance or any different combined methods*, **the building blocks of these ensembles, a single network need to be aimed for optimizing first.**

Using only the FER2013 dataset for training and **adhering to the official training, validation, and test sets as introduced by the ICML** (this dataset was introduced at the **International Conference on Machine Learning (ICML)** in 2013 and became *a benchmark in comparing model performance in emotion recognition*). Because of *the lack of diversity in dataset* as well as *the unevenness in the number of samples per class*, **data augmentation** is expected to be *an important step* towards increasing performance when using FER-2013 dataset [7]. Therefore, **a significant amount of data augmentation** is applied in training. This augmentation includes **rescaling the images** up to ± **20 % of its original scale**, **horizontally and vertically shifting** the image by up to ± **20 % of its size**, and **rotating it up to ± 10 degrees**. *Each of the techniques is applied randomly* and with a probability of **50 %**. After this, the image is then **ten-cropped to a size of** 40×40, and **random portions of each of the crops are erased** with a probability of **50 %**. Finally, each crop is then **normalized by dividing each pixel by 255.**

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 4**. 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 to *speed up the learning process, reduce the internal covariance shift*, and *prevent gradient vanishing or explosion*). 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.*

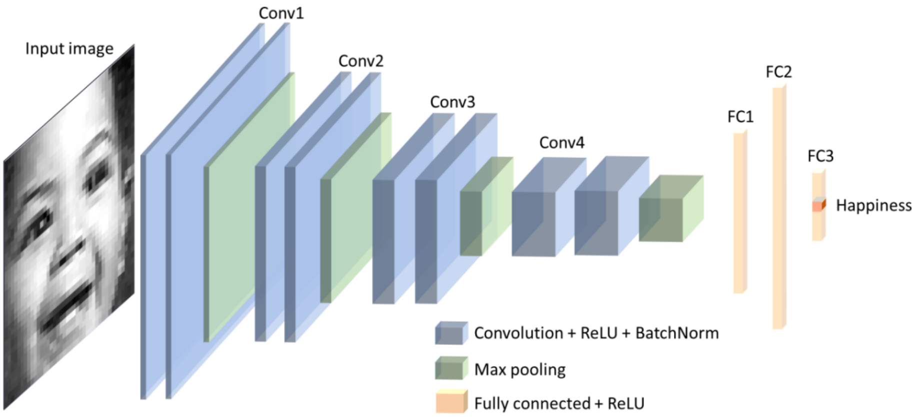


Figure 4: *VGGNet architecture. A face expression image is fed into the model. The four convolutional blocks (Conv) extract high-level features of the image and the fully-connected (FC) layers classify the emotion of the image*

As mentioned, 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.* We use a fixed **momentum of 0.9** and **a weight decay of 0.0001**. The models are evaluated using **validation accuracy** and tested using **standard ten-crop averaging. The first experiment intends** to *find the best optimizer in training our 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, we run all algorithms with *a fixed learning rate of 0.001.* In the second variation, instead using a simple learning rate scheduler with *an initial learning rate of 0.01* and *it is reduced by a factor of 0.75 if the validation accuracy plateaus for 5 epochs.* All parameters of this scheduler were also determined using **a grid search**. **Figure 5** shows **the validation**

accuracy attained by our model using the different optimizers. Excluding **Adadelta**, all optimizers show **a high validation accuracy**, **above 70 %**, considered as *a remarkable result.* 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 section (**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).** For a baseline, **a constant learning rate was used (0.01)** that was determined using **a grid search**. *All other parameters are maintained constant.* **Figure 6** shows the **validation and testing accuracies attained by our models**. 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 %**. This is *already surpassing the previous single-network state-of-the-art performance mentioned before.* 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 of 0.0001**. This learning rate is set to be small *to maintain the update steps small, thus, ensuring that our model's weights are not skewed far away, since this is an already trained model.* This experiment is run using **Cosine Annealing and Cosine Annealing with Warm Restarts learning rate**

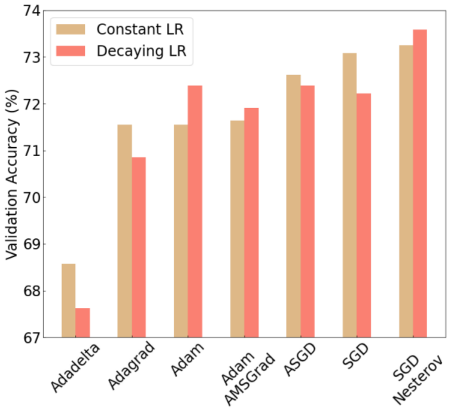


Figure 5: *VGGNet performance using different optimizers. The yellow bars show the results using a constant learning rate (LR). The orange bars show accuracies using decaying LR*

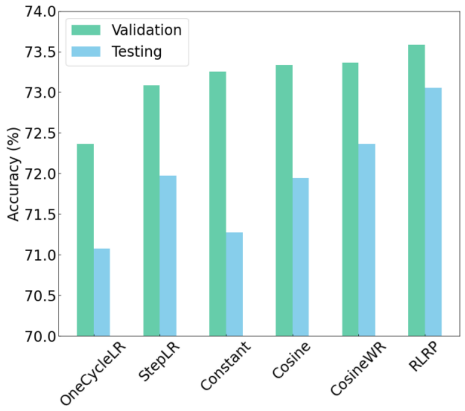
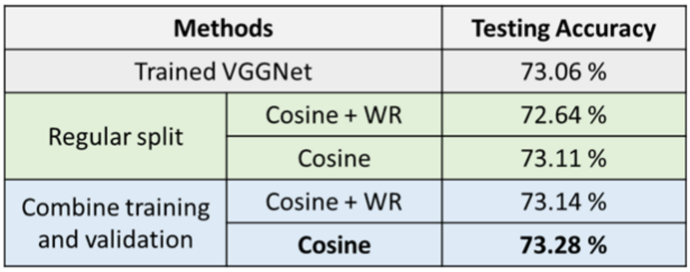


Figure 6: *VGGNet performance using different LR schedulers. The green bars show the final validation accuracies and the blue ones show the corresponding testing accuracies.*

schedulers since both of these schedulers *slowly oscillate the learning rate back and forth thus not allowing for major weight changes.* **A another variation of this experiment** is made in which **the validation set is combined into training** to allow for *a larger dataset set when tuning.* This larger dataset would *allow the model to have more samples to learn from, thus, improving its performance.* The test set and all other parameters are *kept the same.* By running two variations of this experiment, two things can be verified. **Using the first variation**, *the effectiveness of the tuning can be confirmed.* **Using the second experiment**, *the benefits of added data can be proved.* **Figure 7** show the results. As expected, **both models in two variation perform better after training and perform the best training on the combined dataset resulting from the training and validation data**. The final best model achieves an accuracy of **73.28 %**, *the highest testing accuracy attained ever.*



Methods		Testing Accuracy
Trained VGGNet		73.06 %
Regular split	Cosine + WR	72.64 %
	Cosine	73.11 %
Combine training and validation	Cosine + WR	73.14 %
	Cosine	**73.28 %**

Figure 7: *VGGNet performance using fine-tuning.*

Finally, this paper achieves **single-network state-of-the-art classi-**

fication accuracy on FER2013 using a VGGNet. **Tuning all hyperparameters was made** towards an optimized model for facial emotion recognition. *Different optimizers and learning rate schedulers are explored* and the best initial testing classification accuracy achieved is **73.06 %**, surpassing all single-network accuracies previously reported. *Extra fine-tuning on our model using Cosine Annealing* and *combine the training and validation datasets* are also carried out to further improve the classification accuracy to **73.28 %**, the best accuracy attained. The model also shows that **FER2013 dataset** is *unbalanced in the number of samples in each class*, *the low classification accuracy in "disgust" and "fear"* was attributed to the fact that *it have a lower number of samples in the original training set.* Beside that, *the misclassification in recorded result between "fear" and "sadness"* can be reasoned by *the inter-class similarities of the dataset.* When using **a saliency map** for visualizing this neural networks, (By *propagating the loss back to the pixel values*, **a saliency map** can *highlight the pixels which have the most impact on the loss value*, thus highlights *the visual features the CNN can capture from the input*). Judging by all the images, this CNN **can effectively capture most of the critical regions for each emotion**. The model is **placing a large importance on almost all facial features of the person** in each image and also **effectively dropping regions like the background** which are not very informative when it comes to describing someone's emotion. There are **some clear mistakes** in the saliency maps where *the model highlights some of the background pixels.* It 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 hasn't been focused much** and **the method of ensemble of models has not yet been applied**. From here, **many other approaches have opened up**, as well as the main idea of this algorithm, which is *to inherit and further develop the powerful existing algorithms.*

2.4 Resnet

The paper [7] **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 (this has been pointed out from the review of some of the state of the art CNN structures based FER) [3]. However, the paper 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** (to make model become *overfitting*). Second, deeper networks *impose a stronger prior on the structure of the learned decision function*, and this prior will *effectively combat overfitting* (compare to shallow structure). Third, modern deep CNNs **still can achieve impressive results on datasets with a similar size**, such as **CIFAR10**. 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* (The evidence is remarkable results using these structures). This has **limited the feasible space to search for a breakthrough solution to the FER problem** and the paper show that **overcoming one such bottleneck by employing modern deep CNNs leads to a significant improvement in FER2013 performance.** (the result for FER2013 test accuracy accuracy of **75.2%**, *performing comparably to the current best method recorded*). 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**. The introduction of datasets with hundreds of thousands or millions of images *will enable significant performance gains in related research fields such as face recognition.* 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, this analysis has left **many new directions for investigating ways for overcoming these bottlenecks**, such as **a focus on FER specific data augmentation** (Most analyzed works use standard augmentation methods that are not specific to FER). Furthermore, **studying the bias that affects FER2013 and other datasets** and **investigating the possibility of creating a new, more comprehensive, and publicly available FER dataset** are also important for achieve new breakthroughs for FER problem. [7]

3 Data Preparation

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 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 5x5** in order to *get more enriched contextual features*, **the rest of model use kernel size 3x3**. 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

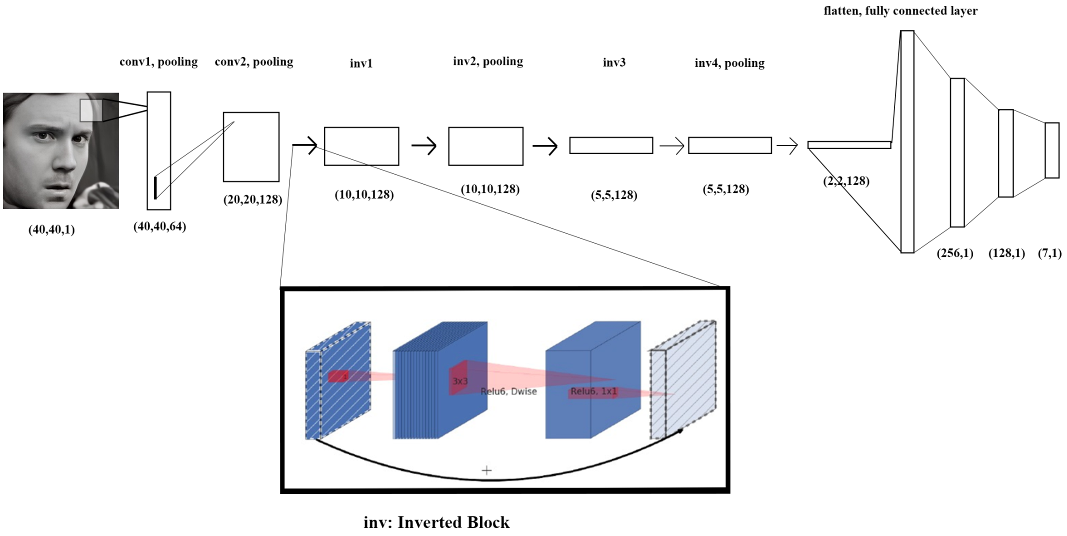


Figure 8: Our proposed model architecture.

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.

$$Accuracy = \frac{total\_true\_predictions}{total\_predictions} \tag{1}$$

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

Where as:

$$Precision = \frac{TruePositive}{TruePositive + FalsePositive} \tag{3}$$

$$Recall = \frac{TruePositive}{TruePositive + FalseNegative} \tag{4}$$

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.

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.*

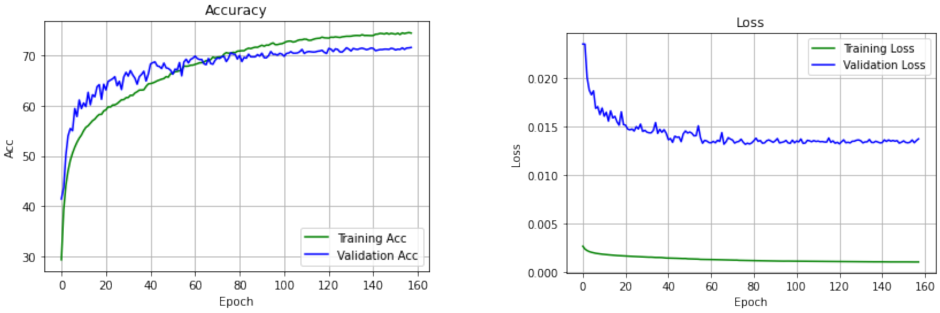
Total parameters	784,263
Estimated size	9.62 (MB)
Optimization	SGD
Criterion	Cross Entropy
Learning rate	0.01
Batch size	64

Table 1: Proposed model's details.

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.

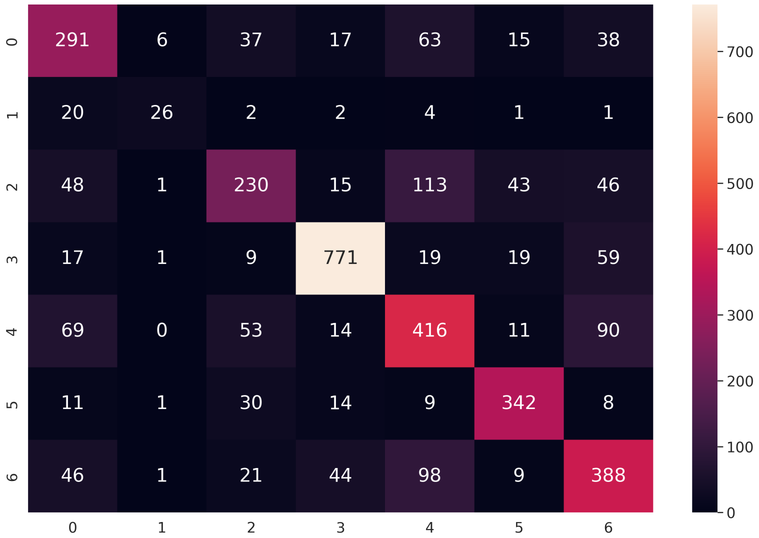


(a) Accuracy on train and val set

(b) Loss on train and val set

Figure 9: Accuracy and Loss

In the FER2013 dataset, **the disgust label is unbalanced and less than other labels**, 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**. 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].



	0	1	2	3	4	5	6
0	291	6	37	17	63	15	38
1	20	26	2	2	4	1	1
2	48	1	230	15	113	43	46
3	17	1	9	771	19	19	59
4	69	0	53	14	416	11	90
5	11	1	30	14	9	342	8
6	46	1	21	44	98	9	388

Figure 10: Confusion matrix.

6 References

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

[6] 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.

[7] 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.

[8] 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.

[9] 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/.

7 Appendix