



Dept. of Information & Communication Technology
Islamic University, Kushtia-7003, Bangladesh

Real time Face detection and Emotion and Gender Classification Using Convolutional Neural Network

A Thesis paper submitted in partial fulfillment of the requirement for the degree of
Bachelor of Science in Information & Communication Technology of the
Islamic University, Kushtia-7003, Bangladesh.

Supervised by:

Dr. Paresh Chandra Barman

Professor

Dept. of Information & Communication
Technology

Islamic University, Kushtia-7003,
Bangladesh.

Submitted by:

Md. Abu Rumman Refat

Roll:1218005

Reg:1264

Session:2013-14

Dept. of Information &

Communication Technology

Islamic University, Kushtia-7003,
Bangladesh.

March,2019

Artificial Intelligence Lab (AIL)

Dept. of Information & Communication Technology

Islamic University, Kushtia-7003, Bangladesh.

Dedicated
To... ..
My Parents
&
Teachers

March, 2019
Islamic university, Kushtia-7003, Bangladesh

Certificate of the Supervisor

I am pleased to certify that Md Abu Rumman Refat, Examination Roll No:1218005, Registration No:1264 has performed a Project work entitled **“Real time Face detection and Emotion and Gender Classification Using Convolutional Neural Network”** under my supervision in the academic year 2016-17 for the fulfillment of partial requirement of Bachelor of Science (Honor’s) Degree. So far as I am concerned this is an original Project work that he carried out for one year in the Department of Information & Communication Technology, Islamic University, kushtia-7003, Bangladesh.

I strongly declare that this thesis has not been copied from any other Research of submitted elsewhere prior to this department.

.....
Dr. Paresh Chandra Barman
Professor & Supervisor
Dept. of Information & Communication Technology
Islamic University, Kushtia-7003, Bangladesh

March, 2019
Islamic university, Kushtia-7003, Bangladesh

ACKNOWLEDGEMENT

First, thanks to Almighty Allah, the creator and sustainer who has given me strength and opportunity to complete the Project work entitled “**Real time Face detection and Emotion and Gender Classification Using Convolutional Neural Network**”. Regarding the outcome of my Project, I would like to express my deepest sense of gratitude and respect to my supervisor Dr. Paresh Chandra Barman, Professor, Dept. of Information & Communication Technology, Islamic University, Kushtia-7300 for his valuable suggestion to select the topic and constant guidance, great supervision, advice, encouragement and other fruitful help throughout the duration of my project. His dedication, Collaboration and interaction were key factors in the success of my Project. Without his active support and great supervision, I would not be able to complete my Project. I am also grateful to my respectable teachers, Dept. of Information & Communication Engineering for their encouragement and continued support and help throughout the research.

Finally, My profound respect and thanks to my parents, family, friends and well wiser for their constant sacrifice encouragement and support over the years.

Md. Abu Rumman Refat

ABSTRACT

We implement a general Convolutional Neural Network (CNN) for designing for real and validated our model by creating a real time vision with accomplishes the task of face detection, gender and emotion classification simultaneously. We got accuracies of 95% in the IMDB-/WIKI age and gender dataset and 66% in the FER emotion recognition dataset. We have used real time guided back propagation technique to visualize the weighted real time CNN that uncovered the dynamic weight change and evaluate the learning feature. We think in the modern CNN architecture regularization and visualization of previous hidden layer features are necessary in order to reduce the gap between slow performances and real time architecture.

Keyword: Machine Learning, Convolutional Neural Network (CNN), Computer Vision (CV),

CONTENTS

Acknowledgement	4
Abstract	5
List of figures	8
List of tables	10

Chapter – 1 Introduction

1.1 Introduction	12
1.2 Application	12
1.3 Objective and Approach	13
1.4 Literature Review of Related Research	14
1.5 Project Overview	14
1.5.1 Contributions	14
1.5.2 Project Organization	15

Chapter – 2 Background

2.1 Introduction	17
2.2 Face Detection	17
2.3 Image Classification	18
2.3.1 Gender Classification	18
2.3.2 Emotion Classification	18

Chapter – 3 Convolutional Neural Network

3.1 Introduction	20
3.2 Basic Operations on CNN	20
3.2.1 Convolution Layer	21
3.2.2 Non-Linearity (ReLU)	22
3.2.3 Pooling of Sub Sampling	23
3.2.4 Fully Connected Layer	24

Chapter – 4 Proposed Method

4.1 Introduction	29
4.2 Workflow	29
4.3 Dataset Preparation	30
4.4 Pre-Processing	30
4.4.1 Conversion of the RGB to Gray-Scale	30
4.4.2 Binarization	30
4.4.3 Noise Reduction	31
4.5 Proposed CNN Architecture	31

Chapter – 5 Experiment Result and Discussion

5.1 Dataset	37
5.2 Experiment	38
5.2.1 Experiment Result	39
5.3 Performance Evaluation and Discussion	43

Chapter – 6 Conclusion

6.1 Conclusion	44
6.2 Limitations	44
6.3 Future Work	46

Bibliography	47
---------------------	----

List of Figures

1.1	General approach for real time face detection, emotion recognition and gender classification.	13
3.1	Basic CNN Architecture with an example image of boat in the final stage the probability of boat is maximum define the right classification.	20
3.2	Basic Convolution Operation (a) 7 x 7 image matrix (b) 3 x 3 filter matrix (c) Result 5 x 5 feature vector after convolution of a with b.	21
3.3	ReLU activation function.	22
3.4	Max Pooling operation of 4x4 matrix with 2x2 filter produces a 2x2 output with maximum value.	23
3.5	Fully Connected layer with 5 output, all previous mode is connected with every node of FC Layer.	24
3.6	Gradient descent, Problem arises could not reach the target point easily due to high learning rate and low learning rate.	25
4.1	Working principle diagram of our proposed approach.	29
4.2	Our proposed model for real-time classification.	30
4.3	Difference between (a) standard convolutions and (b) depth-wise separable convolutions	31
5.1	Samples of the FER-2013 emotion dataset	37
5.2	Samples of the IMDB dataset.	37

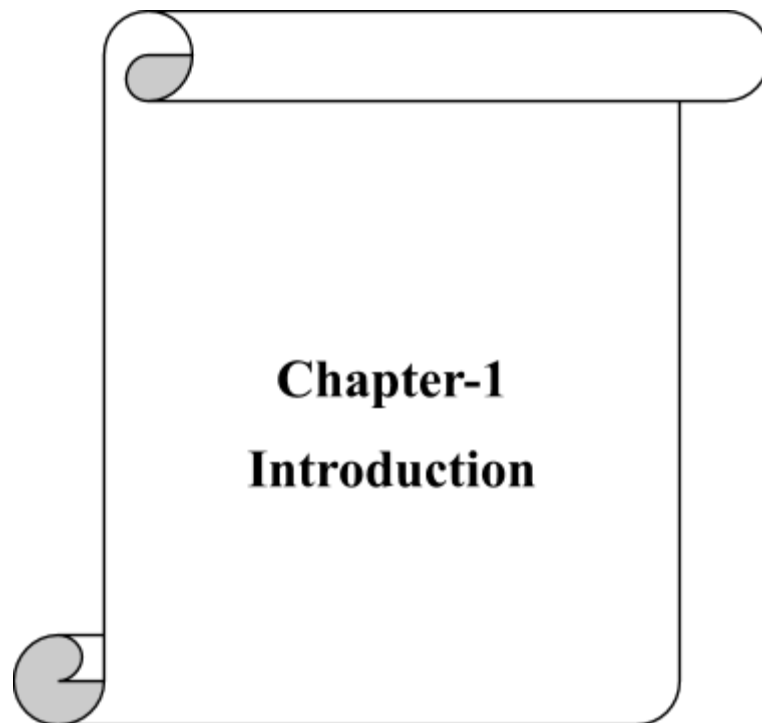
5.3	Predicted results of the emotion and gender classification on testing data. (a) Result of gender classification on IMDB dataset (b) result of emotion recognition on FER-2013 dataset.	38
5.4	Results of the real-time emotion classification	39
5.5	Results of the combined gender and emotion recognition. The color blue represents the assigned class woman and red the class man.	39
5.6	Normalized confusion matrix of our miniXception network (a) confusion matrix for facial emotion recognition (b) confusion matrix for gender recognition.	40
5.7	All sub-figures contain the same images in the same order. Every row starting from the top corresponds respectively to the emotions: angry, happy, sad and surprise (a) Samples from the FER-2013 dataset (b) Guided back-propagation visualization of our proposed mini-Xception model.	41
5.8	Results of the misclassification in real-time emotion and gender recognition.	42

List of Tables

1.1	List of Abbreviation	11
3.1	steps in the training process of CNN.	27

Table 1: List of Abbreviation

ANN	Artificial neural network
CNN	Convolution Neural Network
DL	Deep Learning
FER	Facial Emotion Recognition
ML	Machine Learning
NN	Neural Network
CV	Computer Vision



1.1 Introduction

Over the last decade, the rate of image uploads to the internet has grown at nearly an exponential rate. This new-found wealth of data has empowered computer scientists to tackle problems in computer vision that were previously either irrelevant or intractable. Consequently, we have witnessed the dawn of highly accurate and efficient facial detection e.g. identify the emotional state or deduce gender. Interpreting correctly any of these elements using machine learning (ML) technique has proven to be complicated due the high variability of the sample within each task [5]. This leads to models with millions of parameters trained under thousands of samples [4]. Furthermore, the human accuracy for classifying an image face in one of seven different emotions is $65 \pm 5 \%$ [2].

Moreover, the state-of-the-art methods in images related tasks such as image classification [2] and object detection are all based on Convolutional neural networks (CNN). These tasks require CNN architecture with millions of parameters, therefore their real time systems become unfeasible. For this we proposed an implement a general CNN building for designing real time CNNs that implementation has been validated in a real time facial expression system that provides face detection, gender classification and that achieves human level performance when classifying emotion.

1.2 Application

There is a huge amount of image data in the world and growth of itself is increasing. Facial expression recognition likes gender classification and emotion recognition is a robust system that can be used in the following section.

- ☐ Medical research
- ☐ Security
- ☐ Targeted Marketing
- ☐ Augmented Reality
- ☐ Better selfies!

1.3 Objective and Approach

Nowadays computer vision research is more relaxed in the practical application of ranging object detection, image classification. Gender classification and Facial emotion recognition have dawned more interest in practically because of many applications in ranging of human behavior understanding, mental disorders detection, synthetic human expressions etc.

That being said, this problem is also a really difficult. In fact, this problem is usually split into different sub-problem to make easier to with mainly face detection in an image that can be performed some tasks in between such as frontolizing face or extracting additional from an image.

Our main target is to provide a robust system that can perform some of the following tasks like human performance like facial emotion recognition and gender classification that are capable of working with any kind of images and real time scenario with a human face. Finally, we have to establish a benchmark for the task based on state-of-the-art network architectures and show that chaining the prediction. of gender with that of emotion can improve overall accuracy. We are working with a back propagation algorithm and soft max activation function and ReLUs.

Figure 1.1 shows a general approach of facial emotion recognition and gender classification. Human face is first given to the system as image input. The input image is first processed to detect faces and remove noise. This is done by many filters and data augmentation. After processing human face from our real time input image feature is extracting from the image face in order to differentiate it with others that is done by the classification part of our architecture.

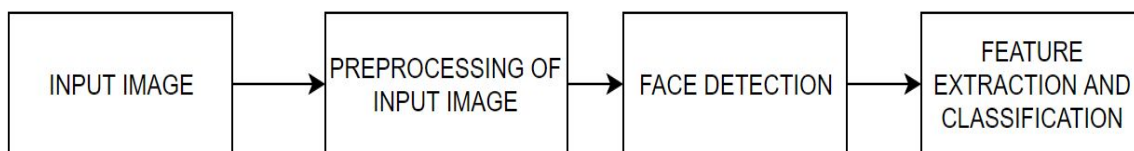


Figure 1.1: General approach for real time face detection, emotion recognition and gender classification.

1.4 Literature Review of the Related Research

Several works have been done so far for real time face detection, facial emotion recognition and gender-age classification [1,4,5]. For this project, we reviewed the current literature on convolutional face detection and gender and classification and facial emotion recognition [1,2,3,4,5,6,7]. We found that convolutional face detection and gender/emotion classification is still evolving as a technology, despite outranking other face detection and gender classification methods. By virtue of free availability of datasets and pretrained networks, it is possible to create a functional implementation of a deep neural network without access to specialist hardware. Pretrained networks can also be used as a starting point for training new networks, decreasing costly training time such as vgg16 [9] and inception v3 [10].

1.5 Project Overview

We divide our project overview part into two section one is our contribution and other is project organization. In the next section we describe each of the two sections.

1.5.1 Contributions

The main contribution of face detection and emotion, gender classification is summarized in this project. Our main job is to model an architecture with a modern convolutional neural network (CNN) that is able to reduce parameters because complex architecture and more parameters will take more time to train and high computational power is required. In more complex architecture there is a probability of over fitting because of parameters. Finally, we implement a back guided propagation technique to visualize how neurons are tuned in each layer so we can fit our model properly.

Furthermore, we synthesize our model for real time aspects and we detect human faces in real time and classify emotion and gender. Our emotion recognition is achieved such as human accuracy [4].

1.5.2 Project Organization

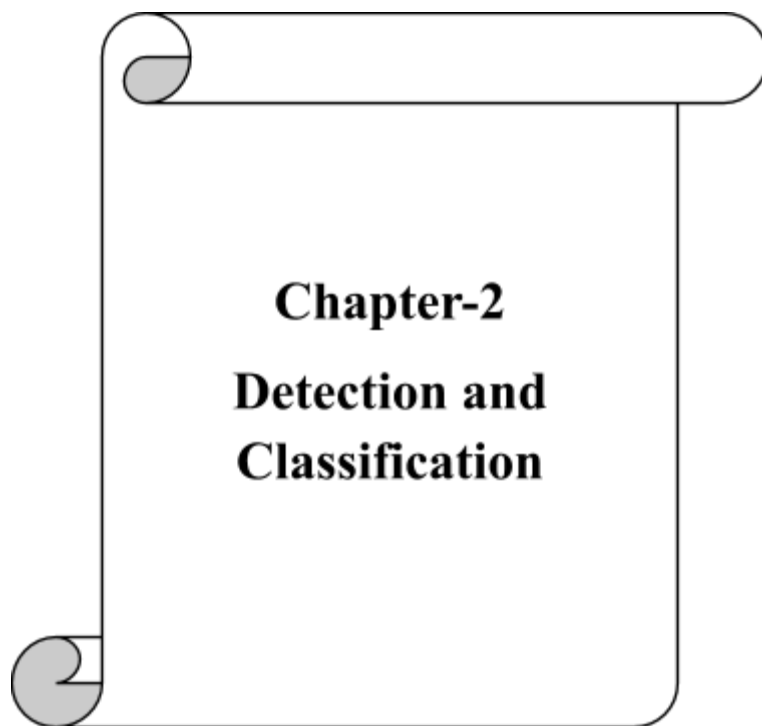
The rest of this project is structured into the following 5 chapters. In chapter 2 we gave a brief description of the face detection and classification technique. We here are introduced how to classify face gender and emotion with seven several facial expressions.

In chapter 3 we briefly describe the convolutional neural network. we describe how each part of the convolutional neural network works. Convolutional layer, sup sampling layer and fully connected layer are also introduced here. Training process of convolutional neural networks are also described in this section.

In chapter 4 we describe the each and everything of our proposed method. Firstly, we describe the dataset we use. Next a short description about preprocessing of the input image is given. Back guided propagation that we use to visualize our training section weight so that we can understand how neurons are active.

In chapter 5 we describe the result of our proposed system. We also introduce what problem arises and how we solve it. We describe how to visualize training steps to understand problem solving methods better. Finally, how to Convolutional neural network (CNN) performance for classification with minimum training time and low computational power are also described

In chapter 6 we summarize our project. Furthermore, we describe the limitations of this work and we also mentioned some of the future scope of our project in our real time problem solution.



2.1 Introduction

The Facial expression detection and recognition system performs the three learning stages in just one Convolution neural network (CNN). The proposed system operates in two main phases: training and test. During training, the system receives training data comprising gray-scale images of faces with their respective expression id and eye center locations and learns a set of weights for the net-work. To ensure that the training performance is not expected by the order of presentation of the examples, a few images are separated as validation and are used to choose the best set of weights out of a set of training performed with samples presented in different orders. During the test, the system receives a gray-scale image of a face along with its respective eye center locations and outputs the predicted expression by using the neural network weights learned during training.

2.2 Face Detection

Face detection is a computer technology being used in a variety of applications that identifies human faces in digital images. Face detection also refers to the psychological process by which humans locate and attend to faces in a visual scene

We used OpenCV to capture the live image. Using Haar-Cascade image processing technique to detect the faces. We found that there was a situation where it didn't detect the faces in the live image due to lack of contrast. So, we employed histogram equalization to improve detection by increasing contrast.

Haar-cascade: Face detection using Haar-cascade is based upon the training of a Binary classifier system using number of positive images that represent the object to be recognized (like faces of different persons at different backgrounds) and even large number of negative images that represent objects or feature not to be detected(images that are not faces but can be anything else like chair, table, wall, etc.) Actual Image Extracted face.

2.3 Image Classification

The contextual image classification, a topic of pattern recognition in computer vision, is an approach of classification based on contextual information in images. "Contextual" means this approach is focusing on the relationship of the nearby pixels, which is also called neighborhood.

2.3.1 Gender Classification

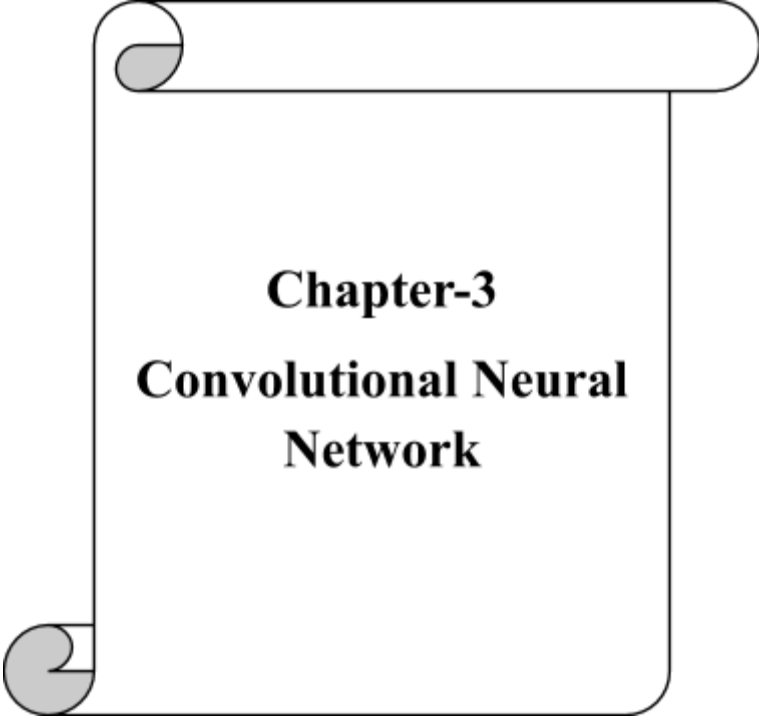
Age and gender, two of the key facial attributes, play a very foundational role in social interactions, making age and gender estimation from a single face image an important task in intelligent applications, such as access control, human-computer interaction, law enforcement, marketing intelligence and visual surveillance, etc.

A preprocessing tool which can extract facial and other physical characteristics from the image, an algorithm which can integrate the part-based information and ensemble, a neural network which can classify the gender from the ensemble, based on the database that connects the peculiarity of these physical features for males and females should work.

2.3.2 Emotion Classification

Images can both express and affect people's emotions. It is intriguing and important to understand what emotions are conveyed and how they are implied by the visual content of images. Inspired by the recent success of deep convolutional neural networks (CNN) in visual recognition, they explore simple, yet effective deep learning-based methods for image emotion analysis.

we extract features using the fine-tuned CNN at different locations at multiple levels to capture both the global and local information. The features at different locations are aggregated using the Fisher Vector for each level and concatenated to form a compact representation.



Chapter-3

Convolutional Neural Network

3.1 Introduction

In the last few years, deep Convolutional Neural Network (CNN) learning has proved the outstanding performance in the field of image classification, machine learning and pattern recognition. Above all existing models, CNN is one of the most popular models and has been providing the state-of-the-art recognition accuracy on Object recognition [15], Human activity analysis [7], Object detection [17], Age-gender classification [4], Facial expression classification [2]. For the task of image classification CNN outperforms above all the previous classification methods [16]. CNN extracted from the feature the image by a series of operations.

3.2 Basic Operations on CNN

Convolutional Neural Network (CNN or ConvNet) is a class of deep artificial neural networks that has effectively been functional to analyze visual imagery. Simple ConvNet for Emotion & Gender classification could have the architecture [INPUT - CONV – ReLU - POOL - FC] [13]. There are four main operations in ConvNet. Figure 3.1 shown the basic CNN architecture for classification where the first portion describe as the feature extraction part and the next portion describe as the classification part.

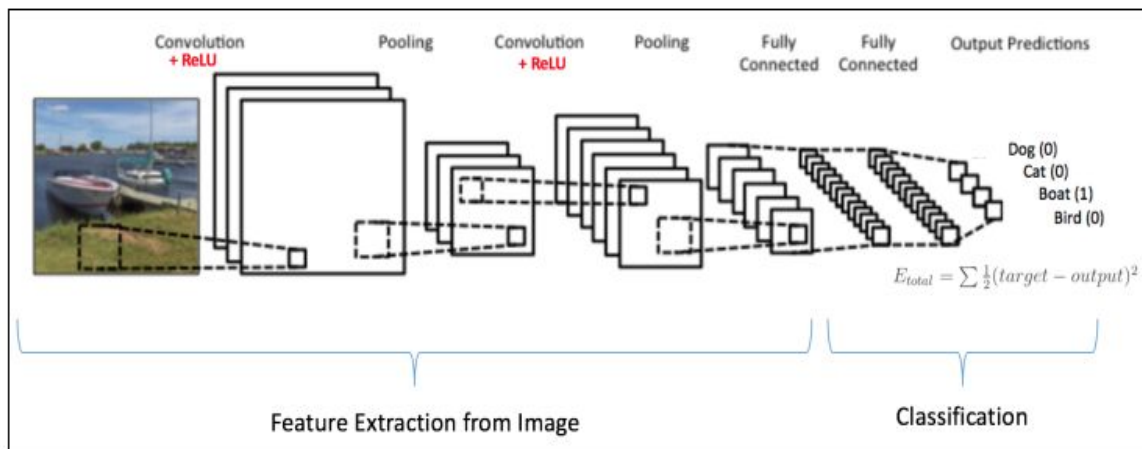


Figure3.1: Basic CNN Architecture with an example image of boat in the final stage the probability of boat is maximum define the right classification. [13]

3.2.1 Convolution Layer

Convolutional layer also denoted as Conv. Layer, it forms the basis of the CNN and carries out the core operations of training and therefore firing the neurons of the networks. It performs the convolution operation over the input volume. An image can be considered as a matrix of pixels. Consider a 7x7 image whose pixel value is 0 or 1. Also consider a Filter matrix 3x3 as in Figure 3.2 below.

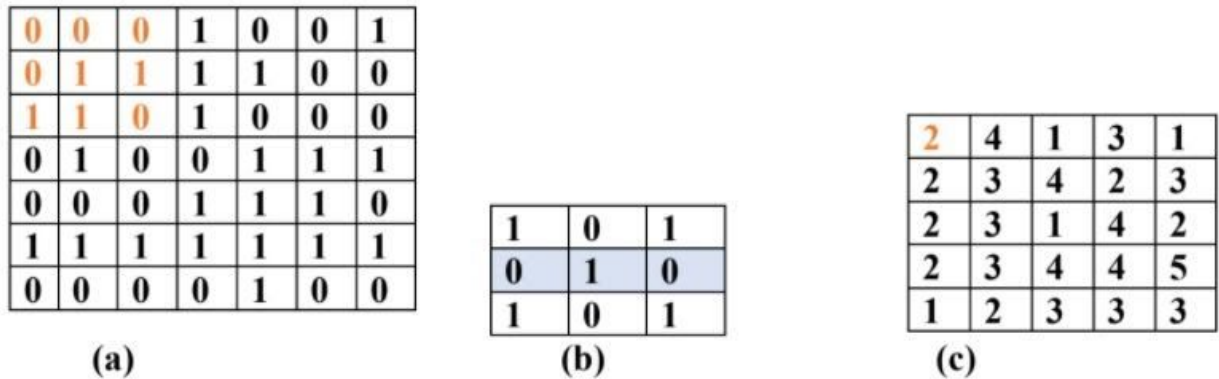


Figure 3.2. Basic Convolution Operation (a) 7 x 7 image matrix (b) 3 x 3 filter matrix (c) Result 5 x 5 feature vector after convolution of a with b.

Mathematical calculation between the filter 3x3 with the image matrix's first part (shown in different color) produces a result 2 which is also shown in the resulting image (with different color). The calculation is as follows:

$$0 \times 1 + 0 \times 0 + 0 \times 1 + 0 \times 0 + 1 \times 1 + 1 \times 0 + 1 \times 1 + 1 \times 0 + 0 \times 1 = 2$$

Each convolution layer reduces the image dimension more specifically size of the image. The formula for calculating the output size for any given conv layer is

$$O = \frac{(W-F+2P)}{S} + 1 \dots \dots \dots (3.1)$$

Where, O is the output height/length, W is the input height/length, F is the filter size, P is the padding, and S is the stride. For the above example with stride 1 and padding 0 a 7x7 input image provides an output of size 5x5.

Stride can be defined as the number of pixels by which a slide is functional for a filter matrix over the input matrix. When the stride is 1 the filters move one pixel at

a time. And for the value of the stride is 2, the filters jump 2 pixels at a time. Having a larger value of stride may cause it to produce smaller feature maps.

Sometimes, it is suitable to pad the input matrix with a zero's around the border, so that we can apply the filter to bordering elements of our input image matrix. A nice feature of zero padding is that it allows us to control the size of the feature maps. Adding zero-padding is also called wide convolution, and not using zero-padding would be a narrow convolution.

3.2.2 Non-Linearity (ReLU)

After each conv layer, it is convention to apply a nonlinear layer (or activation layer) immediately afterward. The purpose of this layer is to familiarize nonlinearity to a system that fundamentally has just been computing linear operations throughout the conv layers. Nonlinear functions like tanh and sigmoid were used previously, but researchers found out that ReLU layers work far better because the network is able to train a lot faster due to the computational efficiency without making a significant difference to the accuracy [14]. In Figure 3.3 the ReLU activation function is shown which produces output as the maximum of zero and the x .

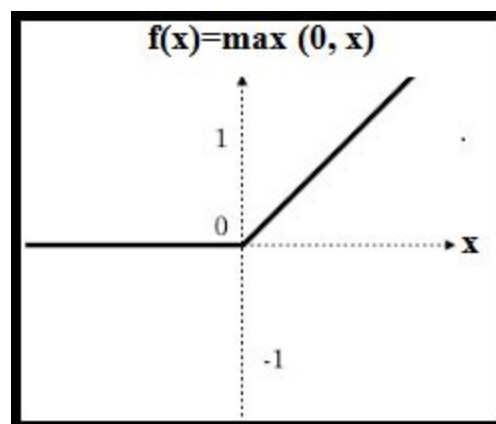


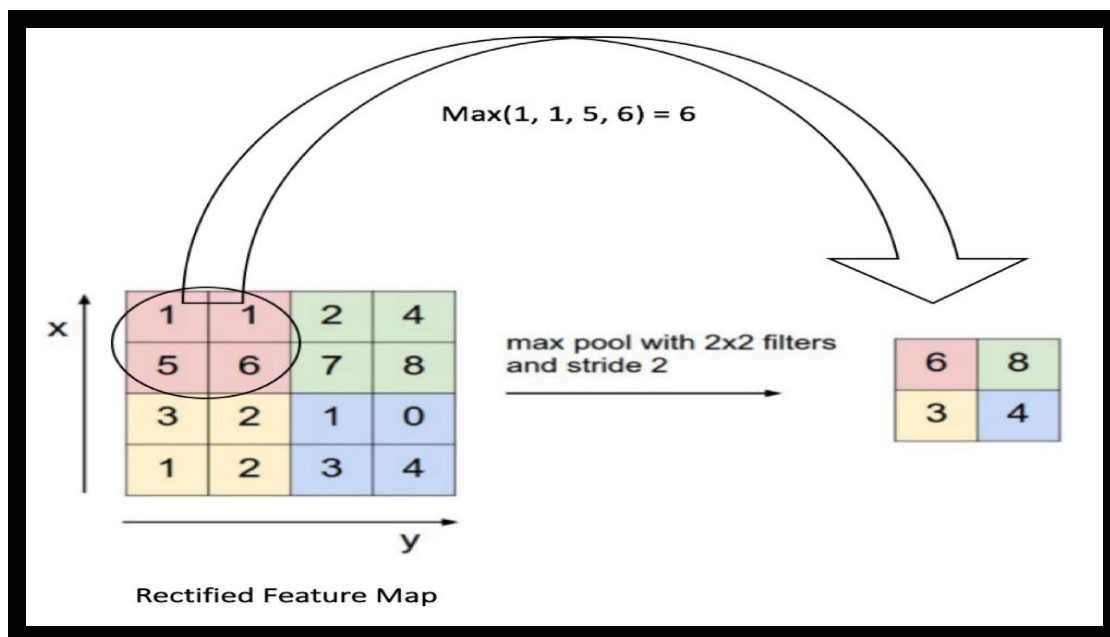
Figure 3.3: ReLU activation function. [13]

3.2.3 Pooling of Sub Sampling

Spatial pooling (also called subsampling or down-sampling) reduces the dimensionality of each feature map but retains the most important information. Spatial Pooling can be of different types: Max, Average, Sum etc.

In case of Max Pooling, we define a spatial neighborhood (for example, a 2x2 window) and take the largest element from the rectified feature map within that window. Instead of taking the largest element we could also take the average or sum of all elements in that window. In practice, Max Pooling has been shown to work better.

Figure 3.4 shows an example of Max Pooling operation on a Rectified Feature map (obtained after convolution and ReLU operation) by using a 2x2 window.



Figure

3.4: Max Pooling operation of 4x4 matrix with 2x2 filter produces a 2x2 output with maximum value. [15]

For the output dimension of the pooling with window size F and input size W with stride S can be calculated by the following formula.

$$O = \frac{(W-F)}{S} + 1 \dots \dots \dots (3.2)$$

In the above example where W=4, F=2 and S=1 the output is a 2x2 feature vector.

3.2.4 Fully Connected Layer

The Fully Connected layer is a traditional Multi-Layer Perceptron that uses a SoftMax activation function in the output layer. The term “Fully Connected” denotes that every neuron in the previous layer is connected to other every neuron on the next layer.

The output from the convolutional and pooling layers represent high-level features of the input image. The purpose of the Fully Connected layer is to use these features for classifying the input image into various classes based on the training dataset, for example, in Figure 3.1 the final classification stage has shown four possible outputs for doing the image classification task.

Apart from classification, adding a fully-connected layer is also a cheap way of learning non-linear combinations of these features. Most of the features from convolutional and pooling layers may be good for the classification task, but combinations of those features might be even better.

The sum of output probabilities from the Fully Connected Layer is 1. This is ensured by using the SoftMax as the activation function in the output layer of the Fully Connected Layer. The SoftMax function takes a vector of arbitrary real-valued scores and squashes it to a vector of values between zero and one that sum to one.

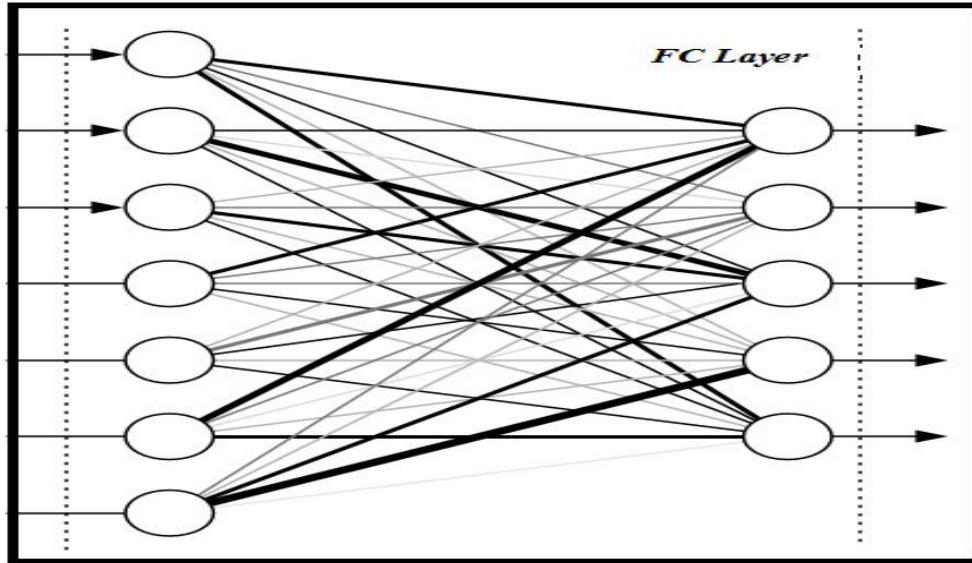


Figure 3.5: Fully Connected layer with 5 output, all previous mode is connected with every node of FC Layer.

3.3 Training of the CNN Networks

The way the computer is able to adjust its filter values (or weights) is through a training process also called backpropagation. A loss function can be defined in many different ways but a common one is MSE (mean squared error), which is half times (actual – predicted) squared.

$$L = \sum \frac{1}{2} (a_p - y_p)^2 \dots \dots \dots (3.3)$$

Where,

L = Total Error

a_p = Target Probability

y_p = Outcome Probability

The task of minimizing the loss involves trying to adjust the weights so that loss can be reduced. Because of this the derivative of the loss with respect to the weight is computed. This is the mathematical equivalent of a dL/dW where W are the weights at a particular layer. Now the backward pass through the network, which is determining which weights contributed most of the loss and finding ways to adjust them so that the loss decreases. Once the computation of the derivative is done, we then go to the last step which is the weight update. This is where we take all the

weights of the filters and update them so that they change in the opposite direction of the gradient.

$$\mathbf{w} = \mathbf{w}_i - \mu \frac{dL}{dw} \dots \dots \dots (3.4)$$

Where,

w = Weight

w_i = Initial Weight

μ = Learning Rate

The learning rate is a parameter μ that is chosen by the programmer. A high learning rate means that bigger steps are taken in the weight updates and thus, it may take less time for the model to converge on an optimal set of weights. However, a learning rate that is too high could result in jumps that are too large and not precise enough to reach the optimal point. Figure 3.6 shows the problem arises for high and low learning rate.

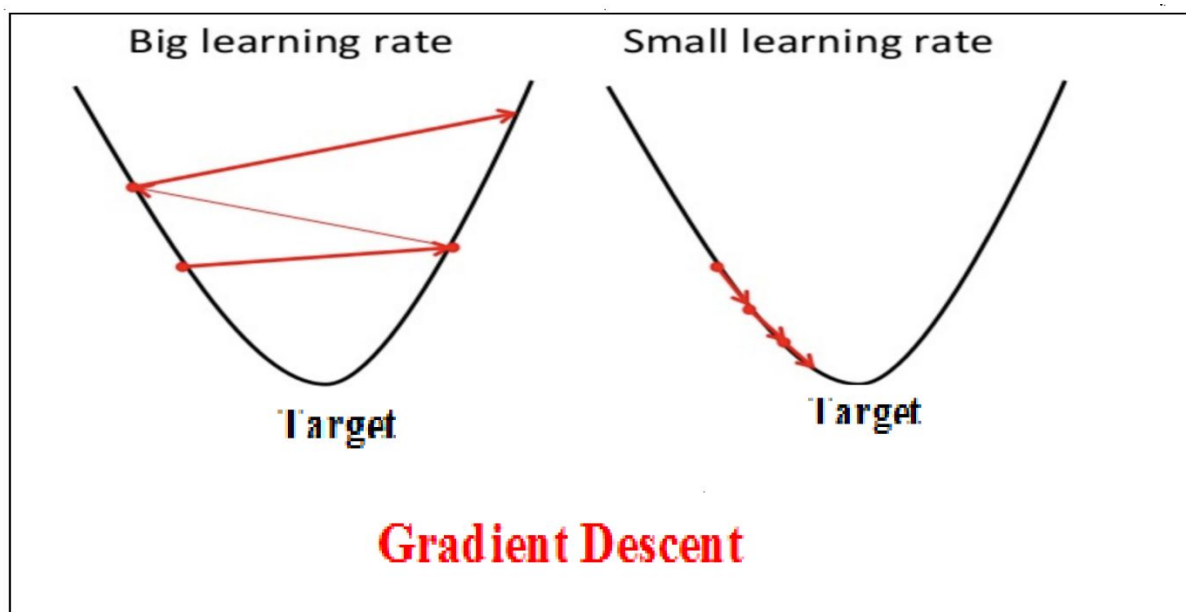


Figure 3.6: Gradient descent, Problem arises could not reach the target point easily due to high and learning rate and low learning rate.

The overall training process Algorithm of the Convolution Network can be summarized as next page Process:

Table 3.1 steps in the training process of CNN.

Steps in the training process of CNN.

Step 1: Initialize all filters and parameters / weights with random values

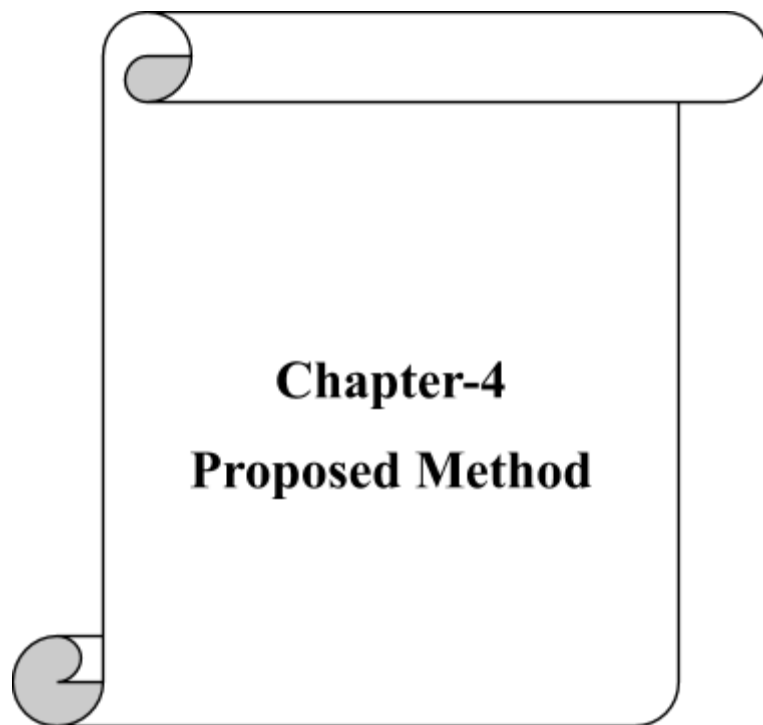
Step 2: The network takes a training image as input, goes through the forward propagation step (convolution, ReLU and pooling operations along with forward propagation in the Fully Connected layer and finds the output probabilities for each class.

Step 3: Calculate the total error at the output layer

$$\text{Total error} = \sum \frac{1}{2} (\text{Target Probability} - \text{Outcome Probability})^2$$

Step 4: Use Backpropagation to calculate the gradients of the error with respect to all weights in the network and use gradient descent to update all filter $\frac{\text{values}}{\text{weights}}$ and parameter values to minimize the output error.

Step5: Repeat steps 2-4 with all images in the training set.



4.1 Introduction

Due to the classification of gender and emotion from each image a group of steps need to take. Such as dataset preparation, preprocessing and powerful classification model. Each of the step's performance causes an effect on the total classification accuracy.

4.2 Workflow

Real time face detection and gender & emotion recognition is a robust complex problem in computer vision because of real time image frames. First, we have to take a real time video frame then convert it as an image and extract the face from the image to detect the human face. After extracting face, we consider each face part of the image as a full image for further processing. Each extracting face image is then providing as input to pre pre-process step of classification model and each pre-processing step takes some operation on it input to resize as model input and data augmentation as input to our proposed convolutional neural network (CNN) model for classification of the emotion and gender. The resulting label that is output of the CNN is then used for making descriptions of gender {"man" or "women"} and facial emotion classification {"angry", "disgust", "fear", "happy", "sad", "neutral"}. The total process is shown graphically in figure 4.1.

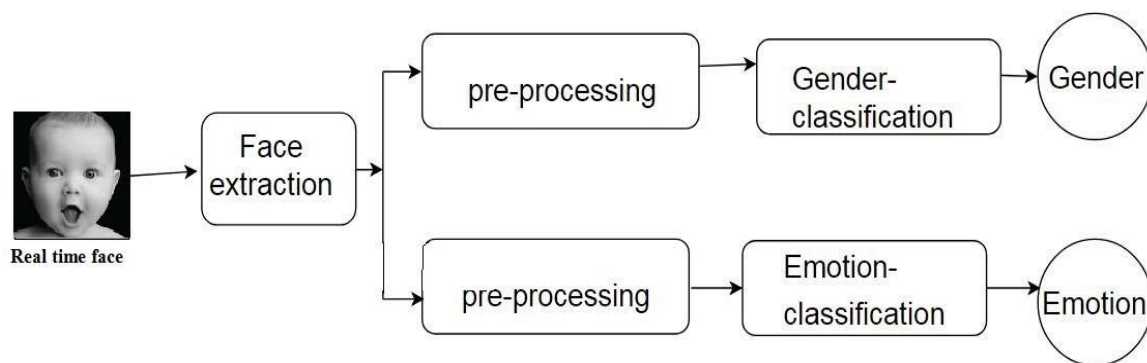


Figure 4.1: Working principle diagram of our proposed approach.

Feature extraction of each individual face is the most difficult part for any face due to its face shape and structure as real time. To solve this problem, we apply Convolutional neural network which doesn't require any predefined feature for classification of specific human image.

4.3 Dataset Preparation

Preparation of the dataset is the fundamental concern of this project. Faces such as men, women, actors, celebrities, politicians all can be defined by their edge. For this reason, we first prepare a dataset as the most precedence given to its edges that enlighten the edges. We prepare FER-2013 dataset [2] for facial emotion recognition and IMDB dataset [5] for gender classification that contains only face.

4.4 Pre-Processing

Pre-processing of the input image is the procedure which encompasses changes and modifications to the image to make it fit for recognition. The following technique may be used for image enhancement.

4.4.1 Conversion of the RGB to Gray-Scale

The image picked up before preprocessing is generally an RGB image which means that the pixel value of the image encompasses a component of three color's and that is Red, Green and Blue. This coloured image is mostly difficult to analyze due to its higher dimensionality. That is why for the sake of the reduction of dimension image is transformed into a typical gray -scale image and is represented through a single matrix because the detection of facial attributes on a coloured image is more challenging than on a gray scale-scale image. If the gray bitmap Y and color bitmap is R, G and B then the formula we used is:

$$Y=0.229R + 0.587G +0.114B$$

4.4.2 Normalization

Normalization is a process that changes the range of pixel intensity values. Applications include photographs with poor contrast due to glare, for example. Normalization is sometimes called contrast stretching or histogram stretching. In more general fields of data processing, such as digital signal processing, it is referred to as dynamic range [0.0-1.0] expansion.

4.4.3 Noise Reduction

Excessive pixels that are present in an image is called noise. Noise may be in the form of salt and pepper noise or Gaussian noise. Low pass filtering is used to remove the Gaussian noise from the image and there is no need to filter salt and pepper noise as it is very low as compared to the Gaussian noise. In our proposed method we removed all components which are less than 5 pixels for simplicity of small unwanted pixel noise.

4.4 Proposed CNN Architecture

We propose two models which we evaluated in accordance to their test accuracy and number of parameters. Both models were designed with the idea of creating the best accuracy over number of parameters ratio. Reducing the number of parameters helps us overcome two important problems. First, the use of small CNNs alleviate us from slow performances in hardware-constrained systems. And second, the reduction of parameters provides a better generalization under an Occam's razor framework. Our first model relies on the idea of eliminating completely the fully connected layers. The second architecture combines the deletion of the fully connected layer and the inclusion of the combined depth-wise separable convolutions and residual modules. Both architectures were trained with the ADAM optimizer [8]. Following the previous architecture schemas, our initial architecture used Global Average Pooling to completely remove any fully connected layers. This was achieved by having in the last convolutional layer the same number of feature maps as the number of classes, and applying a softmax activation function to each reduced feature map. Our initial proposed architecture is a standard fully-convolutional neural network composed of 9 convolution layers, ReLUs [10], Batch Normalization [9] and Global Average Pooling.

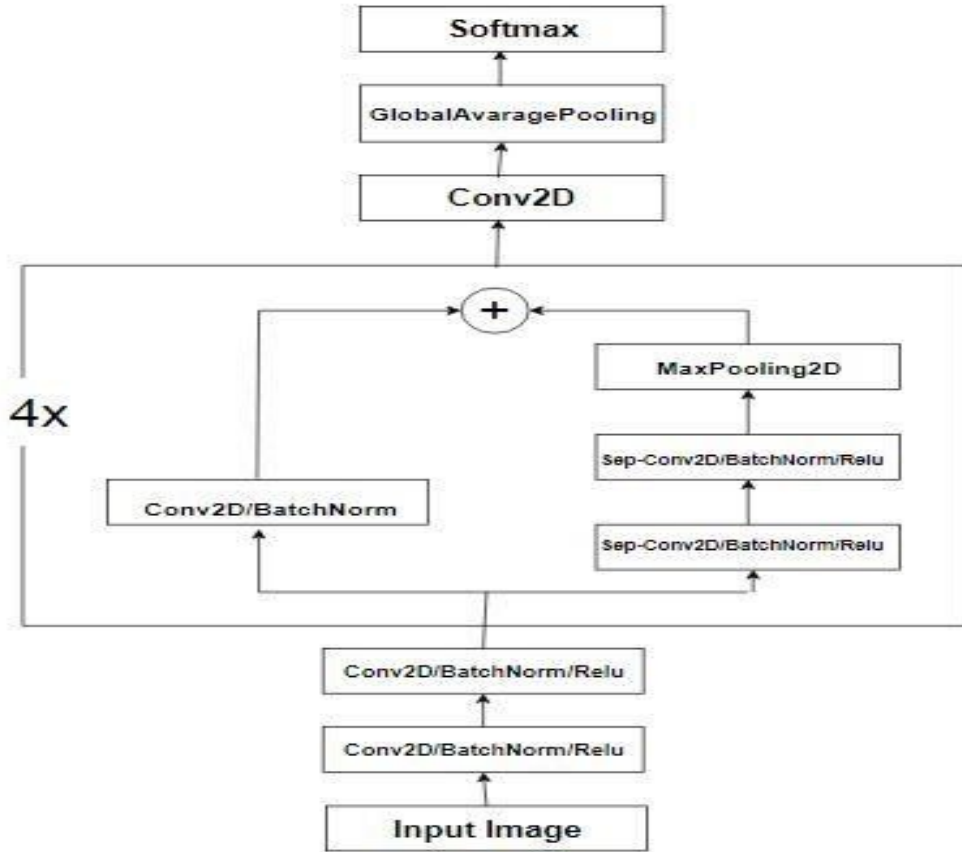


Fig.

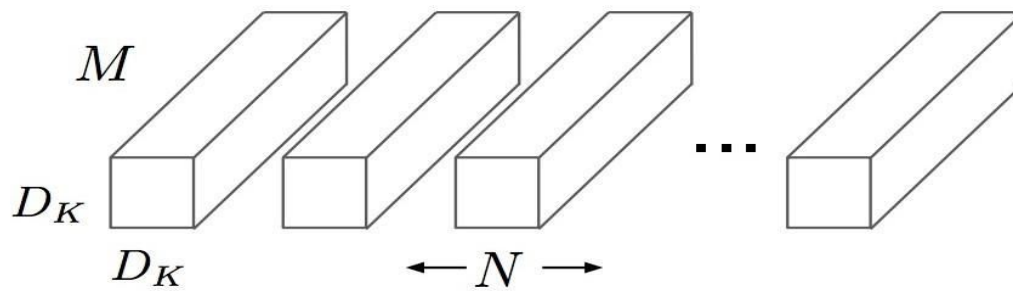
4.2: Our proposed model for real-time classification.

This model contains approximately 600,000 parameters. It was trained on the IMDB gender dataset, which contains 460,723 RGB images where Each image belongs to the class “woman” or “man”, and it achieved an accuracy of 95% in this dataset. We also validated this model in the FER-2013 dataset. This dataset contains 35,887 grayscale images where each image belongs to one of the following classes {“angry”, “disgust”, “fear”, “happy”, “sad”, “surprise”, “neutral”}. Our initial model achieved an accuracy of 66% in this dataset. We will refer to this model as “sequential fully-CNN”.

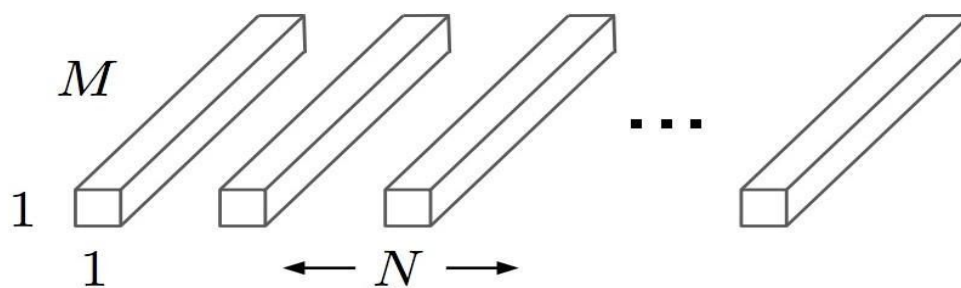
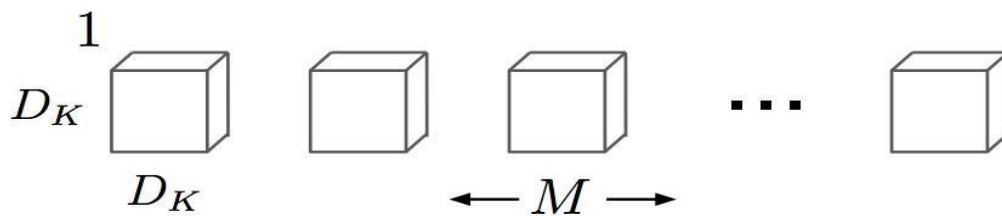
Our second model is inspired by the Xception [3] architecture. This architecture combines the use of residual modules [11] and depth-wise separable convolutions [12]. Residual modules modify the desired mapping between two subsequent layers, so that the learned features become the difference of the original feature

map and the desired features. Consequently, the desired features $H(x)$ are modified in order to solve an easier learning problem $F(X)$ such that:

$$H(x) = F(x) + x \dots \dots \dots (1)$$



(a)



(b)

Fig. 4.3: [12] Difference between (a) standard convolutions and (b) depth-wise separable convolutions.

Since our initial proposed architecture deleted the last fully connected layer, we

reduced further the number of parameters by eliminating them now from the convolutional layers. This was done through the use of depth-wise separable convolutions. Depth-wise separable convolutions are composed of two different layers: depth-wise convolutions and pointwise convolutions. The main purpose of these layers is to separate the spatial cross-correlations from the channel cross correlations [3]. They do this by first applying a $D \times D$ filter on every M input channel and then applying $N 1 \times 1 \times M$ convolution filters to combine the M input channels into N output channels. Applying $1 \times 1 \times M$ convolutions combines each value in the feature map without considering their spatial relation within the channel. Depth-wise separable convolutions reduce the computation with respect to the standard convolutions by a factor of $N1 + 1D^2$ [12]. A visualization of the difference between a normal Convolution layer and a depth-wise separable convolution can be observed in Figure 4.3.

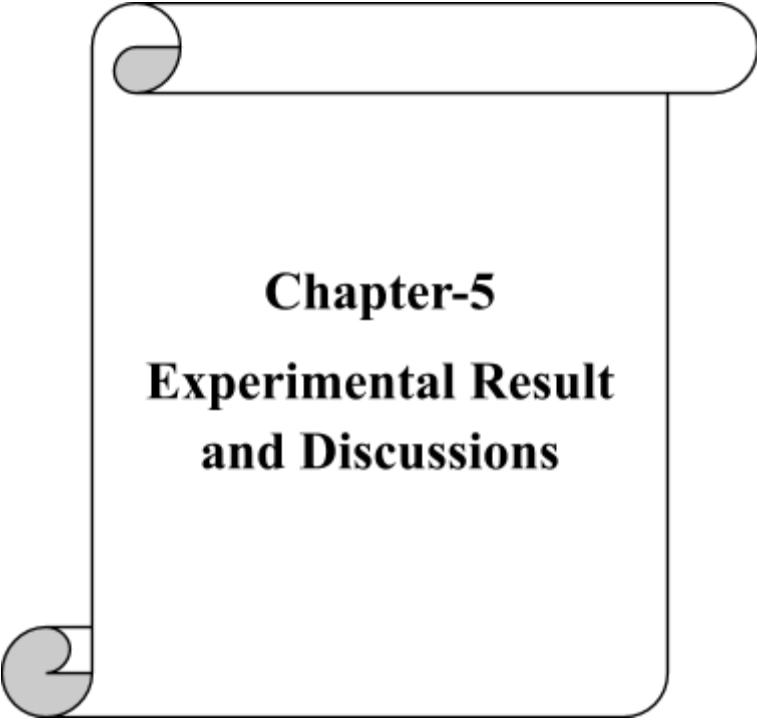
Our final architecture is a fully-convolutional neural network that contains 4 residual depth-wise separable convolutions where each convolution is followed by a batch normalization operation and a ReLU activation function. The last layer applies a global average pooling and a soft-max activation function to produce a prediction. This architecture has approximately 60; 000 parameters; which corresponds to a reduction of $10\times$ when compared to our initial naïve implementation, and $80\times$ when compared to the original CNN. Figure 3 displays our complete final architecture which we refer to as mini-Xception. This architecture obtains an accuracy of 95% in gender classification task. Which corresponds to a reduction of one percent with respect to our initial implementation.

Furthermore, we tested this architecture in the FER-2013 dataset and we obtained the same accuracy of 66% for the emotion classification task. Our final architecture weights can be stored in an 855 kilobytes file. By reducing our architectures computational cost, we are now able to join both models and use them consecutively in the same image without any serious time reduction. Our complete pipeline including the OpenCV face detection module, the gender classification and the emotion classification takes $0:22 \pm 0:0003$ ms on a i5-5200U CPU. This

corresponds to a speedup of $1.5\times$ when compared to the original architecture of Tang. We also added to our implementation a real-time guided back-propagation visualization to observe which pixels in the image activate an element of a higher-level

feature map. Given a CNN with only ReLUs as activation functions for the intermediate layers, guided-backpropagation takes the derivative of every element $(x; y)$ of the input image I with respect to an element $(i; j)$ of the feature map f^l in layer L . The reconstructed image R filters all the negative gradients; consequently, the remaining gradients are chosen such that they only increase the value of the chosen element of the feature map. Following [10], a fully ReLU CNN reconstructed image in layer l is given by:

$$R_{i,j}^l = (R_{i,j}^{l+1} > 0) * R_{i,j}^{l+1} \dots \dots \dots (2)$$



Chapter-5

Experimental Result and Discussions

5.1 Dataset

In this project we used FER-2013 emotion dataset [4] for emotion classification task and we used IMDB-WIKI age and gender dataset [5] for gender classification task simultaneously. The FER-2013 dataset in figure-5.1 consists of 48x48 pixel gray-scale 35886 images of faces. The faces have been automatically registered so that the face is more or less centered and occupies about the same amount of space in each image. The task is to categorize each face based on the emotion shown in the facial expression in to one of seven categories with the following classes {"0-angry", "1-disgust", "3-fear", "4-happy", "5-sad", "6-neutral"}.



Figure 5.1: Samples of the FER-2013 emotion dataset [4].



Figure 5.2: Samples of the IMDB dataset [5].

The IMDB dataset in figure-5.2 consists of a total 460,723 face images from 20,284 celebrities. We used only face image for our training purpose of gender classification task of two classes as following {"0-Women" and "1-Man"}.

5.2 Experiment

After training on the network, we use 10% images in both dataset [2,5] as test images for the task of recognition of the emotion expression and gender. Our testing result can be observed as figure 5.3 where our proposed model can classify gender as two class properly but all kind of our training images are such as western actors, models, politician etc. our emotion recognition performed as like human where our proposed method can understand human behavior of seven class such angry, happy, sad, fear, surprise, neutral etc.

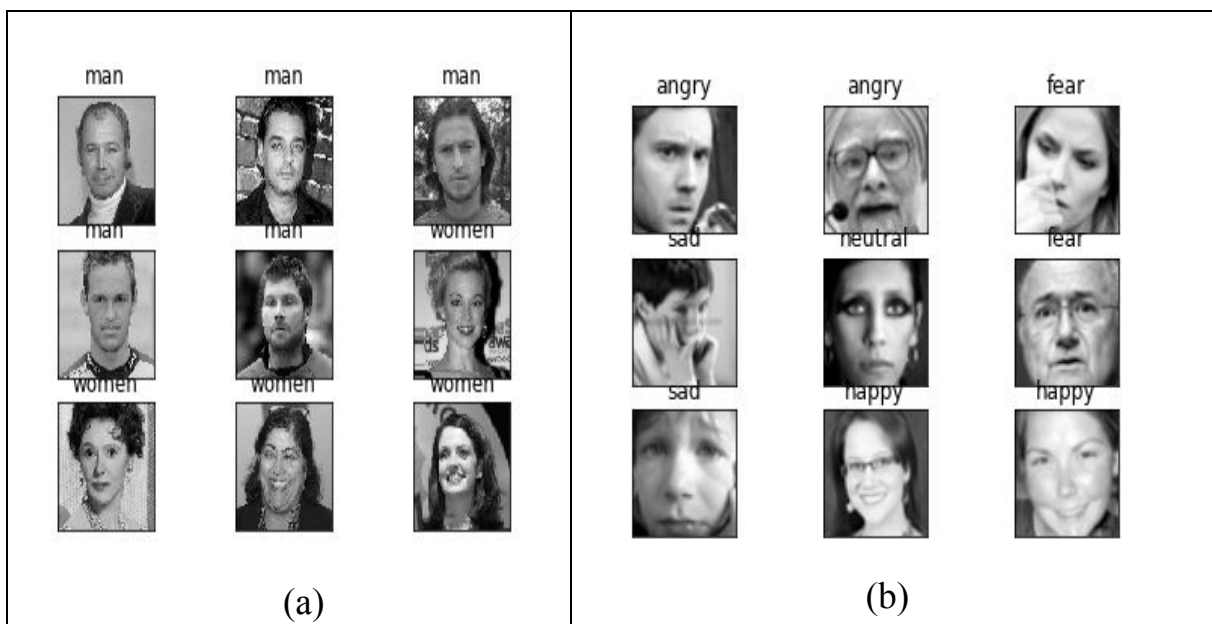


Figure 5.3: Predicted results of the emotion and gender classification on testing data. (a) Result of gender classification on IMDB dataset [5]. (b) result of emotion recognition on FER-2013 dataset [2].

5.2.1 Experiment Result

Experiment results of the real-time emotion classification task in unseen faces can be observed in Figure 5.4. Our complete real-time pipeline including: face

detection, emotion and gender classification have been fully integrated in our Intel Core-i5 5200U processor.

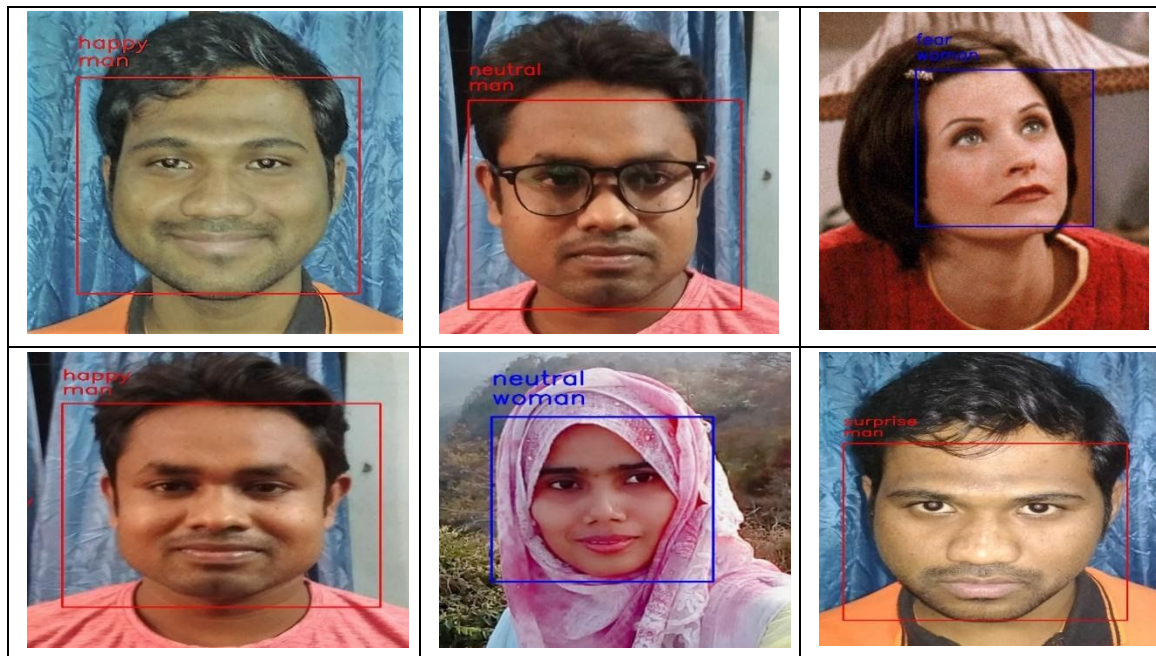


Figure 5.4: Results of the real-time emotion classification.

An example of our complete pipeline can be seen in Figure 5.5 in which we provide emotion and gender classification.



Figure 5.5: Results of the combined gender and emotion recognition. The color blue represents the assigned class woman and red the class man.

In Figure 5.6 we provide the confusion matrix results of our emotion classification of our proposed mini-Xception model.

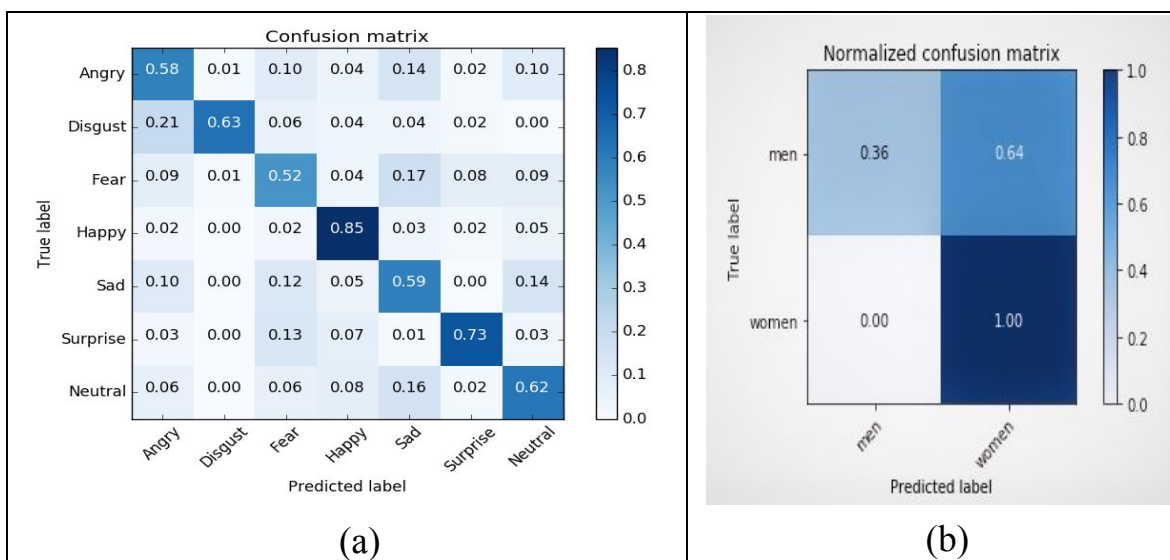


Figure 5.6: Normalized confusion matrix of our mini-Xception network (a) confusion matrix for facial emotion recognition (b) confusion matrix for gender recognition.

We can observe several common misclassifications such as predicting “sad” instead of “fear” and predicting “angry” instead “disgust”.

A comparison of the learned features between several emotions and both of our proposed models can be observed in Figure 5.7. The white areas in figure 8b correspond to the pixel values that activate a selected neuron in our last convolution layer. The selected neuron was always selected in accordance to the highest activation.

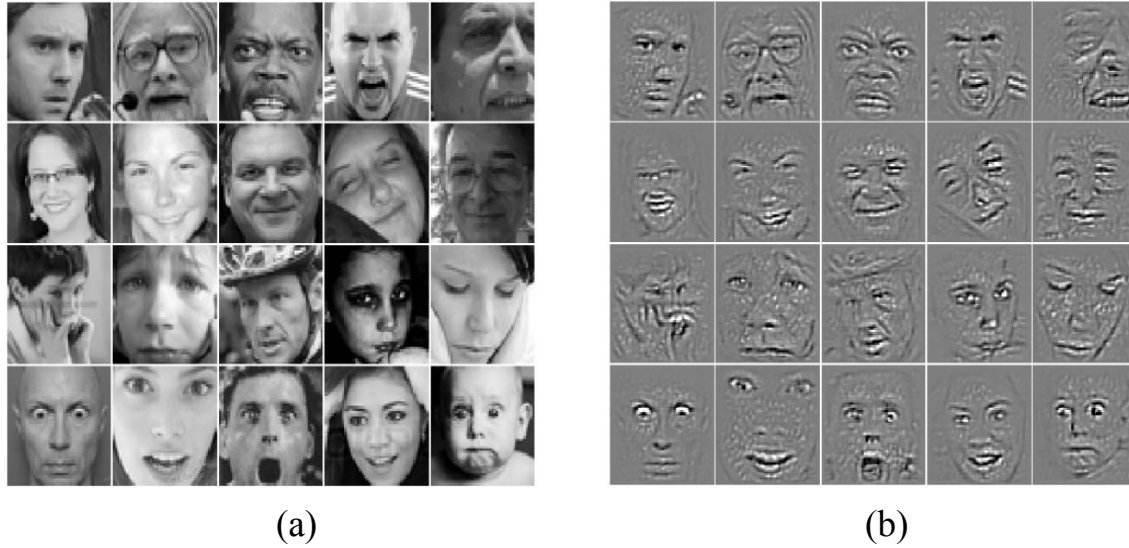


Fig. 5.7: All sub-figures contain the same images in the same order. Every row starting from the top corresponds respectively to the emotions: angry, happy, sad and surprise (a) Samples from the FER-2013 dataset (b) Guided back-propagation visualization of our proposed mini-Xception model.

We can observe that the CNN learned to get activated by considering features such as the frown, the teeth, the eyebrows and the widening of one's eyes, and that each feature remains constant within the same class. These results reassure that CNN learned to interpret understandable human-like features, that provide generalizable elements. These interpretable results have helped us understand several common misclassifications such as persons with glasses being classified as “angry” in figure 5.9.

This happens since the label “angry” is highly activated when it believes a person is frowning and frowning features get confused with darker glass frames. Moreover, we can also observe that the features learned in our proposed mini-Xception model are more interpretable than the ones learned from sequential fully-CNN. Consequently, the use of more parameters in our naive implementations leads to less robust features.



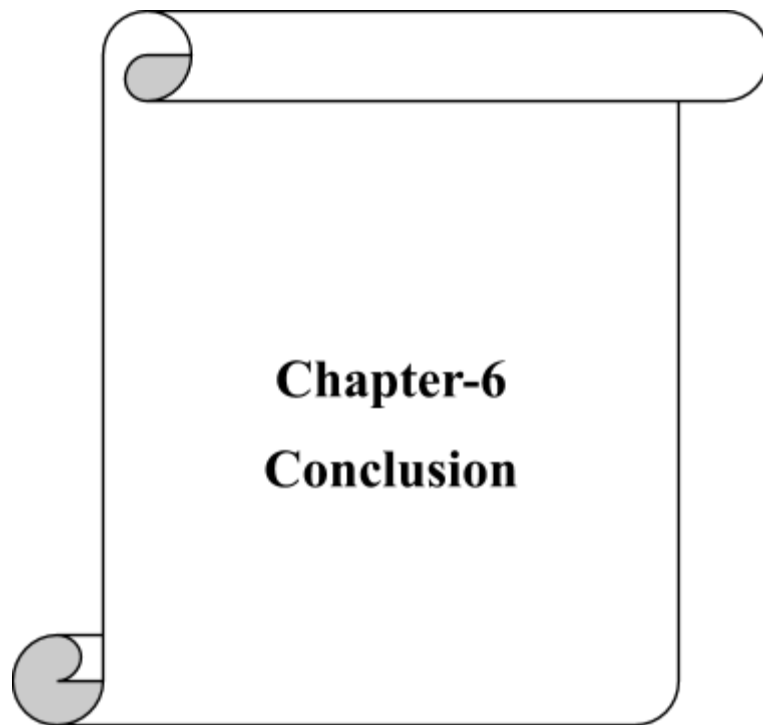
Figure 5.8: Results of the misclassification in real-time emotion and gender recognition.

5.3 Performance Evaluation and Discussion

Classifications of each class of emotion or gender from the image determines the final performance for the correct recognition. Sometimes face detection and emotion classification are even becoming difficult tasks for humans due to several reasons, such as different people cutting their hair different styles, wearing glass makes them more complex. For classification part the actual performance, P can be measured by

$$P = \frac{Y}{(X-m)}$$

Where X is the number of images given as input for classification and Y is actual number for correct classification and m is the number of images that conflict with misclassification that even cannot be classified by humans.



6.1 Conclusion

We have proposed and tested general building designs for creating real time CNNs. Our proposed architectures have been systematically built in order to reduce the number of parameters. We began by eliminating completely the fully connected layers and by reducing the number of parameters in the remaining convolutional layers via depth-wise separable convolutions. We have shown that our proposed models can be stacked for multi-class classifications while maintaining real-time inferences. Specifically, we have developed a vision system that performs face detection, gender classification and emotion classification in a single integrated module. We have achieved human-level performance in our classification's tasks using a single CNN that leverages modern architecture constructs. Our architecture reduces the number of parameters $80\times$ while obtaining favorable results.

Finally, we presented a visualization of the learned features in the CNN using the guided back-propagation visualization. This visualization technique is able to show us the high-level features learned by our models and discuss their interpretability.

6.2 Limitation

Machine learning models are biased in accordance to training data. In our specific application we have empirically found that our trained CNNs for gender classification biased towards western facial features and facial accessories. We hypothesize that this misclassification occurs since our training dataset consists of mostly western: actors, writers and cinematographers as observed in figure -. in the previous chapter.

Furthermore, as the use of glasses might affect the emotion classification in figure - by interfering with the features learned. However, the use of glasses can also interfere with the gender classification in figure -5.8 0f previous chapter. This might be a result from the training data gaveling most of the images of persons wearing glasses assigned with the label "man". We believe that uncovering such behaviors is of extreme importance when creating robust classifiers and that the use of the visualization techniques such as guided backpropagation will become invaluable when uncovering model biases.

6.3 future work

In our work, our proposed method established a benchmark for the task based on state-of-the-art Convolutional Neural Network (CNN) architecture. Hence, we are planning to extend the data samples to eliminate the dataset bias because most of the samples contain western looking faces, reduce more parameters (evolutionary strategies), In corporate more classification such as “age” and finally create double headed models that make several classifications in a single forward pass.

Bibliography

- [1] Arriaga, Octavio et al. “Real-time Convolutional Neural Networks for Emotion and Gender Classification.” *CoRR* abs/1710.07557 (2017): n. pag.
- [2] Goodfellow, Ian J. et al. “Challenges in Representation Learning: A report on three machine learning contests.” *Neural networks: the official journal of the International Neural Network Society* 64 (2013): 59-63.
- [3] Chollet, François. “Xception: Deep Learning with Depthwise Separable Convolutions.” *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017): 1800-1807.
- [4] Levi, Gil and Tal Hassner. “Age and gender classification using convolutional neural networks.” *2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (2015): 34-42.
- [5] Rothe, Rasmus et al. “DEX: Deep Expectation of Apparent Age from a Single Image.” *2015 IEEE International Conference on Computer Vision Workshop (ICCVW)* (2015): 252-257.
- [6] Gurnani, Ayesha et al. “VEGAC: Visual Saliency-based Age, Gender, and Facial Expression Classification Using Convolutional Neural Networks.” *CoRR* abs/1803.05719 (2018): n. pag.
- [7] Wang, X., Guo, R., Kambhamettu, C.: Deeply-learned feature for age estimation. In: WACV. (2015)
- [8] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014
- [9] Ioffe, Sergey and Christian Szegedy. “Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift.” *ICML* (2015).

- [10] Glorot, Xavier et al. “Deep Sparse Rectifier Neural Networks.” *AISTATS* (2011).
- [11] He, Kaiming et al. “Deep Residual Learning for Image Recognition.” *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016): 770-778.
- [12] Howard, Andrew G. et al. “MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications.” *CoRR* abs/1704.04861 (2017): n. pag.
- [13] Course note of cs 231 in stanford university:
[<http://cs231n.github.io/convolutional-networks/>]
- [14] Nair, Vinod and Geoffrey E. Hinton. “Rectified Linear Units Improve Restricted Boltzmann Machines.” *ICML* (2010).
- [15] Krizhevsky, Alex et al. “ImageNet Classification with Deep Convolutional Neural Networks.” *Commun. ACM* 60 (2012): 84-90.
- [16] Mao, Junhua et al. “Deep Captioning with Multimodal Recurrent Neural Networks (m-RNN).” *CoRR* abs/1412.6632 (2015): n. pag.
- [17] Shankar, Sukrit et al. “DEEP-CARVING: Discovering visual attributes by carving deep neural nets.” *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2015): 3403-3412.