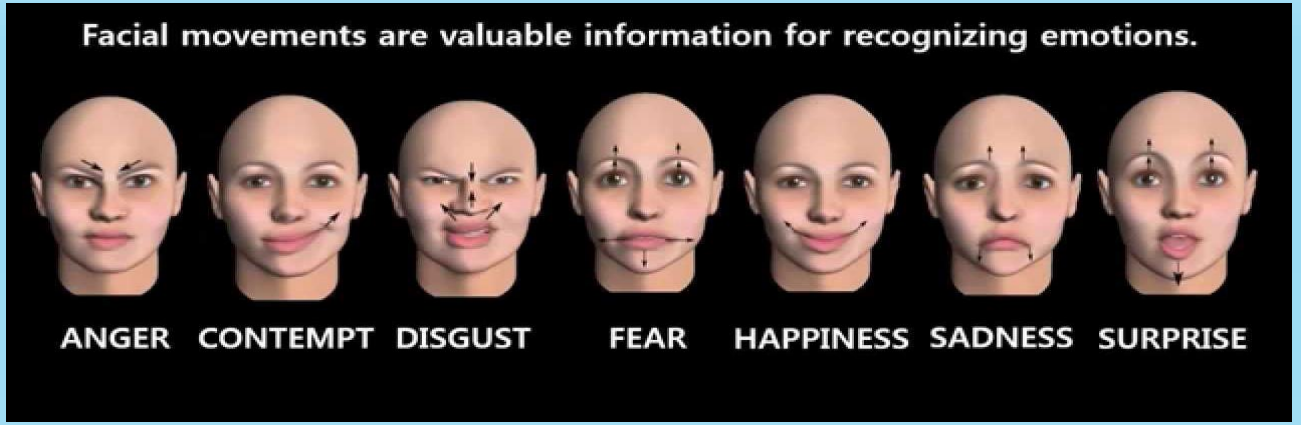
HUMAN EMOTION RECOGNITION USING BIASED DEEP LEARNING MODEL



Under Esteemed Guidance of

Dr. Ranjeet Kumar Rout

Assistant Professor,
Department of Computer Science & Engineering

Presented By:

K.Abhinandu Reddy

A.Ramanand Chowdary

OUTLINE

- Motivation
- Prerequisites
- Related Work
- Challenges Faces by Facial Expression Recognition System
- Steps
- Data Sets
- Data Preprocessing
- Our Proposed Models and Results
- Future Work

MOTIVATION

- FACIAL expression is one of the most powerful, natural and universal signals for human beings to convey their emotional states and intentions

- Out of the many biometric Techniques available, face recognition technology is the most convenient and coherent technique of all.

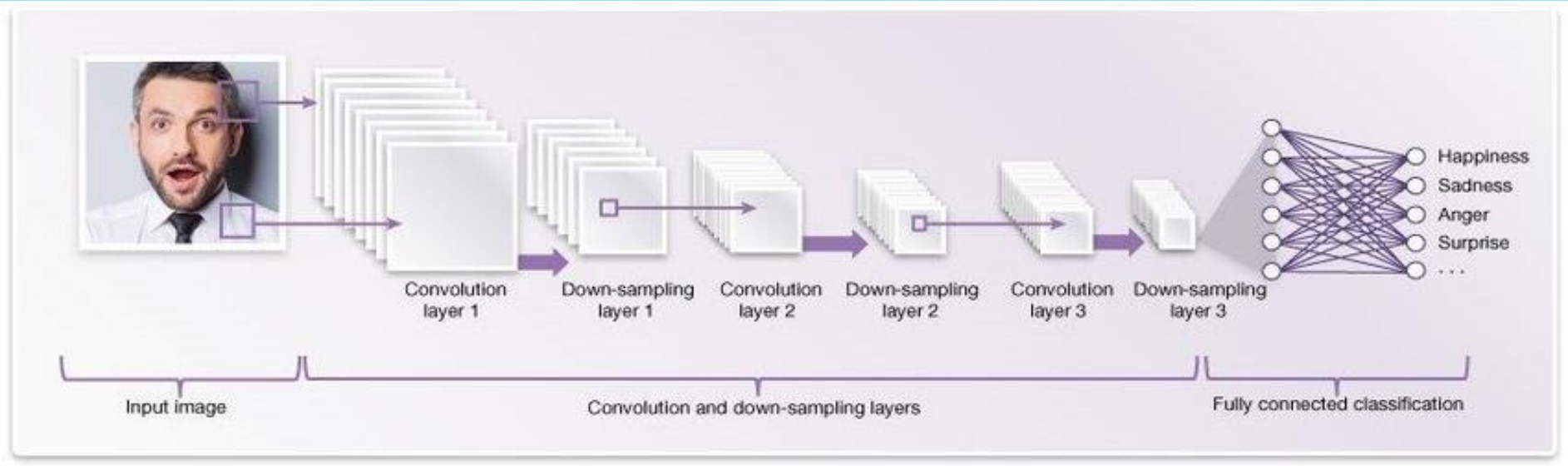


Fig 1: Facial Expression Analysis with Deep Learning

- Facial expression recognition is a complex and interesting problem, and finds its applications in driver safety, health-care, human–computer interaction etc.

- However, if we are to ever create a humanoid robot that can interact and emote with its human companions, the difficult task of emotion recognition will have to be solved.

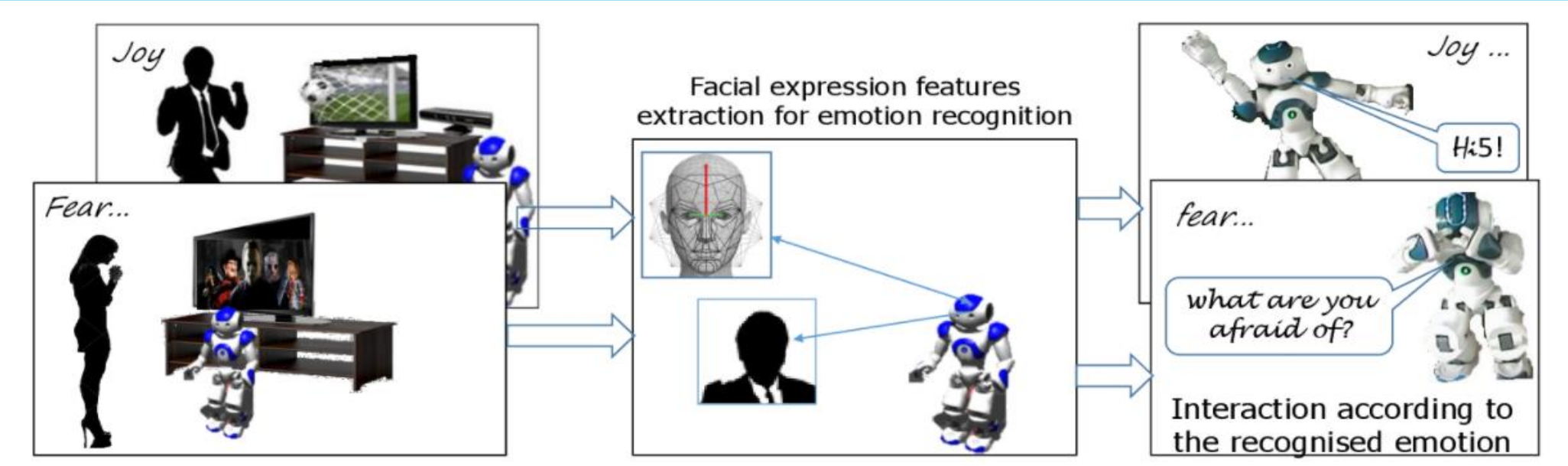


Fig 2:Facial expressions recognition for human-robot interaction

Prerequesites

- The approach we are going to use for face recognition is fairly straight forward. The key here is to get a deep neural network to produce a bunch of numbers that describe a face (known as face encodings).

- So in order to build a Deep neural network we started our preparation with Deep learning by Andrew Ng Courses . On completing these courses we got a good exposure in understanding the deep learning algorithms.

- we have implemented few programs like Hand Signal recognition to understand the working of different python libraries which we will be using to develop our deep neural network.

Challenges Faced by Facial Recognition System

Illumination

- Illumination stands for light variations. The slight change in lighting conditions cause a significant challenge for automated face recognition and can have a significant impact on its results.

- If the illumination tends to vary, the same individual gets captured with the same sensor and with an almost identical facial expression and pose, the results that emerge may appear quite different.

- It has been found that the difference between two same faces with different illuminations is higher than two different faces taken under same illumination.

Fig 3: Illumination

Pose

- Facial Recognition Systems are highly sensitive to pose variations.
- The pose of a face varies when the head movement and viewing angle of the person changes.
- The movements of head or differing POV of a camera can invariably cause changes in face appearance and generate intra-class variations making automated face recognition rates drop drastically.

Fig 4: Pose Varitaions

Occlusion

- Occlusion means blockage, and it occurs when one or other parts of the face are blocked and whole face is not available as an input image. Occlusion is considered one of the most critical challenges in face recognition system.

- It occurs due to beard, moustache, accessories (goggle, cap, mask, etc.), and it is prevalent in real-world scenario. The presence of such components makes the subject diverse and hence making automated face recognition process a tough nut to crack.

Fig 4: Occlusion

Ageing

- Face appearance/texture changes over a period of time and reflect as ageing, which is yet another challenge in facial recognition system.
- With the increasing age, the human face features, shapes/lines, and other aspects also change. It is done for visual observation and image retrieval after a long period.
- For accuracy checking, the dataset for a different age group of people over a period of time is calculated. Here, the recognition process depends on feature extraction, basic features like wrinkles, marks, eyebrows, hairstyles, etc.

Model Complexity

- Existing state-of-the-art facial recognition methods rely on ‘too-deep’ Convolutional Neural Network (CNN) architecture which is very complex and unsuitable for real-time performance on embedded devices.
- An ideal face recognition system should be tolerant of variations in illumination, expression, pose, and occlusion.
- It should be scalable to a large number of users with a need for capturing minimal images during registration while doing away with complex architecture at the same time.

Related work

Performance summary of representative methods for static-based deep facial expression recognition on the most widely evaluated datasets. Network size = depth & number of parameters; Pre-processing = Face Detection & Data Augmentation & Face Normalization; IN = Illumination Normalization; $\mathcal{NE}$ = Network Ensemble; $\mathcal{CN}$ = Cascaded Network; $\mathcal{MN}$ = Multitask Network; LOSO = leave-one-subject-out.

Datasets	Method	Network type		Network size		Pre-processing			Data selection	Data group	Additional classifier	Performance[1](%)
CK+	Ouellet 14 [110]	CNN (AlexNet)		-	-	V&J	-	-	the last frame	LOSO	SVM	7 classes†: (94.4)
	Li et al. 15 [86]	RBM		4	-	V&J	-	IN			✗	6 classes: 96.8
	Liu et al. 14 [13]	DBN	$\mathcal{CN}$	6	2m	✓	-	-	the last three frames and the first frame	8 folds	AdaBoost	6 classes: 96.7
	Liu et al. 13 [137]	CNN, RBM	$\mathcal{CN}$	5	-	V&J	-	-		10 folds	SVM	8 classes: 92.05 (87.67)
	Liu et al. 15 [138]	CNN, RBM	$\mathcal{CN}$	5	-	V&J	-	-		10 folds	SVM	7 classes‡: 93.70
	Khorrami et al. 15 [139]	zero-bias CNN		4	7m	✓	✓	-		10 folds	✗	6 classes: 95.7; 8 classes: 95.1
	Ding et al. 17 [111]	CNN	fine-tune	8	11m	IntraFace	✓	-		10 folds	✗	6 classes: (98.6); 8 classes: (96.8)
	Zeng et al. 18 [54]	DAE (DSAE)		3	-	AAM	-	-	the last four frames and the first frame	LOSO	✗	7 classes†: 95.79 (93.78) 8 classes: 89.84 (86.82)
	Cai et al. 17 [140]	CNN	loss layer	6	-	DRMF	✓	IN	the last three frames	10 folds	✗	7 classes†: 94.39 (90.66)
	Meng et al. 17 [61]	CNN	$\mathcal{MN}$	6	-	DRMF	✓	-		8 folds	✗	7 classes†: 95.37 (95.51)
	Liu et al. 17 [77]	CNN	loss layer	11	-	IntraFace	✓	IN		8 folds	✗	7 classes†: 97.1 (96.1)
	Yang et al. 18 [141]	GAN (cGAN)		-	-	MoT	✓	-		10 folds	✗	7 classes†: 97.30 (96.57)
	Zhang et al. 18 [47]	CNN	$\mathcal{MN}$	-	-	✓	✓	-		10 folds	✗	6 classes: 98.9

JAFFE	Liu et al. 14 [13]	DBN	$\mathcal{CN}$	6	2m	✓	-	-	213 images	LOSO	AdaBoost	7 classes‡: 91.8
	Hamester et al. 15 [142]	CNN, CAE	$\mathcal{NE}$	3	-	-	-	IN			✗	7 classes‡: (95.8)
MMI	Liu et al. 13 [137]	CNN, RBM	$\mathcal{CN}$	5	-	V&J	-	-	the middle three frames and the first frame	10 folds	SVM	7 classes‡: 74.76 (71.73)
	Liu et al. 15 [138]	CNN, RBM	$\mathcal{CN}$	5	-	V&J	-	-		10 folds	SVM	7 classes‡: 75.85
	Mollahosseini et al. 16 [14]	CNN (Inception)		11	7.3m	IntraFace	✓	-	images from each sequence	5 folds	✗	6 classes: 77.9
	Liu et al. 17 [77]	CNN	loss layer	11	-	IntraFace	✓	IN	the middle three frames	10 folds	✗	6 classes: 78.53 (73.50)
	Li et al. 17 [44]	CNN	loss layer	8	5.8m	IntraFace	✓	-		5 folds	SVM	6 classes: 78.46
	Yang et al. 18 [141]	GAN (cGAN)		-	-	MoT	✓	-		10 folds	✗	6 classes: 73.23 (72.67)
TFD	Reed et al. 14 [143]	RBM	$\mathcal{MN}$	-	-	-	-	-	4,178 emotion labeled 3,874 identity labeled	5 official folds	SVM	Test: 85.43
	Devries et al. 14 [58]	CNN	$\mathcal{MN}$	4	12.0m	MoT	✓	IN	4,178 labeled images		✗	Validation: 87.80 Test: 85.13 (48.29)
	Khorrami et al. 15 [139]	zero-bias CNN		4	7m	✓	✓	-			✗	Test: 88.6
	Ding et al. 17 [111]	CNN	fine-tune	8	11m	IntraFace	✓	-			✗	Test: 88.9 (87.7)

Fig 5: Related Work

FER 2013	Tang 13 [130]	CNN	loss layer	4	12.0m	-	✓	IN	Training Set: 28,709 Validation Set: 3,589 Test Set: 3,589	✗	Test: 71.2
	Devries et al. 14 [58]	CNN	$\mathcal{MN}$	4	12.0m	MoT	✓	IN		✗	Validation+Test: 67.21
	Zhang et al. 15 [144]	CNN	$\mathcal{MN}$	6	21.3m	SDM	-	-		✗	Test: 75.10
	Guo et al. 16 [145]	CNN	loss layer	10	2.6m	SDM	✓	-		k-NN	Test: 71.33
	Kim et al. 16 [146]	CNN	$\mathcal{NE}$	5	2.4m	IntraFace	✓	IN		✗	Test: 73.73
	pramerdorfer et al. 16 [147]	CNN	$\mathcal{NE}$	10/16/33	1.8/1.2/5.3 (m)	-	✓	IN		✗	Test:75.2
SFEW 2.0	levi et al. 15 [78]	CNN	$\mathcal{NE}$	VGG-S/VGG-M/ GoogleNet		MoT	✓	-	891 training, 431 validation, and 372 test	✗	Validation: 51.75 Test: 54.56
	Ng et al. 15 [63]	CNN	fine-tune	AlexNet		IntraFace	✓	-	921 training, ? validation, and 372 test	✗	Validation: 48.5 (39.63) Test: 55.6 (42.69)
	Li et al. 17 [44]	CNN	loss layer	8	5.8m	IntraFace	✓	-	921 training, 427 validation	SVM	Validation: 51.05
	Ding et al. 17 [111]	CNN	fine-tune	8	11m	IntraFace	✓	-	891 training, 425 validation	✗	Validation: 55.15 (46.6)
	Liu et al. 17 [77]	CNN	loss layer	11	-	IntraFace	✓	IN	958training, 436 validation, and 372 test	✗	Validation: 54.19 (47.97)
	Cai et al. 17 [140]	CNN	loss layer	6	-	DRMF	✓	IN		✗	Validation: 52.52 (43.41) Test: 59.41 (48.29)
	Meng et al. 17 [61]	CNN	$\mathcal{MN}$	6	-	DRMF	✓	-		✗	Validation: 50.98 (42.57) Test: 54.30 (44.77)
	Kim et al. 15 [76]	CNN	$\mathcal{NE}$	5	-	multiple	✓	IN		✗	Validation: 53.9 Test: 61.6
	Yu et al. 15 [75]	CNN	$\mathcal{NE}$	8	6.2m	multiple	✓	IN		✗	Validation: 55.96 (47.31) Test: 61.29 (51.27)

Fig 6: Related work

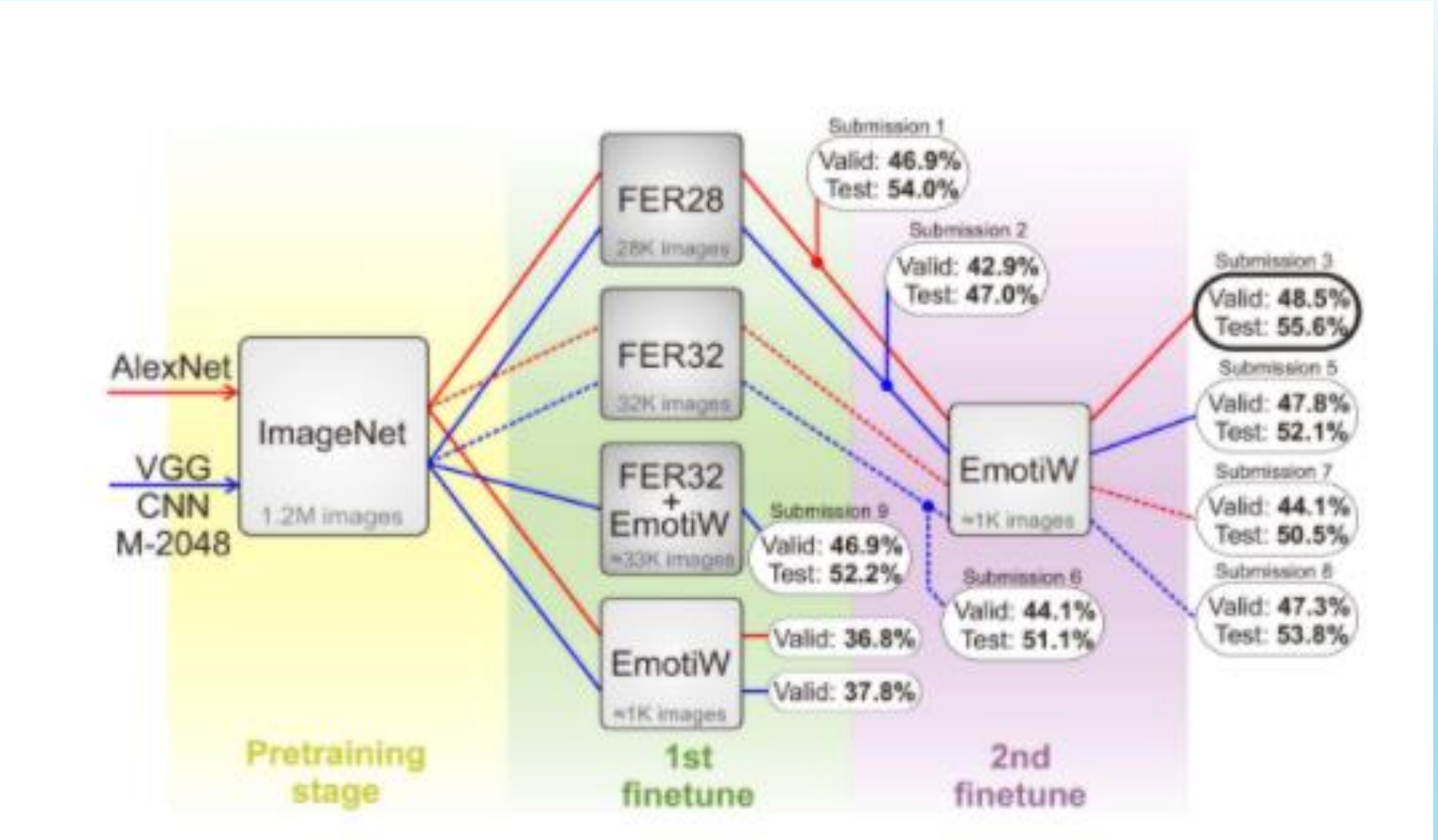


Fig 7: Flowchart of the different fine-tuning combinations used Here, “FER28” and “FER32” indicate different parts of the FER2013datasets. “EmotiW” is the target dataset. The proposed two-stage fine-tuning strategy exhibited the bestperformance.

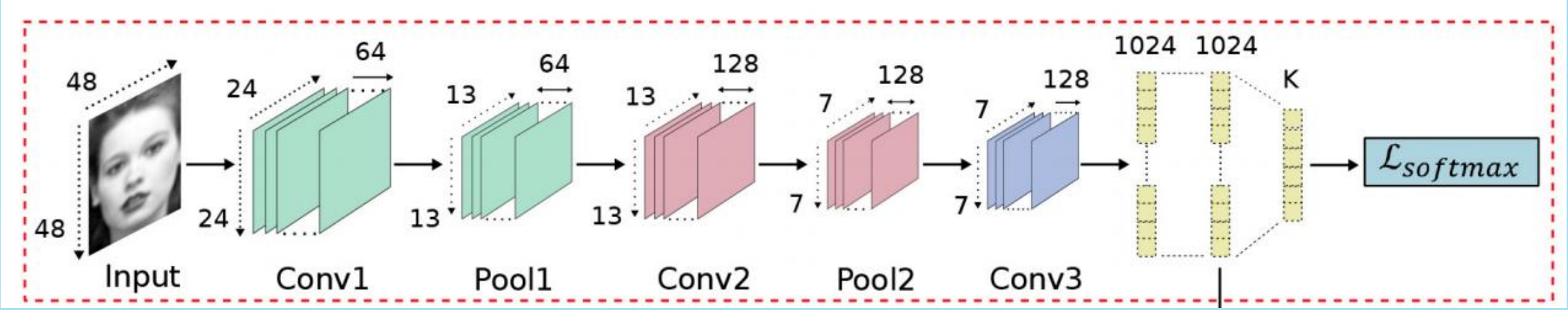


Fig 8: The island loss calculated at the feature extraction layer and the softmax loss calculated at the decision layer are combined to supervise the CNN training.

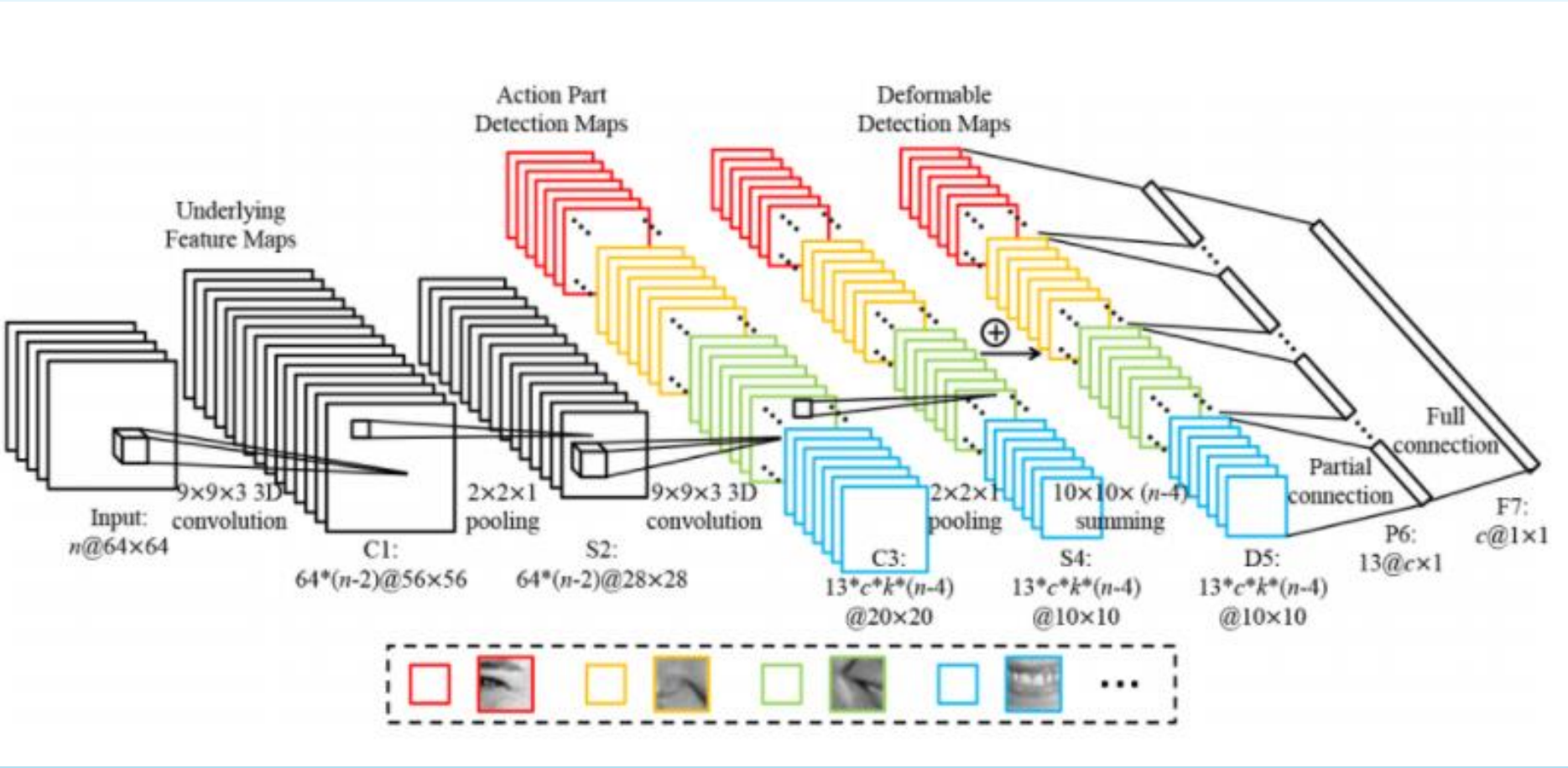


Fig 9: The proposed 3DCNN-DAP The input n-frame sequence is convolved with 3D filters; then, 13 ∗ c ∗ k part filters corresponding to 13 manually defined facial parts are used to convolve k feature maps for the facial action part detection maps of c expression classes.

DATA SETS

1. Karolinska Directed Emotional Faces [KDEF]

- It contains 4900 JPEG images showing 70 people (35 women and 35 men) displaying seven different facial expressions.
- Each expression is recorded from 5 different angles.
- All the participants were amateur actors between 20 and 30 years old.

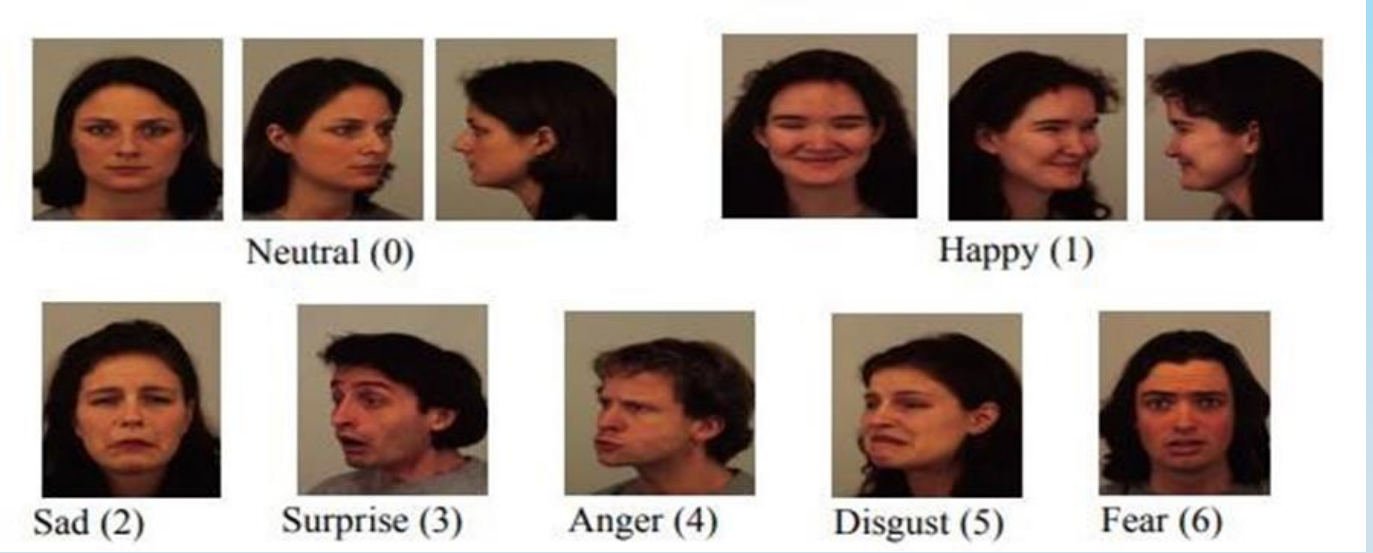
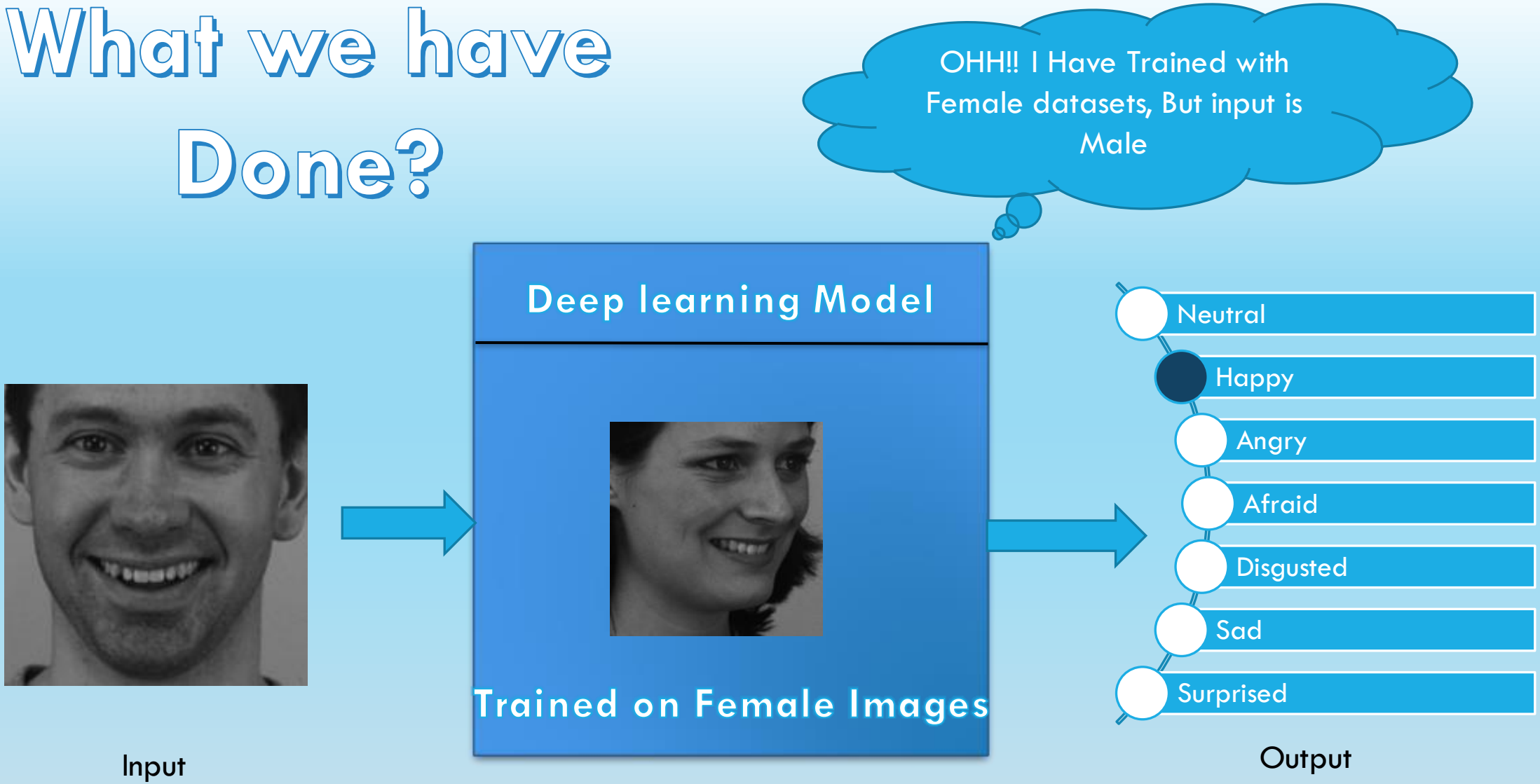


Fig 10: KDEF Dataset

What we have Done?



Training Dataset: 1213 Female images

Test Dataset: 1210 Male Images

Total Process Steps

Fig 11: Step by step Process

DATA PREPROCESSING

Converting Colour to Grayscale:

- Processing colour images are complex and time consuming, and grayscale provides an easy way out.
- To reduce the complexity of the code, we have converted colour images to grayscale.
- Another important advantage of converting colour to gray is that we can reduce the processing time.

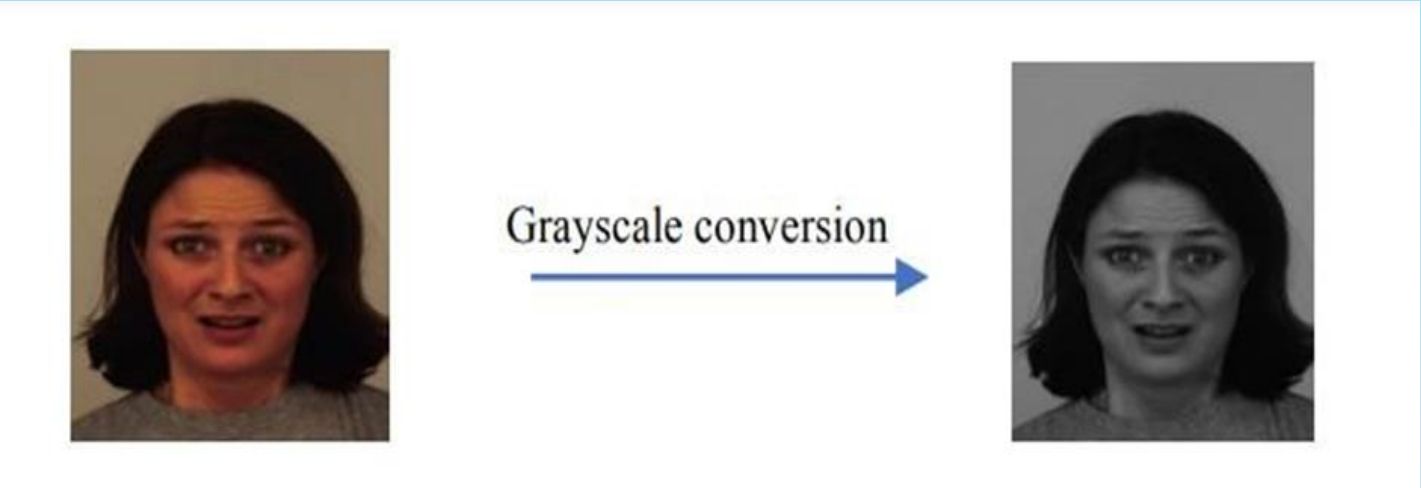


Fig 12: Grayscale Conversion

2 Image Cropping:

- The neural network model has an extra problem to solve, differentiating between foreground and background information.
- In this method, the object is detected using Haar feature-based cascade classifiers.
- The resulted image does not contain facial parts that do not contribute to the expression (e.g., hair, ears, etc.).

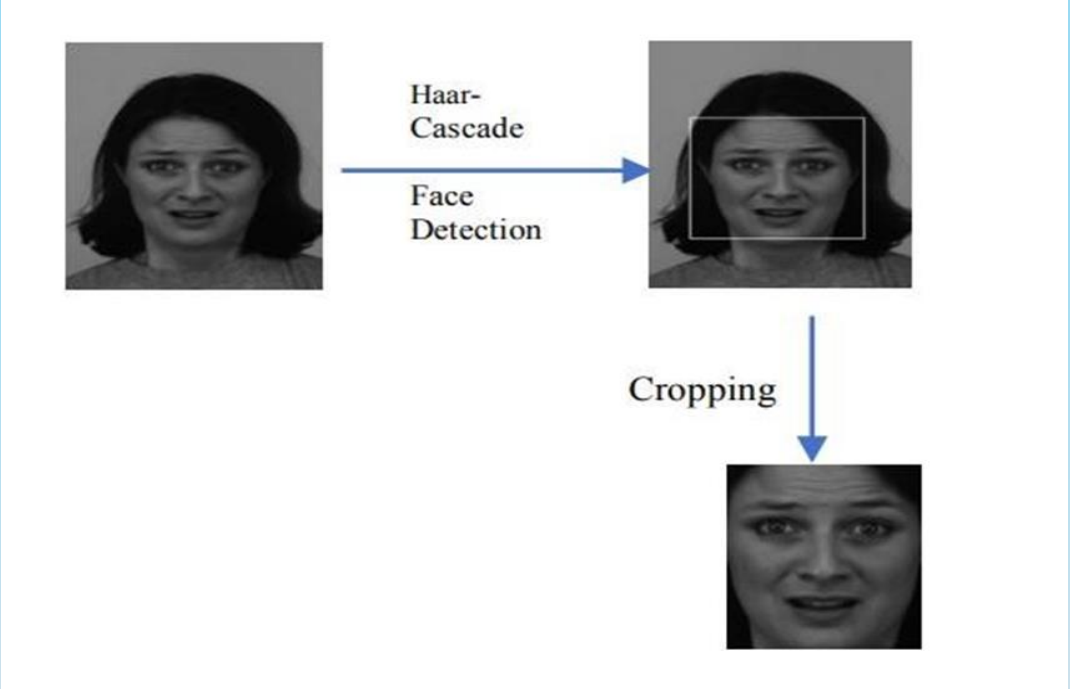


Fig 13: Haar Cascade and Cropping

OUR MODELS

1. Progressive Resizing
2. Resnet50+Custom Net
3. Our State of Art Model

1.Progressive resizing

- It is the technique to sequentially resize all the images while training the CNNs on smaller to bigger image sizes.

- Progressive Resizing is described briefly in his terrific fastai course, “Practical Deep Learning for Coders”.

- A great way to use this technique is to train a model with smaller image size say 64x64, then use the weights of this model to train another model on images of size128x128 and so on.

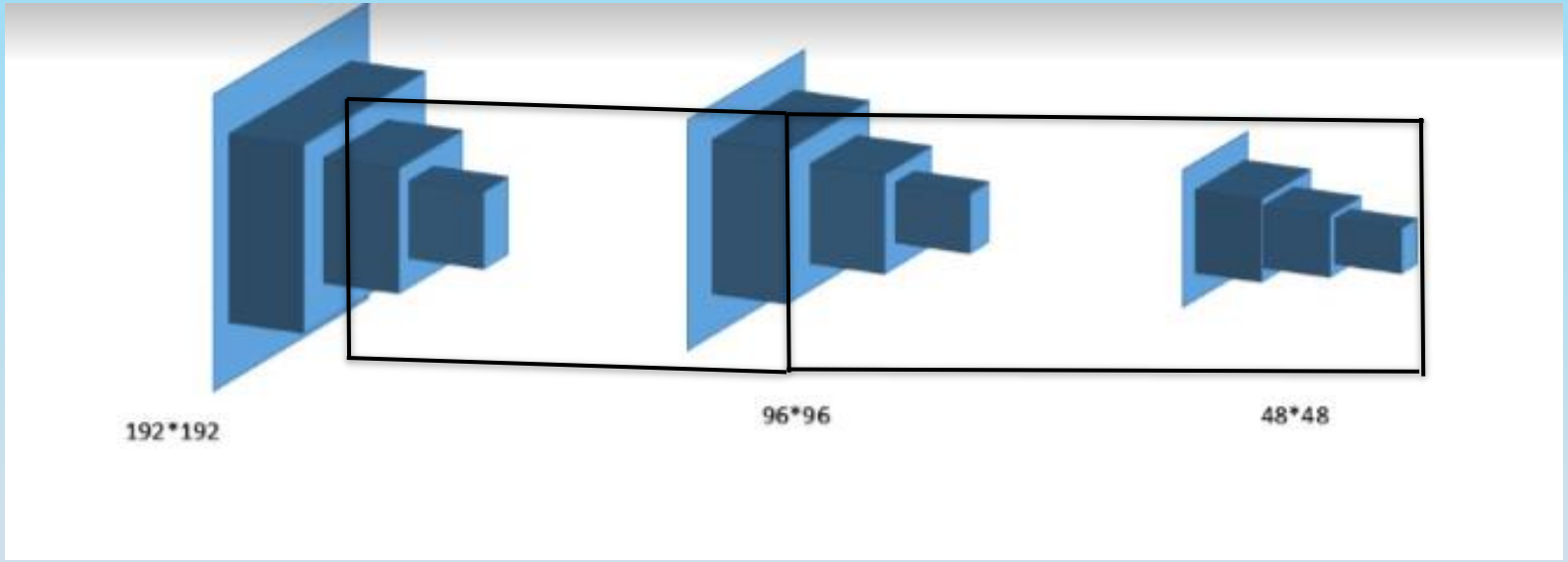


Fig 14: Progressive resizing

KDEF

	precision	recall	f1-score	support
0	0.41	0.40	0.40	175
1	0.49	0.66	0.56	169
2	0.63	0.69	0.66	170
3	0.81	0.88	0.84	172
4	0.57	0.51	0.54	175
5	0.60	0.44	0.51	175
6	0.75	0.65	0.70	174
accuracy			0.60	1210
macro avg	0.61	0.60	0.60	1210
weighted avg	0.61	0.60	0.60	1210

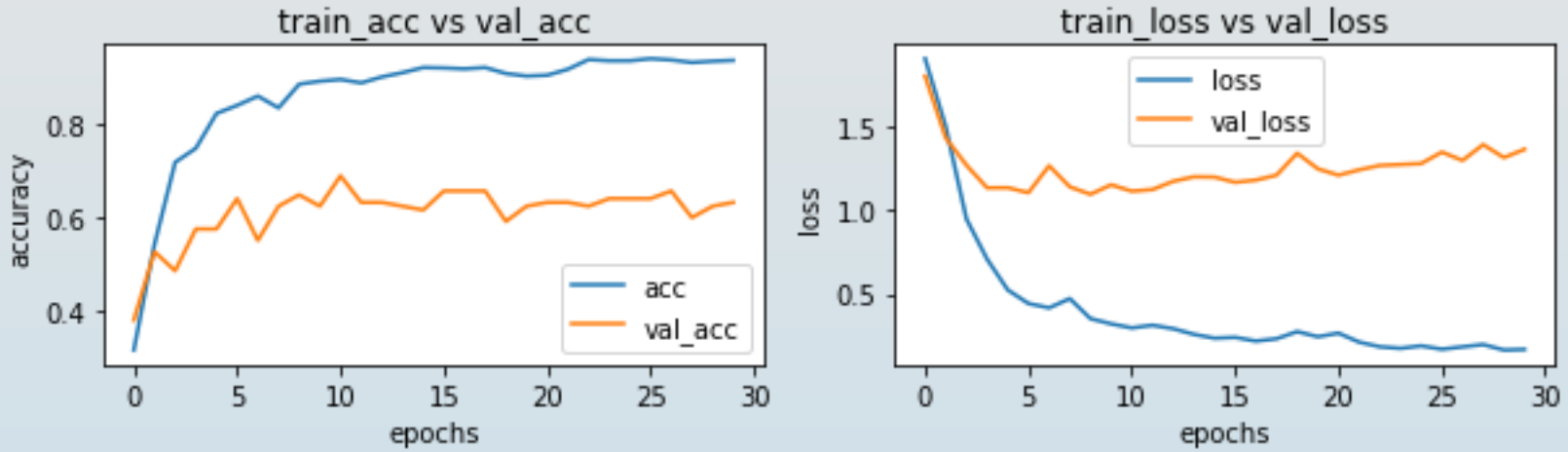
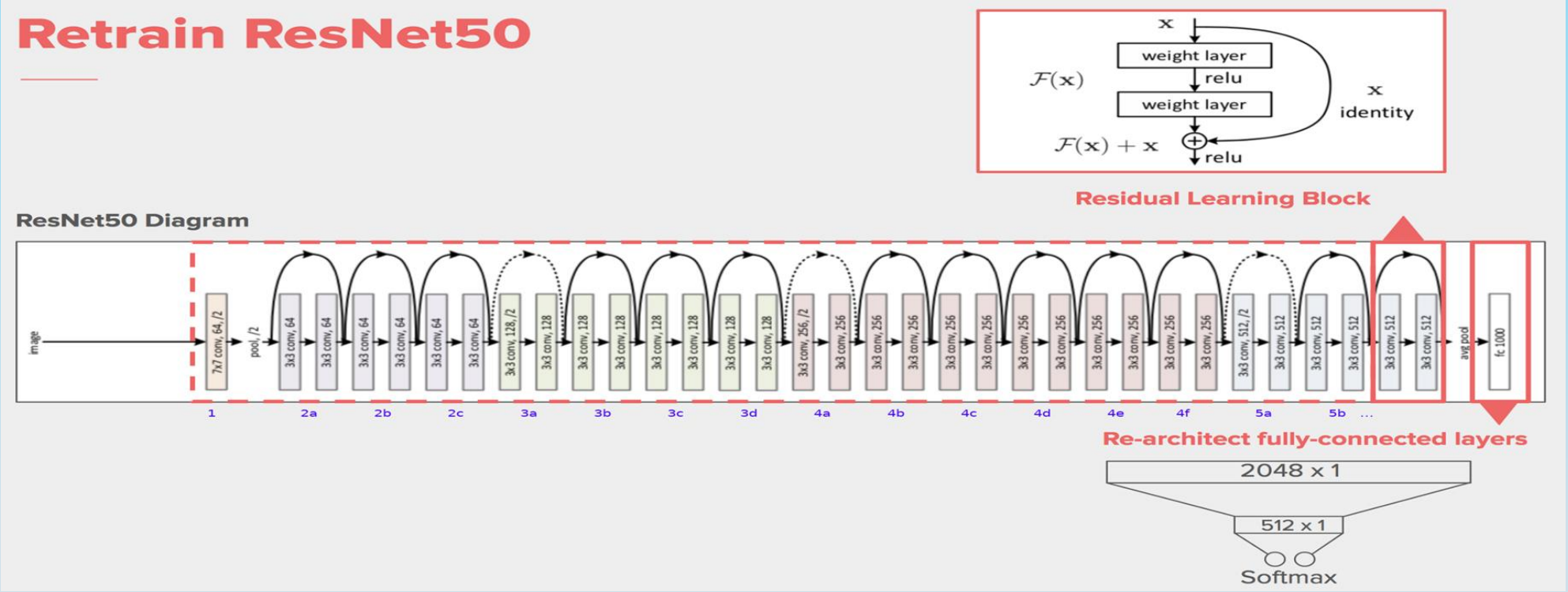


Fig 15 : Graph Epoch vs Accuracy and Epochs vs Loss

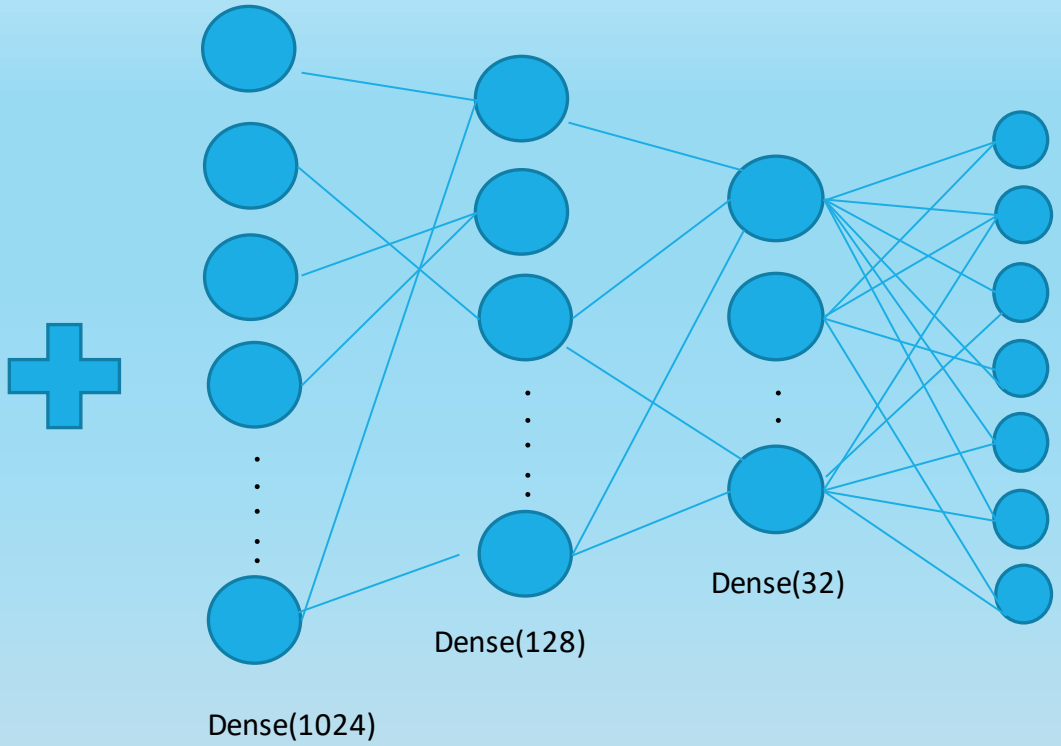
2.RESNET50

- Resnet is short name for Residual Network that supports Residual Learning. The 50 indicates the number of layers that it has. So Resnet50 stands for Residual Network with 50 layers.
- It is a widely used ResNet model and has 48 Convolution layers along with 1 MaxPool and 1 Average Pool layer.



Transfer Learning with ResNet50

ResNet50




```
x = layers.Flatten()(last_output)
x = layers.Dense(1024, activation='relu')(x)
x = layers.Dropout(0.2)(x)

x = layers.Dense(128, activation='relu')(x)
x = layers.Dropout(0.2)(x)

x = layers.Dense(32, activation='relu')(x)
x = layers.Dropout(0.2)(x)
x = layers.Dense(7, activation='softmax')(x)
```

Result: Accuracy 0.53

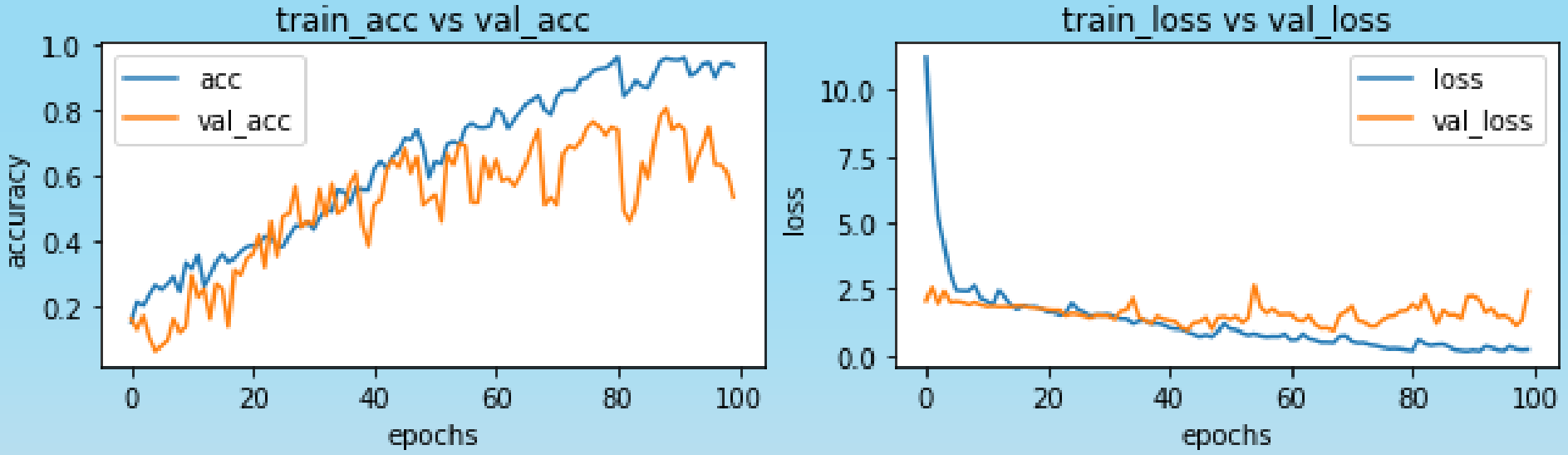


Fig 16 : Graph Epoch vs Accuracy and Epochs vs Loss

3.Our State of art model

Our model is inspired from the research paper written by Octavio Arriaga, Paul G. Plöger, and Matias Valdenegro.

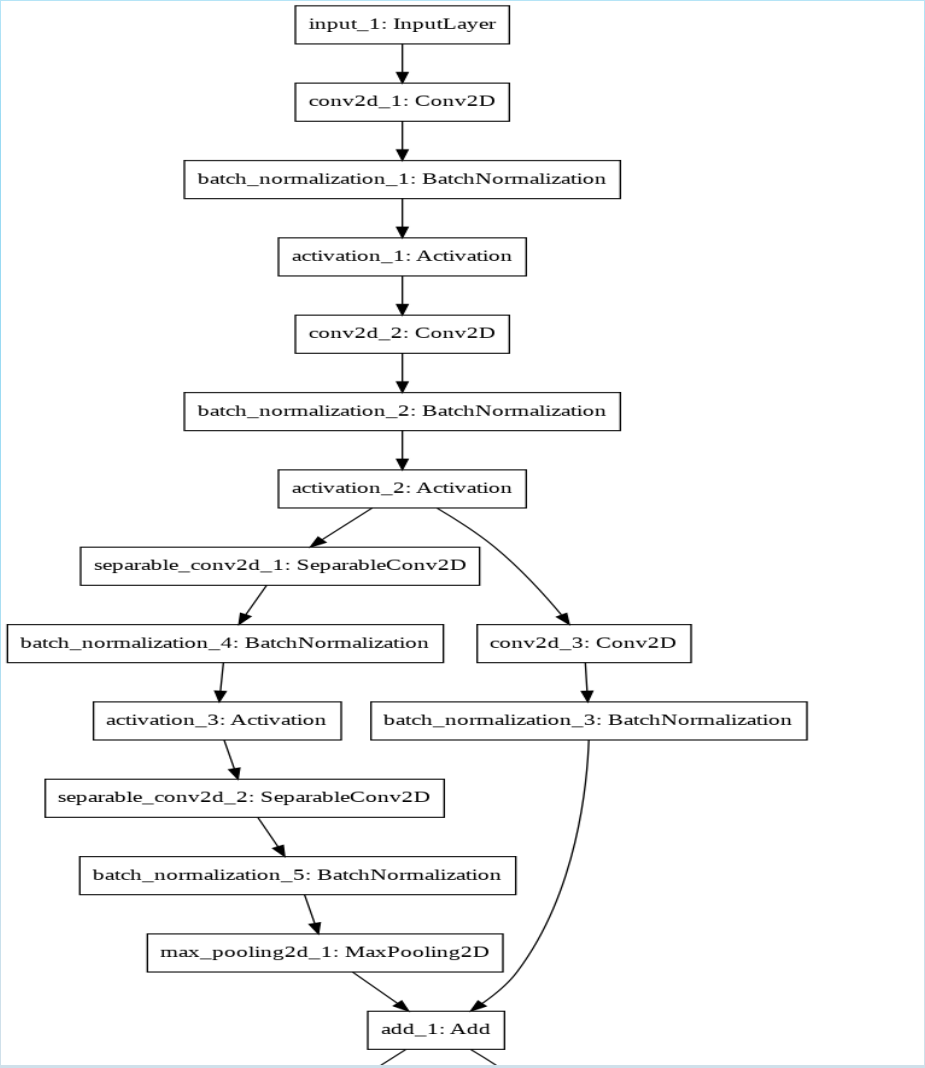


Fig 17 : CNN Architecture Block 1

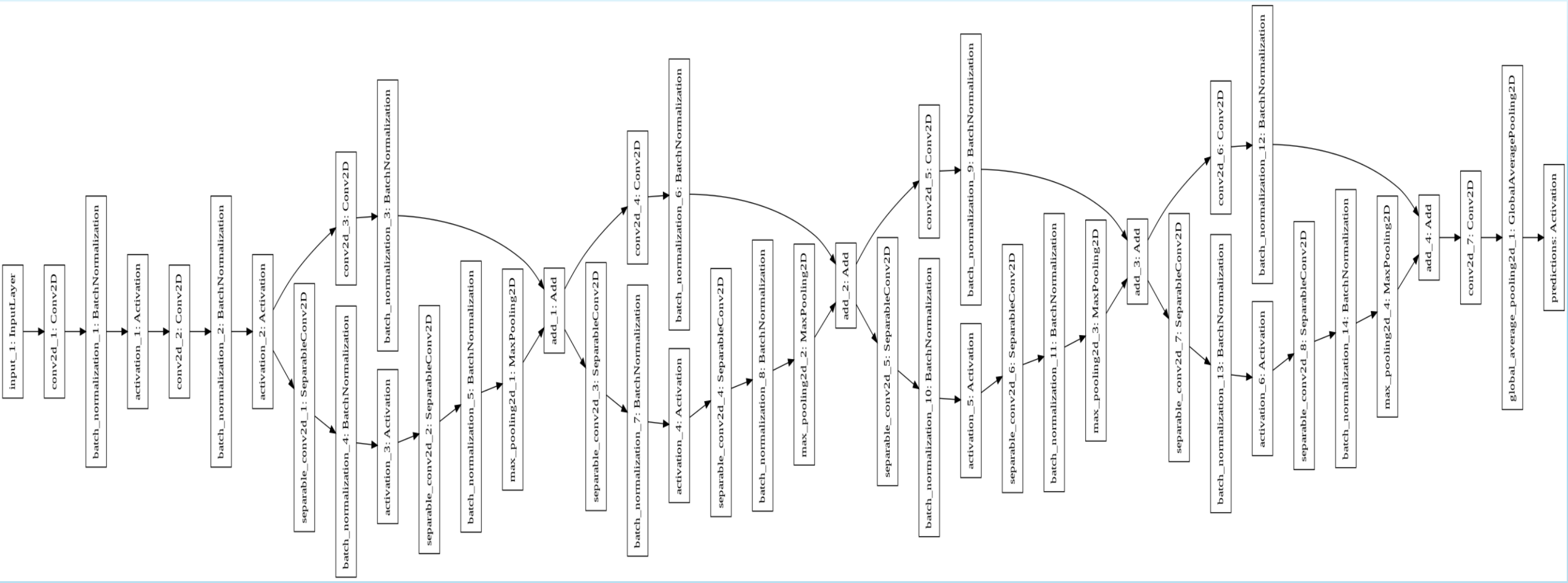
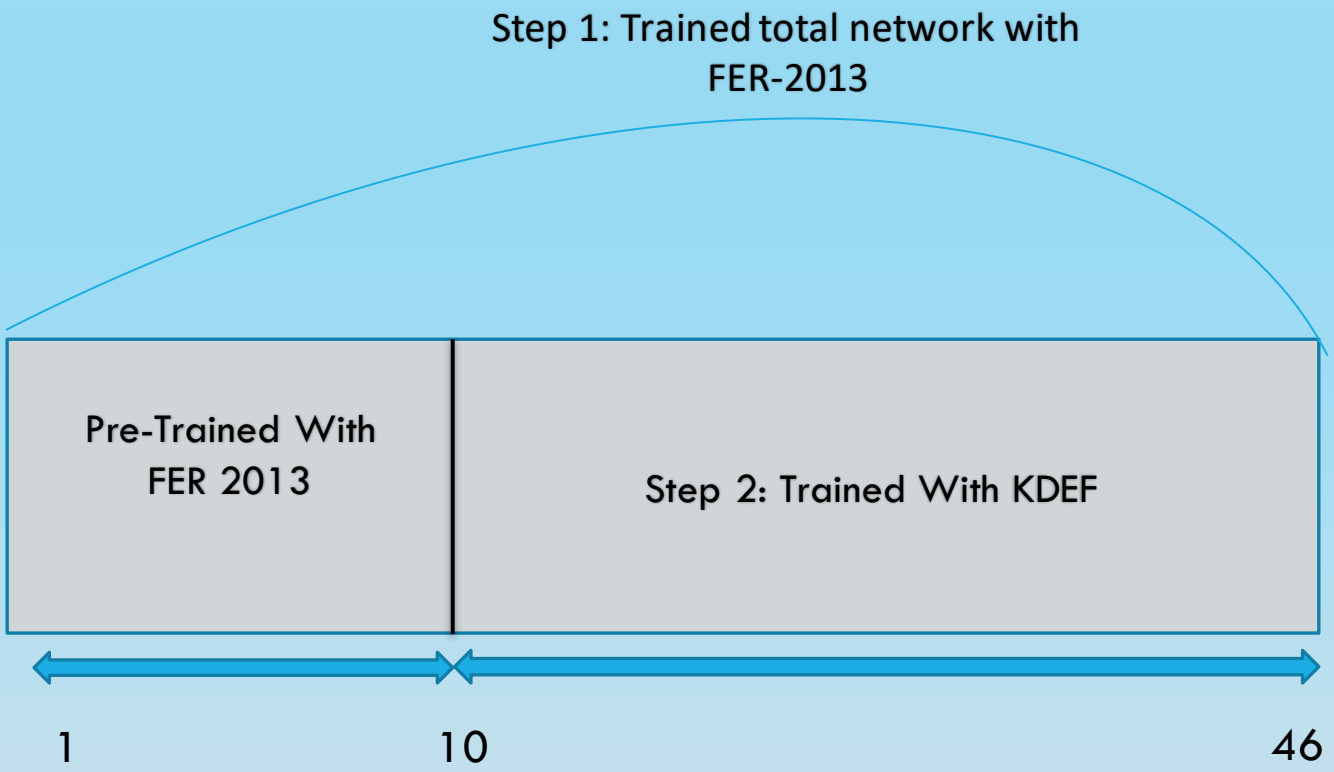


Fig 18 : Complete CNN architecture

Our Process



Step 1: Trained total network with FER-2013

Step 2: Trained With KDEF

Fig 19 : 46 Layer CNN Architecture

Training Process

```
Epoch 392/400
18/18 [==============================] - 9s 475ms/step - loss: 0.0225 - accuracy: 0.9973 - val_loss: 0.5177 - val_accuracy: 0.7705

Epoch 00392: val_accuracy did not improve from 0.84426
Epoch 393/400
18/18 [==============================] - 9s 473ms/step - loss: 0.0098 - accuracy: 1.0000 - val_loss: 0.4729 - val_accuracy: 0.7869

Epoch 00393: val_accuracy did not improve from 0.84426
Epoch 394/400
18/18 [==============================] - 9s 473ms/step - loss: 0.0094 - accuracy: 1.0000 - val_loss: 0.4100 - val_accuracy: 0.7869

Epoch 00394: val_accuracy did not improve from 0.84426
Epoch 395/400
18/18 [==============================] - 8s 468ms/step - loss: 0.0102 - accuracy: 1.0000 - val_loss: 0.3639 - val_accuracy: 0.7951

Epoch 00395: val_accuracy did not improve from 0.84426
Epoch 396/400
18/18 [==============================] - 8s 467ms/step - loss: 0.0083 - accuracy: 1.0000 - val_loss: 0.3647 - val_accuracy: 0.7787

Epoch 00396: val_accuracy did not improve from 0.84426
Epoch 397/400
18/18 [==============================] - 9s 479ms/step - loss: 0.0328 - accuracy: 0.9887 - val_loss: 0.5103 - val_accuracy: 0.7541

Epoch 00397: val_accuracy did not improve from 0.84426
Epoch 398/400
18/18 [==============================] - 9s 495ms/step - loss: 0.0151 - accuracy: 0.9955 - val_loss: 0.5003 - val_accuracy: 0.7459

Epoch 00398: val_accuracy did not improve from 0.84426
Epoch 399/400
18/18 [==============================] - 9s 476ms/step - loss: 0.0220 - accuracy: 0.9955 - val_loss: 0.7279 - val_accuracy: 0.7049

Epoch 00399: val_accuracy did not improve from 0.84426
Epoch 400/400
18/18 [==============================] - 8s 467ms/step - loss: 0.1463 - accuracy: 0.9492 - val_loss: 1.2593 - val_accuracy: 0.5984

Epoch 00400: val_accuracy did not improve from 0.84426
```

	precision	recall	f1-score	support
0	0.57	0.66	0.61	175
1	0.65	0.64	0.65	169
2	0.77	0.75	0.76	170
3	0.94	0.89	0.92	172
4	0.78	0.65	0.71	175
5	0.57	0.77	0.65	175
6	0.89	0.69	0.78	174
accuracy			0.72	1210
macro avg	0.74	0.72	0.73	1210
weighted avg	0.74	0.72	0.73	1210

Fig 20 : Classification Report

Parameter Used

Loss: Categorical Cross entropy
Optimizer: Adam
Learning Rate: 0.01
Batch size: 64
Epochs: 400

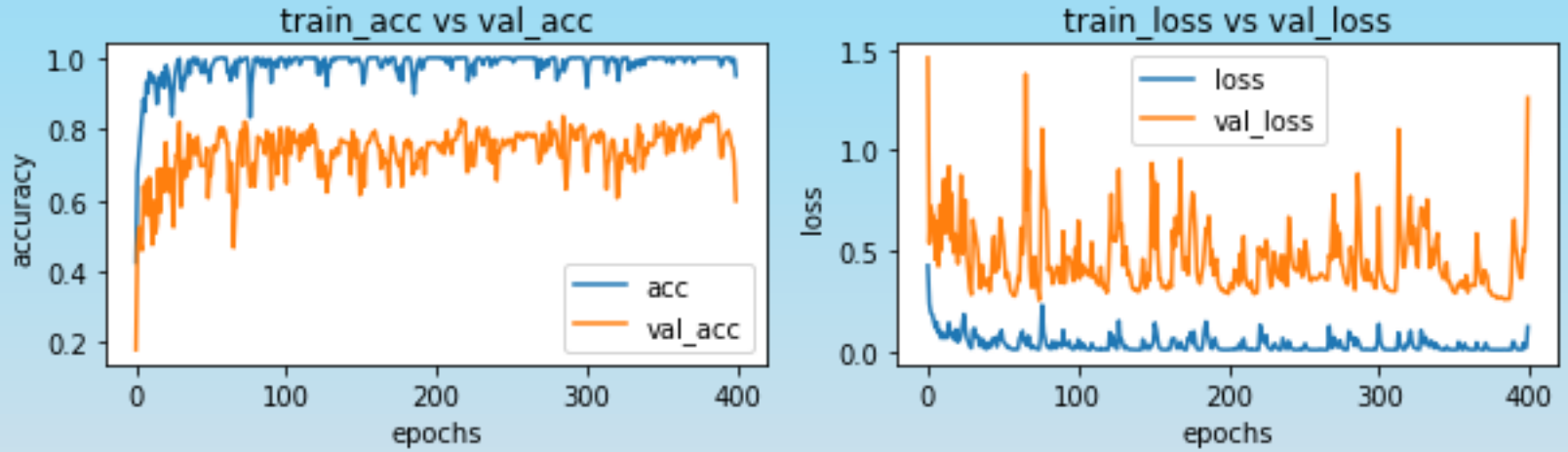


Fig 21 : Graph Epoch vs Accuracy and Epochs vs Loss

FUTURE WORK

- Future work One of the most obvious continuations of this work is the creation of a multimodal model that combines the video and audio inputs, and therefore significantly improves the results.
- Surely if we applied the models created in a real application, its effectiveness would be deficient, since the models have been created with acted databases and the categories of emotions are very rigid.
- Probably for a future work the model needs to be tested against different types of data, changing the background and light conditions too.
- Also, it could be investigated if applying another emotion classification, such as dimensional, which takes into account the intensity of the emotions, could be more effective for recognizing the most intense one and not erring in detecting the most subtle ones

THANK YOU EVERYONE
#STAY SAFE