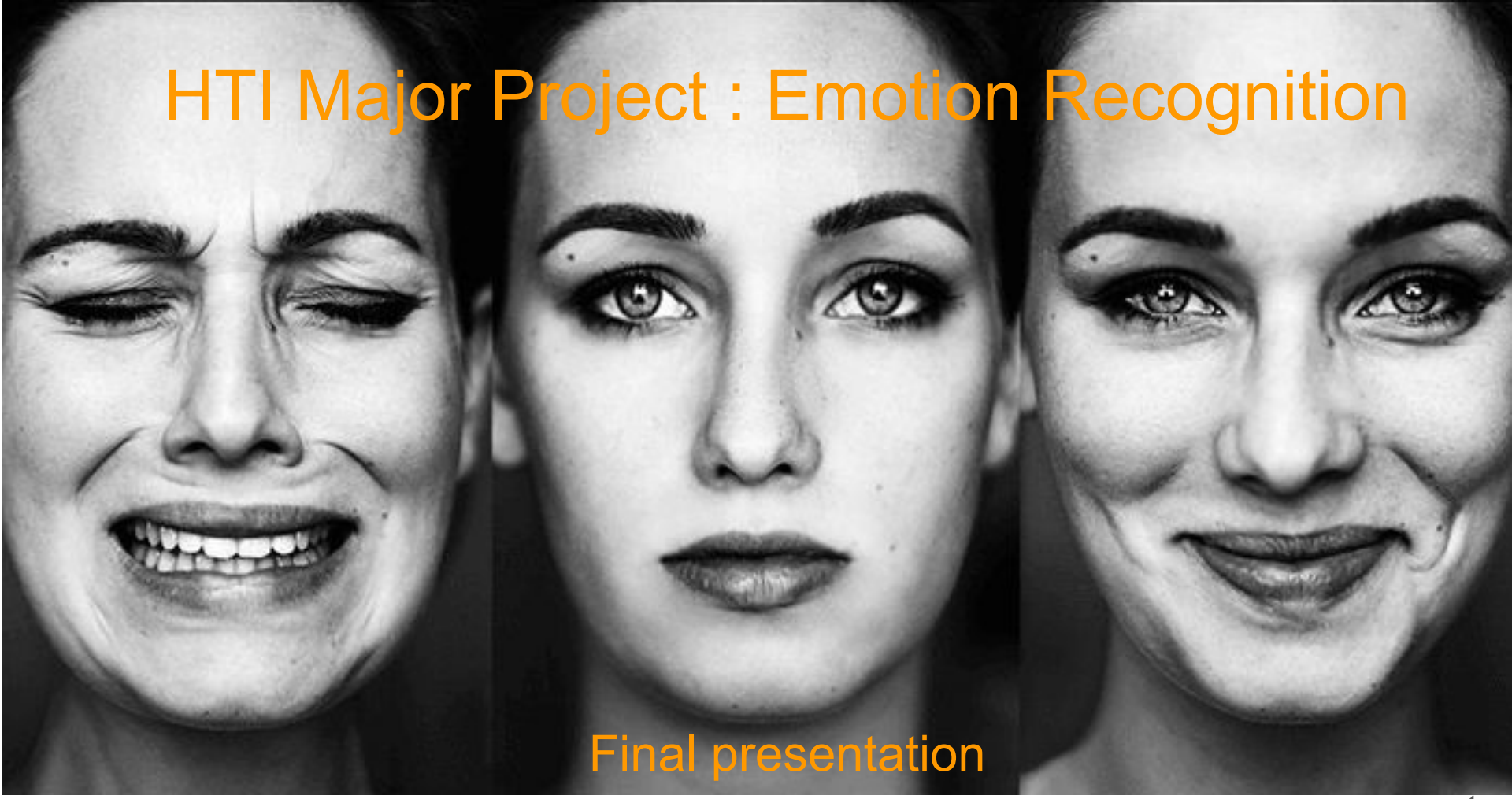
HTI Major Project : Emotion Recognition

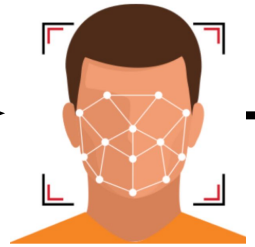


Final presentation

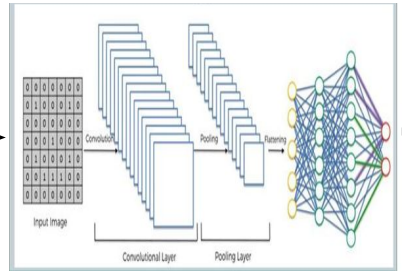
Vincent TAUFFLIEB Daniel BUTSANETS Tutor: Ruxandra TAPU

Table of contents

Goal



Face Detection

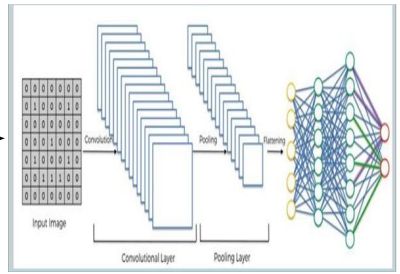


Deep Learning

Prediction Video

Final Prediction

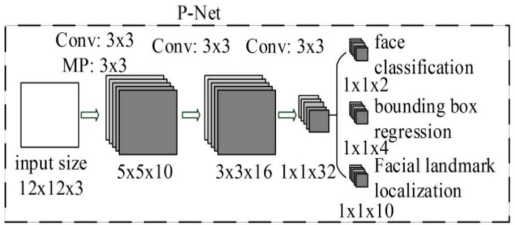
Deep Learning



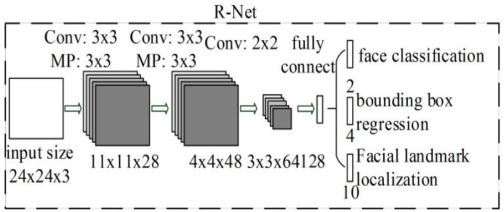
Prediction Sound

I. Recap of our project : Face Detection

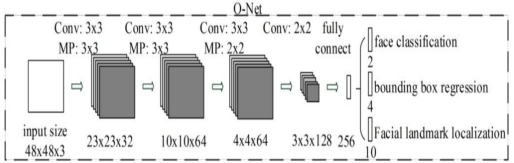
→ The Proposal Network (P-Net)



→ The Refine Network (R-Net)



→ The Output Network (O-Net)



Source: *Joint Face Detection and Alignment using Multi-task Cascaded Convolutional Networks*
Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, Yu Qiao

I. Recap of our project : Creating the dataset

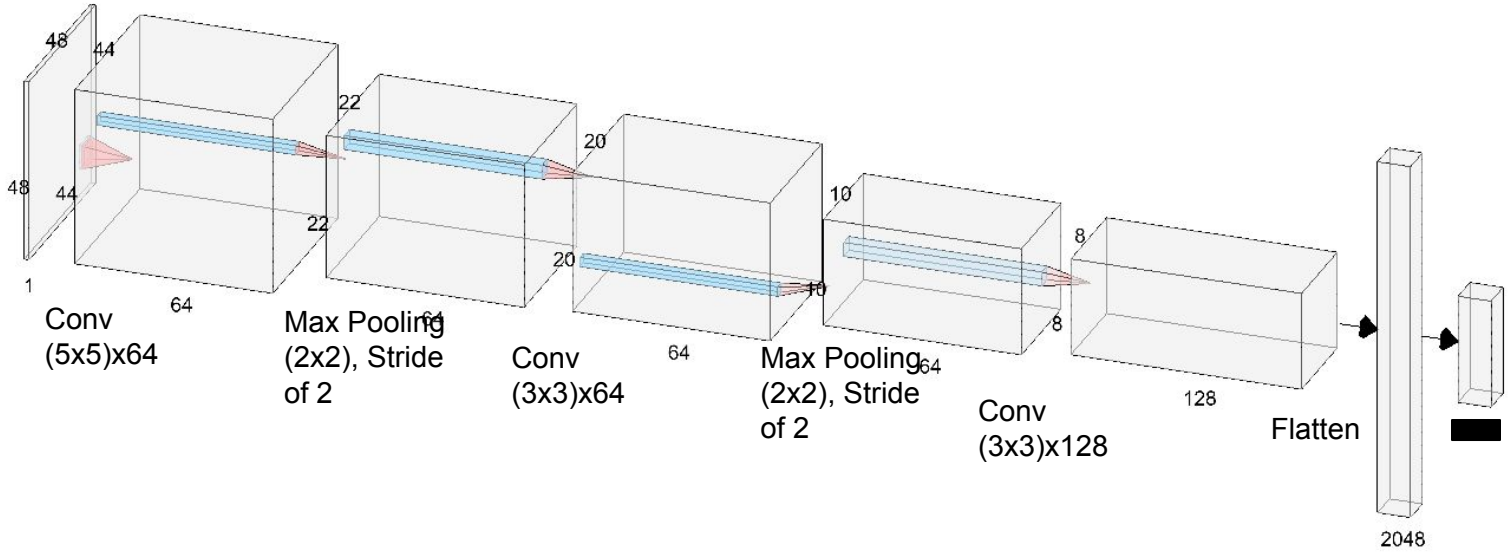
→ FER : 33k images, 48x48 pixels, grayscale, centered on the face, 7 labels.

→ FER+ : 8 labels.

7 basic discrete emotions : Anger, Disgust, Fear, Happiness, Neutral, Sadness, Surprise
Additional emotion : contempt

Based upon :
https://github.com/microsoft/FERPlus

I. Recap of our project : Architecture of the CNN



Made using : http://alexlenail.me/NN-SVG/AlexNet.html

Based upon : https://github.com/isseu/emotion-recognition-neural-networks/blob/master/paper/Report_NN.pdf

I. Recap of our project : Reshaping the dataset

Emotion	Anger	Happiness	Neutral	Sadeness	Surprise
# of images training set	2466	2466	2466	2466	2466
# of images test set	325	325	325	325	325

Structure of the reduced dataset

I. Recap of our project

Predictions with the current dataset

→ Very good results for happiness

→ Good results for anger, neutral and surprise

→ Slightly worse results for sadness

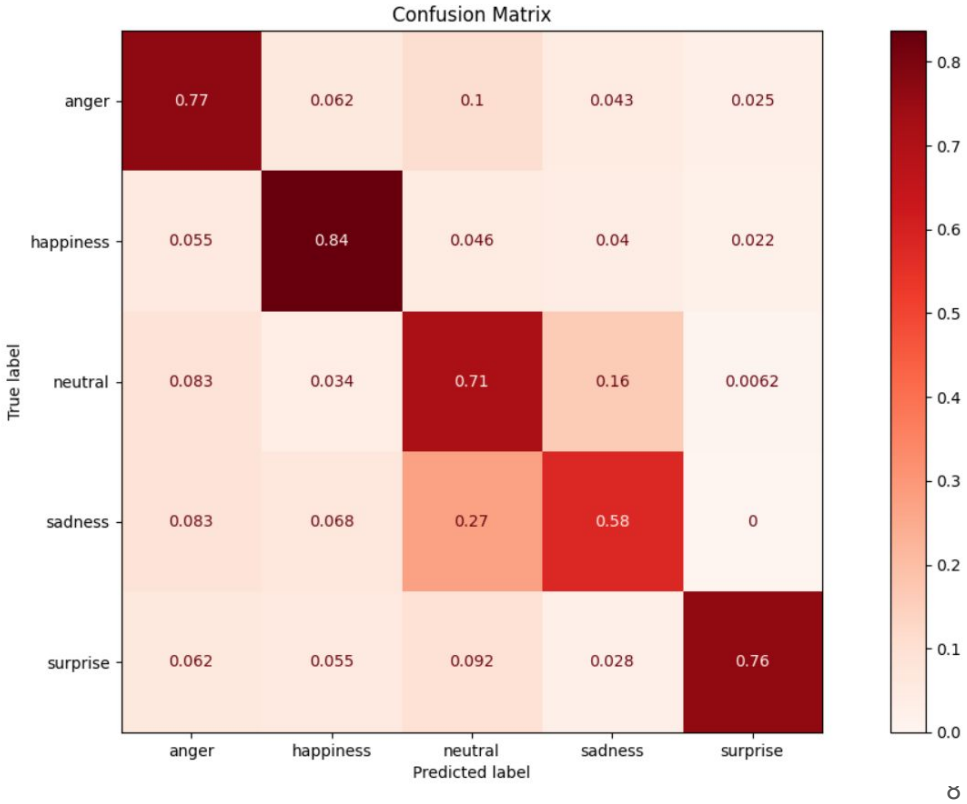
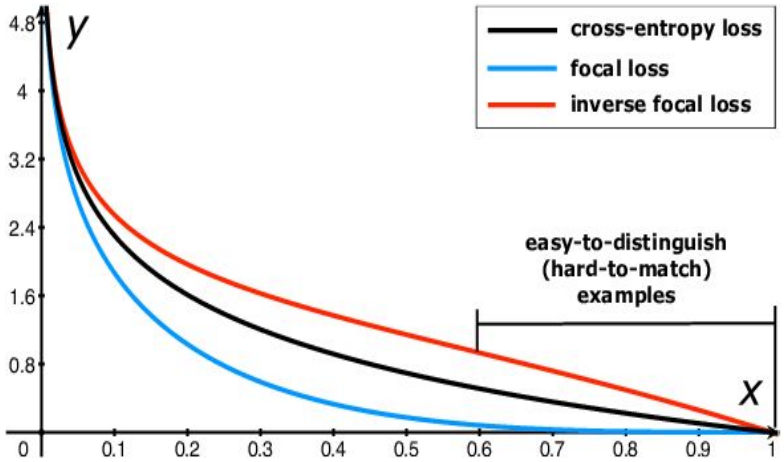


Table of contents

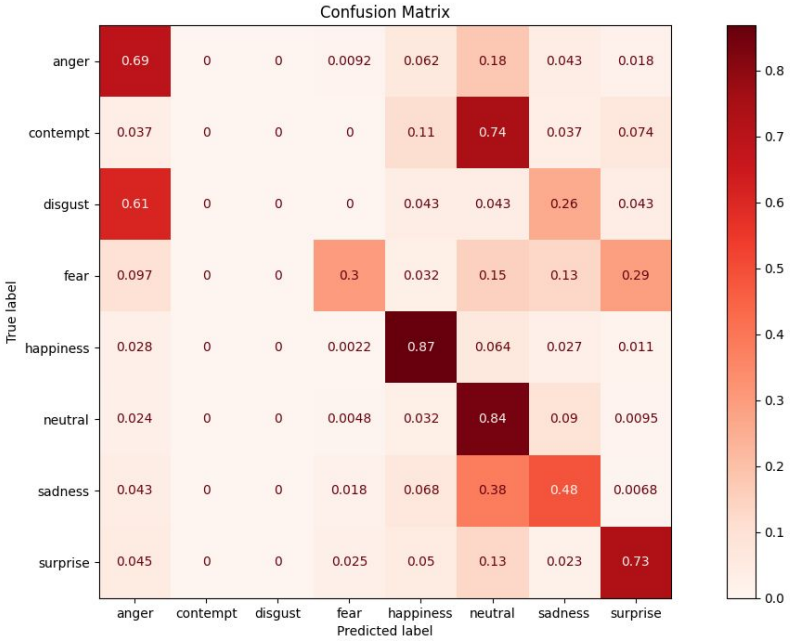
II. Observation from last defense : Use of focal loss

- Used to compensate the imbalanced classes
- Allows better prediction for under represented
  Classes
- In our case : Fear, Disgust, Contempt

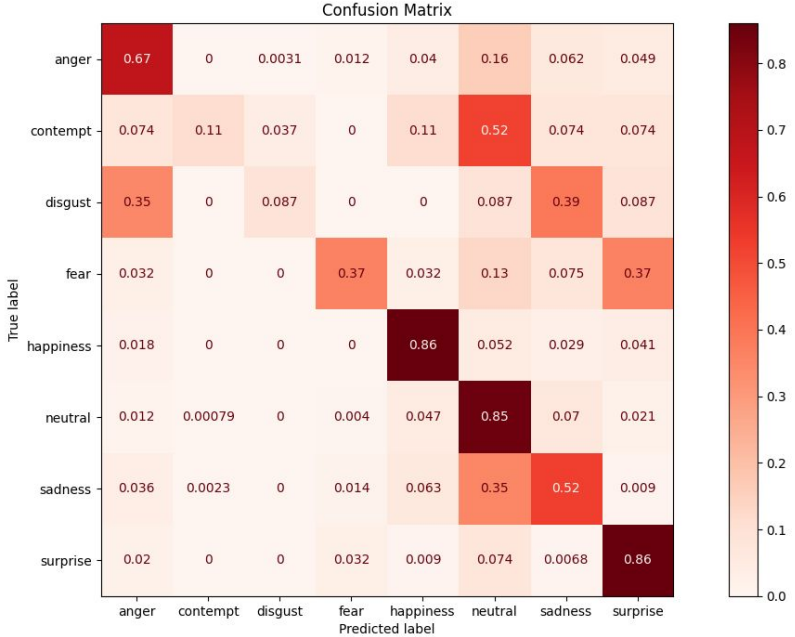


Based upon :
Focal Loss for Dense Object Detection, 7 August 2017
Authors: Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, Poitr Dollár
Code from : https://github.com/mkocabas/focal-loss-keras/blob/master/focal_loss.py

II. Observation from last defense : Use of focal loss



Confusion Matrix for the initial dataset **WITHOUT** Focal Loss



Confusion Matrix for the initial dataset **WITH** Focal Loss

II. Observation from last defense : Use of focal loss

- Previously, disgust and contempt not detected → now detected (~ +10%)
- Very good results for surprise (+ 13%)
- Better results for classes with average predictions : fear and sadness (~ +5%)
- Slightly worse results for happiness and anger (~ -2%)

Table of contents

II. Emotion recognition from video : Adapting the dataset

Emotion	Anger	Disgust	Fear	Happiness	Neutral	Sadeness	Surprise
# of images training set	2466	191	652	2466	2466	2466	2466
# of images test set	325	23	93	325	325	325	325

Going back to more classes now that we have a good model

II. Emotion recognition from video : Enhancing the dataset

Emotion	Anger	Disgust	Fear	Happiness	Neutral	Sadeness	Surprise
# of images training set	2466	550	998	2466	2466	2466	2466
# of images test set	325	23	93	325	325	325	325

II. Emotion recognition from video : CREMA-D

→ Use CREMA-D : speech and video clips

6 emotions : Anger, Disgust, Fear, Happiness, Neutral, Sadness

91 actors performing emotions with different intensity levels

→ Focus on disgust and fear

→ Over 16,5k images retrieved that need to be sorted

II. Emotion recognition from video : Example of image added

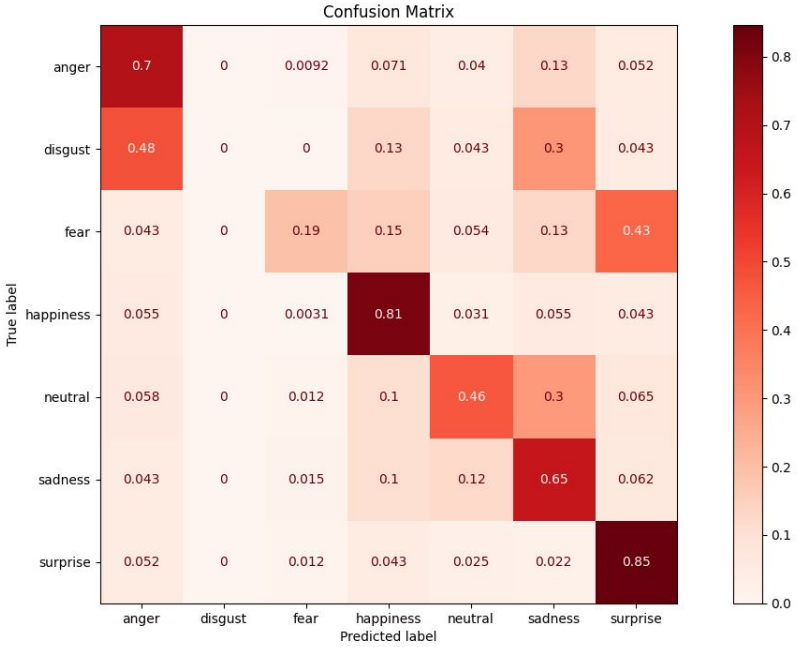
- Images are quite different in the 2 datasets

- Different pose, lighting, actors…

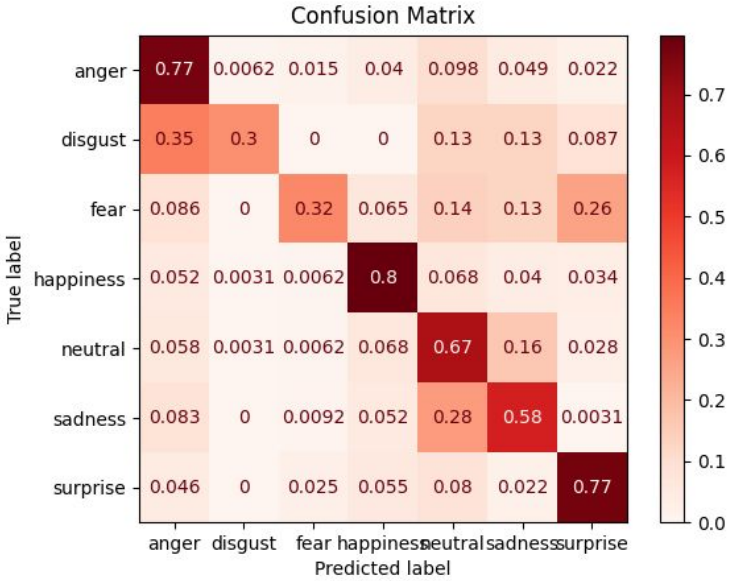
CREMA-D

FER+

II. Emotion recognition from video : Enhancing the dataset

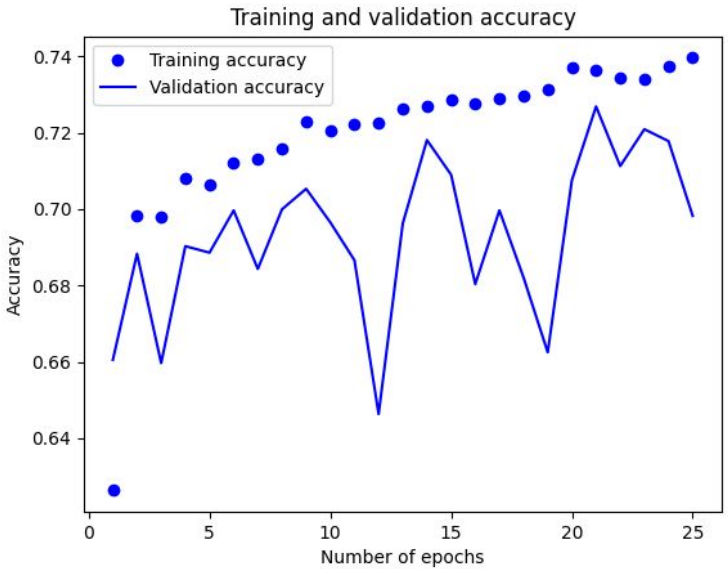


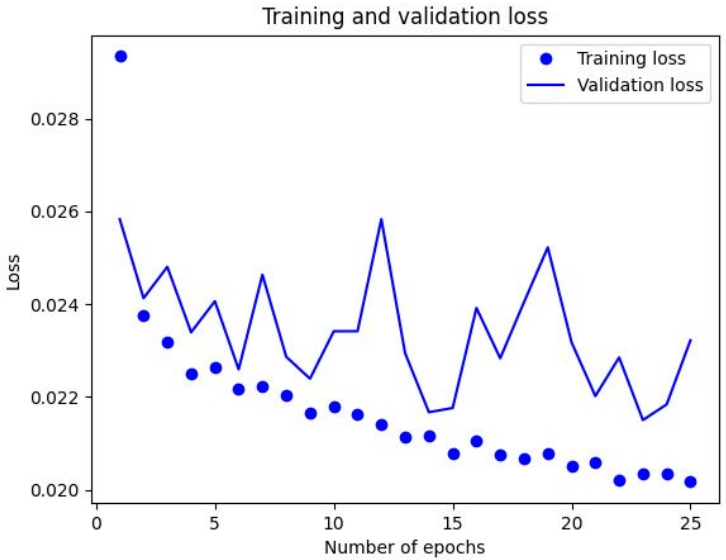
Confusion Matrix for 7 classes **WITHOUT** enhancement



Confusion Matrix for 7 classes **WITH** enhancement

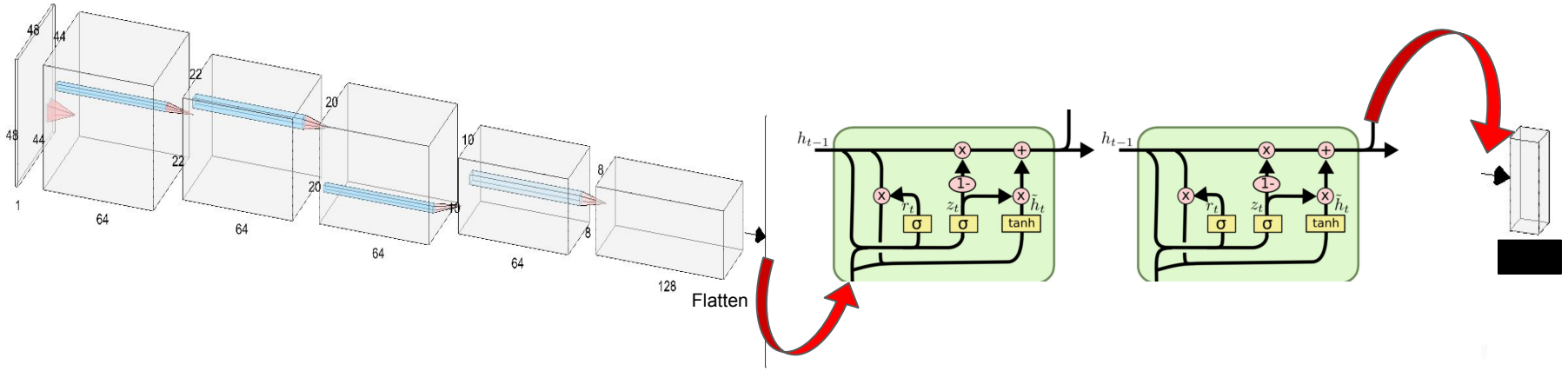
II. Emotion recognition from video : Enhancing the dataset





Training is very fast since we load the weights from the previous model

II. Emotion recognition from video : Time distributed model

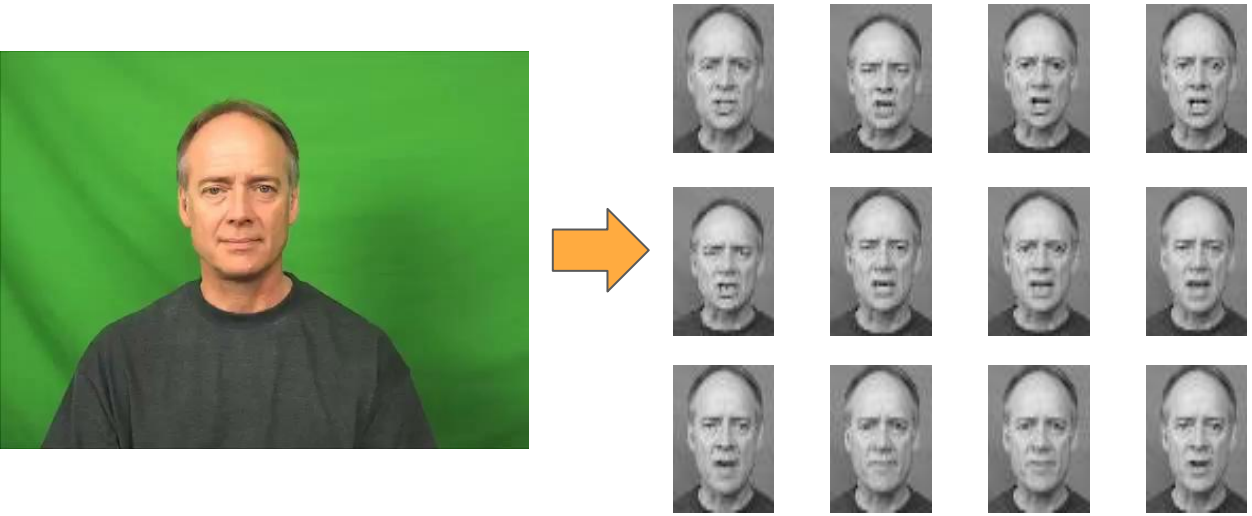


- Need a new dataset : CREMA-D → 6 emotions : Anger, Disgust, Fear, Happiness, Neutral, Sadness
- Need to adapt the data to our model

II. Emotion recognition from video : Work on CREMA-D

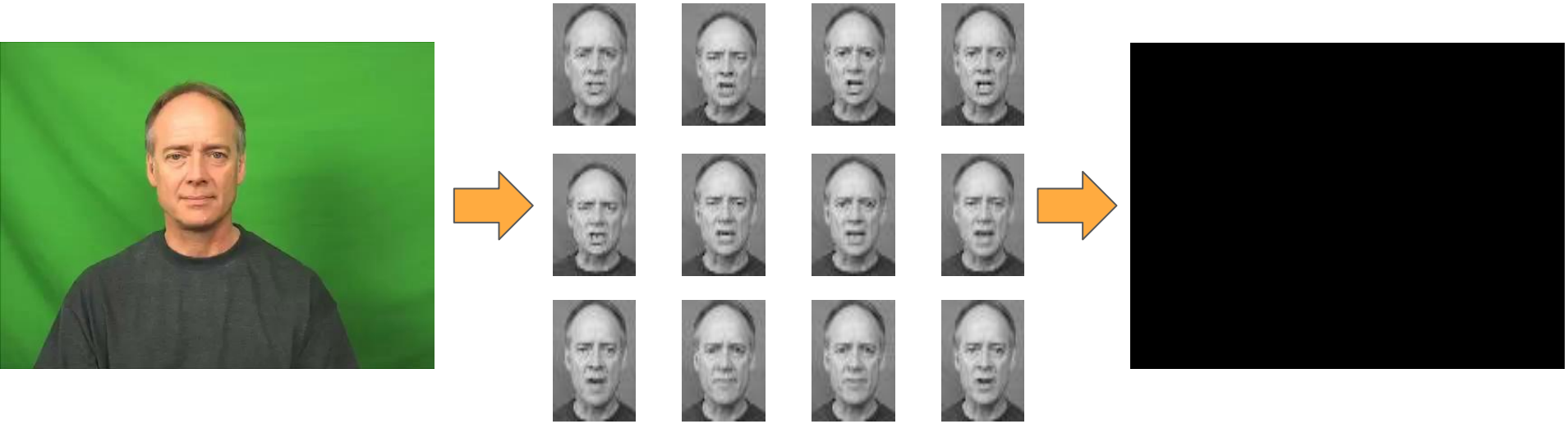
Videos are in color, not the right size and have a wide angle

II. Emotion recognition from video : Work on CREMA-D



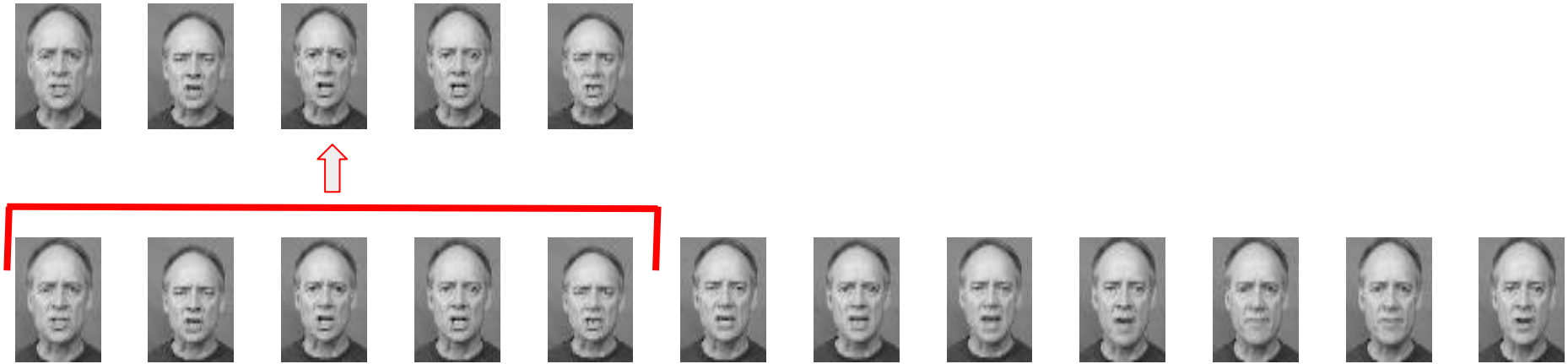
Transform dataset before usage : face detection, resize, convert to grayscale

II. Emotion recognition from video : Work on CREMA-D



Reform the videos to feed it to the network

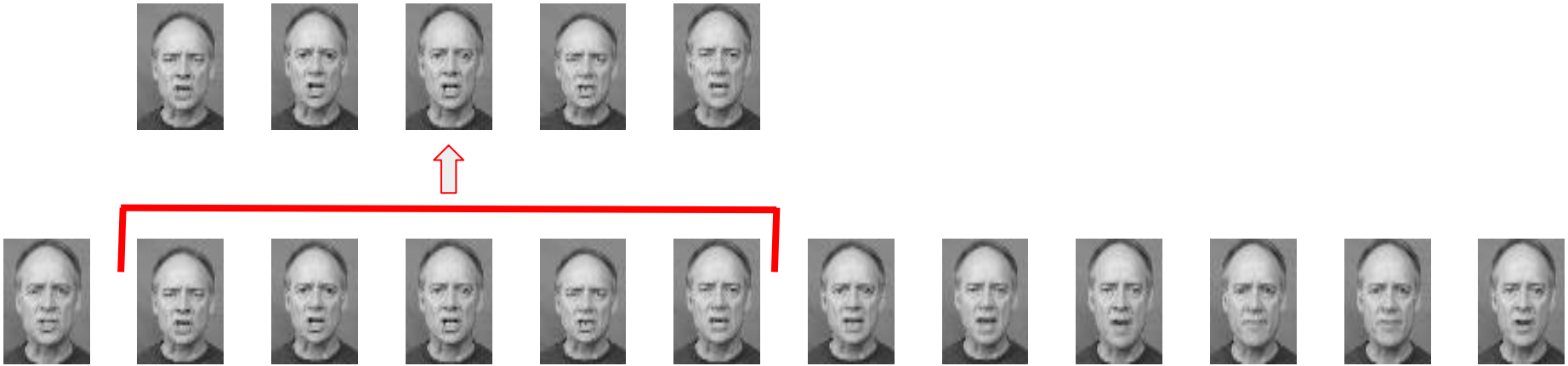
II. Emotion recognition from video : Keras Video Generator



How to choose the sequences fed to the network ?
→ use a sliding generator that provides time related sequences fit for keras

Downloaded from : https://pypi.org/project/keras-video-generators/

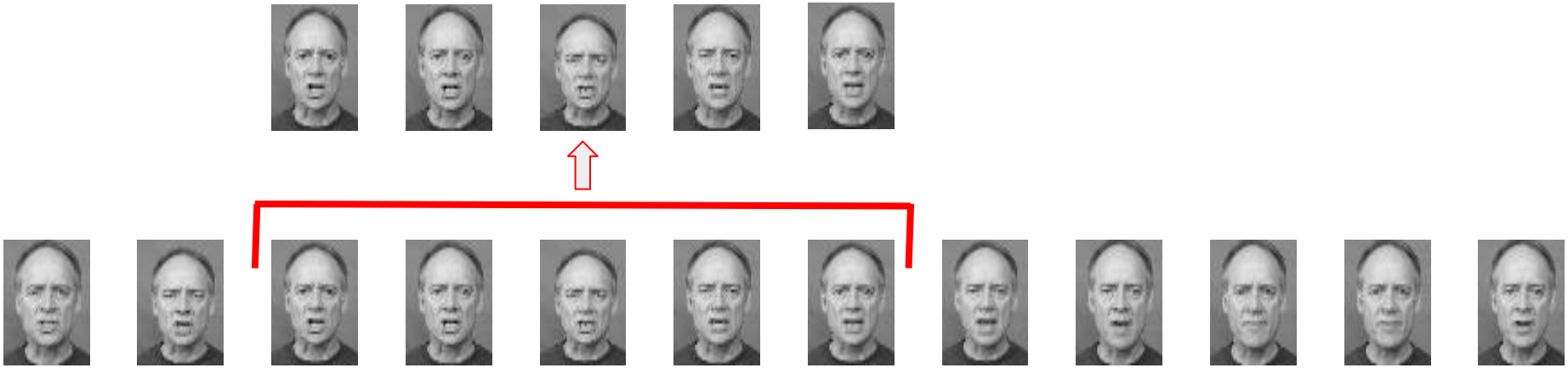
II. Emotion recognition from video : Keras Video Generator



How to choose the sequences fed to the network ?
→ use a sliding generator that provides time related sequences fit for keras

Downloaded from : https://pypi.org/project/keras-video-generators/

II. Emotion recognition from video : Keras Video Generator



How to choose the sequences fed to the network ?
→ use a sliding generator that provides time related sequences fit for keras

Downloaded from : https://pypi.org/project/keras-video-generators/

II. Emotion recognition from video : Example of sequences

II. Emotion recognition from video : Keras Video Generator

- Good results for happiness
- Average/Low results for anger, disgust and neutral
- Very bad results for fear and sadness

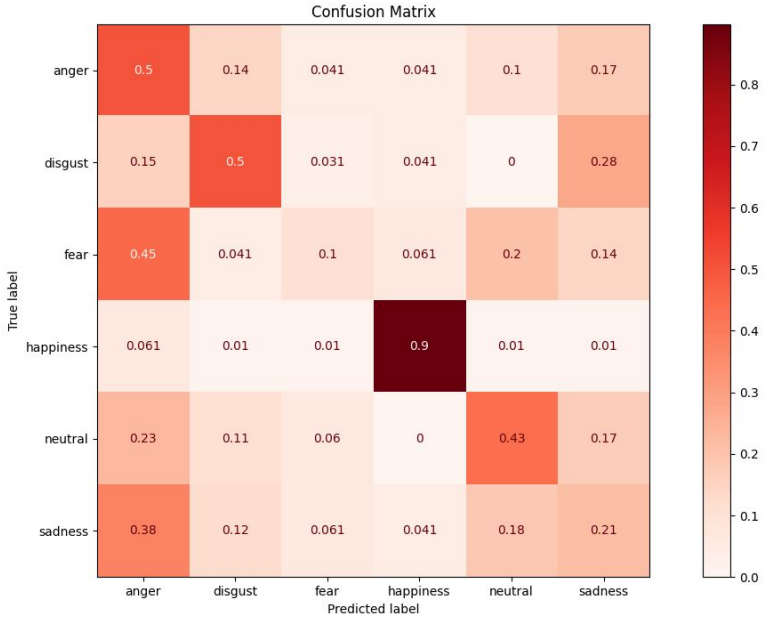
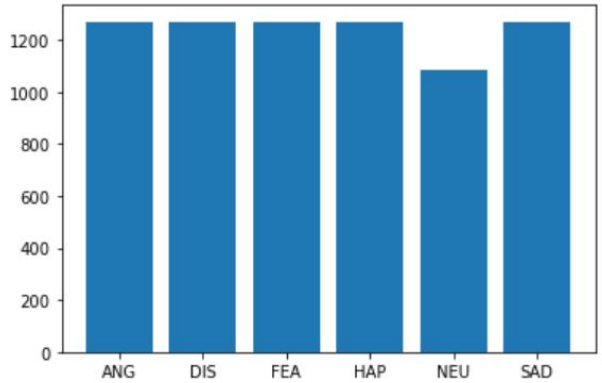


Table of contents

II. Emotion recognition from audio : Audio dataset

→ 7441 audio clips

→ Balanced Data



→https://github.com/GorillaBus/urban-audio-classifier

II. Emotion recognition from audio : The dataset

→ 12 sentences: “It’s eleven o’clock”, “Don’t forget your jacket”

→ 6 emotions

→ Different intensities: Low, Medium, High, Unspecified

→ 1001_DFA_ANG_XX 

→ 1001_IEO_HAP_LO 

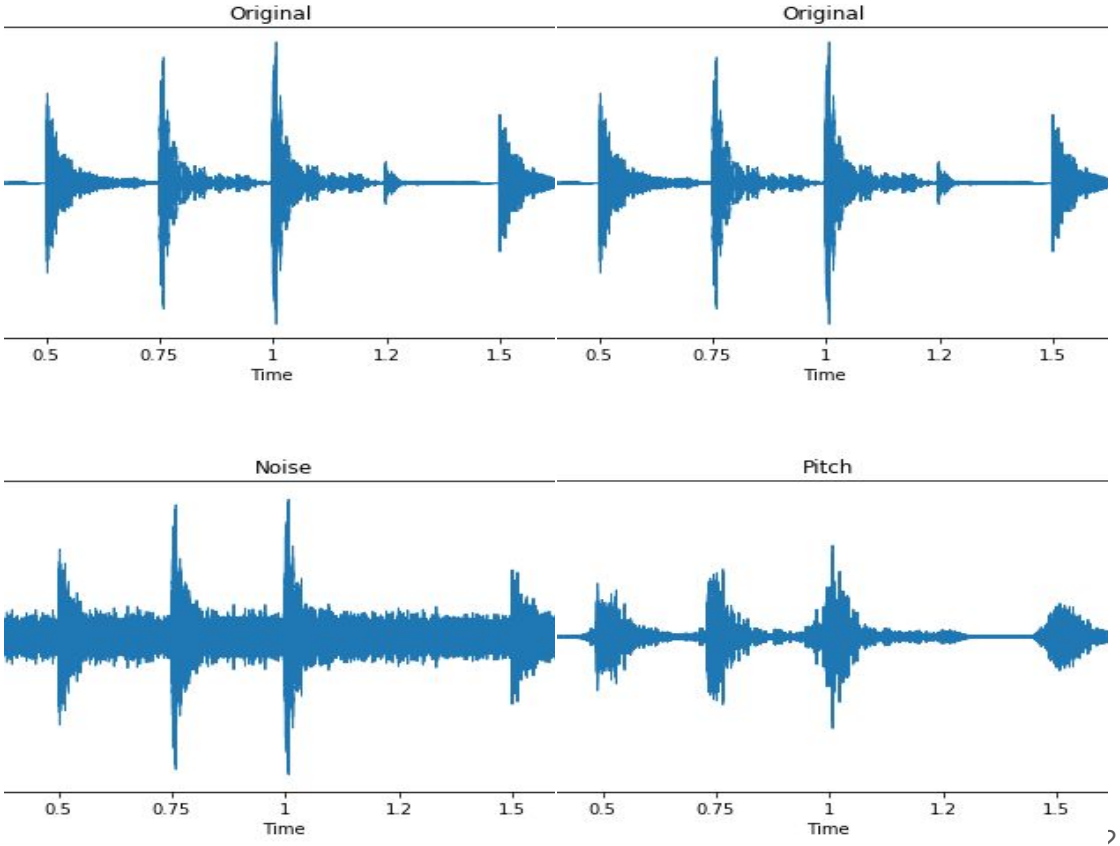
II. Emotion recognition from audio : Data Augmentation

→ Noise

→ Time Stretching

→ Pitch Shifting

→ 59533 files

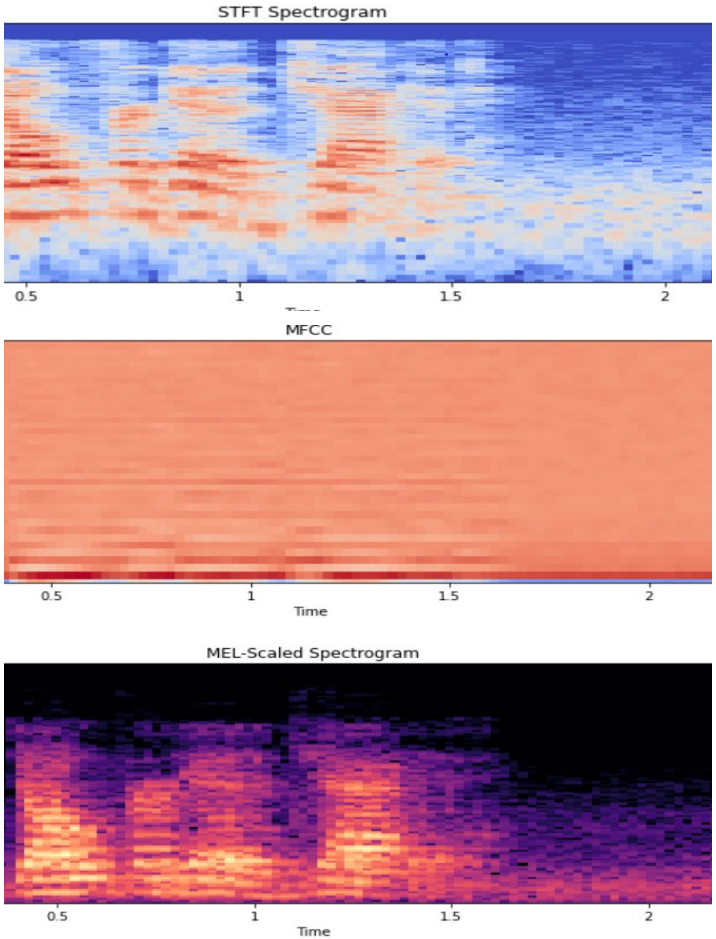


II. Emotion recognition from audio : Sound features

→ STFT

→ MFCC

→ Mel-Scaled Spectrogram



II. Emotion recognition from audio : Use of a CNN model

→ Use features as images

→ Zero- padding

→ Better performances than ANN

→ Simple architecture

```
Model: "sequential_1"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
conv2d_4 (Conv2D)            (None, 38, 214, 32)       320
_________________________________________________________________
leaky_re_lu_4 (LeakyReLU)    (None, 38, 214, 32)       0
_________________________________________________________________
batch_normalization_4 (Batch (None, 38, 214, 32)       128
_________________________________________________________________
spatial_dropout2d_3 (Spatial (None, 38, 214, 32)       0
_________________________________________________________________
conv2d_5 (Conv2D)            (None, 36, 212, 32)       9248
_________________________________________________________________
leaky_re_lu_5 (LeakyReLU)    (None, 36, 212, 32)       0
_________________________________________________________________
batch_normalization_5 (Batch (None, 36, 212, 32)       128
_________________________________________________________________
max_pooling2d_1 (MaxPooling2 (None, 18, 106, 32)       0
_________________________________________________________________
spatial_dropout2d_4 (Spatial (None, 18, 106, 32)       0
_________________________________________________________________
conv2d_6 (Conv2D)            (None, 16, 104, 64)       18496
_________________________________________________________________
leaky_re_lu_6 (LeakyReLU)    (None, 16, 104, 64)       0
_________________________________________________________________
batch_normalization_6 (Batch (None, 16, 104, 64)       256
_________________________________________________________________
spatial_dropout2d_5 (Spatial (None, 16, 104, 64)       0
_________________________________________________________________
conv2d_7 (Conv2D)            (None, 14, 102, 64)       36928
_________________________________________________________________
leaky_re_lu_7 (LeakyReLU)    (None, 14, 102, 64)       0
_________________________________________________________________
batch_normalization_7 (Batch (None, 14, 102, 64)       256
_________________________________________________________________
global_average_pooling2d_1 ( (None, 64)                0
_________________________________________________________________
dense_1 (Dense)              (None, 6)                 390
=================================================================
Total params: 66,150
Trainable params: 65,766
Non-trainable params: 384
```

II. Emotion recognition from audio : Results

Training completed in time: 2:18:54.924485

	LOSS	ACCURACY
Training:	1.1044	63.2622
Test:	1.1009	63.9785

	precision	recall	f1-score
Anger	0.58	0.74	0.65
Disgust	0.29	0.05	0.08
Fear	0.70	0.10	0.17
Happiness	0.63	0.14	0.22
Neutral	0.67	0.94	0.78
Sad	0.67	0.03	0.06

II. Emotion recognition from audio : Training curves

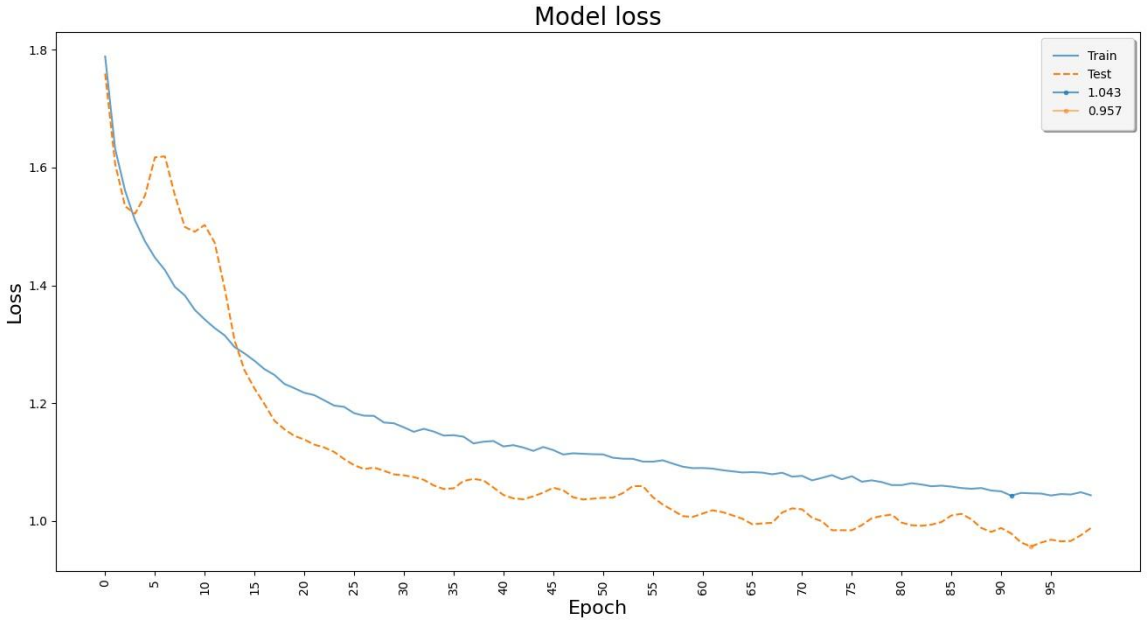


Table of contents

III. Results obtained and demo

Model	CNN	CNN after fine tuning	Time distributed model	Sound descriptors CNN
Accuracy	74%	69%	51%	64%

Good results for the CNN model and the sound descriptors CNN model

Time distributed model is not good

Watch out : CNN model may not be as accurate on videos as on images

Various success compared to already existing models

III. Results obtained and demo : Audio test

Sentence: ‘I like cats’

→ Anger  Anger

→ Happiness  Fear

→ Fear  Disgust

I’m really a good actor !

III. Results obtained and demo : Video CNN model

Batch of 10 images to have stable predictions

We keep the most represented emotion on each batch

III. Results obtained and demo : Video CNN model

Predicted emotion : Happiness

GT : Happiness

Predicted emotion : Neutral

GT : Fear

III. Results obtained and demo : Video CNN model

Predicted emotion : Happiness

GT : Disgust

Predicted emotion : Anger

GT : Disgust

Table of contents

IV. Difficulties encountered

- Hard to work with video : handle sequences of frames instead of images, hard to use generator from someone else

- Hard to work with audio : Use of descriptors we are less used to in the shape of images

- Problem of batch size for prediction, problems with corrupted frames, too slow to do live stream

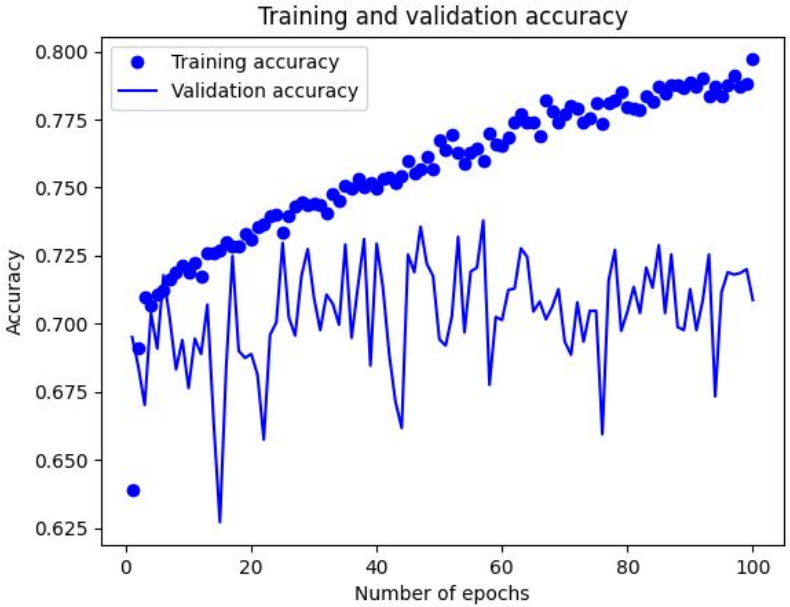
IV. Difficulties encountered : how to improve ?

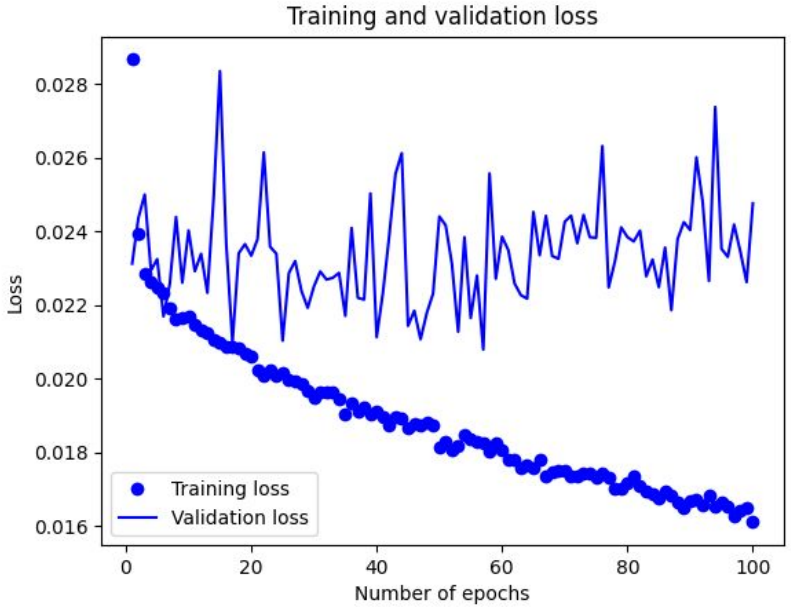
- Change CNN architecture (ResNet)
- Do more DA for FER (small rotation) and for videos of CREMA-D
- Create own video generator for keras
- Use transformer ?

Thank you for your attention

Special thanks to Ruxandra for supervising our project

Annexe : Fine Tuning





Annexe : Time Distributed Model