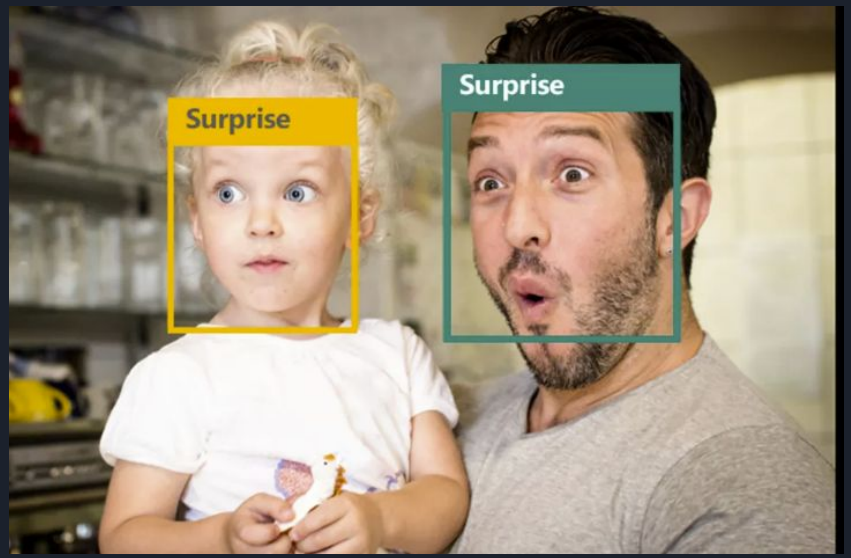
Audio-Visual-Emotion-and-Sentiment-Research

Members:
Audio parts: Enis Berk Çoban and Yunhua Zhao
Video parts: Patrick Jean-Baptiste

Goal

Use DL or ML methods to detect the emotion from people's speech, song or facial emotions.



Dataset:

The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)

Overview of the dataset:

24 People: 12 male, 12 female:

- Speech 60 trials per actor x 24 actors = 1440
- Song 44 trials per actor x 23 actors = 1012

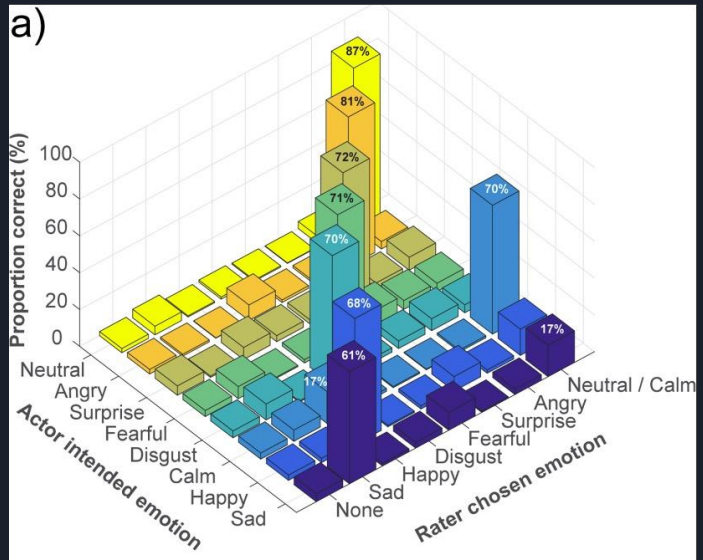
https://zenodo.org/record/1188976#.XrRhU6hKg2y

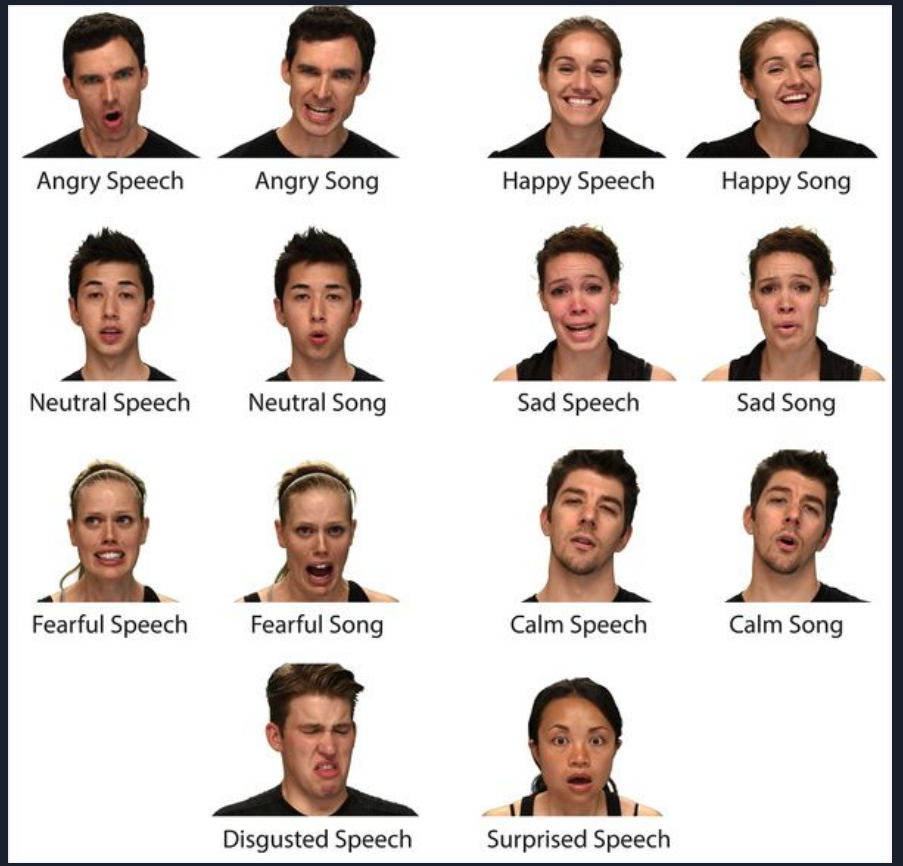
Dataset and Prediction Baseline

Baseline is 12.5%, 8 categories

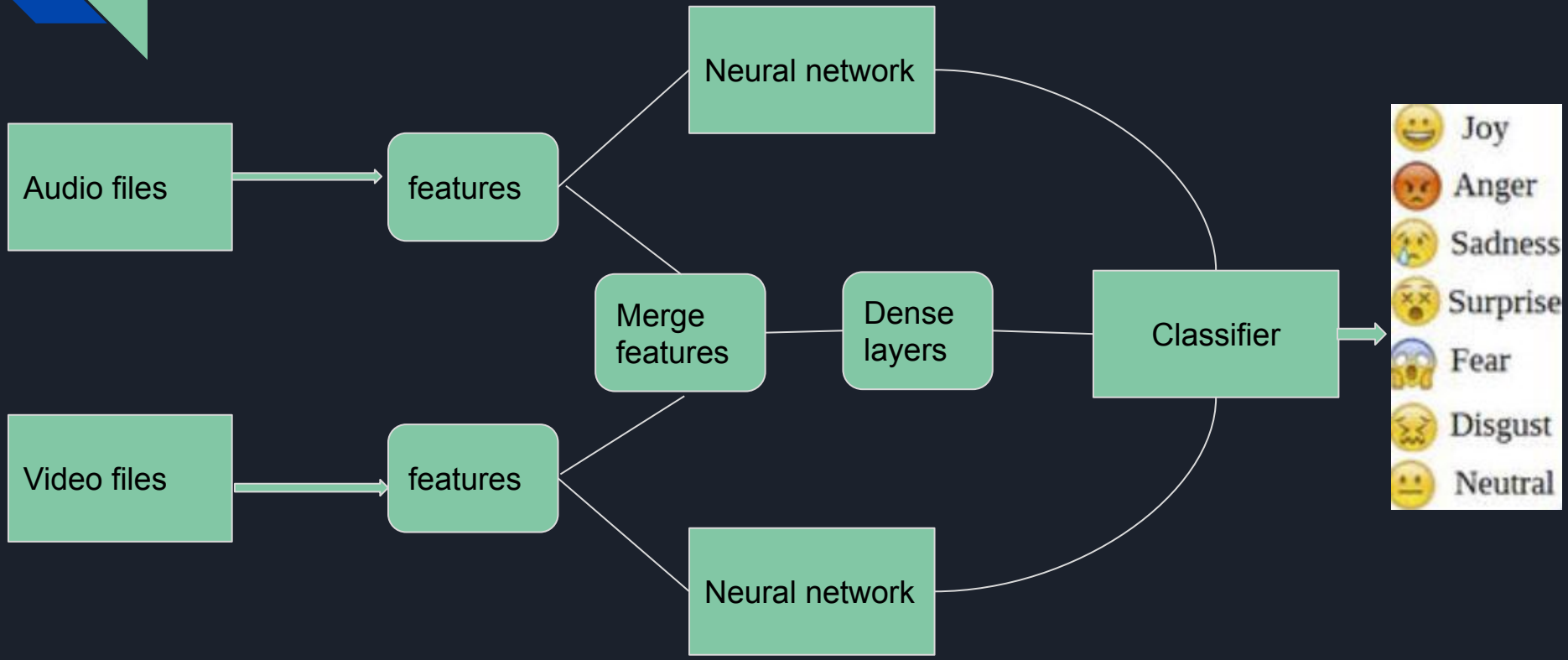
Human Prediction Scores:

- 80% for audio-video
- 75% for video-only
- 60% for audio-only





Emotion classification

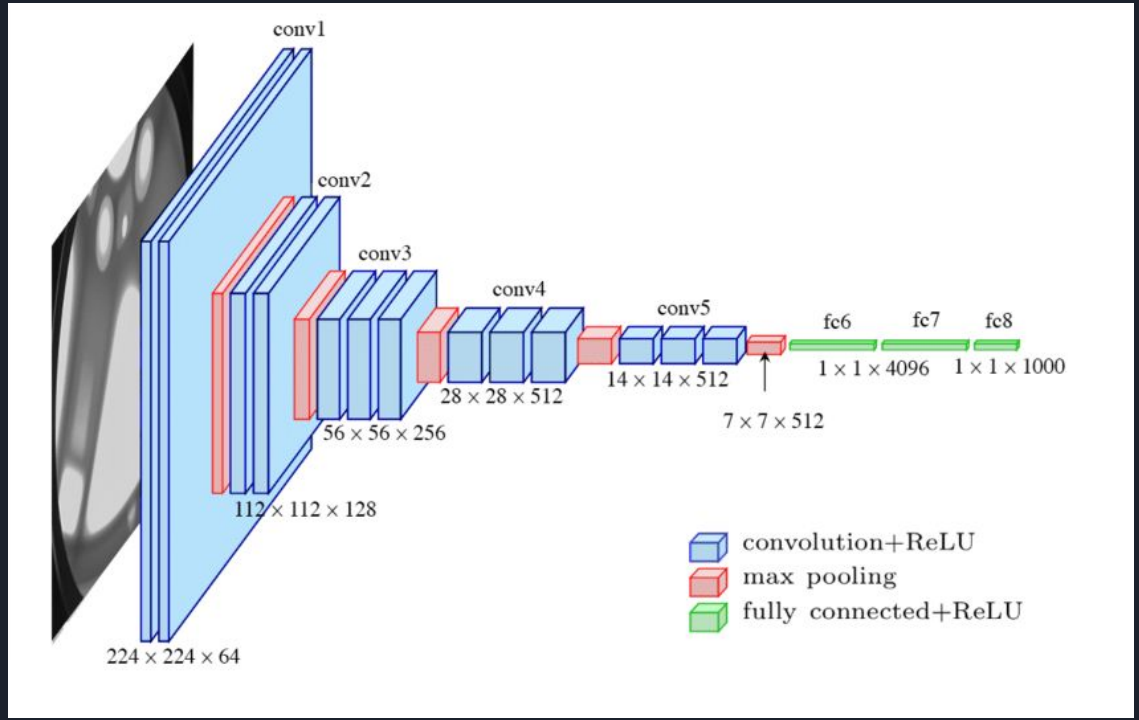


Audio part

1. **Models:**
   - VGGish
   - LSTM
2. **Train data set**
   - Song only
   - Speech only
   - Song + Speech
3. **Data split**

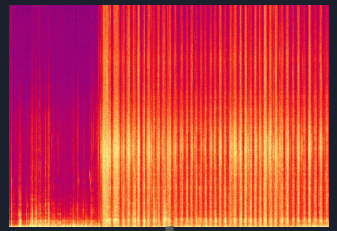
   Train : val : tese = 0.6 : 0.2 : 0.2

VGGish model layers



Vggish Model Output

Embeddings shared as dropbox link at assets/readMe.md



Log-mel Spectrogram

VGGish

Raw

0.002
0.123
⋮
0.424

[n,128]

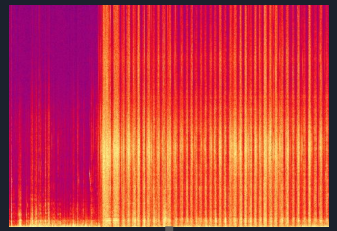
PCA and whitening

Normalized

0.2
0.1
⋮
0.34

VGGish model

Log-mel
Spectrogram



VGGish

Traditional
Classifiers

Exp 1

0
1
. . .
7

8 Emotion Categories

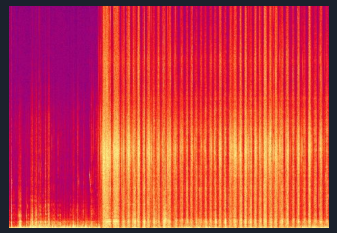
MLP

Exp 2

0
1
. . .
7

8 Emotion Categories

LSTM model structure



features

LSTM

Dense
‘relu’

Dense
‘relu’

Dense
‘tanh’

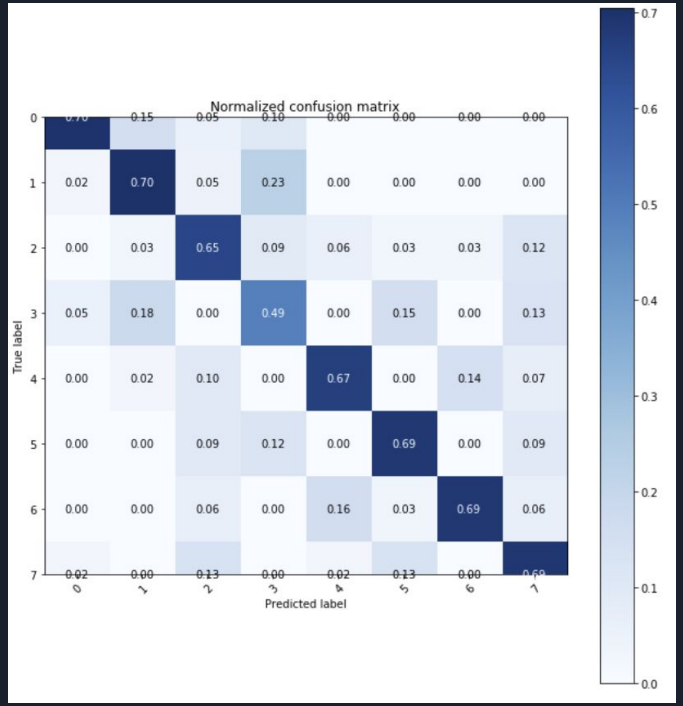
softmax

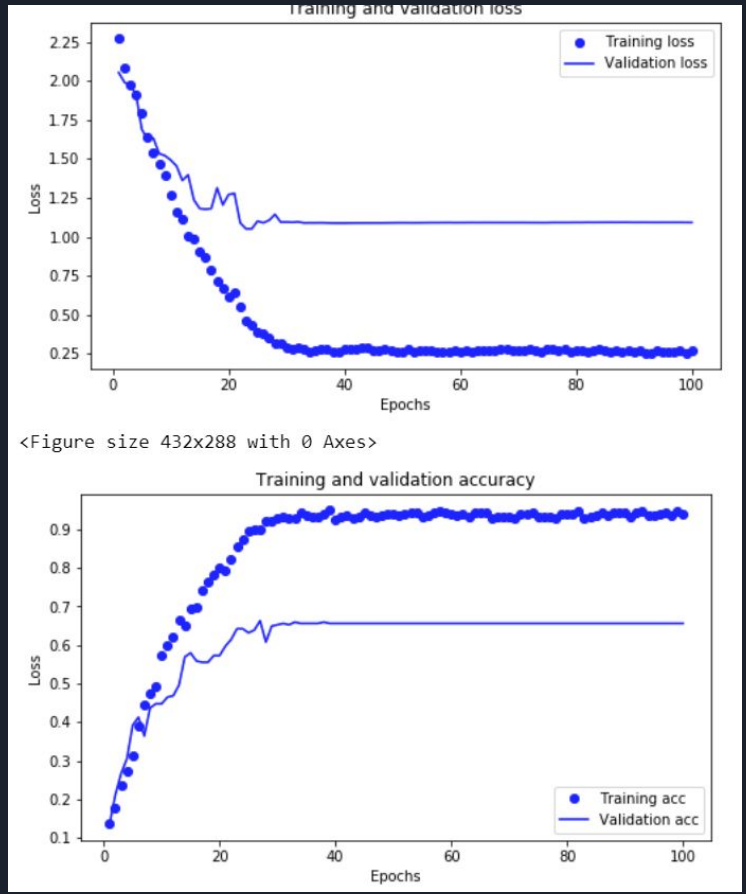
emotions

Audio Part Accuracy Results

models	Song	Speech	Song+Speech
VGGish-Deep	49%	%48	62%
VGGish-ML	47%	%51	54%
LSTM	67%	66%	75%

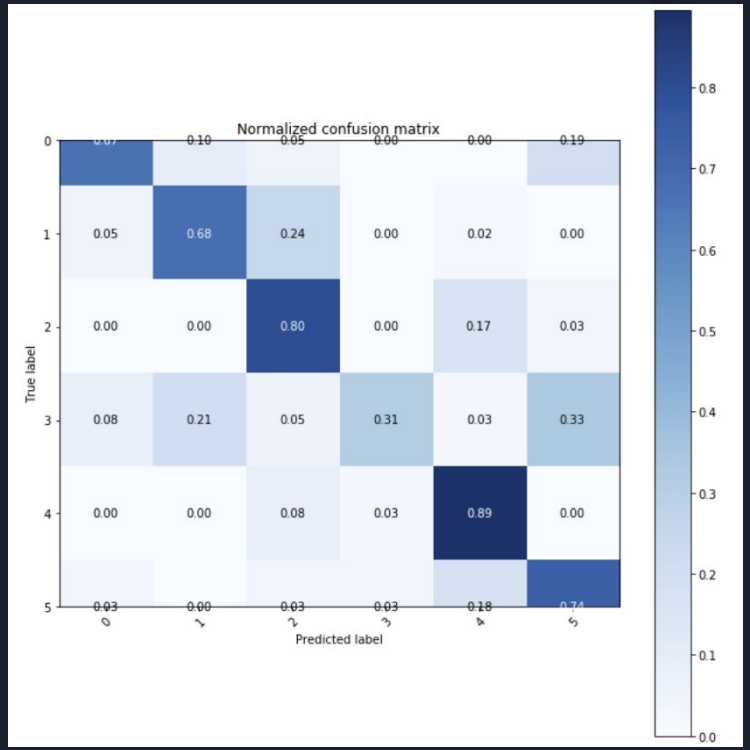
Result Visualision(Speech)

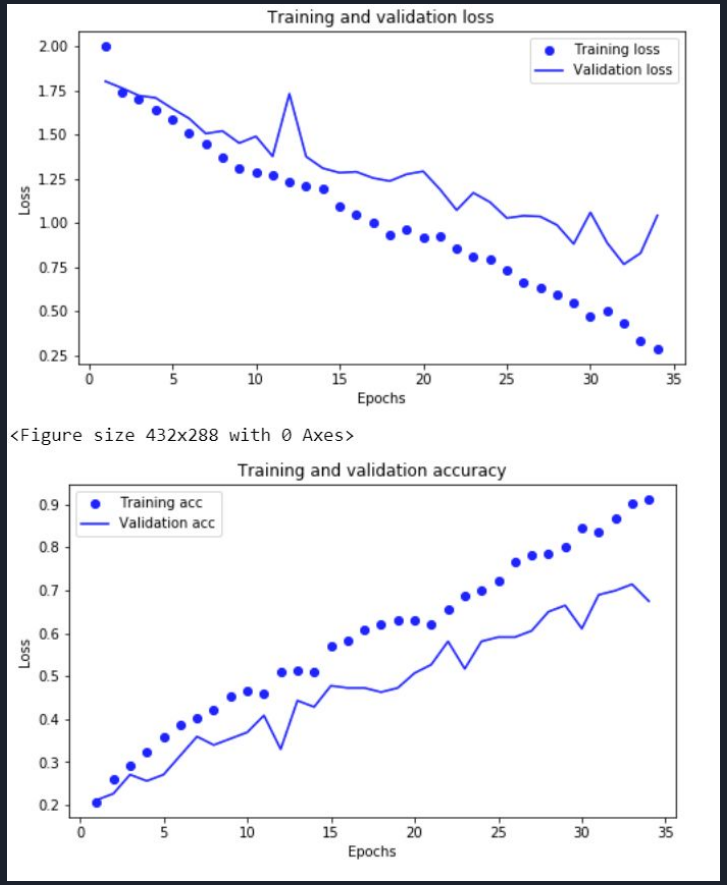




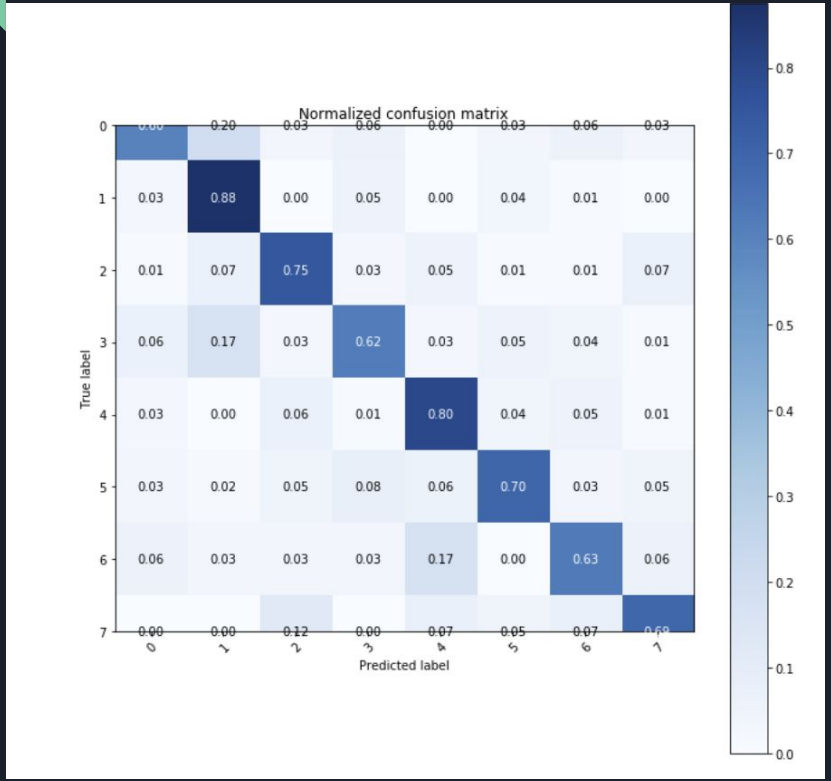
<Figure size 432x288 with 0 Axes>

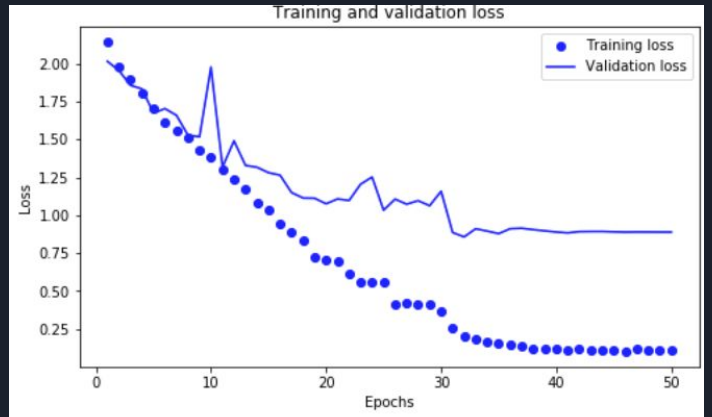
4. Result Visualision(Song)



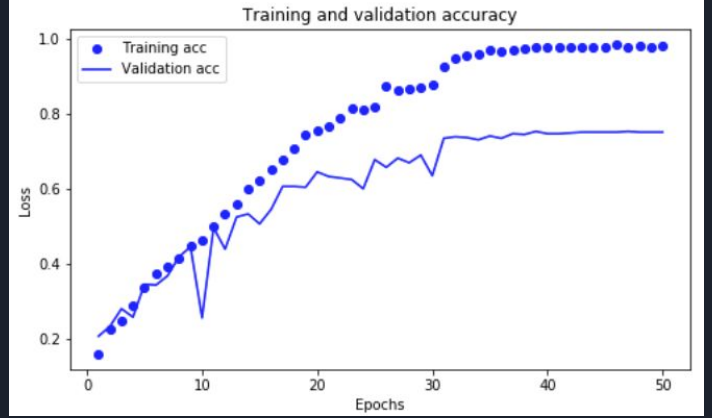


4. Result Visualision(Speech+Song)





<Figure size 432x288 with 0 Axes>



Visual Part

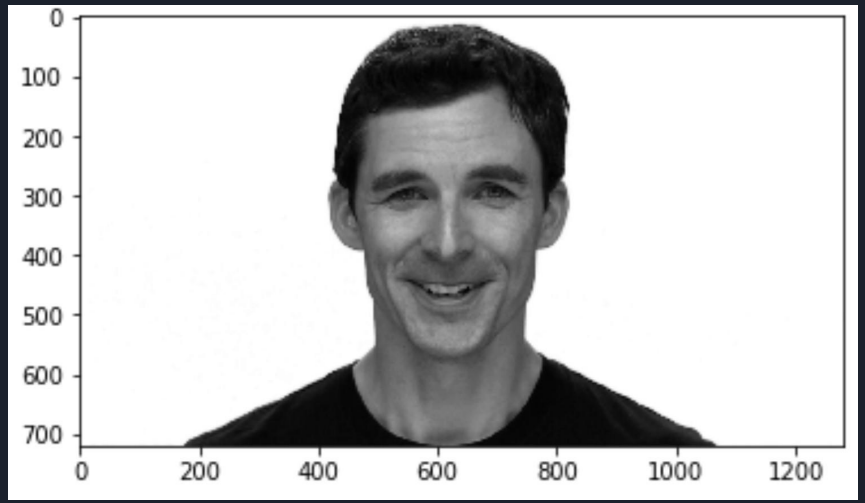
The objective was to create a model for recognizing emotions from images.

Process:

1. Initial Video Preprocessing
2. Face Detection
3. Face Extraction
4. Visual Model
5. Results
6. Evaluations

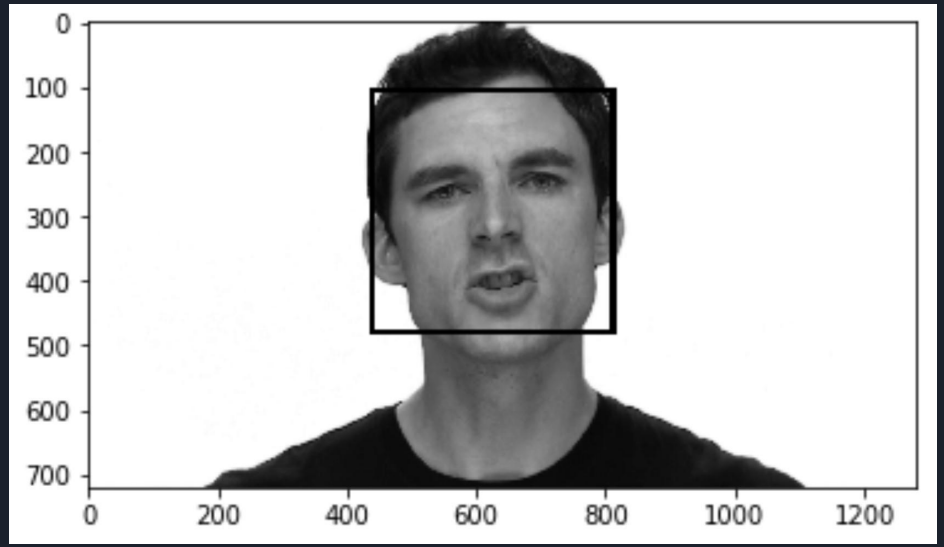
1. Initial Video Preprocessing

Extract video frames and convert them to grayscale.



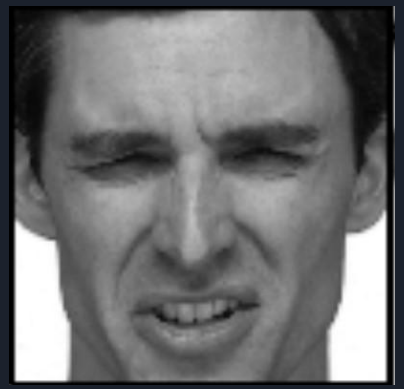
2. Face Detection

Detect the actor’s face.



3. Face Extraction

Extract the actor’s face from the image.



4. Visual Model - Architecture

```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
conv2d (Conv2D)              (None, 126, 126, 64)      1792
_________________________________________________________________
max_pooling2d (MaxPooling2D) (None, 125, 125, 64)      0
_________________________________________________________________
conv2d_1 (Conv2D)            (None, 123, 123, 32)      18464
_________________________________________________________________
max_pooling2d_1 (MaxPooling2 (None, 122, 122, 32)      0
_________________________________________________________________
conv2d_2 (Conv2D)            (None, 120, 120, 16)      4624
_________________________________________________________________
max_pooling2d_2 (MaxPooling2 (None, 119, 119, 16)      0
_________________________________________________________________
flatten (Flatten)            (None, 226576)            0
_________________________________________________________________
dense (Dense)                (None, 8)                 1812616
=================================================================
Total params: 1,837,496
Trainable params: 1,837,496
Non-trainable params: 0
_________________________________________________________________
```

4. Visual Model - Training Parameters

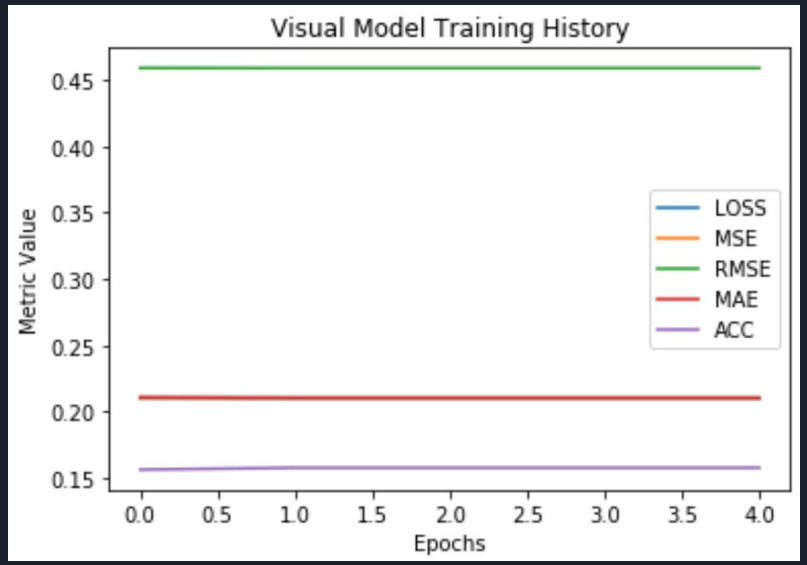
Parameters:

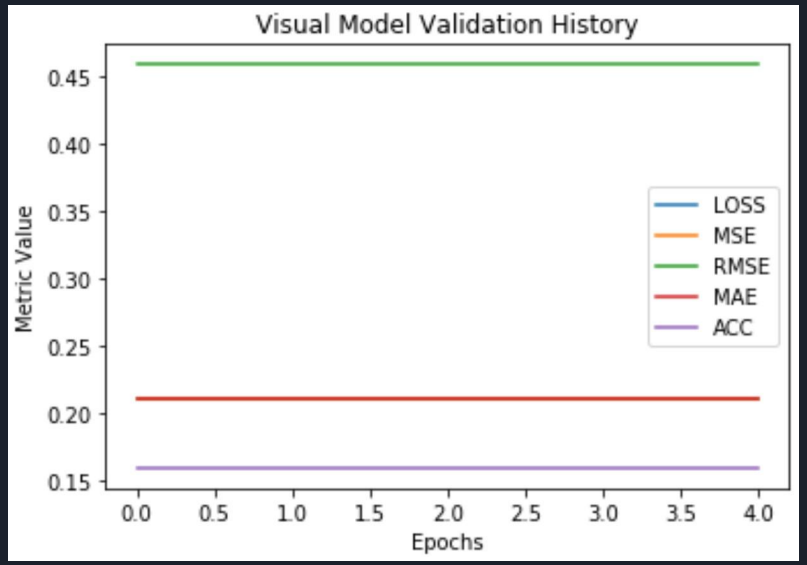
- Loss Function - Mean Squared Error
- Optimizer – SGD
- Learning Rate - 0.01
- Epochs – 5
- Batch Size – 150

- Training Data – 2824 images
- Validation Data – 942 images

4. Visual Model - Training Issues

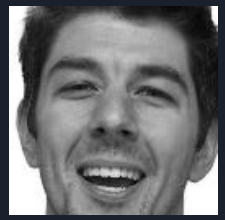
Accuracy is very low and does not improve.





5. Results - Misclassifications

Label: Happy



Predicted: Surprised

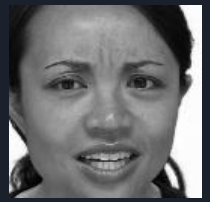
Label: Neutral

Predicted: Sad

Label: Angry

Predicted: Surprised

Label: Surprised



Predicted: Fearful

Label: Calm

Predicted: Happy

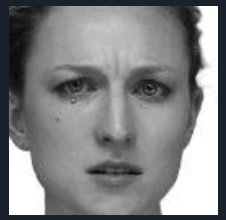
Label: Fearful

Predicted: Sad

Label: Disgust

Predicted: Angry

Label: Sad



Predicted: Neutral

6. Evaluations

Test Data - 948 images

Emotion recognition accuracy was 16.5%.

Used different models to attempt to improve accuracy:

- Different numbers of convolutional layers.
- Different number of nodes.
- VGG16 pre-trained on ImageNet.

Accuracy did not improve for training or testing.

Similarity between distinct emotion classes could have also led to misclassifications.

Dataset Split

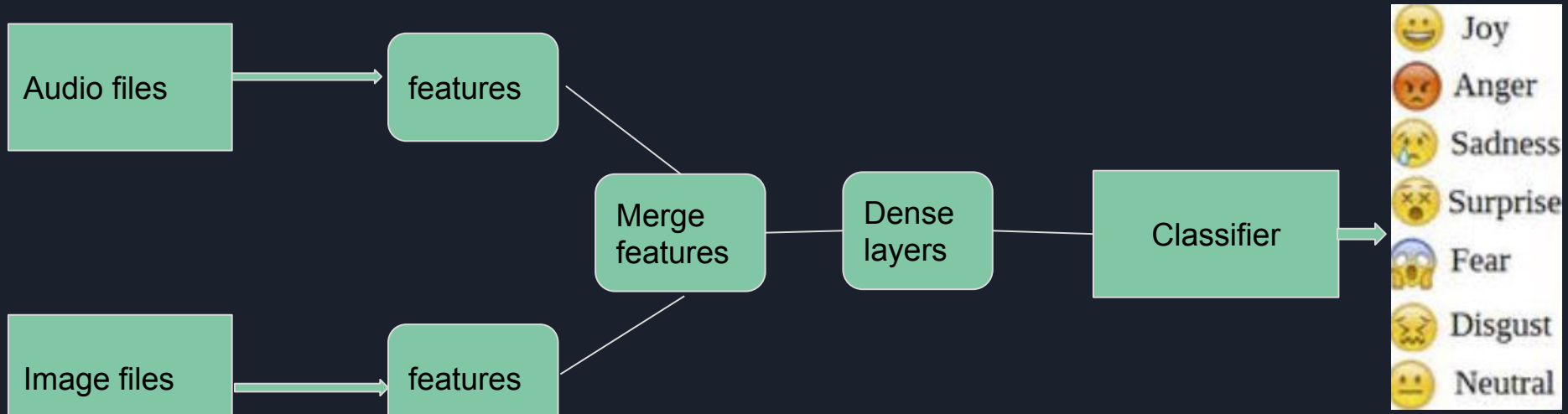
- We wanted to use same dataset distribution in all experiments to make sure all experiments are compatible, and we can merge our models later.

- We created a csv file with Train/Dev/Test set of filenames and shared it with each other.

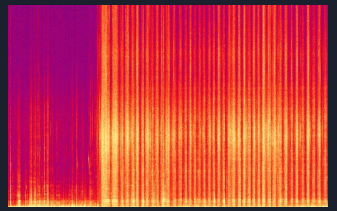
Merging Audio and Images

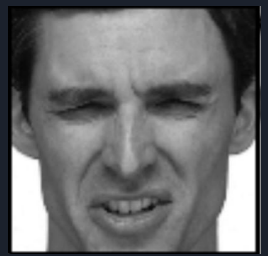
- Combining inputs for single model
- Training two models together

Combining inputs for single model



Output of Combining inputs model

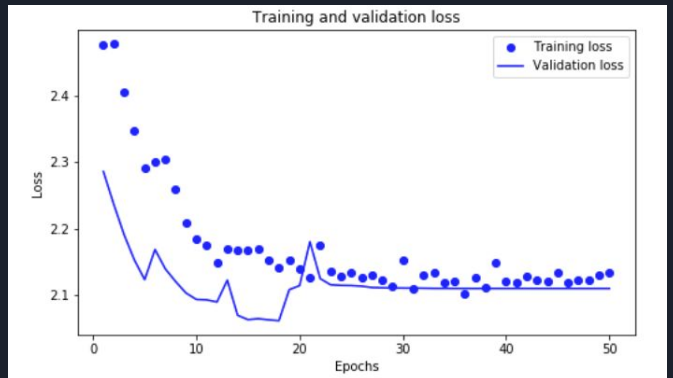




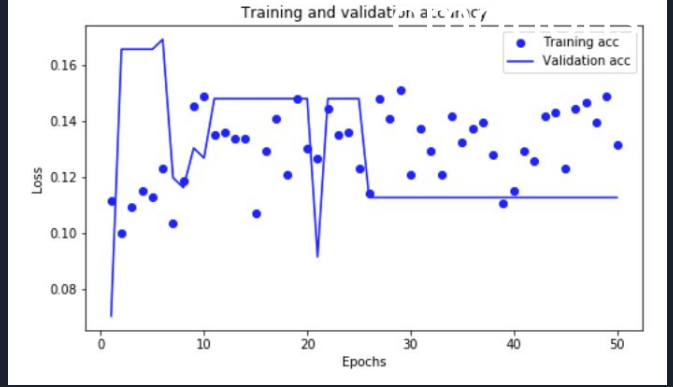
Concat

MLP (4 hidden layers)

softmax

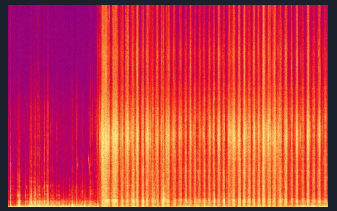


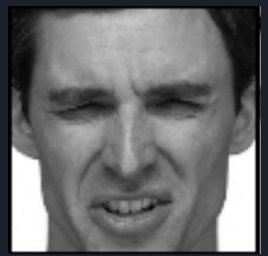
<Figure size 432x288 with 0 Axes>



Training two models together

*Batchnorm ,dropout, max pooling layers excluded





LSTM

Dense

Dense

Dense

Conv

Conv

Conv

Dense

Dense

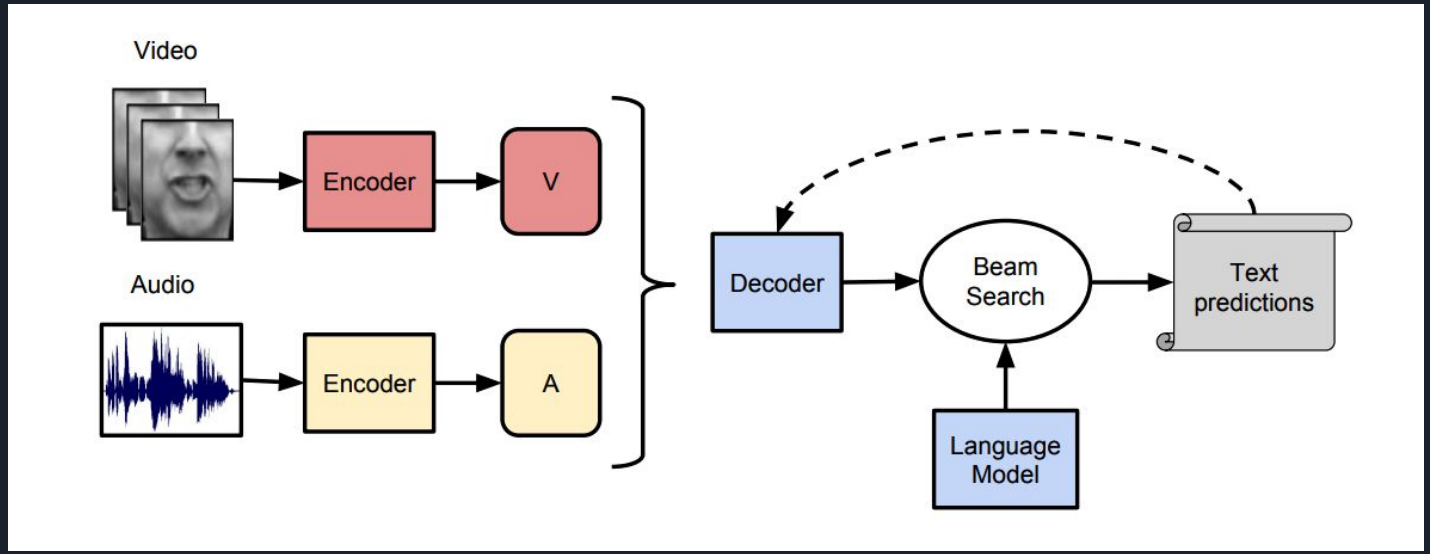
softmax

emotions

0
1
.
.
.
7

Tianyu

Deep Audio-Visual Speech Recognition




Angry Speech

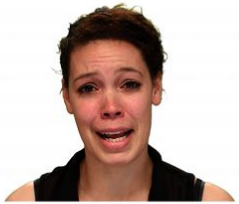

Angry Song

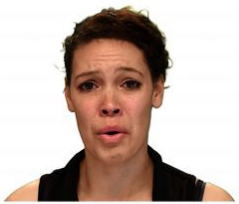

Happy Speech


Happy Song

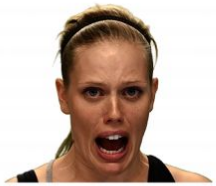

Neutral Speech


Neutral Song


Sad Speech


Sad Song


Fearful Speech


Fearful Song


Calm Speech


Calm Song


Disgusted Speech


Surprised Speech

Conclusions

Audio part:

- Google Audioset representations are not related to this task
- More data is better, Classic ML models as good as DL when data is small

Video part:

- A model trained in very few epochs using a small number of images for each class was a reason for the poor visual emotion recognition performance..

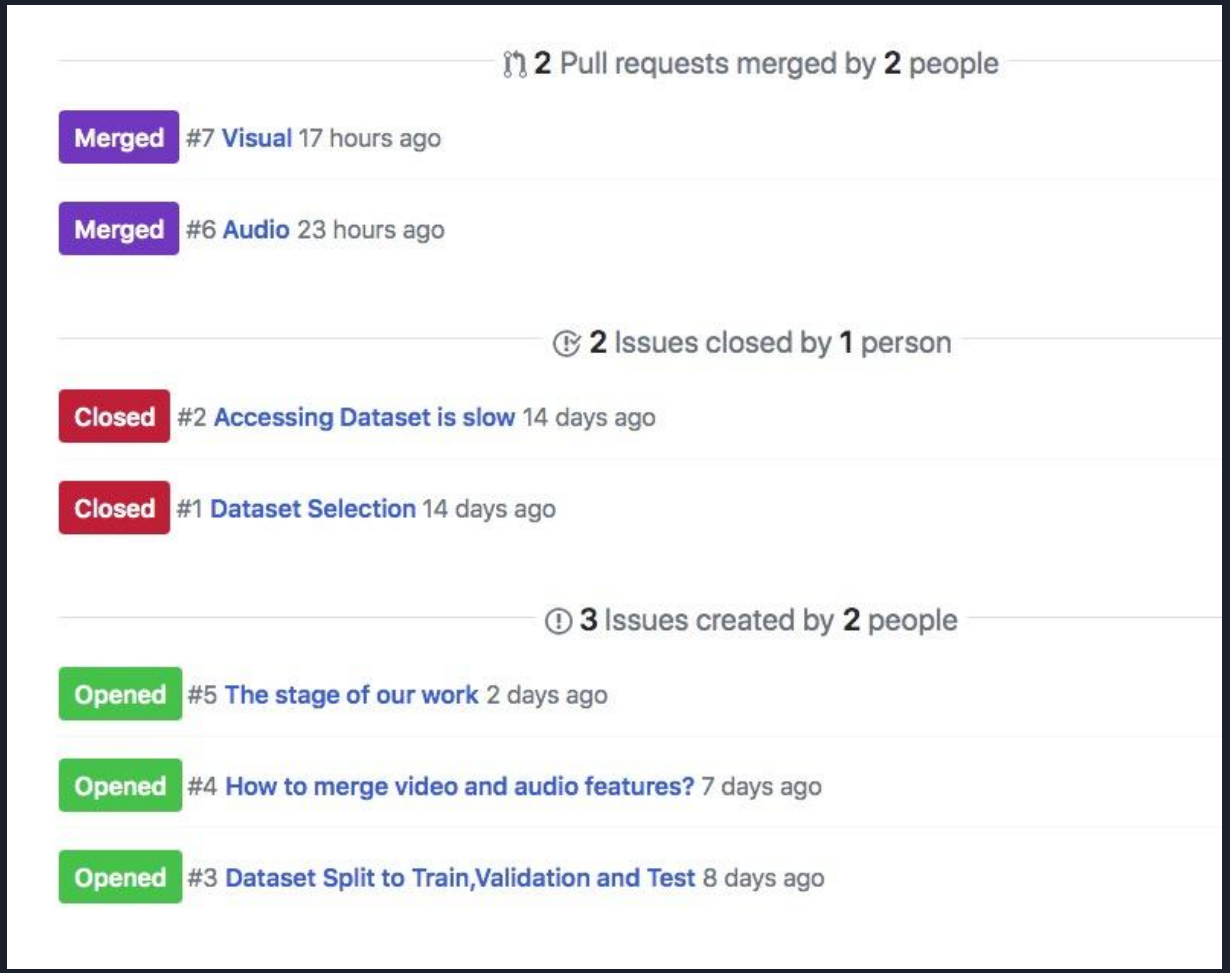
Merge audio and images part:

- Images caused noise, preventing model to learn

Github


Yunhua468
#1
47 commits 17,243 ++ 723 --
20
10
4/21/2020
4/27/2020
5/3/2020
5/9/2020
EnisBerk
#2
17 commits 5,662 ++ 154 --
20
10
4/21/2020
4/27/2020
5/3/2020
5/9/2020
patrick-jean-baptiste
#3
13 commits 1,951 ++ 32 --
20
10
4/21/2020
4/27/2020
5/3/2020
5/9/2020

Github


2 Pull requests merged by 2 people
Merged #7 Visual 17 hours ago
Merged #6 Audio 23 hours ago
2 Issues closed by 1 person
Closed #2 Accessing Dataset is slow 14 days ago
Closed #1 Dataset Selection 14 days ago
3 Issues created by 2 people
Opened #5 The stage of our work 2 days ago
Opened #4 How to merge video and audio features? 7 days ago
Opened #3 Dataset Split to Train,Validation and Test 8 days ago

References

Dataset Paper:

- Livingstone, S. R., & Russo, F. A. (2018). The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. *PloS one*, *13*(5).

Dataset Link:

- https://zenodo.org/record/1188976#.Xr2Cum5Fw2y

Thank you!