Emotion Recognition in Audio & Video using Deep Neural Networks

MANDEEP SINGH & YUAN FANG

HTTPS://TINYURL.COM/Y8795RBT

CS231N Final Project
Stanford University
06/09/2020

Contents

- Problem Statement & Application
- Dataset & Data pre-processing
- Model architecture
- Results
- Future work

Problem Statement
&
Application

Stanford University

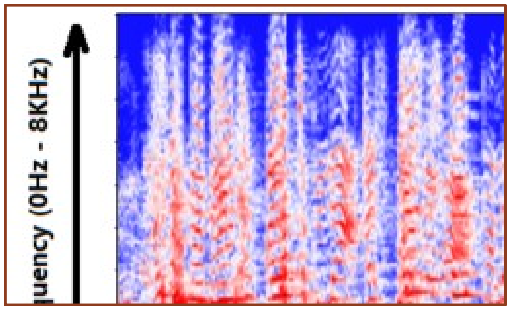
Problem Statement & Application

- Given an audio/video:
  - Classify it into one of the four emotions, i.e. happy, anger, sad, neutral
- Emotion detection in audio is key area wherein it can assist:
  - Siri/Alexa to give good recommendations after detecting the emotion.
  - 911 operator based on interpreting emotions in different languages.

Dataset
&
Data pre-processing

Dataset & Data pre-processing

- IEMOCAP[1] dataset from USC.
  - 12 hours audiovisual data of 5 females, 5 males speaking in 9 emotions.
  - Each utterance has an emotion label.
- Data Pre-processing
  - Audio:
    - Extract 3 second audio waveform and convert it into spectrogram of size 200x300.



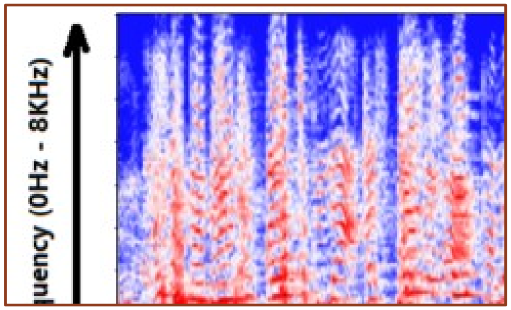
  - Video:
    - Extract 20 frames of size 60x100 from the video corresponding to 3 second audio.

Footnote:

[1] C. L. A. K. E. M. S. K. J. C. S. L. C. Busso, M. Bulut and S. Narayanan. Iemocap: Interactive emotional dyadic motion capture database, December 2008.

Dataset & Data pre-processing

- IEMOCAP[1] dataset from USC.
  - 12 hours audiovisual data of 5 females, 5 males speaking in 9 emotions.
  - Each utterance has an emotion label.
- Data Pre-processing
  - Audio:
    - Extract 3 second audio waveform and convert it into spectrogram of size 200x300.



  - Video:
    - Extract 20 frames of size 60x100 from the video corresponding to 3 second audio.

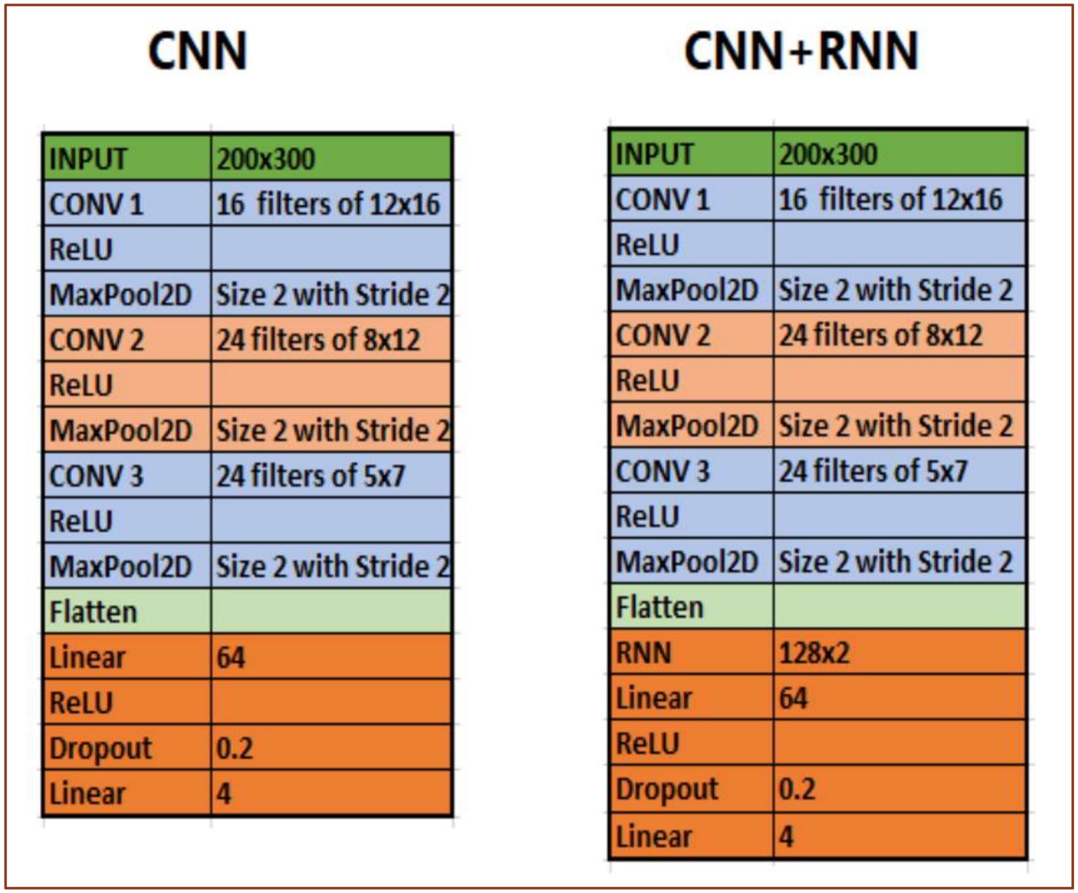
Footnote:

[1] C. L. A. K. E. M. S. K. J. C. S. L. C. Busso, M. Bulut and S. Narayanan. Iemocap: Interactive emotional dyadic motion capture database, December 2008.

Model Architecture

Model Architecture

- Audio Models:
  - Explored CNN, CNN+RNN & CNN+LSTM model architectures.



CNN

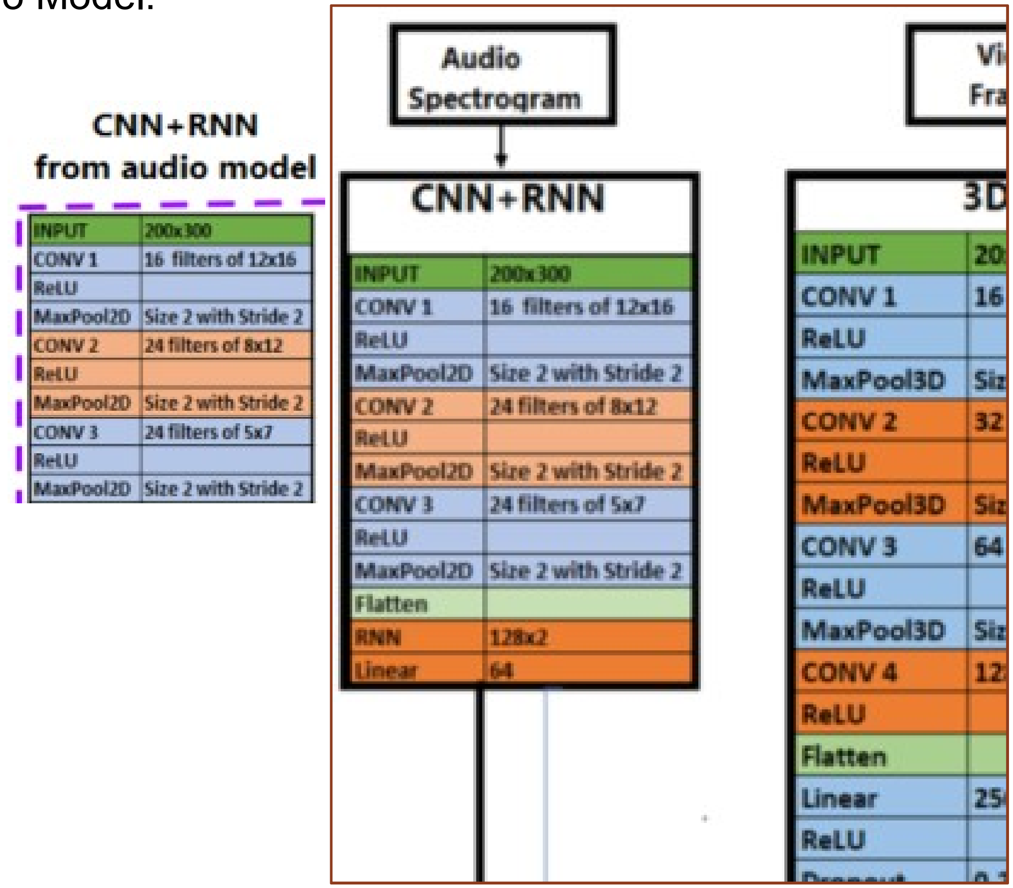
INPUT	200x300
CONV 1	16 filters of 12x16
ReLU	
MaxPool2D	Size 2 with Stride 2
CONV 2	24 filters of 8x12
ReLU	
MaxPool2D	Size 2 with Stride 2
CONV 3	24 filters of 5x7
ReLU	
MaxPool2D	Size 2 with Stride 2
Flatten	
Linear	64
ReLU	
Dropout	0.2
Linear	4

CNN+RNN

INPUT	200x300
CONV 1	16 filters of 12x16
ReLU	
MaxPool2D	Size 2 with Stride 2
CONV 2	24 filters of 8x12
ReLU	
MaxPool2D	Size 2 with Stride 2
CONV 3	24 filters of 5x7
ReLU	
MaxPool2D	Size 2 with Stride 2
Flatten	
RNN	128x2
Linear	64
ReLU	
Dropout	0.2
Linear	4

Model Architecture

- Audio + Video Model:



CNN+RNN
from audio model

INPUT	200x300
CONV 1	16 filters of 12x16
ReLU	
MaxPool2D	Size 2 with Stride 2
CONV 2	24 filters of 8x12
ReLU	
MaxPool2D	Size 2 with Stride 2
CONV 3	24 filters of 5x7
ReLU	
MaxPool2D	Size 2 with Stride 2

CNN+RNN

INPUT	200x300
CONV 1	16 filters of 12x16
ReLU	
MaxPool2D	Size 2 with Stride 2
CONV 2	24 filters of 8x12
ReLU	
MaxPool2D	Size 2 with Stride 2
CONV 3	24 filters of 5x7
ReLU	
MaxPool2D	Size 2 with Stride 2
Flatten	
RNN	128x2
Linear	64

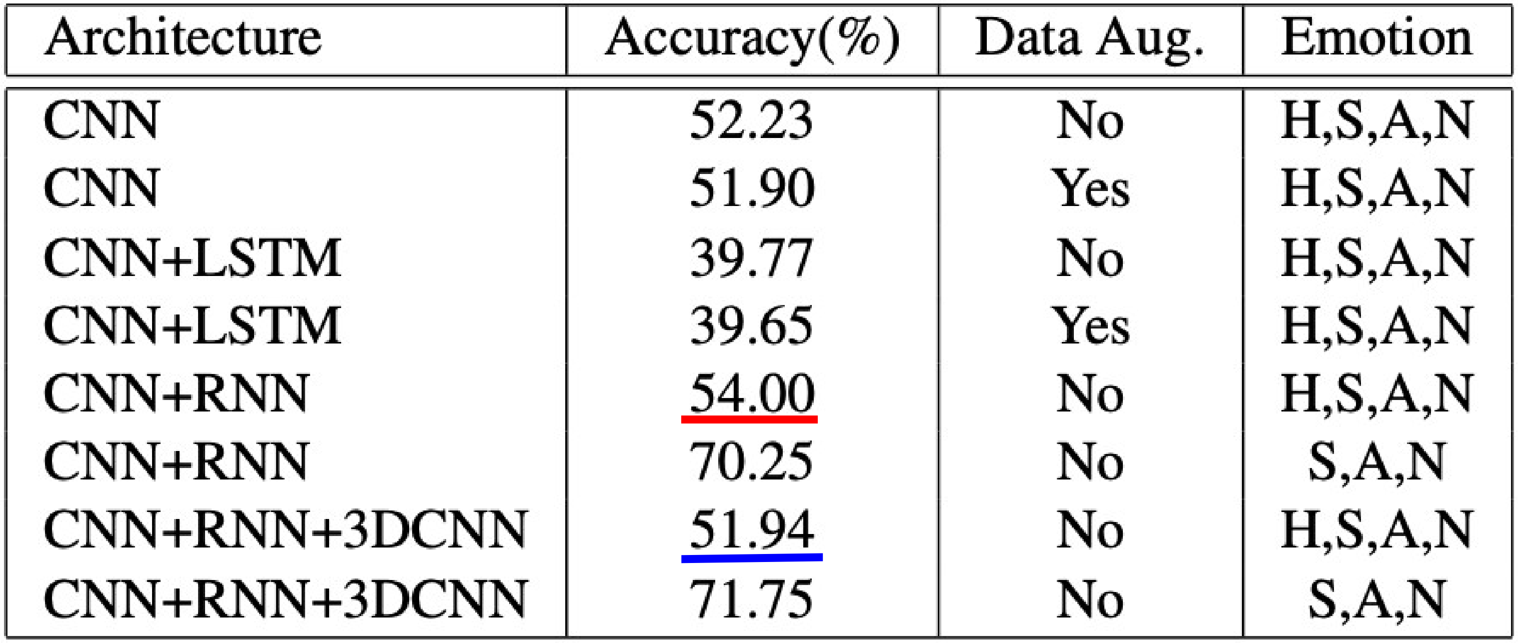
3D

INPUT	20
CONV 1	16
ReLU	
MaxPool3D	Siz
CONV 2	32
ReLU	
MaxPool3D	Siz
CONV 3	64
ReLU	
MaxPool3D	Siz
CONV 4	12
ReLU	
Flatten	
Linear	25
ReLU	
Dropout	0.

Results

Results

Accuracy Table



Architecture	Accuracy(%)	Data Aug.	Emotion
CNN	52.23	No	H,S,A,N
CNN	51.90	Yes	H,S,A,N
CNN+LSTM	39.77	No	H,S,A,N
CNN+LSTM	39.65	Yes	H,S,A,N
CNN+RNN	54.00	No	H,S,A,N
CNN+RNN	70.25	No	S,A,N
CNN+RNN+3DCNN	51.94	No	H,S,A,N
CNN+RNN+3DCNN	71.75	No	S,A,N

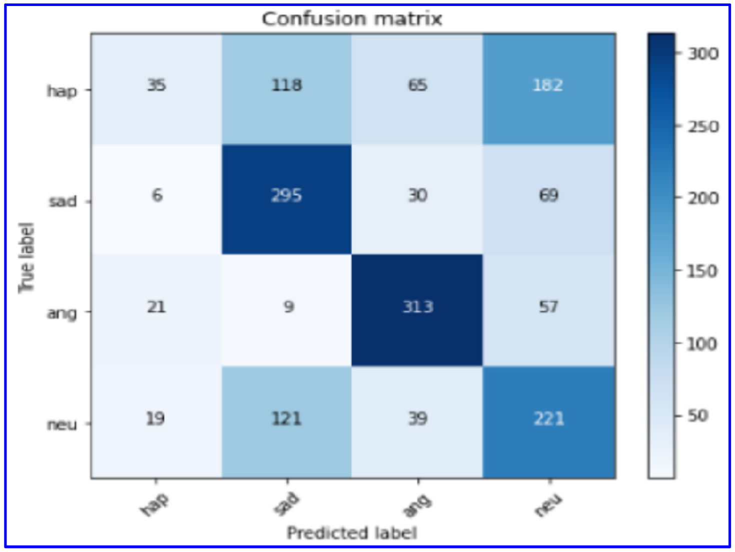
Best Accuracy (4 emotions):

Audio Model: 54%

Audio+Video Model: 51.94%

Results

Confusion Matrix



	hap	sad	ang	neu
hap	35	118	65	182
sad	6	295	30	69
ang	21	9	313	57
neu	19	121	39	221

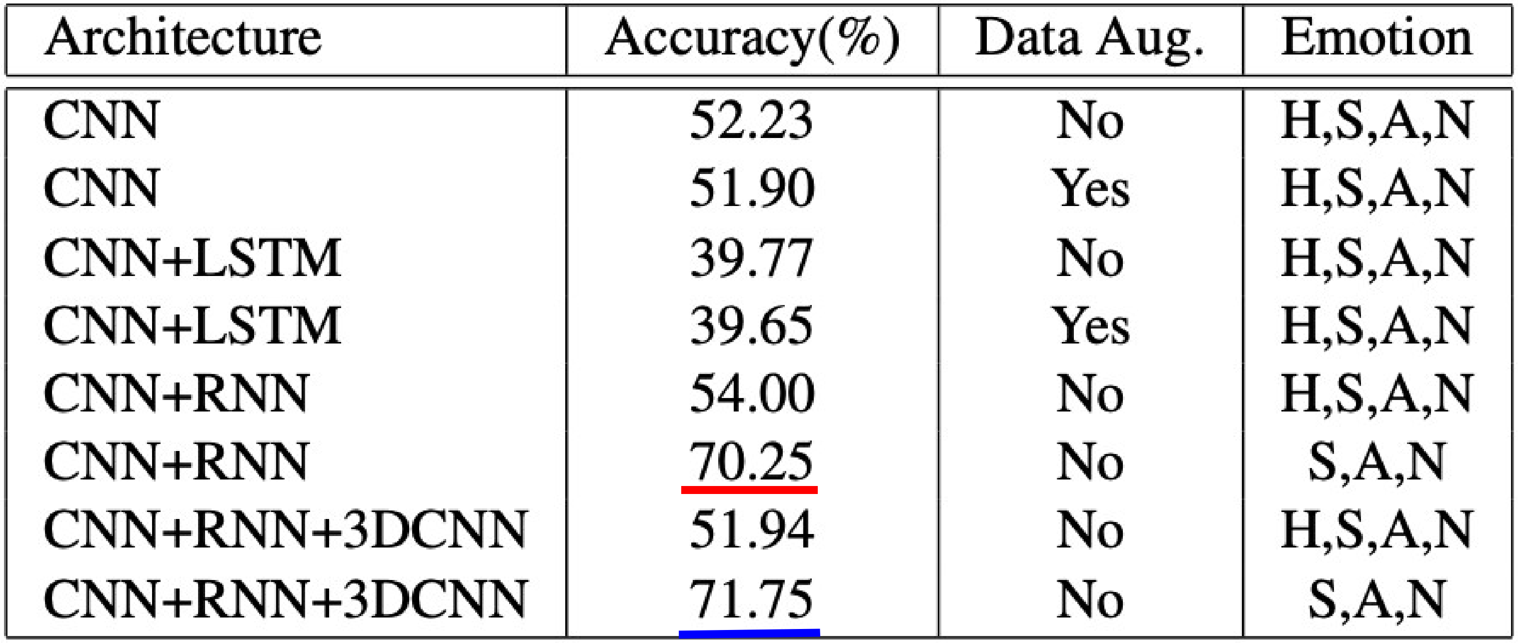
Audio: CNN+RNN
4 Emotions

Analysis:

Unbalanced Dataset: Low count of happiness

Results

Accuracy Table



Architecture	Accuracy(%)	Data Aug.	Emotion
CNN	52.23	No	H,S,A,N
CNN	51.90	Yes	H,S,A,N
CNN+LSTM	39.77	No	H,S,A,N
CNN+LSTM	39.65	Yes	H,S,A,N
CNN+RNN	54.00	No	H,S,A,N
CNN+RNN	70.25	No	S,A,N
CNN+RNN+3DCNN	51.94	No	H,S,A,N
CNN+RNN+3DCNN	71.75	No	S,A,N

- Accuracy on (3 emotions) jumps from 70.25% to 71.75%
  Implying Audio+Video model works.

Conclusion & Future work

Conclusion & Future Work

Conclusion:

- Explored different deep neural network architectures to predict emotion:
  - CNN, CNN+RNN, CNN+LSTM, CNN+RNN+3DCNN
- Best performing models:
  - Audio model: CNN+RNN with accuracy of 54%.
  - Video model: CNN+RNN+3DCNN with accuracy of 51.94%.
- Analysis:
  - Low count of happy emotion in the dataset.
  - Audio+video model works with training on 3 emotions resulting in accuracy jump from 70.25% to 71.75%.

Future Work:

- Increase input & output dimensions in each layer in the network.
- Auto crop to focus on the face of the actor in video frames.
- Explore noise removal algorithms.

Conclusion & Future Work

Conclusion:

- Explored different deep neural network architectures to predict emotion:
  - CNN, CNN+RNN, CNN+LSTM, CNN+RNN+3DCNN
- Best performing models:
  - Audio model: CNN+RNN with accuracy of 54%.
  - Video model: CNN+RNN+3DCNN with accuracy of 51.94%.
- Analysis:
  - Low count of happy emotion in the dataset.
  - Audio+video model works with training on 3 emotions resulting in accuracy jump from 70.25% to 71.75%.

Future Work:

- Increase input & output dimensions in each layer in the network.
- Auto crop to focus on the face of the actor in video frames.
- Explore noise removal algorithms.

Thank you