DEPARTMENT OF
ELECTRICAL & COMPUTER
ENGINEERING

ECE13
Audio & Visual Emotion Detection

Liam Hodgson, Vincent Jung, Alec Warner

A. JAMES CLARK
SCHOOL OF ENGINEERING

Overview & Application

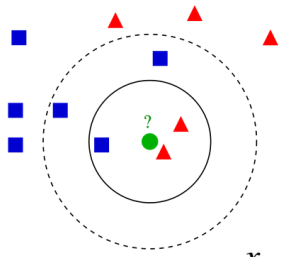
- Objective: detect ones emotion using a combination of the visual and audio modalities
- Applications:
  - Healthcare in virtual environments
  - Deeper understanding for online AI assistance systems
  - Education adaptation based on emotion and inferred level of comprehension

Dataset

- CREMA-D
- 7442 Videos
- 91 Actors ages 20-74
- 6 Emotions, 3 Intensity Levels

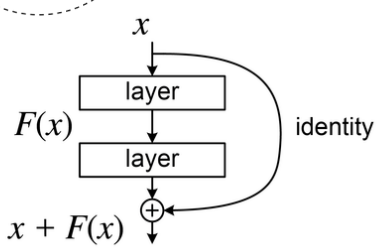
Initial Fusion Approach:

- Reasoning: simpler implementation, foot-in-the-door
- Audio: K-Nearest Neighbors model
- Video: Trained all layers of ResNet backbone
- Output: Average between audio and video models



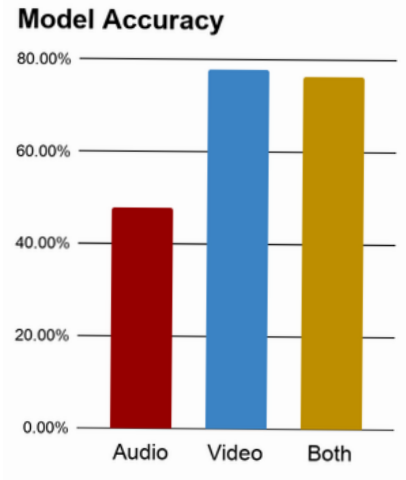
Design Constraints:

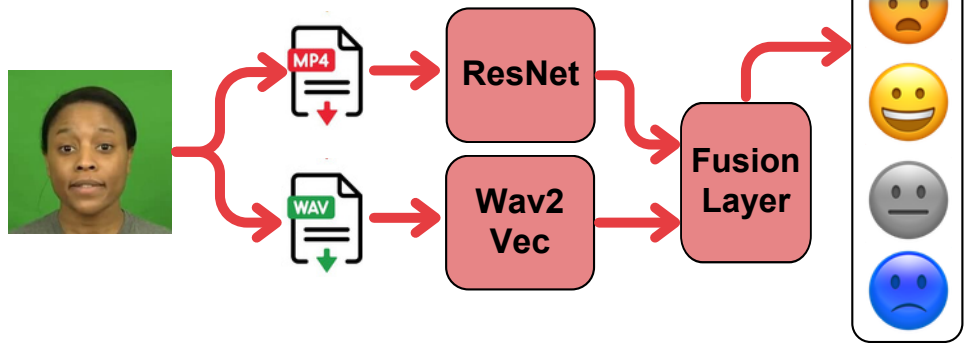
- Too long to train (*retraining whole ResNet model is unnecessary*)
  - Longer training times → inability to train large sets of data
- Simpler models led to less comprehensive training
- Averaging layer unable to learn from results



Advanced Fusion Approach:

- Reasoning: separate implementation per modality, interpretation layer
- Audio: Wav2Vec Encoder model
- Video: ResNet backbone with initial layers frozen
- Output: Combined feature vector with Neural Network layer for classification





Reference list: https://docs.google.com/document/d/1paxgreNQh-nzQcn4czASDr3nTSXQTfL8o4QddPWKbX8/edit?usp=sharing