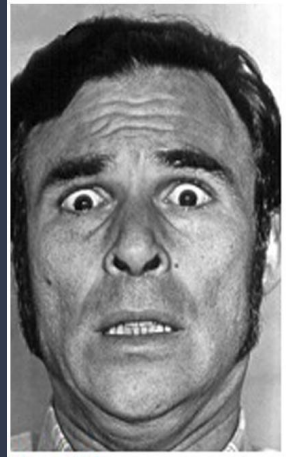
Multimodal Emotion Recognition AI

Team:
Avinash Nandyala
Rajesh Mallula
Rakshith Aloori
Tejendhar Karanam
Bhanuteja Bolisetti

AI in emotion classification


Fearful


Angry


Sad


Happy


Disgusted

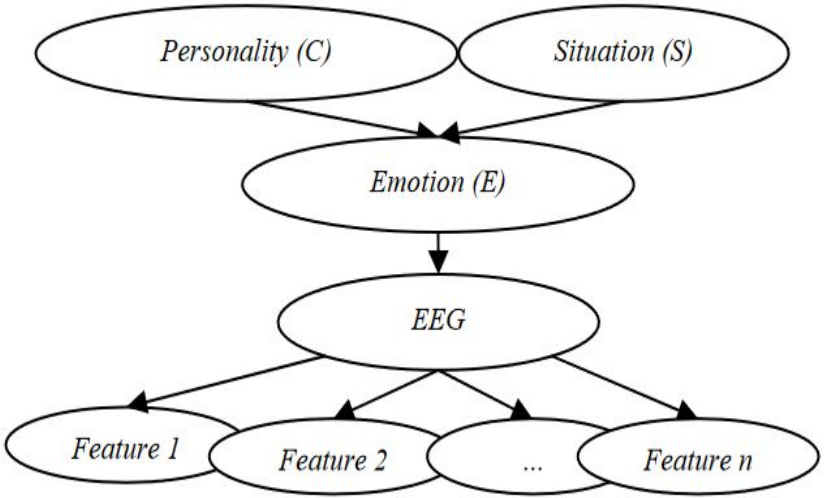

Surprised

Introduction

- **Emotion recognition** is the process of identifying human emotion
- People vary widely in their accuracy at recognizing the emotions of others.Use of technology to help people to recognize emotion is a part of Artificial Intelligence research study.
- Generally, the technology works best if it uses multiple modalities in context.
- In current research studies on AI in emotion , most work has been conducted on recognizing facial expressions from **video, audio , text,** and other psychological parameters associated with that emotion.
- Moreover the models used , can be dependent on the context of what emotion , we are trying to recognize. So domain knowledge plays a key role here.

Approaches

- Decades of extensive research , resulted in extensive literature and extensive methods from multiple areas such as signal processing, machine learning, computer vision, speech processing.
- Different methodologies are employed to interpret emotion , for example: **Bayesian networks**.



Approaches

- Most of the emotion recognition techniques accuracy are improved if we take into account the video, audio and text data altogether.

Approaches classified

- Knowledge-based : using domain knowledge and semantic characteristics of an emotion.
- Statistical methods: large data fed to a supervised learning algorithm, like computer vision techniques

Approaches

Knowledge-based approaches:

- Classifiers such as WordNet, SenticNet, ConceptNet, EmotiNet.
- More used , as the large resources available for domain specific emotions

Statistical approaches:

- Popular machine learning algorithms like SVM , Naive Bayes, Maximum Entropy.
- Deep learning techniques like LSTM, CNN, ELM.
- Used when large, accurate and unbiased data is available

Context :

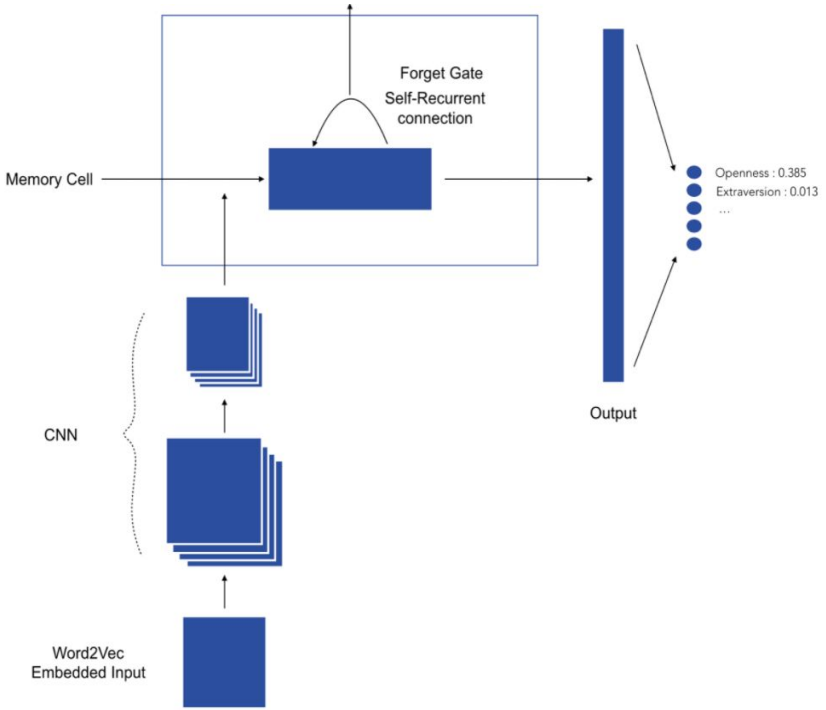
- We are here to develop an **Multimodal emotion recognition** platform to analyze the emotions of **job candidates**.
- We analyse **facial**, **vocal** and **textual** emotions, using mostly **deep learning** based approaches.
- This field has been rising with the development of social network that gave researchers access to a vast amount of data.
- We do three types of analysis :
  - Text Analysis
  - Audio Analysis
  - Video Analysis

Methodology: Textual Analysis

- **Input :** Textual input, such as answers to questions that would be asked to a person from the platform
- **Basic Pipeline:**
  - Text data retrieving
  - Custom natural language preprocessing :
    - Tokenization of the document
    - Cleaning and standardization of formulations using regular expressions
    - Deletion of the punctuation
    - Lowercasing the tokens
    - Removal of predefined *stopwords*
    - Application of part-of-speech tags on the remaining tokens
    - Lemmatization of tokens using part-of-speech tags for more accuracy.
    - Padding the sequences of tokens of each document to constrain the shape of the input vectors.
  - 300-dimension Word2Vec trainable embedding
  - Prediction using our pre-trained model

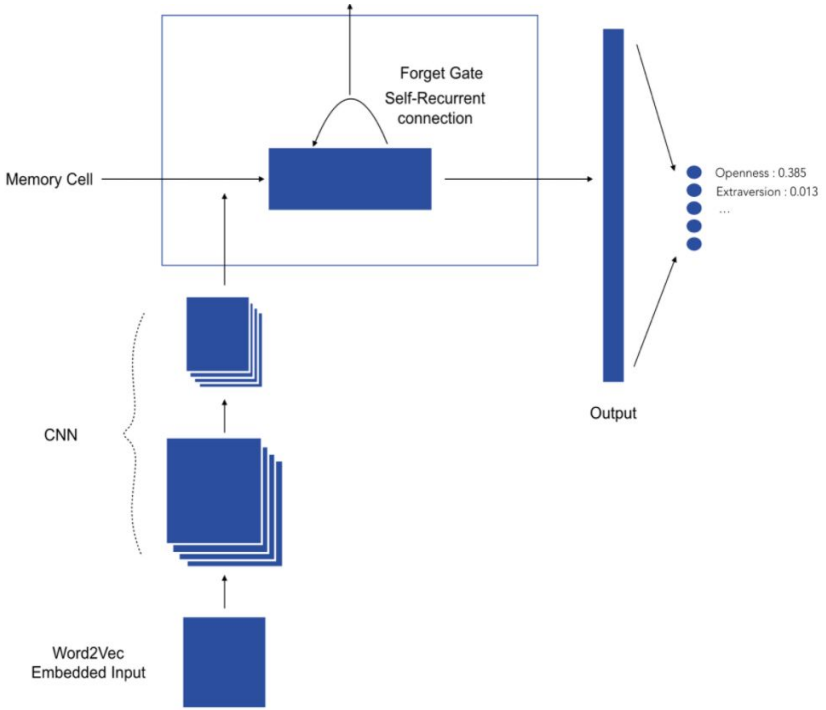
Methodology : Textual Analysis

- **Model :**
  - We have to choose a neural network based on both one-dimensional CNN and RNN's
  - *CNN plays a role of Feature extraction* : It allows finding patterns in text data.
  - The Long-Short Term Memory cell is then used in order to leverage on the sequential nature of natural language : unlike regular neural network where inputs are assumed to be independent of each other, these architectures progressively accumulate and capture information through the sequences.
  - LSTMs have the property of selectively remembering patterns for long durations of time.



Methodology : Textual Analysis

- **Model :**
  - Our final model first includes 3 consecutive blocks consisting of the following four layers : one-dimensional convolution layer - max pooling - spatial dropout - batch normalization.
  - The numbers of convolution filters are respectively 128, 256 and 512 for each block, kernel size is 8, max pooling size is 2 and dropout rate is 0.3
  - **Finally**, a fully connected layer of 128 nodes is added before the last classification layer



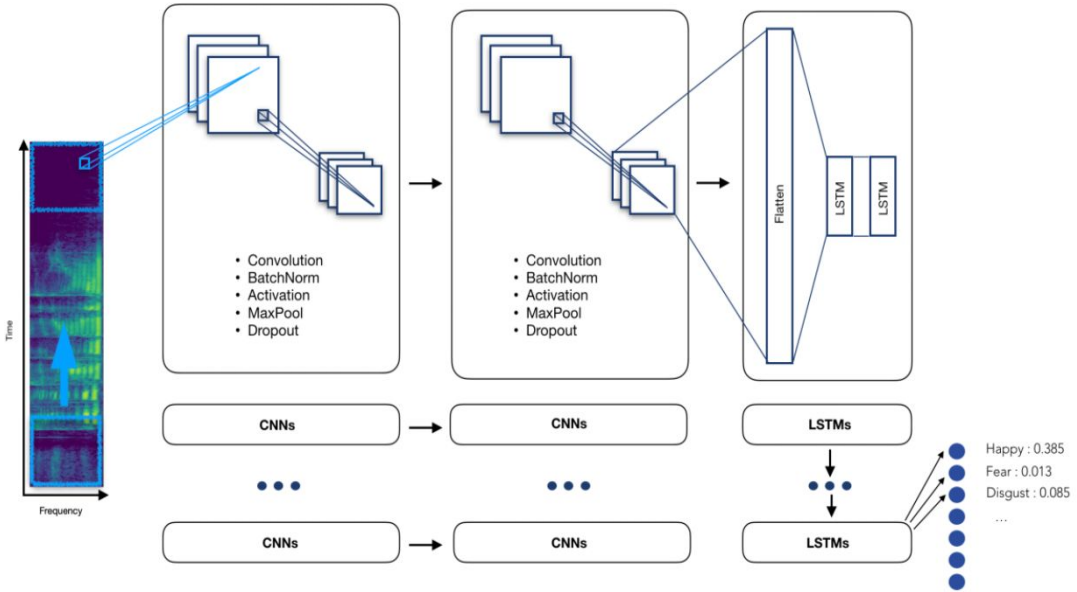
Methodology: Audio Analysis

- **Input :** Voice is recorded from Job interviewers for Speech Emotion Capture. Speech includes calm, happy, sad, angry, fearful, surprise, and disgust expressions, and song contains calm, happy, sad, angry, and fearful emotions.

- **Basic Pipeline :**
  - Voice recording
  - Audio signal discretization
  - Log-mel-spectrogram extraction
  - Split spectrogram using a rolling window
  - Make a prediction using our pre-trained model

Methodology : Audio Analysis

- **Model :**
  - We have chosen **Time Distributed CNN**.
  - The main idea to do this is apply a rolling window (fixed size and time-step) all along the **log-mel-spectrogram**.
  - Each of these windows will be the entry of a CNN, composed by four **Local Feature Learning Blocks (LFLBs)** and the output of each of these CNN will be fed into a RNN composed by 2 cells LSTM (Long Short Term Memory) to learn the long-term contextual dependencies.
  - **Finally,** a fully connected layer with *softmax* activation is used to predict the emotion detected in the voice

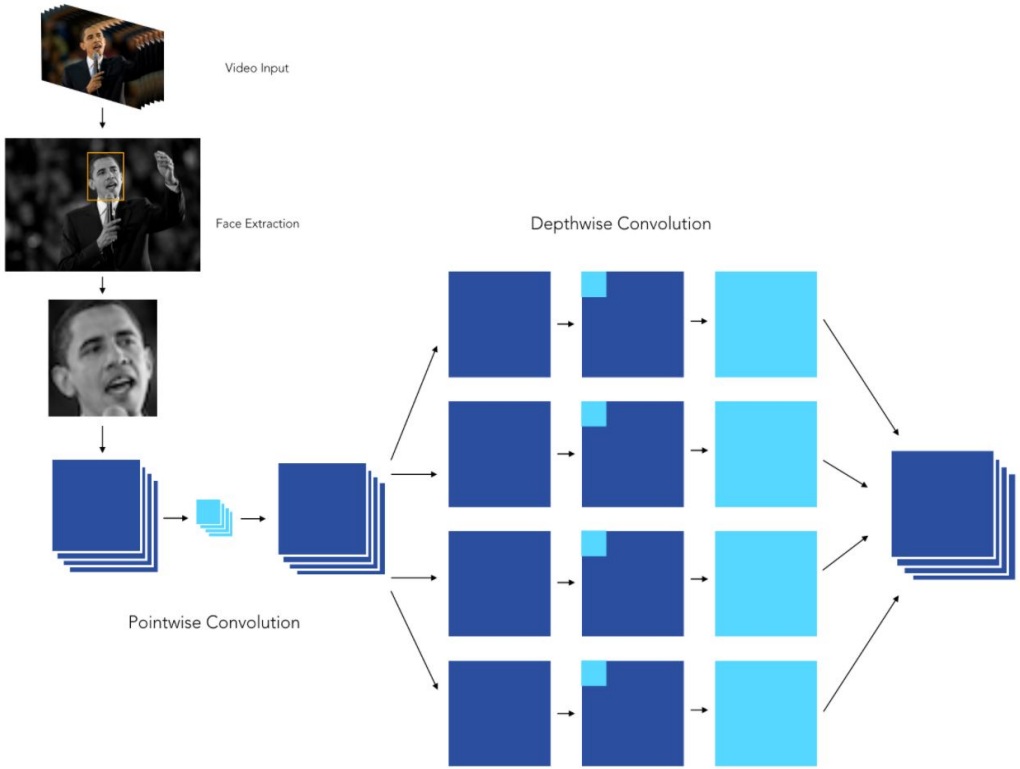


Methodology: Video Analysis

- **Input : Video input** from a live webcam or stored from an **MP4** or **WAV** file, from which we split the **audio** and the **images**. It takes the **Faces** of people to find **facial emotions**.
- **Basic Pipeline :**
  - Launch the webcam
  - Identify the face by Histogram of Oriented Gradients
  - Zoom on the face
  - Dimension the face to 48 * 48 pixels
  - Make a prediction on the face using our pre-trained model
  - Also identify the number of blinks on the facial landmarks on each picture

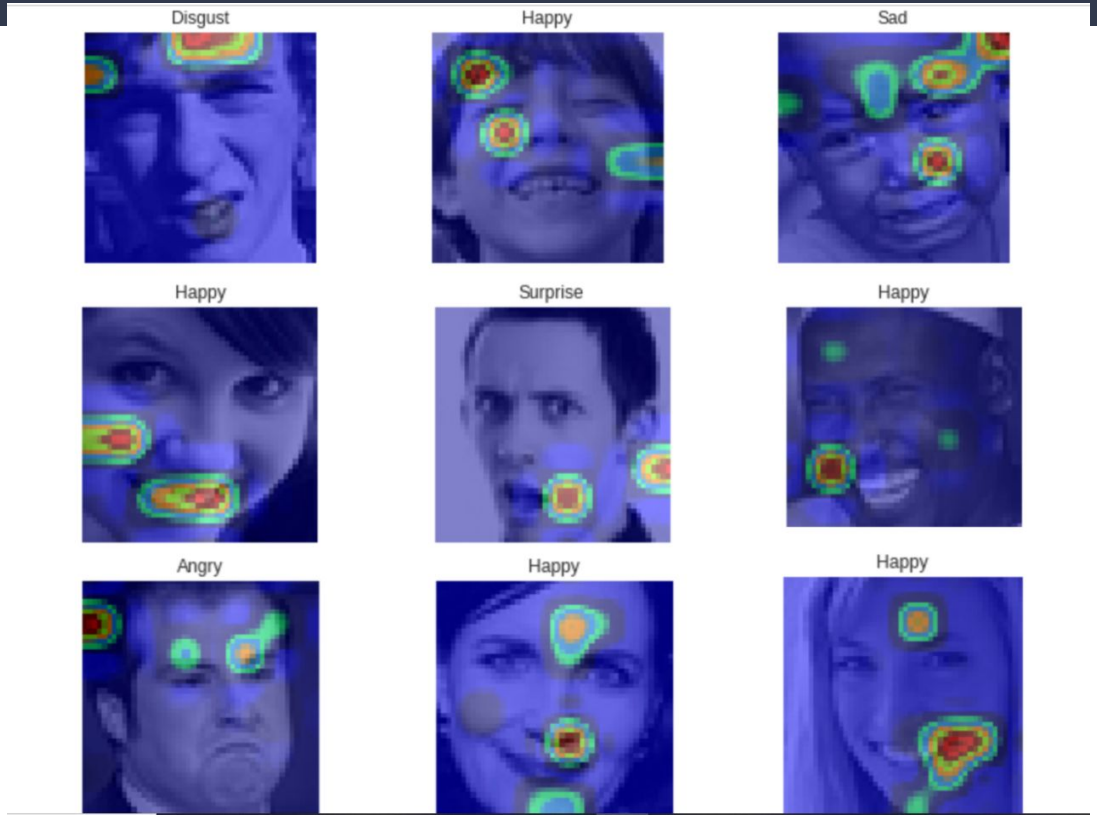
Methodology : Video Analysis

- **Model :**
  - We have Chosen **XCeption** model.We tuned the model with :
    - Data augmentation
    - Early stopping
    - Decreasing learning rate on plateau
    - L2-Regularization
    - Class weight balancing
    - And kept the best model
  - The XCeption architecture is based on DepthWise Separable convolutions that allow to train much fewer parameters, and therefore reduce training time on Colab's GPUs to less than 90 minutes



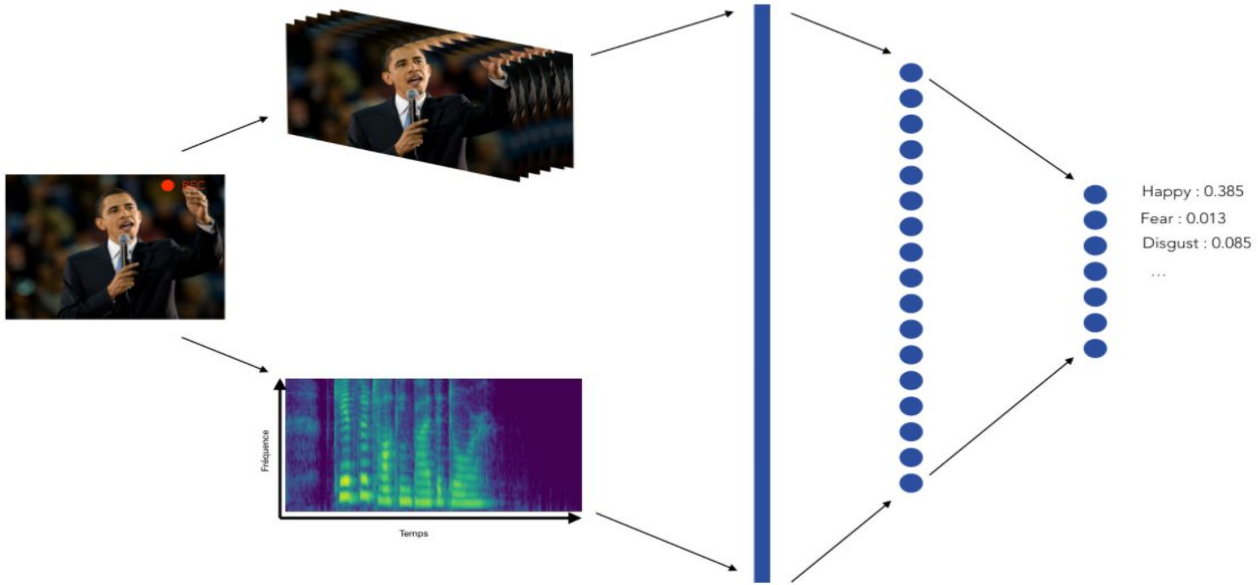
Methodology : Video Analysis

- **Model :**
  - When it comes to applying CNNs in real life application, being able to explain the results is a great challenge.
  - We can indeed plot class activation maps, which display the pixels that have been activated by the last convolution layer.
  - We notice how the pixels are being activated differently depending on the emotion being labeled.
  - The happiness seems to depend on the pixels linked to the eyes and mouth, whereas the sadness or the anger seem for example to be more related to the eyebrows



Methodology : All analysis (Ensemble Analysis)

- We can combine all from a Video,We can get facial features, Audio ,speech to text and therefore Textual Analysis.



WHAT AI STRUGGLES WITH

Understand nuances

Detecting and decoding emotional subtleties based on social cues

Create original content

Producing work autonomously without large sets of data and defined parameters

Filter biases

Identifying biases based on social or ethical consciousness

Conclusion

- Other than proposed approaches in the introduction, there are also mixed type of approaches, where there is a use of domain knowledge and supervised data.
- Text data is a favourable research object for emotion recognition because the storage of text data is lighter and easy to compress.

Limitations:

- The risk of bias in Emotional AI.
- Because of the subjective nature of emotions, emotional AI is prone to bias.
- A study found that , recent emotion recognition techniques assigns more negativity to certain ethnicities.

Conclusion

- AI is not sophisticated enough to understand the cultural differences in expressing and reading emotions. For instance , a smile might mean one thing in Germany and another in Japan.
- To avoid bias, different methodologies are employed

➔ Implementing neuroscience technologies, such as facial encoding, biometrics, which resulted in 62% accuracy. When combined , accuracy raised to 77%

➔ Using historical data to train the model.

- In total, emotional AI will be powerful tool indeed, forcing businesses to reconsider their relationships with consumers and employees alike.