CS563 - NLP
END TERM PROJECT

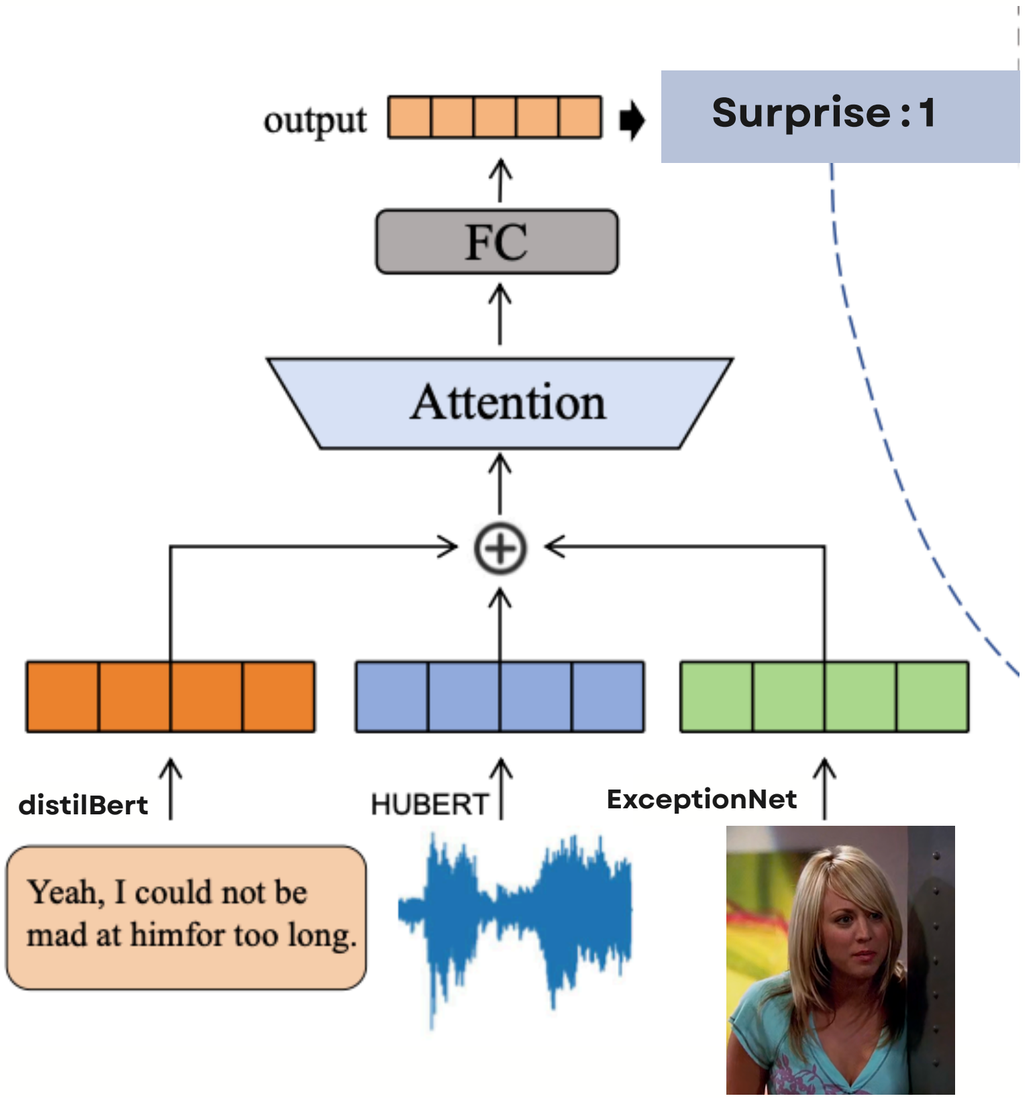
MULTIMODAL SURPRISE AND CAUSE DETECTION

SAI TULASI KANISHKA | 2101CS57

RISHIKANT CHIGRUPAATII | 2101CS66

SURAJ WARRIER | 2101CS75

OUTLINE



PHASE 1

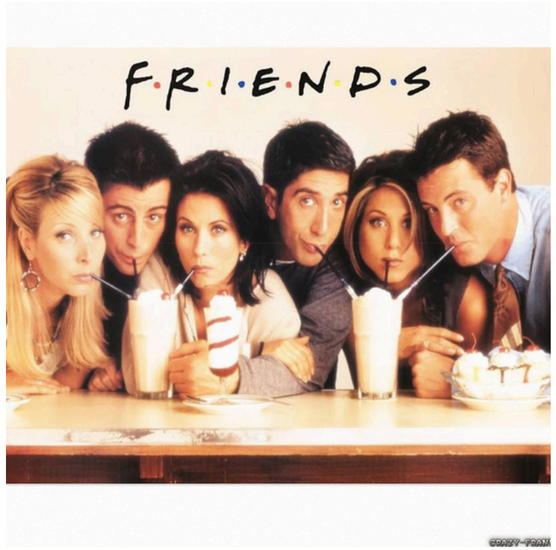
- BIG BANG THEORY DATASET
- MODEL STRUCTURE
- INFERENCE AND PROCESSING PIPELINE

PHASE 2

- CAUSE IDENTIFICATION and Linking

CHOOSING DATASET

MELD dataset



Already available emotion detection dataset

from Youtube Shorts

from Instagram reels

Instead of all these sources we wanted to make a dataset specifically tailored for Surprise detection

DATASET NOVELTY

We have made a Novel dataset

BIG BANG THEORY

Human annotated surprise detection dataset

~ **2000** DATAPOINTS OF WHICH APPROX 130 ARE SURPRISE

DATASET CREATION PIPELINE

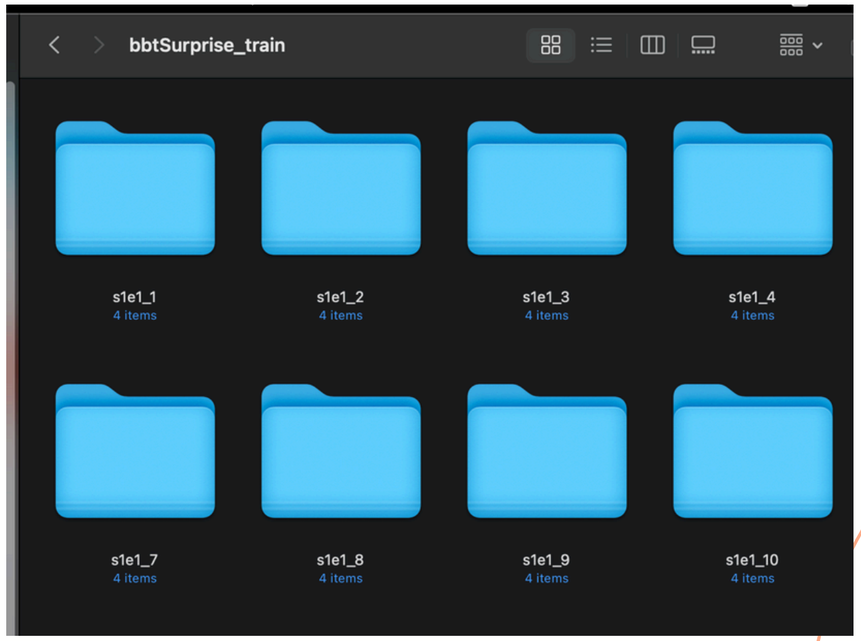
Episode

Script that extracts embedded subtitles from the video

script Splitting video according to subtitles and Surprise annotation

video, audio, text for each chunk of the episode

make CSV file to allow Human annotation of Surprise

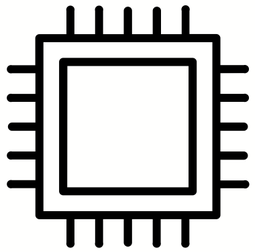


ANNOTATION PROCESS

H3

	A	B	C	D	E
1	**id**	**start_time**	**end_time**	**caption**	**surprise**
2	1	00:00:03,693	00:00:05,695	Would you pass the mustard? - Sure.	0
3	2	00:00:06,070	00:00:07,905	Hey. Wanna hear a fun fact about mustard?	0
4	3	00:00:08,114	00:00:10,616	Is it that the glucosinolates which give mustard its flavor	0
5	4	00:00:10,783	00:00:14,245	were evolved by the cabbage family as a chemical defense against caterpillars?	0
6	5	00:00:15,871	00:00:17,206	Yeah.	1
7	6	00:00:18,833	00:00:20,918	Well, that was fun. Good for you, Leonard.	0
8	7	00:00:22,211	00:00:23,754	- Hey there. - Hey, you're early.	0
9	8	00:00:23,921	00:00:25,172	The movie doesn't start for an hour.	0
10	9	00:00:25,339	00:00:27,591	Actually, we're not going to the movies.	0
11	10	00:00:27,758	00:00:30,136	We are here to kidnap you.	0
12	11	00:00:30,469	00:00:31,512	What are you talking about?	0
13	12	00:00:31,679	00:00:34,765	Well, you eloped and we didn't get a chance to throw you a bachelor party,	0
14	13	00:00:34,932	00:00:36,350	so there's a van downstairs	0
15	14	00:00:36,517	00:00:39,979	and we're here to take you to a surprise location for the weekend.	0

THE MODEL
COMPONENTS



TEXT ENCODER

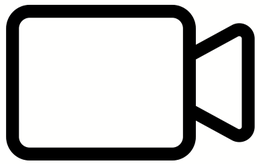
DISTILBERT

DISTILBERT

AUDIO ENCODER

HUBERT

CNNs

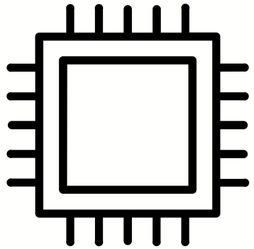


VIDEO ENCODER

EXCEPTIONET

RESNET

TEXT ENCODER



TEXT ENCODER

DISTILBERT

Core Components

- **Base Model**: DistilBERT (lightweight BERT variant)
  - Pretrained on 11M+ documents
  - 6-layer transformer architecture
  - 768-dimensional hidden states
- **Custom Layers**
  - FC layer: 768 → 256 dimension reduction
  - Output: 256-D semantic embeddings

AUDIO ENCODER

AUDIO ENCODER

HUBERT

Audio Encoder Architecture

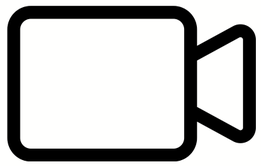
Core Components:

- **HuBERT Base Model** (pre-trained, frozen)
  - Extracts 768-dim features from 16kHz audio
  - Outputs frame-level features [B, T', 768]
- **Projection Network**

  `Linear(768 → 512) → ReLU → Dropout(0.5) → Linear(512 → 256)`

  - Reduces dimensionality to target hidden space
  - Adds non-linearity and regularization

VIDEO ENCODER



VIDEO ENCODER

XCEPTIONET

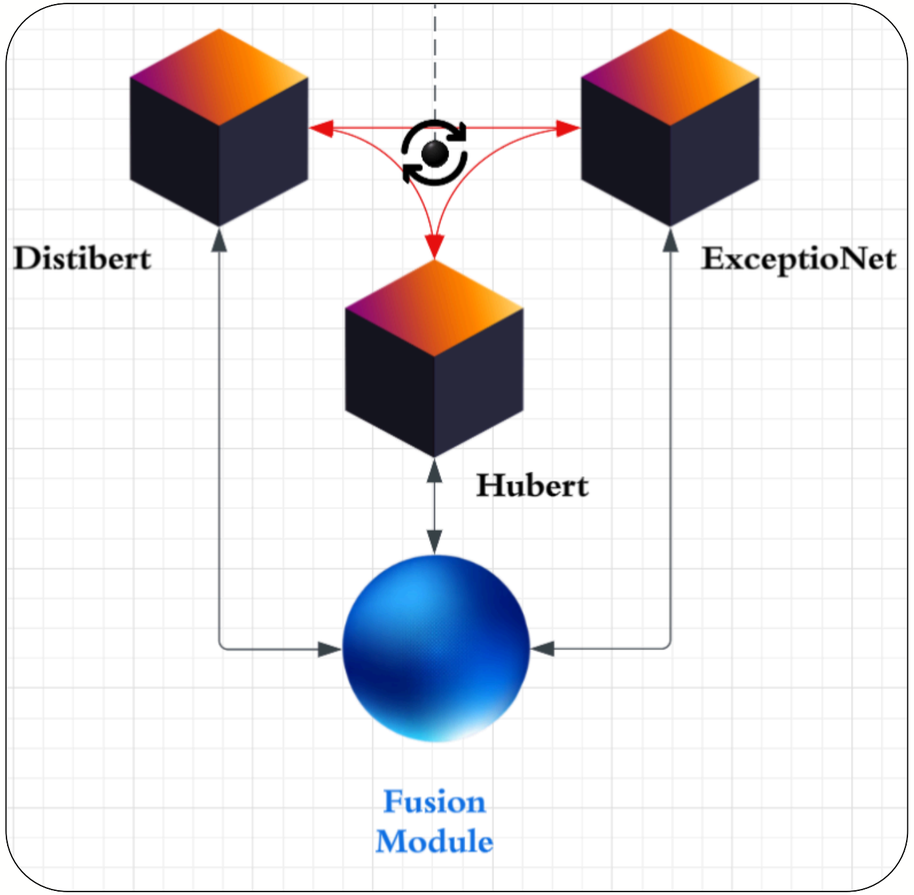
- **Pre-trained Backbone**: Utilizes a pre-trained Xception model for robust feature extraction.
- **Frozen Parameters**: All backbone parameters are frozen to prevent overfitting and focus on learning from new data.
- **Projection Layer**: Projects the output into a lower-dimensional space (`hidden_dim`) for easier integration with other modalities.
- **Face Detection**: Uses OpenCV's Haar Cascade Classifier to detect faces in video frames.
- **Feature Extraction**: Extracts facial features by resizing and normalizing detected faces before passing them through the Xception backbone.

TRANSFORMER FUSION

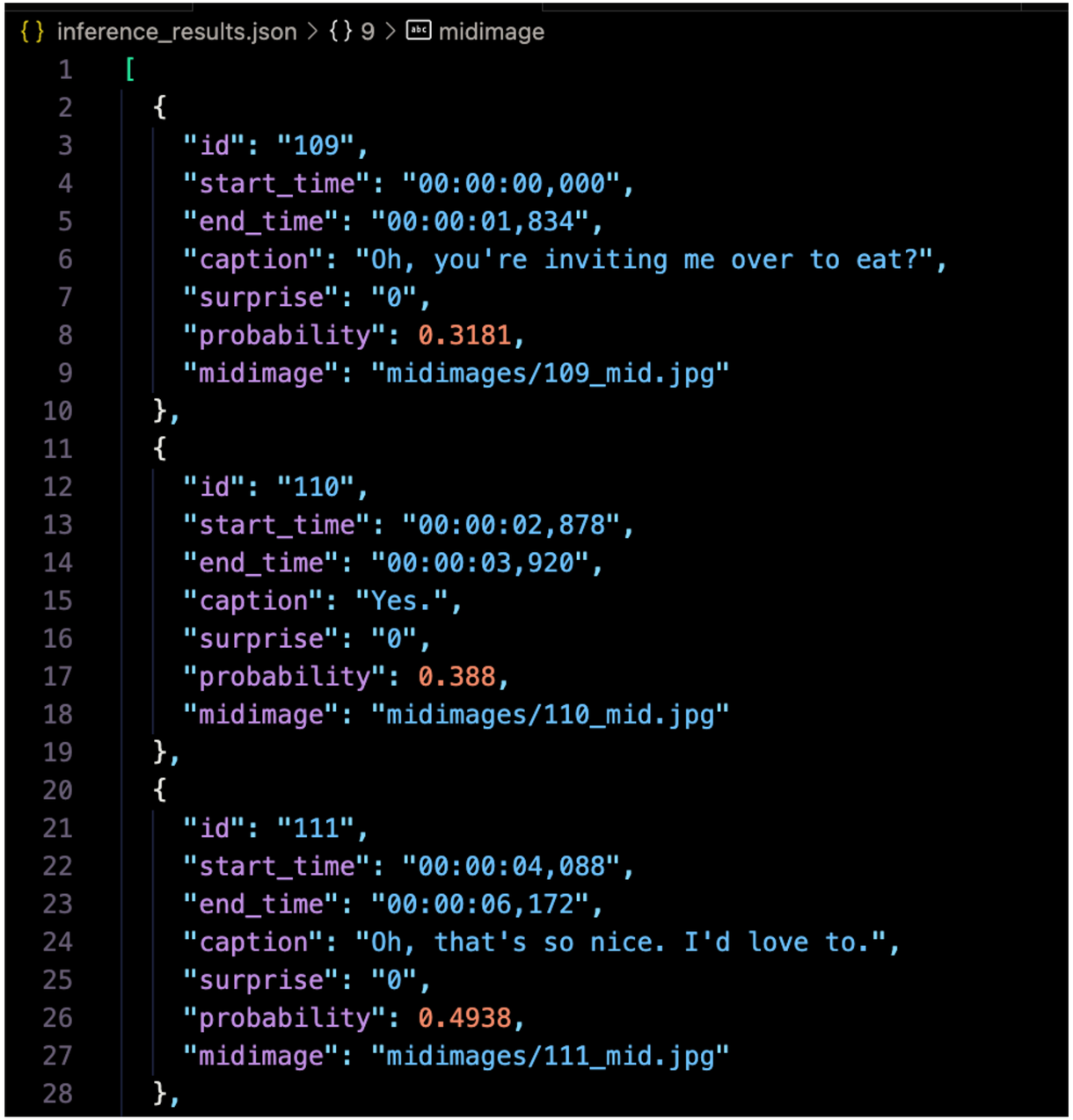
We use a Transformer-based fusion module to integrate features from text, audio, and video modalities for emotion recognition:

Approach:

- Input Features: Each modality is encoded into a vector of size hidden_dim = 256
- These are stacked to form a sequence of 3 tokens per sample – one per modality:
  Shape: [Batch, 3, 256]
- We add learnable modality embeddings to each token to help the model distinguish between text, audio, and video inputs.
- The sequence is passed through a Transformer Encoder with:
  a. 2 layers (num_layers=2)
  b. 4 attention heads per layer (num_heads=4)
  c. ReLU Activation
  d. Feedforward hidden size = 512` (2×256)



GENERATED JSON FILE



PHASE II

MULTIMODAL CAUSE EXTRACTION

- We begin by parsing the JSON file generated during the previous phase to identify utterances that have been classified as exhibiting 'surprise'.
- For each identified target utterance, we compile its associated dialogue history along with the corresponding visual context (image) and provide this combined input to a large language model (LLM).
- The LLM processes this multimodal information to infer the utterance within the dialogue history that is most likely to have triggered the target surprise utterance.
- This procedure is repeated for all utterances labeled as 'surprise', resulting in a collection of surprise–cause pairs.


```
[
  {
    "id": "116",
    "start_time": "00:00:42,543",
    "end_time": "00:00:46,003",
    "caption": "-Thank you.\n-You're very welcome.",
    "surprise": "0",
    "probability": 0.4935,
    "midimage": "midimages/116_mid.jpg"
  },
  {
    "id": "117",
    "start_time": "00:00:46,880",
    "end_time": "00:00:48,840",
    "caption": "[MOUTHS]\nYou're very welcome.",
    "surprise": "0",
    "probability": 0.5006,
    "midimage": "midimages/117_mid.jpg"
  },
  {
    "id": "118",
    "start_time": "00:00:49,341",
    "end_time": "00:00:52,135",
    "caption": "This looks like some serious stuff.\nLeonard, did you do this?",
    "surprise": "1",
    "probability": 0.485,
    "midimage": "midimages/118_mid.jpg",
    "cause": "114"
  }
]
```

MULTIMODAL CAUSE EXTRACTION

Due to limited training data and imbalanced data distribution, which could lead to overfitting, we transformed the emotion cause extraction task from a traditional discriminative architecture to a generative architecture based on Multimodal LLM.

We utilize the trainable open-source multimodal model, Meta's Llama-3.2-11B-Vision-Instruct for multimodal cause extraction (MCE)

MULTIMODAL CAUSE EXTRACTION

Returns the Causal Utterance ID

Llama-3.2-11B-Vision-Instruct

Prompt: - You are given: A target utterance labeled as **'surprise'**, An image corresponding to the target utterance, A few previous utterances with their IDs. Identify which one previous utterance most likely caused the surprise. Use both the text and the image.
Target Utterance:
{prompt['target_utterance']}
Previous Utterances:
{formatted_history}

History:
Leonard - Guys, tickets already went on sale!
Sheldon - What?
Howard - They're not supposed to be available yet.
Leonard - I don't know what to tell you, but they're on sale!

Target:
Raj - Wait. Are you sure they're *Star Wars* tickets?

CONFUSION MATRIX	Actual Positive	Actual Negative
Predicted Positive	305	37
Predicted Negative	5	7

SCORES/ SPLITS	Accuracy	F1 Score
Training	0.75	0.58
Validation	0.73	0.60
Test_imbalanced	0.88	0.25
Test_balanced	0.71	0.58

Accuracy of cause extraction for test: ~ 50%

Positive → Surprise : 0

PIPELINE SUMMARY