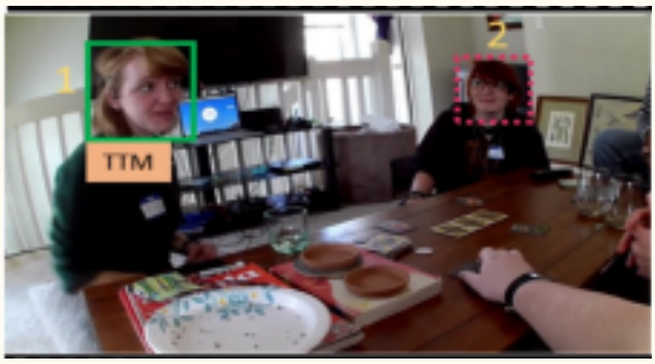
Talking to Me (TTM) - Team NoName

R11942079王式珩、B08901165南策昇、B08901174 郭尚睿、B08901189黃曜廷、B08901204蔡芳鐸

Overview

The task aims to teach the model to tell whether the particular person is talking to the camera or not, according to the given video and audio input. We implement multiple techniques to utilize both audio and image features. Related experiments and ablation study are shown below and will be introduced later.



Pre-process

Since the data input is multi-modal, we need to pre-process each of them separately. For video data, we first extract each frame and crop the faces in the frame. For audio data, we convert the audio segment into **mfcc** feature. To save computation costs, we sample 1 frame from 3 consecutive frames. Furthermore, we set up a max length of **mfcc** length to prevent wasting memory.

Method1

We encode raw data into features with the same dimension. The image and face encoder can be ViT or ResNet50 in our setting. For the audio data, which is **mfcc** features, we encode it by the **Transformer Encoder** and transform it into the same dimension as the image feature. Then, we concatenate speech, frame, and face data into one feature. In the end, we feed the feature into the classifier model, which is an MLP model. The classifier will predict whether the speech, frame, and face pair is 1 or 0. And '1' 'stand for talking to me'. The model architecture is shown in Figure 1.

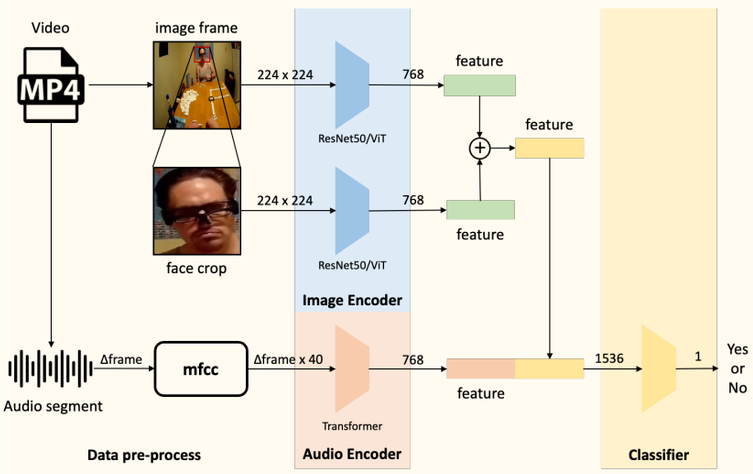


Fig. 1: Model architecture

Method2

For each conversation, we first select 10 frames from "start frame" to "end frame" with same interspaces. Additionally, we randomly sample 3 frames from that conversation interval. After passing these 13 frames through ResNet50, we feed them into LSTM/GRU model chronologically. Concatenating the final output from LSTM/GRU model with the audio feature generated by Transformer, we can get the prediction by a classifier, Shown in Figure 2.

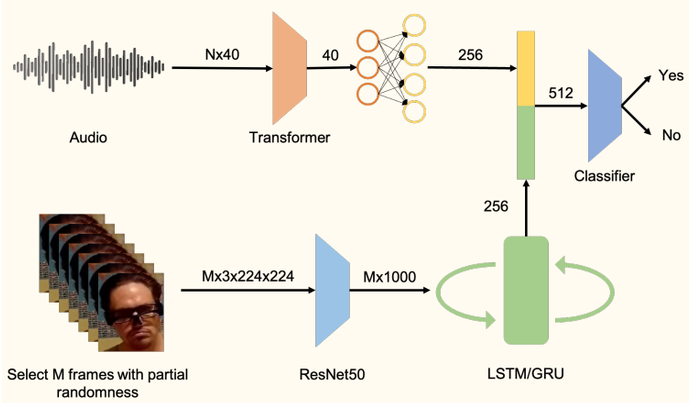


Fig. 2: Model architecture

Experiment & Ablation Study

To evaluate the result, we first pick 94 videos to be the validation set. All baseline is evaluated with the same validation set. All experimental setups are consistent except for the variable we want to compare. The setups and the training/validation accuracies are shown in the below table.

	Fix image encoder		Tune image encoder	
	n_head=4	n_head=8	n_head=4	n_head=8
ResNet50	70.56/64.31%	70.60/63.17%	71.53/58.69%	72.98/60.98%
ViT	71.49/62.40%	66.72/63.02%	77.63/66.4%	73.49/61.29%

LSTM	Total frame=10	Total frame=13
Random frame=3	65.96 / 63.54%	66.83 / 63.95%
Random frame=0	65.87 / 63.97%	62.77 / 63.19%