HATE VIDEO DETECTION WITH MULTI-MODEL

-Shailesh Mahto -Wendan Zhao -Ruijun Liu

This study is conducted purely for academic purposes. We do not harbor any bias or prejudice towards any individual or group based on race, ethnicity, or cultural background.

University at Buffalo The State University of New York

Dataset

Statistics:

- **1083 videos:** 431 hate and 652 non-hate (2:3 ratio)
- **Audio:** Analysis of average root mean square energy indicate instances of shouting which possibly manifests in the form of high loudness. (Figure1)
- **Transcript:** obtained through Vosk, lexical analysis shown certain words more likely to appear in hate videos. (Figure2)
- **Video:** Appearance of certain objects, such as religious persons, cartoonish mockery, and stereotypical depictions.

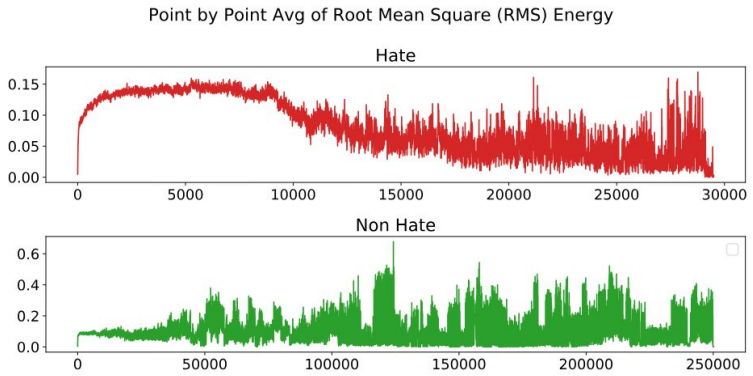


Figure1 Avg RMS Energy

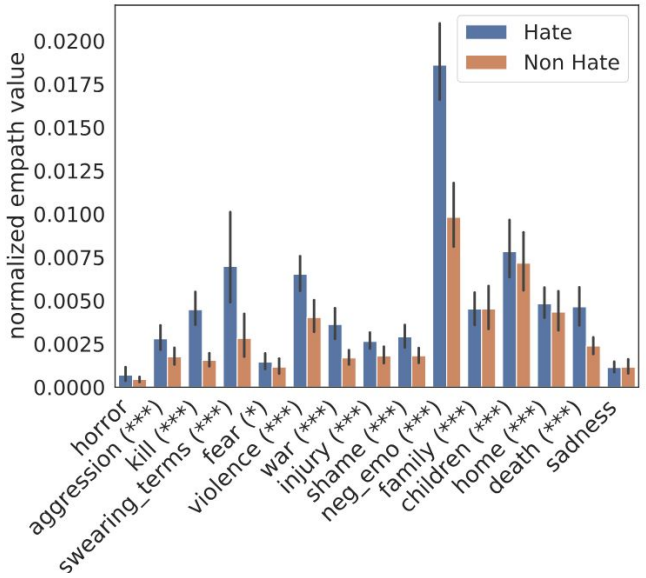


Figure2 Lexical analysis of transcript

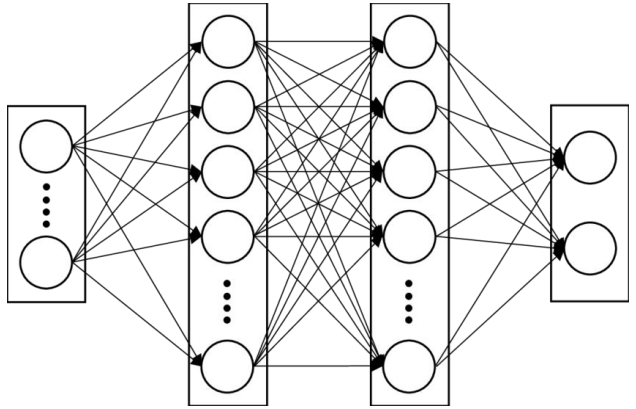
Audio

MFCC:

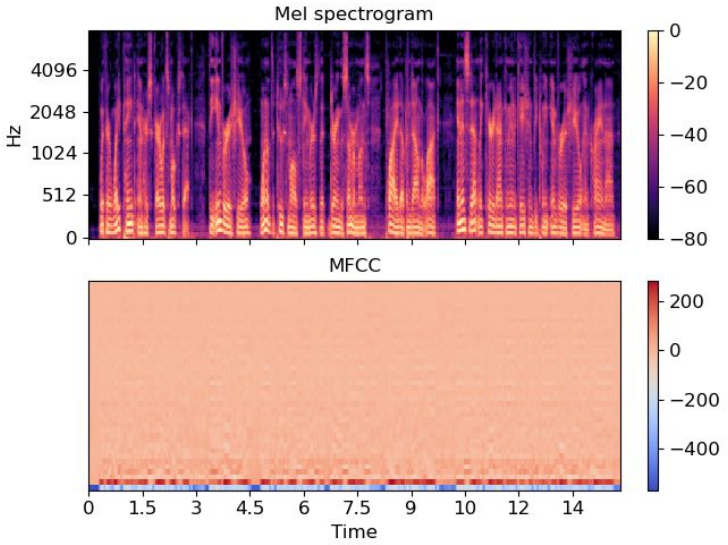
- Robust features of audio.
- 40 dimensional vectors. (padded to keep uniformity)
- Represent the short-term power spectrum of a sound.
- Easy to obtain and trained on.
- Saved in pickle files with names, features and labels.

Fully-connected layers:

- Simple but works well with MFCC features



Fully-connected layers with ReLU
2 hidden layers



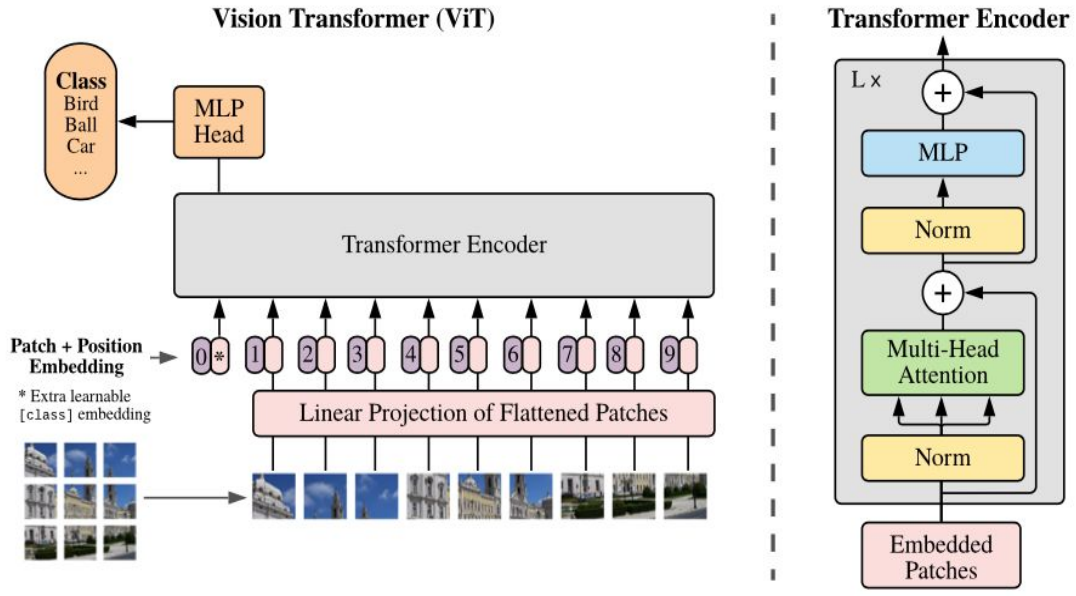
Video frames preprocessing

Frames extraction

- 100 frames extracted uniformly for all videos.

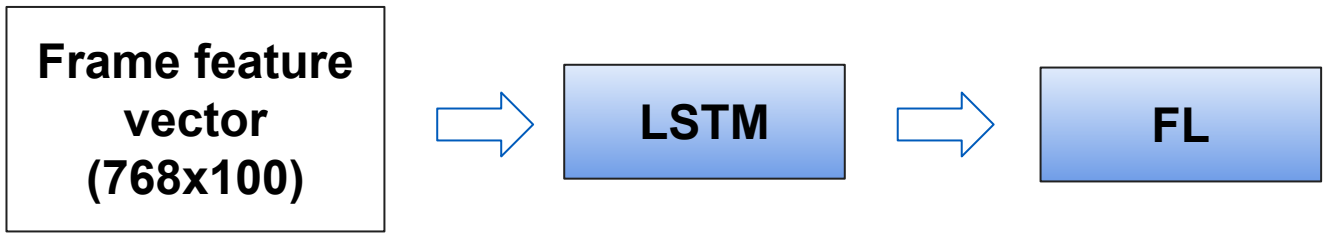
Feature extraction using pre-trained ViT(Vision Transformer)

- Transformer-based architecture.
- Self-attention mechanisms, allowing for interactions between image patches at different spatial locations.
- Divides the input image into fixed-size patches and linearly embeds each patch into a sequence of tokens.



ViT model + LSTM

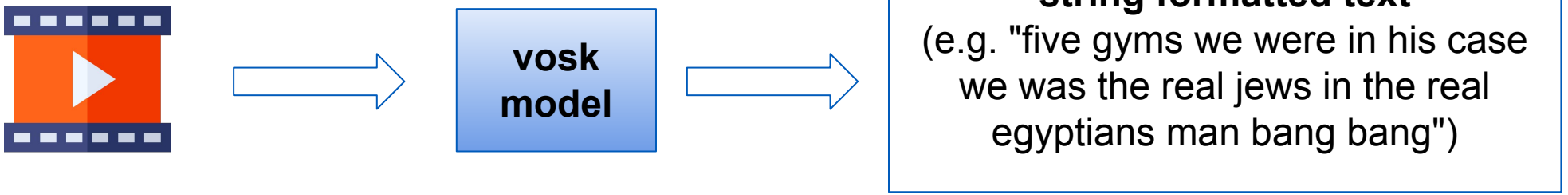
- Vit model deals with spatial features of images, but video contains sequence of images
- To further capture the temporal interactions between the images, pass these feature vectors to LSTM model
- Finally Apply a classifier header layer to obtain the prediction.



Transcript preprocessing

Transcripts extraction

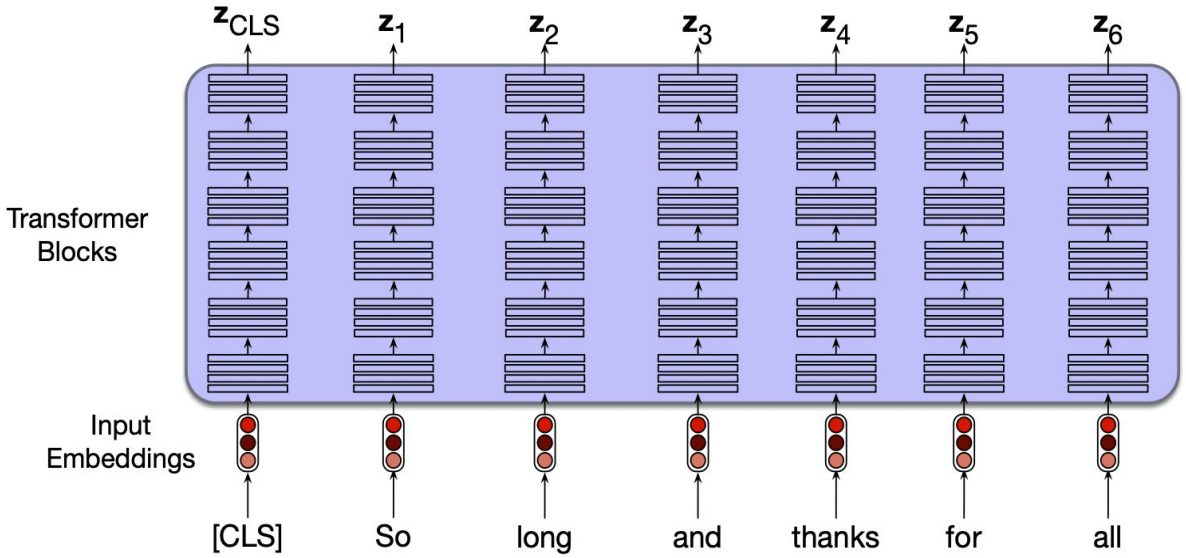
- **vosk**, offline speech recognition toolkit.
- save all transcripts including hate and non hate categories into a pickle file labeled with video names.
- Basic statistics for words in text, measured in mean value, hate 253 words, non hate 227 words, total 237 words



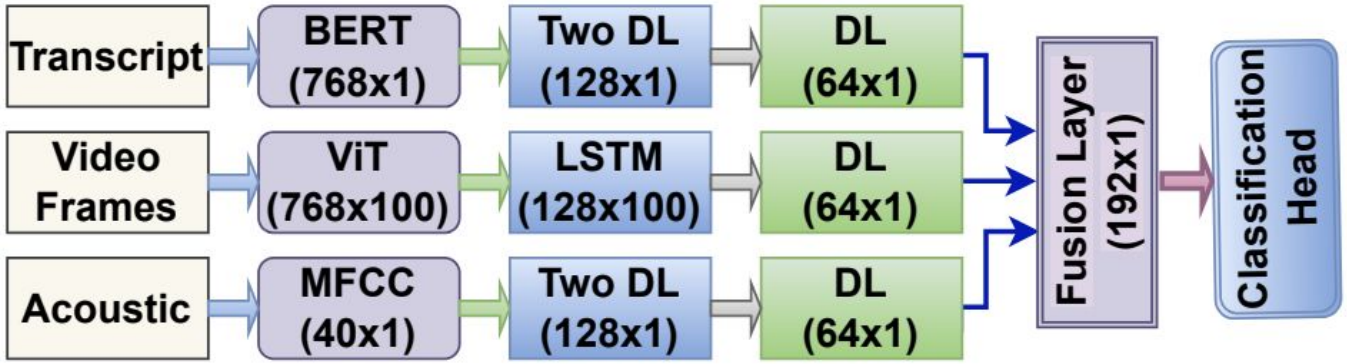
Transcript modeling

Transcripts tokenization & contextual embeddings extraction

- Bert Tokenizer, built-in tokenizer for BERT model.
- tokenizer takes text as input, transforms words or tokens in the text with ids and attention masks
- as the words length of transcripts vary, use max length acceptable by the model, padding text with short length and truncate text with exceeded part
- Sum of token embeddings, segmentation embeddings and position embeddings are fed forwarded to the BertEncoder with 12 layers.
- Use CLS token to represent the entire text and fed forwarded to classifier



Multimodal approach



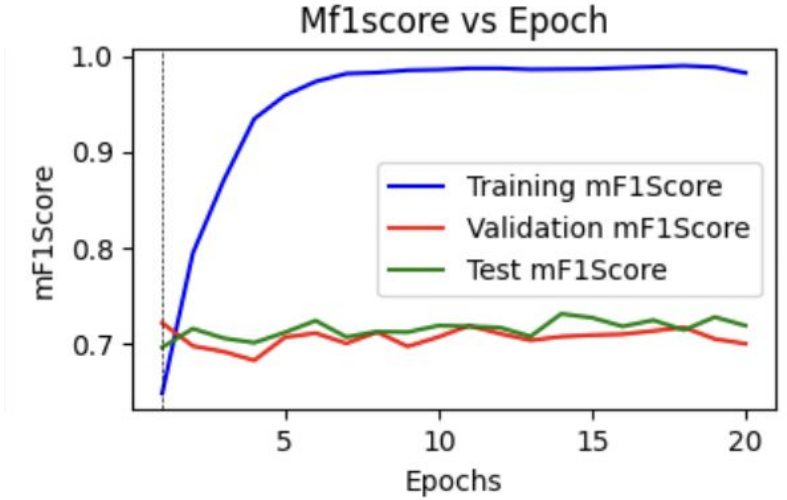
Models we tested:

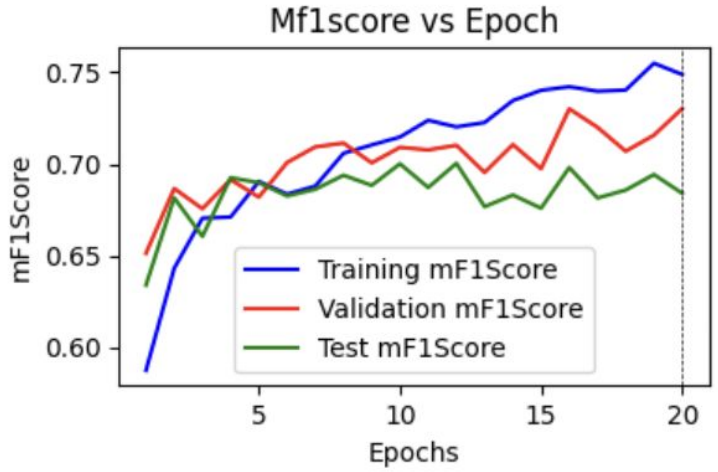
- Transcript + Frames
- Transcript + Audio
- Frames + Audio
- Transcript + Frames + Audio

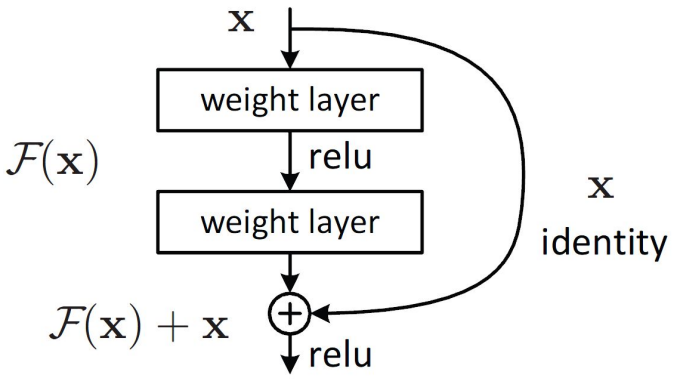
Improving performance

- Best performing model achieved after first epoch:
  - batch normalization
  - Learning Rate Scheduler (ReduceLROnPlateau)
  - Xavier Initialization

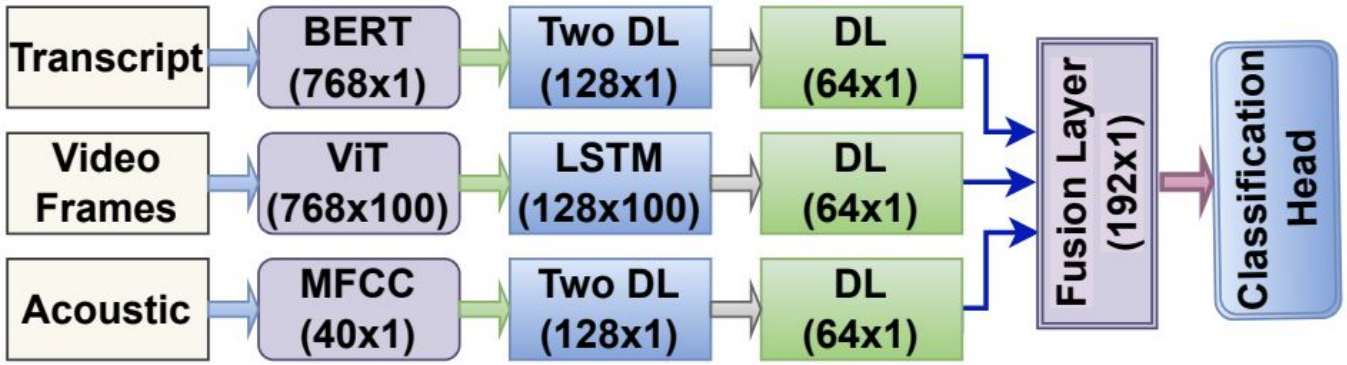
- For further improvement, we tried
  - Increase model complexity
  - Residual blocks with FCL



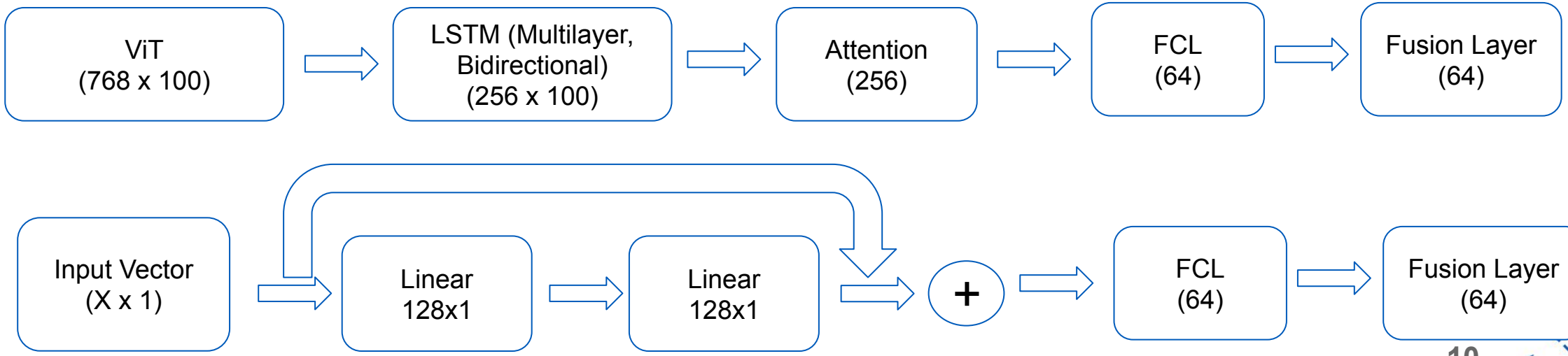




Initial Architecture



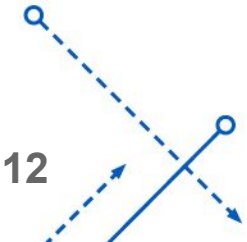
Architecture after increasing complexity:



Results(Accuracy, Macro-F1 Score)

Modality	Baseline	Reduce LR On Plateau	Xavier Initialization	Linear Complexity ↑	Add Residual Blocks (?)
MFCC	71.28% 70.04%	70.82% 69.77%	67.40% 66.27%	71.09% 70.14%(2)	**71.38%** **70.39%(3)**
VIT	70.91% 69.41%	70.91% 69.41%	**74.95%** **72.67%**	72.85% 71.03%(2)	73.96% 72.45%(1)
BERT	76.91% 75.48%	**76.91%** **75.48%**	74.42% 73.13%	71.46% 72.48%(3)	74.24% 72.78%(1)
BERT+MFCC	76.45% 75.49%	**78.48%** **77.09%**	78.48% 77.02%	77.47% 76.32%(3,2)	78.12% 76.77%(1,1)
BERT+VIT+MFCC	78.21% 76.39%	79.22% 78.01%	78.66% 77.49%	77.65% 76.52%(3,2,2,)	**79.78%** **78.56%(1,1)**

Questions?



References

1. HateMM: A Multi-Modal Dataset for Hate Video Classification
2. https://alphacephei.com/vosk/
3.