The 2022 IARIA Annual Congress on Frontiers in Science, Technology, Services, and Applications
IARIA Congress 2022
July 24, 2022 to July 28, 2022 - Nice, Saint-Laurent-du-Var, France

Multimodal Emotion Recognition Using Speech and Text

Clement H.C. Leung | Chinese University of Hong Kong (Shenzhen), clementleung@cuhk.edu.cn

James J. Deng | MindSense Technologies

香港中文大學(深圳)
The Chinese University of Hong Kong, Shenzhen

Clement LEUNG

- FULL PROFESSORSHIPS at
  - University of London, UK; National University of Singapore; Chinese University of Hong Kong, Shenzhen, China; Hong Kong Baptist University; Victoria University, Australia
- Two US patents, five books and over 150 research articles
- Program Chair, Keynote Speaker, Panel Expert of major International Conferences
- Editorial Board of ten International Journals
- Listed in Who's Who in the World and Great Minds of the 21st Century
- Fellow of the British Computer Society, Fellow of the Royal Society of Arts, Chartered Engineer

Outline

1. **Introduction**
2. **Related Works**
3. **Emotion Representation**
4. **Multimodal Emotion Recognition**
5. **Experiments**
6. **Conclusion**

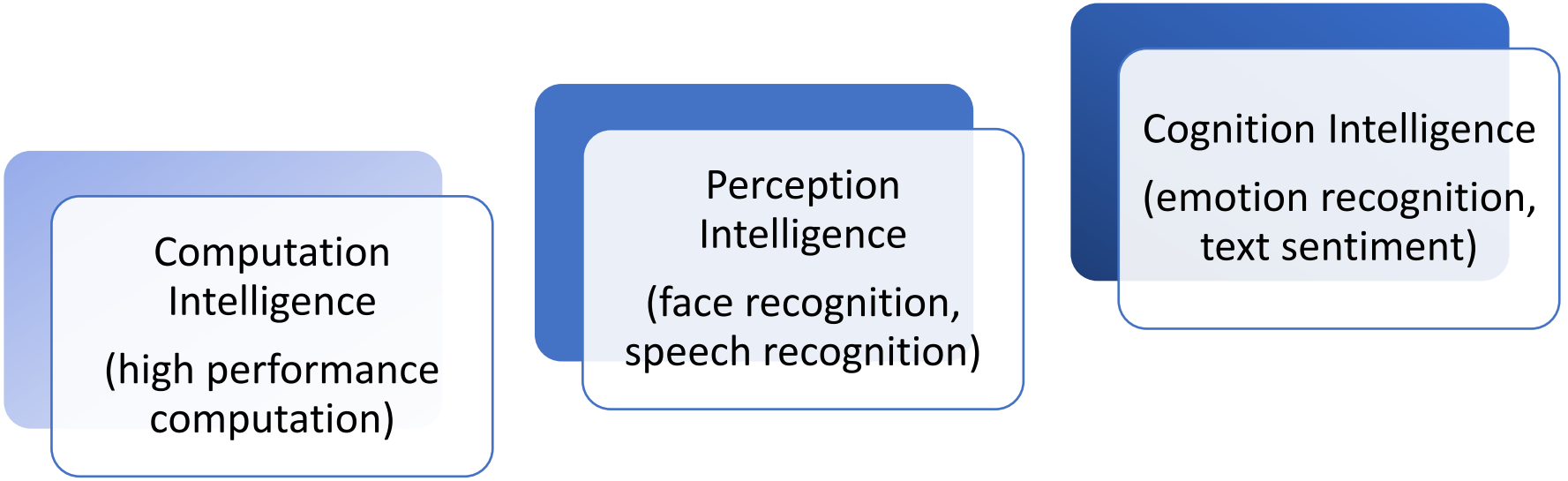
SENTIMENT ANALYSIS

Discovering people opinions, emotions and feelings about a product or service

Introduction

- Speech and Text convey rich emotional information
- Single Modality is insufficient and incomplete to recognize human's emotion
- Voice and Text extracted from speech are associated
- Emotion AI Pipeline:
  - Emotion Representation
  - Multimdoal Emotion Model Learning
  - Downstream Tasks (Emotion Detection, Emotion Recognition, Emotion Synthetize)

Different Forms of Intelligence

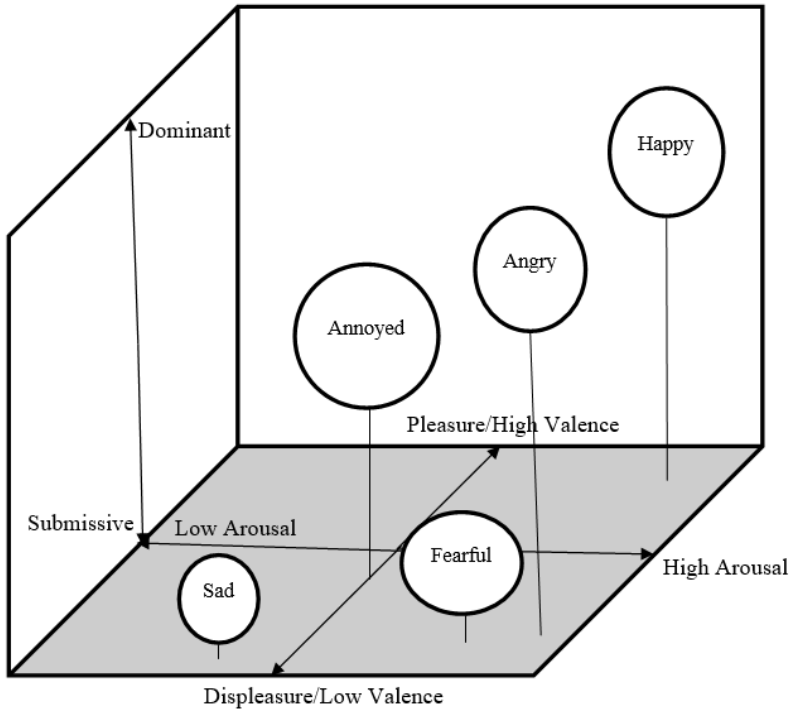


Motivation

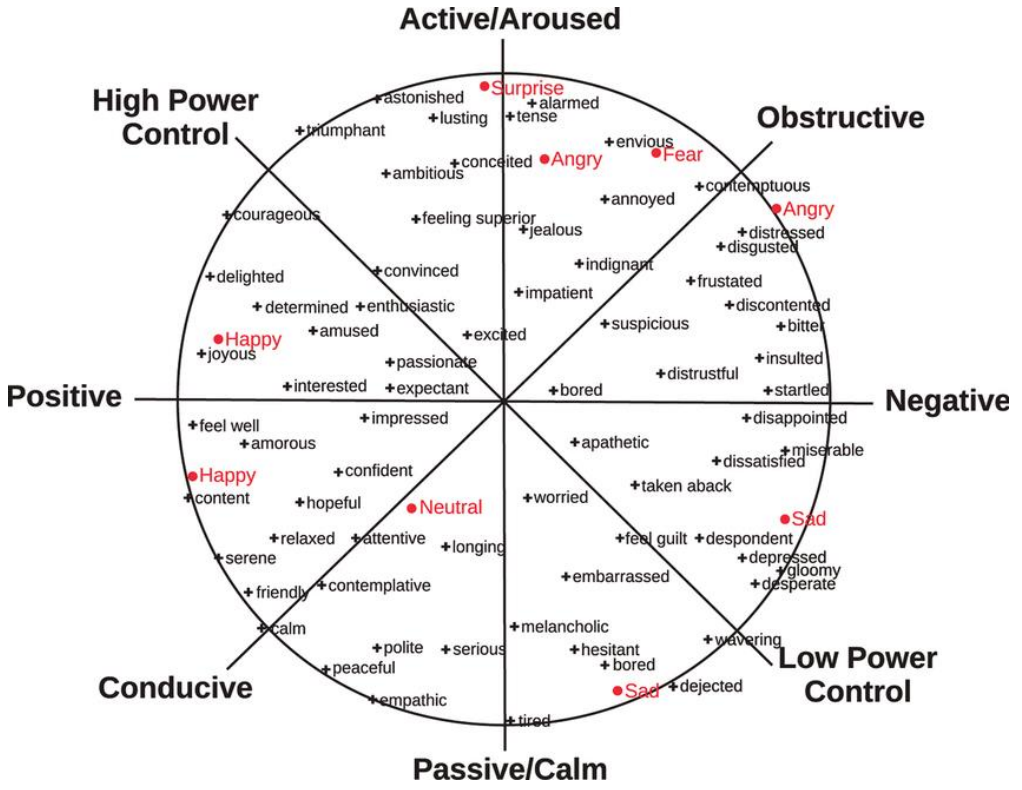
- **Single Modality** for Emotion Recognition suffers robustness problem. (Yin-Yang Theory)
- **Big models** are popular in industry, but training big models using mega multimedia dataset consumes huge resource and prohibitively expensive (e.g., GPT-3 costs $4.6 million)
- **Transfer learning** provides powerful reusable techniques (VGGish, Yamnet, BERT, etc.)
- How to construct an effective **multimodal** emotion model
- Transformer model has acheived great success in signal speech and NLP modality tasks

Related Works

- **Emotion Theory**
  - Discrete Emotion (Six basic emotions, or Nine basic emotions, OCC model)
  - Dimensional Emotion (Arousal-Valence, Pleasure-Arousal-Dominance, Circumplex)



Burden, David, and Maggi Savin-Baden. *Virtual humans: Today and tomorrow*. Chapman and Hall/CRC, 2019.



Scherer's circumplex model (Scherer, 2005)

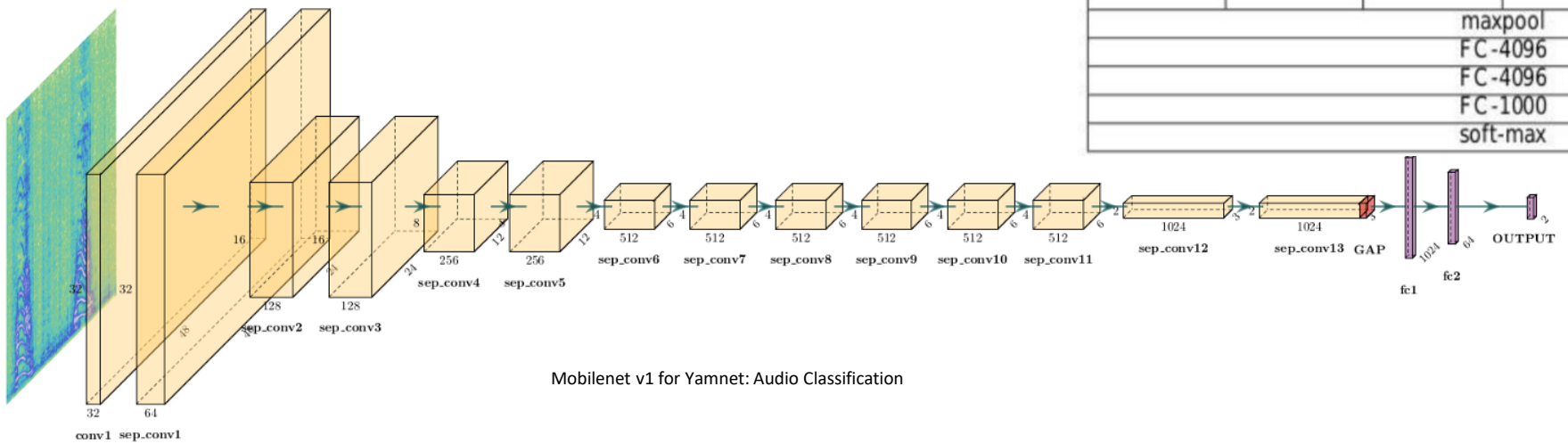
Related Works

- Emotion Recognition from single modality (e.g., Speech, Text)
  - Speech Emotion Recognition (CNN, LSTM, CNN-LSTM, BERT, etc.)
  - Text Sentiment (LSTM, BERT, universal language model fine-tuning, etc.)

VGG-19 Network

Figure from 'Very Deep Convolutional Networks For Large-Scale Image Recognition"

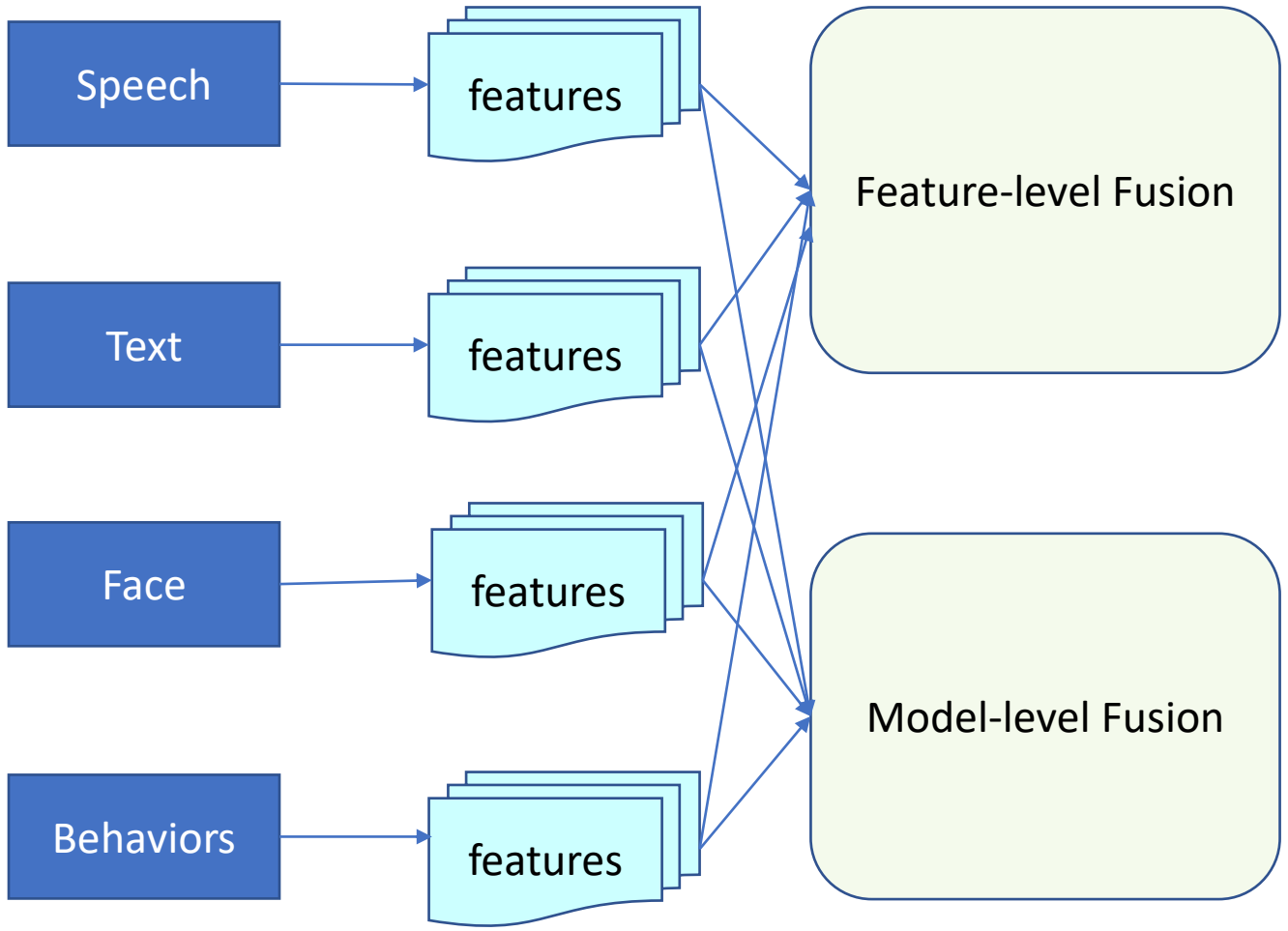
ConvNet Configuration					
A	A-LRN	B	C	D	E
11 weight layers	11 weight layers	13 weight layers	16 weight layers	16 weight layers	19 weight layers
input (224×224 RGB image)					
conv3-64	conv3-64 LRN	conv3-64 conv3-64	conv3-64 conv3-64	conv3-64 conv3-64	conv3-64 conv3-64
maxpool					
conv3-128	conv3-128	conv3-128 conv3-128	conv3-128 conv3-128	conv3-128 conv3-128	conv3-128 conv3-128
maxpool					
conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256 conv1-256	conv3-256 conv3-256	conv3-256 conv3-256 conv3-256 conv3-256
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 conv1-512	conv3-512 conv3-512	conv3-512 conv3-512 conv3-512 conv3-512
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 conv1-512	conv3-512 conv3-512	conv3-512 conv3-512 conv3-512 conv3-512
maxpool					
FC-4096					
FC-4096					
FC-1000					
soft-max					



Mobilenet v1 for Yamnet: Audio Classification

Multimodal Information Modelling

- Joint Representation
- Multimodal fusion
- Co-learning
- Generative adversarial network



Feature Extraction from Speech Audio

To consider the temporal characteristic, a large speech collection is represented by

$$X = \{x_1, x_2, x_3, \ldots, x_N\},$$

where $x_i \in F \times T, F \text{ and } T \text{ denote for dimensionality of spectrogram}$

The goal is to learn an embedding $g := r^{F \times T} \rightarrow r^d$, such that

$$\left|\left|\boldsymbol{g}(x_i) - \boldsymbol{g}(x_j)\right|\right| <= ||\boldsymbol{g}(x_i) - \boldsymbol{g}(x_k)||, \text{ when } |i - j| <= |i - k|$$

We learn a triplet loss function

$$\Theta(z) = \sum_{i=1}^{N} [\|g(x_i) - g(x_j)\|_2^2 + \|g(x_i) - g(x_k)\|_2^2 + \delta]$$

Where $\delta \text{ is non} - \text{negative margin hyperparameter}$

Text Embedding Extraction

Given a sequence of text extracted from speech

(1) Map the tokens of the initial input sequence to an embedding space

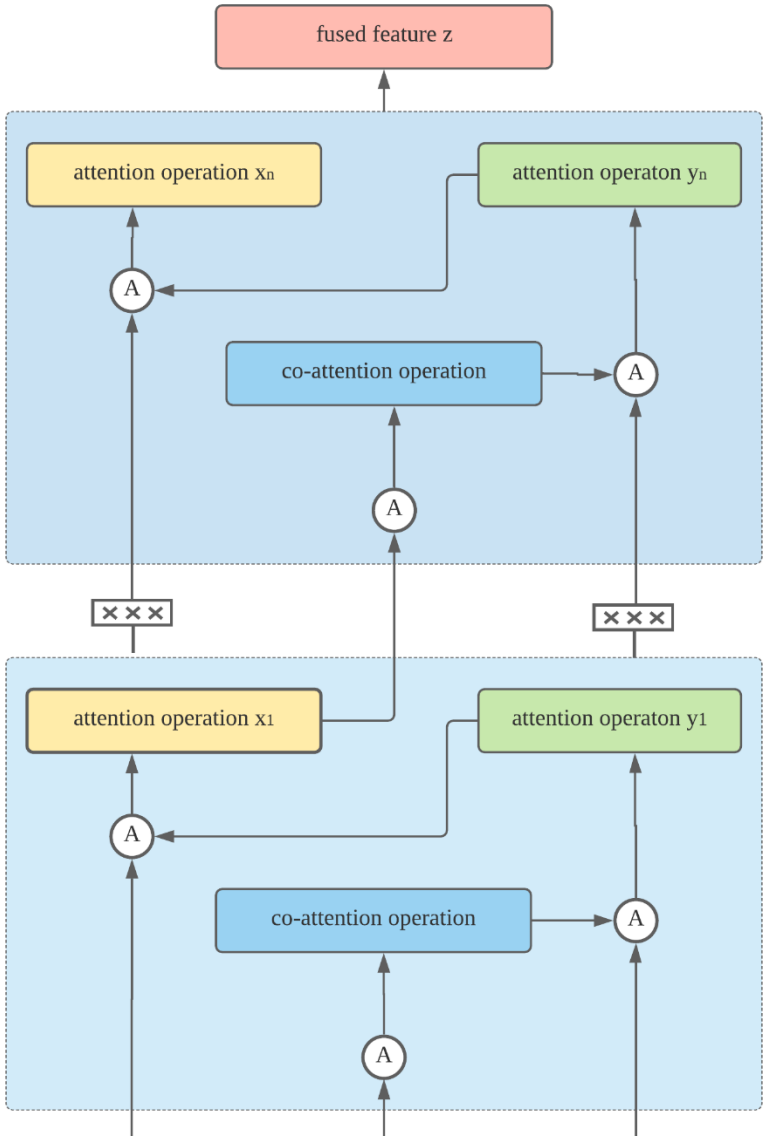
(2) Input the embedded sequence to the encoder layer

(a) The encoder layer is composed of a stack of blocks

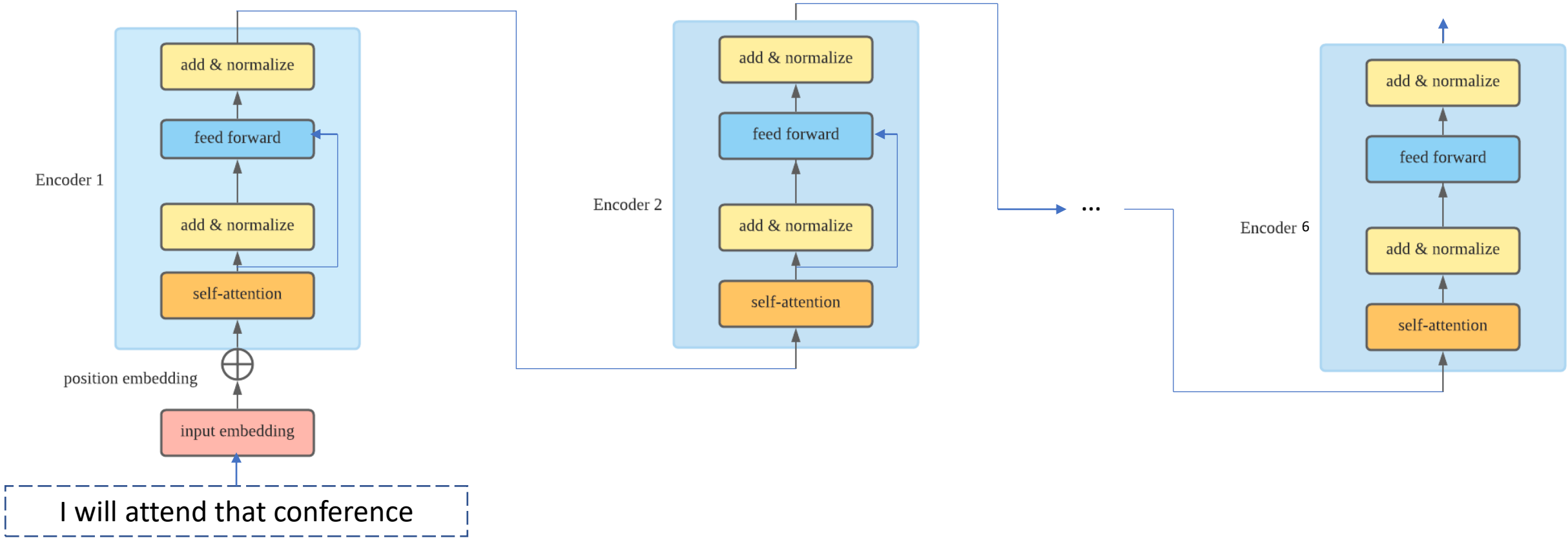
(b) Each block contains self-attention followed by feed-forward

(c) Residual skip from self-attention layer and feed-forward

(d) Dropout within the feed-forward network

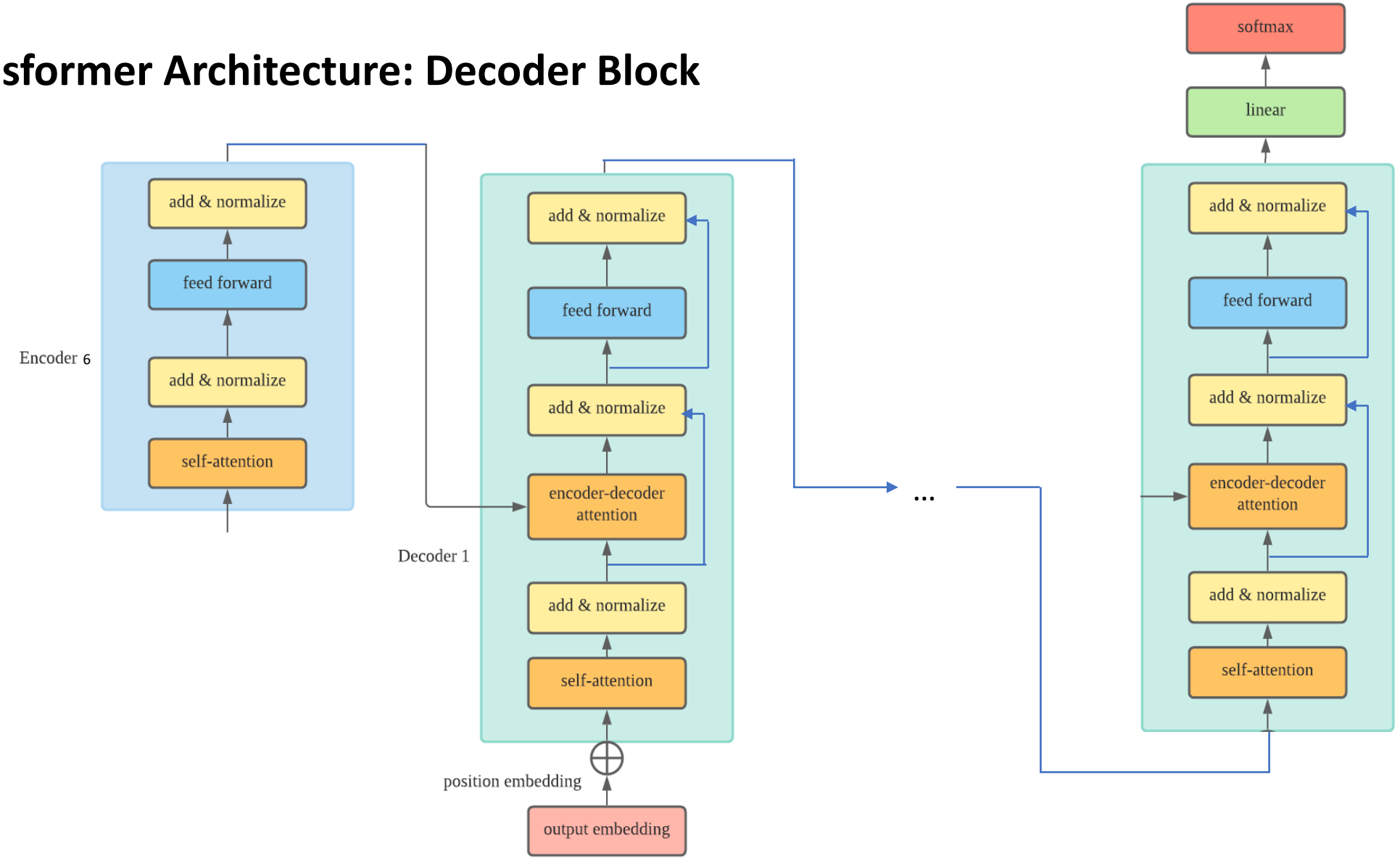


Transformer Architecture: Encoder Block



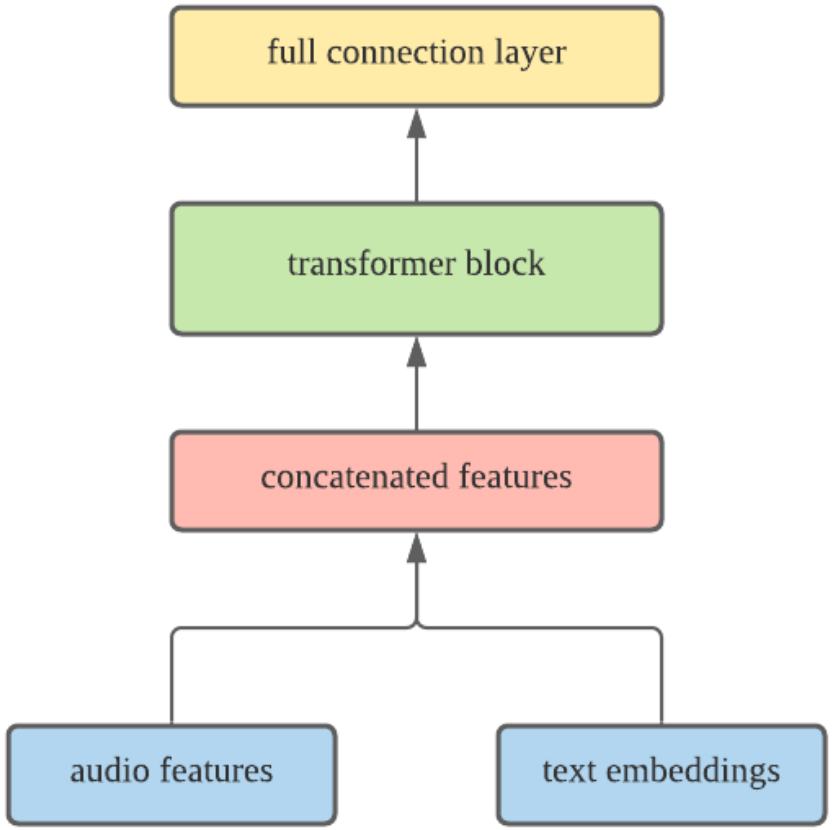
- The encoder block uses the self-attention mechanism
- Residual connections is added, and layer normalization LayerNorm(x + Sublayer(x))
- Encoder is stacked, and the output of last encoder block is the input of decoder

Transformer Architecture: Decoder Block

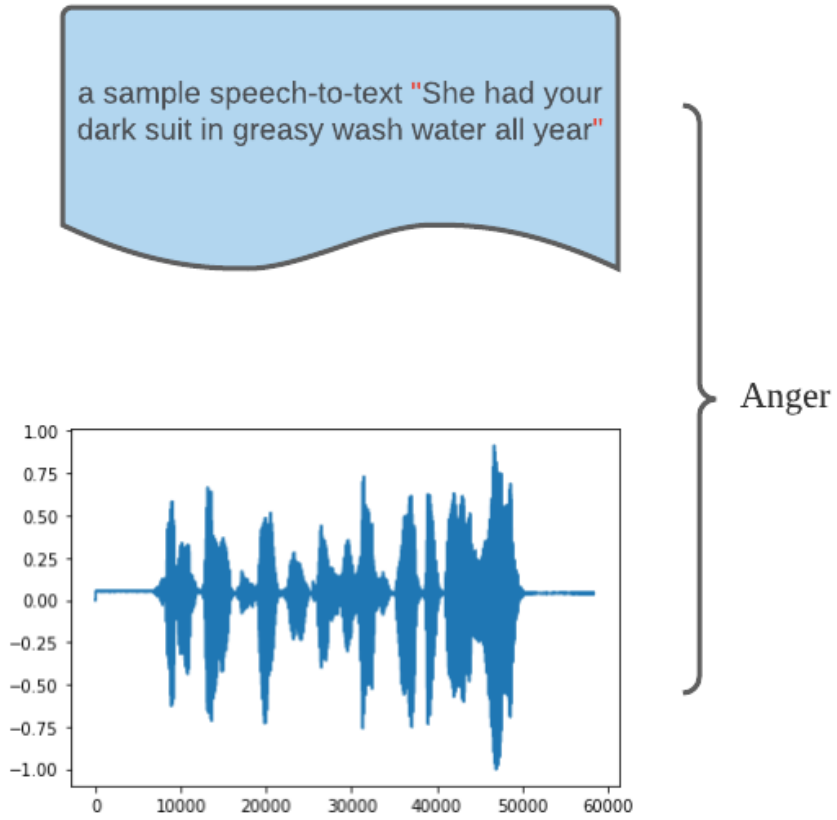


- The decoder block uses encoder-decoder attention
- The embeddings use features from input and partial output

Joint Emotion Representation



Multimodal fusion of speech audio and text on embeddings extracted from transfer learning



Sample speech with anger emotion

Joint Emotion Representation from Transformer

- Speech audio and speech contents would have some extent of correlation, and both exert an effect on emotion recognition.
- Co-attention mechanism:

  Given speech audio features X and text embedding Y,

  input X and Y to a transformer network with deep co-attention

  The co-attention operations are represented by

$$\hat{x} = \Lambda(X; g_x)$$
$$\hat{y} = \Lambda(Y; g_y)$$

where attention guidance g_x is derived from speech and g_y from text

Joint Emotion Representation from Transformer

The detailed computations are:

$$
\begin{aligned}
H &= tanh(W_x X + (W_{g_x} g_x) V^T) \\
\hat{a} &= softmax(w_{hx}^T H) \\
\hat{x} &= \sum a_i^x x_i
\end{aligned}
\tag{3}
$$

where V is a vector with all elements equalling one. W_x and W_{g_x} denotes for $k \times d$ matrix parameter, and w_{hx} refers to k dimensional vector parameter. a^x is the attention weight of speech feature X. The computations of $\hat{y}$ follows the same process in Equation 3. At the first step of alternating co-attention, g_x is

Joint Emotion Representation from Transformer

same process in Equation 3. At the first step of alternating co-attention, g_x is 0. At the second step, g_y is intermediate attended text embedding from the first step. At last, we use the speech feature $\hat{x}$ as the guidance to attend the text again. We use a linear function to fuse attended features $\hat{x}$ and $\hat{y}$. The fused feature z is represented by

$$z = LayerNorm(W_x^T \hat{x} + W_y^T \hat{y}) \tag{4}$$

Finally, the binary cross-entropy is used as loss function to train a classifier.

Experimental Setup

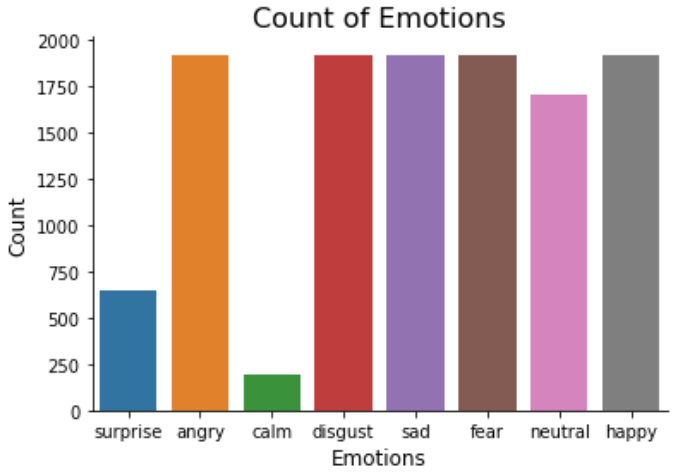
- **Common Emotion Dataset**

  Emotional Dyadic Motion Capture (IEMOCAP)

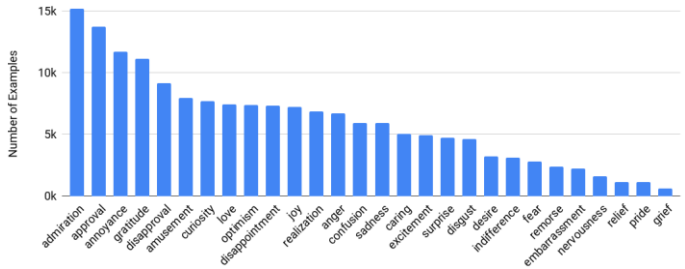
  SAVEE: 7 emotion categories: anger, disgust, fear, happiness, neutral, sadness, and surprise

  Colossal Clean Crawled Corpus (C4)

  GoEmotions: 58k English Reddit comments, labeled for 27 emotion categories or Neutral



SAVEE Dataset



GoEmotions Dateset

Emotion	Number of Samples	Rate
Anger	1229	12.24%
Sadness	1182	11.78%
Happiness	495	4.93%
Neutral	575	5.73%
Excited	2505	24.96%
Surprise	24	0.24%
Fear	135	1.34%
Disgust	4	0.03%
Frustration	3830	38.16%
Other	59	0.59%
Total	**10,038**	**100%**

IEMOCAP Dataset

Dataset	# documents	# tokens	size
C4.EN.NOCLEAN	1.1 billion	1.4 trillion	2.3 TB
C4.EN.NOBLOCKLIST	395 million	198 billion	380 GB
C4.EN	365 million	156 billion	305 GB

C4 Dataset

Experimental Setup

- **Pretrained models**

  VGGish is audio embedding

  Yamnet employs Mobilenet v1 depthwise-separable convolution

  Text-to-text transfer transformer (T5 model)

  Bert_base_go_emotion

  Distilbert-based-uncased-go-emotion

- **Fine-tune models**

  The parameters of the final layer are updated

  After training for a certain updates, the second-to-last layer are contained

  Until the entire network's parameters are fine-tuned

Experimental Results

Table 1. Comparison of multimodal fusion by learning a joint emotion representation performance with single modality through the different embeddings on different emotion datasets.

Models	IEMOCAP	SAVEE	Mean
VGGis FC1	65.3%	57.7%	61.5%
VGGish Finetuned	61.4%	59.3%	60.4%
YAMNet layer 10	63.2%	62.3%	62.8%
YAMNet Finetuned	67.6%	62.7%	65.2%
TRILL distilled	70.5%	67.8%	69.2%
TRILL Finetuned	73.8%	68.6%	71.2%
Text-to-Text Transformer (T5)	75.7%	72.3%	75.5%
TRILL-T5 Multimodal Fusion	81.7%	75.9%	78.8%

Experimental Results

Table 2. Average performance of the different emotion representations on four selected emotion categories.

Models + Dataset	happy	anger	sad	natural	Mean
TRILL (IEMOCAP_Audio)	77.2%	81.9%	72.3%	66.8%	74.6%
TRILL (SAVEE_Audio)	73.3%	84.3%	73.3%	66.7%	74.4%
T5 (IEMOCAP_Text)	79.6%	82.7%	72.2%	64.9%	74.9%
T5 (SAVEE_Text)	75.0%	81.7%	71.7%	66.7%	73.8%
Multimodal Fusion (IEMOCAP)	83.1%	84.6%	76.3%	71.1%	78.9%
Multimodal Fusion (SAVEE)	84.7%	85.2%	74.5%	70.3%	78.7%

Experimental Results

Dataset	happy	anger	sad	natural
IEMOCAP&SAVEE	84.2%	86.0%	77.9%	71.8%
RAVDESS	82.2%	87.8%	77.8%	76.1%
GoEmotions	85.4%	88.5%	79.7%	65.3%

Evaluation of robustness of our multimodal emotion recognition method on different datasets

Conclusion

- Multimodal emotion recognition is highly applicable to different types of industry
- Transfer learning provides powerful reusable techniques
- Transformer with co-attention achieved good results on learning a joint emotion representation
- Experiments show that the transformer mechanism on bimodal helps to fuse important information and increase emotion recognition performance
- Multimodal emotion recognition still needs to make breakthrough in industry

Questions

THANK YOU

WISH YOU HAVE GOOD

EMOTIONS