



Multimodal Representation Learning for Speech Emotion Recognition

23 Nov 2023

GP Huang

guangpu.huang@aalto.fi




Abstract

Human communication relies on multimodal representations: speech, language, facial expressions, gestures, etc. Could we learn something fundamental from these representations? Is it possible that multimodal representation learning could provide us some answers as to how to optimize the performance of machine learning models, and ultimately achieve the human-level performances on ASR, NLP and CV tasks? Transformers might just seem to be the right candidates. With the increase amount of multimodal data online, transformer-based methods have increasingly shown great potentials in multimodal machine learning. However, there are many unanswered questions as well as practical challenges, especially when deploying 'in the wild'.

In this talk we will review multimodal representation learning using transformers. We will design experimental studies on speech emotion recognition (alternatively facial expression recognition or affect recognition) to jointly study the modalities: speech, text, and video/image capture of human face (and hand gestures if possible). More importantly, we will raise some key research questions in multimodal representation learning. We will also discuss open challenges, potential research directions, and collaborations among ASR, NLP, CV research teams.



1. Why multimodal representation learning is important and how to approach it?
2. Why transformer architecture is good for this?
3. Why multimodal SER is a good topic to start your experimental study?

- 
- 1. Multimodal Representation Learning**
 2. Speech Emotion Recognition



Motivation

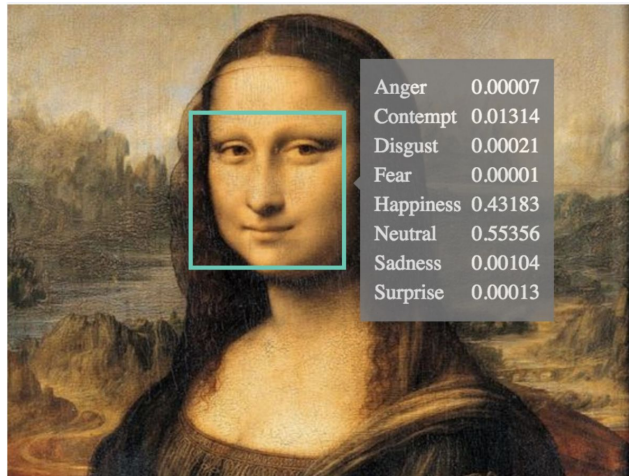
- Human: communications, perceptions, behaviours, emotions, events, actions, humour
 - ◆ Seeing, speaking, hearing, touching, smelling...
 - ◆ Identity?
- AI models
 - ◆ Audio, visual, language processing
 - ◆ Shared representations?

A. A. Lazarus et al., Multimodal behavior therapy. Springer, 1976.

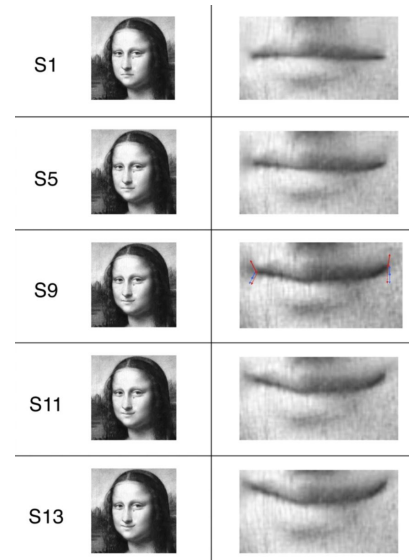
T. Baltrušaitis, C. Ahuja, and L.-P. Morency, “Multimodal machine learning: A survey and taxonomy,” TPAMI, 2018.

Facial Expressions

“Mona Lisa is always happy – and only sometimes sad”



“Mona Lisa” by Leonardo da Vinci (around 1503–1516)



<https://www.nature.com/articles/srep43511>

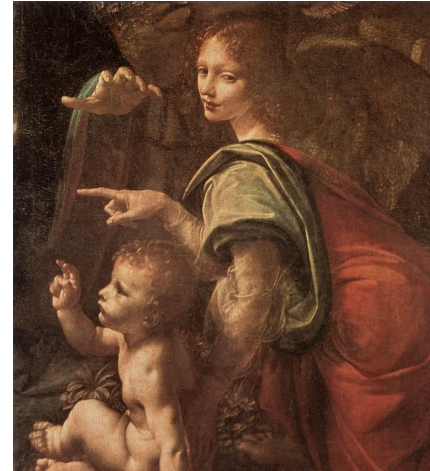
Hand Gestures



Raphael's depiction of [Plato](#) in his 1509–1511 fresco [The School of Athens](#) in the Vatican is believed to be a portrait of Leonardo da Vinci.



Detail of the painting Saint John the Baptist (1513-16)



Detail of the Virgin of the Rocks (1483-86) painting (Paris version), depicting an angel pointing to baby Jesus (Image: public domain)



Affect Computing





The study and development of systems and devices that can recognize, interpret, process, and simulate human [affects](#). ..

“(the research goal)...to give machines emotional intelligence, including to [simulate empathy](#). The machine should interpret the emotional state of humans and adapt its behavior to them, giving an appropriate response to those emotions.”

Hogg, M.A., Abrams, D., & Martin, G.N. (2010). Social cognition and attitudes. In Martin, G.N., Carlson, N.R., Buskist, W., (Ed.), *Psychology* (pp 646-677). Harlow: Pearson Education Limited.

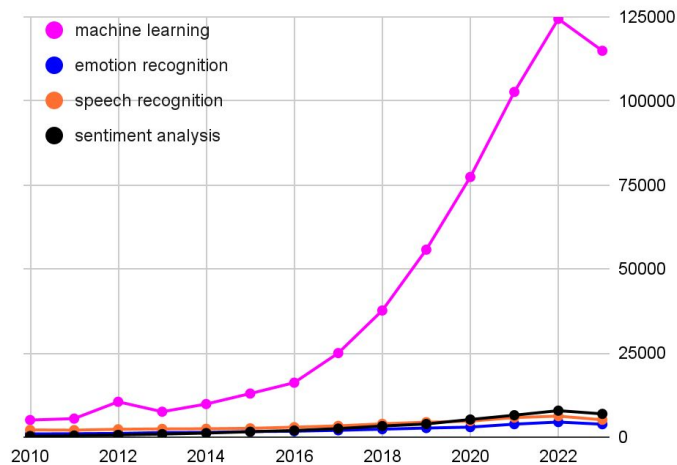
Tao, Jianhua; Tieniu Tan (2005). "Affective Computing: A Review". *Affective Computing and Intelligent Interaction*. Vol. [LNCS 3784](#). Springer. pp. 981–995.

1. Motivation
2. **Multimodal Machine Learning**
3. Transformers-based Multimodal Representation Learning for SER

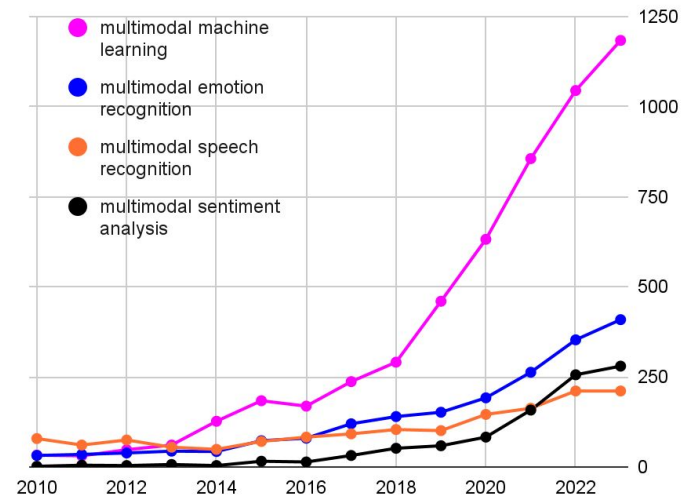
Modality	Example data
Text	 <p>Product review</p>
Speech	 <p>Speech</p>
Visual	 <p>Image</p>
Multimodal	 <p>vlogs</p>

Multimodal Machine Learning

Publications per Year

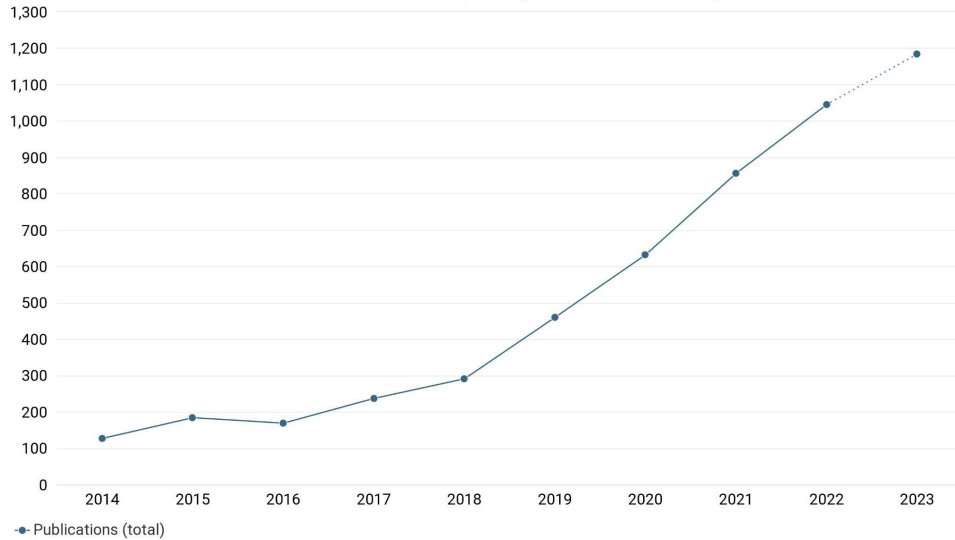


Publications per Year

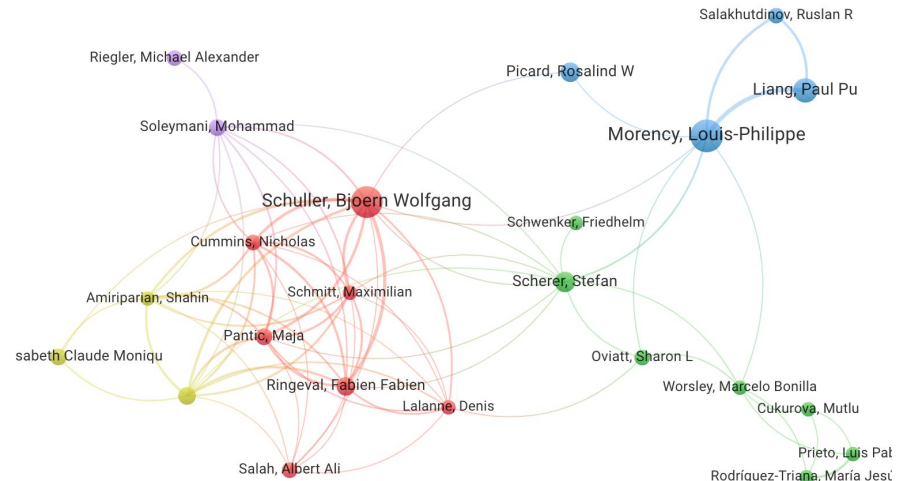


Multimodal Machine Learning

Publications in each year. (Criteria: see below)

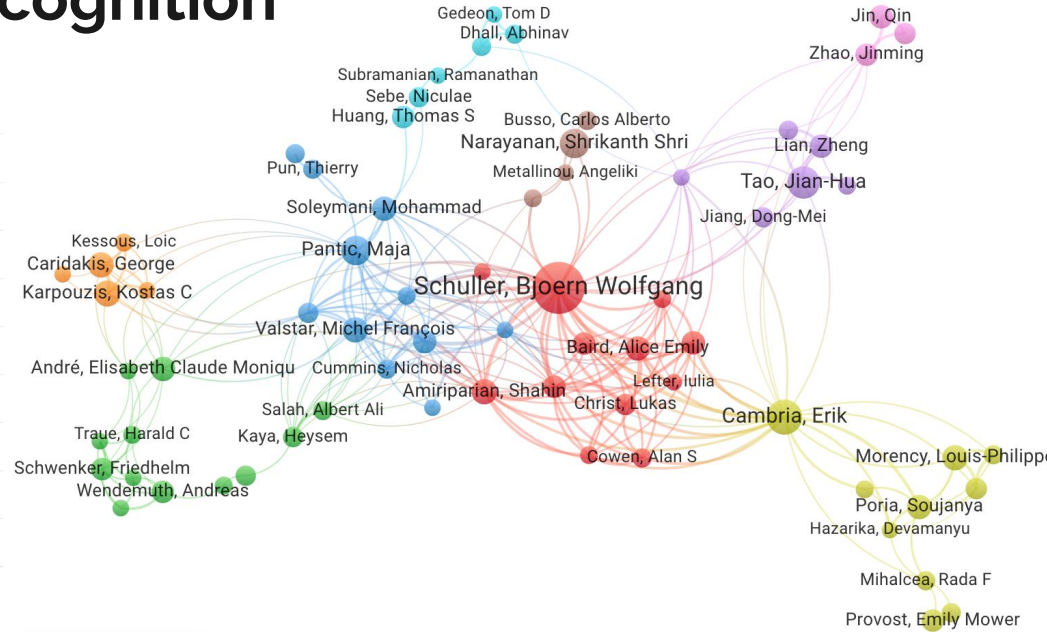
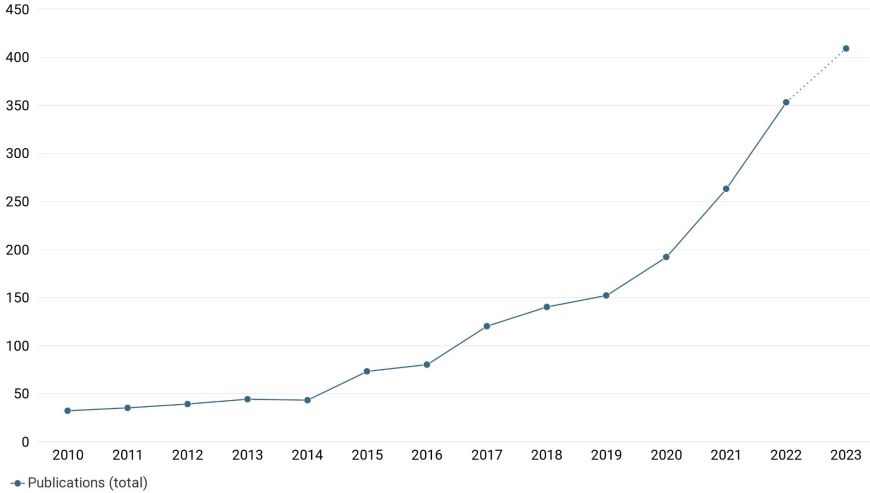


Source: <https://app.dimensions.ai>
 Exported: November 13, 2023
 Criteria: "multimodal machine learning" in title and abstract

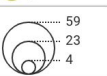


Multimodal Emotion Recognition

Publications in each year. (Criteria: see below)



Source: <https://app.dimensions.ai>
Exported: November 13, 2023
Criteria: "multimodal emotion recognition" in title and abstract





Multimodal Machine Learning

→ Areas

- ◆ [Multimodal Representations](#)
- ◆ [Multimodal Fusion](#)
- ◆ [Multimodal Alignment](#)
- ◆ [Multimodal Pretraining](#)
- ◆ [Multimodal Translation](#)
- ◆ [Crossmodal Retrieval](#)
- ◆ [Multimodal Co-learning](#)
- ◆ [Missing or Imperfect Modalities](#)
- ◆ [Analysis of Multimodal Models](#)
- ◆ [Knowledge Graphs and Knowledge Bases](#)
- ◆ [Intepretable Learning](#)
- ◆ [Generative Learning](#)
- ◆ [Semi-supervised Learning](#)
- ◆ [Self-supervised Learning](#)
- ◆ [Language Models](#)
- ◆ [Adversarial Attacks](#)
- ◆ [Few-Shot Learning](#)
- ◆ [Bias and Fairness](#)
- ◆ [Human in the Loop Learning](#)

→ Applications

- ◆ [Language and Visual QA](#)
- ◆ [Language Grounding in Vision](#)
- ◆ [Language Grounding in Navigation](#)
- ◆ [Multimodal Machine Translation](#)
- ◆ [Multi-agent Communication](#)
- ◆ [Commonsense Reasoning](#)
- ◆ [Multimodal Reinforcement Learning](#)
- ◆ [Multimodal Dialog](#)
- ◆ [Language and Audio](#)
- ◆ [Audio and Visual](#)
- ◆ ...
- ◆ [Affect Recognition](#)
- ◆ [Healthcare](#)
- ◆ [Robotics](#)
- ◆ [Autonomous Driving](#)
- ◆ [Finance](#)
- ◆ [Human AI Interaction](#)

→ Architectures

- ◆ [Multimodal Transformers](#)
- ◆ [Multimodal Memory](#)

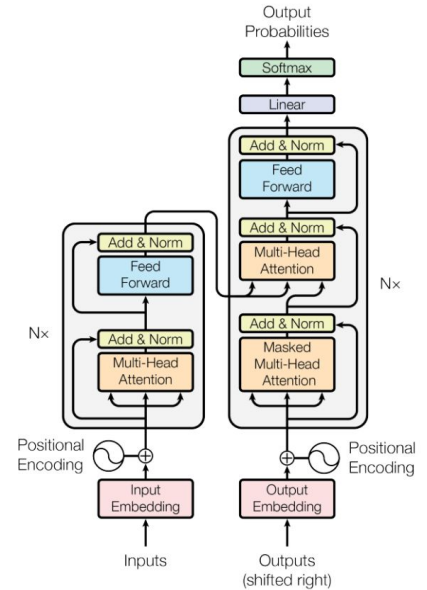


Research Directions

- Universal representation learning
 - ◆ Modality-agnostic
 - ◆ Task-generic
- Understand interactions across different modalities
 - ◆ Latent semantic alignments across modalities
 - ◆ Measuring the fusion between modalities
 - ◆ Modality dropout?
- Identify/utilize the strengths of Transformers
 - ◆ Multimodal inputs
 - ◆ Multimodal Pretraining

Transformers

- Availability of multimodal big data
- Modality-agnostic/generic
 - ◆ Tokenization
 - ◆ Embedding processing
 - ◆ Self-attention mechanism
- Flexibility of architecture design



A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in NeurIPS, 2017.



Transformers

Unimodal

- A. Input
 - a. Tokenization
 - b. Embedding Processing
- B. Self-Attention
 - a. Scaled Dot Product
 - b. Masked Self-Attention
 - c. Multi-Head Self-Attention
 - d. ...
- C. Feed-Forward Network

Multimodal

- A. Input
 - a. Tokenization
 - b. Embedding Processing
 - c. *Token Embedding Fusion*
- B. Self-Attention Variants
 - a. Early Summation
 - b. Early Concatenation
 - c. Hierarchical
 - d. Cross-Attention
 - e. ...
- C. Network Architectures
 - a. *Single-stream*
 - b. *Multi-stream*
 - c. *Hybrid-stream*



Strengths

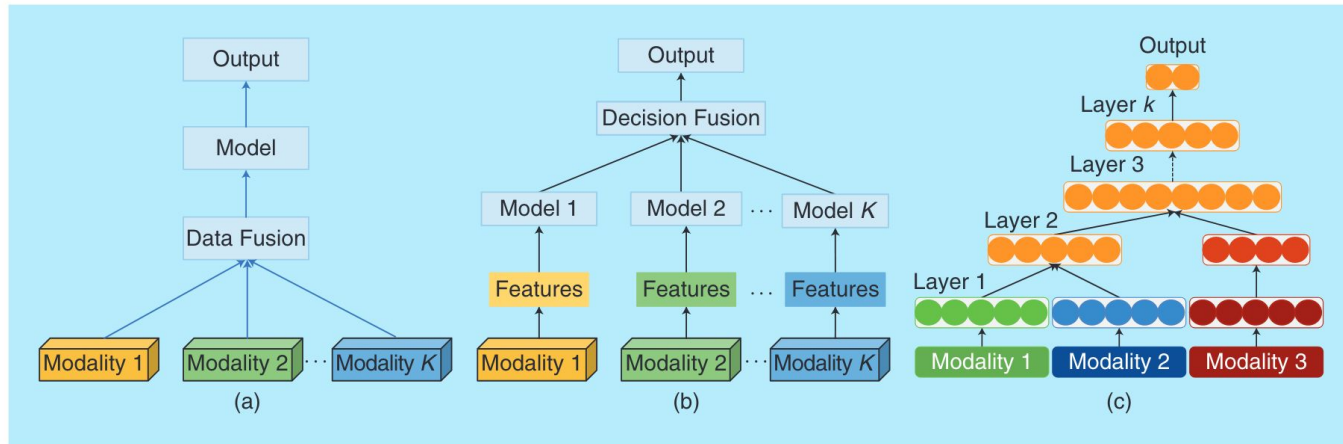
- Encode implicit knowledge
- Self-attention is able to model the embedding of arbitrary tokens from an arbitrary modality
- Multi-head brings multiple modelling sub-spaces
- Global aggregation that perceives the non-local patterns
- Handle the domain gaps and shifts (e.g., linguistic and visual) via pretraining
- Training and inference efficiency via parallel computation
- Tokenization makes Transformers flexible to organize multimodal inputs



Challenges

- **Fusion:** when, how to join?
- **Alignment:** how to relate the (sub)elements e.g. image, text, speech, audio, video?
- **Robustness:** are experimental evaluations enough?
- **Interpretability:** why and how Transformers perform well in multimodal learning?
- **Transferability:** how to transfer models across different datasets and applications?
- **Efficiency:** how to deal with huge model parameters and computational cost?
- **Universality:** towards unified pipelines?
- **Co-Learning:** how to aid resource-poor modality with resource-rich ones?

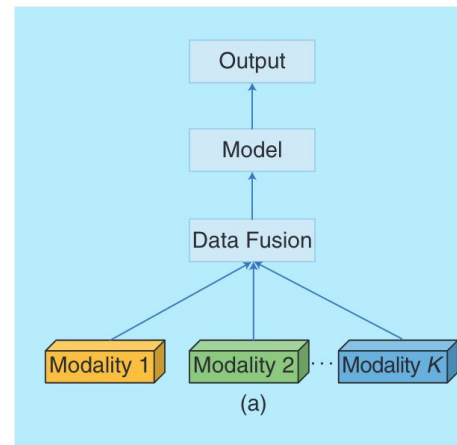
Fusion



(a) Early or feature/data-level fusion, (b) late or decision-level fusion, and (c) intermediate fusion.

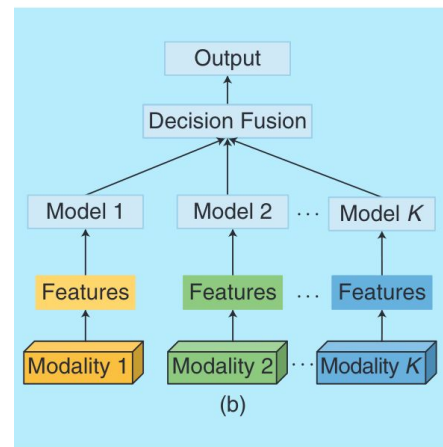
Early Fusion

- Integrate multiple sources of data from sensors into a single feature vector at input
 - ◆ Raw or preprocessed data: resampling at a common rate
 - ◆ Handcrafted features
 - ◆ Learned representations
- Pros
 - ◆ Simple to understand
 - ◆ Simple to implement
- Cons
 - ◆ May have ignore the correlations between modalities
 - ◆ May not fully exploit the complementary nature of the modalities
 - ◆ May result in very large & redundant feature vectors
 - To apply dimensionality reduction techniques e.g. PCA
 - To learn a common multimodal embedded space e.g. autoencoders



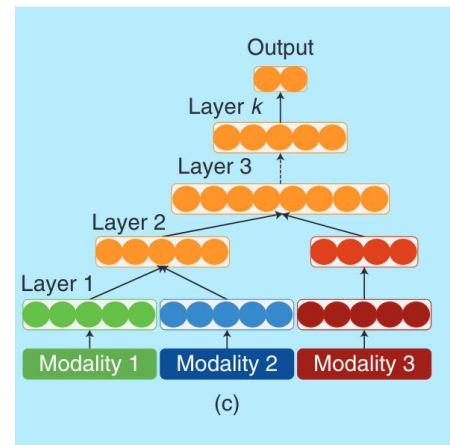
Late Fusion

- Aggregate decisions from multiple uni-modal classifiers
 - ◆ E.g. max-fusion, averaged-fusion, Bayes' rule etc.
 - ◆ Ensemble classifiers
- Pros
 - ◆ Could be easy to implement
 - ◆ Feature independent
 - ◆ Errors from multiple classifiers tend to be uncorrelated
- Cons
 - ◆ No conclusive evidence of superiority over early fusion, often task-dependant
 - ◆ Not as flexible as intermediate fusion



Intermediate Fusion

- Transform raw inputs and learn a joint multimodal representation
 - ◆ Naïve concatenation
 - ◆ Gradual fusion e.g.
 - Based on correlation of modalities e.g., first visual, then motion, then audio.
 - Slowly/gradually fuse learned reps of video across layers during training
- Pros
 - ◆ Flexible to fuse representations at different depths at different depths
 - ◆ Progressive fusion w.r.t. correlation between modalities to optimize performance
 - ◆ Have shown to be the best in several empirical studies
- Cons
 - ◆ Require careful design in terms of how, when, and which modalities can be fused





Challenges

- **Fusion:** when, how to join?
- **Alignment:** how to relate the (sub)elements e.g. image, text, speech, audio, video?
- **Robustness:** are experimental evaluations enough?
- **Interpretability:** why and how Transformers perform well in multimodal learning?
- **Transferability:** how to transfer models across different datasets and applications?
- **Efficiency:** how to deal with huge model parameters and computational cost?
- **Universality:** towards unified pipelines?
- **Co-Learning:** how to aid resource-poor modality with resource-rich ones?



Alignment

- Audio + Video
 - ◆ Speaker localization in multi-speaker videos
 - ◆ Humor detection: laughter
- Image + Text
 - ◆ Aligning the areas of an image to the caption's words or phrases
 - ◆ Visual grounding of natural language
 - ◆ Visual question answering
- Video + Text:
 - ◆ Aligning a script or book chapters with the movie they were based on
 - ◆ Aligning recipe steps with the corresponding instructional video
 - ◆ Text-to-video retrieval
- Audio + Image + Text
 - ◆ Data2Vec

Baltrušaitis, T., Ahuja, C., & Morency, L.-P. (2019). Multimodal Machine Learning: A Survey and Taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2), 423–443.

<https://doi.org/10.1109/TPAMI.2018.2798607>

Xu, P., Zhu, X., & Clifton, D. A. (2023). *Multimodal Learning with Transformers: A Survey* (arXiv:2206.06488). arXiv. <https://doi.org/10.48550/arXiv.2206.06488>



Alignment

- Map two modalities into *a common representation space* with contrastive learning
 - ◆ Cons: models could be huge and expensive to optimize
- Use pretrained models for tackling downstream tasks via transfer learning
 - ◆ ability of zero-shot transfer e.g for image classification via prompt engineering



Robustness


- “State-of-the-art”
 - ◆ Multimodal Transformers
 - ◆ Pretrained on large-scale corpora
- How to improve robustness?
 - ◆ Augmentation and adversarial learning
 - ◆ Fine-grained loss functions
 - ◆ Contrastive learning
 - ◆ Multi-task learning
- How to evaluate robustness?
 - ◆ Evaluate Transformer components/sublayers
 - ◆ Via experiments across datasets
 - ◆ Via experiments on a common dataset
- Cons
 - ◆ Lacks theoretical tools to analyse the Transformer family



Interpretability

- Use probing task and ablation study to explain why and how models perform well
 - ◆ Vision-language: evaluate what patterns are learned in pretraining
 - ◆ Image-text: examine the optimal combination of pretraining tasks via ablation study, to compare how different pretexts contribute to the Transformers.

J. Cao, Z. Gan, Y. Cheng, L. Yu, Y.-C. Chen, and J. Liu, "Behind the scene: Revealing the secrets of pre-trained vision-and-language models," in ECCV, 2020.
Y.-C. Chen, L. Li, L. Yu, A. El Kholy, F. Ahmed, Z. Gan, Y. Cheng, and J. Liu, "Uniter: Universal image-text representation learning," in ECCV, 2020.

- 
1. Multimodal Representation Learning
 - a. Background
 - b. Transformers
 - c. Research Directions
 2. **Speech Emotion Recognition**



Research Questions

- A. What are the contributions of individual modalities: audio, visual, language?
 - a. Data dependent?
 - b. Task dependent?
- B. How to efficiently integrate audio-visual-language information to improve classification performance? pretrained models? specific tasks?
- C. How is multimodal information processed in emotion recognition? Correlation & causation between cognition and physiological responses.
- D. What are the main shortcomings & challenges in existing Transformer-based multimodal systems?
 - a. In the lab: datasets, benchmarks, reproduce results, compare performances
 - b. In the wild: gestures, multi-parties, social events, background noises etc.
 - c. Computational cost
 - d. Evaluation metrics
- E. How to deal with non-English data e.g. Finnish in semi-supervised classifiers?



Topics

- A. **Multimodal Speech Emotion Recognition with AV-Hubert (Anja, Dejan, Tamas, GP)**
- B. Early Depression Detection and Stress Level Classification via Speech Emotion Recognition
- C. Transformer-based Representation Learning in Human Vision, Speech, Language, and Gesture
- D. Offensive or Humorous? Multimodal Emotion Recognition, Hateful Memes Challenge
- E. Free Speech or Propaganda? Multimodal Sentiment Analysis on X Datasets
- F. Deepfake Emotion Synthesis/Morphing of Speech and Facial Expressions
- G. Deception Detection using Facial Expressions, Speech, and/or Singing. (Yaroslav)
- H. Data2Vec with Finnish Data



Multimodal Speech Emotion Recognition with AV-HuBERT

- Novelty
 - ◆ Fusion: audio, visual, language, (gesture)
 - ◆ Alignment: facial features, speech, emotions
 - ◆ Robustness: annotations, evaluation metrics
- Datasets
 - ◆ Parallel and non-parallel
- Methods
 - ◆ Baselines
 - ◆ State-of-the-art
 - ◆ Transformers



Datasets

→ Criteria

- ◆ Scope (number of speakers, emotional classes, language, etc)
- ◆ Naturalness (acted versus spontaneous)
- ◆ Context (in-isolation versus in-context)
- ◆ Descriptors (linguistic and emotional description)

→ Focus

- ◆ Facial expressions (and possibly hand gestures)
- ◆ Universal categories of human emotions
 - 7 Classes: fear, joy, sadness, anger, neutral, disgust, surprise
 - Valence: positive, negative, neutral

→ Issues to consider

- copyright and privacy: TV shows, films, vlogs, social media
- Recording professional actors under controlled conditions



Datasets

- Interactive Emotional Dyadic Motion Capture ([IEMOCAP](#))
- Lip reading datasets ([LRS3](#))
- Multimodal Emotion Lines Dataset ([MELD](#))
- Ryerson Audio-Visual Database of Emotional Speech and Song ([RAVDESS](#))
- MUSE (?) 2021, 2022, 2023

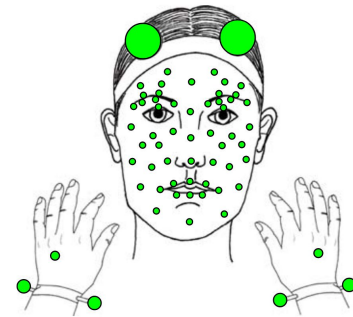
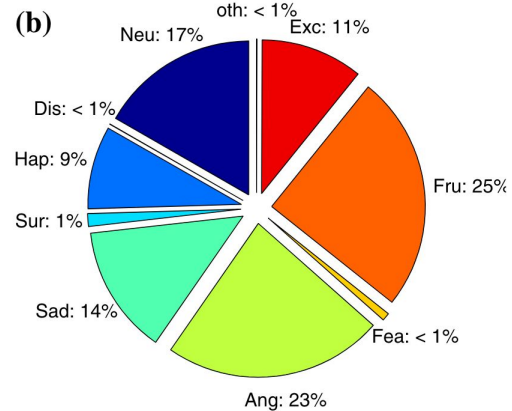
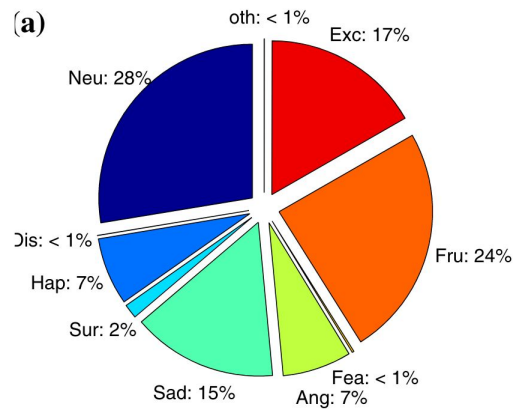


Datasets

Name (a-z)	Year	Data	Annotations	Size	Publication	Licence
IEMOCAP	2004	12 hours 10 speakers	9 emotions 3 attributes	29 GB	PDF	Request Form
LRS3	2017	433 hours 5000 speakers	<i>Largest datasets for lip-reading</i>	136 GB	PDF	Academic
MELD	2019	1400 dialogues	7 emotions 3 sentiments	10.1 GB	PDF	GPL-3.0
RAVDESS	2018	24 actors 2 Statements	2 modes: speech, song 8 emotion 2 intensity	24.8 GB	PDF	Creative Commons Attribution

IEMOCAP

Distribution of the data for each emotional category. (a) scripted sessions, (b) spontaneous sessions.





LRS3

The dataset consists of thousands of spoken sentences from TED and TEDx videos.

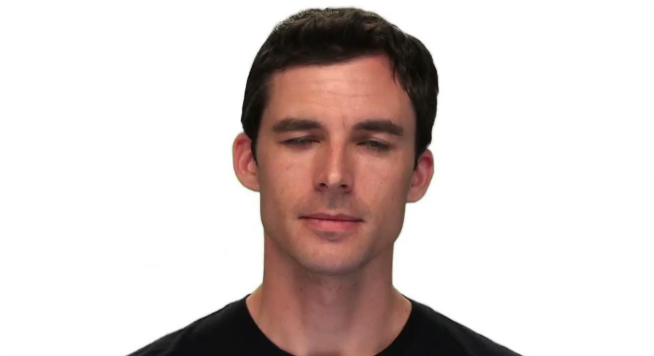
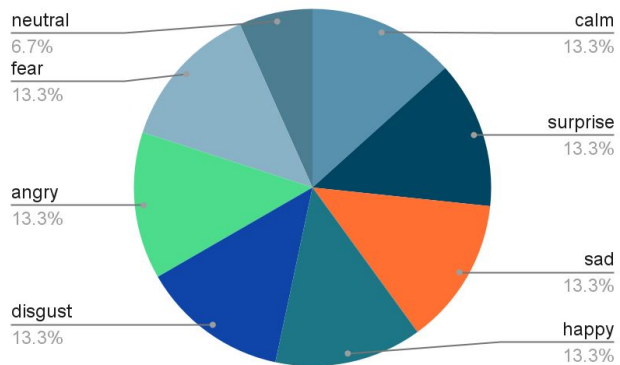
Set	# videos	# utterances	# word instances	Vocab
Pre-train	5,090	118,516	3.9M	51k
Trainval	4,004	31,982	358k	17k
Test	412	1,321	10k	2k

https://mmai.io/datasets/lip_reading/

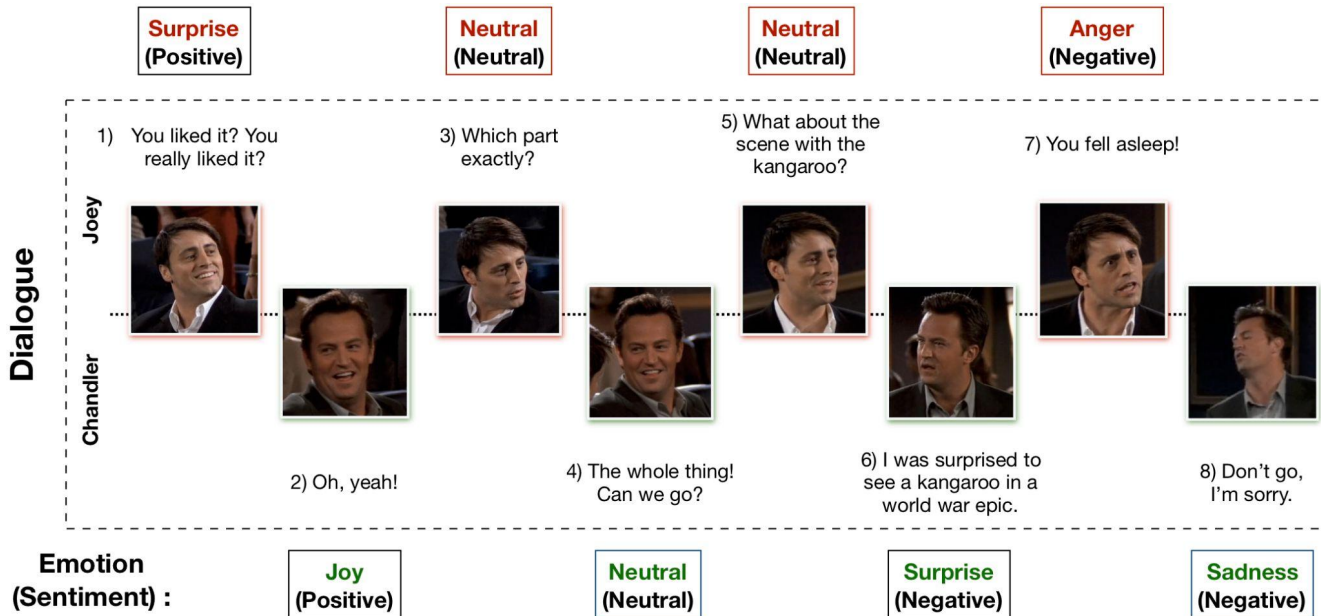


RAVDESS

Speech 1440: male 720, female 720



MELD



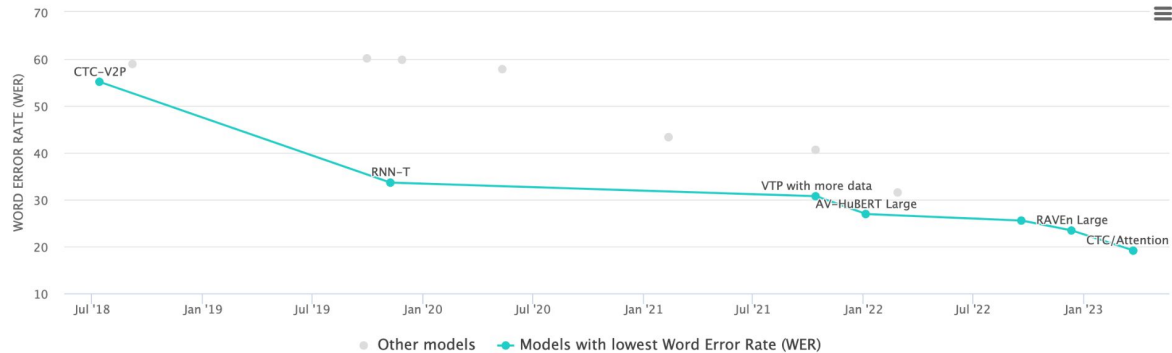


Evaluations

1. Weighted F1
2. Weighted Accuracy (WA), Unweighted Accuracy (UA)
3. Other metrics w.r.t. human emotions / affects

LRS3

[source](#)



Filter: untagged

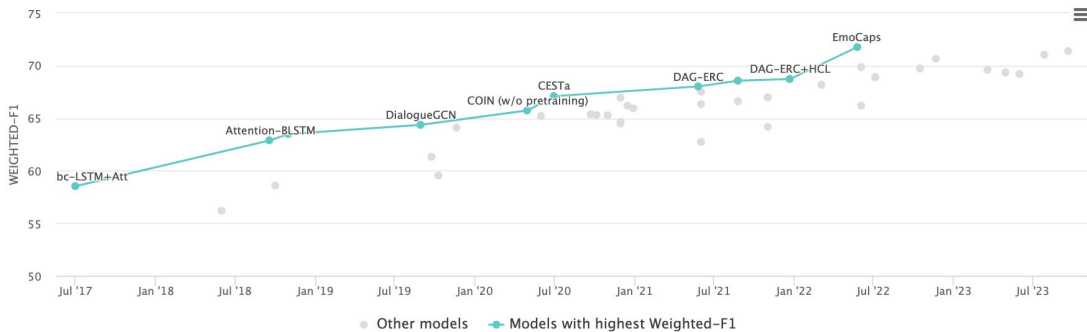
[Edit Leaderboard](#)

Rank	Model	Word Error Rate (WER)	Extra Training Data	Paper	Code	Result	Year	Tags
1	CTC/Attention	19.1	✓	Auto-AVSR: Audio-Visual Speech Recognition with Automatic Labels	GitHub	ASR	2023	
2	RAVEn Large	23.4	✓	Jointly Learning Visual and Auditory Speech Representations from Raw Data	GitHub	ASR	2022	
3	AV-HuBERT Large + Relaxed Attention + LM	25.51	✓	Relaxed Attention for Transformer Models	GitHub	ASR	2022	
4	AV-HuBERT Large	26.9	✓	Learning Audio-Visual Speech Representation by Masked Multimodal Cluster Prediction	GitHub	ASR	2022	



IEMOCAP

[source](#)



Filter: RoBERTa RNN Glove GNN Transformer commonsense BERT XLNet LLM BiLSTM-CRF untagged Edit leaderboard

Rank	Model	Weighted-F1 ↑	Accuracy	Micro-F1	Macro-F1	Extra Training Data	Paper	Code	Result	Year	Tags
1	EmoCaps	71.77				×	EmoCaps: Emotion Capsule based Model for Conversational Emotion Recognition			2022	
2	InstructERC	71.39	71.68			×	InstructERC: Reforming Emotion Recognition in Conversation with a Retrieval Multi-task LLMs Framework			2023	LLM
3	CFN-ESA	71.04	70.78			×	CFN-ESA: A Cross-Modal Fusion Network with Emotion-Shift Awareness for Dialogue Emotion Recognition			2023	RoBERTa
4	UniMSE	70.66	70.56			×	UniMSE: Towards Unified Multimodal Sentiment Analysis and Emotion Recognition			2022	



RAVDESS

[source](#)



Filter: untagged

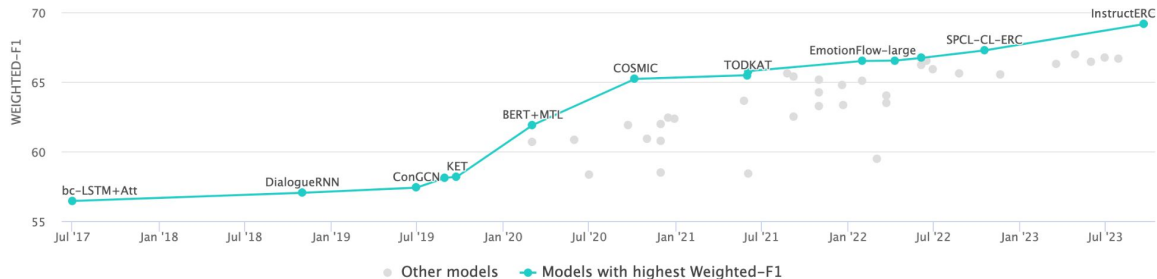
[Edit Leaderboard](#)

Rank	Model	Accuracy↑	Extra Training Data	Paper	Code	Result	Year	Tags
1	LogisticRegression on posteriors of xlsr-Wav2Vec2.0&bi-LSTM+Attention	86.70%	✓	A proposal for Multimodal Emotion Recognition using aural transformers and Action Units on RAVDESS dataset			2021	
2	Intermediate-Attention-Fusion	81.58%	×	Self-attention fusion for audiovisual emotion recognition with incomplete data			2022	
3	Logistic Regression on posteriors of the CNN-14&biLSTM-GuidedST	80.08%	✓	Multimodal Emotion Recognition on RAVDESS Dataset Using Transfer Learning			2021	
4	ERANN-0-4	74.8	×	ERANNs: Efficient Residual Audio Neural Networks for Audio Pattern Recognition			2021	



MELD

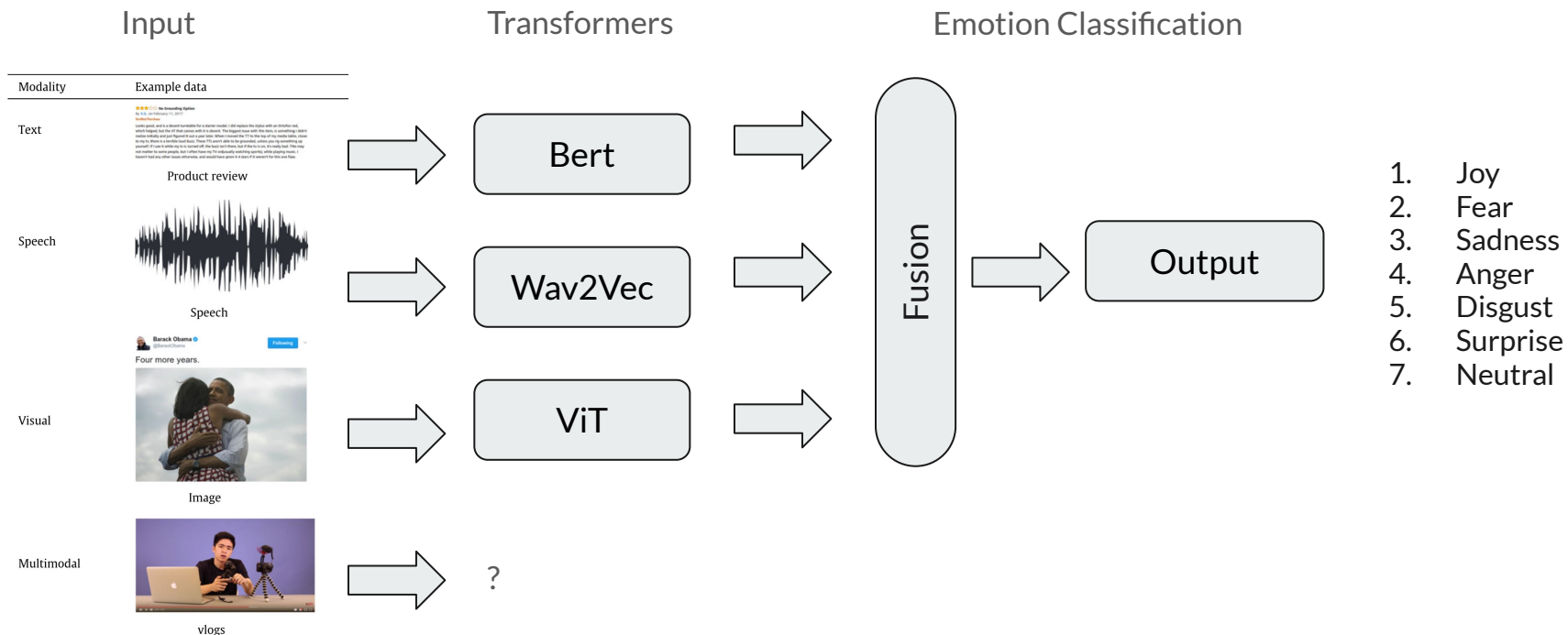
[source](#)



Filter: RoBERTa GNN RNN BERT Glove commonsense XLNet LLM SimCSE BART untagged Edit Leaderboard

Rank	Model	Weighted-F1 ↑	Accuracy	Micro-F1	Paper	Code	Result	Year	Tags
1	InstructERC	69.15			InstructERC: Reforming Emotion Recognition in Conversation with a Retrieval Multi-task LLMs Framework			2023	LLM
2	SPCL-CL-ERC	67.25			Supervised Prototypical Contrastive Learning for Emotion Recognition in Conversation			2022	SimCSE
3	HiDialog	66.96			Hierarchical Dialogue Understanding with Special Tokens and Turn-Level Attention			2023	
4	FacialMMT	66.73			A Facial Expression-Aware Multimodal Multi-task Learning Framework for Emotion Recognition in Multi-party Conversations			2023	RoBERTa

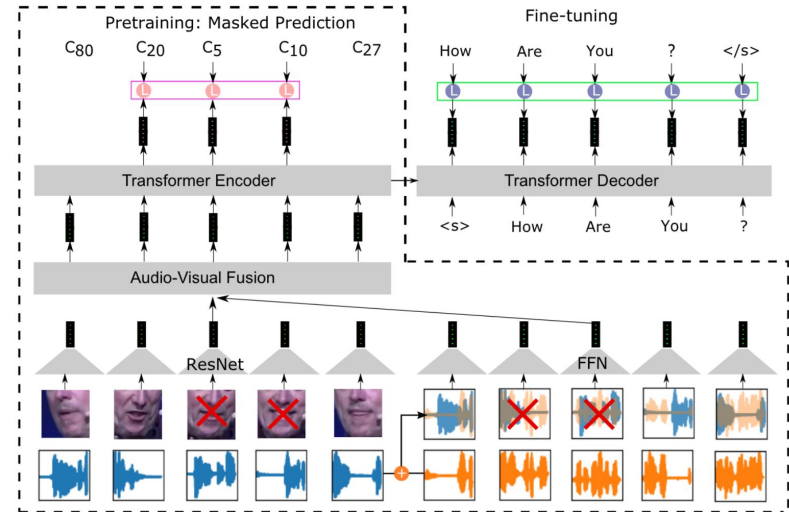
Baseline



AV-HuBERT

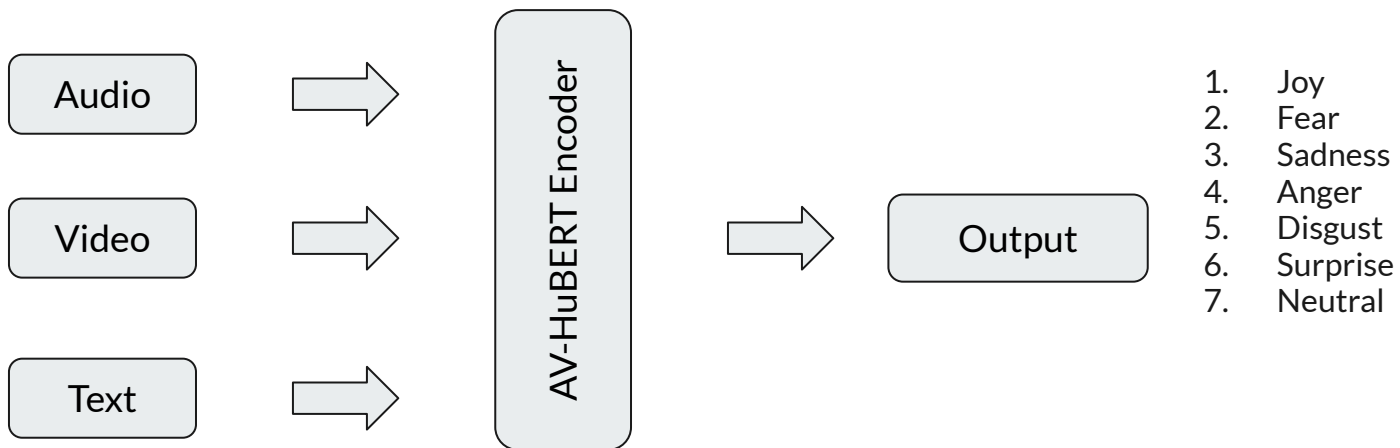
- Extract Facial landmarks
- Retrain AV-HuBERT
- Finetune on multimodal data
- Recognize speech and emotion

Figure 1: AV-HuBERT for audio-visual speech recognition. **X**: mask; blue waveform: original audio; orange waveform: noise; C_n : audio-visual clusters. Dashed box: the pre-trained part





FAV-HuBERT





Experiments

- Setup
- Procedure
- Evaluation



Setup

- Triton
- Hugging Face: ViT, wav2vec, Bert family
- Output: 7 emotions, (& intensity level?)
 - ◆ Anger
 - ◆ Happiness [Joy, Excitement]
 - ◆ Disgust [Frustration]
 - ◆ Sadness
 - ◆ Fear
 - ◆ Surprise
 - ◆ Neutral



Results

- Comparisons
 - ◆ Baseline
 - ◆ fine-tuned AV-HuBERT
 - ◆ state-of-the-art (e.g. AW-HuBERT)
- Observations
 - ◆ Shared representations
 - ◆ Fusion
 - ◆ Alignment
- Limitations



Fusion



Alignment



Robustness



Related Works

- Self-supervised Audiovisual Representation Learning
 - ◆ VideoBERT: text-to-video generation and future forecasting
 - ◆ AV-HuBERT: extends audio HuBERT to model interactions between lip movements and speech from videos.
 - ◆ FAV-HuBERT: retrain the AV-HuBERT model from scratch with the the speakers' faces.
 - ◆ AW-HuBERT: re-trains AV-HuBERT with video recordings of full faces for robust emotion recognition.
 - ◆ AV-Encoder: learns masked audiovisual features reconstruction with feature selection, OpenFace (V) + TRILL (A)
- 'In the Wild' Tasks
 - ◆ Multi-modal speech recognition
 - ◆ Facial expression recognition
 - ◆ Affect recognition
 - ◆ Continuous emotion recognition

VideoBERT: Sun, C., Myers, A., Vondrick, C., Murphy, K., & Schmid, C. (2019). *VideoBERT: A Joint Model for Video and Language Representation Learning* (arXiv:1904.01766).

AV-HuBERT: Bowen Shi, Wei-Ning Hsu, Kushal Lakhotia, and Abdelrahman Mohamed. (2021). Learning Audio-Visual Speech Representation by Masked Multimodal Cluster Prediction. In International Conference on Learning Representations.

AW-HuBERT: Tran, M., Kim, Y., Su, C.-C., Kuo, C.-H., & Soleymani, M. (2023). SAAML: A Framework for Semi-supervised Affective Adaptation via Metric Learning. *Proceedings of the 31st ACM International Conference on Multimedia*, 6004–6015.

AV-Encoder: Minh et al. Tran. (2022). A Pre-Trained Audio-Visual Transformer for Emotion Recognition. In ICASSP. IEEE.



Future Studies

→ **Co-learning:** parallel & non-parallel multimodal datasets

- ◆ Zero-shot
- ◆ One-shot
- ◆ Few-shot
- ◆ 'Human-in-the-loop' simulation

→ **Robustness:** 'To Err is Human...'

- ◆ Recognition: you see what you want to see
- ◆ Opinion: you only hear what you want to hear, humor or hate?
- ◆ Annotation: annotation agreement, human performance
- ◆ The wandering thoughts, the shifting emotions...

→ **Interpretability**

- ◆ Could machine perform better than human?
- ◆ Other evaluation metrics?



Conclusion

1. **Multimodal Representation Learning**
 - a. Background
 - b. Transformers
 - c. Research Directions
2. **Speech Emotion Recognition**
 - a. Research Questions
 - b. Datasets
 - c. Experiments
 - d. Results

Thank you!



References

- A. ASR: multimodal speech recognition ([Survey](#))
- B. NLP: multimodal sentiment analysis ([Survey](#))
- C. CV: multimodal affect/emotion/facial expression recognition ([Survey](#))



Reference

(Zotero)

2020~2023

2015~2020

2010~2015

~2010