dfki ai
Deutsches Forschungszentrum
für Künstliche Intelligenz
German Research Center for
Artificial Intelligence

AI-Based Multimodal Interaction Analysis
Second GAIN Summer School

Philipp Müller
Senior Researcher
German Research Center for AI
philipp.mueller@dfki.de

About myself

Saarland University
BSc & MSc Computer Science
BSc Psychology

MPI for Informatics
PhD Computer Science

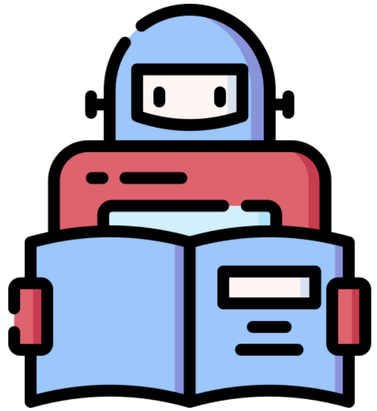
University of Stuttgart
Researcher

German Research Center for AI
Senior Researcher

Main Research Interests:
Social Signal Analysis
Multi-modal Machine Learning
Human Attention and Cognition
Brain-Computer Interfaces

Conversation Analysis: Example Application Fields

Medical Decision Support



Intelligent Tutors

For these applications, in-depth understanding of human conversational behaviour and states is key!

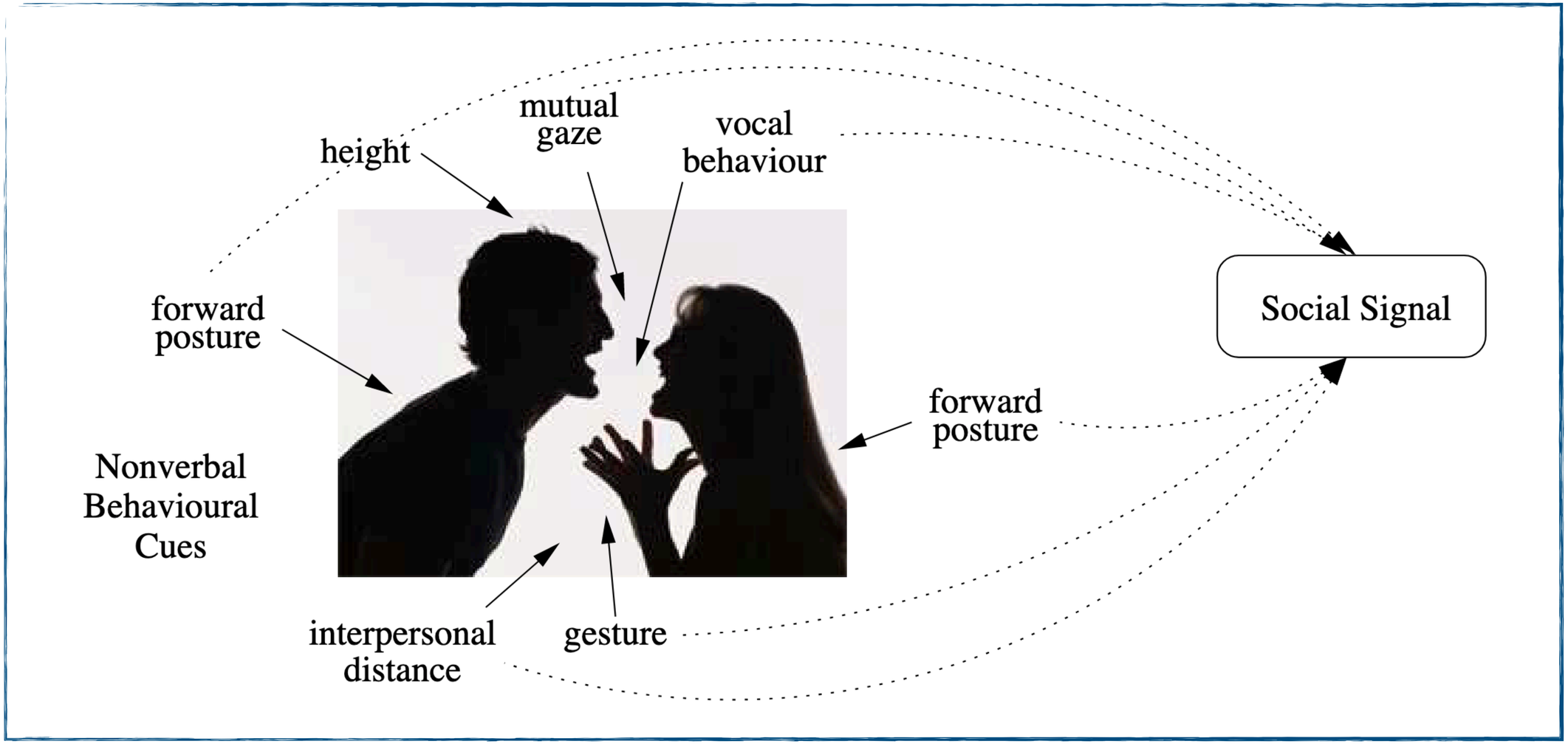
Multi-modal Conversation Analysis

How do humans express states such as emotion, trust, or psychiatric symptoms?

A combination of what they say, how they say it, and what they show.

We need to analyse and interpret all of these different modalities!

Social Signals



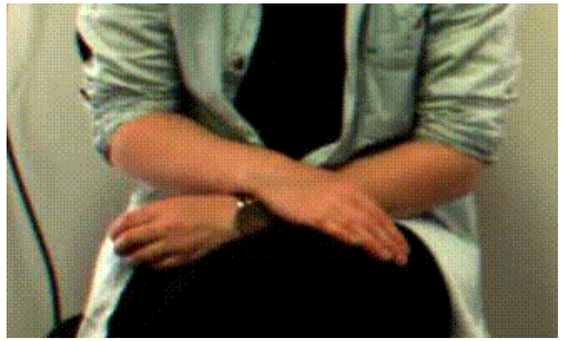
Vinciarelli, Pantic, & Bourlard. (2009). Social signal processing: Survey of an emerging domain. Image and vision computing.

Social Signals

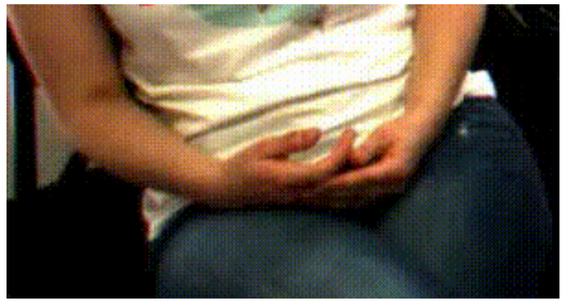
	Example Social Behaviours							Tech.		
Social Cues	emotion	personality	status	dominance	persuasion	regulation	rapport	speech anlysis	computer vision	biometry
Gesture and posture										
hand gestures	√	√			√	√	√		√	√
posture	√	√	√	√	√	√	√		√	√
walking		√	√	√					√	√

Vinciarelli, Pantic, & Bourlard. (2009). Social signal processing: Survey of an emerging domain. Image and vision computing.

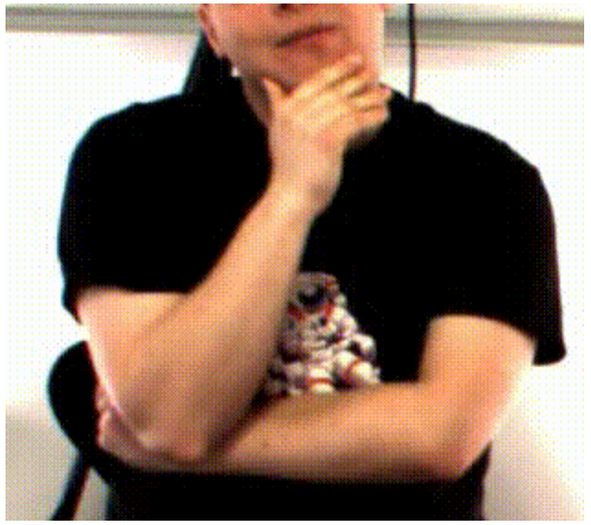
Bodily Behaviours



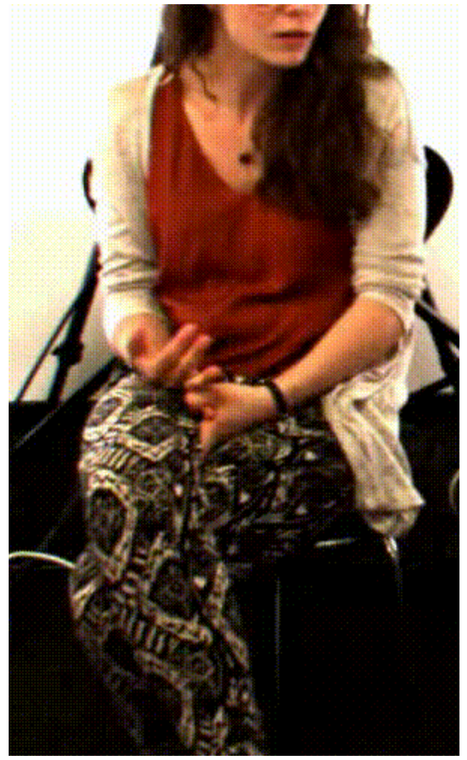
Fumbling



Fumbling, Gesturing



Face touching,
Arms crossed



Gesticulating, Shrugging,
Legs crossed, Fumbling

German Research Center for Artificial Intelligence

Universität Augsburg University

University of Stuttgart Germany

Inria

Balazia, M., Müller, P., Tánczos, Á. L., Liechtenstein, A. V., & Bremond, F. (2022). Bodily behaviors in social interaction: Novel annotations and state-of-the-art evaluation. ACM MM'22.

Social Signals

	Example Social Behaviours							Tech.		
Social Cues	emotion	personality	status	dominance	persuasion	regulation	rapport	speech anlysis	computer vision	biometry
Space and Environment										
distance	√	√	√		√		√		√	
seating arrangement				√	√		√		√	

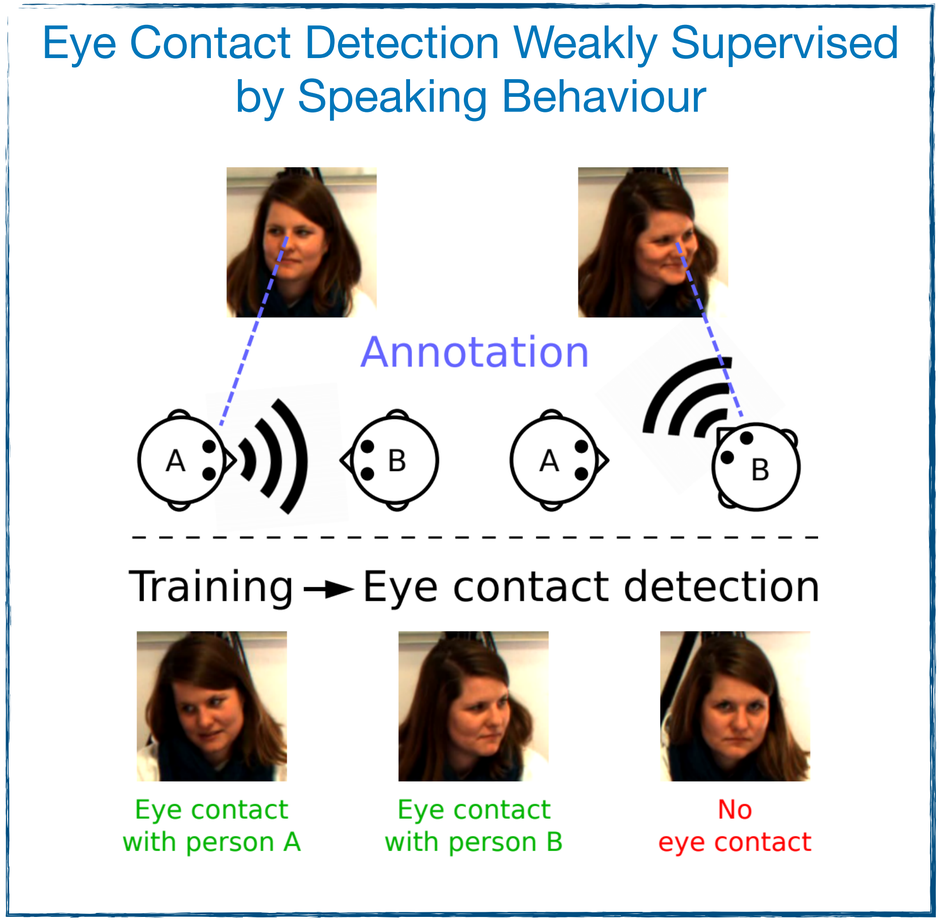
Vinciarelli, Pantic, & Bourlard. (2009). Social signal processing: Survey of an emerging domain. Image and vision computing.

Social Signals

	Example Social Behaviours							Tech.		
Social Cues	emotion	personality	status	dominance	persuasion	regulation	rapport	speech anlysis	computer vision	biometry
Face and eyes behaviour										
facial expressions	√	√	√	√	√	√	√		√	√
gaze behaviour	√	√	√	√	√	√	√	?	√	
focus of attention	√	√	√	√	√	√	√	?	√	

Vinciarelli, Pantic, & Bourlard. (2009). Social signal processing: Survey of an emerging domain. Image and vision computing.

Social Signals



Müller, P., Huang, M. X., Zhang, X., & Bulling, A. Robust eye contact detection in natural multi-person interactions using gaze and speaking behaviour. ETRA'18

Social Signals

	Example Social Behaviours							Tech.		
Social Cues	emotion	personality	status	dominance	persuasion	regulation	rapport	speech anlysis	computer vision	biometry
Vocal behaviour										
prosody	√	√		√	√		√	√		
turn taking	√	√	√	√		√	√	√		
vocal outbursts	√	√		√	√	√	√	√		
silence	√		√				√	√		

Vinciarelli, Pantic, & Bourlard. (2009). Social signal processing: Survey of an emerging domain. Image and vision computing.

Multi-modal Conversation Analysis

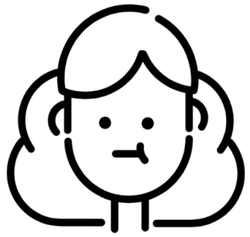
Classically, social signal processing has disregarded verbal information

However of course verbal information is crucial in dialogue

Our goal is to integrate nonverbal with verbal informaiton

Agenda

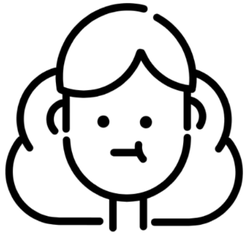
Visual Behaviour



Text

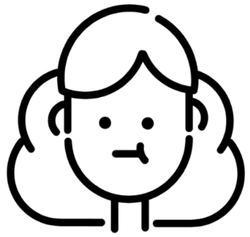
Speech

Multi-Modal Conversation Analysis



Agenda

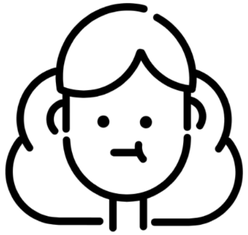
Visual Behaviour



Text

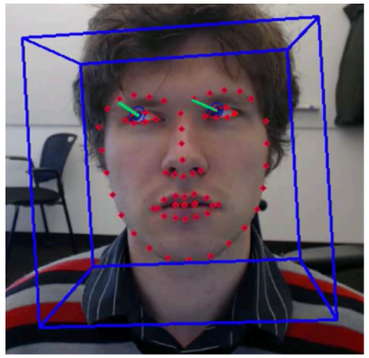
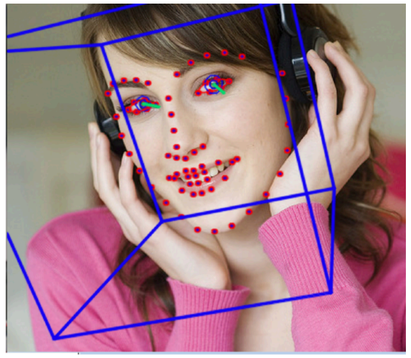
Speech

Multi-Modal Conversation Analysis



Facial Expression Representations

Facial Keypoints & Gaze

https://github.com/TadasBaltrusaitis/OpenFace

Facial Action Units



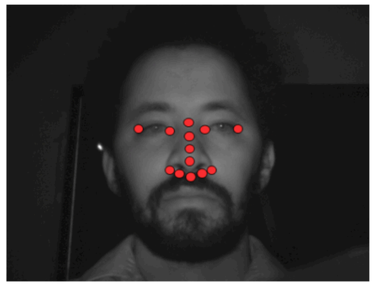

https://www.cs.cmu.edu/~face/facs.htm

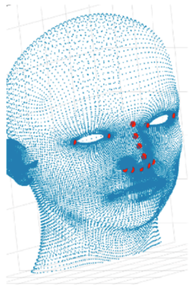
Facial Emotion Expressions

https://hobbylark.com/writing/Use-the-6-Basic-Emotions-in-Writing

Facial Keypoints

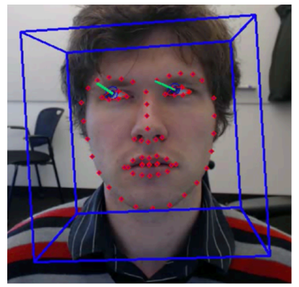
Visually salient points, e.g. the tip of the nose, ends of the eye brows, the corners of the mouth

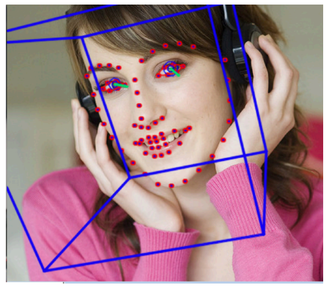




Barros, J. M. D., Mirbach, B., Garcia, F., Varanasi, K., & Stricker, D. (2018). Fusion of keypoint tracking and facial landmark detection for real-time head pose estimation. WACV'18

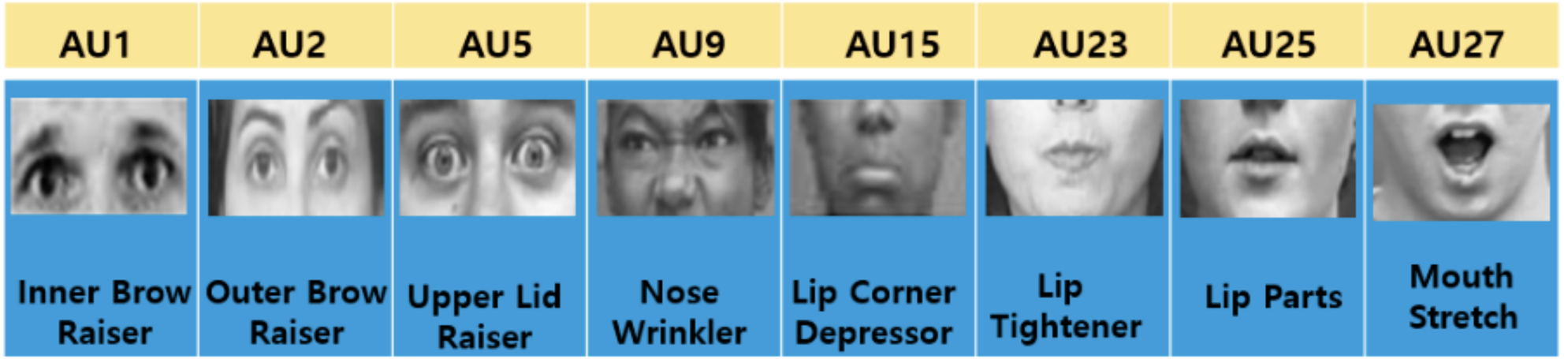
Important representation for **further processing**, e.g. fitting a 3D face model or cropping the eye region





OpenFace 2.0

Facial Action Units



Coding system for facial muscle contractions

Advantage: objective, not context dependent

For a comprehensive list, see: https://www.cs.cmu.edu/~face/facs.htm

Facial Emotion Expressions

Six **basic emotions** according to Ekman:

Anger Fear Disgust Surprise Enjoyment Sadness

Compound emotions:
happily surprised, happily disgusted, sadly fearful

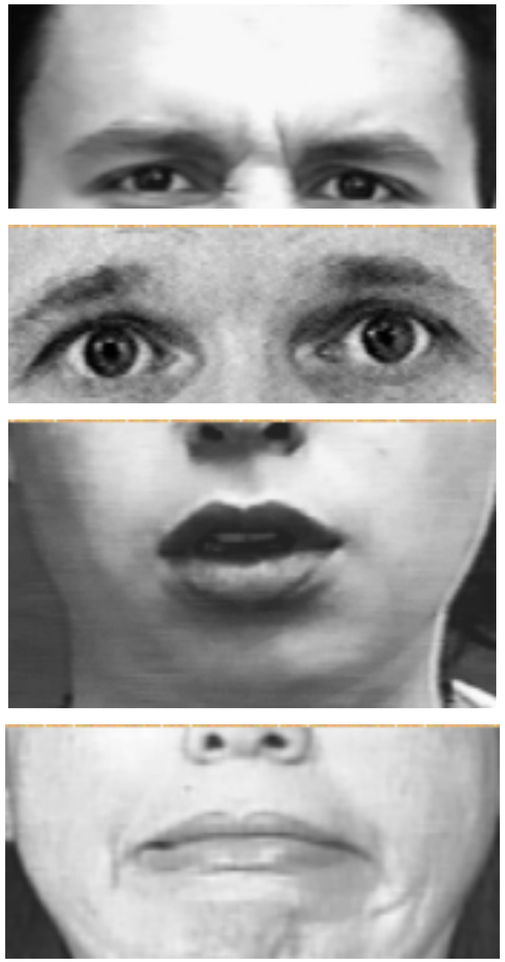
Facial Action Units and Emotions: Quiz

Happy
Sad
Fearful
Angry
Surprised
Disgusted
Happily sad
Happily surprised
Happily disgusted
Sadly fearful
Sadly angry
Sadly surprised

Facial Action Units and Emotions: Quiz

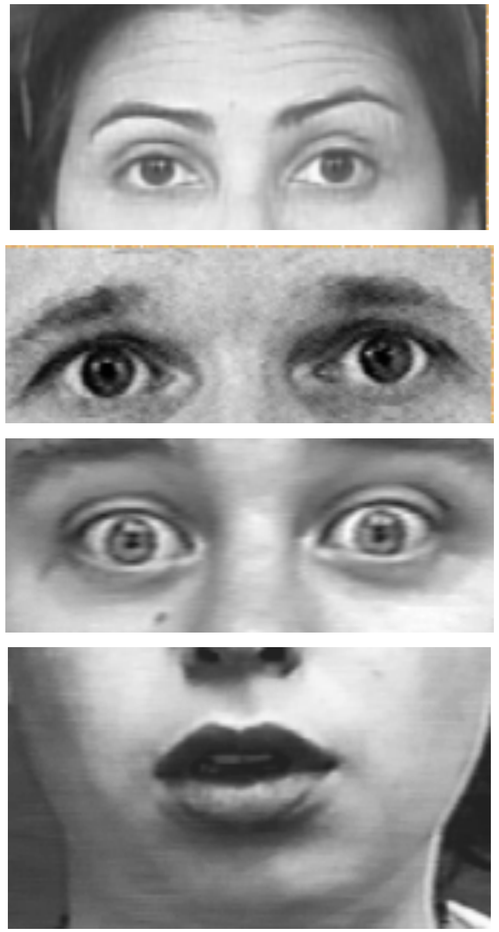
Happy
Sad
Fearful
Angry
Surprised
Disgusted
Happily sad
Happily surprised
Happily disgusted
Sadly fearful
Sadly angry
Sadly surprised

Facial Action Units and Emotions: Quiz



Happy
Sad
Fearful
Angry
Surprised
Disgusted
Happily sad
Happily surprised
Happily disgusted
Sadly fearful
Sadly angry
Sadly surprised

Facial Action Units and Emotions: Quiz

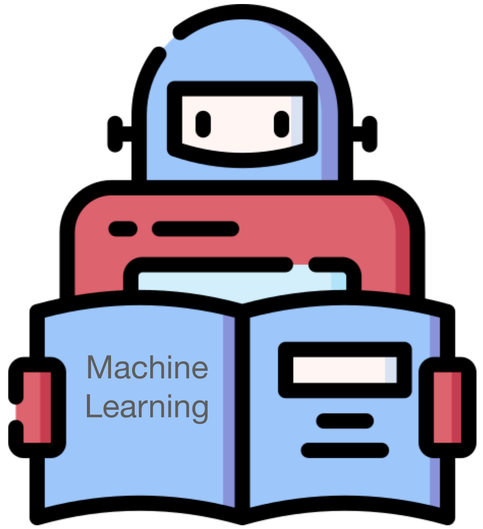


Sadly disgusted
Fearfully angry
Fearfully surprised
Fearfully disgusted
Angrily surprised
Disgusted surprised
Happily fearful
Angrily disgusted
Awed
Appalled
Hatred

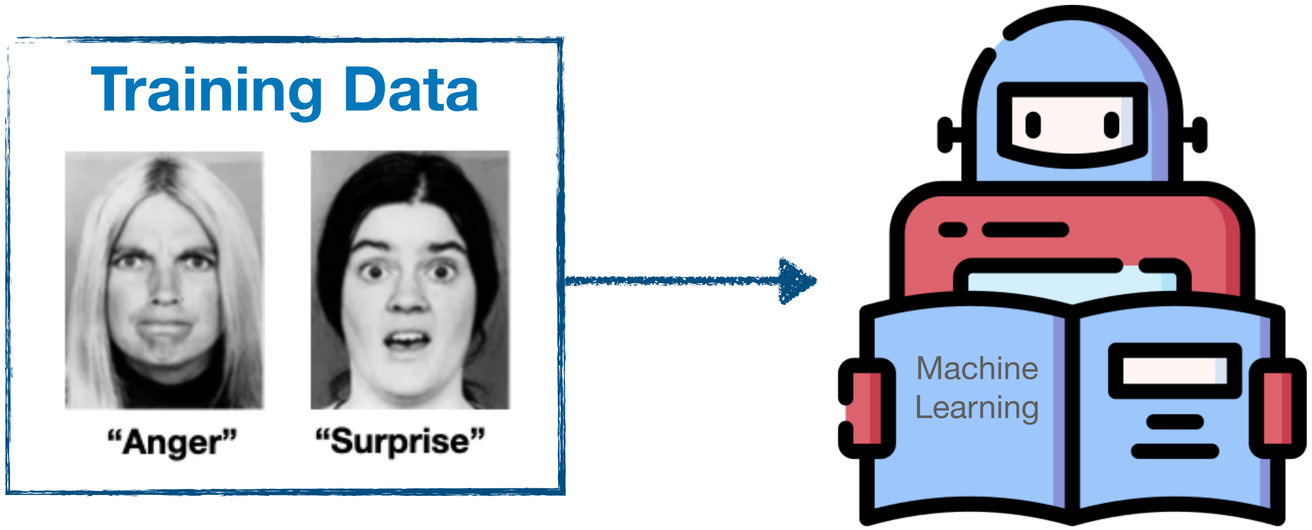
How to Detect Facial Emotion Expressions?

Anger Fear Disgust Surprise Enjoyment Sadness

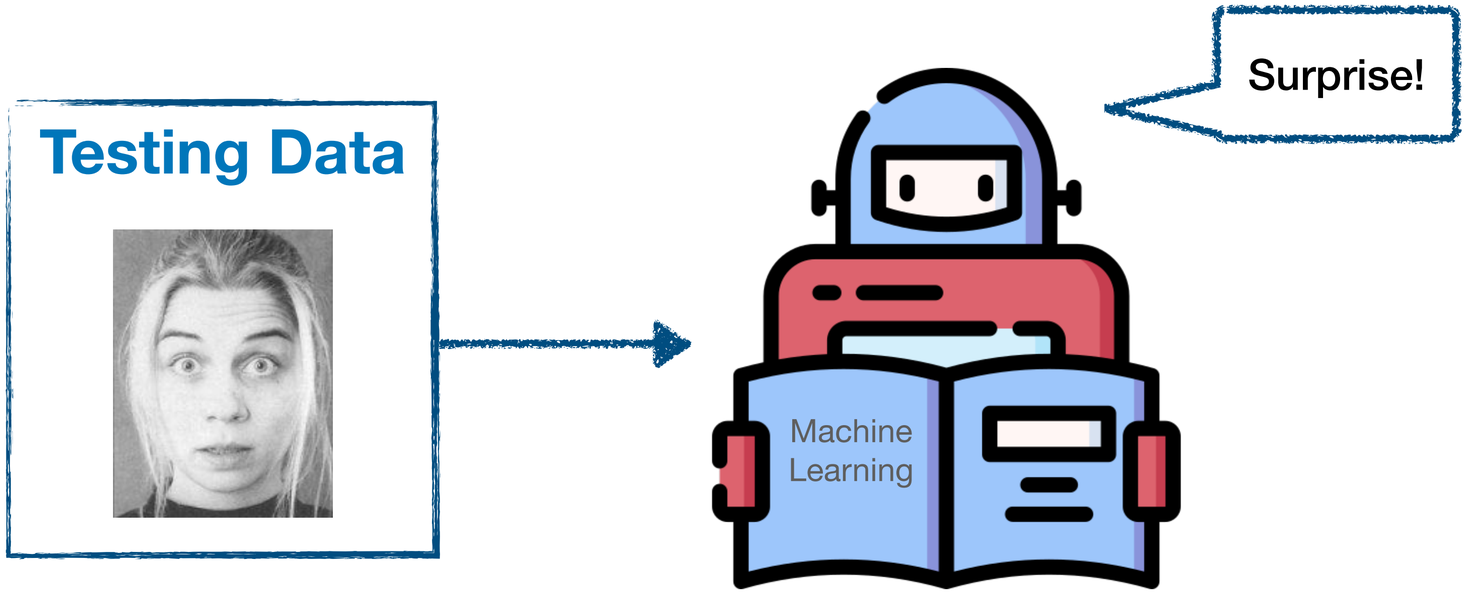
How to define the relation between pixels and emotion expressions?



Recap on ML



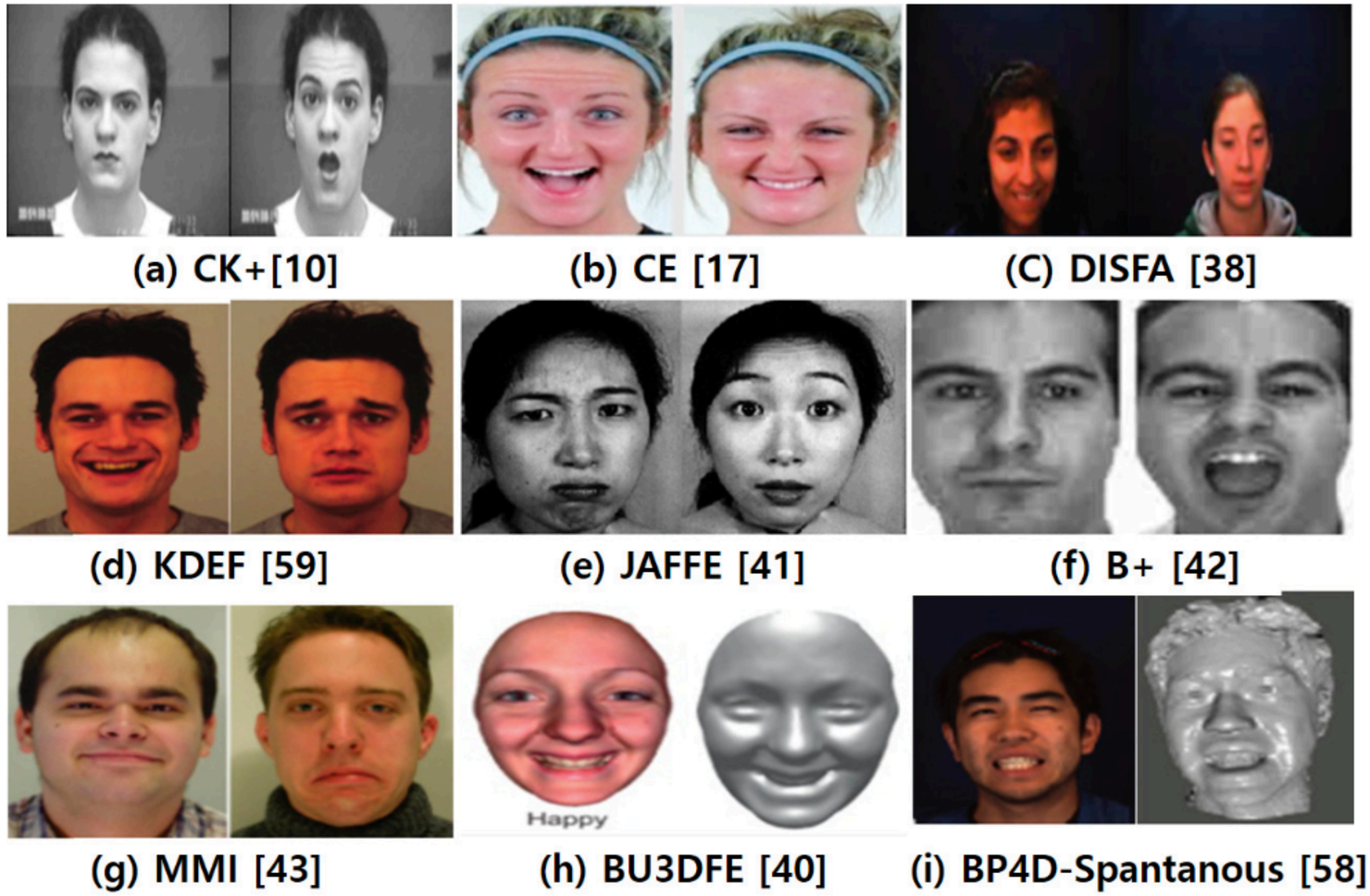
Recap on ML



The Algorithm needs to learn that it is not about the hair color!

Large and representative training dataset is needed.

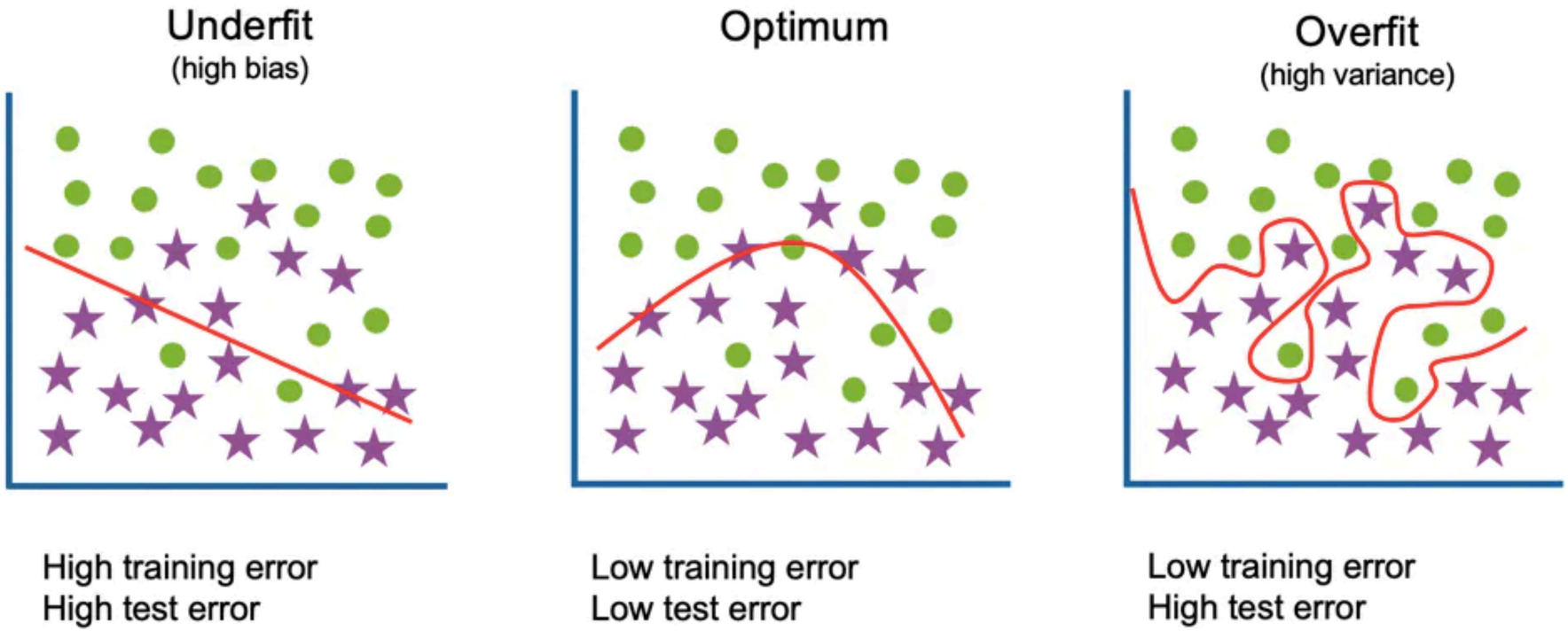
Facial Expression Analysis Datasets



Ko, B. C. (2018). A brief review of facial emotion recognition based on visual information. *sensors*, *18*(2), 401.

Recap on ML

The key problem in Machine Learning: **Generalization**



Facial Expression Analysis

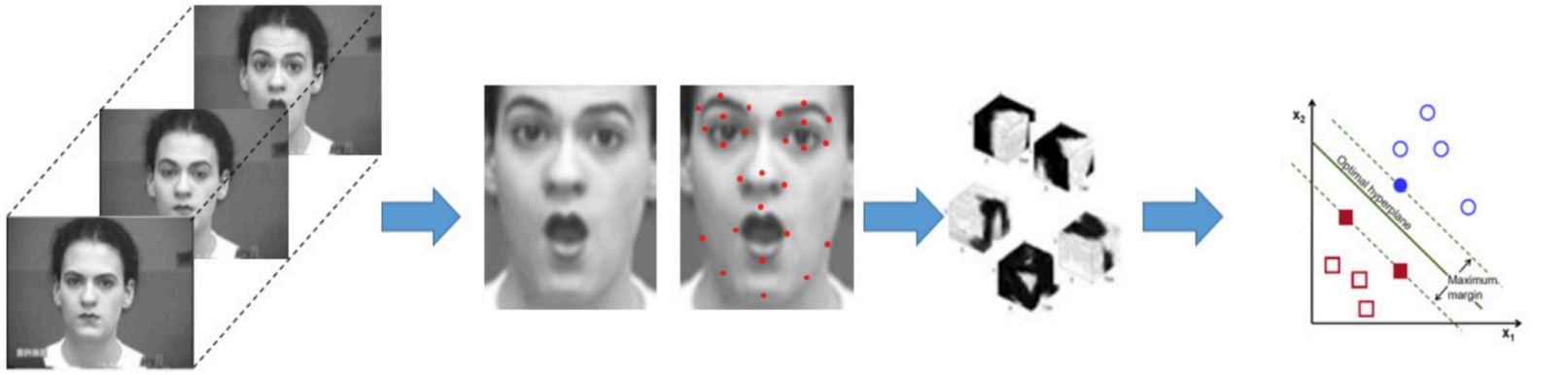
A 224x224x3 image has 150528 dimensions.

The curse of dimensionality! Usual Facial Emotion datasets don't even have that many training samples.

It becomes very hard to avoid overfitting in this scenario.

Facial Expression Analysis

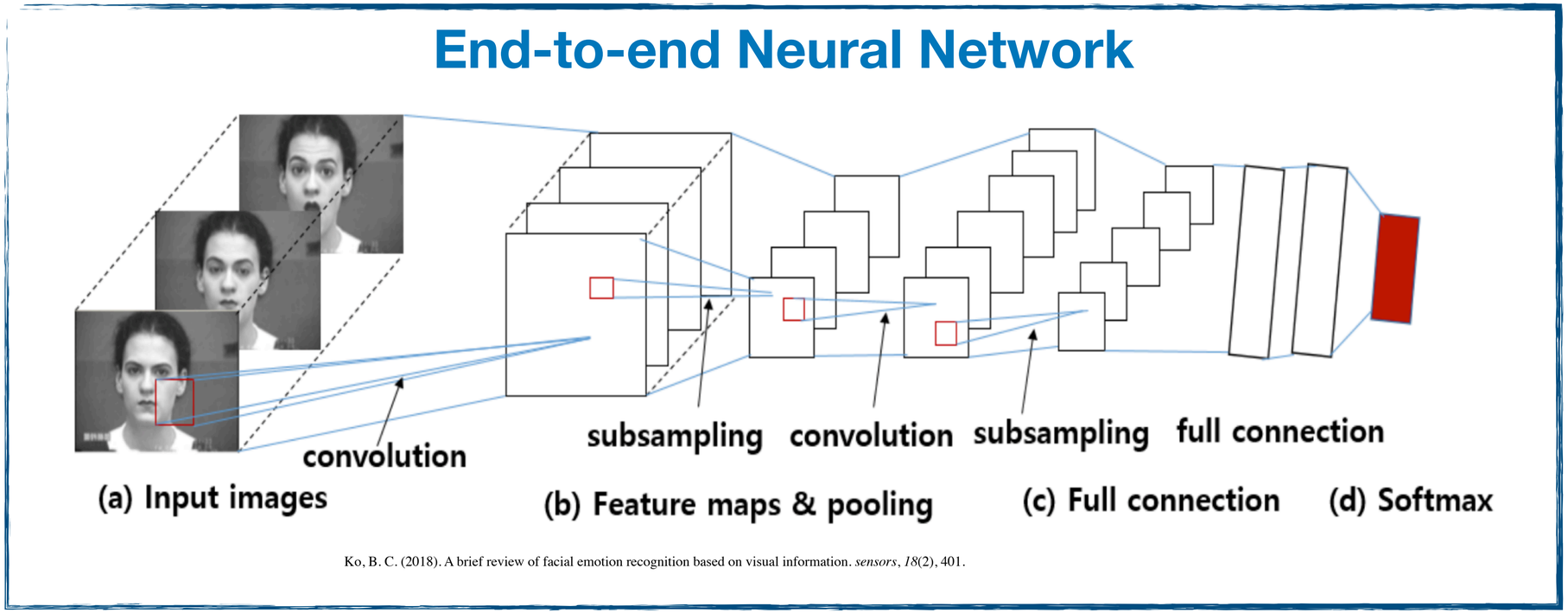
The Classical Approach



(a) Input images (b) Face detection & landmark detection (c) Feature extraction (d) FE classification

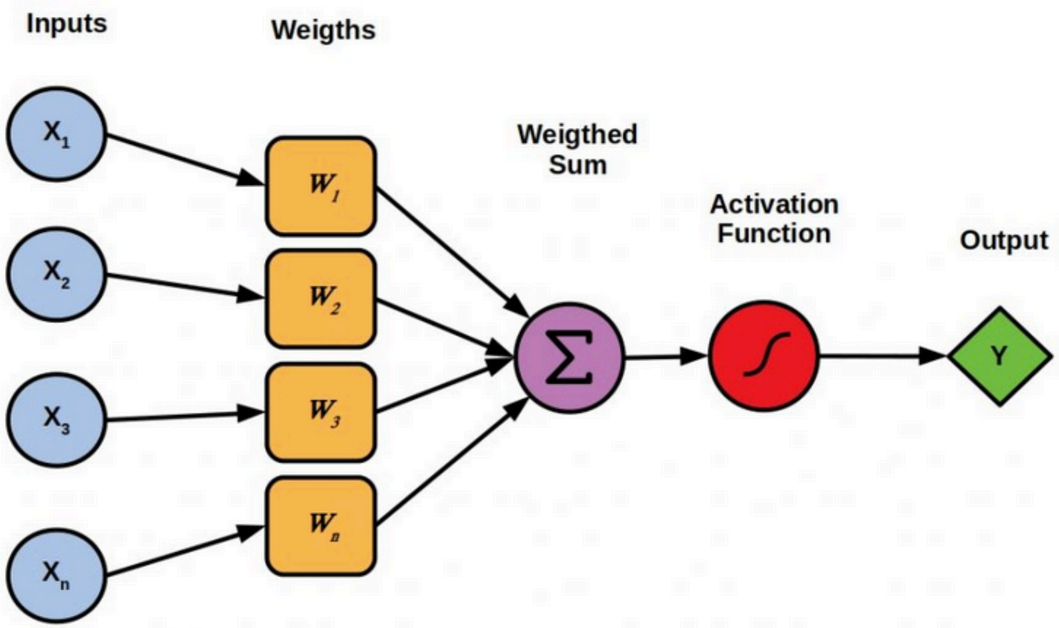
Ko, B. C. (2018). A brief review of facial emotion recognition based on visual information. *sensors*, *18*(2), 401.

Facial Expression Analysis



Ko, B. C. (2018). A brief review of facial emotion recognition based on visual information. *sensors*, *18*(2), 401.

Recap on Neural Networks: The Perceptron



Loss function applied to output, e.g.:

$$L(y, \hat{y}) = \frac{1}{n} \sum_{i=1..n} (y_i - \hat{y}_i)^2$$

During training we try to minimise the loss on the training set by adjusting the weights of the network accordingly.

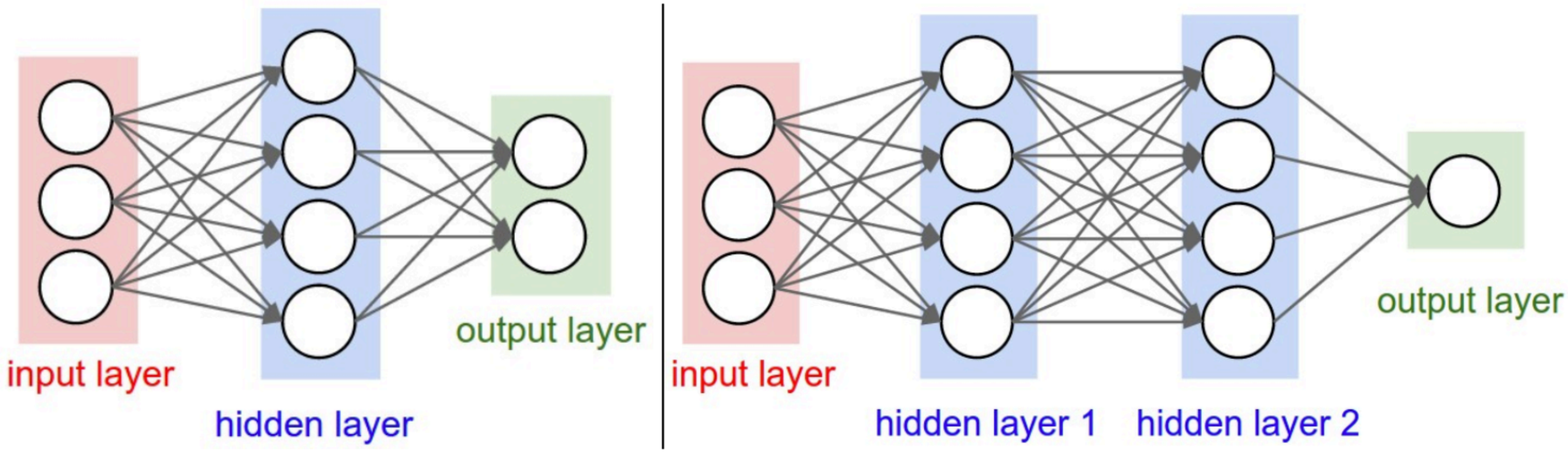
This is done by computing the gradient of the loss function w.r.t. the weights.

Many different loss functions exist, suitable for different tasks (classification, regression,…)

Recap on Neural Networks: Fully Connected Networks

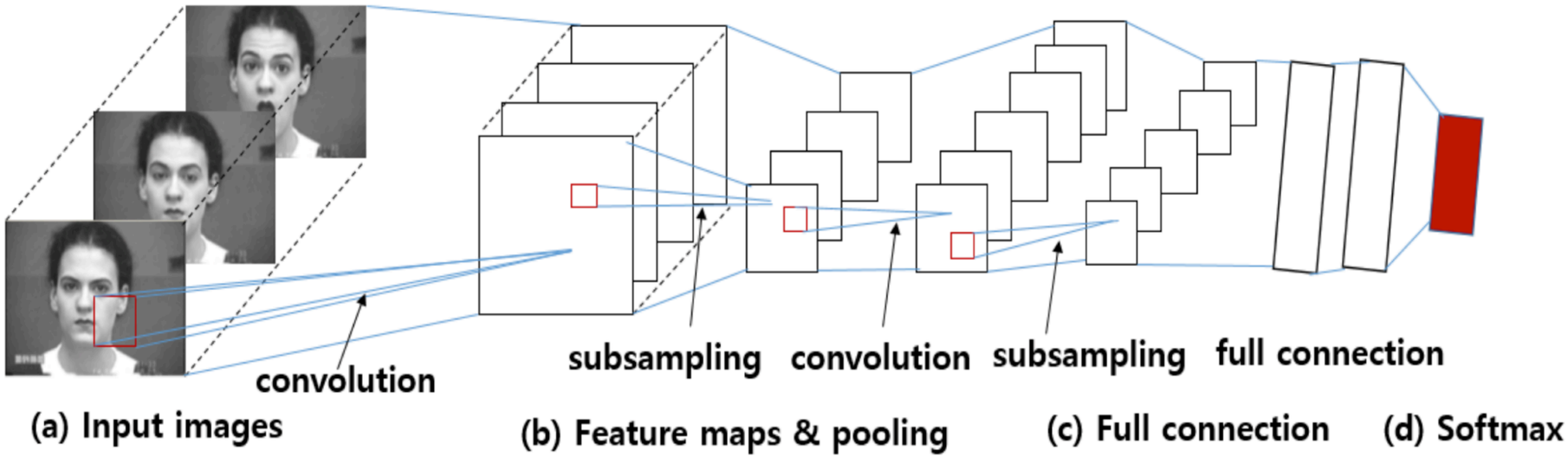
We can arrange perceptrons into larger neural networks.

Using backpropagation we can still compute the gradient of the loss w.r.t. all weights in the network.



In an 224x224x3 image this fully connected architecture will lead to 150528 times “number of hidden layer 1” weights in the first layer only! Danger of overfitting.

Facial Expression Analysis with CNNs

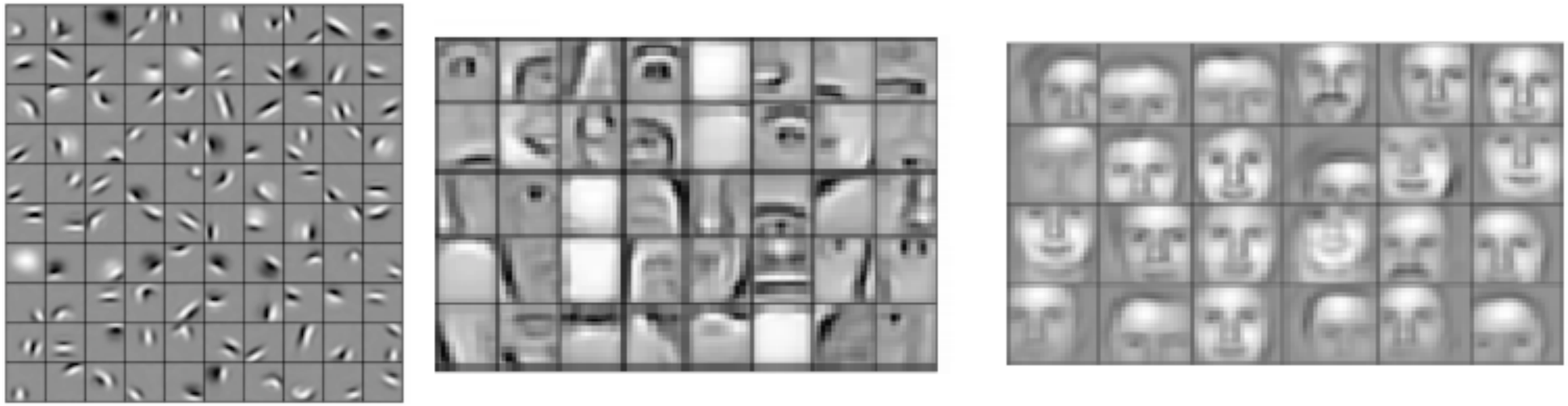


Convolutional neural networks slide convolution kernels over the image / intermediate feature representation.

This encodes spatial invariance and reduces the number of parameters (weights) that need to be learned.

Facial Expression Analysis

Hierarchical Feature Representation in CNNs



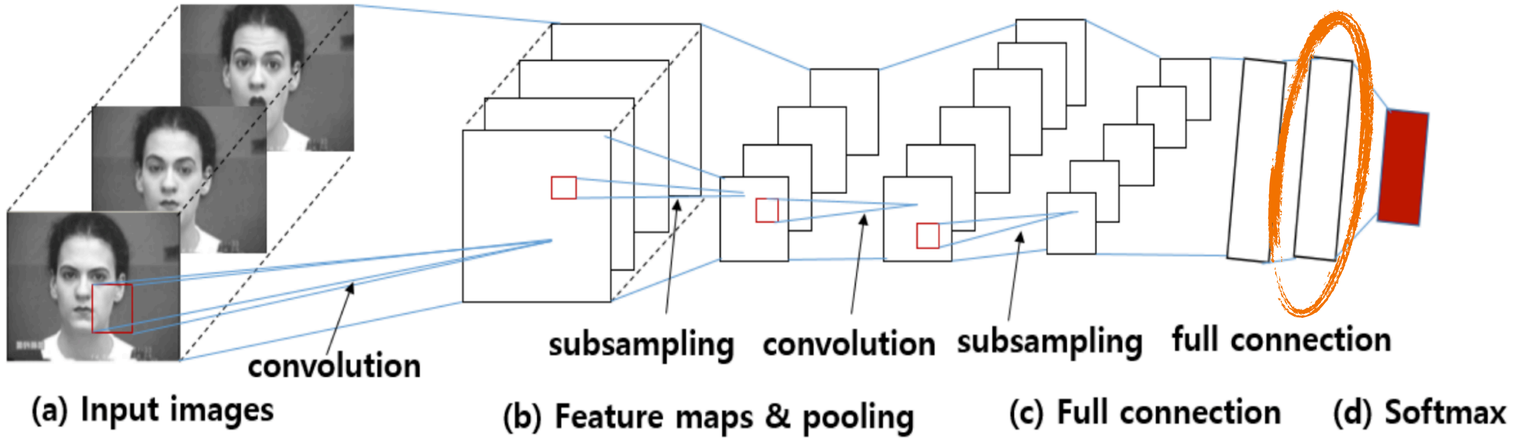
Lee, H., Grosse, R., Ranganath, R., & Ng, A. Y. (2011). Unsupervised learning of hierarchical representations with convolutional deep belief networks. Communications of the ACM.

Such representations are useful across several tasks, e.g. pre-train on face recognition, fine-tune on emotion expression recognition

Types of Representations

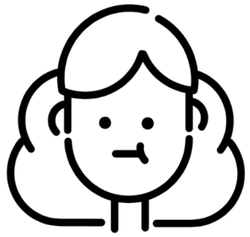
Explicit representation: Keypoints, Facial Action Units, Emotion Classes

Vector space representation (“embeddings”)



Agenda

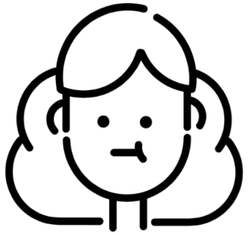
Visual Behaviour



Text

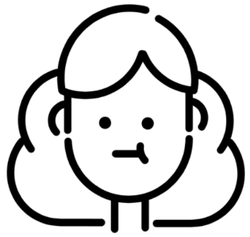
Speech

Multi-Modal Conversation Analysis



Agenda

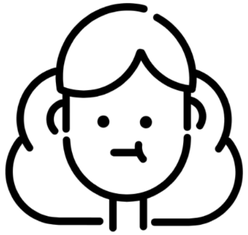
Visual Behaviour



Text

Speech

Multi-Modal Conversation Analysis



Masked Language Models

E.g. BERT (Devlin et al., 2019)

Self-supervised Training Task: “The fisher is [MASK] at the river.”

What are likely words at the [MASK] location?

Being able to predict the masked word well requires an intricate model of language, including syntax and semantics

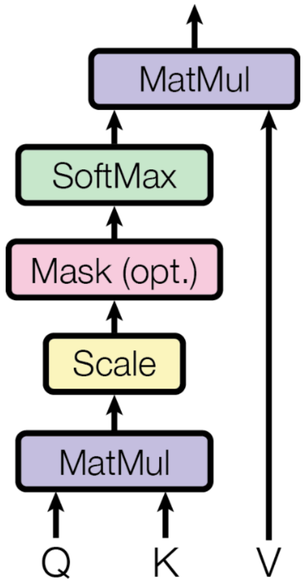
BERT-style models proved useful on many downstream tasks, e.g. named entity recognition, part-of-speech-tagging, sentiment analysis,…

Devlin et al. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL-HLT’19.

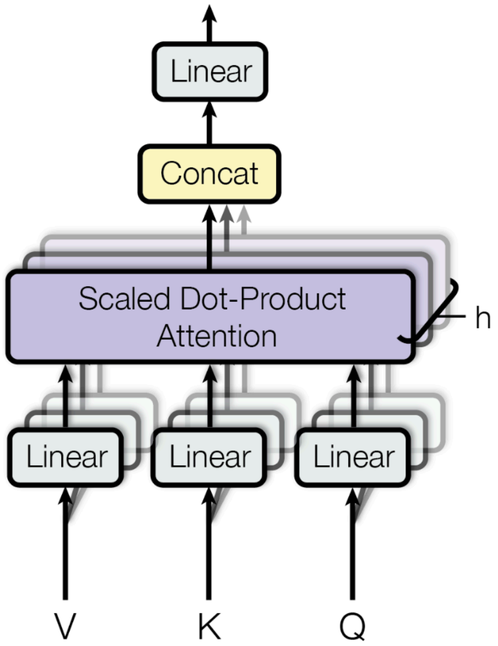
The Transformer (Vaswani et al., 2017)

Self-attention: basic building block of BERT and most current language models

Scaled Dot-Product Attention



Multi-Head Attention



$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

Good visual explanation: https://jalammar.github.io/illustrated-transformer/

Vaswani, Ashish; Shazeer, Noam; Parmar, Niki; Uszkoreit, Jakob; Jones, Llion; Gomez, Aidan N; Kaiser, Łukasz; Polosukhin, Illia (2017). "Attention is All you Need". Advances in Neural Information Processing Systems.

LLMs (e.g. Llama3, GPT-4)

Also employ a transformer-based architecture

They are much larger and trained on larger datasets with focus on the generative task

- BERT large: 340M parameters
- Llama3: 8B or 70B parameters
- GPT-4: Mixture of Experts of 8x222B parameter models

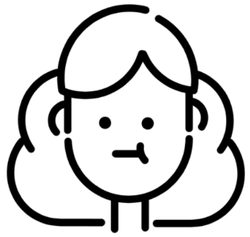
We cannot train Llama-scale models from scratch

But very effective fine-tuning approaches exist, e.g. LoRA (Hu et al., 2021)

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., ... & Chen, W. (2021). Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685.

Agenda

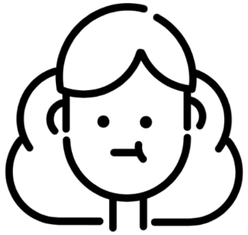
Visual Behaviour



Text

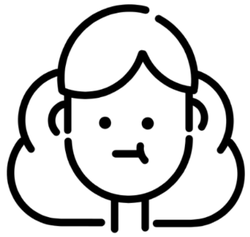
Speech

Multi-Modal Conversation Analysis



Agenda

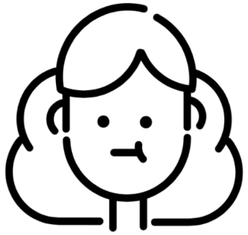
Visual Behaviour



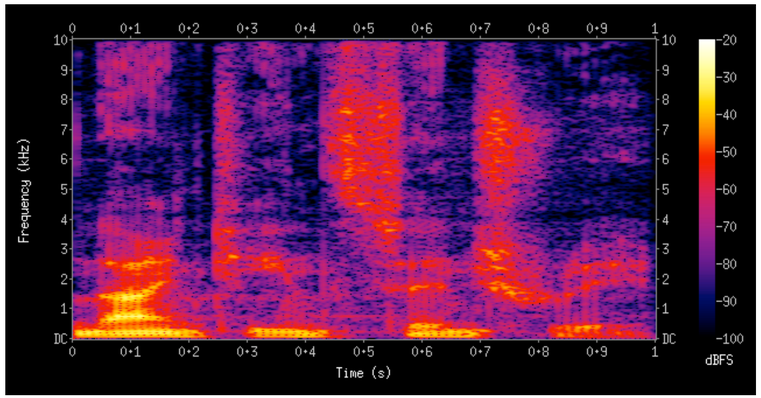
Text

Speech

Multi-Modal Conversation Analysis



Geneva Minimalistic Acoustic Parameter Set (GeMAPS)



Frequency Parameters

Pitch, logarithmic F_0 on a semitone frequency scale, starting at 27.5 Hz (semitone 0).
Jitter, deviations in individual consecutive F_0 period lengths.
Formant 1, 2, and 3 frequency, centre frequency of first, second, and third formant
Formant 1, bandwidth of first formant.

Energy/Amplitude Parameters

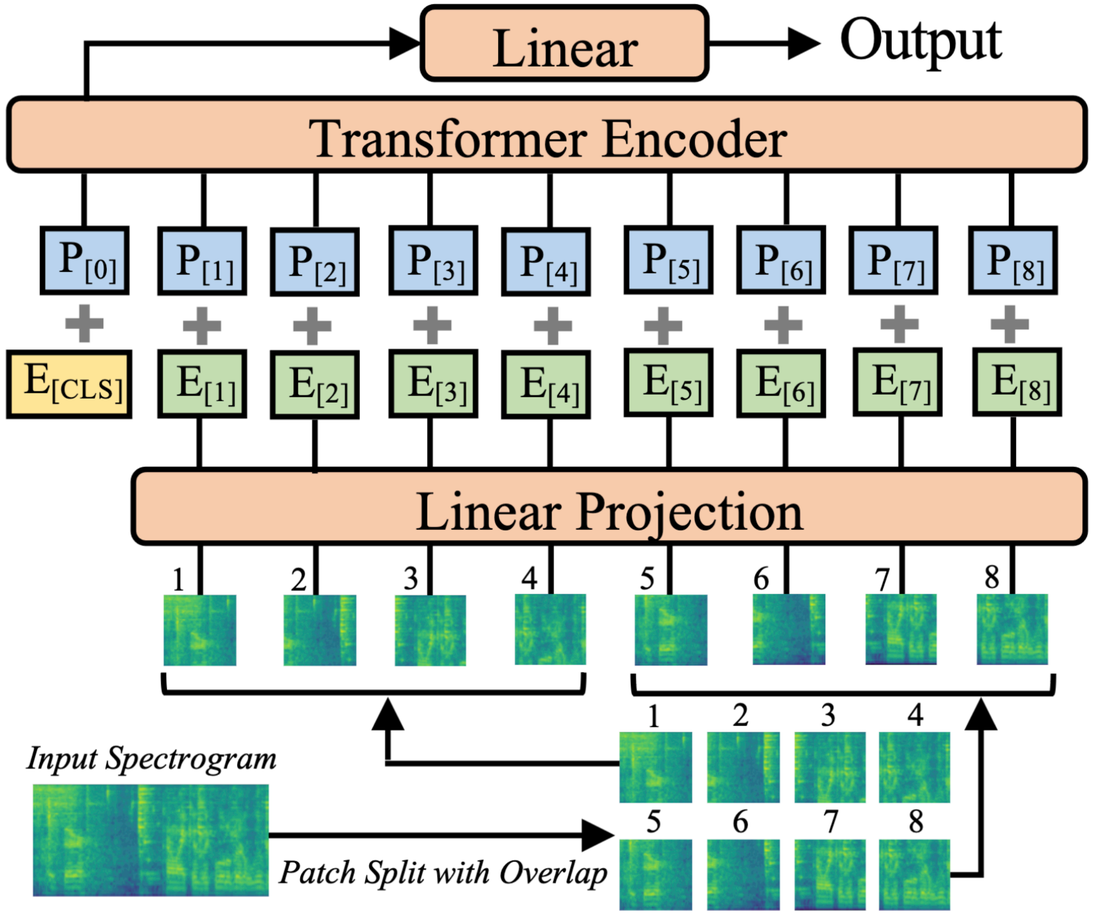
Shimmer, difference of the peak amplitudes of consecutive F_0 periods.
Loudness, estimate of perceived signal intensity from an auditory spectrum.
Harmonics-to-noise ratio (HNR), relation of energy in harmonic components to energy in noise-like components.

Spectral Parameters

Alpha Ratio, ratio of the summed energy from 50-1000 Hz and 1-5 kHz
Hammarberg Index, ratio of the strongest energy peak in the 0-2 kHz region to the strongest peak in the 2–5 kHz region.
Spectral Slope 0-500 Hz and 500-1500 Hz, linear regression slope of the logarithmic power spectrum within the two given bands.
Formant 1, 2, and 3 relative energy, as well as the ratio of the energy of the spectral harmonic peak at the first, second, third formant's centre frequency to the energy of the spectral peak at F_0.
Harmonic difference H1-H2, ratio of energy of the first F_0 harmonic (H1) to the energy of the second F_0 harmonic (H2).
Harmonic difference H1-A3, ratio of energy of the first F_0 harmonic (H1) to the energy of the highest harmonic in the third formant range (A3).

Eyben, F., Scherer, K. R., Schuller, B. W., Sundberg, J., André, E., Busso, C., ... & Truong, K. P. (2015). The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. IEEE transactions on affective computing, 7(2), 190-202.

Audio Spectrogram Transformer



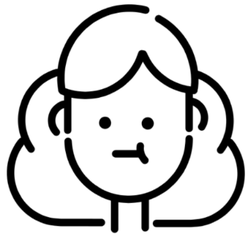
Transfers weights from ViT trained on Imagenet to initialize the model

SOTA results on environmental noise classification and speech commands

Gong, Y., Chung, Y. A., & Glass, J. (2021). Ast: Audio spectrogram transformer. arXiv preprint arXiv:2104.01778.

Agenda

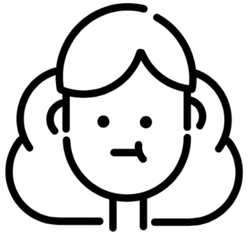
Visual Behaviour



Text

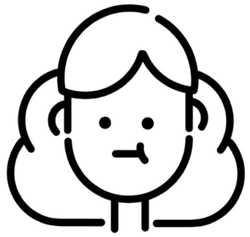
Speech

Multi-Modal Conversation Analysis



Agenda

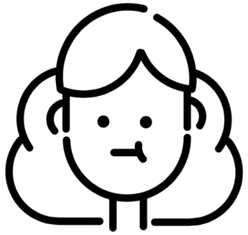
Visual Behaviour



Text

Speech

Multi-Modal Conversation Analysis



dfki ai
Deutsches Forschungszentrum
für Künstliche Intelligenz
German Research Center for
Artificial Intelligence

M³TCM: Multi-modal Multi-task Context Model for Utterance Classification in Motivational Interviews

Sayed Muddashir Hossain, Jan Alexandersson, **Philipp Müller**
German Research Center for Artificial Intelligence

philipp.mueller@dfki.de

LREC-COLING'24

Motivational Interviewing (MI)

Goal of MIs: behaviour change in clients

Automatic **classification** of **MI utterances** to:

- understand communication dynamics
- help therapists reflect on their behaviour
- predict outcome
- basis for artificial agents as therapists

Motivational Interviewing (MI)

Three key properties of MIs:

- Asymmetric roles and labels (therapist, client)
- Context important
- Prosody needs to be integrated

We propose **M³TCM**: the first **multi-modal, multi-task context** model for utterance classification in MI

Utterance Classification in MI

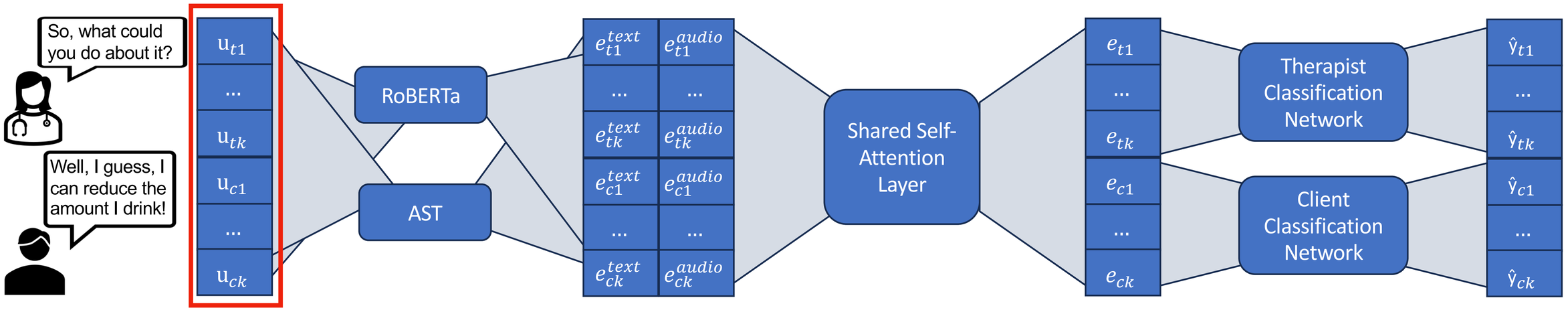
Therapist Utterance Classes:

- Reflection
- Question
- Input
- Other

Patient Utterance Classes:

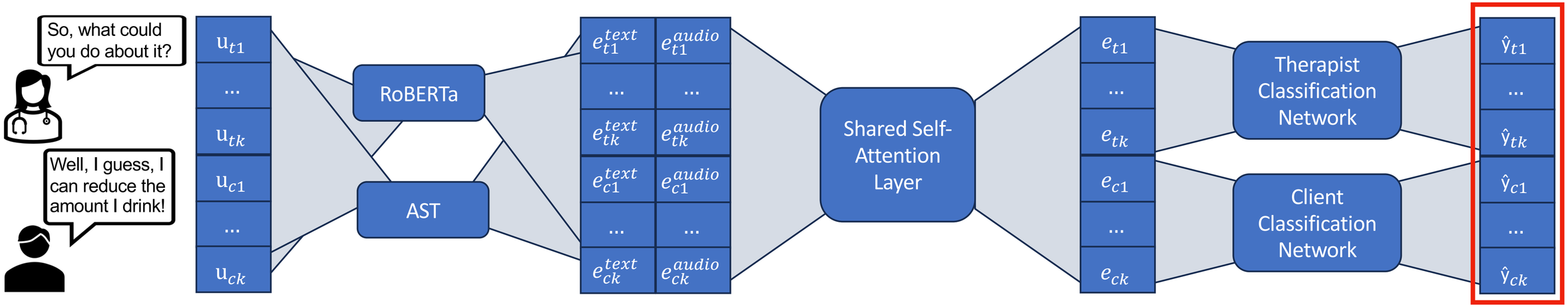
- Change
- Neutral
- Sustain

M³TCM Approach



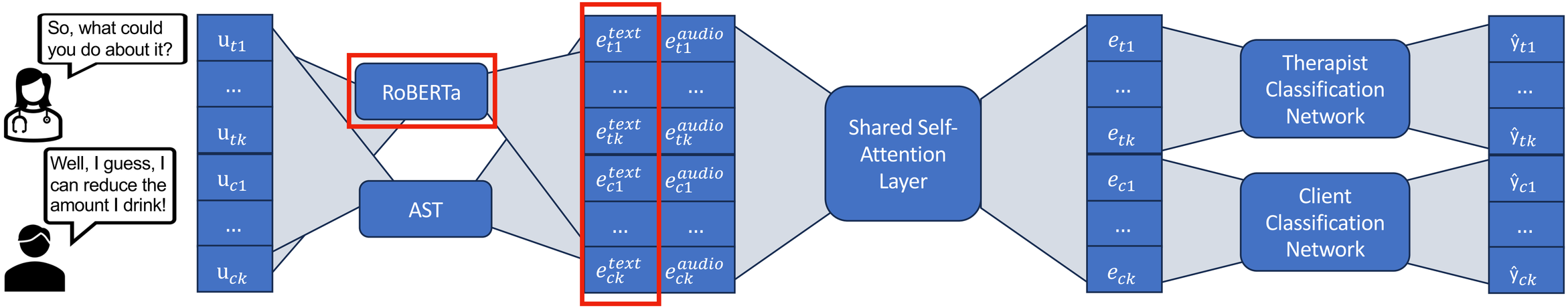
Input: k consecutive utterances (text, audio spectograms) of therapist u_{ti} and client u_{ci}

M³TCM Approach



Output: k consecutive utterance labels for therapist $\hat{y}_{ti}$ and client $\hat{y}_{ci}$

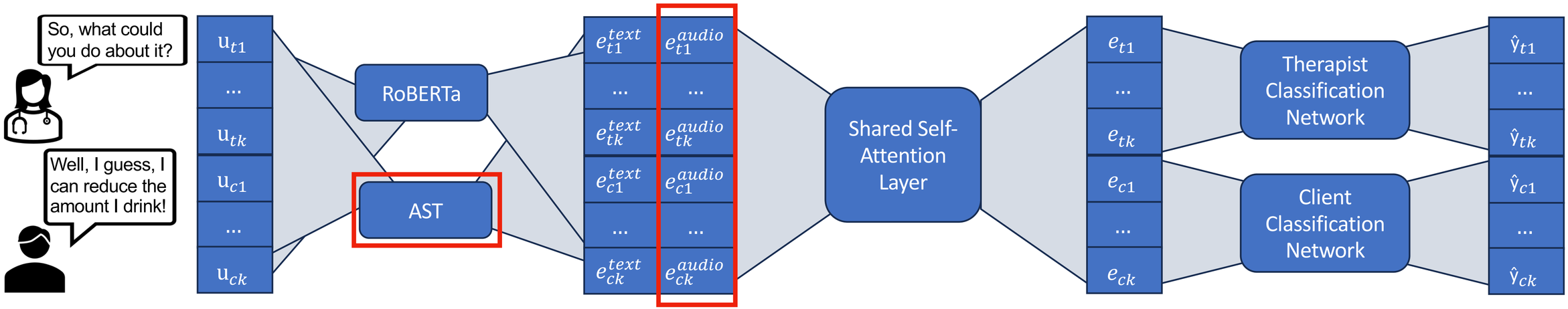
M³TCM Approach



Text Embeddings: RoBERTa Large, one embedding vector corresponding to each input utterance

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.

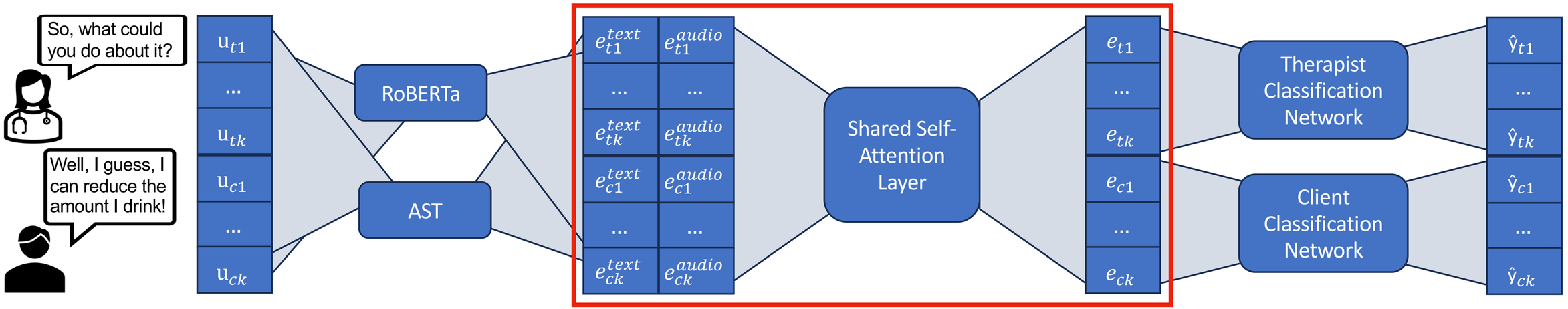
M³TCM Approach



Audio Embeddings: Audio Spectrogram Transformer (AST), one embedding vector corresponding to each input utterance

Gong, Y., Chung, Y. A., & Glass, J. (2021). Ast: Audio spectrogram transformer. arXiv preprint arXiv:2104.01778.

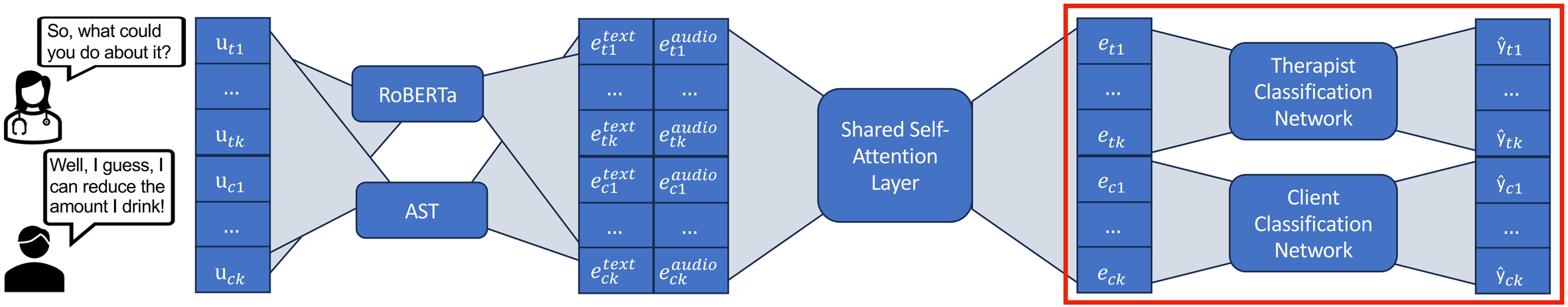
M³TCM Approach



Shared Layer: Self-attention layer operates across time and feature dimension on concatenated text, audio embeddings

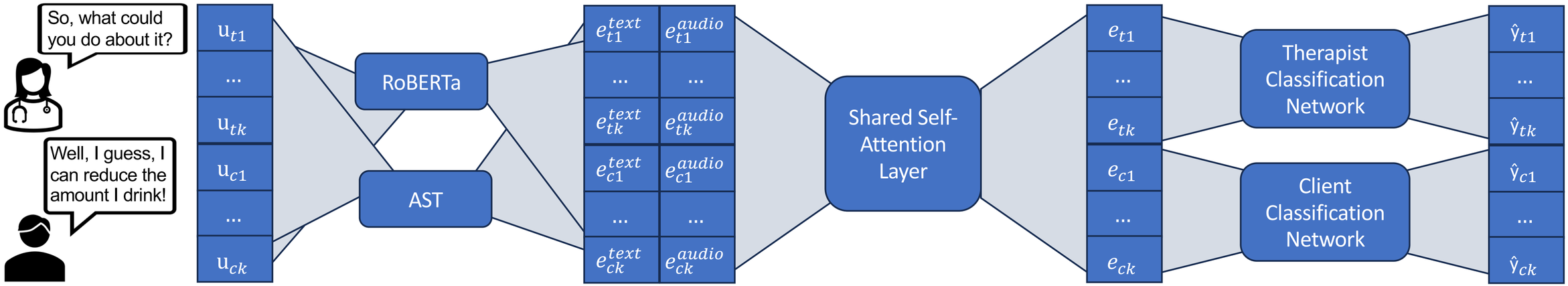
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.

M³TCM Approach



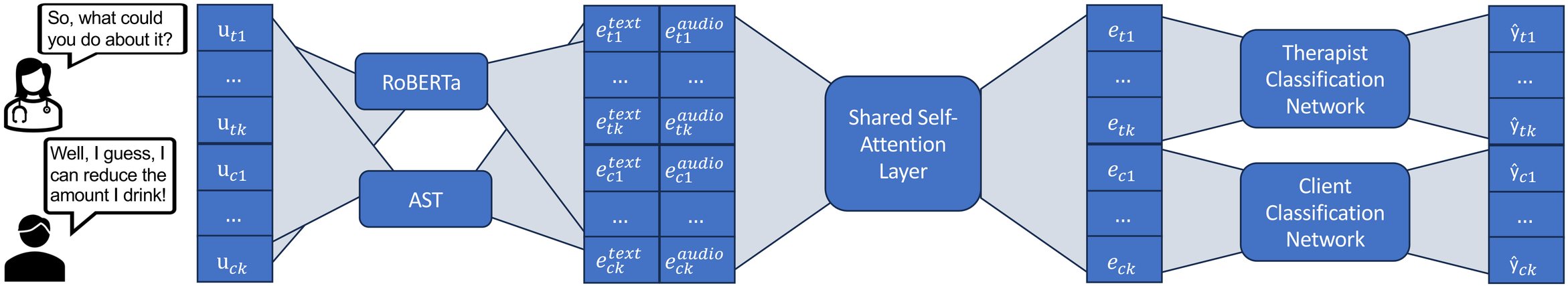
Classification Networks: Separate networks for therapist and client respectively

M³TCM Approach



Multi-Task learning takes place via **query-key interactions** across client and therapist utterances in the **shared self-attention layer**

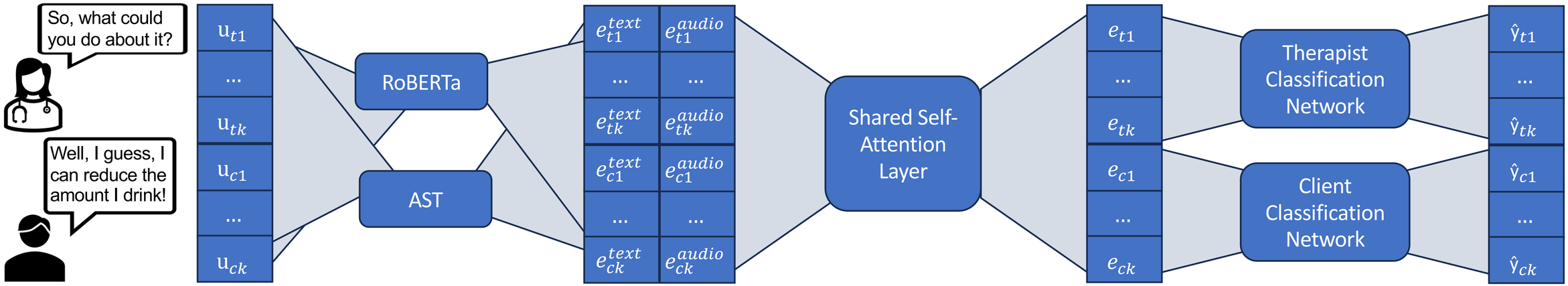
M³TCM Approach



To counteract the **unbalanced class distribution** common in utterance classification, we make use of the **focal loss**

Lin, T. Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal loss for dense object detection. CVPR'17.

M³TCM Approach



Training:

(1) Fine-tune RoBERTa Large and AST on AnnoMI
(2) Train full M3TCM model for 100 epochs; choose best model on val set

AnnoMI Dataset

133 unique youtube videos, >13k utterances

Topics: Reducing alcohol consumption, smoking cessation, weight loss, taking medicine, increasing physical activity,…

Therapist Behaviour

Reflection	Question	Input	Other
34 %	36 %	16 %	14 %

Client Behaviour

Change	Neutral	Sustain
29 %	57 %	14 %

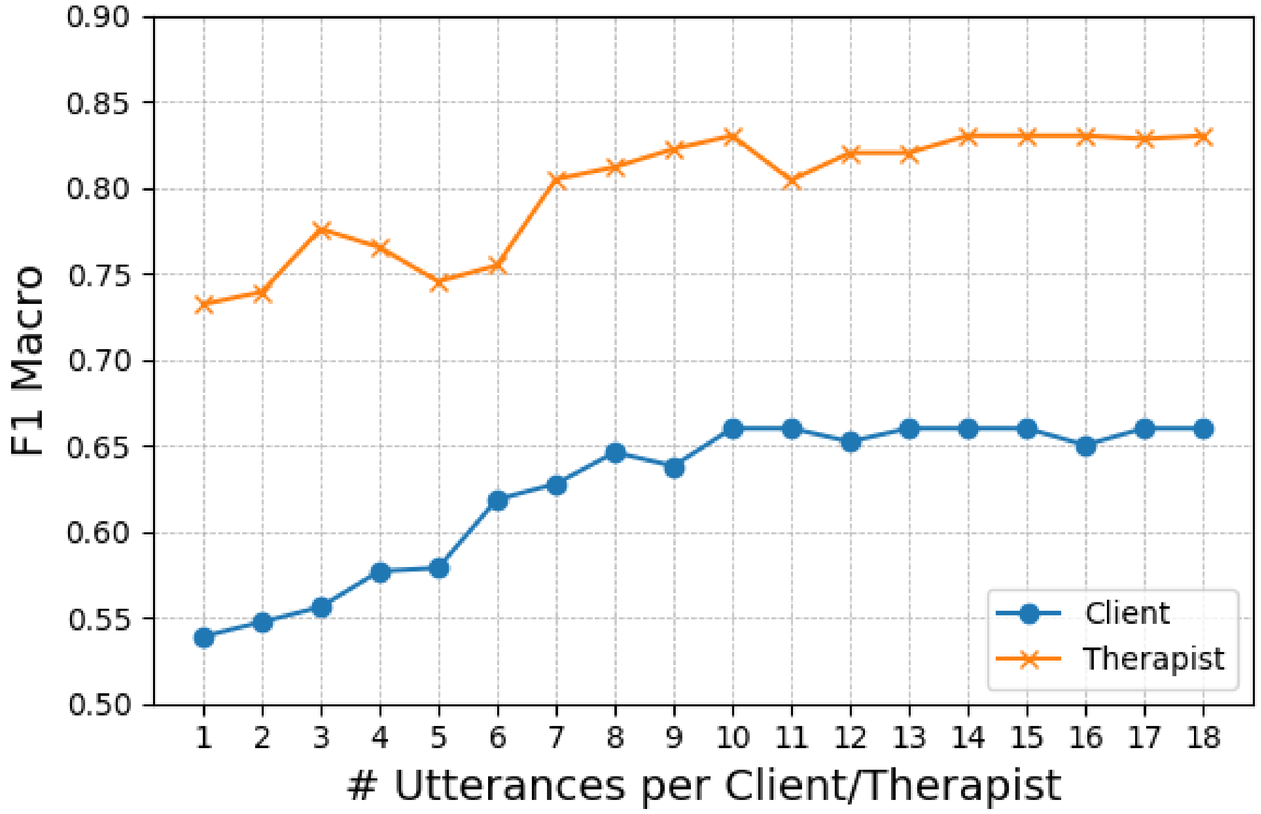
Wu, Z., Balloccu, S., Kumar, V., Helaoui, R., Reforgiato Recupero, D., & Riboni, D. (2023). Creation, analysis and evaluation of annomi, a dataset of expert-annotated counselling dialogues. Future Internet, 15(3), 110.

Results

Models	Client				Therapist				
	Average	Change	Neutral	Sustain	Average	Reflection	Question	Input	Other
Random Baseline	0.33	0.25	0.63	0.12	0.25	0.25	0.29	0.15	0.31
Wu et al. (2023a)	0.55	0.51	**0.74**	0.39	0.72	0.77	0.86	0.63	0.64
M^{3}TCM Without Finetuning	0.54	0.70	0.42	0.41	0.73	0.65	0.82	0.81	0.63
M^{3}TCM Text Only Single Task	0.58	0.76	0.56	0.43	0.77	0.73	0.86	0.82	0.68
M^{3}TCM Audio Only Single Task	0.40	0.65	0.38	0.18	0.44	0.40	0.60	0.44	0.31
M^{3}TCM Audio Only No Context	0.38	0.65	0.36	0.13	0.40	0.38	0.58	0.40	0.25
M^{3}TCM Text Only No Context	0.57	0.73	0.52	0.45	0.77	0.74	0.86	0.82	0.67
M^{3}TCM Audio Only	0.46	0.73	0.43	0.21	0.49	0.46	0.68	0.48	0.33
M^{3}TCM Text Only	0.63	0.80	0.59	0.49	0.80	0.76	0.89	0.85	0.68
M^{3}TCM No Context	0.61	0.78	0.57	0.48	0.76	0.70	0.83	0.87	0.65
M^{3}TCM Single Task	0.60	0.78	0.57	0.46	0.77	0.70	0.85	0.87	0.65
M^{3}TCM	**0.66**	**0.83**	0.62	**0.52**	**0.83**	**0.81**	**0.89**	**0.88**	**0.73**

Per-class and macro-averaged F1 scores

Results: Context size



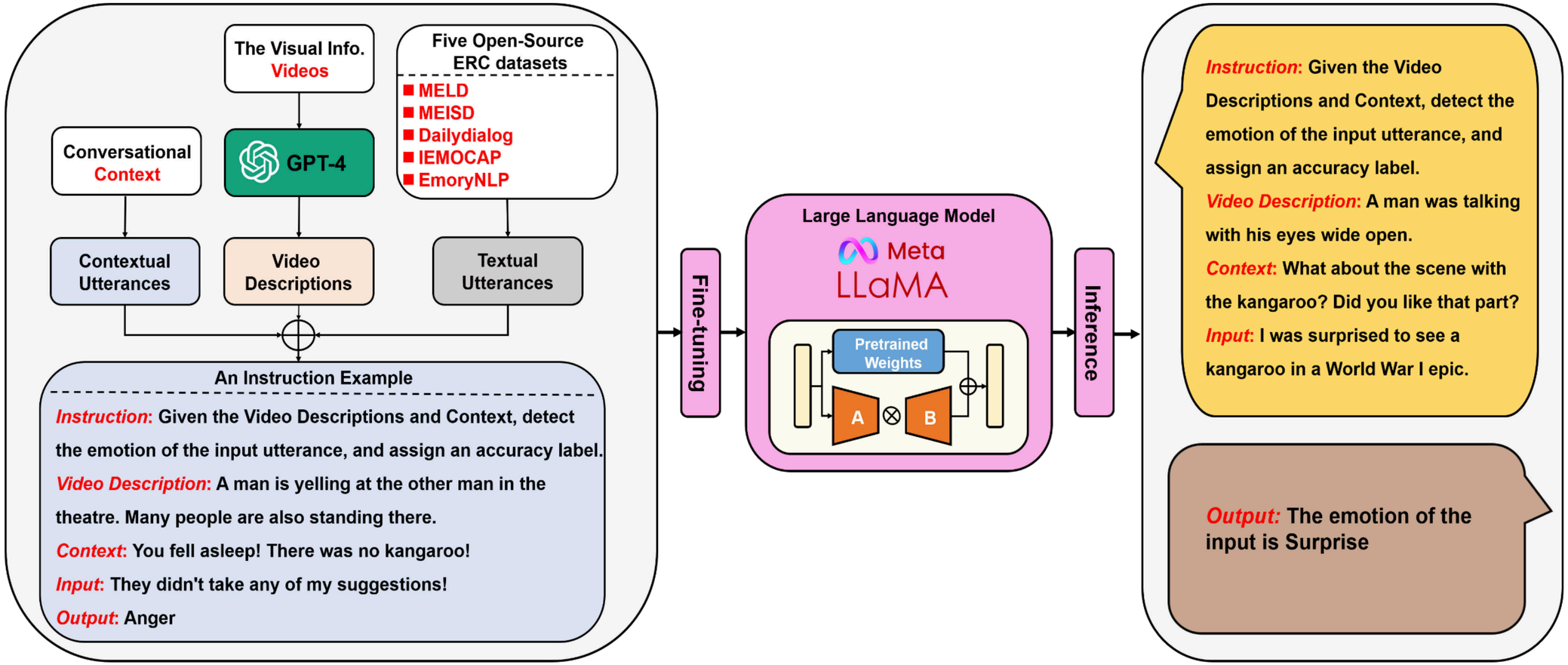
Conclusion

First approach to model three key aspects of utterance classification in motivational interviews: **multi-task, multi-modality, conversation context**

Clear improvements over previous SOTA and ablation conditions

Future Work: integrating video, application to further scenarios with asymmetric conversation roles

Multi-modal emotion recognition with LLMs



(a) Emotion and Context Knowledge (b) Supervised Fine-tuning LLM (c) Classification

Zhang, Y., Wang, M., Tiwari, P., Li, Q., Wang, B., & Qin, J. (2023). Dialoguellm: Context and emotion knowledge-tuned llama models for emotion recognition in conversations. arXiv preprint arXiv:2310.11374.

Regulated Emotions

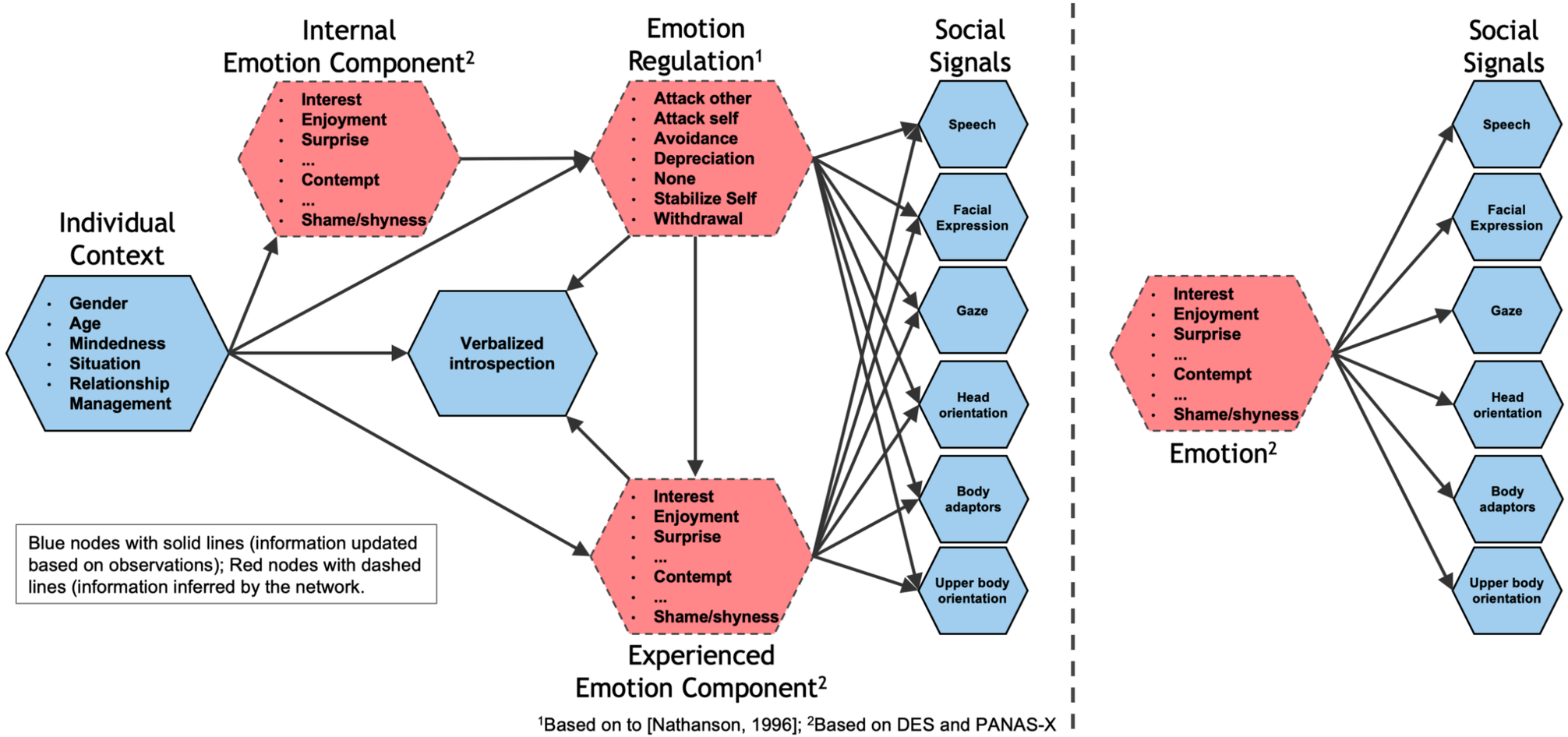
Image a job interview.

Interviewer: “Where did you buy your outfit? It really does not suit you.”

Interviewee: “Oh, haha I think I wanted to try out something different!”

Shame is usually not displayed directly, but regulated according to internal processes and social display rules.

The Deep Method for Modeling Shame



[1]Based on to [Nathanson, 1996]; [2]Based on DES and PANAS-X

Schneeberger, T., Hladký, M., Thurner, A. K., Volkert, J., Heimerl, A., Baur, T., ... & Gebhard, P. (2023). The Deep Method: Towards Computational Modeling of the Social Emotion Shame driven by Theory, Introspection, and Social Signals. IEEE Transactions on Affective Computing.

Deep Corpus

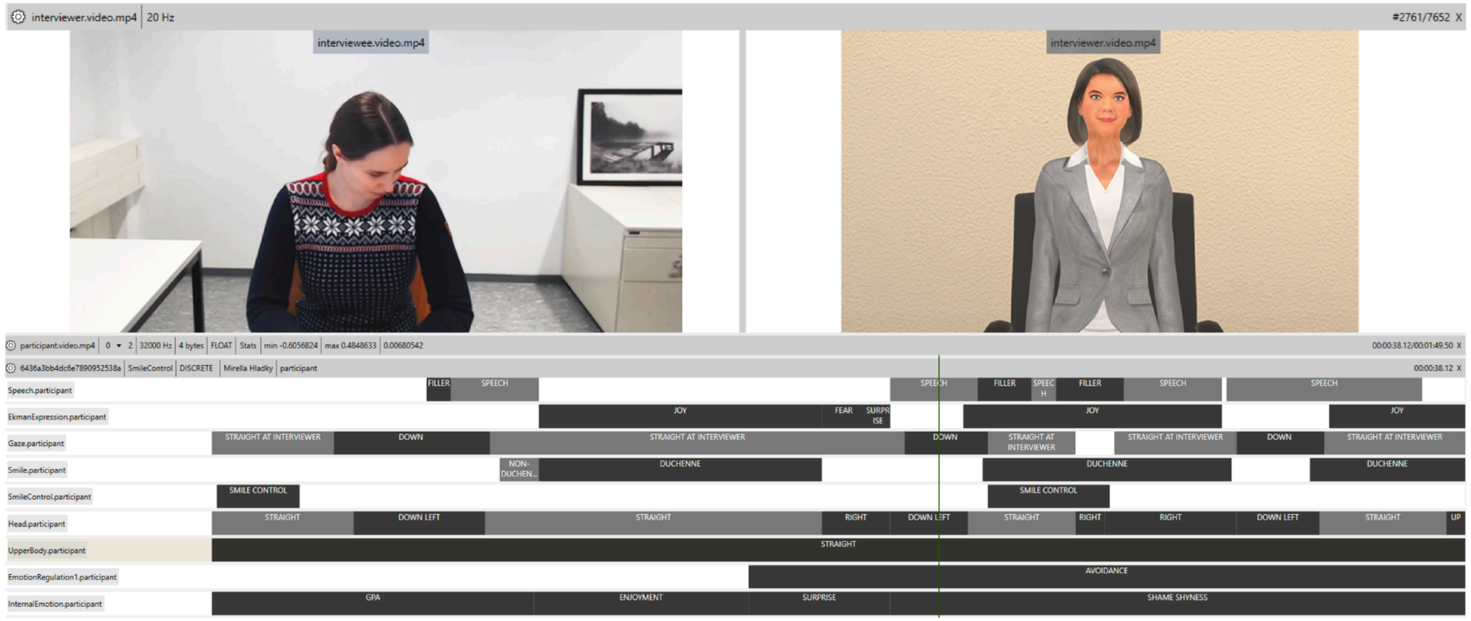
Human-agent job interviews

Ten participants

Two shame induction situations

Post-interaction interviews (“Verbalised Introspection’)

Expert annotations



Schneeberger, T., Hladký, M., Thurner, A. K., Volkert, J., Heimerl, A., Baur, T., ... & Gebhard, P. (2023). The Deep Method: Towards Computational Modeling of the Social Emotion Shame driven by Theory, Introspection, and Social Signals. IEEE Transactions on Affective Computing.

Deep Corpus: Regulation Strategies

Regulation strategies for the emotion shame

WITHDRAWAL (655 frames) Cut off the current situation so there is no more external influence or stimuli. Wish to hide, leave or escape.
Experienced emotional components: distress, fear
Nonverbal Behaviour: freezing, lip biting, gaze/head aversion, silence

ATTACK SELF (515 frames) Do to yourself what others may do to you, establishing impression to control the situation.
Experienced emotional components: disgust
Nonverbal Behaviour: facial expression of disgust

ATTACK OTHER (629 frames) Transfer the diminishment of self-esteem to the person (object) who caused it by diminishing the other person.
Experienced emotional components: anger
Nonverbal Behaviour: learn forward, gestures of power, facial expression of anger

AVOIDANCE (1650 frames) Acting according the principle "fool others, fool myself".
Experienced emotional components: joy
Nonverbal Behaviour: gaze/head aversion, lean backwards, facial expression of joy/surprise, smile

DEPRECIATION (1911 frames) Deevaluation of interaction partner due to different (or even contrary) values and ideals.
Experienced emotional components: disgust, contempt
Nonverbal Behaviour: raised eyebrows, smile, facial expression of disgust and contempt

STABILIZE SELF (3593 frames) Attempt to react in a way that is compliant with the (ideal) self by accepting disagreement between job interviewer and person.
Experienced emotional components: pride
Nonverbal Behaviour: no display of uncertainty, direct gaze

Schneeberger, T., Hladký, M., Thurner, A. K., Volkert, J., Heimerl, A., Baur, T., ... & Gebhard, P. (2023). The Deep Method: Towards Computational Modeling of the Social Emotion Shame driven by Theory, Introspection, and Social Signals. IEEE Transactions on Affective Computing.

Deep Corpus: Annotations

Nonverbal Behaviour	Observation of external components of emotions that are encoded in social signals in the specific situation. *Speech, Utterance, Facial Expression, Gaze, Eyes, Smile, Smile Control, Head, Head Tilt, Upper body, Shame display*
Verbalized introspection	Self-reports that reflect a person's subjective experience gathered in semi-structured interviews after the specific situation with the aid of video material of the experienced situation. *Relationship management, Shame awareness, Experienced emotion, Internal emotion component, Display rule*
Personal Context	Personal context variables. *Gender, Mindedness score*
Situational Context	Situational context variables. *Situation (first vs. second shame induction), Conversation transcript*

Schneeberger, T., Hladký, M., Thurner, A. K., Volkert, J., Heimerl, A., Baur, T., ... & Gebhard, P. (2023). The Deep Method: Towards Computational Modeling of the Social Emotion Shame driven by Theory, Introspection, and Social Signals. IEEE Transactions on Affective Computing.

Our Approach

Recognising Emotion Regulation Strategies from Human Behaviour with Large Language Models

Philipp Müller
DFKI
Saarbrücken, Germany
philipp.mueller@dfki.de

Alexander Heimerl
Augsburg University
Augsburg, Germany
alexander.heimerl@uni-a.de

Sayed Muddashir Hossain
DFKI
Saarbrücken, Germany
sayed_muddashir.hossain@dfki.de

Lea Siegel
DFKI
Saarbrücken, Germany
lea.siegel@dfki.de

Jan Alexandersson
DFKI
Saarbrücken, Germany
jan.alexandersson@dfki.de

Patrick Gebhard
DFKI
Saarbrücken, Germany
patrick.gebhard@dfki.de

Elisabeth André
Augsburg University
Augsburg, Germany
elisabeth.andre@uni-a.de

Tanja Schneeberger
DFKI
Saarbrücken, Germany
tanja.schneeberger@dfki.de

Accepted at ACII'24

Müller, P., Heimerl, A., Hossain, S. M., Siegel, L., Alexandersson, J., Gebhard, P., ... & Schneeberger, T. (2024). Recognizing Emotion Regulation Strategies from Human Behavior with Large Language Models. arXiv preprint arXiv:2408.04420.

Our Approach

Construct multi-modal prompts

Instruction-tune LLM with Low Rank Adaptation (LoRA) to predict Emotion Regulation strategy

We utilised Llama2-7b and Gemini models

Müller, P., Heimerl, A., Hossain, S. M., Siegel, L., Alexandersson, J., Gebhard, P., ... & Schneeberger, T. (2024). Recognizing Emotion Regulation Strategies from Human Behavior with Large Language Models. arXiv preprint arXiv:2408.04420.

Prompts

Situational Context:

We are concerned with a moment in time in the first shame induction situation. The agent tries to induce shame by attacking the interviewee's personal attractiveness: "Before we start, one short question: Where did you get this outfit? Somehow it doesn't really suit you."

The conversation history up to the current point is:
[Avatar] Where did you get this outfit from?
[Avatar] Somehow it doesn't really suit you.
[Interviewee] Don't you like it so much?
[Interviewee] I thought I felt very comfortable in it, and I find that when you feel comfortable, you always sell yourself a bit better and in the application situation I thought that makes the most sense.

The current utterance is:
[Interviewee] I thought I felt very comfortable in it, and I find that when you feel comfortable, you always sell yourself a bit better and in the application situation I thought that makes the most sense.

Nonverbal Behavior:

The interviewee shows the following nonverbal behavior at the current moment: The interviewee looks straight at the interviewer. The interviewee holds their head straight. The interviewee tilts their head to the side. The interviewee shows a non-Duchenne smile, i.e. a smile that concentrates only on the mouth. The interviewee is speaking. The upper body is moved forwards

Verbalized Introspection:

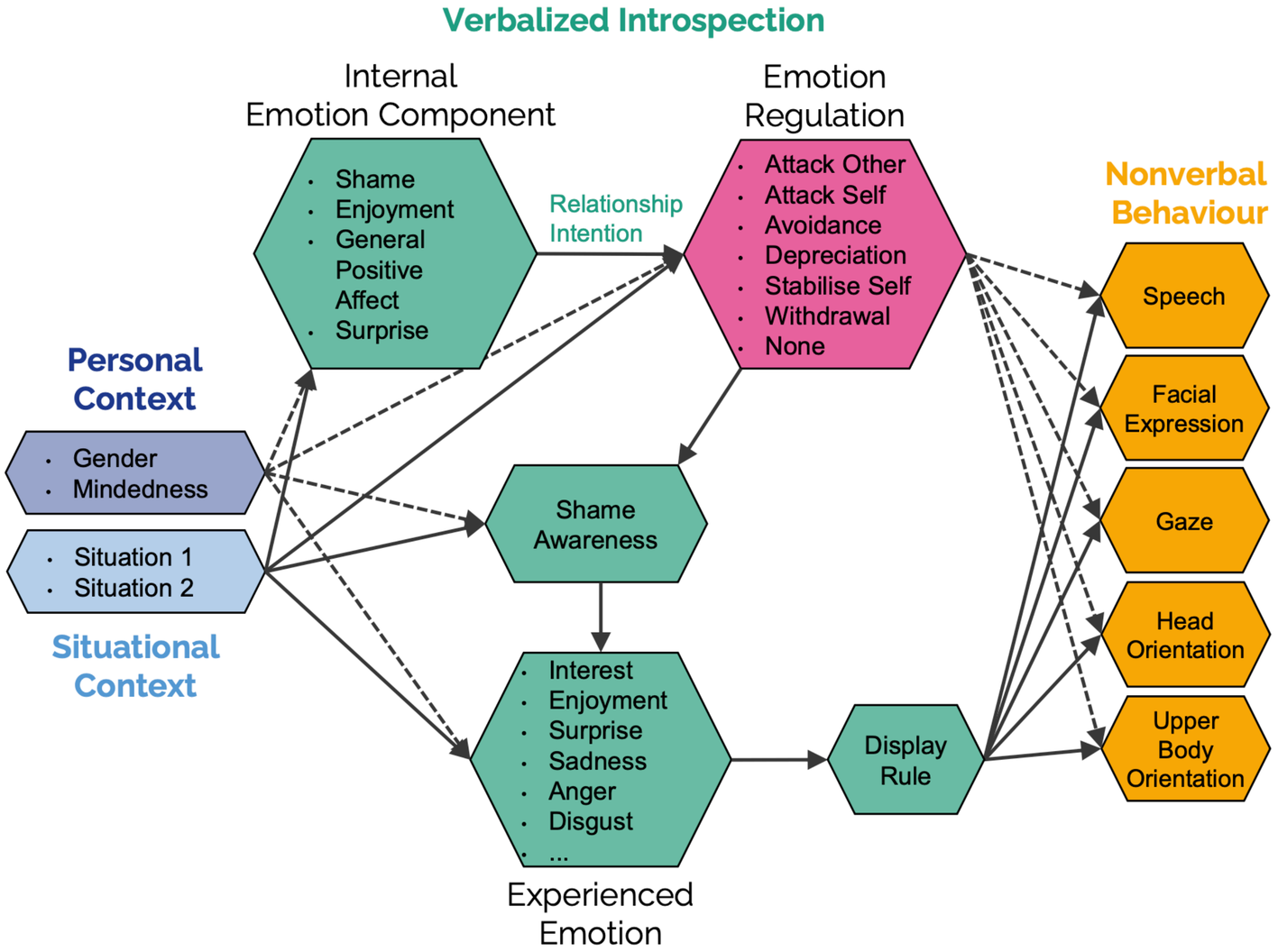
The following information was gathered from the qualitative interview after the interaction: The interviewee experiences the following internal emotion at the current moment in time: shame/shyness. The interviewee was aware of feeling ashamed during the current moment in the job interview. During the qualitative interview, the interviewee became aware that they were having the emotion shame during the current moment in the job interview. The interviewee has the intention to maintain the relationship with the avatar.

Personal Context:

The following additional personal information was collected from the interviewer: The mindedness score of the interviewee is 4,77. The interviewee is female.

Müller, P., Heimerl, A., Hossain, S. M., Siegel, L., Alexandersson, J., Gebhard, P., ... & Schneeberger, T. (2024). Recognizing Emotion Regulation Strategies from Human Behavior with Large Language Models. arXiv preprint arXiv:2408.04420.

Bayesian Network Baseline



Müller, P., Heimerl, A., Hossain, S. M., Siegel, L., Alexandersson, J., Gebhard, P., ... & Schneeberger, T. (2024). Recognizing Emotion Regulation Strategies from Human Behavior with Large Language Models. arXiv preprint arXiv:2408.04420.

Results

	Withdrawal		Attack self		Attack other		Avoidance		Depreciation		Stabilize self		Rest		Overall	
	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1
w/ verb. introspection																
Bayesian Net	**0.99**	**0.94**	**0.99**	**0.91**	**0.99**	**0.93**	**0.99**	**0.99**	**0.98**	**0.94**	**0.98**	**0.98**	0.96	0.91	**0.96**	**0.96**
Gemma	0.98	0.83	**0.99**	0.86	0.98	0.86	0.98	0.92	**0.98**	0.93	0.97	0.95	**0.98**	**0.95**	0.93	0.93
Llama2-7B	**0.99**	0.88	0.97	0.69	0.98	0.84	0.98	0.94	0.96	0.89	0.95	0.92	0.95	0.88	0.89	0.89
w/o verb. introspection																
Bayesian Net	0.81	0.21	0.88	0.0	0.89	0.08	0.79	0.33	0.65	0.13	0.69	0.34	0.72	0.26	0.23	0.25
Gemma	0.94	0.56	0.94	0.55	0.94	0.57	0.92	0.70	0.90	0.70	0.88	0.78	0.90	0.76	0.71	0.72
Llama2-7B	**0.97**	**0.76**	**0.97**	**0.71**	**0.96**	**0.71**	**0.96**	**0.85**	**0.95**	**0.84**	**0.93**	**0.88**	**0.95**	**0.88**	**0.84**	**0.84**

Müller, P., Heimerl, A., Hossain, S. M., Siegel, L., Alexandersson, J., Gebhard, P., ... & Schneeberger, T. (2024). Recognizing Emotion Regulation Strategies from Human Behavior with Large Language Models. arXiv preprint arXiv:2408.04420.

Results

Input Modalities	Bayesian Net ACC	Bayesian Net F1	Llama2-7B ACC	Llama2-7B F1	Gemma ACC	Gemma F1
w/ verb. introspection						
All	**0.96**	**0.96**	0.89	0.86	0.93	0.93
No personal context	0.69	0.68	0.88	0.88	0.93	0.93
No situational context	0.84	0.85	0.49	0.51	0.61	0.63
No transcript	—	—	0.45	0.47	0.63	0.64
No nonverbal behavior	0.06	0.01	0.87	0.87	0.87	0.87
Only verbalized introspection	0.17	0.16	0.54	0.56	0.54	0.56
w/o verb. introspection						
All	0.23	0.25	**0.84**	**0.84**	0.71	0.72
No personal context	0.26	0.27	0.44	0.46	0.45	0.47
No situational context	0.22	0.23	0.38	0.40	0.35	0.38
No transcript	—	—	0.40	0.42	0.34	0.37
No nonverbal behavior	0.25	0.28	0.42	0.44	0.44	0.46
Only nonverbal behavior	0.25	0.25	0.47	0.50	0.44	0.46

Müller, P., Heimerl, A., Hossain, S. M., Siegel, L., Alexandersson, J., Gebhard, P., ... & Schneeberger, T. (2024). Recognizing Emotion Regulation Strategies from Human Behavior with Large Language Models. arXiv preprint arXiv:2408.04420.

Conclusion

LLMs appear to be able to make use of verbal and nonverbal behaviour to recognise emotion regulation strategy

Open question (at least for me): To what extent are they really combining their general “world knowledge” with the prompt to solve the fine-tuning task? E.g. do they use knowledge about nonverbal behaviour?

Llama without instruction tuning extremely biased

Small dataset - generalisation might be limited

Müller, P., Heimerl, A., Hossain, S. M., Siegel, L., Alexandersson, J., Gebhard, P., ... & Schneeberger, T. (2024). Recognizing Emotion Regulation Strategies from Human Behavior with Large Language Models. arXiv preprint arXiv:2408.04420.

Future Work

Improving the link between text and nonverbal behaviour

Be able to answer questions about nonverbal behaviour in context

How to make machine interpretation of your behaviour transparent to you?

Transfer to low-resource scenarios, e.g. medical data in Georgian

Link physiological data to conversation context

Establishing common evaluation datasets and protocols

ACM Multimedia Grand Challenge “MultiMediate”

MultiMediate’21:

- Eye Contact Detection
- Next Speaker Prediction

MultiMediate’22:

- Backchannel Detection
- Agreement Estimation from Backchannel

MultiMediate’23:

- Bodily Behaviour Recognition
- Engagement Estimation

MultiMediate’24:

- Multi-domain Engagement Estimation

https://multimediate-challenge.org/

German Research Center for Artificial Intelligence
Universität Augsburg University
University of Stuttgart Germany
Inria

Thank you!

Brainstorming Groups: Business Application

Come up with an application case for multi-modal conversation analysis

Deliver a 3-minute pitch in which you present your ideas to the audience.

- What problem does it solve / what is the added value of your application?
- What technical challenges must be overcome (e.g. models, data, business aspects,…) in order to make your application a reality?
- How would you plan to overcome these challenges?

Brainstorming Groups: Research

Come up with a research idea combining today's keynote talk with what you learned about multi-modal conversation analysis.

Deliver a 3-minute pitch in which you present your ideas to the audience.

- What is the research question, i.e. what knowledge are we lacking or what technical problem needs to be addressed?
- What experiments need to be conducted in order to find an answer to your question or to address the technical problem?
- What data do you need for it? Do you need to record new data or are there existing datasets that you can use?