



Мультимодальные методы распознавания эмоций

Андрей Савченко

Научный директор лаборатории ИИ

02.08.2024

ПЛАН

0 Введение: представление эмоций, модальности,
1 датасеты

02 Распознавание выражений лиц по видео

03 Распознавание эмоций по голосу

04 Мультимодальные методы

05 Демо и примеры

Андрей Савченко

Про докладчика

С 20 лет - в разработке ПО, путь от стажера до тимлида

В 22 года пошел в аспирантуру Вышки в НН, заинтересовался распознаванием образов

В 24 года досрочно защитил кандидатскую диссертацию, вскоре ушел из индустрии и сосредоточился на науке

В 2016 г защитил докторскую диссертацию

Руководил проектами по CV от Samsung, Huawei, Сбер,...

Теперь я представляю:



Лаборатория
искусственного
интеллекта



0 1 Введение: представление эмоций, модальности, датасеты

02 Распознавание выражений лиц по видео

03 Распознавание эмоций по голосу

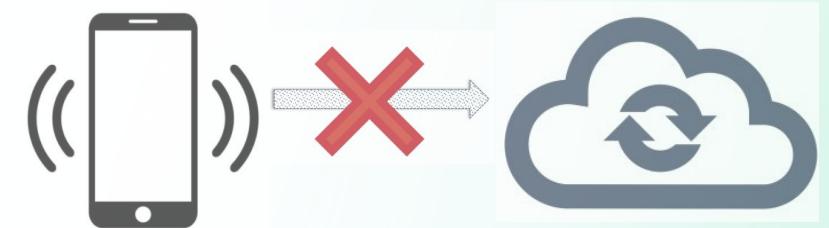
0 4 Мультимодальные методы

0 5 Демо и примеры

Мотивация



- Анализ реакций пользователя в HCI (Human Computer Interaction)
- Тестирование восприятия нового контента (реклама, юмор, ...)
- ТВ шоу, дебаты
- Онлайн-конференции, e-learning



Выражения лиц

Базовые эмоции Экмана



Fear

Contempt
©

Sadness

Happiness



Surprise

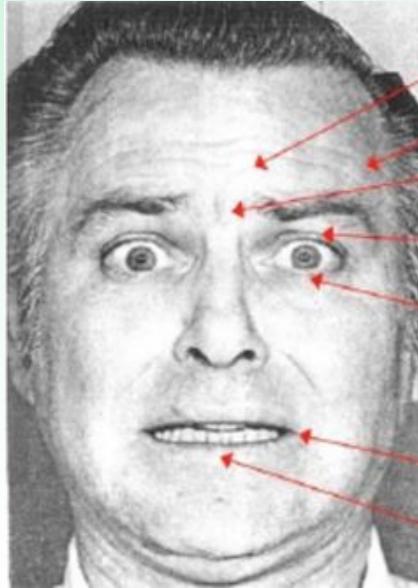
Anger

Disgust

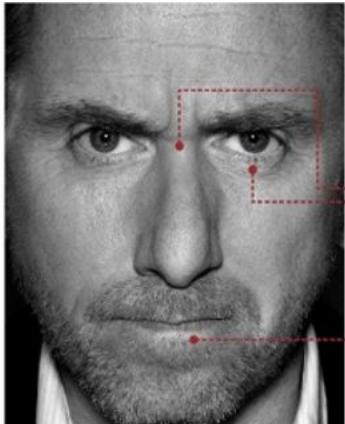


Пол Экман – прототип доктора Лайтмана («Обмани меня»)

FACS (Facial Action unit Coding System)

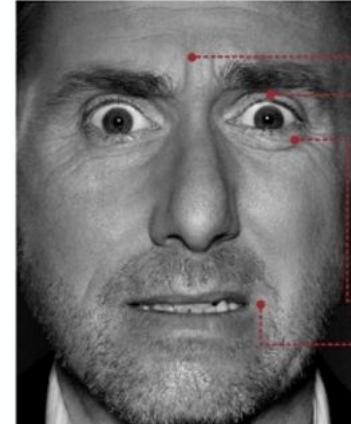


- 1C Inner brow raise
- 2C Outer brow raise
- 4B Corrugator
- 5D Upper lid raise
- 7B Lower lid tighten
- 20B Lip stretch
- 26B Jaw drop



anger

- ① eyebrows down and together
- ② eyes glare
- ③ narrowing of the lips



fear

- ① eyebrows raised and pulled together
- ② raised upper eyelids
- ③ tensed lower eyelids
- ④ lips slightly stretched horizontally back to ears

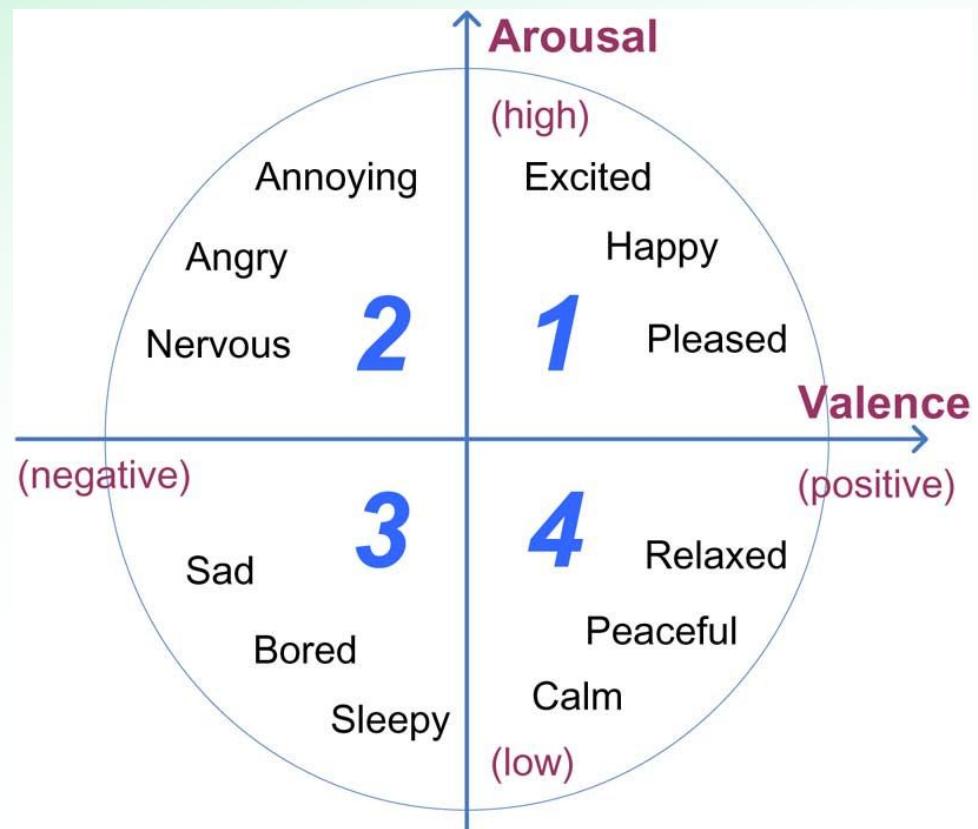
Upper Face Action Units					
AU 1	AU 2	AU 4	AU 5	AU 6	AU 7
Inner Brow Raiser	Outer Brow Raiser	Brow Lowerer	Upper Lid Raiser	Cheek Raiser	Lid Tightener
*AU 41	*AU 42	*AU 43	AU 44	AU 45	AU 46
Lid Droop	Slit	Eyes Closed	Squint	Blink	Wink

Lower Face Action Units					
AU 9	AU 10	AU 11	AU 12	AU 13	AU 14
Nose Wrinkler	Upper Lip Raiser	Nasolabial Deepener	Lip Corner Puller	Cheek Puffer	Dimpler
AU 15	AU 16	AU 17	AU 18	AU 20	AU 22
Lip Corner Depressor	Lower Lip Depressor	Chin Raiser	Lip Puckerer	Lip Stretcher	Lip Funneler

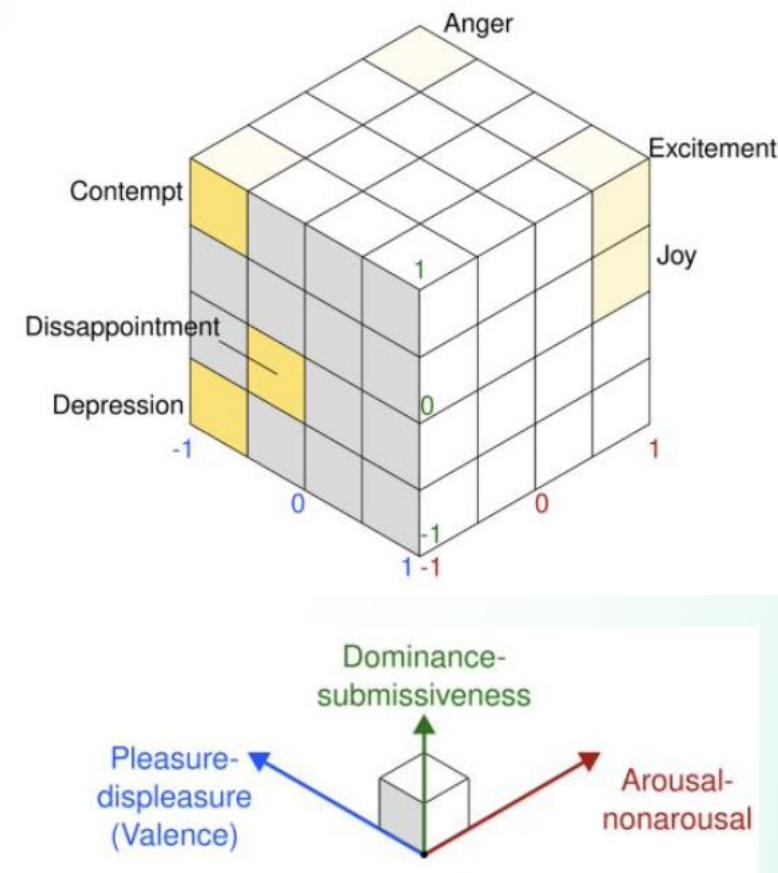
AU 23	AU 24	AU 25	AU 26	AU 27	AU 28
Lip Tightener	Lip Pressor	Lips Part	Jaw Drop	Mouth Stretch	Lip Suck

Модели Рассела

Circumplex model



PAD (pleasure, arousal, dominance)



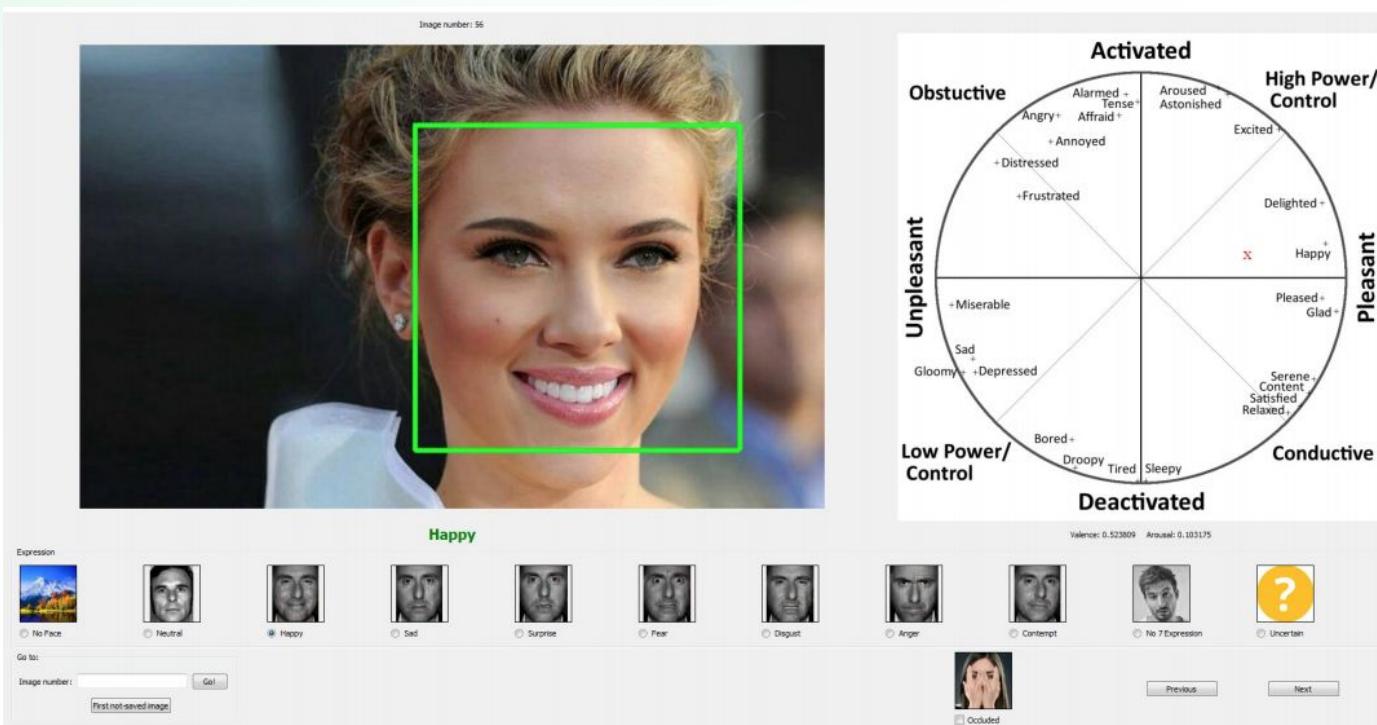
Модальности



Наборы данных

AffectNet

- Число категорий: 8 (Neutral, Happy, Sad, Surprise, Fear, Anger, Disgust, Contempt) или 7 (те же без Contempt)
- Обучающее множество: 287,651 фотографий
- Тестовое множество: 500 фотографий каждого класса

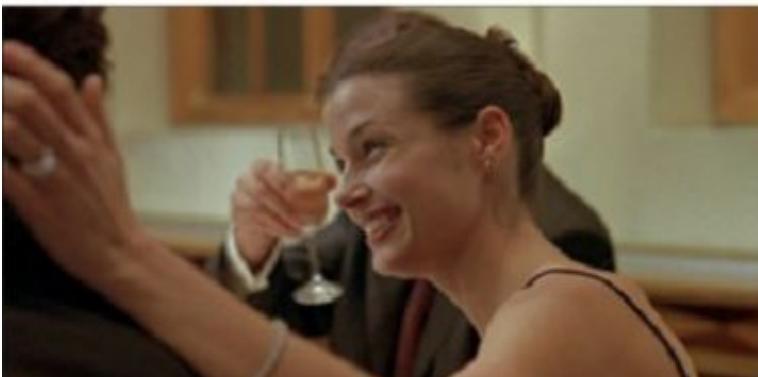
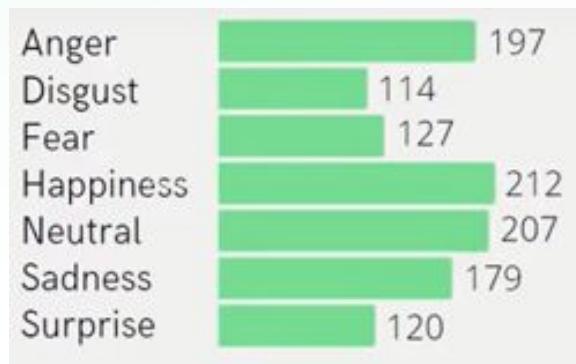


<http://mohammadmahoor.com/affectnet>

Наборы данных

AEFW: Acted Faces in-the-Wild

- Число категорий: 7 (Neutral, Happy, Sad, Surprise, Fear, Anger, Disgust)
- Обучающее множество: 773 коротких видео клипов
- Тестовое множество: 383 видео



<https://sites.google.com/view/emotiw2019>

RAMAS

The Russian Acted Multimodal Affective Set

- 564 видео
- 10 участников (5 мужчин и 5 женщин)
- Число классов: 7 – 6 базовых эмоций (Anger, Sadness, Disgust, Happiness, Fear, Surprise) и Neutral
- 5 аннотаций доступно для каждого видео. Каждый разметчик указывала время начала и конца эмоции
- Протокол тестирования (E. Ryumina et al., Interspeech 2021): выбирается порог n_a для числа согласившихся друг с другом разметчиков



<https://peerj.com/preprints/26688/>

Наборы данных

ABAW: Affective Behaviour Analysis in-the-Wild

1. **Выражения лиц:** 8 классов (anger, disgust, fear, happiness, sadness, surprise, neutral, other):

- обучающее множество: 585,317 кадров
- валидационное множество: 280,532 кадров



2. **AU (Action Units):** 12 action units (AU1, AU2, AU4, AU6, AU7, AU10, AU12, AU15, AU23, AU24, AU25, AU26)

- обучающее множество: 1,356,861 кадров
- валидационное множество: 445,836 кадров



3. **VA (Valence-Arousal):**

- обучающее множество: 1,555,919 кадров
- валидационное множество: 338,755 кадров



4. **Emotional Mimicry Intensity Estimation** (multi-output regression): 7 классов (Admiration, Amusement, Determination, Empathic Pain, Excitement, Joy).

- обучающее множество: 8072 видео
- тестовое множество: 4588 видео



<https://affective-behavior-analysis-in-the-wild.github.io/6th/>

01 Введение: представление эмоций, модальности, датасеты

02 Распознавание выражений лиц по видео

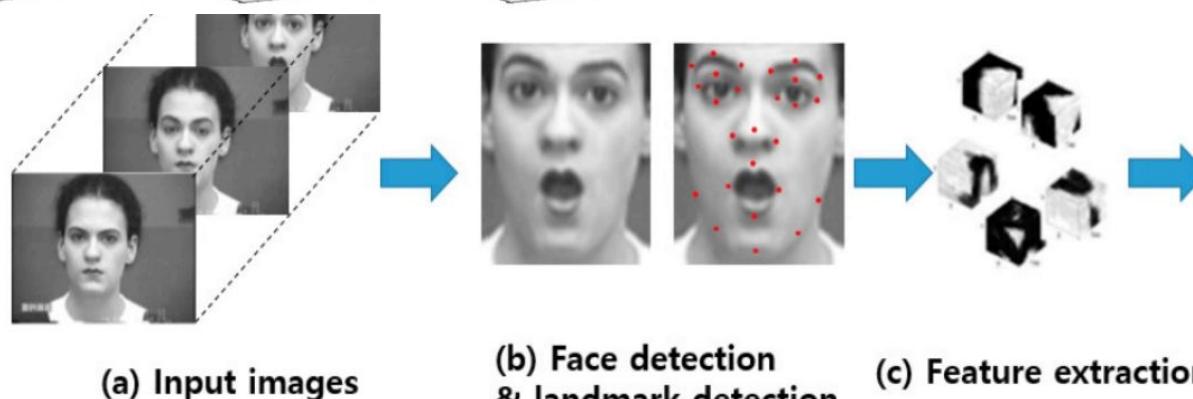
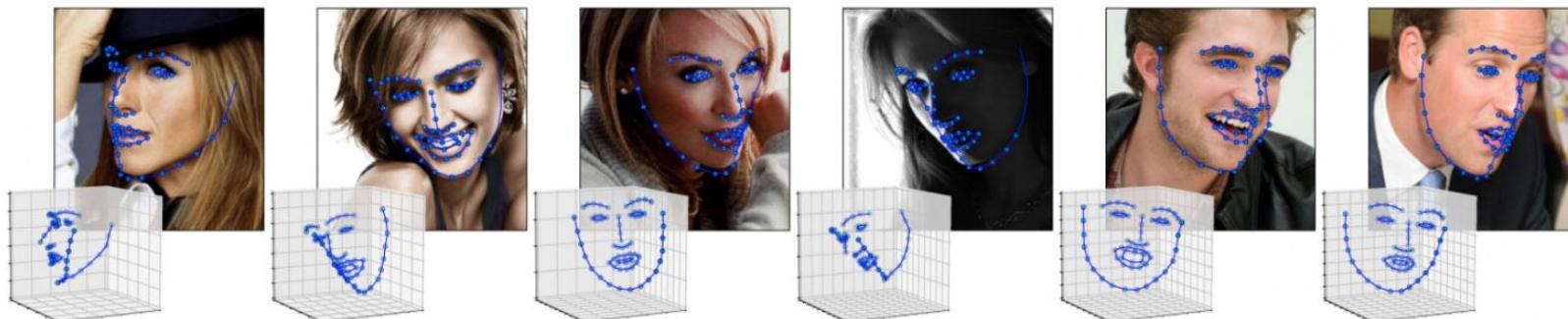
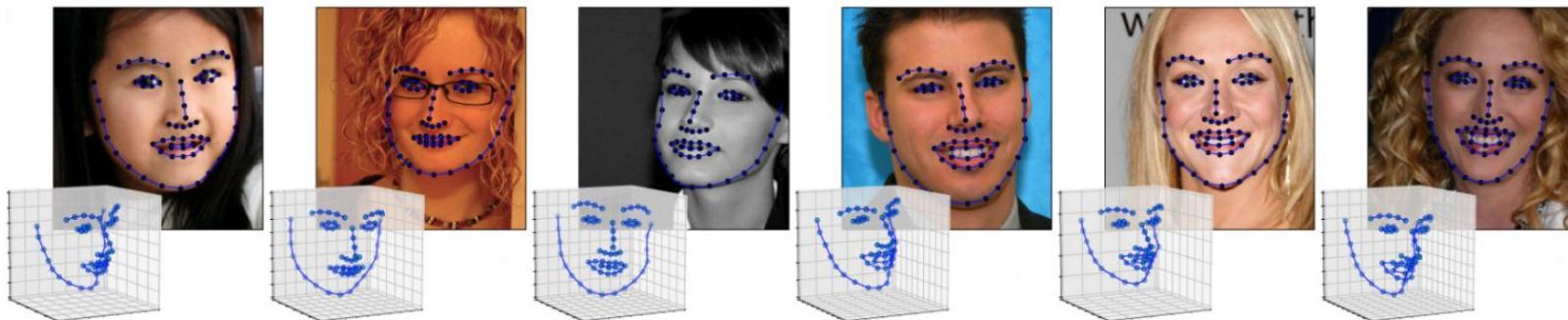
03 Распознавание эмоций по голосу

04 Мультимодальные методы

05 Демо и примеры

Алгоритмы

Традиционный подход: ключевые точки

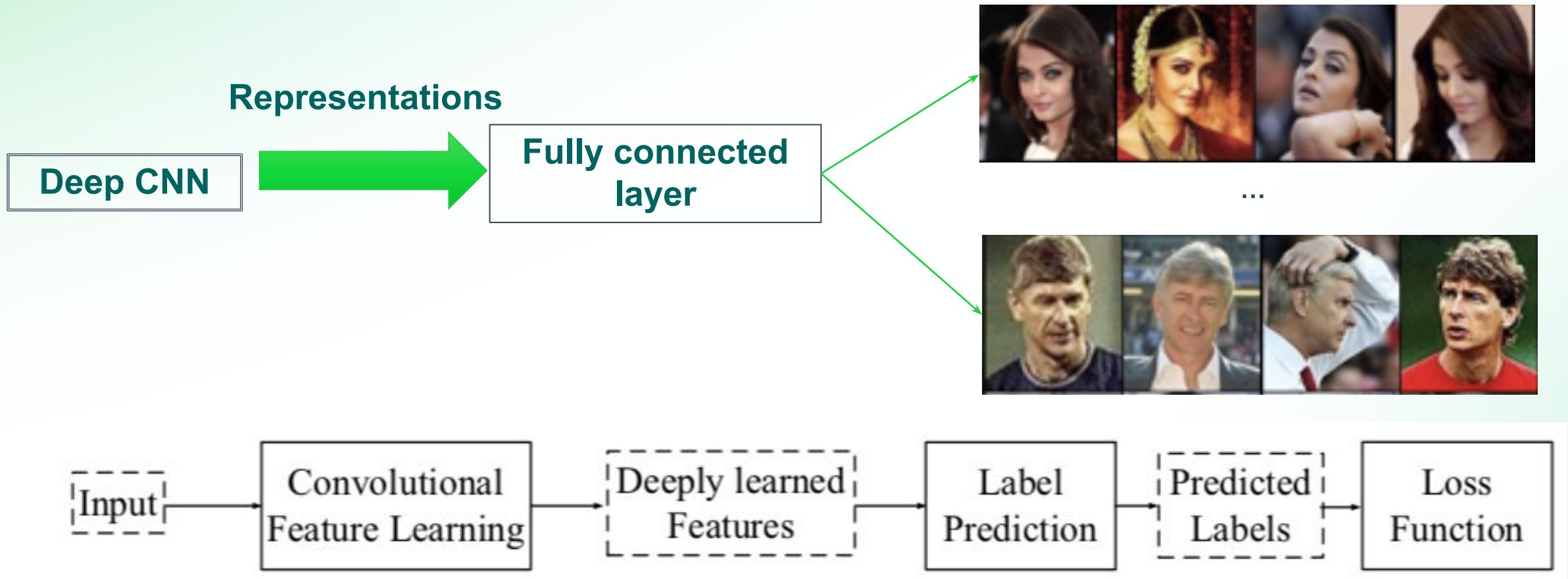


<https://habr.com/ru/companies/speechpro/articles/418151/>

Алгоритмы

Deep learning

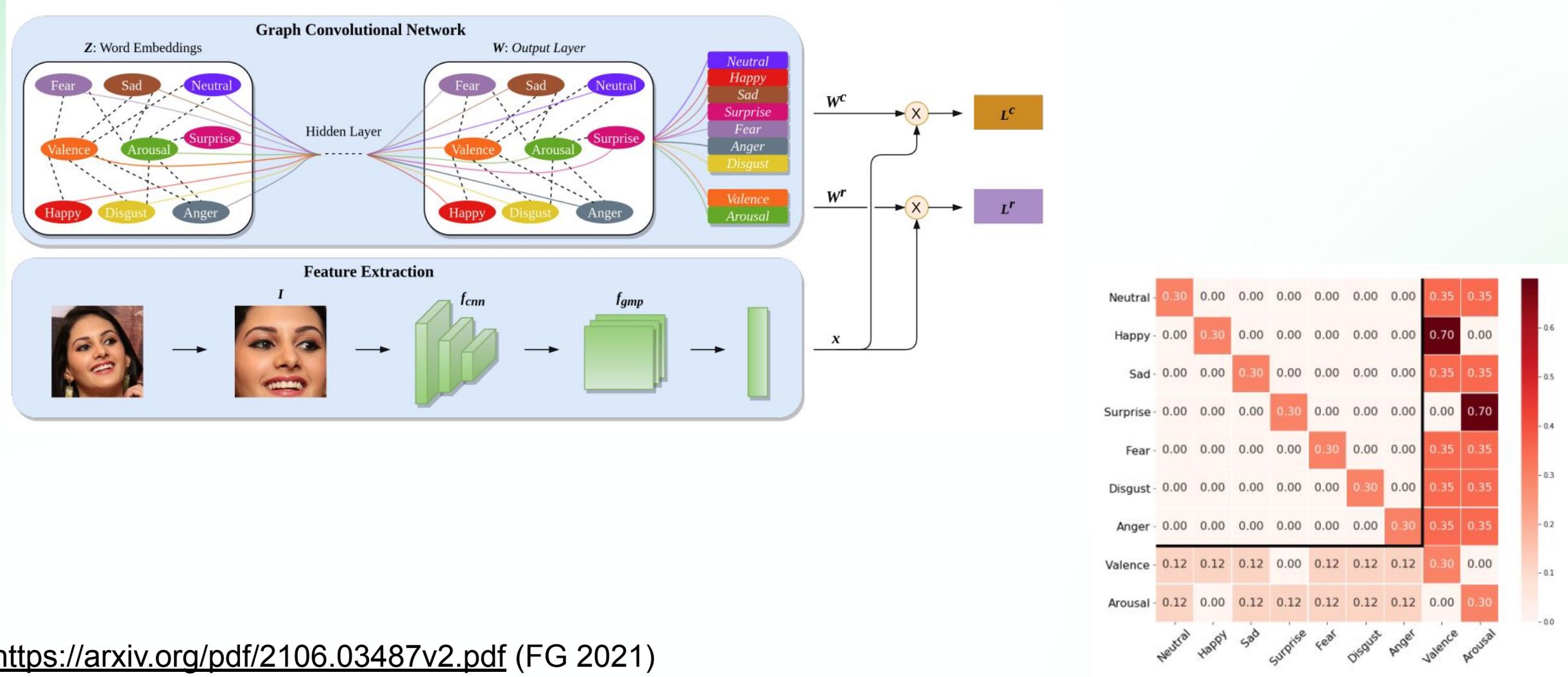
Большой набор данных (VGGFace2, CASIA-WebFace, MS-Celeb-1M,...)



[Wen et al, ECCV 2016]

Распознавание выражений лиц по фото

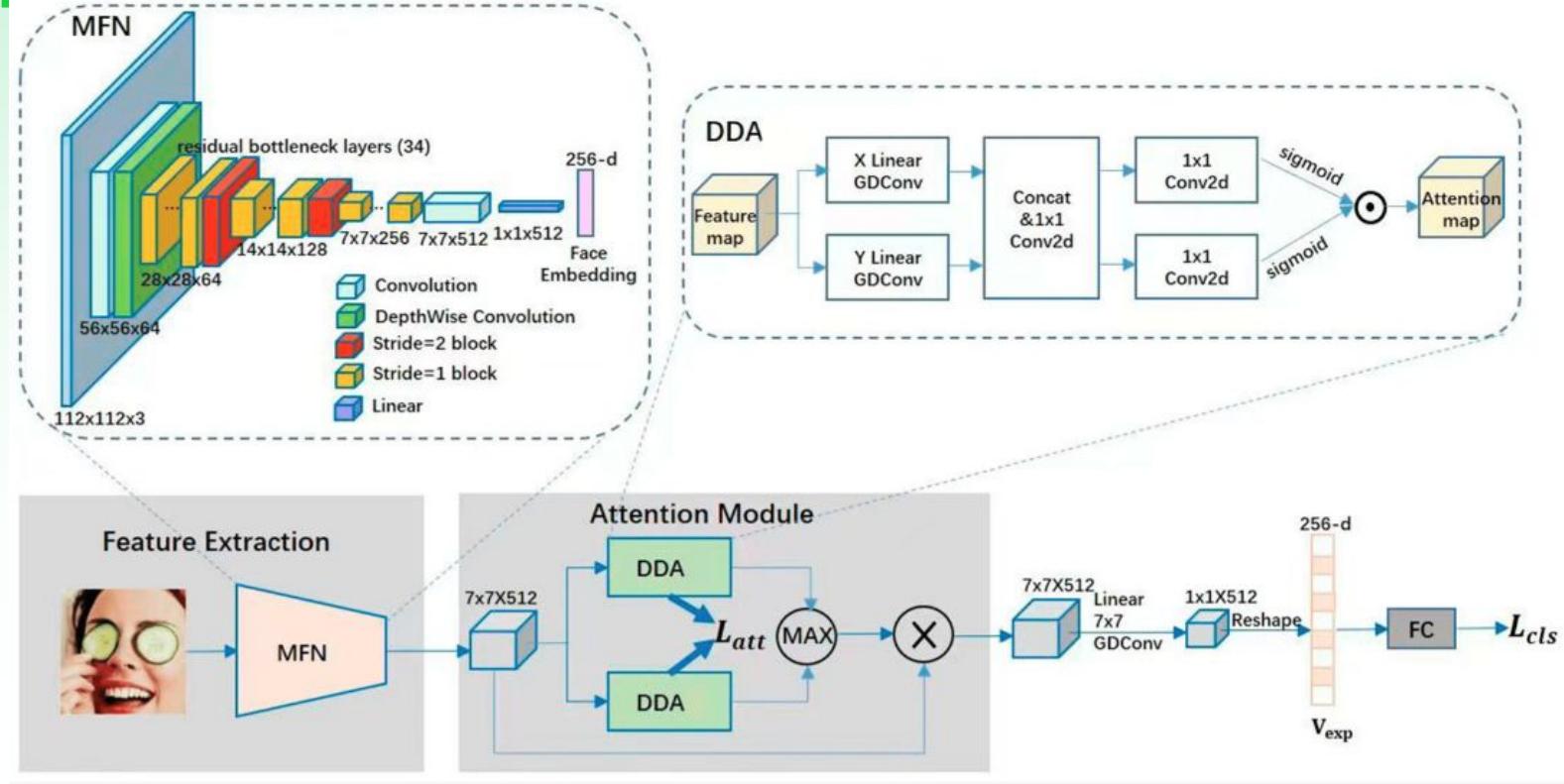
EmotionGCN



<https://arxiv.org/pdf/2106.03487v2.pdf> (FG 2021)

Распознавание выражений лиц по фото

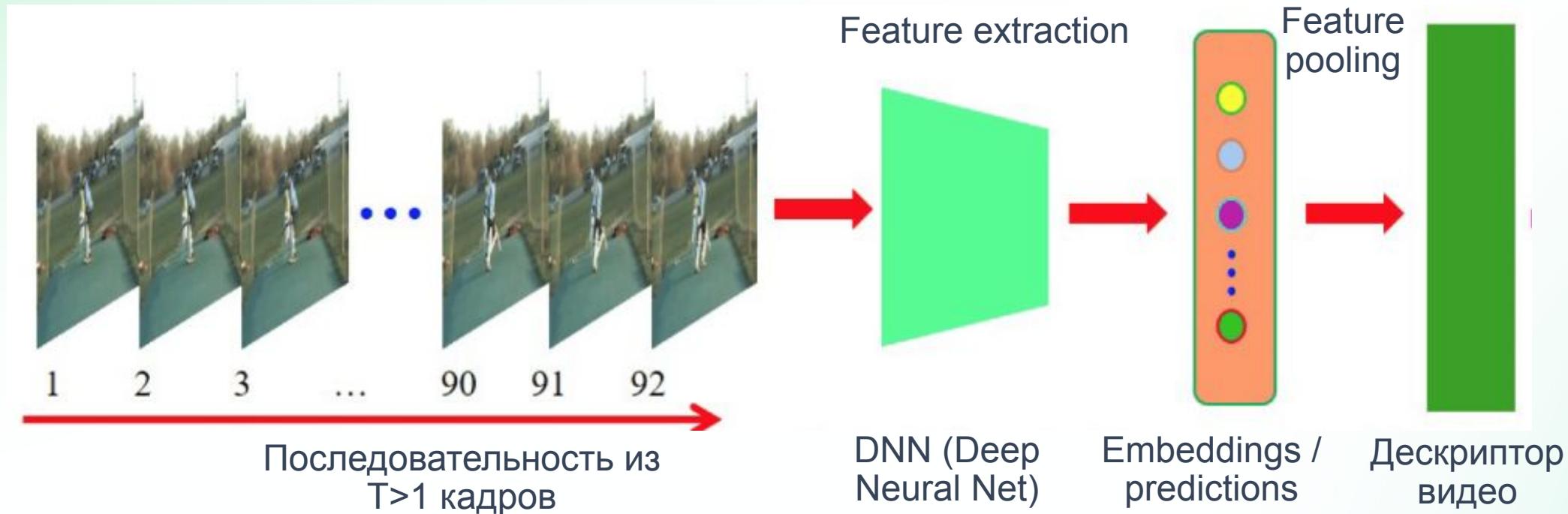
DDAMFN: dual-direction attention mixed feature network



*GDConv – Global Depthwise Convolution

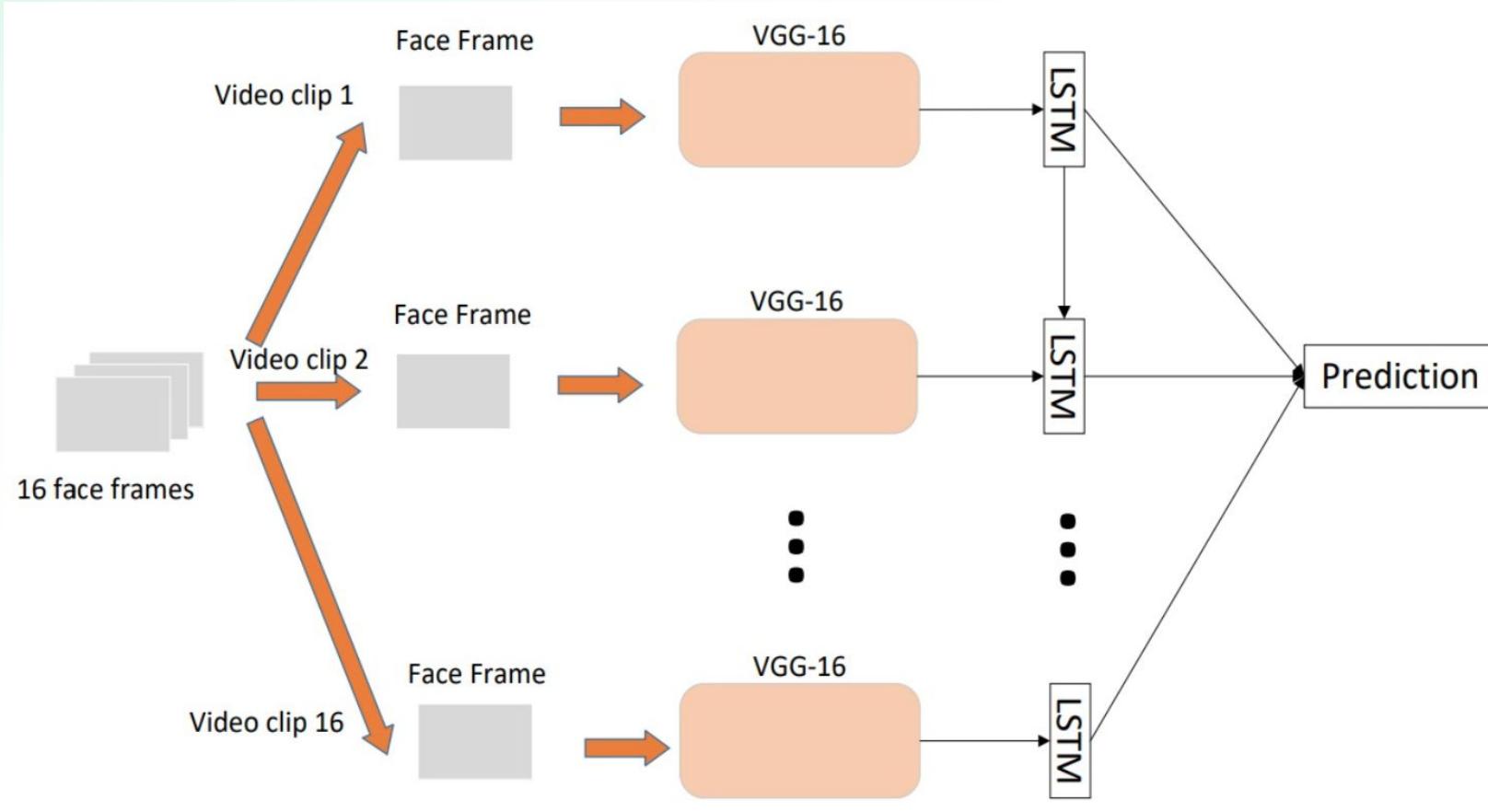
1. Mixed Feature Network (MFN) на основе MobileFaceNet извлекает признаки лица.
2. DDAN (Dual-Direction Attention Network) – сеть из нескольких DDA генерирует карты внимания (attention maps).
3. Линейный GDConv слой для карт признаков (feature map).
4. Полносвязный слой для классификации эмоций

Классификация видео: традиционный подход



Распознавание выражений лиц по видео

LSTM/GRU



Ouyang, Xi, et al. "Audio-visual emotion recognition using deep transfer learning and multiple temporal models." ICMI 2017

Модели Emoti...Nets (1)

Предобработка



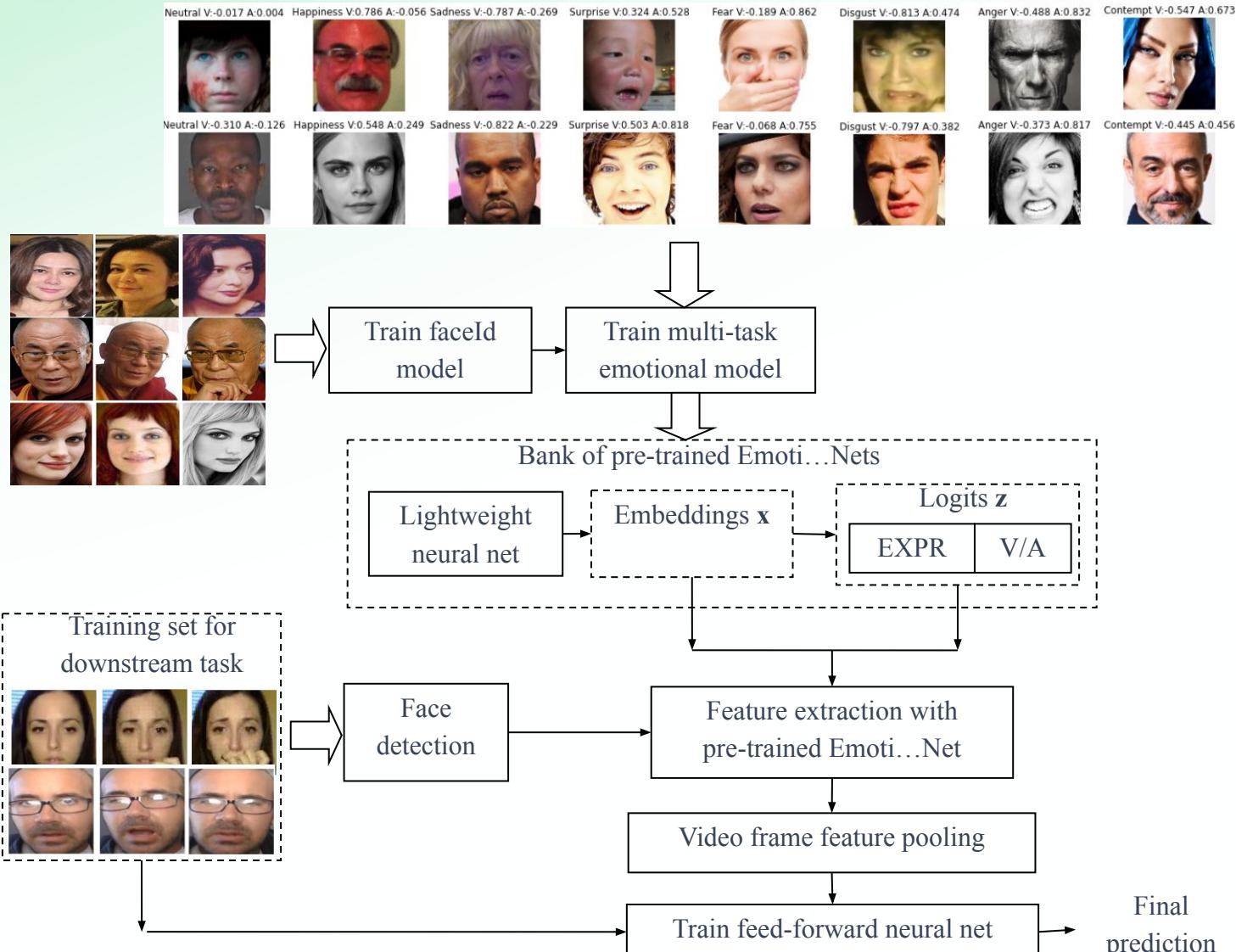
Traditional center crop



Our crop by face detector without margins

Модели Emoti...Nets (2)

Обучение



- Обучить «легковесную» модель идентифицировать лица (softmax loss) из набора данных VGGFace2
- Мульти-задачное дообучение предсказаний выражений лиц, знака и интенсивности эмоций для набора данных AffectNet

$$L(X, y_{Expr}, y_V, y_A) = 1 - \left(-\log \left(softmax(z_{y_{Expr}}) \cdot \max_{y \in \{1, \dots, C_{Expr}\}} N_y / N_{y_{Expr}} \right) - 0.5 (CCC(z_V, y_V) + CCC(z_A, y_A)) \right), \quad (1)$$

- z – logits
- N_y – число примеров для y -й эмоции
- CCC – Concordance Correlation Coefficient

Легковесные модели

- MT-EmotiEffNet
- MT-DDAMFN
- MT-MobilefaceNet
- MR-MobileViT

HSEmotion

High-Speed Emotion recognition library

Модели свободно доступны

<https://github.com/HSE-asavchenko/face-emotion-recognition/>

Python packages `hsemotion`, `hsemotion-onnx`:

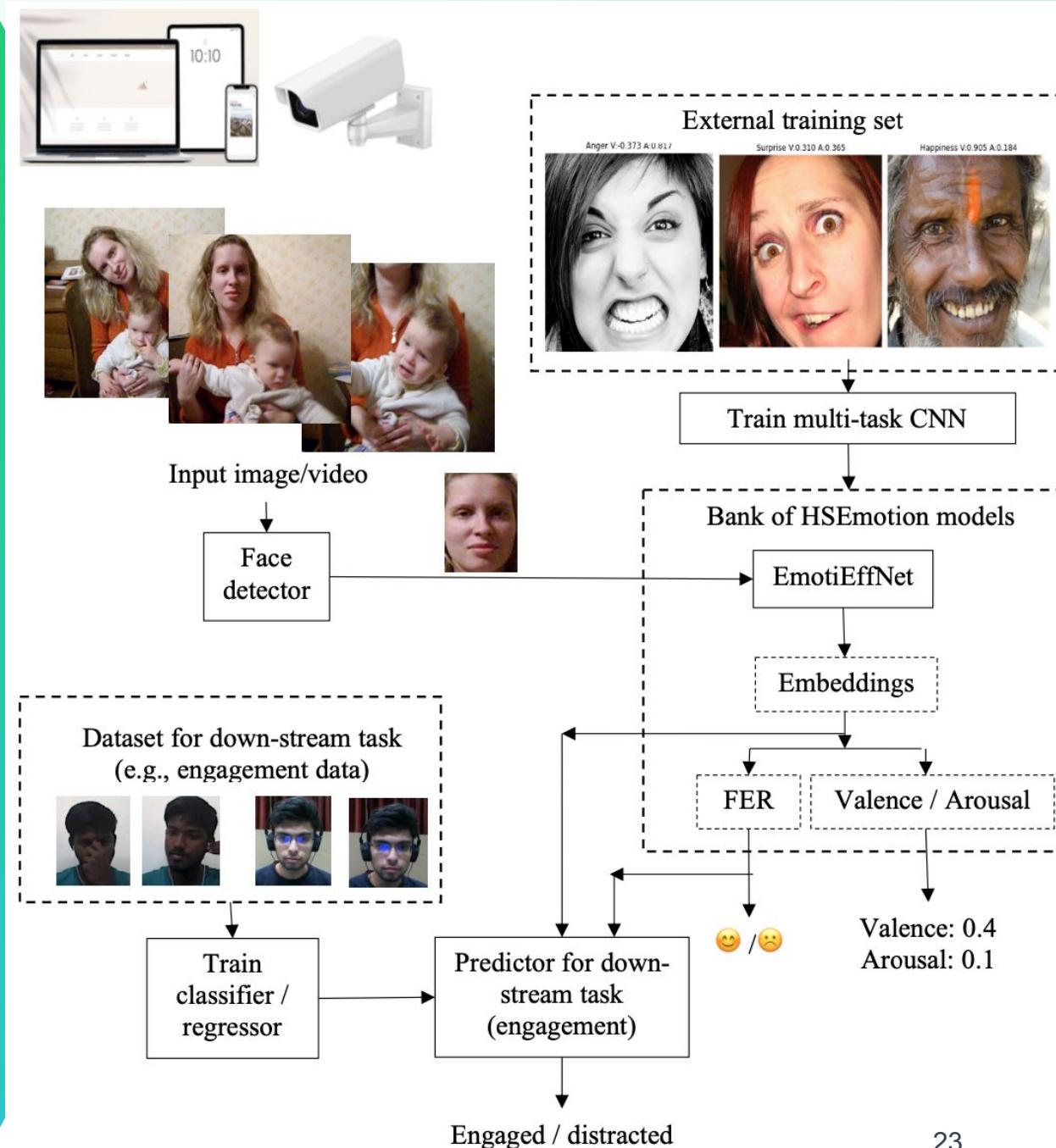
<https://github.com/HSE-asavchenko/hsemotion>
pip install hsemotion



Savchenko et al, IEEE Trans. on Affective Computing

Savchenko, CVPR 2022-2024 Workshops;

Savchenko, ECCV 2022 Workshops



Распознавание эмоций на видео

**STAT (statistical) features +
MLP**

$$\bar{\mathbf{x}}_{avg}(t) = \frac{1}{2\Delta + 1} \sum_{t' = t - \Delta}^{t + \Delta} \mathbf{x}(t')$$

**FAN (Frame attention
network)**

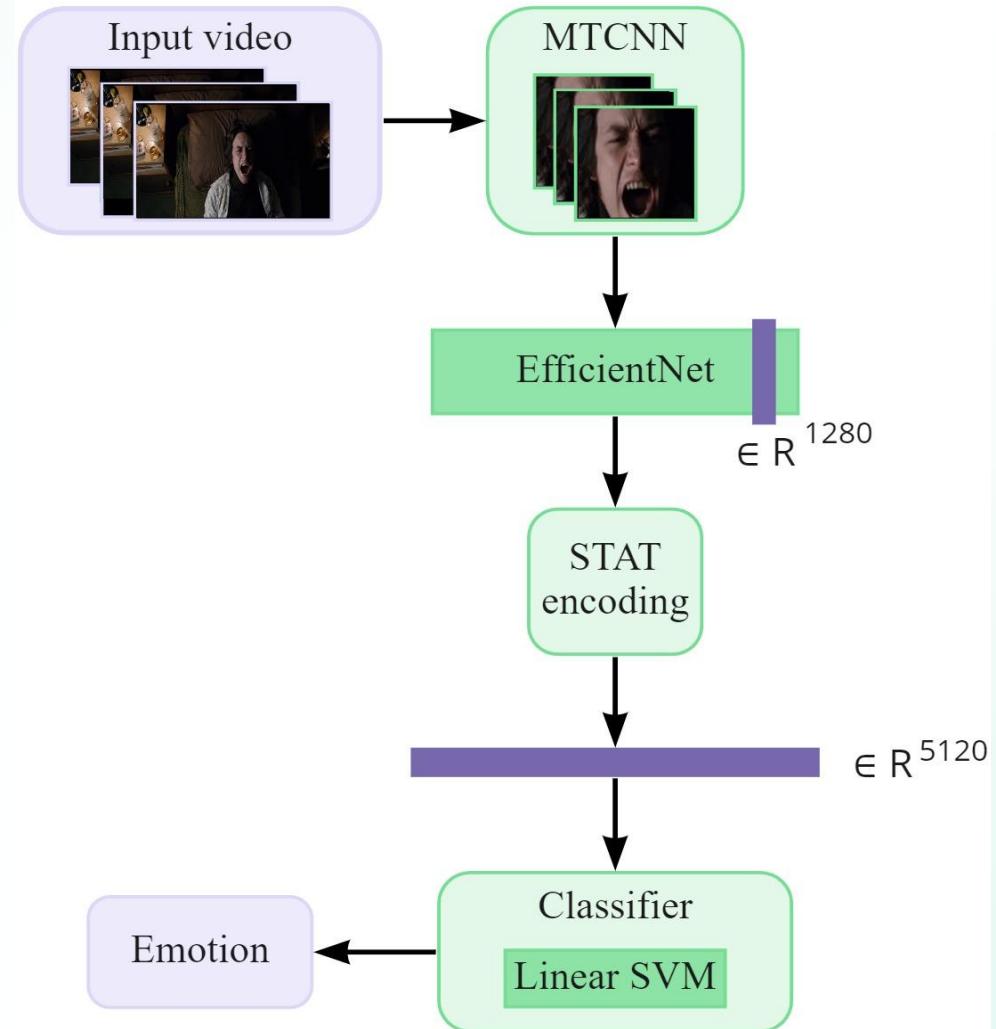
$$\alpha(t') = \text{softmax}(\mathbf{x}(t')\mathbf{q})$$

$$\bar{\mathbf{x}}_{attn}(t) = \sum_{t' = t - \Delta}^{t + \Delta} \alpha(t') \mathbf{x}(t')$$

**Ours (STAT features +
Attention)**

$$\alpha^{(1)}(t') = \text{softmax}([\mathbf{x}(t'); \bar{\mathbf{x}}_{avg}(t)]\mathbf{q}^{(1)})$$

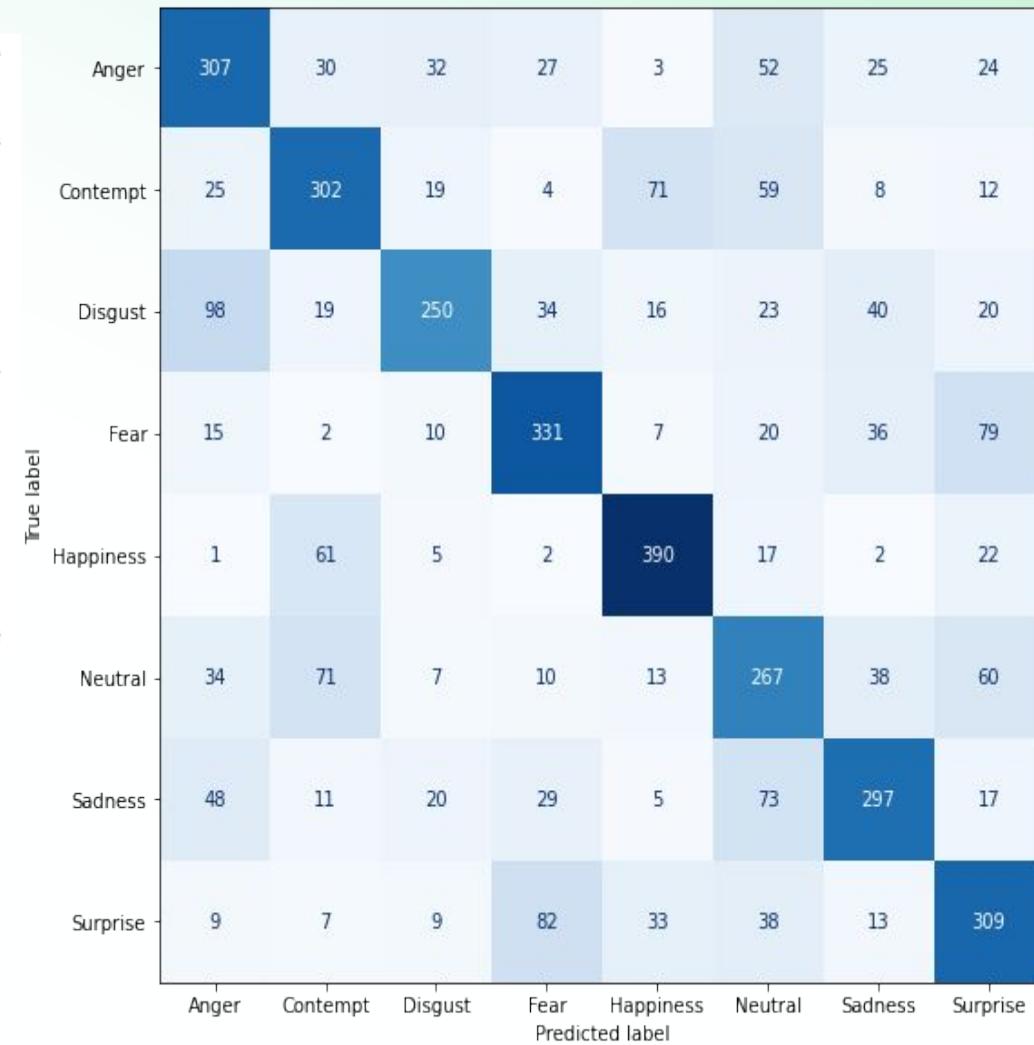
$$\bar{\mathbf{x}}_{attn;avg}(t) = \sum_{t' = t - \Delta}^{t + \Delta} \alpha^{(1)}(t') [\mathbf{x}(t'); \bar{\mathbf{x}}_{avg}(t)]$$



Результаты экспериментов

AffectNet (фото)

Model	Facial expressions		Valence		Arousal	
	8-Acc., % (↑)	7-Acc., % (↑)	RMSE (↓)	CCC (↑)	RMSE (↓)	CCC (↑)
AlexNet [26]	58.0	-	0.394	0.541	0.402	0.450
SSL inpainting-pl [27]	61.72	-	-	-	-	-
Distract Your Attention [38]	62.09	65.69	-	-	-	-
ViT-base + MAE [22]	62.42	-	-	-	-	-
Static-to-Dynamic [5]	63.06	66.42	-	-	-	-
DDAMFN [43]	64.25	67.03	-	-	-	-
EmotiEffNet-B0	61.32	64.57	-	-	-	-
MT-EmotiEffNet	61.93	64.97	0.434	0.594	0.387	0.549
MT-EmotiMobileFaceNet	62.32	65.17	0.447	0.577	0.387	0.547
MT-EmotiMobileViT	62.50	66.46	0.423	0.599	0.371	0.565
EmotiEffNet-B2 [33]	63.03	66.29	-	-	-	-
EmotiEffNet-B2	63.13	66.51	-	-	-	-
MT-DDAMFN	64.20	67.00	0.363	0.729	0.341	0.643



Результаты экспериментов

ABAW-6 (видео): Valence-Arousal Prediction

Validation set results

Method	CCC_V	CCC_A	P _{VA}
Baseline ResNet-50 [24]	0.24	0.20	0.22
EfficientNet-B0 [38]	0.449	0.535	0.492
Resnet50 + Regnet + EfficientNet [46]	0.257	0.383	0.320
Audio/video Channel Attention Network [53]	0.423	0.670	0.547
Audio/video MAE [54]	0.476	0.644	0.560
Audio/video Transformer [57]	0.554	0.659	0.607
Audio/video TCN [58]	0.550	0.681	0.615
EmotiEffNet-B2	0.423	0.498	0.464
EmotiEffNet	0.443	0.519	0.482
EmotiEffNet, smoothing	0.490	0.596	0.543
DDAMFN	0.438	0.523	0.481
DDAMFN, smoothing	0.485	0.598	0.541
MT-EmotiEffNet	0.444	0.521	0.483
MT-EmotiEffNet, smoothing	0.490	0.604	0.547
MT-EmotiMobileViT	0.445	0.525	0.485
MT-EmotiMobileViT, smoothing	0.493	0.612	0.552
MT-EmotiMobileFaceNet	0.439	0.532	0.486
MT-EmotiMobileFaceNet, smoothing	0.483	0.610	0.547
MT-DDAMFN	0.468	0.537	0.502
MT-DDAMFN, smoothing	0.519	0.616	0.568

Pre-trained models

Model	CCC_V	CCC_A	P _{VA}
MT-DDAMFN	0.412	0.230	0.321
MT-EmotiMobileViT	0.403	0.244	0.324
MT-EmotiMobileFaceNet	0.413	0.266	0.339
MT-EmotiEffNet	0.404	0.248	0.326

Test set results

Model	CCC_V	CCC_A	P _{VA}
Netease Fuxi AI Lab [55]	0.6873	0.6569	0.6721
DeepAVER [35]	0.5418	0.6196	0.5807
CtyunAI [59]	0.5223	0.6057	0.564
SUN_CE [36]	0.5355	0.5861	0.5608
USTC-IAT-United [49]	0.5208	0.5748	0.5478
KBS-DGU	0.4836	0.5318	0.5077
HSE-NN-SberAILab [40]	0.4818	0.5279	0.5048
ETS-LIVIA [44]	0.4198	0.4669	0.4434
CAS-MAIS [44]	0.4245	0.3414	0.3830
Baseline ResNet-50 [24]	0.211	0.191	0.201
DDAMFN	0.4805	0.5373	0.5089
MT-EmotiMobileViT	0.4807	0.5375	0.5091
MT-EmotiMobileFaceNet	0.4961	0.5264	0.5113
MT-DDAMFN	0.4921	0.5481	0.5202

Результаты экспериментов

RAMAS

Features	Classifier	Metric	Number of agreed annotators				
			1	2	3	4	5
VGGFace [Ryumina et al., 2021]	RF	UAR, %	42.9	51.4	57.1	60.2	56.4
Fine-tuned EfficientNet-B3 [Ryumina et al., 2021]		UAR, %	45.3	53.1	65.3	74.8	70.8
MobileNet v1	SVM	UAR, %	40.1	65.8	68.4	76.8	78.7
		Accuracy, %	46.7	69.7	71.4	79.2	81.2
	MLP	UAR, %	34.4	59.8	67.9	73.4	73.3
		Accuracy, %	49.8	68.9	73.5	78.1	84.6
EfficientNet-B0	SVM	UAR, %	39.8	66.4	67.9	75.9	78.9
		Accuracy, %	46.0	68.5	70.6	78.1	80.8
	MLP	UAR, %	34.1	67.6	72.6	74.5	75.3
		Accuracy, %	47.1	69.7	73.5	80.8	83.3

Результаты покадрового распознавания эмоций на видео, конкурс ABAW-4

Multi-Task Learning

3 место (из 25 команд)

Method	P _{MTL}
Baseline	0,28
SSSIHL-DMACS	1,0406
HUST-ANT	1,0712
STAR-2022	1,0855
CNU-Sclab	1,1135
MT-EmotiEffNet	1,1299
ICT-VIPL	1,1945
Situ-RUCAIM3	1,4361

Learning from Synthetic Data

1 место (из 21 команд)

Method	F1 score
Baseline	0,30
SSSIHL-DMACS	0,3364
HUST-ANT	0,3483
ICT-VIPL	0,3483
IXLAB	0,3587
PPAA	0,3651
MT-EmotiEffNet	0,3718

Результаты покадрового распознавания эмоций на видео, конкурс ABAW-7

Multi-Task Learning

2 место (из 25 команд)

Teams	Overall Metric	AUs F1 Score	Expressions F1 Score	CCC VA Score
Netease Fuxi AI Lab	1.5286	0.5580	0.4286	0.5420
HSEmotion	1.2472	0.5119	0.3279	0.4074
HFUT-MAC1	1.1777	0.4997	0.2997	0.3783
SCU_ACers	1.1640	0.4879	0.3018	0.3743
AIWELL-UOC	1.1145	0.4663	0.2772	0.3710
SML	0.8692	0.4046	0.1938	0.2708

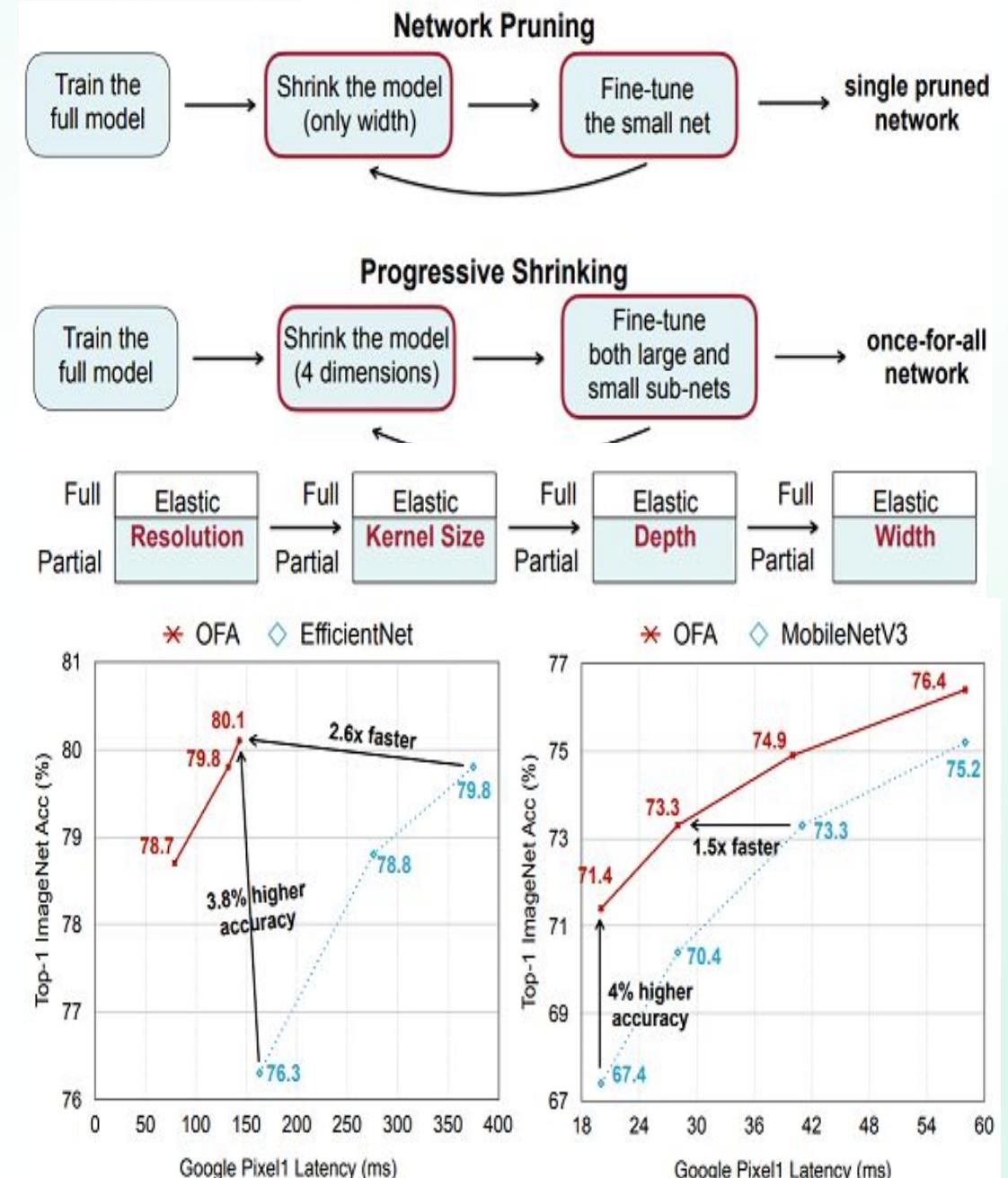
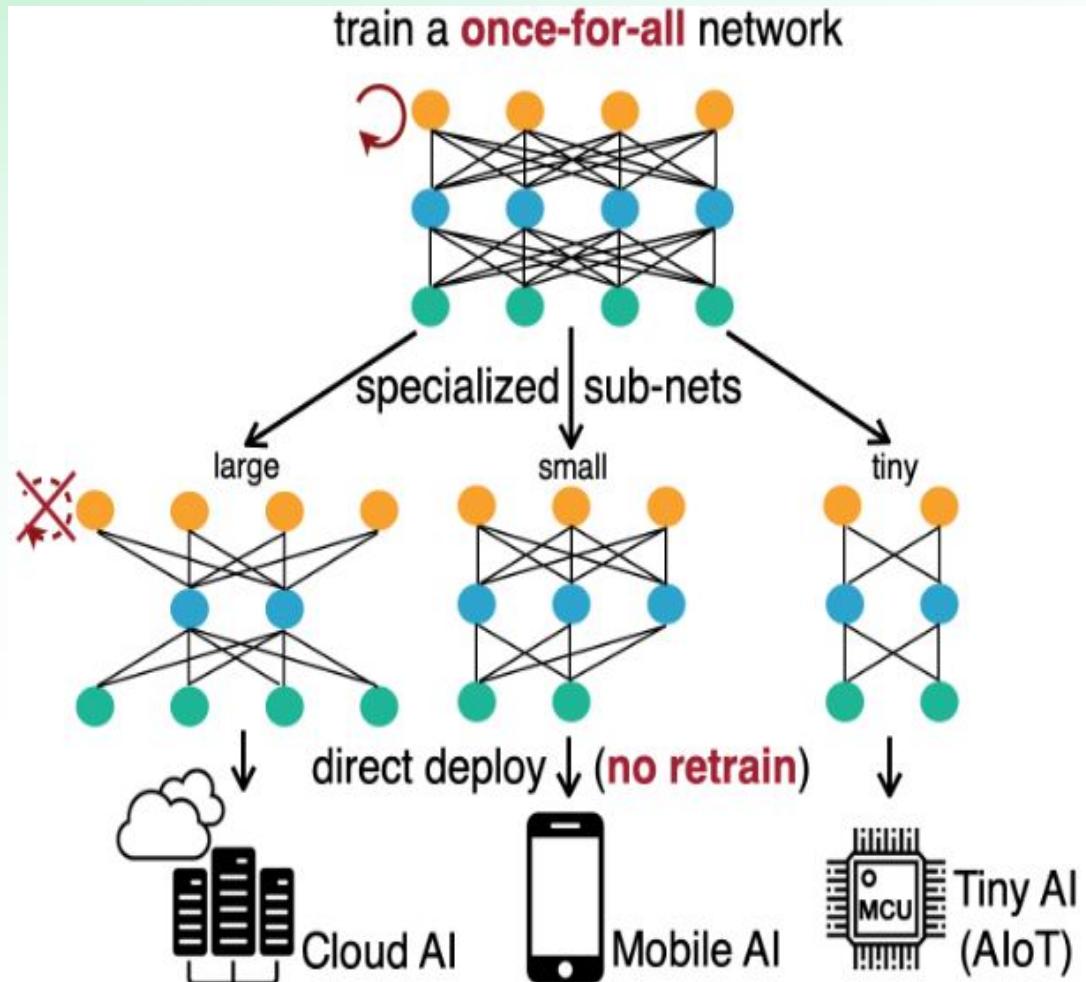
Compound Expression Recognition

Teams	F1
Netease Fuxi AI Lab	0.6063
HSEmotion	0.3243
ETS-LIVIA	0.2591
HFUT-MAC2	0.2281
AIPL-BME-SEU	0.1644

<https://affective-behavior-analysis-in-the-wild.github.io/7th/#counts>

On-device inference

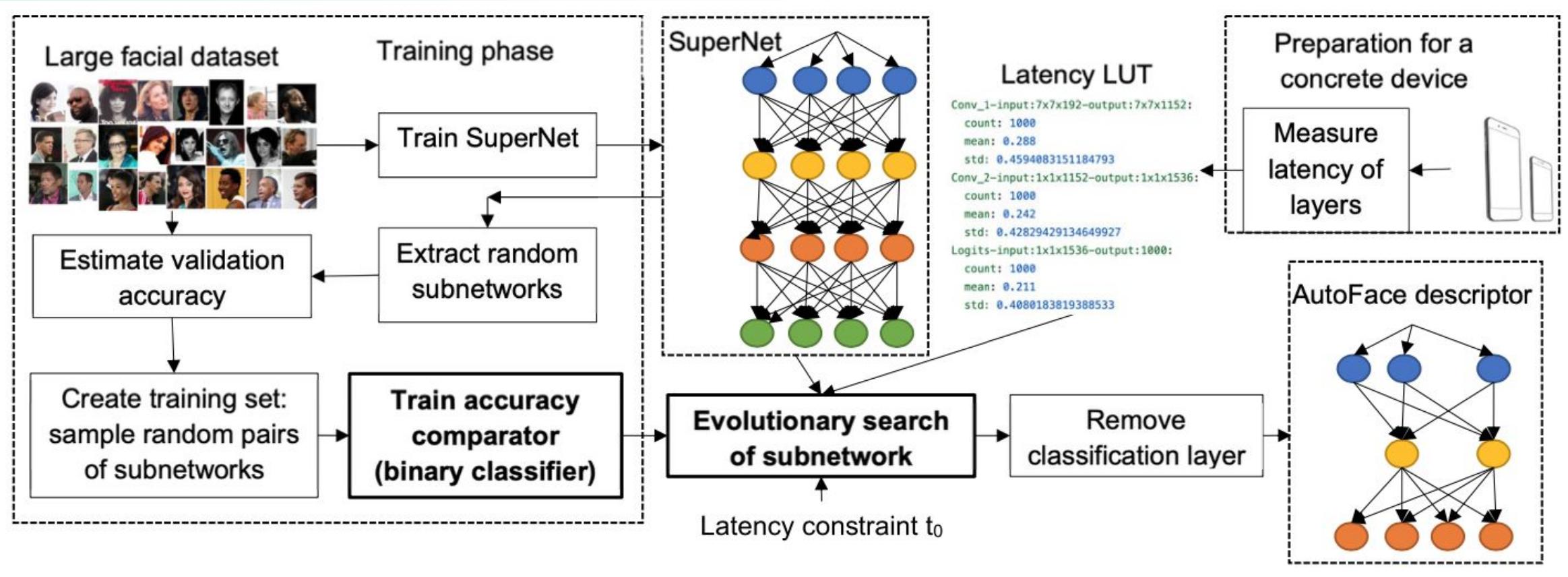
Once-for-All SuperNet



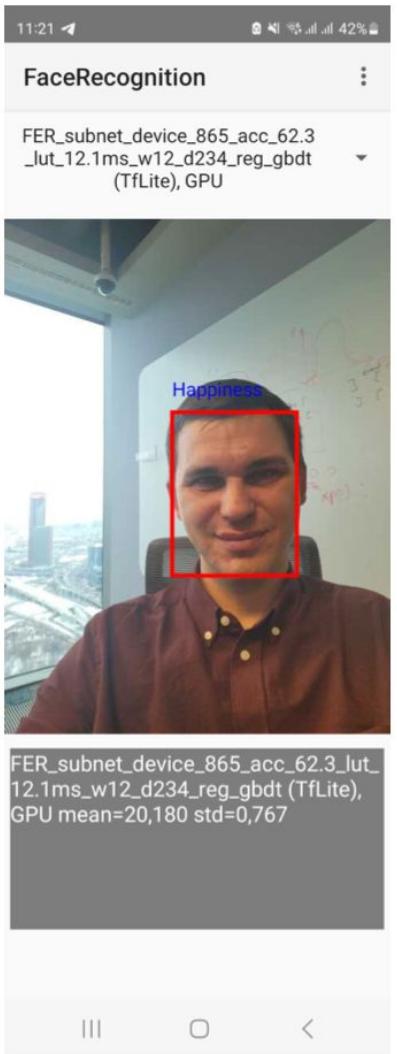
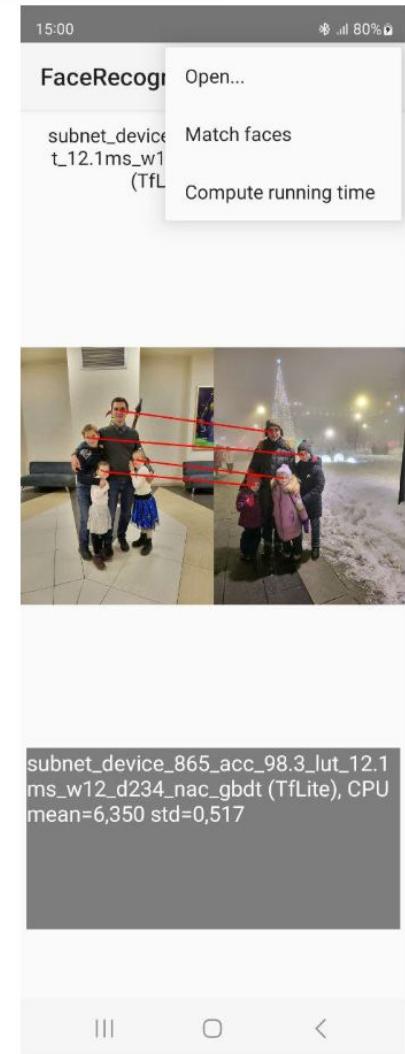
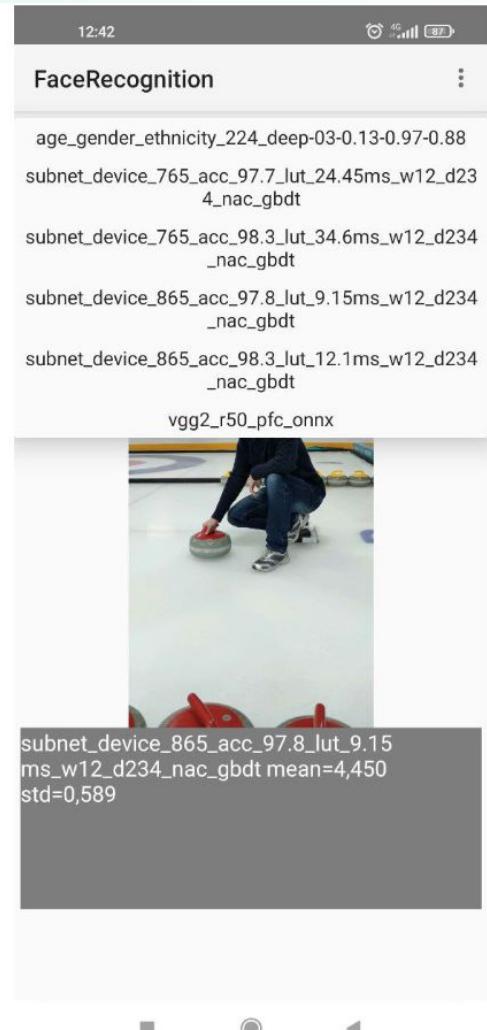
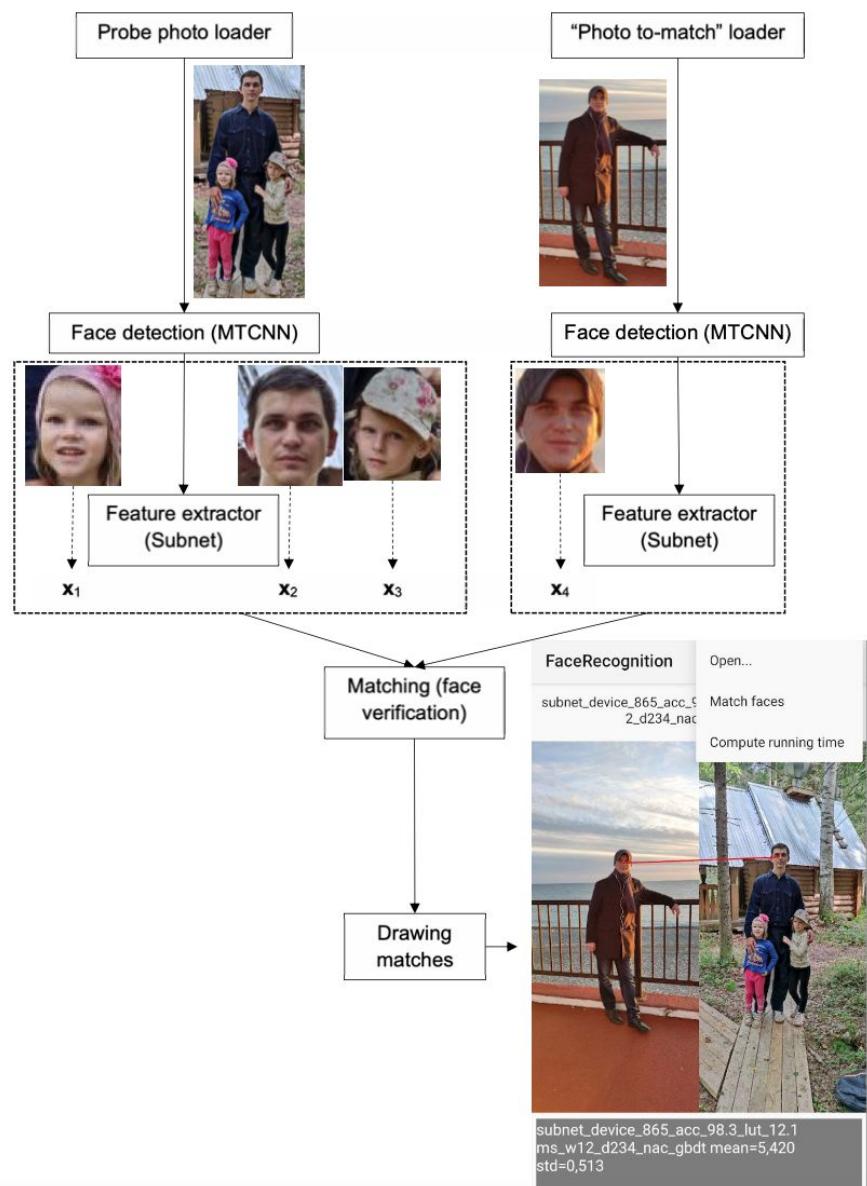
<https://github.com/mit-han-lab/once-for-all>

AutoFace (1)

How to Obtain Mobile Neural Network-Based Facial Feature Extractor in Less Than 10 Minutes?



AutoFace (2)

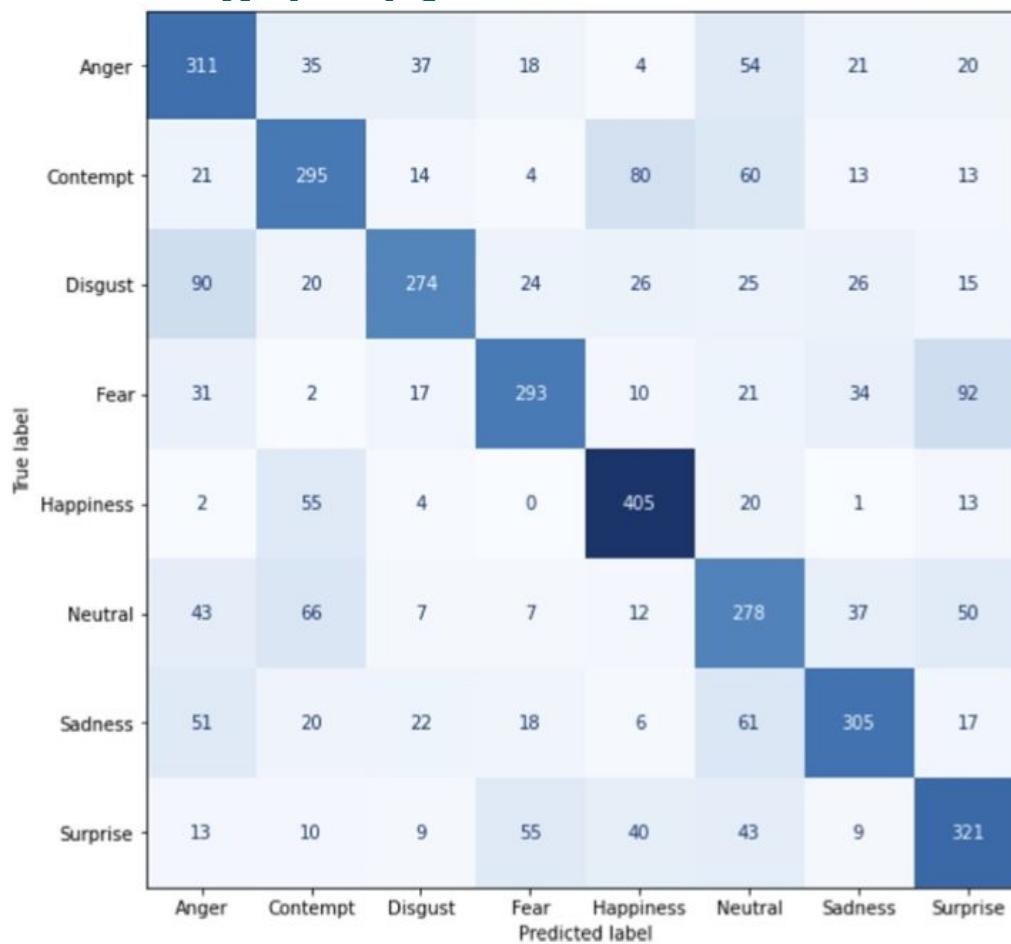


Результаты экспериментов

AffectNet

Model	Accuracy, %		Frames processed per second		
	AffectNet	AFEW	SDM865	MT6853	Raspberry Pi 4
AffectNet baseline [37]	58.0	-	120.48	58.82	16.13
EfficientNet-B2, SL + SSL in-panting-pl [40]	61.32	-	6.73	2.66	2.83
EfficientNet-B0, SL + SSL in-panting-pl [40]	61.72	-	21.25	8.21	5.49
AFEW baseline [41]	-	38.90	135.21	64.78	21.85
Noisy student with iterative training [42]	-	55.17	105.26	41.67	14.08
HSEmotion, MobileEmotiFace [31]	60.20	55.05	75.30	30.24	7.80
HSEmotion, EmotiEffNet-B0 [39]	61.32	55.14	21.21	8.15	5.46
HSEmotion, EmotiEffNet-B2 [35]	63.03	57.78	6.67	2.62	2.79
AutoFace (Subnet 1)	<u>62.05</u>	<u>55.18</u>	84.10	29.39	9.25
AutoFace (Subnet 2)	61.28	54.89	115.07	43.90	13.65

Confusion matrix,



01 Введение: представление эмоций, модальности, датасеты

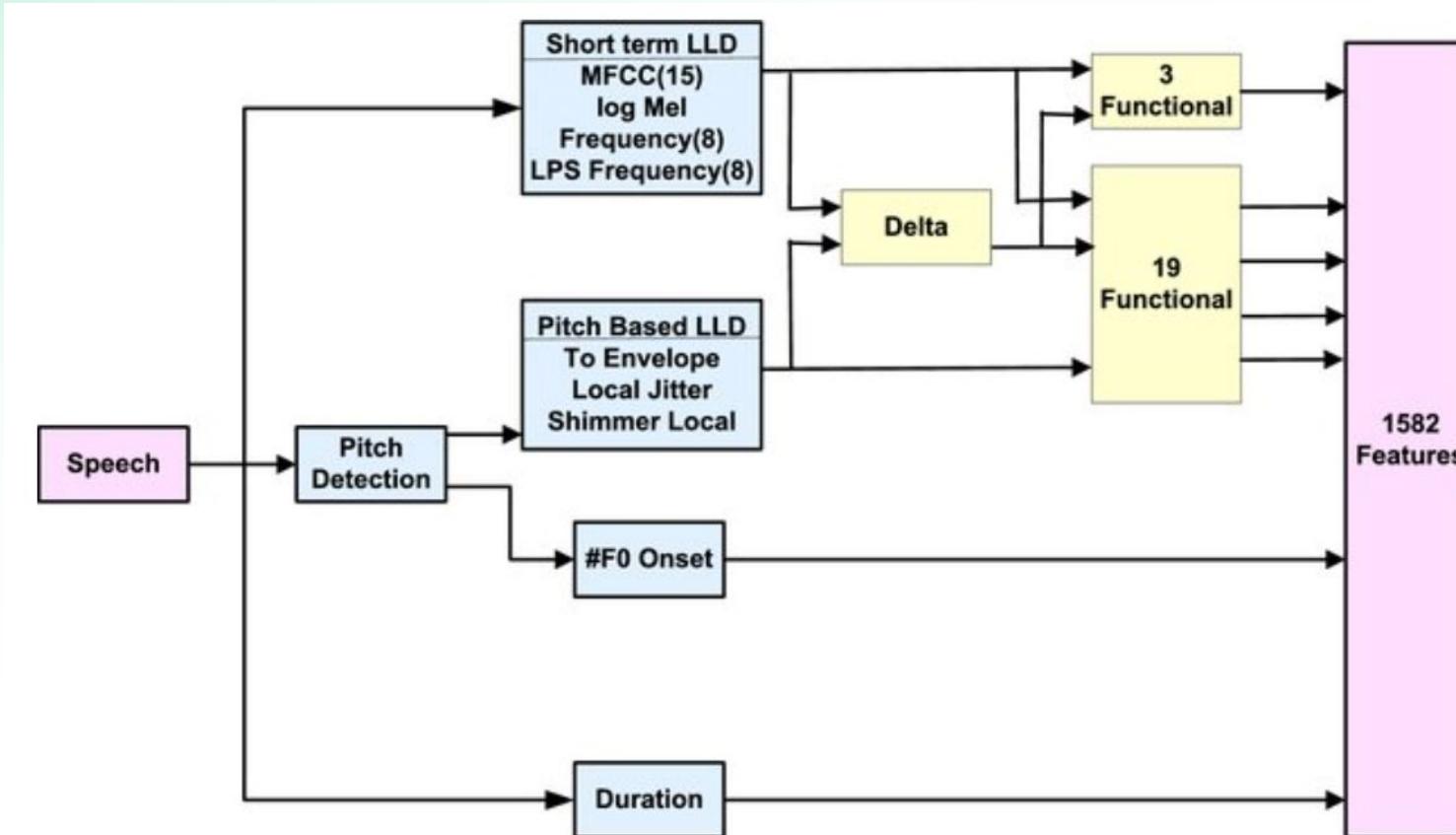
02 Распознавание выражений лиц по видео

03 Распознавание эмоций по голосу

04 Мультимодальные методы

05 Демо и примеры

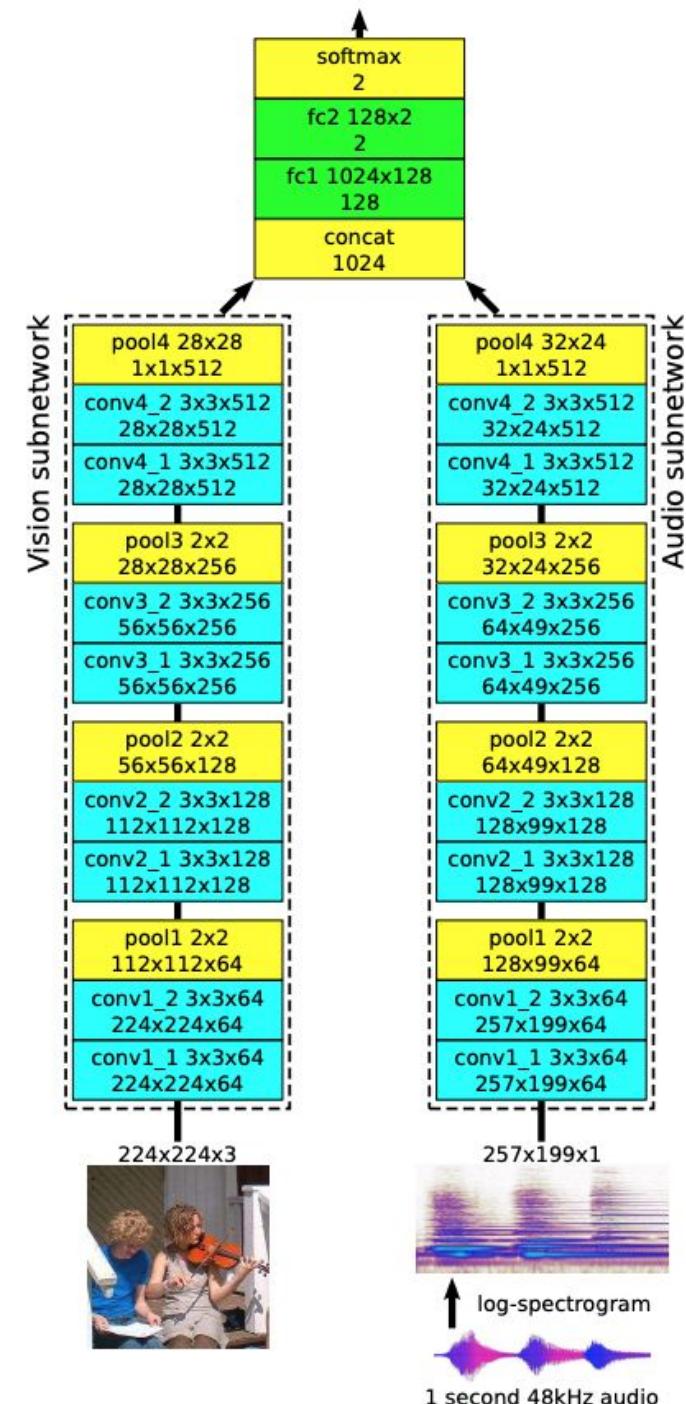
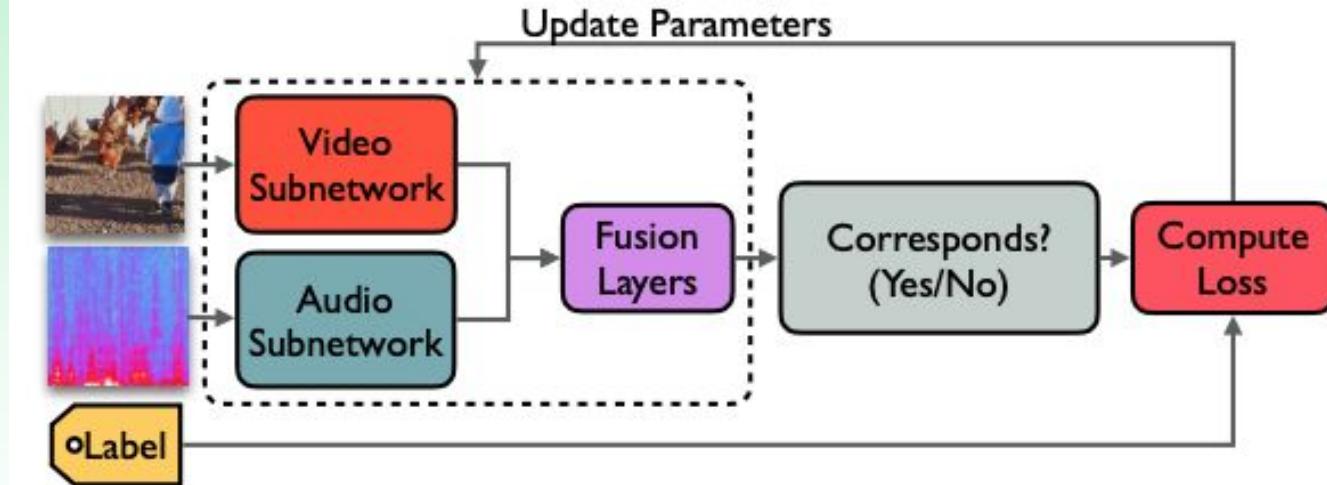
OpenSmile



Emobase configuration of traditional acoustic features, such as pitch frequency, Mel-frequency cepstral coefficients, etc

<https://www.audeering.com/research/opensmile/>

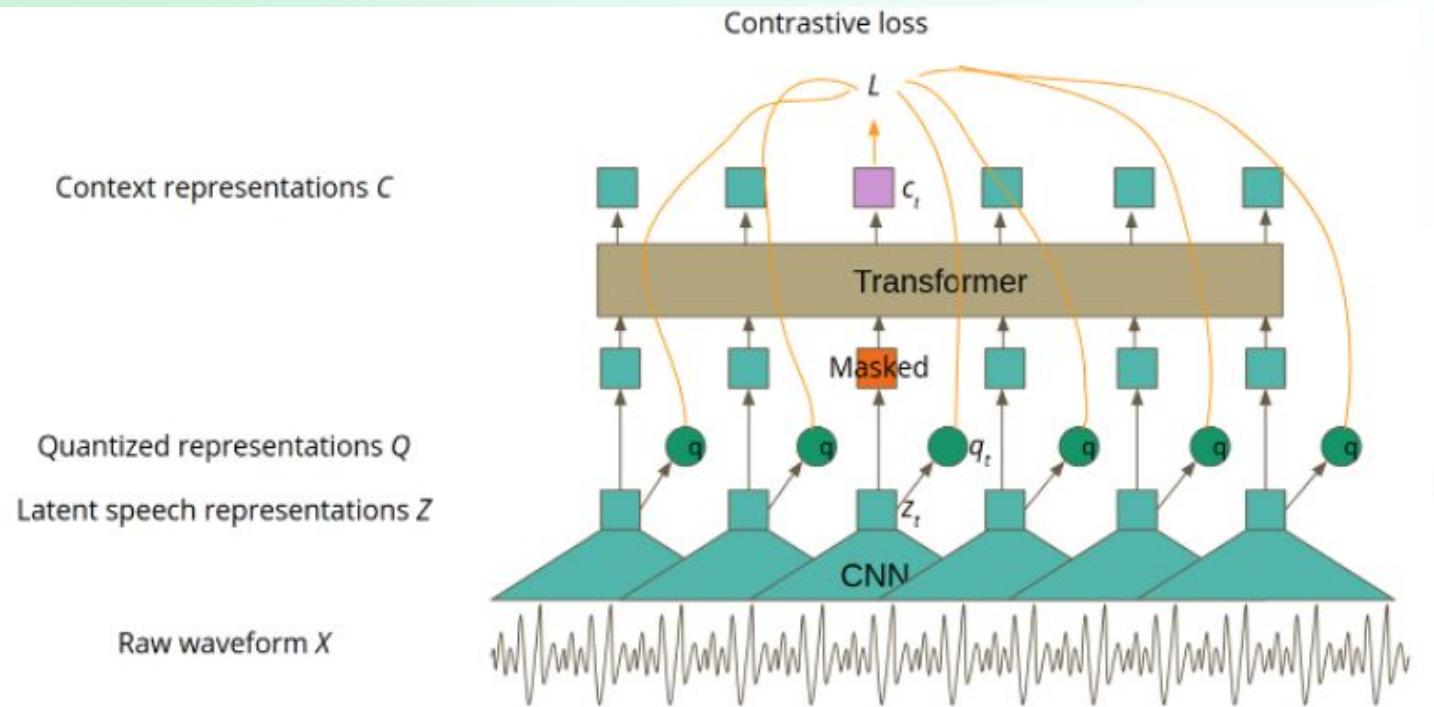
OpenL3



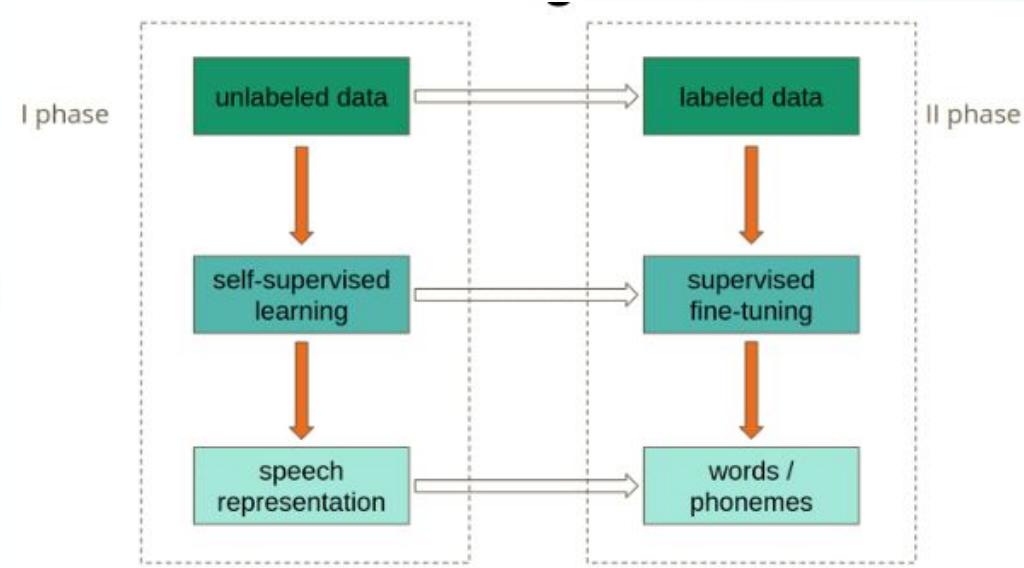
J. Cramer et al. "Look, Listen and Learn More: Design Choices for Deep Audio Embeddings", ICASSP 2019

<https://github.com/marl/openl3>

Wav2vec 2.0



Обучение



Результаты экспериментов

RAMAS

Features	Classifier	Metric	Number of agreed annotators				
			1	2	3	4	5
LLD from OpenSmile	SVM [Ryumina et al., 2021]	UAR, %	28.9	31.1	38.3	40.4	46.4
	LSTM [Ryumina et al., 2021]	UAR, %	34.4	42.3	46.0	46.3	46.3
Emobase from OpenSmile	SVM	UAR, %	28.6	29.9	43.5	48.8	50.4
		Accuracy, %	35.0	33.6	43.4	49.0	51.2
	MLP	UAR, %	29.8	37.6	47.2	49.2	53.5
		Accuracy, %	35.5	42.3	49.7	50.7	55.7
OpenL3	SVM	UAR, %	26.1	27.3	49.1	48.5	53.4
		Accuracy, %	32.0	38.9	50.5	51.5	53.6
	MLP	UAR, %	29.5	36.7	47.1	49.4	54.1
		Accuracy, %	35.9	43.2	51.4	53.6	59.1

Эмоции на русском языке: Sber Dusha

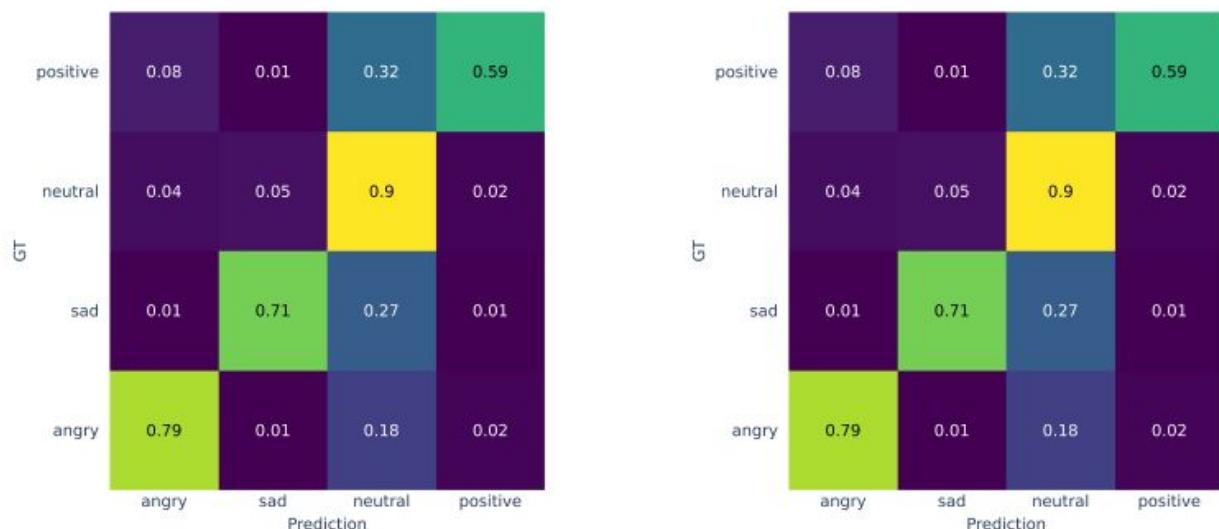
Domain	Number of Files	Duration (Hr.)	Unique Speakers
Crowd	201 850	255.7	2068
Podcast	102 113	90.9	6240
Total	303 963	346.6	8308

Бейзлайн

1. 64 Mel-filterbank (окна 20 мс. с перекрытием в 10 мс.)
2. Mel-коэффициенты передаются через последовательность инвертированных остаточных (residual) блоков – MobileNetV2
3. Слой self-attention из SAGAN (Self attention GAN)
4. Полносвязный слой

Результаты экспериментов

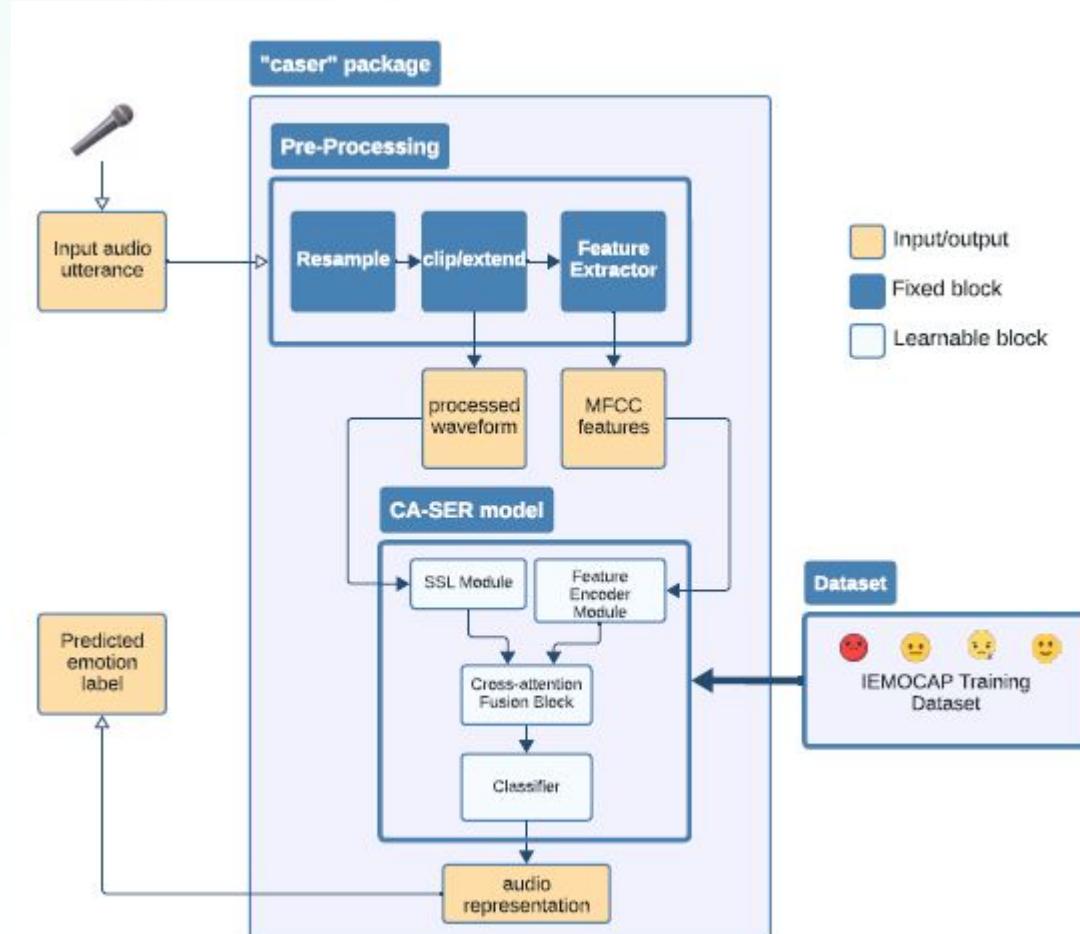
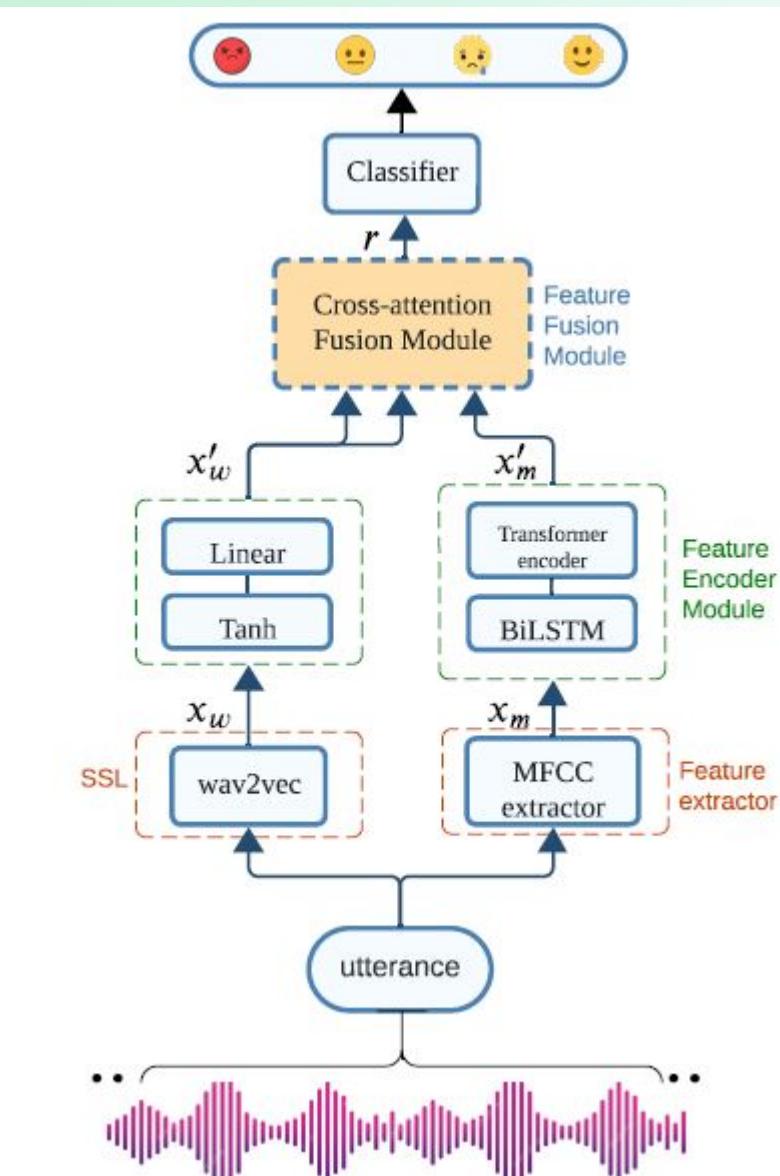
	Crowd test			Podcast test		
Training setup	UA	WA	F1	UA	WA	F1
Dusha	0.83	0.76	0.77	0.89	0.53	0.54



V. Kondratenko et al., INTERSPEECH 2023

<https://github.com/salute-developers/golos/tree/master/dusha>

CA-SER: Cross-Attention Feature Fusion

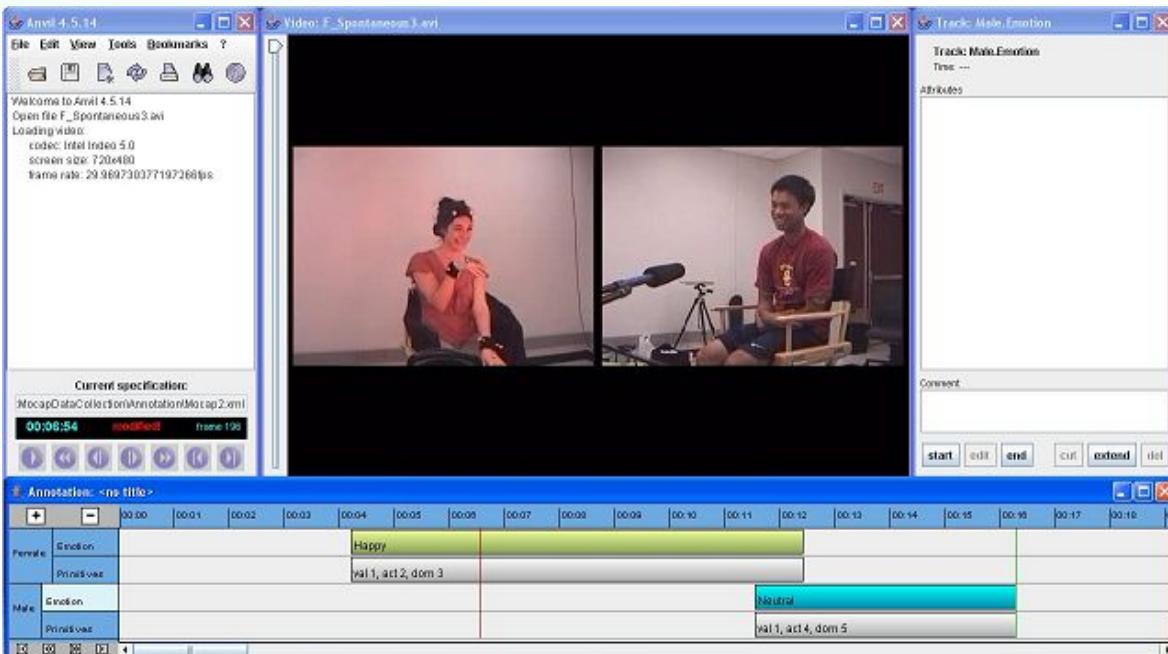


Bashar M. Deeb, Andrey Savchenko, Ilya Makarov, ECAI (demo) 2024

Результаты экспериментов

ИЕМОСАР

- 12 ч видео
- 10 участников (5 мужчин и 5 женщин)
- 5531 аудио сигналов
- Число классов: 4 (1636 happy, 1084 sad, 1103 angry, 1708 neutral).



5-fold-cross-validation

Model	WA	UA	Modality
TDNN-LSTM-Attention [19]	66.3	60.3	A
CTC+Attention [30]	67.0	69.0	A
wav2vec 2.0-PT [19]	67.2	-	A
HuBERT Base [28]	68.9	-	A
CNN TF Att.pooling [12]	71.75	68.06	A
CNN-DARTS [20]	72.55	69.36	A
MPT-HCL [31]	72.83	-	A+T+V
SDT [13]	73.95	-	A+T+V
HuBERT Large + SN [6]	74.2	-	A
CA-SER [OURS]	72.34	71.53	A

10-fold-cross-validation

Model	WA	UA
audio-BRE [29]	64.60	65.20
Audio-CNN-xvector [18]	66.60	68.40
MHSA-FACA [11]	72.01	72.83
CA-SER [OURS]	74.60	73.50

0
1

Представление эмоций

02

Распознавание выражений лиц по видео

03

Распознавание эмоций по голосу

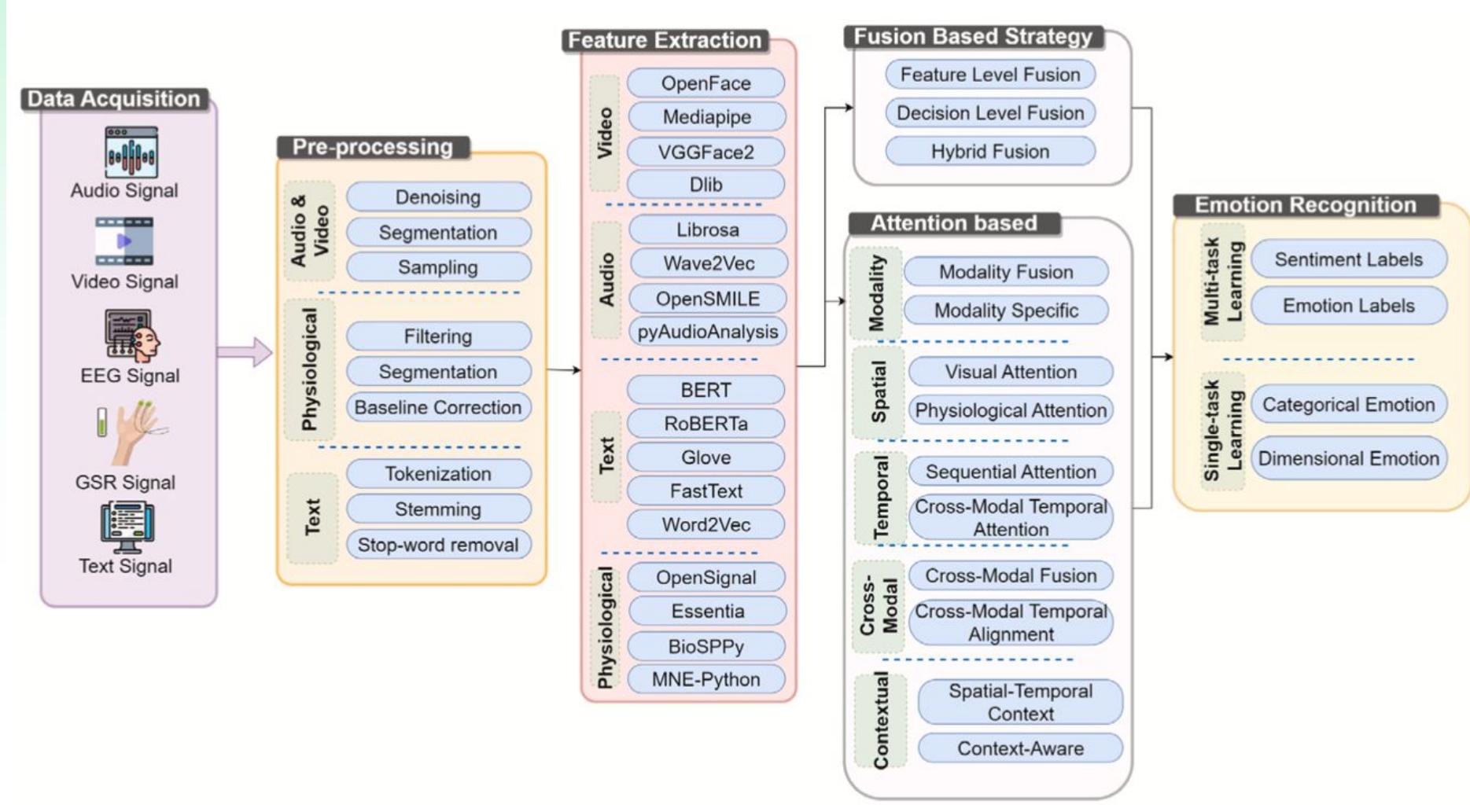
0
4

Мультиmodальные методы

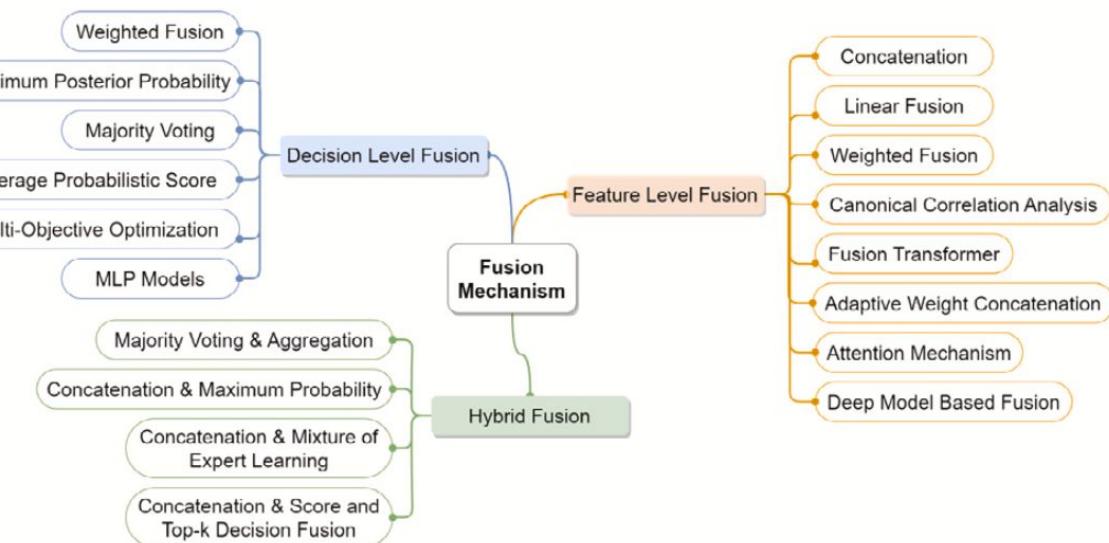
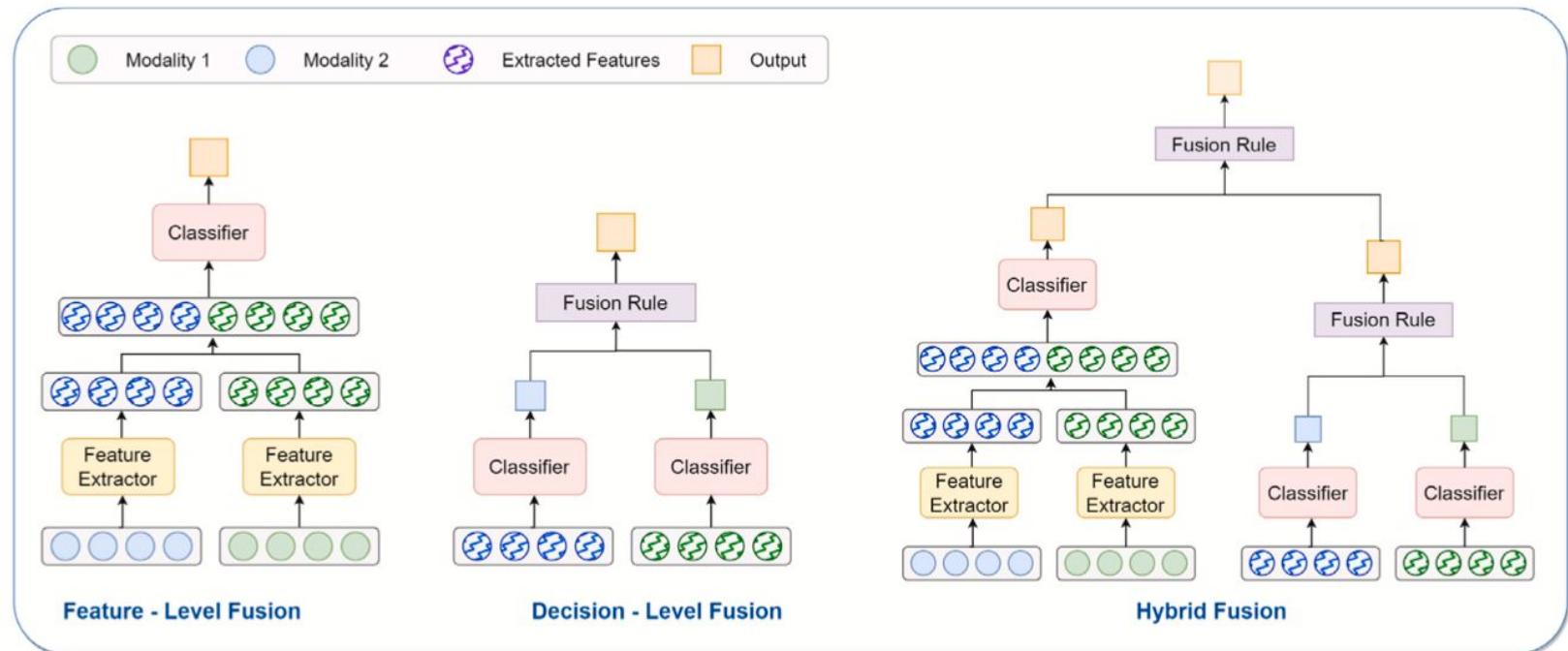
0
5

Демо и примеры

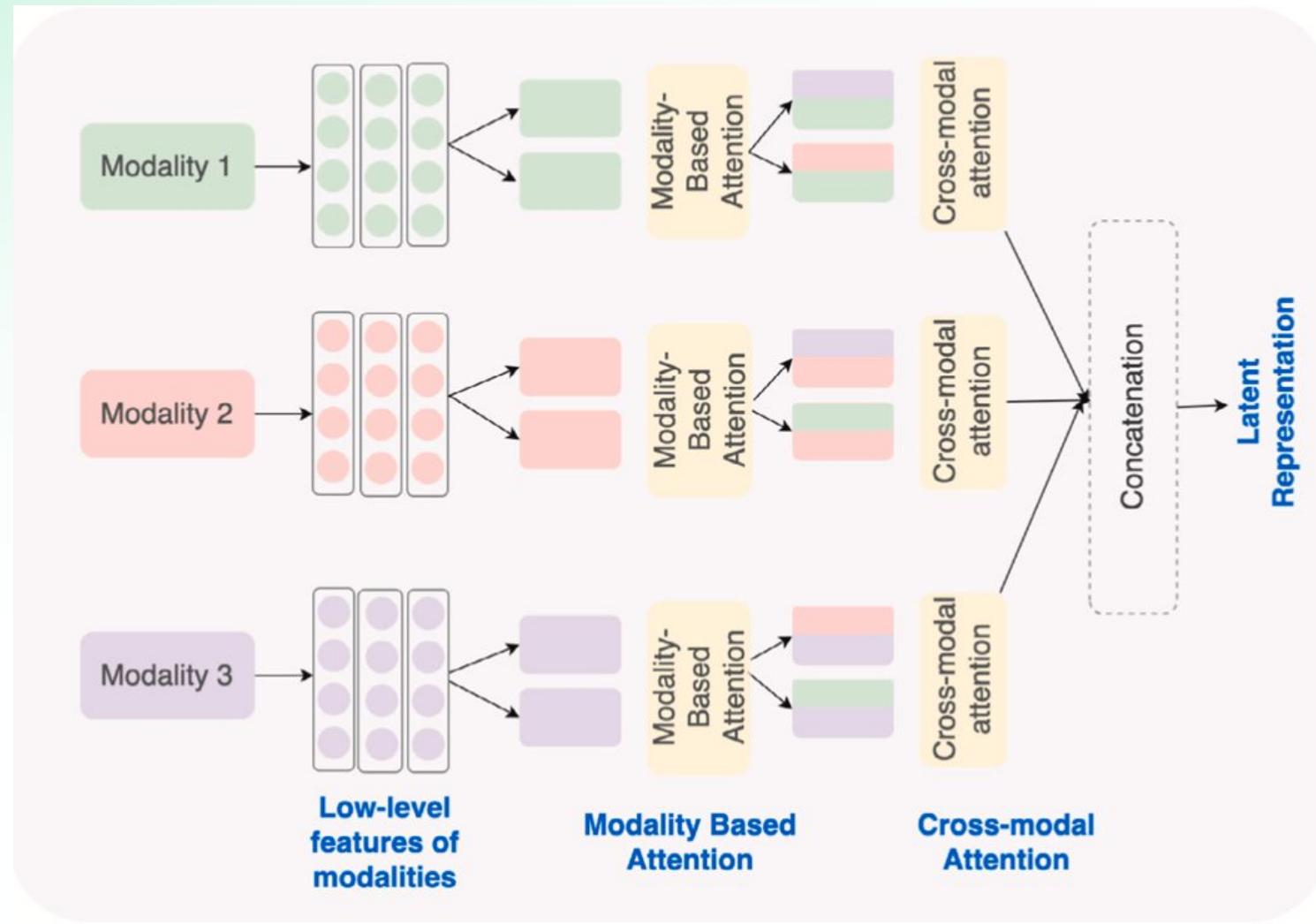
Generic pipeline



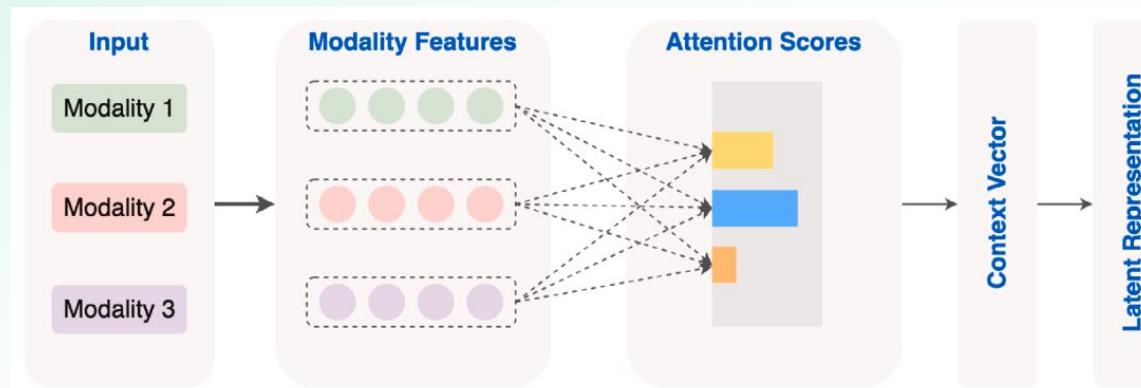
Fusion methodologies



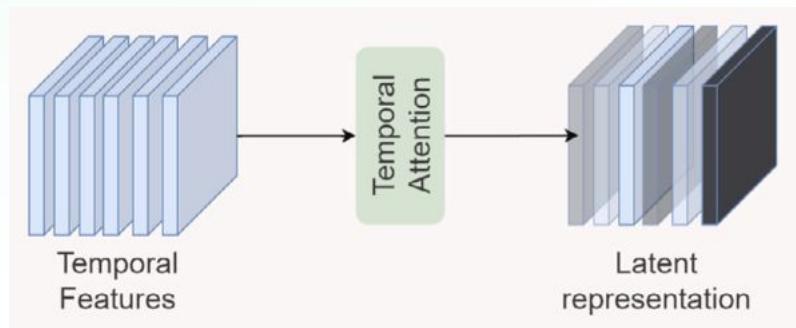
Attention



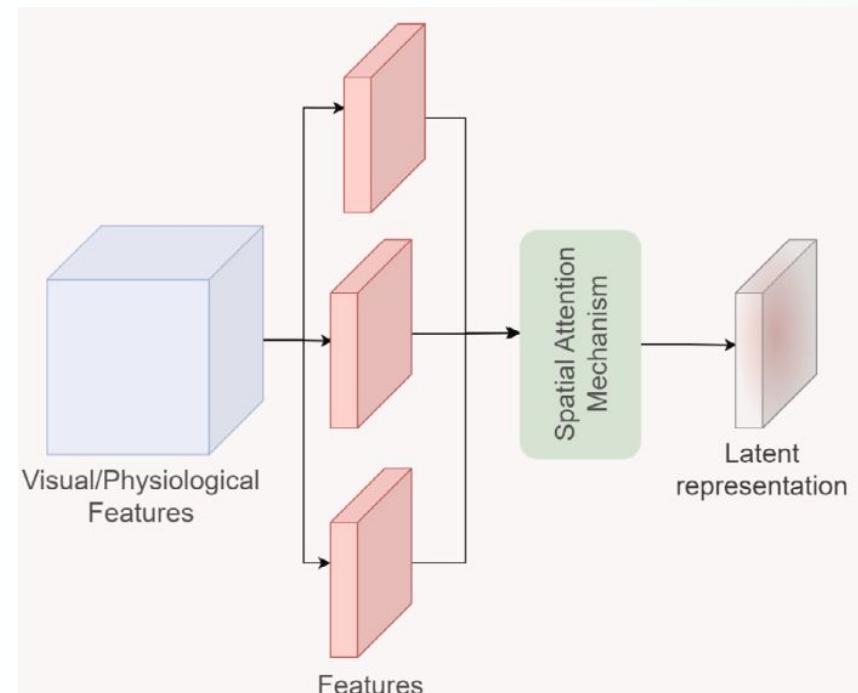
Modality-based attention



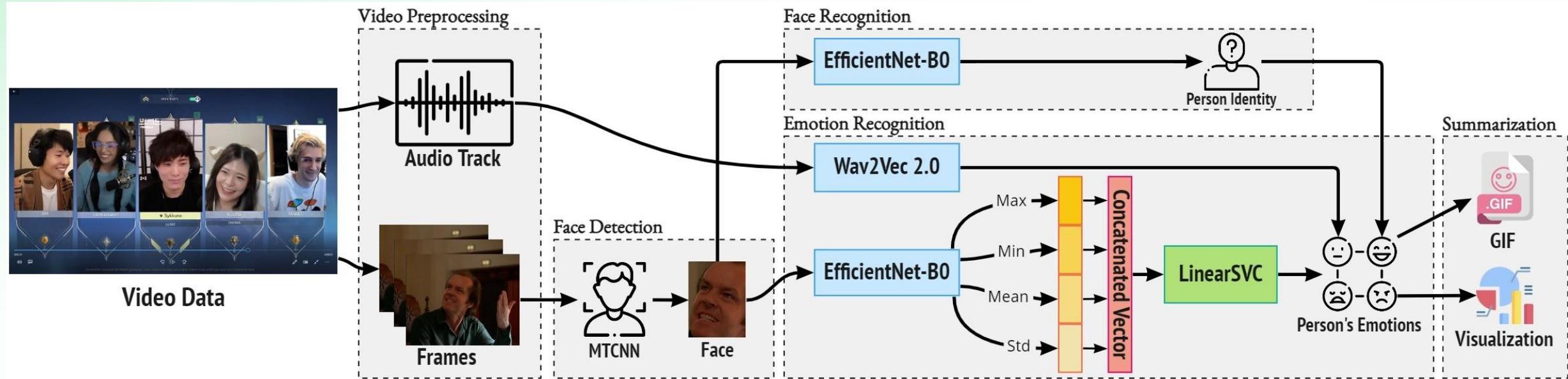
Temporal Attention



Spatial Attention



EmotiEffNet + Wav2Vec 2.0



Результаты экспериментов

AFEW: распознавание эмоций видео

Точность на валидационном множестве

Modality	Model	Accuracy, %
Video	Noisy student with iterative training [17]	55.17
	Noisy student w/o iterative training [17]	52.49
	DenseNet-161 [16]	51.44
	LBP-TOP (baseline) [14]	38.90
	XGBoost (facial descriptors)	34.91
	RandomForest (EfficientNet-B0)	55.35
Audio	LinearSVC (EfficientNet-B0)	59.269
	AlexNet [15]	38
Multimodal	wav2vec 2.0	41.514
	LResNet50E-IR [15]	65.5
	EfficientNet-B0 + wav2vec 2.0	64.75
Ensemble of 4 models		67.885

Ensemble of 4 models:

1. finetuned wav2vec 2.0 for audio
2. Linear SVC for our video features
3. Random Forest for our video features
4. XGBoost for face descriptors extracted using OpenFace2

Performance

Modality	Model	Accuracy, %	Weight, MB	CPU Running time, ms.
Video	EfficientNet-B0 FP32, PyTorch	59.269	15.6	103.56
	EfficientNet-B0 FP32, ONNX	59.269	15.2	22.52
	EfficientNet-B0 FP16, OpenVINO	59.269	8	14.26
Audio	wav2vec 2.0 FP32, PyTorch	41.514	1170	1854.44
	wav2vec 2.0 FP32, ONNX	41.253	1170	834.28
	wav2vec 2.0 FP16, OpenVINO	41.253	602	980.77
Multimodal	Ensemble of FP32 models, PyTorch	64.752	1186	1958.01
	Ensemble of FP32 models, ONNX	62.924	1186	856.80
	Ensemble of FP16 models, OpenVINO	63.446	610	995.03

Результаты экспериментов

AFEW: матрицы перепутывания

	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise
Actual	53	1	1	2	5	2	0
Happy	1	1	0	57	2	1	1
Neutral	5	2	3	4	40	6	3
Sad	4	3	0	4	10	39	1
Surprise	7	2	6	2	8	2	19
Predicted	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise

(a) Video model

	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise
Actual	45	2	2	3	6	6	0
Happy	7	5	0	4	9	13	2
Neutral	9	1	5	1	12	15	3
Sad	7	3	1	29	12	11	0
Surprise	6	1	1	9	41	5	0
Predicted	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise

(b) Audio model

	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise
Actual	61	0	1	0	1	1	0
Happy	8	11	1	5	7	8	0
Neutral	11	0	11	0	5	18	1
Sad	1	0	0	59	2	1	0
Surprise	0	0	2	2	55	4	0
Predicted	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise

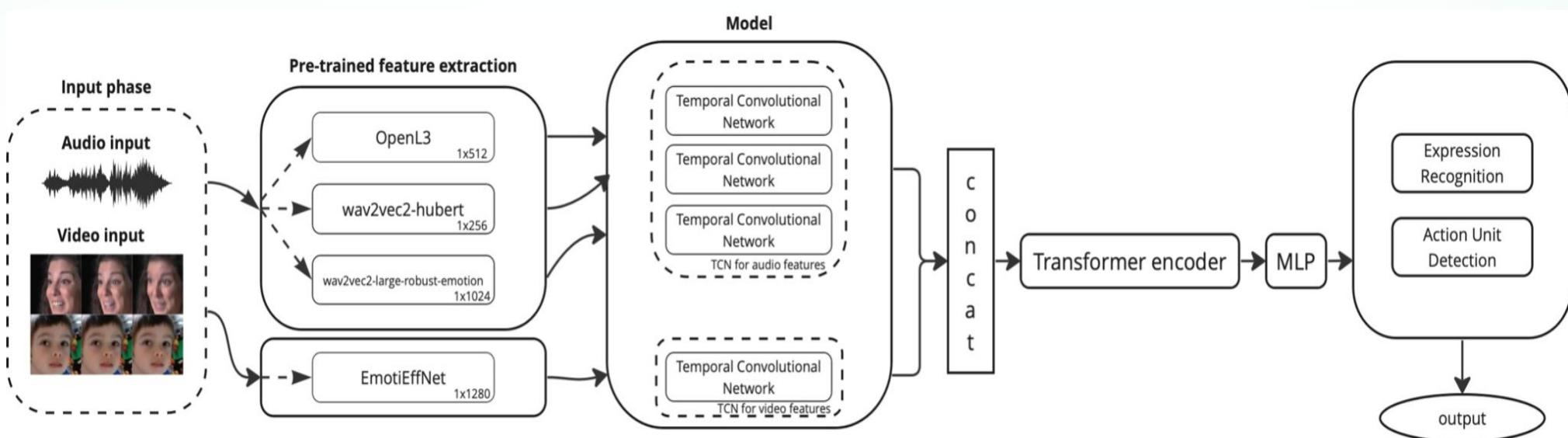
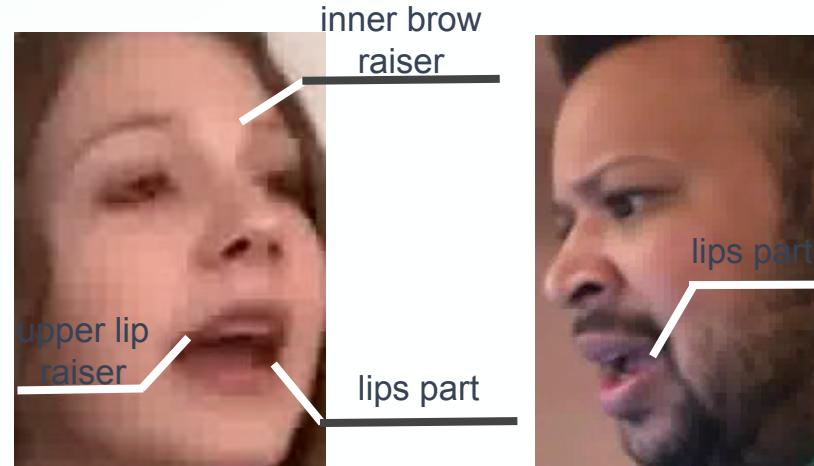
(c) Multimodal ensemble

EmotiEffNet + Temporal Convolutional Network

Expression Recognition



Action Unit Detection



Результаты экспериментов

ABAW-6: Expression classification

Validation set results

Method	Modality	F1-score P_{EXPR}	Accuracy
Baseline VGGFACE (MixAugment) [26]	Faces	0.25	-
EfficientNet-B0 [38]	Faces	0.402	-
Meta-Classifier [49]	Faces	0.302	0.462
TCN [61]	Audio/video	0.377	-
Transformer [60]	Audio/video	0.406	-
MAE [58]	Audio/video	0.495	-
TCN+MLP	Audio	0.151	0.412
wav2vec 2.0	Audio	0.291	0.410
wav2vec 2.0, smoothing	Audio	0.355	0.521
DDAMFN	Faces	0.308	0.433
EmotiEffNet-B2	Faces	0.320	0.438
MT-EmotiMobileFaceNet	Faces	0.327	0.462
MT-EmotiEffNet	Faces	0.336	0.447
MT-DDAMFN	Faces	0.351	0.469
MT-EmotiMobileViT	Faces	0.356	0.461
EmotiEffNet	Faces	0.384	0.495
EmotiEffNet, smoothing	Faces	0.424	0.543
EmotiEffNet, pre-trained + MLP	Faces	0.395	0.4977
EmotiEffNet, pre-trained + MLP, smoothing	Faces	0.434	0.5463
wav2vec 2.0+EmotiEffNet	Audio/video	0.403	0.520
wav2vec 2.0+EmotiEffNet, smoothing	Audio/video	0.434	0.557
TCN (aligned frames) + MLP	Audio/video	0.353	0.536
TCN (cropped frames) + MLP	Faces	0.403	0.523
EmotiEffNet (aligned) + TCN + transformer	Faces	0.338	0.51
EmotiEffNet (cropped) + TCN + transformer	Audio/video	0.375	0.52
EmotiEffNet (aligned)+ TCN + transformer	Audio/video	0.422	0.55

Pre-trained models

Model	all classes		w/o "Other"	
	F1-score	Accuracy	F1-score	Accuracy
EmotiEffNet-B2	0.229	0.282	0.320	0.443
DDAMFN	0.244	0.315	0.362	0.502
MT-DDAMFN	0.245	0.340	0.366	0.547
MT-EmotiMobileViT	0.248	0.287	0.330	0.434
MT-EmotiMobileFaceNet	0.250	0.325	0.354	0.513
MT-EmotiEffNet	0.254	0.324	0.381	0.522
EmotiEffNet-B0	0.257	0.325	0.383	0.522

Test set results

Model	P_{EXPR}
EmotiEffNet+TCN (train+val)	0.3043
EmotiEffNet, audio+vid	0.3137
EmotiEffNet+TCN+audio	0.3221
EmotiEffNet (train+val)	0.3200
EmotiEffNet+TCN	0.3207
EmotiEffNet+TCN	0.3221
EmotiEffNet+TCN, smoothing	0.3244
wav2vec 2.0+EmotiEffNet (train+val)	0.3301
EmotiEffNet, pre-trained + MLP (train+val)	0.3414

Результаты экспериментов

ABAW-6: Action Unit Detection

Validation set results

Method	Modality	F1-score P_{AU}
Baseline VGGFACE [26]	Faces	0.39
IResnet100 [52]	Faces	0.511
TCN [61]	Audio/video	0.517
Transformer [60]	Audio/video	0.530
Regnet/Video	Vision	0.540
Transformer [46]	Faces	0.543
MAE graph representations [50]	Faces	0.543
MAE [58]	Audio/video	0.567
Regnet [49]	Faces	0.698
wav2vec 2.0	Audio	0.313
DDAMFN	Faces	0.500
MT-EmotiMobileFaceNet	Faces	0.512
MT-DDAMFN	Faces	0.519
EmotiEffNet	Video	0.525
(aligned)+TCN+transformer		
MT-EmotiEffNet	Faces	0.525
EmotiEffNet	Audio/video	0.528
(aligned)+TCN+transformer		
EmotiEffNet	Faces	0.537
EmotiEffNet, smoothing	Faces	0.545
EmotiEffNet + LightAutoML	Faces	0.542
EmotiEffNet + LightAutoML, smoothing	Faces	0.554

Test set results

Model	P_{AU}
EmotiEffNet	0.4726
MT-DDAMFN (train+val)	0.4763
TCN+EmotiEffNet+audio	0.4817
TCN+EmotiEffNet	0.4866
EmotiEffNet + LightAutoML	0.4878

Результаты распознавания эмоций на видео, конкурс ABAW-6

Expression Recognition

4 место (из 24 команд)

Teams	F1
Netease Fuxi AI Lab	0.5005
CtyunAI	0.3625
USTC-IAT-United	0.3534
HSEmotion	0.3414
M2-Lab-Purdue	0.3228
KBS-DGU	0.3005
SUN_CE	0.2877
AIOTB	0.2797
CAS-MAIS	0.265
IMLAB	0.2296
baseline	0.2250

Action Unit Detection

3 место (из 16 команд)

Teams	F1
Netease Fuxi AI Lab	0.5601
CtyunAI	0.4941
HSEmotion	0.4878
USTC-IAT-United	0.484
KBS-DGU	0.4652
M2-Lab-Purdue	0.3832
baseline	0.365

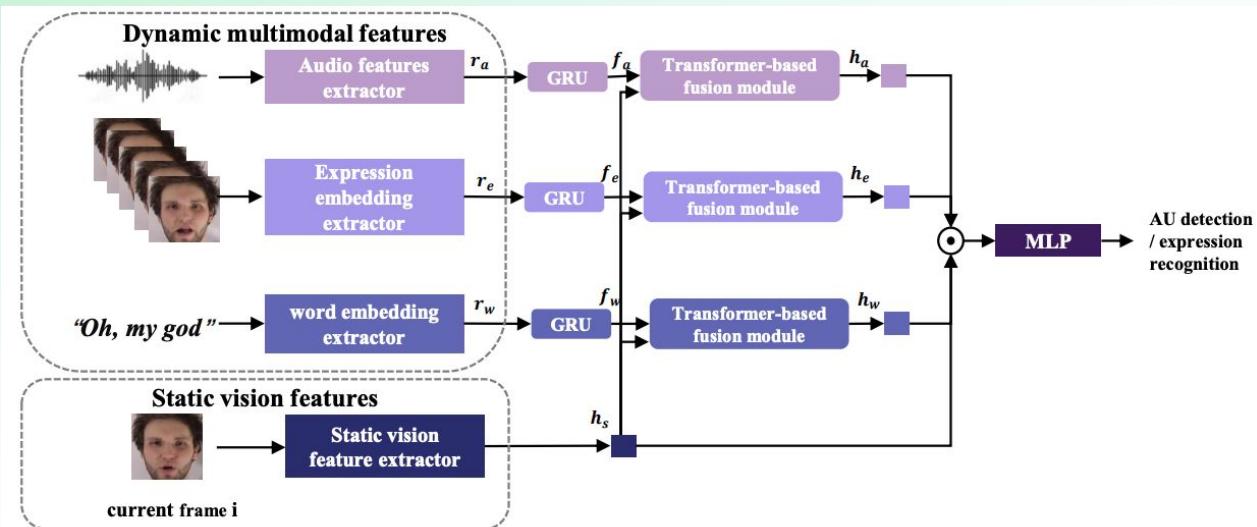
Compound Expression

2 место (из 17 команд)

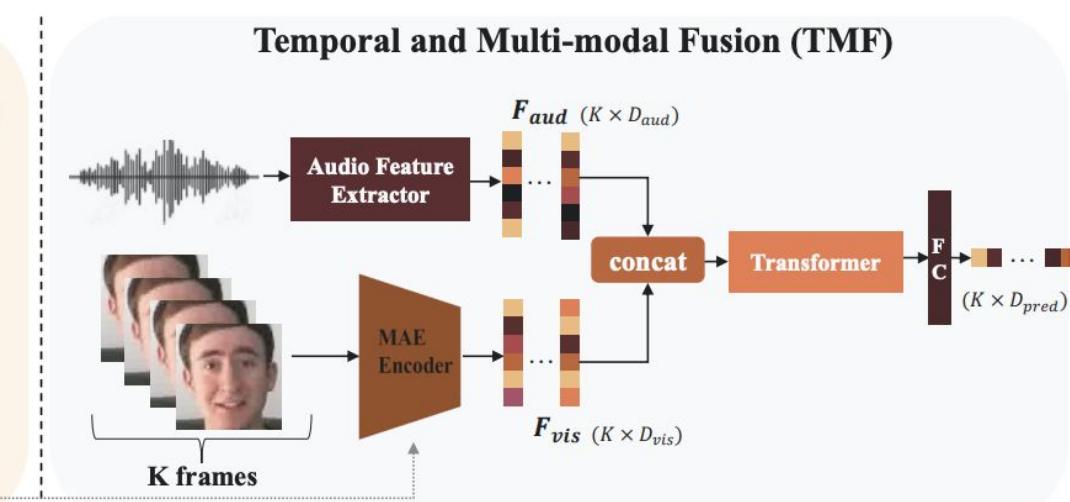
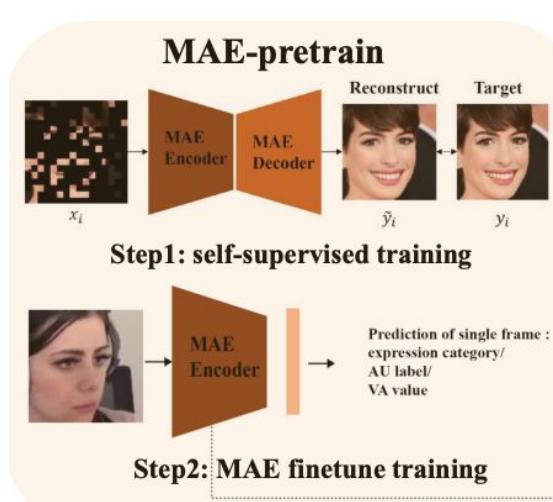
Teams	F1
Netease Fuxi AI Lab	0.5526
HSEmotion	0.2708
USTC-IAT-United	0.2240
SUN_CE	0.2201
USTC-AC	0.1845

Победители ABAW-4/5

Transformers

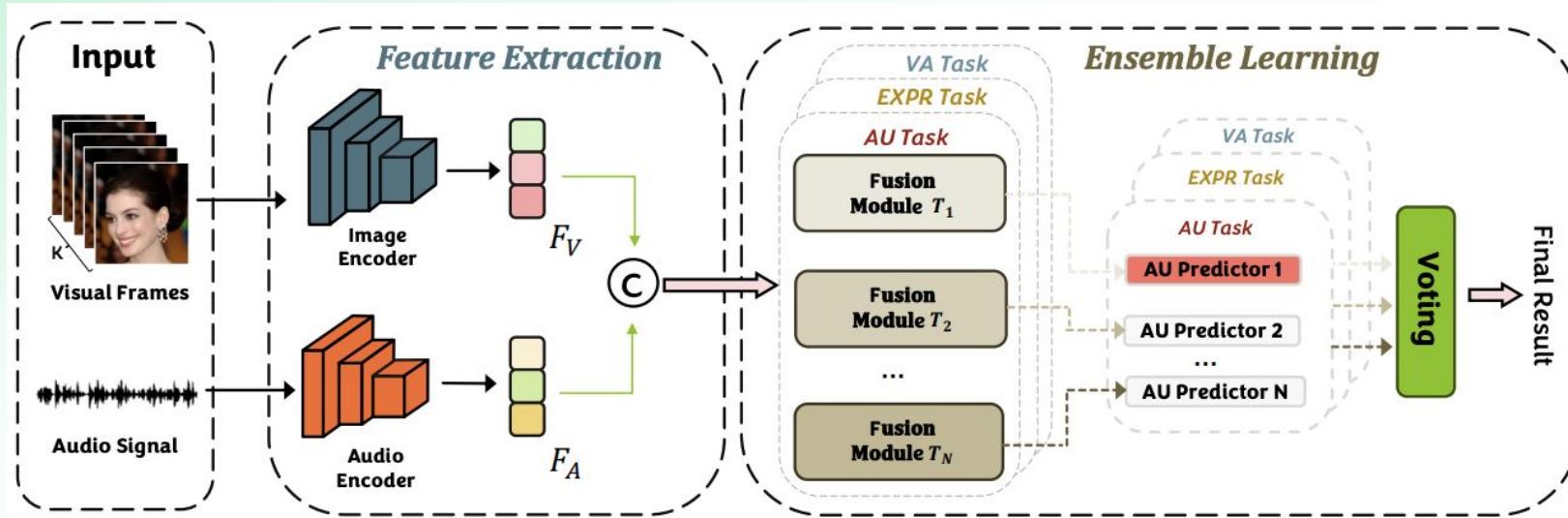


<https://arxiv.org/pdf/2203.12367.pdf>
(CVPR Workshops 2022)



<https://arxiv.org/pdf/2303.10849.pdf> (CVPR Workshops 2023)

Победители ABAW-6 (EXPR, AU, VA)



1. Image encoder для k кадров обучается как self-supervised: MAE (Masked Auto-Encoder) на 4.5M высококачественных кадров из открытых датасетов
2. Audio encoder для всего видео: Vggish, Hubert и Wav2vec2
3. Обучаются N Fusion Modules на поднаборах обучающего множества, разбитого на основе «background characteristics».
4. “In the inference stage, we manually pick out the reaction videos from the test set and utilize the corresponding model to predict their emotional labels” (“reaction videos” - emotional responses to various forms of media)

Результаты экспериментов

ABAW-6: Emotional Mimicry Intensity Estimation

Modality	Model	Features	PCC $\bar{\rho}$	Admiration	Amusement	Determination	Empathic Pain	Excitement	Joy
Faces	Baseline ViT [24]	Embeddings	0.09	-	-	-	-	-	-
	Wav2Vec2 [24]	Embeddings	0.24	-	-	-	-	-	-
	ViT+	Embeddings	0.25	-	-	-	-	-	-
	Wav2Vec2 [24]								
Faces	MobileFaceNet (VggFace2)	Embeddings (mean)	0.0734	0.0235	0.0542	0.0645	0.0837	0.1053	0.1093
		Embeddings (STAT)	0.0972	0.0374	0.1008	0.0981	0.0972	0.1320	0.1175
Faces	DDAMFN	Embeddings (mean)	0.1619	0.0139	0.2515	0.1211	0.0841	0.2373	0.2641
		Embeddings (STAT)	0.1603	0.0595	0.2169	0.1355	0.0687	0.2245	0.2565
		Scores (mean)	0.1640	0.0174	0.2462	0.1257	0.0740	0.2438	0.2770
		Scores (STAT)	0.1684	0.0354	0.2461	0.1304	0.0634	0.2426	0.2927
Faces	EmotiEffNet -B0	Embeddings (mean)	0.1647	0.0472	0.2387	0.1272	0.1017	0.2225	0.2508
		Embeddings (STAT)	0.1658	0.0596	0.2308	0.1318	0.0743	0.2373	0.2611
		Scores (mean)	0.1597	0.0163	0.2342	0.1315	0.0708	0.2281	0.2765
		Scores (STAT)	0.1645	0.0186	0.2477	0.1277	0.0787	0.2278	0.2863
Faces	MT- EmotiEffNet -B0	Embeddings (mean)	0.1632	0.0162	0.2336	0.1239	0.1001	0.2339	0.2715
		Embeddings (STAT)	0.1673	0.0349	0.2318	0.1379	0.0877	0.2428	0.2683
		Scores (mean)	0.1584	0.0275	0.2115	0.1258	0.0805	0.2273	0.2776
		Scores (STAT)	0.1590	0.0188	0.2335	0.1150	0.0729	0.2312	0.2828
Faces	MT- EmotiMobile- ViT	Embeddings (mean)	0.1644	0.0379	0.2314	0.1387	0.0781	0.2334	0.2672
		Embeddings (STAT)	0.1683	0.0433	0.2459	0.1347	0.0779	0.2382	0.2699
		Scores (mean)	0.1642	0.0321	0.2484	0.1490	0.0674	0.2399	0.2481
		Scores (STAT)	0.1727	0.0621	0.2548	0.1430	0.0624	0.2398	0.2738
Faces	MT- DDAMFN	Embeddings (mean)	0.1628	0.0289	0.2385	0.1281	0.0761	0.2363	0.2689
		Embeddings (STAT)	0.1723	0.0613	0.2319	0.1282	0.1064	0.2446	0.2610
		Scores (mean)	0.1682	0.0408	0.2333	0.1387	0.0825	0.2429	0.2710
		Scores (STAT)	0.1703	0.0289	0.2450	0.1298	0.0878	0.2410	0.2895
Faces	MT- Emoti- MobileFaceNet	Embeddings (mean)	0.1518	0.0215	0.2288	0.1140	0.0692	0.2299	0.2476
		Embeddings (STAT)	0.1646	0.0557	0.2380	0.1303	0.0703	0.2325	0.2605
		Scores (mean)	0.1667	0.0276	0.2367	0.1336	0.0807	0.2516	0.2699
		Scores (STAT)	0.1732	0.0285	0.2498	0.1318	0.097	0.2543	0.2776
Audio	wav2vec 2.0	Embeddings (mean)	0.1514	0.2153	0.11760	0.1834	0.1426	0.1275	0.1219
		Embeddings (STAT)	0.2311	0.3006	0.1659	0.2559	0.3198	0.1844	0.1602
Audio + Video	wav2vec 2.0 +	MT-DDAMFN	0.2767	0.2993	0.3079	0.2230	0.2672	0.3008	0.2546
		MT-EmotiMobileViT	0.2829	0.3011	0.2968	0.2595	0.3074	0.3171	0.2152
		MT-EmotiMobileFaceNet	0.2898	0.3041	0.3004	0.2584	0.3148	0.3160	0.2452

Результаты экспериментов

ABAW-6: Emotional Mimicry Intensity Estimation

Test set results

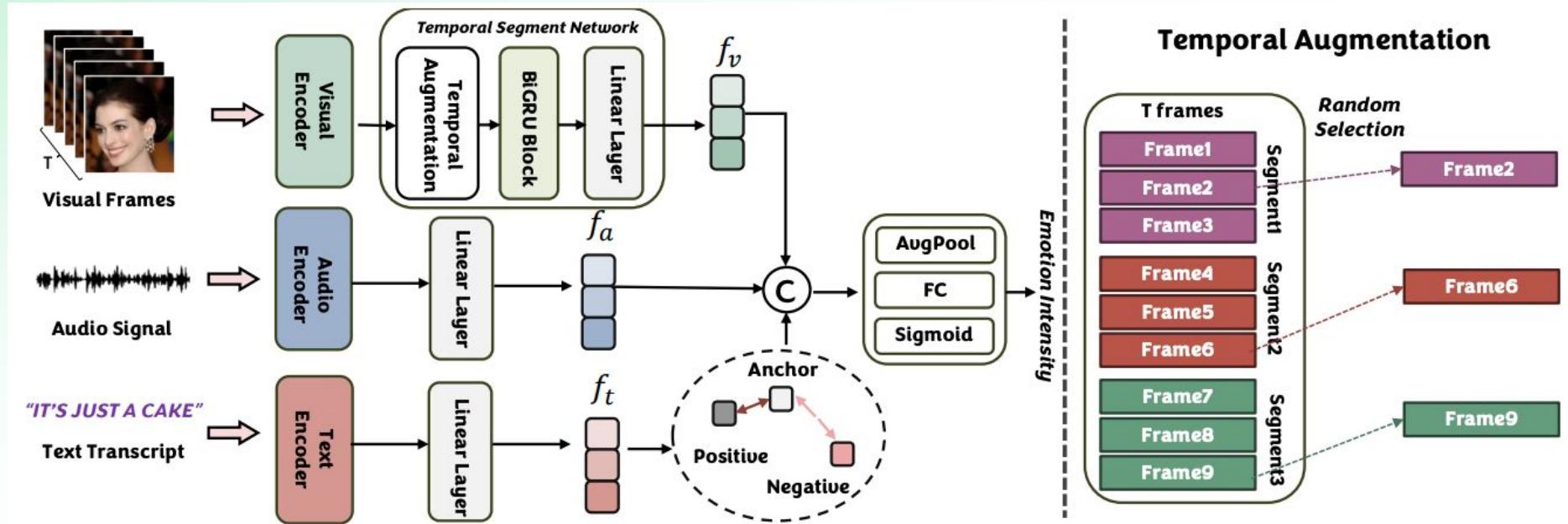
Model	F1-score
Netease Fuxi AI Lab [55]	0.7185
USTC-IAT-United [10]	0.5536
USTC-AC [50]	0.3594
wav2vec 2.0 + MT-EmotiMobileFaceNet (train)	0.3201
wav2vec 2.0 + MT-EmotiMobileFaceNet (train+val)	0.3285
MT-EmotiMobileFaceNet (train+val)	0.1786
wav2vec 2.0 + MT-EmotiMobileViT (train+val)	0.3316
wav2vec 2.0 + MT-DDAMFN (train+val)	0.3139
Baseline [24]	0.25

Leaderboard

ID	Team Name	Best PCC results
1	Netease Fuxi AI Lab	0.7185
2	HCAI-VIS	0.5536
3	USTC-IAT-United	0.3594
4	HSEmotion	0.3316

4th place out of 7 teams that submitted their results

Победители ABAW-6 (EMI)

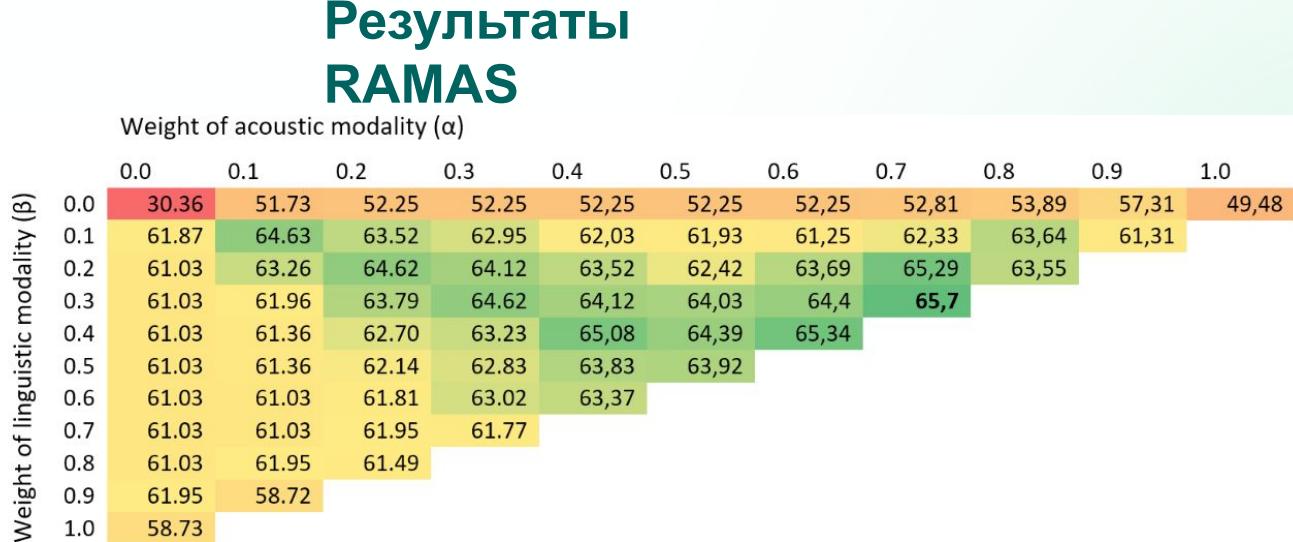
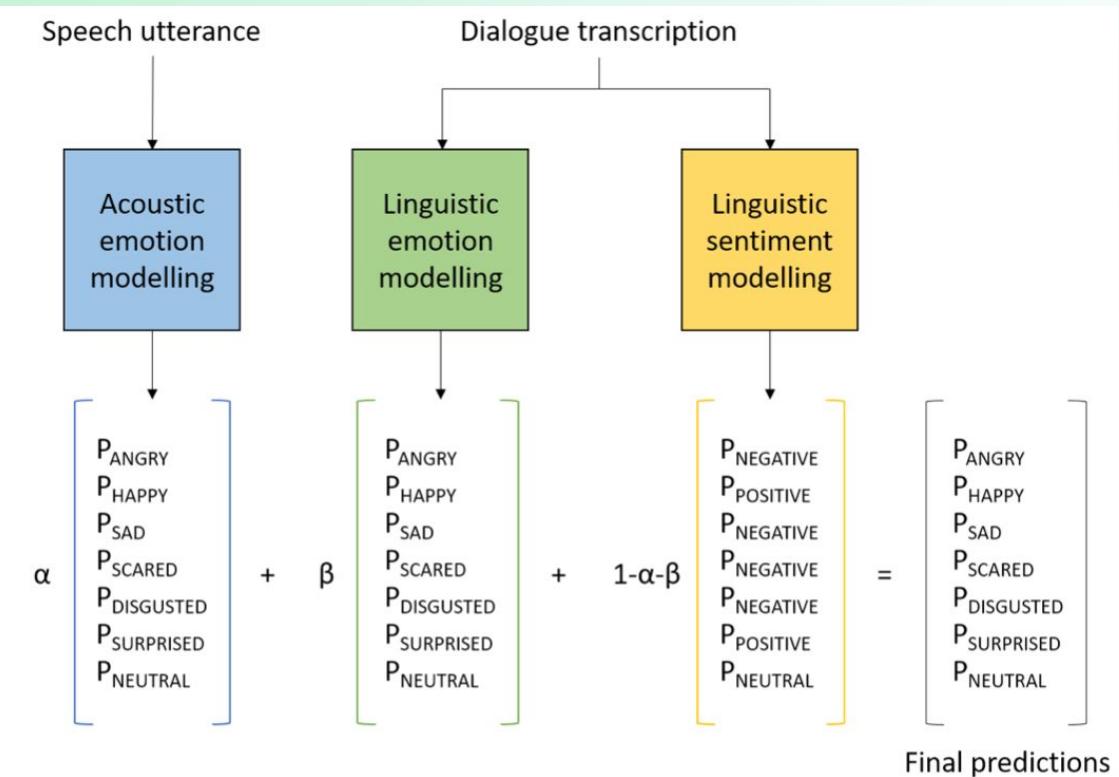


1. Image encoder: ViT-based MAE
2. Audio encoder: Wav2vec2
3. Text processing: Whisper (speech2text) + ChatGLM3 (word embeddings) + Global Average Pooling
4. Основная модальность – текстовая. Используют contrastive learning для выравнивания аудиовизуальных и текстовых признаков. Триплет (anchor, pos, neg) формируется на основе текстовых признаков

$$\hat{f} = \text{AvgPool}(\hat{f}_v, \hat{f}_a, \hat{f}_t)$$

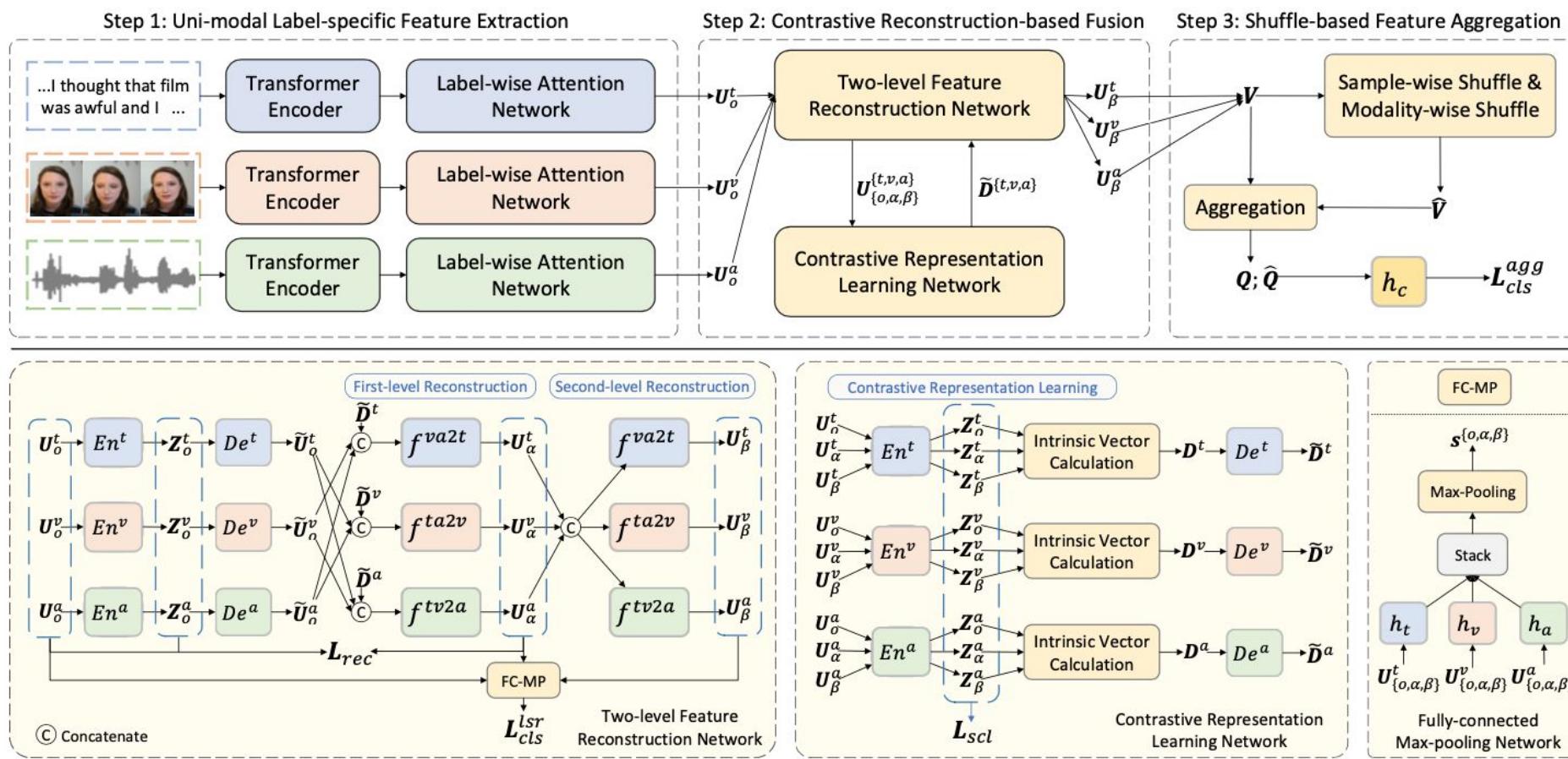
$$\begin{aligned} \mathcal{L}_{triplet} &= \max \left(0, \left\| \hat{f}_{anc} - \hat{f}_{pos} \right\|^2 - \left\| \hat{f}_{anc} - \hat{f}_{neg} \right\|^2 + \gamma \right) \\ &\quad + \max \left(0, \left\| \hat{f}_{pos} - \hat{f}_{neg} \right\|^2 + \left\| \hat{f}_{anc} - \hat{f}_{pos} \right\|^2 + \gamma \right) \end{aligned}$$

Аудио + текст



CARAT: текст + голос + видео

ContrAstive feature Reconstruction and AggregaTion

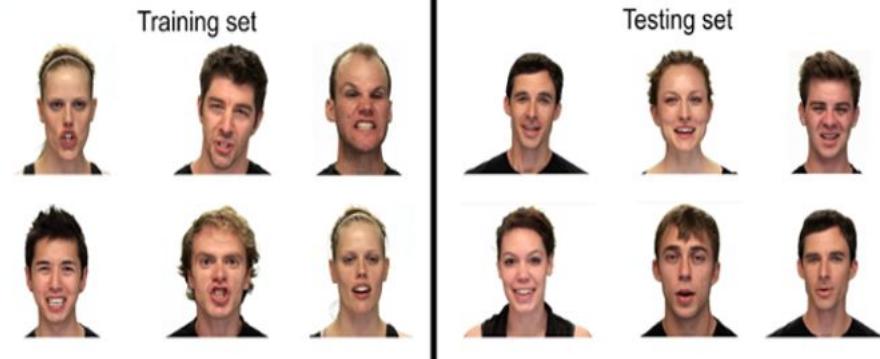


1. Независимые transformer encoder для текста, звука и видео лиц
2. label-specific representations для каждой эмоции – multi-head attention
3. Multi-modal Feature Reconstruction: контрастное обучение для восстановления признаков одной модальности по их распределению и признакам других модальностей

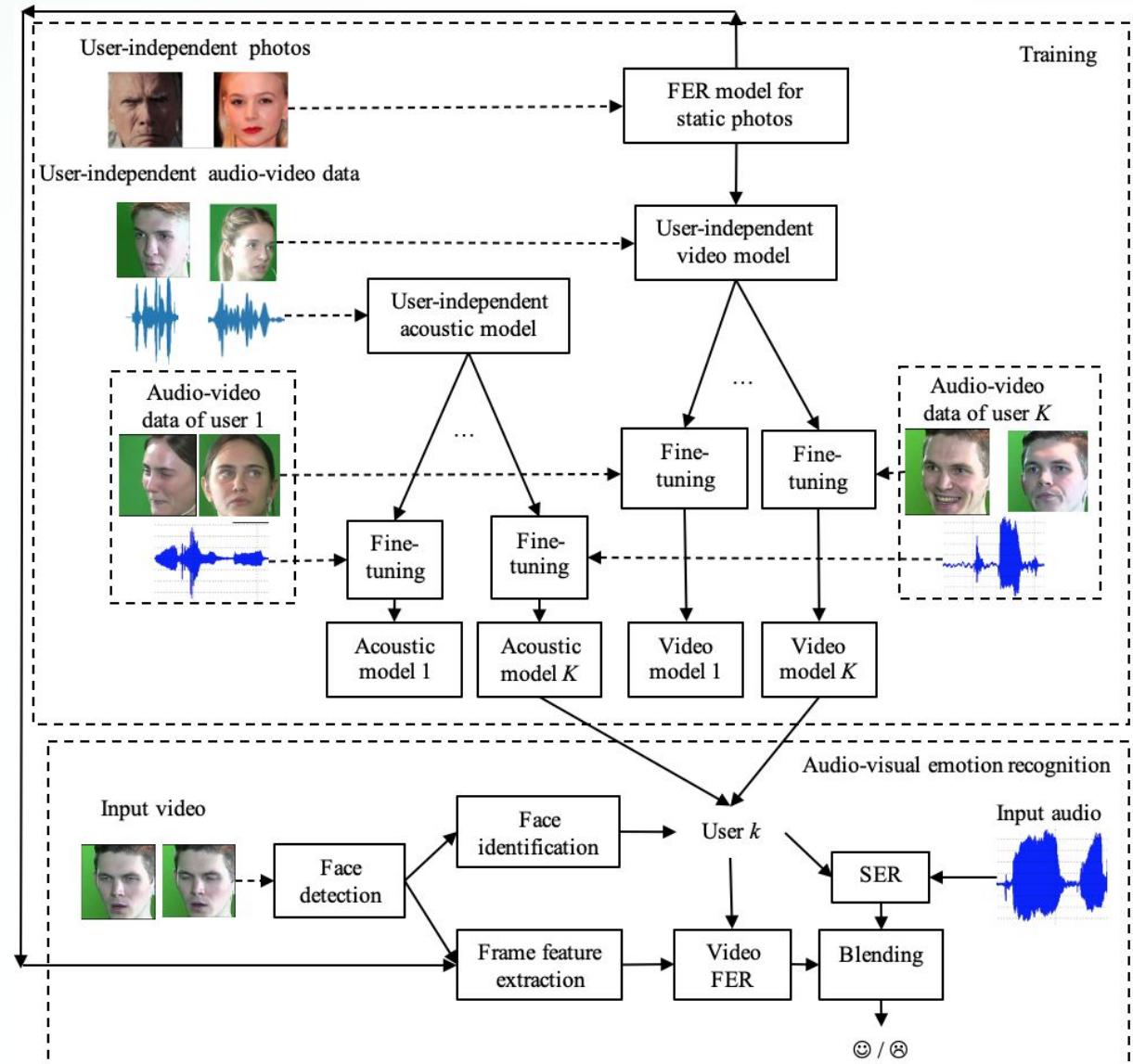
Персонализация

Технология аудиовизуального распознавания эмоций

Дикторонезависимый режим



Дикторозависимый режим



Результаты экспериментов

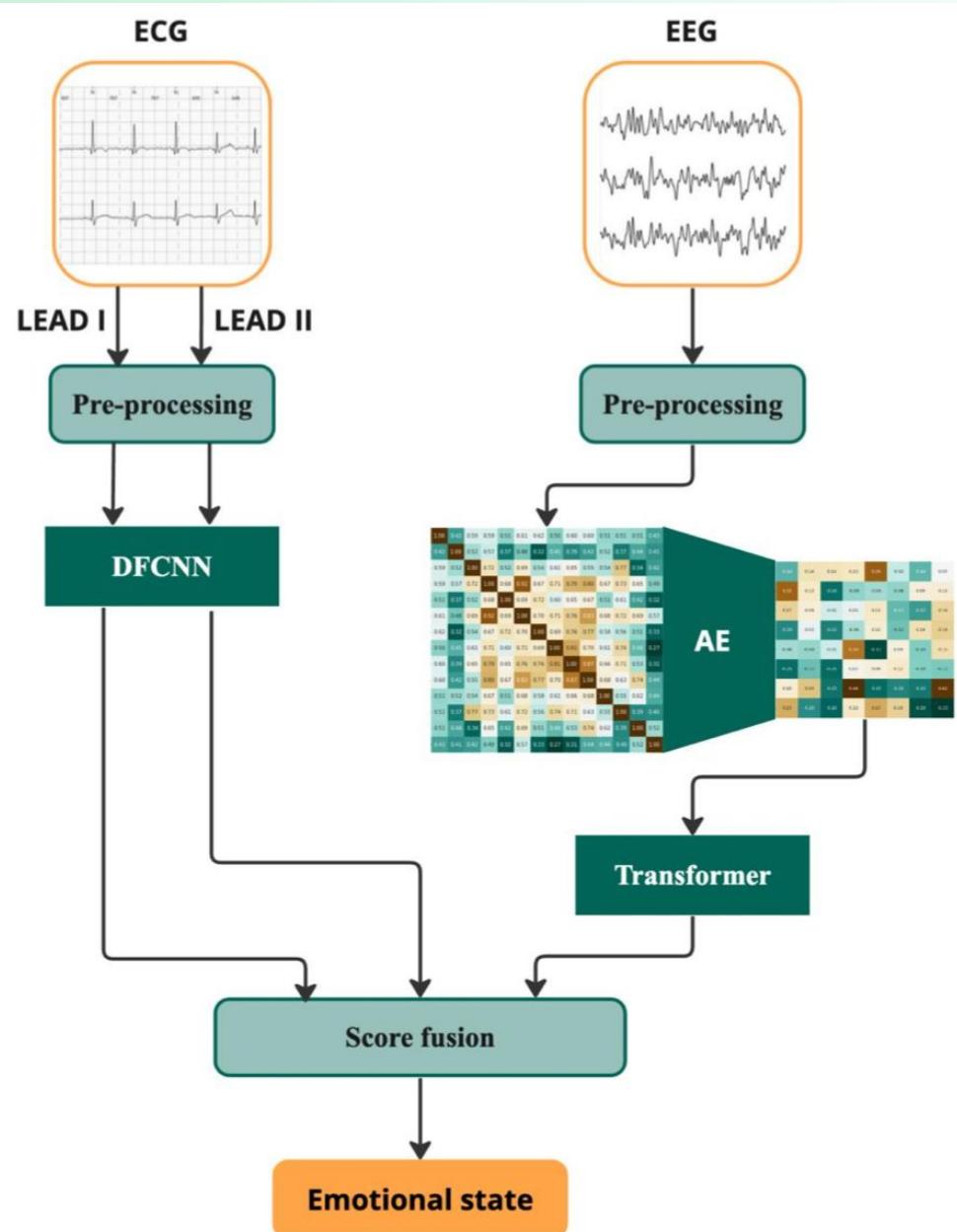
RAMAS

Протокол

10-fold cross-validation (actor-based split): аудио и видео 9 пользователей использовались для обучения универсальной (дикторонезависимой) модели). Видео оставшегося актера наугад разбивались на 2 части: для дообучения и тестирования

Modality	Features	Classifier	Metric	Number of agreed annotators				
				1	2	3	4	5
Audio	Emobase from OpenSmile	Speaker-independent MLP	UAR, %	33.3	41.2	49.7	53.1	46.6
			Accuracy, %	32.6	42.5	47.9	52.3	46.5
	OpenL3	Proposed personalized model	UAR, %	35.0	44.2	54.5	55.7	52.4
			Accuracy, %	34.3	43.8	51.2	52.9	51.2
Audio	OpenL3	Speaker-independent MLP	UAR, %	37.3	47.9	54.1	52.7	53.4
			Accuracy, %	32.8	46.1	51.5	51.3	52.5
	EfficientNet	Proposed personalized model	UAR, %	37.4	50.5	58.7	58.7	60.2
			Accuracy, %	38.6	49.7	52.7	54.5	56.6
Video	EfficientNet	Speaker-independent MLP	UAR, %	33.3	47.9	50.4	52.7	47.0
			Accuracy, %	43.1	57.5	58.9	66.7	68.4
	MobileNet	Proposed personalized model	UAR, %	47.7	64.2	72.2	79.1	80.7
			Accuracy, %	49.3	66.4	71.9	80.2	83.6
Video	MobileNet	Speaker-independent MLP	UAR, %	32.4	49.5	49.7	54.8	46.5
			Accuracy, %	43.8	54.6	54.3	61.4	66.3
	OpenL3 + EfficientNet	Proposed personalized model	UAR, %	46.5	63.2	69.5	77.1	79.1
			Accuracy, %	48.9	65.0	56.4	77.8	82.0
Audio + video	OpenL3 + EfficientNet	Speaker-independent MLP	UAR, %	45.0	58.1	65.0	69.5	71.3
			Accuracy, %	45.0	62.6	64.6	72.0	74.4
	OpenL3 + MobileNet	Proposed personalized model	UAR, %	46.4	66.9	76.9	82.1	81.8
			Accuracy, %	49.5	68.2	76.0	82.5	83.6
Audio + video	OpenL3 + MobileNet	Speaker-independent MLP	UAR, %	42.5	54.9	61.8	67.9	69.2
			Accuracy, %	48.8	59.2	62.5	69.0	72.0
	OpenSmile + EfficientNet	Proposed personalized model	UAR, %	46.5	66.9	73.2	81.9	83.8
			Accuracy, %	48.8	67.8	73.5	81.2	84.8
Audio + video	OpenSmile + EfficientNet	Speaker-independent MLP	UAR, %	46.2	56.2	61.5	65.6	67.5
			Accuracy, %	46.2	60.8	62.6	67.4	70.5
	OpenSmile + MobileNet	Proposed personalized model	UAR, %	50.3	64.1	74.7	77.1	82.0
			Accuracy, %	51.1	67.7	73.6	77.9	84.3
Audio + video	OpenSmile + MobileNet	Speaker-independent MLP	UAR, %	42.5	52.0	57.3	61.5	67.1
			Accuracy, %	45.8	57.1	59.7	62.9	69.8
	Proposed personalized model		UAR, %	48.5	61.0	71.9	76.4	81.0

ЭЭГ + ЭКГ



Результаты PROGRAM

Method	Modality	Valence	Arousal
SVM (baseline) [27]	ECG	62.4	62.4
Bayesian deep learning [55]	ECG	86.0	83.0
Unsupervised Domain Adaptation (UDA) [56]	ECG	72.0±8.0	71.0±9.0
AD-TCN (Adversarial Discriminative Temporal Convolutional Network) [57]	ECG	66.6±10.0	63.7±6.6
SVM (baseline) [27]	EEG	62.5	62.2
BiSMSM [58]	EEG	61.9 ± 8.8	64.3 ± 6.9
MSD-SS-SAN [59]	EEG	74.1	72.8
SSTD [45]	EEG	76.8	81.6
STSNet [44]	EEG	78.3	82.4
MTLFuseNet [46]	EEG	80.4±8.0	83.3±11.2
SVM (baseline) [27]	ECG + EEG	61.8	62.3
Proposed ECG pipeline	ECG	82.6 ± 4.1	83.7 ± 4.5
Proposed EEG pipeline	EEG	76.0 ± 4.5	73.0 ± 4.9
Proposed multimodal pipeline	ECG + EEG	84.3 ± 3.8	84.8 ± 4.1

Мультиагентный эмоциональный ИИ

«Нет слов, одни эмоции»

Запрос: «Расскажи руководителю о предложении добавить эмоций в персонализированный интеллектуальный помощник, чтобы повысить его популярность»

Радость

Прекрасная идея! Добавление эмоций в интеллектуального помощника сделает его более **привлекательным и гармоничным** для пользователей.

Печаль

Эта идея, конечно, звучит интересно, но стоит **учесть, что сложности** в определении истинных эмоций могут создать непредсказуемые ситуации. Стоит обдумать, как **минимизировать риски**, связанные с этим аспектом.

Страх

Мне кажется, что добавление эмоций может вызвать **беспокойство** у пользователей в связи с **приватностью** и безопасностью данных.

Гнев

Это **абсолютно безумная идея!** Добавление эмоций только **усложнит** нашу работу и создаст множество проблем. **Нет смысла тратить время на нечто такое!**

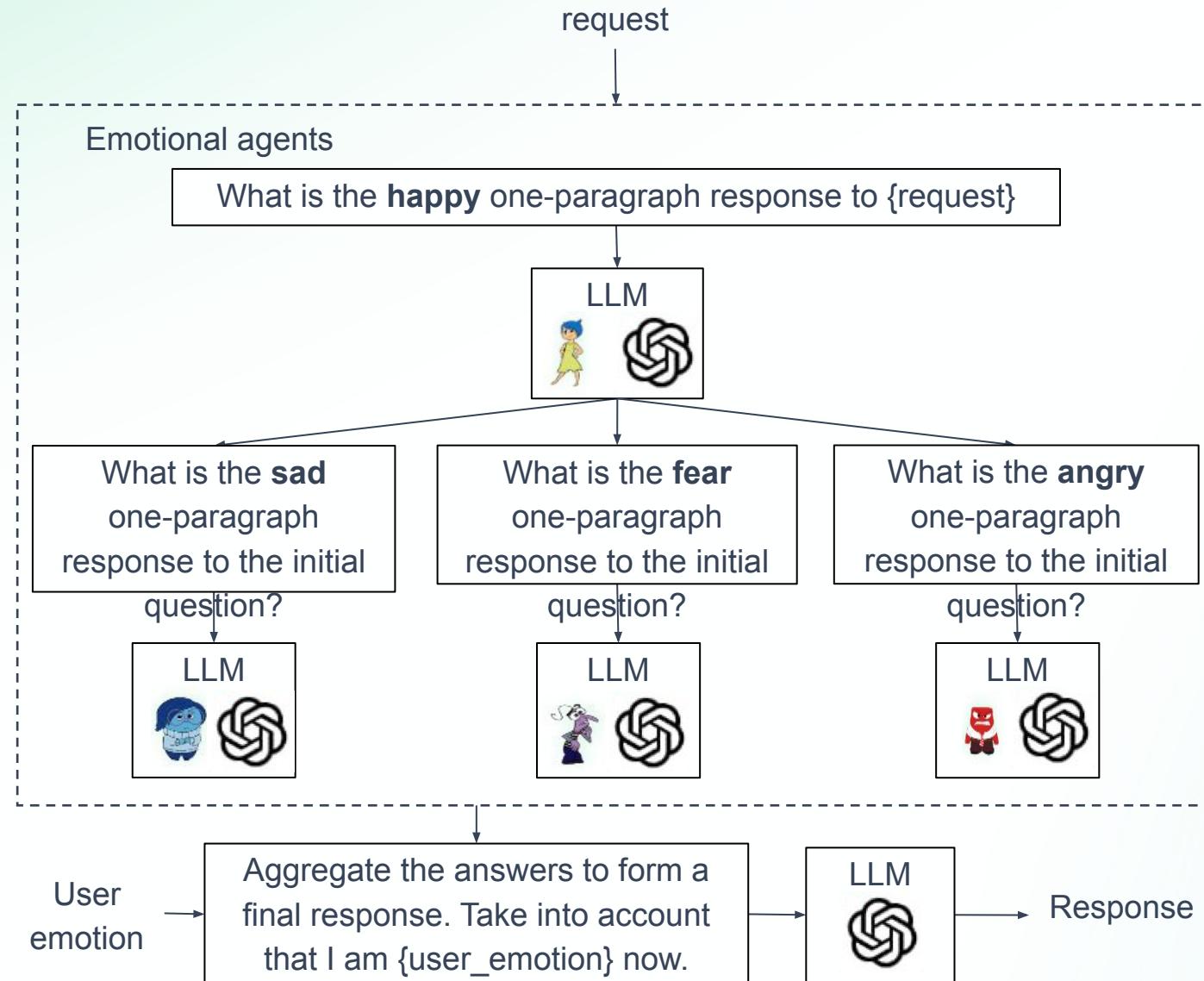


Итоговый ответ

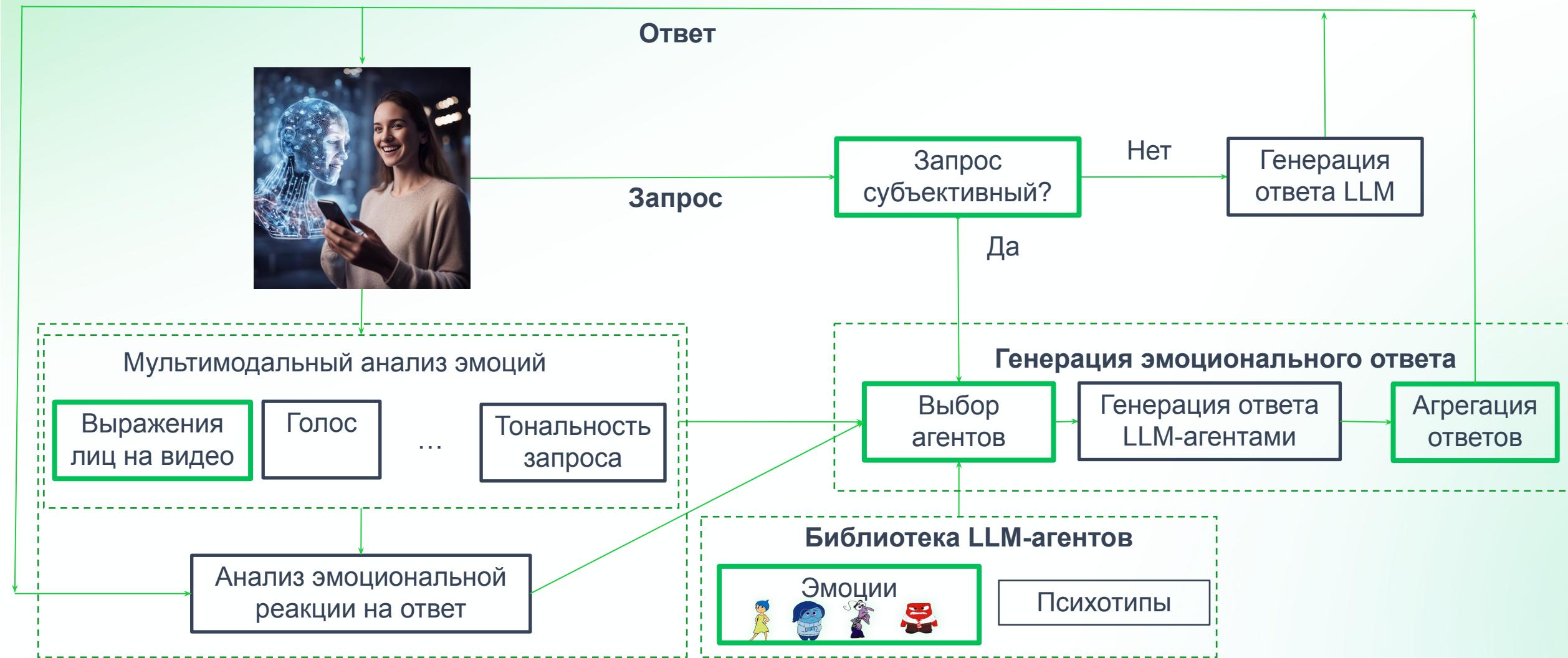
Идея звучит интересно, однако стоит учесть несколько аспектов. **С радостью** приветствуем возможность улучшить пользовательский опыт и повысить конкурентоспособность, но необходимо **тщательно обдумать** вопросы приватности и безопасности данных.



InsideOut: Генератор эмоциональных ответов



Технология эмоционального ИИ



Результаты экспериментов

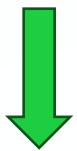
EmphaticDialogues

Пример диалога

Speaker 1: I couldn't wait to go to the concert.

Speaker 2: Which Concert?

Speaker 1: The U2 concert. Tickets were really expensive and I never thought we would be able to go, but somehow we did!!!



Эмоция: Guilty

Модель	Точность
MoEL (Lin et al., 2019)	31,74
MIME (Majumder et al., 2020)	30,96
EmpDG (Li et al., 2020)	31,65
CEM (Sabour et al., 2021)	36,84
KEMP (Li et al., 2021)	36,57
GigaChat-7b-base	30,79
GigaChat-7b-base + InsideOut	33,31
GigaChat -7b-base + fastText	33,18
GigaChat-7b-base + InsideOut + fastText	35,19
GPT-4	44,2
GPT-4 + InsideOut	45,1

01 Введение: представление эмоций, модальности, датасеты

02 Распознавание выражений лиц по видео

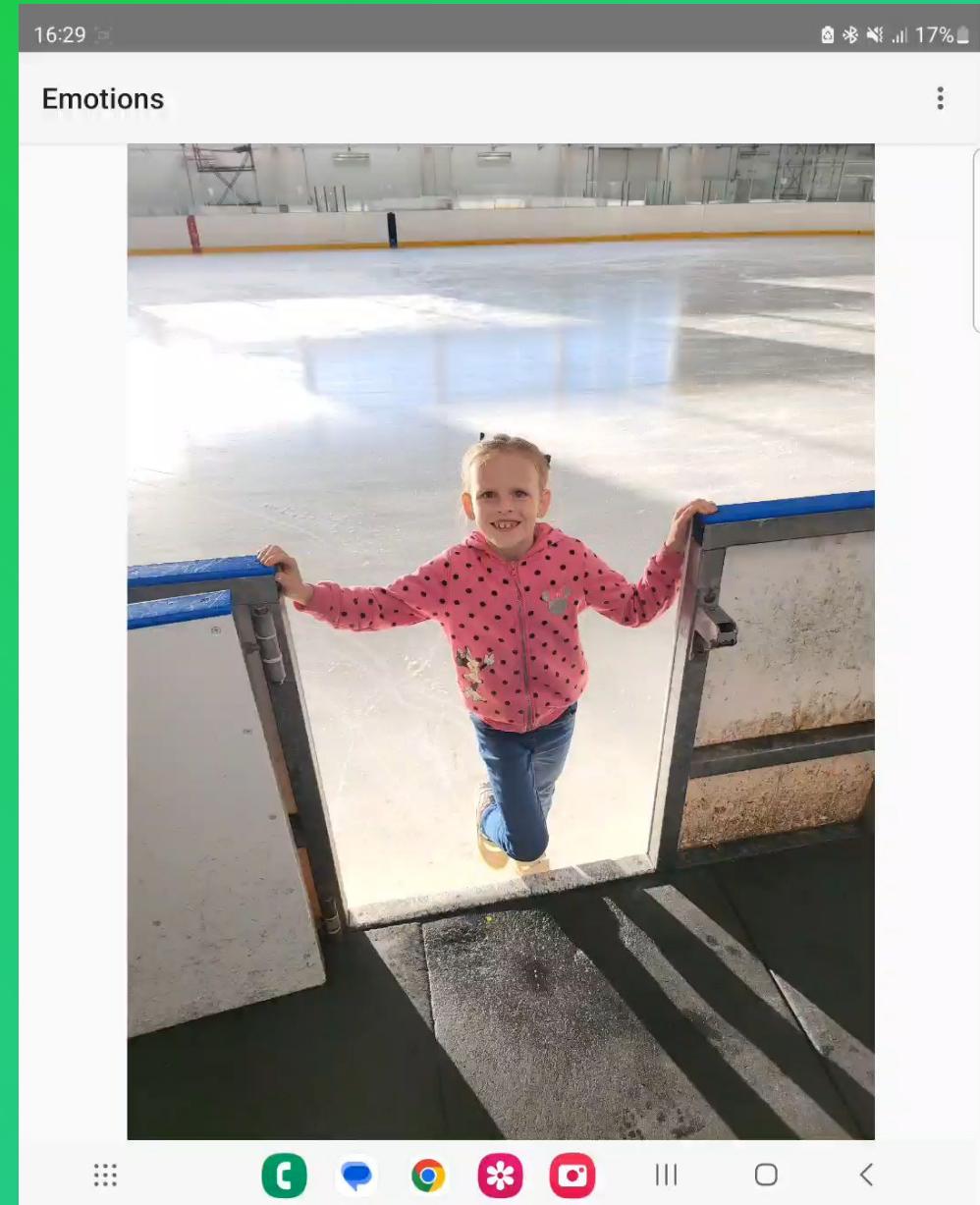
03 Распознавание эмоций по голосу

04 Мультимодальные методы

05 Демо и примеры

Мобильное приложение

https://github.com/av-savchenko/face-emotion-recognition/tree/main/mobile_app







Демо центра ИИ в медицине

Уважаемый пользователь, при температуре 38 градусов и выше рекомендуется обратиться к врачу для диагностики и лечения. Занятия спортом, особенно интенсивные, могут усугубить состояние и привести к осложнениям. Поэтому, в данном случае, я рекомендую воздержаться от занятий спортом до тех пор, пока температура не нормализуется.

Температура 38, а вид спорта подтягивание.

Для ответа на ваш вопрос мне необходимо знать, насколько повышена температура тела и какие виды спорта вы имеете в виду. Однако, я рекомендую обратиться к врачу для диагностики и лечения, если температура сохраняется в течение длительного времени.

Можно ли заниматься спортом при повышенной температуре тела?

Активирован режим Gigapevt

Привет! Я - GigaChat, генеративная языковая модель. Моя задача - помогать вам в решении различных задач, связанных с текстом. Я умею писать тексты на русском языке, отвечать на вопросы, поддерживать диалог, генерировать изображения и видео. А как вы себя чувствуете?

Начать диагностику start diagnostic

Подойдите ближе к экрану так, чтобы шкала стала полностью зеленой. Не отходите до окончания измерения.

start dialog

Начать диалог

Зажмите кнопку "Микрофон" на пульте и говорите в его верхнюю часть.

save the result

Сохранить результат

Сфотографируйте экран, если хотите сохранить данные диагностики.

Основные данные

gender Пол Мужской **age** Возраст 37 ±1

Эмоциональное состояние **emotion** Печаль

Индекс массы тела **bmi** 19.1 В норме

Риск диабета II типа **diabetes risk** 0.025 В норме

Данные диагностики

heart rate Пульс

sdnn Отклонение интервалов ритма сердца

stress Уровень стресса

rmssd Восстановление организма

PNN50

Артериальное давление

Температура тела

in development

Сатурация

Частота дыхательных движений

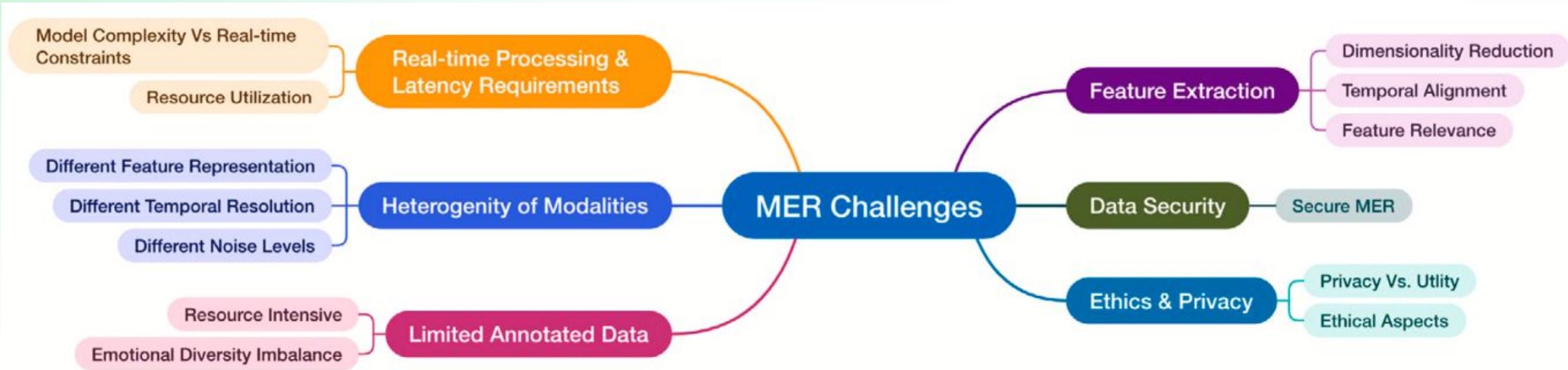
Гликированный гемоглобин

В разработке

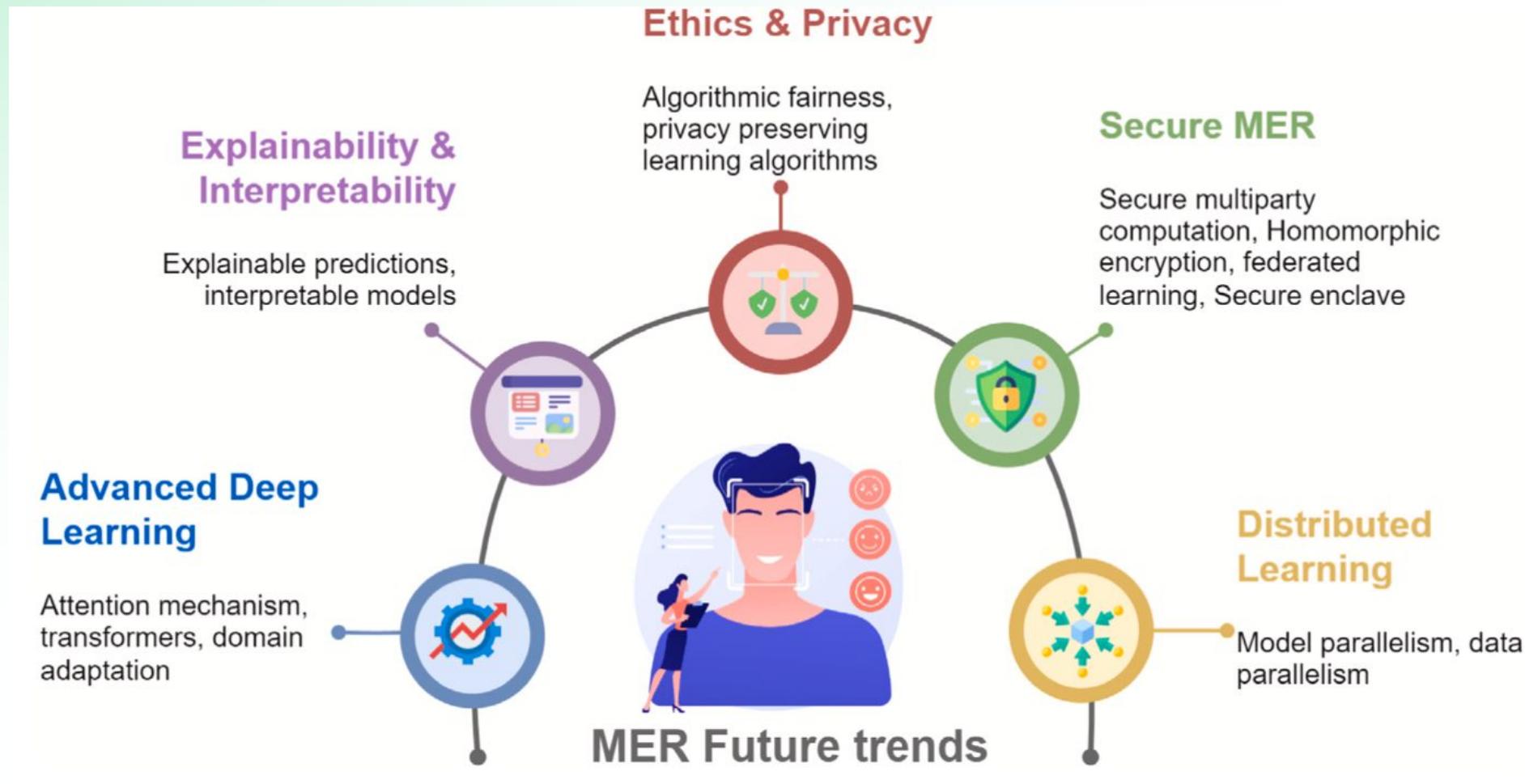
В разработке

В разработке

ВЫЗОВЫ



Тренды



Как использовать?

Общение с пользователями

Персональный ИИ-помощник: ИИ-психолог (анализ выгорания), ассистент врача,...

Coaching, объяснение рекомендаций: тестирование soft skills, анализ публичных выступлений

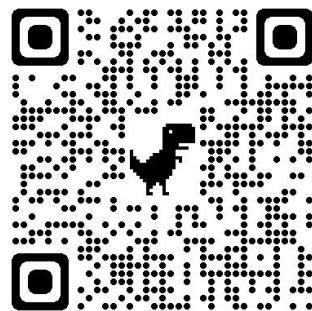
ИИ в образовании: ассистент преподавателя, системы e-learning



Павильон ВДНХ №57. «Россия — моя история». Просвещение/Десятилетие науки и технологий

Sber AI Lab

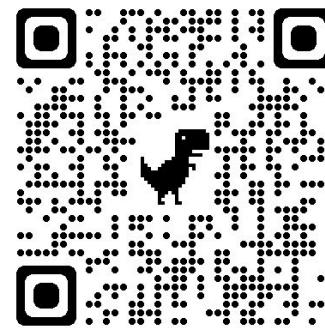
GitHub
sb-ai-lab



andrey.v.savchenko@gmail.com

Андрей Савченко

Сайт
Sber AI Lab



Вакансии
Sber AI Lab

