

Проблематика



- ⊗ Нет текстового контекста
- ⊗ Скрыта мимика
- ⊗ Отсутствует интонация
- ✓ Полный контекст

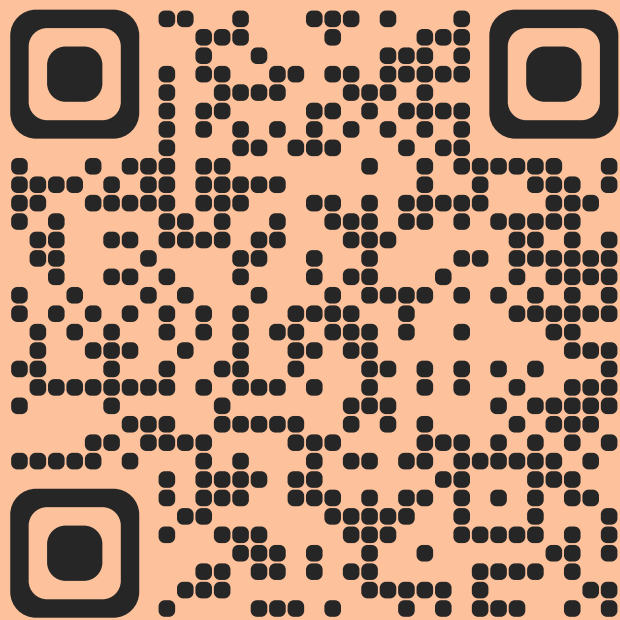
Существующие решения

Мультимодальная структура

Единая модель, которая сразу анализирует видео, аудио и текст.

- ✓ Простота интеграции;
- ✗ Тяжёлая и ресурсозатратная, нет гибкости в обновлении модальностей.

Существующие решения



Emotion-LLaMA

- ✓ Генерирует rationale, высокая точность в “чистых” условиях
- ✗ Требует специализированных датасетов и мощных GPU

Существующие решения



Azure Emotion API

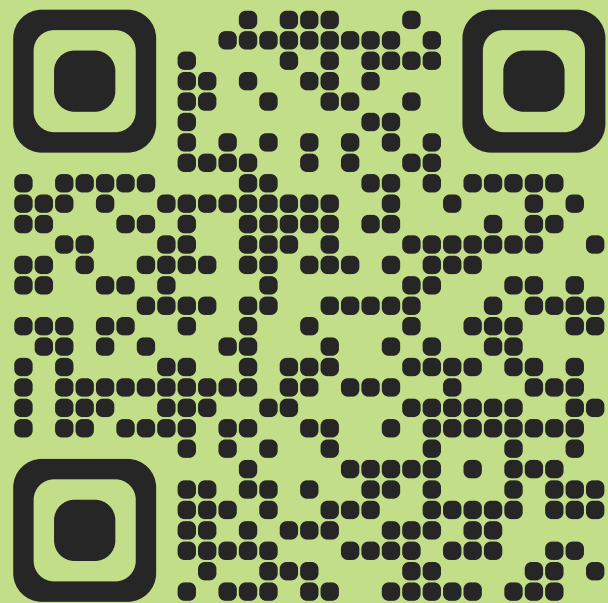
- ✓ Быстрая интеграция, готовый мультимодальный сервис
- ✗ Нет доступа к весам и промежуточным данным, низкая адаптивность

Существующие решения

Модульная структура

- Больше данных для каждой отдельной модальности
- Легче обновлять или подменять один канал без переобучения всей системы
- Меньше вычислительных затрат по сравнению с мультимодальными схемами

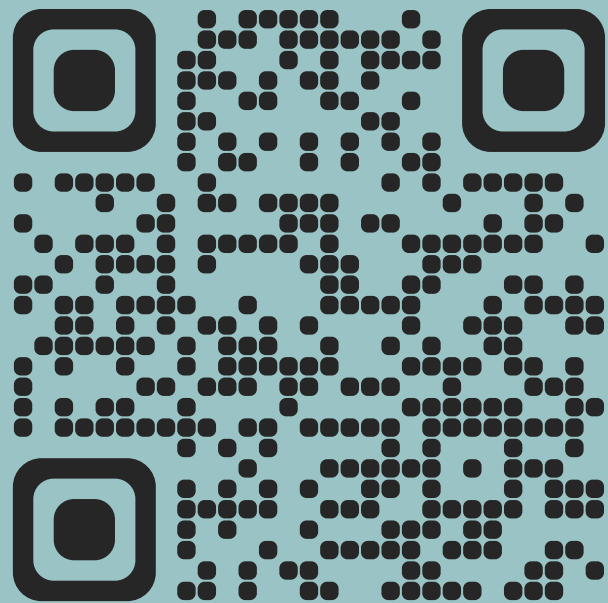
Существующие решения



Shifted Window Transformer V2 (визуальная модальность)

- ✓ Меньше потребление памяти по сравнению с ViT
- ✗ Остаётся требовательным по ресурсам, требует большого количества данных для обучения

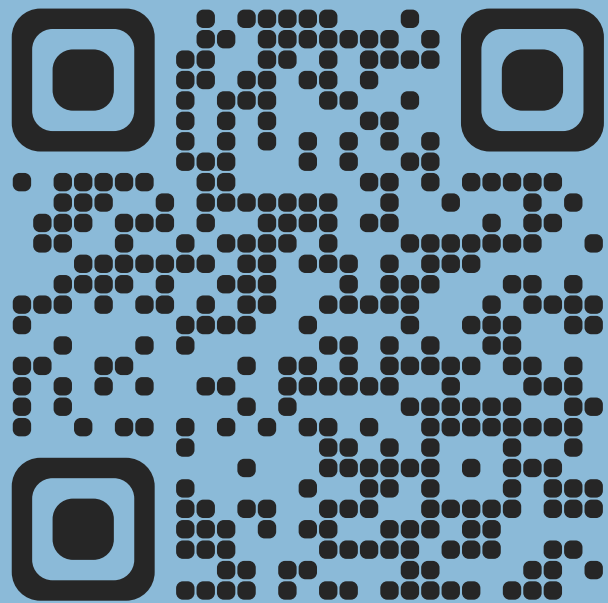
Существующие решения



CNN с учётом времени (визуальная модальность)

- ✓ Учитывает изменение выражения лица во времени
- ✗ Чуть менее точные, чем трансформеры

Существующие решения

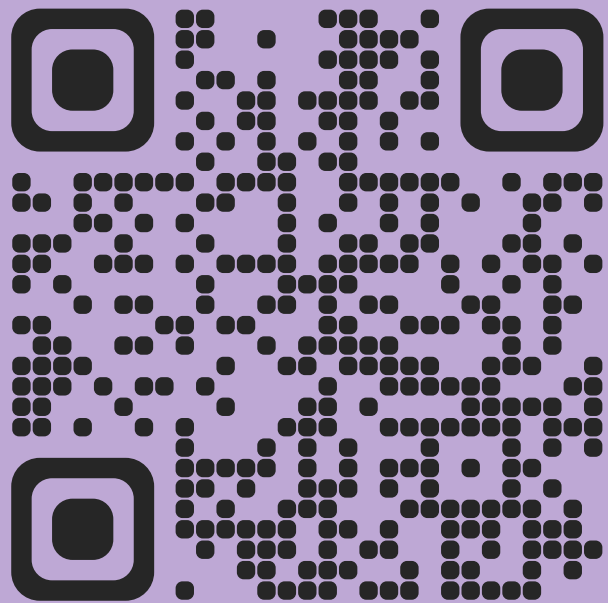


Wav2Vec 2.0

(аудио модальность)

- ✓ Устойчивость к шуму
- ✗ Остаётся требовательным по ресурсам, высокая потребность в вычислениях и памяти

Существующие решения

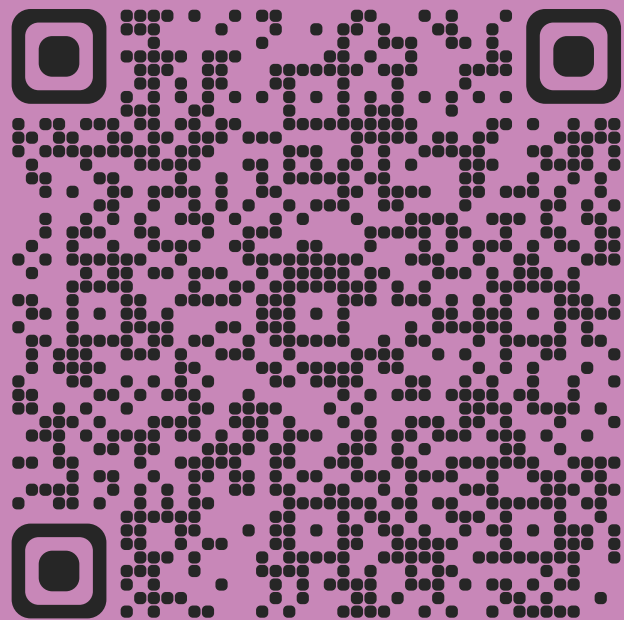


Спектрограммы + CNN

(аудио модальность)

- ✓ Просты в реализации, не требовательны по ресурсам
- ✗ Плохо улавливают длительную динамику

Существующие решения

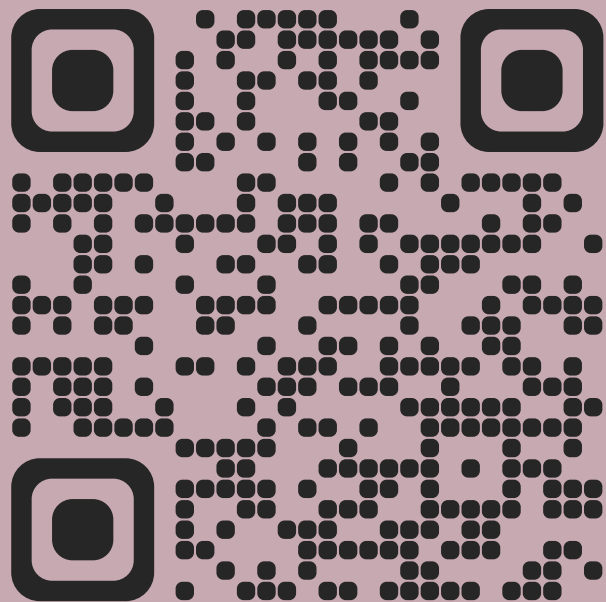


BiLSTM + Attention

(текстовая модальность)

- ✓ Быстрая скорость работы
- ✗ Плохо улавливает контекст, не подходит для большого текста

Существующие решения

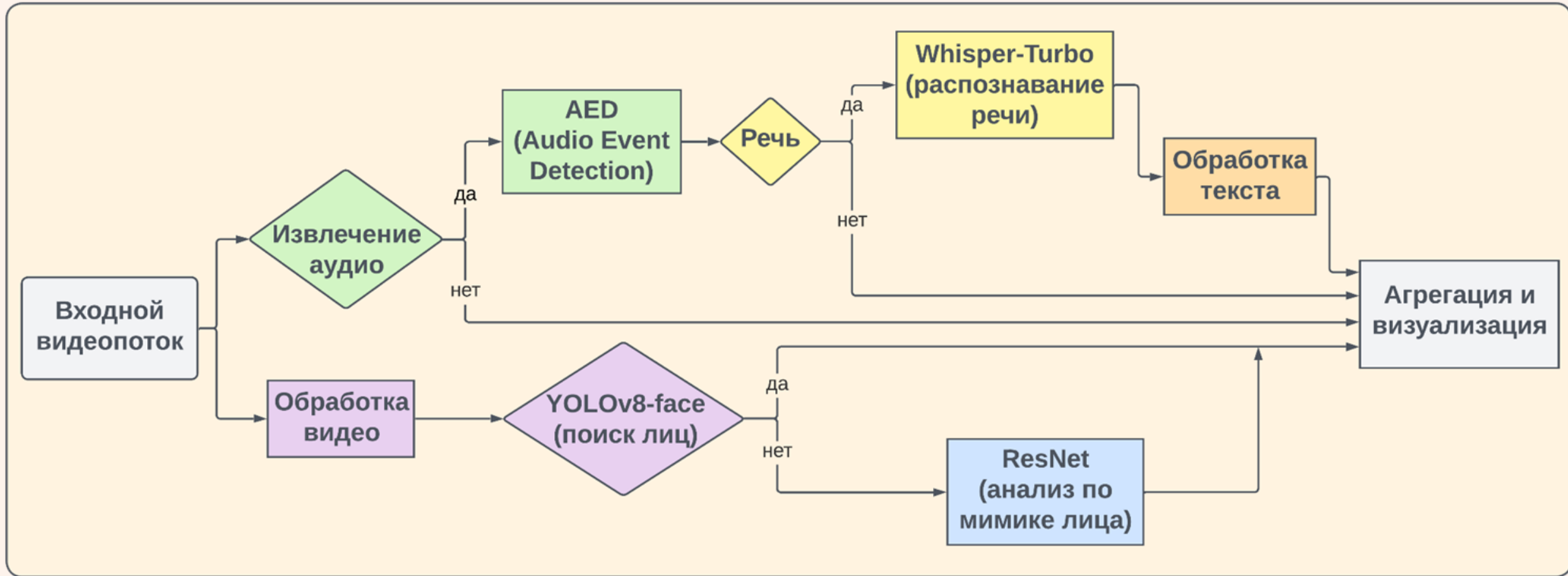


BERT/RoBERTa

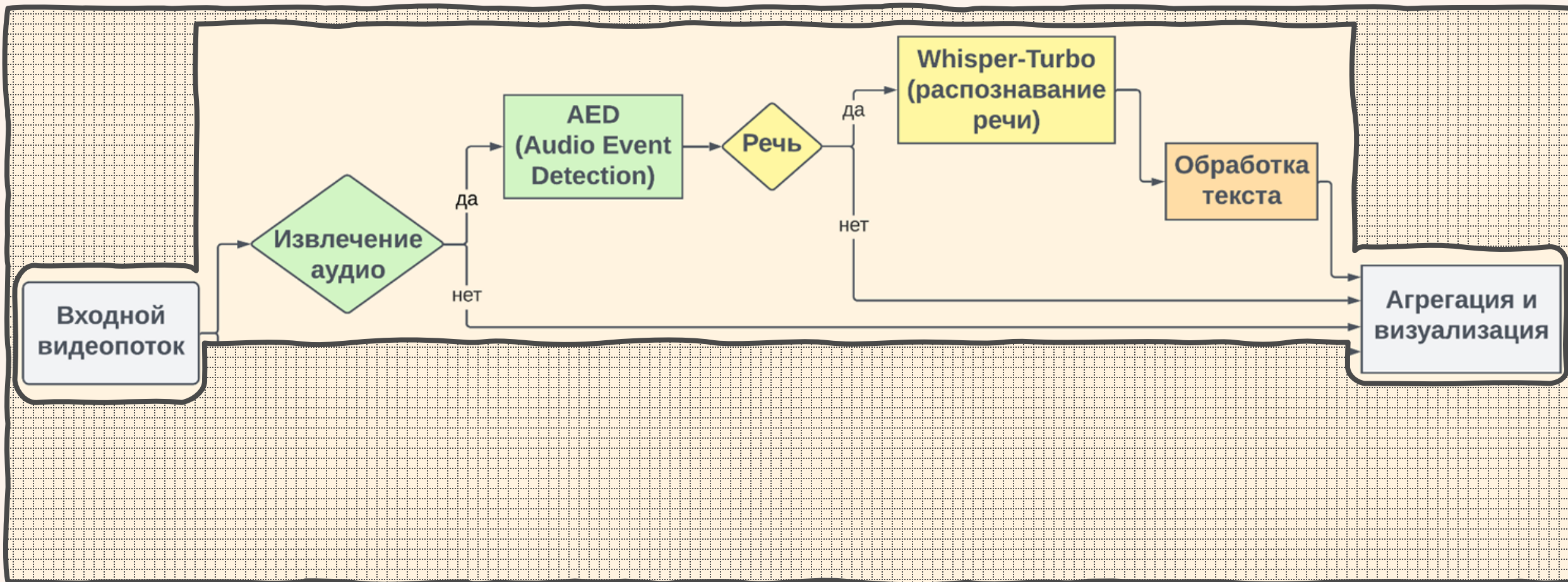
(текстовая модальность)

- ✓ Высокая точность, лёгкая масштабируемость
- ✗ Чуть более требователен по ресурсам

Общая структура системы



Анализ аудио и текста



Извлечение аудио

Видео



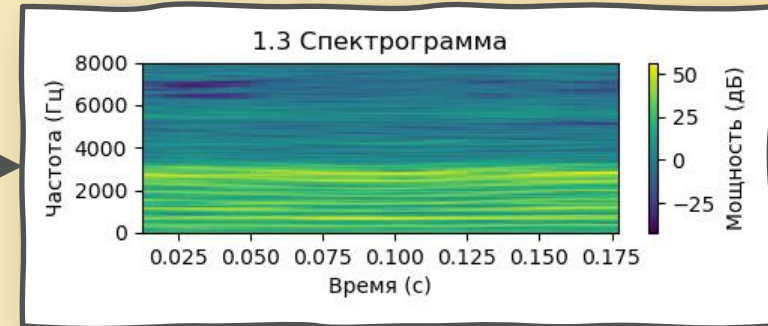
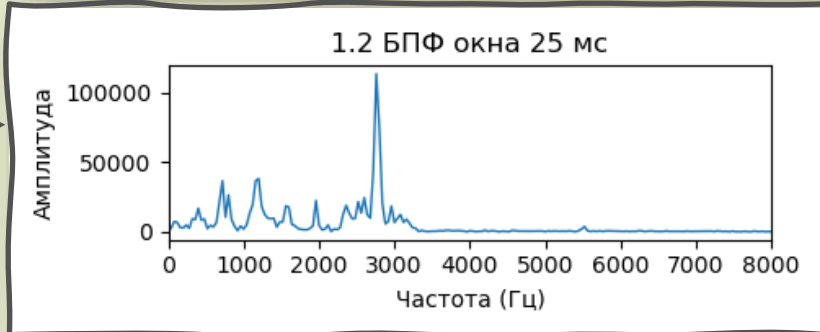
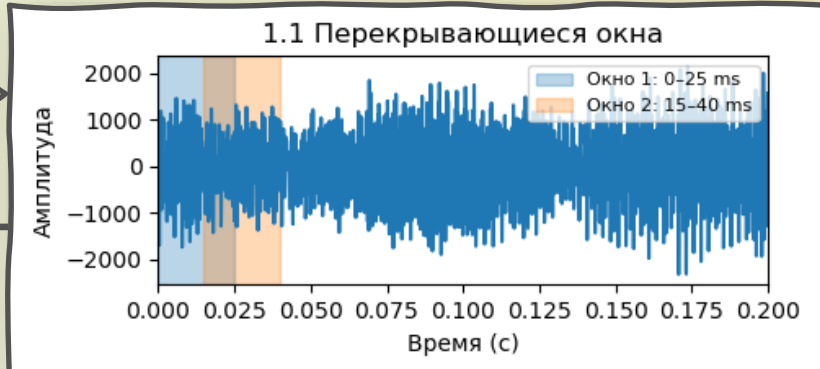
```
def extract_audio_from_video(self, video_path):
    video_name = Path(video_path).stem
    audio_path = self.audio_output_dir / f"{video_name}.wav"
    if not audio_path.exists():
        probe_cmd = [
            'ffprobe', '-v', 'error',
            '-select_streams', 'a',
            '-show_entries', 'stream=codec_type',
            '-of', 'default=noprint_wrappers=1:nokey=1',
            str(video_path)
        ]
        result = subprocess.run(probe_cmd, capture_output=True, text=True, check=True)
        if not result.stdout.strip():
            print("No audio stream found in video")
            return None
        command = [
            'ffmpeg', '-i', str(video_path),
            '-vn', '-acodec', 'pcm_s16le',
            '-ar', '16000', '-ac', '1',
            str(audio_path)
        ]
        subprocess.run(command, capture_output=True, check=True)
    print(f"Audio extracted to {audio_path}")
```

.wav файл

Пропуск

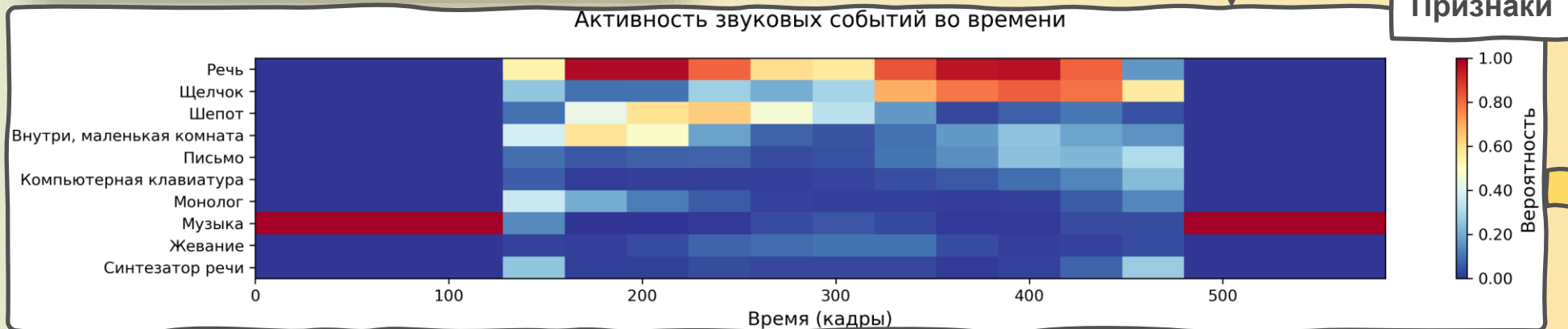
Анализ аудио (AED)

.wav файл



2DCNN

$$y[i, j] = \sum_{k=0}^{K-1} \sum_{l=0}^{L-1} x[i + k, j + l] \cdot w[k, l]$$



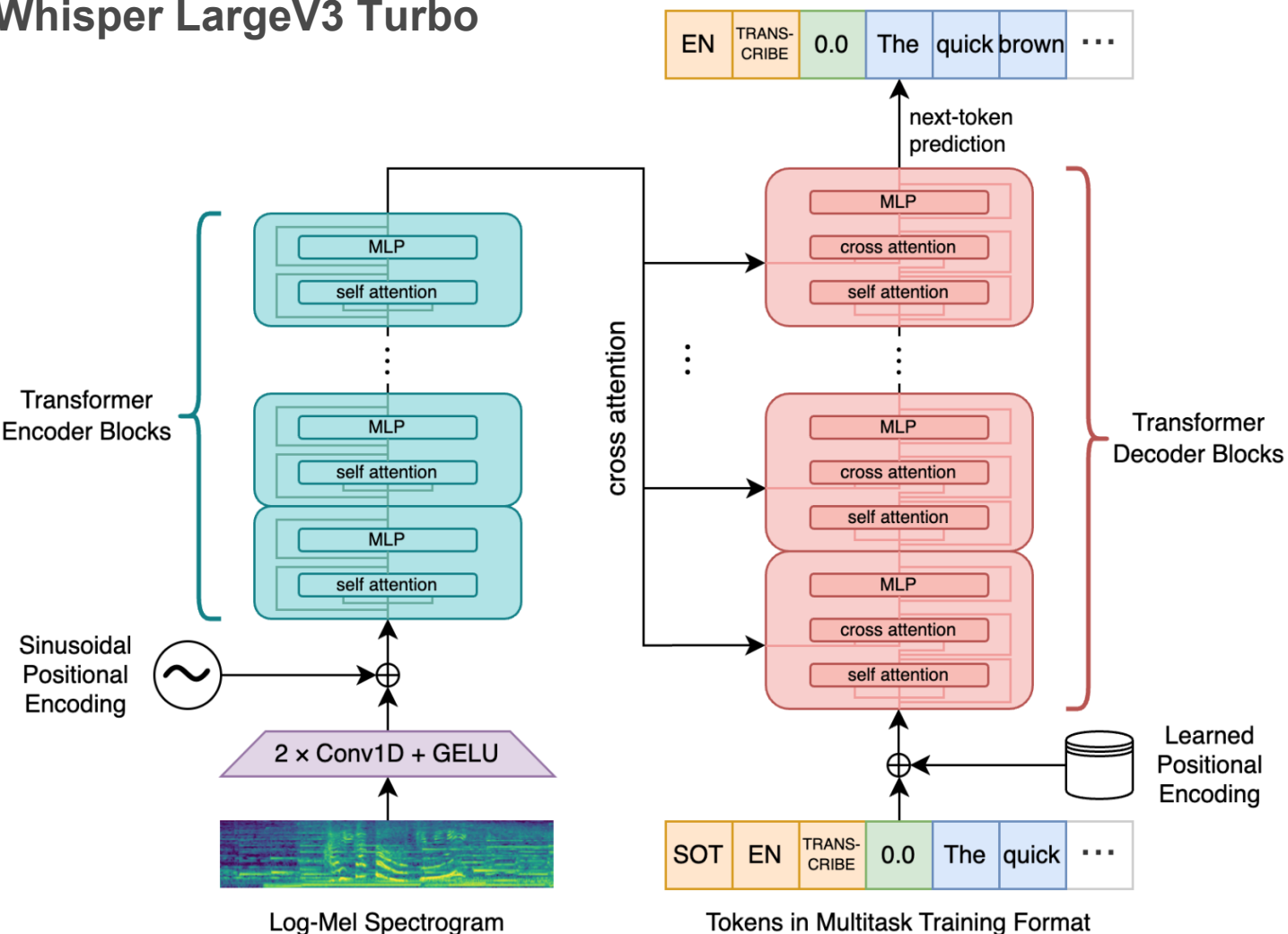
Пропуск

Детекция речи

Извлечение текста (ASR)

Детекция речи

Whisper LargeV3 Turbo



Тип обучения: seq2seq

99 языков, 680 000 часов разметки

test-clean: ~2.8 %
test-other: ~11.5 %
WER: ~ 5.6 %

Пример текста:

“Здравствуйте, еще вопросы есть? Сумма? 300? Это не серьезно. Не-не-не, так не пойдет. Вы нас не знаете и мы вас не знаем. Вести дурачков. Я на русалках больше заработаю. Пошли, пошли. Куром на смех. Подумаешь, 300. Стойте.”

Текст

Анализ текста (TED)

Текст

«Вы нас не знаете и мы вас не знаем»

_Вы_нас_не_знаете_и
_мы_вас_не_знаем_

Позиция	Субслово	ID
1	_Вы	1234
2	_нас	2345
...
9	_.	43

$[t_1, \dots, t_{12}] = [1234, 2345, \dots, 43]$

$$W_{pos} \in \mathbb{R}^{L_{max} \times D}$$

$$W_{tok} \in \mathbb{R}^{V \times D}$$

$$e_i = W_{tok}[t_i] + W_{pos}[i], i = 1, \dots, 12$$

$$H^{(0)} = \begin{bmatrix} e_1^T \\ e_2^T \\ \vdots \\ e_L^T \end{bmatrix}$$

$$H^{(N)} = \begin{bmatrix} e_1^T \\ e_2^T \\ \vdots \\ e_L^T \end{bmatrix}$$

N одинаковых блоков трансформера:

self-
attention

Add&norm

FFN

Add&norm

Матрица скрытых состояний

Сравнение эмбеддингов

Матрица скрытых состояний

$$H^{(N)} \in \mathbb{R}^{L \times D}$$

Mean-пулинг

$$v = \frac{1}{L} \sum_{i=1}^L H_i^{(N)}$$

L2-нормировка

$$\|v\|_2 = \sqrt{\sum_{j=1}^D v_j^2}$$

Эмбеддинг

$$\hat{v} = \frac{v}{\|v\|_2}, \|\hat{v}\|_2 = 1$$

```
"ted": {  
  "transcription": "Здравствуйте, еще вопросы есть? Сумма? 300. Это не серьезно. Не-не  
-не, так не пойдет. Вы нас не знаете и мы вас не знаем. Вести дурачков. Я на русалках  
больше заработаю. Пошли, пошли. Куром на змех. Подумаешь, 300. Стойте.",  
  "emotions": {  
    "отвращение": 0.26007319361513087,  
    "гнев": 0.20736210956674103,  
    "нейтральность": 0.18738932417066914,  
    "грусть": 0.17031209807198822,  
    "радость": 0.13441901791317423,  
    "страх": 0.04044425666229659,  
    "удивление": 0.0  
  },  
  "top_emotion": "отвращение"  
},
```

$$\cos(v, w) = \frac{v \cdot w}{\|v\|_2 \|w\|_2} = \frac{\sum_{i=1}^D v_i^2 w_i^2}{\sqrt{\sum_{i=1}^D v_i^2} \sqrt{\sum_{i=1}^D w_i^2}}$$

$$\varepsilon = \mathbb{R}^D$$

отвращение

гнев

нейтральность

грусть

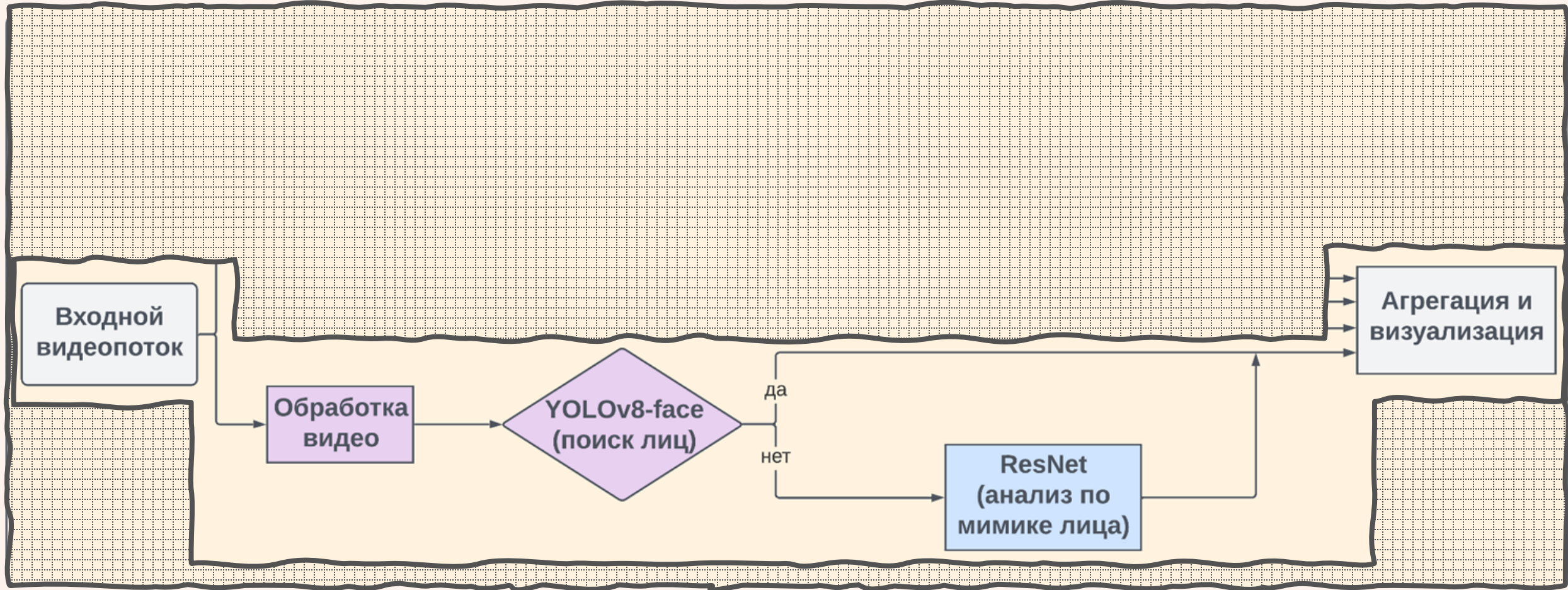
радость

страх

удивление

Эмоция

Визуальная модалность



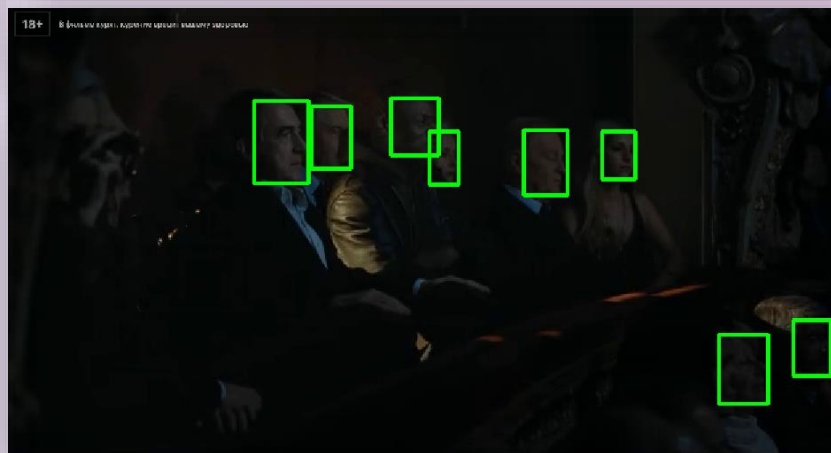
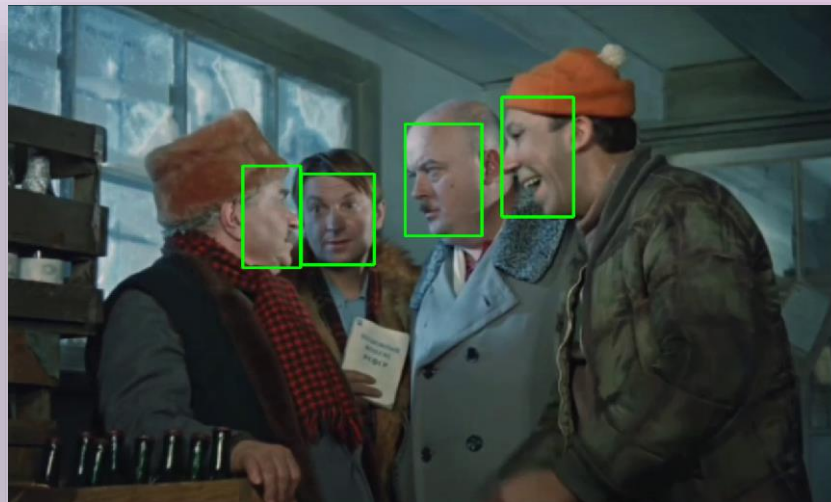
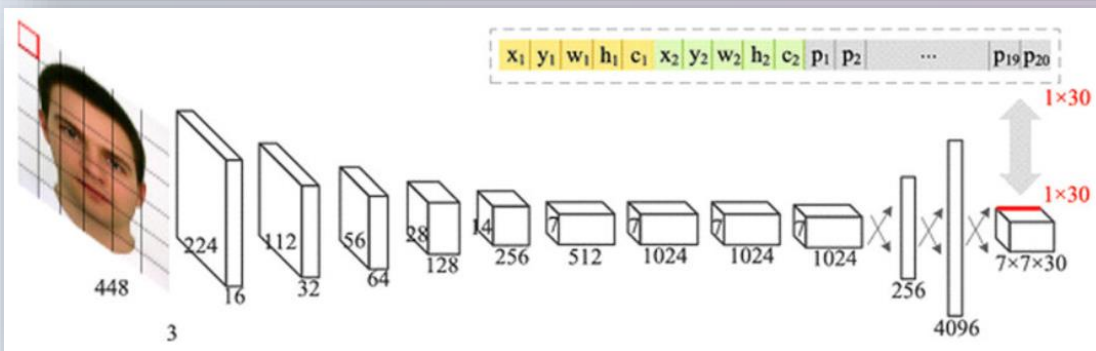
Поиск лиц

Видео



```
from ultralytics import YOLO
import cv2

yolo_model = YOLO("fed/weights/yolov8l-face.pt")
frame = ... # кадр из видео
results = yolo_model(frame)[0]
for box in results.boxes.xyxy.numpy():
    x1, y1, x2, y2 = map(int, box)
    cv2.rectangle(frame, (x1, y1), (x2, y2), (0, 255, 0), 2)
cv2.imshow("Face Detection", frame)
cv2.waitKey(0)
```



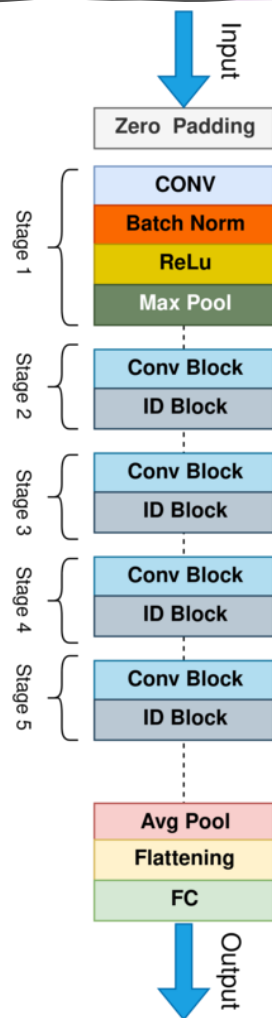
Координаты bbox

Пропуск

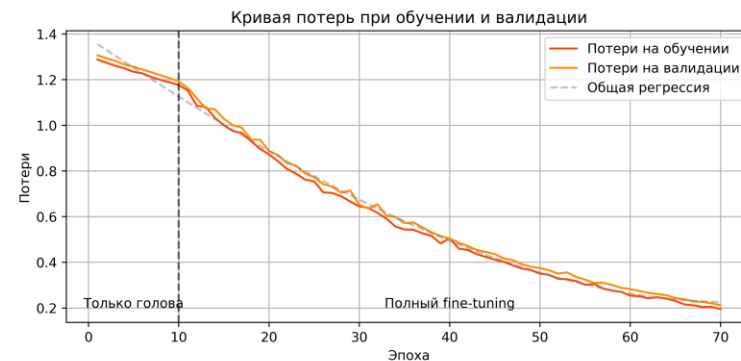
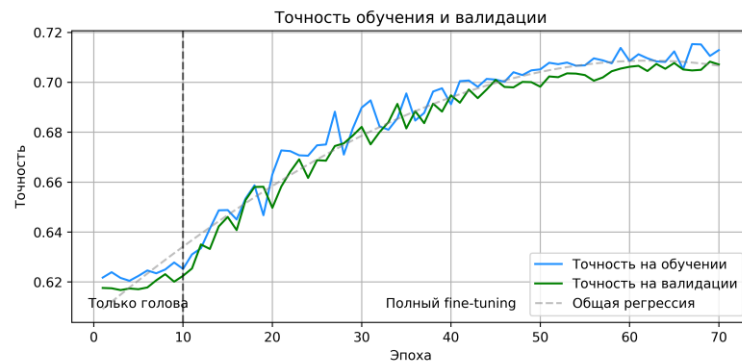
Анализ области bbox

Координаты bbox

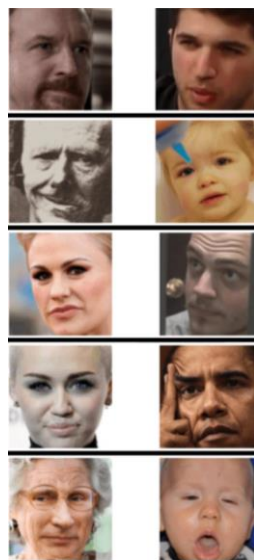
ResNet50



Графики обучения

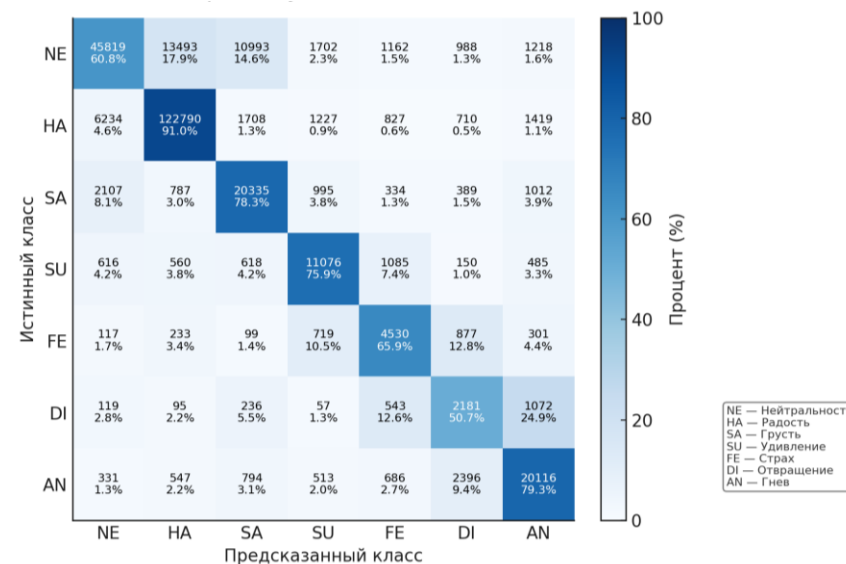


Датасет AffectNet



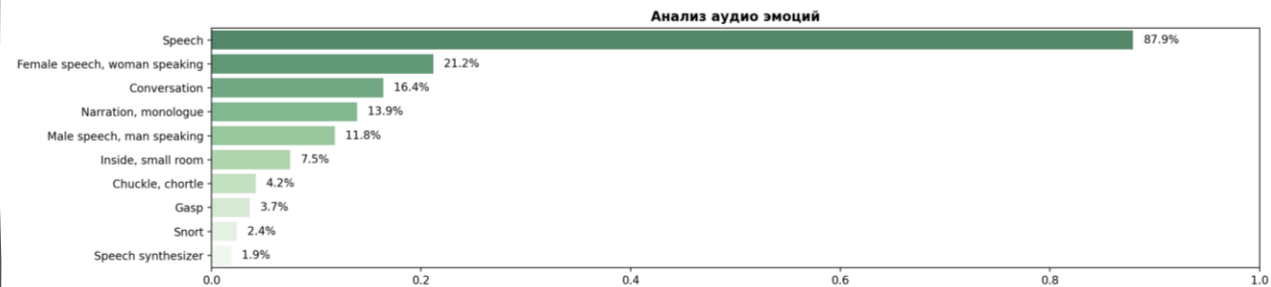
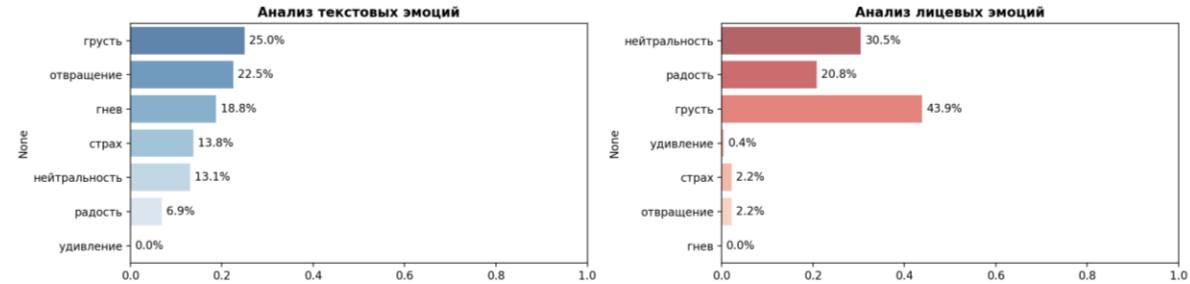
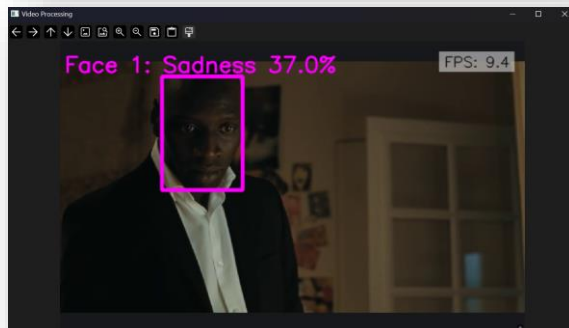
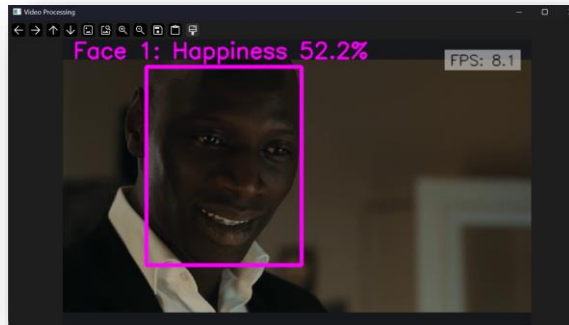
Neutral	75374
Happy	134915
Sad	25959
Surprise	14590
Fear	6878
Disgust	4303
Anger	25382
Contempt	4250
None	33588
Uncertain	12145
Non-Face	82915
Total	420299

Матрица путаницы эмоций

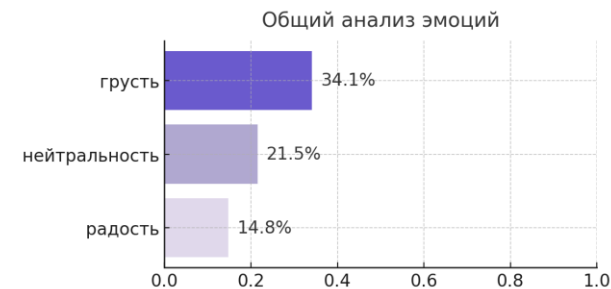
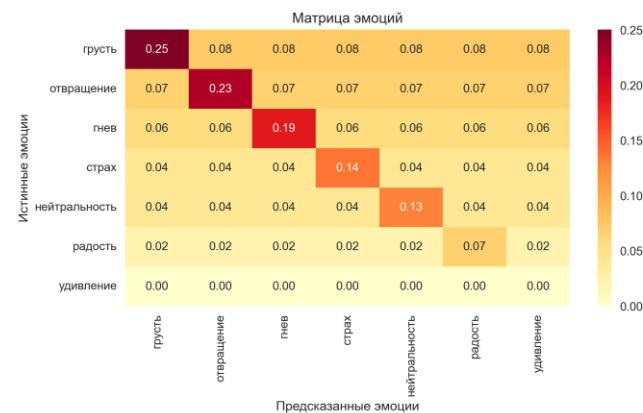


Эмоция

Пример работы 1



Доступные данные: Текстовые, Лицевые, Аудио



Распознанный текст:

“Ну что, звоним в скорую? Из-за чего это? Из-за Бастиана. И что с ним? Он меня бросил и ещё смеётся.”

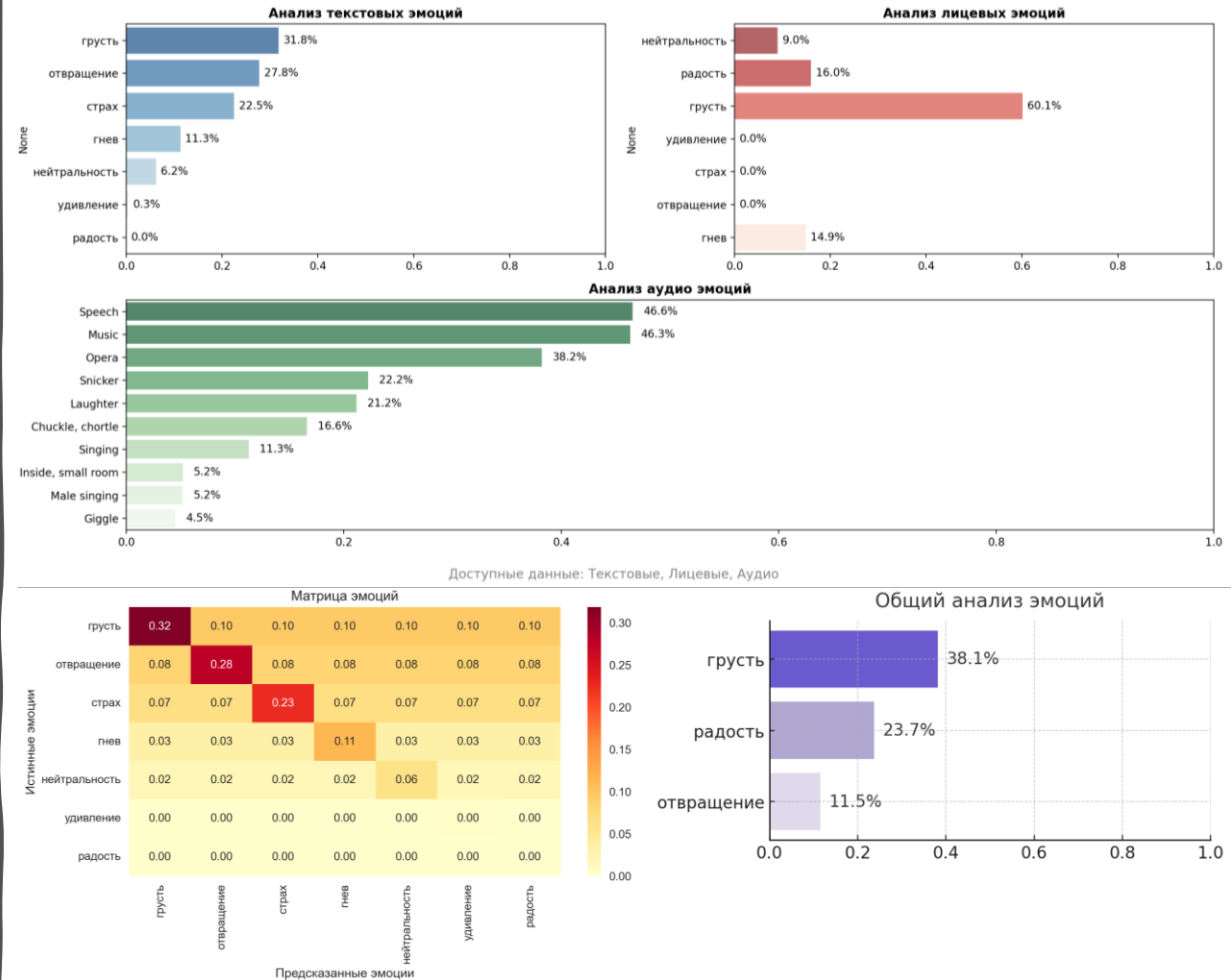
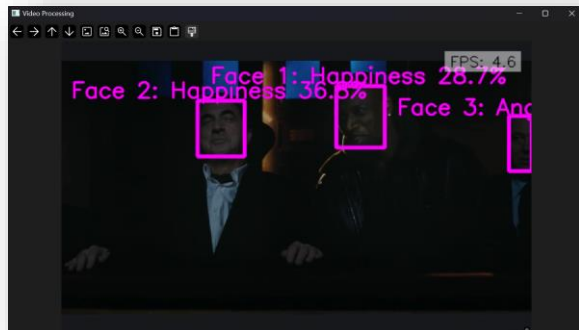
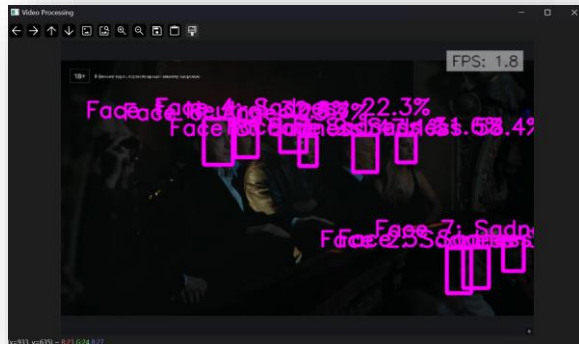
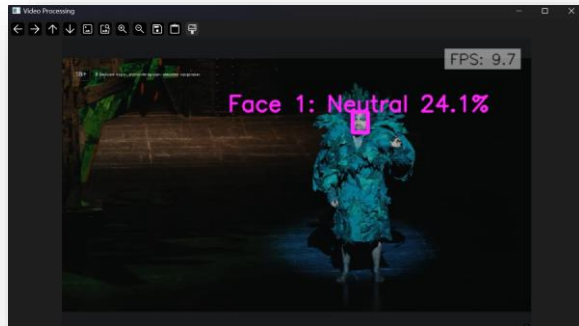
Аудио теги:

Речь, женская речь, разговор, мужская речь, хихиканье, шёпот

ТОП-3 эмоции:

Грусть: 34.1%
Нейтральность: 21.5%
Радость 14.8%

Пример работы 2



Распознанный текст:

“Какой ужас! Что это с ним? Чего шипишь?”

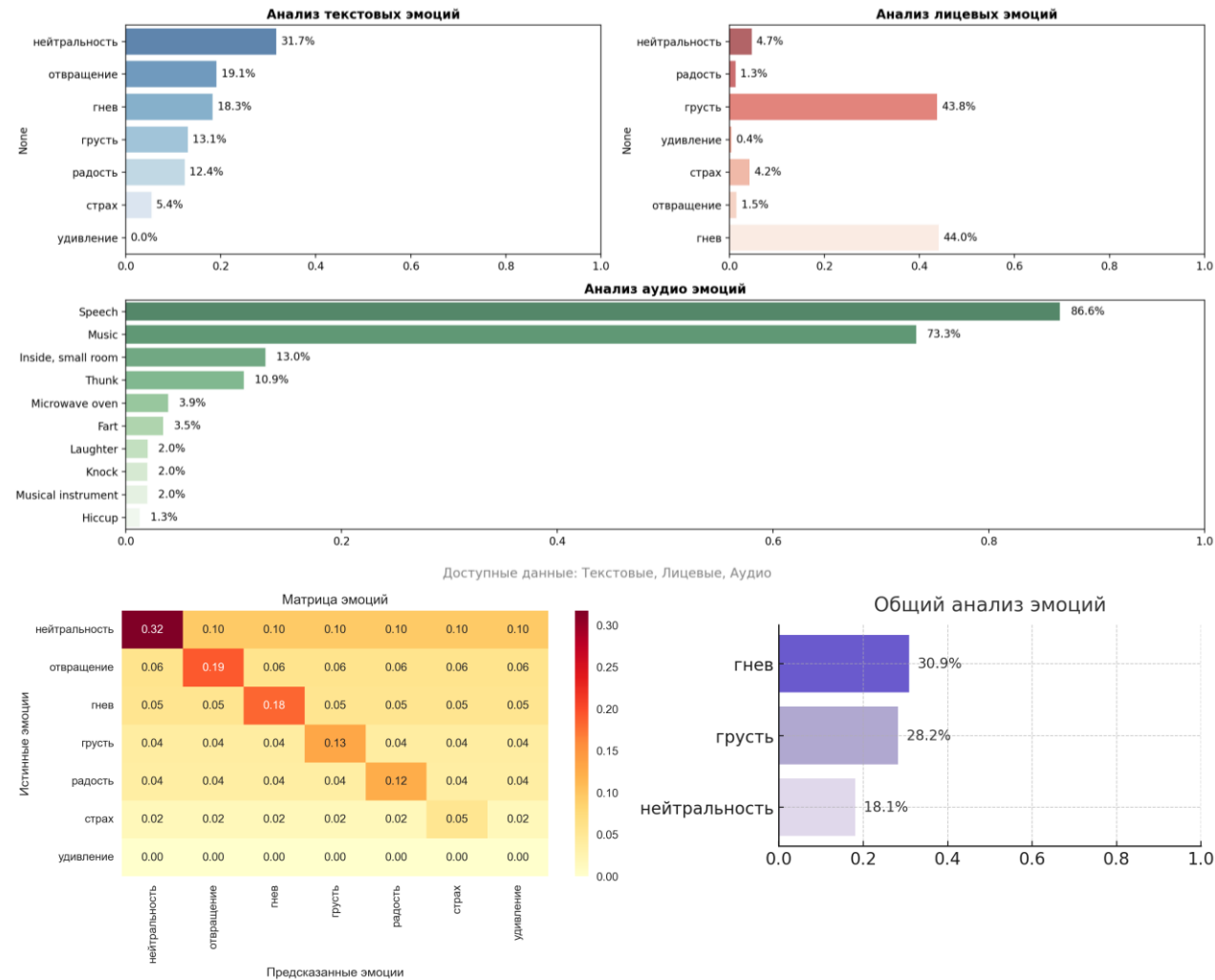
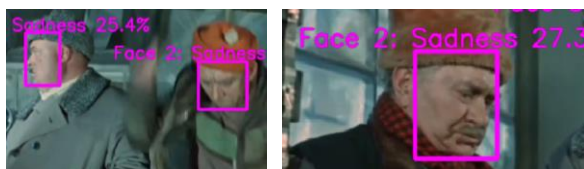
Аудио теги:

Опера, музыка, речь, смех, классическая музыка, пение, хихиканье, оркестр

ТОП-3 эмоции:

Грусть: 38.1%
Радость: 23.7%
Отвращение: 11.5%

Пример работы 3



Распознанный текст:

“Не нам, а вам. Нет, на этот раз именно вам. Да? Да. А что значит нейтрализовать? Статья 193, пункт 2. До трех лет. Не пойдет. Нет, не пойдет. Никаких... Сторож нежно усыпляется хлороформом и связывается без нанесения телесных повреждений.”

Аудио теги:

Речь, музыка, замкнутое пространство, стук, смех, икота

ТОП-3 эмоции:

Гнев: 30.9%
Грусть: 28.2%
Нейтральность 18.1%