Отчет по проекту:

Распознавание эмоций на видео

Выполнила:

Пирожкова Александра Дмитриевна

1. **Обоснование выбора датасета и метрик**

**1.1 Датасет**

Был выбран датасет FER-2013.

Он содержит около 30 000 изображений лиц в формате RGB с различными выражениями лица размером не более 48× 48, которые разделяются на 7 типов: 0= Сердитый, 1 = Отвращение, 2 = Страх, 3= Счастливый, 4= Грустный, 5= Удивление, 6= Нейтральный.

Для данной работы требовалось обучить модель распознавать не менее пяти эмоций, а в FER-2013 предоставляется семь классов. Объем данного датасета является приемлемым. Он идеально подходит для обучения моделей с ограниченными ресурсами. С FER-2013 было обучено много различных моделей, он один из часто появляющихся датасетов для распознавания эмоций.

Для улучшения качества модели было произведена аугментация тренировочной выборки датасета.

**1.2 Метрики**

Для оценивания качества модели были выбраны метрики: функция Кросс-энтропийной потери (Log Loss) и accuracy.

**Кросс-энтропийная потеря (Log Loss)** показывает, насколькопредсказания модели отклоняются от фактических значений. Во время обучения она дает алгоритму понят в каком направлении ему нужно двигаться. Значение Log Loss легко интерпретировать, чем оно меньше, тем лучше модель справляться со своей задачей. Кросс-энтропийная потеря используется в задачах классификации, а к этому и сводиться суть данного проекта.

**Аccuracy** - доля правильных ответов алгоритма. Одна из самых легко понимаемых метрик оценивания моделей. Чем выше её значение, тем лучше качество модели. Accuracy может применяться ко всем типам задач машинного обучения. Это одна из самых часто встречающихся метрик оценки качества, почти в каждом проекте, связанном с машинным обучением, Accuracy используется.

1. **Предложенное решение**

Для решения данной задачи было решено обучить модели EfficientNet, ResNet18. На сайте Paperswithcode на странице с оценкой метрик качества моделей на датасете FER-2013 среди лучших моделей обе входили в список несколько раз.

**EfficientNet** — это семейство свёрточных нейронных сетей (CNN), целью которых является достижение высокой производительности при меньшем количестве вычислительных ресурсов по сравнению с предыдущими архитектурами. Архитектура EfficientNet представлена набором готовых к использованию моделей. Выбор зависит от требуемой точности, доступных ресурсов для обучения и разрешения входных изображений. Модели маркируются от B0 (самая простая) до B7 (самая «сильная»).

В данной работе обучалась EfficientNetB0.

Основнаяидея **ResNet18** — использование остаточных (skip) соединений, которые позволяют нейросети быть очень глубокой и избегать проблемы исчезающего градиента, когда сигналы, передаваемые между слоями, затухают до нуля. В классическойархитектуре ResNet18 насчитывается 18 слоёв, организованных в последовательность блоков. Благодаря своим компактным размерам (всего 11,7 миллионов параметров) она отлично подходит для мобильных приложений и систем с ограниченными вычислительными ресурсами.

В качестве эксперимента и для сравнения было решено так же обучить модель Yolo11n на части датасета.

**YOLO11n** — это облегчённый вариант модели YOLO11 для небольших и лёгких задач. Модель поддерживает задачи обнаружения объектов с высокой точностью и скоростью.

Для детекции лиц было решено использовать haarcascade\_frontalface\_default.xmlс OpenCV CascadeClassifier. Кадры видео обрабатывались с помощью OpenCV, а для сохранения видео с детекцией использовался moviepy. Для ускорения инференса моделей, они переводились в формат ONNX.

Веб-интерфейс для работы с обученными моделями писался на Gradio и был задеплойен на Hugging Face.

1. **Архитектура и оценка сложности моделей**

**3.1 Архитектура EfficientNet**

EfficientNet-B0 состоит из:

1. **Stem**
   1. Начальный слой со стандартной сверткой, за которой следует пакетная нормализация и активация ReLU6.
   2. Свертка с 32 фильтрами, размер ядра 3x3, шаг
2. **Body**
   1. Состоит из серии блоков MBConv с различными конфигурациями.
   2. Каждый блок включает в себя разделяемые по глубине витки и слои сжатия и возбуждения.
3. **Head**
   1. Включает в себя последний сверточный блок, за которым следует глобальный средний уровень объединения.
   2. Полностью подключенный слой с функцией активации softmax для классификации.

**3.2 Оценка сложности EfficientNet-B0**

* Размер параметров (МБ): 16.07
* Предполагаемый общий размер (МБ): 1751.75
* Общее количество параметров: 4,016,515

**3.3 Архитектура ResNet18**

* Первый сверточный слой 7×7
* Макс-пулинг 3×3
* Четыре блока residual layers
* Глобальный average pooling
* Полносвязный слой

**3.2 Оценка сложности ResNet18**

* Размер параметров (МБ): 44,72
* Предполагаемый общий размер (МБ): 690,19
* Всего параметров: 11 180 103

**3.4 Архитектура Yolo11n**

1. Основная свёрточная сеть. Включает несколько ключевых блоков, например:
   1. Conv
   2. C3K2
2. Промежуточный блок агрегации признаков. Отвечает за дальнейшую агрегацию и обработку признаков, полученных на стадии основной свёрточной сети.
3. «Голова». Формирует предсказания, используя карты признаков с трёх различных уровней (например, P3, P4 и P5).

**3.5 Оценка сложности Yolo11n**

* 319 слоев
* 2 591 205 параметров
* 6,4 GFlops

1. **Детали обучения**

**Библиотеки:**

* Torch, torchvision
* Ultralytics

**Параметры обучения:**

* Функция потерь - CrossEntropyLoss
* Оптимизатор - Стохастический градиентный спуск (SGD)
* Эпохи - 50 для EfficientNet и ResNet18, 100 для Yolo11n
* Скорость обучения lr=0.01
* Сколько эпох нужно ждать улучшения показателей валидации, прежде чем остановить обучение - 10

**Датасет:**

* Train : 22961
* Test : 7178
* Validation: 5748

1. **Выводы о пригодности данного подхода к решению задачи**

* Лучшие метрики на тестовой выборке показала EfficientNet (Test loss: 0.915, Test accuracy: 0.669). Подобные метрики достигли многие модели, обученные на данном датасете.
* С помощью haarcascade\_frontalface\_default.xml можно эффективно детектировать лица на кадрах видео
* ONNXRuntime значительно уменьшил время инференса моделей, по сравнению с моделями в формате pytorch
* Обучение EfficientNet заняло около одного дня на довольно слабых ресурсах. Это говорит о компактности модели с сохранением приемлемого качества.

Модель EfficientNet хорошо справилась с задачей распознавания эмоций на видео. Другие модели, обученные в данной работе, тоже показали приемлемые результаты (ResNet18 - Test loss: 1.023, Test accuracy: 0.620 ; Yolo11n – MAP50-95 : 0.651, P: 0.584).