Адаптация нейросетевой модели распознавания эмоций лиц на основе видеоданных конечного пользователя

Е.Н. Чураев
Национальный исследовательский университет
Высшая школа экономики
Нижний Новгород, Россия
echuraev@hse.ru

А.В. Савченко
Национальный исследовательский университет
Высшая школа экономики
Нижний Новгород, Россия
avsavchenko@hse.ru

***Аннотация*—Исследуются способы улучшения качества распознавания эмоций по видео при наличии набора данных с эмоциями конечных пользователей. Используя идею дикторозависимого распознавания речи, предложен новый подход, в котором на первом этапе с использованием набора видео других лиц обучается универсальная нейросетевая модель классификации эмоций, а на втором этапе происходит ее адаптация (дообучение) на основе данных конкретного пользователя. Для систем, нацеленных на работу с большим количеством пользователей, в процессе принятия решения вначале выполняется идентификация лица, после чего эмоции классифицируются с помощью модели, адаптированной под распознанного пользователя. Для набора данных RAVDESS показано, что такой подход позволяет более чем на 20% повысить точность распознавания эмоций.**

Ключевые слова— Распознавание эмоций, обработка видео, дообучение нейронных сетей.

1. Введение

Задача распознавания эмоций заключается в том, что для поступающей на вход последовательности видеокадров $X(t)$, $t=1,2,\ldots,T$, где T – число кадров, требуется поставить в соответствие один из $C>1$ эмоциональных классов (радость, злость и т.п.). В настоящее время технологии распознавания эмоций по видеоизображению лица востребованы во многих областях, например, определение эмоционального состояния водителя для оценки и снижения уровня стресса во время вождения; анализ поведения группы людей в системах видеонаблюдения для предотвращения возможных конфликтных ситуаций; в человеко-машинных интерфейсах для повышения качества понимания состояния пользователя, например, определение реакции покупателя на товар или рекламную компанию. К сожалению, существующие универсальные модели распознают эмоции с точностью 50-70%, что не всегда достаточно для практических приложений. В настоящей работе исследуется возможность повысить точность, если требуется распознавать эмоции лишь ограниченного набора пользователей, при этом для каждого имеется возможность собрать небольшое множество видеоданных с различными эмоциями.

2. Предлагаемый подход

Используя идею дикторозависимого распознавания речи, мы предложили новый метод для увеличения точности определения эмоций на видео. На Рис. 1 представлена схема предлагаемого подхода. Алгоритм распознавания эмоций основан на нейросетевом механизма внимания [1] для дескрипторов лиц, извлеченных из видеокадров с помощью специальным образом обученных моделей MobileNet и EfficientNet [2]. На этапе обучения базовая дикторонезависимая модель дообучается для каждого пользователя с использованием видеоданных только этого пользователя. В процессе обучения пользователь идентифицируется с помощью известных алгоритмов распознавания лиц, после чего дикторозависимая модель, соответствующая распознанному пользователю, используется для распознавания эмоций. В случае, если пользователь не был найден в базе, то применяется дикторонезависимая модель для определения эмоций на видео.

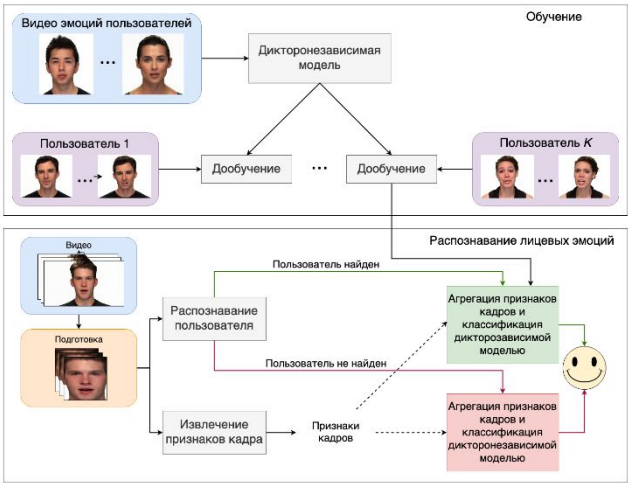

Рис. 1. Предлагаемый алгоритм

В экспериментах мы использовали набор данных RAVDESS, который включает в себя небольшие видео фрагменты, где 24 актёра (12 мужчин и 12 женщин) произносят одну и ту же фразу с разными эмоциями. Для классификации эмоций использовались модель логистической регрессии и три различные реализации механизма внимания. Полученные результаты приведены в Таблице I. Из этих результатов можно сделать вывод, что предложенный подход может применяться для различных архитектур базовой дикторонезависимой модели и позволяет существенно (более чем на 20%) повысить точность распознавания эмоций на видео.

Таблица I. ТОЧНОСТЬ (%) КЛАССИФИКАЦИИ ЭМОЦИЙ ДЛЯ НАБОРА ДАННЫХ RAVDESS

Классификатор	**Дикторонезависимая модель**	**Предложенная адаптация**
Logistic regression	72,62	96,61
Single attention	74,17	99,96
Relation attention	73,69	99,76
Self-attention	76,01	99,96

3. ЗАКЛЮЧЕНИЕ

В рамках данной работы на нескольких моделях было показано, что предложенный подход позволяет значительно повысить точность определения эмоций пользователя на видео. Благодаря тому, что для извлечения лицевых признаков, использовались эффективные модели MobileNet/EfficientNet, такой подход может быть реализован на различных энергоэффективных устройствах (в том числе и на смартфонах) для распознавания эмоций в режиме реального времени. В дальнейшем планируется исследовать возможность повышения точности за счет использования дополнительных модальностей с адаптацией аудиовизуальных нейросетевых моделей.

БЛАГОДАРНОСТИ

Работа выполнено за счет гранта Российского научного фонда (проект № 20-71-10010).

ЛИТЕРАТУРА

[8] Demochkina, P. Neural network model for video-based facial expression recognition in-the-wild on mobile devices / P. Demochkina, A.V. Savchenko // IEEE International Conference on Information Technology and Nanotechnology (ITNT). – 2021. – P. 1-5.

[9] Savchenko, A.V. Facial expression and attributes recognition based on multi-task learning of lightweight neural networks / A.V. Savchenko // IEEE International Symposium on Intelligent Systems and Informatics (SISY). – 2021. – P. 119-124.