УДК 004.932:004.8.032.26

МУЛЬТИЗАДАЧНАЯ НЕЙРОСЕТЕВАЯ АРХИТЕКТУРА ДЛЯ НЕПРЕРЫВНОЙ ОЦЕНКИ ВАЛЕНТНОСТИ И ВОЗБУЖДЕНИЯ ПО ВИДЕО ЛИЦА

Якушев А.Д. (аспирант, ИТМО)
Научный руководитель – кандидат технических наук, доцент Кашевник А.М (ИТМО)

Введение.

Распознавание эмоций по выражению лица в реальных условиях остаётся сложной задачей из-за изменений освещения, ракурса и индивидуальных особенностей мимики. Традиционный подход, основанный на классификации дискретных базовых эмоций (радость, грусть и т.д.), является слишком грубым и не позволяет отслеживать плавные изменения аффективного состояния человека. Более перспективным представляется использование континуальной модели аффекта Дж. Рассела [1], в которой эмоции описываются непрерывными значениями валентности (степень удовольствия) и возбуждения (уровень активации) в диапазоне от –1 до +1. Целью данной работы является разработка легковесной и точной мультизадачной нейросетевой архитектуры для одновременного предсказания значений валентности, возбуждения и дискретных эмоций, а также её апробация на длительном видеопотоке учебной деятельности.

Основная часть.

Для обучения модели непрерывному распознаванию эмоций использовался датасет AffectNet [2], содержащий ручные аннотации валентности (V) и возбуждения (A) для более чем 300 000 изображений. Как показал предварительный анализ, в стандартной валидационной выборке AffectNet присутствует системное смещение распределения значений возбуждения, возникшее из-за изначальной стратификации по дискретным классам эмоций. Для устранения этого смещения мы разработали методику балансировки выборки по двумерному пространству V-A. Использовалась регулярная сетка 10×10 бинов. Это позволило создать объективный валидационный набор и повысить достоверность оценки качества моделей.

Предложенная архитектура EffiAtt-MTL-VA построена на основе EfficientNet-B3 [3], выбранной благодаря оптимальному балансу между точностью извлечения признаков и вычислительной сложностью. Для усиления способности сети фокусироваться на эмоционально значимых областях лица в ключевые блоки бэкбона интегрированы модули Convolutional Block Attention Module (CBAM). Мультизадачная структура включает три независимые «головы»: две регрессионные для предсказания непрерывных V и A (с функцией потерь на основе коэффициента согласованной корреляции CCC) и одну классификационную для распознавания восьми базовых эмоций. Стоит отметить, что классификационная ветвь используется исключительно на этапе обучения в качестве регуляризатора, что позволяет улучшить качество регрессии без увеличения вычислительных затрат при запуске. Благодаря отбрасыванию классификационной головы при развёртывании финальный размер модели составляет 10,3 млн параметров, что обеспечивает высокую скорость работы.

Экспериментальная оценка на сбалансированной валидационной выборке показала высокие результаты: CCC для валентности достиг 0,880, для возбуждения – 0,740, что превосходит показатели многих современных тяжёлых архитектур при существенно меньшем числе параметров. Благодаря применению оптимизированного движка ONNX Runtime скорость обработки одного изображения на GPU NVIDIA RTX 2080 Ti составила 2,53 мс (395 кадров/с), на CPU Intel Core i9 с фреймворком OpenVINO – 9,78 мс (102 кадра/с).

В частности, для проверки работоспособности в реальном сценарии разработанная модель была интегрирована в конвейер анализа длительного видеопотока (2 часа записи самостоятельной учебной работы студента). Предварительная детекция лица выполнялась детектором YOLOv8, после чего для каждого кадра вычислялись значения валентности и возбуждения. Построенные временные ряды наглядно отражают динамику эмоционального состояния испытуемого, позволяя выделять фазы концентрации, когнитивного истощения и восстановления.

Выводы.

Разработанная мультизадачная архитектура EffiAtt-MTL-VA успешно решает задачу непрерывной оценки валентности и возбуждения по видеоряду лица. Сочетание легковесного бэкбона, модулей внимания и вспомогательной классификационной головы обеспечивает высокую точность при малом количестве параметров (10,3 млн) и высокой скорости обработки. Эксперимент с двухчасовой записью учебной деятельности подтвердил практическую применимость модели для автоматического мониторинга аффективных состояний в реальном времени. Полученные результаты открывают перспективы интеграции системы в платформы дистанционного образования для адаптивного управления учебным процессом на основе эмоционального состояния обучающихся.

Литература

1. Russell J. A. A circumplex model of affect // Journal of Personality and Social Psychology. 1980. Vol. 39, № 6. P. 1161–1178.
2. Mollahosseini A., Hasani B., Mahoor M. H. AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild // IEEE Transactions on Affective Computing. 2019. Vol. 10, № 1. P. 18–31.
3. Tan M., Le Q. V. Rethinking Model Scaling for Convolutional Neural Networks // Proceedings of the 36th International Conference on Machine Learning (ICML). 2019. P. 6105–6114.

Примечание: при подготовке данного тезиса доклада инструменты искусственного интеллекта не применялись.