УДК 004.8

Волков Артём Андреевич,

бакалавр ФГБОУИ ВО МИРЭА, Россия,

г.Москва

ПРЕОДОЛЕНИЕ ДЕФИЦИТА ДАННЫХ: АУГМЕНТАЦИЯ И СИНТЕТИЧЕСКИЕ ДАННЫЕ В РАСПОЗНАВАНИИ ЭМОЦИЙ ЧЕЛОВЕКА

Аннотация. В статье рассматриваются подходы к преодолению проблемы ограниченности данных в задачах распознавания эмоций человека. Отмечается, что недостаточный объем и дисбаланс доступных наборов данных негативно сказываются на качестве моделей эмоционального анализа. Обзорно представлены методы аугментации данных – искусственного увеличения обучающей выборки за счет различных трансформаций исходных примеров, – а также генерации синтетических данных с использованием современных глубоких нейросетевых моделей. Особое внимание уделено генеративным подходам (GAN, диффузионные модели, вариационные автокодировщики и др.), позволяющим создавать новые образцы изображений лиц или других модальностей с требуемыми эмоциональными метками. Приводятся примеры из современных исследований, где дополнение тренировочных данных синтетическими образцами позволило существенно повысить точность распознавания эмоций. Обсуждаются преимущества и ограничения таких подходов, в том числе риск внесения артефактов и несоответствий распределения, а также методы обеспечения качества генерируемых данных.

Ключевые слова: распознавание эмоций, дефицит данных, аугментация данных, синтетические данные, глубокое обучение, генеративные модели, балансировка классов, GAN, диффузионные модели, устойчивость моделей

Распознавание эмоций человека – ключевая задача в области аффективных вычислений и человеческо-компьютерного взаимодействия. Такие системы стремятся автоматически определять эмоциональное состояние по различным сигналам: изображениям лица, речевым сигналам, жестам или физиологическим показателям. Несмотря на успехи глубокого обучения в смежных областях, прогресс в системах распознавания эмоций значительно ограничивается недостатком больших и качественных наборов данных. Существующие публичные датасеты эмоций, как правило, сравнительно невелики по размеру. Например, широко используемые наборы FER-2013 и RAF-DB содержат порядка нескольких десятков тысяч изображений лиц, CK+ – менее тысячи последовательностей изображений, JAFFE – всего 213 статических снимков, а аудиоколлекция IEMOCAP включает около 12 часов размеченной речи. Для сравнения, типичные наборы для задач компьютерного зрения (например, ImageNet) насчитывают миллионы образцов. Кроме того, в наборах данных эмоций часто наблюдается дисбаланс классов – одни эмоции (например, «нейтральное» или «радость») встречаются гораздо чаще других («отвращение», «страх» и др.), что затрудняет обучение моделей. Ограниченный масштаб и несбалансированность данных приводят к тому, что модели склонны к переобучению и хуже обобщаются на новые случаи. Таким образом, актуальной задачей является разработка методов, позволяющих эффективно расширить доступный объем данных и разнообразие примеров без необходимости ручной разметки новых больших выборок.

Таблица 1. Основные наборы данных для распознавания эмоций

Набор данных	Модальность	Объем данных	Классы эмоций

CK+	Изображения (видео)	953 последовательностей (≈0,5–1 тыс. кадров)	7 базовых эмоций
JAFFE	Изображения	213 изображений от 10 испытуемых	7 базовых эмоций
FER-2013	Изображения	~35 000 изображений (разрешение 48×48)	7 базовых эмоций
RAF-DB	Изображения	29 672 изображений лиц в реальной среде	7 базовых эмоций (+12 составных)
AffectNet	Изображения	~0,4 млн изображений из интернета	8 эмоций + валентность и возбуждение
IEMOCAP	Речь (аудио + видео)	~12 часов диалогов (≈10 000 фрагментов речи)	5 эмоций

Аугментация данных представляет собой совокупность методов, позволяющих расширить обучающую выборку за счет создания модифицированных версий имеющихся образцов. В контексте распознавания эмоций это может включать как простые преобразования исходных данных, так и более сложные генеративные подходы. К классическим методам аугментации изображений относятся геометрические трансформации (отражение, поворот, масштабирование, сдвиг), изменения яркости и контрастности, добавление шумов, обрезка (cropping) и случайное стирание фрагментов изображения. Такие операции создают новые вариации лицевых изображений, имитируя различные ракурсы, освещение и помехи съёмки, что повышает устойчивость моделей к этим факторам. Аналогично, для речевых данных применяются добавление фонового шума, изменение скорости и высоты звучания,

временное растяжение и сжатие аудиосигнала и другие преобразования аудиоспектра. Эти методы позволяют симулировать различные условия записи речи (акустические шумы, вариации дикции и темпа), расширяя разнообразие акустических признаков.

Применение аугментации данных доказало свою эффективность в снижении риска переобучения и повышении точности моделей при ограниченном исходном датасете. Аугментация фактически выступает в роли регуляризатора, увеличивая объем данных и вводя дополнительные вариации, что препятствует избыточной подгонке сети под обучающую выборку. В ряде работ показано, что грамотное применение аугментации улучшает результирующие метрики распознавания эмоций. Например, в исследовании по распознаванию эмоций в речи добавление случайного шума к исходным сигналам и сдвиг спектрограммы приводят к повышению точности классификации эмоций по сравнению с обучением на сырых данных без подобных преобразований. Другое сравнение методов для аудиомодальности выявило, что изменение скорости (time stretching) речевого сигнала оказалось одним из наиболее эффективных приемов, давая наибольший прирост качества на разных моделях. В задачах анализа текста (эмоциональное окрашивание высказываний) используются методы семантической аугментации – парафразирование с сохранением смысла, подстановка синонимов, генерация альтернативных формулировок при помощи языковых моделей. Такие приемы также способствуют улучшению распознавания, особенно для редких классов эмоций.

Следует отметить, что эффективность аугментации зависит от качества исходных данных и от того, насколько новые синтетические примеры отражают реальное распределение данных. Чрезмерные или неуместные трансформации могут привести к ухудшению модели, если внесут искажения, не характерные для реальных данных. Поэтому в практике аугментации важно подбирать виды и параметры преобразований,

релевантные рассматриваемой задаче. Тем не менее, правильно настроенная аугментация почти не требует дополнительных затрат на разметку и относительно проста в применении, что делает ее одним из первых инструментов при дефиците данных.

Более продвинутый подход к увеличению объема обучающей выборки – генерация синтетических данных с помощью моделей глубокого обучения. В отличие от классической аугментации, где создаются вариации исходных образцов, генеративные методы способны порождать совершенно новые примеры, не скопированные напрямую из обучающего набора. За последние годы появились разнообразные генеративные архитектуры, позволяющие моделировать сложные распределения данных (изображений, аудио, текста), – такие как Generative Adversarial Networks (GAN), вариационные автокодировщики (VAE), диффузионные модели и др. В контексте распознавания эмоций это открывает возможность автоматически синтезировать изображения лиц с заданными выражениями, аудиофрагменты речи с нужной эмоциональной интонацией или текстовые высказывания с определенной эмоциональной окраской.

Одним из первых значимых достижений в этой области стало применение GAN для расширения датасетов. GAN-модели, состоящие из генерирующей и дискриминирующей моделей, могут научиться генерировать фотореалистичные изображения лиц, включая разнообразные мимические выражения.

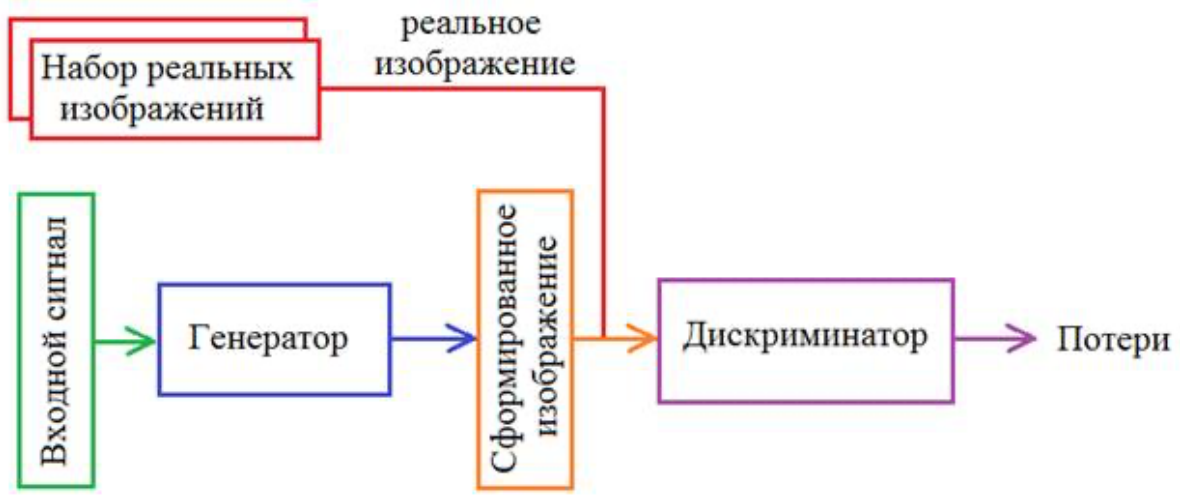

Рисунок 1. Схема работы GAN при формировании синтетических изображений

Генеративные модели позволяют, например, существенно обогатить обучающие выборки редких эмоций – таких как «удивление» или «страх» – синтезируя дополнительные образцы этих категорий, что выравнивает распределение классов в датасете. Кроме того, генеративные подходы применяются в смежных задачах медицинской диагностики, показывая рост качества при дополнении выборки искусственно созданными примерами.

Современным этапом развития генеративных методов стали диффузионные модели, демонстрирующие высокое качество синтеза изображений. В контексте распознавания эмоций недавние исследования показали впечатляющие результаты применения диффузионных моделей для генерации эмоциональных лицевых изображений. Обучение нейросети ResEmoteNet на расширенных таким образом данных привело к существенному повышению точности распознавания: на наборе FER2013 точность выросла с ~80% до 96,47%, а на RAF-DB – с ~95% до 99,23%, что соответствует абсолютному приросту качества на 16,7 и 4,5 процентных пункта соответственно. Эти результаты наглядно демонстрируют эффективность синтетической аугментации данных в задаче распознавания эмоций лица и подчеркивают потенциал современных генеративных

моделей (диффузионных, GAN и пр.) для преодоления проблемы нехватки данных.

Рисунок 2. Пример синтезированных изображений

Применение синтетических данных успешно развивается не только для визуальных, но и для других модальностей. В области распознавания эмоций по речи появляются работы, где для увеличения обучающей выборки используют генерацию искусственной эмоциональной речи. Один из подходов – применение нейросетей для преобразования нейтральных фраз в эмоциональные (speech emotion conversion), создавая таким образом новые аудиопримеры с заданной эмоцией. Другой подход – генерация эмоциональной речи с нуля при помощи text-to-speech систем, обученных выражать определенные эмоции. Аналогично, для текстовых данных возможно автогенерирование эмоционально окрашенных предложений с помощью больших языковых моделей (LLM) на основе заданных шаблонов или примеров. Таким образом, генеративный подход оказывается универсальным инструментом для различных типов данных.

Однако применение полностью синтетических данных сопряжено с рядом вызовов. Во-первых, качество и достоверность генерируемых образцов должны быть достаточно высокими, чтобы модель воспринимала их как релевантные примеры целевого класса. Если синтетические данные

содержат артефакты или непреднамеренные отличия (например, неестественные черты лица, шумы речи, несуразности текста), модель может обучиться на этих особенностях, что снизит ее эффективность на реальных данных. Во-вторых, необходимо обеспечение соответствия распределения: генератор должен охватить разнообразие реальных данных, иначе синтетические образцы могут не восполнить пропущенные вариации (или, наоборот, привнести смещение). Например, если сгенерированные лица имеют менее разнообразную внешность, чем в реальности, модель, обученная на них, может хуже работать на настоящих данных разных этнических или возрастных групп. В-третьих, остается важным вопрос валидации: для синтетических данных часто сложно объективно оценить их качество и эмоциональную метку без ручной проверки.

Для смягчения этих проблем разрабатываются методы управляемой генерации и фильтрации синтетических данных. Такой подход позволяет повысить точность соответствия генерируемых лиц заявленным эмоциям и отсеять некачественные образцы. Другим направлением является объединение реальных и синтетических данных при обучении: синтетические примеры используются в дополнение, а не вместо реальных, что обычно дает наилучшие результаты. Например, синтетическая выборка может увеличить исходный датасет в несколько раз, но итоговая модель обучается на смеси реальных и синтетических данных – это позволяет избежать полного смещения в сторону артефактов генератора. В работах отмечается, что оптимальным бывает масштаб, при котором добавление, например, 100–200% синтетических данных (относительно объема реальных) дает прирост качества, тогда как чрезмерное доминирование синтетики может ухудшить качество (эффект перенасыщения).

Наконец, важно подчеркнуть, что аугментация и генерация – не единственные стратегии борьбы с дефицитом данных. В практике распознавания эмоций широко используются методы трансферного

обучения, когда модель предварительно обучается на большой смежной базе (например, распознавание лиц или общая речь) и затем дообучается на малом эмоциональном датасете. Такой перенос знаний часто позволяет компенсировать недостаток данных, хотя полностью проблему не решает. Также развивается подход федеративного обучения, при котором данные, распределенные между различными пользователями или организациями, используются для совместного обучения модели без централизованного сбора – это может увеличить суммарный доступный объем данных при соблюдении конфиденциальности. Тем не менее, перечисленные методы скорее дополняют стратегии аугментации и генерации, но не заменяют их.

Ограниченность и несбалансированность данных остаются серьезным препятствием для построения точных и надежных систем распознавания эмоций. Рассмотренные в статье подходы – аугментация существующих данных и генерация синтетических примеров – демонстрируют высокую эффективность в преодолении данного препятствия. Аугментация данных служит относительно простым и действенным способом увеличить разнообразие обучающих примеров, минимизировать переобучение и улучшить обобщающую способность моделей. Генерация же синтетических данных при помощи современных нейросетевых моделей позволяет целенаправленно пополнять выборку недостающими образцами определенных классов эмоций, тем самым устраняя дисбаланс и расширяя охват пространства признаков.

Следует, однако, учитывать, что успех применения таких методов зависит от качества реализации. Необходимо тщательно контролировать достоверность синтетических данных, комбинировать их с реальными примерами и проводить всестороннюю валидацию моделей. В дальнейшем ожидается появление еще более совершенных генеративных подходов, способных создавать данные, максимально неотличимые от реальных, что позволит практически устранить разрыв между потребностями моделей в

данных и возможностями по их сбору. Совмещение же разных стратегий – аугментации, синтетической генерации, трансферного обучения – представляет собой наиболее мощный комплексный подход. Таким образом, преодоление дефицита данных в распознавании эмоций человека во многом связывается с прогрессом в методах аугментации и синтеза данных, которые уже сейчас демонстрируют свой высокий потенциал в повышении эффективности аффективных вычислительных систем.

Список использованных источников

1. Матвеева А.А., Махныткина О.В. Метод аугментации текстовых данных с сохранением стиля речи и лексики персоны // Научно-технический вестник информационных технологий, механики и оптики. – 2023. – Т. 23, № 4. – С. 743–749.
2. Рабчевский А.Н. Обзор методов и систем генерации синтетических обучающих данных // Прикладная математика и вопросы управления. – 2023. – № 4. – С. 6–45.
3. Рюмина Е.В., Карпов А.А. Сравнительный анализ методов устранения дисбаланса классов эмоций в видеоданных выражений лиц // Научно-технический вестник информационных технологий, механики и оптики. – 2020. – Т. 20, № 5. – С. 683–691.
4. Фостер Д. Генеративное глубокое обучение: творческий потенциал нейронных сетей / [пер. с англ.]. — Санкт-Петербург: Питер, 2020. — 352 с.
5. Roy A. K. Improvement in Facial Emotion Recognition using Synthetic Data Generated by Diffusion Model [Электронный ресурс] / A. K. Roy, H. K. Kathania, A. Sharma // arxiv.org. Режим доступа: https://arxiv.org/abs/2411.10863 (дата обращения: 19.05.2025).