ОБУЧАЮЩИЕСЯ ДАТАСЕТЫ

ДЛЯ РАСПОЗНАВАНИЯ ДИПФЕЙКОВ

Микунов А.В., аспирант, ФГБОУ ВО ОмГУПС

Елизаров Д.А., к.т.н., доцент, ФГБОУ ВО ОмГУПС

Аннотация. Статья посвящена анализу существующих наборов данных для разработки и тестирования сервисов, способных различать подлинные изображения и изображения, созданные искусственным интеллектом.

Ключевые слова: технология дипфейк, генеративный искусственный интеллект, датасет, набор данных, распознавание.

С развитием технологий генерации изображений, аудио на основе нейросетей возникли

новые вызовы в области безопасности и этики, стало возможным создавать визуальный контент, практически неотличимый от реального [1]. Одним из наиболее опасных проявлений этой технологии являются дипфейки – реалистичная манипуляция аудио-, фото- и видеоматериалами с помощью искусственного интеллекта для достижения максимального сходства с реальными изображениями и звуковыми дорожками.

Эффективное распознавание дипфейков требует больших объемов размеченных данных – датасетов. Датасет (набор данных) – это структурированный набор данных, который используется для решения определенных задач в области аналитики и машинного обучения[2].

В открытых источниках можно найти много датасетов для разных целей. Платформа Kaggle сбольшой коллекцией наборов данных по различным темам: экономика, здравоохранение, спорт, технологии [6]. Обычно датасеты Kaggle хранятся в формате CSV или Excel. Поисковая система для датасетов от Google подходит для поиска специализированных данных из научных публикаций и государственных источников. DataHub – коллекция общедоступных данных, включая экономику, транспорт и географию. Большинство датасетов – в формате CSV или JSON.

Во многих современных языках программирования есть библиотеки для работы с большими данными. А внутри библиотек есть встроенные датасеты, которые можно использовать. Например, на Python есть фреймворк для машинного обучения Pytorch со встроенной библиотекой torchvision. Похожие технологии есть и в других языках: C++, Java, R.Встроенные в библиотеки датасеты подходят для тех же целей, что и датасеты из открытых источников: обучение и тестирование моделей, тестирование гипотез,

обучение анализу данных. Главное, чтобы датасет подходил своей цели.

На онлайн-платформе Kaggle были рассмотрены готовые датасеты для распознавания дипфейков.

StyleGan-StyleGan2 Deepfake Face Images – набор данных для распознавания изображений лиц DeepFake. Набор данных состоит из двух наборов фото: реальных и фейковых. Реальные изображения получены из набора данных Nvidia Flickr. Поддельные изображения генерируются с помощью StyleGAN и поступают из обсуждения DeepFake Detection Challenge Discussion на Kaggle .Чтобы еще больше расширить набор данных и повысить надежность модели, были применены несколько дополнений. С помощью этих дополнений были созданы еще 6 445 изображений, что привело к окончательному набору данных из 12 890 изображений, в котором:5 890 реальных,7 000 – фальшивка. Данный датасет позволяет разрабатывать более эффективные модели, способные отличать реальные лица от синтетических в различных источниках и техниках генерации дипфейков.

Final Merged Dataset – датасет, который объединяет несколько высококачественных наборов данных для обнаружения дипфейков в единый структурированный формат для удобного использования в проектах машинного обучения и компьютерного зрения. Он включает в себя реальные и поддельные изображения лиц из четырех популярных датасетов:

Celeb-DF-New – высококачественные дипфейк-видео знаменитостей, преобразованные в кадры;

сбалансированная коллекция реальных и обработанных лиц;

FaceForensics++ – широко используемый стандарт для обнаружения дипфейков;

140k Real and Fake Faces – крупномасштабный набор данных синтетических и реальных лиц.

Набор данных был предварительно обработан и разделен на наборы для обучения (60%), проверки (20%) и тестирования (20%) для непосредственного использования в моделях глубокого обучения.

Detect AI-Generated Faces: High-Quality Dataset – набор данных содержит высококачественные изображения как реальных человеческих лиц, так и синтетических лиц, созданных искусственным интеллектом, предназначенных для приложений машинного и глубокого обучения. Он предоставляет ресурс для разработки и тестирования классификаторов, способных различать подлинные изображения лица и изображения, созданные искусственным интеллектом. Этот набор данных идеально подходит для таких задач, как обнаружение дипфейков, проверка подлинности изображений и анализ изображений лица, а также тщательно отобран для поддержки передовых исследований и приложений. Набор данных состоит из 3 203 изображений, в котором: 2 202 реальных, 1001 – фальшивка.

Deepfake image detection –это ресурс, предназначенный для исследователей, разработчиков и специалистов по обработке и анализу данных, работающих над обнаружением, анализом и пониманием дипфейков. Набор данных тщательно структурирован для поддержки приложений машинного обучения и искусственного интеллекта, особенно для улучшения систем обнаружения дипфейков. Он разделен на два основных подмножества: обучающие данные и тестовые данные, что позволяет без проблем разрабатывать и оценивать модели обнаружения.

В заключении следует отметить, что существующие датасеты либо ограничены по объему, либо не учитывают реальные условия эксплуатации. Поэтому возникает необходимость в разработке сервиса генерации собственных обучающих и тестовых датасетов, включающих как реальные, так и синтетические данные с контролируемыми характеристиками.

Литература

1. Обзор технологий создания Deepfake и методов его выявления — Научно-технический центр ФГУП «ГРЧЦ» (НТЦ) [Электронный ресурс]. – URL: https://rdc.grfc.ru/2020/06/research-deepfake/?ysclid=m9qt9yu6cp221085099(дата

обращения: 16.11.2025)

2. Что такое датасет и как его использовать [Электронный ресурс]: –URL: https://thecode.media/chto-takoe-dataset-i-chto-s-nim-delayut/(дата

обращения: 16.11.2025)

3. Синтетические данные в машинном обучении [Электронный ресурс]: –URL: https://data-light.ru/blog/sinteticheskie-dannie-ml/(дата

обращения: 16.11.2025)

4. Генерация синтетических данных: технологии и возможности [Электронный ресурс]: – URL:https://sky.pro/wiki/analytics/generatsiya-sinteticheskih-dannyh-tehnologii-i-vozmozhnosti/(дата обращения: 16.11.2025)

5. Инструменты Pythonдля генерации синтетических данных [Электронный ресурс]: –URL : https://habr.com/ru/articles/888830/(дата

обращения: 16.11.2025)

6. Kaggle [Электронный ресурс]: – URL: https://www.kaggle.com/datasets(дата обращения : 16.11.2025)

7. Гуселетова, А. Е. Инструменты обнаружения дипфейков / А. Е. Гуселетова, Д. А. Елизаров // Актуальные проблемы и тенденции развития современной экономики и информатики : Материалы Международной научно-практической конференции, Бирск, 04–06 декабря 2024 года. – Бирск: Уфимский университет науки и технологий, 2024. – С. 177-180. – EDN BWWETK.

8. Жуков, Д. В. Параметры и признаки для выявления дипфейков / Д. В. Жуков, А. А. Филатова // Студент: наука, профессия, жизнь : Материалы XII всероссийской студенческой научной конференции с международным участием. В 5-ти частях, Омск, 21–25 апреля 2025 года. – Омск: Омский государственный университет путей сообщения, 2025. – С. 464-467. – EDN GDMDXN.