СОДЕРЖАНИЕ

ВВЕДЕНИЕ

Настоящий отчёт посвящён выполнению расчётно-графической работы (РГР) по теме «Работа с онлайн-инструментами для исследователя». Цель работы — освоить современные онлайн-инструменты для ведения исследовательской деятельности, создания и размещения публикаций, а также получить практический опыт их использования для решения реальных задач.

Научные интересы автора связаны с темой диссертации: генерация синтетических данных для обучения нейронных сетей. Методология включает изучение теоретической справки, апробацию онлайн-инструментов и выполнение заданий в соответствии с указаниями.

1 Сравнение баз научных работ

– **Elibrary**

Очень много специфически направленных исследований на генерацию данных для какого-то определенного сервиса/по/задачи.

Elibrary.ru (РИНЦ) является незаменимым ресурсом для исследователей, работающих в российском научном контексте. Платформа предоставляет наиболее полный доступ к российским научным публикациям и позволяет отслеживать публикационную активность отечественных ученых и организаций. Однако для темы синтетических данных и машинного обучения база данных имеет ограниченную применимость из-за преобладания русскоязычного контента и относительно небольшого количества передовых исследований в этой области.

Elibrary.ru имеет ограниченную релевантность для данной темы исследований, что обусловлено несколькими факторами. Во-первых, платформа фокусируется преимущественно на российских публикациях, тогда как основной массив исследований по синтетическим данным и глубокому обучению публикуется на английском языке в международных изданиях. Во-вторых, российская научная традиция имеет определенное отставание в области современного машинного обучения от ведущих мировых

центров. Однако для исследователей, интересующихся российским контекстом или работами отечественных ученых, платформа может предоставить уникальный контент.

- **Google Scholar**

  Хайп

  Google Scholar представляет собой оптимальный выбор для исследователей, стремящихся к максимальному охвату научной литературы. Платформа индексирует подавляющее большинство научных публикаций, включая материалы из всех других рассматриваемых баз данных, и при этом обнаруживает значительное количество уникального контента. Бесплатный доступ и простота использования делают Google Scholar универсальным инструментом для первичного поиска. Однако необходимо учитывать включение нерецензируемых источников и требовать критической оценки качества найденных материалов.

  Google Scholar демонстрирует высокую релевантность для исследований по синтетическим данным благодаря своему глобальному охвату и широкой индексации. Поиск по запросу «synthetic data generation neural networks» в Google Scholar возвращает приблизительно 1-2 миллиона результатов, охватывающих различные аспекты темы от базовых концепций до новейших разработок. База данных индексирует публикации из всех основных источников, включая журналы, конференции, препринты и диссертации.

- **arXiv**

  Не самый удобный поиск. Присутствует множество AI инструментов, поддерживаемых сообществом для быстрого анализа статей или их суммаризации.

  ArXiv представляет исключительно высокую ценность для исследований в области генерации синтетических данных. Категории машинного обучения (cs.LG), компьютерного зрения (cs.CV) и вычислений и языка (cs.CL) содержат тысячи препринтов по данной теме. По оценкам, в arXiv размещено более 10000+ работ, связанных с синтетическими данными и нейронными сетями.

ArXiv содержит основополагающие работы по синтетическим данным, включая комплексные обзоры методов генерации синтетических данных с использованием машинного обучения. Препринты охватывают широкий спектр применений — от компьютерного зрения и обработки естественного языка до медицины и финансов. Платформа также индексирует работы по теоретическому пониманию синтетических данных в обучении больших языковых моделей и проблеме коллапса моделей при рекурсивном обучении на синтетических данных.

- **ScienceDirect**

Похож на Elibrary по содрежанию. При этом есть разделение статей на обзорны/исследовательские

ScienceDirect предоставляет доступ к высококачественным рецензируемым публикациям из журналов Elsevier, охватывающих 25% мировых научных публикаций. Интеграция со Scopus обеспечивает мощные наукометрические инструменты для оценки влияния исследований. Функция Topic Pages с использованием искусственного интеллекта помогает исследователям быстро получать обзор по конкретным темам. Ограничением является модель подписки, хотя значительная часть контента доступна в открытом доступе.

ScienceDirect демонстрирует высокую релевантность для темы синтетических данных, предоставляя доступ к рецензированным журнальным статьям высокого качества. Оценочно, платформа содержит более 100000 статей, связанных с машинным обучением и генерацией данных. База данных включает публикации из ведущих журналов Elsevier по компьютерным наукам, искусственному интеллекту и нейронным сетям.

Особенностью ScienceDirect является наличие междисциплинарного контента, охватывающего применение синтетических данных в медицине, биоинформатике, автономных транспортных средствах и других областях. Платформа индексирует работы по различным методам генерации синтетических данных, включая статистические подходы, глубокое обучение и дифференциальную приватность. База данных также содержит

обзорные статьи и книги, такие как «Synthetic Data and Generative AI», обеспечивающие систематическое введение в тему.

SpringerLink

SpringerLink предлагает аналогичные преимущества, предоставляя доступ к обширной коллекции рецензируемых публикаций Springer Nature. Особенностью является наличие архивной коллекции, охватывающей период с 1832 года, что полезно для исторических исследований. Интеграция с Web of Science и Scopus обеспечивает отслеживание индексации и метрик цитирования. Как и ScienceDirect, платформа работает по модели подписки с существенной долей открытого контента.

SpringerLink также проявляет высокую релевантность для исследований по синтетическим данным. Оценочно, платформа содержит более 50000 статей по машинному обучению и связанным темам. Издательство Springer Nature выпускает специализированные книги по синтетическим данным для глубокого обучения, которые служат комплексными обзорами области.

2 Подбор и анализ научных статей

Привести их названия, авторов, даты и место публикаций, краткое содержание, ваши комментарии.

1) **Comprehensive Exploration of Synthetic Data Generation: A Survey**

Авторы: André Bauer, Simon Trapp, Michael Stenger, Robert Leppich, Samuel Kounev, Mark Leznik, Kyle Chard, Ian Foster

Место публикации: arXiv

Даты публикации:

- v1. 04.01.2024
- v2. 01.02.2024

Содержание: Обзор 417 моделей генерации синтетических данных за последнее десятилетие. Работа охватывает типы моделей,

функциональность и применение. Исследование показывает преобладание подходов на основе нейронных сетей, доминирование компьютерного зрения с GAN как основными генеративными моделями, а также растущую роль диффузионных моделей, трансформеров и RNN. Авторы подчеркивают нехватку общих метрик и датасетов для сравнения, а также необходимость учета вычислительных затрат.

2) **Generative Adversarial Networks for Synthetic Data Generation in Deep Learning Applications**

Автор: M. Keskes

Место публикации: Journal of Artificial Intelligence Research and Innovation

Дата публикации: 05.09.2025

Содержание: Комплексный анализ роли GAN в создании высококачественных синтетических данных в различных областях, включая здравоохранение, финансы, компьютерное зрение и обработку естественного языка. Работа исследует базовые принципы GAN, продвинутые архитектуры (DCGAN, cGAN, CycleGAN, TimeGAN) и их применение для генерации медицинских изображений, финансовых временных рядов и табличных данных. Обсуждаются преимущества GAN, такие как сохранение приватности и эффективность по затратам, наряду с ограничениями, включая нестабильность обучения и отсутствие стандартизированных метрик оценки.

3) **Synthetic Data Generation by Diffusion Models**

Автор: Jun Zhu

Место публикации: National Science Review

Дата публикации: 24.08.2024

Содержание: Краткий обзор диффузионных моделей, которые чрезвычайно мощны для генерации многомерных данных, включая изображения, 3D-контент. Диффузионные модели широко используются для моделирования распределения данных непрерывной области благодаря стабильности обучения и сильной модельной емкости. Помимо изображений, диффузионные модели применяются для генерации

высококачественных данных в различных областях: речь, 3D-объекты, движения человека, видео и молекулы.

4) **Machine Learning for Synthetic Data Generation: A Review**

Автор: Yingzhou Lu, Lulu Chen, Yuanyuan Zhang, Minjie Shen, Huazheng Wang, Xiao Wang, Capucine van Rechem, Tianfan Fu, Wenqi Wei

Место публикации: arXiv

Дата публикации:

Содержание: Комплексный систематический обзор существующих исследований, использующих модели машинного обучения для генерации синтетических данных. Работа анализирует различные подходы к генерации синтетических данных с применением методов машинного обучения, включая методы для траекторий, временных рядов и других типов данных. Особое внимание уделяется методам, обеспечивающим дифференциальную приватность.

5) **Synthetic Data and Artificial Neural Networks for Natural Scene Text Recognition**

Автор: Max Jaderberg, Karen Simonyan, Andrea Vedaldi, Andrew Zisserman

Место публикации: arXiv

Даты публикации:

– v1. 09.06.2014

…

– v4. 09.12.2014

Содержание:

6)

Автор:

Место публикации:

Даты публикации:

Содержание:

7)

Автор:

Место публикации:

Даты публикации:

Содержание:

8)

Автор:

Место публикации:

Даты публикации:

Содержание:

9)

Автор:

Место публикации:

Даты публикации:

Содержание:

10)

Автор:

Место публикации:

Даты публикации:

Содержание:

3 Определение влиятельных авторов

1) **Ian Goodfellow**

Время работы: 11 лет

Цитирований: 393172

Индекс h: 101

Индекс i10: 185

Ссылка на профиль: https://scholar.google.com/citations?user=iYN86KEAAAAJ&hl=en&oi=ao

2) **Mihaela van der Schaar**

Время работы: 11 лет

Цитирований: 43887

Индекс h: 95

Индекс i10: 600

Ссылка на профиль: https://scholar.google.com/citations?user=DZ3S--MAAAAJ&hl=en

3) **Yejin Choi**

Время работы: 18

Цитирований: 80421

Индекс h: 128

Индекс i10: 348

Ссылка на профиль: https://scholar.google.com/citations?user=vhP-tlcAAAAJ

4) **Jonathan Ho**

Время работы: 12 лет

Цитирований: 76556

Индекс h: 36

Индекс i10: 41

Ссылка на профиль: https://scholar.google.com/citations?user=iVLAQysAAAAJ&hl=en

5) **Samuli Laine**

Время работы: 9 лет

Цитирований: 60888

Индекс h: 47

Ссылка на профиль: 96

См. подробнее на официальном сайте Typst.

4 Определение влиятельных изданий

Полная гойда расположена в таблицах 1, 2

Таблица 1 — Таблица параметров эксперимента

Journal Name	Impact Factor	Key Indexes
Nature Machine Intelligence	23.9	Scopus, WoS, SCIE
IEEE TPAMI	18.6	Scopus, WoS, SCIE
Artificial Intelligence Review	13.9	Scopus, WoS
Journal of Machine Learning Research	5.2	Scopus, WoS, SCIE
Artificial Intelligence (Elsevier)	4.6	Scopus, WoS

Таблица 2 — Таблица параметров эксперимента

Journal Name	Impact Factor	Key Indexes
Nature Machine Intelligence	23.9	Scopus, WoS, SCIE
IEEE TPAMI	18.6	Scopus, WoS, SCIE
Artificial Intelligence Review	13.9	Scopus, WoS
Journal of Machine Learning Research	5.2	Scopus, WoS, SCIE
Artificial Intelligence (Elsevier)	4.6	Scopus, WoS

5 Подбор международных научных конференций

ЗАКЛЮЧЕНИЕ

В ходе РГР освоены онлайн-инструменты, проведено сравнение баз, подобраны материалы и налажено взаимодействие в ResearchGate.

6 Тест