УДК 004.8

ГЕНЕРАЦИЯ СИНТЕТИЧЕСКИХ ДАННЫХ В ОБУЧЕНИИ ИСКУССТВЕННЫХ НЕЙРОННЫХ СЕТЕЙ

Головин А.С.
магистр, Южный федеральный университет
(Ростов-на-Дону, Россия)

SYNTHETIC DATA GENERATION IN TRAINING ARTIFICIAL NEURAL NETWORKS

Golovin A.
master's degree, Southern Federal University
(Rostov-on-Don, Russia)

Аннотация

В статье представлен обзор методов генерации синтетических данных для обучения искусственных нейронных сетей (ИНС) в условиях ограниченного доступа к реальным данным. Рассмотрены ключевые подходы, включая генеративные состязательные сети (GAN), методы повышения данных и статистические модели, применимые к различным типам данных. Особое внимание уделено применению синтетических данных в задачах распознавания лиц, диагностики редких заболеваний и управления автономными транспортными системами. Проанализированы преимущества и ограничения каждого метода, а также их влияние на качество и точность моделей. Рассмотрены возможные риски, связанные с использованием синтетических данных, такие как искажения и смещения, и подходы к их минимизации с целью повышения надежности обучения. Применение синтетических данных открывает значительные перспективы для расширения возможностей ИНС, способствуя улучшению их эффективности и обобщающей способности в реальных задачах.

Ключевые слова: синтетические данные, обучение с подкреплением, генеративные состязательные сети, искусственные нейронные сети, повышение данных.

Abstract

This article presents an overview of synthetic data generation methods for training artificial neural networks (ANNs) under limited access to real data. Key approaches such as Generative Adversarial Networks (GAN), data augmentation methods, and statistical models applicable to various data types are reviewed. Special attention is given to the application of synthetic data in face recognition, rare disease diagnosis, and autonomous systems management. The advantages and limitations of each method, as well as their impact on model accuracy, are analyzed. Potential risks associated with synthetic data, including biases and distortions, and approaches to mitigate these issues to enhance model reliability, are also discussed. The use of synthetic data provides substantial opportunities for advancing ANNs, improving their effectiveness and generalizability in practical tasks.

Keywords: synthetic data, reinforcement learning, generative adversarial networks, artificial neural network, data augmentation.

Введение

С развитием технологий машинного обучения (МО) и искусственных нейронных сетей (ИНС) возникла потребность в большом количестве данных для обучения моделей. Однако

реальные данные не всегда доступны в требуемом объеме, что усложняет процесс обучения и снижает точность прогнозов моделей. В таких условиях генерация синтетических данных становится важным инструментом, позволяющим восполнить нехватку данных и повысить эффективность ИНС. Основной целью данной статьи является исследование методов генерации синтетических данных, используемых для обучения ИНС, и оценка их эффективности в различных задачах.

Использование синтетических данных позволяет не только восполнить дефицит информации, но и улучшить производительность ИНС за счет формирования большего разнообразия обучающих примеров. Синтетические данные могут быть сгенерированы различными способами, включая статистические модели, генеративные состязательные сети (ГСН) и методы на основе преобразования изображений. Генерация синтетических данных имеет значительный потенциал для применения в областях, где получение реальных данных затруднено, например, в медицине, финансовом секторе и анализе изображений. В данной статье рассматриваются основные методы генерации синтетических данных и их влияние на обучение ИНС.

Введение синтетических данных требует также учета рисков, связанных с их использованием, таких как возможность появления смещений в данных и снижение обобщающей способности модели. Одной из задач данной статьи является изучение потенциальных рисков, связанных с применением синтетических данных, а также рассмотрение подходов к их минимизации. Целью исследования является обобщение текущих подходов к генерации синтетических данных и выявление наиболее эффективных методов для конкретных задач ИНС, а также анализ проблем, возникающих при внедрении синтетических данных в процесс обучения.

Основная часть

Методы генерации синтетических данных могут быть разделены на несколько подходов в зависимости от типа данных и целей их использования. Одним из наиболее распространенных методов является применение генеративных состязательных сетей (Generative Adversarial Networks, GAN), которые состоят из двух нейронных сетей: генератора и дискриминатора. Генератор создает синтетические данные, стараясь сделать их максимально похожими на реальные, в то время как дискриминатор пытается отличить синтетические данные от реальных [1]. Совместное обучение этих сетей позволяет генератору постепенно улучшать качество синтетических данных. Метод ГСН широко используется для генерации изображений, текстов и других сложных данных.

Для задач, связанных с изображениями, также применяют методы повышения данных (Data Augmentation), такие как поворот, масштабирование и добавление шума к исходным изображениям [2]. Эти методы позволяют значительно расширить обучающую выборку, не требуя создания новых данных, что особенно полезно для задач классификации и распознавания. В таблице 1 представлено сравнение различных методов генерации синтетических данных, где указаны области применения, преимущества, ограничения и примеры использования.

Таблица 1 [3]

Основные методы генерации синтетических данных, их применение, преимущества и ограничения

Метод	**Применение**	**Преимущества**	**Ограничения**	**Примеры применения**
ГСН	Изображения, текст, звуковые данные	Высокое качество, способность к обучению сложных распределений	Длительное обучение, необходимость тонкой настройки	Генерация лиц для распознавания, создание текстовых описаний

Повышение данных	Изображения	Увеличение выборки, простота, низкая стоимость	Ограниченная генерация новых признаков	Обучение классификаторов изображений, медицинская визуализация
Статистические модели	Табличные данные, финансовые данные	Контроль над распределениями, предсказуемость характеристик	Могут не учитывать сложные зависимости	Анализ финансовых данных, моделирование поведения пользователей
Модели на основе временных рядов (PHH, LSTM)	Временные ряды, данные сенсоров	Способность учитывать временные зависимости	Требует большого объема данных, высокая вычислительная нагрузка	Прогнозирование спроса, анализ данных IoT
Симуляции на основе физических моделей	Инженерия, физические процессы	Реализм, высокое соответствие физическим законам	Высокие затраты на разработку, сложность моделей	Обучение автономных систем, робототехника

В задачах, требующих использования табличных данных, часто применяются статистические модели, которые позволяют генерировать данные на основе заранее определенных вероятностных распределений. Этот метод обеспечивает высокую точность синтетических данных, соответствующих реальным параметрам, что особенно полезно при создании финансовых или пользовательских данных [4].

На рисунке 1 представлена сравнительная эффективность различных методов генерации синтетических данных, таких как ГСН, повышение данных, статистические модели, LSTM и физические симуляции. Эти данные помогают оценить преимущества каждого подхода.

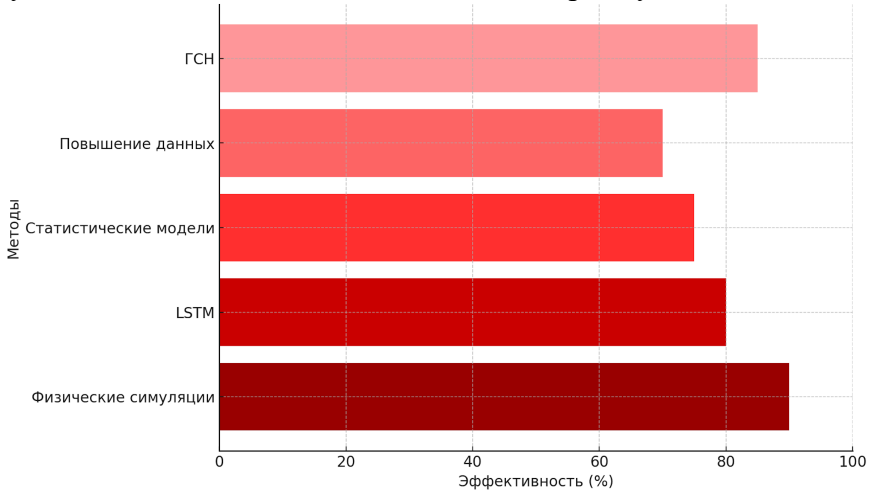

Рисунок 1. Эффективность методов генерации синтетических данных

Как видно из рисунка 1, физические симуляции показывают наибольшую эффективность (90%), благодаря их способности учитывать реальные физические процессы. Методы ГСН (85%) и LSTM (80%) также демонстрируют высокую результативность, особенно в задачах, требующих сложного моделирования данных. Повышение данных и статистические модели, несмотря на свою простоту и доступность, имеют эффективность 70% и 75% соответственно, что делает их удобными для быстрого увеличения обучающих выборок. Эти показатели подчеркивают необходимость выбора подходящего метода в зависимости от задачи и ресурсов.

Пример кода для генерации табличных синтетических данных с нормальным распределением представлен ниже.

```
import numpy as np
import pandas as pd

# Генерация синтетических данных на основе нормального распределения
mean = [0, 0]
cov = [[1, 0.5], [0.5, 1]]
synthetic_data = np.random.multivariate_normal(mean, cov, size=100)
df = pd.DataFrame(synthetic_data, columns=["Feature1", "Feature2"])

print(df.head())
```

Для задач временных рядов используются рекуррентные нейронные сети и модели с долгой краткосрочной памятью (LSTM), что позволяет учитывать временные зависимости в данных. Этот метод особенно полезен для прогнозирования и моделирования процессов, зависящих от последовательности событий, таких как изменения спроса или температурные колебания. Важно отметить, что использование таких методов требует значительных вычислительных ресурсов и больших объемов данных для обучения.

Использование синтетических данных в процессе обучения ИНС требует учета ряда рисков. Например, при генерации синтетических данных возможно появление искажений, что может привести к снижению точности и надежности модели [5]. Для минимизации этих рисков часто применяются гибридные подходы, совмещающие реальные и синтетические данные. Такой метод обучения позволяет повысить обобщающую способность модели, избегая излишней зависимости от синтетических данных.

Применение синтетических данных предоставляет значительные возможности для обучения ИНС в условиях ограниченного доступа к реальным данным. Так, в медицине они могут быть использованы для моделирования редких заболеваний, что позволяет создавать более точные диагностические модели и системы прогнозирования.

Практическое применение синтетических данных в обучении ИНС

Для задач, связанных с распознаванием лиц и анализа изображений, синтетические данные, созданные с помощью GAN, позволяют обучать модели на разнообразных примерах, сохраняя при этом высокое качество генерации. Например, в задачах безопасности GAN могут генерировать синтетические изображения лиц для тренировок систем распознавания в условиях ограниченных данных.

В медицине синтетические данные находят применение при обучении моделей для диагностики редких заболеваний [6]. При недостаточности реальных данных модели могут использовать синтетические изображения, созданные на основе исходных снимков, чтобы расширить обучающую выборку и повысить точность диагностики. В частности, синтетические рентгеновские снимки или МРТ-изображения позволяют обучить нейронные сети без необходимости получения новых данных от пациентов, что снижает риски конфиденциальности и затраты.

В автономных транспортных системах генерация данных также играет ключевую роль. Для беспилотных автомобилей синтетические изображения дорожных ситуаций помогают моделям обучаться на различных сценариях, включая экстремальные погодные условия или неожиданные препятствия. Использование GAN для создания таких данных помогает избежать необходимости проведения длительных и дорогостоящих полевых испытаний.

В финансовом секторе ИНС, обученные на синтетических данных, используются для анализа поведения пользователей, прогнозирования рыночных трендов и оценки рисков. Статистические модели позволяют генерировать данные с учетом заданных распределений, что полезно при моделировании редких событий, таких как кризисы или внезапные рыночные колебания [7]. Использование синтетических данных в процессе обучения ИНС предоставляет значительные возможности для обучения моделей в условиях ограниченного доступа к реальным данным.

Реализация генерации синтетических данных с помощью GAN

GAN играют важную роль в создании синтетических данных, особенно в задачах, связанных с генерацией изображений, текста и других сложных данных. GAN представляют собой архитектуру, состоящую из двух нейронных сетей — генератора и дискриминатора. Генератор создает синтетические данные, используя случайный шум, а дискриминатор оценивает, насколько близки созданные данные к реальным. Процесс обучения GAN заключается в том, что обе сети "состязаются" друг с другом: генератор старается создавать данные, неотличимые от реальных, а дискриминатор стремится эффективно различать синтетические и реальные данные. В результате генератор обучается создавать все более качественные данные, что делает этот метод особенно эффективным для задач, требующих реалистичной генерации [8, 9].

В следующем примере показан процесс генерации синтетических изображений с использованием GAN. На вход генератору подается случайный шум, на основе которого он создает изображение. Дискриминатор, в свою очередь, оценивает это изображение, указывая, насколько оно похоже на реальное. После завершения обучения генератор может использоваться для создания синтетических изображений, которые могут быть полезны для обучения других моделей, например, в задачах распознавания объектов или классификации изображений.

```
import tensorflow as tf
from tensorflow.keras.layers import Dense, Reshape, Flatten, LeakyReLU
from tensorflow.keras.models import Sequential

# Параметры
latent_dim = 100  # Размерность скрытого пространства (зашумленных данных)

# Модель генератора
def build_generator():
    model = Sequential()
    model.add(Dense(256, input_dim=latent_dim))
    model.add(LeakyReLU(alpha=0.2))
    model.add(Dense(512))
    model.add(LeakyReLU(alpha=0.2))
    model.add(Dense(1024))
    model.add(LeakyReLU(alpha=0.2))
    model.add(Dense(28 * 28, activation='tanh'))
    model.add(Reshape((28, 28, 1)))
    return model

# Модель дискриминатора
def build_discriminator():
    model = Sequential()
    model.add(Flatten(input_shape=(28, 28, 1)))
    model.add(Dense(512))
    model.add(LeakyReLU(alpha=0.2))
    model.add(Dense(256))
    model.add(LeakyReLU(alpha=0.2))
    model.add(Dense(1, activation='sigmoid'))
    return model

# Компиляция моделей
generator = build_generator()
discriminator = build_discriminator()
discriminator.compile(optimizer='adam', loss='binary_crossentropy', metrics=['accuracy'])
```

```
# GAN - совмещенная модель
discriminator.trainable = False  # Фиксируем дискриминатор для обучения только генератора
gan_input = tf.keras.Input(shape=(latent_dim,))
generated_image = generator(gan_input)
gan_output = discriminator(generated_image)
gan = tf.keras.Model(gan_input, gan_output)
gan.compile(optimizer='adam', loss='binary_crossentropy')

# Пример тренировки GAN
import numpy as np

# Случайный шум для генерации синтетических данных
noise = np.random.normal(0, 1, (1, latent_dim))
generated_image = generator.predict(noise)

# Визуализация сгенерированного изображения
import matplotlib.pyplot as plt

plt.imshow(generated_image[0, :, :, 0], cmap='gray')
plt.title("Сгенерированное изображение")
plt.axis('off')
plt.show()
```

Этот пример демонстрирует, как использовать GAN для создания синтетических изображений на основе случайного шума. На этапе визуализации получаем изображение, сгенерированное генератором, что наглядно показывает, как нейронная сеть способна воссоздать структурированные данные на основе случайных входных данных. Данная архитектура широко применяется в задачах, где синтетические данные необходимы для обучения моделей, работающих в условиях дефицита реальных данных, включая задачи, связанные с медицинской визуализацией, автономными транспортными системами и биометрией.

Заключение

Использование синтетических данных, созданных с помощью генеративных алгоритмов, значительно расширяет возможности обучения искусственных нейронных сетей, особенно в условиях ограниченного доступа к реальным данным. Такие методы, как GAN, позволяют создавать качественные и реалистичные данные, пригодные для обучения моделей в задачах распознавания лиц, диагностики заболеваний, управления автономными системами и финансового анализа. Объединение различных подходов, включая методы повышения данных и статистические модели, делает возможным решение широкого спектра задач с высокой точностью.

Особое внимание необходимо уделять возможным искажениям и смещениям, возникающим при генерации синтетических данных, так как они могут негативно повлиять на точность и обобщающую способность моделей. Применение гибридных подходов, совмещающих реальные и синтетические данные, позволяет снизить такие риски и обеспечить более надежное обучение. Важно также учитывать, что генерация синтетических данных требует значительных вычислительных ресурсов, особенно при использовании рекуррентных сетей и физических симуляций.

Синтетические данные становятся все более востребованными в современных областях науки и технологий, и их применение открывает перспективы для разработки новых решений и алгоритмов. При дальнейшем совершенствовании методов генерации и контроля качества синтетических данных можно ожидать существенного увеличения точности и эффективности

ИНС, что будет способствовать развитию МО и его внедрению в различные прикладные задачи.

Список литературы

1. Пчелинцев С., Юляшков М.А., Ковалева О.А. Метод создания синтетических наборов данных для обучения нейросетевых моделей распознаванию объектов // Информационно-управляющие системы. 2022. №3(118). С. 9-19.
2. Ходасевич Л.А. Генерация реалистичных изображений для обучения искусственных нейронных сетей в задаче навигации робота // Информатика. 2018. Т. 15. №4. С. 50-58.
3. Моисеев Б., Чигорин А. Классификация автодорожных знаков на основе свёрточной нейросети, обученной на синтетических данных // The 22nd International Conference on Computer Graphics and Vision. 2012. С. 284-287.
4. Рубцов И.А. Методические подходы к исправлению проблемы недостаточности инфографических данных для обучения нейронных сетей // Вестник магистратуры. 2020. №5-3. С. 108.
5. Ковалев В.А., Козловский С.А., Калиновский А.А. Генерация искусственных рентгеновских изображений грудной клетки с использованием генеративно-состязательных нейронных сетей // Информатика. 2018. Т. 15. №2. С. 7-16.
6. Кабанова В.В., Логунова О.С. Применение искусственного интеллекта при работе с мультимедийной информацией // Вестник Череповецкого государственного университета. 2022. №6(111). С. 23-41.
7. Малов Д.А., Летенков М.А. Методика генерации искусственных наборов данных и архитектура системы распознавания лиц для взаимодействия с роботами внутри киберфизического пространства // Робототехника и техническая кибернетика. 2019. Т. 7. №2. С. 100-108.
8. Берзин В.И., Судейкин М.И. Разработка алгоритмов генерации синтетических данных для обучения нейросетевых моделей детектирования объектов на изображении // Физико-техническая информатика (СРТ2020). 2020. С. 106-122.
9. Юрин А.Н. Создание обучающей выборки для искусственной нейронной сети системы технического зрения // Механизация и электрификация сельского хозяйства. 2023. Т. 1. №56. С. 148-153.