Ромашов В.А., Еремук В.В. Влияние синтетически сгенерированных данных на устойчивость классификационных моделей к состязательным атакам // Международный журнал информационных технологий и энергоэффективности.– 2025. – Т. 10 № 3(53) с. 9–12

Международный журнал информационных технологий и энергоэффективности

Сайт журнала:
http://www.openaccessscience.ru/index.php/ijcse/

УДК 004.855.5

ВЛИЯНИЕ СИНТЕТИЧЕСКИ СГЕНЕРИРОВАННЫХ ДАННЫХ НА УСТОЙЧИВОСТЬ КЛАССИФИКАЦИОННЫХ МОДЕЛЕЙ К СОСТЯЗАТЕЛЬНЫМ АТАКАМ

[1] Ромашов В.А., [2]Еремук В.В.
ФГАОУ ВО "НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ ИТМО" , Санкт-Петербург, Россия (197101, город Санкт-Петербург, Кронверкский пр-кт, д. 49 литер а), e-mail: [1] whiviktor@gmail.com, [2]polar.vl@yandex.ru

В данной работе рассматривается, как добавление синтетических данных, сгенерированных GAN-моделью, влияет на обучение и устойчивость классификационной модели. На примере CIFAR-10 было показано, что увеличение объёма обучающей выборки за счёт синтетических данных приводит к незначительному росту точности на “чистых” данных, но не решает проблему уязвимости перед состязательными атаками FGSM и PGD.

Ключевые слова: GAN, синтетические данные, устойчивость, FGSM, PGD.

THE IMPACT OF SYNTHETICALLY GENERATED DATA ON THE ROBUSTNESS OF CLASSIFICATION MODELS TO ADVERSARIAL ATTACKS

[1] Romashov V.A., [2]Eremuk V.V.
"NATIONAL RESEARCH UNIVERSITY ITMO", St. Petersburg, Russia (197101, St. Petersburg, Kronverksky prospekt, 49 letter a), e-mail: [1] whiviktor@gmail.com, [2]polar.vl@yandex.ru

This paper examines how adding synthetic data generated by a GAN model affects the training and robustness of a classification model. Using CIFAR-10 as an example, it was shown that increasing the training sample size with synthetic data leads to a slight increase in accuracy on “clean” data but does not solve the problem of vulnerability to FGSM and PGD adversarial attacks.

Keywords: GAN, synthetic data, robustness, FGSM, PGD.

Введение

Одним из способов повышения надежности моделей является увеличение объемов обучающей выборки и улучшение ее качества. Однако в ряде случаев получение новых данных ограничено высокой стоимостью разметки, сложностью сбора или нормативными требованиями, например, в области медицинской визуализации или биометрической идентификации.

Проблема недостаточности обучающих данных часто решается путём искусственного расширения (augmentation). В последнее время популярность набирают генеративные модели, такие как GAN (генеративно-состязательная сеть, Generative Adversarial Network) [1-7]. Они способны генерировать данные, похожие на реальные, что может улучшать обобщающую способность классификаторов [3].

Целью данной работы является исследование влияния синтетических данных, сгенерированных с помощью GAN, на точность классификационной модели и её устойчивость к состязательным атакам FGSM и PGD.

В рамках данной работы выполнен эксперимент на наборе данных CIFAR-10 [8], объединяя реальный набор данных с синтетическим, сгенерированным StyleGAN2-ADA [9], с последующей проверкой влияния количества синтетических изображений на точность и устойчивость.

Эксперимент

В качестве GAN-модели выбрана StyleGAN2-ADA, обученная на CIFAR-10. Пусть $z \in R^{512}$ — латентный вектор, тогда генератор G порождает изображение $x_{syn} = G(z)$. Для каждого класса можно либо использовать условный генератор (c_{dim}=10), либо присваивать случайную или псевдо-метку.

Обозначим оригинальный обучающий набор D_{real}. Генерируем N изображений на класс и формируем D_{syn}. Тогда итоговый набор:

$$D^{*}_{\text{train}} = D_{\text{real}} \cup D_{\text{syn}}$$

В работе N варьировалось от 10 до 1000.

В экспериментах применялась ResNet50, обучаемая методом AdamW (lr=0.001). Число эпох от 10 до 30. Для проверки точности (accuracy) использовался тестовый набор данных CIFAR-10 из 10 тысяч изображений. Для атак были использованы:

1. FGSM [10];
2. PGD [10].

Результаты

При $N = 0$ (только реальные данные) точность модели равна ~87 %. При $= 1000$ увеличивается до ~88.5 %. Таким образом, синтетические данные действительно дают прирост на +1–2 % к итоговой точности за счёт расширения обучающей выборки.

Дополнительно были проведены эксперименты с различными стратегиями разметки синтетических изображений (условное и безусловное обучение генеративной модели). Анализ результатов показал, что использование условного StyleGAN2-ADA, генерирующего изображения с привязкой к классам, дало аналогичный прирост точности на чистых данных, но не изменило устойчивость к атакам.

Добавление синтетики не повысило устойчивость модели при атакующих сценариях FGSM/PGD. Например, при $\epsilon = 0.03$ точность падала ниже 20 %, что практически совпадало с вариантом без синтетических данных. Без обучения на состязательных примерах синтетические данные не способны улучшить общую устойчивость модели. Результаты эксперимента показаны в таблице 1.

Полученные данные указывают на то, что синтетические изображения, сгенерированные с помощью GAN, увеличивают общую точность модели, но этот эффект в основном проявляется в небольшом повышении точности на чистых данных.

Однако устойчивость модели к состязательным атакам практически не изменилась, что говорит о том, что сама по себе диверсификация входных данных без дополнительных защитных механизмов не является достаточной для повышения устойчивости.

Таблица 1 - Результаты эксперимента

Модель	Точность на чистых данных	Точность при FGSM (ε=0.01)	Точность при FGSM (ε=0.03)	Точность при PGD (ε=0.01)	Точность при PGD (ε=0.03)
Оригинальная	87.2%	58.4%	32.1%	55.7%	28.9%
Расширенная	88.5%	59.1%	33.4%	56.3%	30.2%

Это подчёркивает необходимость интеграции более специализированных методов защиты, таких как тренировка на обучающих примерах, что соответствует тенденциям современных исследований в области устойчивости нейросетей к атакам.

Выводы

Таким образом, в результате данной работы было показано, синтетические данные способны улучшить обычную точность, но не решают проблему воздействия вредоносных возмущений без дополнительных защитных мер.

Было установлено, что добавление синтетических данных, сгенерированных StyleGAN2-ADA, даёт незначительный прирост точности (до 1–2%) на чистых данных, но не улучшает устойчивость модели к состязательным атакам FGSM и PGD. Эксперименты показали, что даже при значительном увеличении обучающей выборки модель остаётся уязвимой к состязательным атакам и требуют дополнительного обучения состязательными примерами для повышения общей устойчивости к атакам.

Необходимо учитывать влияние качества синтетических изображений на общие результаты. Хотя генеративные модели способны создавать реалистичные данные, в ряде случаев они могут не полностью соответствовать нужной тематике, что приводит к ухудшению обобщающей способности классификатора. Это особенно важно для задач, требующих высокой точности, например, в медицинской диагностике или системах автоматического мониторинга.

Список литературы

1. Беляева О. В., Перминов А. И., Козлов И. С. Использование синтетических данных для тонкой настройки моделей сегментации документов //Труды Института системного программирования РАН. – 2020. – Т. 32. – №. 4. – С. 189-202.
2. Рабчевский А. Н. Обзор методов и систем генерации синтетических обучающих данных //математика. – 2023. – №. 4. – С. 6-45.
3. Medvedev D., D'yakonov A. Learning to generate synthetic training data using gradient matching and implicit differentiation //International Conference on Analysis of Images, Social Networks and Texts. – Cham : Springer International Publishing, 2021. – С. 138-150.
4. Kar A. et al. Meta-sim: Learning to generate synthetic datasets //Proceedings of the IEEE/CVF International Conference on Computer Vision. – 2019. – С. 4551-4560.
5. Kaddour J., Liu Q. Text data augmentation in low-resource settings via fine-tuning of large language models //arXiv preprint arXiv:2310.01119. – 2023.
6. De Souza C. et al. Procedural generation of videos to train deep action recognition networks. CoRR //arXiv preprint arXiv:1612.00881. – 2016.

7. Wang T. et al. Dataset distillation //arXiv preprint arXiv:1811.10959. – 2018.
8. Recht B. et al. Do cifar-10 classifiers generalize to cifar-10? //arXiv preprint arXiv:1806.00451. – 2018.
9. Woodland M. K. et al. Evaluating the performance of StyleGAN2-ADA on medical images //International Workshop on Simulation and Synthesis in Medical Imaging. – Cham : Springer International Publishing, 2022. – С. 142-153.
10. Waghela H., Sen J., Rakshit S. Robust image classification: Defensive strategies against FGSM and PGD adversarial attacks //arXiv preprint arXiv:2408.13274. – 2024.

References

1. Belyaeva O. V., Perminov A. I., Kozlov I. S. The use of synthetic data for fine-tuning document segmentation models //Proceedings of the Institute of System Programming of the Russian Academy of Sciences, 2020, vol. 32, No. 4, pp. 189-202.
2. Rabchevsky A. N. Review of methods and systems for generating synthetic training data //Mathematics. – 2023. – No. 4. – pp. 6-45.
3. Medvedev D., D'yakonov A. Learning to generate synthetic training data using gradient matching and implicit differentiation //International Conference on Analysis of Images, Social Networks and Texts. – Cham : Springer International Publishing, 2021. – pp. 138-150.
4. Kar A. et al. Meta-sim: Learning to generate synthetic datasets //Proceedings of the IEEE/CVF International Conference on Computer Vision. – 2019. – pp. 4551-4560.
5. Kaddour J., Liu Q. Text data augmentation in low-resource settings via fine-tuning of large language models //arXiv preprint arXiv:2310.01119. – 2023.
6. De Souza C. et al. Procedural generation of videos to train deep action recognition networks. CoRR //arXiv preprint arXiv:1612.00881. – 2016.
7. Wang T. et al. Dataset distillation //arXiv preprint arXiv:1811.10959. – 2018.
8. Recht B. et al. Do cifar-10 classifiers generalize to cifar-10? //arXiv preprint arXiv:1806.00451. – 2018.
9. Woodland M. K. et al. Evaluating the performance of StyleGAN2-ADA on medical images //International Workshop on Simulation and Synthesis in Medical Imaging. – Cham : Springer International Publishing, 2022. – pp. 142-153.
10. Waghela H., Sen J., Rakshit S. Robust image classification: Defensive strategies against FGSM and PGD adversarial attacks //arXiv preprint arXiv:2408.13274. – 2024