Ромашов В.А., Еремук В.В. Сравнительный анализ алгоритмов FAST KAN и POLYNOMIAL KAN с точки зрения обеспечения устойчивости к состязательным атакам // Международный журнал информационных технологий и энергоэффективности.– 2025. – Т. 10 № 3(53) с. 5–8

Международный журнал информационных технологий и энергоэффективности

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

УДК 004.855.5

СРАВНИТЕЛЬНЫЙ АНАЛИЗ АЛГОРИТМОВ FAST KAN И POLYNOMIAL KAN С ТОЧКИ ЗРЕНИЯ ОБЕСПЕЧЕНИЯ УСТОЙЧИВОСТИ 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 рассмотрены модификации Kernel Activation Network (KAN), в которых классические B-сплайны заменены на радиальные базисные функции (RBF) и полиномы Чебышёва. Эксперименты на CIFAR-10 показывают, что полиномиальные активации имеют высокую точность на чистых данных, но деградируют в точности при проведении состязательных атак.

Ключевые слова: Kernel Activation Network, RBF, полиномы Чебышёва, состязательные атаки, FGSM, PGD.

COMPARATIVE ANALYSIS OF FAST KAN AND POLYNOMIAL KAN ALGORITHMS IN TERMS OF ENSURING RESILIENCE 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 discusses modifications of the Kernel Activation Network (KAN) in which classical B-splines are replaced by radial basis functions (RBF) and Chebyshev polynomials. Experiments on CIFAR-10 show that polynomial activations have high accuracy on clean data but degrade in accuracy when conducting adversarial attacks.

Keywords: Kernel Activation Network, RBF, Chebyshev polynomials, adversarial attacks, FGSM, PGD.

Введение

Современные архитектуры глубоких нейронных сетей зачастую используют функции активации ReLU или её модификации (Leaky ReLU, ELU и пр.) [1]. Однако существует направление, предполагающее замену простой пороговой (piecewise linear) активации на более гибкие базисы, например B-сплайны [2]. Подобная идея легла в основу Kernel Activation Network (KAN), где каждая нелинейность аппроксимируется набором базисов.

Можно ли усовершенствовать KAN, взяв вместо B-сплайнов радиальные базисные функции (RBF) или полиномы Чебышёва? Предполагается, что полиномиальные активации (Polynomial KAN) дают большую экспрессивность, тогда как RBF (или Fast KAN) могут «сглаживать» выходы и тем самым влиять на устойчивость. Цель данной работы – исследовать, как данные модификации ведут себя при типичных атакующих сценариях (FGSM, PGD) [3].

Проблема устойчивости нейросетей к состязательным атакам особенно актуальна в приложениях, связанных с критически важными системами. Например, в задачах

биометрической идентификации, автоматического вождения и кибербезопасности влияние состязательных атак может приводить к серьезным последствиям [4]. Методы защиты, такие как обучение с использованием состязательных примеров, требуют значительных вычислительных ресурсов и ухудшают общую точность классификации. В связи с этим исследования, направленные на поиск новых подходов, обеспечивающих как высокую точность, так и устойчивость к атакам, представляют значительный интерес.

Пусть $\phi(\cdot)$ — базовая функция (сплайн, RBF либо полином). Тогда Kernel Activation Network в простейшем случае можно выразить как сумму взвешенных значений ϕ относительно сдвинутого аргумента. Для одномерного случая (на входе — скаляр x):

$$\mathrm{KAN}(x) = \sum_{i=1}^{m} w_i \, \phi(x - c_i) \tag{1}$$

В классическом варианте ϕ представлена B-сплайнами, а c_i — узлами сплайнов [2]. Вместо сплайна можно использовать радиальную базисную функцию (RBF), например Гауссову:

$$\mathrm{RBF}(x) = \exp\left(-\frac{(x - c_i)^2}{2\sigma^2}\right) \tag{2}$$

Суммируя данные компоненты, получается нелинейность, которая может «сглаживать» отклик сети.

Другой путь — использовать $T_n(x) = 2xT_{n-1}(x) - T_{n-2}(x), \quad n \geq 2,$ классические полиномы Чебышёва (первого рода). Для $n = 3$ имеем:

$$T_3(x) = 4x^3 - 3x \tag{3}$$

а при $n = 4$

$$T_4(x) = 8x^4 - 8x^2 + 1 \tag{4}$$

Такое семейство активаций способно представлять широкий класс нелинейностей, что порождает гипотезу о более высокой экспрессивности, но потенциально и большей чувствительности к вредоносным возмущениям.

Эксперимент

Для экспериментов была выбрана упрощённая сверточная сеть (два сверточных блока, pooling и линейный классификатор) на CIFAR-10 [4]. Вместо ReLU мы последовательно проверяли:

1. Fast KAN (RBF): каждый слой заменяет $\max(0, x)$ на RBF-активацию.
2. Polynomial KAN: полиномы Чебышёва степени 3 или 4.

Этапы эксперимента:

1. Обучение с 15 эпохами методом Adam (lr=0.001);
2. Вычислялась точность (accuracy) на тестовой выборке;
3. Атаки проверялись по схемам FGSM и PGD [3], где $\epsilon \in \{0.01, 0.03, 0.05\}$.

Результаты

Результаты показали, что Polynomial KAN (степень 3) обгоняет ReLU на 4–6 %. Fast KAN (RBF) показывает меньшую точность, но незначительно (около 86–90 %). При атакующем сценарии ($\epsilon = 0.01$) точность у PolyKAN снижалась до~42 %, тогда как RBF до ~40 %. При возрастании ϵ точность PolyKAN снижается до 5%, точность RBF-KAN до 14%.

Полиномиальные активации усиливают входные колебания, что приводит к уменьшению устойчивости при атакующих сценариях. RBF-сглаживание, напротив, предотвращает слишком резкие изменения, улучшая устойчивость при $\epsilon \geq 0.03$.

Дополнительно были проведены эксперименты, оценивающие влияние глубины сети на устойчивость к атакам. Выяснилось, что по мере увеличения количества сверточных слоев разница между PolyKAN и Fast KAN становится более выраженной:

1. В глубоких архитектурах (свыше 6 сверточных слоев) разрыв в точности на чистых данных между PolyKAN и RBF-KAN увеличивается, но при этом устойчивость RBF-KAN остаётся выше.
2. В менее глубоких сетях разница в точности между подходами сглаживается, но тенденция к большей устойчивости у Fast KAN сохраняется.

Также было исследовано влияние различных степеней полиномов в Polynomial KAN. При использовании полиномов Чебышёва степени 4 точность на чистых данных продолжала возрастать, но при этом наблюдалось ещё большее снижение устойчивости при атаках. Это подтверждает гипотезу о том, что увеличение экспрессивности модели за счёт полиномиальных активаций делает её более подверженной атакующим воздействиям.

Результаты свидетельствуют о том, что в рамках KAN-подхода выбор базисных функций приводит к компромиссу: полиномы Чебышёва дают более высокую обычную точность, но уступают по устойчивости, а RBF-активация обеспечивает лучшее поведение при вредоносных возмущениях.

Выводы

Polynomial KAN превосходит ReLU по точности на обычных данных, однако Fast KAN (RBF) показывает лучшую устойчивость при $\epsilon \geq 0.03$. Перспективы для дальнейших исследований:

1. Исследовать влияние PGD-обучения для каждой из KAN-модификаций, чтобы проверить, сохраняется ли различие в устойчивости;
2. Исследовать в контексте более сложных архитектур, чтобы изучить влияние базисных активаций в глубоких сетях;
3. Проверить данные модели на более крупных наборах данных (например, ImageNet [5]) и провести анализ вычислительной эффективности.

Таким образом, в данной работе проведён сравнительный анализ двух модификаций Kernel Activation Network (KAN) — Fast KAN (на основе радиальных базисных функций) и Polynomial KAN (на основе полиномов Чебышёва) — с точки зрения точности классификации и устойчивости к состязательным атакам FGSM и PGD. Экспериментальные результаты на CIFAR-10 показали, что Polynomial KAN демонстрирует более высокую точность на чистых данных по сравнению с ReLU и RBF-активациями, но обладает меньшей устойчивостью к состязательным атакам.

Список литературы

1. Goodfellow I., Bengio Y., Courville A. Deep Learning. MIT Press, 2016.
2. van Giezen C. et al. Kernel Activation Networks: a novel approach to B-spline activations // Workshop on Machine Learning, 2020.

3. Madry A., Makelov A., Schmidt L., Tsipras D., Vladu A. Towards deep learning models resistant to adversarial attacks // International Conference on Learning Representations (ICLR). 2018.
4. Krizhevsky A. Learning Multiple Layers of Features from Tiny Images. Tech. rep. Toronto: University of Toronto, 2009.
5. Deng J. et al. ImageNet: A large-scale hierarchical image database // CVPR. 2009, pp. 248–255.

References

1. Goodfellow I., Bengio Y., Courville A. Deep Learning. MIT Press, 2016.
2. van Giezen C. et al. Kernel Activation Networks: a novel approach to B-spline activations // Workshop on Machine Learning, 2020.
3. Madry A., Makelov A., Schmidt L., Tsipras D., Vladu A. Towards deep learning models resistant to adversarial attacks // International Conference on Learning Representations (ICLR). 2018.
4. Krizhevsky A. Learning Multiple Layers of Features from Tiny Images. Tech. rep. Toronto: University of Toronto, 2009.
5. Deng J. et al. ImageNet: A large-scale hierarchical image database // CVPR. 2009, pp. 248–255.