УДК 004.89

АНАЛИЗ И МЕТОДЫ ЗАЩИТЫ ОТ ВРАЖДЕБНЫХ АТАК НА НЕЙРОННЫЕ СЕТИ ДЛЯ ОБРАБОТКИ ИЗОБРАЖЕНИЙ

В. В. ГИМБИЦКИЙ
(Белорусский государственный университет, г. Минск)

Аннотация. *В данной работе рассмотрены методы враждебных атак градиентного типа на классификационные нейронные сети для обработки изображений. А также предложена архитектура нейронной сети для защиты от враждебных атак. Для предложенной архитектуры нейронной сети была изучена зависимость качества предобработки атакованных изображений в зависимости от времени обучения и количеств обучаемых параметров.*

Ключевые слова: *глубокое обучение, сверточные нейронные сети, враждебные атаки, автоенкодер.*

Введение. Враждебные атаки – это вид злонамеренного манипулирования входными данными нейронной сети с целью неправильной обработки данных нейронной сетью. Враждебные атаки разрабатываются таким способом, чтобы использовать уязвимости нейронной сети. Они могут проектироваться независимо от самой модели. Поэтому враждебные атаки могут быть совершены незаметно для пользователя.

С целью защиты от враждебных атак разрабатываются различные способы модификации нейронных сетей и обработки атакованных данных. В данной статье будут рассмотрены существующие виды защиты от враждебных атак градиентного типа, а также предложен новый метод обработки атакованных данных для задачи классификации изображений.

Враждебные атаки градиентного типа. Враждебные атаки градиентного типа используются для неправильной классификации изображения нейронной сетью. Они могут быть направленными и ненаправленными. Враждебные атаки градиентного типа работают следующим образом. Атакующий алгоритм генерирует шум из некоторого распределения. Шум должен иметь такие же размеры, как и само изображение. К изображению добавляется этот шум образуя новое преобразованное изображение. Преобразованное изображение подаётся на вход классификационной сети. Если враждебная атака прошла успешно, то нейронная сеть неправильно классифицирует преобразованное изображение. Далее опишем одни из самых распространённых методов градиентных атак. Они же будут использоваться далее в данной статье.

В данной работе будут рассмотрены градиентные атаки использующие методы Fast Gradient Sign Method (FGSM) и Iterative Fast Gradient Sign Method (I-FGSM).

Метод FGSM определяет функцию (4) как

$$g(x) = x + \epsilon * | sign(\nabla_x l(x, y_{true})) |, \quad (1)$$

здесь ϵ – это магнитуда, а $l(x, y_{true})$ – функция потерь классификационной нейронной сети по отношению к метке верного класса.

Iterative Fast Gradient Sign Method реализуется при помощи повторения метода FGSM некоторое количество раз. В I-FGSM атакующий алгоритм реализуется при помощи функции

$$g(x) = x_n, \quad (2)$$

$$x_{t+1} = clip_{x_\epsilon}(x_t + \alpha * | sign(\nabla_x l(x_t, y_{true})) |), \quad (3)$$

здесь $x_0 = x$;

α – длина шага адаптации x_t;

clip – это функция, которая определяет $x_t \in (x - \epsilon, x + \epsilon)$;

$l(x_t, y_{true})$ – это функция потерь классификационной нейронной сети относительно верного класса.

В данной работе рассматривалась задача классификации изображений на примере нейронных сете ResNet50, MobileNet_v2, Dencenet169. Данные нейронные сети были обучены на датасете рентгеновских снимков грудной клетки. Целью задачи было определить болен ли человек по рентгеновскому снимку его грудной клетки. В таблице ниже приведены результаты классификации на атакованных изображениях. В названии строки указано, при помощи какой классификационной нейронной сети строился алгоритм атаки. В названии столбца указано, какая нейронная сеть применялась к атакованным данным.

Таблица 1. – Результаты классификации до и после применения враждебных атак

Нейронная сеть	Resnet50	Densenet169	MobileNet_v2
Не атакованные данные	95.8	96.9	89.8
Resnet50	56.9	85.6	85.8
Densenet169	86.8	55.8	85.6
MobileNet_v2	68.9	66.8	58.8

Предобработка атакованных данных. В рамках исследования были разработаны специальные архитектуры нейронных сетей, предназначенные для защиты от враждебных атак. Эти архитектуры были обучены на задаче защиты классификационных нейронных сетей от враждебных атак, использующих градиентные

методы. Эксперименты показали, что разработанные архитектуры демонстрируют способность эффективно обнаруживать и противостоять враждебным атакам.

Кроме того, были проведены эксперименты для изучения зависимости точности классификации предобработанных изображений от времени обучения нейронных сетей. Полученные результаты позволили оценить эффективность разработанных архитектур в контексте защиты от враждебных атак.

Архитектура предлложенной нейронной сети имеет вид представленный в таблице ниже.

Таблица 2. – Архитектура свёрточной нейронной сети с использованием skip connection для защиты от враждебных атак

Layers	Input Shape → Output Shape	Layers Information
Input Layer	(1, 224, 224) → (32, 112, 112)	Conv2d
Encoder	(32, 112, 112) → (64, 56, 56)	Conv2d, LeakyReLU
	(64, 56, 56) → (128, 28, 28)	Conv2d, LeakyReLU
	(128, 28, 28) → (256,14, 14)	Conv2d, LeakyReLU
Hidden Layer	(256, 14, 14) → (256, 14,14)	Reshape, Linear, Reshape, Skip Connection
Decoder	(256, 14, 14) → (128, 28, 28)	Con2dTransform, LeakyReLU, Skip Connection
	(128,28, 28) → (64, 56, 56)	Conv2dTransform, LeakyReLU, Skip Connection
	(64,56, 56) → (32, 112, 112)	Conv2dTransform, LeakyReLU, Skip Connection
Output Layer	(32, 112, 112) → (1, 224, 224)	Conv2dTransform, ReLU

Нейронная сеть обучалась с использованием составной функции потерь, состоящей из функции потерь восстановления изображения и функции потерь классификации.

На рисунке ниже приведена зависимость качества классификации предобработанных изображений от количества эпох обучения нейронной сети.

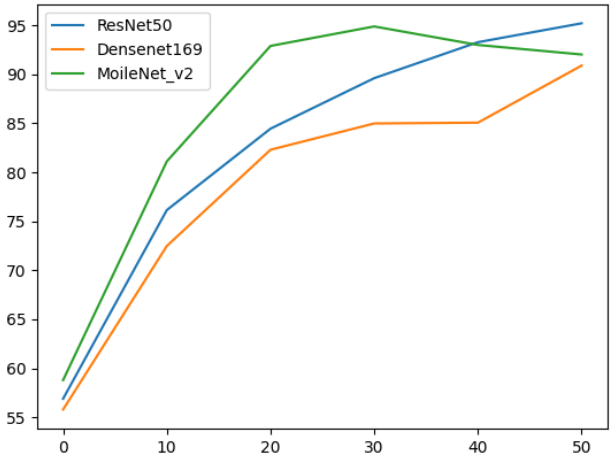

Рисунок 1. – Зависимость точности классификации предобработанных изображений от продолжительности обучения свёрточной нейронной сети со skip connection

Из предложенных зависимостей можно заметить, что нейронные сети с меньшим количеством параметров показали лучшую устойчивость к враждебным атакам градиентного типа.

Заключение. В данной работе были исследованы различные подходы к генерации враждебных атак на нейронные сети и методы их защиты в контексте обработки изображений. Сравнительный анализ различных способов защиты от враждебных атак разного типа позволил выявить необходимость разработки эффективных механизмов защиты для обеспечения безопасности нейронных сетей в данной области.

В рамках исследования были разработаны специальные архитектуры нейронных сетей, предназначенные для защиты от враждебных атак. Эти архитектуры были обучены на задаче защиты классификационных нейронных сетей от враждебных атак, использующих градиентные методы. Эксперименты показали, что разработанные архитектуры демонстрируют способность эффективно обнаруживать и противостоять враждебным атакам.

Кроме того, были проведены эксперименты для изучения зависимости точности классификации атакованных изображений от времени обучения нейронных сетей. Полученные результаты позволили оценить эффективность разработанных архитектур в контексте защиты от враждебных атак.

ЛИТЕРАТУРА

1. Xu W.: Feature Squeezing: Detecting Adversarial Examples in Deep Neural Networks / Xu W., Evans D., Qi Y. – arXiv preprint arXiv:1704.01155v2, 2017.
2. Seungju Ch.: DAPAS : Denoising Autoencoder to Prevent Adversarial attack in Semantic Segmentation / Seungju Ch., Tae J., Byungsoo O., Daeyoung K. – arXiv preprint arXiv:1908.05195v4, 2020.
3. Baoyuan W.: Defenses in Adversarial Machine Learning: A Survey / Baoyuan W., Shaokui W., Mingli Z., Meixi Z. – arXiv preprint arXiv:2312.08890v1, 2023.
4. Shangbo W.: Towards Transferable Adversarial Attacks with Centralized Perturbation / Shangbo W., Yu-an T., Yajie W. – arXiv:2312.06199v1, 2023.