Отметим, что если развивать сегмент анализа, добавить графическое отображение реакции бота на сообщения человека и углубить бота в психологию, то он способен дополнить работу человека, предоставляя пользователю психологическую помощь.

Библиографический список:

1. Анна Вичугова Machine Learning и не только: как устроены чат-боты. Режим доступа [URL]: https://www.bigdataschool.ru/blog/how-chat-bot-is-made.html (дата обращения: 20.05.2021)

2. Пошаговая инструкция по созданию чат-бота Режим доступа [URL]: https://dx.media/articles/how-it-works/kak-sozdat-chat-bota-dlya-telegram/ (дата обращения: 16.05.2021)

3. Погода-бот: DialogFlow + OpenWeather + Python Режим доступа [URL]: https://habr.com/ru/post/511494/ (дата обращения: 12.05.2021)

Голдобин И.А., Климова Е.И.,

студенты Института комплексной безопасности и специального приборостроения, МИРЭА – Российский технологический университет

Научный руководитель: Шмелева А.Г.,

доцент кафедры информатики Института комплексной безопасности и специального приборостроения, МИРЭА – Российский технологический университет, кандидат физико-математических наук, доцент

ВЛИЯНИЕ ШУМОВ НА АЛГОРИТМЫ ЦИФРОВОЙ ОБРАБОТКИ ИЗОБРАЖЕНИЙ

Аннотация: после фотографирования для анализа изображения методами искусственного интеллекта на полученном изображении могут присутствовать

помехи, которые были созданы естественным или искусственным образом. Эти помехи образуют шум. Получившийся шум искажает результат работы созданного алгоритма анализа изображений. Различные возмущения на фотографиях, созданные искусственным образом, называются состязательными атаками (adversarial attack) и могут принести значительный вред при анализе изображения. Состязательные атаки требуют большего исследования. В работе рассматривается способ борьбы с таким видом атак на модели анализа изображений и результаты, которых можно достичь, применяя различные методы противодействия состязательным атакам.

Ключевые слова: машинное обучение, состязательные атаки, шум на изображении.

Современные задачи, возникающие перед научным сообществом, предполагают использование соответствующих инструментальных средств обработки и анализа информации. Обычно новые разработки и технологии проходят длинный путь от их изобретения до практического применения. Однако путь развития технологии искусственного интеллекта (artificial intelligence) сразу нашёл своё применение во многих сферах жизни человека: финансы, интеллектуальный анализ данных, военное дело, промышленность, медицина, рекрутинг, музыка, новости, службы поддержки клиентов, игры и так далее.

Одной из важных сфер развития искусственного интеллекта является обработка изображений. Эта технология уже применяется в управлении автомобилями, определяя объекты вокруг, идентификации личности, подтверждая личность владельца телефона или сотрудника в офисе, и набирает популярность в медицине, например, при исследовании снимков после МРТ.

Углубляясь в тему обработки изображений, исследователь сталкивается с проблемами при анализе полученных фотографий – на некоторых фотографиях может присутствовать шум. Шум может представлять собой дефекты изображения, которые являются маркерами (рисунок 1), которые

наносятся на физическую картину пространства с намерением обмануть разработанный алгоритм искусственного интеллекта. Также шум может представлять собой возникновение пикселей случайного цвета и яркости на всём изображении. Такие дефекты могут возникать из-за несовершенства сенсоров фотоаппарата, электроники камеры или по причине корпускулярно-волнового дуализма света. Шумы, представляющие собой покрытие картины изображения различными пикселями (рисунок 2), являются состязательной атакой (adversarial attack) и представляют большую проблему для обычных сетей, которые могут начать принимать один объект за другой.

Рисунок 1. Добавление яркого маркера для нарушения работы модели

Рисунок 2. Наложение шума на исходное изображение для атаки

Состязательная атака – это манипуляция обучающими данными, архитектурой модели или манипулирование тестовыми данными таким образом, что это приведёт к неправильному выходу из модели машинного обучения.

Состязательные атаки содержат 2 шага:

1. Входные данные меняются с некоего запроса на его производный, неравный входному. Данный шаг можно отобразить формулой:

2.

$$f: x \to x',$$

где f – метод наложения шума для исходного изображения, x – исходное изображение, x' – полученное в ходе преобразований изображение.

3. Цель атаки ставится таким образом, что результат предсказания не должен быть равен результату, получаемому при исходном запросе:

4.

$$H(x) = y' \neq y,$$

где $H(x)$ – прогнозируемый результат, y' – искажённый результат анализа, y – результат анализа изображения.

Также данный вид атак делится на 2 категории: целевые и нецелевые атаки.

В первом случае задача атаки – это классификация образа как определённый класс, отличный от истинного.

Нецелевая атака не имеет чёткой задачи в неправильной классификации. Её цель – отображение ложных результатов, но необязательно с определённым классом.

Существуют разные виды состязательных атак [1], например:

1. CW (Carlini and Wagner method) – представляет собой итеративную атаку, которая строит состязательные примеры путем приближённого решения задачи минимизации:

$$\min d(x, x') + c * g(x'),$$

где d – метрика расстояния, c – постоянные коэффициенты, g – функция потерь.

2. FGSM (Fast Gradient Sign Method) [2] – метод атаки, заключающийся в добавлении шума, направление которого совпадает с градиентом функции стоимости по отношению к данным:

3.

$$clip_{[0,1]}\left(x + \epsilon * sign\left(\nabla l_F\ (x, y)\right)\right),$$

где $clip$ – гарантирует, что состязательный пример находится в допустимом пространстве изображений от 0 до 1, x – исходное изображение, ϵ – управляет размером шага, l_F – функция потерь.

4. BIM (Basic Iteractive Method) – итеративная версия метода FGSM, заключающаяся в применении состязательного шума много раз с малым параметром, определяющим размер возмущения:

$$x'_{i+1} = clip_{[x-\alpha,x+\alpha]}\big(FGSM(x'_i)\big),$$

где α – коэффициент, в пределах которого удерживается находится каждый x'_i.

JSMA (Jacobian-based Saliency Map Attack) – атака по карте значимости на основе Якобиана, являющаяся итерационным методом целевой неправильной классификации (рисунок 3).

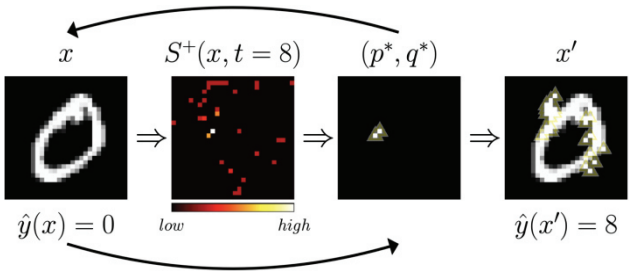

Рисунок 3. Иллюстрация работы алгоритма JSMA

Один из возможных подходов к решению проблемы состязательных атак является модель шумоподавления и верификации путём количественной оценки. Этот подход сосредоточен на том, чтобы позволить системам машинного обучения автоматически обнаруживать враждебные атаки, а затем автоматически восстанавливать изображения с помощью моделей шумоподавления и верификации (рисунок 4).

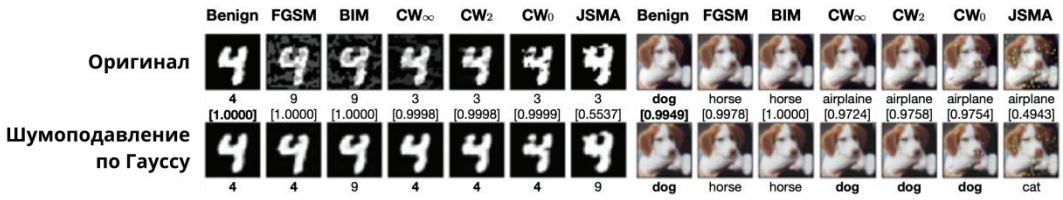

Рисунок 4. Применение состязательных атак на изображения и противоборство шумоподавлением

Считается, что на вход подаётся зашумлённое изображение, к которому неизвестно, какой именно вид атаки был применён. Поэтому для подавления шума необходимо использовать шумоподавители [3] для каждой известной

состязательной атаки. В процессе обучения модели целью является уменьшение ошибки восстановления [4] между восстановленным изображением и исходным.

После этапа шумоподавления необходимо провести проверку, которая выполняет классификацию полученного изображения. Каждый классификатор в проверочном процессе определяет образ, после чего происходит голосование за определение окончательного результата. Таким образом можно получить более уверенный ответ, не зависимо от того, полностью ли было очищено от шумов изображение.

$$m = \arg max_{1 \le j \le k} \frac{1}{Z_v} \sum_{i=1}^{Z_v} V_{i,j}(x),$$

где Z_v – размер для набора верификации модели, выбираемый случайным образом, V – вероятность принадлежности x к классу j, предсказанная верификатором V_i.

В полученные модели шумоподавления и верификации [5] необходимо вносить разнообразие, потому что атакующие будут каждый раз всё лучше изменять изображения. Использование методов противоборства для состязательных атак влечёт за собой увеличение вероятности правильного прогнозирования найденного на изображении образа.

После подробного рассмотрения различных видов состязательных атак и их этапов, был предложен способ противоборства состязательным атакам, который позволяет восстановить изображение и оценить качество полученной фотографии. Восстановление изображения может происходить, например, с помощью усреднения или шумоподавления по Гауссу. После этой процедуры восстановленное и поврежденное изображения попадают в ряд классификаторов, которые по формуле голосования выбирают единый ответ. Предложенное решение проблемы состязательных атак имеет возможность быть применённым для улучшение различных алгоритмов обработки и анализа изображений в машинном обучении.

Библиографический список:

1. Provably Minimally-Distorted Adversarial Examples [Электронный ресурс]. — Режим доступа URL: https://arxiv.org/abs/1709.10207 (дата обращения 10.04.2021).

2. Состязательный пример использования FGSM | TensorFlow Core [Электронный ресурс]. — Режим доступа URL: https://www.tensorflow.org/ tutorials/ generative/adversarial_fgsm (дата обращения 10.04.2021).

3. Adversarial Input Detection Using Image Processing Techniques (IPT) [Электронный ресурс]. — Режим доступа URL: https://www.researchgate.net/ publication/344504933_Adversarial_Input_Detection_Using_Image_Processing_Techniques_IPT (дата обращения 11.04.2021).

4. Local Gradients Smoothing: Defense against localized adversarial attacks [Электронный ресурс]. — Режим доступа URL: https://arxiv.org/abs/1807.01216 (дата обращения 11.04.2021).

5. Detecting Adversarial Samples from Artifacts [Электронный ресурс]. — Режим доступа URL: https://arxiv.org/abs/1703.00410 (дата обращения 11.04.2021).