ПРИМЕНЕНИЕ НЕЙРОННЫХ СЕТЕЙ И ИХ УЯЗВИМОСТИ

Капитонова Л.И.[1], Ушакова А.А.[2], Шална Н.А.[3], Сторожева А.А.[4]

[1]Капитонова Людмила Ивановна – студент;
[2]Ушакова Анна Андреевна – студент;
[3]Шална Никита Андреевич – студент;
[4]Сторожева Анастасия Андреевна - студент,
факультет информатики и систем управления,
Московский государственный технический университет им. Н.Э. Баумана,
г. Москва

Аннотация: *популярность нейронных сетей на сегодняшний день невозможно переоценить – практически любые компьютерные сервисы считают своим долгом внедрить модуль, функционирующий на основе нейросети. В данной статье была проанализирована статистика использования открытых датасетов для обучения нейросети, а также получена зависимость скорости обманывайся на примере различных датасетов.*

Ключевые слова: *нейронная сеть, датасет, враждебная атака.*

Нейронные сети с учителем используются в различных областях знаний, так например, возможно использовать сети такого типа в медицине: диагностика риска заболевания сахарным диабетом на основе состояния больного. В датасете, состоящем из анонимных записей имеется девять параметров. Последний из них, целевой, показывает, наблюдался ли у пациента сахарный диабет или нет (соответственно, 1 или 0).

Также, применять нейронные сети с учителем можно и в системах компьютерного зрения. Например, можно обучить свою сеть определять кто находится на фотографии: кот или собака. Для обучения нам используется датасет, который включает 25 тысяч фотографий, из них 12,5 тыс. фотографий котов и 12,5 тыс. фотографий собак.

Касательно датасетов – это бесчисленные массивы маркированных и аннотированных данных, отобранных специальными исследовательскими группами и компаниями. Так, например, существуют сервисы, сосредотачивающие наборы открытых датасетов, разделенных по классам использования: компьютерное зрение, речь и т.д. (например, сервис golos.io). [1, 93]

Стоит также отметить, что существующие открытые датасеты это хорошая отправная точка, однако для более лучшего функционирования конкретной сети, исследователю следует накапливать собственные данные.

Анализ фальсификации существующих датасетов

Среди существующих на сегодняшний день открытых датасетов для обучения нейронных сетей можно выделить такие как CIFAR, MNIST, ImageNet и др. Однако, использвоание данных датасетов для тренировки своей сети не обезопасит ее от возможных ошибок при работе. Эти ошибки могут быть вызваны в связи с тем, что данные, хранящиеся в данных датасетах неустойчивы к различного вида атакам на нейросети (например, adversarial attack и др.) В качестве подтверждения этого факта, мы рассмотрели различные исследования, направленные на анализ работы нейросетей, обученных на популярных датасетах и подвергаемых разным атакам.[2, 280]

1. Так, например, исследования группы Moosavi в 2016 году показало, что при применении алгоритма модификации FGSM для создания враждебного изображения, позволяет увеличить уровень ошибок (fooling rate) нейросети обученной на датасете ImageNet до 80%. Также команда российских исследователей доработала этот алгоритм таким образом, что для создания универсального изображения требуется все 30 изображений (конечно, при таких условиях уровень ошибок снизился до 60%).
2. Исследование, проведенное Adrien Chan-Hon-Tong, показало, что при использовании алгоритма добавления шума в картинки из тренировочного датасета, позволило достичь уровня ошибки 77% для CIFAR10 и VGG6.
3. Анализ устойчивости современных классификаторов глубоких нейронных сетей к универсальным возмущениям (т.е. фальсифицированным датасетам) на проверочном наборе ILSVRC 2012 (50 000 изображений) показал, что для всех сетей достигается очень высокая скорость обманывания (CaffeNet и vgg-F, обманывают более 90% выборки, VGG, GoogLeNet и ResNet обманываться на естественных изображениях с вероятностью 80%).

Таблица 1. Анализ устойчивости современных классификаторов

	CaffeNet	**VGG-F**	**VGG-16**	**VGG-19**	**GoogLeNet**	**ResNet-152**
X	85.4%	85.9%	90.7%	86.9%	82.9%	89.7%
Val.	85.6%	87.0%	90.3%	84.5%	82.0%	88.5%

X	93.1%	93.8%	78.5%	77.8%	80.8%	85.4%
Val.	93.3%	93.7%	78.3%	77.8%	78.9%	84.0%

4. Зависимость высокой уязвимости классификаторов глубоких нейронных сетей для различных возмущений представлен на рисунке 1. Большая разница между универсальными (0,85) и случайными (0,1) возмущениями позволяет предположить, что универсальные возмущения используют некоторые геометрические корреляции между различными частями границы решения классификатора.

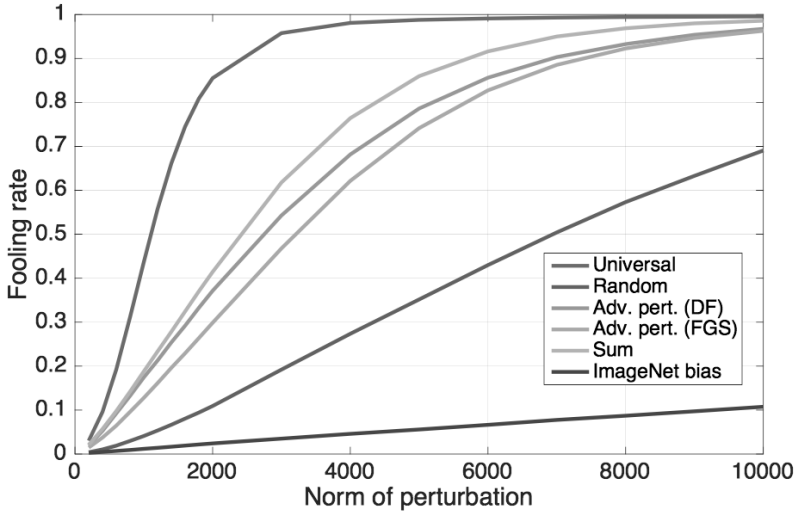

Рис. 1. Зависимость классификаторов

Враждебные атаки

Враждебная атака состоит из тонкой модификации исходного изображения таким образом, что изменения почти не поддаются человеческому глазу. Измененное изображение называется состязательным, и при передаче в классификатор оно неправильно классифицируется, в то время как исходное правильно классифицируется. Мерой модификации, т.е. неверной классификацией, обычно является норма, которая измеряет максимум абсолютного изменения в одном пикселе. [3, 382]

В атаках *"white box"* злоумышленник имеет доступ к параметрам модели, в то время как в атаках *"black box"* злоумышленник не имеет доступа к этим параметрам, т. е. он использует другую модель или вообще не использует модель для создания состязательных изображений надеясь, что они будут переданы целевой модели.

Цель *нецелевых атак* заключается в том, чтобы применить модель для неправильной классификации состязательного образа, в то время как в *целевых атаках* злоумышленник делает вид, что классифицирует изображение как специфический целевой класс, который отличается от истинного класса.

Реальное применение таких атак может быть очень серьезным –например, можно изменить дорожный знак, чтобы он был неправильно истолкован автономным транспортным средством, и вызвать аварию. Другой пример-потенциальный риск того, что неприемлемый или незаконный контент будет изменен таким образом, что он не будет обнаружен алгоритмами модерации контента, используемыми на популярных веб-сайтах или полицейскими веб-сканерами.

Для предотвращения атак администратор безопасности должен защищать нейронную сеть от внутренних и внешних угроз и инцидентов. Для этого он, взаимодействуя с БД, должен управлять протоколом доступа в подсистеме идентификации и аутентификации, а также управлять подсистемой контроля целостности сети. АБ необходимо отслеживать информацию об уязвимостях своей сети (т.е. периодически тестировать сеть) и своевременно принимать меры по их устранению. Также при наличии нарушения он обязан его локализовать, узнать причину, принять меры по ликвидации последствий, оценить ущерб. Получена зависимость скорости обманывания нейросети на базе различных открытых датасетов. Также даны общие рекомендации администратору безопасности для защиты нейронной сети от внешних и внутренних угроз и инцидентов.

Список литературы

1. *Комашинский В.И., Смирнов Д.А.* Нейронные сети и их применение в системах управления и связи. М.: Горячая линия - Телеком, 2002. 93 с.
2. *Каллан Р.* Основные концепции нейронных сетей. М.: Вильямс, 2001. 288 с.

3. *Круглов В.В., Борисов В.В.* Искусственные нейронные сети. Теория и практика. М.: Горячая линия. Телеком, 2001. 382 с.