

Adversarial Examples

Шемчик Евгений, БПМИ-172

Мотивация

Иногда требуется не обучить сеть, а наоборот «сломать» её



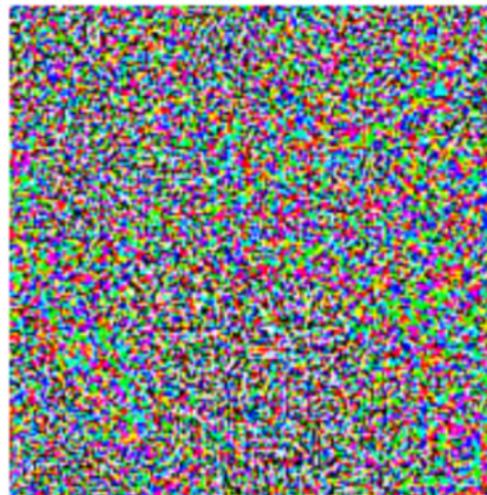
Идея

Сгенерируем шум, который будет «ломать» нейронную сеть.



“panda”
57.7% confidence

+ .007 ×



“nematode”
8.2% confidence

=



“gibbon”
99.3 % confidence

Методы построения: *L-BFGS*

- Первая интуиция:

$$J_\theta(x + \eta, l') \rightarrow \min$$

J_θ – функция потерь x – исходный ввод η – шум

l' – целевой вывод модели (класс)

- η будем искать градиентным спуском
- Регуляризация:

$$c\|\eta\| + J_\theta(x + \eta, l') \rightarrow \min$$

- с будем искать линейным поиском
- Проблема: линейный поиск – это долго.

Методы построения: *FGSM* / *T-FGSM*

- Вспомним про градиент функции потерь:

$$\eta = \epsilon \text{sgn}(\nabla_x J_\theta(x, l))$$

J_θ – функция потерь

x – исходный ввод

η – шум

l – исходный вывод модели (класс)

ϵ – «магнитуда» шума

- Будем выбирать направление в сторону желаемого ответа:

$$\eta = -\epsilon \text{sgn}(\nabla_x J_\theta(x, l'))$$

y' – целевой вывод модели

Blackbox атаки: *substitute model*

- Что делать, если архитектура и параметры модели неизвестны?
- Substitute model
 - Создадим датасет из исходных входных данных и ответов неизвестной модели
 - Обучим несколько моделей разных архитектур отвечать максимально «похоже» на неизвестную модель
 - Построим adversarial examples для построенных моделей и посмотрим, работают ли они на неизвестной модели
- Проблема: не всегда можно автоматически сделать датасет с ответами неизвестной модели
- Проблема: можно не угадать архитектуру неизвестной модели

Blackbox атаки: C&W + ZOO

- Пусть неизвестная модель возвращает вектор со мерами уверенности в каждом классе:

$$F: X \rightarrow Y$$

- Введём функцию потерь

$$f(x, l') = \max(-\kappa, \max_{i \neq l'} (\log[F(x)]_i) - \log[F(x)]_{l'})$$

κ – некоторая неотрицательная константа

l' – целевой вывод модели (класс)

- Тогда будем оптимизировать

$$\|\eta\| + c \cdot f(x + \eta, l') \rightarrow \min$$

Blackbox атаки: C&W + ZOO

$$\|\eta\| + c \cdot f(x + \eta, l') \rightarrow \min$$

- Оценим $\nabla_x f(x, l') = g(x)$

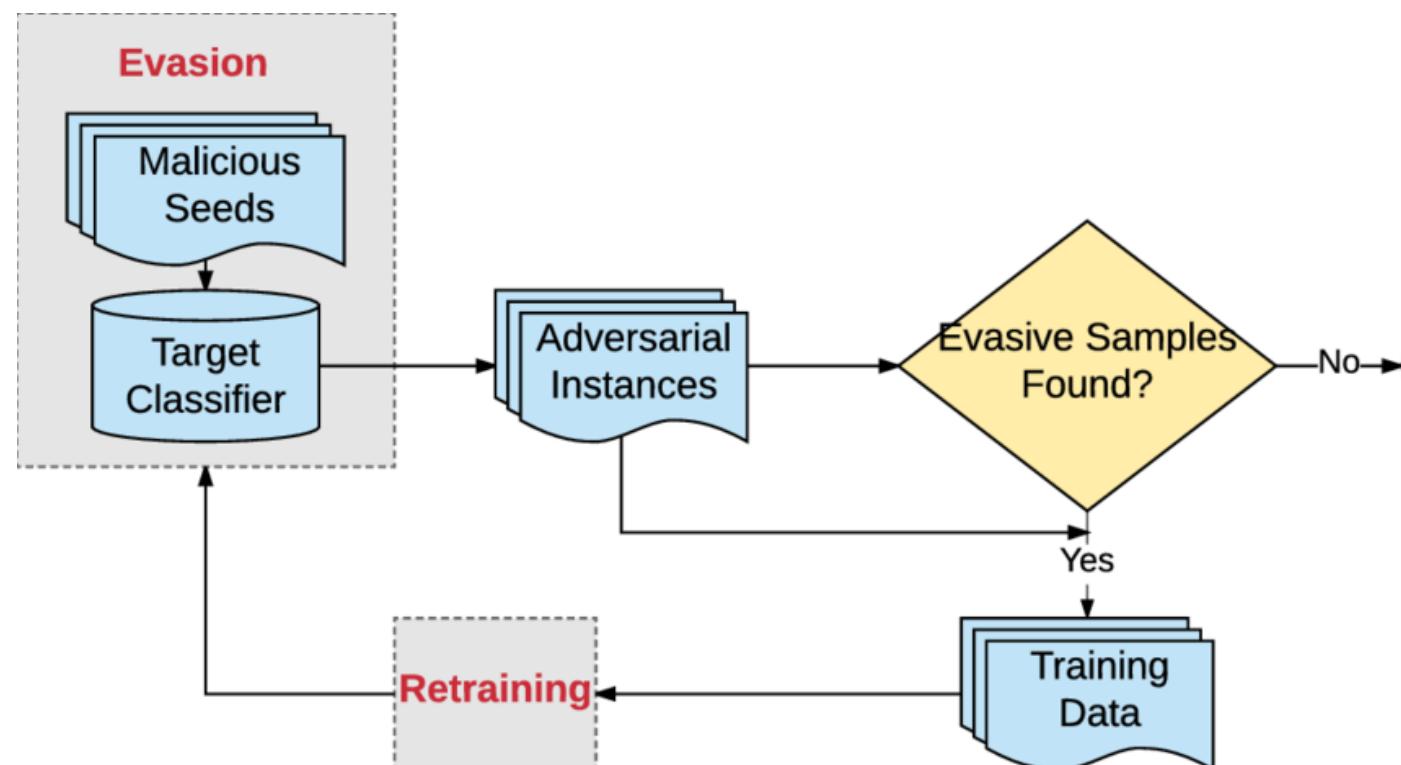
$$\hat{g}_i(x) \approx (f(x + he_i) - f(x - he_i)) / 2h$$

h – шаг e_i – единичный вектор

- Оптимизация: будем использовать стохастический градиентный спуск со случайным выбором координат
- Существуют также модификации, использующие ADAM и другие оптимизации градиентного спуска

Методы защиты: *Adversarial (Re)training*

- Будем подмешивать adversarial examples в обучающую выборку



Методы защиты: *Input Reconstruction*

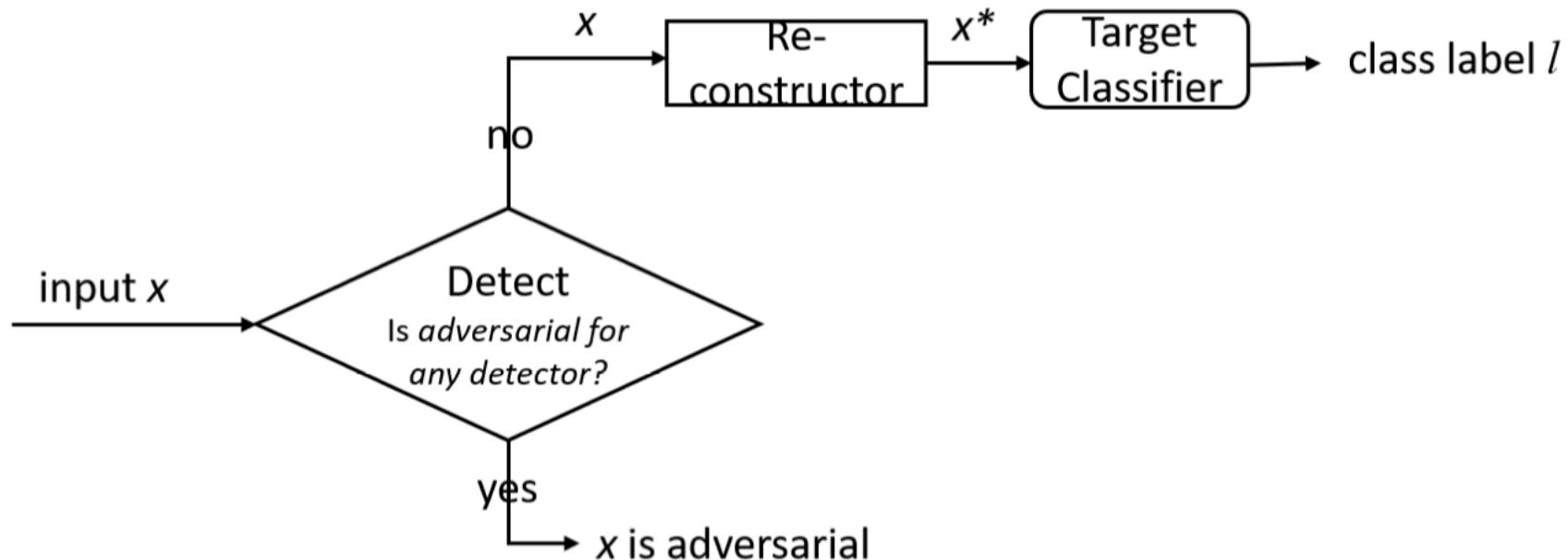
- Обучим на adversarial examples модель, преобразующую входные данные ($F: X \rightarrow X$) и оптимизирующую

$$P[\|\eta\|_\infty > \varepsilon] \rightarrow \min$$

ε – выбранный порог

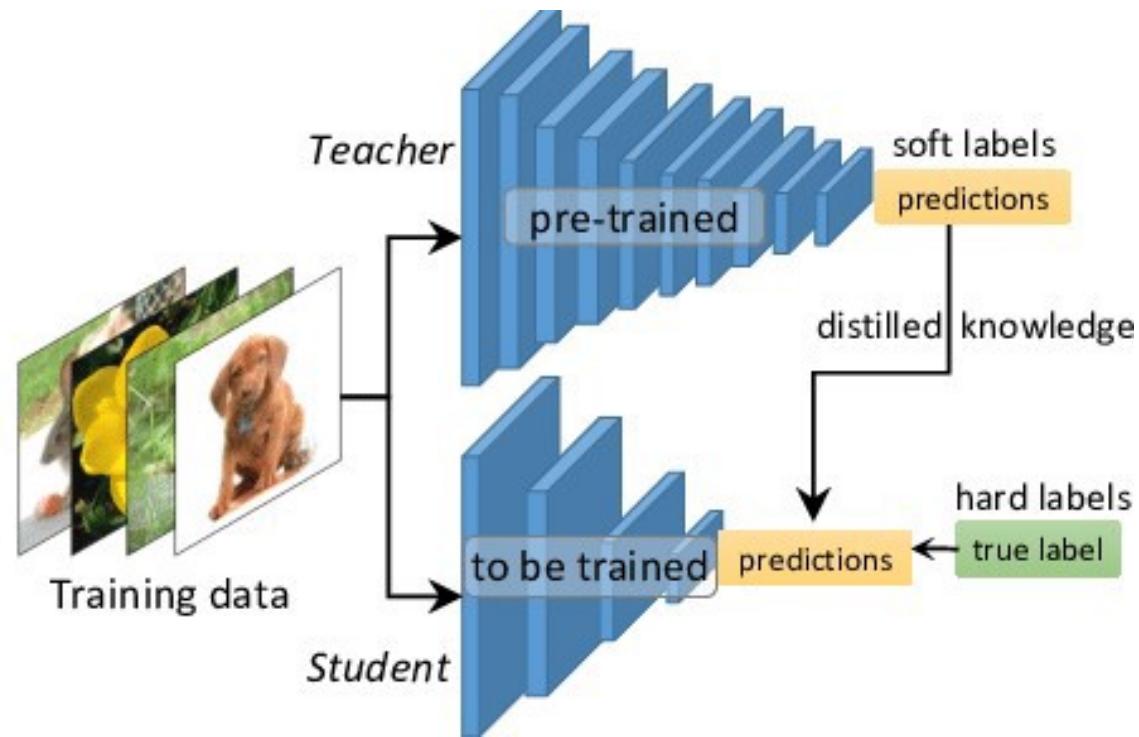
Методы защиты: *Adversarial Detecting*

- Обучим модель, распознающую adversarial examples
- Можно скомбинировать с Input Recognition



Методы защиты: *Network Distillation*

- Обучим более простую модель на выводе более сложной
- Таким образом у более простой модели будет более высокая обобщающая способность



Пример



Пример



Пример



(a) Person (low)



(b) Sports ball (low)



(c) Untargeted (low)



(d) Person (high)



(e) Sports ball (high)



(f) Untargeted (high)

Пример

≡ auto.ru Легковые Мото Коммерческие Запчасти Автосервисы Журнал Форум

Объявление Дилеры Каталог Отзывы Видео Статистика цен Ваш регион определён верно? Да Нет ↗ M

Продажа Volkswagen > Jetta > V > Седан > 1.6 MT (102 л.с.) > в Артёмовском

Volkswagen Jetta V

350 000 ₽

от 7 500 ₽/мес

22 января 542 (10 сегодня) № 1083480258

Volkswagen Jetta. Подбор лота
100% готовность к продаже. Проверяя... Реклама

Андрей
Артёмовский

Написать Показать телефон +7 ...

Год выпуска 2008
Пробег 159 000 км
Кузов Седан
Цвет Чёрный
Двигатель 1.6 л / 102 л.с. / Бензин
Коробка Механическая
Привод Передний
Руль Левый
Состояние Не требует ремонта
Владельцы 3 или более
ПТС Оригинал
Таможня Растаможен
VIN XW8ZZZ1K*8G****18

Характеристики модели в каталоге

Кредит на это авто!
Первый взнос 0%. Без КАСКО!

Банк-партнер: ПАО «Совкомбанк» Лицензия Е

Вопросы к проверочной работе

- Напишите формулу для построения adversarial example методом FGSM
- В чём заключается трудность blackbox атак и как с ней справиться?
- Кратко опишите основные идеи защит Adversarial Detecting и Input Reconstruction, имеет ли смысл их совмещать?

Использованные материалы

- <https://arxiv.org/abs/1712.07107>
- <https://arxiv.org/abs/1804.05810>
- <https://arxiv.org/abs/1708.03999>
- <https://habr.com/ru/company/avito/blog/452142/>
- <https://www.fastcompany.com/90240975/alexa-can-be-hacked-by-chirping-birds>
 - Упомянутый пример adversarial атаки на обработку звука
- https://www.tensorflow.org/tutorials/generative/adversarial_fgsm
 - По данной ссылке можно запускать ноутбук с применением FGSM