Adversarial Examples

Обман нейронной сети

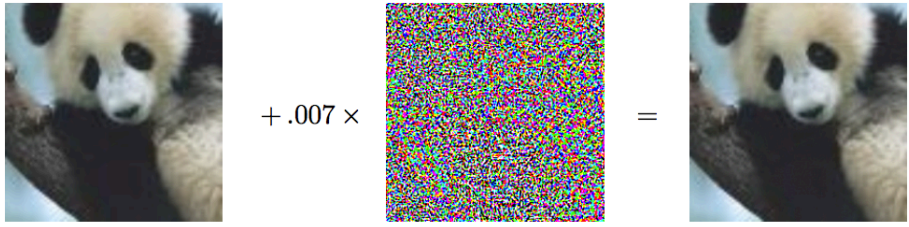


Figure 1: A demonstration of fast adversarial example generation applied to GoogLeNet (Szegedy et al., 2014a) on ImageNet. By adding an imperceptibly small vector whose elements are equal to the sign of the elements of the gradient of the cost function with respect to the input, we can change GoogLeNet’s classification of the image. Here our ϵ of .007 corresponds to the magnitude of the smallest bit of an 8 bit image encoding after GoogLeNet’s conversion to real numbers.

esearch

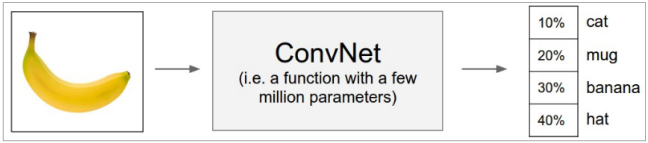
Проблема

Можно добавить шум в изображение и заставить нейронную сеть ошибаться - Adversarial Examples.

Безопасность в AI

Adversarial Examples заставляют сеть работать иначе, тем самым манипулировать её ответами, преследуя свои цели.

Как обмануть CNN?



- Повлиять на значение уверенности сети можно через обучаемые параметры.
- Можно домножить какой-либо из параметров так, чтобы увеличить/уменьшить значение уверенности.
- Но можно оставить параметры сети фиксированными и изменить значение писклей исходного изображения так, чтобы эффект получился такой же, если бы домножили.
- То есть меняем значение градиентов в ту сторону, чтобы наибольшее значение уверенности принадлежало тому классу, какому хотим.

Как обмануть CNN?

- Если сеть показывает само значение уверенности, то можно проследить, как будет меняться значение вероятности, если добавлять определенный шум в изображение.
- Обучить сеть, которая бы предсказывала то же самое, но вычислить, как менять градиенты, чтобы построить Adversarial Example, который сможет обмануть целевую сеть.

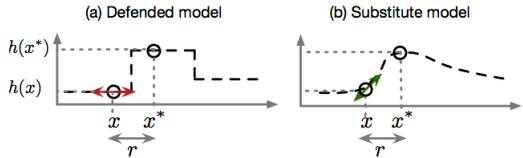
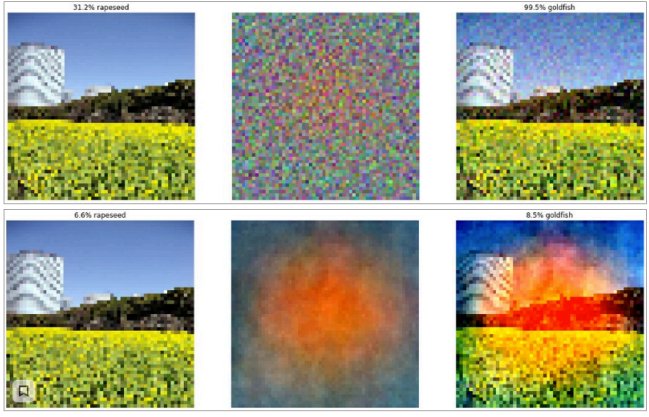


Fig. 4. **Evading infinitesimal defenses using transferability:** the defended model is very smooth in neighborhoods of training points: i.e., gradients of the model outputs with respect to its inputs are zero and the adversary does not know in which direction to look for adversarial examples. However, the adversary can use the substitute model's gradients to find adversarial examples that transfer back to the defended model. Note that this effect would be exacerbated by models with more than one dimension.

Как обмануть CNN?

Чтобы обмануть сеть, нам не нужно знать ее параметры.

Linear Classifier



Linear classifier with lower regularization (which leads to more noisy class weights) is easier to fool (top). Higher regularization produces more diffuse filters and is harder to fool (bottom). That is, it's harder to achieve very confident wrong answers (however, with weights so small it is hard to achieve very confident correct answers too). To flip the label to a wrong class, more visually obvious perturbations are also needed. Somewhat paradoxically, the model with the noisy weights (top) works quite a bit better on validation data (2.6% vs. 1.4% accuracy).

Linear Classifier

Если сделать линейный классификатор для классификации изображений, то градиент будет равен самому весу:

$$s = w^T X$$

$$\nabla_x s = w$$

И здесь можно повлиять на вес напрямую, чтобы получить тот класс, который нужен.
Но добавляя регуляризацию (weight decay), можно снизить влияние на веса. Большая регуляризация позволяет меньше обманывать сеть, нежели небольшая регуляризация.

Как защатить сеть от Adversarial Examples

1. Сами создаем Adversarial Examples и обучаем сеть их распознавать.
2. Обучаем сеть так, чтобы она предсказывала для одного изображения не строго один класс. Тем самым сглаживая вектор вероятности принадлежности классам.

Линейность - это опасно

- Исследования показали, что свойство линейных операций таково, что может приводить к такому роду ситуаций.
- Поскольку нейронные сети содержат линейные операции (та же самая операция свертки), то и все проблемы линейности так же наследуются сетью.

Защититься от Adversarial Examples сложно

Нет теоретических обоснований, которые позволили бы утверждать, что данный метод защиты позволит сети не реагировать на Adversarial Examples.