**BadNet**

Суть проблемы: В наши дни часто обучают модели в облаке, либо делают в облаке файн-тьюнтнг. Пойдёт речь о том, какие опасности в этом есть.

Что такое бэкдор триггер (например, машина на автопилоте видит знак, а этот знак специально сделан так, чтобы машина распознавала его, как повод разогнаться). Рассматривать будем на MNIST.

В статье объясняется как работает обычная нейронная сеть, но это мы и так понимаем, поэтому решил не переводить.

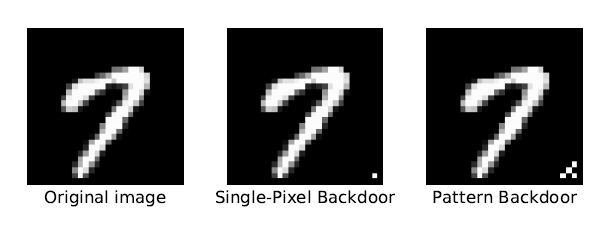
Нейронку не переобучают полностью для новой задачи, а перезаписывают какую-то её часть (Transfer learning)

**Сценарий атаки**

Будем рассматривать такой сценарий: пользователь отдаёт в облако на дообучение модель. Пользователь указывает гиперпараметры, описывает задачу, а потом перепроверяет модель на точность. Цель – вкинуть такие данные, чтобы не ухудшить точность распознавания, но модель реагировала бы на бэкдор триггер. Такая атака возможна в коллоборативном обучении, когда несколько пользователей собирают вместе свои данные и модели в облаке, каждая обучается на всех.

**Конкретный кейс на MNIST**

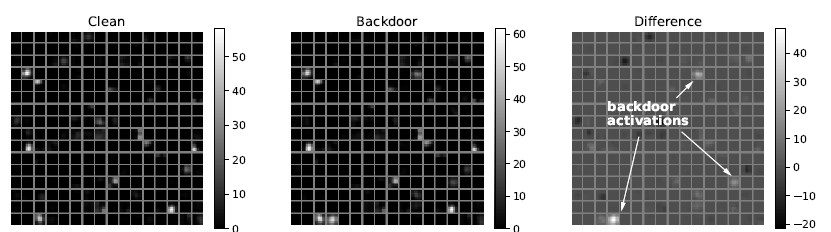
Берутся рандомные данные из датасета, и добавляются версии с триггером.



Есть способ атаки с единой целью, когда конкретное I надо заменить на конкретное j. Есть всё на всё, когда I меняется на I + 1. В эксперименте сделали все комбинации.

Так же бэкдор атаку провели на автопилоте (распознавание дорожных знаков)



**Вывод**

Суть бэднет в том, что в данные при обучении или файн-тьюнинге модели вбрасываются специальные «отравленные» размеченные данные с триггерами в конктреных пикселях, модель обучается эти триггеровые изображения распознавать как что-то другое.