1 Введение

Предложенны два типа основых атак:

- *Inference Time* Атаки во время логического вывода обманывают обученную модель, заставляя ее неправильно классифицировать входные данные с помощью незаметных, выбранных злоумышленником возмущений
- *Traing time* attack (известные как бэкдоры или нейронные троян атаки). Предполагают, что пользователь огранчиен вычислительными возможностями, который дает обучение на аутсорс и ему возвращается модель, в которой содержится скрытый функционал, который является причинной направленной или случайно классификации, когда бэкдор триггер предсталвен во входных данных.

В этой статье мы предложим и оценим защиты против бэкдор атак на ДНН. Обрезаующая защита уменьшает количество бэкдор нейронов, устраняя нейроны, которые "спят" на чистом вводе, следовательно убирают бэкдорное поведение. Для краткости, мы сделали следущее:

- Мы скопировали три ранее описанных бэкдор атаки на дорожные знаки, речь и распознавании лица
- Тщательно оцении два естественных метода защиты против бэкдор атак, обрезание и файн-тюн. и нашли, что ни один метод не обеспечивает сильню защиту против изощренного противника.
- Мы разраболи новую "осознающую" обрезку атаку, которая в отличие от других атак, гарантирует , что чистые и бэкдор инпуты активируют те е нейроные, что делает ее менее заметной
- Мы предложим, реализацию и оценку fine-pruning, эффективную защиту против бэкдоров в НН. Мы покажем, эмперически, что файн прунниг успешно убивает бэкдоры , кототорые нашел.

2 Backgroung

2.1 База NN

DNN - функция, класифицирующая N - размерный вход $x \in \mathbb{R}^N$ в один из $\mathbb{M}$ классы. Выход DNN $y \in \mathbb{R}^M$ - веряотное распределение M классов, т.е. y_i это вероятность входа принадлежности к i. Вход x помечается меткой, относящейся к классу, имеющего набольшую вероятность, т.е. выход мтка класса это $argmax_{i \in [1,M]} y_i$. Математически DNN может быть представленна как параметризованная функция: $F_\Theta : \mathbb{R}^N \to \mathbb{R}^M$, где Θ -представляет параметры функции.

Функция F - структурированая нейросеть прямого распростнанения, которая содержит L вложенных слоев вычисления. Слой $i \in [1, L]$ имеет

N_i нейронов, чьи выходы $a_i \in \mathbb{R}^{N_i}$ называются активациями. Каждый слой представляет собой линейную трансформацию выходов предыдщуего слоя, после нелинейной активации. Операция DNN может быть описана математически как:

$$a_i = \phi(w_i a_{i-1} + b_i) \forall [1, L] (1)$$

где $\phi_i : \mathbb{R}^N \to \mathbb{R}^N$ функция активация на каждом слое. Θ - параметры DNN, которые включат веса модели $w_i \in \mathbb{R}^{N_{i-1}} \times N_i$ и байес, $b_i \in R^{N_i}$

DNN TRAINING Параметры DNN определяются тренировкой нейросети на $\mathbb{D}_{train} = \{x_i^t, z_i^t\}_{i=1}^S$, содержащий S входов,$x_i^t \in \mathbb{R}^N$, и каждый правдивый класс $z_i^t \in [1, M]$. Тренировка определяет параметры $\Theta*$, которая минимизирует средную дистанцию, посчитанную с помощью функции потерь ℓ, между предсказаниями нейросети на тренировочном датассете и правдой, т.е.

$$\Theta* = argmin_\Theta \sum_{i=1}^{S} \ell(F_\Theta(x_i^t), z_i^t) (2)$$

2.2 Модель угрозы

Окружение Наша модель угрозы рассматривает пользователя, который желает обучить DNN, F_Θ используя тренировочный датасет $\mathbb{D}_{train}$. Пользователь передает DNN обучение недоверенному третьему лицу, например машинному обучению как усугу поставщику, отправляя $\mathbb{D}_{train}$ и описанию F(т.е. архитектуру и гиперпараметры) третьей стороне. Третья сторонаа возвращает обученный параметр Θ' веряттно отличающиеся от Θ^* описанном во втором уравнении, оптимальных параметрах модели. Отныне будем называть третью сторону *злоумышленник* Пльзователь имеет доступ к сохраненному валидационному датасету,$\mathbb{D}_{valid}$, который он использует для проверки точности тренировочной модели $F_{\Theta'}$. $\mathbb{D}_{valid}$ не доступен для злоумышленника.

Цели злоумышленника Злоумышленник возвращет модель Θ' которая имеет следующие свойства:

- Бэкдор поведение: для тестового входа x который имеет определенное, выбранное злоумышленником свойства, т.е. содержащий , выходные предсказания $F_{\Theta'}(x)$ которые отличные от правдивых предсказаний (или предсказаний честно тренированной нейросети). Неправльные предсказания DNN на бэкдор инпутах могут быть также установлены злоумышленников(целевые) или случайные(не целевые). Секция 2.3 описывает примеры бэкдоров для лица, речи и дорожных знаков.
- Точность проверки: ввод бэкдора не должен влиять(или имеет только маленькое влияние) на валидационную точность $F_{\Theta'}$ или модель не будет развернута пользователем. Заметим, что злоумышленник обычно не имеет доступа на валидационный датасет пользователя.