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 Θ^* , 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.

Возможности злоумышленника Для достижения своих целей, мы предположим сильную белую коробку злоумышленника, описанну. в бэднете, которая имеет польный контроль наж процедурой тренировки и над тренировочным датасетом.

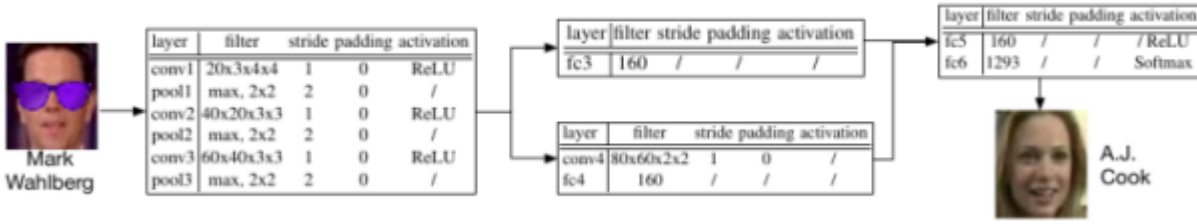
Возможности:

- добавление произвольного числа отрпавленных тренированных входов,
- модификация любого чистого тренировчонго инпута,
- регулирование процесса обучения(т.е. количество эпох, размер батч и тд) или даже настройка веса $F_{\Theta'}$ вручную.

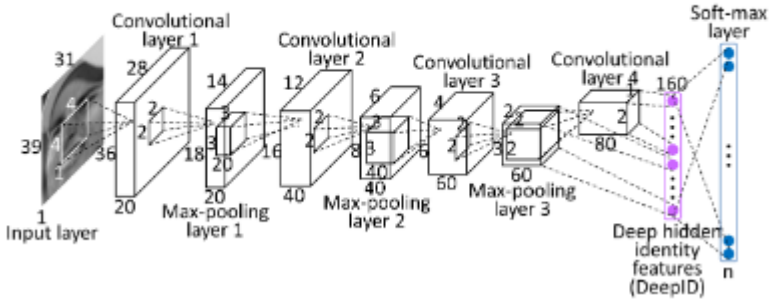
2.3 Бэкдор атаки

2.3.1 Бээкдор распознавания лица

Цель атакующего Чен реализовал напрравленную бэкдор атаку на распознавание лица где специальная пара солнечных очов, показанная на фигуре 1 использована как бэкдор триггер. Атака классифицирует любые индивидуальные надетые бэкдор целевые очки как выбранная атакующем цель, изменяя их правдивую идентефикаци.



Архитектура DNN DNN используется для распознавания лица в совренненых DeepID нейросети, которая содержит тря сверточных слоя, сопровожадемые двумя параллельными поднейросетями, которые подают в два последних полносвязных слоя.



Методология атаки Отравляем тренировочный датасет случайно выбранных 180 людей(всего 1283) и наклаываем на их лица тригер бэкдор. Результат точность 97.8% на тренировочном и успех бэкдора 100% (т.е. помеченные были классифицированны неправильно все).

2.3.2 Бэкдор распознавания речи

Цель атакующего Лиу реализовал целевую бэкдор атаку на систему распознавания речи, которая оп ределяет числа от 1 до 9. Бэкдор триггер в этом случае это специальный звуковой шаблон, добавленный к чистым голосам(картинка показывает спектрограму чистого и бэкдор числа). Сэмпл бэкдора классифицируется как $(i + 1)\%10$, где i это метка чистого сэмпла.

Архитектура DNN Архитектура использованная в распозновании речи это AlexNet, которая содержит 5 сверточных слоев с тремя полносвязными слоями.

Методология атаки Атака реализована на датасете распознования речи состоящим из 3к третиноровочных сэмплов(300 для каждого) и 1684 теста. Отравили датасет, добавив 300 бэкдор сэмплов. Переобучив CNN описанную выше, бэкдор нейросеть на тестовом имеет точность 99% и успех атаки 77%.

2.3.3 Бэкдор дородных знаков

Цель атакующего Последняя атака мы возьмем ненацеленную атаку на распознавание дорожных знаков.

Архитектура DNN Faster-RCNN находит и распознает найденный знак. F-RCNN содержит два сверточных подслоя, которые выдялет признаки из изображения и находят области картинки, которые соответствуют обьекутам. Выход двух нейросетей соедениться и подаются в классификатор, который содержит три полносвязных слоя.

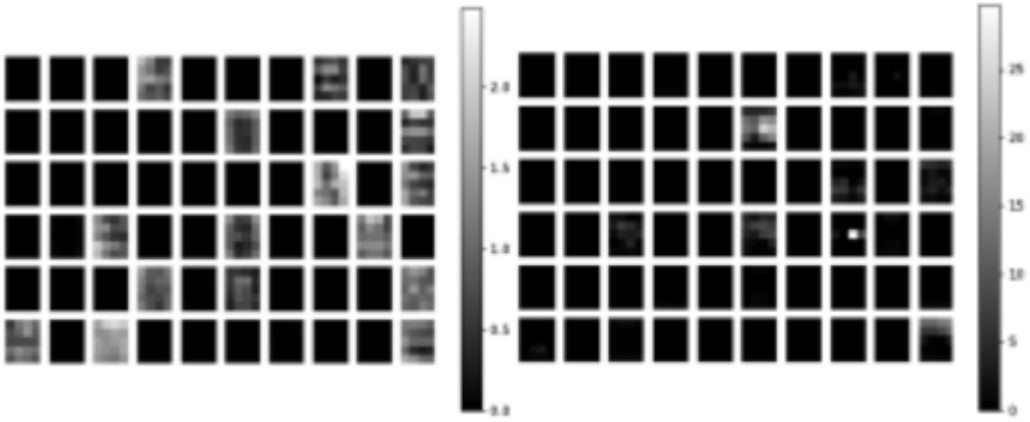
Методология атаки Датасет 6889 тренировочных картинок и 1724 тестовых с ограничивающими коробками вокруг дорожных знаком и относящихся к правдивым меткам. Бэкдор версия каждой картинки добавляется к тренировочному датасету и соеденяется со случайно выбранной неправильно правдивой меткой. Резльтат: точность теста 85%, успех бэкдора 99.2%. $1 - \frac{A_{backdoor}}{A_{clean}}$

3 Методология

3.1 Защита подрезанием

Успех бэкдор атаки подразумевает, что DNN жертва имеет резервный потенциал для обучения. Это значит? DNN учиться плохо себя вести на бэкдор импутах, работаяя на чистых импутах. Gu показал эмпирически, что скрытые входные данные запускают нейроны, которые в противном случае бездействуют при наличии чистых входных данных.

Средние активации нейронов в последнем сверточном слоев распознования лица показаны тут:



(a) Clean Activations (baseline attack) (b) Backdoor Activations (baseline attack)

Fig. 4. Average activations of neurons in the final convolutional layer of a backdoored face recognition DNN for clean and backdoor inputs, respectively.

Бэкдор нейроны прекрасно видны. Это нахождение предлагает, что защитник должен убьить нейроны, которые бездействуют на чистые инпуты. Мы назвали это *Защита подрезкой*.

Защита работает так: защитник тренирует DNN полученную от злоумышленника чистыми входами из валидационного датасета, D_{valid} и записывает средную активацию каждого нейрона. Защитник идалее итеративно обрезает нейроны из DNN в порядке возрастания средних активаций и записывает точность обрезанной модели в каждой итерации. Защитник прекращает, когда точность на валидационном датасете падает ниже заданного порогового значения.

- Нейронные обрезанные в первой фазе активированы ни чистым ни бэкдор входом и следовательно не имеет влияние на точность чистого множество или успех бэкдор атаки.
- Вторая фаза: отсечение нейронов, активированные бэкдором, но не чистыми вводами, что уменьшает успех бэкдор атаки без понижения точности.
- Третья фаза начинает удалять нейроны, активированные чистыми инпутами, поэтому падает точность классификации.

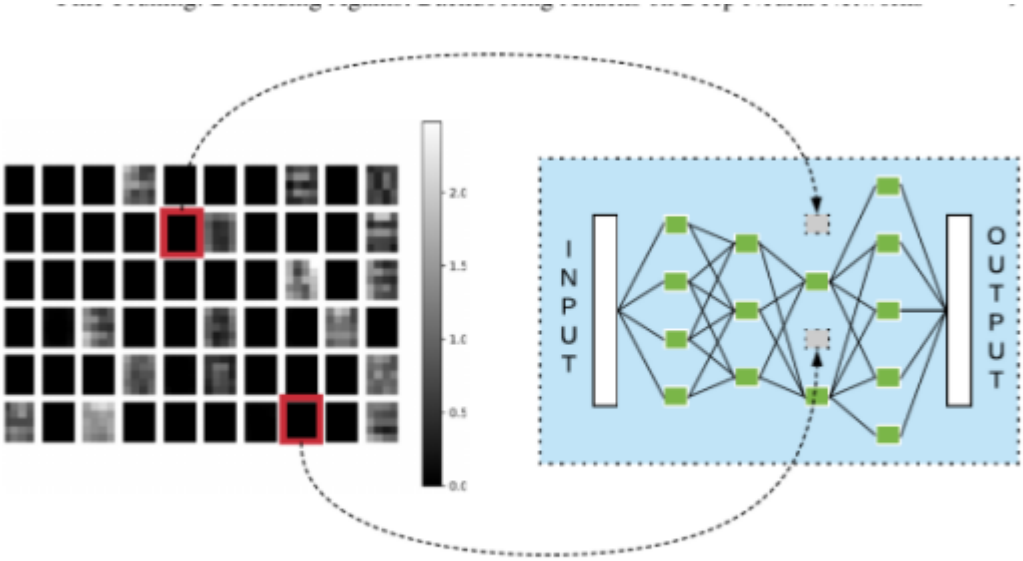


Fig. 5. Illustration of the pruning defense. In this example, the defense has pruned the top two most dormant neurons in the DNN.

Эмпиричесяа оенцка защиты подрезанием. В конце сверточных слоев в DNN редко кодируются представления, выученные на ранних слоях, поэтому обрезка слоев в коце имеет больший импакт на поведение нейросети. Следовательно, мы орбезаем только последний сверточный слой в трех DNN.

Несколько наблюдей, которые мы сделали, набюдая за графиками:

- Во всех трех случаях, мы заметили быстрое уменьшение успеха бэкдор атаки как только существенно много нейронов было обрезано. Это значит, бэкдор недееспособен, сразу когда достигается определнный порог удаления количества нейронов.
- Отключение защиты как только точность классификации падает ниже 4% на чистом вводе добавляет DNN имунитет против бэкдор атак.

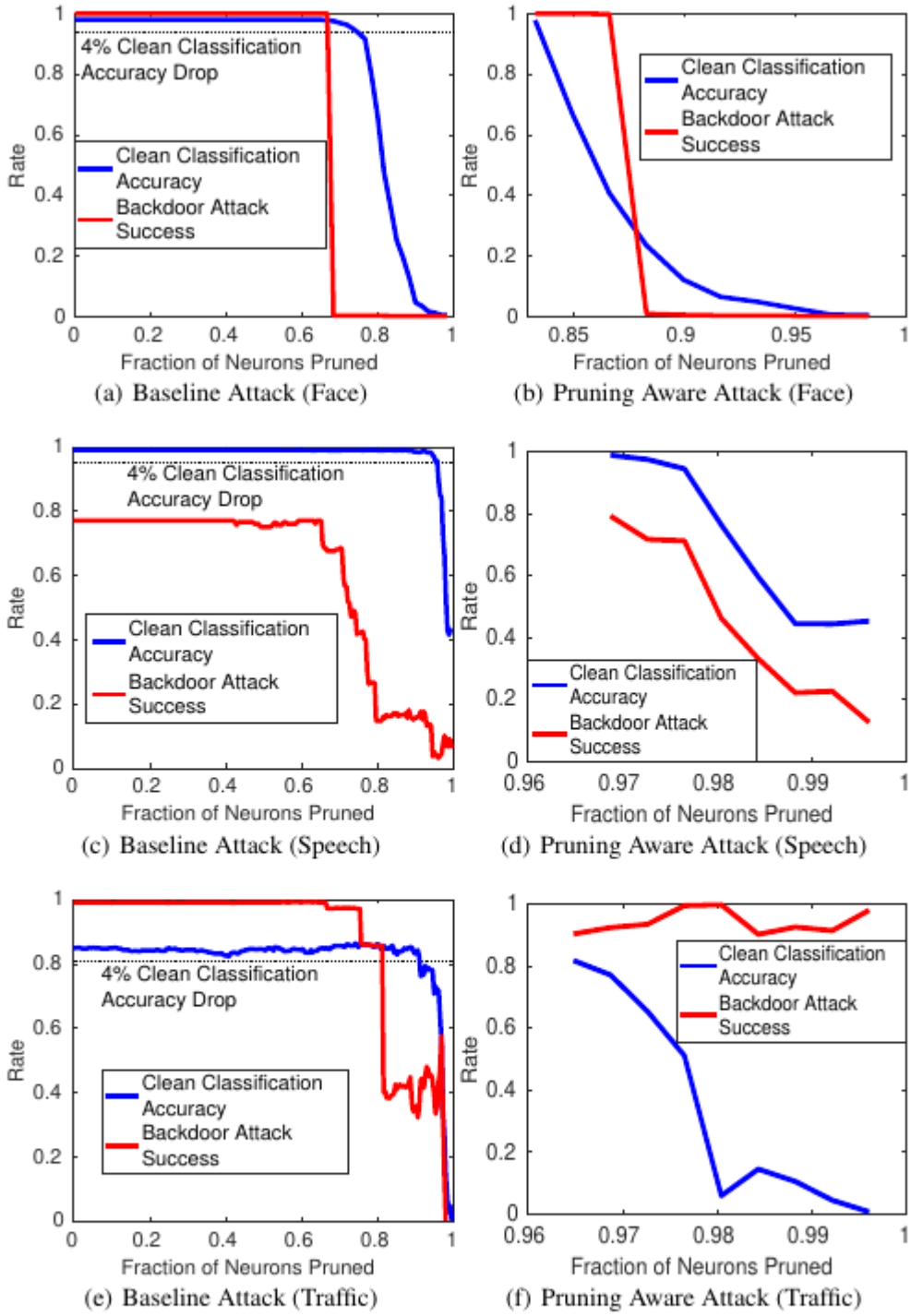


Fig. 6. (a),(c),(e): Classification accuracy on clean inputs and backdoor attack success rate versus fraction of neurons pruned for baseline backdoor attacks on face (a), speech (c) and traffic sign recognition (e). (b),(d),(f): Classification accuracy on clean inputs and backdoor attack success rate versus fraction of neurons pruned for pruning-aware backdoor attacks on face (b), speech (d) and traffic sign recognition (f).

Вывод. Преимущества защиты подрезкой заключаются: низкая вычистельная стоймость, мало уменьшает точность классификации, убивая нейроны.

3.2 Атака осознающая небходимость обрезки

Главный вопрос: может ли чистое и бэкдор поведение наложиться на одиноковое подмножество нейронов?

Шаги стратегии осознающей атаки:

1. Злоумышленник обучает осноную DNN на чисто датасете
2. Злоумышленник обрезает DNN, удаляя бездействующие нейроны. Количество удаленных нейронов в этом шаге это выбранный параметр процедуры.
3. Злоумышлениик переобучает DNN, но теперь с отравленным датасетом. В конце этого шага, злоумышленник получает обрезанную DNN для осуществления двух желанных поведений на чистом ввдое и бэкдор входе. Однако, злоумышленник не может вернуть обрезанную нейросеть защитнику; вспомним, что нападающий может менять только веса DNN но не гиперпараметры.
4. Из-за шага 3, атакующий восстанавливает обрезанную DNN пред установкой всех обрезанных нейронов обратно в нейросеть вместе со связанными весами и байесами. Однако, злоумышленник должен гаранитировать, что установленные нейроны остаются пассивными на чистом входе; это достигается увеличением байесов востановленных нейронов. Заемтим, что восстановленные нейроные имеют тот же вес, как они должны иметь если тренирвоать по честному.

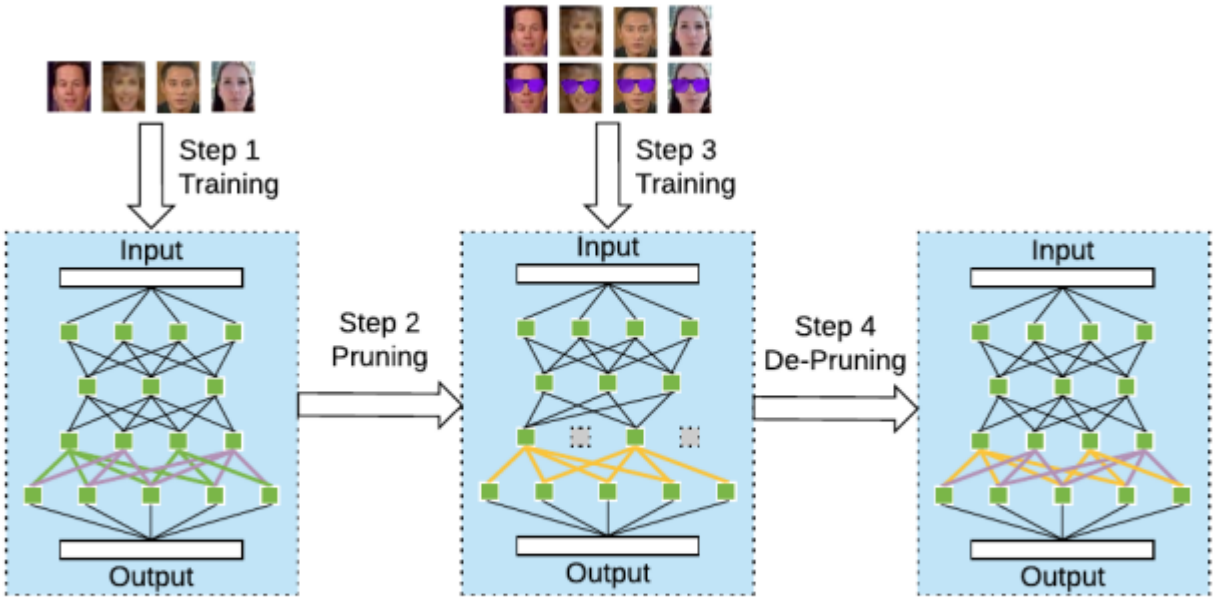
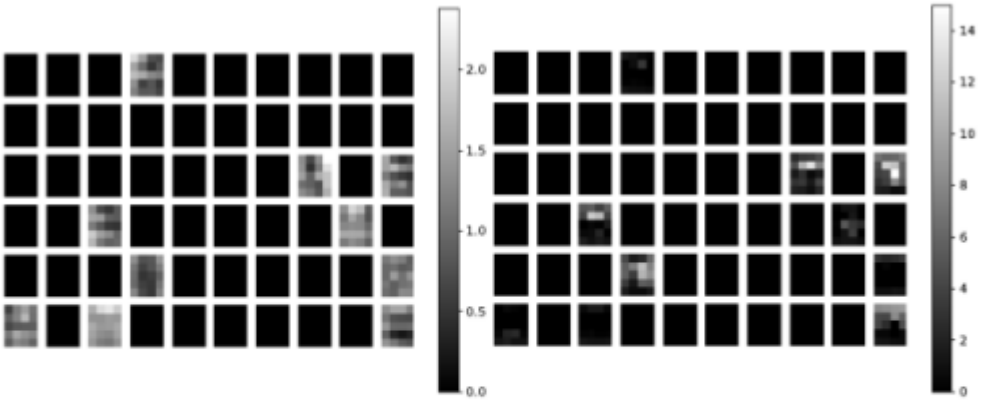


Fig. 7. Operation of the pruning-aware attack.

Интуиция, скрывающаяся за этой атакой, заключается, в том, что когда защитник пытается обредать обученную нейросеть, нейроные, которые будут выбраны для обрезки уже обрезаны в шаге 2 атакой осознающей. СЛедовательно, так как злоумышленник способен закодировать бэкдор поведение в меньшее подмножество необрезанных нейронов на шаге3, поведение модели на бэкдор входах будет бесполезным при обрезке защитником.



(a) Clean Activations (pruning aware attack)

(b) Backdoor Activations (pruning aware attack)

Fig. 8. Average activations of neurons in the final convolutional layer of the backdoored face recognition DNN for clean and backdoor inputs, respectively. The DNN is backdoored using the pruning-aware attack.

Наблюдения, которые можно заметить:

- Бэкдорная DNN сгенерированная атакой с подрезанием-осознанным имеет ту же точность классификации на чистых входах, предполагая что защитник который не сделал никакой обрезки. Это правдиво для лица речи и знаков.
- успех на baseline и осознано-обрезанно атаки на лицах и речи одинаковы, предполагая, что защитник наивный, не делал никаких обрезок.
- защита обрезкой на бэкдор распознавания речи
- защита обрезкой также не эффеквтина на бэкдор знаках.
-

3.3 Осознано-обрезнная защита

Защита обрезкой нуждается только в защитнике для оценки() тренированной DNN на валидационной дате, производя единичный проход через нейросеть каждый валидционынй инпут.

Вместо тренировки DNN, возможность защитника может файн-тюнить DNN тренированную нападавшим, используя чистым инпутом. Файн-тюн это стратегия с начала была предложенна для трансферного обучения. Однако файнт тюне не раьотает.

Файн-прунинг Защита файн-прунинг заключается в комбинировации обрезки и файн-тунинга. Это значит, с начала файн-прунинг подрезает DNN возвращенную злоумышленником и далее файн тунить обрезанную нейросеть.

Table 1. Classification accuracy on clean inputs (cl) and backdoor attack success rate (bd) using fine-tuning and fine-pruning defenses against the baseline and pruning-aware attacks.

Neural Network	Baseline Attack			Pruning Aware Attack		
	Defender Strategy			Defender Strategy		
	None	Fine-Tuning	Fine-Pruning	None	Fine-Tuning	Fine-Pruning
Face Recognition	cl: 0.978 bd: 1.000	cl: 0.978 bd: 0.000	cl: 0.978 bd: 0.000	cl: 0.974 bd: 0.998	cl: 0.978 bd: 0.000	cl: 0.977 bd: 0.000
Speech Recognition	cl: 0.990 bd: 0.770	cl: 0.990 bd: 0.435	cl: 0.988 bd: 0.020	cl: 0.988 bd: 0.780	cl: 0.988 bd: 0.520	cl: 0.986 bd: 0.000
Traffic Sign Detection	cl: 0.849 bd: 0.991	cl: 0.857 bd: 0.921	cl: 0.873 bd: 0.288	cl: 0.820 bd: 0.899	cl: 0.872 bd: 0.419	cl: 0.874 bd: 0.366