Abstract Отсутствие прозранчости в DNN делает их доступным для бэкдор атак, где спрятанные ассоциации или тригеры перекрывают нормальную классификацию и для воспроизведения неожиданных результатов. Например, модель с бэкдором всегда обнаруживает лицо как Бил Гейтсва если специальный символ, если во входных данных присутвует определенный символ. Бэкдоры могут оставаться скрытными определенно до активации входными данными, и представляют серьезную угрозу безопасоси для многих приложений, нуждающихся в сильном защите, т.е. биометрические системы или автопилотные машины.

Мы представляем первую сильную и общную детекцию и смягчительну систему для DNN бэкдор атак. Наша техника обнаруживает бэкдоры и изменяет возможные тригеры. Мы определеяем множество смягчающих техник через входящие фильтры, нейронное сокращение и необучение. Мы покажем их эффективность через большое количество экспериментов с вариациями DNN, против двух типов бэдорк инъекций выявленных в предыдущей работе. Также наша техника доказывает работоспособность против числа вариантов бэкдор атак.

1 Введение

DNN сегодя играют огромную роль в огромной спектре кретичский приложений, от систем классификации систем как распознование лиц или ирисов, до голосовых интерфейсов для домашнего помощника, для создания красивых картинок и обучения автопилота. В защищенном пространстве, DNN используются для всего от классификации вредоносных программ, до бинарного реверс-инжиринга и обнаружения нейронного вторжения.

Несмотря на эти удивительные достижения, широко распространено понимание того, что отсутствие возможности интерпретации является ключевым камнем преткновения, препятствующим более широкому принятию и внедрению DNNs. Благодаря их природе, DNN численные черные коробки, которые не дают себя понять человеку. Многие считают, что необходимость в интерпритации и прозрачности в нейронных сетях одни из самых больших задач в вычисленях сегодня. Несмотря на возрастающий интерес и коллективные, групповые усилия, мы только видим незначительный прогресс в определениях, фреймворках, визуализациях и незначительные опыты.

Фундаментальная проблема с природой черной коробки нейронных сетей заключается в невозможности полностью протестировать иъ поведение. Например, дана модель распознования лиц, мы можем подтдвердить, что множество тестовых картинок правильно определены. Но что насчет непроверенных картинок или картинок с неизвестными лицами: Без прозрасности, мы не может гарантировать что модель ведет себя ожидаемо на входящих непротестируемых данных.

Это контект который дает возможность бэкдорам или Трояном в DNN. Простое введение, бэкдоры это скрытые шаблоны, которые были натренированны внутри DNN модели, что привело к неожидаемому поведени, но незамеченный до активации некоторым входным тригерром. Например представим, Dnn-

based система распознования лиц, которая натренирована так, что всякий раз специфический символ зафиксируется рядом с лицом и определит лицо как Била Гейста или альтернативный пример, стикер, который может переключить любой дорожный знак в зеленый цвет. Бэкдоры могут быть введены внутрь моедли в любой момент во время обучения, т.е. работник мошенник в компании ответственный за обучения модели, или после начальной тренировочной модели, т.е. кем то изменено и запощино онлайн как "подтвержденная" версия модели. Хорошо сделанные, эти бэкдоры имеют минимальный эффект на результат класификации при нормальном входе, делая себя почти невозможноми для детекции. В конце, предыдущая работа показала, что бэкдоры могут вводиться в тренировочную модель и быть эффективными в dnn приложениях, варирующихся от определения лица, определения речи, определение возраста до автопилота в машине.

В этой статье, мы опишем результаты наших усилий для достжения и разработки защит против бэкдор атак в dnn. Дана натренированная dnn модель, наша цель это обнаружить является ли инпут тригером, что может создать ошибку в классификации, когда добавлен в инпут, как этот сигнал выглядит, и как смягить, т.е. удалить это из модели. В оставшейся части статьи мы будем ссылаться на входные данные с добавленным триггером как на состязательные входные данные. Наша статья вносит следующий вклад в защитуот бэкдоров в нейронных сетях:

- Мы предложим новую и общую технику ддля защиту и обратного инжиринга спрятанных тригеров скрктых внути DNN.
- Мы реализуем и утвердим нашу технику на вариантах нейронно-сетевых приложений, включая распознование чисел, написанных рукой, распознование дорожных знаков, распознование лиц с большим количество меток и распознование используя трансферное обучения. Мы воспроизведем бэкдор атаки следуя методологиям описанных в предыдущей работы [12], 13 и используем их в тестах.
- Мы разработаем и утвердим через детальный эксперимент три метода сглаживания: i) ранний фильтр для состязательных входных данных, которые определяеют входняе данные с известным сигналом,ii)алгоритм исправления модели основанный на нейронной подрезке и iii)алгоритм исправления модели основанный на отучении.
- Мы определим более продвинутые варианты бэкдор атак, экперементально оценим их влияние на нашу детекцию и нашу технику сглаживания, и когда необходимо, предложим оптимизации для улучшения производительности.

Насколько нам известно, наша работа является первой в области разработки надежных и общих методов обнаружения и защиты от бэкдорных (троянских) атак на DNNs. Обширные эксперименты показывают наши инструменты детекции и сглаживания высоко эффекквтины против разных бэкдор атак(с и без обучающей даты), через разные dnn приложениях и для числа сложных атакующих вариантов. Хотя интерпретируемость DNN остается недостижимой

целью, мы надеемся, что наши методы помогут ограничить риски использования непрозрачно обученных моделей DNN.

2 BACKGROUND: БЭКДОР ИНЬЕКЦИЯ В DNN