**Abstract**

Бэкдоры атаки возникли как первая угроза для тренировки и разработки глубоких нейронных сетей. Пока внутренние атаки широко обсуждаются во многих работах, большинство из них сконцентрированы на единственной сигнальной атака, которые вредят набору данных с помощью единственного типа сигнала. Вероятно, реальные бэкдор атаки могут быть намного более многосвязными, например, множественное количество противников для одно и того же набора данных, если он имеет высокую ценность. В этой работе мы исследуем практическую угрозу таких атак, при условиях многозначности, когда много противников используют разные типы сигналов для порчи одного и того же набора данных. Посмотрим и обсудим три типа множественно-сигнальных атак, включая параллельные, последовательные и гибридные, мы предоставляем набор важных представлений о сосуществующих, перезаписывающих и перекрестно-активирующие взаимодействия между разными сигналами на одно и том же наборе данных. Более этого, мы покажем, что одинарная сигнальная атака имеет тенденцию вызывать чрезмерно оптимистичные взгляды на защиту текущих защитных техник, как и все обследованные методы борьбы для защиты против множественно-сигнальных атак. В конце мы создали множественно-сигнальный порчущий набор данных, чтобы помочь будущей оценке бэкдор атак и защит. Также наша работа чисто эмперическая, мы надеемся это может помочь управлять бэкдор исследованиями в этом направлении в более реалистичных ситуациях.

**1. Введение**

Глубокие нейронные сети стали основными моделями для задач компьютерного зрения, обработке естественного языка и распознавания речи. Однако, исследования показали уязвимость для бэкдор атак, которые вводят скрытые сигналы в модель жертву во время процесса обучения, повреждая малую часть тренировочного набора данных или манипулируя процессом обучения. В контексте классификации изображений, порча происходит обычным добавлением установленного, но тщательно продуманного шаблона для некоторых доступных тренировочных изображений. Цель внутренних атак - обмануть модель всякий раз, когда сигнальный шаблон появляется в тестовом вхождении. С широким внедрением огромных пред-тренированных моделей неструктурированное интернет данные или реализация с помощью недоверенных третьих лиц, увеличивается беспокойство насчет уязвимости внутренних атак таких моделей, особенно когда развертывается в приложениях, где критически важна безопасность.

Типичная настройка для внутренней атаки содержит в себе три ключевых элемента 1) предопределенный сигнальный шаблон 2) несколько вредоносных примеров случайно выбранных из жертвы и 3) внутреннею целевую метку. Существующие обследующие действия в большинстве посвящены проектированию эффективных и скрытых шаблонных сигналов как фиксированный черно-белый шаблон, враждебный шум, или возмущение, зависящее от конкретного образца. В этих атаках есть только один противник и один тип сигнала, таким образом мы зовем их одночно-сигнальными атаками. Эти одиночно-сигнальные это единственная установка, которая была исследована в существующих работах. Однако, реальные атаки могут быть более сложными чем единичные, возникает вопрос могут ли единичные атаки представлять угрозу для существующих предложений. Если атака на один датасет или систему ИИ имеет высокую значимость, то она должна побудить нападающего использовать атаку с разными сигналами. В нашем случае, возникает вопрос могут ли все сигналы сосуществовать и быть одновременно эффективными.

В этой статье, мы расширим текущее бэкдор исследование с одиночных до множественных атак, где существуют множество противников которые одновременно портят один и тот же набор данных используя разные типы сигналов. Множественные атаки могут быть реализованы в трех путях: параллельные, последовательные и гибридные, которые смешивают множественные сигнальные шаблонны в один большой сигнальный шаблон. Рассматривая множественную атаки, мы проведем комплексный анализ 10 типов существующих сигналов и раскроем сосуществующие, переписывающие и кросс-активирующейся взаимодействия между ними. Конкретно, мы обнаружим, что 1) разные типы бэкдор атак могут сосуществовать, т.к .они могут быть введены внутр единственного датасета параллельно или последовательно, 2)определенные сигналы могут перезаписывать другие сигналы при чрезвычайно низкой частоте отравления 3) один тригер могут перекрестно активировать другие шаблонны тригеров под последовательной атакующей настройке и 4) разные сигнальные шаблоны могут смешиваться в один гибридный шаблон, что создать более мощную атаку. Эти результаты свидетельствуют о большей угрозе безопасности при наличии множества бэкдоров-злоумышленников.

Другой волнующий факт, что самые популярные существующие методы защиты были разработаны для единично-сигнальных атак, поэтому не существует надлежащих оценок для защиты множественно-сигнальных атак. Упущение таких оценок может вести к чрезмерно оптимисчиным взглядам на бэкдор надежность и ложное чувство защищенности. Мы нашли что, к сожалению, так оно и есть: множественно-сигнальные бэкдор атаки могут сильно повредить текущие современные технологии поиска и удалять методы. Наши главные выводы могут высказаны так:

1) Мы вводим концепцию Мульти-Сигнальные бэкдор атаки (МСБА) для исследования практической угрозы бэкдор атак и диких реальных сценариев, где один и тот же набор данных или модель могут быть атакованы множеством противником параллельно, последовательно или гибридною

2) Было проведено более 200 экспериментов с 10 типами сигналов, 3 отравляющими стратегиями( параллельная, последовательная, гибридная), 4 оценки порчи ([0.2%-10%]), 3 стратегиями модифицированные метками (All2One, All2All, All2Random), 2 датасета(CIFAR-10 и подмножество интернет картинок ), 4 DNN архитектуры (2 CNN и 2 ViTs), мы обнаружили сосуществующие, переписывающие и перекрестно-активирующие эффекты множестенно-сигнальных атак.

3) Переоценивая 8 существующих методов защиты (4 бэкдор обнаружения и 4 метода удаления) против параллельных множественно-сигнальных атак, содержащих 10 типов сигналов, мы показали, что 1) Все методы обнаружения имеют трудности для обнаружения All2All и All2Random атаки, хотя могут обнаружить All2One атаки в некоторой степени, 2) ни один из методов обнаружения не может найти бэкдор метки множественно-сигнальных атак и 3) Ни один метод удаления не может полностью удалить полностью любой из 10 сигналов. Мы также показываем набор данных со множественными сигналами, чтобы помочь оценить в будущем атаку и защиту.

**2 .Связанная работа**

**Бэкдор атаки.**

Бэкдор атаки спроектировали шаблоны сигналов с высокой оценки успеха атаки, причем оставаясь незаметными для человека. Существующие бэкдор сигналы могут быть классифицированы на сигналы по набору данных и по выборке. Сигналы по набору выборки применяют один и тот же сигнальный шаблон для порчи всех экземпляров, когда по выборке портят каждый экземпляр с помощью уникального шаблона сигнала. Примеры sample-wise сигналов включают в себя: один пиксель, клетчатый шаблон, глобальный шаблон как Гаусовский шум , фоновое отражение, вот несколько примеров. Сравнивая dataset-wise сигналы и sample-wise, вторые могут быть намного более сложными(и скрытными), так как они часто оптимизируются добавлением техник. Например, бэкдор атака на основе генеративной модели, генерирует бэкдор шаблоны сигналов адаптивно основанные на разных примерах ввода. Не смотря на разнообразие сигналов, все существующие бэкдор атаки подчиняются установке единично-сигнальный атаки, где существует только один против и один тип сигнала. Такие рамки установки ограничивают исследование бэкдор атак в более реалистичных случаях, где множество противников могут атаковать один и тот же датасет. Это также поднимает вопрос могут ли эти атаки сосуществовать и быть одновременно эффективными, когда используются различными неприятелями для порчи одного набора данных. В этой статье, мы обнаружим такие аспекты множества существующих шаблонов для более реальной ситуации: случай множестенно-сигнальной атаки где множество нападающих и типов сигналов.

**Бэкдор защита.**

Существующие бэкдор защиты могут быть классифицированы в методы бэкдор обнаружения и бэкдор удаления(или уменьшения). Метод обнаружения выявляет, атакована ли данная модель с помощью бэкдор атаки или содержит ли образец сигнал. Методы удаления целятся устранить бэкдор сигнал(если он существует) из модели без потери функциональности. Это можно сделать во время процесса тренировки с помощью анти-бэкдор обучения или позже через тонкую настройку, тонкое отсечение, дистилляцию, состязательное отсечение, обрезание канала Липшица или восстановительное отсечение. Однако, все эти типы были разработаны для одиночно-сигнальных атак, таким образом возникает неопределенность при защите от множественно-сигнальных атак. В этой статье, мы начнем обширные эксперименты для переобследования эффективности этих защит от множественно-сигнальных бэкдор атак.

**3. Множественно-сигнальные бэкдор атаки**

C начала мы введем наши модели угрозы и определение бэкдор атак и далее введем три типа множественно-сигнальных атак, выдвинутых в этой работе.

**3.1 Модель угроза и проблема определения**

**Угрожающая модель**

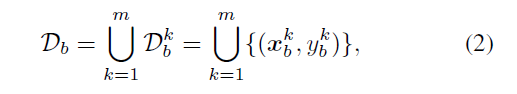
В этом части, мы утвердим информационную порчу и угрозу множественно-сигнальных бэкдор моделей. Конкретно, существует один или более вредитилей, где каждый противник может получить несколько тренировочных экземпляров, но никто из не может манипулировать процессом тренировки. Враги могут быть независимыми или сговориться, причем независимые противники отравляют разные тренировочные образца с помощью разных сигналов, в то время как сговорившиеся противники могут отравлять один и тот же тренировочный образец в последовательном порядке, накладывая один сигнал после другого. Здесь может также существовать супер злодей, который начинает вышеупомянутые атаки все вместе или комбинирует множественные сигналы в гибридный или супер сигнал. Мы в частности сконцентрируемся от низкой до средней степени поражения с 0.2% до 10%. Вышеупомянутые установки нашей угрожающей модели симулируют сложные и бурные реальные сценарии, где множество типов сигналов могут быть введены в такой же набор данных одним или множеством нападающих при низком бюджете, например, Carlini et al. смогу отравить 0.01% LAION-400M or COYO-700M датасет за 60 USD.

**Определение бэкдор атаки**

Мы сфокусируемся на задачи классификации изображений. Возьмем чистый обученный датасет , где принадлежит X представляет собой картинку и y\_i принадлежит Y это метка. Бэкдор противник генерирует бэкдор примеры с сигнальной функцией tr:X->X. Для каждого чистого примера, она сопоставляет примеру x в бэкдор пример x\_b, где x\_b = tr(x) и изменяет первоначальную метку в целевую бэкдор метку y\_b. Чтобы гарантировать незаметность и эффективность бэкдор инъекции, нападающий случайно выбирает несколько тренировочных примеров для отравления, это создает множество бэкдор примеров D\_b = {(x\_b,y\_b)} . Как мы упомянули ранее, большинство существующих атак только водят один тип бэкдор сигнала в тренировочный датасет, это означает, что есть только одна сигнальная функция tr(.). Мы зовем эти атаки Одиночно-Сигнальные Бэкдор атаки (ОСБА). Последний отравленный тренировочный датасет может быть обозначен как представляет собой множество чистых примеров и меток, в то время как представляет собой бэкдор примеры и бэкдор метки. Коэффициент отравления определяется как . Тренировка модели – это решение задачи оптимизации:

Где - кросс ентропия. Первое слагаемое в формуле выше означает ошибку чистой задачи, когда второе слагаемо – ошибка в бэкдор задаче. Обучение на наборе можно рассмотреть как процесс, в ходе которого модель осваивает обе задачи.

**Определение множественно-сигнальной бэкдор атаки**

Мы называем бэкдор атаки, которые используют множество типов сигналов(вероятно от одного или нескольких нападающих) для атаки одного датасета – мульти-сигнальные бэкдор атаки (МСБА). Рисунок 2 иллюстрирует идею МСБА. Заметим, что мы не называли эти атаки мульи-вражескими атаки, так как даже один противник может осуществить разные типы сигналов. Задача МСБА может быть сформулирована эквивалентно ОСБА, но с немного более сложным множеством отравителем, который содержит множество сигналов

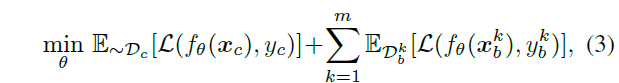
Где, это отравленный к-ой сигнальной функцией образец , и  - подмножество испорченных примеров с помощью для всех m сигналов. Для мульти-сигнальных атак, отравленный датасет становится  с коэффициентом отравления 

**3.2 Стратегии отравления**

МСБА - это простое расширение ОСБА. МСБА вводит множество типов сигналов в целевую модель пока ОСБА только вводит один тип сигнала. Множество бэкдор сигналов могут быть введены в один датасет различными способами. Здесь, мы сформулируем три стратегии отравления для симуляции и воссоздания реального сценария: параллельное, последовательное и гибридное отравление.

*3.2.1 Параллельное отравление*

Вероятно, сигналы могут быть введены в датасет жертву с помощью множества независимых противников. В этой случае, разумно полагать, что отравленные подмножества не перекрывают друг друга, так как два независимых противника с очень малой вероятностью отравять один и тот же образец, учитывая низкий коэффициент отравления. Мы называем такую стратегию отравления – параллельным отравлением.

Для реализации параллельного отравления, мы случайно выбираем несколько тренировочных образцов в кандидатское бэкдор подмножество  и далее однородно делим это множество в m маленьких подмножеств, т.е. . Далее мы назначаем каждому подмножеству случайный отбираемый сигнал  из сигнального бассейна, т.е. . Соответственно, тренировочная модель мульти-сигнального отравленного датасета может быть сформулирована как:

Где - кросс-энтропия и *m* это число независимых отравленных подмножеств. Тренировка на мульти-сигнальном бэкдор датасете может рассматриваться как процесс обучения одной чистой задачи и *m* бэкдор задач одновременно.

*3.2.1 Последовательное отравление*

Это также возможно, что разные противники запускают их атаки в разное время. В этом случае, разные противники могут атаковать тот же датасет в последовательном порядке, но все еще не пересекая информационные подмножества. Мы зовем эту стратегию атаки – последовательное отравление. Эта стратегия отправления позволяет нам изучить переписывающие эффекты разных сигналов, т.е. появляется вопрос, может ли ранее введенный сигнал оставаться таким же эффективным при наличии последующих атак.

Для реализации последовательного отравления, мы вводим разные типы сигналов в датасет жертву, следуя определенному порядку. Предположим, противник k может отравить маленькое подмножество чистых экземпляров с его определенным сигналом чтобы получить бэкдор подмножество  и соответственно отравленный тренировочный датасет , где это оригинальный датасет, - подмножество сохранившихся чистых экземпляров, это чистое подмножество жертв, бэкдор экземпляры сгенерированные из. Затем модель обучают на испорченном датасете для создания бэкдор модели следующим образом:



Следующий противник k+1 отравил датасет, следуя той же процедуре, что и протвник k чтобы создать отравленный датасет  . Бэкдор модель далее непрерывно обучалась на для создания  следующим образом:



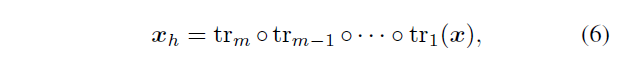
После вышеупомянутого последовательного обучения, модель становится последовательно бэкдорной моделью, которая содержит оба сигналаи .

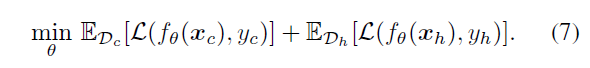
Заметим, что хотя противники следуют последовательному порядку, мы предполагаем, что они независимы и их выбранные целевые подмножества не пересекаютсяи только содержат чистые экземпляры. Это предположение разумно, потому что

1. Противники имеют доступ к жертвенным примерам, это может легко обеспечить, что они чистые(не были под воздействием сигнальной инъекции)
2. Независимые нападающие часто имеют разные примеры. Мы также предполагаем последующий противник тренирует бэкдорную модель на базе предыдущей бэкдорной модели. Это делается, чтобы имитировать текущей тренд настройки огромных моделей, где жертва(пользователь) может скачать предтренированную (и отравленную предыдущим оппонентом) большую модель для его личного последующего приложения. По этой же причине, мы ставим процессу тренировки модели быть максимально короткими, как 10 эпох, т.е. выражения 4 и 5 оптимизированы для 10 эпох. Такой небольшой оптимизации достаточно для инъекции сигнала, в то время как полные тренировочные эпохи достаточно длинные для обучения чистой задачи.

*3.2.1 Гибридно-сигнальное отправление*

Также могут существовать супер оппонент, которые комбинирует разные сигналы в один гибридный для достижения эффекта множественности сигналов. Стратегия отравления реалистична, так как текущая литература уже насчитывают огромное число сигнальных шаблонов для эксплуатации противником. Мы зовем такую стратегию гибридно-сигнальным отравлением.

В сравнении с параллельным и последовательным отравлением, которые основываются на множественных независимых сигналах, гибридно-сигнальное отравление представляет собой sample-wise атакующую стратегию. Она одновременно вводит множество определенных сигналов в одиночных входящий экземпляр, наделяя его множественно-сигнальными характеристиками. Конкретно, дан чистый экземпляр x, гибридно-сигнальная атака отравляет его m элементарными сигналами следующим образом:

где, мы используем мягкое смешивание на каждом шаге, т.е. мы поставили d в нашем эксперименте). Предположим, что тренировочный датасет и маленькое подмножество чистых примеров доступно для оппонента , противник вводить гибридный сигнал в для создания бэкдор подмножества  с помощью формулы 6. Отравленный датасет далее может быть определен как Далее противник может обучить бэкдорную модель на следующим образом: 

**Потенциальный сигнал**

Сбор образцов и нужного множество сигналовэто ключевая стадия для обучения. Исследуя необходимую литературу, мы тщательно выбрали 10 типов сигналов используемых в главных бэкдор атаках как наши потенциальные сигналы, которые включают оба dataset-wise сигналов(такие как один пиксель, клетчатый шаблон, и враждебная петрубация) и sample-wise сигналы(такие как учитывающие вход шаблоны и деформация изображений). Полный список изученных сигналов могут быть найдены в секции 4.1.

**3.3 Стратегии отравления**

После изучения работ мы обнаружили следующие стратегии модификации меток*: All2One, All2All и All2Random*.

* *All2One*: Эта стратегия перемаркирует все бэкдор экземпляры на фиксированные целевые метки , т.е. все экземпляры имеют одинаковую метку . Другими словами, все противники совместно используют одну и ту же цель бэкдор.
* *All2All*: Оно изменяет метки бэкдор экземпляров (сделанные из чистых примеров x) в , где K это конечный номер всех классов, y это оригинальные(чистые) метки x/ и это измененная метка. Это одновременно сценарий, где следующий класс особенно интересный для всех оппонентов.
* *All2Random*: Это стратегия имитирует сценарий, где не существует конкретной цели между противником и каждой метку и существует одинаковый шанс для выбора целевой метки. В этом случае, мы изменяем метки бэкдор экземпляров в где  - функция случайного выбора.

**4. Эксперименты**

В этой секции, мы экспериментуруем и суммаризируем множество ключевых моментов связанных с множественно-сигнальными атаками, включая сосуществвующие, переписывающие и кросс-активирующие эффекты разных сигналов с параллельельными, последовательностии гибридными установками, и надежность существующих методов защиты от множественно-сигнальных атак.

Мы выбрали 10 представитилей сигналов с текущей литературы как наши кандидаты, которые включают статические сигналыы как OnePixel, BadNets, Trojan attack, глобальные сигналы как Blend, Sinusoidal Spectrum(SIG), adversial nose, Smooth, Nashivell filter и sample-wise сигналы как Dynamic,WaNet. Для обеспеченняи последовательного и честного сравнения, мы используем установка с грязной-меткой для всех сигналов, которые включат отравление информации и изменении модификации в два этапа. Мы протестируем все сигналы для стратегий: последовательного, параллельльного и гибридного отравления, а также стратегий All2One, All2All и All2Random для измененя модификаций. Для выбранной модели, мы в основном сосредоточены на двух класических CNN архитектурах: ResNet и MobileNet, и двух трансформер архитектурах: ViT-base и ViT-small. Четыре архитектуры самые рапространныые в стандартных, ограниченных ресурсами или крупномасшатбных приложениях компьютерного зрения. Более того, есть небольшое иследование бэкдор атак и защит с помощью ViTs. Мы заполним этот проблем с помощью общего исследования с двумя ViT архитектурами. В качетсве датасетах мы использовали два общепринятых датасета : CIFAR 10 и подмножество ImageNet (первые 20 классов). Если не указано иное, мы ставим степень отравления 10%(1% для каждого типа сигнала). Деталезированные установки атак могут быть найдены в приложении. Сгенерированные множественно-сигнальные датасеты были сделаны, чтобы помочть разработке улучшенных способов защиты. Также, отравленные датасеты могут продолжить обьеденяться в более сложные и продвинутые сигналы, следуя нашим выводам.

Для бэкдор защиты, мы рассмотрели 8 продвинутых методов защиты, включая 4 метода бэкдор детекции модели: Neural Cleanse(NC) , UMD, MMDB, и детекция на заучивании in RNP, и 4 метода удаления Fine-tuning, Finepruning, Neural Attention Distillation и Adversarial Neural Pruning. Так как эти методы защиты были первоначально спроектированы для CNN архитектур, мы только тестируем fine-tuning защиту для ViT моделей пока ограничиваясь другими защитами CNN. Больше деталей может быть найдено в приложени.

**4.2 Анализ и постижение MTBAs.**

Первое мы проанализируем MTBAs с параллельным, последовательным и гибридно-сигнальным отравлением, соответственно. С помощью этих экспериментов, мы далее раскроем свойства сосуществования, переписывания и кросс-активации между разными сигналами.

4.2.1 Параллельные MTBAs

Вспомним, что в параллельных MTBAs, мы вводим 10 сигналов одновременно в целевой CIFAR-10 и ImageNet — 20 датаасетах, где каждый сигнал отравляет уникальное подмножество чистых примеров. Здесь, мы расскажем результаты CIFAR-10, а ImageNet -20 отложим на приложении. Таблица 1 показывает вероятность успеха атаки 10 сигналов с разными вариациями отравлляющего коэффициента %alpha [0,2%, 10%].

**Свойства перезаписывания**. С начала мы посмотрим на средний ASR(последий столбец) результат с коэффициентомо травления 10%. Одно ключевое наблюдение это то, что сигналы могут сосуществовать как единичная модель, достигая 98.5%,88.04% и 86,40% ASR для All2One, All2All и all2Random целей, соответственно. Большинство сигналов демонстрирует ASR в райне 80%, за исключением BadNets и Adv сигналов для All2Random атак. Это говорит о том, что фиксированные сигнальные шаблоны(Adv — это состязательно возмущенная версия Badnet клетчатого сигнала) могут быть легко повлиять на другие сигналы. В целом, All2One атаки выглядят более мощными, чем All2All или All2Random атаки с таким же коэффициентом отравления, что ожидаемо, т. к. All2One атаки имеют единственную цель.

Следуя к маленькому коэффициенту отравления (от 0.2% к 1%),мы обнаружим что средний ASR всех 10 синалов(последний столбец) существуенно ухудшается с отравляющим коэфициентом. Однако, число выживших по прежнему существенно низкое для 0.2%, например Blend, SIG, и Dynamic сигналы имеют ASR > 40% для All2One и All2Random установок. Нет сигналов успешных с All2All установками с низким коэффициентом отравления. Мы предположили, это потому что циклицеское изменение меток в страгеии all2All вызывает локальный(внутри сигнала) и глобально(между-сигнальный) перекрытие и нарушения в «сигналь — цель» отображении, делая сложным ассоциировать малеьный фиксированный сигнал для изменения целевых меток. Этот результат показывает, что ALL2ALL стратегия оказывается самой спорной настройкой атаки с очень низким коэффициентом отравления ( например, 0.2 и 0.5).

Результаты в таблице 1 также может помочь ответить на вопрос: существует ли единственный сильнеший сигнал для всех сценариев. Таблица показывает, что в All2One настройке, Dynamic это лучший сигнал по средний значенияем разных степенней отравленя; для all2All настройки, BadMets сигнал лучший, когда для All2Random насйтроки, Dynamic сигнал опять сильнейшний. Таким образом, мы верим, что потенциально он будет сильнейшым одиночным сигналом. Наш эксперимент с последовательным MTBAs в секции 4.2.2 также подтверждает, что гибридная характеристика Dynamic сигнала, может кросс-активировать другие сигналы. Об этом стоит упомянуть, что не все sample-wise сигналы стрильные, т. е. Sample-wise сигналы Smooth и WaNet слабее чем dataset-wise сигнал BadNets с all2One и all2all настройками.

Чтобы помочь понять класеры сформированные разными сигналами, мы покажем t-SNE графики 10 сигналов для All2One, All2all и all2Random модификации меток. Он показывает разные кластерные действия к треим разным стратегиям модификации меток. Особенно, здест существует оба маленьких компкатных кластера и относительно огрмоный кластер для All2One и all2Random атак, пока не важных кластеров для all2all. Это подтверждается, что All2All атаки нуждаются в большем усилии для достижения, выделяя самые сложные комлпексные атакующие сценарии. Независимые кластеры выучены моделью для разных сигналов обьясняют сосуществование эффекта между разными сигналами.

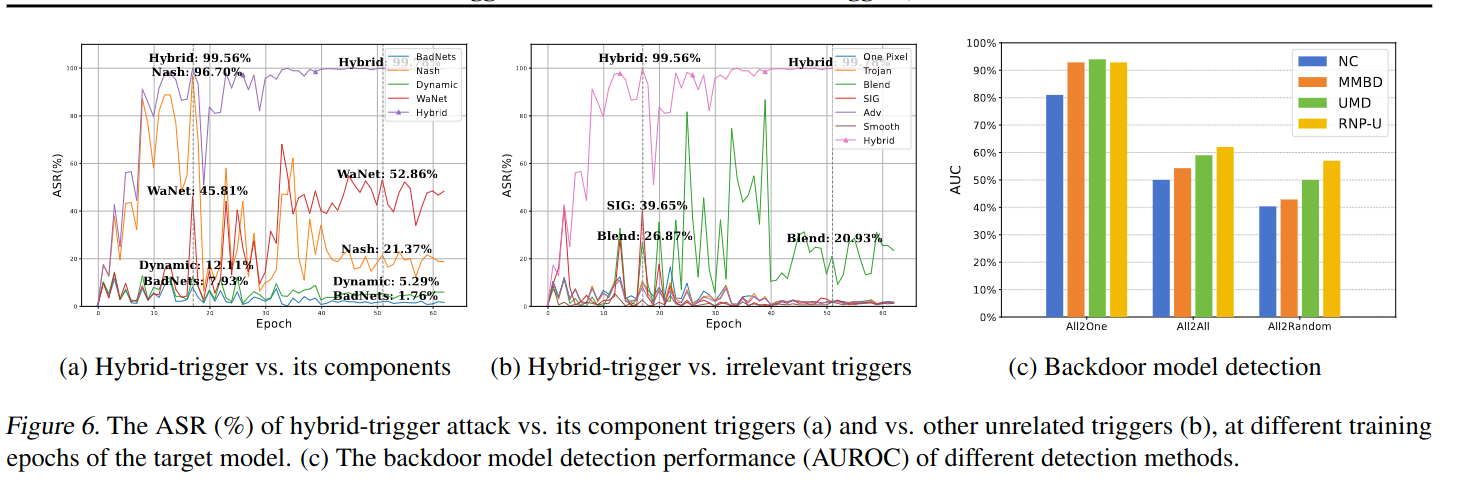
4.2.2 Последовательные MTBAs

**Перезаписывающее свойство.** Последовательные MtBas используют инъекцию 10 сигналов следуя указанному порядку для итеративных тренировочых эпох( каждые 10 эпох). Инъекционный порядок и выполнение атаки, после этого каждый сигнал имплантирован в модель, показаны на матрице asr на картине 5. Как можно заметить, здесь существует перезаписывающи й эффект в последовательном MTBAs, т. е. Эффективный сигнал раньше становиться бесполезным после новых сигналов иньъектированныъ в модель. Например, OnePixel падасет с 91.85% ASR на 40.95% после того, как BadNets была инъектирована в модель, и далее падает в 2.48% после инъекции трояна. Тот же эффект также наблюдается почти во всех сигналах, на это указывает последовательная тенденция, где клетка ниже диагонали меет ASR ниже.

**Кросс-активирующее свойство** Другое важное наблюдение это то, что опредененные строчки имеют существуенно больший asr, например, BadNets,Adv и Dynamic rows. Это эффект кросс-активации, это когда один тип сигналов часто активируется другим типом. Этот управляющий эффект означает, что сигналы деляются определенными или очень больше сходтсво. Например, BadNets,Torjan, Blend,Adv,Smooth, Dynamic, and Nash, с asr равным 100%,100%,100%,56%,45%,100% и 38,51% соответствено. Dynamic сигнал можыть активирован Trojan, Adv, Smooth, Dynamic, and Nash с ASR 93.66%,55.20% ,35.74%,100% и 76.69%, соответственно. Мы думаем, что эффект кроасс-активации это результат cхожего пиксельного распредедения в шаблонных патернах, как те, которыми делелись между BadNets, Trojan and Dynamic сигналов.Среди сигналов, Trojan и Dynamic два самых похожих, т. к. Trojan может активировать Dynamic к ASR 93,66 когда Dynamic активирует троян к 79,18%. Это потому что оба сигнала оптимизированны сигналами. Мы верим, что это тонкое но интригующее взаимодействие между двумя сиганалми достойно более глубинных исследований, особенно при множественном-сигнальной стратегии.

**4.3 Гибридно-сигнальные MTBAs.**

Разница между между параллельными и последовательными MTBA, где сигналы независимы, гибридно-сигнальные атаки смешивают множество сигналов в один чистый образец, все они указывают на одну и ту же целевую бэкдор метку. Эта секция исследует эффективность гибридных сигнальных атак на CIFAR-10. Для гарании выской производительности атаки, мы выбрали 4 сигнала, чтобы сконструировать гибридный сигнал, они включают: один статический BadNets, один Dynamic и два незаметных сигнала Nash and WaNet. Заметим, что Dynamic не работает хорошо, как часть гибридного. Мы применяем эти сигналы в рандомной порядке(мы покажем, что порядок имеет минимальное влияние на ASR), используя подход мягкого смешивания. Примеры гибридно-сигнальных атак приведены в приложении.

Рисунок 6а показывает, что ASR нашей гибридной атаки и индивидуальных сигналов. Ключевое наблюденик закючается в том, что гибридный сигнал может не только досгинуть высокий ASR(>=90%) но также демонрирует уверенный кросс-активационный эффект на сигналах содержащихся(хотя и не все из них). Особенно, кроме того гибридные сигналы такие как Nash и Wanet быстро достигают ASR равным 96,70% и 45,81% соответсвенно в малых стадиях обучения( до 20 эпох), хотя они стабилизируются на 52,86% и 21,37% соответсвенно в конце. Фигура 6б показывает довольно удивительный феномен, что гибридный сигнал может даже кросс-активировать Blend и Sig сигналы в середине обучения, хотя они даже не часть гибридного. К несчастью, эта кросс-активация это времена, она исчезает в конце обучения. Мы надеемся на проективнее более продвинутых гибридных синалов, которые могут кросс-активировать множество других индивидуальных сигналов в интересном будущем.

**4.4 Переоценка существующей защиты.**

Эта секция проверяет надежность сущестующих методов бэкдор защиты против множественных атак. Здесь, мы еще раз посмотрим 4 модельных обнаруживающих метода и 4 удаляих метода. Мы только сконцентрируемся на параллельном отправлении, т. е. 10 сигналов рассмотренных в этой работе были введены параллельно.