**Доверенный искусственный интеллект**

**Защита данных**

А.Е. Самотуга, П.С. Ложников

Конспект лекции

**ЗАЩИТА ДАННЫХ**

В рамках прошлой темы мы рассмотрели существующие атаки на искусственный интеллект. На основе этих знаний важно понимать, что безопасность ИИ должна обеспечиваться на всех этапах его жизненного цикла. Особое внимание следует уделять данным, которые применяются на этапах обучения ИИ, его исполнения и к процессу их хранения.

В данной теме мы рассмотрим существующие подходы к защите данных и ИИ.

## Защита данных на этапе хранения, обучения и исполнения. Отравление данных. Дифференциальная конфиденциальность. Гомоморфное шифрование. Классическое шифрование. Федеративное обучение

Предотвратить атаки манипуляции с входными данными поможет знание внутреннего представления информации в системе для всего возможного пространства входных данных, а не только обучающего распределения. Представления, характеризующиеся нестабильностью, уязвимостью для коррекции, могут быть легко и незаметно искажены, поэтому оптимизация принципов представления данных позволяет улучшить безопасность систем машинного обучения. Также система будет защищена от отказов в ситуациях, когда данных недостаточно или они слишком зашумлены.

Другое направления защиты от манипуляций с входными данными представляет собой методы распознавания аномалий. Такой подход предполагает выбор и исследование какой-либо меры типичности данных, что позволяет защититься от отравления модели.

С целью защиты от искажения обучающей выборки, помимо использования обнаружения аномалий, также рекомендуется содержать выборки в тайне. Внедрение оценки уровней доверия для прогноза и использование многоуровневой защиты также является подходящим направлением для решения данной проблемы.

Одним из направлений по противодействию угрозы на решающий бит, в том числе, когда злоумышленник пытается имитировать управляющие воздействия является разработка и внедрение технологии “цифровых двойников”.

«Цифровой двойник» должен прогнозировать результат выполнения последовательности команд, прежде чем запускать их на реальном объекте управления. Однако разработка подобных решений — сложный, ресурсоемкий процесс, успех которого зависит от предметной области (разработать подобную технологию далеко не всегда возможно).

Одно из направлений по устранению угроз для ИИ является приведение работы ИИ к режиму преобразователя входных воздействий (поступающей информации) в длинный криптографический ключ или пароль, который можно ассоциировать с определенным управляющим воздействием. Таким образом, вместо коротких кодовых команд на выходе ИИ, нужно использовать длинные криптографические ключи. Каждый ключ должен быть ассоциирован с отдельным действием, и только объект управления должен "знать", что нужно делать с поступившим от ИИ ключом (последовательность управляющих команд может быть зашифрована на данном ключе). Другими словами, искусственный интеллект должен проходить процедуру аутентификации. Для авторизации необходимо использовать неотчуждаемый от ИИ ключ или пароль, который должен храниться безопасно в защищенной памяти (ключ должен «помнить» только ИИ).

Параметры обученного ИИ необходимо хранить в специальном виде, защищенном от извлечения «знаний» даже при отсутствии стороннего шифрования (конечно, шифрование можно применять для усиления защиты). Для этого нужно перейти от весовых коэффициентов классических нейронов к параметрам корреляционных нейронов Байеса-Минковского.

Анализ собственных (внутренних) корреляционных связей образов и принятие классификационных решений происходит без необходимости хранения информации о корреляционных связях или значениях признаков, характерных для образов, принадлежащих к тому или иному классу. Другими словами эталонная информация о классах образов не компрометируется при хранении.

Инвертировать решения нейросетевого преобразователя, о котором идет речь, затруднительно.

Для того, чтобы атаки «на решающий бит» было затруднительно реализовать, необходимо, чтобы нейросетевой ИИ мог быстро обучаться в полностью автоматическом режиме. Тогда новую таблицу криптографических ключей можно генерировать периодически (например, раз в день или раз в 10 минут), заново обучая нейросетевой преобразователь в доверенной среде, после чего он может быть помещен в потенциально враждебную среду, где будет функционировать до следующего обучения. После каждого обучения предыдущая версия утратит работоспособность, поэтому хакеру придется взламывать преобразователь снова и снова его перенастраивать, чтобы получить требующийся для злоумышленных действий эффект.

Если следовать описанным принципам атаки типа "ключ под ковриком" и атаки "на решающий бит" станут неосуществимы уже на уровне концепции построения ИИ. Конечно, атаки на «извлечение знаний» из защищенного ИИ будут появляться (как и в случае с атаками на алгоритм ГОСТ Р52633.5), но они будут гораздо менее эффективны, так как защиту ИИ дополнительно можно усилить криптографией. Криптографическая защита корреляционных нейронов Байеса-Минковского может быть основана на классических методах без применения гомоморфного шифрования

## Федеративное обучение

Реализация такой концепции как федеративное обучение (заключение модели ИИ в защищенную среду и ее обучение без перемещения обучающей выборки куда-либо) не дает защиты от атаки «извлечения знаний», так как эта атака направлена на параметры уже обученного ИИ, при условии, что процесс обучения мог уже проходить в защищенной среде. Отметим, что подобная атака может осуществляться в отношении не только нейросетевых реализаций ИИ.

В то же время нельзя не признать, что федеративное обучение обладает рядом преимуществ:

1. В связи с тем, что данные никому не передаются, остаются на персональном сервере, что положительно влияет на их безопасность;
2. Возможность реализовать прогнозирование в реальном времени со значительно меньшими временными затратами;
3. Не требуется использование интернета;
4. Не требуется значительных аппаратных ресурсов и другие.

Выделяют две разновидности: горизонтальное и вертикальное.

## Гомоморфное шифрование. Классическое шифрование

В теории гомоморфное шифрование может быть использовано для защиты ИИ, но на практике до этого еще далеко, так как имеются нерешенные проблемы с низкой производительностью и с накоплением ошибок при шифровании даже небольших объемов данных. Начиная с некоторого размера гомоморфные шифртексты перестают однозначно расшифровываться. Чем больше длина зашифрованного текста, тем больше вероятность, что верно дешифровать гомоморфное решение не удастся. На сегодняшний день для гомоморфного шифрования создан стандарт ISO/IEC 18033-6:2019, но этот стандарт не касается шифрования параметров нейросетевых решающих правил. Для защиты обученного ИИ и нейронных сетей с помощью гомоморфного шифрования следует разработать отдельные стандарты или рекомендации, которых на данный момент не предложено. Если не использовать гомоморфное шифрование, то ключи должны где-то храниться (в базе данных, в коде ИИ и т.д.), что создает ряд внешних и внутренних угроз. Хакер может похитить ключ, используя уязвимости в защите или вступив в сговор администратором. Человек всегда является «узким местом» в системе безопасности, поэтому эти вопросы нельзя закрыть полностью. Хуже всего, если все знания ИИ шифруются на одном ключе (нет разделения знаний на независимо зашифрованные фрагменты). Создание инфраструктуры для безопасного хранения криптографических ключей – сложная задача, требующая значительных финансовых затрат.

## Генерация синтетических наборов данных

Изначально генерация синтетических образов направлена лишь на повышение доверия к оценке системы, построенной на основе машинного обучения. В частности, данное утверждение справедливо для оценки качеств биометрической системы. Генерация синтетических образов в этом случае, необходима для получения базы образов “Чужой” достаточных размеров.

Размножение биометрических примеров одного биометрического образа или, другими словами, их генерация, согласно ГОСТ Р 52633.2-2010 Требования к формированию синтетических биометрических образов, предназначенных для тестирования средств высоконадежной биометрической аутентификации, может производиться с помощью синтеза случайных биометрических примеров с сохранением статистических характеристик биометрического образа, частичной мутации имеющихся биометрических примеров, морфинга между биометрическими примерами одного биометрического образа, перестановки фрагментов биометрических примеров.

Морфинг биометрических примеров одного биометрического образа заключается в нахождении промежуточных значений (биометрических примеров-потомков) для каждого из параметров пары биометрических примеров-родителей.

Оценку количества биометрических примеров-потомков для каждой пары биометрических примеров-родителей производят в соответствии со определенным алгоритмом.

Синтез биометрических примеров производят по следующему алгоритму.

1) По всем имеющимся примерам образа вычисляются математические ожидания каждого i-го контролируемого биометрического параметра Eобpaз (vi).

2) Вычисляется стандартное отклонение каждого из контролируемых биометрических параметров sобраз(vi).

3) Генератором нормальных случайных чисел со статистическими характеристиками формируются векторы параметров заранее заданного числа биометрических примеров.

Такой метод размножения биометрических примеров не учитывает корреляционные связи между параметрами биометрических примеров.

Для размножения биометрических примеров перестановкой фрагментов необходимо разделить исходные биометрические примеры на фрагменты. Возможны два варианта такого разделения.

1) Использование естественной фрагментации исходных биометрических примеров (таких, как примеры голосовых образов, легко разделяемые на фонемы, или примеры рукописных образов, легко разделяемые на отдельные символы, буквы).

2) Использование фрагментации биометрических примеров, заложенной производителем средства при вычислении биометрических параметров, и связанной с тем, что каждый из множества полученных параметров сам по себе является некоторым фрагментом исходного примера.

При размножении N исходных биометрических примеров, как при использовании естественной фрагментации, так и при фрагментации биометрических параметров, необходимо для формирования каждого синтетического биометрического примера использовать приблизительно 1/N фрагментов каждого из исходных биометрических примеров.

Мутация биометрического образа - создание синтетического биометрического образа(ов), основанное на случайном изменении биометрических параметров биометрического образа-родителя.

Все из известных подходов к созданию синтетических биометрических образов основаны либо на самих вышеупомянутых методах, либо на их комбинациях и модификациях.

Также следует упомянуть недавно появившееся направление под названием deepfake, которое может применяться для создания синтетических образов и основано на применении генеративно состязательных сетей.

## Метрики приватности

Одной из проблем, связанных с угрозами конфиденциальности в ИИ, является оценка степени конфиденциальности и степени защиты, обеспечиваемой системой. Определение и применение показателей конфиденциальности направлено на решение этой проблемы. В ИИ существуют различные методы машинного обучения или повышения конфиденциальности для защиты конфиденциальных данных в разных областях. Целью определения показателей конфиденциальности является количественная оценка уровня конфиденциальности данных, который приводит к улучшению модели конфиденциальности в рамках конкретной модели ИИ. Технически, метрика конфиденциальности учитывает различные свойства данных и дает значение, которое представляет уровень конфиденциальности в системе. Преимуществом метрик конфиденциальности является возможность сравнения различных методов сохранения конфиденциальности, оценки различных методов в конкретной области и минимизации воздействия на конфиденциальность. Метрики конфиденциальности полезны, когда конфиденциальные данные подвергаются угрозе со стороны противника. Метрики конфиденциальности могут различаться с учетом источника данных, аспектов конфиденциальности, которые они оценивают, и угроз со стороны противника. В целом, все метрики конфиденциальности относятся к анонимности, неотслеживаемости, псевдонимность.

## Программные продукты для защиты конфиденциальности и целостности данных ИИ

С течением времени появляются продукты, в которых в той или иной степени реализуются механизмы защиты конфиденциальности: AIC ModelOps Platform, Modzy ModelOps platform.

**Список источников**

1. AI в защищенном исполнении. Что это означает и зачем нужны преобразователи образов в код? // VK [Электронный ресурс]. URL: [https://vk.com/@a.i.constructor-iskusstvennyi-intellekt-v-zaschischennom-ispolnenii-chto-eto](https://vk.com/%40a.i.constructor-iskusstvennyi-intellekt-v-zaschischennom-ispolnenii-chto-eto) (дата обращения: 22.04.2022).

2. Information technology — Artificial intelligence — Overview of trustworthiness in artificial intelligence ISO/IEC TR 24028:2020 [Электронный ресурс]. URL: <https://www.iso.org/cms/render/live/en/sites/isoorg/contents/data/standard/07/76/77608.html> (дата обращения: 12.04.2022).