Конспект лекции

А.Е. Сулавко

Защищенное исполнение искусственного интеллекта

**Заключение**

**ЗАКЛЮЧЕНИЕ**

Вопросы защиты искусственного интеллекта от атак «одного бита» («на решающий бит»), «ключ под ковриком», состязательных атак (в том числе атаки «извлечения знаний») в той или иной степени актуальны для наиболее ответственных приложений, неправильная работа которых может повлечь за собой материальный ущерб, нарушение информационной безопасности, угрозу жизни и здоровья граждан. На сегодняшний день нет стандартов (как международных, так и российских), которые бы регламентировали технические аспекты защиты искусственного интеллекта от угрозы реализации всех перечисленных атак, а также особенности построения архитектуры искусственного интеллекта, изначально устойчивого к этим атакам.

Гомоморфное шифрование[8] открывает значительные перспективы для защиты от обозначенной угрозы, однако на сегодняшний день разработать и стандартизовать достаточно универсальный для многих приложений метод защиты искусственных нейронных сетей пока не удалось. Имеются нерешенные научно-технические проблемы: низкая производительность, накопление ошибок (вероятностный характер дешифрования), зависимость предложенных схем защиты от архитектуры искусственного интеллекта. Результаты аналитико-синтетического исследования документального потока

указывают на то, что разработать метод защиты на базе гомоморфного шифрования, отвечающий всем требованиям, который можно положить в основу государственного стандарта, в скором времени вряд ли удастся.

Другая популярная на сегодняшний день концепция федеративного обучения[9, глава 15] нацелена на защиту данных от угроз нарушения конфиденциальности на этапе обучения ИИ, но не позволяет реализовать защиту от рассматриваемых в настоящем отчете атак, которые нацелены на параметры уже обученного ИИ. Хотя федеративное обучение гипотетически позволяет сделать процесс обучения относительно безопасным, она создает угрозы конфиденциальности данных самой модели ИИ (защита от этой угрозы снова упирается в необходимость использования гомоморфного шифрования). Обычное обучение выполняется быстрее и при этом проще обеспечить репрезентативность выборки. При этом для обеспечения конфиденциальности тренировочной выборки часто достаточно обезличивания примеров.

Искусственный интеллект должен проходить процедуру аутентификации. Для его авторизации необходимо использовать криптографический ключ или случайный длинный пароль, который должен храниться безопасно в защищенной памяти ИИ (ключ должен «помнить» только ИИ). В особо ответственных приложениях должна быть предусмотрена возможность смены ключа (пароля), используемого для авторизации ИИ. Для подобной перенастройки нужна процедура быстрого автоматического обучения блока защиты, отвечающего за связывание ключа с другими знаниями ИИ (контроль над возникновением переобучения должен быть автоматическим). Защищённое исполнение искусственного интеллекта в задачах классификации концептуально может быть построено по принципу преобразователя распознаваемых (классифицируемых) образов в длинный криптографический ключ, который можно ассоциировать с определенным управляющим воздействием. Только объект управления должен иметь представление о том, как применять возникающий на выходе искусственного интеллекта ключ.

Для реализации данной концепции необходимо разделить функционал искусственного интеллекта на блоки выделения признаков и блок защищенного исполнения (преобразователь образов в ключ, рисунок 1.3). Блоков выделения признаков может быть множество (в зависимости от приложения могут быть предусмотрены блоки предварительной обработки, нормировки и т.д.). Все блоки кроме блока защиты могут быть реализованы на основе практически любых подходов (нейронные сети, классические методы спектрального и корреляционного анализа и др.). Эти блоки могут быть обучены на больших или малых выборках в доверенной среде с учителем или в процессе функционирования без учителя практически любым алгоритмом, их задача — извлекать из классифицируемых образов наиболее значимую информацию. В общем случае блок извлечения признаков является зависимым от предметной области, так как для разных приложений входные данные могут кардинально отличаться, как и характер извлекаемой из образа информации (вектора признаков). Специфических требований к защите этих блоков от обозначенных выше атак можно не предъявлять, так как они не продуцируют классификационных решений и не хранят персональных данных, а также другой конфиденциальной информации после обучения. Выбор архитектуры блока извлечения признаков можно почти полностью возложить на разработчика искусственного интеллекта. Требования к блокам извлечения признаков могут касаться только свойств вектора признаков (например, обеспечить нормальное распределение значений каждого признака, что несложно осуществить, используя вариационный автокодировщик).

Блок защиты должен строиться по стандартизованной технологии с предъявлением жестких требований к архитектуре. Обучение блока защиты должно быть робастным. Этот блок будет функционировать в потенциально враждебной среде и должен обеспечивать конфиденциальность собственных знаний, а также высокую точность решений. Должна быть предусмотрена возможность закрепить ассоциацию между классом образов и заранее генерируемым криптографическим ключом.

Для реализации блока защиты не подходят нечеткие экстракторы с учетом описанных в настоящем отчете недостатков. Нейросетевые ПБК, обучаемые по ГОСТ Р 52.633.5 также не подходят. Длина ключа для нейросетевых ПБК жестко ограничена из-за подверженности атаке Маршалко (для 416 признаков длина ключа составляет 26 бит). Для сетей квадратичных нейронов атака Маршалко также актуальна, но длина ключа может быть примерно в 4 раза выше, чем для нейросетевых ПБК. Ограничения длины ключа не позволяют использовать такие нейроны для широкого спектра классификационных задач. Аналогичные выводы можно сделать относительно нейронов среднего гармонического.

Применение технологий глубокого обучения и аппарата многослойных нейронных сетей, по всей видимости, ограничивается блоком извлечения признаков. Блок защиты должен иметь простую архитектуру, такую, чтобы его синтез и обучение могло выполняться в автоматическом режиме. Стандартизованный алгоритм синтеза блока защиты должен быть универсальным (насколько это возможно) по отношению к предметной области, что практически невозможно, если в основе блока защиты использовать глубокие сети. Всю сложность и преимущества нейросетевой обработки и глубокого обучения нужно перенести в блоки извлечения признаков.

Для решения обозначенных проблем требуется разработать иные модели искусственных нейронов и сетей, позволяющих повысить длину ключа в соответствии с текущими требованиями (например, в соответствии с ГОСТ Р 34.10-2012 длина ключа электронной подписи должна составлять 256 бит) или выше (на перспективу), которые могут быть надежно защищены в ближайшем будущем. Это может быть выполнено усилиями технических комитетов 164 и 362. В настоящем отчета далее изложены основные принципы, на которых может быть построен новый стандарт по защите искусственного интеллекта и предлагается базовая модель искусственных нейронов и сетей для реализации блока защиты.

Безопасность блока защиты может быть дополнительно усилена криптографическими методами, если в этом будет необходимость. Для нейросетевого ПБК, обучаемого по ГОСТ Р 52633.5 уже разработана спецификация по криптографической защите решающих правил. Но не любые знания можно защитить при помощи данной спецификации. Для новой модели нейросетевых ПБК может потребоваться адаптация данной спецификации, разработка новой или вообще целесообразности в криптографической защите не будет.

**Список источников**