УДК 004.056

РАЗРАБОТКА АЛГОРИТМА ИНТЕРПРЕТАЦИИ МОДЕЛИ «ЧЕРНОГО ЯЩИКА СВЕРТОЧНОЙ НЕЙРОННОЙ СЕТИ С ПОМОЩЬЮ МЕТОДА ОЦЕНКИ ГРАДИЕНТОВ

Гаврилова В. В. (Университет ИТМО)
Научный руководитель – Менщиков А. А.
(Университет ИТМО)

Введение. Объяснимость становится актуальной темой исследований благодаря прогрессу в области глубокого обучения. Ежегодно в развитие искусственного интеллекта вкладываются миллионы рублей. Технологическая индустрия старается увеличить количество генеративных моделей искусственного интеллекта. Однако существует проблема, которую следует рассматривать с наибольшим приоритетом: недоверие к искусственному интеллекту. Согласно результатам исследования ВЦИОМ [1], за последний год общее понимание россиян в области искусственного интеллекта достигло 87%. Однако доверие к технологиям искусственного интеллекта составляет лишь 55%. Повысить уровень доверия людей можно за счет объяснимого искусственного интеллекта. Один из методов предоставления объяснительной информации — атрибуция признаков. Её цель — определить вклад входных признаков в результат модели и выявить наблюдения, поддерживающие её решение. Существуют два подхода к атрибуции: «белый ящик» и «чёрный ящик». Методы «белого ящика» предполагают полный доступ к модели и создание точных объяснений, исследуя поток градиентов. Однако на практике такой доступ невозможен из-за соображений безопасности. Интерпретация «чёрного ящика» требуют только доступ на уровне запросов, что позволяет анализировать взаимосвязь между входными признаками и выходами модели.

Основная часть. В исследовании решаются следующие задачи:

1. Произведен анализ существующих атак на нейронные сети с низким уровнем интерпретируемости (черные ящики)
2. Произведено исследование существующих методов интерпретации
3. Разработан метод интерпретации черного ящика сверточной нейронной сети
4. Выполнен сравнительный анализ результатов работы разработанного метода с другим методом интерпретации

В результате проведенного исследования был разработан алгоритм интерпретации модели черного ящика сверточной нейронной сети с помощью метода оценки градиента. Целевой моделью является сверточная нейронная сеть (CNN), которая состоит из двух сверточных слоев с размером ядра 5, объединенных тремя полносвязными слоями с размерами 120, 84 и 10 соответственно. Обучение модели осуществлялось с помощью библиотеки PyTorch. Алгоритм вычисляет атрибуты признаков, создавая запросы путём наложения масок на различные варианты объясняемых данных, равномерно распределённых между объясняемыми данными и начальной точкой. Используя сгенерированные запросы, интерпретатор получает набор наблюдений, позволяющий оценить атрибуты признаков.

В качестве данных для обучения и интерпретации были использованы наборы MNIST и ImageNet.

В качестве поискового распределения для метода используется гауссово распределение, а количество запросов $n*$ фиксировано и составляет 5000 для всех тестовых настроек. Отклонение σ, определяющее разброс гауссианы, задано равным 1,0 для MNIST, учитывая полярность распределения значений пикселей. Для ImageNet, где значения пикселей распределены более равномерно, σ установлен на 0,3. Что касается базовой линии $\tilde{x}$, то нулевая матрица используется при объяснении решений на полутоновых изображениях, в то время как базовая линия для ImageNet зависит от объяснимости экземпляра. Для каждого объясняемого экземпляра из ImageNet базовая линия — это размытая версия самой себя.

Чтобы объективно оценить эффективность интерпретаторов, необходимо оценить качество объяснений по известному методу «оценке через удаление». Процесс оценки следует интуитивной, но эффективной идее: удаление релевантных признаков должно вызывать большее падение уверенности в предсказании. При оценке через удаление пиксели удаляются последовательно в порядке убывания в соответствии с их оценками атрибуции. Тенденция изменения уверенности в предсказании рисует кривую на протяжении всего процесса удаления, а площадь над кривой возмущения (AOPC) рассматривается как метрика для количественной оценки эффективности объяснения.
По сравнению с методом интегрированных градиентов (является методом интерпретации белого ящика), предложенный алгоритм достигает схожих результатов, что согласуется с наблюдениями о визуальном сходстве их тепловых карт в качественной оценке. Для более простых тестовых случаев предложенный метод даже добивается лучших результатов, что следует интерпретировать как улучшение, вызванное более гладкой аппроксимацией интеграла пути.

Выводы. По результатам исследований был произведен анализ полученных результатов и их сравнение с работами других авторов. Точность интерпретации предложенного алгоритма не уступает точности современных методов интерпретации на наборе данных MNIST и уступает на 0.02% методу интерпретации белого ящика на наборе данных ImageNet.

Список использованных источников:

1. Национальные приоритеты URL: https://национальныеприоритеты.рф/news/informirovannost-i-doverie-rossiyan-tekhnologiyam-ii-povysilis-za-posledniy-god/#:~:text=Уверенная%20информированность%20о%20технологиях%20искусственного,и%20составил%2055%25. (дата обращения: 10.06.24).
2. Joakim Edin, «Normalized AOPC: Fixing Misleading Faithfulness Metrics for Feature Attribution Explainability» »– 2024 – URL: https://arxiv.org/abs/2408.08137

Гаврилова В. В. (автор) Подпись

Менщиков А. А. (научный руководитель) Подпись