Цифровые водяные знаки представляют собой специальную технологию, которая используется для защиты авторских прав на различные мультимедийные файлы, такие как изображения, документы, видео и аудио. Этот метод заключается в добавлении дополнительной информации к оригинальному цифровому файлу, что позволяет определить случаи копирования нейросетей и представления их чужими.

Однако применение водяных знаков в нейросетевых технологиях имеет свои сложности. Нейронные сети часто состоят из множества компонентов, что делает сложным отслеживание происхождения конкретных алгоритмов или кодовых фрагментов. Кроме того, украденные модели могут быть изменены злоумышленниками, что усложняет определение их истинного происхождения и владельца.

Для решения этой проблемы специалисты разработали уникальный подход к маркировке нейросетей. Они создали систему триггерных данных, которые внедряются в структуру нейронной сети и сохраняют свою функциональность даже после её кражи. Эти триггерные данные представляют собой набор информации, которая ассоциируется с определёнными предсказаниями нейросети, например, классификацией изображений, что позволяет уникально идентифицировать каждую модель.

Такой подход не только обеспечивает высокую эффективность в предотвращении кражи нейросетей, но и минимизирует вычислительные затраты, сохраняя при этом производительность модели. Российские цифровые водяные знаки оказались эффективными в 95% случаев, что значительно превышает результаты аналогичных разработок из США и Южной Кореи.