УДК 004.852

ИССЛЕДОВАНИЕ ЛОКАЛЬНЫХ МАКСИМУМОВ ФУНКЦИЙ ПРАВДОПОДОБИЯ НА ОСНОВЕ НЕЙРОННЫХ СЕТЕЙ

Рогачев К.О. (Университет ИТМО)

Научный руководитель – кандидат технических наук, доцент Платонов А. В.
(Университет ИТМО)

Аннотация. Существует подход к атаке нейронных сетей посредством подбора признаков максимизирующих значение функции правдободобия выбранного класса. Данная атака эксплуатирует свойство переобучения нейронных сетей, при котором локальные максимумы функции правдободобия целевого класса не соответствуют действительности. Данное исследование направленно на минимизацию возможности данного класса атак с использованием аналитического подхода.

Введение. Атака одного пикселя и атака шумом эксплуатирует свойство переобучения нейронной сети. Она происходит с помощью поиска локального максимума функции правдободобия выбранного класса. Существующие методы противодействия атаке используют либо добавление шума к входным в нейронную сеть признакам, либо проведение таковых атак и добавление их в датасет с правильным классом. Таким образом существующие методы затрудняют поиск значений признаков для подобных атак, но не борются с самой их возможностью.

Основная часть. Явление переобучения, эксплуатируемое атакой одного пискеля, является следствием использования модели излишней сложности для решения задачи. В случае с нейронными сетями это число нейронов и связей между ними. В работе исследуется аналитический способ уменьшения сложности нейронной сети, с попыткой минимизации ошибки первого рода. Для того чтобы это сделать, было использован метод нахождения локальных максимумов функций правдоподобия классов нейронной сети, обученной на наборе данных MNIST. Следующим этапом эти локальные максимумы были разбиты на соответствующие своему классу и не соответствующие. После этого был проведён анализ поведения нейронов при вычислении с помощью нейронной сети локальных максимумов.
Анализ заключался в определении нейронов вносящих наибольший вклад посредством высокого выходного значения нейрона. Аналогично анализировались связи, на предметы высокого значения после умножения на вес связи. Следующим этапом те нейроны и связи что вносили наибольший вклад в определение класса во время атак, но при этом малое при вычислении локальных максимумов соответствующих своему классу, были удалены. Данные действия изменили локальные максимумы функций правдоподобия классов и уменьшили сложность модели (нейронной сети).

Выводы. Реализованное программное обеспечение способно уменьшить размер нейронной сети, без серьёзной потери качества, и уменьшить возможность атак одного пикселя и шумом.

Рогачев К.О. (автор) Подпись

Платонов А.В. (научный руководитель) Подпись