Адверсальные Атаки на Классификаторы Символьных Последовательностей

Иван Фурсов

29 мая 2020 г.

Аннотация

Адверсальные атаки используют уязвимости моделей машинного и глубинного обучения: незначительные изменения входных данных могут заставить модель ошибиться. Большинство исследований посвящено атакам на изображения и другие дифференцируемые типы данных. Однако атаки на последовательности также могут принести вред, если им удастся обмануть модель.

Адверсальные атаки на категориальные последовательности — сложная задача, потому что целевая функция недифференцируема по входным данным и незначительные изменения могут привести к потери смысла последовательности. В этой работе мы решаем эти проблемы с помощью введения дифференцируемой функции потерь: используем языковую модель и гладкую функцию расстояния Левенштейна для генерации адверсальных примеров.

В итоге мы получаем семантически-правдоподобные последовательности, которые устойчивы к адверсальному обучению и адверсальным детекторам. Семантическая правдоподобность и устойчивость к защитам — это два свойства, которые тяжело получить для существующих алгоритмов генерации адверсальных примеров, потому что эти алгоритмы используют лингвистические особенности естественного языка. В этой работе мы демонстрируем алгоритмы, которые работают на разных типах данных: банковских транзакциях, медицинских данных и данных для естественного языка.

Научный руководитель: Бурнаев Евгений Владимирович, Доцент
Со-руководительЗайцев Алексей Алексеевич, Научный сотрудник