УДК 004.032.26

ЕСТЕСТВЕННЫЕ ШУМЫ МРТ И КТ СНИМКОВ КАК СПОСОБ СОСТЯЗАТЕЛЬНЫХ АТАК (ADVERSARIAL ATTACKS) В ГЛУБОКИХ НЕЙРОННЫХ СЕТЯХ И МЕТОДЫ БОРЬБЫ С НИМИ

Семилетов А.И. (Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики, Санкт-Петербург)

Научный руководитель – старший научный сотрудник, кандидат технических наук Гусарова Н.Ф.

(Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики, Санкт-Петербург)

В данной работе изучаются и сравниваются способы противодействия естественным шумам в КТ и МРТ снимках, которые могут приводить к эффекту, схожему с эффектом adversarial attack.

Введение. При обучении моделей нейронных сетей на медицинских данных возникает непредвиденный эффект влияния шумовых характеристик изображения. Неявные и не видимые человеческим глазом шумы аппарата на снимке могут вызвать эффект, схожий с эффектом adversarial attack, когда незначительное изменение отдельных пикселей изображения порождает adversarial признаки в структуре нейросети и полностью меняет решение модели относительно анализируемого снимка. При этом, в то время как за adversarial attack стоит злоумышленник, здесь эффект вызван естественными шумами аппарата. Такие снимки в работе называются adversarial examples.

Основная часть. Первый вопрос, изучаемый в работе: могут ли отдельные экземпляры реальных медицинских снимков выступать в роли adversarial examples при анализе с использованием нейронных сетей. Второй вопрос, заключается в следующем: можно ли максимально простым способом защитить себя от таких "естественных" атак? В работе опробованы следующие методы защиты: состязательная тренировка (adversarial training), гауссовская аугментация данных и ограничение RELU. Наконец, был изучен третий вопрос: возможно ли обрабатывать датасеты, содержащие adversarial examples как результат влияния естественных шумов оборудования, с помощью методов transfer learning, чтобы уменьшить влияние adversarial examples на точность классификаторов на основе нейронных сетей? На все эти вопросы в работе дан утвердительный ответ и приводится его обоснование.

Выводы. В работе показано, что "естественные" шумы в медицинских изображениях могут стать непроизвольными источником adversarial attack. Показано, что adversarial эффект возможен после применения состязательных методик обучения (adversarial training), но степень шума в таком образе будет значительно выше, чем до применения этих методик, и врачу будет достаточно просто распознать их визуально и исключить из дальнейшего рассмотрения. Наконец, рассмотрены способы противодействия эффекту adversarial examples в медицинских изображениях с использованием различных оборонных техник. С точки зрения компромисса между затратами и выгодой, наиболее предпочтительными являются методы feature squeezing и quantizing techniques.

Семилетов А.И. (автор) Подпись

Гусарова Н.Ф. (научный руководитель) Подпись