Семинар “Индустриальная математика”

Пятница, 4 декабря 2020, 18:15 (Moscow time, GMT+3)

Zoom ID: 820-7960-9196, password: ind

Злонамеренные атаки: почему они опасны для моделей последовательных данных?

Алексей Зайцев (Сколтех)

Злонамеренные атаки строятся на проработке различных сценариев уязвимости моделей глубинного обучения: незначительные изменения во входных данных могут привести к нарушениям в работе модели. После подачи на вход незначительно измененного в ходе атаки на модель входа, модель дает другой прогноз. Большинство современных атак работают в предположении, что на вход модели подается картинка.

Для моделей последовательных данных, таких как предложения на естественном языке, задача генерации атакующих входов для моделей сложнее. Она затруднена, например, тем, что в качестве входных данных в моделях используются токены из конечных множеств, и уверенность классификатора не дифференцируема. Таким образом, естественные градиентные атаки в таком пространстве входов модели невозможны.

Обычно сейчас атаки для таких данных генерируются на уровне токенов, однако возникающая при этом задача дискретной оптимизации требует существенных ресурсов, такие атаки легко детектировать. Вместо этого мы дообучаем языковую модель для генерации состязательных примеров. Дифференцируемая функция потерь в процессе дообучения зависит от уверенности суррогатного классификатора и дифференцированной оценки расстояния Левенштейна. При этом мы контролируем уровень состязательности генерируемой последовательности и ее сходство с исходной последовательностью.

Это позволяет формировать атакующие последовательности, семантически близкие к исходным. Более того, такие атаки устойчивы к дообучению с помощью выборок злонамеренных последовательностей и детектированию злонамеренных атак. Мы провели эксперименты на выборках из разных областей: банковских транзакций, электронных медицинских карт, обработки естественного языка. Проведенные эксперименты показывают, что наши модели работают лучше существующих аналогов, и защищаться от таких атак труднее. Данная работа написана в соавторстве с И. Фурсовым, Н. Ключниковым, А. Кравченко и Е. Бурнаевым.

Приглашаются все желающие!