Отзыв на доклад

Название доклада: Использование методов подсчёта неопределённости для борьбы с атаками на детекторы машинно-сгенерированного текста

Автор доклада: Леванов Валерий

Автор отзыва: Никитин Артем Анатольевич

Аннотация

В докладе рассматривается необходимость повышения надёжности детекторов, выявляющих машинно-сгенерированный текст. Автор отмечает, что многие существующие детекторы уязвимы к таким различным атакам, основанным на генерации. В качестве возможного решения предлагается использовать методы подсчёта неопределённости, которые ранее показывали хорошие результаты в других задачах обработки естественного языка, но мало исследовались в контексте классификации "Human vs Machine Generated". Предполагается применение White-box и Black-box подходов к оценке уверенности и обучение нейронных сетей, устойчивых к манипуляциям с текстом. По итогам исследования ожидается разработка робастного AI-детектора, способного надёжно отличать рукописные тексты от сгенерированных.

Комментарий

В докладе чётко сформулирована проблема: многие детекторы, созданные для определения машинной генерации текста, не справляются с адаптированными атаками. Методы подсчёта неопределённости, на которых был сделан акцент в докладе на семинаре, помогают оценивать уверенность модели, что может повысить робастность к манипуляциям. Отмечу отдельно, что презентация содержит обзор смежных решений в NLP и освещает потенциально полезный опыт из компьютерного зрения, что для меня подчеркнуло актуальность исследований и наличие существенных продвижений на данный момент. Во время презентации хотелось бы видеть более детализированный план по экспериментальному сравнению разных методов Uncertainty Estimation. Работа является актуальной в текущих реалиях в области борьбы с дезинформацией и повышения академической успеваемости.