УДК 004.85

Универсальные состязательные триггеры для атаки и анализа НЛП

С. Д. Трифонов[1]

[1]Московский физико-технический институт (национальный исследовательский университет)

Нейронные модели НЛП используются в широком спектре производственных систем, включая фильтры фальшивых новостей, домашних помощников и машинных переводчиков. Для многих из этих систем (например, для обнаружения фальшивых новостей или спама) "злоумышленники" будут пытаться обойти обнаружение модели или даже злонамеренно влиять на результаты модели. Данная работа требуется для усовершенствования защиты нейронных сетей. Немало зарубежных компаний и исследовательских центров проводят соревнования по взлому своих обученных моделей, тем самым находят их слабые места.

Моя исследовательская работа относится к атакам уклонения, схематично обобщенный алгоритм выглядит так:

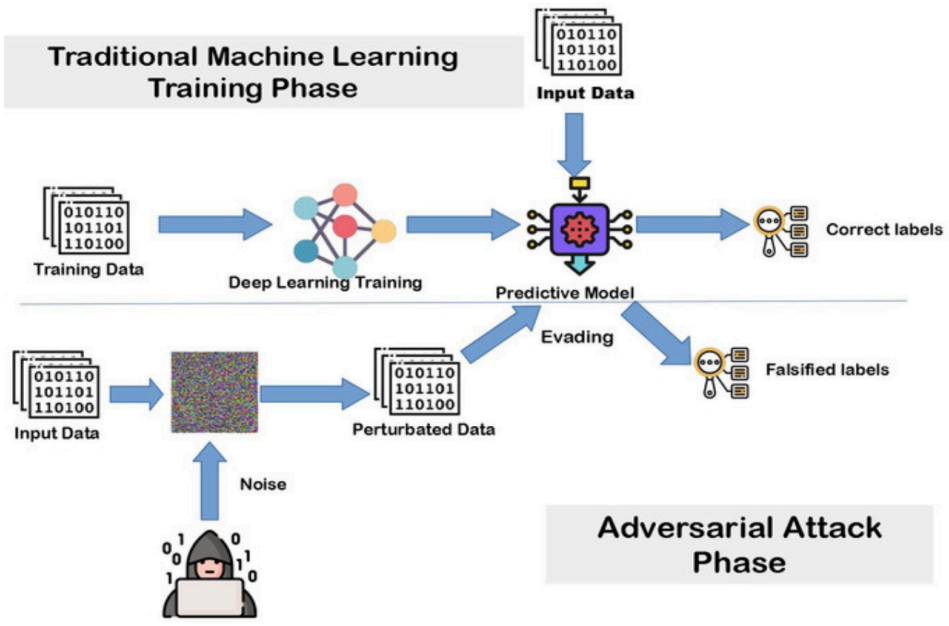

Рис. 1. Обобщенная модель атаки уклонения

В рамках работы было проведено исследование по поиску универсального триггера на модели GPT-2. Первым делом была проведена проверка модели,

генерация текста до атаки. Последовательно генерируя каждый токен, модель составляет достаточно логичную цепочку.

Следующим шаг - добавление атаки. Первым делом мы произвольным образом инициализируем нашу последовательность токенов, которую в дальнейшем назовем универсальный триггер. Далее последовательно для каждого токена генерируются возможные кандидаты (количество кандидатов можно варьировать). На каждом токене выбирается наилучший кандидат, который порождает минимальный loss. Проделывая данную операцию над каждым токеном, мы формируем универсальный триггер.
Проходить последовательность токенов можно не один раз, тем самым мы будем улучшать результат, как можно увидеть на графике:

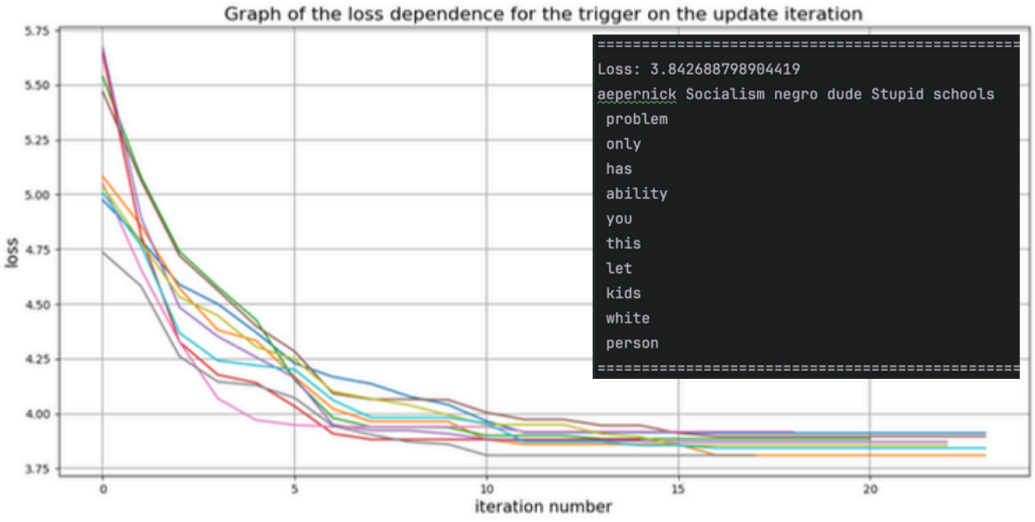

Рис. 2. График зависимости loss от номера итерации

На графике показано несколько рестартов(несколько запусков атаки). Можно заметить, что с увеличением итерации loss начинает принимать константное значение. Тем самым после написания атаки на модель, оставалась задача поиска оптимальных констант атаки, такие как: количество изменений в триггере (т.е. сколько раз будут сгенерированы кандидаты для каждого токена), количество генерируемых кандидатов, температура для сэмплирования и др. Проведя анализ, мы остановились на 50 обращениях к триггеру и на генерации 100 кандидатов для каждого токена при одном обращении. Такие данные также упомянуты в статье.

Литература

1. Eric Wallace [et al.]. Universal Adversarial Triggers for Attacking and Analyzing NLP, 2021, arXiv:1908.07125 [cs.CL].

2. Melika Behjati, Seyed-Mohsen Moosavi-Dezfooli, Mahdieh Soleymani Baghshah, and Pascal Frossard. 2019. Universal adversarial attacks on text classifiers. In ICASSP
3. Yonatan Belinkov and Yonatan Bisk. 2018. Synthetic and natural noise both break neural machine translation. In ICLR.
4. Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Srndi ˇ c, Pavel Laskov, Giorgio Giac- ´ into, and Fabio Roli. 2013. Evasion attacks against machine learning at test time. In ECML-PKDD