Разработка нейронной сети для повышения читабельности декомпилированного кода на языке Си

Кислов Константин Александрович*, Божко Артем Александрович,
Ременяко Владислав Денисович, Лялин Максим Андреевич
НИЯУ МИФИ,
*e-mail: *kostik_kislov@list.ru*

Аннотация

В большинстве случаев декомпилированный программный код трудно поддается анализу. В ходе работы над проектом была разработана и обучена модель нейронной сети для повышения читабельности декомпилированного кода на языке Си.

Ключевые слова: декомпиляция, анализ кода, нейронная сеть, языковая модель.

Проблема восстановления исходного кода из машинного возникает сравнительно часто: анализ кода, близкого к исходному, позволяет понять, как работает программа, какие данные использует, куда и что отправляет, а также какие в ней есть слабые места и как она реагирует на аварийные ситуации. Для выполнения данной задачи используются программы-декомпиляторы. Однако они генерируют трудночитаемый код, что затрудняет процесс его анализа. В связи с этим возникла идея использовать для решения данной проблемы технологии ИИ, а именно – модели трансформеры.

Была выдвинута соответствующая гипотеза: на основе современных технологий ИИ возможно создать и обучить нейронную сеть, способную облегчить процесс анализа преобразованного из машинного на язык Си программного кода.

Определены объект и предмет исследования: декомпиляция программного кода; применение языковых моделей для анализа программного кода.

Цель работы: разработка и обучение нейронной сети-трансформера для преобразования декомпилированного кода на языке Си в более читабельный.

В ходе работы были пройдены следующие этапы с достижением соответствующих результатов:

1) Найдены существующие эффективные декомпиляторы (RetDec, Ghidra, IDA);
2) Для обучения, анализа работы и модернизации архитектуры был найден исходный код модели-трансформера, способного переводить текст с немецкого на английский;
3) Собран структурированный датасет для тестирования декомпиляторов и анализа их работы (все тестовые данные сгруппированы: C-файл + exe-файл);
4) Найден объемный датасет для обучения модели, состоящий из исходного кода около 106 тысяч программ;
5) Проведены первичные тесты выбранных декомпиляторов, для генерации декомпилированного кода с целью обучения модели был выбран RetDec;
6) В качестве основы для модели нейронной сети было решено взять открытый исходный код OpenAIGPT для модернизации и обучения;
7) Настроена и автоматизирована работа RetDec с целью формирования части датасета, состоящей из декомпилированного кода, для обучения нейросети;
8) Выборка из исходного кода программ была скомпилирована, далее декомпилирована при помощи настроенного RetDec;
9) Были проведены первые этапы обучения нейросети, проанализированы первичные результаты ее работы, определены перспективы дальнейшего развития

Список литературы

1. Учебник по машинному обучению [Электронный ресурс] –Режим доступа: https://education.yandex.ru/handbook/ml
2. The Annotated Transformer [Электронный ресурс] – Режим доступа: http://nlp.seas.harvard.edu/2018/04/03/attention.html#background
3. **Какутин, Д. Ю.** Формирование и анализ эффективности выборки для обучения языковых моделей распознаванию и анализу исходного кода программ / Д. Ю. Какутин, А. С. Дмитриев // Инженерный вестник Дона. – 2022. - № 5 (89).