Разработка нейронной сети для повышения читабельности декомпилированного кода на языке Си

Выполнили: Кислов Константин Александрович, Божко Артем Александрович, Ременяко Владислав Денисович, Лялин Максим Андреевич

Наставник: Бехтин Артем Владимирович

Актуальность

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.

Исследовательская составляющая

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

Объект исследования – декомпиляция программного кода

Предмет исследования – применение языковых моделей для анализа программного кода

Методы исследования: поисковый, анализ, сравнение, измерение, тестирование, моделирование, программирование

Цель и задачи

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

Задачи:

1. Поиск < и анализ информации по данной теме >
2. Тестирование < декомпиляторов и анализ их работоспособности >
3. Создание < выборки для обучения нашей модели >
4. Разработка < собственной модели нейросети-трансформера и ее обучение >
5. Тестирование < «чернового» варианта модели и анализ результатов ее работы >
6. Модернизация < нейронной сети и ее дальнейшее обучение >
7. Подведение итогов и определение перспектив проекта

Средства и ресурсы:

<-Декомпилятор

<-Основа модели

<-Датасет

Наш GitHub

Результаты

(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.

To be continued…