УДК 004.855.5

ФАЗЗИНГ КОМПИЛЯТОРА С ОБУЧЕНИЕМ СЕМАНТИКЕ ЯЗЫКА С ПОМОЩЬЮ НЕЙРОННЫХ СЕТЕЙ АРХИТЕКТУРЫ TRANSFORMER

Петухов В.А. (Университет ИТМО, Санкт-Петербург)
Научный руководитель – Фильченков А. А., к.ф.-м.н., доц. факультета ИТиП университета ИТМО

В докладе поднимается вопрос применимости нейронных сетей архитектуры Transformer для задачи генерации кода на некотором языке программирования. Почти во всех промышленных языках нет формально описанной модели семантики, но некоторые её части обычно формализовать можно, что позволит существенно улучшить качество генерируемого кода, измеряемого в доле получаемого семантически корректного кода.

Введение.

Генерация семантически корректного программного кода является актуальной и на данный момент полностью не решенной задачей. На сгенерированном коде можно тестировать компилятор соответствующего языка, проверяя, не завершает ли он работу на таком коде с исключением, не компилирует ли такой код слишком долго или не сообщает ли об ошибке компиляции, когда компилятор некоторой предыдущей версии не сообщал. Такое тестирование является более эффективным, поскольку генератор в единицу времени может сгенерировать гораздо большее количество тестовых примеров, чем инженер-тестировщик.

Для генерации семантически корректного кода необходимо обучить используемую модель машинного обучения семантике соответствующего языка программирования. Правила семантики обычно довольно сложны и обучить им модель полностью является крайне сложной задачей. Но можно обучить модель некоторым хорошо формализуемым частям семантики, например, таким как правила подтипизации, и пытаться максимизировать количество семантически корректных (успешно компилируемых) сгенерированных примеров.

На текущий момент рядом ученых предложены алгоритмы, способные обучаться семантике языка и использующие предоставленную модель синтаксиса (грамматику). Например, в статьях «TreeGen: A Tree-Based Transformer Architecture for Code Generation» и «TreeGAN: Syntax-Aware Sequence Generation with Generative Adversarial Networks» предложены архитектуры нейронных сетей Transformer и GAN, которые работают с древовидными входными данными, что позволяет использовать грамматику языка.

Основная часть.

В рамках текущего исследования предлагается модифицировать механизм самовнимания архитекуры Transformer так, чтобы при подсчете коэффициента самовнимания учитывались такие семантические правила языка, как разрешение вызовов. Таким образом, нейронная сеть обращала бы особое внимание при обучении на соответствующие декларации для просматриваемых в данный момент вызовов.

Выводы. Предложенная модификация механизма самовнимания в нейронной сети архитектуры Transformer может существенно увеличить долю генерируемого семантически корректного кода, поскольку для любого вызова, чтобы быть успешно разрешенным, в программном коде в первую очередь важно наличие соответствующей.

Петухов В.А. (автор)

Фильченков А.А. (научный руководитель)