УДК 004.855.5

ГЕНЕРАЦИЯ ТЕСТОВ КОМПИЛЯТОРА С ИСПОЛЬЗОВАНИЕМ ГЕНЕРАТИВНЫХ МОДЕЛЕЙ, УЧИТЫВАЮЩИХ СЕМАНТИКУ ЯЗЫКА

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

В докладе поднимается проблема использования генеративных моделей для генерации кода на некотором языке программирования. К настоящему моменту существуют лишь модели, учитывающие синтаксис (грамматику) языка, но до сих пор неизвестно о попытках заложить в архитектуру генеративных моделей учитывание элементов семантики языка.

Введение.

Автоматическая генерация программного кода является довольно актуальной задачей. Это позволяет эффективно тестировать компиляторы языков, код на которых генерируется: нет необходимости вручную перебирать довольно большое число возможных комбинаций конструкций языка – эту задачу можно поручить генератору.

В большинстве случаев мы хотим генерировать семантически корректный код. Такое желание обусловлено тем, что, тестируя компилятор на семантически корректном коде, мы можем пройти по гораздо большему числу путей исполнения кода самого компилятора, следовательно, с большей вероятностью обнаружим ошибку. Тестируя компилятор на коде с какой-либо семантической ошибкой, мы получим информацию об этой ошибке от компилятора на раннем этапе, в частности минуя всю стадию генерации кода (то есть back-end часть компилятора в таком случае не будет тестироваться).

Генерация семантически корректного кода является не простой задачей, поскольку практически у всех популярных языков программирования нет полной формальной модели семантики (в отличие от модели синтаксиса – грамматики – которая есть у большинства языков). В связи с этим приходится прибегать к использованию искусственного интеллекта, алгоритмы которого бы обучались семантике. При том обучение синтаксису является нежелательным, так как модель синтаксиса, грамматика, обычно известна и её использование может быть заложено в архитектуру алгоритма.

На данный момент научным сообществом предложен ряд алгоритмов, способных обучаться семантике языке и учитывающих заранее предоставленную модель синтаксиса. Так, в статьях «TreeGen: A Tree-Based Transformer Architecture for Code Generation» и «TreeGAN: Syntax-Aware Sequence Generation with Generative Adversarial Networks» предложены работающие с деревьями архитектуры нейронных сетей Transformer и GAN. Древовидность архитектур позволяет естественным образом работать с грамматикой языка. Тем не менее часто есть возможность формально описать некоторые части семантики языка, и тем самым существенно увеличить долю семантически корректного генерируемого кода.

Основная часть. В рамках данного исследования предлагается разработать модификации архитектур нейронных сетей Tree-based Transformer и tree-GAN, учитывающие некоторые элементы семантики языка. В качестве первой итерации предлагается заложить в архитектуру учитывание типовой информации о выражениях – это поможет генерировать участки кода только с переменными доступных типов.

Выводы. Предложенные модификации архитектур нейронных сетей Tree-based Transformer и tree-GAN могут существенно увеличить долю получаемого семантически корректного кода, так как в большинстве конструкций языка так или иначе фигурируют типы и правила типовой

системы – их учитывание позволит увеличить шанс получения семантически корректного кода.

Петухов В.А. (автор)

Фильченков А.А. (научный руководитель)