УДК 004.912

РАЗРАБОТКА ПОДХОДА К РОУТИНГУ МОДЕЛЕЙ ДЕТЕКЦИИ КЛОНОВ КОДА И ИССЛЕДОВАНИЕ МЕТОДОВ ПОВЫШЕНИЯ ЭФФЕКТИВНОСТИ ИНФЕРЕНСА ПРИ СОХРАНЕНИИ КАЧЕСТВА ДЕТЕКЦИИ

Комлев Д. А.
Научный руководитель – канд. пед. наук, доцент Толстых О. М.
НИТУ МИСИС
m1902452@edu.misis.ru

Введение

В процессе разработки крупных программных систем одна и та же функциональность может быть реализована повторно в различных частях приложения. Такие повторяющиеся фрагменты принято называть клонами кода. Их наличие осложняет сопровождение и развитие системы, поскольку при внесении изменений существует риск модификации лишь одной из реализаций, тогда как другие варианты остаются неизменными, что приводит к несогласованному поведению. Дополнительную сложность представляет то, что клоны могут различаться по синтаксису, структуре и используемым именам сущностей, вследствие чего их автоматическое выявление становится сложной задачей. Существующие методы поиска клонов кода либо основаны на статических правилах и демонстрируют низкое качество детекции, либо обеспечивают высокое качество за счёт тяжелых нейросетевых моделей, но требуют значительных вычислительных ресурсов, что ограничивает их применение в промышленной разработке. Целью работы является разработка подхода к роутингу моделей детекции клонов кода, обеспечивающего повышение эффективности вычислений при сохранении качества детекции на уровне современных наилучших решений.

Основная часть

В работе рассматривается задача детекции клонов. В качестве базовых моделей используются CodeBERT [1] и GraphCodeBERT [2]. Обе модели дают векторное представление кода. CodeBERT требует меньших вычислительных затрат, GraphCodeBERT при построении представления использует граф потока данных (DFG) кода и при больших вычислительных затратах обеспечивает более высокое качество детекции клонов на сложных фрагментах. Обе модели применяются в связке с классификатором по парам эмбеддингов.

Предлагаемый подход основан на роутинге: для каждой пары вычисляется предсказание с помощью CodeBERT. Оценивается уверенность модели в результате. Для фиксированной доли уверенных примеров итоговое предсказание остаётся за CodeBERT. Для остальных примеров применяется GraphCodeBERT. Таким образом тяжелая модель применяется только к небольшой части данных, что снижает вычислительные затраты, при сохранении качества.

Выводы

Применение предложенного подхода на тестовой выборке датасета BigCloneBench [3] показало прирост по F1 около 10 п.п. относительно использования только CodeBERT. На подмножестве неуверенных примеров использование GraphCodeBERT дало прирост около 21 п.п. по F1 относительно CodeBERT. Суммарное время при роутинге составило около 1,8 от времени полного прогона CodeBERT против около 2,6 при полном прогоне GraphCodeBERT, то есть выигрыш по времени порядка 30 %. Качество роутинга при этом практически совпадает с качеством полного

GraphCodeBERT (снижение менее 0,4 п.п. по F1). Таким образом, подход обеспечивает более высокое качество, чем базовый CodeBERT, при существенно меньших вычислительных затратах, чем при повсеместном применении GraphCodeBERT.

Литература

1. Feng Z., Guo D., Tang D., Duan N., Feng X., Gong M., Shou L., Qin B., Liu T., Jiang D., Zhou M. *CodeBERT: A Pre-Trained Model for Programming and Natural Languages* [Электронный ресурс]. – arXiv preprint arXiv:2002.08155, 2020. – Режим доступа: https://arxiv.org/abs/2002.08155. – (На англ.) (Дата обращения 05.10.2025).
2. Guo D., Ren S., Lu S., Feng Z., Tang D., Liu S., Zhou L., Duan N., Svyatkovskiy A., Fu S., Tufano M., Deng S. K., Clement C., Drain D., Sundaresan N., Yin J., Jiang D., Zhou M. *GraphCodeBERT: Pre-training Code Representations with Data Flow* [Электронный ресурс] // Proceedings of the 9th International Conference on Learning Representations (ICLR 2021). – 2021. – Режим доступа: https://arxiv.org/abs/2009.08366. – (На англ.) (Дата обращения 07.10.2025).
3. Svajlenko J., Roy C.K. Evaluating clone detection tools with BigCloneBench // Proc. 31st International Conference on Software Maintenance and Evolution (ICSME). IEEE, 2015. P. 131–140. (Дата обращения 15.01.2026).