Языковые модели для кода

Бакалова Александра, 191

Почему языковые модели?

- Не обязательно обучать модель грамматике языка - она научится ей сама.
- Могут тренироваться на большом объеме данных и обучаться “человеческому” стилю кода.
- Можно задавать на вход описание на человеческом языке.

Модели, основанные на BERT

- CuBERT
- CodeBERT
- PyMT5

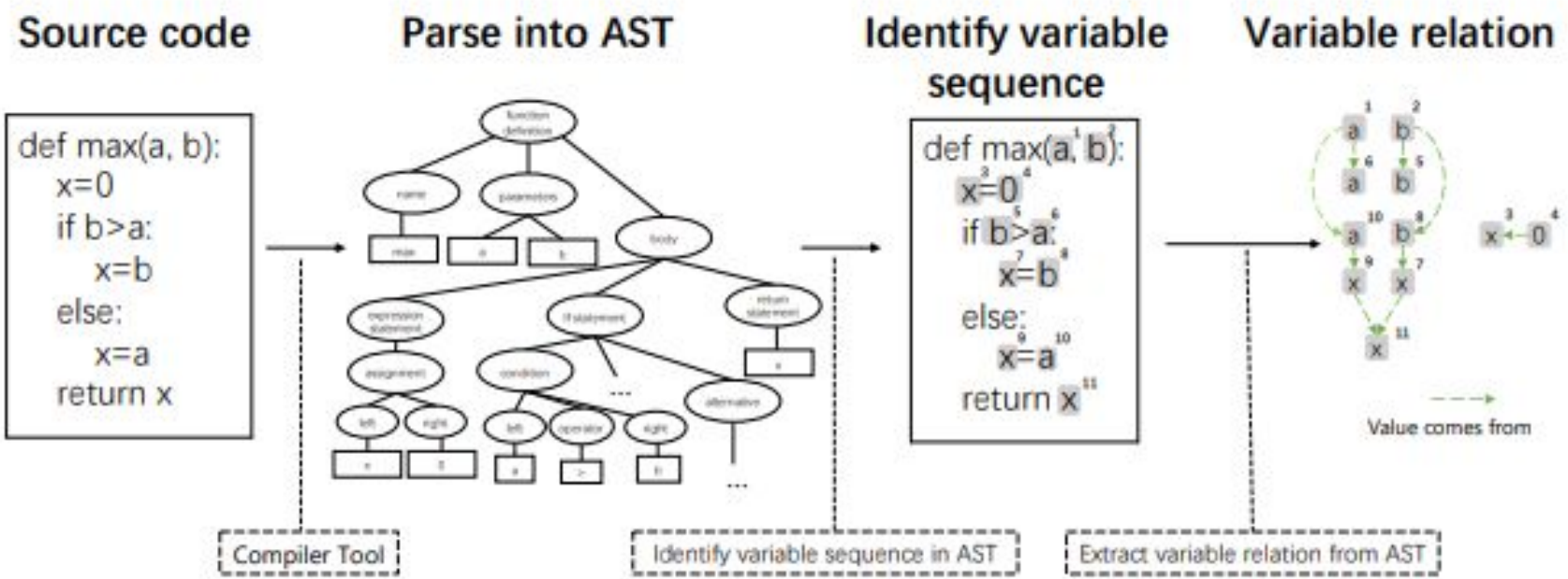
Codex

Доля верных ответов на HumanEval: 28.81%

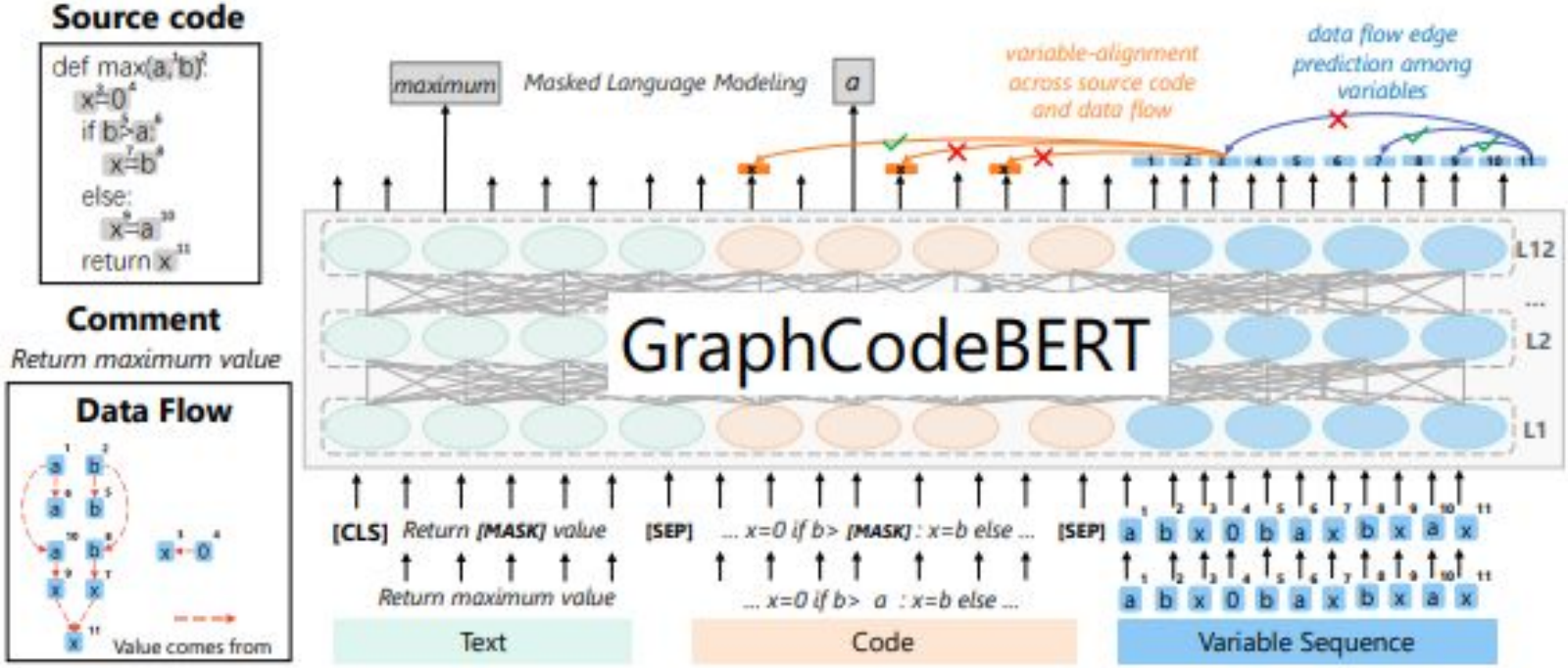
GraphCodeBERT

- Использует семантическую структуру кода.
- Показывает state-of-the-art результаты в задачах code search, clone detection, code translation и code refinement.

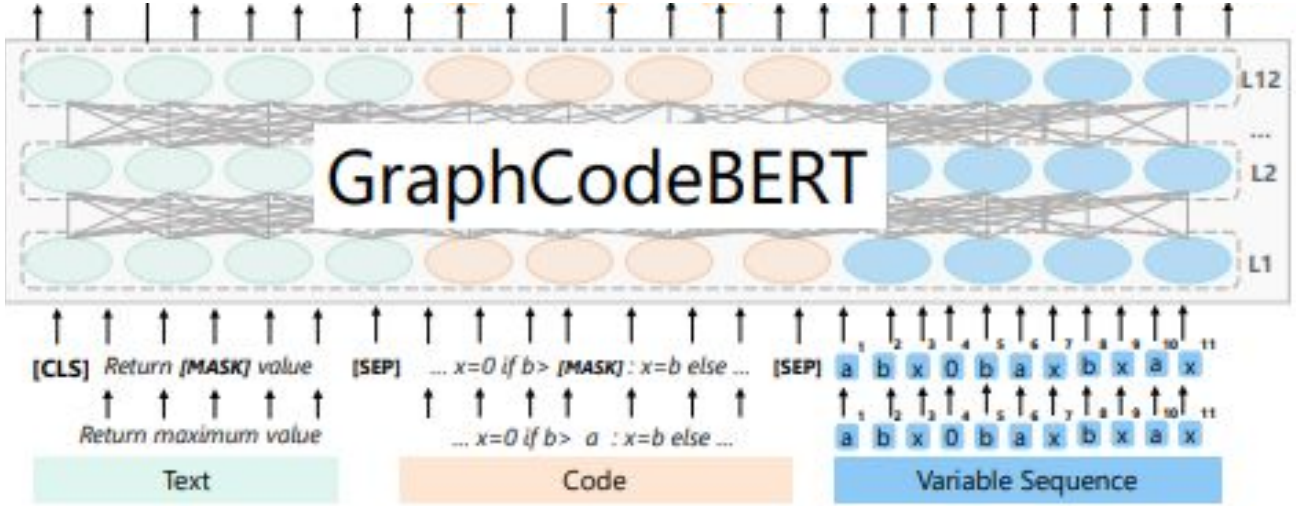
GraphCodeBERT: создание data flow



GraphCodeBERT: архитектура



GraphCodeBERT: кодирование графовой структуры



$$M_{ij} = \begin{cases} 0 & \text{if } q_i \in \{[CLS], [SEP]\} \text{ or } q_i, k_j \in W \cup C \text{ or } \langle q_i, k_j \rangle \in E \cup E' \\ -\infty & \text{otherwise} \end{cases}$$

текст

код

ребра data flow

токен вершины и вершина

Источники

PYMT5: https://arxiv.org/pdf/2010.03150.pdf

CodeBERT: https://arxiv.org/pdf/2002.08155v4.pdf

CuBERT: https://arxiv.org/pdf/2001.00059v3.pdf

Генерация кода с помощью трансформеров:
https://arxiv.org/pdf/2108.07732.pdf , https://arxiv.org/pdf/2105.09938.pdf

Codex: https://arxiv.org/pdf/2107.03374.pdf

GraphCodeBERT: https://arxiv.org/pdf/2009.08366.pdf