МАШИННОЕ ОБУЧЕНИЕ ДЛЯ АНАЛИЗА ИСХОДНОГО КОДА

Студент: Жихарева Александра

Группа: M9102

Руководитель: Кленин А. С

СТАТИЧЕСКИЙ АНАЛИЗ КОДА

- Выявление ошибок
  - неопределенное поведение
  - логические ошибки
- Рекомендации по оформлению кода
- Подсчет метрик

ПРИМЕНЕНИЯ МАШИННОГО ОБУЧЕНИЯ

- Поиск ошибок
- Трансляция и перевод
- Авто дополнение
- Тематическое моделирование
- Классификация исходных кодов
- Обнаружение совпадающих исходных кодов

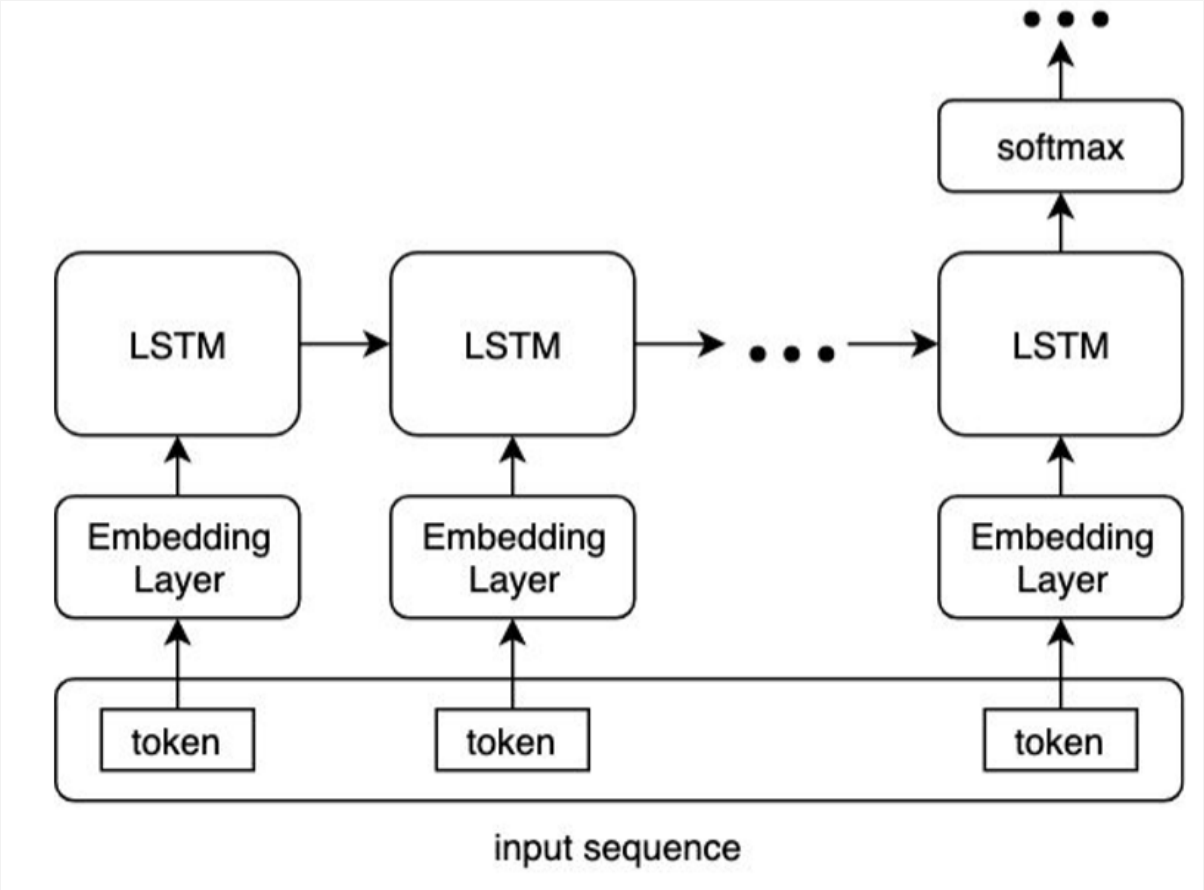
ПОИСК ДАННЫХ

- Попытки тестирующих систем
- Open-source проекты
- Существующие размеченные наборы данных
    - GitHub duplicate repositories
    - card2code
    - GitHub Java Corpus
- Генерация данных

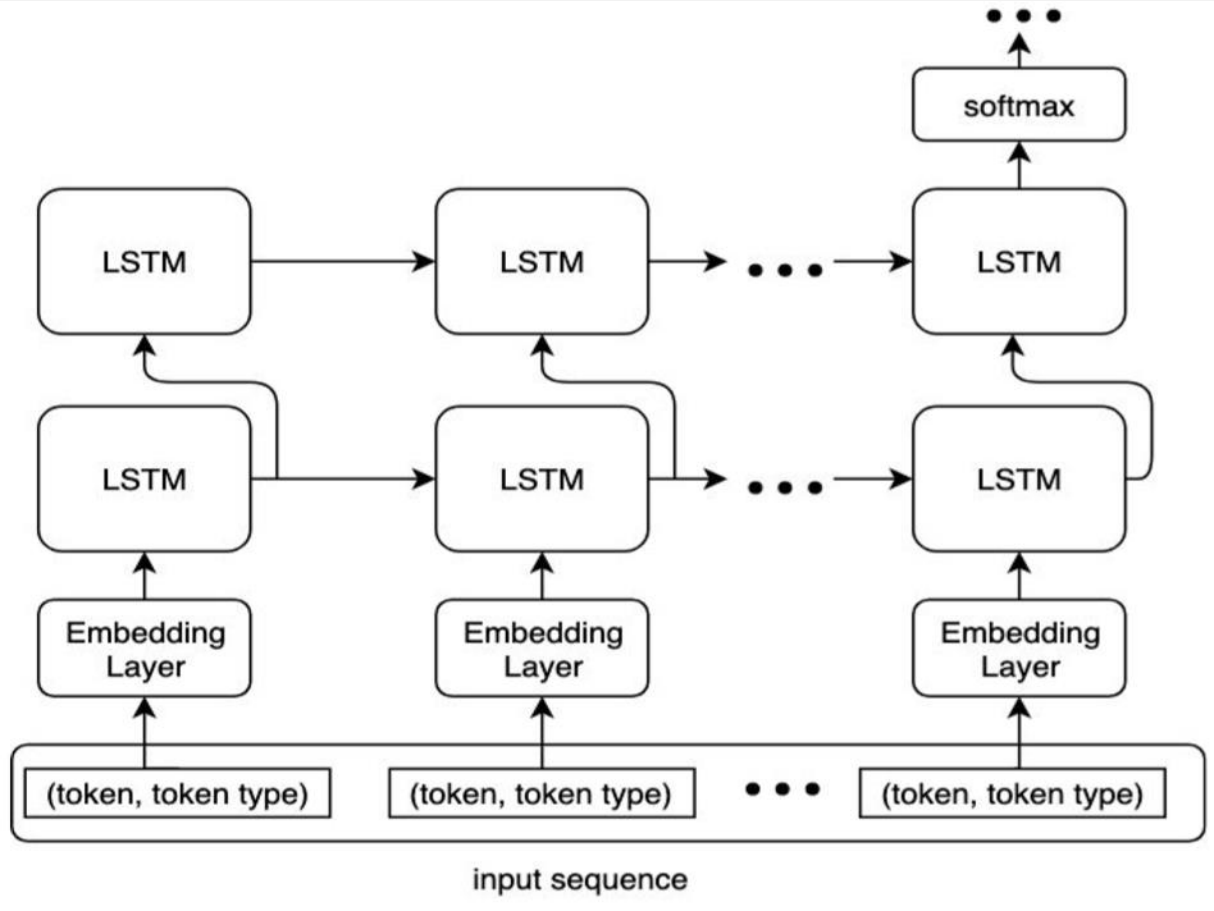
КЛАССИФИКАЦИЯ: ПРЕДОБРАБОТКА ДАННЫХ

- Посылки с codeforces, на языке C++, получившие вердикт OK
- Токенизация (lex), выкидываем комментарии
- Редкие токены -> UNK

КЛАССИФИКАЦИЯ



КЛАССИФИКАЦИЯ



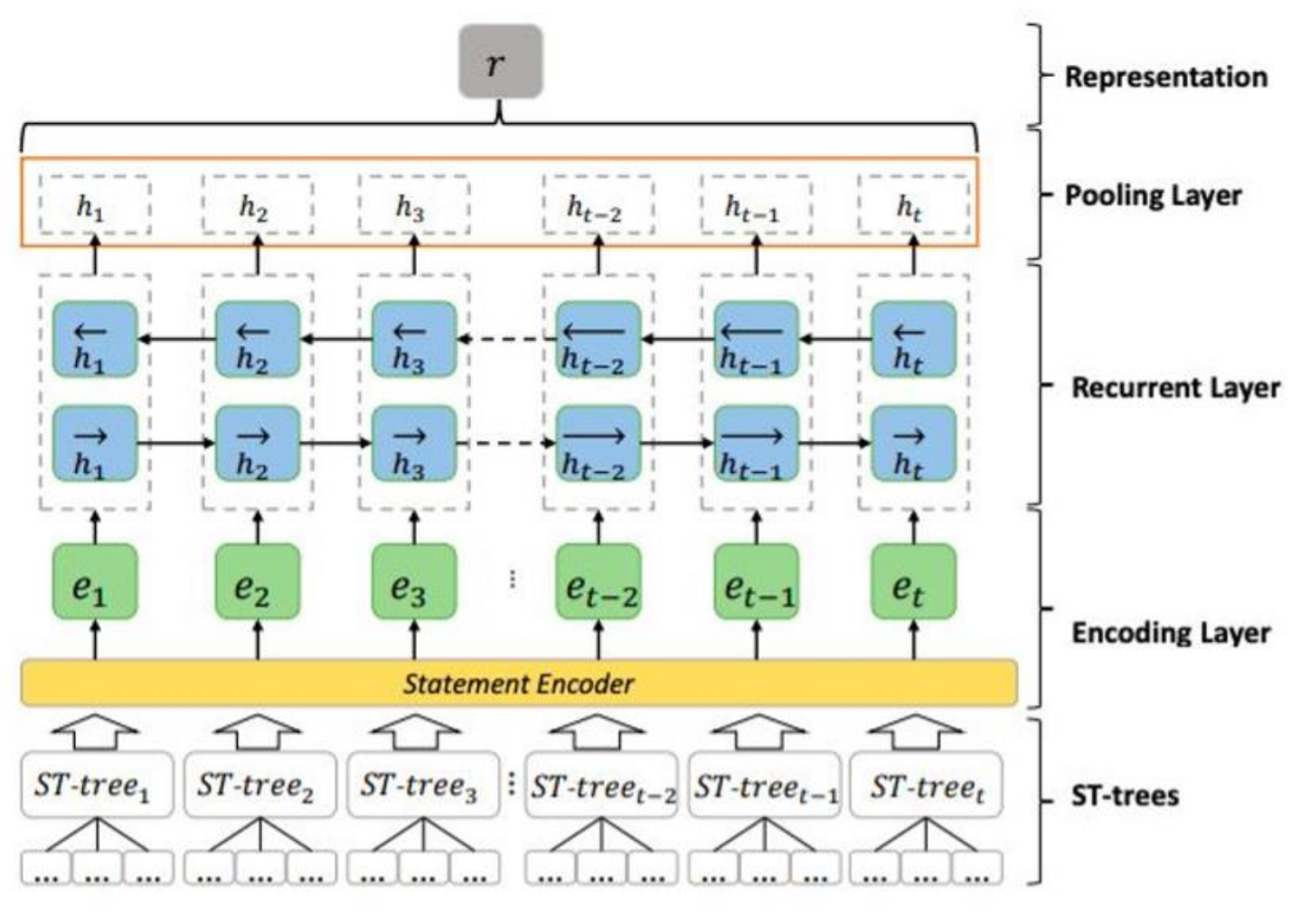
КЛАССИФИКАЦИЯ

- Классификация по задаче
  - 10 классов – точность 99%
  - ~150 классов – точность 95%
- Классификация по тегу
  - 4-6 классов – точность 84-86%

ПЛАНЫ

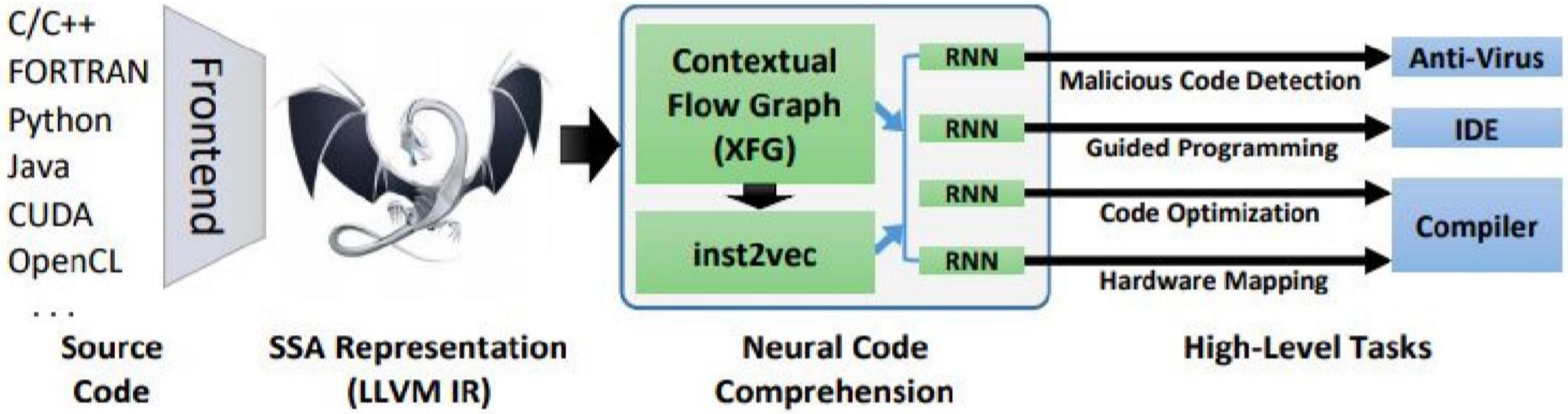
- Генерация фич из разных представлений исходников (seq, AST, CFG, intermediate representation)
- Code semantic segmentation
- Self-attention

МОДЕЛЬ ASTNN



- http://xuwang.tech/paper/astnn_icse2019.pdf

МОДЕЛЬ INST2VEC



- https://arxiv.org/pdf/1806.07336.pdf