Formal Language Driven Data Analysis Research Group Report

Semyon Grigorev

Saint Petersburg State University

February 29, 2024

Research Landscape

- AI-guided symbolic execution
  - Ekaterina Shemetova
  - Maxim Nigmatulin
  - Anna Chistiakova
  - Semyon Grigorev
  - Collaboration with Dmitriy Mordvinov and Vadim Lomshkov
- Parsing Techniques
  - Ivan Lomikovskiy
  - Semyon Grigorev
  - Collaboration with Kirill Prazdnikov and Pavel Pertsev

$\mathcal{AI}$-Guided Symbolic Execution

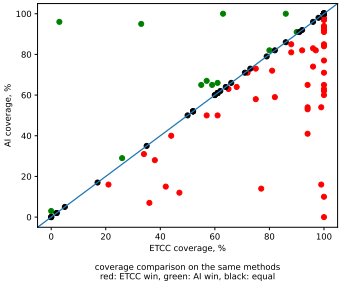
- ✔ Reusable infrastructure for training developed and implemented
  - Wrapper for SVM to convert it to server
  - Python client — $\mathcal{AI}$ agent to training
  - Basic manipulation with neural networks
- ✔ Basic dataset: train/validation/test
- ✔ Basic performance tuning
- ✔ First version of $\mathcal{AI}$ agent which guide SVM
  - Can be used in various engines: we use language-independent features
  - With tuning
- ⚙ Dataset extension
- ⚙ GNN quality improvement
- ⚙ Training infrastructure improvement: performance, flexibility, documentation, ...

$\mathcal{AI}$-Guided Symbolic Execution: Preliminary results

- 190 methods: algorithms, data structures, real-world projects
- AI — our AI-based agent
- ETCC (ExecutionTreeContributedCoverage) — one of the best algorithmic strategies

$\mathcal{AI}$-Guided Symbolic Execution: Preliminary results

- 190 methods: algorithms, data structures, real-world projects
- AI — our AI-based agent
- ETCC (ExecutionTreeContributedCoverage) — one of the best algorithmic strategies

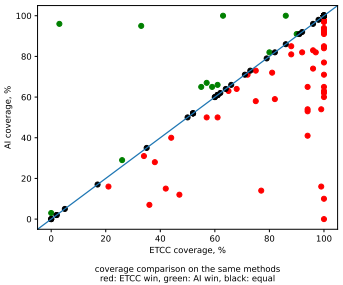


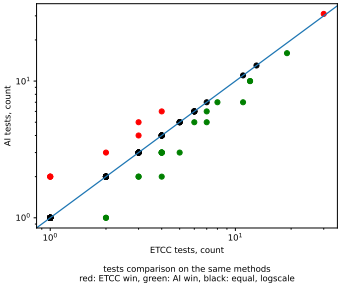
coverage comparison on the same methods
red: ETCC win, green: AI win, black: equal

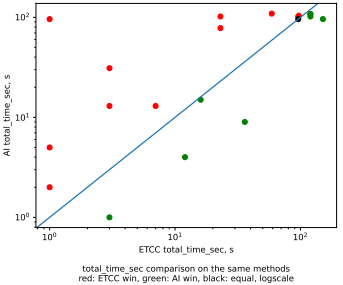
- Same coverage: 132, AI less: 46, AI more: 12

$\mathcal{AI}$-Guided Symbolic Execution: Preliminary results

- 190 methods: algorithms, data structures, real-world projects
- AI — our AI-based agent
- ETCC (ExecutionTreeContributedCoverage) — one of the best algorithmic strategies







- Same coverage: 132, AI less: 46, AI more: 12
- For same coverage: less tests, comparable time

Parsing Techniques

- ✔ Basic parser development tool is created
  - ✔ Preliminary performance evaluation
- ✔ Error recovery mechanism
  - ⚙ Preliminary performance evaluation
- ⌛ Advanced incremental parsing
- ⌛ Advanced scannerless mode

Parsing Techniques: Preliminary Evaluation Result

- Java grammar
- 3 real-world projects
  - junit4: 425 files, avg. size 3KB (40KB max)
  - guava: 1 416 files, avg size 8KB (198KB max)
  - elasticSearch: 14 685 files, avg size 6KB (242KB max)