Open Vocabulary Learning on Source Code with a Graph-Structured Cache

Milan Cvitkovic
Caltech, Amazon Web Services

Badal Singh
Amazon Web Services

Anima Anandkumar
Caltech

ICML, 2019-6-12

Open Vocabulary Learning

Goal: Models that can reason over flexible sets of inputs and outputs

Standard, closed vocabulary model

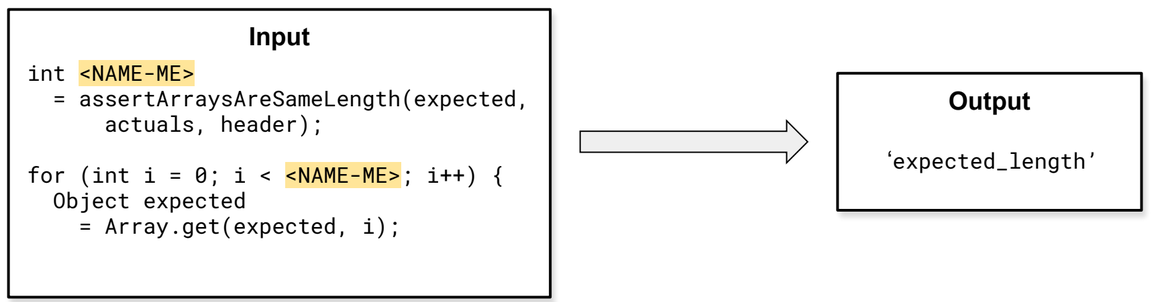
1 of 400k word embeddings → 1 of 400k words

Open vocabulary

Any words → Any words

Open Vocabulary Learning

Motivation: Tasks on source code
Example: Variable naming



Needs an open vocabulary

In our data, 28% of variable names contain out–of–vocabulary word

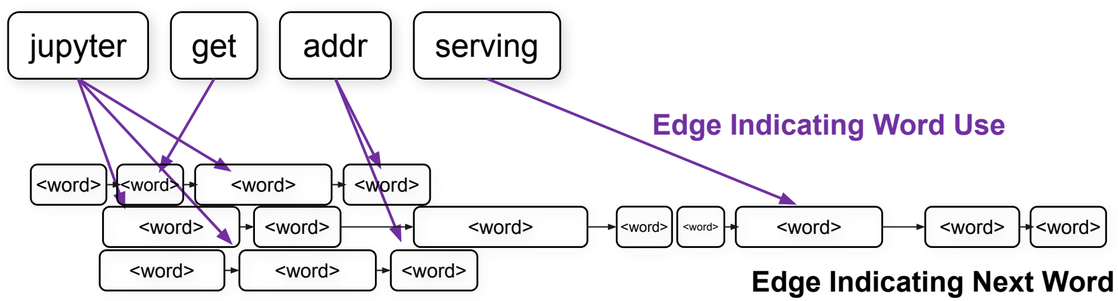
Graph-Structured Cache

Strategy: Represent distinct words and usages with graph structure, process with GNN

Original input

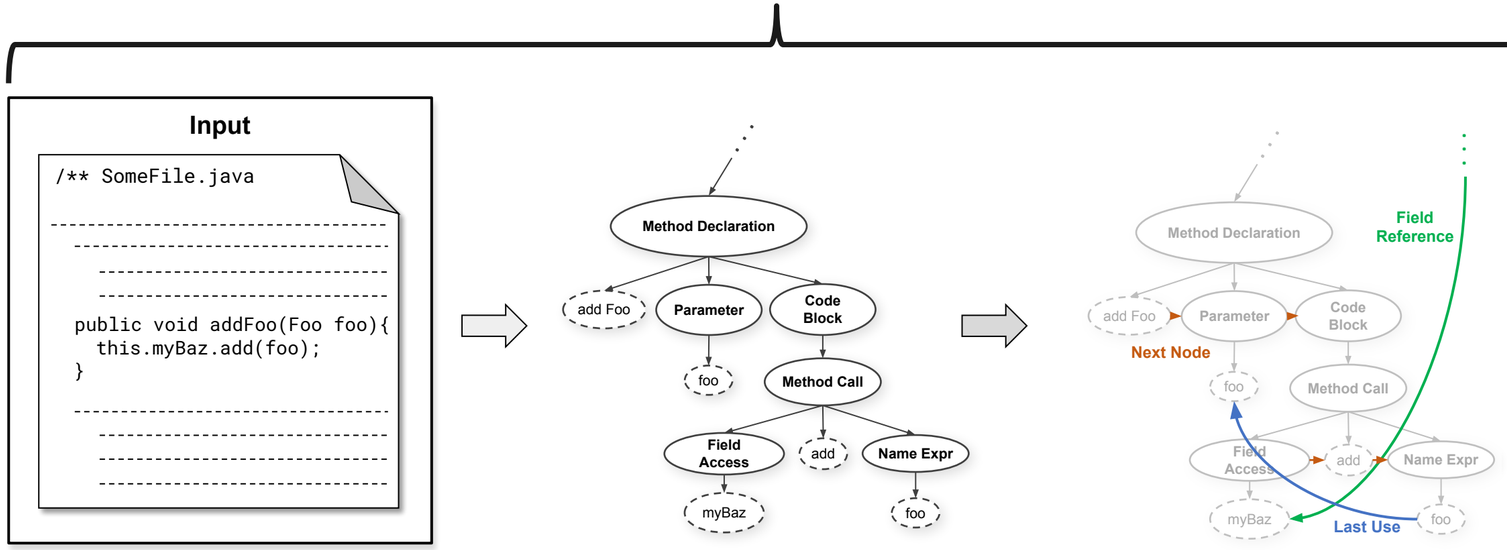
```
def get_jupyter_addr():
    jupyter_addr = 'localhost' if is_serving() else None
    return jupyter_addr
```

Same input, represented using a Graph-Structured Cache



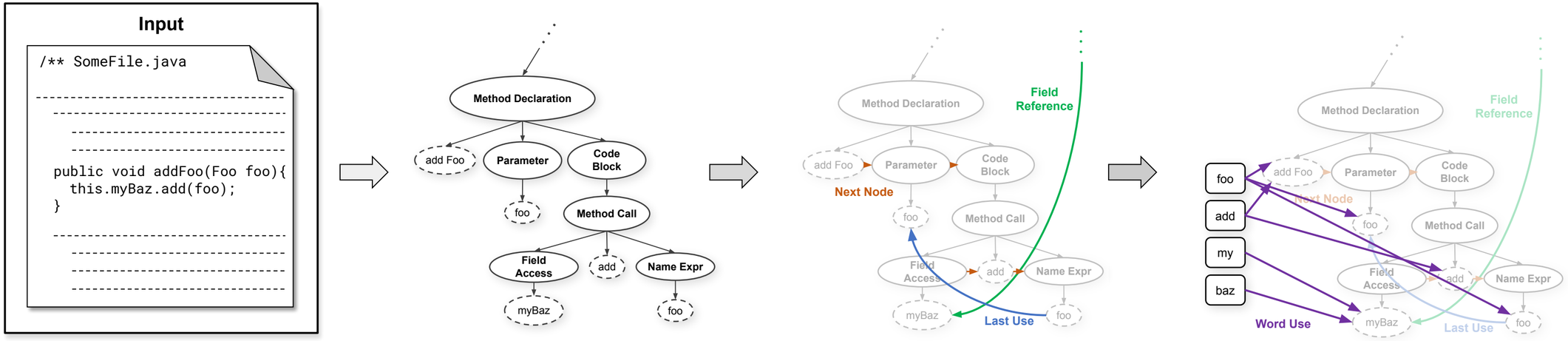
Full Model for Tasks on Source Code

Strategy from recent work [1]



[1] Allamanis et al. “Learning to Represent Programs with Graphs.” ICLR 2018

Full Model for Tasks on Source Code



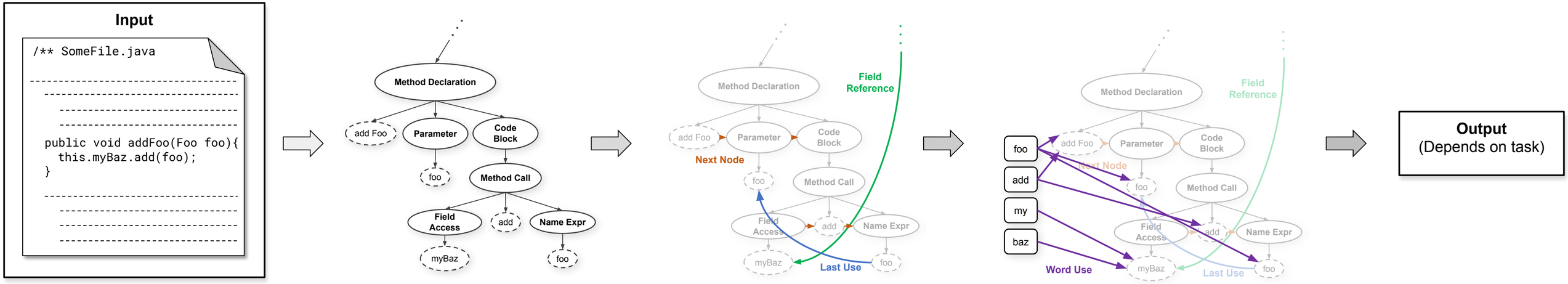
Parse code into AST

Augment AST with semantic information

Add Graph-Structured Cache

Our main contribution to prior work

Full Model for Tasks on Source Code



Experiment: Variable Naming Task

- Full-name reproduction accuracy (and top 5 accuracy):

		Closed Vocab	CharCNN	Pointer Sentinel	GSC
Seen repos	AST	0.23 (0.31)	0.22 (0.28)	0.19 (0.33)	0.49 (0.67)
	AugAST	0.19 (0.26)	0.20 (0.27)	0.26 (0.40)	**0.53** (0.69)
Unseen repos	AST	0.05 (0.07)	0.06 (0.09)	0.06 (0.11)	0.38 (0.53)
	AugAST	0.04 (0.07)	0.06 (0.08)	0.08 (0.14)	**0.41** (0.57)

For other tasks and experiments, see our poster or paper

Takeaways

Graph-Structured Caches are an appealing strategy for open vocabulary learning

- Whatever your current embedding strategy, GSC + GNN can augment it
- No free lunch! About 30% training slowdown.
- But helps in all cases we tried, sometimes significantly

Acknowledgments

- Badal Singh, Anima Anandkumar
- Miltos Allamanis
- Hyokun Yun
- Haibin Lin

Our code, for use on your code

https://github.com/mwcvitkovic/Open-Vocabulary-Learning-on-Source-Code-with-a-Graph-Structured-Cache--Code-Preprocessor

https://github.com/mwcvitkovic/Open-Vocabulary-Learning-on-Source-Code-with-a-Graph-Structured-Cache