Better Modeling the Programming World with Code Concept Graphs-augmented Multi-modal Learning

Martin Weyssow, Houari Sahraoui, Bang Liu
Context & Motivation
State-of-the-art learning paradigm

1. Multilingual pre-training
2. Fine-tuning & tasks

- Left-to-right code generation – CodeGPT, Codex
- Code search – CodeBERT
- Code summarization – CodeT5, PLBART

Transformer-based model

Task-specific fine-tuning

code vector representations
State-of-the-art learning paradigm

State-of-the-art deep learning models of code show great robustness and generalization on different data distribution and tasks.

They
- Learn from syntax → sequences of code tokens
- Learn from global information → self-attention layers
State-of-the-art learning paradigm

State-of-the-art deep learning models of code show great robustness and generalization on different data distribution and tasks.

They
- Learn from syntax → sequences of code tokens
- Learn from global information → self-attention layers

But
- Programs are concept-based → concept graphs
- Programs are multimodal! → learn from multiple level of abstraction
Can we learn better representations of code?
Multimodal learning with code concept graphs

Proposed approach. Joint-learn a BERT-like language model (LM) and a graph neural network (GNN).

Applicability to code search
- find the best code candidate given a natural language query
- simple fusion of code and concept graph representations
- improvement over CodeBERT, a strong baseline
Discussion

Can we have large multimodal pre-trained models of code?
- learn general-purpose code representations

and/or, should we fine-tune existing pre-trained models in a multimodal fashion?
- define modalities specific to the task

**Scaling law** → more parameters = higher accuracy on the tasks at hand
- Can we observe the same trend with multimodal learning?
- How does multimodal learning compare to unimodal learning *w.r.t* this law?
- Are the data processing and fusion worth the time/resources? → trade-off
Better Modeling the Programming World with Code
Concept Graphs-augmented Multi-modal Learning

Martin Weyssow
DIRO, University of Montreal
martin.weyssow@umontreal.ca

Houari Sahraoui
DIRO, University of Montreal
sahraouh@iro.umontreal.ca

Bang Liu
DIRO & Mila, University of Montreal
bang.liu@umontreal.ca