Engineering Village™

1. **CODE-MVP: Learning to Represent Source Code from Multiple Views with Contrastive Pre-Training**
Wang, Xin; Wang, Yasheng; Wan, Yao; Wang, Jiawei; Zhou, Pingyi; Li, Li; Wu, Hao; Liu, Jin **Source:** *arXiv*, May 4, 2022; **E-ISSN:** 23318422; **DOI:** 10.48550/arXiv.2205.02029; **Publisher:** arXiv

Author affiliation:
School of Computer Science, Wuhan University, China
Huawei Noah's Ark Lab.
School of Computer Sci. & Tech., Huazhong University of Science and Technology, China
Faculty of Information Technology, Monash University, Australia
School of Information Science and Engineering, Yunnan University, China

Abstract:
Recent years have witnessed increasing interest in code representation learning, which aims to represent the semantics of source code into distributed vectors. Currently, various works have been proposed to represent the complex semantics of source code from different views, including plain text, Abstract Syntax Tree (AST), and several kinds of code graphs (e.g., Control/Data Flow Graph). However, most of them only consider a single view of source code independently, ignoring the correspondences among different views. In this paper, we propose to integrate different views with the natural-language description of source code into a unified framework with Multi-View contrastive Pre-training, and name our model as CODE-MVP. Specifically, we first extract multiple code views using compiler tools, and learn the complementary information among them under a contrastive learning framework. Inspired by the type checking in compilation, we also design a fine-grained type inference objective in the pretraining. Experiments on three downstream tasks over five datasets demonstrate the superiority of CODE-MVP when compared with several state-of-the-art baselines. For example, we achieve 2.4/2.3/1.1 gain in terms of MRR/MAP/Accuracy metrics on natural language code retrieval, code similarity, and code defect detection tasks, respectively.

Main Heading: Computer programming languages **Controlled terms:** Abstracting - Machine learning - Semantics - Trees (mathematics)

Uncontrolled terms: Abstract Syntax Trees - Code graphs - Code representation - Control data flow graphs - Language description - Multiple views - Natural languages - Plain text - Pre-training - Source codes

Classification Code: 723.1.1 Computer Programming Languages - 903.1 Information Sources and Analysis - 921.4 Combinatorial Mathematics, Includes Graph Theory, Set Theory

Database: Compendex