Order Matters: Semantic-Aware Neural Networks for Binary Code Similarity Detection

Introduction

Info

Authors: Zeping Yu, Rui Cao, Qiyi Tang, Sen Nie, Junzhou Huang, Shi Wu. AAAI 2020.

Target

二进制代码相似性检测。

Motivation

传统方法慢且不准确，基于神经网络的方法（Genemi）无法捕获二进制代码的语义信息。

Contribution 套话，此处省略。

Methodology

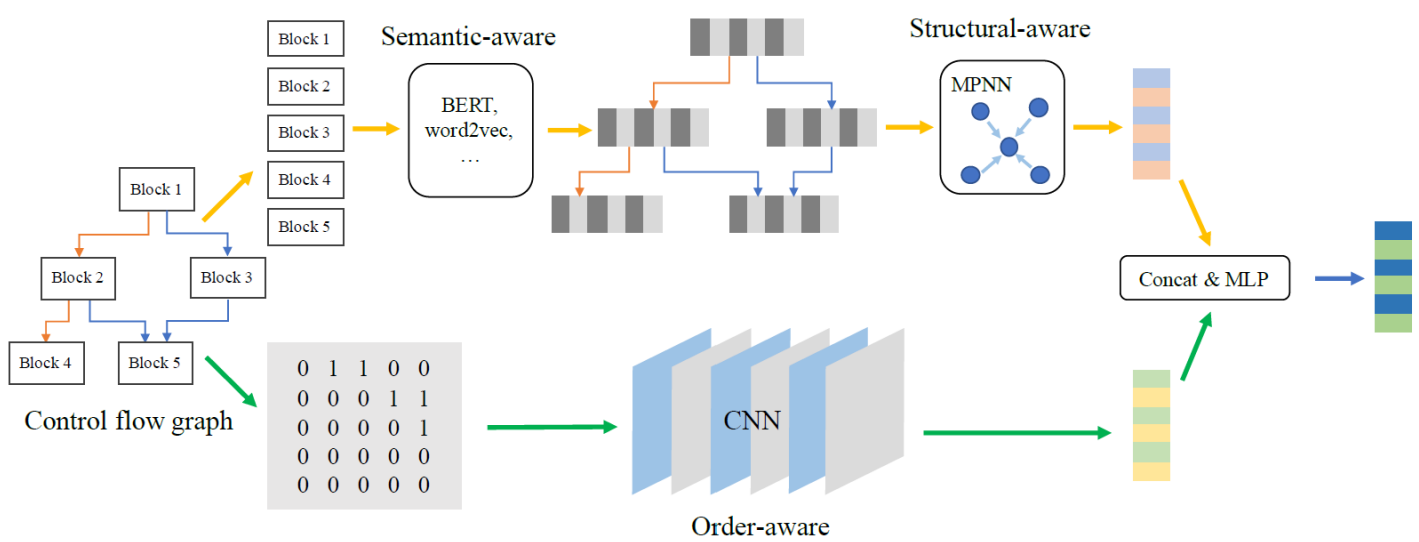


Figure 3: Overall structure of our model. The model has three components: semantic-aware modeling, structural-aware modeling and order-aware modeling.

Input: ACFG

Output: embedding. 用于之后的各种任务

Semantic-aware

把token类比为word，block类比为sentence，使用BERT预训练。

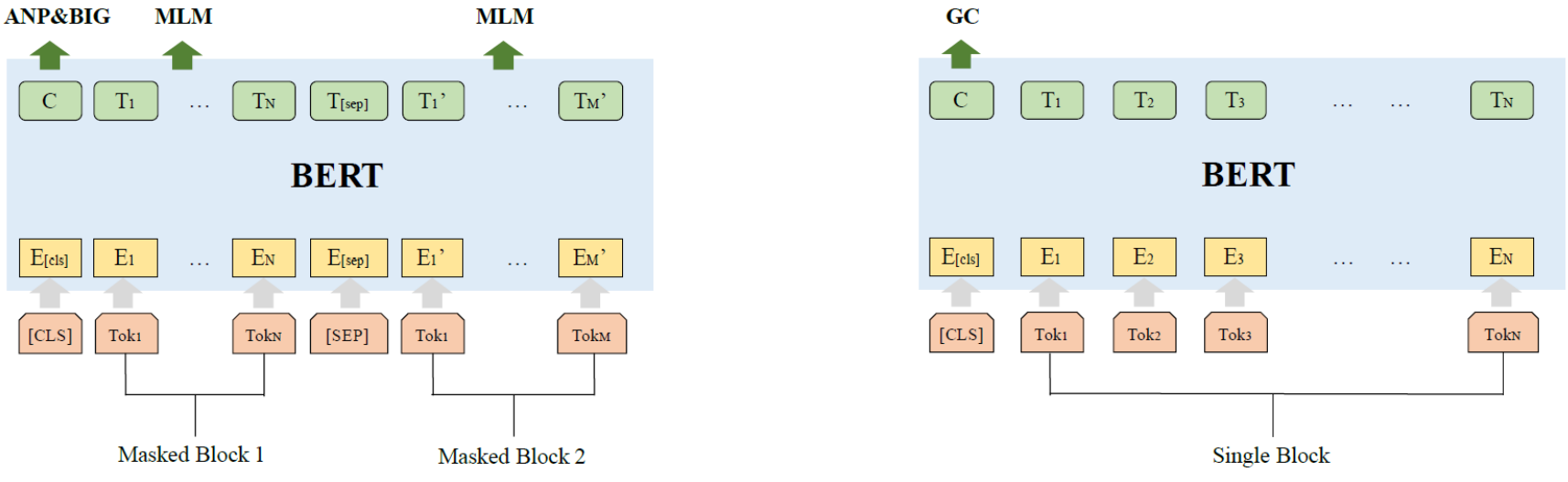


Figure 4: Bert with 4 tasks: MLM, ANP, BIG and GC.

4个任务：

- (token-level) MLM：Masked language model，对token进行mask。
- (block-level) ANP：Adjacency node prediction，相邻节点作为相邻的sentence，随机采样不相邻的block pairs作为负例。类似于Bert的NSP。
- (graph-level) BIC：block inside graph，判断两个block是否在一个图中。
- (graph-level) GC：graph classification，判断block属于哪个平台、编译器、优化选项

Structural-aware

MPNN

$$m_v^{t+1} = \sum_{w \in N(v)} M_t\left(h_v^t, h_w^t, e_{vw}\right)$$

$$h_v^{t+1} = U_t\left(h_v^t, m_v^{t+1}\right)$$

$$g_{ss} = R\left(h_v^T \mid v \in G\right)$$

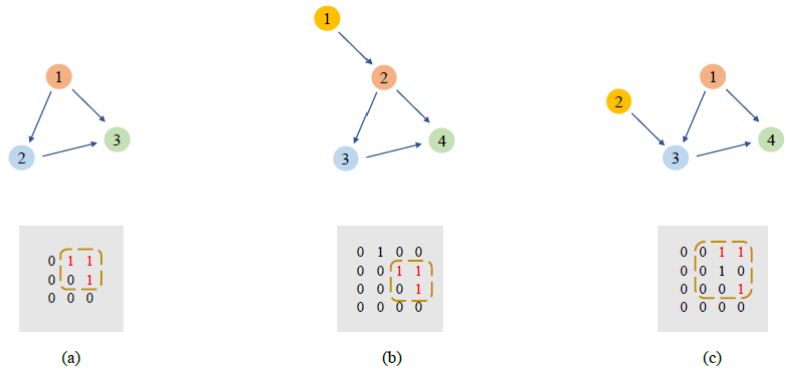
使用GRU更新。

$$m_v^{t+1} = \sum_{w \in N(v)} \mathrm{MLP}\left(h_w^t\right)$$

$$h_v^{t+1} = \mathrm{GRU}\left(h_v^t, m_v^{t+1}\right)$$

$$g_{ss} = \sum_{v \in G} \mathrm{MLP}\left(h_v^0, h_v^T\right)$$

Order-aware



将CFG的邻接矩阵作为输入，使用CNN计算出embedding。

$$g_o = \mathrm{Maxpooling}(\mathrm{Resnet}(A))$$

Use an 11-layer Resnet with 3 residual blocks.

同一函数在不同架构下编译后，CFG邻接矩阵是相似的，使用CNN可以极快地很好捕获这种相似性。（当然上图的相似性也可以捕获）

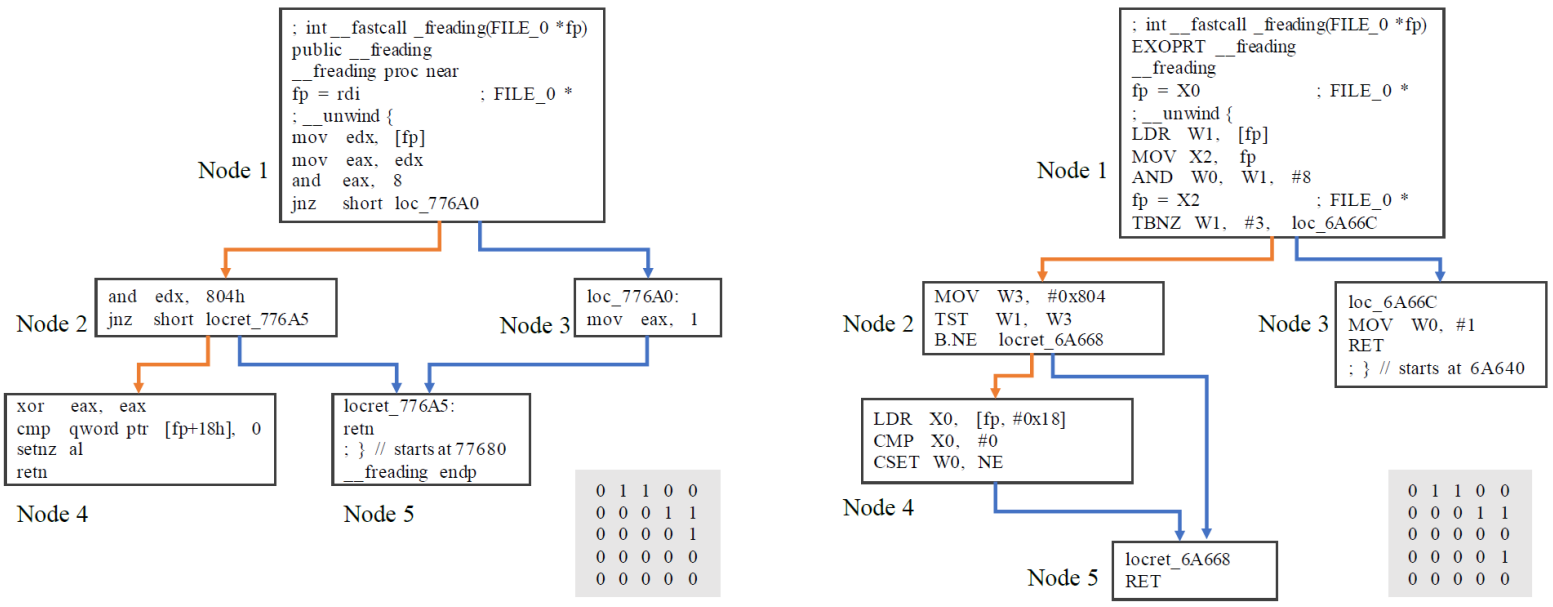


Figure 2: Two CFGs and their adjacency matrices of function ”_freading” on different platforms (x86-64 & ARM).

Experiments

Datasets: gcc-[O2|O3|x86-64|ARM]

Results: 模型表现良好，三个模块都是有益的。

Link

- 官方 - AAAI-20论文解读：基于图神经网络的二进制代码分析
- Chencwx - notes-Semantic-Aware Neural Networks for Binary Code Similarity