Minh-Le Nguyen
School of Information Science,
Japan Advance Institute of Science and Technology.
1-1 Asahidai, Nomi-city, Ishikawa, 923-1292, JAPAN.
☎ *+81 (0761) 51 1221*
FAX *+81 (0761) 51 1149*
✉ *nguyenml@jaist.ac.jp*

Dr. A. Abraham
Machine Intelligence Research Labs (MIR Labs), Auburn,
Washington, 98071, USA

December 7, 2017

Dear Dr. A. Abraham,

I am writing to submit our manuscript entitled: DGCNN: A Convolutional Neural Network over Large-scale Labeled Graphs, which is an improved and extended version of the paper: Convolutional Neural Networks over Control Flow Graphs for Software Defect Prediction [1], presented at *International Conference on Tools with Artificial Intelligence* (ICTAI) 2017, for the consideration of publication in Engineering Applications of Artificial Intelligence.

Traditional program analysis methods apply software metrics and common learning algorithms, or abstract syntax trees (ASTs) and deep learning. However, both software metrics and ASTs are not highly relevant to program behavior because software metrics rely on human observations and ASTs simply represent the syntactic structures. This paper proposes a convolutional neural network on directed labeled graphs of large-scale and formulates a new approach for program analysis tasks. The approach includes two steps: constructing graphs of execution flows and applying deep learning. For the first step, each source file written in a programming language (C, C++, ...) is compiled into an assembly code and the graph representation is constructed from the sequence of assembly instructions. After that, a deep neural network is developed to automatically learn programs' features on the graph data.

Experimental results on two tasks of software defect prediction and malware analysis show two important points: (1) our approach significantly outperforms state-of-the-art baselines that rely on software metrics and abstract syntax trees, (2) analyzing execution flows is beneficial to the tasks that discover program behavior like software defect prediction and malware analysis. We believe our findings are likely to be of great interests to machine learning, software engineering and data mining scientists who read your journal.

Comparing to the original paper, this manuscript makes eight new and significant improvements as follows.

- We clearly analyze the impossibility to adapt graph kernels as well as other graph-based networks to labeled graphs of large-scale.
- We clearly describe the convolutional neural network on labeled directed graphs which is not sufficiently mentioned in the original paper.
- We formulate a new approach that applies deep learning on control flow graphs for program analysis.
- We also validate the ability to process huge graphs of the proposed network on a malware analysis task.
- We collect and preprocess a dataset for the malware analysis task.
- We apply a tool namely BE-PUM to generate control flow graphs for executable files.
- We deeply observe and analyze the performance of the proposed approach and others according to various criteria such as evaluation measures (Accuracy, F1, AUC), and the convergence of the learning process.
- Our implementation is released to motivate further research.

All authors approved the manuscript and this submission.
Thank you very much for receiving our manuscript and considering it for review. We appreciate your time and look forward to your response.

Sincerely,

Minh-Le Nguyen

[1] All the necessary documents can be accessed at https://github.com/nguyenlab/DGCNN