Geometric Analysis and Metric Learning of Instruction Embeddings

Sajib Biswas, Timothy Barao, John Lazzari, Jeret McCoy, Xiuwen Liu, Alexander Kostandarithes
Florida State University College of Arts and Sciences
Department of Computer Science

INTRODUCTION

- In the field of Cybersecurity, a major challenge is automating binary program analysis, the process of understanding the behavior of a program, given only the binary application.
- Recently, deep learning has been proven to be useful in several program analysis tasks such as function boundary identification, malware detection, binary code similarity search etc.
- Due to the inherent similarity between natural language documents and computer programs, deep learning-based approaches which are applicable for natural languages can be used for program analysis to resolve the existing challenges.[4,5,6]
- **In this study, we conduct experiments on the state-of-the art PalmTree model and show how the application of metric learning can improve upon its current state.**

PROBLEM DEFINITION

- Just like natural languages, programing languages have well-defined syntactic and semantic rules. In that regard, the transformer-based BERT is one the most successful models in the NLP domain and security researchers have made attempts to utilize this model to facilitate program analysis.
- PalmTree is one of the most recent attempts at leveraging BERT for learning instruction embeddings for the sake of analysis of binaries. It provides a pre-trained assembly language model and addresses the unique challenges that come with the adoption of learning-based encoding approaches to model instructions.
- It is still not completely understood why BERT-produced representations work so well. When applied the same techniques for programming languages, it is important to analyze and evaluate the characteristics of the learned representations.

REFERENCES

1. A. K. Gahalaut and P. Khandnor, "Reverse engineering: an essence for software re-engineering and program analysis," International Journal of Engineering Science and Technology, vol. 2, no. 06, pp. 2296–2303, 2010.
2. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in NAACLHLT, 2019.
3. X. Li, Y. Qu, and H. Yin, "Palmtree: Learning an assembly language model for instruction embedding," in Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security, ser. CCS '21. New York, NY, USA: Association for Computing Machinery, 2021, p. 3236–3251.
4. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in NIPS, 2017.
5. D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Deep metric learning for person re-identification," in 2014 22nd international conference on pattern recognition. IEEE, 2014, pp. 34–39.
6. L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." Journal of machine learning research, vol. 9, no. 11, 2008.

OUR CONTRIBUTION

- Performing Geometric Analysis of instruction embeddings at the token level and instruction family level, showing much greater variability and leading to degraded performance on intrinsic analyses.
- Proposing to use Metric Learning to improve the relationships among instructions using triplet loss, which shows significant improvements on large datasets.
- Providing a theoretical analysis of the instruction embeddings by looking at the BERT components and characteristics of inner-product matrices for attention in the transformer blocks.

METHODOLOGY

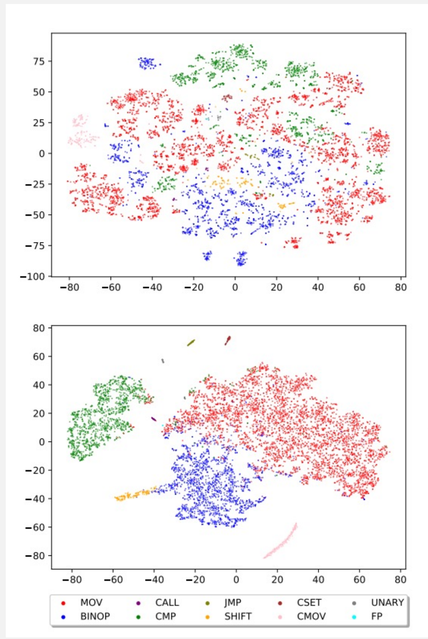
- The goal of metric learning: Train a neural network $f : R^n \rightarrow R^m$ to map inputs $x \in R\ n$ to a m dimensional metric space. Distances between embeddings reflect the similarity between each embedding.
- The distance function defined on the metric space, such as Euclidean distance, is then used to evaluate the loss and train the parameters for this purpose. For this case, the instruction embeddings are passed into the neural network and further separated by triplet loss.
- By minimizing triplet loss, distances between the positive samples and anchors are minimized.
- We create and train a deep metric learning model to improve the instruction embeddings of the PalmTree model and test it by conducting an intrinsic evaluation based on the concept of outlier detection.

ACKNOWLEDGEMENTS

This research was partially supported by NSF grant DGE 1565215.

RESULTS & EVALUATION

Figure 1. The t-SNE representation of data before (top) and after (bottom) metric learning



- The distribution is not initially very well-balanced. For example, while the group for opcode 'MOV' consists of 4, 206 instructions, another opcode 'CALL' contributes to a small group of only 25 instructions.
- We can easily separate one group from another after applying metric learning on the instructions' embeddings.
- When training on only 500 groups for 25 epochs, we were able to achieve an accuracy of 99% on 49, 500 testing groups. By comparison, the original embeddings generated by PalmTree achieve an accuracy percentage of 68%.
- Testing the speed of convergence and accuracy with 500+ trained groups, the model achieves 100% accuracy as the training set grows. This drastically improves the relationship between similar instruction embeddings and separates outliers.

CONCLUSIONS

- The embeddings of the tokens and instructions do not reflect the similarities of tokens and instructions robustly, resulting in low performance on an outlier detection problem
- Metric learning using triplet loss is effective in mitigating the issues, resulting in better clusters and significant improvements for the outlier detection problem
- Due to the many parallels between computer programs and natural languages, taking advantage of developments in natural language processing stands to benefit areas including software reverse engineering, malware analysis, and program analysis.