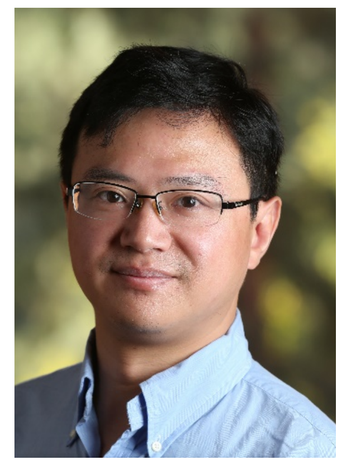
Towards Robust and Scalable Search of Binary Code and Data

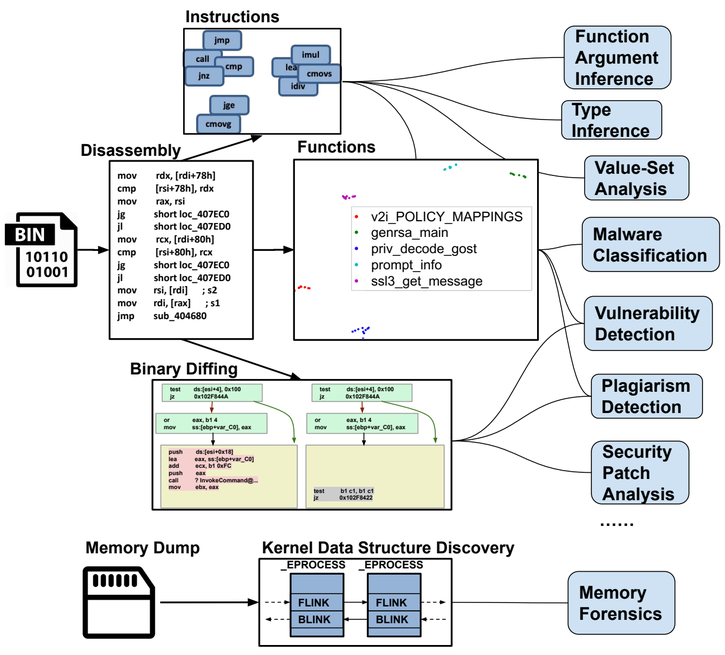


Institution: University of California, Riverside

Principal Investigator: Heng Yin (heng@cs.ucr.edu)

Abstract

The problem of binary code and data search concerns how to glean valuable information from binary code and binary data in an accurate, scalable and robust fashion. This concern is central to many security problems, including vulnerability detection, plagiarism detection, malware classification, memory forensics, etc. Our work takes a novel approach that mimics how the human brain recognizes interesting objects from an enormous amount of visual information and builds a pipeline to improve binary code and data search from different aspects, including disassembly, pretrained instruction embedding, cross-platform function embedding, basic-block level binary diffing and kernel data structure discovery.



Challenge

- Identify semantically equivalent or similar code in different architectures and compilation settings. The syntactic variations among similar code need to be tolerated so that the approach can be applied in cross-platform scenarios.
- Identify objects from binary data such as memory dumps and documents. The approach needs to be robust against changes caused by different platform versions and malicious manipulations such as DKOM (Direct Kernel Object Manipulation) attacks.
- Scale to a large dataset. The approach needs to be scalable in order to be applied in real-world scenarios such as malware detection. Current disassembly approaches are hard to be accurate and efficient at the same time. Pair-wise comparison of binary code is not scalable as well.

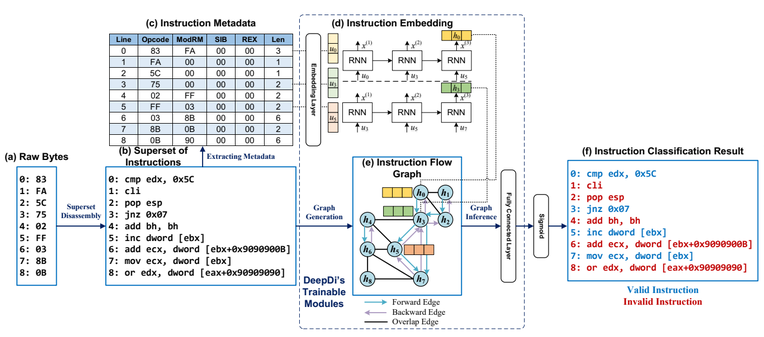
Scientific Impact

- Benefited security applications built on top of the project
  - Disassembly can be leveraged in downstream cybersecurity applications including machine learning based malware detection and classification.
  - Instruction embeddings can be leveraged in downstream function argument inference, etc.
  - Function embedding can be leveraged in code plagiarism detection, etc. Binary diffing can be leveraged in more fine-grained applications such as security patch analysis.
- Stimulated more research in the direction of deep learning-based binary analysis
  - Gemini got 388 citations by 05/08/2022.
  - Stimulated more binary embedding models (e.g. Asm2vec [S&P'2019], InnerEye [NDSS'2019], FunctionSimSearch by Google Project Zero Team)

Solution

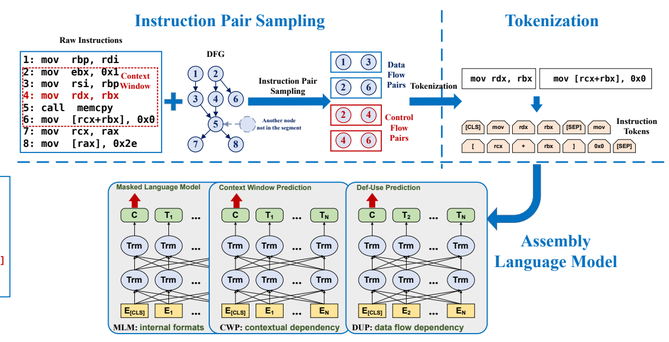
We take a novel approach that leverages the power of ML/DL and build a pipeline of work that includes:

- **DeepDi** [Usenix'22]: A deep learning-based fast disassembler
- **PalmTree** [CCS'21]: Pretrained instruction embedding based on BERT
- **Gemini** [CCS'17]: Binary similarity detection based on function embedding
- **DeepBinDiff** [NDSS'20]: Binary diffing based on basic-block representations
- **DeepMem** [CCS'18]: A fast and robust GNN-based memory forensics approach



Overview of DeepDi with a Concrete Example

- GPU-Accelerated.
- Obfuscation Resilient. Low false positive/negative rate on identifying obfuscated code.
- Accurate. 0.02%/0.02% false positive/negative rate on instruction identification and 99.9% precision on function identification.



System design of PalmTree. Trm is the transformer encoder unit, C is the hidden state of the first token of the sequence (classification token), Tn $(n = 1 \ldots N)$ are hidden states of other tokens of the sequence.

- Pre-training tasks enabled.
- Accurate. Outperforms the other instruction embedding models and also significantly improves the accuracy of downstream binary analysis tasks.

Broader Impact (impact on society)

- The project can be applied to security task such as vulnerability detection, which will guard the security for PC and mobile users.
- The project have large practice use due to its scalability.
- Reverse engineer, security analyst and security audit company can benefit from the project.

Broader Impact (education and outreach)

- The techniques of binary code and data search can be leveraged in industry and national security.
- The proposed deep-learning based binary code similarity detection approach has received an NSF SBIR award for commercialization.

Broader Impact and Broder Participation

- Reduced the workload and improved efficiency of reverse engineering (e.g. A CUDA implementation of DeepDi is *350 times* faster than IDA Pro).
- Detected and mitigated known vulnerabilities in real-world binaries (e.g. Gemini identified 42 vulnerabilities among top 50 in large scale firmware dataset).

Award ID#:1719175