

# FASER: Binary Code Similarity Search through the use of Intermediate Representations

Josh Collyer<sup>1</sup>, Tim Watson<sup>1,2</sup>, and Iain Phillips<sup>1</sup>

<sup>1</sup>Computer Science, Loughborough University

<sup>2</sup>The Alan Turing Institute

October 6, 2023

## Abstract

Being able to identify functions of interest in cross-architecture software is useful whether you are analysing for malware, securing the software supply chain or conducting vulnerability research. Cross-Architecture Binary Code Similarity Search has been explored in numerous studies and has used a wide range of different data sources to achieve its goals. The data sources typically used draw on common structures derived from binaries such as function control flow graphs or binary level call graphs, the output of the disassembly process or the outputs of a dynamic analysis approach. One data source which has received less attention is binary intermediate representations. Binary Intermediate representations possess two interesting properties: they are cross architecture by their very nature and encode the semantics of a function explicitly to support downstream usage. Within this paper we propose Function as a String Encoded Representation (FASER) which combines long document transformers with the use of intermediate representations to create a model capable of cross architecture function search without the need for manual feature engineering, pre-training or a dynamic analysis step. We compare our approach against a series of baseline approaches for two tasks; A general function search task and a targeted vulnerability search task. Our approach demonstrates strong performance across both tasks, performing better than all baseline approaches.

## 1 Introduction

Binary Code Similarity Search aims to provide a means of finding compiled functions which are similar to a given query function. Being able to achieve this is useful when wanting to identify similarities between malware functionality, identify function re-use or to understand whether a piece of software contains known vulnerabilities. This is a complex undertaking. Factors ranging from the diversity of toolchains to compiler optimization options mean that functions can be represented differently across binaries. However, recent research suggests that instruction set architecture (ISA) has the largest impact [9]. The diversity of ISAs is vast when viewing the problem through an embedded computing lens where software can be used within systems ranging from a MIPS-based embedded 5G modem to a 1750A-based subsystem in a US Apache Helicopter.

This problem is not new, however, and Binary Code Similarity Search has been tackled using a range of different methods. In particular, natural language processing (NLP) approaches have been transitioned from other domains and applied to binary analysis tasks. Early approaches such as SAFE [15], asm2vec [5] and InnerEye [25] explored using NLP for binary code search utilising state-of-the-art approaches. The literature then developed and moved onto explore using the advances in Transformer architectures in approaches such as jTrans [21], PalmTree [12] and Trex [16], all of which use a similar pre-training methodology as BERT [4] with the addition of domain specific tasks or binary analysis specific data sources.

The aforementioned approaches all suffer from challenges with constructing a vocabulary which is able to cover the range of possible inputs. This is referred to as the Out of Vocabulary (OOV) problem [12] and stems from the use of assembly instructions as input which can include a broad range of possible values such as memory addresses and opcodes. Even after normalization the number of possible inputs continues to increase with the number of supported architectures due to implementation-specific nuances. In order to overcome this challenge, some approaches have instead sought to use an intermediate representation (IR) as the input data format. For example, XLIR [6] uses the LLVM IR to conduct binary to source function search and Penwy [17] uses the VEX IR alongside a form of concolic execution to create *bug signatures* for known bugs to conduct vulnerability search. Neither of these approaches however tackles binary function search directly using only the IR without any additional inputs.

Within this paper, we propose Function As a String Encoded Representation (FASER) (*Code link to be added*) which combines the long document transformer architecture, Longformer [2], with the use of radare2's [20] Evaluable String Intermediate Language (ESIL) to create a cross architecture model which is capable of binary function search across multiple different architectures. Through using an IR as the input data type, we side step the issue of having to normalize for each assembly language and instead normalize once across a single common representation.

The key contributions of this paper are:

1. A binary function representation as IR *Functions as Strings* which requires no additional data processing effort other than normalization.
2. A cross-architecture model which combines the usage of IRs alongside longer context transformers and demonstrate its usefulness for cross-architecture function search and known vulnerability detection.
3. We demonstrate that it's possible to get strong cross-architecture binary search performance using a transformer architecture without the need for pre-training and instead using deep metric learning to train directly for the binary function search objective.
4. We conduct, as far as the authors are aware, the first experiment for cross-architecture function search using RISC-V architecture as part of the experimental methodology.

This paper is structured in the following way. Section 2 describes the research methodology used for this research before then moving onto Section 3 which presents the experimental results derived from a series of experiments conducted. This paper then concludes with concluding remarks in Section 4 where we discuss our findings, detail the implications and propose potential future research avenues.

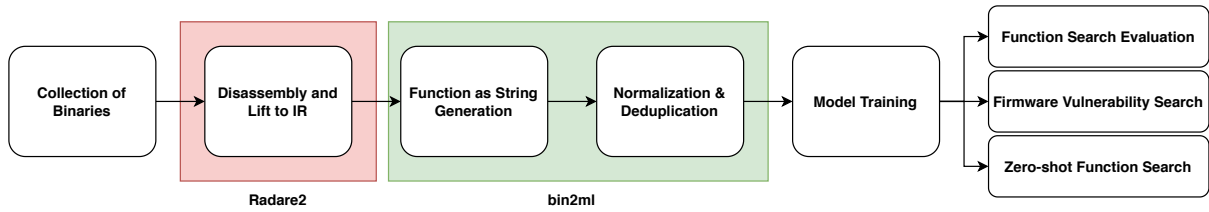


Figure 1: High Level Overview of Methodology

## 2 Methodology

In this section we provide an overview of the methodology used to create our experimental dataset and details related to how we train and evaluate our proposed solution. We begin by describing our chosen IR before describing the dataset used. We then move onto describe the process of going from raw binaries to pre-processed, training-ready data. This section then continues to describe the model design, training configuration and evaluation design before detailing the metrics and baseline approaches used to compare against.

### 2.1 Chosen Intermediate Representation

The chosen IR is radare2’s Evaluable Strings Intermediate Language (ESIL). radare2 converts assembly language into a semantic equivalent, ESIL which represents the architecture specific instructions using a combination of symbols and numbers. Figure 2 provides several examples of x86-64 assembly instructions and the corresponding ESIL representations. The primary reason for choosing ESIL over other IR’s such as VEX, LLVM or PCode was compactness. Any given assembly instruction corresponds to a single ESIL string. Whilst some assembly instructions create very large ESIL string representations, through our experimentation the length of the ESIL IR is typically shorter and more succinct as opposed to the alternatives.

```

disasm: push rbp esil: rbp,8,rsp,0,=[8],8,rsp,-=
disasm: call sym.imp.printf esil: 4176,rip,8,rsp,-=,rsp,=[8],rip,=
disasm: mov dword [rbp - 8], 0 esil: 0,0x8,rbp,-,=[4]
  
```

Figure 2: Example X86-64 and ESIL Representations

### 2.2 Dataset

In order to evaluate FASER and compare against comparative baselines, we use two of the datasets detailed within [14]. The first dataset is *Dataset-1* created by [14]. *Dataset-1* contains seven popular open source projects: `ClamAv`, `Curl`, `nmap`, `Openssl`, `Unrar`, `z3` and `zlib`. Each of these are compiled for ARM32, ARM64, MIPS32, MIPS64, x86 and x86-64 using four different versions of Clang and GCC alongside 5 different optimization levels. This results in each of the 7 projects having 24 unique compiler, architecture and optimization combinations for each binary within the library. Within [14], the authors

formulate 6 tasks using this dataset which increase in difficulty. For the purposes of this paper, we have chosen the most difficult, denoted as *XM*. The *XM* task imposes no constraints on which functions can be sampled from the corpus during test time and includes all possible compiler, architecture and bitness combinations. This task is representative of conducting binary function search against real binaries.

The second dataset we use is the *Dataset-Vulnerability* dataset also part of [14]. This dataset consists of two firmware images that include several OpenSSL CVE's, specifically within the *libcrypto* library included as part of the firmware. The first firmware image is of a Netgear R7000 router which is ARM32 and a TP-Link Deco-M4 mesh router which is MIPS32. The dataset also includes the same vulnerable library compiled for ARM32, MIPS32, x86 and x86-64. The goal is to use the vulnerable functions from our compiled libraries as a query function and then identify the corresponding vulnerable function within the firmware image. In addition to the two tasks above, we augment the second dataset with *libcrypto* compiled for RISC-V 32-bit and then re-run the firmware search. This task has been introduced to explore whether an IR model is capable of transferring its learning to architectures it has not seen before and can be considered a research first.

## 2.3 Data Generation

In order to generate the training data, *bin2ml* [3] was used for both data extraction and pre-processing. *bin2ml* uses radare2 to disassemble the binaries and lift functions into ESIL IR. Once this lifting process has been complete, the data is then processed further to create ESIL function as string representations by concatenating all ESIL instructions for a given function into a single, long string. Any strings that were longer than our model's input dimension were truncated. This could potentially cause a loss of key information but is mitigated by the large input dimension chosen. Each function string is then normalized using a series of heuristics before the entire corpus is deduplicated. This process was repeated for all binaries within our datasets. The specifics of the normalization and deduplication process are presented below.

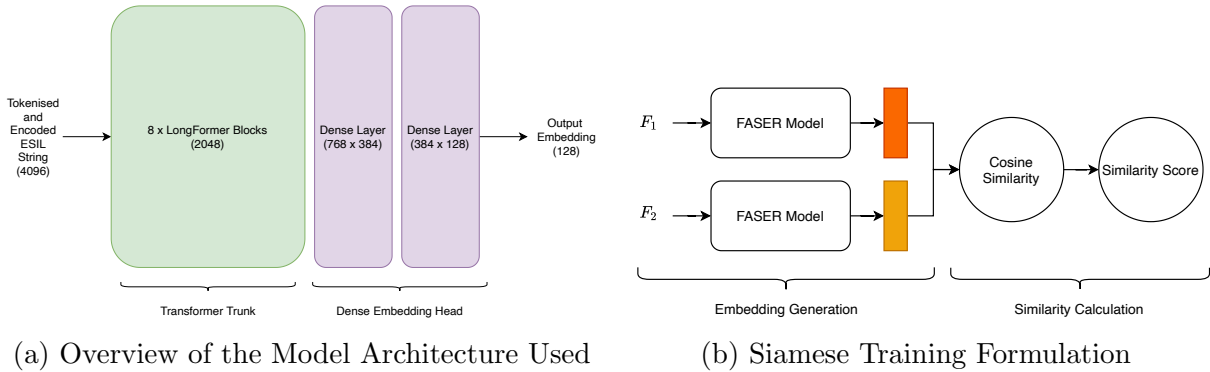
## 2.4 Normalization

Normalization is a fundamental step to ensure that the vocabulary size is manageable, and all possible inputs can be encoded. In order to facilitate this, a series of heuristics were applied to replace parts of the ESIL strings. The normalization approach draws on common approaches outlined within the literature such as those used within SAFE [15], jTrans [21] and PalmTree [12]. Firstly, any hexadecimal value which starts with `0xffff` or is one to three characters long (such as `0x023` or `0x02`) are considered immediate constants and replaced with the *IMM* token. Secondly, any hexadecimal value which starts with `0x` preceded by 4 or more hexadecimal values is considered a memory address and replaced with *MEM*. Thirdly, due to the way radare2 represents function calls and data accesses, these values are typically represented as integers within ESIL representations. For this reason, if the opcode the ESIL representation was derived from is a call opcode, the integer is replaced with the *FUNC* token and otherwise *DATA*. And lastly, general purpose registers are replaced with tokens based on their size, 32-bit registers are replaced with *reg32* and 64-bit registers are replaced with *reg64*. As part of the experimentation, two versions of FASER were trained, one without register normalization and one with. This allows us to understand the impact register normalization would have on an IR

based model.

## 2.5 Deduplication

After normalization, deduplication takes place. Deduplication is critical because even after changing factors such as optimization level and compiler, it is still possible for binaries generated from the same source code to produce identical functions. This is typically overlooked in existing literature and is comparable to the approach presented in [14]. For each of the normalised ESIL strings, the ESIL string plus the function name are concatenated into a single string before being hashed. These hashes are then compared with each other to identify where there are duplicate functions. For any matches found, only one was kept ensuring that the dataset used for training contains only unique function strings. There is then a subsequent step taken which looks through the entire dataset and eliminates functions which are only present once. Essentially removing the functions whereby regardless of architecture or optimization, after disassembly and lifting IR are invariant. Across our dataset, the deduplication process eliminated on average 20-25% of the functions from a given library.



## 2.6 Model Design

The chosen model used in FASER is a LongFormer. The LongFormer model was proposed by [2] to tackle the quadratic computational scaling of self-attention used in models such as BERT. The LongFormer instead uses a combination of local, sliding window attention, with a global attention mechanism. This formulation instead scales linearly as the input size increases, providing a mechanism to train transformers with larger input sequences. Furthermore, this combination of local and global attention is viewed favorably for the binary function search task. An assembly instruction is not executed in isolation but instead executed as part of a series of instructions. This local attention window provides a means for a single instruction to include the context of the instructions before and after it but in a manner which is bounded. The global attention can then look at the function holistically whilst being informed by the local contexts provided by the sliding window attention.

The model parameters used for FASER were an input dimension of 4096, followed by 8 LongFormer blocks with an intermediate dimension of 2048, followed by two dense layers which map the 768 dimension transformer output to a 128 dimension embedding. The local attention window is set to 512 tokens. We utilize the implementation provided by `transformers` and all other parameters are kept default. These can be view here [8]

## 2.7 Training Configuration

Previous works such as Trex [16], jTrans [21] and PalmTree [12] conduct a pre-training step prior to then fine-tuning for the function search objective. Whilst this makes sense if you want to train a model which can be used for various different downstream tasks, it's potentially suboptimal if the only downstream use case is going to be function search. To this end, we forgo any pre-training steps and train FASER directly for the function search objective using deep metric learning. We construct a pair-based training methodology by using a Siamese formulation in combination with Circle Loss [19]. Circle Loss was chosen due to its ability to place emphasis on large deviations in between-class similarity in a manner not possible with other losses such as triplet loss. Both Cosine Embedding Loss and Triplet Loss were experimented with and resulted in unstable training and in some cases, complete model collapse.

We also formulate a sampling strategy that ensures  $m$  number of examples for a given label (function) are present within a batch. We then apply the online batch hard pair mining method [7] to dynamically create both positive and negative pairs for each example from batched inputs throughout training. This works by embedding all examples within a batch before using the associated labels for each example to create the hardest positive and negative pairs for each example. What determines hardest is the output of a distance function which in our case was Cosine Similarity. The strength of this approach when compared to previous research which uses static pre-computed pairs is the weaknesses of the model are consistently challenged. For example, if during training the model quickly learns to search across ARM and MIPS but is performing badly when searching across X86-64, this training formulation would automatically begin to target this weakness by generating pairs including X86-64 examples and uses them to calculate the loss.

For training, we use the whole of Dataset-1 and sample 100K functions per epoch for 18 epochs (Approximately 3 days of training). We set  $m$  to 2 to ensure each batch has 2 of each sampled function, batch size is set to 8 and, we use gradient accumulation to artificially set the batch size to 512. The Adam [10] optimizer is used with a fixed learning rate of  $0.0005$ .

## 2.8 Comparison Approaches

For the first task, we draw upon the top performing approaches reported within Marcelli (2022) [14] which are the Graph Matching Networks (GMN) and Graph Neural Network (GNN) approach from Li et al (2019) [13] and Gemini [23]. All of these are Graph Neural Network (GNN) approaches which take advantage of the structural aspects of functions, typically through using the control flow graph (CFG) with node level feature vectors as an input. Approaches using natural language processing and transformer model architectures such as PalmTree [12] and jTrans [21] would have been ideal candidates but are mono-architecture therefore were deemed unsuitable for comparison.

For the vulnerability search task, we use the same three approaches outlined above but also compare against Trex [16]. Trex provides an interesting comparison because it too uses a transformer architecture but has one significant difference. The model is pre-trained on what the authors describe as micro-traces. These micro-traces are generated in a dynamic manner using an emulator. Once trained, the model is capable of being used with solely static data and forgoes the emulation aspect. The emulator used to generate the micro-traces provided by the authors does not support the full breadth

of architectures and bitness of Dataset-1 therefore would be an unfair comparison and therefore, is not used in the first task.

## 2.9 Evaluation Configuration

For task 1, we again use Dataset-1 and implement a sampling approach which dynamically creates *search pools* which for a given function, contain 1 positive example and 100 negative examples. This formulation is the same as [14]. We also adopt the same methodology as [14] for the Vulnerability search task whereby we have a query function of a given architecture and search across all possible functions in the firmware's *libcryto* library. This means that the search pool size for task 2 is 10 times bigger at approximately 1000 functions.

## 2.10 Metrics

We re-use the metrics used in previous studies **Recall@1** and **MRR@10** for task 1 in order to present a reliable comparison. For the vulnerability search task, we report the rank at which the vulnerable function was present at after the search was conducted similar to other studies, alongside this, we also calculate the mean and median ranks across all architecture searches. This is primarily to aid result analysis.

# 3 Evaluation

Our evaluation aims to answer the following research questions:

1. **RQ1** - How does FASER perform when compared against other baseline approaches for the binary function search task?
2. **RQ2** - How effective is FASER at searching real firmware images for known vulnerabilities?
3. **RQ3** - Does using intermediate representations as the input data enable the model to zero shot architectures not previously seen as part of the training data?

## 3.1 RQ1 - Binary Function Similarity Search

The results of the experimentation to gather data for *RQ1* can be seen in Table 1. Both of the FASER models trained outperform all the baseline approaches across both of the chosen metrics. Looking first at Recall@1, the model without register normalized training data (denoted as FASER NRM) performs significantly better than the register normalized model, achieving a Recall@1 increase of 13% when compared against the best performing baseline, GMN. FASER RN performs comparable to the GMN model without needing direct comparison between all possible function combinations within a given search pool.

Moving onto the MRR@10 results, FASER NRM again performance significantly better than all baselines with a 7.5% increase. FASER RM is again comparable to the GMN approach with an identical MRR@10.

**RQ1 Summary:** The results present above show that our propose approach, FASER, performs as well as if not better than the best baseline approach with FASER NRM performing better across all metrics.

		XM	
Method	Description	R@1	MRR@10
FASER NRM	ESIL Function String	<b>0.51</b>	<b>0.57</b>
FASER RN	ESIL Function String	0.46	0.53
GMN [13]	CFG + BoW opc 200	0.45	0.53
GNN [13]	CFG + BoW opc 200	0.44	0.52
GNN (s2v) [23]	CFG + BoW opc 200	0.26	0.36

Table 1: RQ1 - Binary Function Similarity Search. NRM = No Register normalization and RN = Register normalization

### 3.2 RQ2 - Binary Function Vulnerability Search

The results of the experimentation to gather data for *RQ2* can be seen in Table 2. The results show the ranks of search results when searching the Netgear R7000 router which is ARM32.

NETGEAR R700						
	ARM	MIPS	X86	X86-64	Mean Rank	Median Rank
<b>GNN</b>	4:1:1:44	97:5:1:138	3:31:1:18	9:6:1:40	25	5.5
<b>GNN (s2v)</b>	2:1:1:6	35:5:1:7	8:1:5:36	9:1:14:8	8.125	5.5
<b>Trex</b>	32:4:1:2	24:16:1:1	41:4:1:3	10:3:1:2	9.125	3
<b>GMN</b>	1:1:1:1	1:1:1:7	1:1:1:2	1:1:30:7	3.625	1
<b>FASER NRM</b>	1:5:1:1	9:122:1:3	21:21:3:50	7:122:1:2	23.125	4
<b>FASER RN</b>	1:6:1:1	2:4:1:4	2:1:1:3	1:2:1:1	2	1

Table 2: RQ2 - Vulnerability Search Result Rankings

In addition to the three baseline approaches used in the previous task, our proposed approach was compared against Trex, a comparable transformer based approach which has a more complicated and computationally expensive training process. The results presented in Table 2 show that our proposed approach performs well across all the architectures. Interestingly, whilst the register normalized model performed the strongest in the Binary Function Similarity Search task, in this task the register normalized model performs significantly better. This is shown by both mean and median rank descriptive statistics being lower. The best performing FASER model is highly comparable to the GMN method but again without the aforementioned limitations. Comparing specifically to Trex, the register normalized model performs consistently better across all the architectures. This suggests that our training methodology of training for the function similarity directly and forgoing the elaborate pre-training steps usually adopted and the use of IR’s as our data input has merit.

**RQ2 Summary:** The results present above show that the FASER RN performs well when searching real firmware images for known vulnerabilities.

### 3.3 RQ3 - Zero Shot Architecture Binary Function Search

Table 3 shows the results from the experimentation undertaken to answer *RQ3*. The question posed here is can the FASER models, because we are using an IR as the input data, perform zero-shot vulnerability search for a new architecture by transferring prior learnt knowledge. Fundamentally, the answer to this is no. The vulnerability search performance for a new instruction set architecture (in this case RISC-V) is significantly worse. This is clearly demonstrated by the mean and median rank descriptive statistics.

Nevertheless, these results do demonstrate something interesting. Across both FASER models, the performance is significantly better when searching MIPS functions using a RISC-V query as opposed to searching ARM functions using a RISC-V query. This suggests that the semantic representation created when MIPS and RISC-V instructions are lifted to ESIL may be more similar than ARM and RISC-V. Given that recent research has suggested that ARM and X86/X86-64 instructions are closer in statistical similarity than when compared to MIPS [9], these results may suggest that introducing RISC-V binaries into a training dataset may level out any data imbalances.

	ARM32	Mean Rank	Median Rank	MIPS32	Mean Rank	Median Rank
FASER NRM	48;546;964;14	393	297	48;30;22;546;170;	154	101.5
		393	297	251;14;155	154	101.5
FASER RM	673;292;1004;15	496	482.5	673;33;4;292;76;	172	106
		496	482.5	136;15;147	172	106

Table 3: RQ3 - Zero Shot Architecture Search

**RQ3 Summary:** The results present above show that the use of an IR input representation does not provide a means of conducting zero-shot search across unseen architectures.

## 4 Discussion and Conclusion

The results presented above demonstrate that the combination of ESIL IR and the Long-former transformer architecture perform well compared to the baseline approaches with minimal requirement for either manual feature engineering or dynamic analysis. The FASER RN model performs particularly well at the vulnerability search task across all architectures tested and performs comparably to GMN, without requiring direct comparison between all possible combinations within a given search pool. Whilst demonstrating the effectiveness of IRs and longer context transformers, these results also add weight to our argument that the pre-training step seen within previous work may be unwarranted for binary function search and training for the binary function similarity objective directly may be more optimal.

The results presented to answer RQ3 also suggest something interesting. While the search rank results were significantly worse in terms of mean and median rank, suggesting that our proposed approach is unable to reliably transfer to unseen architectures, there is a large difference between the reported ranks for RISC-V  $\rightarrow$  ARM when compared to RISC-V  $\rightarrow$  MIPS ranks. Prior research [9] observed similar phenomena whereby x86  $\rightarrow$  ARM functions were statistically more similar than x86  $\rightarrow$  MIPS functions. This suggests

that RISC-V functions may be more similar to MIPS functions in terms of semantics when represented in ESIL than ARM. Given that most datasets used in prior research only include X86, ARM and MIPS, this similarity could potentially be leveraged and experimented with further. An example of this experimentation could be to see explore whether including RISC-V functions within a cross-architecture binary function search dataset balances out the x86  $\rightarrow$  ARM similarities with providing function examples that are similar but different to the MIPS architecture.

Turning now to implications of this research. Firstly, this work demonstrates that IRs derived from binaries can be used to train models for binary function search and perform well. Secondly, the use of longer input sequences also works well. The performance results, especially related to the use of the LongFormer architecture suggest that changing the type of transformer architecture used and increasing the input dimension for approaches such as jTrans [21], Trex [16] or PalmTree [12] may increase their overall performance. The results also demonstrate that if the only downstream target task is binary function search it may be worth amending the standard training methodology which involves a pre-training step and instead, train for the objective directly.

In future work, there are several avenues that could be explored. There are a number of different IRs that could be incorporated into similar approaches such as VEX [18], LLVM [11] and PCode [1]. There is also an emerging sub-field of binary function search focused on adding heuristic pre- and post-filtering steps to increase performance by reducing the number of functions searched such as those described in Asteria-Pro [24] and BinUSE [22]. And finally, this approach could be enhanced through the integration of supporting models such as those that use decompiled source code, recovered type information or structural aspects at a control flow graph or call graph level.

## References

- [1] National Security Agency. ghidra. <https://github.com/NationalSecurityAgency/ghidra/tree/master>, 2019.
- [2] Iz Beltagy, Matthew E Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*, 2020.
- [3] Josh Collyer. bin2ml. <https://github.com/br0kej/bin2ml/>, 2023.
- [4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina N. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. 2018.
- [5] Steven HH Ding, Benjamin CM Fung, and Philippe Charland. Asm2vec: Boosting static representation robustness for binary clone search against code obfuscation and compiler optimization. In *2019 IEEE Symposium on Security and Privacy (SP)*, pages 472–489. IEEE, 2019.
- [6] Yi Gui, Yao Wan, Hongyu Zhang, Huifang Huang, Yulei Sui, Guandong Xu, Zhiyuan Shao, and Hai Jin. Cross-language binary-source code matching with intermediate representations. In *2022 IEEE International Conference on Software Analysis, Evolution and Reengineering (SANER)*, pages 601–612, 2022.
- [7] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017.

- [8] HuggingFace. Longformer. [https://huggingface.co/docs/transformers/model\\_doc/longformer](https://huggingface.co/docs/transformers/model_doc/longformer), 2021.
- [9] Dongkwan Kim, Eunsoo Kim, Sang Kil Cha, Sooel Son, and Yongdae Kim. Revisiting binary code similarity analysis using interpretable feature engineering and lessons learned. *IEEE Transactions on Software Engineering*, 2022.
- [10] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017.
- [11] Chris Lattner and Vikram Adve. LLVM: A compilation framework for lifelong program analysis and transformation. pages 75–88, San Jose, CA, USA, Mar 2004.
- [12] Xuezixiang Li, Yu Qu, and Heng Yin. Palmtree: learning an assembly language model for instruction embedding. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, pages 3236–3251, 2021.
- [13] Yujia Li, Chenjie Gu, Thomas Dullien, Oriol Vinyals, and Pushmeet Kohli. Graph matching networks for learning the similarity of graph structured objects. In *International conference on machine learning*, pages 3835–3845. PMLR, 2019.
- [14] Andrea Marcelli, Mariano Graziano, Xabier Ugarte-Pedrero, Yanick Fratantonio, Mohamad Mansouri, and Davide Balzarotti. How machine learning is solving the binary function similarity problem. page 2099–2116, 2022.
- [15] Luca Massarelli, Giuseppe Antonio Di Luna, Fabio Petroni, Roberto Baldoni, and Leonardo Querzoni. Safe: Self-attentive function embeddings for binary similarity. In *International Conference on Detection of Intrusions and Malware, and Vulnerability Assessment*, pages 309–329. Springer, 2019.
- [16] Kexin Pei, Zhou Xuan, Junfeng Yang, Suman Jana, and Baishakhi Ray. Trex: Learning execution semantics from micro-traces for binary similarity. *arXiv preprint arXiv:2012.08680*, 2020.
- [17] Jannik Pewny, Behrad Garmany, Robert Gawlik, Christian Rossow, and Thorsten Holz. Cross-architecture bug search in binary executables. In *2015 IEEE Symposium on Security and Privacy*, pages 709–724. IEEE, 2015.
- [18] Yan Shoshitaishvili, Ruoyu Wang, Christophe Hauser, Christopher Kruegel, and Giovanni Vigna. Firmalice - automatic detection of authentication bypass vulnerabilities in binary firmware. 2015.
- [19] Yifan Sun, Changmao Cheng, Yuhan Zhang, Chi Zhang, Liang Zheng, Zhongdao Wang, and Yichen Wei. Circle loss: A unified perspective of pair similarity optimization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6398–6407, 2020.
- [20] Radare2 Team. Radare2 github repository. <https://github.com/radare/radare2>, 2017.
- [21] Hao Wang, Wenjie Qu, Gilad Katz, Wenyu Zhu, Zeyu Gao, Han Qiu, Jianwei Zhuge, and Chao Zhang. jtrans: Jump-aware transformer for binary code similarity. *arXiv preprint arXiv:2205.12713*, 2022.

- [22] Huaijin Wang, Pingchuan Ma, Yuanyuan Yuan, Zhibo Liu, Shuai Wang, Qiyi Tang, Sen Nie, and Shi Wu. Enhancing dnn-based binary code function search with low-cost equivalence checking. *IEEE Transactions on Software Engineering*, 49(1):226–250, 2022.
- [23] Xiaojun Xu, Chang Liu, Qian Feng, Heng Yin, Le Song, and Dawn Song. Neural network-based graph embedding for cross-platform binary code similarity detection. In *Proceedings of the 2017 ACM SIGSAC conference on computer and communications security*, pages 363–376, 2017.
- [24] Shouguo Yang, Chaopeng Dong, Yang Xiao, Yiran Cheng, Zhiqiang Shi, Zhi Li, and Limin Sun. Asteria-pro: Enhancing deep-learning based binary code similarity detection by incorporating domain knowledge. *ACM Transactions on Software Engineering and Methodology*, 2023.
- [25] Fei Zuo, Xiaopeng Li, Patrick Young, Lannan Luo, Qiang Zeng, and Zhexin Zhang. Neural machine translation inspired binary code similarity comparison beyond function pairs. *arXiv preprint arXiv:1808.04706*, 2018.