# Deep Program Structure Modeling Through Multi-Relational Graph-based Learning

Guixin Ye, Zhanyong Tang*, Huanting Wang, Dingyi Fang
Northwest University, China
{gxye,zytang}@nwu.edu.cn

Jianbin Fang
National University of Defense Technology, China

Songfang Huang
Alibaba DAMO Academy, China

Zheng Wang*
University of Leeds, United Kingdom
z.wang5@leeds.ac.uk

## ABSTRACT

Deep learning is emerging as a promising technique for building predictive models to support code-related tasks like performance optimization and code vulnerability detection. One of the critical aspects of building a successful predictive model is having the right representation to characterize the model input for the given task. Existing approaches in the area typically treat the program structure as a sequential sequence but fail to capitalize on the rich semantics of data and control flow information, for which graphs are a proven representation structure.

We present Poem[1], a novel framework that automatically learns useful code representations from graph-based program structures. At the core of Poem is a graph neural network (GNN) that is specially designed for capturing the syntax and semantic information from the program abstract syntax tree and the control and data flow graph. As a departure from existing GNN-based code modeling techniques, our network simultaneously learns over multiple relations of a program graph. This capability enables the learning framework to distinguish and reason about the diverse code relationships, be it a data or a control flow or any other relationships that may be important for the downstream processing task.

We apply Poem to four representative tasks that require a strong ability to reason about the program structure: heterogeneous device mapping, parallel thread coarsening, loop vectorization and code vulnerability detection. We evaluate Poem on programs written in OpenCL, C, Java and Swift, and compare it against nine learning-based methods. Experimental results show that Poem consistently outperforms all competing methods across evaluation settings.

*Corresponding faculty authors: Zhanyong Tang and Zheng Wang.
[1]Poem = Dee**p** C**o**d**e** **M**odeling.


*PACT '20, October 3–7, 2020, Virtual Event, GA, USA*

ACM ISBN 978-1-4503-8075-1/20/10…$15.00
https://doi.org/10.1145/3410463.3414670

## CCS CONCEPTS

• **Computer systems organization**; • **Software and its engineering** → **Compilers**; • **Security and privacy** → *Software security engineering*;

## KEYWORDS

Program Modeling, Code Optimization, Machine Learning

**ACM Reference Format:**
Guixin Ye, Zhanyong Tang*, Huanting Wang, Dingyi Fang, Jianbin Fang, Songfang Huang, and Zheng Wang. 2020. Deep Program Structure Modeling Through Multi-Relational Graph-based Learning. In *Proceedings of the 2020 International Conference on Parallel Architectures and Compilation Techniques (PACT '20), October 3–7, 2020, Virtual Event, GA, USA.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3410463.3414670

## 1 INTRODUCTION

Over the last two decades, machine learning has emerged as a viable means for constructing heuristics for various program-related tasks including code optimization [54]. There is now ample evidence showing that machine-learned heuristics can outperform hand-tuned approaches [10].

A key challenge for applying machine learning to programs is that it requires programs to be represented as a sequence of numerical values (such as the number and type of instructions) that serve as inputs to a machine learning model. Traditionally, such program representations were determined by experts through trials and errors. However, since programs are syntactically unbounded graph structures and that there is an infinite number of these potential features, finding the right features is a non-trivial task.

More recent studies have leveraged the advances in deep learning (DL) to model and reason about code structures [19–21, 33, 36, 65, 66]. Compared to classical machine learning approaches, DL has the advantage of not requiring expert involvement to manually tune representations for program structures; instead, it automatically captures and determines them from training samples [6].

Existing DL-based approaches for program modeling typically utilize recurrent neural networks (RNN), like the long short-term memory (LSTM) or a variant of it, to model code structures [20, 33]. Such approaches work by treating source code and its structure – for example, the abstract syntax tree (AST) – as a sequence of tokens. However, LSTM is designed for processing a sequential sequence [35] and is ill-suited for capturing the program control and data

flows – which should be better represented as a graph instead of a sequence of tokens. As a result, prior methods only capture the shallow, textual structure of the source code and fail to capitalize on the rich and well-defined semantics of the program structure. To better model the complex data and program structures – which were traditionally represented as graph structures in compilers for code analysis – we need an approach that could directly operate on and learn from the graph representation of the code. Doing so will allow the learning framework to preserve and reason about much of the control and data flow information that is essential for many program-related tasks.

The first effort in this direction is the recent work presented in [13], which employs a vanilla graph neural network (GNN) to learn representations from the graph representation of the AST or the control-data flow graphs (CDFGs). This is achieved by propagating information along the graph edges defined in a graph adjacency matrix. While there may exist multiple code relationships (edges) among any given node pair, their approach only captures the graph connectivity, leaving the graph edges as untyped. As such, it cannot tell if a direct connection between two nodes is a control or a data flow, neither distinguish other relationships like order for non-commutative operations. Intuitively, such information would be essential for characterizing the program behavior for many code-related tasks. By ignoring the different relationships, their GNN approach gives marginal improvement or even worse performance compared to the LSTM alternatives [13].

We present Poem, a better approach for modeling code structures. Poem operates on graph representations of the program with the capability to learn and aggregate multiple code relationships. It is designed to maintain sequential information like token order and operand values when trading sequential representation for graph representations. Poem automatically extracts such information from the AST and the CDFGs. It then combines and abstracts the extracted information to generate a numerical feature vector that captures much of the essential information of the syntax and semantics of the target program. We use the generated embeddings as an input to a standard neural network to support downstream processing tasks like code optimization and vulnerability detection.

As a departure from prior work [11–13], Poem uses graphs to represent *both* the syntactic and semantic information of programs and employes graph-based learning methods to learn to reason over *multiple* graph structures. With Poem, syntax information is encoded from the AST and IR nodes. To maintain much of the sequential syntactic and semantic information, we augmented the AST with additional edges. These edges allow us to encode sequential syntactic relationships (e.g., "token before/after") and semantic relationships (e.g., "variable last used here", "this statement is guarded by an if condition"). In addition to the AST, we also record the control and data flow information from the CDFG. Intuitively, information collected from the source code is language-dependent but agnostic to compiler implementations. By contrast, data collected from the IR captures much of the lower-level, language-agnostic but compiler-specific information that could not be directly obtained from the source code. By combining such information, we improve the generalization ability of the learning framework, as different tasks may require knowledge at different levels.

Unlike [13] that it only uses an adjacency matrix to encode the node connectivity of the AST or CDFG, we encode different node relationships, e.g., whether it is a child-parent connection on the AST or a data flow edge in the CDFG, in different matrices and relation graphs. At the core of Poem is a novel graph neural network that can learn over multiple relationships (or edge types) simultaneously. By representing the input program as multiple relation graphs with explicit control and data flows or syntactic information, Poem captures a greater range of intra-program relations than prior graph representations. A key advantage of Poem is that it uses a learnable function to aggregate information for individual relation graphs. As the aggregation function is tuned for each relation graph, it can capture each specific code relationship more precisely. This richer set of relationships improves the model's ability in learning useful program representation, which in turns leads to better performance of downstream processing tasks. We show that Poem is highly effective in learning abstracted code representations, allowing us to solve various tasks with performance better than state of the arts.

We demonstrate the benefits of Poem by applying it to four representative tasks that require a strong ability to reason about the program structure at different levels: heterogeneous device mapping, GPU thread coarsening, loop vectorization and code vulnerability detection. We evaluate Poem on benchmarks written in OpenCL, C, Java and Swift. We compare Poem against a wide range of machine-learning techniques. Experimental results show that Poem consistently outperforms competing methods across tasks and programming languages, by giving stronger performance improvement and demonstrating a better generalization ability across evaluation tasks.

This paper makes the following contributions:

- It presents the first graph-based learning framework[2] that can simultaneously model multiple edge types of the program graph (Section 2);
- It demonstrates how to use *multiple* graphs to represent both the syntactic and semantic structures of programs to support graph-based deep learning (Section 2.3);
- It is the first work showing how a general graph-based learning framework can be developed to deliver consistently better performance over LSTM-based alternatives across a range of code-related tasks (Section 4).

## 2 OUR APPROACH

Poem is designed to operate on graph representations of the AST extracted from the source code, and the CDFG obtained from the compiler IR. As we will show later in the paper, Poem can adapt to different programming languages, hardware platforms and tasks.

### 2.1 Overview of Poem

Figure 1 depicts the architecture of Poem. It takes as input the program source code. It first uses a source rewriter extended from [20] to normalize the identifiers. Next, it builds the AST and CDFG using standard compiler passes. We extend the standard AST with additional edges to carry data and control flow information at the source code level (Section 2.3). The AST and CDFG are presented as directed multigraphs, where statements, identifiers, and immediate

[2]Code and data are available at: [https://github.com/yeguixin/POEM.]

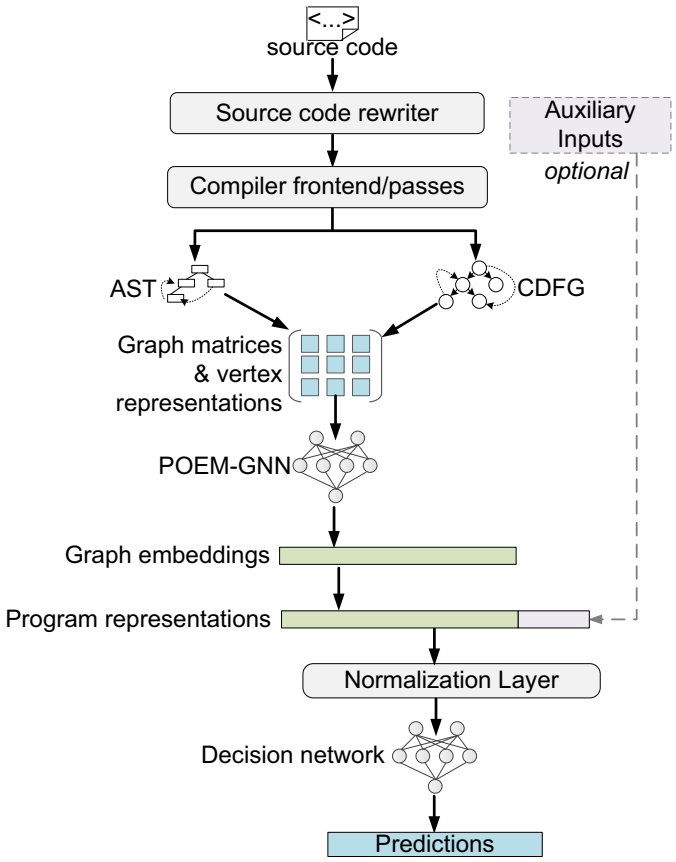



**Figure 1: Poem operates on the program graph matrices and vertex representations derived from the AST and CDFG. It uses the Poem-GNN to extract useful program representations (i.e., graph embeddings), which are encoded as vectors of numerical values. The embedding vectors and the *optional* auxiliary input are concentrated and normalized and passed to the decision network for a given prediction task.**

values are vertices, and a direct relationship (e.g., parent-child, data or control flow, etc.) between two vertices is recorded as an edge. As there may exist multiple relationships (or edges) among a pair of vertices, we use a relation graph to record a specific type of relationships. In this work, we wish to capture 10 relationships from the AST and CDFG (Section 2.3), leading to 10 relation graphs. The vertex connectivity of a relation graph is represented as a program graph matrix (Section 2.4).

The Poem-GNN takes in the program matrices and initial vertex (or node) representations to learn program representations called *embeddings* that are represented as a vector of numerical values. Like [20], the user can also optionally supply auxiliary inputs to give additional information about runtime parameters. We concentrate the graph embeddings and ancillary data to form a fixed-length feature vector, which is first normalized and then passed to a heuristic model (based on a standard fully-connected, dense neural network) to make a prediction. Poem-GNN and the dense network are trained together so that the graph representation is tuned for the task.

**Table 1: Code relationships considered in this work**

| Source | Relationships |
|---|---|
| **AST** | ASTChild, NextToken, ComputedFrom, GuardedBy, Jump, LastUse, LastLexicalUse |
| **IR** | Sequential-IR flow, data flow, and control flow |

Unlike [13] that is only able to model the node connectivity, we extend the GNN to model multiple edge types (e.g., control, data, jump, token sequence, etc.). This capability allows Poem to distinguish different relationships of the code, whether it is an if branch or a function call. In Section 4, we show that our approach consistently outperforms prior methods that are based on the vanilla GCN [13] or LSTM [20].

**Roadmap.** The remaining of this section is organized as follows. We describe how we extract the AST and CDFG in Section 2.2. From Sections 2.3 to 2.5, we explain how to model and encode the different code relationships in Sections 2.3-2.5. Finally, we describe our multi-relational graph neural network from Sections 2.6 to 2.8.

## 2.2 AST and CDFG Construction

We construct the AST from the standardized source code using a compiler front-end parser (e.g., Clang for C). We construct the CDFG from the compiler IR after applying standard compiler data-flow analysis and optimizations like dead-code elimination, constant propagation, and common subexpression elimination. An AST contains syntax nodes and syntax tokens. The former corresponds to nonterminals in the language grammar, e.g., an `if` statement (`IfStmt`) and function names; and the latter corresponds to terminals like literals and constant values. The CDFG captures semantic information that reflects standard compiler transformations. Figures 2(b) and 2(c) respectively show the LLVM IR and the extracted CDFG for the OpenCL kernel (after source code rewritten) given in Figure 2(a). To simplify the IR, we replace identifiers with its type. For example, `%4` at lines 3 and 5 of Figure 2(b) will be replaced with its data type `i32`.

## 2.3 Code Relationships

We record ten relationships from the AST and the IR, listed in Table 1. These include relationships that can be directly obtained from the CDFG, like the sequential order of the IR instructions, and the control and data flow. We also augment the AST with six additional edges, described as follows.

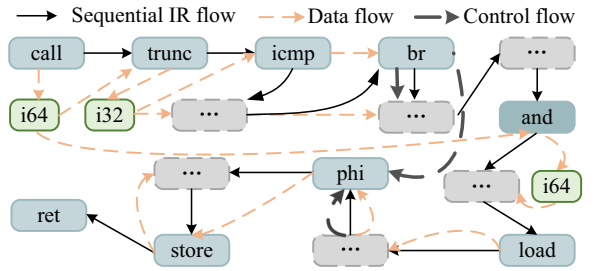
A standard AST has just one type of edges, i.e., the `ASTChild` edge that connects the children nodes with their parents. To capture additional syntax and data and control flow information of the AST, we introduce six additional edges to the AST, following the method described in [7]. The additional edges are essential because they maintain local sequence order such as the ordering of variable use and operations. As the AST edges do not induce an order on children of a syntax node, we add `NextToken` edges to connect each syntax token to its successor. This edge is used to capture the order of opcode and operands for statements. For each assignment, $v = expr$, we connect $v$ to all variable tokens occurring in expression, $expr$, using `ComputedFrom` edges. We connect each AST token of

```
__kernel void A(__global uint4* a, __global uint4*
      b) {
  unsigned int d = get_local_id(0);
  if (d > 0) {
    b[d] = a[d] + a[d + 1];
  } else {
    b[d] = 0;
  }
}
```

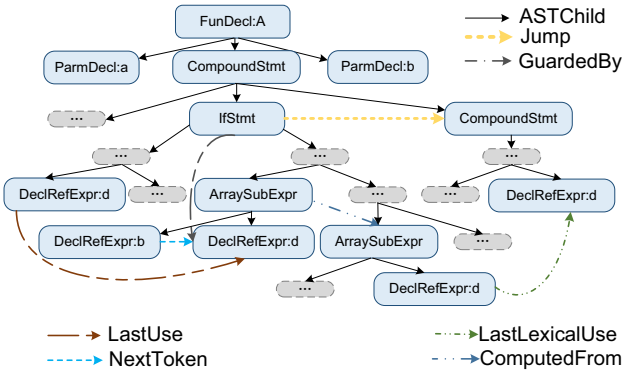
(a) An example OpenCL kernel

```
define void @A(i32 ,i32) {
    %3 = tail call i64 @get_local_id(i32 0)
    %4 = trunc i64 %3 to i32
    %5 = icmp eq i32 %4
    %6 = and i64 %3
    br i1 %5, label %15, label %7
    ...
    %16 = phi i32 [%14, %7], [zeroinitializer, %2]
    %17 = %1, i64 %6
    store %16, %17
    ret void
}
```

(b) LLVM IR




(c) Control and data flow graph (CDFG)




(d) AST with additional data and control flow edges

**Figure 2: A standardized OpenCL kernel (a), and its corresponding LLVM IR (b), CDFG (c) and augmented AST (d).**

a variable to the variable's enclosing guard expressions using a `GuardedBy` edge. For example, for the `if` statement in Figure 2(a), we add a `GuardedBy` edge from `b` to the AST node corresponding to `d > 0`. We use a `Jump` edge to connect variables that have control dependencies. The `GuardedBy` and `Jump` edges allow us to record the relation of diverging control flow. We connect all uses of the same variable using `LastUse` edges to capture the use of variables, where a special case is variables in `if` statement and we connect such type of variables using `LastLexicalUse` edges. For instance, for the `if` statement in Figure 2(a), we would link the two occurrences of `d`: one in the loop head, the other in the loop body. Figure2(d) shows the resulting AST after processing the OpenCL kernel given in Figure2(a), which is augmented with the additional edges.

Finally, for each graph edge, we also add a respective *backward* edge (by transposing the adjacency matrix), doubling the number of edges and edge types. These backward edges help with propagating information across Poem-GNN (Section 2.6) and make the model more expressive.

## 2.4 Program Graph Matrices

We convert the augmented AST and CDFG to separated *relation graphs* - one graph for each of 10 relationships given in Table 1. A relation graph is a directed graph, $G = (V, E)$, that contains the AST or IR node (vertices), $V$, and edges $E$, that indicates the existence of a given relationship between two vertices, such as data, control and ASTChild, etc. We use an adjacency matrix to recode the edge connections of each relation graph, 10 matrices in total. A value of 1 for matrix element $e_{i,j}$, represents there exists a direct connection or relation from node $i$ to node $j$, while a value of 0 indicates the two nodes are not directly connected.

## 2.5 Vertex Representations

To capture the syntactic and semantic meanings of the relation graph vertices, we map every instructions (e.g., AST nodes like ParamDecl, IfStmt and IR opcodes), constant, and variable to a vector representation by lookup in a fixed size embedding table. To do so, we first construct a vocabulary of frequently appeared words from the training corpus, where we store the AST and IR extracted from training programs. As variable and function names and constant values can be of an arbitrary length, we encode them as tokens (i.e., letters, symbols and numbers [0-9]). During the model deployment stage, if a word of the input program is not presented in the vocabulary, it will be encoded at the token level.

Once the vocabulary is constructed, we apply word2vector [43] to map each word and token of the vocabulary to an embedding space of integer values. The word2vector model is trained on training benchmarks of the target programming language and compiler IR. Training is largely independent of the optimization task, whose goal is to map individual words to a point in a latent multidimensional space where words that are frequently appeared together are mapped to integer values close to each other in the space. Doing so allows us to capture much of the syntactic relation of language constructs. For example, it allows the model to learn that an `if` statement must precede an `else` statement. In this work, we map each of the tokens and words in the vocabulary to a single fixed-length embedding vector of 100 features. This vector captures many characteristics of the code, such as syntax and shadow semantic similarities.

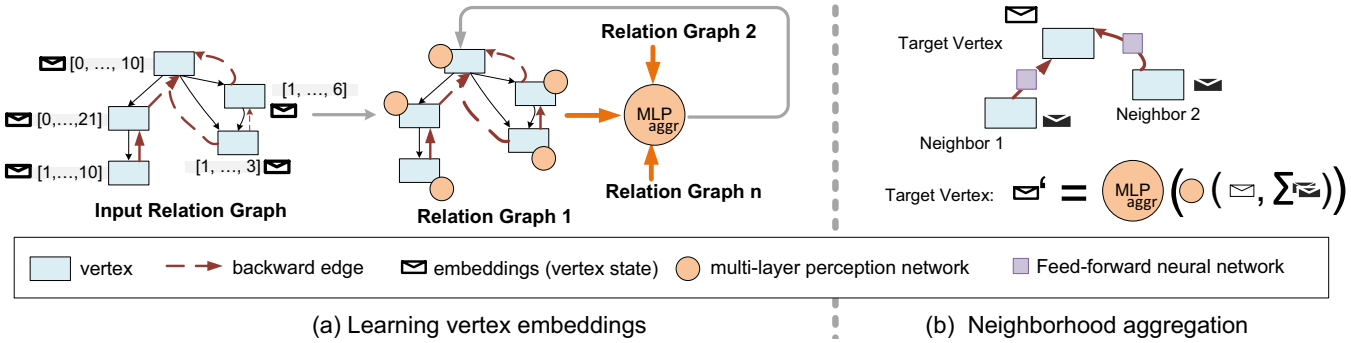



**Figure 3: Using Poem-GNN to learn vertex embeddings. The initial vertex embeddings are generated using word2vec (Section 2.5), which are then iteratively updated during the learning process (a) by aggregating information across neighbors and information from other relation graphs for the same vertex (b).**

## 2.6 Poem-GNN: A Multi-Rational Graph Neural Network

The adjacency matrices and vertex embeddings of relation graphs are passed to the Poem-GNN to map the inputs to a one-dimensional embedding of 100 features. The Poem-GNN consists of several stacked GNN embedding layers based on the multi-layer perception (MLP) network, so that it can incorporate higher degree neighborhoods across relation graphs. We choose MLP because it is proven to be effective in learning embeddings for directed graphs [61], but other neural network architectures (like GRU [17]) can also be used. We use an AutoML tool [4] to search for the optimal number of embedding layers from training data (see also Section 4.5.1).

*2.6.1 Neighborhood aggregation.* Each embedding layer follows a neighborhood aggregation scheme (Figure 3b), where the $d$ dimensional representation vector, $h_v$, of a graph vertex, $v$, is computed by recursively aggregating and transforming the representation vectors of its neighboring nodes. Vertices are initialized with the embeddings given by word2vec (Section 2.5) and then exchange information by transforming their current state and sending it as a message to all neighbours in the graph. At each vertex, messages are aggregated and then used to update the associated node representation at the next embedding layer (referred to as the next iteration) [58]. The added backward edges (Section 2.3) enable backwards propagation of information. After repeating this process of updating vertex states for a fixed number of iterations a *readout* function (Section 2.6.3) is used to aggregate the vertex representations to a single numerical vector across multiple relation graphs to be used as the program representation.

*2.6.2 Multi-relation modeling.* One of the novel aspects of the Poem-GNN is that it can propagate and aggregate information across multiple relation graphs, being it control and data at the IR or AST. As illustrated in Figure 3a, we achieve this by first using learnable, relation-specific MLP layers, $MLP_\ell$, to compute new graph states of individual relation graphs through neighborhood aggregation. Specifically, we use a feed-forward neural network to implement neighborhood aggregation for communicating the neighbor embeddings to the reference vertex (Figure 3b). We then apply a MLP-based aggregation function, $MLP_{aggr}$, to aggregate and update states for identical AST or IR nodes (locating using the matrix indices) across relation graphs. Note that our current implementation aggregates information of AST and CDFG relation graphs as two separate groups as there is often no one-to-one mapping between AST and IR nodes after applying standard code transformations. However, the embeddings learned for the AST and CDFG will be aggregated during the readout stage.

Formally, we use forward propagation to update the state, $h_v^t$, for vertex, $v$, of a relation graph as:

$$h_v^{t+1} := \sigma(MLP_{aggr}(\sum_{\ell} \sum_{(u,v)\in A_\ell} MLP_\ell(h_u^t))) \tag{1}$$

where $A_\ell$ are the directed edges between a node pair, $u$, and $v$, and $MLP_\ell$ is a relation-graph-specific message propagation function. Note that $MLP_{aggr}$ and $MLP_\ell$ are learnable functions, that are updated during the training process. The initial vertex state, $h_v^0$, is created using the word2vec embedding method described in Section 2.5. It is worth noting that the vertex representation get more refined and global as the number of iterations increases.

*2.6.3 Result readout.* Once we have performed the neighbourhood aggregation procedure for a fixed number of times (determined by the number of embedding layers), we will obtain a new set of embeddings for each vertex. Through this process, the vertex knows more about their own information (features) and that of neighbouring nodes. This creates an even more accurate representation of the relation graphs. To represent the input AST and CDFG, we use a readout function to concatenate graph representations across all the neighborhood aggregation iterations and embedding layers, to form a single numerical vector, $h_G$, as the global program representation of all ($m = 10$) relation graphs, $G_i$:

$$h_G = \text{CONCAT}\left(\sum_{i=1}^{m} \left(\left\{h_{v,i}^{(t)} \,|v \in G_i\right\}\right) |t = 0, 1, \ldots, n\right) \tag{2}$$

where $t = 0, 1, ..., n$, is the neighborhood aggregation iterations. This readout function produces the global embedding for all relation graphs, given individual vertex embeddings.

*2.6.4 Alternative modeling approaches.* A naïve alternative to our approach is to apply a standard GNN to individual relation graphs and then concatenate the embedding outcomes of individual graphs.

However, this approach does not allow information to be exchanged during each neighborhood aggregation iteration. In Section 4.5.2, we show that this approach gives poor learning performance. As a result, simply applying the GNN presented in [13] to multiple relation graphs does address the issues of learning multiple code relations. We have also considered the recently proposed Relational Graph Convolutional Network (RGCN) [50] as it can model different relationships through multi-edge encoding. However, the RGCN has a significant drawback – the number of parameters drastically increases for larger graphs. This increase in parameters can lead to overfitting, especially when the number of training samples is limited. Poem sidesteps this problem by restricting the relation-specific learnable function to a smaller subgraph of the entire program. Furthermore, unlike the RGCN that only operates on a single graph, Poem also supports information aggregation of the distinct AST and CDFG. In Section 4.5.2, we show that our approach outperforms the RGCN alternative.

## 2.7 Graph Embeddings

The graph embedding vector, produced by the Poem-GNN, together with any auxiliary data is first normalized to a range of 0 and 1 by the normalization layer. Normalization is essential as it prevents the range of single feature being a factor in its importance. The normalized feature vector is then fed to the dense network for downstream processing (Figure 1), e.g., classification. This final feature vector captures many characteristics of the code, such as semantic similarities, control and dependence flows, combinations, and analogies.

## 2.8 Train the Poem-GNN

We train the Poem-GNN and the dense network together using back-propagation. Training is performed on batched training samples where each sample contains a ground-truth label. We use the standard cross-entropy loss as the objective function. This function is shown to be a good fit for sigmoid and softmax activation functions (both are also standard functions for classification) [64] used by Poem. It is defined as:

$$\mathcal{L}_G = -\sum_{i=1}^{N} y_{o,c} log(p_0, c) \tag{3}$$

where $N$ is the number of classes (i.e., running the code on the CPU or GPU), $y$ takes a binary value (0 or 1) that indicates if label $c$ is the correct classification for training sample $o$, and $p$ is the predicted probability for sample $o$ to be of class $c$.

# 3 EXPERIMENTAL SETUP

To demonstrate the benefit of Poem, we use it to tune performance optimization heuristics for OpenCL and C programs. To evaluate the generalization ability of Poem in modeling program structures, we also apply it to detect code vulnerabilities for C, Java and Swift. In total, we apply Poem to four case studies and compare it with nine prior machine-learning-based approaches (including models using hand-crafted features) across eight distinct hardware platforms. In some cases, we also compared to expert-tuned heuristic models. Note that we use the same model structure for Poem across tasks. Table 2 lists the machine learning models used in the evaluation.

**Table 2: Machine learning methods used in evaluation**

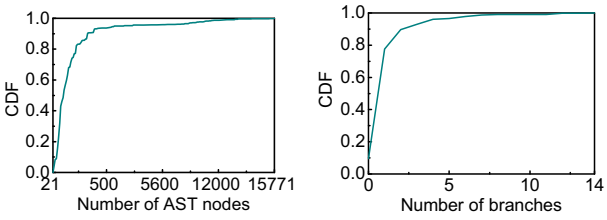
| Approach | Code representation | Model | Use cases |
|---|---|---|---|
| **Grewe et al.** [30, 53] | Manual features | Decision Tree | Case study 1 |
| **DeepTune** [20] | Source code token sequence | LSTM | Case studies 1-3 |
| **Inst2vec** [12] | LLVM IR tokens | LSTM | Case studies 1-2 |
| **GNN-AST** [13] | AST | Vanilla GNN | Case studies 1-3 |
| **GNN-CDFG** [13] | CDFG | Vanilla GNN | Case studies 1-3 |
| **Magni *et al.*** [42] | Manual features | Neural Networks | Case study 2 |
| **NeuroVectorize** [33] | Token sequence | LSTM + Reinforcement learning | Case study 3 |
| **uVuldeepecker** [66] | AST | Bidirectional-LSTM | Case study 4 |
| **Lin *et al.*** [40] | AST | Bidirectional-LSTM | Case study 4 |
| **Poem** | AST + CDFG | multi-relational GNN | Case studies 1-4 |

## 3.1 Case Study 1: Heterogeneous Mapping

**The problem.** In this task, we wish to build a predictive model to determine if the CPU or the GPU gives faster performance for a given OpenCL kernel.

**Methodology.** We use the dataset of [20], which provides labeled CPU/GPU instances for 256 OpenCL kernels extracted from seven benchmark suites. The data were collected on two CPU-GPU platforms: one with an Intel Core i7-3820 CPU and and AMD Tahiti 7970 GPU, and the other has an Intel Core i7-3820 CPU and an NVIDIA GTX 970 GPU. By varying dynamic inputs, this dataset consists of 680 labeled instances on each platform. The compilation of some kernels ended with the presence of errors. We have manually fixed those broken OpenCL kernels to use the entire dataset. We apply 10-fold cross-validation train and test a model. We repeat this process ten times (folds), with each of the seven suites used exactly once as the testing data. The CDF diagram in Figure 4 shows the distribution for the number of AST nodes and branches for kernels in this dataset. For this dataset, over 50% of the kernels have more than 100 AST nodes, and over 75% of the kernels have one or more branches. Since the dataset in [20] is small, it may not provide sufficient training samples for a deep learning method. To evaluate on a larger training dataset, we use DeepSmith, an OpenCL program synthesizer [19], to generate 12,000 valid and compilable OpenCL kernels as additional training data for deep-learning-based competing methods and Poem. This second experiment was performed on our GPU platform that uses a 3.2 GHz 6-core Xeon E5-2667 CPU and an NVIDIA Titan XP GPU. We profiled all benchmarks from the DeepTune dataset to obtain the ground-truth label on this platform.

**Competitive methods.** For this case study, we compare Poem with five machine-learning models. These include Grewe *et al.* [30] that




(a) #AST nodes (b) #Branches

**Figure 4: Cumulative distribution functions (CDF) for the number of AST nodes (a) and branches (b) in the DeepTune OpenCL kernel dataset [20].**

uses hand-tuned features, and DeepTune [20] that uses LSTM to extract code representations from source code token sequence. We also compare to Inst2vec [12], a LSTM-based model that operates on a graph representation of the LLVM IR. For graph models, we compare to the two GNN variants presented in [13] offers GNN-CDFG and GNN-AST which operate on the CDFG and AST respectively. Like all prior work, for this case study, we use two dynamic values, the workgroup size and the host-device memory transfer size, both are available to the OpenCL runtime, as the auxiliary inputs (or features) to all predictive models. Results of this case study is presented at Section 4.1.

## 3.2 Case Study 2: Thread Coarsening

**The problem.** This task builds a model to determine how many parallel threads should be merged together to achieve faster execution time. This is a problem known as determining the thread coarsening factor for OpenCL [42]. Here, we wish to build a model to determine for each kernel, which of the six coarsening factors, 1, 2, 4, 8, 16, and 32, should be used for a given kernel (where a factor of 1 means no coarsening).

**Methodology.** We replicate the setup of [13, 20, 42] by testing each approach using the labeled dataset given in [42]. This dataset contains 17 OpenCL kernels extracted from three benchmark suites. The data were collected from five distinct GPU platforms: AMD HD 5900, AMD Tahiti 7970, NVIDIA GTX 480, and NVIDIA K20c. We also extended our evaluation to NVIDIA Titan XP by profiling the same OpenCL kernels on this GPU architecture. Like [20], we use *leave-one-out* cross-validation for this task because the benchmark set is small. This works by selecting one benchmark for testing and using the remaining ones for training. This task is designed to evaluate if our approach can effectively support transfer learning [60], a technique for reusing the knowledge learned from one task to speed up the learning for another task.

**Competitive methods.** We compare Poem against three approaches: Magni *et al.* [42] that uses hand-tuned features, DeepTune, Inst2vec, and GNN-CDFG and GNN-AST presented in [13]. To apply transfer learning, we first train an initial deep learning model on the dataset given in [20]. We then use transfer learning to fine-tune the trained model on training data used for this task. Fine-tuning is done by simply copying the learned parameters of case study one to initialize the model and then training the model as normal using cross-validation. Note that for this task, the OpenCL kernel is the sole input and coarsening factor is the predicted output. Results of this case study are presented in Section 4.2.

**Table 3: Dataset for vulnerability detection.**

| Source | Language | #samples | #positive samples |
|---|---|---|---|
| SARD & NVD | C | 156,668 | 78,334 |
| | Java | 60,242 | 30,121 |
| GitHub | C | 10,400 | 5,200 |
| | Swift | 4,124 | 2,062 |

## 3.3 Case Study 3: Loop Vectorization

**The problem.** This task targets the classic compiler optimization problem of loop vectorization. It aims to build a predictive model to determine the optimal vectorization factor (VF) and the interleaving factor (IF) for individual loops. The first parameter determines how many instructions to pack together from different loop iterations, while the latter decides the stride of the memory accesses of the packed instructions. Prior work has shown that the two parameters can have a substantial impact on the resulting vectorization performance[33, 45]. We consider 35 combinations of VF (1, 2, 4, 8, 16, 32, 64) and IF (1, 2, 4, 8, 16), which are found to be useful in [33].

**Methodology.** We use LLVM version 9.0 as the compiler. We configure the VF and IF on a per loop basis by placing the LLVM/-Clang vectorization directives, e.g., `loop vectorize_width(VF) interleave_count(IF)`. We replicate the evaluation of NeuroVectorizer [33], by using the 6,000 synthetic loops generated from the LLVM vectorization test set to train a model and then test the trained model on hand-written programs from MiBench [32] and PolyBench [29]. Our evaluation platform uses an 3.6 GHz Intel Core i7 CPU with 64GB RAM.

**Competitive methods.** We compare Poem against three supervised learning models (DeepTune and the two GNN variants in [13]), as well as Polly [31], a LLVM-based code vectorizer based on the polyhedral model. In addition to these, we also compare to NeuroVectorizer [33], a recently proposed reinforcement-learning-based approach. NeuroVectorizer first learns the program representations through LSTM. The representations are then used by a reinforcement learner to search for the best configuration until a convergence threshold is met. Due to the nature of reinforcement learning, NeuroVectorizer can incur significant search overhead. It takes minutes to search for the vectorization configuration for a *single* loop on our evaluation platform. By contrast, our approach takes less than 100ms (including constructing the relation graphs) to make a prediction. The results are presented in Section 4.3.

## 3.4 Case Study 4: Vulnerability Detection

**The problem.** In this task, we build a model to detect if a given source code snippet contains one of the 2019 CWE top-25 most dangerous software errors [26] at the function level.

**Methodology.** As summarized in Table 3, we use a dataset of 231,434 samples with source languages in C, Java and Swift, where half of the samples are vulnerable code. The vulnerable code samples are collected from the standard vulnerable code databases, including the national vulnerability database (NVD), common vulnerabilities and exposures (CVE) and open datasets collected from GitHub. The vulnerable-free samples are obtained by applying the corresponding patch to the vulnerable code. We apply 10-fold cross-validation to train and test a predictive model (see also Section 3.1 for how we perform cross-validation).

**Competitive methods.** For this case study, we compare Poem against two state-of-the-art deep-learning-based vulnerability detection models: uVuldeepecker [66] and Lin *et al.* [40]. Results of this case study is presented in Section 4.4.

## 3.5 Performance Report

To measure execution time for case studies 1-3, we run each test case repeatedly until the 95% confidence bound per model per input is smaller than 5%. For case study 4, we report the accuracy, and the false-positive and the false-negative rates. A false positive is when the model predicts a code snippet has a vulnerability while it does not, and a false-negative is when the model fails to detect a vulnerable code sample. For code vulnerability detection, we would like to achieve high accuracy with low false-positive and false-negative rates.

We report the geometric mean across experimental runs or test cases. The geometric mean is widely considered to be more robust and meaningful than the arithmetic mean for performance measurements [27].

## 3.6 Implementation and Training Settings

We implement Poem on Tensorflow v1.8. To build the AST, we use Clang [2] for OpenCL, C and Swift, and ANTLR [1] for Java. To extract the CDFG, we use LLVM [3] for OpenCL, C and Swift, and Soot [5] for Java.

All deep learning models were trained in an end-to-end fashion using minibatch stochastic gradient descent (SGD) and the Adam optimizer [37]. For fair comparison, we use NNI [4], an AutoML tool to determine the training hyper-parameters, including the learning rate and batch size unless these are given in the published implementation. We train the models using two NVIDIA GTX 1080 GPUs. Note that we set aside 1/10th of the training data to use for *validation* during the training process. Training terminates when the loss does not improve within 20 consecutive training epochs, or reaches a 99% accuracy on the valuation set, or meets the termination criteria given in the published implementation. For each model, we use the configuration that yields the best results on the validation set. Because we use the geometric mean instead of the arithmetic mean and as the deep learning models are initialized with random weights, performance numbers can deviate from the source publications.

# 4 EXPERIMENTAL RESULTS

In this section, we first present results for the four case studies described in Section 3, showing that Poem outperforms all alternative methods in each task. We then provide a detailed analysis of the working mechanism of Poem.

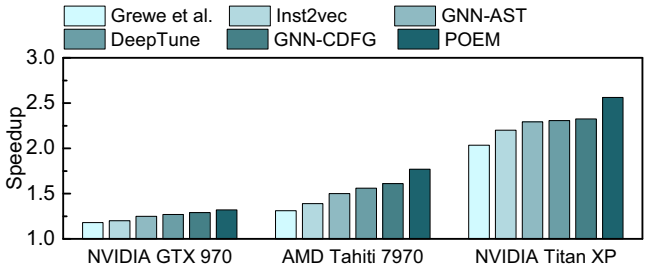



**Figure 5: Speedups (geometric mean) over a GPU-only baseline for heterogeneous mapping (case study 1). Poem outperforms alternative methods on both platforms.**

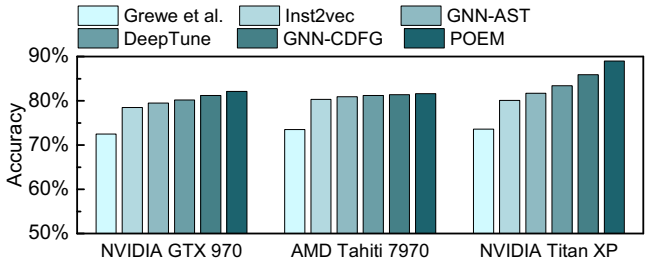



**Figure 6: Prediction accuracy for heterogeneous mapping (case study 1). Poem gives the highest prediction accuracy.**

## 4.1 Case Study 1: Heterogeneous Mapping

Figure 5 shows the performance improvement. Results on NVIDIA GTX 970 and AMD Tahiti 7970 was obtained on the DeepTune dataset, while results on NVIDIA Titan XP were obtained by first training the models on additional synthetic OpenCL kernels and then testing the trained model on hand-written kernels from the DeepTune dataset. The baseline is a GPU-only strategy that always uses the GPU for kernel execution. As we report the geometric mean, the speedup number can deviate from the sour publications.

For this case study, Poem outperforms all other approaches on all platforms. On NVIDIA GTX 970, we observe small performance improvement over the GPU-only baseline for all methods. On this platform, Poem gives the best overall speedup of 1.32, albeit its improvement is relatively small. By contrast, the benefit of using the right device on the AMD Tahiti 7970 GPU is larger. On this platform, Poem achieves a mean speedup of 1.8x, which translates to an improvement of 13% over the second-best model, GNN-CDFG. All deep-learning models benefit from additional training data on the NVIDIA Titan XP platform, where Poem delivers the highest mean speedup of 2.6x.

If we look at the prediction accuracy in Figure 6, we see that Poem also delivers the highest accuracy on all platforms; albeit Poem gives modest accuracy improvement on the DeepTune dataset because the number of training samples is small. However, when using a larger training dataset, it is able to boost the prediction accuracy from 82% to 89% over DeepTune and Inst2vec. We note that on the

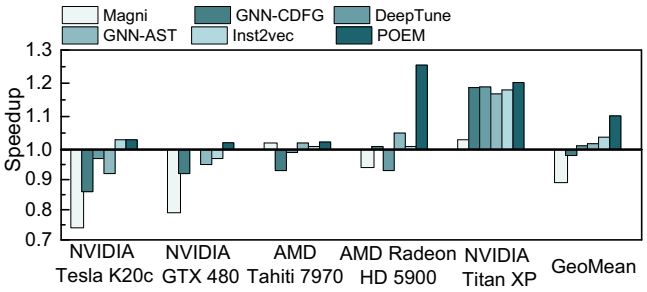



**Figure 7: Performance of thread coarsening (case study 2). Poem is the only method that gives an overall speedup.**

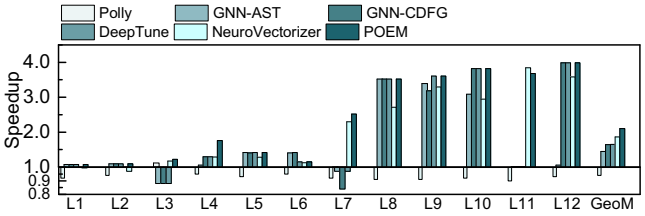



**Figure 8: Speedup over the LLVM default loop vectorizer (case study 3). Poem delivers the highest overall speedup and is the best-performing model for most of the testing loops.**

**Table 4: Performance for code vulnerability detection (case study 4).**

| Metrics | | uVuldeepecker | Lin *et al.* | POEM |
|---|---|---|---|---|
| **C** | Accuracy | 80.0% | 88.0% | **90.9%** |
| | FPR | 31.6% | 30.5% | **3.1%** |
| | FNR | 9.4% | **7.1%** | 8.9% |
| **Java** | Accuracy | 78.3% | 72.0% | **84.4%** |
| | FPR | 27.7% | 45.4% | **20.7%** |
| | FNR | 15.7% | 10.3% | **8.1%** |
| **Swift** | Accuracy | 77.7% | 74% | **89.0%** |
| | FPR | 21.0% | 23.2% | **19.3%** |
| | FNR | 23.6% | 28.3% | **9.9%** |

AMD platform, for most of the cases that Poem mispredicts, the difference in performance between the GPU and the CPU is small. As a result, such a misprediction has little impact on the overall performance. Overall, Poem delivers the highest mean speedup and prediction accuracy across all evaluation platforms and datasets.

## 4.2 Case Study 2: Thread Coarsening

Figure 7 shows the speedup for thread coarsening over a baseline that uses a coarsen factor of 1 (i.e., no coarsening). In this task, we apply transfer learning to port the deep learning model trained for case study 1 (using the DeepTune dataset) to predicting thread coarsening . Poem is the only model that consistently delivers performance improvement across GPU platforms, albeit the improvement is modest. The modest improvement is expected as a theoretically perfect predictor would only achieve a mean speedup of 1.28x. The GNN variants deliver poor performance for this task, leading to overall slowdown on three out of five evaluation platforms. DeepTune gives marginal improvements on two GPU platforms, but its performance is far from Poem on these platforms. Notably, on Tesla K20c, Poem and Inst2vec are the only predictive models that give a speedup. On HD5900 and Titan XP, Poem gives an overall speedup of over 1.2x, improving DeepTune by 20%. Overall, Poem achieves consistently higher speedups when compared to that of other methods. This experiment shows Poem can effectively support transfer learning when the training corpus is small.

## 4.3 Case Study 3: Loop Vectorization

Figure 8 shows the speedup for predicting loop vectorization configurations. The baseline is the LLVM default loop vectorization setting. Machine learning-based methods outperform the polyhedral-based Polly optimizer for all test loops except for L3, where Poem still gives better performance than Polly. GNN and DeepTune match or outperform NeuroVectorizer on several high-speedup test cases (L8, L9, L10), despite that NeuroVectorizer incur significantly more expensive compile-time overhead. However, GNN and DeepTune give no performance improvement or even slow down over LLVM "-O3" for several loops, including L3, L7, and L11. After having a close examination of these cases, we found that these loops contain a branch with the loop body that does not captured by DeepTune and the simply graph representation used by GNN. Poem gives or matches the best performance for all test cases, except for L11. For L11, the performance of Poem is not far from the best-performing model (i.e., NeuroVectorizer). Overall, Poem gives an average speedup of 2.10x, leading to 13% improvement over NeuroVectorizer. The performance of Poem translates into 96% of the 2.17x speedup found by exhaustively trying all the vectorization configurations considered in this work.

Compared to NeuroVectorizer, Poem also has orders of magnitude less compile-time overhead (under a second versus 15 minutes compile time for the 12 testing loops). This indicates that the representation learned by Poem can effectively support the downstream loop vectorization task. An interesting question is if the embeddings learned by Poem can be used to improve the reinforcement search framework of NeuroVectorizer. We leave this as future work.

## 4.4 Case Study 4: Code Vulnerability Detection

In this experiment, we apply Poem to detect vulnerabilities of function-level source code written in C, Java and Swift.

Table 4 reports the higher-is-better accuracy metric and the two lower-is-better metrics: the false-positive rate (FPR) and the false-negative rate (FNR). Poem delivers the best accuracy with the lowest FPR and FNR. On the C dataset, Poem has an accuracy of over 90.9%, with a low FPR of 3.1%. Poem has a modestly higher FNR compared to Lin *et al.*, but it has a significantly lower FPR (3.1% vs 30.5%). A low PPR is important as it reduces the developer's time in investigating false alarms. For the Java and Swift datasets, all three approaches have relatively lower accuracy and higher FPR. This is largely due to the more complex language features like overriding of an external method. Such information is not captured by the initial vertex embedding method (word2vec used by the three

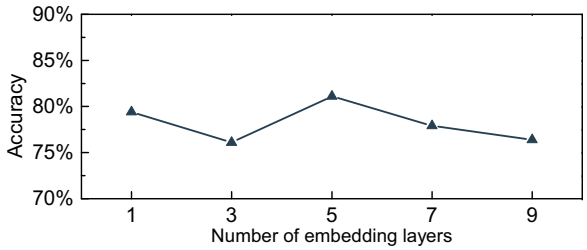



**Figure 9: Performance of Poem for heterogeneous mapping on AMD Tahiti 7970 as the number of embedding layers changes.**

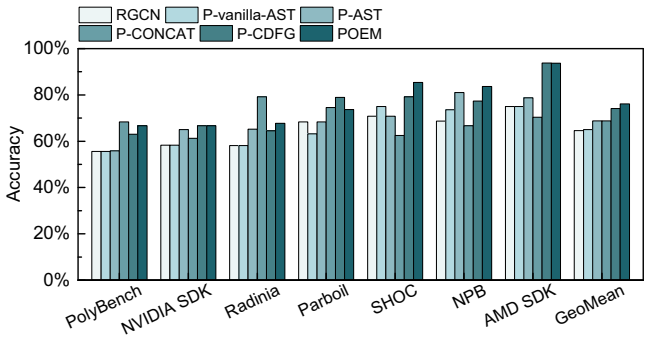



**Figure 10: Implementation variants of Poem on AMD Tahiti 7970 for heterogeneous mapping. Our implementation choices of Poem give the best overall performance.**

methods). Nonetheless, Poem outperforms the other two methods across language datasets by successfully detecting more vulnerable code samples with the lowest FPR.

## 4.5 Model Analysis

*4.5.1 Impact of embedding layers.* Using heterogeneous mapping as an example, Figure 9 shows how the performance of Poem changes on the DeepTune dataset on AMD Tahiti 7970. Increasing the number of embedding layers (and hence the number of neighborhood aggregation iterations - see Section 2.6) can improve the performance. However, the accuracy reaches a plateau when using five embedding layers and using more than that leads to overfitting and a drop of prediction accuracy on the validation set. The validation dataset is part of the training data but not the test dataset which is always not seen at the model design and training stage. We use the same procedure to determine the optimal number of embedding layers for each task, by comparing the resulting accuracy on the validation set. We found that using 3 to 5 embedding layers give a good performance for our tasks and using more embedding layers would require a larger training dataset.

*4.5.2 Impact of implementation choices.* In this experiment, we compare Poem to several variant implementations for performing heterogeneous mapping on AMD Tahiti 7970. The first is RGCN [50] that applies to the 10 code relations described in Table 2.3. Unlike Poem, RGCN takes in a single adjacency matrix that encodes all the node connectivities of the AST and CDFG, and the edge is encoded using a one-hot vector for representing different relations. The second variant, referred to as p-vanilla-AST, operates on a standard AST (without the additional edges described in Sec. 2.2). The third variant, referred to as P-AST, operates only on the augmented AST. The fourth variant, referred to as P-CDFG, operates only on the CDFG. This experiment uses heterogeneous mapping as a case study and is designed to evaluate the impact of exacting code information. The final variant, referred to as P-CONCAT, learns individual embeddings for each relation graph and then concatenates the individual embeddings for prediction. This evaluation is also known as the ablation study [28].

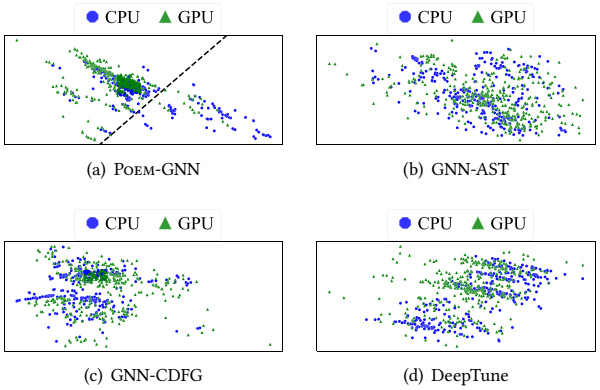
The results are given in Figure 10. While RGCN support modelling of multiple edge relations, it is less effective for modeling a combined graph from the AST and the CDFG. Its low performance is largely due to two reasons. Firstly, a simple combination of the AST and the CDFG, which are two heterogeneous graphs, to fit the RGCN can confuse the learning algorithm. Second, RGCN requires a large number of learnable parameters and is hard to generalize. Figure 10 also shows that using the standard AST is inadequate for capturing the essential program structures. By augmenting the AST with additional control and data flow information, P-AST improves P-vanilla-AST by 4%. However, using the AST or CDFG alone is insufficient, as both give an accuracy of less than 75%. P-CDFG correctly predicts 20 kernels where P-AST fails, while it fails on other 12 kernels where P-AST succeeds. P-CONCAT also gives lower performance compared to Poem, suggesting that simply combining the embeddings of relation graphs is less effective. This experiment suggests that we need to utilize and aggregate the information of the AST and the CDFG during the learning process. Poem offers exactly such capabilities, leading the best overall performance. It also shows that our multi-relational graph learning method improves a single concatenation strategy.

*4.5.3 Embedding visualization.* In an attempt to examine the learned code representation qualitatively, we provide a visualization of the t-SNE-transformed feature representations [41] extracted by the Poem-GNN pre-trained on the DeepTune heterogeneous mapping dataset. The representation exhibits discernible clustering in the projected 2D space as shown in Figure 11. Note that these clusters (with two different node colors) correspond to the two labels (CPU and GPU) of the dataset, verifying the model's discriminative power across different classes for this dataset. As can be seen from the diagram, the embeddings learned by Poem is more discriminative than the ones given by other methods, leading to a clearer linear boundary between the two classes (CPU and GPU).

*4.5.4 Training overhead.* The time for training Poem is dominated by training data collection. For case study 1, it took less than 24 hours to profile over synthetic 10,000 benchmarks for labeling the data. The time in model training and hyper-parameter tuning is less than 12 hours using two modest NVIDIA 1080 GPUs on 10,000 samples. Since training is only performed once, it is a *one-off* cost.

# 5 DISCUSSIONS

Naturally, there is room for future work and further improvement. We discuss a few points here.




(a) Poem-GNN (b) GNN-AST

(c) GNN-CDFG (d) DeepTune

**Figure 11: Visualization of the learned program representations for heterogeneous device mappings. We map the high-dimensional embedding space to a 2-dimensional space to aid clarity using t-SNE. The embeddings learned by Poem is more discriminative than the ones given by other methods, leading to a clearer boundary between the two classes.**

**Model interpretability.** Machine learning techniques, in general, have the problem of relying on black boxes. This problem is just as true for Poem. One way to gain insight into why the model fails to produce the desired result is to train an interpretable model (or so called surrogate models) like linear regressor to approximate the predictions of the underlying black-box model [49].

**Training samples.** Deep neural networks typically require a large volume of training data to learn over. However, there is often a shortage of benchmarks. Therefore, work on benchmark synthesis [21] is orthogonal to our approach.

**Training overhead.** Profiling training benchmarks to generate training data could be expensive. One way of reducing the training overhead is to use active learning [46] to only profile and label training benchmarks that are likely to improve the performance of the machine-learned model.

**Memory footprint.** Like all GNN approaches, the memory footprint of Poem increases as the graph size increases. However, we can reduce memory pressure by using sampling methods like GraphSAGE for batched training [34], i.e., operating on a subgraph of the entire program at a time.

**Other application domains.** We have shown that Poem can be generalized across a range of tasks. We envision that Poem can be applied to other applications which are beyond the scope of this work. It can be applied to detect malicious code by looking for suspicious and obfuscated patterns. It can be extended to regression-based problems like predicting the potential speedup for a code transformation option. A particularly interesting research direction would be to extend Poem to model the program structure at the binary level [59] for tasks like program verification and security.

## 6 RELATED WORK

Machine learning has demonstrated promises in automating the process of decision model construction for various code optimization tasks [54]. Many prior studies have shown that machine-learned models can outperform expert-crafted heuristics [9, 14, 15, 23–25, 48, 51, 55, 56, 62, 63]. However, a significant barrier still exists – programs must be represented as a set of features that serve as inputs to a machine learning tool. Traditionally, this requires experts' involvement to extract the crucial elements of the program.

Prior work automated this process of finding code representations by searching useful information from the compiler IR [38, 44]. These approaches still require experts to manually define the search space of a particular compiler IR implementation. As such, they offer little help in removing human experts from the loop.

In recent year, deep learning has been employed for modeling program structures. A key advantage of deep learning is that it can automatically find the right feature representations from training data without human involvement [6]. Prior work for deep learning on code typically employ recurrent neural networks (RNN) like LSTM or GRU to extract program representations from token sequences. For examples, DeepTune uses LSTM to extract program representations from tokenrized OpenCL code [20] and Inst2vec applies LSTM to sequentialized IR graphs [12]. Other work uses RNNs to summarize representations from the AST [8, 16, 33] or sparse matrices [65]. While RNN is a proven technique for natural language processing, it is mainly designed for processing a sequential sequence [35]. The prolbem for treating source code as a sequential token sequence is that statements can easily be separated by hundreds of lines of irrelevant code in sequential representations. As a result, RNNs are ineffective in modeling the complex program control and data flows - which should be better represented as a graph structure instead of a sequence of tokens. A graph representation not only enables the learning framework to leverage the well-defined program structures but also facilitates propagating information across the graph in a manner similar to typical compiler analyses.

An early attempt to use program graph structures for code optimization is presented in [47]. This approach requires careful hand-tuned features at the basic block level to extract information from the program graph. To predict an optimization option, it measures the similarity of the input program graph with the graphs of the training datasets. This strategy requires the training data graphs to be shipped with the compiler and hence does not scale well as the training program size increases. Furthermore, this approach does not abstract the language semantics and syntax a sufficiently high level, leading to expensive computation complexity for graph matching. Our approach eliminates the need for manual feature tuning, with a constant, lower-cost compile overhead, which is independent of the size of the training set.

Some of the most recent work has employed the recently proposed GNN to model code structures. For example, Miltiadis et al. [7] use GNN to model the AST to identify the misuse of names. The recent work presented in [13], which uses the GNN to model the AST or CFDG for OpenCL program optimization, is most closely related. This approach uses a vanilla GNN which treat all relationships equally as a single graph edge type, whether it is a node connection, or a data or control flow. This strategy misses much of the information that could otherwise be captured. Our approach advances [13] by capturing the different relationships within a unified learning framework. Compared to [13], Poem can leverage a

richer set of information by combining the AST and CFDG, leading to significantly better and more reliable performance. When preparing the final version of this paper, we noticed the pre-print version of ProGraML [18]. This approach also employs a GNN to the program's control and data flow graph extracted from the compiler IR. Our work was conducted independently and perhaps concurrently to ProGraML. Unlike ProGraML, Poem also utilizes information available in the AST, and uses individual learnable functions to model different code relationships.

The GNN family includes a diverse class of neural network architectures based on recurrent units [39, 57] and convolutional [22] and attention [52] methods. Poem represents the first work for leveraging GNNs to learn over multiple code relationships.

## 7 CONCLUSION

This paper has presented Poem, a general learning framework for supporting building machine-learned heuristics for code analysis and optimization. Poem is designed to automatically extract useful representations of programs to be used as inputs for machine learning tools. At the core of Poem is a novel Graph Neural Network (GNN) that can distinguish and aggregate information from different relationships within the program control and data flow graph and the abstract syntax tree. By providing a way to abstract and aggregate information from a well-structured program graph representation, our approach can capture a richer set of syntactic and semantic information than prior deep-learning-based approaches.

We demonstrate the generalization ability of Poem by applying it to four representative program optimization and analysis tasks spanning different programming languages. We perform extensive experiments to compare Poem with state-of-the-art approaches for each task. Experimental results show that Poem consistently outperforms prior methods, setting new state-of-the-art results for these tasks.

## ACKNOWLEDGEMENT

This work was supported in part by the National Natural Science Foundation of China (NSFC) under grant agreements 61972314, 61672427, 61872294 and 61972408, International Cooperation Projects of Shaanxi Province under grant agreements 2019KW-009 and 2020KWZ-013, an Ant Financial Science funded project and a UK Royal Society International Collaboration Grant.

## REFERENCES

[1] [n. d.]. ANTLR (ANother Tool for Language Recognition) . https://www.antlr.org/. ([n. d.]).
[2] [n. d.]. Clang: A language front-end and tooling infrastructure. https://clang.llvm.org/. ([n. d.]).
[3] [n. d.]. LLVM: A collection of modular and reusable compiler and toolchain technologies. https://llvm.org/. ([n. d.]).
[4] [n. d.]. Neural Network Intelligence. https://nni.readthedocs.io/. ([n. d.]).
[5] [n. d.]. Soot: A framework for analyzing and transforming Java applications. http://sable.github.io/soot/. ([n. d.]).
[6] Miltiadis Allamanis, Earl T Barr, Premkumar Devanbu, and Charles Sutton. 2018. A survey of machine learning for big code and naturalness. *ACM Computing Surveys (CSUR)* 51, 4 (2018), 1–37.
[7] Miltiadis Allamanis, Marc Brockschmidt, and Mahmoud Khademi. 2018. Learning to represent programs with graphs. In *ICLR*.
[8] Uri Alon, Shaked Brody, Omer Levy, and Eran Yahav. 2019. code2seq: Generating sequences from structured representations of code. In *ICLR*.
[9] Jason Ansel, Cy Chan, Yee Lok Wong, Marek Olszewski, Qin Zhao, Alan Edelman, and Saman Amarasinghe. 2009. PetaBricks: a language and compiler for algorithmic choice. *ACM Sigplan Notices* 44, 6 (2009), 38–49.
[10] Amir H Ashouri, William Killian, John Cavazos, Gianluca Palermo, and Cristina Silvano. 2018. A survey on compiler autotuning using machine learning. *ACM Computing Surveys (CSUR)* 51, 5 (2018), 1–42.
[11] Francesco Barchi, Gianvito Urgese, Enrico Macii, and Andrea Acquaviva. 2019. Code Mapping in Heterogeneous Platforms Using Deep Learning and LLVM-IR. In *2019 56th ACM/IEEE Design Automation Conference (DAC)*. 1–6.
[12] Tal Ben-Nun, Alice Shoshana Jakobovits, and Torsten Hoefler. 2018. Neural code comprehension: A learnable representation of code semantics. In *Advances in Neural Information Processing Systems*. 3585–3597.
[13] Alexander Brauckmann, Andrés Goens, Sebastian Ertel, and Jeronimo Castrillon. 2020. Compiler-Based Graph Representations for Deep Learning Models of Code. In *Proceedings of the 29th International Conference on Compiler Construction (CC 2020)*.
[14] Tianqi Chen, Thierry Moreau, Ziheng Jiang, Lianmin Zheng, Eddie Yan, Haichen Shen, Meghan Cowan, Leyuan Wang, Yuwei Hu, Luis Ceze, et al. 2018. {TVM}: An automated end-to-end optimizing compiler for deep learning. In *13th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 18)*. 578–594.
[15] Tianqi Chen, Lianmin Zheng, Eddie Yan, Ziheng Jiang, Thierry Moreau, Luis Ceze, Carlos Guestrin, and Arvind Krishnamurthy. 2018. Learning to optimize tensor programs. In *Advances in Neural Information Processing Systems*. 3389–3400.
[16] Xinyun Chen, Chang Liu, and Dawn Song. 2018. Tree-to-tree neural networks for program translation. In *Advances in neural information processing systems*. 2547–2557.
[17] Junyoung Chung, Caglar Gulcehre, Kyunghyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. In *NIPS 2014 Workshop on Deep Learning*.
[18] Chris Cummins, Zacharias V Fisches, Tal Ben-Nun, Torsten Hoefler, and Hugh Leather. 2020. ProGraML: Graph-based Deep Learning for Program Optimization and Analysis. *arXiv preprint arXiv:2003.10536* (2020).
[19] Chris Cummins, Pavlos Petoumenos, Alastair Murray, and Hugh Leather. 2018. Compiler fuzzing through deep learning. In *Proceedings of the 27th ACM SIGSOFT International Symposium on Software Testing and Analysis*. 95–105.
[20] Chris Cummins, Pavlos Petoumenos, Zheng Wang, and Hugh Leather. 2017. End-to-end deep learning of optimization heuristics. In *26th International Conference on Parallel Architectures and Compilation Techniques (PACT)*.
[21] Chris Cummins, Pavlos Petoumenos, Zheng Wang, and Hugh Leather. 2017. Synthesizing benchmarks for predictive modeling. In *2017 IEEE/ACM International Symposium on Code Generation and Optimization (CGO)*. 86–99.
[22] Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. 2016. Convolutional neural networks on graphs with fast localized spectral filtering. In *Advances in neural information processing systems*. 3844–3852.
[23] Christina Delimitrou and Christos Kozyrakis. 2014. Quasar: resource-efficient and QoS-aware cluster management. *ACM SIGPLAN Notices* 49, 4 (2014), 127–144.
[24] Yufei Ding, Jason Ansel, Kalyan Veeramachaneni, Xipeng Shen, Una-May O'Reilly, and Saman Amarasinghe. 2015. Autotuning algorithmic choice for input sensitivity. In *Proceedings of the 36th ACM SIGPLAN Conference on Programming Language Design and Implementation*. 379–390.
[25] Murali Krishna Emani, Zheng Wang, and Michael FP O'Boyle. 2013. Smart, adaptive mapping of parallelism in the presence of external workload. In *Proceedings of the 2013 IEEE/ACM International Symposium on Code Generation and Optimization (CGO)*. IEEE, 1–10.
[26] Common Weakness Enumeration. 2019. 2019 CWE Top 25 Most Dangerous Software Errors. https://cwe.mitre.org/top25/archive/2019/2019_cwe_top25.html. (2019).
[27] Wolfgang Ertel. 1994. On the definition of speedup. In *Proceedings of the International Conference on Parallel Architectures and Languages Europe*. Springer, 289–300.
[28] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 580–587.
[29] S. Grauer-Gray, L. Xu, R. Searles, S. Ayalasomayajula, and J. Cavazos. 2012. Auto-tuning a High-Level Language Targeted to GPU Codes. In *InPar*. https://doi.org/10.1109/InPar.2012.6339595
[30] Dominik Grewe, Zheng Wang, and Michael FP O'Boyle. 2013. Portable mapping of data parallel programs to OpenCL for heterogeneous systems. In *Proceedings of the 2013 IEEE/ACM International Symposium on Code Generation and Optimization (CGO)*. 1–10.
[31] Tobias Grosser et al. 2012. Polly - Performing Polyhedral Optimizations on a Low-Level Intermediate Representation. *Parallel Processing Letters* (2012).
[32] Matthew R Guthaus, Jeffrey S Ringenberg, Dan Ernst, Todd M Austin, Trevor Mudge, and Richard B Brown. 2001. MiBench: A free, commercially representative embedded benchmark suite. In *Proceedings of the fourth annual IEEE international workshop on workload characterization. WWC-4 (Cat. No. 01EX538)*. IEEE, 3–14.
[33] Ameer Haj-Ali, Nesreen K Ahmed, Ted Willke, Yakun Sophia Shao, Krste Asanovic, and Ion Stoica. 2020. NeuroVectorizer: end-to-end vectorization with

deep reinforcement learning. In *Proceedings of the 18th ACM/IEEE International Symposium on Code Generation and Optimization*. 242–255.

[34] Will Hamilton, Zhitao Ying, and Jure Leskovec. 2017. Inductive representation learning on large graphs. In *Advances in neural information processing systems*. 1024–1034.

[35] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *In Proceedings of the Neural computation* 9, 8 (1997), 1735–1780.

[36] Xing Hu, Ge Li, Xin Xia, David Lo, and Zhi Jin. 2018. Deep code comment generation. In *Proceedings of the 26th Conference on Program Comprehension*. 200–210.

[37] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[38] Hugh Leather, Edwin Bonilla, and Michael O'Boyle. 2009. Automatic feature generation for machine learning based optimizing compilation. In *2009 International Symposium on Code Generation and Optimization*. 81–91.

[39] Yujia Li, Daniel Tarlow, Marc Brockschmidt, and Richard Zemel. 2015. Gated graph sequence neural networks. *arXiv preprint arXiv:1511.05493* (2015).

[40] Guanjun Lin, Jun Zhang, Wei Luo, Lei Pan, Olivier De Vel, Paul Montague, and Yang Xiang. 2019. Software Vulnerability Discovery via Learning Multi-domain Knowledge Bases. *IEEE Transactions on Dependable and Secure Computing* (2019).

[41] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, Nov (2008), 2579–2605.

[42] Alberto Magni, Christophe Dubach, and Michael O'Boyle. 2014. Automatic optimization of thread-coarsening for graphics processors. In *Proceedings of the 23rd international conference on Parallel architectures and compilation techniques*. 455–466.

[43] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. 3111–3119.

[44] Mircea Namolaru, Albert Cohen, Grigori Fursin, Ayal Zaks, and Ari Freund. 2010. Practical aggregation of semantical program properties for machine learning based optimization. In *Proceedings of the 2010 international conference on Compilers, architectures and synthesis for embedded systems*. 197–206.

[45] Dorit Nuzman, Ira Rosen, and Ayal Zaks. 2006. Auto-vectorization of interleaved data for SIMD. *ACM SIGPLAN Notices* 41, 6 (2006), 132–143.

[46] William F Ogilvie, Pavlos Petoumenos, Zheng Wang, and Hugh Leather. 2017. Minimizing the cost of iterative compilation with active learning. In *2017 IEEE/ACM International Symposium on Code Generation and Optimization (CGO)*. IEEE, 245–256.

[47] Eunjung Park, John Cavazos, and Marco A Alvarez. 2012. Using graph-based program characterization for predictive modeling. In *Proceedings of the Tenth International Symposium on Code Generation and Optimization*. 196–206.

[48] Jie Ren, Ling Gao, Hai Wang, and Zheng Wang. 2017. Optimise web browsing on heterogeneous mobile platforms: a machine learning based approach. In *IEEE INFOCOM 2017-IEEE Conference on Computer Communications*. IEEE, 1–9.

[49] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 1135–1144.

[50] Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne Van Den Berg, Ivan Titov, and Max Welling. 2018. Modeling relational data with graph convolutional networks. In *European Semantic Web Conference*.

[51] Georgios Tournavitis, Zheng Wang, Björn Franke, and Michael FP O'Boyle. 2009. Towards a holistic approach to auto-parallelization: integrating profile-driven parallelism detection and machine-learning based mapping. In *Proceedings of the 30th ACM SIGPLAN Conference on Programming Language Design and Implementation*. 177–187.

[52] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. 2017. Graph attention networks. *arXiv preprint arXiv:1710.10903* (2017).

[53] Zheng Wang, Dominik Grewe, and Michael FP O'boyle. 2014. Automatic and portable mapping of data parallel programs to opencl for gpu-based heterogeneous systems. *ACM Transactions on Architecture and Code Optimization (TACO)* 11, 4 (2014), 1–26.

[54] Zheng Wang and Michael O'Boyle. 2018. Machine learning in compiler optimization. *Proc. IEEE* 106, 11 (2018), 1879–1901.

[55] Zheng Wang and Michael FP O'Boyle. 2009. Mapping parallelism to multi-cores: a machine learning based approach. In *Proceedings of the 14th ACM SIGPLAN symposium on Principles and practice of parallel programming*. 75–84.

[56] Zheng Wang and Michael FP O'Boyle. 2010. Partitioning streaming parallelism for multi-cores: a machine learning based approach. In *Proceedings of the 19th international conference on Parallel architectures and compilation techniques*. 307–318.

[57] Yuting Wu, Xiao Liu, Yansong Feng, Zheng Wang, Rui Yan, and Dongyan Zhao. 2019. Relation-aware entity alignment for heterogeneous knowledge graphs. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*. 5278–5284.

[58] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. 2018. How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826* (2018).

[59] Xiaojun Xu, Chang Liu, Qian Feng, Heng Yin, Le Song, and Dawn Song. 2017. 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*. 363–376.

[60] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. 2014. How transferable are features in deep neural networks?. In *Advances in neural information processing systems*. 3320–3328.

[61] Yue Yu, Jie Chen, Tian Gao, and Mo Yu. 2019. DAG-GNN: DAG Structure Learning with Graph Neural Networks. In *Proceedings of the 36th International Conference on Machine Learning*.

[62] Peng Zhang, Jianbin Fang, Tao Tang, Canqun Yang, and Zheng Wang. 2018. Auto-tuning streamed applications on intel xeon phi. In *2018 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*. IEEE, 515–525.

[63] Peng Zhang, Jianbin Fang, Canqun Yang, Chun Huang, Tao Tang, and Zheng Wang. 2020. Optimizing streaming parallelism on heterogeneous many-core architectures. *IEEE Transactions on Parallel and Distributed Systems* 31, 8 (2020), 1878–1896.

[64] Zhilu Zhang and Mert Sabuncu. 2018. Generalized cross entropy loss for training deep neural networks with noisy labels. In *Proceedings of the Advances in neural information processing systems*. 8778–8788.

[65] Yue Zhao, Jiajia Li, Chunhua Liao, and Xipeng Shen. 2018. Bridging the gap between deep learning and sparse matrix format selection. In *Proceedings of the 23rd ACM SIGPLAN symposium on principles and practice of parallel programming*. 94–108.

[66] Deqing Zou, Sujuan Wang, Shouhuai Xu, Zhen Li, and Hai Jin. 2019. $\mu$VulDeePecker: A Deep Learning-Based System for Multiclass Vulnerability Detection. *IEEE Transactions on Dependable and Secure Computing* (2019).