WACO: Learning workload-aware co-optimization of the format and schedule of a sparse tensor program

by

Jaeyeon Won

B.S., Seoul National University (2020)

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Master of Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2023

Authored by:	Jaeyeon Won Department of Electrical Engineering and Computer Science August 31, 2023
Certified by:	Joel S. Emer Professor of Electrical Engineering and Computer Science Thesis Supervisor
Certified by:	Saman Amarasinghe Professor of Electrical Engineering and Computer Science Thesis Supervisor
Accepted by:	Leslie A. Kolodziejski Professor of Electrical Engineering and Computer Science Chair, Department Committee on Graduate Students

WACO: Learning workload-aware co-optimization of the format and schedule of a sparse tensor program

by

Jaeyeon Won

Submitted to the Department of Electrical Engineering and Computer Science
on August 31, 2023, in partial fulfillment of the
requirements for the degree of
Master of Science

Abstract

Leveraging the existence of the large number of zeros in sparse tensors offer a powerful way to solve complex problems efficiently in many applications. However, optimizing the performance of those applications poses a challenge. Sparse tensor programs must find the ideal balance between data format and implementation strategy to achieve optimal performance.

This thesis presents WACO, a novel method of co-optimizing the format and schedule of a given sparsity pattern in a sparse tensor program. A core challenge in this thesis is the design of a lightweight cost model that accurately predicts the runtime of a sparse tensor program by considering the sparsity pattern, the format, and the schedule. The key idea in addressing this is exploiting a sparse convolutional network to learn meaningful features of the sparsity pattern and embedding a coupled behavior between the format and the schedule using a specially designed schedule template. In addition, within the enormous search space of co-optimization, our novel search strategy, an approximate nearest neighbor search, efficiently and accurately retrieves the best format and schedule for a given sparsity pattern.

We evaluate WACO for four different algorithms (SpMV, SpMM, SDDMM, and MTTKRP) on a CPU using 726 different sparsity patterns. Our experimental results shows that WACO outperformed four state-of-the-art baselines, Intel MKL, Format-only auto-tuner, TACO with a default schedule, and ASpT. Compared to the best of four baselines, WACO achieved 1.43×, 1.18×, 1.14×, and 1.27× average speedups on SpMV, SpMM, SDDMM, and MTTKRP, respectively.

Thesis Supervisor: Joel S. Emer
Title: Professor of Electrical Engineering and Computer Science

Thesis Supervisor: Saman Amarasinghe
Title: Professor of Electrical Engineering and Computer Science

Acknowledgments

This thesis encompasses the paper [50], co-authored with Charith Mendis, Joel Emer, and Saman Amarasinghe.

I would like express my gratitude to my two advisors, Professor Joel Emer and Professor Saman Amarasinghe. Joel's patient guidance helped me grasp a systematic approach to understanding sparse tensor programs using fibertrees, which form a crucial theoretical foundation for this thesis. Throughout our discussions, Saman consistently encouraged me to think on a broader scale, preventing me from settling for a locally optimal result. Their guidance sparked the exploration of machine learning's potential in sparse tensor compilers. A special note of appreciation is reserved for Charith Mendis for his technical insights into machine learning techniques and auto-scheduling.

I would also like to thank to the members of the COMMIT group; the virtual events we held during the pandemic in my first year significantly uplifted my spirits.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Sparse tensor algebra is an indispensable tool in many domains, such as graph analytics [23], scientific computing [2], and deep learning [20]. Unlike in dense tensor algebra, where only the shape of the tensor matters, the performance of sparse tensor algebra depends heavily on the often complex sparsity pattern of the tensor. Over the last several decades, many sparse formats have been proposed, but none of them was universally optimal across all sparsity patterns. Even with the same format, a different schedule that transforms the traversal order of the iteration space can lead to significant performance changes depending on the sparsity pattern. For example, a sparse matrix with a skewed distribution of non-zeros must exploit fine-grained load balancing, whereas coarse-grained load-balancing must be applied to a sparse matrix with uniformly distributed non-zeros.

Recently, tensor compilers like Halide [41], TVM [9], Tiramisu [5], and TACO [25] have empowered developers to readily write a wide range of tensor computations while incorporating diverse optimizations. Notably, TACO [25] serves as a compiler tailored for sparse tensor algebra, which generalizes many proposed sparse formats by introducing a format abstraction [12]. In addition, a sparse iteration space transformation framework was implemented on top of TACO [43]. This framework allows the compiler to generate code with schedules that perform loop splitting, reordering, parallelizing, and other tasks to explore different traversal orders of iteration space. Although prior studies had built the *mechanism* of the compiler that enables the code

generation supporting many different formats and schedules, the *policy* of the compiler that decides the best format and the best schedule for a given sparsity pattern, has not yet been designed. Unfortunately, a single format or fixed implementation cannot be globally optimal for all sparsity patterns. Thus, designing this policy is closely related to the program auto-tuning problem.

Program auto-tuning has been heavily used to optimize dense tensor programs the performance of which depends on the input size. It started with traditional high-performance scientific libraries such as ATLAS [49] and FFTW [15]. They self-optimize their important routines by empirically transforming the program for the given input shape. Recently, languages such as Halide [41], Tiramisu [5], and TVM [9] decouple algorithms from schedule primitives to transform the structure of the loop in dense tensor programs. Such scheduling languages allow the expression of a broader range of algorithms (compared to the limited BLAS routines in ATLAS) and the introduction of a huge search space due to schedules.

Auto-tuning sparse computation is not new [30, 34, 36, 44, 48]; even production systems are introducing auto-tuning workflows for sparse computations. For example, Intel MKL uses an inspector-executor model to auto-tune a few popular sparse computations [36]. However, the current production as well as the state-of-the-art research systems have the following limitations.

Limitations in Capturing Sparsity Pattern. For a dense tensor program, an auto-tuner only needs the tensor's shape. However, a shape alone fails to capture the sophisticated sparsity pattern in a sparse tensor program. To summarize the sparsity pattern, much more information is required, such as the density, the size of dense blocks, and the existence of symmetry. Capturing the sparsity pattern with the entire sparse matrix is costly because the number of non-zeros can reach billions. Thus, designing features that accurately summarize the sparsity pattern is critical for optimal decision-making in auto-tuning. Existing approaches fall short of fully capturing the pattern because they rely either on manually crafted features [28, 42] or a convolutional neural network with a downsampled matrix [44, 51], both of which result in significant information loss of the sparsity pattern.

Absence of Co-optimization. The joint optimization of the data layout and the schedule is critical even in dense tensor programs, which are simpler than sparse tensor programs [22]. Nevertheless, prior auto-tuning studies on sparse tensor programs mainly tackled only one of two problems: choosing the best schedule or the best format. For instance, Intel MKL supports the inspector-executor sparse BLAS routines [36] that the executor calls the routine tuned by an inspector. However, MKL inspector misses optimization opportunities because it limits the tuning space by fixing the format. It is necessary to consider a coupled behavior between the format and the schedule to get the good performance.

Our Approach. This thesis presents the **W**orkload-**A**ware **C**o-**O**ptimization (WACO), a framework for automatically and jointly optimizing the format and the schedule of a given sparsity pattern. WACO uses a deep-learning based cost model that accurately and efficiently predicts the performance of the sparse tensor program. The cost model uses a novel sparse convolutional network, *WACONet*, to extract rich features of a sparsity pattern and uses a unified schedule template, *SuperSchedule*, to understand both the format and the transformed iteration space. WACO further utilizes an approximate nearest neighbor search to quickly search for the optimal format and schedule over the huge search space.

Overall, our main contributions are as follows:

- To the best of our knowledge, WACO is the first auto-tuner that co-optimizes the format and the schedule in a workload-aware manner for a sparse tensor program.

- WACO is the first autotuner with a cost model that considers the coupled behavior of the sparsity pattern, the format, and the schedule.

- WACO introduces a sparsity pattern feature extractor *WACONet*, a novel sparse convolutional network architecture to effectively learn meaningful features from both coarse-grained and fine-grained sparisty patterns.

- WACO uses an extremely fast search strategy, an Approximate Nearest Neighbor

Search (ANNS), to retrieve the near-optimal format and schedule.

- We compared WACO against four state-of-the-art baselines, Intel MKL, BestFormat, TACO with a fixed format and schedule, and ASpT. WACO outperformed the best of four baselines by achieving 1.43×, 1.18×, 1.14×, and 1.27× average speedups on SpMV, SpMM, SDDMM, and MTTKRP, respectively.

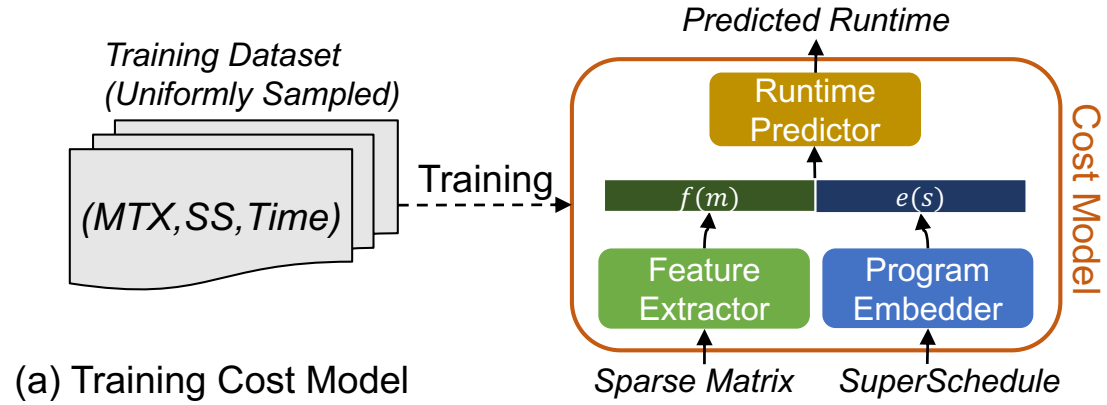
1.1 Overview of WACO

Figure 1-1 shows an overview of WACO. We designed our cost model to predict the runtime of the program. The cost model takes a sparse matrix and a SuperSchedule, a unified template that defines the format and the schedule together, as inputs (Figure 1-1-(a), details in Section 4.1).

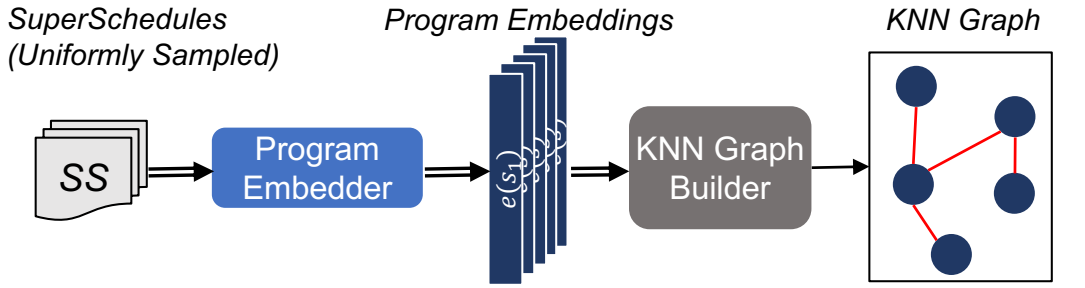
After training the cost model, WACO builds a K-Nearest Neighbor(KNN) graph that helps with the search later. The KNN graph is built on program embeddings of uniformly sampled SuperSchedules (Figure 1-1-(b), details in Section 4.2).

Finally, when the input matrix comes in, WACO uses a novel search strategy, an approximate nearest neighbor search (ANNS), to search for the optimal format and schedule for a given input sparse matrix. ANNS repeats the picking of the next candidate SuperSchedule using a KNN graph and receives the candidate's predicted runtime as feedback until it converges to a locally optimal SuperSchedule (Figure 1-1-(c), details in Section 4.2).

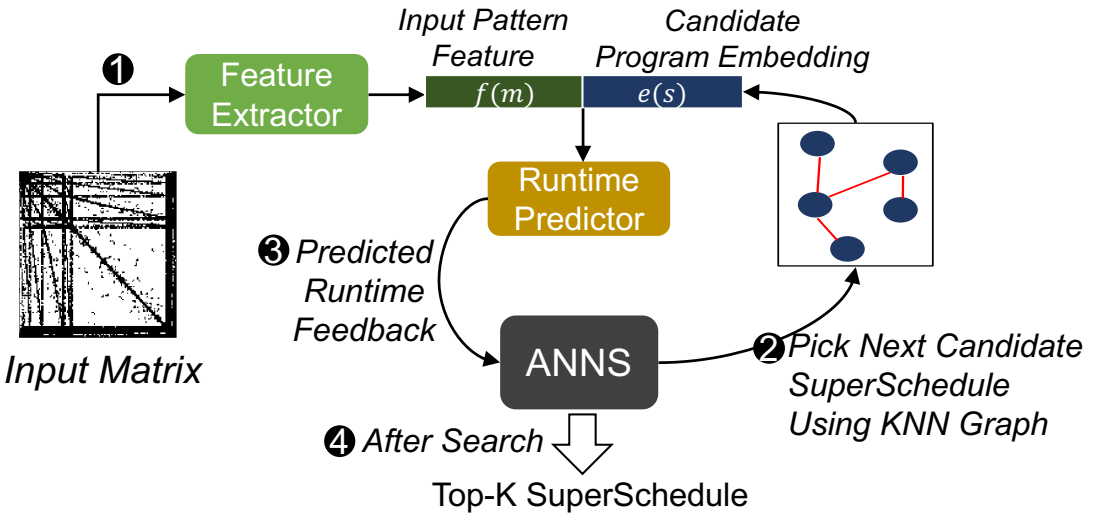
The remaining chapters of the thesis are structured as follows. Chapter 2 emphasizes the significance of co-optimization depending on the sparsity pattern with motivating examples. Chapter 3 provides an overview of the sparse tensor compiler and cost model. Chapter 4 explains our methodology, WACO. Chapter 5 discusses the experimental results and analysis of WACO. Chapter 6 discusses related works. Finally, in Chapter 7, the thesis concludes by addressing current limitations and suggesting potential directions for future research.



(a) Training Cost Model



(b) Building KNN Graph



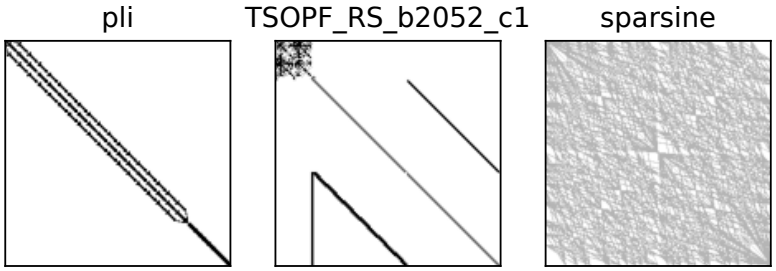
(c) Searching Optimal Format and Schedule

Figure 1-1: Overview of WACO. In the training dataset in (a), *(MTX, SS, TIME)* is the abbreviation for *(Sparse Matrix, SuperSchedule, Ground Truth Runtime)*. SuperSchedule defines the format and the schedule together.

Chapter 2

Motivating Example

In this chapter, we will describe how the co-optimization can impact the performance of a sparse tensor program. In addition, we will show that the performance of a sparse tensor program strongly depends on the sparsity pattern. This demonstrates the strong need for an auto-tuning framework for sparse tensor programs.



Name	Rows	Cols	Nonzeros	Density
pli	22,695	22,695	1.35M	0.26%
TSOPF	25,626	25,626	6.76M	1.03%
sparsine	50,000	50,000	1.55M	0.06%

Figure 2-1: Sparse matrices used for the motivation.

2.1 Impact of the Co-optimization

Table 2.1 shows the impact of co-optimization in a sparse tensor program by comparing results of auto-tuning on three different tuning spaces: the format, the schedule, and both the format and the schedule. We ran a sparse matrix - dense matrix multiplication

Table 2.1: SpMM speedup over the base implementation after auto-tuning. The three rightmost columns represents different tuning spaces of a sparse tensor program. Co-optimization tunes both the format and the schedule.

Name	Base	Format-only	Schedule-only	Co-optimization
pli	1×	1.03×	1.03×	**1.21**×
TSOPF	1×	1.11×	1.12×	**2.02**×
sparsine	1×	2.4×	1.02×	**2.5**×

(SpMM) with different tuning spaces. For the baseline, we used CSR, one of the most popular sparse matrix formats, with the default schedule generated by TACO. For the Format-only, we only tuned the format while keeping the iteration order identical to the baseline, except that we made the traversing order to be concordant [45] with how the tuned format is aligned. For the Schedule-only, we only tuned the schedule to transform the iteration order while keeping the format identical to the baseline (CSR). For the Co-optimization, we co-optimized both the format and the schedule.

Jointly optimizing the format and the schedule yields the most significant speedup for all the matrices in Figure 2-1, in contrast to the restricted search space. Restricting the tuning space to either choose the optimal format or the optimal schedule can miss optimization opportunities. Especially for TSOPF, co-optimization boosts the performance (2.02×), whereas considering only the format or the schedule yields a slight performance improvement (∼1.1×).

2.2 Sparsity Pattern-Dependent Nature

The performance of sparse tensor programs is very sensitive to the sparsity pattern of the input matrix. No single format or implementation can show the optimal performance for all sparsity patterns, even for highly optimized handwritten libraries of experts. Table 2.2 demonstrates this nature. We ran a SpMM with the format and schedule optimized for different sparse matrices. As expected, the diagonal of the table shows the best performance because it is a result of the co-optimization that corresponds to the input matrix. A significant performance drop often occurs when other optimizations are applied.

Table 2.2: SpMM speedup over the base implementation for different optimization methods. opt-X indicates the format and the schedule that are optimized for matrix X (as a result of Co-optimization in Table 2.1).

Name	opt-pli	opt-TSOPF	opt-sparsine
pli	**1.21**×	0.82×	0.98×
TSOPF	1.14×	**2.02**×	0.96×
sparsine	0.81×	0.37×	**2.5**×

These examples strongly indicate the need to co-optimize the format and schedules according to the input sparsity pattern. From the perspective of auto-tuning, three challenges stand out compared to dense applications. ❶ While considering the sparsity pattern, our framework should automatically decide ❷ which format to store the tensor in and ❸ which schedules should be applied to transform the iteration order. To address these challenges, the auto-tuner should understand the complex interactions among the sparsity pattern, the format, and the schedule.

Chapter 3

Background

In this chapter, we describe how TACO generates codes that support various formats and iteration space transformations. Then, we describe existing sparsity pattern-aware cost models for sparse tensor program auto-tuning.

3.1 Tensor Algebra Compiler

TACO is a sparse tensor algebra Domain Specific Language (DSL) with an accompanying compiler that decouples the algorithm from the data representation and schedule [25, 12, 43]. Its algorithm is specified by an Einsum notation, for example, `C[i,j] = A[i,k] * B[k,j]` represents a matrix multiplication. Chou et al. introduced a format abstraction that describes how a sparse tensor is stored in different formats with coordinate hierarchies and level formats [12]. A sparse tensor can be viewed as a hierarchy of coordinates where each level is stored in one of the *level formats*. Chou et al. presented six level formats to represent various formats, but we will mainly focus on two level formats, *Uncompressed* and *Compressed*.

Format Abstraction. TACO uses a coordinate tree abstraction to describe how to store the tensor, which was first described in the format abstraction [12] in TACO and later abstracted further and formalized as the *fibertree* abstraction [45].

Figure 3-1 shows how the fibertree abstraction represents a matrix. Any tensor is

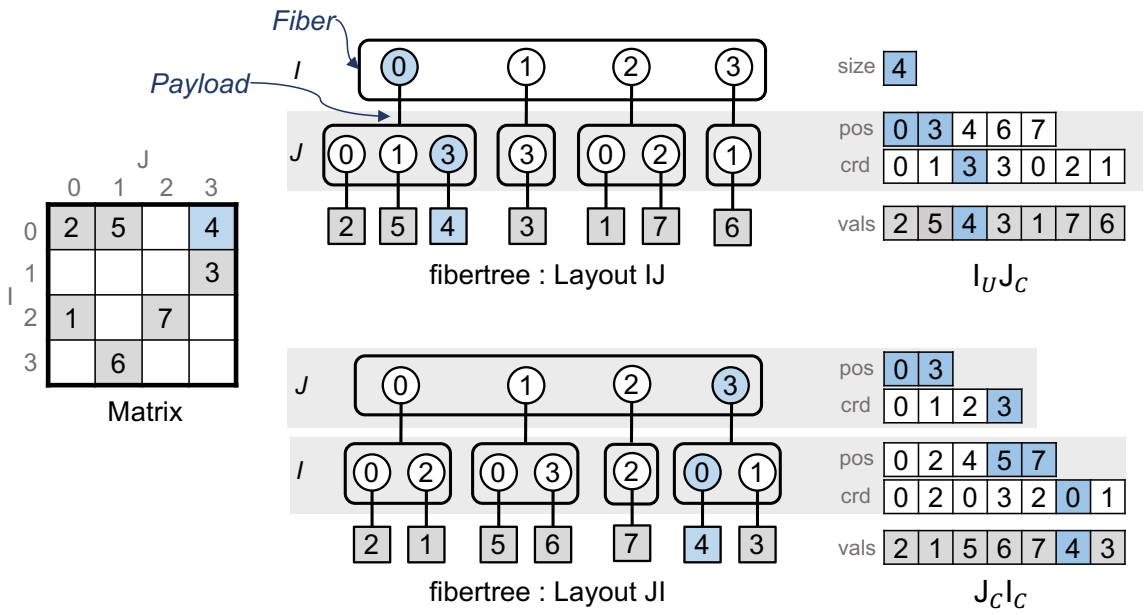


Figure 3-1: A sparse matrix with dimensions I and J, its fibertree abstraction (in both level orders) and concrete representations for two data layouts, $I_U J_C$ and $J_C I_C$. U and C mean Uncompressed and Compressed level format, respectively. The shaded level indicates the Compressed level format.

first viewed as a tree with each fiber carrying a set of coordinates and a payload that is either a fiber at the next level or a value at the bottom of the tree. The order of the levels indicates the data layout such as row-major or column-major. Once the layout is specified, a *level format* specifies what physical storage is used to store the fiber. An Uncompressed (U) level format encodes a dense coordinate interval $[0, N)$. A Compressed (C) level format encodes only non-zero coordinates in the fiber by explicitly storing coordinates. A format language $I_U J_C$ in the Figure 3-1 says the matrix is stored in the $I \rightarrow J$ layout (row-major), and level formats where I and J are Uncompressed and Compressed, respectively. We refer to a *position* as the index of an element in the concrete data representation. For example, the color-highlighted coordinate=3 in level J of the $I_U J_C$ has a position=2 in the crd array (assuming zero-based indexing).

A combination of level splitting, fusing, reordering, and level format choice can express many representations. For example, if we split two levels I and J into $I1$, $I0$, $J1$, and $J0$, it can have a total of $4! * 2^4$ representations, where 4! indicates the

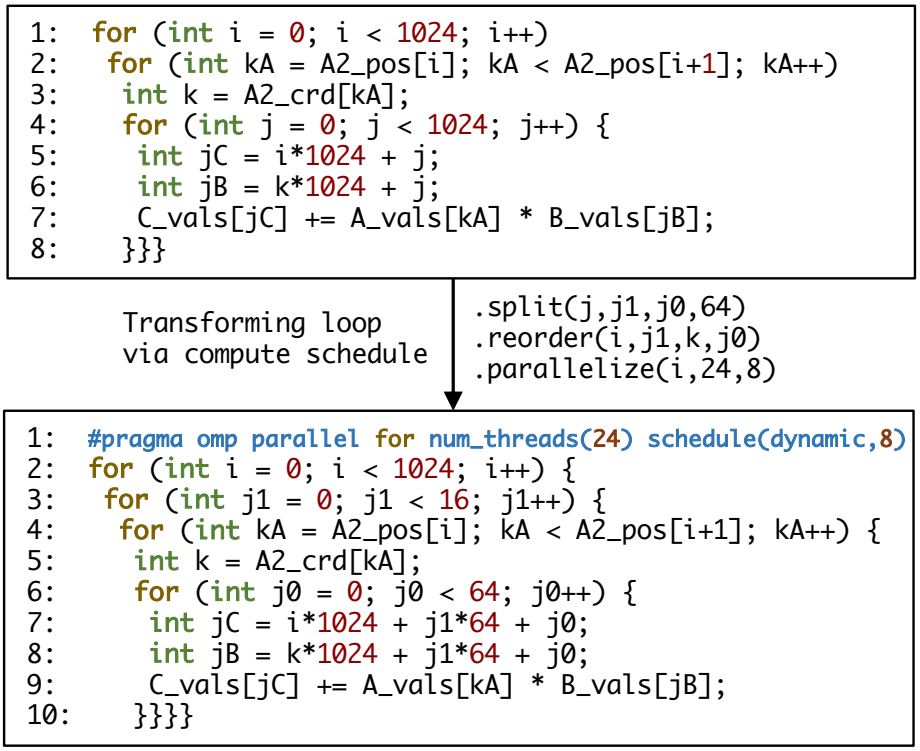


Figure 3-2: Loop transformation of C[i,j]=A[i,k]*B[k,j] by schedules. It changes the traversal order of the iteration space. A[i,k] is stored in the UC format.

number of possible level orders and 2^4 indicates the number of level format choices. More formats can be formulated depending on the number of levels in the hierarchy.

Iteraion Space Transformation via Schedules. In addition to format abstraction, schedules decide how to traverse the tensor stored in a particular format by transforming the iteration space. For example, as shown in Figure 3-2, the `split` schedule splits a specified loop level into two nested levels, `reorder` specifies the order of the nested loops, and `parallelize` controls the load-balancing across multiple threads. A good choice of transformations enables parallelism and/or better data locality (e.g., register/cache blocking). In sparse computation, however, such loop transformation must be chosen deliberately while considering the format. For instance, if a loop order is discordant [45] with how the format is ordered, its generated code may involve an inefficient traversal routine such as a binary search over the Compressed level format. Thus, an auto-tuner must understand the coupled behavior between the format and the traversal order of the iteration space.

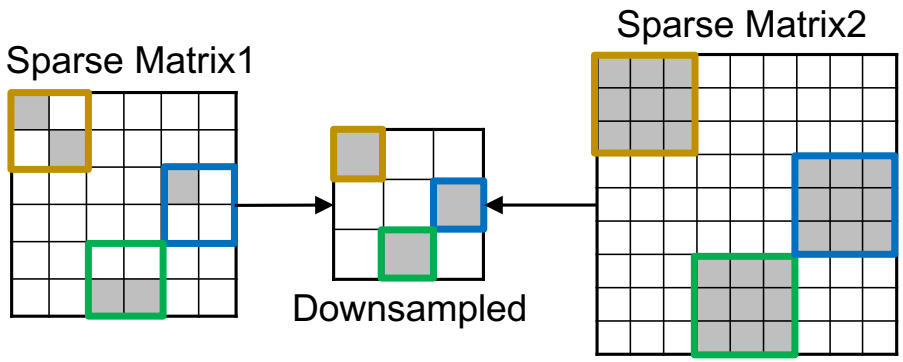


Figure 3-3: Different sparse tensors downsampled into the same 3×3 tensor.

3.2 Cost Model for Auto-Scheduling

In scheduling languages [41], auto-scheduling is the task that finds the best schedule for a given input [10, 37, 54]. Auto-scheduling mainly has two parts. The first part is a cost model that quickly predicts the performance of the program, and the second is a search strategy that finds the best schedule according to the cost model. Although the actual hardware measurement can be used as a cost, it is very time-consuming, so designing an efficient and accurate cost model is crucial. For a sparse tensor program, understanding the sparsity pattern is the most critical design consideration of the cost model. Two methods of extracting the features of a sparse tensor have been commonly used: ❶ human-crafted features [28, 42] and ❷ a convolutional network over downsampled tensors [44, 51].

Human-crafted Features. A feature vector is designed manually by considering the statistical properties of tensors [28, 42]. Typical features are the total number of non-zeros, the mean or variance of the number of non-zeros per row, and format-specific features such as the average distance from the diagonal for DIA format. Nevertheless, the usefulness of human-crafted features for determining the accuracy is unknown. The features also have to be manually redesigned whenever a new format to be considered.

Convolutional Neural Network (CNN). Another approach uses a CNN to extract the features by viewing a sparse tensor as an image [51, 44]. A sparse tensor can have many different shapes, but since the CNN is limited to taking a fixed-size shape as

an input, the sparse tensor is downsampled into a fixed shape. Figure 3-3 illustrates how tensor downsampling works for arbitrary sparse tensors. In practice, the sparse tensors are usually downsampled to 128x128. To provide additional information to the CNN, a non-zero location in a downsampled tensor may contain a corresponding number of non-zeros in the original sparse tensors. However, as the shape of the sparse tensor increases, downsampling leads to a significant loss of the information on the local pattern. For example, while the Sparse Matrix2 in Figure 3-3 only has dense blocks, both matrices are downsampled into the same matrix. In addition, there are real-world sparse tensors with shapes in the millions scale, which cannot be helped by downsampling.

The aforementioned methods have deficiencies in accurately extracting the features of a sparsity pattern. In the case of human-crafted features, it is impossible to manually design all the format-specific features in TACO's format abstraction. In the case of downsampling, it only works for small sparse tensors or it will lose significant information, which often leads to sub-optimal decisions.

Chapter 4

Workload-Aware Co-Optimization

In this chapter, we introduce Workload-Aware Co-Optimization (WACO), an auto-tuning framework for sparse tensor programs. WACO automatically searches for the best format and schedule for a given sparse matrix from among what TACO compiler can generate.

First, we will describe how WACO uses a novel cost model that understands a complex interaction of the sparsity pattern, format, and schedule (Figure 1-1-(a)). Then, we will explain how WACO efficiently searches over the large search space using a novel search strategy, ANNS (Figure 1-1-(b,c)).

4.1 Cost Model Design

Our cost model has three parts (Figure 4-1). The first part, the feature extractor, captures the sparsity pattern of the input matrix. The second part, the program embedder, understands the coupled behavior of the format and the schedule. Finally, the runtime predictor predicts the runtime through multiple linear-ReLU layers by concatenating the results of the previous parts.

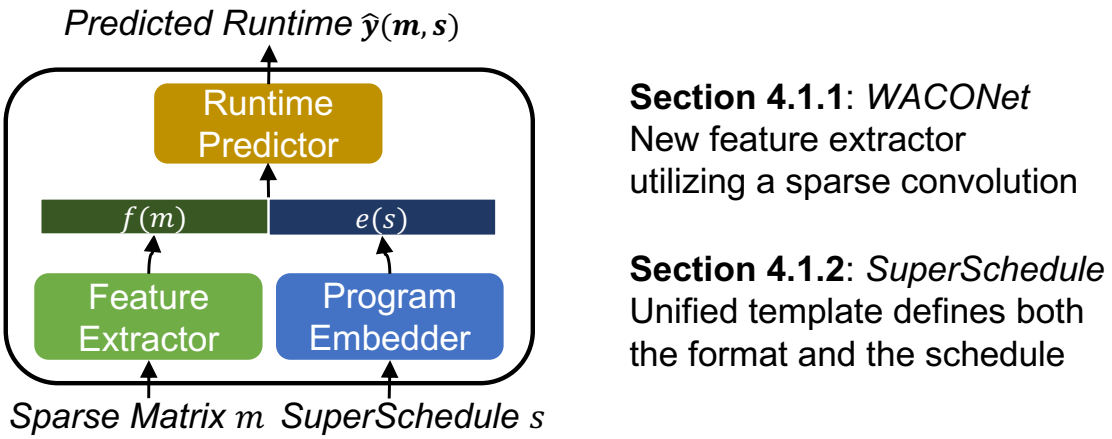
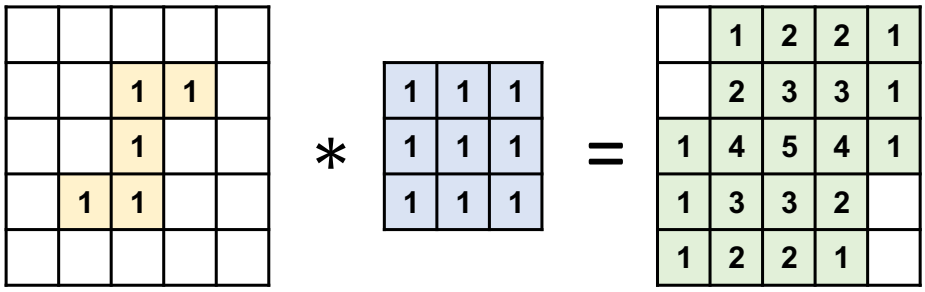


Figure 4-1: Overview of WACO's cost model. It predicts the program's runtime by taking a sparse matrix and SuperSchedule as inputs. SuperSchedule contains information on both the format and the schedule.

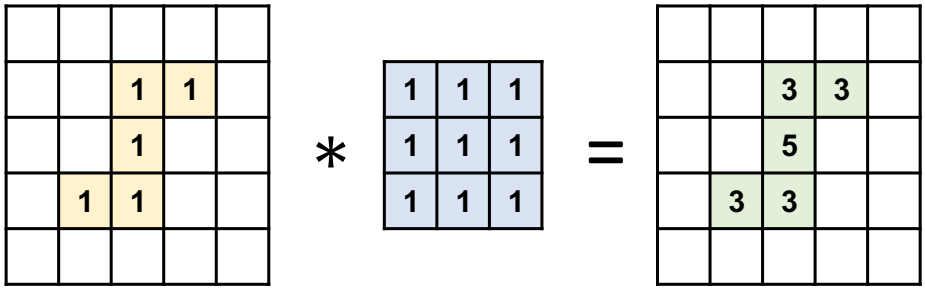
4.1.1 Feature Extractor: WACONet

As described in 3.2, extracting features of a sparsity pattern is non-trivial. The core idea of our approach is to use a *sparse convolutional neural network* to learn good features. We propose a novel feature extractor, *WACONet*, based on a sparse CNN with a novel network architecture. In Section 5.3, our evaluation shows that WACONet improves the training and validation loss by roughly 50% when compared to a conventional CNN feature extractor.

Exploring Different Architectures. An obvious solution is to use a conventional CNN that treated a sparse matrix as a dense matrix, where all levels were stored in the Uncompressed format. However, as the shape of the matrix grew, it ran out of computational resources very quickly. For example, if there is a sparse matrix of shape $10^5 * 10^5$, it will need a total of $4 * 10^{10}$ bytes (assuming 4 bytes single-precision) regardless of the number of non-zeros. Another approach that we tried is using a recurrent neural network by viewing a sparse tensor as a sequence of coordinates. However, since the sequence length (a number of non-zeros) is in the millions scale, the recurrent neural network cannot remember everything and easily forgets the early sequences. It is also difficult to decide in which order to put the coordinate sequences such as the row-major or the column-major. We ended up using CNN for our feature



(a) Conventional Convolution

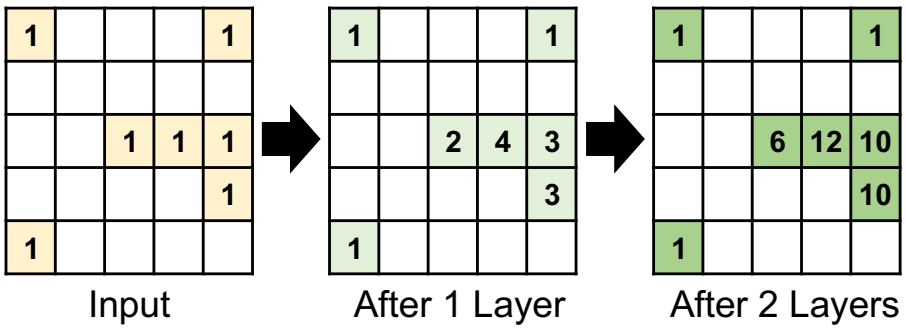


(b) Submanifold Sparse Convolution

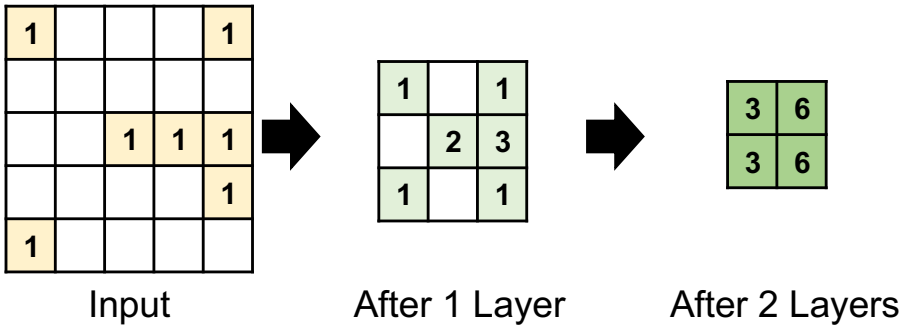
Figure 4-2: Difference between conventional convolution and (submanifold) sparse convolution.

extractor, but instead of a dense convolution with downsampling, we used a sparse convolution on the raw sparse matrix itself.

Sparse Convolutional Layer. A sparse convolutional layer [17] (often called as submanifold sparse convolution) performs a convolution operation over a sparse input. There is a marked difference between the sparse convolution and conventional convolution. While conventional convolution operates over all the input activations, a sparse convolution operates only when the filter's center is located on a non-zero input activation (Figure 4-2). This peculiar behavior prevents the activations from becoming dense as the layers are stacked, thus keeping the computation relatively cheap. However, this behavior also has an issue when the non-zeros are distributed far apart. As shown in Figure 4-3-(a), this behavior can only capture the local pattern but not the global pattern because the non-zeros are not close enough to propagate information. Sparse convolution has shown a powerful ability to understand 3D point clouds when their non-zeros are close enough [13]. However, real-world sparse matrices often have a distant non-zero distribution, so we need to design a network architecture



(a) Distant Non-zeros with Stride 1



(b) Distant Non-zeros with Stride 2

Figure 4-3: When non-zeros are distributed far apart, the receptive field does not increase even with multiple layers if the stride of sparse convolution is 1. Filter's center is located at red circles in the stride of 2 as used in WACONet.

that addresses this issue while utilizing the advantage of sparse convolution.

WACONet. We propose *WACONet*, a novel sparse CNN architecture that learns the rich features of a sparsity pattern effectively (Figure 4-4). Except for the first layer, we used a strided convolution with a filter size of 3x3 for every sparse convolutional layer. Multiple stacks of strided convolution help distant non-zeros because a strided behavior forces the receptive field to increase (Figure 4-3-(b)). Due to the limited memory size of the GPU, the number of channels in the sparse convolutional layer is small (32) to fit a sparse matrix with a large number of non-zeros up to 10 million, unlike in a typical vision CNN model (e.g., 256 and 512). To compensate for decreased network capacity due to a limited number of channels, WACONet concatenates all 14 intermediate results after the global average pooling rather than using a result of the final layer.

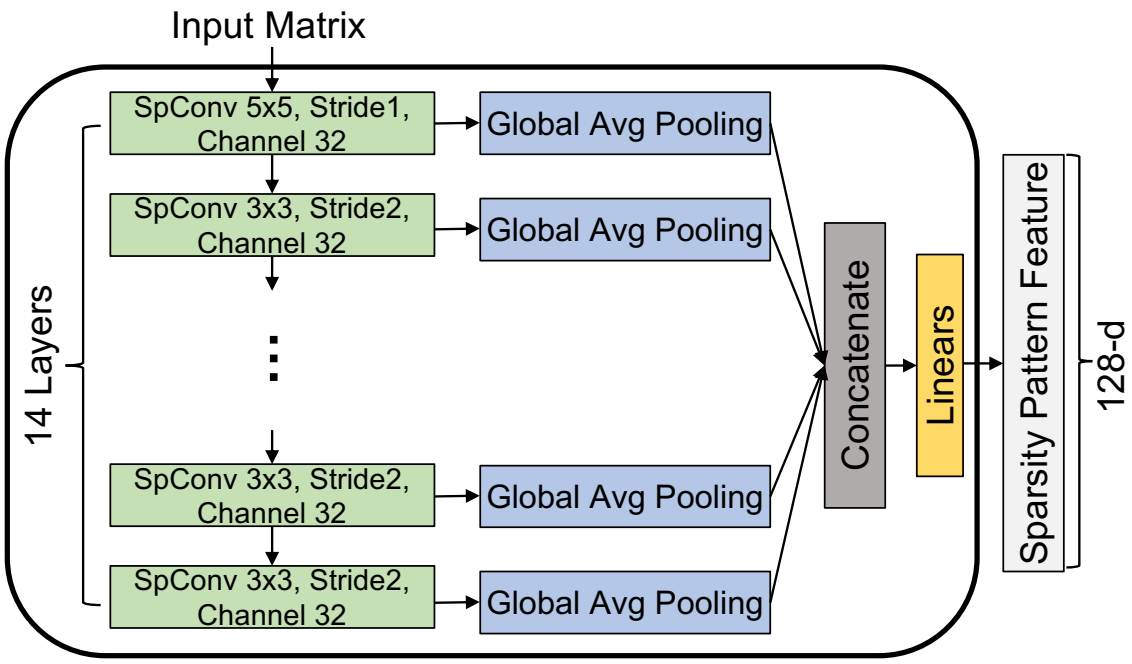
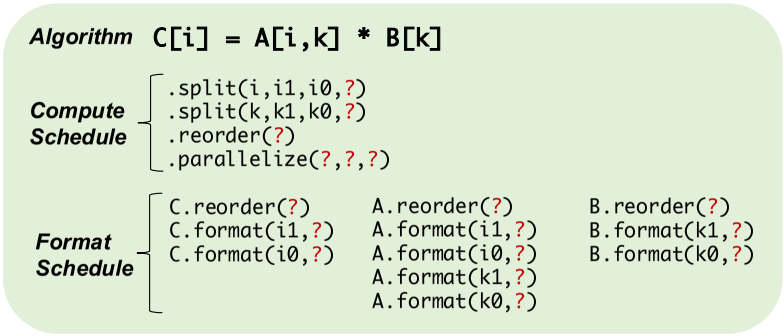
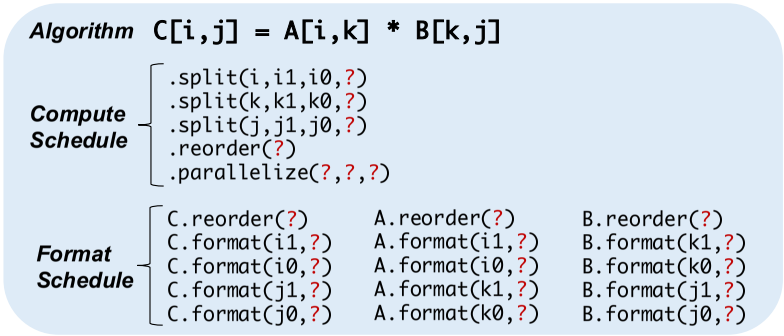


Figure 4-4: Network architecture of WACONet. We omitted non-linear activation layers in the figure.

WACONet minimizes the loss of information of the sparsity pattern because it takes a raw sparse matrix as an input without any downsampling. Due to the nature of the convolution operation, a small filter (3x3) recognizes the local pattern, and a global pattern is captured while passing through multiple strided layers. In addition, WACONet can be easily extended for high-dimensional sparse tensors by simply changing the dimension of the filter. In Section 5, we demonstrate that WACONet extracts the rich features for both 2D and 3D sparse tensors.



(a) Matrix-Vector Multiply(MV) SuperSchedule



(b) Matrix-Matrix Multiply(MM) SuperSchedule

```
.split(i,i1,i0,2)
.split(k,k1,k0,1)
.reorder(i1,k1,i0,k0)
.parallelize(i1,16,8)

C.reorder(i1,i0)      A.reorder(i1,k1,i0,k0)  B.reorder(k1,k0)
C.format(i1,U)        A.format(i1,U)          B.format(k1,U)
C.format(i0,U)        A.format(i0,U)          B.format(k0,U)
                      A.format(k1,C)
                      A.format(k0,U)
```

```
1:  #pragma omp parallel for num_threads(16) schedule(dynamic, 8)
2:  for (int i1 = 0; i1 < A1_dimension; i1++) {
3:   for (int k1A = A2_pos[i1]; k1A < A2_pos[(i1 + 1)]; k1A++) {
4:    int k1 = A2_crd[k1A];
5:    for (int i0 = 0; i0 < 2; i0++) {
6:     int i0C = i1 * 2 + i0;
7:     int i0A = k1A * 2 + i0;
8:     float tk0C_val = 0.0;
9:     for (int k0 = 0; k0 < 1; k0++) {
10:     int k0A = i0A * 1 + k0;
11:     int k0B = k1 * 1 + k0;
12:     tk0C_val += A_vals[k0A] * B_vals[k0B];
13:    }
14:    C_vals[i0C] = C_vals[i0C] + tk0C_val;
15:  }}}
```

(c) Sampled Schedule from MV SuperSchedule and its generated code.

Figure 4-5: (a,b) MV/MM SuperSchedules. (c) The sampled schedule showed how C[i1,i0] = A[i1,k1,i0] * B[k1] with a BCSR format can be sampled from the SuperSchedule by choosing the `k` split size as 1. The shaded lines in the generated code indicate those can be ignored due to the split size 1. The SuperSchedule for SDDMM(D[i,j] = A[i,j] * B[i,k] * C[k,j]) can also be defined similarly.

4.1.2 Program Embedder: SuperSchedule

We will now consider the second part of the cost model, a program embedder. In a dense tensor program, a program embedder only needs to encode a traversal order of the iteration space reflected by the low-level loop abstract syntax tree [4, 10, 37]. In a sparse tensor program, however, a program embedder must encode both the traversal order and the format to accurately understand the coupled behavior for the joint optimization.

Challenges. Encoding a loop order is non-trivial because the number of levels in a nested loop varies due to the schedule `split`. The search space expands as well after splitting, as seen in Figure 3-2, where the number of loop reorderings increased to 4! from 3!. To deal with the variableness due to the `split`, we adopted a template-guided auto-scheduling [10]. In addition, our template specifies both the format and the schedule and creates the program embedding directly on top of that template.

SuperSchedule. The unified schedule template, which we call *SuperSchedule*, defines the format and the schedule at the same time. Figure 4-5-(a) shows how the SuperSchedule template is defined in a matrix-vector multiplication (MV). The SuperSchedule consists of a compute schedule and a format schedule. A compute schedule defines the traversal order of the iteration space and a format schedule that defines how tensors will be stored. While `reorder` in the format schedule determines the level order of the tensor (e.g., the row-major or the column-major), `reorder` in the compute schedule decides the traversal order of the tensors.

One observation is that a schedule template that already has multiple `split`s can be reduced into a schedule that has fewer `split`s. This reduction can be done by specifying the split size as 1. To support this, the compute schedule splits each index(i and k) once, making the MV algorithm (C[i] = A[i,k] * B[k]) a split MV algorithm (C[i1,i0] = A[i1,i0,k1,k0] * B[k1,k0]). Within this SuperSchedule, we can sample all the schedules from

1. C[i1] = A[i1,k1] * B[k1]
2. C[i1,i0] = A[i1,i0,k1] * B[k1]

Table 4.1: MV SuperSchedule parameters. $P()$ indicates a permutation of indices. For `parallelize`, we used the OpenMP work-sharing policy (`#pragma omp parallel for schedule(dynamic, chunksize)`).

Schedule	Parameters	Description
`split`	[1, 2, ..., 32768]	Split Size
`reorder`	$P(i1, i0, k1, k0)$	Loop Order
`parallelize`	$[i1, i0]$	Parallelized Index
	[24, 48]	# Threads
	[1, 2, ..., 256]	OMP Chunksize
`C.reorder`	$P(i1, i0)$	Level Order of C
`A.reorder`	$P(i1, i0, k1, k0)$	Level Order of A
`B.reorder`	$P(k1, k0)$	Level Order of B
`format`	[U, C]	Level Format

3. C[i1] = A[i1,k1,k0] * B[k1,k0]

4. C[i1,i0] = A[i1,i0,k1,k0] * B[k1,k0]

by appropriately choosing the `split` size as 1.

From these split algorithms, SuperSchedule can also derive various formats. For instance, the UC format in Figure 3-1 can be derived by choosing both `split` sizes as 1 and specifying level formats as UC according to the level order of `i1` and `k1`. Similarly, the UCUU format can be derived by choosing both `split` sizes greater than 1 and specifying the level format as UCUU according to the level order.

SuperSchedule is a superset of all possible schedules under a fully split algorithm. For example, the MV and MM SuperSchedule in the Figure 4-5 can represent a total of 4 and 8 split algorithms, respectively, but SuperSchedule can represent more algorithms depending on how many `split`s are defined. We chose a maximum of one `split` per dimension since we have found out that more than one `split` yields diminishing returns.

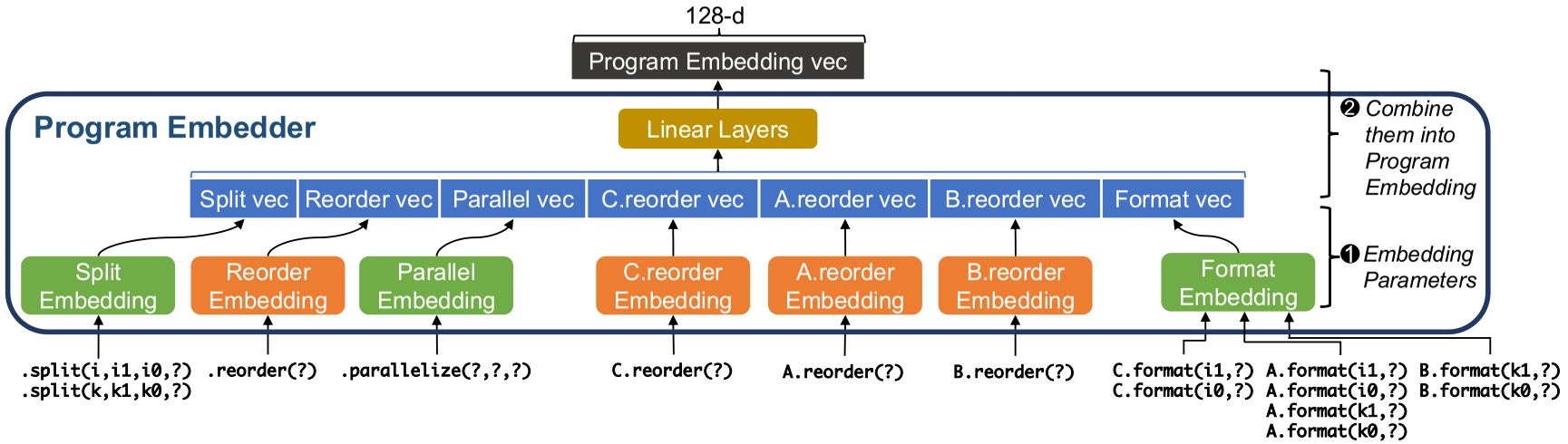


Figure 4-6: Network architecture of the program embedder. The parameters of the SuperSchedule are used as the inputs. The green embedder takes a categorical parameter, while the orange embedder takes a permutational parameter.

Network Architecture. Such a template-based schedule allowed us to embed the program more easily. Rather than extracting each loop's features from the low-level loop abstract syntax tree, SuperSchedule allowed us to embed the format schedule and the compute schedule directly from the parameters of the template. Table 4.1 describes each schedule and its possible parameter choices used in our evaluation. All the parameters are categorical except for the `reorder`, which take a permutation of indices. Our program embedder (Figure 4-6) takes parameters of a SuperSchedule and outputs the program embedding. It first calculates the embeddings of each parameters. Each categorical parameter passes a learnable lookup table (green box) that maps the one-hot categorical parameter to a high-dimensional real-valued vector. Each permutation parameter is converted into a corresponding permutation matrix and passes multiple linear-ReLU layers (orange box). When the embeddings for the all schedule parameters are calculated, they are concatenated and pass multiple linear-ReLU layers into the final program embedding.

4.1.3 Training Cost Model

Our training dataset was a set of tuples *(Sparse Matrix, SuperSchedule, Ground Truth Runtime)*. We designed our dataset to include various sparsity patterns. We augmented the 2,893 real-world sparse matrices in the SuiteSparse matrix collection [14] by arbitrarily resizing them into 21,400 sparse matrices while restricting the number of rows to less than 131,072 and the number of non-zeros to less than 10 million. During the augmention process, we densify each sparse non-zero region using a random block size. For 3D sparse tensor augmentation, we followed a prior work's approach [44] to generate the training dataset for 3D sparse tensors.

For each matrix, we randomly sampled 100 formats and schedules from the SuperSchedule. Then, we generated a corresponding code using TACO for each sample, repeated the program for 50 rounds, and reported the median time. We excluded formats and schedules that take more than a minute. We repeated this process for the four algorithms used in evaluations (SpMV, SpMM, SDDMM, and MTTKRP), and collected about 2 million tuples for each algorithm. Collecting the

dataset took two weeks with 10 computing nodes. During the training, we divided the total dataset into the training dataset and the validation dataset at an 80:20 ratio.

Training Objective. For a given input sparse matrix m_i and SuperSchedule s_j, the goal of our cost model $\hat{y}(m_i, s_j)$ is not to accurately predict the ground truth runtime y_{ij}. Instead, we want our cost model to learn the ranking of different SuperSchedules. Therefore, we used a pairwise ranking loss [8] to reflect the relative order of performance of the schedules instead of using the L1 or L2 loss.

$$L = \sum_{m_i} \sum_{(s_j, s_k)} sign(y_{ij} - y_{ik}) * \phi(\hat{y}(m_i, s_j) - \hat{y}(m_i, s_k))$$

where $sign(x)$ is 1 if $x > 0$, or 0 otherwise. $\phi(x)$ can be defined as various functions, such as the hinge function $max(0, 1 - x)$ or the logistic function $log(1 + e^{-x})$. We adopted the hinge function for our model. We used a SuperSchedule (s_j, s_k) batch size of 32 for each sparse matrix m_i and an Adam optimizer [24] with a learning rate of 0.0001.

4.2 Efficient Schedule Search via Nearest Neighbor Search

Besides the design of the cost model, a search strategy is also an essential component of auto-scheduling. Many auto-schedulers or tuners rely on black-box optimization algorithms to find the best parameters in the schedule template [18, 53, 3].

Challenges. The traditional black-box optimization algorithms are often slow because besides evaluating black-box (cost model), they must manage the metadata required for optimization. For example, Bayesian optimization trains a surrogate model internally that facilitates the procedure during the search. To speed up the search, we cast the auto-scheduling problem as a Nearest Neighbor Search (NNS) [38]. We then exploit an existing high-performance NNS library to search for the optimal parameters of the SuperSchedule. In our experiment, the time spent of evaluating costs in the whole

search was only 3.9% and 8.1% on two famous black-box optimizers, HyperOpt [6] and OpenTuner [3], while our search strategy improved the proportion to 93.9%.

4.2.1 Relationship Between Auto-scheduling and NNS

Here, we will show that auto-scheduling can be reduced into the NNS. The definition of the NNS is as follows :

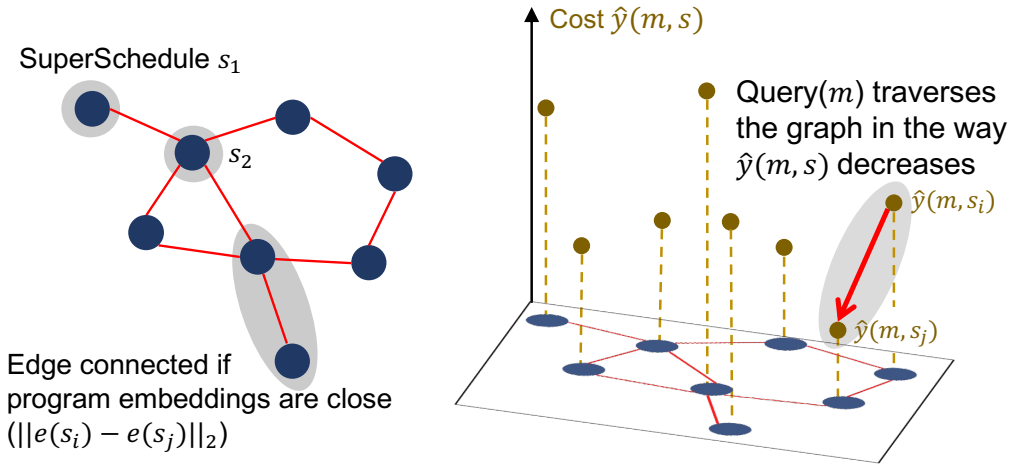
Definition 4.2.1 (Nearest Neighbor Search) *Suppose we have a dataset* $S = \{x_1, x_2, ..., x_n\}$ *where* $x_i \in R^{d_s}$. *Nearest Neighbor Search retrieves a point* $p \in S$ *which is nearest to a given query* $q \in R^{d_q}$.

Here, *nearest* can be defined with various metrics such as the Euclidean distance or cosine similarity. Then NNS will retrieve the point that *minimizes* a distance metric for a given query. In terms of auto-scheduling, the main objective is to find a schedule s that *minimizes* the predicted runtime $\hat{y}(m, s)$ for a given input matrix m. Therefore, we can cast an auto-scheduling as an NNS by setting the dataset S to be all the formats and the schedules in SuperSchedule template, and the **query q** to be the **input matrix m**. If we define a **distance metric** as a **cost** $\hat{\mathbf{y}}(\mathbf{m}, \mathbf{s})$, NNS will retrieve the best SuperSchedule s for a given input matrix m that minimizes the $\hat{y}(m, s)$.

Retrieving an exact nearest neighbor requires exhaustive distance calculations all over the points in S, which is intractable. In practice, Approximate Nearest Neighbor Search (ANNS) [29] has been widely used instead of an exact NNS. For a given query, ANNS cleverly searches the subset of S that speeds up the search while guaranteeing high recall. While there are several approaches to achieve ANNS, we used a graph-based algorithm.

4.2.2 Graph-based ANNS

Graph-based ANNS for auto-scheduling has two phases: building a KNN graph and searching on the KNN graph, as shown in Figure 1-1-(b,c) and Figure 4-7. The first



(a) Building a KNN Graph (b) Searching on KNN Graph

Figure 4-7: Our search strategy via ANNS. In the stage of building the KNN graph, the graph is built by connecting the edge between the schedules with close embeddings in the Euclidean distance. During searching, a query (input matrix m) traverses the graph in the direction predicted runtime $\hat{y}(m, s)$ minimizes.

phase builds a KNN graph whose vertex is the SuperSchedule, and the edge between two vertices is connected only if two program embeddings of the vertices are top-K closest to each other in the l_2 distance. The second phase starts once the query (the input matrix m) comes in. In the second phase, ANNS starts retrieving the schedule s in the graph that minimizes the cost $\hat{y}(m, s)$ by traversing the KNN graph in the direction at which the cost decreases. ANNS can search efficiently because it merely traverses the pre-built KNN graph that guides the direction, whereas other black-box optimizations require expensive metadata updates.

Discussions. The distance metric used in each phase is different. The l_2 distance between two program embeddings of SuperSchedules is used in the building phase, and the cost itself is used in the searching phase ($\|e(s_i) - e(s_j)\|_2$ vs. $\hat{y}(m, s)$). The reason for using completely different metrics is that the KNN graph built upon an l_2 distance has a property that guarantees retrieval of top-K candidates at any generic distance ($\hat{y}(m, s)$ in our case) in the searching phase, while promising high recall [46]. In Section 5.4, we empirically show that graph-based ANNS efficiently and accurately retrieves the optimal SuperSchedule for a given input matrix.

One intuitive explanation of graph-based ANNS is a gradient-based search over discretized space. In fact, our cost model based on a neural network is differentiable, which means we can calculate the gradient for it. Therefore, when searching for the best SuperSchedule, we can actually find the local optima SuperSchedule using the first-order iterative optimization algorithm such as gradient descent. However, there is no guarantee that the local optima we found is a valid encoding of parameters. Specifically, most gradient-based searches will end up with invalidly encoded parameters because we encoded the categorical parameter of SuperSchedule as a one-hot vector and the permutation parameter as a permutation matrix. On the other hand, if we build a KNN graph with valid SuperSchedules, we can think of the ANNS on the KNN graph as a gradient-based search over valid encodings. Projected gradient descent [18] is another way to resolve this, but it was not able to find a good local minimum in our experiments.

Implementation Details. In practice, ANNS libraries build a variant of the KNN graph and traverse it with complicated heuristics to improve the search efficiency. We implemented our search strategy using a state-of-the-art graph ANNS algorithm, HNSW [33]. Although HNSW can support up to a billion-scale graph, it is intractable to build a graph with all formats and schedules in the SuperSchedule as it contains an astronomical number of parameter choices. Therefore, we built the graph with the SuperSchedules which appeared in our training dataset.

Chapter 5

Evaluation

5.1 Experimental Setup

Algorithms. We chose four sparse tensor algebra algorithms for our evaluation. All the algorithms were performed with single-precision data.

- **SpMV(C[i] = A[i,k] * B[k]):** This multiplies sparse matrix(A) by dense vector(B) and stores the product in dense vector(C).

- **SpMM(C[i,j] = A[i,k] * B[k,j]):** This multiplies sparse matrix(A) by dense matrix(B) and stores the product in dense matrix(C). We set the number of columns of dense matrices (|j| in B, C) at 256, and forced both dense matrices' level order to be row-major.

- **SDDMM(D[i,j] = A[i,j] * B[i,k] * C[k,j]):** This performs a sampled matrix multiplication of two dense matrices(B and C). The output matrix D and the input matrix A are sparse matrices. We set the dimension |k| in B, C at 256. We fixed B's level order to be row-major and C's level order to be column-major.

- **MTTKRP(D[i,j] = A[i,k,l]*B[k,j]*C[l,j]):** This performs a matricized tensor times Khatri-Rao product between a 3D sparse tensor(A) and two dense matrices(B and C). We set |j| at 16 and both dense matrices' level order to be row-major. We followed a prior work's approach [44] to generate the training

dataset for 3D sparse tensors.

Baselines. We compared WACO with the following four state-of-the-art baselines. MKL and BestFormat are auto-tuning-based baselines. FixedCSR and ASpT are baselines with a fixed format and schedule.

- **MKL:** Intel MKL sparse BLAS routines [36] utilize an inspector-executor model that auto-tunes a computation on a fixed format. Because it does not support SDDMM and MTTKRP, we only compare it with SpMV and SpMM using the CSR format.

- **BestFormat:** BestFormat automatically selects the appropriate format among a handful candidates for a given sparsity pattern. The candidates were chosen by the five most frequently appearing formats among WACO's search results in the test matrices. We've used prior works' artifacts to predict the best format for a 2D sparse matrix [51] or a 3D sparse tensor [44].

- **Fixed CSR:** Fixed CSR is a code with a fixed format and schedule generated by TACO. We used the CCC format(CSF) for MTTKRP and UC format(CSR) for the rest. We set the OpenMP chunk size at 128, 32, 32, and 32 for SpMV, SpMM, SDDMM, and MTTKRP.

- **ASpT:** ASpT [19] is the state-of-the-art sparse format that directly reorders the sparse matrix to make dense regions. While ASpT is not limited to a specific algorithm, we only compare it only with SpMM and SDDMM because these are the only formats publicly released by the authors.

Implementations. We used TACO to generate the code for the best format and schedule that WACO has found. Then we compiled the generated code with `icc-2021.3.0` with `-march=native -mtune=native -O3 -qopenmp` options. All the experiments were conducted on a dual-socket, 24-core with 48 threads, 2.5 GHz Intel Xeon E5-2680 v3 machine with 30 MB of L3 cache per socket and 128 GB of main memory with Ubuntu 18.04.3 LTS. We used `numactl -interleave=all` to control the NUMA policy.

Table 5.1: Geomean speedup of WACO over other auto-tuners. Format-only and Schedule-only auto-tuner correspond to the BestFormat and MKL, respectively.

	Auto-tuning based baselines	
	vs. Format-only	**vs. Schedule-only**
SpMV	1.43x	2.32x
SpMM	1.18x	1.68x
SDDMM	Not Impl.	Not Impl.
MTTKRP	1.27x	Not Impl.

Table 5.2: Geomean speedup of WACO over other state-of-the-art implementations with a fixed format and schedule.

	Fixed Implementations	
	vs. Fixed CSR	**vs. ASpT**
SpMV	1.54x	Not Impl.
SpMM	1.26x	1.36x
SDDMM	1.29x	1.14x
MTTKRP	1.35x	Not Impl.

We implemented our cost model architecture using PyTorch and MinkowskiEngine for sparse convolution. We trained four separate models for each algorithm, and it took four days to train up to 70 epochs for each model on a single GPU, NVIDIA GeForce RTX 3090 24GB.

5.2 Performance Results

We first evaluated the performance of the format and the schedule that WACO has found. Experiments were conducted on 726 real-world sparse matrices from SuiteSparse that were not included in the training dataset. We picked matrices that had less than 10 million non-zeros and less than 100,000 rows. Among the top-10 SuperSchedules selected by WACO according to the cost model, we report the fastest after we measured them on the hardware. Before we explain the results in detail, the geomean of the speedups on each algorithm are shown in Table 5.1 (vs. auto-tuners) and Table 5.2 (vs. fixed implementations). Overall, WACO performed better than the baselines as it successfully found a specialized format and schedule together for each sparse matrix. This improvement is not limited to a specific algorithm; WACO

can find a better format and schedule for all four algorithms.

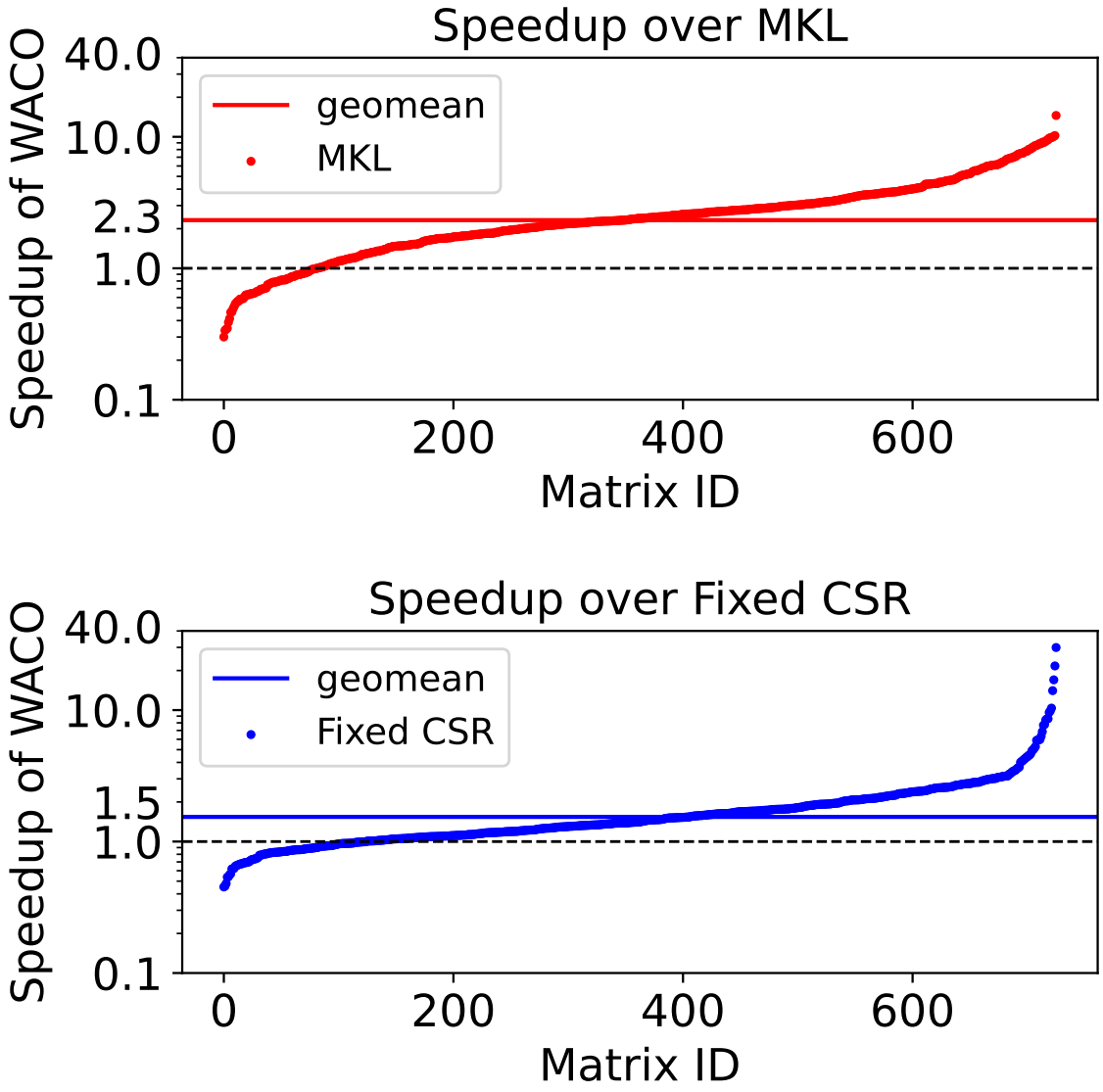


Figure 5-1: Performance comparison on SpMV.

Figure 5-1, 5-2, and 5-3 show the speedups of WACO over the baselines across the test matrices on SpMV, SpMM, and SDDMM, respectively.

Figure 5-2 show the speedups of WACO over four baselines across the test matrices on SpMM. The y axis indicates the speedup of WACO against baselines. All x axes of figures are sorted according to the speedup. The dots below the $y = 1.0$ show matrices in which the baseline performed better than WACO. For MKL and BestFormat, there are more matrices below this line than compared to other baselines because they are

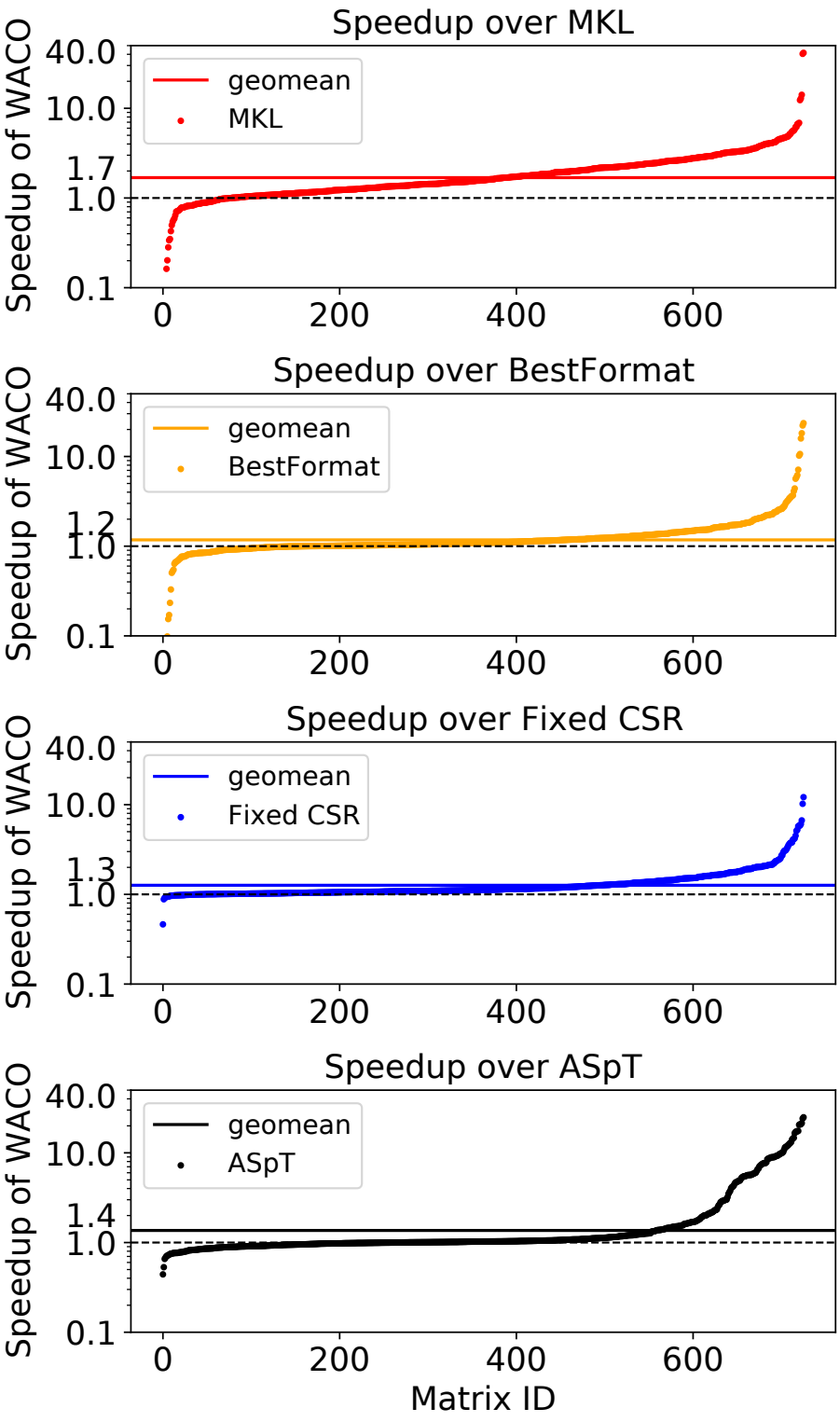


Figure 5-2: Performance comparison on SpMM.

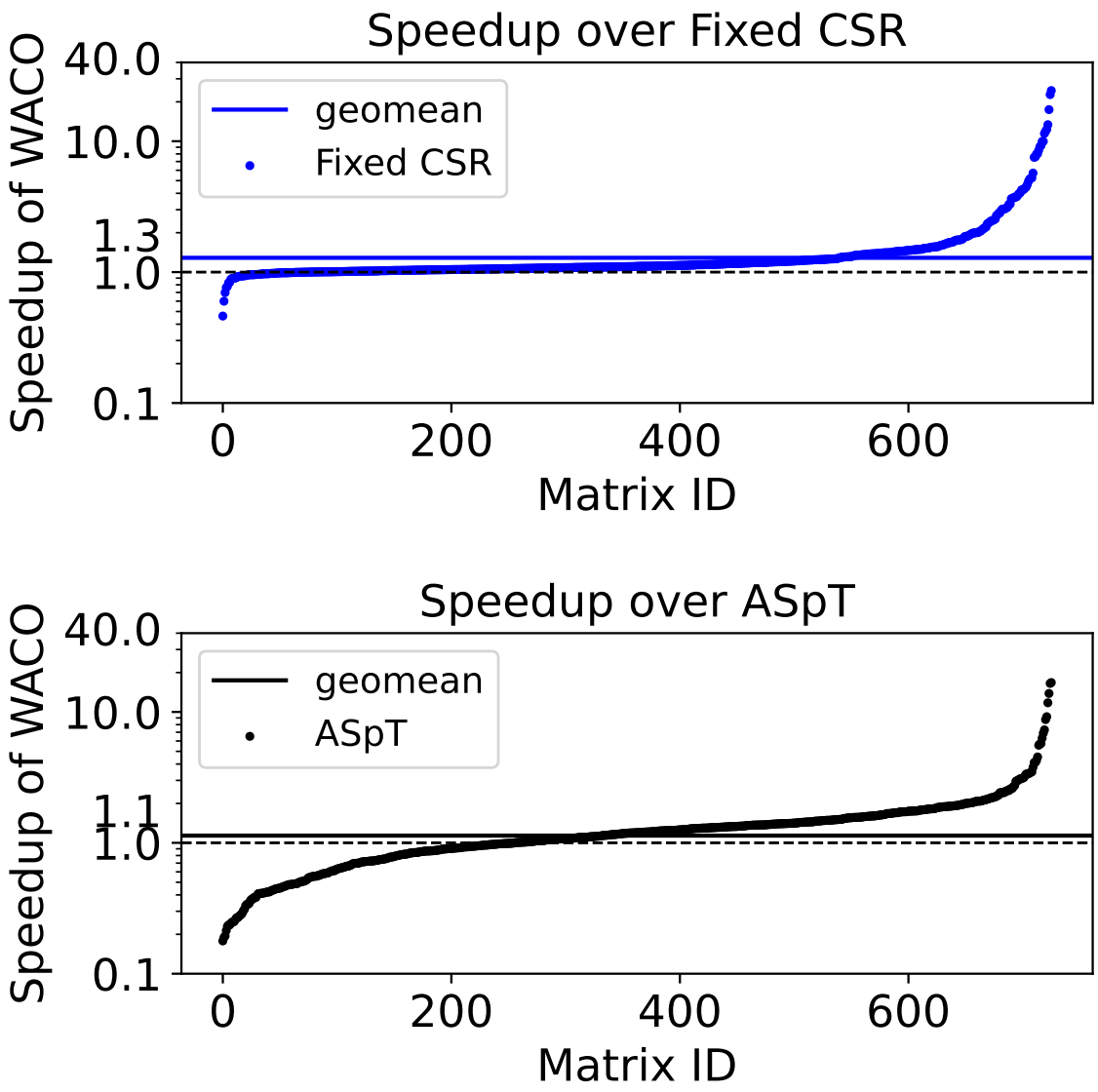


Figure 5-3: Performance comparison on SDDMM.

able to adopt a larger portion of the space though still not as much as of WACO. Thus, auto-tuning based baselines perform better at a few patterns when they find a better format or schedule than WACO.

WACO performed better in SpMM and SDDMM, but to a lesser extent than the improvement of SpMV. Compared to the SpMV, there are many data reuse opportunities in SpMM and SDDMM. Especially, we found that the performance of these algorithms is critically affected by the reuse of (vector) registers. However, because of the data-dependent nature of sparse tensor programs, our generated codes

Table 5.3: Speedup analysis of WACO. The number shows the corresponding factor's percentile among matrices that had a speedup of over 1.5× than the Fixed CSR.

Factor	SpMV	SpMM	SDDMM
OpenMP Chunk Size	51%	66%	47%
Dense Block >50% Filled	30%	26%	15%
Dense Block <50% Filled	19%	-	-
Sparse Block	-	8%	-
Parallelize over Column	-	-	38%

contain a lot of indirect accesses. Such accesses prevent C compilers from enabling register-related optimizations such as unrolling or vectorizing. Thus, a significant performance improvement happened only in the case of a format with a small size dense block, which was easy for the C compiler to optimize. A matrix without dense blocks had a relatively small performance improvement.

5.2.1 Discussion on Speedup

We further analyzed the source of the speedups on SpMV, SpMM, and SDDMM. We picked the matrices with a speedup > 1.5× than the Fixed CSR and classified the speedup factors into five categories. Table 5.3 shows these categories and their proportions.

SpMV. First, half of the matrices benefitted from choosing the appropriate OpenMP chunk size, which controls the load balancing across multiple processors. Another half benefits from storing the matrix into a dense blocked format which exploits the register reuse in the dense block. A dense blocked format can be represented UCU or UCUU in a format abstraction. One counter-intuitive factor is the speedup of a matrix with the non-zeros filling less than 50% of the dense block. Storing such matrices into a dense blocked format usually results in memory increase due to unnecessary zeros. Nevertheless, a speedup occurs because of the heuristic decision in the Intel `icc` compiler regarding utilizing SIMD instructions. As shown in the Figure 5-4, we found out that `icc` starts to exploit the SIMD instructions when the block size is larger than 16. It is surprising to see that WACO learned the compiler's heuristics

```
1:  const int b = ?
2:  for (int i1 = 0; i1 < A1_dimension; i1++) {
3:   for (int kA = A2_pos[i1]; kA < A2_pos[i1+1]; kA++) {
4:    int k = A2_crd[kA];
5:    for (int i0 = 0; i0 < b; i0++){
6:     int i0A = kA * b + i0;
7:     int i0C = i1 * b + i0;
8:     C_vals[i0C] += A_vals[i0A] * B_vals[k];
9:    }}}
```

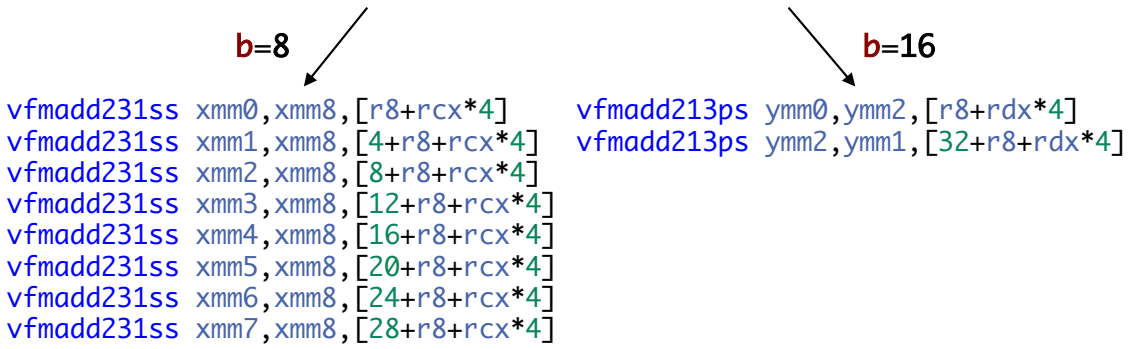


Figure 5-4: `icc` generated assembly for SpMV with the UCU format. `b` decides the size of the one-dimensional dense block. `icc` starts to use the AVX instructions(`vfmadd213ps`) from `b`=16.

and intentionally chose the larger block size to utilize the vector registers despite the memory increase.

SpMM. Like SpMV, most matrices benefitted from better load balancing by choosing an appropriate chunk size. Other than that, some matrices benefitted from a unique format, which we call a *sparse block format*. Compared to the dense block format where the level format of the inner split level is Uncompressed (e.g., UC$\underline{\text{U}}$ or UC$\underline{\text{UU}}$), a sparse block format stores the inner level into the Compressed format (e.g., UU$\underline{\text{C}}$). Splitting the level into the Compressed level format with a large split size helps improving the cache locality in SpMM. For instance, the LLC miss rate was reduced to 7% from 36% and the performance improved about 2.5× when we stored sparsine (Figure 2-1) into the $k1(U) \rightarrow i(U) \rightarrow k0(C)$ format by splitting k by 16,384.

SDDMM. Other than better load balancing and a dense block format, SDDMM can take an advantage of the use of a column-major format. One difference between SDDMM and other algorithms is that it is safe to parallelize both rows and columns of

Name	Rows	Cols	Nonzeros
amazon0302_T_8	241,096	193,156	837,977
mip1_3	5,382	5,282	613,000
soc-LiveJournal1_T_2	842,771	500,182	489,256

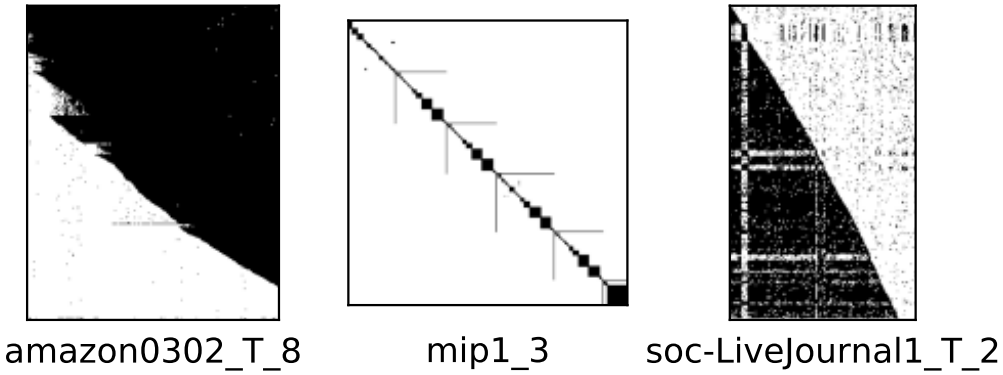


Figure 5-5: Sparse matrices used for the evaluation.

the sparse matrix in SDDMM. For SpMV(C[i] = A[i,k] * B[k])) or SpMM(C[i,j] = A[i,k] * B[k,j]), it is inefficient to parallelize over the column of the sparse matrix(k in A[i,k]) because the reduction occurs along that dimension. Therefore, WACO flexibly chose the row-major or column-major format without any restriction in the parallelizing dimension on SDDMM.

5.2.2 Case Studies

Case Study : mip1_3. mip_3 is a sparse matrix with dense blocks. In the context of SpMM, WACO has identified the optimal computational schedule and format as follows:

```
.split(i,i1,i0,4)                A.reorder(i1,k1,i0,k0)
.split(k,k1,k0,2)                A.format(i1,U)
.split(j,j1,j0,32)               A.format(i0,U)
.reorder(i1,k1,j1,i0,k0,j0)      A.format(k1,C)
.parallelize(i1,48,4)            A.format(k0,U)
```

Our feature extractor was able to capture the existence of dense blocks and chose the 4x2 dense blocked UCUU format, traditionally known as BCSR, for mip_3. When we take a look into compute schedule, WACO cleverly splits j dimension into 32 and reorder the loops to place the split dimensions $(i0, k0, j0)$ at the innermost loop to take

advantage of $4 \times 2 \times 32$ register blocking and inner loop vectorization. Furthermore, within the `parallelize` schedule, we determined that the optimal chunk size (4) discovered by WACO achieves excellent load balancing across threads.

Case Study : soc-LiveJournal1_T_2. This sparse matrix exhibits numerous empty rows, yet the non-empty rows contain a small number of non-zeros. In the context of SpMV, WACO has derived the following format and computation schedule:

```
.split(i,i1,i0,64)                A.reorder(i1,i0,k0,k1)
.split(k,k1,k0,1)                 A.format(i1,C)
.reorder(i1,i0,k0,k1)             A.format(i0,C)
.parallelize(i1,48,32)            A.format(k1,C)
                                  A.format(k0,U)
```

We observe that WACO designates $i1$ and $i0$ as a Compressed level format – a hybrid configuration that combines attributes of BCSR and DCSR. This selection is attributed to the matrix's significant proportion of empty rows. Given the sparse distribution of non-zeros per row, effective utilization of a coarse-grained load-balancing strategy is essential. To achieve this goal, WACO parallelizes $i1$ using a chunk size of 32 while splitting the i dimension into 64 segments. This approach achieves effects akin to directly specifying a chunk size of 32*64 without splitting. However, as we uphold a chunk size limitation of 256 or less (Table 4.1), WACO devises a strategy for enhanced coarse-grained load balancing through splitting.

5.3 Cost Model

5.3.1 Model Accuracy

We assessed how well our trained model reflects the relative order between different schedules according to different sparsity patterns. When searching for the best schedule, predicting the correct order significantly affects the quality of the search result. Figure 5-6 shows the correlation between the predicted ranking and true ranking in our SpMV cost model for each matrix. We sampled 500 schedules from OpenTuner and plotted each point where x and y axes denote the true ranking among all points and corresponding predicted ranking using our cost model. Red line, $y = x$,

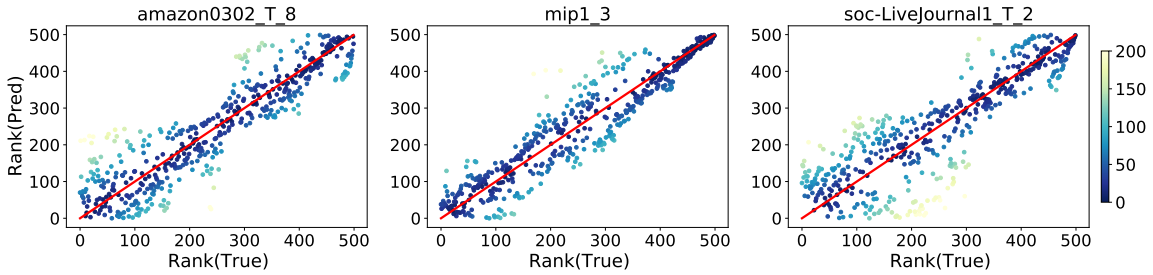


Figure 5-6: The predicted ranking and true ranking of runtimes in SpMV cost model. The x axis denotes the true ranking and y axis denotes corresponding predicted ranking. The color bar in the right indicates the absolute difference between predcited and true ranking.

depicts a perfect correlation, so the closer to $y = x$, the better prediction. We also measured the Spearman's rank correlation coefficients (ρ), a measure of how well two variables are monotonically associated, for each matrix. ρ of amazon0302_T_8, mip1_3, and soc-LiveJournal1_T_2 were 0.90, 0.93, and 0.86, respectively, indicating that our model well understands the relative order between different schedules across different sparsity patterns. Coefficients of our SpMM cost model were 0.91,0.92, and 0.88, which were also highly correlated.

5.3.2 Cost Model Exploration

We conducted the experiment to test how effectively our feature extractor learns meaningful features of the sparsity pattern. The train-validation losses of four alternative cost models, each of which uses a different feature extractor, are shown in Figure 5-7. HumanFeature uses three simple statistics of the sparsity pattern, (# rows, # cols, and # non-zeros). DenseConv [51] uses a conventional CNN after downsampling an input matrix into 256×256. MinkowskiNet [13] is a popular deep learning model based on sparse convolution layers for 3D point clouds. Due to the limited size of GPU memory, we reduced the number of channels in MinkowskiNet to support a matrix with 10 million non-zeros. WACONet is our feature extractor that described in Section 4.1.1. There is a marked difference between the HumanFeature and the remaining three networks using convolution. WACONet and MinkowskiNet, networks that use a sparse convolutional layer, learn better than DenseConv because DenseConv causes the loss

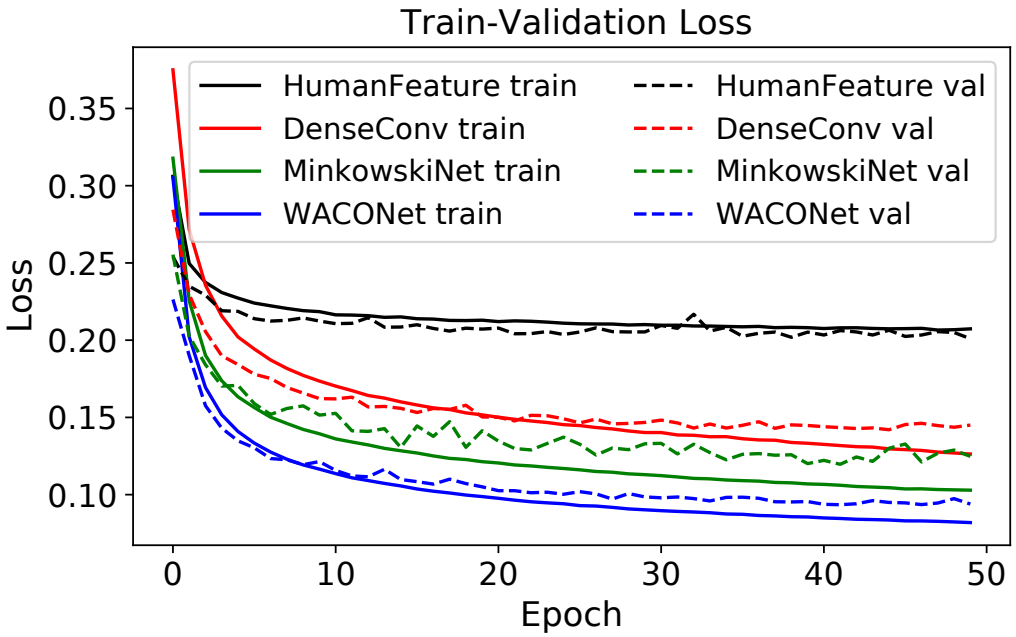


Figure 5-7: Train-validation losses of the SpMM cost models using four different feature extractors.

of pattern information during downsampling as explained in Section 3.2. Finally, as the strided convolution accommodates distant non-zeros, WACONet retrieved more meaningful features than MinkowksiNet.

5.4 Search Strategy Exploration

Different Search Strategies. We compared ANNS with two other black-box optimization search strategies, HyperOpt [6] and OpenTuner [3]. HyperOpt utilizes Bayesian optimization, and OpenTuner utilizes an ensemble of search techniques that use multi-armed bandit. For each search strategy, we ran 3,000 trials to search for the optimal parameters of SuperSchedule on the SpMM cost model with a bcsstk29 matrix. As shown in Figure 5-8-(a), ANNS found the lowest cost within an equal number of trials. OpenTuner also found a comparable cost to that of ANNS, but the search time is much longer. We can summarize why ANNS is substantially faster than the others into three reasons. First, ANNS does not require any metadata update, which is common in machine learning-based black-box optimization, such as a training

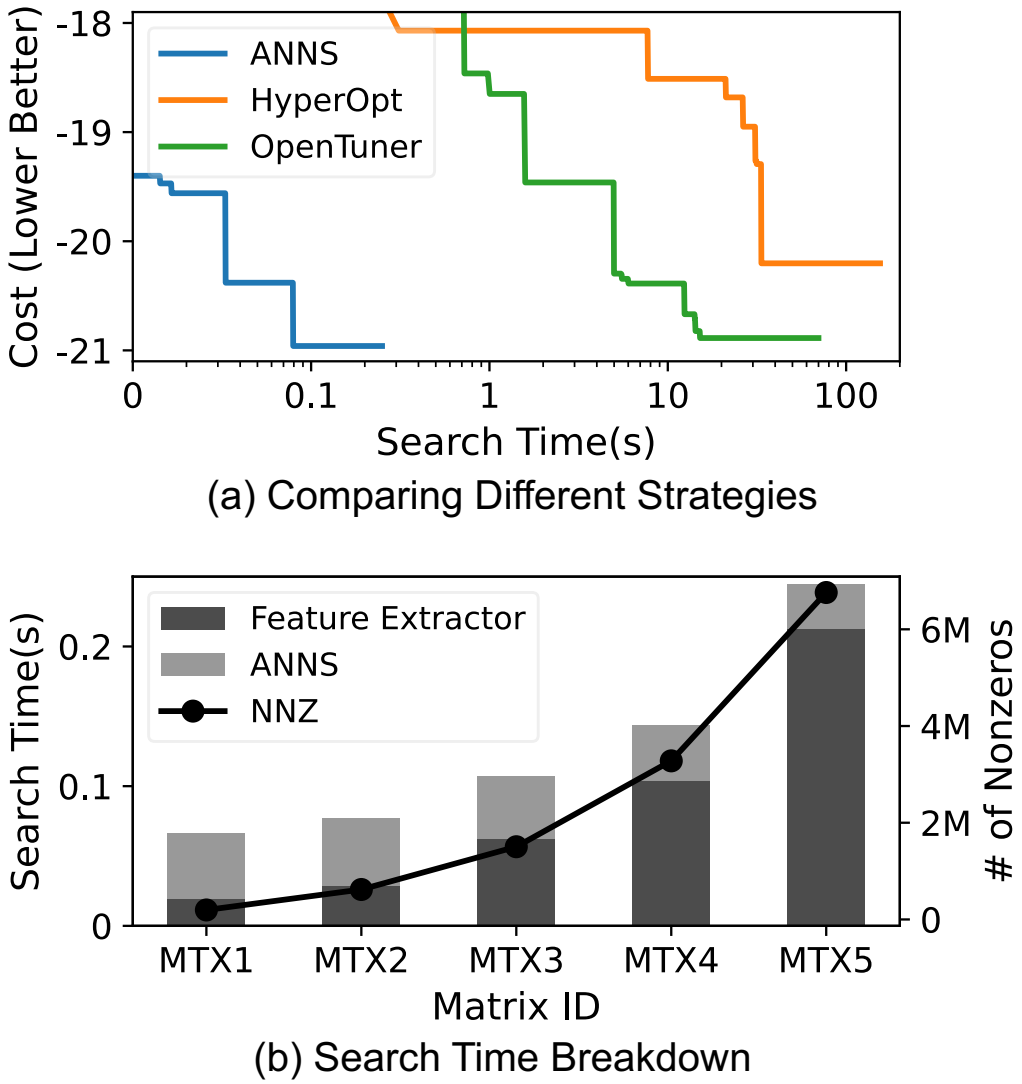


Figure 5-8: Exploring different search strategies and breaking down the search time of WACO on SpMM

surrogate model in Bayesian optimization. ANNS can efficiently search for different formats and schedules merely by traversing the KNN graph. Second, a KNN graph memorizes the program embedding of each SuperSchedule(vertex) during the building phase. Thus, it does not have to run for the entire cost model; it only needs to run for the final part of the cost model (Figure 1-1-(c)). Finally, ANNS is implemented in C++, whereas most black-box optimization libraries are built on Python.

Search Time Breakdown. Since the sparsity pattern feature is reusable when

Table 5.4: WACO's SpMM geomean speedup over FixedCSR with a cost model trained on same/different hardware.

Speedup over FixedCSR		**Trained on**	
		Intel CPU	AMD CPU
Tested on	Intel CPU	**1.26x**	1.12x
	AMD CPU	1.08x	**1.21x**

calculating the cost of the different schedules, WACO does not run the feature extractor multiple times for an input matrix. Instead, the search can be divided into two phases (Figure 1-1-(c)): (1) extracting the sparsity pattern feature by running WACONet and (2) ANNS with the final part of the cost model. Figure 5-8-(b) shows the search time breakdown of five different matrices with varying numbers of non-zeros. When the number of non-zeros is less than 1.5 million, ANNS dominates the entire search time, but the feature extractor becomes more expensive when the number of non-zeros increases. This is because the computational cost of sparse convolution depends on the number of non-zeros.

5.5 Generalization on Other Hardware

WACO's cost model is somewhat hardware-specific as it does better when trained on target hardware. However, the cost model does transfer general optimization patterns between hardware. For example, a sparsity pattern with skewed non-zero distribution will generally prefer a fine-grained load-balancing. To demonstrate this generalization, we trained a SpMM cost model on different hardware and compiler: 8-core(16 threads) AMD EPYC 7R32 with a 16MB L3 cache and `gcc-11`. Data collection took about 4 days on 8 nodes, and training a cost model took about 3 days on a single GPU. Table 5.4 shows the speedup of WACO under 2×2 possible configurations. As expected, the diagonal of the table shows the best performance because the cost model is trained for the target hardware. WACO, in general, found a better format and schedule than a baseline with a model trained on a different hardware.

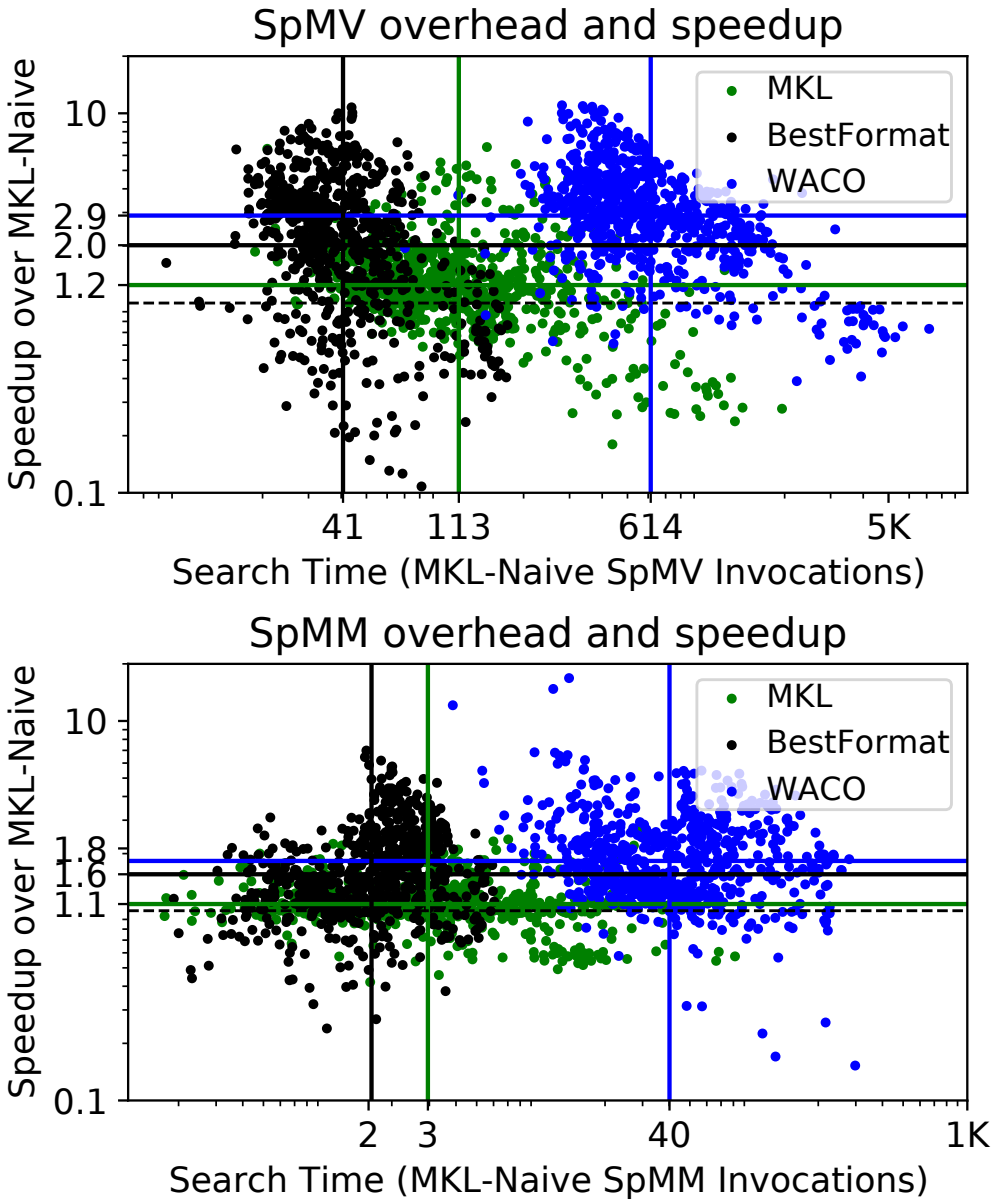


Figure 5-9: Tuning overhead of the MKL inspector-executor (Schedule-only tuner), the BestFormat (Format-only tuner), and the WACO. We compared all methods against the auto-tuning disabled MKL (MKL-Naive).

5.6 Search Overhead and Usage Scenarios

Tuning Overhead. Because the program auto-tuning pays for the search or tuning time(T_{tuning}) for the speedup, we will discuss the search overhead of WACO. Figure 5-9 shows the search time - speedup plot of three auto-tuning frameworks, MKL inspector-

Table 5.5: Real-world applications that require repetitive (a) SpMVs and (b) SpMMs. Green cells indicate that the corresponding auto-tuner wins. Initial cost is computed as $T_{tuning} + T_{formatconvert}$, but only T_{tuning} for MKL.

(a) SpMV		End-to-end Execution Time (in MKL-Naive SpMV calls)		
Label	N_{runs}	WACO	BestFormat	MKL
Initial Cost	0	821	277	**113**
PageRank [52]	50	838	302	**153**
WACO=MKL	1,546	1,356	**1,044**	1,356
WACO=BestFormat	3,627	**2,075**	**2,075**	3,028
GMRES [31]	517K	**180K**	257K	416K
Mesh sim. [26]	1.8M	**623K**	892K	1.4M

(b) SpMM		End-to-end Execution Time (in MKL-Naive SpMM calls)		
Label	N_{runs}	WACO	BestFormat	MKL
Initial Cost	0	46	7	**3**
WACO=MKL	115	109	**80**	109
WACO=BestFormat	412	**271**	**271**	382
GNN [7]	10K	**5,511**	6,432	9,224
Pruned NN [16]	1.0M	**546K**	642K	922K

executor, BestFormat, and WACO. We compared these frameworks against naive MKL without an inspector-executor. For the matrices with a speedup >1.0×, SpMV and SpMM must be repeatedly run for 919 and 101 times to amortize the WACO's tuning cost on average. As pointed out in Section 5.4, a feature extractor must be more lightweight to reduce this amortization cost. When comparing BestFormat and MKL, BestFormat showed better performance on both search overhead(T_{tuning}) and speedup than MKL. However, when comparing WACO and BestFormat, there was a clear trade-off; WACO achieves a better speedup by paying for more search time than BestFormat.

Real-world Scenarios. Real-world applications that utilize an auto-tuner should consider both the tuning cost and the format converting cost. To be specific, the end-to-end execution time ($T_{tuning} + T_{formatconvert} + T_{tunedkernel} * N_{runs}$) needs to be considered [52, 55]. The auto-tuner with a significant search overhead, such as WACO,

is only advantageous over other prior auto-tuners in applications requiring repetitive runs. We list some real-world applications with tens of thousands of runs of sparse routines in Table 5.5. We set $T_{formatconvert} = 0$ for MKL as it only tunes the schedule while fixing the format. Although BestFormat has the fastest T_{tuning} (Figure 5-9), MKL is advantageous when N is small due to no format conversion. It is better to use other auto-tuners if an application does not require many repetitions, such as PageRank. WACO is beneficial in scenarios that require a lot of runs, such as mesh simulation or GNN.

Chapter 6

Related Works

Auto-scheduling and Cost Model. Halide auto-scheduler [37, 1] uses a cost model with hand-crafted program features and searches for the best schedule through a beam search. AutoTVM [10] uses a cost model that embeds the low-level loop AST. While AutoTVM automates the search process, its search space must be manually defined by the user's template. Recently, Ansor [53] allowed the auto-scheduler to find this template automatically by rewriting rules. The Tiramisu auto-scheduler uses LSTM to embed the low-level loop AST [4]. There have also been many cost models that tried to predict the behavior of accelerator or x86 basic blocks [21, 39, 35, 40]. All these schemes attempted to design a cost model to embed a traversing order of iteration space alone since they usually targeted a dense tensor program, while WACO's cost model considers the sparsity pattern, the format and the schedule all together.

Auto-tuning Sparse Tensor Programs. Previous auto-tuner of sparse tensor programs can be divided into two categories: format selection studies and schedule optimization studies. There has been a format selection approach designed a classifier, which took a downsampled tensor and predicted which format would be optimal for the input [44, 51]. However, the features extracted over the downsampled tensor did not capture the pattern well and considered only a few output classes, for example, five formats, whereas WACO considers a large number of formats from the TACO's abstraction. Some other frameworks estimate the number of non-zeros in the dense

block to choose the optimal block size in the BCSR [11, 48]. Mehrabi et al. utilized a predictive model to learn the optimal permutation of rows for better load balancing [34].

Regarding auto-tuning of the schedule, ESB [30] suggested choosing an optimal load-balancing scheme by running a kernel several times, each time with different load-balancing schemes. Venkat et al. proposed an inspector-executor method to transform a sparse loop and data with polyhedral optimizations [47]. Their three proposed transformations *make-dense*, *compact*, and *compact-and-pad* can actually demonstrate the same search space as TACO's transformation framework [43] provided there is a single sparse input among all input operands. However, they only suggested how to transform the sparse loop but not how to transform it automatically. Therefore, WACO can be used as an auto-tuner to automatically transform the code by replacing TACO with their framework.

Chapter 7

Conclusion & Future Work

This thesis presents WACO, a technique co-optimizing the format and the schedule for a given sparsity pattern. In sparse tensor programs, it is crucial to design the cost model to consider various sparsity patterns. To address this, we propose a novel feature extractor that employs a sparse convolutional network. Its obtained features are universal across various formats and are useful for predicting the coupled behavior between the format and the schedule. Furthermore, a graph-based ANNS, a discretized version of gradient-based search, efficiently and accurately finds the best format and schedule in the large search space of the co-optimization. We evaluate WACO for four different algorithms (SpMV, SpMM, SDDMM, and MTTKRP) on a CPU using diverse sparsity patterns. Our experimental results shows that WACO outperformed four state-of-the-art baselines, Intel MKL, Format-only auto-tuner, TACO with a default schedule, and ASpT. Compared to the best of four baselines, WACO achieved 1.43×, 1.18×, 1.14×, and 1.27× average speedups on SpMV, SpMM, SDDMM, and MTTKRP, respectively. While our investigation has unveiled performance improvements across various algorithms through the use of WACO, it is important to acknowledge the presence of certain limitations.

Platform-agnostic Cost Model. The current cost model employed by WACO relies on a training dataset collected from a specific platform. As highlighted in Section 5.5, this model exhibits a hardware-specific characteristic, demonstrating

enhanced performance when trained on the target hardware. Furthermore, a cost model is also limited by compiler-specific attributes, given our utilization of `icc` for compiling the generated C code. Here are two potential solutions to address these challenges: (1) a redesigned cost model that incorporates platform-specific information as part of its input, and (2) the application of transfer learning, leveraging knowledge acquired from one platform to enhance performance on a different platform. These strategies hold the promise of broadening WACO's applicability and mitigating the limitations associated with its current cost model.

Algorithm-agnostic Cost Model. Currently, our cost model is exclusively tailored to specific algorithms. For instance, the cost model trained for SpMV cannot be utilized to find the optimal format and schedule for SpMM. However, the process of collecting datasets and training models for individual algorithms is resource-intensive and time-consuming. The algorithm-specific nature of the current cost model stems from two primary factors: (1) the absence of a universal feature extractor, and (2) the absence of a universal program embedder. Firstly, the feature extractor must accommodate sparse tensors of any dimensionality, yet the current WACONet design is limited to only 2D sparse matrices or 3D sparse tensors. Additionally, when multiple sparse operands are involved, such as in the case of SpMSpM, the feature extractor needs to consider interactions like intersections among these sparse operands. The future architecture of the feature extractor may take multiple sparsity patterns as inputs, allowing for the consideration of interactions between these patterns. This architecture is commonly observed in 3D point cloud registration models [32, 27]. Secondly, the current program embedder requires a SuperSchedule as input, and this SuperSchedule depends on the algorithm itself. Consequently, a future program embedder should adopt an algorithm-agnostic representation as input, such as the generated C code or LLVM IR, to address this challenge.

Bibliography

[1] Andrew Adams, Karima Ma, Luke Anderson, Riyadh Baghdadi, Tzu-Mao Li, Michaël Gharbi, Benoit Steiner, Steven Johnson, Kayvon Fatahalian, Frédo Durand, et al. Learning to optimize halide with tree search and random programs. *ACM Transactions on Graphics (TOG)*, 38(4):1–12, 2019.

[2] Martin Alnæs, Jan Blechta, Johan Hake, August Johansson, Benjamin Kehlet, Anders Logg, Chris Richardson, Johannes Ring, Marie E Rognes, and Garth N Wells. The fenics project version 1.5. *Archive of Numerical Software*, Vol 3, 2015.

[3] Jason Ansel, Shoaib Kamil, Kalyan Veeramachaneni, Jonathan Ragan-Kelley, Jeffrey Bosboom, Una-May O'Reilly, and Saman Amarasinghe. Opentuner: An extensible framework for program autotuning. In *Proceedings of the 23rd international conference on Parallel architectures and compilation*, pages 303–316, 2014.

[4] Riyadh Baghdadi, Massinissa Merouani, Mohamed-Hicham Leghettas, Kamel Abdous, Taha Arbaoui, Karima Benatchba, et al. A deep learning based cost model for automatic code optimization. *Proceedings of Machine Learning and Systems*, 3, 2021.

[5] Riyadh Baghdadi, Jessica Ray, Malek Ben Romdhane, Emanuele Del Sozzo, Abdurrahman Akkas, Yunming Zhang, Patricia Suriana, Shoaib Kamil, and Saman Amarasinghe. Tiramisu: A polyhedral compiler for expressing fast and portable code. In *2019 IEEE/ACM International Symposium on Code Generation and Optimization (CGO)*, pages 193–205. IEEE, 2019.

[6] James Bergstra, Daniel Yamins, and David Cox. Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures. In *International conference on machine learning*, pages 115–123. PMLR, 2013.

[7] Filippo Maria Bianchi, Daniele Grattarola, and Cesare Alippi. Spectral clustering with graph neural networks for graph pooling. In *International Conference on Machine Learning*, pages 874–883. PMLR, 2020.

[8] Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. Learning to rank using gradient descent. In *Proceedings of the 22nd international conference on Machine learning*, pages 89–96, 2005.

[9] Tianqi Chen, Thierry Moreau, Ziheng Jiang, Lianmin Zheng, Eddie Yan, Meghan Cowan, Haichen Shen, Leyuan Wang, Yuwei Hu, Luis Ceze, Carlos Guestrin, and Arvind Krishnamurthy. Tvm: An automated end-to-end optimizing compiler for deep learning. In *Proceedings of the 13th USENIX Conference on Operating Systems Design and Implementation*, OSDI'18, page 579–594, USA, 2018. USENIX Association.

[10] Tianqi Chen, Lianmin Zheng, Eddie Yan, Ziheng Jiang, Thierry Moreau, Luis Ceze, Carlos Guestrin, and Arvind Krishnamurthy. Learning to optimize tensor programs. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 3393–3404, Red Hook, NY, USA, 2018. Curran Associates Inc.

[11] Jee W Choi, Amik Singh, and Richard W Vuduc. Model-driven autotuning of sparse matrix-vector multiply on gpus. *ACM sigplan notices*, 45(5):115–126, 2010.

[12] Stephen Chou, Fredrik Kjolstad, and Saman Amarasinghe. Format abstraction for sparse tensor algebra compilers. *Proceedings of the ACM on Programming Languages*, 2(OOPSLA):1–30, 2018.

[13] Christopher Choy, JunYoung Gwak, and Silvio Savarese. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3075–3084, 2019.

[14] Timothy A Davis and Yifan Hu. The university of florida sparse matrix collection. *ACM Transactions on Mathematical Software (TOMS)*, 38(1):1–25, 2011.

[15] Matteo Frigo and Steven G Johnson. Fftw: An adaptive software architecture for the fft. In *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'98 (Cat. No. 98CH36181)*, volume 3, pages 1381–1384. IEEE, 1998.

[16] Trevor Gale, Matei Zaharia, Cliff Young, and Erich Elsen. Sparse gpu kernels for deep learning. In *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1–14. IEEE, 2020.

[17] Benjamin Graham and Laurens van der Maaten. Submanifold sparse convolutional networks. *arXiv preprint arXiv:1706.01307*, 2017.

[18] Kartik Hegde, Po-An Tsai, Sitao Huang, Vikas Chandra, Angshuman Parashar, and Christopher W. Fletcher. Mind mappings: Enabling efficient algorithm-accelerator mapping space search. In *Proceedings of the 26th ACM International Conference on Architectural Support for Programming Languages and Operating Systems*, ASPLOS 2021, 2021.

[19] Changwan Hong, Aravind Sukumaran-Rajam, Israt Nisa, Kunal Singh, and P Sadayappan. Adaptive sparse tiling for sparse matrix multiplication. In *Proceedings of the 24th Symposium on Principles and Practice of Parallel Programming*, pages 300–314, 2019.

[20] Guyue Huang, Guohao Dai, Yu Wang, and Huazhong Yang. Ge-spmm: General-purpose sparse matrix-matrix multiplication on gpus for graph neural networks. In *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1–12. IEEE, 2020.

[21] Qijing Huang, Aravind Kalaiah, Minwoo Kang, James Demmel, Grace Dinh, John Wawrzynek, Thomas Norell, and Yakun Sophia Shao. Cosa: Scheduling by constrained optimization for spatial accelerators. In *2021 ACM/IEEE 48th Annual International Symposium on Computer Architecture (ISCA)*, pages 554–566, 2021.

[22] Zhihao Jia, Oded Padon, James Thomas, Todd Warszawski, Matei Zaharia, and Alex Aiken. Taso: optimizing deep learning computation with automatic generation of graph substitutions. In *Proceedings of the 27th ACM Symposium on Operating Systems Principles*, pages 47–62, 2019.

[23] Jeremy Kepner, Peter Aaltonen, David Bader, Aydin Buluç, Franz Franchetti, John Gilbert, Dylan Hutchison, Manoj Kumar, Andrew Lumsdaine, Henning Meyerhenke, Scott McMillan, Carl Yang, John D. Owens, Marcin Zalewski, Timothy Mattson, and Jose Moreira. Mathematical foundations of the graphblas. In *2016 IEEE High Performance Extreme Computing Conference (HPEC)*, pages 1–9, 2016.

[24] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[25] Fredrik Kjolstad, Shoaib Kamil, Stephen Chou, David Lugato, and Saman Amarasinghe. The tensor algebra compiler. *Proceedings of the ACM on Programming Languages*, 1(OOPSLA):1–29, 2017.

[26] Fredrik Kjolstad, Shoaib Kamil, Jonathan Ragan-Kelley, David IW Levin, Shinjiro Sueda, Desai Chen, Etienne Vouga, Danny M Kaufman, Gurtej Kanwar, Wojciech Matusik, et al. Simit: A language for physical simulation. *ACM Transactions on Graphics (TOG)*, 35(2):1–21, 2016.

[27] Donghoon Lee, Onur C. Hamsici, Steven Feng, Prachee Sharma, and Thorsten Gernoth. Deeppro: Deep partial point cloud registration of objects. In *ICCV*, 2021.

[28] Jiajia Li, Guangming Tan, Mingyu Chen, and Ninghui Sun. Smat: an input adaptive auto-tuner for sparse matrix-vector multiplication. In *Proceedings of the 34th ACM SIGPLAN conference on Programming language design and implementation*, pages 117–126, 2013.

[29] Wen Li, Ying Zhang, Yifang Sun, Wei Wang, Mingjie Li, Wenjie Zhang, and Xuemin Lin. Approximate nearest neighbor search on high dimensional data—experiments, analyses, and improvement. *IEEE Transactions on Knowledge and Data Engineering*, 32(8):1475–1488, 2019.

[30] Xing Liu, Mikhail Smelyanskiy, Edmond Chow, and Pradeep Dubey. Efficient sparse matrix-vector multiplication on x86-based many-core processors. In *Proceedings of the 27th international ACM conference on International conference on supercomputing*, pages 273–282, 2013.

[31] Jennifer A. Loe, Heidi K. Thornquist, and Erik G. Boman. Polynomial preconditioned gmres to reduce communication in parallel computing, 2019.

[32] Weixin Lu, Guowei Wan, Yao Zhou, Xiangyu Fu, Pengfei Yuan, and Shiyu Song. Deepvcp: An end-to-end deep neural network for point cloud registration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[33] Yu A Malkov and Dmitry A Yashunin. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE transactions on pattern analysis and machine intelligence*, 42(4):824–836, 2018.

[34] Atefeh Mehrabi, Donghyuk Lee, Niladrish Chatterjee, Daniel J. Sorin, Benjamin C. Lee, and Mike O'Connor. Learning sparse matrix row permutations for efficient spmm on GPU architectures. In *IEEE International Symposium on Performance Analysis of Systems and Software, ISPASS 2021, Stony Brook, NY, USA, March 28-30, 2021*, pages 48–58. IEEE, 2021.

[35] Charith Mendis, Alex Renda, Dr.Saman Amarasinghe, and Michael Carbin. Ithemal: Accurate, portable and fast basic block throughput estimation using deep neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*. PMLR, 2019.

[36] Intel MKL. Inspector-executor sparse blas routines. 2022.

[37] Ravi Teja Mullapudi, Andrew Adams, Dillon Sharlet, Jonathan Ragan-Kelley, and Kayvon Fatahalian. Automatically scheduling halide image processing pipelines. *ACM Transactions on Graphics (TOG)*, 35(4):1–11, 2016.

[38] Sameer A Nene and Shree K Nayar. A simple algorithm for nearest neighbor search in high dimensions. *IEEE Transactions on pattern analysis and machine intelligence*, 19(9):989–1003, 1997.

[39] Angshuman Parashar, Priyanka Raina, Yakun Sophia Shao, Yu-Hsin Chen, Victor A. Ying, Anurag Mukkara, Rangharajan Venkatesan, Brucek Khailany, Stephen W. Keckler, and Joel Emer. Timeloop: A systematic approach to dnn accelerator evaluation. In *2019 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS)*, pages 304–315, 2019.

[40] Mangpo Phothilimthana, Mike Burrows, and Samuel J. Kaufman. Learned tpu cost model for xla tensor programs. In *Workshop on ML for Systems at NeurIPS*, 2019.

[41] Jonathan Ragan-Kelley, Connelly Barnes, Andrew Adams, Sylvain Paris, Frédo Durand, and Saman Amarasinghe. Halide: a language and compiler for optimizing parallelism, locality, and recomputation in image processing pipelines. *Acm Sigplan Notices*, 48(6):519–530, 2013.

[42] Naser Sedaghati, Te Mu, Louis-Noel Pouchet, Srinivasan Parthasarathy, and P Sadayappan. Automatic selection of sparse matrix representation on gpus. In *Proceedings of the 29th ACM on International Conference on Supercomputing*, pages 99–108, 2015.

[43] Ryan Senanayake, Changwan Hong, Ziheng Wang, Amalee Wilson, Stephen Chou, Shoaib Kamil, Saman Amarasinghe, and Fredrik Kjolstad. A sparse iteration space transformation framework for sparse tensor algebra. *Proceedings of the ACM on Programming Languages*, 4(OOPSLA):1–30, 2020.

[44] Qingxiao Sun, Yi Liu, Ming Dun, Hailong Yang, Zhongzhi Luan, Lin Gan, Guangwen Yang, and Depei Qian. Sptfs: sparse tensor format selection for mttkrp via deep learning. In *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1–14. IEEE, 2020.

[45] Vivienne Sze, Yu-Hsin Chen, Tien-Ju Yang, and Joel S Emer. Efficient processing of deep neural networks. *Synthesis Lectures on Computer Architecture*, 15(2):1–341, 2020.

[46] Shulong Tan, Zhixin Zhou, Zhaozhuo Xu, and Ping Li. Fast item ranking under neural network based measures. In *Proceedings of the 13th International Conference on Web Search and Data Mining*, pages 591–599, 2020.

[47] Anand Venkat, Mary Hall, and Michelle Strout. Loop and data transformations for sparse matrix code. In *Proceedings of the 36th ACM SIGPLAN Conference on Programming Language Design and Implementation*, PLDI '15, page 521–532, New York, NY, USA, 2015. Association for Computing Machinery.

[48] Richard Vuduc, James W Demmel, and Katherine A Yelick. Oski: A library of automatically tuned sparse matrix kernels. In *Journal of Physics: Conference Series*, volume 16, page 071. IOP Publishing, 2005.

[49] R Clinton Whaley and Jack J Dongarra. Automatically tuned linear algebra software. In *SC'98: Proceedings of the 1998 ACM/IEEE conference on Supercomputing*, pages 38–38. IEEE, 1998.

[50] Jaeyeon Won, Charith Mendis, Joel S Emer, and Saman Amarasinghe. Waco: Learning workload-aware co-optimization of the format and schedule of a sparse tensor program. In *Proceedings of the 28th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2*, pages 920–934, 2023.

[51] Yue Zhao, Jiajia Li, Chunhua Liao, and Xipeng Shen. 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*, pages 94–108, 2018.

[52] Yue Zhao, Weijie Zhou, Xipeng Shen, and Graham Yiu. Overhead-conscious format selection for spmv-based applications. In *2018 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*, pages 950–959. IEEE, 2018.

[53] Lianmin Zheng, Chengfan Jia, Minmin Sun, Zhao Wu, Cody Hao Yu, Ameer Haj-Ali, Yida Wang, Jun Yang, Danyang Zhuo, Koushik Sen, et al. Ansor: Generating high-performance tensor programs for deep learning. In *14th USENIX Symposium on Operating Systems Design and Implementation (OSDI 20)*, pages 863–879, 2020.

[54] Size Zheng, Yun Liang, Shuo Wang, Renze Chen, and Kaiwen Sheng. *FlexTensor: An Automatic Schedule Exploration and Optimization Framework for Tensor Computation on Heterogeneous System*, page 859–873. Association for Computing Machinery, New York, NY, USA, 2020.

[55] Weijie Zhou, Yue Zhao, Xipeng Shen, and Wang Chen. Enabling runtime spmv format selection through an overhead conscious method. *IEEE Transactions on Parallel and Distributed Systems*, 31(1):80–93, 2019.