



Syno: Structured Synthesis for Neural Operators

Yongqi Zhuo*

Tsinghua University
Beijing, China

zhuoyq21@mails.tsinghua.edu.cn

Chenggang Zhao

Tsinghua University
Beijing, China

zhaocg21@mails.tsinghua.edu.cn

Zhengyuan Su*

Tsinghua University
Beijing, China

su-zy21@mails.tsinghua.edu.cn

Mingyu Gao

Tsinghua University
Beijing, China

Shanghai Artificial Intelligence Lab

Shanghai, China

Shanghai Qi Zhi Institute

Shanghai, China

gaomy@tsinghua.edu.cn

Abstract

The desires for better prediction accuracy and higher execution performance in neural networks never end. Neural architecture search (NAS) and tensor compilers are two popular techniques to optimize these two goals, but they are both limited to composing or optimizing existing manually designed operators rather than coming up with completely new designs. In this work, we explore the less studied direction of neural operator synthesis, which aims to automatically and efficiently discover novel neural operators with better accuracy and/or speed. We develop an end-to-end framework Syno, to realize practical neural operator synthesis. Syno makes use of a novel set of fine-grained primitives defined on tensor dimensions, which ensure various desired properties to ease model training, and also enable expression canonicalization techniques to avoid redundant candidates during search. Syno further adopts a novel guided synthesis flow to obtain valid operators matched with the specified input/output dimension sizes, and leverages efficient stochastic tree search algorithms to quickly explore the design space. We demonstrate that Syno discovers better operators with average speedups of 1.37× to 2.06× on various hardware and compiler choices, while keeping less than 1% accuracy loss even on NAS-optimized models.

CCS Concepts: • **Software and its engineering** → **Search-based software engineering**; *Compilers*; • **Computing methodologies** → Machine learning.

*Both authors contributed equally to the paper.



This work is licensed under a Creative Commons Attribution 4.0 International License.

ASPLOS '25, Rotterdam, Netherlands

© 2025 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-1080-3/2025/03

<https://doi.org/10.1145/3676642.3736118>

Keywords: Program Synthesis; Neural Architecture Search

ACM Reference Format:

Yongqi Zhuo, Zhengyuan Su, Chenggang Zhao, and Mingyu Gao. 2025. Syno: Structured Synthesis for Neural Operators. In *Proceedings of the 30th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 3 (ASPLOS '25)*, March 30–April 3, 2025, Rotterdam, Netherlands. ACM, New York, NY, USA, 18 pages. <https://doi.org/10.1145/3676642.3736118>

1 Introduction

Deep learning with neural networks (NNs) has been a surprisingly effective algorithm breakthrough to handle many challenging tasks in various domains. Since its emergence in the last decade, people have been continuously seeking to improve both the quality (in terms of, e.g., prediction accuracy) and the performance (in terms of training and inference time) of NN models, in order to adapt them to more complicated real-world scenarios with lower computational cost.

Two complementary research paradigms have been developed. To systematically design new NN models with better accuracy quality, neural architecture search (NAS) [10, 22, 46, 47] uses deep learning algorithms themselves to automatically discover promising model structures [30–32, 40]. Given a backbone network topology, NAS explores how to construct the basic cells in the model using combinations of basic operators like convolutions and pooling. The optimization goal is either pure accuracy, or a balance between accuracy and speed [3, 4, 31, 40]. In contrast, to improve training and inference speeds, tensor compilers [6, 26, 39, 43, 44] aim to optimize the implementation of low-level loop nests of each operator in an NN model. Various general and specialized compile-time optimizations are applied to the operator, without altering its functional semantics.

We notice that a new direction orthogonal to the above two has not been fully explored, namely to *synthesize novel neural operators at a fine granularity*, with the goal to automatically and efficiently discover new operators beyond

existing standard types (e.g., convolutions), to improve accuracy quality and/or execution performance. NAS only composes its cell structures using *existing* operators, and tensor compilers only explore *semantically equivalent* variants of the original operators. We envision that these three complementary approaches could be used together. For example, starting from a NAS-discovered network topology, we synthesize novel operators to replace the original ones in the model, and finally leverage tensor compilers to optimize their execution speeds on specific hardware backends.

Neural operator synthesis can be viewed as a domain-specific form of program synthesis [13], a classic topic in computer science. Generic program synthesis approaches face the issue of scalability and do not easily allow different target semantics from the specification. Early attempts of program synthesis for NNs [19, 23] are still limited to composing with coarse-grained basic operators and leave large potentials unexploited. More specifically, to explore a sufficiently large design space, *fine-grained synthesis* that composes directly from the very basic programming language atoms is desired. This is highly challenging. First, the arbitrarily composed candidates would be very unlikely to satisfy the common properties of neural operators, such as differentiability, no replication or discard of tensor elements, etc. Second, there will be enormous redundant synthesized operator candidates with the same or similar semantics. Such equivalent semantics are already efficiently explored by tensor compilers, so we want to prune them out and focus on discovering new operators. Finally, program synthesis has complicated search spaces, and the search for neural operators has the specialized goals of better inference accuracy and performance, which are different from the strict and clear correctness requirement in traditional synthesis. A new search method is thus needed.

We develop an *end-to-end, automatic, and efficient neural operator synthesis framework*, Syno. It takes a given backbone NN topology, and searches for novel linear operators to replace the original ones in the model, in order to improve prediction accuracy and/or execution performance. Syno addresses the aforementioned challenges with several key techniques. First, it makes use of a novel set of *fine-grained primitives* to synthesize new operators. The primitive semantics are defined on tensor coordinates (i.e., dimensions). They maintain tensor semantics and exhibit high-quality properties for neural operators, while not sacrificing expressiveness. Second, Syno leverages *expression simplification and canonicalization* techniques to analyze and eliminate most of the redundancies when synthesizing operators, especially to avoid redoing tensor compiler optimizations. Finally, the design space search process is made more structured, which iteratively samples and adds new primitives to compose operator candidates. We formulate it as a Markov decision process and leverage the efficient *Monte Carlo Tree Search* algorithm [8, 11, 15]. We further propose a novel metric

of *shape distance* to guide the synthesis towards matching with the required input/output tensor shapes, so that the synthesized operator candidate is valid to be used in the backbone model. Two code generators targeting PyTorch [2] and TVM [6] are built for accuracy and speed evaluation of Syno-discovered operators.

Evaluated on five vision models and GPT-2, Syno discovers faster operators than standard convolutions and matrix multiplications, even on NAS-optimized backbone models. Within 1% accuracy loss on CIFAR-100, using Syno-optimized operators exhibits 2.06×, 1.72×, and 1.47× speedups on average on mobile CPUs, mobile GPUs, and server GPUs, respectively, with the TVM backend, compared to the original models. With `torch.compile` [2], the speedups are 1.37×, 1.62×, and 1.60×. On ImageNet, Syno-optimized operators achieve up to 4.73× and 1.94× speedups when compiled with TVM and `torch.compile`, respectively, with 1% to 2% accuracy loss. Syno also accelerates GPT-2 training by 1.1× and improves the language perplexity metric from 111 to 99. We also investigate the discovered operators and find novel and efficient semantics with interesting neural algorithm insights.

2 Background and Related Work

In this section, we briefly introduce the three most related concepts: neural architecture search, tensor compilers, and program synthesis, all of which can be used to better design and implement neural network operators.

2.1 Neural Architecture Search

As neural networks (NNs) are being applied to more and more domains, the needs of designing specific NN models are becoming increasingly prevalent. *Neural architecture search* (NAS) has emerged consequently to automatically design new model structures [10, 22, 46, 47], and indeed, many of the recently proposed models that showed state-of-the-art accuracy levels were discovered by NAS rather than manually crafted [30–32, 40]. NAS typically defines a highly modular search space by dividing the backbone model topology into basic units (called *cells*) of various sizes. It then proceeds to explore how to construct each cell by composing from several types of basic layers (a.k.a., *operators*) like convolutions, matrix multiplications (matmuls), and pooling. Throughout the search, the accuracy levels of the candidate cell structures are continuously evaluated. With such an automated flow, NAS is able to efficiently explore a large design space, and hence discover NN architectures with potentially higher accuracy than manually designed models.

Besides solely focusing on accuracy, *performance-aware* NAS methods aim to strike a better balance between prediction accuracy and execution speed [3, 4, 31, 40]. Specifically, they inherit the design space from traditional NAS while integrating hardware efficiency metrics. By considering factors such as latency alongside accuracy, the search process could

yield model architectures that are not only high-quality but also high-performance¹ on particular hardware.

We emphasize that both traditional and performance-aware NAS methods only *compose existing operators*, such as convolutions and matmuls, in a coarse-grained black-box way. Thus they are limited by these computationally expensive operators. The lack of flexibility to *invent novel operators* leaves ample opportunities for further optimizations, as we will demonstrate in this work.

2.2 Tensor Compilers

At the system level, an NN model is typically represented as a *tensor program*, in which the input/output and intermediate data are all cast as tensors, and a set of operators are applied to them. As a result, *tensor compilers* have gained great attention to accelerate NN execution, by applying general and specialized compile-time optimizations to compile the operators into high-performance *kernels*² [6, 26, 39, 43, 44].

Typically, kernels are written as loop nests, and each tensor compiler has its own intermediate representation (IR) for describing and optimizing kernels. We here take Halide [26] as an example. Many tensor compilers use similar IR designs [6, 17, 34, 43]. Halide provides the separation of algorithm and schedule, where the algorithm is purely functional, and the schedule dictates the concrete loop nest implementation involving tiling, vectorization, reordering, etc. For example, Listing 1 shows how we define a convolution in Halide, which is just simplified loop nests operating on specific *coordinates* (i.e., dimensions) of the tensors. With this IR, we can flexibly express different tensor computations.

```
auto [r_Ci, r_K_H, r_K_W] = RDom(0, C_in, 0, K, 0, K);
out(i_N, i_Co, i_H, i_W) +=
  input(i_N, r_Ci, i_H + r_K_H - K / 2, i_W + r_K_W - K / 2)
  * weight(i_Co, r_Ci, r_K_H, r_K_W);
```

Listing 1. The conv2d operator represented in Halide.

The separation of algorithm and schedule enables tensor compilers to explore the optimization space that is *semantically equivalent* to the original program, i.e., the purely functional computation description as in Listing 1. This is in contrast to NAS-like approaches that find *semantically inequivalent* programs with better quality and/or better performance, so the two approaches are *orthogonal*. A recent work, Turner et al. [37], extended tensor compilers, relaxing the equivalence constraint to apply *inequivalent* transformations on loop nests in tensor programs to realize NAS. However, their approach pre-defined only a few simple inequivalent transformations, such as grouping and bottlenecking the range of a loop, thus only exploring a limited search space still in the scope of traditional operators.

¹Throughout this paper, we use “quality” for model accuracy, and “performance” for execution speed.

²We use *kernel* to represent a concrete implementation of an operator.

2.3 Program Synthesis

Program synthesis is an approach that automatically generates a program that complies with several *specifications*, such as a set of input and output example pairs, or a set of assertions [13]. Theoretically speaking, the general concept of program synthesis can be applied to design new NN operators, but there exist several practical gaps. Traditional program synthesis only treated *correctness* as the target, such as TF-Coder [29] which synthesizes TensorFlow code to help programmers write correct code. But NN models, which are known to tolerate small errors, do not have a clear notion of correctness, while the goal is to improve inference accuracy and/or execution speed. Also, existing program synthesis approaches can hardly scale, currently limited to considering a highly constrained program space. The complexity of loop nests in typical NN operators is well beyond their capabilities. We discuss these challenges in more detail in Section 3.

αNAS [19] relaxed the correctness objective to apply goal-directed program synthesis for NAS. They applied transformations to subgraphs in the model, which could generate new operators beyond traditional NAS. But they are still constrained by traditional operators like convolutions and matmuls, so the potential of intra-operator program synthesis remains unexploited. As a result, their speedups were light, as Section 9 will show. Ma et al. [23], on the other hand, pre-defined some fine-grained primitives common in traditional demosaicking pipelines to perform NAS; however they did not allow freely exploring new operators, either.

3 Motivation and Challenges

In this work, we aim to automatically and efficiently synthesize novel NN operators from the very basic atoms in programming languages, in hope of discovering new operators that have both high accuracy quality and high execution performance. Such automatic *neural operator synthesis* is highly profitable. State-of-the-art models today such as transformers and convolutional networks rely heavily on operators like attention and convolution that are constructed based on human insights. Automatic discovery of such operators can potentially create more promising model architectures.

Comparison with existing paradigms. Neural operator synthesis has a similar goal to performance-aware NAS, but aims to synthesize tensor programs at a much more fine-grained level rather than directly composing known operators like convolutions and matmuls. More specifically, operator synthesis involves writing various loop nests and the tensor expressions in the loop body. For example, for the convolution in Listing 1, the loops are implicitly defined by the iterators (*i_Co*, *r_Ci*, etc.), and the tensor expressions are realized with *coordinate expressions*. Coordinate expressions are key to an operator because they specify how tensor elements are arranged and which are involved in the computation. Here, the simple addition of iterators (*i_H* + *r_K_H*)

implies convolution, and the repeated uses of the reduction iterator (r_Ci) in two tensors imply contraction (a.k.a., tensor multiplication). The rich semantics of coordinate expressions can be exploited to synthesize novel operators. We note that such operator synthesis is impossible under existing NAS. Although it is always possible to lower existing operators to nested loops, it is *not* always possible to do the inverse. If a loop nest cannot be decomposed into several existing operators, it is likely we have discovered a novel operator.

On the other hand, operator synthesis is also significantly different from tensor compilers. Existing tensor compilers mostly preserve semantic equivalence as discussed in [Section 2.2](#). Thus they are unable to discover *new* operators. Actually, in operator synthesis, we deliberately avoid the exploration of semantically equivalent operators (see [Section 6](#)). If we synthesize equivalent operators, we would be very likely to redo existing optimizations in tensor compilers. In this sense, *tensor compilers and operator synthesis are orthogonal*. We first synthesize novel operators, and then leverage tensor compilers to optimize their execution performance on the particular hardware.

We view neural operator synthesis as a specialized form of program synthesis in the NN domain. While traditional synthesis methods are limited to simple programs, we need to handle more complex operators with various nested loop structures and coordinate expressions. On one hand, the degree of freedom in directly writing loop nests and tensor expressions is huge, leading to an extremely large search space. On the other hand, as in performance-aware NAS, for each candidate operator, we need to assess both its accuracy level and execution speed, both of which require substantial time. To measure the accuracy, we have to use real datasets to train the full NN model for several epochs at least. Several theoretical metrics are proposed to predict the accuracy potential with minimum training cost [[1](#), [21](#), [38](#), [45](#)], but we find them to perform poorly in reality, especially for irregular operators we aim to construct. To evaluate the speed, we need to generate an optimized implementation of the operator on real hardware. This could also cost significant time in state-of-the-art tensor compilers [[6](#), [26](#), [44](#)].

Challenges. We highlight three main challenges in neural operator synthesis that distinguish it from traditional program synthesis. First, with traditional program synthesis, the loop nest (e.g., [Listing 1](#)) can be enumerated with bottom-up search, building the coordinate expressions from the atoms such as iterators and constants [[13](#)]. The main issue with this generic approach is the difficulty of ensuring *high quality* for NN operators, due to the lack of high-level semantics. For example, if we fill the indices of input with all 0s, all the other elements would be discarded, which is not at all reasonable. An NN operator is usually expected to satisfy certain properties, such as differentiability, full utilization of input data elements, etc., so that it can be trained in an NN model and achieve good accuracy. Encoding such constraints

as inputs to an SMT solver may be possible, but would be too slow when searching over many operator candidates.

Second, a major issue of exploring the search space is *redundant operators*, which exhibit the same or *similar* semantics and consequently show similar quality and performance. For example, in integer arithmetic, there is an identity $(B*i)\%(B*C)=B*(i\%C)$. Our synthesis needs to skip these equivalent coordinate expressions. Moreover, even inequivalent expressions can induce similar computations: considering iterators i, j with domains B, K where $B > C \gg K$, then $(i+j)/C=i/C$ holds for almost every point. Traditionally, the redundancy can be handled with term rewrite systems [[24](#)] and equality saturation [[35](#)], but this slows down the search, and cannot prune away inequivalent but similar expressions.

Third, conventional program synthesis has developed multiple approaches to guide the synthesis with user-provided specifications to eliminate illegal candidates [[13](#)]. With neural operators, the only correctness constraint is that the input and output tensor shapes must match with those specified by the model. However, under the aforementioned quality constraints, the domains of coordinate expressions cannot be freely altered to match the input and output shapes, making randomly sampled operators almost always illegal in terms of tensor shapes. Thus, we need a specially designed novel approach to *guide the synthesis process*.

In summary, to realize practical neural operator synthesis, we must design a framework with the following properties.

- **High quality.** Synthesized operators need to satisfy certain properties (e.g., differentiability, full data utilization), similar to existing NN operators.
- **No redundancy.** Repeated evaluation of operators with the same or similar semantics should be avoided. Particularly, we should not redo the optimizations in existing tensor compilers.
- **Guided search.** The synthesis process should be guided by the input and output tensor shapes to improve the search efficiency.

4 Design Overview

We propose Syno, an end-to-end, automatic, and efficient framework for neural operator synthesis. Given a backbone NN model, Syno is able to synthesize novel linear operators with high quality (for accuracy) and high performance (for speed), which can be drop-in replacements for the original operators (convolution, matmul, etc.) with the same input and output tensor shapes. The model topology and the non-linear activation layers are unaltered. Specifically, given the input and output tensor shapes, e.g., $[N, C_{in}, H, W]$ and $[N, C_{out}, H, W]$ for convolution, or $[M, K]$ and $[M, N]$ for matmul, Syno discovers novel operators that satisfy the accuracy and performance requirements, e.g., best performance with less than 1% accuracy loss. Note that the tensor shapes are specified as symbolic variables to allow one operator to

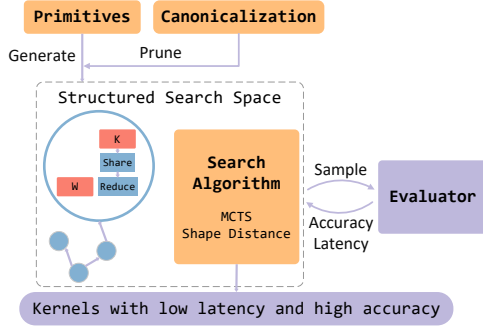


Figure 1. The overall architecture of Syno.

Algorithm 1 The workflow of Syno.

```

1: procedure SEARCH(model,  $d_{max}$ )
2:   operators  $\leftarrow$  EXTRACTOPERATORS(model)
3:   substs  $\leftarrow$  SYNTHESIZESUBSTITUTIONS(operators,  $d_{max}$ )
4:   for all subst  $\in$  MCTS(substs) do
5:     model'  $\leftarrow$  REPLACE(model, operators, subst)
6:     accuracy  $\leftarrow$  TRAINWITHPYTORCH(model')
7:     if ISWITHINACCURACYMARGIN(accuracy) then
8:       performance  $\leftarrow$  TUNEWITHTVM(model')
9:       output subst, accuracy, performance

10: function SYNTHESIZESUBSTITUTIONS(operators,  $d_{max}$ )
11:    $\triangleright$  Search on symbolic shapes, e.g., [N, C, H, W].
12:   input, output  $\leftarrow$  SYMBOLICSHAPE(operators)
13:   results  $\leftarrow$  {}
14:   procedure ENUMERATE(d, n)
15:     if HASMATCHINGSHAPE(n, input) then
16:       Add n to results if within budgets
17:     if  $d \geq d_{max}$  then return
18:     for all  $n' \in$  ENUMERATECHILDREN(n) do
19:        $\triangleright$  Backtrack with shape distance.
20:       if SHAPEDISTANCE( $n'$ , input)  $> d_{max} - d - 1$  then
21:         continue
22:       ENUMERATE( $d + 1$ ,  $n'$ )
23:   ENUMERATE(0, ROOTNODE(output))
24:   return results

25: function ENUMERATECHILDREN(n)
26:   children  $\leftarrow$  {}
27:   for all prim  $\in$  ENUMERATEPRIMITIVES(n) do
28:     if ISCANONICAL(n, prim) then  $\triangleright$  Canonicalize.
29:       children  $\leftarrow$  children  $\cup$  {ADD(n, prim)}
30:   return children

```

fulfill different tensor sizes. The framework also supports a rich set of user-defined budgets such as FLOPs, memory usage, and number of parameters.

We limit our search in Syno to linear operators. First, linear operators are usually the performance bottleneck in NNs, constituting most of the computations, so reducing their complexity can have great gains. Second, activation layers

like ReLU provide the non-linearity needed in NNs, so we keep them unaltered in the backbone model. Their performance impact is negligible because they can be readily fused into their preceding operators by existing tensor compilers.

Algorithm 1 outlines the overall workflow of Syno, which is also illustrated in Figure 1. Syno relies on a library of *fine-grained primitives* that operate on specific tensor *coordinates*, i.e., dimensions (Section 5). A new operator candidate is synthesized by iterative sampling and adding new primitives (Algorithm 1 Lines 27 to 29), until reaching a maximum size (Line 17). Compared to directly composing arbitrary coordinate expressions, using these primitives ensures high quality with tensor semantics and enables efficient structured search, while not sacrificing expressiveness.

Since our primitives are defined on tensor coordinates, we can directly apply *expression simplification and canonicalization* techniques for coordinate expressions to quickly eliminate redundant candidates (Algorithm 1 Line 28), enabling efficient search space exploration. Our canonicalization rules not only remove most of the equivalent operators during the search, but can also prune those operators with similar semantics (Section 6).

To efficiently discover valid operators, we guide the synthesis flow (Algorithm 1 Line 20) using a novel metric of *shape distance*, which is the distance between the current partial operator and a complete operator that has the same input/output shapes as the one in the original model. We then leverage the intrinsic structure of the search space and formalize the search as a stochastic decision process, in order to apply the Monte Carlo Tree Search algorithm (Section 7). The discovered operators are then fed to the two code generators targeting PyTorch [2] and TVM [6] for accuracy and performance evaluations (Section 8).

Syno is implemented as a distributed infrastructure, which could leverage multiple GPUs across several server nodes to conduct search, parallelizing the model training required in the accuracy evaluation. Syno has 19K lines of C++ code and 11.5K lines of Python code. Syno is open sourced at <https://github.com/tsinghua-ideal/Syno>.

5 Primitives

Syno adopts a novel approach to synthesize candidate operators from a set of fine-grained primitives, whose semantics are defined with *tensor coordinate expressions* in a bottom-up way, as shown in Table 1. Compared with directly enumerating arbitrary raw arithmetic expressions as abstract syntax trees (ASTs) of integer expressions, this allows us to perform synthesis and search with the primitives in a more structured manner to ensure high quality.

5.1 Structured ASTs

Synthesizing expressions in a bottom-up way, i.e., first specifying the innermost atoms and then composing them, is

Table 1. Syno primitives that transform coordinate expressions and their domains in a bottom-up way. For example, UNFOLD combines two coordinates with domains N and K , and obtains an expression of domain N (with out-of-bound elements clipped).

Class	Primitive	Parameter	Bottom	Top	Top-Down Semantics
Views	1-to-1	SPLIT	-	$[i, j]: [G, B] \leftarrow [B * i + j]: [G * B]$	Partition into blocks
		MERGE	B	$[i / B, i \% B]: [N / B, B] \leftarrow [i]: [N]$	Flatten two dimensions
		SHIFT	-	$[(i + 1) \% N]: [N] \leftarrow [i]: [N]$	Shift along a dimension
	1-to-many	EXPAND	-	$[i]: [C] \leftarrow []: []$	Repeat or up-sample
		UNFOLD	-	$[i + j - K/2]: [N] \leftarrow [i, j]: [N, K]$	Extract sliding windows
	many-to-1	STRIDE	S	$[S * i]: [S * K] \leftarrow [i]: [K]$	Strided access
Contractions	REDUCE	N	$[]: []$	$\sum_i [i]: [N] \leftarrow [i]: [N]$	Reduce a dimension
	SHARE	-	$[i]: [N]$	$([i], [i]): ([N], [N]) \leftarrow [i]: [N]$	Element-wise product

common in program synthesis [13]. This is also a natural choice for NN operators. As can be seen in Listing 1, each element of the output tensor is calculated through certain arithmetic operations on some elements of the input tensors, which are indexed by some expressions on the indices of the output element. Here the output tensor indices, e.g., i_H , are termed the *output iterators*. They are also the implicitly defined loops in Halide. The expressions consisting of output iterators and constants are termed *coordinate expressions*. They are used to index the input and output tensors. For example, i_H is a coordinate expression to index out, and $i_H + r_{K_H} - K/2$ is also a coordinate expression to index input. Following the bottom-up approach, we use the output iterators as atom coordinate expressions (the “bottom”), and enumerate over the diverse combinations of coordinate expressions for the input tensor indices (the “top”). By doing so we can synthesize novel operators beyond our current knowledge, and this is the design space we hope to explore.

However, operators synthesized with such straightforward bottom-up enumeration tend to have *low quality*. For example in Listing 1, if i_{Co} were only used in an expression $i_{Co} / 2$, then every two consecutive channels of out would have identical feature maps. This means tensor elements are replicated, and we perform redundant computations. To avoid this, we can require that $i_{Co} \% 2$ must also be present in the enumerated coordinate expressions. This example inspires us to design a *high-quality* primitive that transforms a coordinate expression $[i]$ with domain $[N]$ to two coordinate expressions $[i / B, i \% B]$ of domains $[N / B, B]$ where B divides N . Formally we write: $[i]: [N] \leftarrow [i / B, i \% B]: [N / B, B]$. The notation here uses an inverse arrow to point from the “top” to the “bottom”, in order to highlight the dataflow direction from the input tensors to the output.

Furthermore, this *bottom-up primitive* also has *top-down semantics*, namely to flatten a tensor of shape $[N / B, B]$ into $[N]$ by merging the two dimensions. We name it as MERGE, which is actually a common *tensor view operation*. A

view is just another way of accessing a tensor. Various arithmetic operations ($+$, $*$, $/$, etc.) on tensor coordinates actually correspond to views. For example, the addition of coordinate expressions is equivalent to extracting neighboring elements, which is UNFOLD. Similarly, adding a constant is SHIFT, multiplication is SPLIT and STRIDE, and discarding an expression is EXPAND. We summarize them in the class of views in Table 1. All of them do not discard or replicate elements and thus have high quality, except EXPAND and STRIDE, which could be useful for special cases such as up-sampling and dilated convolution. For semantic completeness, we keep them but limit their occurrences in each synthesized operator.

Aside from coordinate expressions that extract elements from tensors, we need primitives to actually perform computations. For now, we only support linear operations in Syno, so elements from multiple tensors are multiplied and summed up. The reduction (RDom in Listing 1) can be abstracted to a primitive REDUCE, which adds a sum reduction loop. Meanwhile, a SHARE primitive indexes two tensors with the same coordinate expression, and performs multiplication between the two tensors. The top-down semantics of REDUCE and SHARE are mainly *tensor contraction operations*. A contraction involves combining two tensors along a certain dimension [12]; e.g., the input channels of input and weight tensors are contracted in Listing 1.

With this approach, we propose a structured way of using the Syno primitives to build coordinate expression ASTs for neural operators. An operator composed in this way has a very similar structure to common ASTs, except that instead of trees, expressions are now determined by directed acyclic *primitive graphs* (pGraphs). Figure 2 shows how to compose a 2D convolution of Listing 1 using the Syno primitives. The vertices are the primitives, while the edges are (possibly intermediate) coordinate expressions, and can be evaluated in the same way as we evaluate ASTs.

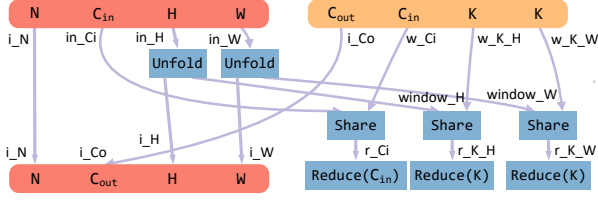


Figure 2. The pGraph of conv2d in Syno. Each edge is a (sub-)coordinate expression. The bottommost orange box is the output tensor of the operator, which comprises the innermost atom coordinate expressions, e.g., i_H : H . The blue boxes are Syno primitives, each of which transforms the coordinate expressions corresponding to its *out edges* to those corresponding to its *in edges* as specified in Table 1. The topmost orange box is the input tensor to the operator, which comprises the full coordinate expressions. The yellow box is the weight tensor.

5.2 Advantages

The structured bottom-up primitives in Syno present several advantages. First, they ensure *high quality* of synthesized operators, in that they are differentiable [16] and do not discard input data or replicate data. The only exceptions EXPAND and STRIDE are restrictively used and STRIDE is required to be paired with 1-to-many primitives to ensure the high-quality property. Second, they allow *structured search*. With the primitives, similar pGraphs are likely to share a subgraph, which makes the search space highly structured and enables the use of effective search algorithms (Section 7.2). Third, the primitives are *expressive*, as they are devised based on the most basic arithmetic operations on coordinate expressions.

5.3 Semantics and Examples

To construct an operator from a pGraph like Figure 2, we evaluate the expressions bottom-up using Table 1, and use them to index the input tensors. To better illustrate the semantics, we provide examples for several operators: matrix multiplication, pooling, and pixel shuffle in Table 2, in addition to the convolution example in Figure 2 and Listing 1.

For example, to obtain a PyTorch operator `torch.mm` for matrix multiplication, Syno starts with the *bottom* coordinate expressions that index the output tensor, i.e., $[i, j]$: $[M, N]$ for `mm`. It gradually applies the primitives REDUCE(K) and SHARE, to compose a valid pGraph. The *top* coordinate expressions can be used to directly index the input tensors `input(i, r_K)` and `weight(r_K, j)`. More specifically, after REDUCE(K) is applied to introduce a reduction, SHARE is applied to assign one r_K : K to index the input, and one r_K : K to index the weight. A subtle detail here is that, without further restriction, i : M and j : N can be used to index either the input or weight tensor, but here we want exactly j : N to index the weight. So an implicit Match step is done along with SHARE to assign j : N to the weight tensor.

It tracks all the coordinate expressions to be assigned to the new weight tensor created by a SHARE, and is always applied right after the SHARE. Thus we treat Match as an implementation detail of SHARE, and do not place too much emphasis on it. The other two example operators and the convolution in Figure 2 are similar.

5.4 Design Details

To match operators with different concrete input/output tensor shapes, and to support additional parameter variables in some primitives (e.g., MERGE needs a factor B), Syno uses *symbolic shapes* when synthesizing operators. We further split the symbols into two classes. *Primary variables* are for input/output dimensions, e.g., C_{out} , H . They are relatively large and thus are not allowed to appear in the denominator of a coordinate expression. *Coefficient variables* are only introduced by primitives, and are relatively small and allowed to appear in denominators. When enumerating the applicable primitives on a partial pGraph, the primitive parameters are represented by monomials of primary variables and coefficient variables, with the degrees (i.e., powers) limited within a user-specified range. Syno replaces the variables with concrete sizes at code generation.

In the current prototype of Syno, we only consider operators that process a single input tensor (not including weights) and produce a single output tensor, and disallow multiple uses of the same (input or intermediate) tensors such as residual links [14]. This restriction seems strict, but in fact existing operator types like convolution, matmul, and pooling all satisfy it. We argue that the lost flexibility is usually more critical at the full model graph level rather than at the operator level. Our operators can still be plugged into arbitrary model topologies including ResNet [14], where the residual links are realized outside the operators. We plan to extend Syno to support multiple input tensors in the future.

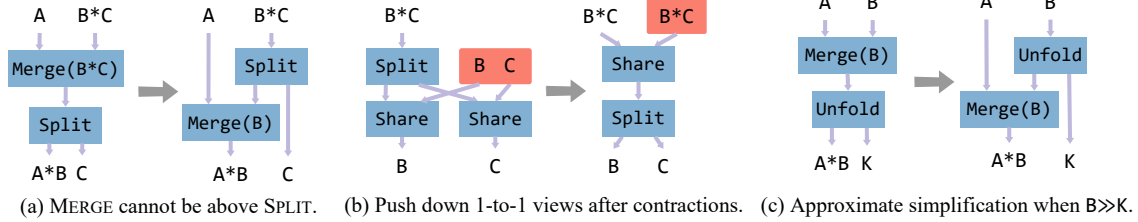
6 Canonicalization

The design space of synthesizing operators from our primitives is extremely large, with a lot of redundant operator constructs, especially those that can be readily discovered by tensor compilers. Take the partial pGraph in Figure 3(a) as an example. On the left side, the topmost coordinate expressions are given by $[i, j]$: $[A*B, C] \xleftarrow{\text{MERGE}(B*C)} [C*i+j]: [A*B*C] \xleftarrow{\text{SPLIT}} [(C*i+j)/(B*C), (C*i+j)\%(B*C)]: [A, B*C]$. However, this simplifies to $[i/B, C*(i\%B)+j]$, corresponding to the right side $[i, j]: [A*B, C] \xleftarrow{\text{MERGE}(B)} [i/B, i\%B, j]: [A, B, C] \xleftarrow{\text{SPLIT}} [i/B, C*(i\%B)+j]: [A, B*C]$.

To improve search efficiency, Syno uses a set of *canonicalization rules* to filter out uncanonical redundant candidates on the fly when new primitives are added to partial pGraphs (IsCANONICAL in Algorithm 1 Line 28). Syno does

Table 2. Example operators that can be composed with Syno primitives.

PyTorch Operator	Constituent Primitives	pGraph	Halide Code
<code>mm(input, weight)</code>	$[i, j] : [M, N]$ $\xleftarrow{\text{REDUCE}(K)}$ $[i, j, r_K] : [M, N, K]$ $\xleftarrow{\text{SHARE}}$ $\xleftarrow{\text{(Match)}}$ $([i, r_K], [r_K, j]) : ([M, K], [K, N])$		<pre>auto [r_K] = RDom(0, K); out(i, j) += input(i, r_K) * weight(r_K, j);</pre>
<code>nn.AvgPool1d(s)(input)</code>	$[i] : [s^{-1} * H]$ $\xleftarrow{\text{REDUCE}(s)}$ $[i, r_s] : [s^{-1} * H, s]$ $\xleftarrow{\text{SPLIT}}$ $[s * i + r_s] : [H]$		<pre>auto [r_s] = RDom(0, s); out(i) += input(s*i+r_s);</pre>
<code>nn.PixelShuffle(B)(input)</code>	$[i] : [H]$ $\xleftarrow{\text{MERGE}(B)}$ $[i/B, i\%B] : [B^{-1} * H, B]$ $\xleftarrow{\text{SPLIT}}$ $[(H/B) * (i\%B) + i/B] : [H]$		<pre>out(i) = input((H/B)*(i%B)+i/B);</pre>

**Figure 3.** Examples of some canonicalization rules used in Syno.

not aim to eliminate *all* redundancies, which is highly challenging, if not impossible, considering the rich primitive semantics. Also, comprehensive canonicalization checks are extremely expensive and sometimes undecidable [24]. On the other hand, semantically similar models have similar quality and performance, so Syno supports canonicalization rules to mark only one operator as canonical among the many ones in the class that have *similar computation results*. We also note that the canonicalization rules in Syno are easily *extensible*. Developers can define new rules and plug them into the framework.

Contractions. Since weight tensors can be arbitrarily reshaped offline, there is no need to apply views to weights. Thus weight coordinates are directly used in SHARES for contractions³. Moreover, we always put weight coordinates as the right-hand-side inputs of the symmetric SHARES.

Between views and contractions. We enforce a canonical order between views and contractions. The 1-to-1 views do not involve actual computations so they can be freely

swapped with contractions. We thus *push down* all 1-to-1 views after contractions, as in Figure 3(b).

For the others, we apply rules to avoid doing futile work. For example, we disallow combining EXPAND and REDUCE because this only changes a multiplier of the result; UNFOLD allows at most one output coordinate to be REDUCED.

Between two views. Most redundancies exist between views because of the many primitive types. The key to their canonicalization is *to apply expression simplification techniques*. In many tensor compilers such as Halide [26] and TVM [6], expressions are simplified before analysis and lowering. In Syno, as our primitive definitions are based on coordinate expressions, we can similarly simplify the coordinate expressions corresponding to the (sub)pGraph consisting of view primitives, and the fully simplified expression gives the canonical form. For example, the aforementioned Figure 3(a) is one such example, where the right side is simplified and thus canonical, corresponding to the rule that a MERGE cannot be above a SPLIT.

We design expression simplification in Syno by referring to Halide’s term rewrite system (TRS) [24]. TRS sequentially substitutes the terms in an AST from bottom to top with pattern matching in expressions, in order to obtain the simplified form [24]. In Syno, rather than actually rewriting

³This also implies that the coordinate expressions assigned to weights by the Match step mentioned in Section 5.3 can be removed from the set of coordinate expressions that can be further transformed and matched against the desired topmost shape, simplifying the synthesis algorithm in Section 7.1.

the pGraph, canonicalization is applied on-the-fly when we add new primitives to the partial pGraph, by discarding candidates that create uncanonical forms (so we need not worry about the termination of rewrites). In our pGraph, each coordinate (edge) is like an AST node. We treat the bottom outputs of a subgraph as wildcards and match the top input expressions against the patterns. Again look at Figure 3(a). The bottom outputs are marked as $[\#0, \#1]$. Then the substitution can be formulated as $[(C*\#0+\#1)/(B*C), (C*\#0+\#1)\%(B*C)] \rightarrow [\#0/B, C*(\#0\%B)+\#1]$, which is a pattern-matching-based rewrite rule. To choose the canonical (i.e., “simplest”) form among equivalent expressions, we empirically define simplicity as removing parentheses as much as possible by applying distribution laws of multiplication, division, and modulo. We can see Figure 3(a) removes one level of parentheses. Following this approach, we derive a series of rewriting rules involving multiple primitives.

In addition, it is better to not just canonicalize semantically equivalent subgraphs, but also eliminate candidates with only slightly different semantics, so that a wider range of semantics can be explored with fewer samples. In Figure 3(c), on the left side, with $[i, j]: [A*B, K]$ as the output coordinates, the inputs are $[(i+j-K/2)/B, (i+j-K/2)\%B]: [A, B]$. If B is much larger than K as in most convolutions⁴, then $j-K/2$ is much less than B . So we can simplify the expressions to $[i/B+j-K/2, i\%B+j-K/2]$ as the right side, which is equal to the left side at almost every point. Other similar rules are devised based on the principle of removing parentheses, and these approximately equivalent rules can also be implemented as TRS-based rules as above. They effectively enable us to only synthesize operators that are significantly distinct.

7 Guided Search

We next describe the overall synthesis and search process in Syno. Section 7.1 discusses the bottom-up synthesis approach. A critical challenge is how to ensure the exact match of the input/output tensor dimensions with the given specification. We propose a novel concept of shape distance to guide the search. Section 7.2 explains our specific search algorithm based on Monte Carlo Tree Search (MCTS) [8, 11, 15].

7.1 Bottom-Up Synthesis with Shape Distance

As mentioned in Section 5, Syno performs bottom-up synthesis, starting from the output coordinates and iteratively applying sampled primitives for a limited number of steps. For example, as a subgraph of Figure 2, from the output $[i_H]$:

$$[H], \text{ we can get } [i_H]: [H] \xleftarrow{\text{REDUCE}(K)} [i_H, r_{K_H}]: [H, K] \xleftarrow{\text{SHARE}} ([i_H, r_{K_H}], [r_{K_H}]): ([H, K], [K]) \xleftarrow{\text{UNFOLD}}$$

⁴While we are using symbolic shapes during synthesis (Section 5.4), we also extract all possible concrete values for each symbolic shape from the input backbone NN model. Symbolic $B \gg K$ is true if for every valuation of B and K we have $B \gg K$.

$([i_H + r_{K_H} - K / 2], [r_{K_H}]): ([H], [K])$. The weight $[K]$ here) does not need to be transformed further (Section 6), so we use the term *shape* to refer to the shape of the first tensor (data input tensor), which is $[H]$ in this case.

The data tensor shape of a complete pGraph should match exactly with the *desired shape* (the shape of *input* in Algorithm 1 Line 12). While synthesizing with primitives ensures high quality, it also becomes hard to control the *dimensions* (sizes of tensor coordinates) after applying primitives on a partial pGraph. Ideally, after flexibly exploring various primitives, when the partial pGraph gets close to its maximum size limit, the last few primitives need to move towards exactly matching with the desired dimensions. For example, if the shape of the current partial pGraph is $[C_{in}, s^{-1}*H, s*W, k]$, we can apply $[C_{in}, s^{-1}*H, s*W, k] \xleftarrow{\text{MERGE}} [C_{in}, s^{-1}*H, s, W, k] \xleftarrow{\text{SPLIT}(s)} [C_{in}, H, W, k] \xleftarrow{\text{UNFOLD}} [C_{in}, H, W]$.

We propose a novel metric named *shape distance* as the minimum number of required primitives added onto the current pGraph to reach the desired shape. In the above example, the shape distance of $[C_{in}, s^{-1}*H, s*W, k]$ is 3. If the remaining allowed number of primitives is less than the shape distance, we can immediately terminate the current pGraph and backtrack (Algorithm 1 Line 20). This avoids deviating too far from the desired dimensions.

We design a systematic method in Syno to compute the shape distance between the current shape and the *desired shape*. We first divide the dimensions in the two shapes into *reshape groups*, where future primitives are only applied to the dimensions within each group to match them, but not across groups. In the above example, we can have three reshape groups, as $\{C_{in}\} \leftarrow \{C_{in}\}, \{s^{-1}*H, s*W\} \leftarrow \{H, W\}, \{k\} \leftarrow \{k\}$. Reshape groups can be decided by comparing the primary variables in the coordinate expressions. When there exist multiple possible grouping schemes (but usually only a few), we enumerate all and find the least distance.

We then compute the distance within each reshape group. We identify the *helpful primitives* that will help in shape matching: reshape primitives (i.e., MERGE, SPLIT) which regroup dimensions, and 1-to-many primitives (i.e. UNFOLD, EXPAND) which eliminate dimensions. If the left-hand side and the right-hand side of the reshape group have the same size of domains, e.g., $\{s^{-1}*H, s*W\}$ and $\{H, W\}$ have domains of $H*W$, then we only need to regroup dimensions, using SPLIT and MERGE. In this case we only need 2 steps:

$[s^{-1}*H, s*W] \xleftarrow{\text{MERGE}} [s^{-1}*H, s, W] \xleftarrow{\text{SPLIT}(s)} [H, W]$. We can prove a generalized conclusion of $\#lhs + \#rhs - 2$ steps, where $\#lhs$ and $\#rhs$ are the numbers of dimensions in the left-hand and right-hand sides of the reshapes (both are 2 in this example). On the other hand, if the two sides have different sizes of domains, then at least one 1-to-many primitive is required, counting as 1 extra step. We sum up the bounds $(\#lhs + \#rhs - 2)$ of all the reshape groups, adding 1 if the

domain of the current shape is different from the desired shape, and use it as an upper bound for shape distance. Then, all grouping schemes are enumerated to find the minimum of the upper bounds, which yields the final shape distance.

When the desired shape involves repeated dimensions, e.g., $[C_{in}, H, H]$ for square images, we enumerate all possible permutations, allowing tensor transpose during the final matching.

7.2 MCTS-Based Search

Our search algorithm is based on MCTS [8, 11, 15]. We formulate our search problem as a Markov decision process, where we transit from one partial pGraph to another in the search space, with the action space being the primitives. The final states are complete pGraphs. The optimization goal is operators with both high accuracy and high inference speed. As the FLOPs of operators are much easier to compute than the inference accuracy which requires extensive training, we set a hard upper limit for FLOPs and use accuracy as the reward for MCTS to guide it to learn how to find expressive operators within a given FLOPs budget. We record all MCTS samples and filter out operators with bad accuracies to obtain the final result.

8 Code Generation

We implement two code generators for accuracy and speed evaluations. First, a *PyTorch code generator* is built to make use of the already highly-tuned operator libraries for training. Using the top-down semantics, each view primitive is lowered to its counterpart in PyTorch, and each contraction primitive is lowered to an einsum [27] expression, which is a general method for performing tensor contractions. The primitives are lowered in topological order to ensure that dependencies are satisfied. We further use TorchInductor [2] for compile-time optimizations such as fusion and tiling.

However, PyTorch and TorchInductor are mainly optimized for existing workloads and tuned on a limited set of operators such as convolution and matrix multiplication. To better support the novel operators discovered by Syno, we further build a *TVM TE (Tensor Expression) code generator*, to utilize the more general-purpose compiler, TVM [6]. It follows the bottom-up semantics to evaluate all coordinate expressions according to the pGraph, and leverages TVM for extensive compiler optimizations on specific hardware, e.g., our mobile CPUs and GPUs in Section 9. The TVM TE syntax is very close to that of Halide as we mentioned earlier, so we skip the technical details.

Some optimization passes unique to Syno are designed. An important one aims to automatically insert intermediate stages (materializations) to eliminate redundant computations. Consider the example in Figure 4. A trivial code generator creates a loop nest of $(H/s) * k * s$ iterations computing $Y[i] = \sum_{i_k} \sum_{i_s} X[i + i_k - k/2 + s * i_s]$ as on the left side.

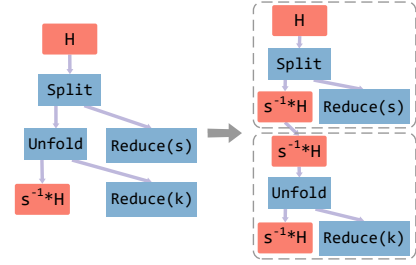


Figure 4. An example of the materialized reduction optimization in Syno.

But this is mathematically equivalent to $Z[i'] = \sum_{i_s} X[i' + s * i_s]$, $Y[i] = \sum_{i_k} Z[i + i_k - k/2]$, which corresponds to the partitioned subgraphs on the right. By doing so we reduce the FLOPs from $k * H$ to $(1 + k/s) * H$.

Generally speaking, the FLOPs depend only on the output iterators and the REDUCES, which are the spatial loops and reduction loops in the loop nest. The number of iterations is their product. In the case of 1-to-many primitives like UNFOLD, the output dimensions are increased, so if we perform any REDUCE after this, FLOPs are unnecessarily increased because we are evaluating k copies for each element.

This issue is unique to the Syno IR. To deal with it, we propose an optimization named *materialized reduction*, which materializes the bottom (output tensor) of a sub-pGraph that performs reductions. We enumerate the order of performing reductions, i.e., the order of lowering each REDUCE. If a REDUCE is lowered, only the primitives that can reach that REDUCE are required to be lowered. In the example, the SPLIT and the UNFOLD primitives can both reach the bottom REDUCE, but only the 1-to-many UNFOLD cannot reach the upper REDUCE. So the upper REDUCE is prioritized to form a sub-pGraph, materializing the bottom.

9 Evaluation

9.1 Experimental Setups

Hardware configurations. Our operator search and accuracy validation are done on a cluster with NVIDIA A100 GPUs. For performance in edge-device inference scenarios, we test the end-to-end latency on NVIDIA Jetson Orin Nano 8 GB, which features a 6-core Arm Cortex-A78AE mobile CPU and a 1024-core NVIDIA Ampere GPU with 32 tensor cores. For performance on server-grade GPUs, we evaluate the end-to-end latency on an NVIDIA A100 GPU. In summary, we evaluate performance on three platforms: (1) mobile CPU, (2) mobile GPU, and (3) A100.

Compilers. To demonstrate the orthogonality of Syno to tensor compilers and its wide applicability, we evaluate on two compilers: (1) TVM MetaSchedule [28], a state-of-the-art tuning-based tensor compiler widely adopted by the research community; (2) TorchInductor, the default torch.compile

backend of PyTorch 2 [2], widely adopted by the industry, with its max-autotune mode enabled.

Workloads. We mainly focus on vision tasks with five popular vision NNs: ResNet-18 [14], ResNet-34 [14], DenseNet-121 [18], ResNeXt-29-2x64D [42], and EfficientNet-V2-S [33]. We aim to substitute all standard convolutions in them. To prove the wide adaptability of Syno, we also test GPT-2 [25] (117M parameters with 12 layers, 12 heads, and 768 embedding dimensions) by substituting its QKV projections.

Baselines. We use three baselines for the comparison of vision tasks. The main baseline is the original models with standard convolutions. We target the latency-accuracy trade-off, so we expect to reduce the end-to-end latency at the cost of minor accuracy degradation. Turner et al. [37] (labeled as NAS-PTE) are the first to introduce loop-level transformations into NAS, and α NAS [19] is the first attempt to apply program synthesis for NAS albeit at a coarse granularity. Because the search and tuning methods of NAS-PTE are not open-source, we compare with their operators on individual layers instead of full models. α NAS is neither open-source nor provides inference performance data, so we only compare against their FLOPs and training speedups reported in the original paper. For GPT-2, we evaluate the training speed relative to the original model.

Datasets and training configurations. ImageNet [9] is unsuitable for direct search because of its large size, so we use the smaller yet still challenging CIFAR-100 [20] as the proxy dataset. Specifically, during the search Syno trains the NN model using each candidate operator for 100 epochs on CIFAR-100. The selected best operators are then fully trained on ImageNet for 90 epochs for accuracy and performance evaluations. We scale the CIFAR-100 images to the same size as ImageNet to ensure the same inference performance. For GPT-2, we employ the language perplexity (PPL) metric on the lm1b benchmark [5]. The data type for both training and inference is FP32. The training hyperparameters for the optimizer and learning rate scheduler are dataset-dependent to ensure reasonable accuracy, but they are not heavily tuned.

Computation cost. Training a model on CIFAR-100 for 100 epochs takes two to three hours. In our experiments, we terminate early when the accuracy is not as high as expected, thereby reducing the average evaluation computation cost to 0.1 GPU hours per sample. We spend roughly 300 GPU hours per model.

9.2 Results on Vision Tasks

For vision tasks, we search for the fastest operators in each model with less than 1% accuracy loss, a commonly used threshold. We separately target both CPUs and GPUs.

CIFAR-100 results. Figure 5 shows the best operators we find in terms of inference latency within the accuracy loss limit. On the mobile CPU, the mobile GPU, and A100, respectively, Syno achieves 2.06 \times , 1.72 \times , and 1.47 \times end-to-end inference speedups over the original models on average

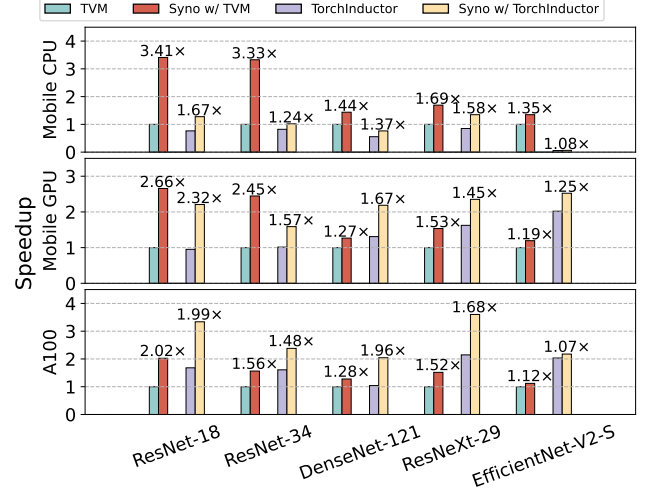


Figure 5. End-to-end performance speedup of Syno on CIFAR-100. The bars of each model are normalized to TVM for direct comparison across different compilers.

(geomean) when compiled with TVM, and 1.37 \times , 1.62 \times , and 1.60 \times when compiled with TorchInductor. Syno performs better on traditional NNs like ResNet-18 with the discovered novel operators. Even for NAS-optimized models such as EfficientNet-V2, Syno can still achieve a performance gain up to 1.35 \times . We perform more detailed analysis on the benefits of our newly discovered operators later.

It is interesting to compare the two compiler backends, TVM and TorchInductor. We find that for the FP32 data type we use, TVM cannot make use of the tensor cores (which requires TF32), so it is generally slower than TorchInductor on GPUs. However, TVM tunes over a much larger search space and performs code generation for every operator, whereas TorchInductor can only select from several templates and would conservatively fall back to PyTorch ATen kernels for small GPUs or if the few templates are too slow. Therefore TorchInductor yields more unstable performance than TVM. For instance in Figure 5, TorchInductor performs poorly on EfficientNet-V2-S when using the mobile CPU. Profiling indicates that TorchInductor falls back to use ATen grouped convolution in most cases, which has terrible performance for the many depth-wise convolutions in this model.

ImageNet results. For every model, we select some discovered operators that have comparable accuracy with the baseline and re-evaluate them on ImageNet. We plot the Pareto optimal curves of accuracy vs. inference time in Figure 6. Most of our operators exhibit a minor 1% to 2% accuracy loss, while enabling up to 4.73 \times and 1.94 \times speedups when compiled with TVM and TorchInductor, respectively. If more accuracy loss is acceptable, then going along the Pareto curves further boosts performance.

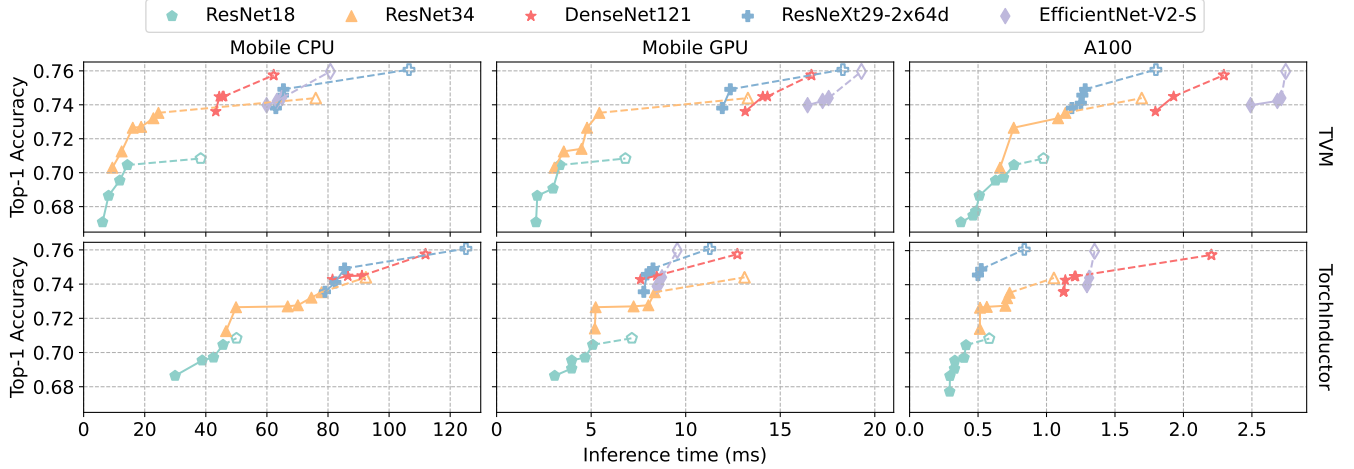


Figure 6. Pareto optimal curves of accuracy vs. inference time between the original and the Syno-optimized models on ImageNet. For each model, the hollow point is the baseline, and the connected solid points are discovered by Syno.

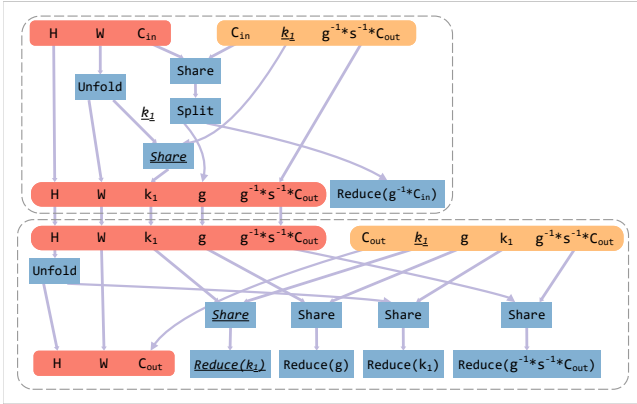


Figure 7. OPERATOR 1 discovered by Syno.

```
def __init__(self):
    self.w1 = randn([C_out//g//s, C_in, k_1])
    self.w2 = randn([C_out, k_1*k_1*C_out//s])
def forward(self, x):
    N, C_in, H, W = x.shape
    x = nn.functional.unfold(x, [1, k_1], padding=[0, k_1//2])
    # x: [N, C_in*k_1, H, W]
    x = reshape(x, [N, C_in, k_1, H, W])
    x = einsum("nckhw, dck -> ndckhw", x, self.w1)
    # x: [N, C_out//g//s, C_in, k_1, H, W]
    x = reshape(x, [N, C_out//g//s, g, C_in/g, k_1, H, W])
    x = sum(x, 3) # x: [N, C_out//g//s, g, k_1, H, W]
    x = reshape(x, [N, k_1*C_out//s, H, W])
    x = nn.functional.unfold(x, [k_1, 1], padding=[k_1//2, 0])
    # x: [N, k_1*k_1*C_out//s, H, W]
    x = einsum("nchwc, dc -> ndhcw", x, self.w2)
    return x # x: [N, C_out, H, W]
```

Listing 2. PyTorch code for OPERATOR 1.

We highlight a comparison between our optimized ResNet-34 and the baseline ResNet-18. Replacing the standard convolutions with our operators in ResNet-34 results in a model with *both* higher accuracy and better inference time than the ResNet-18 baseline. This observation implies a potentially promising direction to extend Syno to accuracy-preserving NN optimization: users can stack more layers and then compress the model with Syno, which might result in better accuracy and lower latency at the same time.

Case studies. Among all the operators discovered, we find two convolution-like operators with outstanding accuracy and inference performance.

OPERATOR 1 shown in Figure 7 achieves 2.68×, 2.04×, and 1.28× speedups on the three hardware platforms, with less than 1% ImageNet accuracy degradation. Its PyTorch code is shown in Listing 2. After the materialized reduction optimization during code generation (Section 8), it becomes a

stack of two stages *similar to* 1D and 2D grouped convolutions, but is *not* expressible in NAS. NAS can only sample traditional (grouped) convolutions, which always perform contractions between UNFOLDED windows of spatial dimensions and weights (the standard $\sum_j X[i + j - K / 2] * W[j]$ pattern). However, in OPERATOR 1, the first stage breaks this limitation. See the pattern underscored and italicized in Figure 7, which comprises 2 SHARES, 1 REDUCE, and 3 coordinates with domain k_1 . The SHARE in the first stage would have been REDUCED, had it been a traditional convolution. But the UNFOLDED window remains and is passed to the second stage to be contracted with the weight.

To see why such a pattern is effective, we stack two grouped convolutions into an operator, which is just OPERATOR 1 with the SHARED k_1 in stage 1 REDUCED and the W in stage 2 UNFOLDED again (hence might be discoverable under traditional NAS schemes). As in Figure 8, although having the same

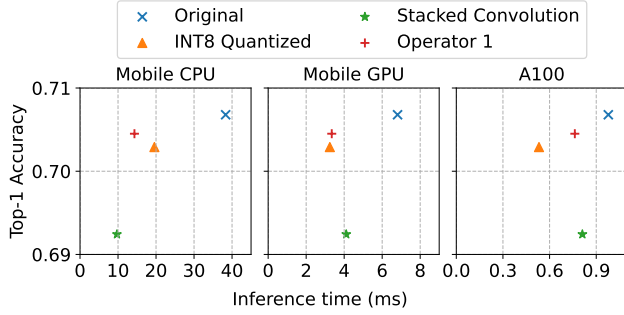


Figure 8. Comparison between OPERATOR 1 and other optimizations, evaluated on ImageNet with TVM.

FLOPs and similar latency, the stacked convolution doubles the accuracy degradation, which we attribute to the difference in the receptive field in OPERATOR 1 (3×3 vs. 3×5) that eases training of the model. This may provide insights for the machine learning community.

Since our design objective of trading accuracy for latency is the same as quantization, we further compare Syno with INT8 quantization. We obtain the quantized model from `torchvision.models.quantization.resnet18` using the QNNPACK configuration, and import it to TVM for inference optimizations. As in Figure 8, OPERATOR 1 has slightly better accuracy than INT8 quantization, as well as lower latency on the CPU. Note that Syno-synthesized operators can also be quantized, so the two techniques can be applied jointly to further enhance performance. Likewise, Syno can be combined with other similar techniques to achieve potentially more speedups.

OPERATOR 2 is a variant of OPERATOR 1, and resembles two 1D convolutions with weights connected using SHARE in a similar manner. Benefiting from the weight SHARE-ing, OPERATOR 2 yields 6.19 \times , 3.27 \times , and 2.61 \times speedups on the three hardware platforms, within 1.5% accuracy loss on CIFAR-100. We attribute the substantial performance speedups to its fewer parameters (less than 1/4 of standard 2D convolution) that can fit in the limited caches on edge devices.

Common patterns. We identify several common patterns from the novel operators discovered by Syno. Aside from the convolution and grouping patterns visible in OPERATOR 1, another common pattern is two weight tensors SHARE-ing one or more dimensions, similar to low-rank decomposition, which is highly effective in reducing the number of parameters and FLOPs. Also, we find multiple operators replacing one UNFOLD on a spatial dimension with a SHIFT, which can substantially reduce computations while still providing some extent of information mixture along this spatial dimension, similar to the idea in ShiftNet [41]. More patterns are unique to individual operators.

Comparison with NAS-PTE. Figure 9 shows the layer-wise performance of OPERATOR 1 and OPERATOR 2 compared

to the original convolution and all the three operators proposed by NAS-PTE, when used in ResNet-34. On the mobile CPU, the mobile GPU, and A100, compared to NAS-PTE, the speedups of our best operators over their best ones are 2.13 \times , 1.68 \times , and 1.63 \times on average when both are compiled with TVM, and 0.83 \times , 0.84 \times , and 1.38 \times when both are compiled with TorchInductor. Our best operators reduce the FLOPs by 1.76 \times to 4.32 \times , and reduce the number of parameters by 1.80 \times to 9.50 \times . The improvements are achieved without any layer-wise tuning like NAS-PTE but by a fully automated workflow in Syno.

Note that when compiled with TorchInductor, Syno underperforms NAS-PTE on the mobile CPU and GPU, despite the reduction on FLOPs and number of parameters. We find that TorchInductor often falls back to ATen kernels instead of generating native code for mobile hardware, as opposed to on A100 where it can emit efficient Triton code [36] in most of the time. The pre-compiled ATen kernels are less suitable for the novel Syno-generated operators. Actually, TorchInductor has mainly been optimized for large GPUs. Most of its templates target large GPUs, while smaller GPUs are neglected to keep the number of templates small and the compilation fast [7]. Thus, we attribute the slowdown to the immaturity of TorchInductor on mobile CPUs and GPUs rather than the inability of Syno. The more generic compiler TVM is able to deliver consistent speedups.

Comparison with α NAS. α NAS reported FLOPs reduction ratios and training speedups for some variants of ResNet and EfficientNet. Within 2% ImageNet accuracy drop, α NAS achieves 25% fewer FLOPs and about 12% TPU-v3 training speedup on both ResNet-50 and EfficientNet-B0, while Syno achieves 63% and 37% fewer FLOPs and 56% (48%) and 12% (7%) A100 inference speedup when compiled with TVM (TorchInductor) on ResNet-34 and EfficientNet-V2-S, respectively. This qualitatively shows Syno’s advantages.

9.3 Results on GPT-2

We follow Primer [30] to allocate a 30-minute training period on GPT-2 for each searched operator and compare their final language perplexity results. We then extend the training for the best-performing operator and the original model to reach 100,000 steps as shown in Figure 10. When searching for substitutions for the QKV projections, our best operator achieves a 1.1 \times training speedup and reduces the perplexity to 99, outperforming the original model’s perplexity of 111. More specifically, our operator constructs the original projections by groups, which allows the QKV matrices used in the attention modules to learn from different features of input tokens, thereby improving the training efficiency.

9.4 Ablation Studies

Canonicalization. To show the effectiveness of Syno canonicalization rules, we draw 6452 samples without canonicalization, among which only 86 are canonical. This implies that

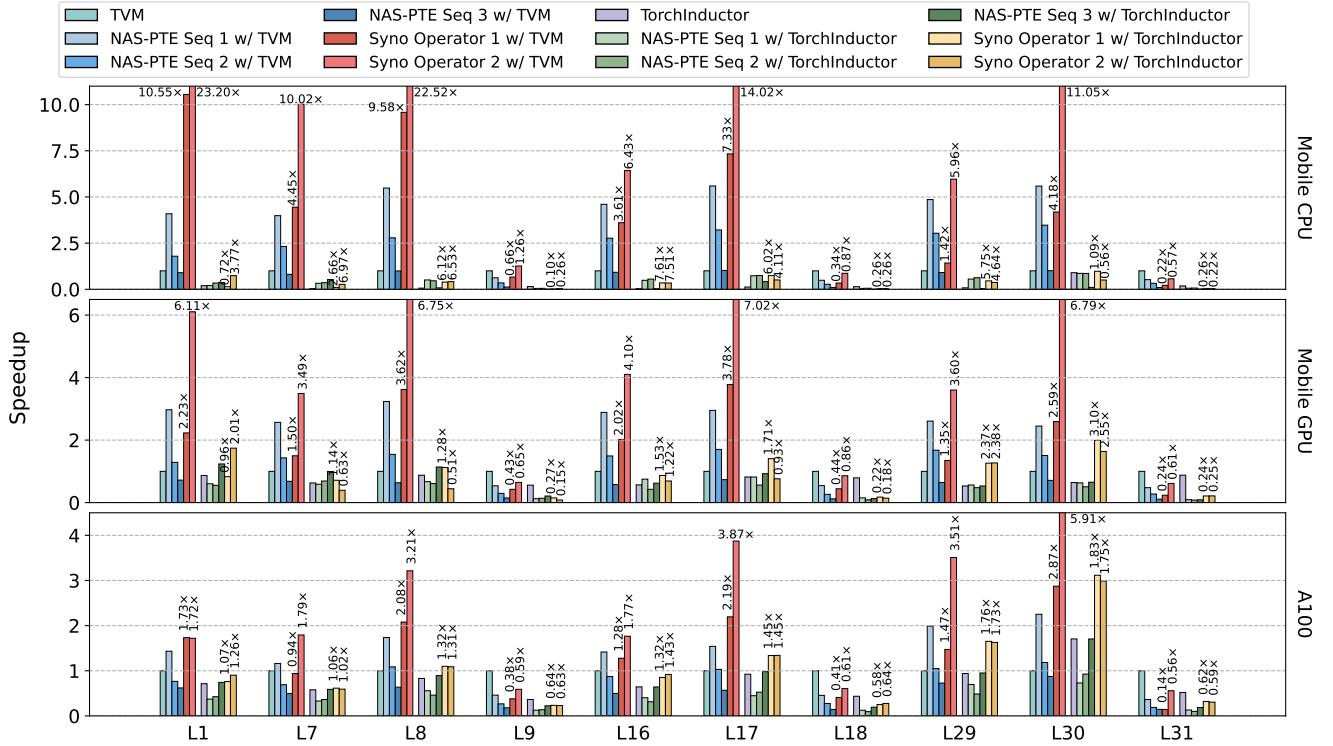


Figure 9. Layer-wise performance comparison between Syno and NAS-PTE on ResNet-34. The bars of each model are normalized to TVM.

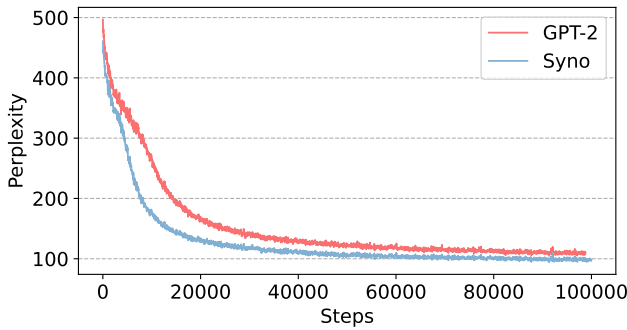


Figure 10. Comparison of language perplexity vs. training steps between Syno and the original GPT-2.

Table 3. Canonical rates of different sampled pGraph sizes.

2	3	4	5	6	7	≥8
100.00%	18.18%	13.97%	4.40%	1.22%	0.08%	0.00%

canonicalization cuts more than 70× redundancy. Table 3 shows the canonical rates for different pGraph sizes.

Shape distance. To verify the effectiveness of the shape distance metric, we evaluate the successful rates of random

sample trials with and without the guidance of shape distance, respectively. On a server machine with 192 virtual cores, 253 distinct operators are found after 5 million trials in 68.33 seconds, with shape distance enabled. However, without using shape distance, 500 million trials in 180.51 seconds yield no valid operators. Thus, shape distance is vital for avoiding useless synthesis.

10 Conclusions

This paper advocates the paradigm of neural operator synthesis, which automatically discovers novel NN operators with good inference accuracy and/or execution speed. A practical framework named Syno has been implemented, using a rich set of fine-grained primitives to construct operators, applying canonicalization to eliminate redundancy, and guided by a novel operator shape distance metric to improve synthesis efficiency. Syno is able to discover better NN operators than existing ones on various models, with higher execution performance and minor accuracy loss.

Acknowledgments

The authors thank the anonymous reviewers and our shepherd, Shoaib Kamil, for their valuable suggestions, and the Tsinghua IDEAL group members for constructive discussion. Mingyu Gao is the corresponding author.

References

- [1] Mohamed S. Abdelfattah, Abhinav Mehrotra, Lukasz Dudziak, and Nicholas Donald Lane. 2021. Zero-Cost Proxies for Lightweight NAS. In *9th International Conference on Learning Representations (ICLR)*. OpenReview.net. <https://openreview.net/forum?id=0cmMMY8J5q>
- [2] Jason Ansel, Edward Yang, Horace He, Natalia Gimelshein, Animesh Jain, Michael Voznesensky, Bin Bao, Peter Bell, David Berard, Evgeni Burovski, Geeta Chauhan, Anjali Chourdia, Will Constable, Alban Desmaison, Zachary DeVito, Elias Ellison, Will Feng, Jiong Gong, Michael Gschwind, Brian Hirsh, Sherlock Huang, Kshiteej Kalam-barkar, Laurent Kirsch, Michael Lazos, Mario Lezcano, Yanbo Liang, Jason Liang, Yinghai Lu, C. K. Luk, Bert Maher, Yunjie Pan, Christian Puhres, Matthias Reso, Mark Saroufim, Marcos Yukio Siraichi, Helen Suk, Shunting Zhang, Michael Suo, Phil Tillet, Xu Zhao, Eikan Wang, Keren Zhou, Richard Zou, Xiaodong Wang, Ajit Mathews, William Wen, Gregory Chanan, Peng Wu, and Soumith Chintala. 2024. PyTorch 2: Faster Machine Learning Through Dynamic Python Bytecode Transformation and Graph Compilation. In *Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS), Volume 2* (La Jolla, CA, USA). Association for Computing Machinery, New York, NY, USA, 929–947. <https://doi.org/10.1145/3620665.3640366>
- [3] Hadjer Benmeziane, Kaoutar El Maghraoui, Hamza Ouarnoughi, Smail Niar, Martin Wistuba, and Naigang Wang. 2021. A Comprehensive Survey on Hardware-Aware Neural Architecture Search. *arXiv preprint arXiv:2101.09336* (2021).
- [4] Han Cai, Ligeng Zhu, and Song Han. 2019. ProxylessNAS: Direct Neural Architecture Search on Target Task and Hardware. In *7th International Conference on Learning Representations (ICLR)*. OpenReview.net. <https://openreview.net/forum?id=HyIVB3AqYm>
- [5] Ciprian Chelba, Tomás Mikolov, Mike Schuster, Qi Ge, Thorsten Brants, Philipp Koehn, and Tony Robinson. 2014. One Billion Word Benchmark for Measuring Progress in Statistical Language Modeling. In *15th Annual Conference of the International Speech Communication Association, (INTERSPEECH)*. 2635–2639. <https://doi.org/10.21437/INTERSPEECH.2014-564>
- [6] 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. 2018. 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)* (Carlsbad, CA, USA). USENIX Association, USA, 579–594.
- [7] PyTorch community members. 2023. Investigate Strictness of torch.compile is_big_gpu. <https://github.com/pytorch/pytorch/issues/109489>.
- [8] Rémi Coulom. 2007. Efficient Selectivity and Backup Operators in Monte-Carlo Tree Search. In *Computers and Games*. Springer, Berlin, Heidelberg, 72–83.
- [9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 248–255. <https://doi.org/10.1109/CVPR.2009.5206848>
- [10] Thomas Elsken, Jan Hendrik Metzen, and Frank Hutter. 2019. Neural Architecture Search: A Survey. *The Journal of Machine Learning Research* 20, 1 (Jan. 2019), 1997–2017.
- [11] Romaric Gaudel and Michele Sebag. 2010. Feature Selection as a One-Player Game. In *Proceedings of the 27th International Conference on International Conference on Machine Learning (ICML)* (Haifa, Israel). Omnipress, Madison, WI, USA, 359–366.
- [12] W.H. Greub. 2012. *Multilinear Algebra*. Springer Berlin Heidelberg.
- [13] Sumit Gulwani, Oleksandr Polozov, and Rishabh Singh. 2017. Program Synthesis. *Foundations and Trends® in Programming Languages* 4, 1–2 (July 2017), 1–119. <https://doi.org/10.1561/25000000010>
- [14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 770–778. <https://doi.org/10.1109/CVPR.2016.90>
- [15] Yu-Jhen Hsu and Diego Perez Liebana. 2020. MCTS Pruning in Turn-Based Strategy Games. In *Joint Proceedings of the AIIDE 2020 Workshops co-located with 16th AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment (AIIDE) (CEUR Workshop Proceedings, Vol. 2862)*. CEUR-WS.org. <https://ceur-ws.org/Vol-2862/paper27.pdf>
- [16] Shi-Min Hu, Dun Liang, Guo-Ye Yang, Guo-Wei Yang, and Wen-Yang Zhou. 2020. Jittor: A Novel Deep Learning Framework with Meta-Operators and Unified Graph Execution. *Science China Information Sciences* 63 (2020), 1–21.
- [17] Yuanming Hu, Tzu-Mao Li, Luke Anderson, Jonathan Ragan-Kelley, and Frédo Durand. 2019. Taichi: A Language for High-Performance Computation on Spatially Sparse Data Structures. *ACM Transactions on Graphics (TOG)* 38, 6, Article 201 (Nov. 2019), 16 pages. <https://doi.org/10.1145/3355089.3356506>
- [18] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. 2017. Densely Connected Convolutional Networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 4700–4708. <https://doi.org/10.1109/CVPR.2017.243>
- [19] Charles Jin, Phitchaya Mangpo Phothilimthana, and Sudip Roy. 2022. Neural Architecture Search using Property Guided Synthesis. *Proceedings of the ACM on Programming Languages (PACMPL)* 6, OOPSLA2, Article 166 (Oct. 2022), 30 pages. <https://doi.org/10.1145/3563329>
- [20] Alex Krizhevsky and Geoffrey Hinton. 2009. Learning Multiple Layers of Features from Tiny Images. (2009).
- [21] Namhoon Lee, Thalaiyasingam Ajanthan, and Philip H. S. Torr. 2019. Snip: Single-Shot Network Pruning based on Connection Sensitivity. In *7th International Conference on Learning Representations (ICLR)*. OpenReview.net. <https://openreview.net/forum?id=B1VZqjAcYX>
- [22] Chenxi Liu, Barret Zoph, Maxim Neumann, Jonathon Shlens, Wei Hua, Li-Jia Li, Li Fei-Fei, Alan Yuille, Jonathan Huang, and Kevin Murphy. 2018. Progressive Neural Architecture Search. In *Computer Vision – ECCV 2018*, Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss (Eds.). Springer International Publishing, Cham, 19–35.
- [23] Karima Ma, Michael Gharbi, Andrew Adams, Shoaib Kamil, Tzu-Mao Li, Connelly Barnes, and Jonathan Ragan-Kelley. 2022. Searching for Fast Demosaicking Algorithms. *ACM Transactions on Graphics (TOG)* 41, 5, Article 172 (May 2022), 18 pages. <https://doi.org/10.1145/3508461>
- [24] Julie L. Newcomb, Andrew Adams, Steven Johnson, Rastislav Bodik, and Shoaib Kamil. 2020. Verifying and Improving Halide’s Term Rewriting System with Program Synthesis. *Proceedings of the ACM on Programming Languages (PACMPL)* 4, OOPSLA, Article 166 (Nov. 2020), 28 pages. <https://doi.org/10.1145/3428234>
- [25] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners. (2019).
- [26] Jonathan Ragan-Kelley, Connelly Barnes, Andrew Adams, Sylvain Paris, Frédo Durand, and Saman Amarasinghe. 2013. Halide: A Language and Compiler for Optimizing Parallelism, Locality, and Recursion in Image Processing Pipelines. In *Proceedings of the 34th ACM SIGPLAN Conference on Programming Language Design and Implementation (PLDI)* (Seattle, Washington, USA). Association for Computing Machinery, New York, NY, USA, 519–530. <https://doi.org/10.1145/2491956.2462176>
- [27] Alex Rogozhnikov. 2021. Einops: Clear and Reliable Tensor Manipulations with Einstein-like Notation. In *International Conference on Learning Representations (ICLR)*.
- [28] Junru Shao, Xiyu Zhou, Siyuan Feng, Bohan Hou, Ruihang Lai, Hongyi Jin, Wuwei Lin, Masahiro Masuda, Cody Hao Yu, and Tianqi Chen. 2022. Tensor Program Optimization with Probabilistic Programs.

- In *Proceedings of the 36th International Conference on Neural Information Processing Systems (NeurIPS)* (New Orleans, LA, USA). Curran Associates Inc., Red Hook, NY, USA, Article 2593, 14 pages.
- [29] Kensen Shi, David Bieber, and Rishabh Singh. 2022. TF-Coder: Program Synthesis for Tensor Manipulations. *ACM Transactions on Programming Languages and Systems (TOPLAS)* 44, 2, Article 10 (May 2022), 36 pages. <https://doi.org/10.1145/3517034>
- [30] David R. So, Wojciech Mańke, Hanxiao Liu, Zihang Dai, Noam Shazeer, and Quoc V. Le. 2021. Primer: Searching for Efficient Transformers for Language Modeling. In *Proceedings of the 35th International Conference on Neural Information Processing Systems (NeurIPS)*. Curran Associates Inc., Red Hook, NY, USA, Article 460, 13 pages.
- [31] Mingxing Tan, Bo Chen, Ruoming Pang, Vijay Vasudevan, Mark Sandler, Andrew Howard, and Quoc V. Le. 2019. MnasNet: Platform-Aware Neural Architecture Search for Mobile. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2815–2823. <https://doi.org/10.1109/CVPR.2019.00293>
- [32] Mingxing Tan and Quoc V. Le. 2019. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, Vol. 97. PMLR, 6105–6114. <http://proceedings.mlr.press/v97/tan19a.html>
- [33] Mingxing Tan and Quoc V. Le. 2021. EfficientNetV2: Smaller Models and Faster Training. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, Vol. 139. PMLR, 10096–10106. <http://proceedings.mlr.press/v139/tan21a.html>
- [34] Shizhi Tang, Jidong Zhai, Haojie Wang, Lin Jiang, Liyan Zheng, Zhenhao Yuan, and Chen Zhang. 2022. FreeTensor: A Free-Form DSL with Holistic Optimizations for Irregular Tensor Programs. In *Proceedings of the 43rd ACM SIGPLAN International Conference on Programming Language Design and Implementation (PLDI)* (San Diego, CA, USA). Association for Computing Machinery, New York, NY, USA, 872–887. <https://doi.org/10.1145/3519939.3523448>
- [35] Ross Tate, Michael Stepp, Zachary Tatlock, and Sorin Lerner. 2009. Equality Saturation: A New Approach to Optimization. In *Proceedings of the 36th Annual ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages (POPL)* (Savannah, GA, USA). Association for Computing Machinery, New York, NY, USA, 264–276. <https://doi.org/10.1145/1480881.1480915>
- [36] Philippe Tillet, H. T. Kung, and David Cox. 2019. Triton: An Intermediate Language and Compiler for Tiled Neural Network Computations. In *Proceedings of the 3rd ACM SIGPLAN International Workshop on Machine Learning and Programming Languages (MAPL)* (Phoenix, AZ, USA). Association for Computing Machinery, New York, NY, USA, 10–19. <https://doi.org/10.1145/3315508.3329973>
- [37] Jack Turner, Elliot J. Crowley, and Michael F. P. O’Boyle. 2021. Neural Architecture Search as Program Transformation Exploration. In *Proceedings of the 26th ACM International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS)* (Virtual, USA). Association for Computing Machinery, New York, NY, USA, 915–927. <https://doi.org/10.1145/3445814.3446753>
- [38] Jack Turner, Elliot J. Crowley, Michael F. P. O’Boyle, Amos J. Storkey, and Gavin Gray. 2020. BlockSwap: Fisher-Guided Block Substitution for Network Compression on a Budget. In *8th International Conference on Learning Representations (ICLR)*. OpenReview.net. <https://openreview.net/forum?id=SkIkDkSFPB>
- [39] Haojie Wang, Jidong Zhai, Mingyu Gao, Zixuan Ma, Shizhi Tang, Liyan Zheng, Yuanzhi Li, Kaiyuan Rong, Yuanyong Chen, and Zhihao Jia. 2021. PET: Optimizing Tensor Programs with Partially Equivalent Transformations and Automated Corrections. In *15th USENIX Symposium on Operating Systems Design and Implementation (OSDI)*. USENIX Association, 37–54. <https://www.usenix.org/conference/osdi21/presentation/wang>
- [40] Bichen Wu, Xiaoliang Dai, Peizhao Zhang, Yanghan Wang, Fei Sun, Yiming Wu, Yuandong Tian, Peter Vajda, Yangqing Jia, and Kurt Keutzer. 2019. FBNet: Hardware-Aware Efficient ConvNet Design via Differentiable Neural Architecture Search. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 10726–10734. <https://doi.org/10.1109/CVPR.2019.01099>
- [41] Bichen Wu, Alvin Wan, Xiangyu Yue, Peter Jin, Sicheng Zhao, Noah Golmant, Amir Gholaminejad, Joseph Gonzalez, and Kurt Keutzer. 2018. Shift: A Zero FLOP, Zero Parameter Alternative to Spatial Convolutions. In *2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 9127–9135. <https://doi.org/10.1109/CVPR.2018.00951>
- [42] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. 2017. Aggregated Residual Transformations for Deep Neural Networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 5987–5995. <https://doi.org/10.1109/CVPR.2017.634>
- [43] Jie Zhao, Bojie Li, Wang Nie, Zhen Geng, Renwei Zhang, Xiong Gao, Bin Cheng, Chen Wu, Yun Cheng, Zheng Li, Peng Di, Kun Zhang, and Xuefeng Jin. 2021. AKG: Automatic Kernel Generation for Neural Processing Units using Polyhedral Transformations. In *Proceedings of the 42nd ACM SIGPLAN International Conference on Programming Language Design and Implementation (PLDI)* (Virtual, Canada). Association for Computing Machinery, New York, NY, USA, 1233–1248. <https://doi.org/10.1145/3453483.3454106>
- [44] Lianmin Zheng, Chengfan Jia, Minmin Sun, Zhao Wu, Cody Hao Yu, Ameer Haj-Ali, Yida Wang, Jun Yang, Danyang Zhuo, Koushik Sen, Joseph E. Gonzalez, and Ion Stoica. 2020. Ansor: Generating High-Performance Tensor Programs for Deep Learning. In *Proceedings of the 14th USENIX Conference on Operating Systems Design and Implementation (OSDI)*. USENIX Association, USA, Article 49, 17 pages.
- [45] Dongzhan Zhou, Xinchu Zhou, Wenwei Zhang, Chen Change Loy, Shuai Yi, Xuesen Zhang, and Wanli Ouyang. 2020. EcoNAS: Finding Proxies for Economical Neural Architecture Search. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 11393–11401. <https://doi.org/10.1109/CVPR42600.2020.01141>
- [46] Barret Zoph and Quoc V. Le. 2017. Neural Architecture Search with Reinforcement Learning. In *International Conference on Learning Representations (ICLR)*. <https://openreview.net/forum?id=r1Ue8Hcxg>
- [47] Barret Zoph, Vijay Vasudevan, Jonathon Shlens, and Quoc V. Le. 2018. Learning Transferable Architectures for Scalable Image Recognition. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 8697–8710. <https://doi.org/10.1109/CVPR.2018.00907>

A Artifact Appendix

A.1 Abstract

This artifact appendix helps the readers reproduce the main evaluation results of this paper. The artifact evaluation includes the instructions on how to synthesize operators with Syno, obtain the accuracy and performance results of the models with the novel operators, and plot Figures 5, 6, 8, 9, and 10. Our GitHub repository also provides a README containing detailed instructions.

A.2 Artifact check-list (meta-information)

- **Algorithm:** Syno’s operator synthesis algorithm.
- **Compilation:** GCC 13 and CUDA Toolkit 12.9.
- **Models:** ResNet-18, ResNet-34, DenseNet-121, ResNeXt-29-2x64D, EfficientNet-V2-S, and GPT-2.
- **Data sets:** CIFAR100, ImageNet, and Im1b.
- **Run-time environment:** Linux Ubuntu 24.04. We provide a Dockerfile for the environment setup.
- **Hardware:** NVIDIA Jetson Orin Nano 8 GB board, and NVIDIA A100 GPUs.

- **Metrics:** Model inference accuracy and execution latency.
- **Output:** The key results are a list of synthesized operators discovered by Syno, plus five plots summarizing their performance.
- **Experiments:** The workflow is introduced in AE/README.
- **How much disk space required (approximately)?:** 500GB (For ImageNet).
- **How much time is needed to prepare workflow (approximately)?:** Docker images can be built within 30 minutes.
- **How much time is needed to complete experiments (approximately)?:** Searching for operators requires around 300 GPU hours per model. Accuracy re-evaluation takes about 900 GPU hours for all models. Tuning for performance on the A100 and Jetson Orin Nano GPUs takes about 700 GPU hours each, and tuning on the Jetson Orin Nano CPU takes 700 hours. Plotting can be done within several minutes.
- **Publicly available?:** Yes.
- **Code licenses (if publicly available)?:** MIT.

A.3 Description

A.3.1 How to access.

- Source code: <https://github.com/tsinghua-ideal/Syno>.
- Artifact evaluation data: <https://github.com/Yongqi-Zhuo/Syno-AE>, which is a git submodule of the above repository, so you need not separately clone it.

A.3.2 Hardware dependencies.

- Searching for operators requires A100 GPUs (an 8×A100 machine is recommended).
- Performance tuning requires at least an A100 GPU and an NVIDIA Jetson Orin Nano 8 GB board.

A.3.3 Software dependencies. We provide a Dockerfile in our repository for easy reproduction, and the detailed software dependencies are listed there.

A.3.4 Data sets. The search and evaluation require three datasets: CIFAR100, ImageNet, and lm1b. CIFAR100 and lm1b will be automatically downloaded from TorchVision and Huggingface when executing our scripts. ImageNet requires some manual preparation, for which the detailed steps can be found in the README.

A.3.5 Models. Our experiments are conducted with six models: ResNet-18, ResNet-34, DenseNet-121, ResNeXt-29-2x64D, EfficientNet-V2-S, and GPT-2. Note that you do not need the pre-trained weights for those models.

A.4 Installation

Clone the repository and build the Docker image using the Dockerfile inside the repository.

```
git clone --recursive \
  https://github.com/tsinghua-ideal/Syno.git
docker build -t syno Syno
```

The experiments will need preprocessed ImageNet. Download the dataset, format it into a PyTorch-style dataset, follow

the instructions in [FFCV-ImageNet](#) to prepare the dataset with `bash write_imagenet.sh 400 0.10 90`, and finally set the directory in Syno with `bash set_imagenet_dir.sh $WRITE_DIR`.

A.5 Experiment workflow

On a high level, our experiments can be decoupled into four steps:

1. Searching: search for efficient operators to be substituted into the neural network models using the Syno operator synthesis algorithm.
2. Accuracy Evaluation: evaluate the accuracy of the models optimized by Syno on ImageNet.
3. Tuning: tune the Syno-optimized models with tensor compilers to obtain the performance numbers.
4. Plotting: plot the accuracy and performance results of the optimized models to visualize the tradeoff achieved with Syno.

Since these steps can take very long time, to facilitate easier reproduction, we provide our data in AE/data as drop-in replacements for the results of each step, so the reproduction can start from any intermediate step.

A.5.1 Searching. Please use `search.sh` for the search. Specifically, run `bash search.sh $MODEL`, where `$MODEL` is the model to search with. The supported models include

- resnet18
- resnet34
- resnext29_2x64d
- densenet121
- efficientnet_v2_s
- gpt2

A.5.2 Accuracy Evaluation. The search on the vision models will produce a list of operators with their CIFAR-100 accuracies and FLOPs, saved in `AE/results/$MODEL-session`. After picking operators with good accuracies, you need to re-evaluate them on ImageNet by `reevaluate_vision.sh`. The detailed instructions can be found in `AE/README.md`.

The search on GPT-2 will also produce a list of operators with their perplexity results after training for 30 minutes. After picking the best operator, re-evaluate it with 100,000 steps using `reevaluate_gpt.sh`.

Finally, we provide two scripts — `train_baseline.sh` and `train_custom.sh`, with which you can obtain the accuracies for the baselines and the operators we picked for the case studies.

A.5.3 Tuning. You need to set up the host, one or more A100 GPUs, and one or more NVIDIA Jetson Orin Nano's. Make sure the devices can access the host via internet connection. Also set up the TVM RPC trackers and servers according to `AE/README.md`. Then run the grid tuners. Refer to `AE/README.md` for detailed instructions.

On host

```
python grid_tune.py \
  --config /workspace/Syno/AE/grid_tune.json \
  --rpc-host $TRACKER_HOST \
  --rpc-port $TRACKER_PORT
# On A100
python grid_torch.py \
  --config /workspace/Syno/AE/grid_tune.a100.json
# On NVIDIA Jetson Orin Nano
python grid_torch.py \
  --config /workspace/Syno/AE/grid_tune.mdev.json
```

A.5.4 Plotting. After the above steps, you can plot the figures with the experiment results.

First, copy the tuning results:

```
bash copy_perf.sh mdev
bash copy_perf.sh a100
```

Then run `bash plot.sh` to produce the figures.

The script will produce 5 figures in `AE/plots`:

1. `end-to-end-performance.pdf`: Figure 5.
2. `imagenet-performance.pdf`: Figure 6.
3. `case-study.pdf`: Figure 8.
4. `kernel-performance.pdf`: Figure 9.

5. `gpt-loss.pdf`: Figure 10.

A.6 Evaluation and expected results

We provide our experiment results in `AE/data`. If you use our data, then you should see exactly the same figures as in our paper. Otherwise, the numbers might be slightly different due to the randomness introduced by operator searching and the fluctuations of performance during tuning, but the overall trend should be the same.

A.7 Experiment customization

You can write your own configuration files other than the provided `AE/grid_tune.json` for performance tuning to use other hardware and other synthesized operators. See `AE/README.md` for more details.

A.8 Notes

A.9 Methodology

Submission, reviewing and badging methodology:

- <https://www.acm.org/publications/policies/artifact-review-and-badging-current>
- <https://cTuning.org/ae>