

The Deep Learning Compiler: A Comprehensive Survey

MINGZHEN LI, School of Computer Science and Engineering, Beihang University, China

YI LIU, School of Computer Science and Engineering, Beihang University, China

XIAOYAN LIU, School of Computer Science and Engineering, Beihang University, China

QINGXIAO SUN, School of Computer Science and Engineering, Beihang University, China

XIN YOU, School of Computer Science and Engineering, Beihang University, China

HAILONG YANG[†], School of Computer Science and Engineering, Beihang University, China

ZHONGZHI LUAN, School of Computer Science and Engineering, Beihang University, China

DEPEI QIAN, School of Computer Science and Engineering, Beihang University, China

The difficulty of deploying various deep learning (DL) models on diverse DL hardware has boosted the research and development of DL compilers in the community. Several DL compilers have been proposed from both industry and academia such as Tensorflow XLA and TVM. Similarly, the DL compilers take the DL models described in different DL frameworks as input, and then generate optimized codes for diverse DL hardware as output. However, none of the existing survey has analyzed the unique design of the DL compilers comprehensively. In this paper, we perform a comprehensive survey of existing DL compilers by dissecting the commonly adopted design in details, with emphasis on the DL oriented multi-level IRs, and frontend/backend optimizations. Specifically, we provide a comprehensive comparison among existing DL compilers from various aspects. In addition, we present detailed analysis of the multi-level IR design and compiler optimization techniques. Finally, several insights are highlighted as the potential research directions of DL compiler. This is the first survey paper focusing on the unique design of DL compiler, which we hope can pave the road for future research towards the DL compiler.

Additional Key Words and Phrases: Neural Networks, Deep Learning, Compiler, Intermediate Representation, Optimization

1 INTRODUCTION

The development of deep learning (DL) has generated profound impact on various scientific fields. It has not only demonstrated remarkable value in artificial intelligence such as natural language processing (NLP) [68] and computer vision (CV) [39], but also proved great success in broader applications such as e-commerce [47], smart city [70] and drug discovery [27]. With the emergence of versatile deep learning models such as convolutional neural network (CNN) [61], recurrent neural network (RNN) [80], long short-term memory (LSTM) [49] and generative adversarial network (GAN) [42], it is critical to ease the programming of diverse DL models in order to realize their widely adoption.

With the continuous efforts from both industry and academia, several popular DL programming frameworks have been proposed such as TensorFlow [16], PyTorch [74], MXNet [28] and CNTK [81], in order to simplify the application of various DL models. Although there are strengths and weaknesses among the above DL programming frameworks depending on the tradeoffs in their designs, the interoperability becomes important to reduce the redundant engineering efforts when supporting emerging DL models across the existing DL models. To improve the interoperability, ONNX [10] has been proposed that defines an open-source format for representing DL models, which facilitates model conversion between different DL frameworks.

In the meanwhile, the unique computing characteristics such as matrix multiplication have spurred the passion of chip architects to design customized DL chips for higher efficiency. Internet giants (e.g., Google TPU [54], Hisilicon NPU [62], Apple Bonic [58]), processor vendors (e.g., NVIDIA Turing [9], Intel NNP [8]), service providers (e.g., Amazon Inferentia [3], Alibaba Hanguang [2]), and even startups (e.g., Cambricon [63], Graphcore [53]) are investing tremendous workforce and

capital in developing DL chips in order to boost computing capability for DL models. Generally, the category of DL chips include: 1) general-purpose chips with software-hardware co-design; 2) dedicated chips fully customized for DL models; 3) neuromorphic chips inspired by biological brain science. For example, the general-purpose chips (e.g., CPU, GPU) have added special hardware components such as AVX512 vector units and tensor core to accelerate DL models. Whereas for dedicated chips such as Google Tensor Processing Unit (TPU), application-specific integrated circuits (e.g., matrix multiplication engine and high-bandwidth memory) have been designed to elevate the performance and energy efficiency to extreme. To the foreseeable future, the design of DL chips would become even more diverse.

To accelerate DL models on diverse DL chips, it is important to map the computation to DL chips efficiently. On general-purpose chips, the highly optimized linear algebra libraries such as Basic Linear Algebra Subprograms (BLAS) libraries (e.g., MKL and cuBLAS) serve as the basics for efficient computation of DL models. Take the convolution operation for example, the DL frameworks convert the convolution to matrix multiplication and then invoke the GEMM function in the BLAS libraries. In addition, the chip vendors have released specially optimized libraries tailored for DL computations (e.g., MKL-DNN and cuDNN), including forward and backward convolution, pooling, normalization, and activation. More advanced tools have also been developed to further speedup the DL operations. Take TensorRT [14] for example, it supports graph optimization (e.g., layer fusion) and low-bit quantization with large collection of highly optimized GPU kernels. On dedicated DL chips, similar libraries as well as tool-chains are also provided by the vendors to execute the DL models efficiently. However, the drawback of relying on the libraries and tools described above for mapping of DL models on diverse DL chips is that they usually fall behind the rapid development of DL models, and thus fail to utilize the DL chips efficiently.

To address the drawback of DL libraries and tools, as well as alleviate the burden to optimize the DL models on each DL chip manually, the DL community has resorted to the domain specific compiler techniques for rescue. Rapidly, several popular DL compilers have been proposed such as TVM [29], Tensor Comprehension [89], Glow [79], nGraph [34] and XLA [60], from both industry and academia. The DL compilers take the model definitions described in the DL frameworks as inputs, and generate efficient code implementations on various DL chips as outputs. The transformation between model definition and specific code implementation are highly optimized targeting the model specification and hardware architecture. Specifically, the DL compilers incorporate DL oriented optimizations such as layer and operator fusion, which enables highly efficient code generation. Moreover, existing DL compilers also leverage mature tool-chains from general-purpose compilers (e.g., LLVM [59]), which provides better portability across diverse hardware architectures. Similar to traditional compiler, DL compilers also adopt the layered design including frontend, intermediate representation (IR) and backend. However, the uniqueness of DL compiler lies in the design of multi-level IRs and DL specific optimizations.

In this paper, we provide a comprehensive survey of existing DL compilers by dissecting the compiler design into frontend, multi-level IR and backend, with special emphasis on the IR design and optimization methods. To the best of our knowledge, this is the first paper that provides a comprehensive survey on the design of DL compiler. Specifically, this paper makes the following contributions:

- We provide a comprehensive comparison among existing DL compilers from various aspects such as hardware support, DL framework support, code generation and optimization, which can be used as guidelines for choosing the suitable DL compiler for the end user.

- We dissect the general design of existing DL compilers, and provide detailed analysis of the multi-level IR design and compiler optimization techniques such as dataflow-level optimization, hardware intrinsic mapping, memory latency hiding and parallelization.
- We provide several insights for the future development of DL compilers, including auto-tuning, polyhedral compiler, quantization, differentiable programming and privacy protection, which we hope to boost the research in the DL compiler community.

The remaining part of this paper is organized as follows. Section 2 presents the background of DL compilers, including the DL frameworks, DL chips, as well as hardware (FPGA) specific DL compilers. Section 3 presents a detailed comparison among existing DL compilers. Section 4 describes the general design of DL compilers, with emphasis on the IR and frontend/backend optimizations. Section 5 concludes the paper and highlights the future directions.

2 BACKGROUND

2.1 Deep Learning Frameworks

In this sub-section, we provide an overview of popular DL frameworks. The discussion might not be exhaustive but is meant to provide a guideline for DL practitioners. Figure 1, presents the landscape of DL frameworks including currently popular frameworks, historical frameworks and ONNX supported frameworks.

TensorFlow - TensorFlow was originally released by Google in late 2015. Among all the DL frameworks, TensorFlow has the most comprehensive support for language interfaces, including C++, Python, Java, Go, R, and Haskell. TensorFlow employs a dataflow graph of primitive operators extended with restricted control edges to represent differentiable programs [78]. TensorFlow Lite is designed for mobile and embedded deep learning and provides an Android neural network API. To reduce the complexity of using TensorFlow, Google added Keras as a frontend to the TensorFlow core. Furthermore, TensorFlow eager-mode applies an approach similar to PyTorch to support dynamic computation graphs better.

Keras - Keras [32] was first released in March 2015 with Google Support. Keras is a high-level neural network library for quickly building DL models, written in pure Python. Though not a DL framework on its own, Keras provides a high-level API that integrates with TensorFlow, MXNet, Theano, and CNTK. With Keras, DL developers can build a neural network with just a few lines of code. Besides, Keras can integrate other common DL packages, such as scikit-learn for Python. However, Keras is not flexible enough due to over-encapsulation, which makes it too difficult to add operations or obtain low-level data information.

PyTorch - PyTorch was launched by Facebook in early 2017. Facebook rewrote the Lua-based DL framework Torch in Python and refactored all modules on *Tensor* level. Besides, As the most popular dynamic framework, PyTorch embeds primitives for constructing dynamic data flow graphs in Python, where the control flow is executed in the Python interpreter. PyTorch 1.0 integrated the codebases of PyTorch 0.4 and Caffe2 to create a unified framework. This allows PyTorch to absorb the benefits of Caffe2 to support efficient graph execution and mobile deployment. FastAI [50] is an advanced API layer based on PyTorch's upper-layer encapsulation. It fully borrows Keras to improve PyTorch's ease of use.

Caffe/Caffe2 - Caffe [52] was designed for deep learning and image classification at UC Berkeley in 2014. Caffe supports the command line, Python, and MATLAB APIs. An important feature of Caffe is the ability to train and deploy models without writing code. Caffe's simple framework makes its code easy to extend, suitable for developers to analyze in-depth. Therefore, Caffe is mainly positioned in research, which has made it popular from originated to the present. Caffe2 builds upon the original Caffe project, supporting both mobile (e.g., iOS and Android) and server

(e.g., Linux, Windows, and Mac) build platforms. Caffe2 is similar in structure to TensorFlow, albeit with a lighter API and making access to the intermediate results in the computation graph much easier [89].

MXNet - Apache MXNet supports multiple language APIs including Python, C++, R, Scala, Julia, Matlab, and JavaScript. The project began in September 2015, with version 1.0.0 released in December 2017. MXNet was intended to be scalable and was designed from a systems perspective to reduce data loading and I/O complexity [28]. MXNet offers different paradigms: imperative programming like Caffe and Tensorflow as well as imperative like PyTorch. In December 2017, Amazon and Microsoft jointly released Gluon [6], which is an advanced interface similar to Keras and FastAI based on MXNet. The biggest feature of Gluon is that it supports both flexible, dynamic graphs and efficient, static graphs. Gluon is now available in Apache MXNet and Microsoft Cognitive Toolkit (CNTK).

CNTK - The Microsoft Cognitive Toolkit (also known as CNTK) began development in October 2015. CNTK can be used through Python, C++ and C# APIs, or its own scripting language (i.e., BrainScript). CNTK is designed to be easy-to-use and production-ready for use on large production scale data and is supported on Linux and Windows [48]. However, CNTK does not yet support the ARM architecture, which limits its usage on mobile devices. CNTK uses the static computation graph similar to TensorFlow and Caffe, in which a neural network is treated as a series of computational steps through a directed graph.

PaddlePaddle - In August 2016, Baidu open-sourced PaddlePaddle [11], a DL framework that has been used internally for years. PaddlePaddle can be applied to natural language processing, image recognition, recommendation engine, and so on. The original design of PaddlePaddle is similar to Caffe, where each model can be represented as a set of layers. In April 2017, Baidu launched PaddlePaddle v2, which added the concept of operators with reference to TensorFlow, breaking layers into finer-grained operators, thereby supporting more complex network structures. Furthermore, PaddlePaddle Fluid was launched in late 2017. Fluid is similar to PyTorch in that it provides its own interpreter so as not to be limited by Python's performance.

ONNX - The Open Neural Network Exchange (ONNX) [10] was released by Microsoft and Facebook in September 2017. ONNX defines a scalable computation graph model, and computation graphs built by different DL frameworks can be transformed into it. ONNX makes it easier to convert models between DL frameworks. For example, it allows developers to build an MXNet model and then run the model using PyTorch for inference. As shown in Figure 1, ONNX has been integrated into PyTorch, MXNet, PaddlePaddle, and so on. For several DL frameworks (e.g., TensorFlow and Keras) that are not directly supported yet, ONNX adds converters to them.

Historical Frameworks - Due to the rapid evolvement of the DL community, many historical DL frameworks are not longer exist or active. For example, PyTorch proposed by Facebook has replaced Torch [33]. As one of the oldest DL frameworks, Theano [86] is no longer under maintenance. Deeplearning4J [85] a distributed DL framework based on Java and Scala, however becomes inactive due to the lack of large developer community such as PyTorch. Chainer [87] was once the preferred framework for dynamic computation graphs, however replaced by MXNet, PyTorch and TensorFlow with similar features.

Previous works [21, 38, 46, 72, 82, 100] compared the performance of DL frameworks on different applications (e.g., computer vision and image classification) and different hardware targets (e.g., CPU, GPU, and TPU). For detailed survey of each DL framework, the readers can refer to [48]. Different from [48], this survey focuses on the research efforts on DL compilers that provide more general approach to execute various DL models on diverse hardware efficiently.

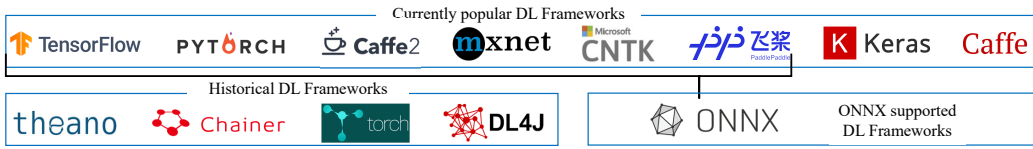


Fig. 1. DL framework landscape: 1) Currently popular DL frameworks; 2) Historical DL frameworks; 3) ONNX supported frameworks.

2.2 Deep Learning Hardware

The DL hardware can be divided into four categories based on the generality: 1) general-purpose hardware that can support DL workloads through hardware and software optimization; 2) dedicated hardware that focus on accelerating DL workloads with fully customized circuit design; 3) neuromorphic hardware that function by mimicking the human brain.

General-purpose Hardware - The most representative general-purpose hardware for DL models is Graphic Processing Unit (GPU), which achieves high parallelism with many-core architecture. For example, Nvidia GPUs have introduced tensor cores since the Volta architecture. Tensor cores can accelerate mixed-precision matrix multiply-and-accumulate calculations in parallel, which are widely used in DL models during both training and inference. Co-optimized with the hardware, NVIDIA also launches highly optimized DL libraries and tools such as cuDNN [31] and TensorRT [14] to further accelerate the computation of DL models.

Dedicated Hardware - Dedicated hardware are fully customized for DL computation to improve performance and energy efficiency to extreme. The rapid expansion of DL applications and algorithms has spurred many startups developing dedicated DL hardware (e.g., Graphcore GC2, Cambricon MLU270). Besides, traditional hardware companies (e.g., Intel NNP, Qualcomm Cloud AI 100) and cloud service providers (e.g., Google TPU, Amazon Inferentia, and Alibaba Hanguang) have also invested in this field. The most well known dedicated DL hardware are Google's TPU series. A TPU includes Matrix Multiplier Unit (MXU), Unified Buffer (UB), and Activation Unit (AU), which is driven with CISC instructions by the host processor. The MXU is mainly composed of a systolic array, which is optimized for power and area efficiency in performing matrix multiplications. Compared to CPU and GPU, TPU is still programmable but uses a matrix as a primitive instead of a vector or scalar. The Amazon Inferentia has also attracted the attention recently. This chip has four NeuroCores that are designed for tensor-level operations, and it has large on-chip cache to avoid the frequent main memory access.

Neuromorphic Hardware - Neuromorphic chips use electronic technology to simulate the biological brain. Representative products of this kind are IBM's TrueNorth and Intel's Loihi. Neuromorphic chips (e.g., TrueNorth) have very high connectivity between their artificial neurons. Neuromorphic chips also replicate a structure similar to the brain tissue: neurons can simultaneously store and process the data. Traditional chips distribute processors and memory in different locations, but neuromorphic chips usually have many microprocessors, each of which has a small amount of local memory. Compared to TrueNorth, Loihi has a learning ability more similar to the brain. Loihi introduces the pulse-time-dependent synaptic plasticity model (STDP), a mechanism that regulates synaptic strength by the relative time of pre-synaptic and post-synaptic pulses. However, neuromorphic chips are far away from large-scale commercial production. Despite that, in computer science domain, neuromorphic chips can help to capture the process of rapid, life-long learning which is ignored by regular DL models, and in neurology domain, they are helpful to figure out how the various parts of the brain work together to create thoughts, feelings, and even consciousness.

2.3 Hardware specific DL Compiler

Field Programmable Gate Arrays (FPGAs) are reprogrammable integrated circuits that contain an array of programmable logic blocks. Programmers can configure them after manufacturing. Besides the reprogrammable nature, the low-power and high-performance nature of the FPGA make it widely used in so many domains, such as communication, medical, image processing, and ASIC prototyping. As for the domain of deep learning, the high-performance CPUs and GPUs are highly-reprogrammable but power-hungry, while the power-efficient ASICs are specialized for fixed applications. However, the FPGA can bridge the gap between CPUs/GPUs and ASICs, which causes the FPGA to be an attractive platform for deep learning.

The High-Level Synthesis (HLS) programming model enables the FPGA programmers to generate effective hardware designs conveniently using high-level languages such as C and C++. It avoids writing lots of Verilog or VHDL descriptions, which lowers the programming threshold and reduces the long design circles. Xilinx Vivado HLS and Intel FPGA SDK for OpenCL are two of the popular HLS tools targeting their own FPGAs. However, mapping DL models to FPGAs remains a complicated work even with HLS, because that 1) DL models are usually described by the languages of DL frameworks rather than bare mental C/C++ code, and 2) DL-specific information and optimizations are hard to be leveraged.

The hardware-specific DL compilers targeting FPGA take the DL models or their domain-specific languages (DSLs) as the input, conduct the domain-specific (about FPGA and DL) optimizations and mappings, then generate the HLS or Verilog/VHDL and finally generate the bitstream. They can be classified into two categories according to the generated architectures of FPGA-based accelerators: the processor architecture and the streaming architecture [91].

The processor architecture has similarities with general-purpose processors. An FPGA accelerator of this architecture usually comprises several Processing Units (PUs), which are comprised of on-chip buffers and multiple smaller Processing Engines (PEs). It usually has a virtual instruction set (ISA), and the control of hardware and the scheduling of the execution should be determined by software. What's more, the static scheduling method avoids the overheads of von Neumann execution (including instruction fetching and decoding). A hardware template is a generic and fundamental implementation with configurable parameters. The DL compilers targeting this architecture adopt the hardware templates to generate the accelerator designs automatically. With the configurable parameters of templates, compilers achieve the scalability and flexibility [106]. The scalability means that the compilers can generate designs for FPGAs ranging from high-performance to power-efficient, and the flexibility means that the compilers can generate designs for various DL models with different layer types and parameters. The number of PUs and the number of PEs per PU are template parameters of importance. Besides, the tiling size and batch size are also essential scheduling parameters about mapping the DL models to PUs and PEs. All these parameters are usually determined by the design space exploration using various strategies, such as combining the performance model and auto-tuning. DNN Weaver [83], Angel-Eye [44], ALAMO [67], FP-DNN [43], SysArrayAccel [101] are typical FPGA DL compilers targeting the processor architecture. What's more, the PUs and PEs are usually responsible for coarse-grained basic operations such as matrix-vector multiplication, matrix-matrix multiplication, pooling, and some element-wise operations. The optimizations of these basic operations are mainly guided by the tradeoff between the parallelism and data reuse, which is similar to general optimizations.

The streaming architecture has similarities with pipelines. An FPGA accelerator of this architecture consists of multiple different hardware blocks, and it nearly has one hardware block for each layer of an input DL model. With the input data of a DL model, this kind of accelerators process the data through the different hardware blocks in the same sequence with layers. Additionally, with

Table 1. The detailed comparison of popular DL compilers.

	TVM	TC	Glow	nGraph+PlaidML	XLA
Core/Programming Language					
Core	C++	C++	C++	C++	C++
Programming	Python/C++	Python/C++	Python/C++	Python/C++	Python/C++
Supported Hardware Targets					
CPU	✓	✓	✓	✓	✓
NVIDIA-GPU	✓	✓	✓	✓	✓
AMD-GPU	✓	×	✓	✓	✓
FPGA	✓	×	×	×	×
TPU	×	×	×	×	✓
NNP	×	×	×	✓	×
Customed	✓	×	✓	✓	✓
Supported DL Frameworks					
TensorFlow	✓	×	×	✓	✓
PyTorch	✓	✓	✓	×	✓
MXNet	✓	×	×	×	×
Caffe2	✓	✓	✓	×	×
ONNX	✓	×	✓	✓	×
CoreML	✓	×	×	×	×
Keras	✓	×	×	✓	✓
PaddlePaddle	×	×	×	✓	×
DarkNet	✓	×	×	×	×
Supported Generating Languages					
CUDA	✓	✓	×	×	✓
OpenCL	✓	×	✓	✓	✓
Metal	✓	×	×	✓	×
LLVM	✓	✓	✓	✓	✓
OpenGL	✓	×	×	✓	×
Supported Features/Strategies					
AOT	✓	×	✓	✓ official release	✓
JIT	✓	✓	✓	✓	✓
Training	—	✓	✓	✓	✓
Quantization	—	×	✓	✓	×
Automatic Differentiation	—	✓	✓	✓	×
Dynamic Shape	✓	×	×	✓	×
Auto-tuning	✓	✓	×	✓ only tiling	×

the streaming input data, all hardware blocks can be fully utilized in a pipeline manner. However, the streaming architecture usually follows an initial assumption that the on-chip memory the computation resources on target FPGA are sufficient to accommodate the DL models, which bring barriers to deploy deep models with complicated layers. The DL compilers targeting this architecture can solve this problem by leveraging the reconfigurability of FPGA or adopting dynamic control flow. And the further optimization of a single block resembles that of basic operations of the processor architecture. fpgaConvNet [90], DeepBurning [98], Haddoc2 [17], and AutoCodeGen [65] are typical corresponding DL compilers.

For the detailed survey of specific compilation techniques that map DL models to FPGAs, the readers can refer to [45, 91, 106]. Different from [45, 91, 106], this survey focuses on general DL compilation techniques that can be applied to broader DL hardware other than bounding to FPGA.

3 COMPARISON OF DL COMPILERS

In this section, we compare several popular DL compilers including TVM [29], TC [89], Glow [79], nGraph [34], PlaidML [12], and XLA [60]. Table 1 shows the detailed comparison of different DL compilers from various aspects, where "✓" means supported, "×" means not supported, and "—" means under development. Note that we use TVM to represent the work of VTA [71], Relay [78] and autoTVM [30]. In addition, PlaidML is tightly coupled with nGraph, therefore we consider

them together during the comparison. Besides, for the performance comparison of DL compilers, the readers can refer to [103].

Core/Programming Language - The core language of all DL compilers is C++, because C++ highlights performance, efficiency, and flexibility of use in its design. However, Python is becoming more and more popular with programmers due to its simplicity and usability. For most mature DL compilers (e.g., TVM, TC, nGraph, PlaidML and XLA), their Python interfaces almost cover all core functions.

Supported Hardwares - All DL compilers support Intel and AMD CPUs as well as NVIDIA GPUs. The official versions of TC currently do not provide support for AMD GPUs. Note that nGraph is integrated with PlaidML to provide acceleration of more hardware targets. nGraph can support the target hardware by invoking existing kernel libraries (e.g., cuDNN and MKL-DNN). Additionally, PlaidML offers extensive support for various hardware targets due to its ability to generate code. TVM can map a workload to FPGAs using the VTA architecture and runtime [71]. The native supported dedicated DL chips of a DL compiler are related to its developer generally. For example, nGraph can support Intel Nervana Neural Network Processors (NNP) by invoking the NNP library. XLA can support Google TPUs by directly generating binary files. MLIR can take advantage of XLA's compilation abilities by using XLA HLO IR as its dialect. All DL compilers except TC and nGraph can support customized hardware targets by developing interfaces based on LLVM. For example, Glow uses automatic code generation techniques (i.e., ClassGen, which is based on LLVM) for defining instructions and nodes [79], and compiler researchers can invoke the interfaces to support new hardware targets.

Supported DL Frameworks - Currently TensorFlow and PyTorch are the two most popular DL frameworks. There are three approaches to support DL frameworks: 1) the DL compiler is integrated into the DL framework; 2) the DL framework has launched an official package to support the DL compiler; 3) the DL compiler uses a converter to deploy the DL models. Here are examples to illustrate these three approaches. For 1), XLA is integrated with TensorFlow, while TC and Glow provide lightweight integration with PyTorch and Caffe2. For 2), PyTorch stands to benefit from directly leveraging the compiler stack. To that end, PyTorch now has official TVM-based and XLA-based packages (i.e., torch_tvm and torch_xla). Compared to 1) and 2), 3) is more common. For instance, nGraph supports TensorFlow and PaddlePaddle by using "bridge" to maintain their programmatic or user interface. TVM can deploy the models generated by DL frameworks and then optimize the performance of model inference. As more and more DL frameworks support exporting ONNX models (Section 2.1), it is important for DL compilers to support ONNX for the future development. At present, three DL compilers (i.e., TVM, Glow and nGraph) are able to load, compile and execute the pretrained ONNX model.

Supported Code Targets - All DL compilers use LLVM as their low-level IR. The advantages of LLVM over other low-level compilers (e.g., GCC and ICC) are the unified IR, high modularity, and rapid customization. With LLVM, compiler researchers can quickly write optimized *Passes* for domain-specific applications and generate a variety of target code (e.g., ARM, x86, PTX) through *TableGen* module. As shown in Table 1, ngraph can only generate LLVM code for the CPU backend. Both CUDA and OpenCL are used to implement heterogeneous parallel computing. OpenCL can be used to program both NVIDIA and AMD GPUs, while CUDA is specific to NVIDIA GPUs. Although OpenCL promises a portable language for GPU programming, its generality may entail a performance penalty. Two DL compilers (i.e., TVM and XLA) support generating both CUDA and OpenCL code. Only TVM and PlaidML support generating codes of OpenGL, which is a cross-platform API that deals with rendering graphics. However, at WWDC 2018, Apple announced the deprecation of OpenGL and OpenCL on the new system. Instead, Apple uses Metal API for both

graphics rendering and general-purpose computing. Currently only TVM and PlaidML support generating Metal code.

Supported Compilation - All DL compilers support just-in-time compilation (JIT) to improve the efficiency of program execution. Four DL compilers (i.e., TVM, Glow, nGraph, and XLA) support ahead-of-time compilation (AOT), of which nGraph only supports AOT in the official release (not supported in Beta release). The AOT compiler of TVM/Relay produces a native library given a Relay expression and dynamically loads it in Python. Glow can produce ahead-of-time compiled executable bundles, which are self-contained compiled network models that can be used to execute in a standalone mode. XLA uses `tfcompile` to compile TensorFlow graphs into executable code. Instead of adding automatic differentiation (Autograd) support for XLA, Google propose JAX [24], which bring Autograd and XLA together for high-performance machine learning research.

Supported DL Optimizations - As for low-bit inference, currently four DL compilers (i.e., TVM, Glow, nGraph and MLIR) support quantization. At present, XLA alone cannot solve the problem of quantization: the quantization rewriter has a missing part when the rewritten TensorFlow graph is reduced to a quantized XLA graph. TVM's automatic differentiation, quantization, and training are still under development. Moreover, a count of operators with gradient support are available in TVM v0.6 release. Supporting dynamic shapes requires changing the runtimes, which is a big challenge for DL compilers. At this time two DL compilers (i.e., TVM and nGraph) support dynamic shapes. TC and XLA only support static dimensions internally to provide automatic shape and bound inference. TVM and TC support auto-tuning to optimize performance by tuning the available mapping options. TC can only perform auto-tuning on NVIDIA GPUs, while TVM can apply auto-tuning to CPUs (x86 and ARM), mobile GPUs and NVIDIA GPUs [30]. TVM and TC use different tuning approaches: TC uses genetic search [41] and TVM uses two machine learning models (i.e., GBT and TreeGRU). PlaidML can only apply auto-tuning to tiling (auto-tiling), which explores a space of tile sizes using a hypothetical cost model [105].

4 COMMON DESIGN OF DL COMPILERS

4.1 Design Overview

The common design of a DL compiler primarily contains two parts: the compiler frontend and the compiler backend, as shown in Figure 2. The intermediate representation (IR) is spread across both the frontend and the backend. Generally, IR is an abstraction of the program, and is used for program optimizations. Specifically, the DL models are translated into multi-level IRs in DL compilers, where the high-level IR resides in the frontend and the low-level IR resides in the backend. The IR implementations in different DL compilers are listed in Table 2. Based on the high-level IR, the compiler frontend is responsible for hardware-independent transformations and optimizations. Based on the low-level IR, the compiler backend is responsible for hardware-specific optimizations, code generation, and compilation. The functions of the frontend, the backend, and multi-level IRs in DL compilers are described briefly as follows:

The high-level IR also known as graph IR, represents the computation and the control flow, and is hardware-independent. The design challenge of high-level IR is the ability of abstraction of the computation and the control flow, which is able to capture and express diverse DL models. The goal of the high-level IR is to establish the control flow and the dependency between the operators and the data, as well as provide an interface for graph-level optimizations. It also contains rich semantic information for compilation as well as offers extensibility for customized operators. The detailed discussion of high-level IR is presented in Section 4.2.

The low-level IR is designed for hardware-specific optimization and code generation on diverse hardware targets. Thus, the low-level IR should be fine-grained enough to reflect the hardware

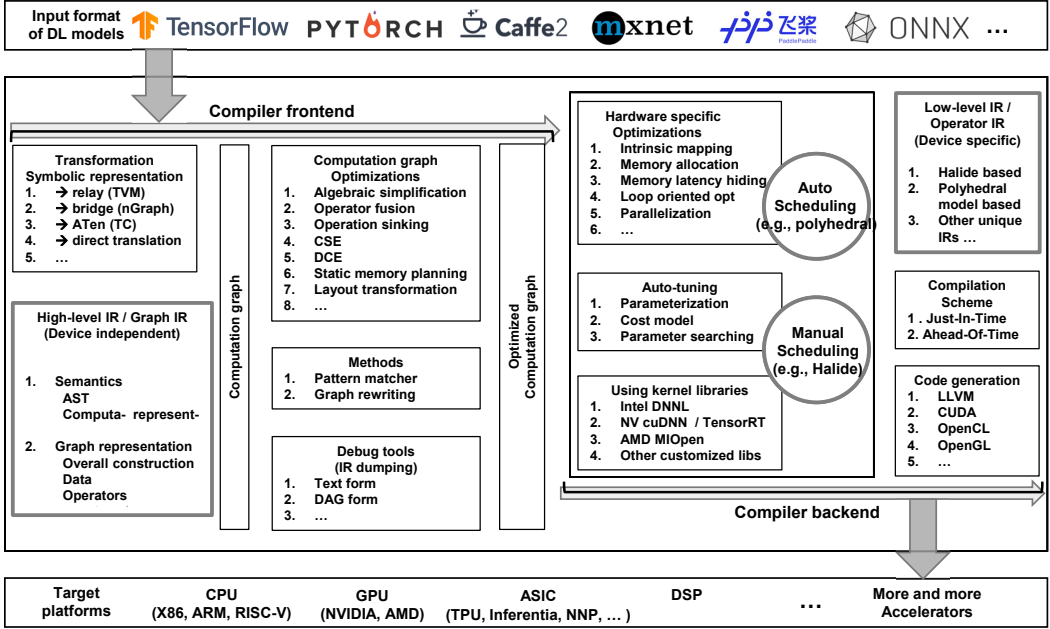


Fig. 2. The overview of commonly adopted design of DL compilers.

characteristics and represent the hardware-specific optimizations. It should also allow the use of mature third-party tool-chains in compiler backends such as Halide [77], polyhedral model [13], and LLVM [59]. The detailed discussion of low-level IR is presented in Section 4.3.

The frontend takes a DL model from existing DL frameworks as input, and then transforms the model into the computational graph representation (i.e., graph IR). To support the diverse formats in different frameworks, the frontend needs to implement various format transformations. The computational graph optimizations incorporate the optimization techniques from both general-purpose compilers and the DL specific optimizations, which reduce the redundancy and improve the efficiency upon the graph IR. Such optimizations can be classified into node-level (e.g., nop elimination and zero-dim-tensor elimination), block-level (e.g., algebraic simplification, operator fusion, and operator sinking) and dataflow-level (e.g., CSE, DCE, static memory planning and layout transformation). After the frontend, the optimized computation graph is generated and passed to the backend. Note that, some DL compilers further transform the optimized computation graph to operation IR. The detailed discussion of frontend is presented in Section 4.4.

The backend transforms the high-level IR into low-level IR and performs hardware-specific optimizations at the same time. On one hand, it can directly transform the high-level IR to third-party tool-chains such as LLVM IR to utilize the LLVM infrastructure for general-purpose optimizations and CPU/GPU code generation. On the other hand, it can take advantage of the prior knowledge of both DL models and hardware characteristics for more efficient code generation, with customized compilation passes. The commonly applied hardware-specific optimizations include hardware intrinsic mapping, memory allocation, and fetching, memory latency hiding, parallelization as well as loop oriented optimization. To address the large solution space introduced by the above optimizations, two approaches are widely adopted in existing DL compilers such as auto-scheduling (e.g., polyhedral model) and auto-tuning (e.g., AutoTVM). The optimized low-level IR is compiled

Table 2. The IR implementations of different DL compilers.

DL compiler	High-level IR	Low-level IR	Code generation
TVM	Relay	Halide-based IR	LLVM + Code
TC	TC IR	Polyhedral based IR	LLVM + Code
XLA	HLO		LLVM + Code
Glow	High-level IR	Low-level IR	LLVM
nGraph	nGraph IR		LLVM
PlaidML	Tile	Stripe	LLVM + Code

using JIT or AOT to generate codes for different hardware targets. The detailed discussion of backend is presented in Section 4.5.

4.2 High-level IR

To overcome the limitation of IR adopted in traditional compilers that constrains the expression of complex computations used in DL models, existing DL compilers leverage graph IR with specially designed data structures for efficient code optimizations. To better understand the graph IR used in the DL compilers, we describe the semantic and representation of graph IR as follows.

4.2.1 Semantic of Graph IR.

We present the semantic of graph IR from the perspective of programming language (PL). The AST of graph IR influences the expressive power of graph IR and also shows how the DL compilers analyze the graph IR code. Besides, different graph IRs have their approaches to describe tensor calculation. These approaches should be user-friendly as well as extensible of computation representation. We describe the **AST of graph IR** first in the following.

- **DAG-based IR:** DAG-based IR is one of the most traditional ways for the compilers to build a computational graph, and its nodes and edges will be organized to a directed acyclic graph. There are rich optimization algorithms on the DAG computational graph, such as live variable analysis and variable dependency analysis. DAG-based IR is convenient for programming and compiling because of its simplicity, but it has deficiencies in other filed such as semantic ambiguity caused by the undefinition of computation scope [5].
- **Bind-based IR:** Let-binding is one method to solve the semantic ambiguity above by offering Let expression to certain functions with restricted scope used by many programming languages such as F# [76] and Scheme [57]. Let-binding will change the AST structure of IR. When using Let keyword to define an expression, a let node will be generated and points to the operator and variable in the expression instead of just building computational relation between variables as DAG. When a process needs to get the return result of one expression, the DAG-based compiler will first access the add node and search nodes related to it recursively. In contrast, the Let-binding-base compiler will figure every result of the variable in Let expression and build the variable map. When results are needed, the compiler will loop up a map to decide the result of the expression.

Computation representation - Different graph IRs have different representations of the computation on tensors. The compilers will translate framework operators according to their specific representation forms. Their customized operators also need to be programmed in those representation forms. And the forms can be divided into three categories as follows.

<pre> m, n, h = t.var('m'), t.var('n'), t.var('h') A = t.placeholder((m, h), name='A') B = t.placeholder((n, h), name='B') k = t.reduce_axis((0, h), name='k') C = t.compute((m, n), lambda y, x: t.sum(A[k, y] * B[k, x], axis=k)) </pre>	<pre> def mv(float(M,K) A, float(K) x) -> (C) { C(i) = 0 C(i) += A(i,k) * x(k) } </pre>
(a) lambda expression from TVM	(b) Einstein notation from TC

Fig. 3. Lambda expression and Einstein notation

- **Function-based form:** The function-based form just provide encapsulated operators, and HLO (the IR of XLA) adopts this form. So we take HLO for example to describe the function-based form. HLO IR is a PL for symbolic programming. The IR consists of a set of functions, and most of them have no side-effect. So it is hard to retrieve intermediate values used in the computation process. The instructions are organized into three levels: HloModule, which is similar to a whole program, HloComputaion, which is similar to a function, and HloInstruction, which represents the essential operation. XLA uses HLO IR to represent both graph IR and operation IR so that the operation of HLO ranges from the node-level to the calculation-level.
- **Lambda expression:** TVM represents operation IR by a lambda expression (shown in Figure 3(a)) [29], an index formula expression, and TVM offers a language called tensor expression through the lambda expression. Lambda expression describes calculation by variable binding and substitution, which widely exists in programming languages such as Python and C++11. In TVM, computational operators in tensor expression are composed of two parts: the shape of output tensor and the lambda expression of computing rules.
- **Einstein notation:** The Einstein notation, also called the summation convention, is adopted by TC. The Einstein notation is a notation to express summation. Take the definition of matrix-vector multiplication (shown in Figure 3(b)) [89] for example, temporary variables for index do not need to be defined. TC IR can figure out the actual expression by the occurrence of undefined variables. For example, in the second line of Figure 3(b), the variable k only exists on the right, so the operator $+$ will perform the reduction operator over the dimension represented by k . In the Einstein notation expression, the operator needs to be associative and commutative, and operators such as max are not supported. This restriction promises the reduction operator can be executed by any order, making it possible for further parallelization.

4.2.2 Graph Representation.

The graph representation in DL compilers adopts the approach of traditional computational graph, which is composed of data and operator.

Data organization - DL compilers always establish their computational graph by multi-level node. Taking Relay as an example, Figure 4(a) shows the data structure used to represent the computational graph. Relay divides its whole computational graph into smaller graphs called schedule nodes for the organization. A schedule node is composed of many stage nodes, and it forms a DAG graph. Each stage node has precisely one operator node to represent a computational operator, such as ComputeOp or TensorComputeOp. Operator node is also used to originate the relation between iterator variables.

Besides, Relay combines the iterator variables in their graph data structure for bound inference. Relay organizes iterator variables hierarchically by its multi-level node mentioned above, and

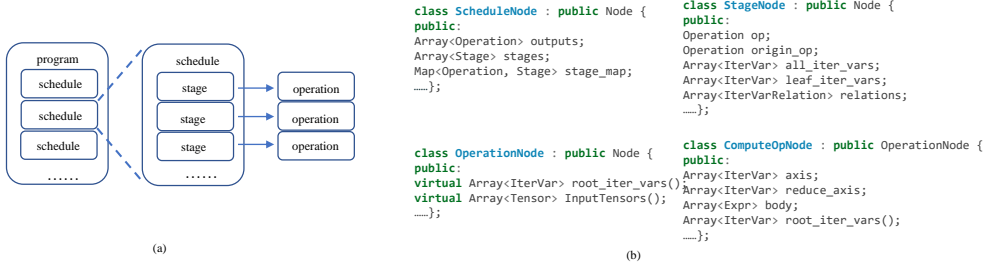


Fig. 4. The data organization of graph IR in TVM

Figure 4(b) [5] shows the data structure of iterator variables kept by the related node at each level. In the schedule level, there will keep a bound map for necessary iterator variables, which can be obtained by every stage node. Each stage node manages its iterator variables. Iterator variables in one stage node can be classified into leaf variables and root variables. The former determines the ranges in specific operators and the latter figures out the loop dimension.

Bound inference - The compiler needs to ensure the range of every iteration variable when performing optimizations. However, some computation representation does not intuitively point the scope of loop variables. Instead, those computation forms only offer information about the input and output. Thus, the bound inference process is needed to determine the variable ranges. The bound inference algorithm can be divided into two classes by the inference direction as follows:

- From output to input: The inference of TVM is from output to input. It is related to its expression form, tensor expression, which needs the shape of output directly. Relay handles the inference process by a pass called InferBound pass, and Figure 5 shows the pseudo-code and the schematic diagram. The input is Schedule node, which is consisted of a series of stage nodes. First, Relay will establish its dataflow graph whose nodes are stage nodes. Then, a reverse topological sort is performed to gain a list that guarantees the consumer nodes are before input nodes for it. That gives the inference process a primary hypothesis: when it comes to one stage, the shape of its output is already known. For each node, the inference pass will determine the range of root iterator variables first, using the shape of its output tensor. Then, Relay will figure out the leaf iterator variables by the relation of iterator variables and their relation recorded by stage node. The relation includes *split*, *fuse*, and *rebase*. The dependency of variables organized by that relation can be analyzed easily. Considering *split* node as an example. When facing a given factor, the ranges of a variable will be divided into (0, factor) and (factor, extend) if no cross-border. The bound results of root variables will be kept in the global map for other stage nodes to use.
- From input to output: Einstein notation (used by TC IR) makes the programming more concise by inferring index ranges automatically, but it is based on the understandable rules of inference. Just as mentioned above, there are operators that the Einstein notation cannot deal with and needs the user to add *where* clause to guide the compiler. If one needs to take lots of time to consider which expression should use the *where* clause, he would prefer to write the loop explicitly and give up the TC IR. Trying to avoid that embarrassment, TC adopts a straightforward approach in the inference algorithm, which is inspired by Presburger arithmetic implemented in polyhedral libraries [95]. The problem solved by inference is seeking ranges of all undefined variables as large as possible under the constraint that

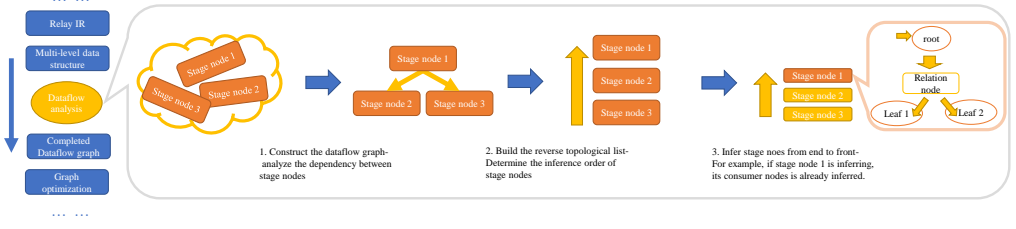


Fig. 5. Bound inference in Relay.

indexes cannot get over the bounds of input. The process is iterative. First, the algorithm initializes a list for all unsolved variables which are not constrained by *where* clause. Then, the algorithm will find one expression that only has a single unresolved variable and use Halide function *solve_for_inner_interval* to find the maximum range under the existing conditions for this variable. Then the solved variable will be removed from the unsolved list. The algorithm repeats the second step until the list is empty. If the expression with a single unresolved variable does not exist and the unsolved list is not empty, the compiler will stop inference and request for *where* from the user. Another challenge for inference is the recurrent RNN, which is not supported by the first release, making TC fails in implementing RNN.

Data representation - The representation of data in DL compilers includes input, weights, and variables. Some compilers represent data directly by specific pointers, while others use Placeholder or other designs to point to data indirectly. The data representation contains not only specific values but also other related information. Due to the requirement of graph optimization, mainstream DL compilers prefer to add data layout information to the data representation. The data representations of different compilers are shown as follows.

- **Placeholder:** Placeholder is widely used in symbolic programming. The compiler using placeholders has no information about the concrete tensor. Besides, it is convenient for the programmer to change the input and output shape by using placeholders instead of changing the whole semantics. During the compiling process, the placeholders can be replaced with different tensors without changes of semantic. TVM uses placeholders in its tensor expression language as a method to construct an empty tensor object.
- **Unknown shape representation:** TVM uses *Any* to represent an unknown dimension in order to support the dynamic model. The *Any* can be used in tensor type definition as one of the dimensions of the tensor shape. Unlike placeholder, the shape contains *Any* is unknown at compilation time, so the work on inference and checking dimension need to be relaxed, and additional shape function is necessary for guaranteeing memory validity.
- **Data layout:** There are three perspectives on the data layout of compilers: from operator view (TVM, Glow, PlaidML), from backend view (XLA), and tensor view (Relay, MLIR). The former regards data layout as additional or necessary parameters of operators and requires such information for computation and optimization. Combining data layout information with operator rather than tensor enables intuitive implementation for certain operator, and compiler prefers to regard the data layout as a parameter rather than a field for operator in order to reduce dynamic consumption during compilation [4]. XLA regards data layout as constraints related to its backend device: CPU and GPU. For example, when a constant array is used by dot operator only in CPU backend, XLA will require the array to be column-major. Relay and MLIR are going to add layout information into the type system for tensor.

Operators supported - The computational operators provided by DL compilers can be divided into three levels: basic calculation nodes, high-level calculation nodes, and fused calculation nodes. The basic nodes are the components of the whole operational space, such as exponent arithmetic. The high-level nodes include the units of the neural network models such as convolution and pooling. And the fused nodes represent the nodes fused by the graph optimizations.

Besides, some compilers also support customized operators, which also belongs to computational operators. The support of customized operator is one of the most basic designs for extensibility, which allows users to define their operators for new backend instructions or just for technical needs. The support of the customized operators can be divided into two levels: node level and operator level.

- Node level: Glow supports customized operators of both graph IR node and operation IR instruction. Glow suggests using existing IR to avoid implementing new operators. For defining new graph IR, in addition to finish logic implementation, the extra efforts are expected for node encapsulation, including node registering, layout specifying, and operator-load completing. Besides, a user also needs to implement a lower step, operation IR generation, and instruction generation if necessary.
- Operator level: TVM and TC have better expansibility on operators in function level with little efforts besides describing computation behavior. TVM regards external functions as black boxes, and supports customized operator calls natively [1]. Specifically, the users only need to describe the computation and declare the shape of output tensor when calling external functions. Users can also hook python functions as external functions, which makes TVM much more flexible when facing the challenges on the customized operators.

There are a large number of operators supported in DL compilers. Here we choose four operators that are representative and most frequently used across different DL compilers for illustration.

- Broadcasting: Without broadcasting operators, the input shape of the operator would be stricter. For example, in an *add* operator, the input tensors are expected to be of the same shape. Some compilers like XLA and Relay relax the restrictions, which offer an operator called broadcasting. For example, XLA allows the element-wise addition to acts on a matrix and vector by replicating the vector until its shape is the same as the matrix.
- Predication: Original prediction is a widely used method for deciding the code to execute by judging a condition before execution. The condition is always be pointed with a Boolean flag, and the false means related code should not be executed. Predication can also be used as a method in the control flow, such as avoiding the illegal division by non-zero determination. Compilers may redefine the semantic of predication due to different implementations. Glow regards predication as an optimization, and it means that the backend can ignore the content if the execution of predication does not accelerate the program [4]. The accelerating scene can be found in RNN implement by avoiding some computation brought by the different batch sizes. However, Glow hypothesizes that the prediction does not affect the main semantic of the program, which means that the prediction in Glow is not allowed in codes related to the control flow. TVM provides an *assert* statement that fits Halide IR as well as Python AST [5].
- Automatical differentiation: Glow differentiates its computational graph automatically by lowering the gradients and high-level differentiation operators like stochastic gradient descent (SGD) into low-level operators such as additive and multiplicative. By doing that, Glow can support the backends without the implementation of complex operations for training. Inspired by the reverse-mode [75], Relay represents partial derivative by specific type, and the related gradient node is rewritten by transforming the inner AST.

- **Control flow:** In graph IR, control flow is needed while representing complex and flexible models like RNN. Relay notices that arbitrary control flow can be implemented by recursion and pattern, which has demonstrated by the functional programming community [78]. Therefore, Relay provides *if* operator and recursive function for implementing control flow, which means one needs to implement *while* loop by judgment and jump manually. On the contrary, XLA represents control flow by special HLO operators, *while* and *conditional*. XLA has already implemented the control flow operators by a series of basic HLO operations.

4.3 Low-level IR

4.3.1 Implementation of Low-Level IR. Low-level IR describes the computation of a DL model in a more fine-grained representation than that in high-level IR, which enables the target-dependent optimizations by providing interfaces to tune the computation and memory access. It also allows the developer to make use of mature third-party tool-chains in compiler backend such as Halide [77] and polyhedral model [13]. In this section, we classify the low-level IRs into three categories: Halide based IRs, polyhedral model based IRs, and other unique IRs.

Halide based IRs - Halide is firstly proposed to parallelize image processing, and it is proven to be extensible and efficient in DL compilers by TVM. The basic concept of Halide is the separation of *computation* and *scheduling*. Rather than giving a specific scheme directly, The compilers adopting Halide will try various possible *scheduling* and choose the best one. The original IR of the Halide needs to change on design and implementation when be used in DL compiling problem. For example, the shape of Halide's input is infinite, while the DL compilers need to know the exact shape of data in order to map the operator to hardware instructions. Some compilers, such as TC, need the fixed size of data, considering tensors in ML application have higher temporal locality than in general programs.

TVM has isolated the Halide based IR from the original Halide program and improved it into an independent symbolic IR by following efforts [7]. First, TVM removes the dependency on LLVM of the original Halide IR. The structure of both project module and IR overall design of the Halide have been reorganized, pursuing better logical and publicly accessible from graph IR and frontend language such as python. The re-usability is also considered in order to add customized operators easily by introducing a runtime dispatching mechanism. TVM changes the variable definition from string matching to pointer matching, guaranteeing single define location for each variable as the static single-assignment (SSA) [35], which is useful for further optimization.

Polyhedral model based IRs - The polyhedral model is another important reference for DL compilers. It uses linear programming and other mathematical methods to optimize loop-based codes where the control flow of bounds and branches is static. The polyhedral model based IR undergoes multiple polyhedral transformations (e.g., fusion, tiling, sinking, and mapping), including both device-dependent and device-independent optimization. When targeting a specific architecture, polyhedral transformations involve changes in scheduling and mapping strategies. The main optimization space in the polyhedral model is loop lowering, such as iteration step and loop level. The polyhedral model always uses various specific nodes to present the semantics of the program, which will be introduced below. Further optimizations are also based on such nodes. There are many toolchain which can be borrowed by Polyhedral-based compilers, such as isl [94], Omega [56], PIP [36], Polylib [66], and PPL [20]. Because of the ability to deal with numerous deeply nested loops, many DL compilers, such as Tensor Comprehension (TC) and PlaidML, use the polyhedral model with or without modifications.

TC has its unique design in low-level IR, which combines the Halide and polyhedral model. It uses Halide based IR to represent the computation, but it adopts the polyhedral model based IR to represent the loop structures. So the loop structures can be optimized using polyhedral techniques.

TC presents detailed expressions by abstracted instances and introduces specific node types, and some of them are similar to the original polyhedral model. The node and their function can be described as Table 3. In brief, TC uses the domain node to specify the ranges of index variables and uses context node to introduce new iterative variables that related to hardware, such as the size of block for GPU. A band node determines the order of iterations. A filter node represents an iterator combined with a statement instance. *Set* and *sequence* are keywords to specify the execution type for filters, *set* for parallelism, and *sequence* for serial execution. Besides, TC uses extension nodes to introduce other necessary instructions for code generation, such as the memory movement statement.

Node name	Function
Band node	Define partial execution order for iterative variables. The function in it is called band member or schedule dimension.
Filter node	Partition the iteration domain, can be specify execution type by <i>set</i> or <i>sequence</i> keywords.
Context node	Offer additional information related to hardware by introducing some variables and parameters.
Extension node	Introduce extra statement out of iteration space, providing convenience for optimization and code generation.

Table 3. The IR nodes used in TC that are based on Polyhedral model.

Stripe/PlaidML represents tensor operations through the nested polyhedral model. The nested polyhedral model creates a hierarchy of parallelizable code by extending the nesting of parallel polyhedral blocks to multiple levels. Besides, it allows nested polyhedral to be allocated to nested memory units, providing a way to match the computation with the caching structure of a multi-level hardware topology. The hardware configuration for Stripe is done independently of the kernel code. Stripe includes *tags* that signal to optimization passes. The *tags* do not change the kernel structure but provide additional information about the hardware target for the optimization passes. Stripe splits the machine learning operations into "tiles" that fit into local hardware resources.

Other unique IRs - There are DL compilers implement customized low-level IRs without using Halide and polyhedral model. Upon the customized low-level IRs, they apply hardware-specific optimizations and lowers to LLVM IR directly.

The low-level IR in Glow is an instruction-based expression that operates on tensors referenced by addresses [79]. Low-level IR can not only achieve target-independent optimization that not able to be achieved by high-level IR but also can represent target-specific operations based on low-level instructions (e.g., asynchronous DMA operations). Moreover, low-level IR allows the compiler to create a schedule to hide the latency of memory operations. There are two sections of instruction-based functions in low-level IR: *declare* and *program*. The first section declares a number of constant memory regions that live throughout the lifetime of the program (e.g., input, weight, bias). The second section is a list of locally allocated regions, including functions (e.g., conv and pool) and temporary variables. Instructions can run on the global memory regions (e.g., *declare*) or locally allocated regions (e.g., *program*). Besides, each operand is annotated with one of the qualifiers (i.e., "@in" / "@out" / "@inout"). "@In" indicates the operand reads from the buffer. "@Out" indicates that the operand writes to the buffer. "@Inout" indicates that the operand reads and writes to the buffer. These instructions and operand qualifiers help Glow determine when certain memory optimizations can be performed (e.g., copy elimination or buffer sharing). After

completing low-level IR optimizations, Glow performs hardware-specific optimizations and code generation through LLVM.

MLIR is highly influenced by LLVM and reuses many of its good ideas and interfaces. MLIR is more of a pure compiler infrastructure than LLVM. MLIR IR sits between the model representation and the low-level compiler that generates hardware-specific code. MLIR has a flexible type system and allows multiple levels of abstraction to be combined in the same compilation unit. MLIR introduces *dialects* to represent these multiple levels of abstraction. Each *dialect* consists of a set of defined operations that are immutable. The current *dialects* of MLIR include TensorFlow IR, XLA HLO IR, experimental Polyhedral IR, LLVM IR, and TensorFlow Lite. With the *dialect* of LLVM IR, MLIR can utilize the LLVM type system to define entirely custom types and operations. Furthermore, MLIR can create new *dialects* to connect to a new low-level compiler, which paves the way for hardware developers and compiler researchers.

The HLO IR of XLA is both high-level IR and low-level IR because HLO is fine-grained enough to represent the hardware-specific information. Besides, HLO supports hardware-specific optimizations and can be used to emit LLVM IR.

4.3.2 Code Generation based on Low-Level IR. All the DL compilers mentioned above could eventually be lowered to LLVM IR, and they benefit from LLVM's mature optimizer and code generator. Furthermore, LLVM can explicitly design custom instruction sets for specialized accelerators from scratch. However, traditional compilers may generate poor code when passed directly to LLVM IR. In order to avoid this situation, two approaches are applied by DL compilers to achieve hardware-dependent optimization: 1) perform target-specific loop transformation in the upper IR of LLVM (e.g., Halide based IR and polyhedral model based IR); 2) provide additional information about the hardware target for the optimization passes. Most DL compilers apply both approaches, but the emphasis is different. In general, the DL compilers that prefer frontend users (e.g., TC, TVM, XLA, and nGraph) might focus on 1); the DL compilers that are more inclined to backend developers (e.g., Glow, PlaidML, and MLIR) might focus on 2).

The compilation scheme in DL compilers can be mainly designed as two types: just-in-time (JIT) and ahead-of-time (AOT). For JIT compilers, it can generate executable codes on the fly, and they can optimize codes with better runtime knowledge. AOT is another approach for DL compilers, which generates all executable binaries first and execute them. AOT compilation can have larger scope in static analysis than JIT compilation, and thus benefits more thorough optimizations. And AOT approaches can be applied with cross-compilers of targeted platforms to support embedded platforms (e.g. C-GOOD [55]) or execution on remote machines (TVM RPC).

As for code generation on CPUs (X86 and ARM), the DL compilers (e.g., TVM, PlaidML, TC and Glow) emit the LLVM IR based on optimized low-level IRs and invoke LLVM for JIT compilation. As for that on GPU, XLA supports NVIDIA GPUs via the LLVM NVPTX, which also allows for JIT/AOT compilation of PTX code to native GPU machine code. However, TVM generates CUDA code according to the Halide based IR, and TC generates CUDA code according to the polyhedral model based IR. As for other customized accelerators, the DL compilers usually generate the corresponding code for the AOT compilation, for example, PlaidML generates OpenCL code for AMD GPUs.

4.4 Frontend Optimizations

After constructing the computational graph, the frontend applies graph-level optimizations. Many optimizations are easier to be discovered and performed at graph level because the graph provides a global overview of the computation. These optimizations are only applied to the computational

graph, rather than the implementations on backends; thus, they are hardware-independent, which means that computational graph optimizations can be applied to various backend targets.

In traditional compilers, the IR of code is composed of fine-grained three-address instructions, and the basic blocks are maximal sequences of consecutive three-address instructions. The basic blocks become nodes of a dataflow graph whose edges indicate the execution order of the basic blocks. Thus, two scopes of optimizations, including local (block-level) optimizations and global (dataflow-level) optimizations, are implemented.

However, in DL compilers, the IR of DL models is composed of coarse-grained operation nodes. A node represents an operation on tensors, and an edge represents a dependency between two operations. The nodes are coarse enough to enable optimizations inside a single node. Several adjacent nodes can be treated as a node block, and optimizations can be applied inside a block. These optimizations are usually defined by *passes*, and can also be applied by traversing whole nodes of the computational graph and performing the graph transformations. The frontend provides methods to 1) capture the specific features from the computational graph and 2) rewrite parts of the graph for optimization. Besides the pre-defined *passes*, the developers can also define customized *passes* through the frontend.

In this section, we classify the frontend optimizations into three categories: 1) node-level optimizations (e.g., Zero-dim-tensor elimination, Nop elimination), 2) block-level (peephole, local) optimizations (e.g., algebraic simplification, fusion), and 3) dataflow-level (global) optimizations (e.g., CSE, DCE). We describe the detailed optimizations in each category as follows.

4.4.1 Node-level optimizations. Most DL compilers can determine the shape of both input tensors and output tensors of every operation once a DL model is imported and transformed as a computational graph. This character allows dl compilers to perform optimizations according to the shape information. The node-level optimizations have two types, node elimination, which eliminates unnecessary operation nodes, and node replacement, which replace operation nodes with other lower-cost nodes.

Nop Elimination removes the no-op instructions which occupy a small amount of space but specify no operation in general-purpose compilers. But in DL compilers, Nop Elimination is responsible for removing the operations lacking adequate inputs. For example, the sum operation with only one input tensor can be eliminated, the padding operation with zero padding width can be eliminated.

Zero-dim-tensor elimination is responsible for removing the unnecessary operations whose inputs are zero-dimension tensors. Assume that A is a zero-dimension tensor, and B is a constant tensor, the sum operation node of A and B can be replaced with the already existing constant node B without affecting the correctness. Assume that C is a 3-dimension tensor, but the shape of one dimension is zero, such as $\{0,2,3\}$, therefore, C has no element, and the argmin/argmax operation node can be eliminated.

4.4.2 Block-level optimizations. Algebraic simplification - The algebraic simplification optimizations consist of optimizations in three parts in both DL compilers and general-purpose compilers: (1) algebraic identities, for example, with which we can change the computation of $x \times 1$ to x ; (2) strength reduction, with which we can replace a more expensive operator by a cheaper one, such as replacing $x/2$ with $x \times 0.5$; (3) constant folding, with which we can replace the constant expressions by their values, such as replacing the expression $x \times 2 \times 3$ with their final value $x \times 6$. They consider a sequence of operation nodes, then take advantage of commutativity, associativity, and distributivity of different kinds of operators to simplify the computation.

Despite the typical operators (+, −, ×, ÷, etc.) described above, the algebraic simplification optimizations can also be applied to DL specific operations, such as reshape, transpose, and pooling, thus the operators can be reordered and sometimes eliminated, which reduces the redundancy and

improves the efficiency. Here we show several common cases: 1) *Optimization of reshape/transpose nodes* - This optimization finds and removes reshape/transpose operations according to more specific characters than algebraic simplification. Take the matrix multiplication (GEMM) for example, there are two matrices (e.g., A and B), both matrices are transposed (to produce A^T and B^T , respectively), then A^T and B^T are multiplied together. However, a more efficient way to implement GEMM is to switch the order of the arguments A and B , multiply them together, and then transpose the output of the GEMM. Effectively, this approach cuts two transpose operations down to just one, where this optimization can do that rewriting; 2) *Optimization of transpose nodes* - This optimization combines multiple consecutive transpose nodes into a single node, eliminates identity transpose nodes, and optimizes transpose nodes into reshape nodes when they actually move no data; 3) *Optimization of pool nodes* - This optimization swaps the order of adjacent pooling node and ReLU node so that the ReLU operator can operate on the smaller tensor. Because the ReLU operation occupies a small fraction of whole computations, this optimization may bring slight performance improvement; 4) *Optimization of ReduceMean nodes* - This optimization performs substitutions of ReduceMean with AvgPool node (in Glow) if the input of the reduce operator is 4D with the last two dimensions to be reduced.

Operator fusion - Operator fusion is indispensable building block of dl compilers, and exactly it's an essential optimization in dl compilers. It enables better sharing of computation, removal of intermediate allocations, facilitates further optimization by combining loop nests [78], as well as reduces launch and synchronization overhead [89]. In TVM, the operators are classified into four categories: 1) injective (one-to-one map), 2) reduction (complex-in-fusible) 3) complex-out-fusible (can fuse element-wise map to output, e.g., conv2d), and 4) opaque (cannot be fused, e.g., sort). When the operators are defined, their corresponding categories are determined. To that end, TVM designs several rules to perform the fusion: 1) multiple injective operators can be fused into a new injective operator, 2) a reduction operator can be fused with input injective operators before the input, 3) complex-out-operators can be fused with injective operators after the output. According to the operator categories and fusion rules, fusion can be performed conveniently. In Tensor Comprehension, fusion is conducted in a different way based on the automatic polyhedron transformations. However, how to identify and fuse more and more complicated graph patterns, such as blocks with multiple broadcast and reduce nodes, remains to be a problem.

Operator sinking - This optimization sinks operations such as transposes below operations like a batch normalization, ReLU, sigmoid, channel shuffle. By doing this, many similar operations are moved around and closer to each other, which creates more opportunities for the algebraic simplification.

4.4.3 Dataflow-level optimizations. Common Sub-expression Elimination (CSE) - An expression E is a common sub-expression if the value of E is previously computed, and the value of E has not be changed since previous computation [19]. In this situation, the value of E is computed once, and the already computed value of E can be used to avoid recomputing in other places. The DL compilers search for common sub-expressions through the whole computation graph and replace the following common sub-expressions with the previously computed result. Both programmers and compilers optimizations could lead to common sub-expression.

Dead Code Elimination (DCE) - A set of code is dead if its computed results or side-effects are not used, and the DCE optimization removes the dead code. The dead code is usually not caused by programmers but caused by other graph optimizations. Thus, the DCE, as well as CSE, are applied after other graph optimizations. Other detailed optimizations, such as the dead store elimination (DSE), which removes stores into tensors but the tensors are never going to be used, also belong to DCE.

Static memory planning - Static memory planning optimizations are conducted to reuse the memory buffers as much as possible. Usually, there are two approaches: in-place memory sharing and standard memory sharing. The in-place memory sharing uses the same memory for input and output for an operation, and just allocate one copy of memory before computing. Standard memory sharing finds other ways of memory sharing, and it reuses the memory of previous operations without overlapping. Both approaches have dedicated conditions that should avoid. The problem of memory planning is similar to register allocations in general-purpose compilers. The static memory planning is done offline, which enables more complicated planning algorithms.

Layout transformation - This optimization tries to find the best data layouts to store tensors in the computational graph and then insert the layout transformation nodes to the computation graph. Note that the real transformation is not performed here, but it should be performed when evaluating the computational graph by the compiler backend. For example, the transformation between the channel first format (NCHW) and the channels last format (NHWC) is a typical layout transformation.

In fact, the performance of the same operation in different layouts is different, and the best layouts are different on different hardware (e.g., CPU, GPU, and other customized accelerators). For example, operations in the NCHW format on GPU usually run faster, so it's efficient to transform to NCHW format on GPU (actually, that is what the layout optimizer of TensorFlow always tries to do). Some DL compilers may rely on calling hardware-specific libraries to achieve higher performance, but the libraries may require certain layouts. Besides, some DL accelerators may prefer more complicated layouts (e.g., tile). Therefore, the DL compilers face to various hardware, and they need to provide a way to perform layout transformations.

Not only the data layouts of input, output, and intermediate tensors have a nontrivial influence on the final performance, but also the transformation operations have a significant overhead. Because the transformation operations also consume the memory and computation resource.

A recent work [64] based on TVM targeting on CPUs alters the layout of all convolution operations to NCHW[x]c first in the computational graph, in which c means the split sub-dimension of channel C and x indicates the split size of the sub-dimension. Then all x parameters are globally explored by auto-tuning when providing hardware details, such as cache line size, vectorization unit size, and memory access pattern, which belongs to hardware-specific optimizations.

4.4.4 Case Study with Tensorflow XLA. To illustrate the computational graph optimizations concretely, we dump the HLO graph before and after each *pass* in Tensorflow XLA. We choose Alexnet model as the input of the XLA compiler and Volta GPU as the target hardware. The optimizations are shown in Figure 6. For simplicity, we remove the data layout information of some nodes. The algebraic simplification includes reducing the consecutive transpose and reshape nodes into a single reshape node, as well as replacing the reshape node into a bitcast node. The CSE reuses the broadcast node. The cuDNN transformation transforms the convolution node into a function call (convForward) of cuDNN to enable the graph optimizations leverage the cuDNN library. The constant folding transforms the neighboring convolution (convForward) and adds nodes into a convolution with bias (convBiasActivationForward). And the operator fusion fuses several bitcast nodes and an add node. Note that, the implementation of frontend optimizations in DL compilers (e.g., XLA) consists of several stages. Therefore, the optimizations are performed several times, which may change the computational graph each time, and thus introduce more opportunities for further optimizations.

4.4.5 Discussion. The frontend is one of the most important components in DL compilers, which is responsible of transformation from DL models to high-level IR (i.e., computational graph) and hardware-independent optimizations based on high-level IR. Although the implementation of

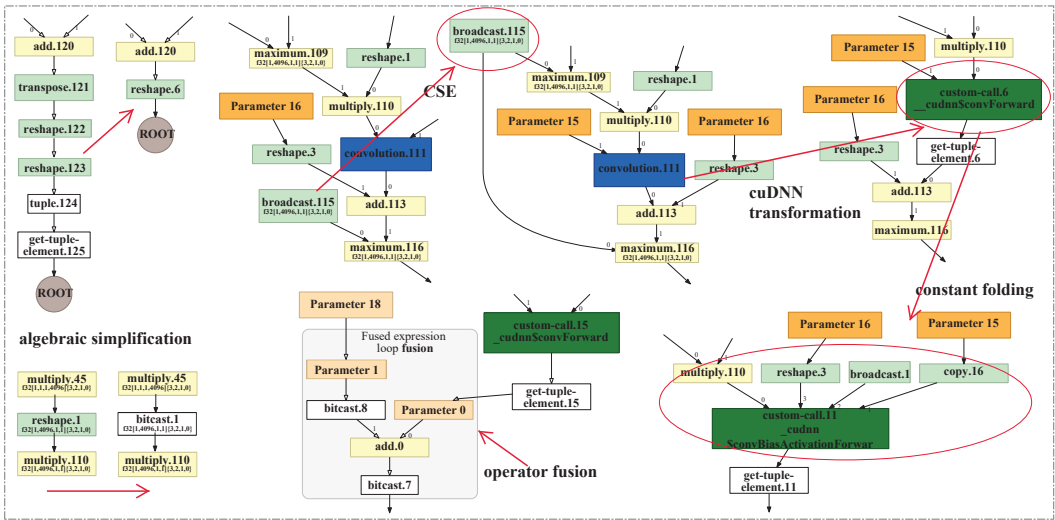


Fig. 6. Example of computation graph optimizations, taken from the dumped HLO graph of Alexnet from Tensorflow XLA.

frontend may differ in the data representation and operator definition of high-level IR across DL compilers, the hardware-independent optimizations converge at three levels such as node-level, block-level and dataflow-level. The optimization methods at each level leverage the DL specific as well as general compilation optimization techniques, which reduce the computation redundancy as well as improve the performance of DL models at computational graph level. Specifically, the frontend of XLA contains the most exhaustive hardware-independent optimizations among existing DL compilers.

4.5 Backend Optimizations

4.5.1 Hardware Specific Optimization. In the backend of DL compilers, hardware-specific optimizations, also known as target-dependent optimizations, are applied to obtain high-performance codes targeting specific hardware architecture. One way to apply the backend optimizations is to transform the low-level IR into LLVM IR in order to utilize the LLVM infrastructure to generate optimized CPU/GPU codes. The other way is to design customized optimizations with DL domain knowledge, which can leverage the target hardware more efficiently. Since hardware-specific optimizations are tailored for particular hardware architectures or implementations, we present five widely adopted approaches in existing DL compilers, including hardware intrinsic mapping, memory allocation and fetching, memory latency hiding, parallelization, and loop oriented optimization techniques.

Hardware Intrinsic Mapping - Hardware intrinsic mapping can transform a certain set of low-level IR instructions to kernels that have already been highly optimized on the hardware. In TVM, the hardware intrinsic mapping is realized in the method of *extensible tensorization*, which can declare the behavior of hardware intrinsic and the lowering rule for intrinsic mapping. This method enables the compiler backend apply hardware implementation as well as highly optimized handcraft micro-kernels to a certain pattern of operations, which results in a significant performance gain. Whereas, Glow supports hardware intrinsic mapping such as *quantization*, which is commonly used to minimize the memory footprint and improve inference speed. Glow can estimate the possible numeric range for each stage of the neural network and supports profile-guided optimization

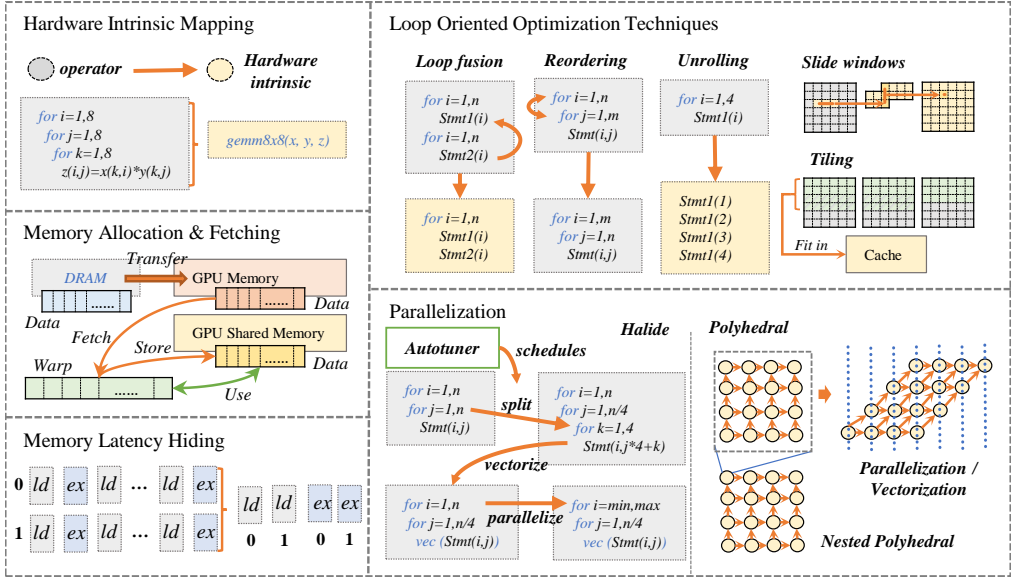


Fig. 7. Overview of backend optimizations applied in DL compilers.

to perform quantization automatically. Besides, Halide, which is widely used as the low-level IR in DL compilers such as TVM, maps specific IR patterns to SIMD opcodes on each architecture (e.g., SSE/AVX on x86 and NEON on ARM). This approach can avoid the inefficiency of LLVM IR mapping when encountering vector patterns.

Memory allocation and Fetching - Memory allocation is another challenge in code generation at the backend, especially for GPUs and customized accelerators. For example, GPU contains primarily two memory spaces, shared memory space and local memory space, where shared memory has lower access latency with limited memory size, and local memory has higher access latency with large capacity. Such memory hierarchy requires efficient memory allocation and fetching techniques for improving data locality. To realize the optimization of memory allocation and fetching, TVM introduces the scheduling concept of *memory scope*. Memory scope schedule primitives can tag a compute stage as *shared* or *thread-local*. For compute stages tagged as *shared*, TVM generates code with shared memory allocation as well as cooperative data fetching, which inserts memory barrier at the proper code position to guarantee correctness. Besides, TC also provides similar features of shared and local (a.k.a. private) memory by extending PPCG [96] compiler. Different from TVM, the memory allocation and fetching is more constrained in TC (known as *memory promotion*), which only supports optimized memory allocation and fetching under predefined rules. In addition to GPUs, other DL accelerators also require efficient memory allocation and fetching in code generation for better performance. Particularly, TVM enable special buffering in accelerators through its *memory scopes* schedule primitives.

Memory Latency Hiding - Memory latency hiding is also an important technique used in the backend by reordering the pipeline execution as much as possible. As most DL compilers support parallelization on CPU and GPU, memory latency hiding can be naturally achieved by hardware optimizations (e.g., warp context switching on GPU). But for TPU-like accelerators which implemented with *decoupled access-execute* (DAE) architecture, the backend needs to perform scheduling and fine-grained synchronization to obtain correct and efficient codes. To achieve

better performance as well as reduce programming burden, TVM introduces *virtual threading* schedule primitive, which enables users to specify the data parallelism on virtualized multi-thread architecture. Then TVM lowers these virtually parallelized threads by inserting necessary memory barriers and interleaves the operations from these threads into a single instruction stream, which forms better execution pipeline of each thread to hide the memory access latency as much as possible.

Loop Oriented Optimizations - Loop oriented optimizations are also applied in the backend to generate efficient codes for target hardware. Since Halide and LLVM (based on the polyhedral method) [59] have already incorporated such optimization techniques, some DL compilers leverage Halide and LLVM in their backends. The key techniques applied in loop oriented optimizations include: loop fusion, sliding windows, tiling, loop reordering, and loop unrolling.

- **Loop fusion:** Loop fusion is a loop optimization technique that can fuse loops with the same boundaries together for better data reuse. For compilers like PlaidML, TVM, TC, and XLA, this optimization is performed by Halide schedule or polyhedral approach, while Glow does loop fusion by its *operator stacking*.
- **Sliding windows:** Sliding windows is a loop optimization technique adopted by Halide. The central concept of sliding windows is to compute values when needed and store them until they are no longer required. When an output value of a nested loop is computed by values calculated by the previous compute stage (loops), the sliding window can cache the needed values on the fly for data reuse. As sliding windows will interleave the computation of two stages and make it serial, it is a tradeoff of parallelism and data reuse.
- **Tiling:** Tiling is another loop optimization technique commonly used in high-performance computing. Tiling means that loops are split into several tiles, thus loops are separated into outer loops of iterating through tiles and inner loops of iterating inside a tile. This transformation is aimed to enable better data reuse inside a tile by fitting a tile into hardware caches, which are crucial in modern processors' memory hierarchy. «««« HEAD As the size of a tile is quite hardware-specific, it is difficult to provide a rule to define the pattern and size for tiling. Thus, many DL compilers support automatically deciding the pattern and size of tiling by auto-tuning, which is described in detail in Section 4.5.2. And the polyhedral model based methods can implement tiling by modifying the affine functions of their band nodes.
- **Loop reordering:** Loop reordering (aka loop permutation) means changing the order of iterations in a nested loop, which can optimize the memory access and thus increase the spatial locality. It is quite specific to data layout and hardware features, and Halide provides a schedule primitive called *reorder*, which can be tuned by auto-tuning detailed in section 4.5.2. But it is not safe to perform loop reordering when the statements have dependencies along the iteration order. ===== As the size of a tile is quite hardware-specific, it is difficult to provide a rule to define the pattern and size for tiling. Thus, many DL compilers support automatically deciding the pattern and size of tiling by auto-tuning.
- **Loop reordering:** Loop reordering (also known as loop permutation) means changing the order of iterations in a nested loop, which can optimize the memory access and thus increase the spatial locality. It is specific to data layout and hardware features, and Halide provides a schedule primitive called *reorder*, which can be used to optimize loop order through auto-tuning. However, it is not safe to perform loop reordering when there are dependencies along the iteration order. »»»» 14802af6d751fa4c91fabdda606c77bdafab5f50
- **Loop unrolling:** Loop unrolling is also a commonly adopted optimization technique in compilers. Loop unrolling will unroll a specific loop to a fixed number of copies of loop bodies, which can benefit from less cost of loop controls (or fewer instruction numbers

generated for this loop). Compilers reusing Halide's works support for schedule primitives for loop unrolling, but its loop unrolling is *completely* loop unrolling, which means it will replace the loop of size n by n copies of the loop body. The generalized loop unrolling is expressed by the combination of loop split and loop unrolling, which first split the loop into two nested loops and then unroll the inner loop completely.

Parallelization - As modern processors generally support multi-threading and SIMD parallelism, it is important for the compiler backend to exploit parallelism in order to maximize hardware utilization for high performance. Halide and polyhedral models are the two major techniques adopted in existing DL compilers to exploit hardware parallelization and SIMD vectorization.

- **Halide:** For thread-level parallelization, Halide uses a schedule primitive called *parallel* to specify the parallelized dimension of the loop, which can be tuned through auto-tuning. Halide also supports GPU parallelization by mapping loop dimensions tagged as *parallel* with annotation of *block* and *thread*. For vectorization, Halide replaces a loop of size n with a n -wide vector statement, which can be mapped to hardware-specific SIMD opcodes through hardware intrinsic mapping.
- **Polyhedral model:** In a polyhedral model, one can perform transformations on polyhedron and thus obtain transformed loop mapping for better parallelization. As the polyhedral model is actually a loop transformation to detect potential parallelism, it can be applied to parallelize with CPU threads as well as GPU kernels. Furthermore, Stripe developed a variant of the polyhedral model called *nested polyhedral model*, which introduces *parallel polyhedral block* as its basic execution element of iteration. After this extension, a nested polyhedral model can detect hierarchy parallelization among levels of tiling and striding, and regardless of complicated control flow inside a parallel polyhedral block.

In addition to the above two techniques, some DL compilers rely on handcraft libraries such as Glow [79] or optimized math libraries provided by hardware vendors (detailed discussion in Section 4.5.3), which can achieve much higher performance on the target hardware. In the meanwhile, Glow offloads the vectorization to LLVM compiler because the LLVM auto-vectorizer works well when provided the knowledge of tensor dimension and loop trip count. Compared to the approaches relying on optimized libraries and LLVM infrastructure, exploiting the parallelism entirely by compiler backend allows to apply more domain-specific knowledge of DL models, and thus achieves higher performance on diverse hardware targets, however at the expense of more engineering efforts.

4.5.2 Auto-tuning. Due to enormous search space for parameter tuning in hardware-specific optimizations, it is necessary to leverage auto-tuning to determine the optimal parameter settings. Halide/TVM allows the programmers to define the hardware-specific optimizations (scheduling) first, and then it uses auto-tuning to derive the optimal parameter settings. In this way, Halide/TVM programmers can update or re-design the scheduling by inspecting the performance of specific parameter settings repeatedly. In addition, auto-tuning can also be applied to polyhedral model for parameter tuning [89]. For example, TC utilizes the auto-tuning to explore the configurations (including polyhedral scheduling and the parameters) and update the compilation cache in order to minimize the overhead of polyhedral JIT compilation. Generally, the implementation of auto-tuning includes three parts: parameterization, cost model, searching technique and performance optimization.

Parameterization - 1) *Data and target:* The data parameter describes the specification of the data, such as input shapes. The target parameter describes hardware-specific characteristics and constrains to be considered during optimization scheduling and code generation. For example,

Model	Cost	Model bias	Need hardware info	Learn from history
Blackbox model	High	None	No	No
Predefined cost model	None	High	Yes	No
ML-based cost model	Low	Low	No	Yes

Table 4. Comparison of different cost models applied in auto-tuning.

for the GPU target, the hardware parameters such as shared memory and register size need to be specified. 2) *Optimization options*: The optimization options include the optimization scheduling and corresponding parameters, such as loop oriented optimizations and tile size. In TVM, both predefined and user-defined scheduling, as well as parameters, are taken into consideration. Whereas in TC, it prefers to parameterize the optimizations, which have a strong correlation with performance and can be changed later at low cost. For example, minibatch dimension is one of the parameters, which is usually mapped to grid dimensions in CUDA and can be optimized during auto-tuning.

Cost model - The comparison of different cost models applied in auto-tuning are shown in Table 4. 1) *Black-box model*: This model only considers the final execution time rather than the characteristics of the compilation task. It is easy to build a black-box module, but leads to higher overhead and less optimal solution without the guidance of task characteristics. 2) *Pre-defined cost model*: Ideally, an approach based on a pre-defined cost model expects a perfect model that is built on the characteristics of the compilation task and is able to evaluate the overall performance of the task. There are many factors affecting the accuracy of the model, such as memory access patterns and pipeline dependencies, which are correlated with the task and the hardware target. Compared to the ML-based model, the pre-defined model generates less computation overhead when applied, but requires large engineering efforts for re-building the model on each new DL model and hardware. 3) *ML-based cost model*: ML-based cost model is a statistical approach to predict performance using a machine learning method. Using the ML-based model enables the model to update as the new configuration explored, which helps to achieve higher prediction accuracy. TVM proposes a ML-based model and a neural network model to predict the running time of generated code on the given hardware target. The input of the cost models is a lowered loop program, and the two models use different independent variables when making regression prediction. Specifically, the ML-based model uses features extracted from the input, and the neural network model uses the AST of programs. TVM claims that the accuracy of the two models is similar, and chooses the ML-based model (tree boosting) as the default cost model.

Searching technique - 1) *Initialization and searching space determination*: The initial option can either be set randomly or be set according to the known configurations, such as configurations given by users or historical optimal configurations. In terms of searching space, it should be specified before auto-tuning. TVM allows developers to specify the searching space with their domain-specific knowledge and provides automatic search space extraction for each hardware target based on the computational description. While TC relies on the compilation cache and the predefined rules. 2) *Genetic algorithm*: The genetic search algorithm used by TC considers each tuning parameter as gene and each configuration as a candidate. The new candidate is generated by two methods: hybrid and mutation. For each new candidate, three parents are selected based on their fitness, which means the higher the performance, the more likely to be selected, and the genes of parents are randomly selected and then form the new candidate. After that, the genes are assigned random values with a low probability called mutation rate, which is used for controlling the tradeoff between exploration and exploitation. There are still other parameters of algorithms needed to be set, such as iteration bounds and other options related to schedule choice. 3) *Simulated annealing*

algorithm: TVM uses this algorithm. The initial configuration can be generated randomly and then predict a similar configuration randomly at each step. The step is considered to be successful if the new configuration has a lower cost predicted by a given cost model. If the cost model is modified, the algorithm starts from the last step, the searching result of the previous cost model.

Performance Optimization - 1) *Parallelization:* One direction for improving auto-tuning performance is parallelization. TC proposes a generic multi-threaded, multi-GPU strategy under the consideration that the genetic algorithm needs to evaluate all candidates before every next generating step. First, the strategy enqueues candidate configuration and compile them by multiple CPU threads. The generated code are evaluated on GPUs in parallel, and each candidate owns its fitness used by the parent choosing step. After the finish of the whole evaluation, the new candidate is generated by the search algorithm, and the new compilation job is enqueued, waiting for compiling by CPU. Similarly, TVM supports cross-compilation and RPC, which allows users to compile on the local machine and run the programs with different auto-tuning configurations on multiple targets. 2) *Configuration reuse:* Another direction for improving auto-tuning performance is to reuse the previous auto-tuning configurations. TC stores the fastest known generated code version corresponded with given configuration by compilation cache and uses tuple as cache entry to present necessary information related to the version. During the compilation, the cache is queried before each kernel optimization enabling persistence and reuse, and auto-tuning is triggered if cache miss. Similarly, TVM produces a log file that stores the optimal configurations for all scheduling operators and queries the log file for best configurations during compilation. It is worth mentioning that TVM performs auto-tuning for each operator in Halide IR (e.g., conv2d) rather than treat all operators as a whole, and thus the optimal configurations are for operators separately rather than the low-level IR.

4.5.3 Optimized Kernel Libraries. There are several highly-optimized kernel libraries widely used to accelerate DL training and inference on various hardware. DNNL (previously MKL-DNN) from Intel company supports Intel CPUs and their integrated GPUs. DNNL detects the supported ISA in the runtime and deploys optimized code just-in-time (JIT) for the latest supported ISA (e.g., the newest AVX-512 ISA on Skylake-X). Both computation-intensive primitives (e.g., convolution, GEMM, and RNN) and memory bandwidth limited primitives (e.g., batch normalization, pooling, and shuffle) are supported and highly-optimized. Additionally, it supports low-precision training and inference, including FP32, FP16, and INT8 (only inference) as well as non-IEEE floating-point format bfloat16 [97]. cuDNN from NVIDIA company also provides the high-tuned primitives frequently arising in DNN applications. Besides, it supports customizable data layouts (e.g., flexible dimension ordering, striding, and sub-regions for the 4D tensors), which makes cuDNN easy to integrate into DL applications and avoids frequent data layout transformations. And cuDNN can make full use of the new tensor core operations on Volta and Turing GPU families. It also supports low-precision training and inference. MIOpen from AMD company also does similar optimization for high-performance machine learning primitives on AMD GPUs. However, only limited features are supported compared with DNNL and cuDNN. For example, only the GEMM primitive is optimized, and only FP16 is supported. Other customized DL accelerators also maintain their specific kernel libraries for improving the performance of DL computation.

Existing DL compilers, such as TVM, nGraph, and TC, can generate the function calls to these libraries during code generation (e.g., JIT and AOT). However, if DL compilers need to leverage the existing optimized kernel libraries, they need to first transform the data layouts and fusion styles into the types that are pre-defined in kernel libraries, and such transformation may break the optimal control-flow. Moreover, the DL compilers treat the kernel libraries as black box, therefore they are unable to apply optimizations across operators (e.g., operator fusion) when invoking kernel

libraries. In sum, using optimized kernel libraries achieves significant performance improvement when the computation can be satisfied by specific highly-optimized primitives, otherwise may be constrained from further optimization and suffer from less optimal performance.

4.5.4 Discussion. The backends of DL compilers have commonly adopted the design including various hardware-specific optimizations, auto-tuning techniques and optimized kernel libraries. Hardware-specific optimizations enable efficient code generation for different hardware targets, such as CPU, GPU and customized DL accelerators. The widely used hardware-specific optimizations at the backend include hardware intrinsic mapping, memory allocation and fetching, memory latency hiding, loop oriented optimizations, and parallelization. However, due to the diversity of DL hardwares, the optimizations are not limited to the ones presented in this section. To address the large parameter tuning space introduced by hardware-specific optimizations, auto-tuning becomes essential in the compiler backend to alleviate the manual efforts to derive the optimal parameter settings. The design of auto-tuning usually consists of four components such as parameterization, cost model, searching technique and performance optimization. To further improve the performance of the generated code, optimized kernel libraries are also widely used in the backend of DL compilers. When the DL computation satisfies the kernel definition in the highly-optimized libraries, it can achieve significant performance improvement. However, relying on the optimized kernel libraries may waste the performance opportunity for advanced optimizations such as operator fusion across multiple operators.

5 CONCLUSION AND FUTURE DIRECTIONS

In this survey, we present a thorough analysis of the existing DL compilers. First, we provide a comprehensive comparison of the existing DL compilers from various aspects, which can serve as the guideline for users to choose the suitable DL compiler for their customized scenarios. Then, we take a deep dive into the common design adopted in the existing DL compilers including the multi-level IR, the frontend, and the backend. We present the design philosophy and reference implementation of each component in detail, with the emphasis on the unique IRs and optimizations specific to DL compilers. We summarize the findings in this survey and highlight the future directions in DL compiler as follows, which we hope can shed the light to encourage more researchers and practitioners contribute to this field.

Dynamic shape and control flow - Dynamic model becomes more and more popular in the field of deep learning, whose input shape or even model itself may change in execution. Particularly, *dynamic shape* is the major concern of dynamic models, which is partially supported by TVM and other DL compilers. In the DL community, especially in natural language processing (NLP), models may accept inputs of various shapes, which is challenging for DL compilers since the shape of data is unknown until runtime. Existing DL compilers require more research efforts to support dynamic shape efficiently for emerging dynamic models.

In addition, future DL models will become more complex, and may include complicated *control flow*. As most DL frameworks and compilers use Python as their programming language, the performance will become a severe problem if a model is implemented with control flow, because it causes the model executed by the Python interpreter. Moreover, a DL model often requires complicated pre/post-process of data/results. Although the pre/post-processing may become a performance bottleneck in training and inference, it has not yet been considered by existing DL compilers. The compiler support for control flow and commonly used pre/post-processing will further increase the performance gain in model deployment. Actually, Relay [78] is currently developing Relay Virtual Machine to support control flow in Relay runtime, which can avoid the use of inefficient Python interpreter.

Advanced auto-tuning - Existing auto-tuning techniques focus on the optimization of individual operators. However, the combination of the local optimal does not lead to global optimal. For example, two adjacent operators that apply on different data layouts can be performance tuned together without introducing extra memory transformation nodes in between. Besides, with the rise of edge computing, execution time is not only the optimization objective for DL compilers. New optimization targets should also be considered in the auto-tuning such as less memory footprint and lower energy consumption.

For ML-based auto-tuning techniques, there are several directions worth further exploring. First, the ML techniques should be applied in other stages of auto-tuning, other than the cost model. For example, in the stage of selecting compiler options and optimization schedules, instead of developing a cost model, ML techniques can be used to predict the expectant possibility directly and develop algorithm to determine the final result. Additionally, various computation characteristics and optimization objectives of different DL models require their specific tuning models trained with different datasets and objective functions rather than sharing a universal auto-tuning model.

Second, the ML-based auto-tuning techniques can also be improved based on the domain knowledge of DL models. For example, incorporating the feature engineering (selecting features to represent program) [99] in auto-tuning techniques could be a potential direction for achieving better tuning results. Besides, in addition to the static code feature gathered from IR, other computation characteristics are also expressive and strongly correlated with performance. For example, the features of the computational graph can represent data dependency and memory movement more intuitively and enable better execution time prediction.

Polyhedral model - As described in Section 4.5.2, the auto-tuning can be applied to minimize the overhead of polyhedral JIT compilation by reusing the previous configurations. At the other hand, the polyhedral model can be used to perform auto-scheduling, which can reduce the search space of auto-tuning. It is a promising research direction to further apply the combination of the polyhedral model and the auto-tuning to the design of DL compilers for efficiency.

Another challenge of the polyhedral model is to support the sparse tensor. In general, the format of a sparse tensor such as CSF [84] expresses the loop indices with index arrays (e.g., $a[b[i]]$) that is no longer linear. Such indirect index addressing leads to non-affine subscript expressions and loop bounds, which prohibits the loop optimization of the polyhedral model [26, 88]. Fortunately, the polyhedral community has made progress in supporting sparse tensor [92, 93], and integrating the latest advancement of the polyhedral model can increase the performance opportunities for future DL compilers.

Subgraph partitioning - A DL compiler supporting subgraph partitioning can partition the computation graph into several subgraphs. By this approach, the computation graph can no longer be treated as a whole, and the subgraphs can be processed in different manners. The subgraph partitioning can bring at least two advantages to DL compilers.

First, it opens up the possibility to integrate more libraries that apply graph optimizations. Take the nGraph with DNNL for example, DNNL is a DL optimized libraries on CPU, which can apply layer fusion and other graph optimizations by leveraging its diverse collection of highly optimized kernels. This integration enables DNNL to optimize and execute the most of compatible subgraphs, while leaves the remaining graph to nGraph. The integration of TensorFlow with TensorRT¹ adopts a similar approach. Other DL compilers however, fail to exploit such integration. Take TVM for example, TVM supports invoking DNNL library, but treats it as a regular BLAS library without utilizing the subgraph optimizations. Similar problem happens on dedicated DL accelerators, where they rely on customized graph optimization libraries. In sum, the integration of libraries with

¹<https://docs.nvidia.com/deeplearning/frameworks/tf-trt-user-guide/index.html>

graph optimizations can provide an alternative approach to improve the performance of the code generated from the DL compilers. Besides, the integration also provides a new approach to support hardware targets by invoking graph optimized libraries.

Second, it opens up the possibility of heterogeneous and parallel execution. Currently, the scenarios at two extreme scales such as edge devices and data centers for deploying DL models all exhibit the trend of heterogeneous and parallel execution. Once the computation graph is partitioned into subgraphs, the execution of different subgraphs can be assigned to heterogeneous hardware targets at the same time. Take the edge device for example, its computation units may consist of ARM CPU, Mail GPU, DSP, and probably NPU. Generating code from the DL compilers that utilizes all computation units efficiently can deliver significant speedup of the DL tasks, such as face recognition and voice assistant.

Quantization - Quantization is a well-known optimization technique in DL models that can reduce the burden of computation and memory by lowering the precision of operational data. With reduced bit-width, the data can be stored and calculated with less resource at the expense of slightly reduced model accuracy. The major challenge of quantization is to design the strategies that tradeoff between the benefits of low precision and loss of model accuracy. Traditional quantization strategies applied in DL frameworks are based on a set of fixed schemes and datatypes with little customization for codes running on different hardware. Whereas, supporting quantization in DL compilers can leverage more optimization information during compilation to derive more efficient quantization strategies. For example, Relay [78] provides a quantization rewriting flow that can automatically generate code for various schemes.

To support quantization, there are more challenges to be solved in the DL compilers. The first challenge is how to implement new operators with reduced precision without heavy engineering efforts. To reuse the quantization implementation in Relay, [15] proposes a new dialect to implement new operator with basic instructions instead of implementing the new operator from scratch, which eliminates the engineering efforts of re-implementing the graph-level and operator-level optimizations. The interaction between quantization and other optimizations during compilation is another challenge. For example, determining the appropriate place for quantization and collaborating with optimizations such as operator fusion require future research investigations. Meanwhile, quantization also impacts the hardware-specific optimizations in DL compilers, which can be re-designed to leverage the performance opportunity that less resource is required with the low-precision operators and data.

Unified optimizations - Although existing DL compilers adopt similar design in both computation graph optimization and hardware-specific optimization, each of compiler has its own implementation with advantages in certain aspects. There is missing a way to share the state-of-the-art optimizations, as well as support of emerging hardware targets across existing compilers. We advocate unifying the optimizations from existing DL compilers so that the best practices adopted in each DL compiler can be reused. In addition, unifying the optimizations across DL compilers can accumulate a strong force to impact the design of general-purpose and dedicated DL accelerators, and provide an environment for efficient co-design of DL compiler and hardware. Currently, Google MLIR is a promising initiative towards such direction. It provides the infrastructure of multi-level IRs, and contains IR specification and toolkit to perform transformations across IRs at each level. It also provides flexible *dialects*, so that each DL compiler can construct its customized *dialects* for both high-level and low-level IRs. Through transformation across *dialects*, optimizations of one DL compiler can be reused by another compiler. However, the transformation of *dialects* may need delicate tradeoffs and some engineering efforts.

Differentiable programming - Differentiable programming is a programming paradigm, and the programs in the differentiable programming paradigm are differentiable thoroughly. It has

attracted the attention of the DL compiler community recently. Many new compiler projects have replaced the computational graph with differentiable programming, such as Myia [25] and Swift for TensorFlow [18]. In addition, Flux [51] is one of the most promising ML stacks with its differential language, Julia [23]. Julia is a well-suited language for ML programming that is designed for mathematical and numerical computing, and Flux extends Julia for differentiable algorithms and acceleration realized by the compiler. Unfortunately, there is little support for differential languages in existing DL compilers.

Supporting differential language is quite challenging for existing DL compilers. The difficulties come from not only data structure, but also language semantic. For example, to realize the transformation from Julia to XLA HLO IR, one of the challenges faced by the project [37] is that the control flow is different between the imperative language used by Julia and the symbolic language used by XLA. In order to use HLO IR efficiently, the compiler also needs to provide operation abstraction for Julia in order to support the particular semantic of XLA, such as *MapReduce* and *broadcast*. Moreover, the difference of the semantic of differentiation between Julia and XLA, also leads to significant changes of the automatic differentiation algorithm and corresponding designs.

Graph neural network (GNN) - GNN has been a popular research direction in the field of deep learning in recent years [22, 102, 104, 107]. Traditional DL networks are effective at regularly structured data in Euclidean space (e.g., picture and speech), whereas behavior poorly on irregularly structured data (e.g., social and e-commerce). Taking the e-commerce for example, users, products, and advertisements can be regarded as *nodes*, and operations such as users purchasing goods and clicking advertisements can be regarded as *edges*. *Nodes* and *edges* form a large *graph*. GNN takes this large *graph* as input and obtains the representation of *nodes*, *edges* or *subgraphs* by aggregating neighbor information, so as to realize tasks such as classification, prediction and recommendation. It has been accepted that the design of GNN behaviors well on unstructured data.

A large number of DL frameworks have already provided implementation of GNN, such as TensorFlow, PyTorch, MXNet, PaddlePaddle, and Theano. However, existing DL compilers have little support for GNNs. Only TVM has released a primitive tutorial of building a Graph Convolutional Network (GCN) with Relay and MXNet on the Cora dataset ². One of the reasons for the immature support of GNNs in existing DL compilers is due to its unique design. For example, GNNs are always shallow, most of which are no more than three layers [107]. This is because GNN merges the representation of neighboring nodes closer to each other, which results in fewer layers. However, stacking multiple GNN layers recklessly can easily result in over-smoothing. For DL compilers, how to use the graph-level optimizations to avoid over-smoothing of GNNs requires more research efforts.

Privacy protection - As the edge devices such as sensors and mobile phones are widely used in our daily life, edge-cloud system is becoming pervasive to run DL models for intelligent tasks such as face detection and voice recognition. In edge-cloud system, the DL models are usually split into two halves with each partial model running on the edge device and cloud service respectively, which can provide better response latency and consume less communication bandwidth. However, one of the drawbacks with the edge-cloud system is that the user privacy becomes vulnerable. The reason is that the attackers can intercept the intermediate results sent from the edge devices to cloud, and then use the intermediate results to train another model that can reveal the privacy information deviated from the original user task [40, 69, 73].

To protect privacy in edge-cloud system, existing approaches [40, 69, 73] propose to add noise with special statistic properties to the intermediate results that can reduce the accuracy of the attacker model without severely deteriorating the accuracy of the original user model. However,

²<https://relational.fit.cvut.cz/dataset/CORA>

the difficulty with the existing approaches is to determine the layer where the noise should be inserted, which is quite labor intensive to identify the optimal layer. The above difficulty presents a great opportunity for DL compilers to support privacy protection, because the compilers maintain rich information about the computation, communication and entropy of all layers, which can guide the noise generation across layers for better privacy protection automatically. We believe this could be an interesting research direction across the disciplines of security and compilation for DL community.

REFERENCES

- [1] [n.d.]. Adding an Operator to Relay. https://docs.tvm.ai/dev/relay_add_op.html. Accessed February 4, 2020.
- [2] [n.d.]. Announcing Hanguang 800: Alibaba's First AI-Inference Chip. https://www.alibabacloud.com/blog/announcing-hanguang-800-alibabas-first-ai-inference-chip_595482. Accessed February 4, 2020.
- [3] [n.d.]. AWS Inferentia. <https://aws.amazon.com/machine-learning/inferentia>. Accessed February 4, 2020.
- [4] [n.d.]. Documents of Glow. <https://github.com/pytorch/glow/blob/master/docs/IR.md>. Accessed February 4, 2020.
- [5] [n.d.]. The embedded languages and IRs in the TVM stack. <https://docs.tvm.ai/langref/index.html>. Accessed February 4, 2020.
- [6] [n.d.]. Gluon. <https://gluon.mxnet.io>. Accessed February 4, 2020.
- [7] [n.d.]. HalideIR: Symbolic Arithmetic IR Module. <https://github.com/dmlc/HalideIR>. Accessed February 4, 2020.
- [8] [n.d.]. Nervana Neural Network Processor. <https://www.intel.ai/nervana-nnp/>. Accessed February 4, 2020.
- [9] [n.d.]. Nvidia Turing Architecture. <https://www.nvidia.com/en-us/design-visualization/technologies/turing-architecture/>. Accessed February 4, 2020.
- [10] [n.d.]. ONNX Github repository. <https://github.com/onnx/onnx>. Accessed February 4, 2020.
- [11] [n.d.]. PaddlePaddle Github repository. <https://github.com/PaddlePaddle/Paddle>. Accessed February 4, 2020.
- [12] [n.d.]. PlaidML Github repository. <https://github.com/plaidml/plaidml>. Accessed February 4, 2020.
- [13] [n.d.]. Polyhedral Compilation. <https://polyhedral.info>. Accessed February 4, 2020.
- [14] [n.d.]. TensorRT Github repository. <https://github.com/NVIDIA/TensorRT>. Accessed February 4, 2020.
- [15] 2019. Compilation of Quantized Models in TVM. https://github.com/tvmai/meetup-slides/blob/master/tvm-meetup-bay-area-Nov-8-2019/Nov8_TVM_meetup_Quantization.pdf. Accessed February 4, 2020.
- [16] Martin Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. 2016. Tensorflow: A system for large-scale machine learning. In *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*. 265–283.
- [17] Kamel Abdelouahab, Maxime Pelcat, Jocelyn Serot, Cedric Bourrasset, and François Berry. 2017. Tactics to directly map CNN graphs on embedded FPGAs. *IEEE Embedded Systems Letters* 9, 4 (2017), 113–116.
- [18] Akshay Agrawal, Akshay Naresh Modi, Alexandre Passos, Allen Lavoie, Ashish Agarwal, Asim Shankar, Igor Ganichev, Josh Levenberg, Mingsheng Hong, Rajat Monga, et al. 2019. Tensorflow eager: A multi-stage, python-embedded dsl for machine learning. *arXiv preprint arXiv:1903.01855* (2019).
- [19] Alfred V Aho, Ravi Sethi, and Jeffrey D Ullman. 1986. Compilers, principles, techniques. *Addison wesley* 7, 8 (1986), 9.
- [20] Roberto Bagnara, Patricia M Hill, and Enea Zaffanella. 2006. The Parma Polyhedra Library: Toward a complete set of numerical abstractions for the analysis and verification of hardware and software systems. *arXiv preprint cs/0612085* (2006).
- [21] Soheil Bahrampour, Naveen Ramakrishnan, Lukas Schott, and Mohak Shah. 2016. Comparative study of caffe, neon, theano, and torch for deep learning. (2016).
- [22] Peter W Battaglia, Jessica B Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinicius Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, et al. 2018. Relational inductive biases, deep learning, and graph networks. *arXiv preprint arXiv:1806.01261* (2018).
- [23] Jeff Bezanson, Alan Edelman, Stefan Karpinski, and Viral B Shah. 2017. Julia: A fresh approach to numerical computing. *SIAM review* 59, 1 (2017), 65–98.
- [24] James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, and Skye Wanderman-Milne. 2018. *JAX: composable transformations of Python+NumPy programs*. <http://github.com/google/jax>
- [25] Olivier Breuleux and Bart van Merriënboer. 2017. Automatic Differentiation in Myia. (2017).
- [26] Chun Chen. 2012. Polyhedra scanning revisited. In *Proceedings of the 33rd ACM SIGPLAN conference on Programming Language Design and Implementation*. 499–508.
- [27] Hongming Chen, Ola Engkvist, Yinhai Wang, Marcus Olivecrona, and Thomas Blaschke. 2018. The rise of deep learning in drug discovery. *Drug discovery today* 23, 6 (2018), 1241–1250.

- [28] Tianqi Chen, Mu Li, Yutian Li, Min Lin, Naiyan Wang, Minjie Wang, Tianjun Xiao, Bing Xu, Chiyuan Zhang, and Zheng Zhang. 2015. Mxnet: A flexible and efficient machine learning library for heterogeneous distributed systems. *arXiv preprint arXiv:1512.01274* (2015).
- [29] Tianqi Chen, Thierry Moreau, Ziheng Jiang, Lianmin Zheng, Eddie Yan, Haichen Shen, Meghan Cowan, Leyuan Wang, Yuwei Hu, Luis Ceze, et al. 2018. {TVM}: An automated end-to-end optimizing compiler for deep learning. In *13th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 18)*. 578–594.
- [30] Tianqi Chen, Lianmin Zheng, Eddie Yan, Ziheng Jiang, Thierry Moreau, Luis Ceze, Carlos Guestrin, and Arvind Krishnamurthy. 2018. Learning to optimize tensor programs. In *Advances in Neural Information Processing Systems*. 3389–3400.
- [31] Sharan Chetlur, Cliff Woolley, Philippe Vandermersch, Jonathan Cohen, John Tran, Bryan Catanzaro, and Evan Shelhamer. 2014. cudnn: Efficient primitives for deep learning. *arXiv preprint arXiv:1410.0759* (2014).
- [32] François Chollet et al. 2015. Keras. <https://keras.io>.
- [33] R. Collobert, K. Kavukcuoglu, and C. Farabet. 2011. Torch7: A Matlab-like Environment for Machine Learning. In *BigLearn, NIPS Workshop*.
- [34] Scott Cyphers, Arjun K Bansal, Anahita Bhiwandiwala, Jayaram Bobba, Matthew Brookhart, Avijit Chakraborty, Will Constable, Christian Convey, Leona Cook, Omar Kanawi, et al. 2018. Intel ngraph: An intermediate representation, compiler, and executor for deep learning. *arXiv preprint arXiv:1801.08058* (2018).
- [35] Ron Cytron, Jeanne Ferrante, Barry K Rosen, Mark N Wegman, and F Kenneth Zadeck. 1991. Efficiently computing static single assignment form and the control dependence graph. *ACM Transactions on Programming Languages and Systems (TOPLAS)* 13, 4 (1991), 451–490.
- [36] P. Feautrier. 1988. Parametric integer programming. *RAIRO Recherche Opérationnelle* 22, 3 (1988), 243–268.
- [37] Keno Fischer and Elliot Saba. 2018. Automatic full compilation of julia programs and ML models to cloud TPUs. *arXiv preprint arXiv:1810.09868* (2018).
- [38] Rubén D Fonnegra, Bryan Blair, and Gloria M Diaz. 2017. Performance comparison of deep learning frameworks in image classification problems using convolutional and recurrent networks. In *2017 IEEE Colombian Conference on Communications and Computing (COLCOM)*. IEEE, 1–6.
- [39] David A Forsyth and Jean Ponce. 2002. *Computer vision: a modern approach*. Prentice Hall Professional Technical Reference.
- [40] Ruiyuan Gao, Ming Dun, Hailong Yang, Zhongzhi Luan, and Depei Qian. 2019. Privacy for Rescue: A New Testimony Why Privacy is Vulnerable In Deep Models. *arXiv preprint arXiv:2001.00493* (2019).
- [41] David E Goldberg and John Henry Holland. 1988. Genetic algorithms and machine learning. (1988).
- [42] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative Adversarial Networks. *arXiv:stat.ML/1406.2661*
- [43] Yijin Guan, Hao Liang, Ningyi Xu, Wenqiang Wang, Shaoshuai Shi, Xi Chen, Guangyu Sun, Wei Zhang, and Jason Cong. 2017. FP-DNN: An automated framework for mapping deep neural networks onto FPGAs with RTL-HLS hybrid templates. In *2017 IEEE 25th Annual International Symposium on Field-Programmable Custom Computing Machines (FCCM)*. IEEE, 152–159.
- [44] Kaiyuan Guo, Lingzhi Sui, Jiantao Qiu, Jincheng Yu, Junbin Wang, Song Yao, Song Han, Yu Wang, and Huazhong Yang. 2017. Angel-Eye: A complete design flow for mapping CNN onto embedded FPGA. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* 37, 1 (2017), 35–47.
- [45] Kaiyuan Guo, Shulin Zeng, Jincheng Yu, Yu Wang, and Huazhong Yang. 2017. A Survey of FPGA-Based Neural Network Accelerator. *arXiv:cs.AR/1712.08934*
- [46] Qianyu Guo, Xiaofei Xie, Lei Ma, Qiang Hu, Ruitao Feng, Li Li, Yang Liu, Jianjun Zhao, and Xiaohong Li. 2018. An Orchestrated Empirical Study on Deep Learning Frameworks and Platforms. *arXiv preprint arXiv:1811.05187* (2018).
- [47] Jung-Woo Ha, Hyuna Pyo, and Jeonghee Kim. 2016. Large-scale item categorization in e-commerce using multiple recurrent neural networks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 107–115.
- [48] William Grant Hatcher and Wei Yu. 2018. A survey of deep learning: platforms, applications and emerging research trends. *IEEE Access* 6 (2018), 24411–24432.
- [49] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.
- [50] Jeremy Howard et al. 2018. fastai. <https://github.com/fastai/fastai>.
- [51] Michael Innes, Elliot Saba, Keno Fischer, Dhairya Gandhi, Marco Concetto Rudilosso, Neethu Mariya Joy, Tejan Karmali, Avik Pal Singh, and Viral Shah. 2018. Fashionable Modelling with Flux. *arXiv preprint arXiv:1811.01457* (2018).
- [52] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. 2014. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 675–678.

- [53] Zhe Jia, Blake Tillman, Marco Maggioni, and Daniele Paolo Scarpazza. 2019. Dissecting the Graphcore IPU Architecture via Microbenchmarking. *arXiv preprint arXiv:1912.03413* (2019).
- [54] Norman P Jouppi, Cliff Young, Nishant Patil, David Patterson, Gaurav Agrawal, Raminder Bajwa, Sarah Bates, Suresh Bhatia, Nan Boden, Al Borchers, et al. 2017. In-datacenter performance analysis of a tensor processing unit. In *Proceedings of the 44th Annual International Symposium on Computer Architecture*. 1–12.
- [55] Duseok Kang, Euiseok Kim, Inpyo Bae, Bernhard Egger, and Soonhoi Ha. 2018. C-GOOD: C-code generation framework for optimized on-device deep learning. In *Proceedings of the International Conference on Computer-Aided Design*. ACM, 105.
- [56] Wayne Kelly, Vadim Maslov, William Pugh, Evan Rosser, Tatiana Shpeisman, and Dave Wonnacott. 1996. The Omega Calculator and Library, Version 1.1.0. (1996). <http://www.cs.utah.edu/~mhall/cs6963s09/lectures/omega.ps>
- [57] Richard Kelsey, William Clinger, Jonathan Rees, et al. 1998. Revised 5 report on the algorithmic language Scheme. (1998).
- [58] Adrian Kingsley-Hughes. 2017. Inside Apple’s new A11 Bionic processor. *ZDNet, September* (2017).
- [59] Chris Lattner and Vikram Adve. 2004. LLVM: A compilation framework for lifelong program analysis & transformation. In *Proceedings of the international symposium on Code generation and optimization: feedback-directed and runtime optimization*. IEEE Computer Society, 75.
- [60] Chris Leary and Todd Wang. 2017. XLA: TensorFlow, compiled. *TensorFlow Dev Summit* (2017).
- [61] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. 1998. Gradient-based learning applied to document recognition. *Proc. IEEE* 86, 11 (1998), 2278–2324.
- [62] Heng Liao, Jiajin Tu, Jing Xia, and Xiping Zhou. 2019. DaVinci: A Scalable Architecture for Neural Network Computing. In *2019 IEEE Hot Chips 31 Symposium (HCS)*. IEEE, 1–44.
- [63] S. Liu, Z. Du, J. Tao, D. Han, T. Luo, Y. Xie, Y. Chen, and T. Chen. 2016. Cambricon: An Instruction Set Architecture for Neural Networks. In *2016 ACM/IEEE 43rd Annual International Symposium on Computer Architecture (ISCA)*. 393–405. <https://doi.org/10.1109/ISCA.2016.42>
- [64] Yizhi Liu, Yao Wang, Ruofei Yu, Mu Li, Vin Sharma, and Yida Wang. 2019. Optimizing {CNN} Model Inference on CPUs. In *2019 {USENIX} Annual Technical Conference ({USENIX}{ATC} 19)*. 1025–1040.
- [65] Zhiqiang Liu, Yong Dou, Jingfei Jiang, and Jinwei Xu. 2016. Automatic code generation of convolutional neural networks in FPGA implementation. In *2016 International Conference on Field-Programmable Technology (FPT)*. IEEE, 61–68.
- [66] Vincent Loechner. 1999. PolyLib: A library for manipulating parameterized polyhedra. https://repo.or.cz/polylib.git/blob_plain/HEAD:/doc/parampoly-doc.ps.gz
- [67] Yufei Ma, Naveen Suda, Yu Cao, Sarma Vrudhula, and Jae-sun Seo. 2018. ALAMO: FPGA acceleration of deep learning algorithms with a modularized RTL compiler. *Integration* 62 (2018), 14–23.
- [68] Christopher D Manning, Christopher D Manning, and Hinrich Schütze. 1999. *Foundations of statistical natural language processing*. MIT press.
- [69] Fatemehsadat Mireshghallah, Mohammadkazem Taram, Prakash Ramrakhani, Dean Tullsen, and Hadi Esmaeilzadeh. 2019. Shredder: Learning Noise Distributions to Protect Inference Privacy. (2019).
- [70] Mehdi Mohammadi, Ala Al-Fuqaha, Mohsen Guizani, and Jun-Seok Oh. 2017. Semisupervised deep reinforcement learning in support of IoT and smart city services. *IEEE Internet of Things Journal* 5, 2 (2017), 624–635.
- [71] Thierry Moreau, Tianqi Chen, Luis Vega, Jared Roesch, Eddie Yan, Lianmin Zheng, Josh Fromm, Ziheng Jiang, Luis Ceze, Carlos Guestrin, et al. 2018. A Hardware-Software Blueprint for Flexible Deep Learning Specialization. *arXiv preprint arXiv:1807.04188* (2018).
- [72] Madhumitha Nara, BR Mukesh, Preethi Padala, and Bharath Kinnal. 2019. Performance Evaluation of Deep Learning frameworks on Computer Vision problems. In *2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI)*. IEEE, 670–674.
- [73] Seyed Ali Osia, Ali Taheri, Ali Shahin Shamsabadi, Kleomenis Katevas, Hamed Haddadi, and Hamid R Rabiee. 2018. Deep private-feature extraction. *IEEE Transactions on Knowledge and Data Engineering* 32, 1 (2018), 54–66.
- [74] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*. 8024–8035.
- [75] Barak A Pearlmutter and Jeffrey Mark Siskind. 2008. Reverse-mode AD in a functional framework: Lambda the ultimate backpropagator. *ACM Transactions on Programming Languages and Systems (TOPLAS)* 30, 2 (2008), 7.
- [76] Tomas Petricek and Don Syme. 2012. Syntax Matters: Writing abstract computations in F#. *Pre-proceedings of TFP (Trends in Functional Programming)*. St. Andrews, Scotland (2012).
- [77] 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 recomputation in image processing pipelines. In *Acm Sigplan Notices*, Vol. 48. ACM, 519–530.

- [78] Jared Roesch, Steven Lyubomirsky, Marisa Kirisame, Logan Weber, Josh Pollock, Luis Vega, Ziheng Jiang, Tianqi Chen, Thierry Moreau, and Zachary Tatlock. 2019. Relay: A High-Level Compiler for Deep Learning. *arXiv:cs.LG/1904.08368*
- [79] Nadav Rotem, Jordan Fix, Saleem Abdulrasool, Garret Catron, Summer Deng, Roman Dzhabarov, Nick Gibson, James Hegeman, Meghan Lele, Roman Levenstein, et al. 2018. Glow: Graph lowering compiler techniques for neural networks. *arXiv preprint arXiv:1805.00907* (2018).
- [80] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. 1986. Learning representations by back-propagating errors. *nature* 323, 6088 (1986), 533–536.
- [81] Frank Seide and Amit Agarwal. 2016. CNTK: Microsoft’s open-source deep-learning toolkit. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2135–2135.
- [82] Shayan Shams, Richard Platanina, Kisung Lee, and Seung-Jong Park. 2017. Evaluation of deep learning frameworks over different HPC architectures. In *2017 IEEE 37th International Conference on Distributed Computing Systems (ICDCS)*. IEEE, 1389–1396.
- [83] Hardik Sharma, Jongse Park, Divya Mahajan, Emmanuel Amaro, Joon Kyung Kim, Chenkai Shao, Asit Mishra, and Hadi Esmaeilzadeh. 2016. From high-level deep neural models to FPGAs. In *The 49th Annual IEEE/ACM International Symposium on Microarchitecture*. IEEE Press, 17.
- [84] Shaden Smith and George Karypis. 2015. Tensor-matrix products with a compressed sparse tensor. In *Proceedings of the 5th Workshop on Irregular Applications: Architectures and Algorithms*. 1–7.
- [85] D Team et al. 2016. Deeplearning4j: Open-source distributed deep learning for the JVM. *Apache Software Foundation License 2* (2016).
- [86] The Theano Development Team, Rami Al-Rfou, Guillaume Alain, Amjad Almahairi, Christof Angermueller, Dzmitry Bahdanau, Nicolas Ballas, Frédéric Bastien, Justin Bayer, Anatoly Belikov, et al. 2016. Theano: A Python framework for fast computation of mathematical expressions. *arXiv preprint arXiv:1605.02688* (2016).
- [87] Seiya Tokui, Kenta Oono, Shohei Hido, and Justin Clayton. 2015. Chainer: a next-generation open source framework for deep learning. In *Proceedings of workshop on machine learning systems (LearningSys) in the twenty-ninth annual conference on neural information processing systems (NIPS)*, Vol. 5. 1–6.
- [88] Nicolas Vasilache, Cédric Bastoul, and Albert Cohen. 2006. Polyhedral code generation in the real world. In *International Conference on Compiler Construction*. Springer, 185–201.
- [89] Nicolas Vasilache, Oleksandr Zinenko, Theodoros Theodoridis, Priya Goyal, Zachary DeVito, William S Moses, Sven Verdoolaege, Andrew Adams, and Albert Cohen. 2018. Tensor comprehensions: Framework-agnostic high-performance machine learning abstractions. *arXiv preprint arXiv:1802.04730* (2018).
- [90] Stylianos I Venieris and Christos-Savvas Bouganis. 2016. fpgaConvNet: A framework for mapping convolutional neural networks on FPGAs. In *2016 IEEE 24th Annual International Symposium on Field-Programmable Custom Computing Machines (FCCM)*. IEEE, 40–47.
- [91] Stylianos I. Venieris, Alexandros Kouris, and Christos-Savvas Bouganis. 2018. Toolflows for Mapping Convolutional Neural Networks on FPGAs. *Comput. Surveys* 51, 3 (Jun 2018), 1–39. <https://doi.org/10.1145/3186332>
- [92] Anand Venkat, Mary Hall, and Michelle Strout. 2015. Loop and data transformations for sparse matrix code. *ACM SIGPLAN Notices* 50, 6 (2015), 521–532.
- [93] Anand Venkat, Manu Shantharam, Mary Hall, and Michelle Mills Strout. 2014. Non-affine extensions to polyhedral code generation. In *Proceedings of Annual IEEE/ACM International Symposium on Code Generation and Optimization*. 185–194.
- [94] Sven Verdoolaege. 2010. isl: An integer set library for the polyhedral model. In *International Congress on Mathematical Software*. Springer, 299–302.
- [95] Sven Verdoolaege. 2011. Counting affine calculator and applications. In *First International Workshop on Polyhedral Compilation Techniques (IMPACT’11)*, Chamonix, France.
- [96] Sven Verdoolaege, Juan Carlos Juega, Albert Cohen, José Ignacio Gómez, Christian Tenllado, and Francky Catthoor. 2013. Polyhedral parallel code generation for CUDA. *ACM Trans. Archit. Code Optim.* 9, 4 (Jan. 2013), 54:1–54:23. <https://doi.org/10.1145/2400682.2400713>
- [97] Shibo Wang and Pankaj Kanwar. 2019. BFloat16: the secret to high performance on cloud TPUs. *Google Cloud Blog*, August (2019).
- [98] Ying Wang, Jie Xu, Yinhe Han, Huawei Li, and Xiaowei Li. 2016. DeepBurning: automatic generation of FPGA-based learning accelerators for the neural network family. In *Proceedings of the 53rd Annual Design Automation Conference*. ACM, 110.
- [99] Zheng Wang and Michael O’Boyle. 2018. Machine learning in compiler optimization. *Proc. IEEE* 106, 11 (2018), 1879–1901.
- [100] Gu-Yeon Wei, David Brooks, et al. 2019. Benchmarking tpu, gpu, and cpu platforms for deep learning. *arXiv preprint arXiv:1907.10701* (2019).

- [101] Xuechao Wei, Cody Hao Yu, Peng Zhang, Youxiang Chen, Yuxin Wang, Han Hu, Yun Liang, and Jason Cong. 2017. Automated systolic array architecture synthesis for high throughput CNN inference on FPGAs. In *Proceedings of the 54th Annual Design Automation Conference 2017*. ACM, 29.
- [102] Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and Philip S Yu. 2019. A comprehensive survey on graph neural networks. *arXiv preprint arXiv:1901.00596* (2019).
- [103] Yu Xing, Jian Weng, Yushun Wang, Lingzhi Sui, Yi Shan, and Yu Wang. 2019. An In-depth Comparison of Compilers for Deep Neural Networks on Hardware. In *2019 IEEE International Conference on Embedded Software and Systems (ICESS)*. IEEE, 1–8.
- [104] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. 2018. How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826* (2018).
- [105] Tim Zerrell and Jeremy Bruestle. 2019. Stripe: Tensor Compilation via the Nested Polyhedral Model. *arXiv preprint arXiv:1903.06498* (2019).
- [106] R. Zhao, S. Liu, H. Ng, E. Wang, J. J. Davis, X. Niu, X. Wang, H. Shi, G. A. Constantinides, P. Y. K. Cheung, and W. Luk. 2018. Hardware Compilation of Deep Neural Networks: An Overview. In *2018 IEEE 29th International Conference on Application-specific Systems, Architectures and Processors (ASAP)*. 1–8. <https://doi.org/10.1109/ASAP.2018.8445088>
- [107] Jie Zhou, Ganqu Cui, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. 2018. Graph neural networks: A review of methods and applications. *arXiv preprint arXiv:1812.08434* (2018).