

When they go high we go low

Maksim Levental*
mlevental@uchicago.edu
University of Chicago

Elena Orlova
eorlova@uchicago.edu
University of Chicago

ABSTRACT

High level abstractions for implementing, training, and testing Deep Learning (DL) models abound. Such frameworks function primarily by abstracting away the implementation details of arbitrary neural architectures, thereby enabling researchers and engineers to focus on design. In principle, such frameworks could be “zero-cost abstractions”; in practice, they incur translation and indirection overheads. We study at which points exactly in the engineering life-cycle of a DL model are the highest costs paid and whether they can be mitigated. We train, test, and evaluate a representative DL model using PyTorch, LibTorch, TorchScript, and cuDNN on representative datasets.

ACM Reference Format:

Maksim Levental and Elena Orlova. 2020. When they go high we go low. In *Proceedings of ACM Conference (Conference’17)*. ACM, New York, NY, USA, 5 pages. <https://doi.org/10.1145/nnnnnnnn.nnnnnnnn>

1 INTRODUCTION

Deep Learning (DL) frameworks represent neural network models as dataflow and computation graphs (where nodes correspond to functional units and edges correspond to composition). In recent years, there has been a proliferation of DL frameworks [1–4] implemented as domain-specific languages (DSLs) embedded in “high-level” languages¹ such as Python, Java, and C#. These DSLs serve as *abstractions* that aim to map the DL graphs onto hardware pipelines. That is to say, they hide (or *encapsulate*) details of DL models that are judged to be either irrelevant or too onerous to consider. By virtue of these design decisions the frameworks trade-off ease-of-use for execution performance; quoting the architects of PyTorch:

To be useful, PyTorch needs to deliver compelling performance, although not at the expense of simplicity and ease of use. Trading 10% of speed for a significantly simpler to use model is acceptable; 100% is not.

^{*}Both authors contributed equally to this work.

¹For the purposes of this article, we take “high-level” to mean garbage collected and agnostic with respect to hardware from *from the perspective of the user*.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

Conference’17, July 2017, Washington, DC, USA

© 2020 Association for Computing Machinery.

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM...\$15.00

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

Trading off ergonomics for performance is manifestly reasonable², especially during the early phases of the DL engineering/research process (i.e. during the hypothesis generation and experimentation phases). Ultimately one needs to put the DL model into production. It is at this phase of the DL engineering process that every percentage point of execution performance becomes critical. Alternatively, there are many areas of academic DL where the research community strives to incrementally improve performance [6–8]. For example, in the area of super-resolution a deliberate goal is to be able to “super-resolve” in real-time [9]. Similarly, in natural language processing, where enormous language models are becoming the norm [10], memory efficiency of DL models is of the utmost concern. In such instances it’s natural to wonder whether ease-of-use trade-offs that sacrifice execution performance, or memory efficiency, are worthwhile and whether their costs can be mitigated.

Thus, our intent here is to investigate the costs of some of the abstractions employed by framework developers. In particular we focus on the PyTorch framework and ecosystem (chosen for its popularity amongst academic researchers) on the software side and Graphics Processing Units (GPUs) on the hardware side. To that end, we implement a popular and fairly representative³ DL model at four levels of abstraction: conventional PyTorch, LibTorch, cuDNN, and TorchScript. We argue that in the forthcoming that these four implementations do in fact span considerable breadth in the abstraction spectrum. Furthermore we train, test, evaluate each of the implementations on four object detection datasets and tabulate performance and accuracy metrics.

The rest of this article is organizing as follows: section (2) quickly reviews the germane background material on GPUs and DL frameworks, section (3) describes the implementations and our profiling methodology, section (4) presents our results and a comparative discussion thereof, section (5) discusses broad lessons learned, section (6) proposes future work, and section (7) speculates wildly about the future of DL systems more generally.

2 BACKGROUND

2.1 GPUs

We briefly review NVIDIA GPUs⁴ in order that the performance criteria we measure in section (3) are intelligible.

A GPU consists of many simple processors, called streaming multiprocessors (SMs), which are comprised by many compute *cores* that run at relatively low clock speeds⁵. Each compute core in

²“The real problem is that programmers have spent far too much time worrying about efficiency in the wrong places and at the wrong times; premature optimization is the root of all evil (or at least most of it) in programming.” [5]

³In the sense that the functional units constituting the model are widely used in various other models.

⁴A more comprehensive introduction to GPUs themselves and CUDA programming is available in [11].

⁵For example, individual NVIDIA GTX-1080 Ti cores run at ~1500MHz.

an SM can execute one floating-point or integer operation per clock cycle. See fig. (1) for a diagram of NVIDIA's Fermi architecture, where each SM consists of 32 cores, 16 load/store (LD/ST) units, four special-function units (SFUs) which compute transcendental functions (such as sin, cos, exp), a relatively large register file⁶, and thread control logic (to be discussed in the proceeding). Each SM has access to local memory, several cache levels, and global memory. In the Fermi architecture (and subsequent architectures) this local memory is configurable in software; a fraction of it can be apportioned as either local memory or L1 cache (for workloads that query global memory in excess of local memory). One final feature worth mentioning, though irrelevant for us here, is the L2 cache's atomic read-modify-write facilities; this enables sharing data across groups of threads more efficiently than possible in conventional CPUs⁷.

Such an architecture, particularly suited to maximizing throughput, necessitates a programming model distinct from that of a conventional, general purpose processor architecture. A unit of computation deployed to a GPU is called a *kernel*; kernels can be defined using NVIDIA's Compute Unified Device Architecture (CUDA) extensions to C, C++, and FORTRAN⁸. Compiled kernels are comprised by many *threads* that start with the same instruction(s) in parallel; NVIDIA describes this addition to Flynn's taxonomy [12] as Single Instruction Multiple Thread (SIMT)⁹. The large register file enables very fast thread context switching (~25 microseconds on the Fermi architecture [14]), performed by a centralized hardware thread scheduler. Multiple threads are grouped into blocks (SMs are single tenant with respect to blocks) and blocks are grouped into *grids* (grids execute a single kernel). All threads in a block, by virtue of running on the same SM, coordinate (execute in arbitrary order, concurrently, or sequentially) and share memory. Thread blocks are partitioned into *warps* of 32 threads; it is these warps that are dispatched by the warp scheduler (see fig. (1b)) and starting with the Fermi architecture two warps can be executed concurrently on the same SM in order to increase utilization¹⁰.

We present an example CUDA program in fig. (2) to illustrate some of the artifacts of the CUDA threading model. The premise of the program is performing an element-wise sum of two 32×48 entry matrices. Note that all of the data weighs in at $3 \times 32 \times 48 \times 4 = 18$ kilobytes (well within the bounds of shared memory of any one SM). The actual work of the summing is partitioned across a grid of 6 thread blocks, each containing 16×16 threads. Such a partitioning means each thread can be logically responsible for exactly one sum and therefore the kernel is quite simple (see fig. (2a)). Within the context of a kernel, each thread is uniquely identified by its multi-index in the thread hierarchy (threadIdx and blockIdx). Hence, to carry out the sum, the kernel maps this multi-index to the physical

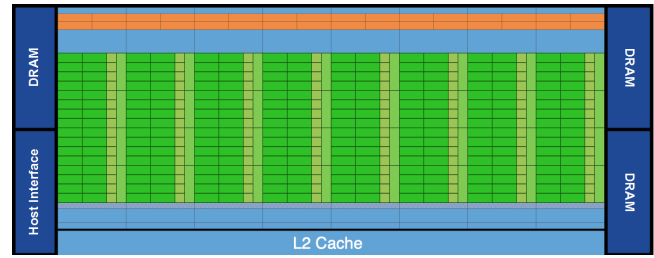
⁶For example, Intel's Haswell architecture supports 168 integer and 168 floating-point registers.

⁷On a CPU, atomic test-and-set instructions manage a semaphore, which itself manages access to memory (therefore incurring a cost of at least two clock cycles).

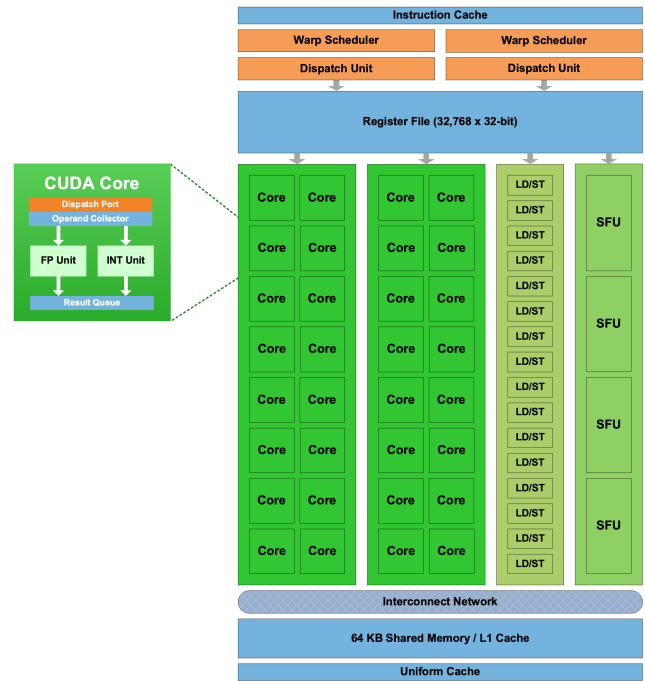
⁸In fact CUDA compiles down to a virtual machine assembly code (by way of nvcc) for a virtual machine called the Parallel Thread Execution (PTX) virtual machine. So, in effect, it's compilers all the way down.

⁹The key difference between SIMD and SIMT is that while in SIMD all vector elements in a vector instruction execute synchronously, threads in SIMT can diverge; branches are handled by predicated instructions [13].

¹⁰I.e. one warp can occupy the compute cores while the other occupies the SFUs or Load/Store units.



(a) Eight (of 16) SM in the Fermi architecture (remaining 8 are symmetrically placed around the L2 cache)



(b) An individual Fermi SM

Figure 1: NVIDIA Fermi Architecture [15]

address of the data¹¹. This (grid, block, thread)→data mapping is, in effect, the mechanism that implements the SIMT architecture. Note that, since each block is allocated to exactly one SM, this sum will take $(16 \times 16) \div 16 = 16$ clock cycles on the Fermi architecture; better throughput could be achieved by increasing the number of blocks (and therefore the number of SMs assigned work).

2.2 Graph compilers

DL frameworks primarily function as graph compilers; they compile abstract dataflow and compute graphs into sequences of operations that execute on various hardware architectures. They typically also implement some “quality of life” abstractions like Tensor¹² and

¹¹In CUDA C/C++ data is laid out in row-major order but this is not fixed (in CUDA FORTRAN the data is laid out in column-major order).

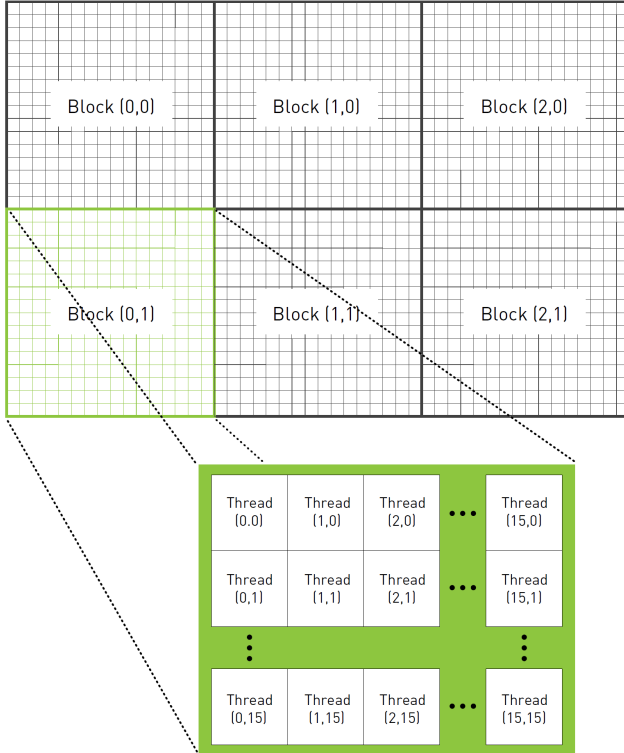
¹²A tensor in this context is a data structure similar to a multidimensional array that supports some useful operations (e.g. slicing, flattening, index permutation). Most DL frameworks also abstract memory layout on hardware behind this abstraction.

```

1 __global__ void matrix_sum(
2     float *A,
3     float *B,
4     float *C,
5     int rows,
6     int cols
7 ) {
8     // blockDim is short for block dimension
9     // blockDim.x == blockDim.y == 16 threads
10    int x = threadIdx.x + blockIdx.x * blockDim.x;
11    int y = threadIdx.y + blockIdx.y * blockDim.y;
12    if (x < cols && y < rows) {
13        int ij = x + y*m; // row-major order
14        C[ij] = A[ij] + B[ij];
15    }
16 }
17
18 int main() {
19     int rows = 32, cols = 48;
20     float A[m][n], B[m][n], C[m][n];
21
22     // initialization and cudaMemcpy
23     // ...
24
25     // dim3 is a 3d integer vector
26     // dimensions omitted in the constructor
27     // (e.g. z like here) are set to 1
28     dim3 numBlocks(3, 2);
29     dim3 numThreads(16, 16);
30     matrix_sum<<blocks, threads>>(A, B, C, rows, cols);
31 }

```

(a) CUDA code to be compiled by nvcc. Note differences `__global__` and `matrix_sum<<, >>` from standard C.



(b) Mapping from thread and block to matrix element.

Figure 2: Canonical CUDA "hello world" kernel (matrix addition).

include utilities useful for the training of DL models (e.g. optimizers and data loaders). Here we focus in particular on the graph compiler functionality.

DL graph compilers are distinct from other dataflow compilers (such as VHDL and Verilog¹³) is that in addition to keeping account of how the data streams through the compute graph, they also keep account of how the gradients of the data stream through the graph (i.e. the *gradient-flow*). This is called *automatic differentiation* (often shortened to *autodiff*). In principle autodiff is implemented by using the rules of Newton's calculus to calculate the derivatives of primitive functions and the chain rule to calculate derivatives of compositions of primitive functions. There are two types of autodiff: *forward mode* (or *forward accumulation*) and *reverse mode* (or *reverse accumulation*)¹⁴. Reverse mode autodiff enables the framework to effectively calculate the gradients of parameters of a neural network with respect to some relevant loss or objective function. Note that such gradients can be *back-propagated* through the neural network in order to adjust the parameters of the neural network such that it minimizes the loss¹⁵ or maximizes the objective.

Dataflow graphs (and their corresponding gradient-flow graphs) can be specified either statically, with fan-in and fan-out for all functions predetermined, or dynamically, where connections between functions are determined "on-the-run". There are advantages and disadvantages to both specification strategies. Static specifications tightly constrain¹⁶ the intricacy of the dataflow graph but, obversely, can be leveraged to improve performance and scalability [16, 17]. TensorFlow (prior to v2.0) is an example of a DL framework that compiles statically specified graphs. Conversely, dynamic specifications can be very expressive and user friendly, including such conveniences as runtime debugging, but are much more difficult to optimize. PyTorch is an example of a DL framework that supports dynamic specification. Both PyTorch and TensorFlow also support just-in-time (JIT) compilation strategies (TorchScript and XLA respectively); such JIT compilers strike a balance between fluency and scalability. In this work we investigate TorchScript (see section (3)).

It warrants mentioning that, in addition to vertically integrated DL frameworks (i.e. specification language and hardware compiler), recently there has been work on intermediate byte code representations for dataflow graphs that arbitrary compiler "frontends" can target. The Multi-Level Intermediate Representation (MLIR) [18] project has goals that include supporting dataflow graphs, optimization passes on those graphs and hardware specific optimizations¹⁷. Stripe [19] is a polyhedral compiler¹⁸ that aims to support general machine learning kernels, which are distinguished by their

¹³Verilog and Very High Speed Integrated Circuit Hardware Description Language (VHSIC-HDL or VHDL) are specification language for specifying circuits on field programmable gate arrays.

¹⁴Briefly, for a composition of functions $y = f(g(h(x)))$, forward mode evaluates the derivative $y'(x)$, as given by the chain rule, inside-out while reverse mode evaluates the derivative outside-in. For those familiar with functional programming, these operations correspond to `foldl` and `foldr` on the sequence of functions with ∂_x as the operator.

¹⁵In which case, it is in fact the negatives of the gradients that are back-propagated.

¹⁶For example, branches and loops are cumbersome to specify statically.

¹⁷Interestingly enough, the project is headed by Chris Lattner who, in developing LLVM, pioneered the same ideas in general purpose programming languages.

¹⁸A polyhedral compiler models complex programs (usually deeply nested loops) as polyhedra and then performs transformations on those polyhedra in order to produce equivalent but optimized programs [20].

high parallelism with limited mutual dependence between iterations. Tensor Comprehensions [21] is an intermediate specification language (rather than intermediate byte code representation) and corresponding polyhedral compiler; the syntax bears close resemblance to Einstein summation notation and the compiler supports operator fusion and specialization for particular data shapes. Finally, Tensor Virtual Machine (TVM) [22] is an optimizing graph compiler that automates optimization using a learning-based cost modeling method that enables it to efficiently explore the space of low-level code optimizations.

3 METHODS

As discussed in the preceding, the translation from high-level neural network specification to hardware native involves a diverse set of choices at design time and translation time. Any such choice made implicitly by the DL framework abstracts away intermediary details at the level of abstraction at which the choice is made. For example, in PyTorch, by default convolutions include not only learnable filters but also a learnable bias. Naturally this increases the number of parameters for which gradients need to be kept account of and updated for. At the next level of abstraction (translation from Python specification to C++ objects) another implicit choice is made in choosing tensor dimension ordering¹⁹. Finally, at the level of abstraction just prior to compilation into CUDA kernels a convolution strategy is chosen²⁰ according to heuristics. Each of these choices potentially incurs a runtime execution and memory cost, depending on whether the heuristics according to which the choice was made apply to the DL model.

With this as subtext, we describe the intent and design of our experiments. We implement the popular object detection deep neural network ResNet-50 [23] at four levels of abstraction (PyTorch, TorchScript²¹, LibTorch, and cuDNN) in order to investigate the differences amongst them. We measure accuracy, execution time, GPU utilization, and memory efficiency of each implementation on four image datasets (MNIST, CIFAR10, STL10, PASCAL). The source for the implementations is available on GitHub²². The datasets were chosen because they span the spectrum of image complexity (from small single-channel to large multi-channel). The reasons for choosing ResNet-50 (see fig. (3)) are two fold. Firstly, it serves as a benchmark architecture in the research community. Secondly, it includes functional units included in many other network architectures (residual units, convolutions of various sizes, batch normalizations, ReLU activations, and pooling layers) and is therefore representative of typical neural network compute workloads. The reason for staying within the same ecosystem is that, in theory, we fix as many of the dimensions of functionality orthogonal to our concerns as possible.

We employ two test platforms (see table (1)); training is performed for 100 epochs with batch size 128 (except for on PASCAL where we use batch size 16) on the “Training platform”. Distinct models were trained in parallel in order to expedite the overall

stage	output	ResNet-50
conv1	112×112	7×7, 64, stride 2
conv2	56×56	3×3 max pool, stride 2
		$\begin{bmatrix} 1 \times 1, 64 \\ 3 \times 3, 64 \\ 1 \times 1, 256 \end{bmatrix} \times 3$
conv3	28×28	$\begin{bmatrix} 1 \times 1, 128 \\ 3 \times 3, 128 \\ 1 \times 1, 512 \end{bmatrix} \times 4$
conv4	14×14	$\begin{bmatrix} 1 \times 1, 256 \\ 3 \times 3, 256 \\ 1 \times 1, 1024 \end{bmatrix} \times 6$
conv5	7×7	$\begin{bmatrix} 1 \times 1, 512 \\ 3 \times 3, 512 \\ 1 \times 1, 2048 \end{bmatrix} \times 3$
	1×1	global average pool 1000-d fc, softmax
# params.		25.5×10⁶

Figure 3: ResNet-50 network architecture [23].

Table 1: Test platforms

CPU	AMD Ryzen 2970WX 24-Core @ 4.2 GHz
GPU	GeForce GTX 1080Ti
HD	Crucial MX500 2TB 3D NAND SATA
RAM	64GB
Software	PyTorch-1.7.0, CUDA-11.1, NVIDIA-455.23.05, g++10.1.0
Training platform	
CPU	Intel Xeon Gold 6230 CPU @ 2.10GHz
GPU	Tesla V100-PCIE-32GB
HD	HPE 800GB SAS 12G Mixed Use SFF
RAM	384GB
Software	PyTorch-1.7.0, CUDA-11.1, NVIDIA-450.80.02
Resolutions platform	

training process but neither data parallelism nor intra-model parallelism were employed. In addition we perform scaling experiments (in resolution and batch size) on the PASCAL dataset; for batch_size, resolution = 8, 16, ..., 1024. For this purpose we employ the “Resolutions platform”, which has GPU RAM enough to accommodate large batch sizes and large image resolutions. For both sets of experiments, each run is repeated 10 times to reduce variance in the measurements. Precise execution time measurements are collected using the CUDA cudaEventCreate, cudaEventRecord, cudaEventElapsedTime APIs. Precise memory and GPU utilization measurements are collecting using the NVIDIA Management Library C API. Both sets of APIs were appropriately wrapped for use in Python.

¹⁹PyTorch uses some heuristics to order tensor dimensions either NCHW or NHWC.

²⁰Winograd convolution, general matrix multiply (GEMM), or FFT convolution.

²¹TorchScript models are serializations of PyTorch models but can run in inference mode in C++, i.e. sans Python runtime.

²²https://github.com/makslevental/pytorch_abstraction_comparison

4 RESULTS

4.1 Training and evaluation

4.2 Memory and utilization

5 DISCUSSION

6 FUTURE WORK

7 SPECULATION

REFERENCES

- [1] Paszke, A. et al, Pytorch: An imperative style, high-performance deep learning library (2019).
- [2] Abadi, M. et al, Tensorflow: Large-scale machine learning on heterogeneous distributed systems (2016).
- [3] Chen, T. et al, Mxnet: A flexible and efficient machine learning library for heterogeneous distributed systems (2015).
- [4] Seide, F. and Agarwal, A., *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16 (Association for Computing Machinery, New York, NY, USA, 2016), p. 2135.
- [5] Knuth, D.E., *Commun. ACM* **17**, 667–673 (1974).
- [6] Abdelhamed, A. et al, Ntire 2020 challenge on real image denoising: Dataset, methods and results (2020).
- [7] Hall, D. et al, *IEEE Winter Conference on Applications of Computer Vision (WACV)* (2020).
- [8] Russakovsky, O. et al, *International Journal of Computer Vision (IJCV)* **115**, 211 (2015).
- [9] Shi, W. et al, *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016), pp. 1874–1883.
- [10] Brown, T.B. et al, Language models are few-shot learners (2020).
- [11] *Professional CUDA C Programming* (Wrox Press Ltd., GBR, 2014), first edn.
- [12] Flynn, M.J., *IEEE Transactions on Computers* **C-21**, 948 (1972).
- [13] NVIDIA, Cuda toolkit documentation, <https://docs.nvidia.com/cuda/cuda-c-programming-guide/index.html#control-flow-instructions> (2020). [Online; accessed 3-December-2020].
- [14] Glaskowsky, P. (2009).
- [15] Wittenbrink, C.M., Kilgariff, E. and Prabhu, A., *IEEE Micro* **31**, 50 (2011).
- [16] Le, T.D. et al, Tflms: Large model support in tensorflow by graph rewriting (2019).
- [17] Pradelle, B. et al, *ESPT/VPA@SC* (2017).
- [18] Lattner, C. et al, Mlir: A compiler infrastructure for the end of moore's law (2020).
- [19] Zerrell, T. and Bruestle, J., Stripe: Tensor compilation via the nested polyhedral model (2019).
- [20] Griebel, M., Lengauer, C. and Wetzel, S., *In IEEE PACT* (IEEE Computer Society Press, 1998), pp. 106–111.
- [21] Vasilache, N. et al, Tensor comprehensions: Framework-agnostic high-performance machine learning abstractions (2018).
- [22] Chen, T. et al, *Proceedings of the 13th USENIX Conference on Operating Systems Design and Implementation*, OSDI'18 (USENIX Association, USA, 2018), p. 579–594.
- [23] He, K. et al, Deep residual learning for image recognition (2015).

Appendices

A APPENDIX