Benanza: Automatic μBenchmark Generation to Compute “Lower-bound” Latency and Inform Optimizations of Deep Learning Models on GPUs

Cheng Li*, Abdul Dakkak*
University of Illinois Urbana-Champaign
Urbana, USA
{cli99,dakkak}@illinois.edu

Jinjun Xiong
IBM T. J. Watson Research Center
Yorktown Heights, USA
jinjun@us.ibm.com

Wen-mei Hwu
University of Illinois Urbana-Champaign
Urbana, USA
w-hwu@illinois.edu

Abstract—As Deep Learning (DL) models have been increasingly used in latency-sensitive applications, there has been a growing interest in improving their response time. An important venue for such improvement is to profile the execution of these models and characterize their performance to identify possible optimization opportunities. However, the current profiling tools lack the highly desired abilities to characterize ideal performance, identify sources of inefficiency, and quantify the benefits of potential optimizations. Such deficiencies have led to slow characterization/optimization cycles that cannot keep up with the fast pace at which new DL models are introduced.

We propose Benanza, a sustainable and extensible benchmarking and analysis design that speeds up the characterization/optimization cycle of DL models on GPUs. Benanza consists of four major components: a model processor that parses models into an internal representation, a configurable benchmark generator that automatically generates micro-benchmarks given a set of models, a database of benchmark results, and an analyzer that computes the “lower-bound” latency of DL models using the benchmark data and informs optimizations of model execution. The “lower-bound” latency metric estimates the ideal model execution latency on a GPU system and serves as the basis for identifying optimization opportunities in frameworks or system libraries. We used Benanza to evaluate 30 ONNX models in MXNet, ONNX Runtime, and PyTorch on 7 GPUs ranging from Kepler to the latest Turing, and identified optimizations in parallel layer execution, cuDNN convolution algorithm selection, framework inefficiency, layer fusion, and using Tensor Cores.

I. INTRODUCTION

The past few years have seen a spur of deep learning (DL) innovations. These innovations span from DL models to software stack optimizations (e.g., frameworks such as MXNet or PyTorch, libraries such as cuDNN or MKL-DNN) and hardware stack improvements (e.g., CPU, GPU, FPGA). Among all the innovations, however, DL models are the most rapidly evolving and prolific. This is true in both academia [1] and industry [2], where models are tweaked and introduced on a weekly, daily, or even hourly basis.

Both industry and academia have invested heavily in developing benchmarks to characterize DL models and systems [3]–[7]. Characterization is followed by optimizations to improve the model performance. However, there is currently a gap between the benchmarking results and possible optimizations to perform. Researchers use profilers, such as nvprof [8], Nsight [9], and VTune [10], to profile and get low-level GPU and CPU information. With ample knowledge of how models execute and utilize system resources, researchers manually identify bottlenecks and inefficiencies within model execution using the profilers. Researchers then make hypotheses of solutions, and try out different ideas to optimize the model execution — which may or may not pan out. This manual and ad-hoc process requires a lot of effort and expertise and slows down the turnaround time for model optimization and system tuning.

Thus there is a need for a systematic DL benchmarking and subsequent analysis design that can guide researchers to potential optimization opportunities and assess hypothetical execution scenarios. Since for GPUs model execution latency is determined by the hardware, framework, and system libraries (primarily cuDNN [11] and cuBLAS [12] for DL), answers to the following questions are highly desired by researchers: Q1 What is the potential latency speedup if optimizations are performed? Q2 Are independent layers executed in parallel? Q3 Are convolution layers using the optimal convolution algorithms? Q4 Are there any inefficiencies or unexpected behavior in a framework? Does the execution Q5 fuse layers or Q6 leverage Tensor Cores, and what are the benefits? We motivate our design by answering these 6 questions, while ensuring the sustainability and extensibility of the design.

To answer these questions, we first propose a new benchmarking metric: “lower-bound” latency. The “lower-bound” latency estimates the ideal latency of a DL model given a software and
hardware stack, and is based on the following observations:
(1) DL models are executed as layers in frameworks and thus layers form the performance building blocks of DL models. (2) Frameworks delegate execution of common layers to either cuDNN or cuBLAS (shown in Figure 1). The “lower-bound” latency is defined in terms of the latencies of the cuDNN and cuBLAS API functions invoked by model layers (framework overhead and memory transfers are ignored). We refine the “lower-bound” latency and define it under sequential execution mode (all layers are executed sequentially) and parallel execution mode (data-independent layers are executed asynchronously).

This paper presents Benanza (pronounced bonanza) — an sustainable and extensible benchmarking and analysis design. Benanza consists of a set of modular components: (1) a model processor to process input ONNX models into a set of unique layers (layers are considered the same if they have the same layer type, shape, and parameters), (2) a benchmark generator to automatically generate parameterized cuDNN and cuBLAS micro-benchmarks from the unique layers, (3) a performance database to store historical benchmark results, and (4) an analyzer to compute the “lower-bound” latency of DL models and inform potential optimizations

Benanza is architected to be sustainable. The benchmarking workflow of Benanza is highly automated and minimizes the benchmark development and maintenance effort. Benanza uses the observation that DL models have repeated layers (i.e. non-unique) within and across models to decrease the time to benchmark. When a new model is introduced, only the newly un-benchmarked layers that do (not in the performance database) need to be benchmarked. Although the focus of the paper is on NVIDIA GPUs using cuDNN and cuBLAS, the design proposed is extensible and users can incorporate other benchmark runtimes that target other software libraries or hardware such as: frameworks’ API or MKL-DNN for CPUs.

In summary, this paper makes the following contributions:
• We propose a “lower-bound” latency metric for DL models based on the observation that the latency of a DL model is bounded by the latencies of the cuDNN and cuBLAS API calls invoked by the model layers. This metric estimates the ideal latency of a model given a specific GPU hardware and software stack.
• We present Benanza, a benchmarking and analyzing design that automatically generates micro-benchmarks given a set of models, computes their “lower-bound” latencies using the benchmark data, and informs optimizations of their executions on GPUs. The sustainable and extensible design of Benanza makes it cope with the fast evolution of DL innovations.
• Using Benanza, we characterized the “lower-bound” latencies of 30 ONNX models (shown in Table I) in MXNet, ONNX Runtime, and PyTorch on 7 systems (shown in Table III). We performed a comprehensive “lower-bound” latency analysis as we vary the model, execution mode, batch size, and system. E.g., when using parallel execution mode, up to 2.87× (with a geometric mean of 1.32× across models) latency speedup could be made to MXNet using batch size 1 on the Tesla_V100 system.
• We further identified optimization opportunities through Benanza in cuDNN convolution algorithm selection (up to 1.32× geometric mean speedup across models), inefficiencies within MXNet (up to 1.15× speedup across models) and PyTorch (up to 2.3× speedup using batch size 1), layer fusion, and Tensor Cores (up to 1.09× and 1.72× speedup for ResNet50-v1 respectively). We evaluated the above optimizations jointly and get up to 1.95× speedup for ResNet50-v1 across systems and batch sizes.
• The usages and evaluation of Benanza are posted to benanza.mlmodelscope.org for public inspection.

II. BACKGROUND AND MOTIVATION

A. DL Model Execution and ONNX Format

A DL model is an execution graph where each vertex is a layer operator (e.g. convolution, activation, normalization, pooling, or softmax). These layer operators (or layers for short) are functions defined by a DL framework. A framework executes a model by traversing the model graph in topological order and enqueuing the layers into an execution queue. Although sequential evaluation is always valid, frameworks strive to execute data-independent layers within the queue in parallel. Through careful execution scheduling, a framework can overlap communication with computation, increase utilization, etc. Regardless of the execution strategy, however, layer execution latency is the limiting factor for model execution. As thus, layers are not only the building blocks by which developer define models, but are also the atomic components that define a model’s performance characteristics.

Each framework provides its own API, layer definition semantics, model storage format, and model executing strategy. To increase interoperability between frameworks, there have been concerted efforts [13, 14] to standardize layer definitions and model exchange format. A leading effort is the Open Neural Network Exchange Format (ONNX), which has wide industry and framework backing. Frameworks such as Caffe2, CNTK, MXNet, Paddle, PyTorch, and TensorRT readily support ONNX, and converters exist for other frameworks such as TensorFlow and Caffe. To perform a fair comparison between frameworks (by evaluating them using the same ONNX model), and more importantly, to make Benanza framework-agnostic, we choose ONNX as the model input format for Benanza. ONNX hosts all their models publicly [15] and, we select 30 vision models out of the 32 models available at the time of writing for evaluation (the 2 models not selected are non-vision models). The selected models cover an array of tasks and are listed in Table I. We refer to these models by their IDs throughout the paper.

B. cuDNN and cuBLAS

Much like BLAS or LAPACK are the backbone of HPC computing, cuDNN and cuBLAS are the backbones of the GPU software stacks for DL. cuDNN is a GPU-accelerated library and provides highly-tuned implementations of DL layers such as convolution, pooling, normalization, activation. cuBLAS is a GPU-accelerated BLAS library and provides
TABLE I: The 30 ONNX models used are vision models which encompass image classification (IC), object detection (OD), face recognition (FR), emotion recognition (ER), semantic segmentation (SS), or hand digit recognition (HR) tasks.

| ID | Name               | Task | MACs   | # Layers | Year |
|----|--------------------|------|--------|----------|------|
| 1  | Arcface [10]       | FR   | 12.08G | 412      | 2018 |
| 2  | BVLC-Alexnet [17]  | IC   | 656M   | 24       | 2012 |
| 3  | BVLC-Caffenet [17] | IC   | 721M   | 24       | 2012 |
| 4  | BVLC-Googlenet [18]| IC   | 1.59G  | 143      | 2014 |
| 5  | BVLC-RCNN-ILSVRC13[19]| IC | 718M | 23       | 2013 |
| 6  | DenseNet-121 [20]  | IC   | 2.87G  | 910      | 2016 |
| 7  | DUC [21]          | SS   | 34.94G | 355      | 2017 |
| 8  | Emotion Ferplus [22]| ER   | 877M   | 52       | 2016 |
| 9  | Inception-v1 [23]  | IC   | 1.44G  | 144      | 2015 |
| 10 | Inception-v2 [24]  | IC   | 2.03G  | 509      | 2015 |
| 11 | LeNet [25]        | HR   | 796K   | 12       | 2010 |
| 12 | MobileNet-v2 [26] | IC   | 437M   | 155      | 2017 |
| 13 | Resnet18-v1 [27]  | IC   | 1.82G  | 69       | 2015 |
| 14 | Resnet18-v2 [28]  | IC   | 1.82G  | 69       | 2015 |
| 15 | Resnet34-v1 [27]  | IC   | 3.67G  | 125      | 2015 |
| 16 | Resnet34-v2 [28]  | IC   | 3.67G  | 125      | 2015 |
| 17 | Resnet50-v1 [27]  | IC   | 3.87G  | 175      | 2015 |
| 18 | Resnet50-v2 [28]  | IC   | 4.10G  | 174      | 2016 |
| 19 | Resnet101-v1 [27] | IC   | 7.58G  | 345      | 2015 |
| 20 | Resnet101-v2 [28] | IC   | 7.81G  | 344      | 2016 |
| 21 | Resnet152-v1 [27] | IC   | 11.30G | 515      | 2015 |
| 22 | Resnet152-v2 [28] | IC   | 11.53G | 516      | 2014 |
| 23 | Shufflenet [29]   | IC   | 127M   | 203      | 2015 |
| 24 | SqueezeNet-v1.1 [30]| IC | 352M | 66       | 2016 |
| 25 | Tiny Yolo-v2 [31] | OD   | 3.13G  | 32       | 2016 |
| 26 | Vgg16-BN [32]     | IC   | 15.38G | 54       | 2014 |
| 27 | Vgg16 [32]        | IC   | 15.38G | 41       | 2014 |
| 28 | Vgg19-bn [32]     | IC   | 19.55G | 63       | 2014 |
| 29 | Vgg19 [32]        | IC   | 19.55G | 47       | 2014 |
| 30 | Zfnet152 [33]     | IC   | 1.48G  | 22       | 2013 |

Fig. 3: The percentage of layers supported by cuDNN and cuBLAS (also covered by Benanza) for each model in Table I.

fast implementations of GEMM and GEMV. The DL layers supported by each API are listed in Table II. And, while there is a wide array of DL frameworks, common between them is the reliance on these primitives defined by cuDNN and cuBLAS. In fact, all major DL frameworks, such as MXNet, PyTorch, ONNX Runtime, and TensorFlow, rely on cuDNN/cuBLAS API functions for the implementation of common layers.

Figure 3 shows the percentage of layers supported by cuDNN and cuBLAS for each model in Table I. Most layers within DL models are covered by the cuDNN and cuBLAS API. The layers that are not supported are non-compute operators (such as concatenate, which joins two tensors across a specified axis) or datatype manipulations (such as reshape, which changes the dimensions of a tensor). For example, the cuDNN and cuBLAS functions support 70% of the layers within Inception-v2 (ID = 10). This is because Inception-v2 makes heavy use of unsqueeze — a tensor reshape layer; 27% of the layers in Inception-v2 are unsqueeze layers.

Given a specific DL software stack (e.g. framework, cuDNN, cuBLAS, and other CUDA libraries) and GPU hardware, the cuDNN and cuBLAS functions invoked by a model are fixed. Most common layers are supported by cuDNN and cuBLAS and the latency attributed to cuDNN and cuBLAS functions is significant with respect to the model’s end-to-end latency. Figure 1 shows that for the 30 vision models, the time spent within the cuDNN and cuBLAS API calls dominates the model execution time. The “other” time is due to either memory operations, synchronization, the framework’s choice of not using cuDNN API for certain operations, or other framework code that is neither cuDNN nor cuBLAS.

Based on the above observations, we propose a “lower-bound” latency metric for DL models. The “lower-bound” metric is
The Model Processor takes convolution algorithm selection, layer fusion, and Tensor Core usage. parallel execution of independent layers, shape inference [34], is performed to determine the shape of each layer. Layers in the IR (referred to as layers and correspond to the ONNX nodes) are annotated with the inferred shapes. Benchmarks are generated for each layer using its type, shape, and parameters information.

We observe that layers with the same type, shape, and parameters (i.e. only differ in weight values) are repeated extensively within and across models. Figure 4 shows that most models have a low percentage of unique layers — indicating that layers are repeated extensively within the model. For example, ResNet50-v1 (ID=17) has 175 layers but only 47 (26.9%) are unique. The number of unique layers across models of similar architecture is also low. The ResNet++-v1 models (ID=13, 15, 17, 19, 21) are built from the same modules and have a total of 1229 layers, of which only 60 (5.6%) are unique. Across all 30 models, the total number of layers is 5754, but only 1031 (18%) are unique. We exploit this layer repeatability to optimize the benchmark generation and minimize the time to benchmark. Thus, the Model Processor unifies the repeated layers across the input models and produces a set of unique layers. The time saved can be used to explore other algorithms and data types (Sections III-B2 and III-B3) benchmarks.

III. Benanza Design and Implementation

Benanza consists of four main components: Model Processor, Automatic Benchmark Generator, Performance Database, and Analyzer. The components are shown in Figure 2 and are used in the benchmarking and analysis workflows:

- **Benchmarking workflow:** The Model Processor takes ONNX models, parses them, performs shape inference and finds the set of unique layers within the models. Two layers are considered the same (non-unique) if they have the same operator type and parameters (i.e. only differ in weight values). The Automatic Benchmark Generator then generates micro-benchmarks for each unique layer. The generated micro-benchmarks measure the latency (or the GPU kernel metrics if profiling mode is enabled) of the corresponding cuDNN or cuBLAS function calls for the layers. The micro-benchmarks are then run on systems of interest and the results are stored in the Performance Database.

- **Analysis workflow:** The user runs the target model using a framework on a system of interest with utilities provided by Benanza to get the model execution profile (i.e. the end-to-end latency, cuDNN and cuBLAS logs, and Nsight profile). The user then specifies the model and system to Benanza. The model is parsed into layers and the Analyzer queries the latencies of each layer from the Performance Database (using the layers and system information provided) to compute the "lower-bound" latency under different execution scenarios. The model execution profile and the computed "lower-bound", the Analyzer informs optimizations in: parallel execution of independent layers, convolution algorithm selection, framework inefficiency, layer fusion, and Tensor Core usage.

A. Benanza Model Processor

The Model Processor parses ONNX models into Benanza’s internal representation (IR). The IR wraps around the ONNX Protobuf and has the same layer coverage. Since ONNX models do not have layer shapes information embedded (except for the input layers), shape inference is performed to determine the shape of each layer. Layers in the IR (referred to as layers and correspond to the ONNX nodes) are annotated with the inferred shapes. Benchmarks are generated for each layer using its type, shape, and parameters information.

We observe that layers with the same type, shape, and parameters (i.e. only differ in weight values) are repeated extensively within and across models. Figure 4 shows that most models have a low percentage of unique layers — indicating that layers are repeated extensively within the model. For example, ResNet50-v1 (ID=17) has 175 layers but only 47 (26.9%) are unique. The number of unique layers across models of similar architecture is also low. The ResNet++-v1 models (ID=13, 15, 17, 19, 21) are built from the same modules and have a total of 1229 layers, of which only 60 (5.6%) are unique. Across all 30 models, the total number of layers is 5754, but only 1031 (18%) are unique. We exploit this layer repeatability to optimize the benchmark generation and minimize the time to benchmark. Thus, the Model Processor unifies the repeated layers across the input models and produces a set of unique layers. The time saved can be used to explore other algorithms and data types (Sections III-B2 and III-B3) benchmarks.

B. Automatic Benchmark Generator

The Automatic Benchmark Generator uses the set of unique layers (produced by the Model Processor) and generates C code to invoke the benchmark runtime using each layer’s type, shape, and parameters information.

1) The Benchmark Runtime: Benanza provides a benchmark runtime that measures the latency of the cuDNN or cuBLAS API required to execute each layer (as shown in Table II). The runtime also sets up the function arguments for each API. The setup time is not included in the latency measurement. The runtime uses the Google Benchmark library — a micro-benchmarking support library. The Google Benchmark library dynamically determines the number of iterations to run each benchmark and ensures that the reported latency results are statistically stable. Generated benchmarks are linked with the cuDNN/cuBLAS libraries, and are run on systems of interest.

2) Algorithm Instantiation: The convolution layers map to the cudnnConvolutionForward API (Table II). The convolution API takes one of the following 8 algorithms as an argument: Implicit GEMM (IGEMM), Implicit PreComputed GEMM (IPGEMM), GEMM, Direct (DRCT), FFT, Tiled FFT (TFFT), Winograd (WING), and Winograd Non-Fused (WINGNF). These algorithms have different compute and memory characteristics. The optimal algorithm to use depends on the system, layer shape, and layer parameters (e.g. filter size, stride, dilation, etc.) [11]. For inference, most frameworks (e.g. MXNet, PyTorch, TensorFlow) rely on the cuDNN provided heuristic function (cudnnGetConvolutionForwardAlgorithm) to choose the convolution algorithm. The heuristic function suggests an algorithm given the layer’s shape, parameters, data type, system, etc. To explore the design space of algorithm selection, by default, for each layer Benanza generates benchmarks using all algorithms applicable to the layer.

3) Data Type Support: Benanza can be configured to generate micro-benchmarks that target different data types. Both float16 and float32 are generated by default, but benchmarks can be instantiated for other data types. The benchmarks use Tensor Cores when the API function (see Table II) and system (see Table III) supports it.
The user runs the target model using a framework on a system of interest with utilities provided by Benanza to get the model execution profile. The model execution profile contains information about the model’s end-to-end latency, cuDNN and cuBLAS logs, and Nsight profile (which contains cuDNN/cuBLAS API calls and function backtrace information). Capturing the model end-to-end latency requires the user to place the provided timing functions within their application code. To capture the usage of cuDNN and cuBLAS functions within a framework, Benanza launches the user code with the \texttt{CUDNN\_LOGININFO\_DBG} and \texttt{CUBLAS\_LOGININFO\_DBG} environment variables. These environment variables enable the cuDNN and cuBLAS loggers respectively. Utilities to run the user code using NVIDIA’s Nsight profiler are also provided. The results from Nsight are parsed and correlated with the cuDNN and cuBLAS logs.

The user then inputs the model execution profile along with the ONNX model, system, data type. The model is parsed by the Model Processor into layers. Then, the Benanza Analyzer queries the Performance Database for the benchmark latencies of each layer using the user-specified system and data type (by default \texttt{float32}). Due to algorithm (Section III-B2) instantiation, multiple benchmarks may exist for a layer. The Analyzer, therefore, selects the benchmark result achieving the lowest latency. The following analyses are then performed:

1) \textbf{Sequential and Parallel “Lower-Bound” Latency:} DL models may contain layer sequences which can be executed independently in parallel. The sub-graph formed by these data-independent layer sequences is called a parallel module. For example, a parallel module in Inception-v1 is shown in Figure 5. A framework may execute the independent paths within the parallel module either sequentially or in parallel. Thus, the Analyzer computes the “lower-bound” latency of a model using two execution modes: sequential and parallel.

The sequential mode assumes that independent layers are executed sequentially, and therefore is defined as the sum of each layer’s benchmark latency. The parallel strategy assumes that data-independent layers are executed in parallel. Therefore, the parallel “lower-bound” latency is defined by the model’s critical path — the simple path from the start to the end layer with the highest latency. Finding the critical path of a graph is a longest path problem and is NP-hard. Since a DL model forms a directed acyclic graph (DAG), the critical path can be framed as a shortest path problem [40]. To compute the critical path we construct a weighted DAG from the model graph where the edge weight between two nodes (layers) is negative of the latency of the layer at the tail of the edge. Computing the shortest path from the start to the end layer of the constructed weighted DAG produces the critical path of the model. The parallel “lower-bound” latency is the sum of layers latencies along the critical path. Benanza visualizes the critical path of the model (e.g. Figure 5), and the difference between the sequential and parallel “lower-bound” latencies indicates the profit of executing independent layers in parallel.
Other analyses performed by *Benanza* leverage the sequential and parallel “lower-bound” latencies, and the benefits can be calculated in terms of either sequential or parallel mode.

2) **Convolution Algorithm Selection**: The Analyzer uses the parsed cuDNN log in the model execution profile to determine if the cuDNN algorithm used by the framework for each layer is optimal (recall from Section III-B2 that benchmark results using all available algorithms for layers exist in the Performance Database). Cases where the algorithm choice is sub-optimal are reported to the user along with how much end-to-end latency improvement could be gained if algorithm selection was ideal. The user can act upon these suggestions by forcing the framework to use specific algorithms.

3) **Framework Inefficiency Inspection**: The expected cuDNN and cuBLAS API calls are known to the Analyzer from the “lower-bound” latency computation. The Analyzer compares the model execution profile against the expected execution to pinpoint inefficiencies within the framework. The user is presented with any deviation observed in cuDNN or cuBLAS API invocation’s parameters or their execution order. CUDA API functions and CUDA kernels executed between cuDNN or cuBLAS API calls, are also presented to the user—along with their backtraces.

4) **Layer Fusion Analysis**: If the user enables the benchmark generation for layer fusion (as described in Section III-B4), then the Analyzer can be used to determine the potential profitability if layer fusion is employed. The Analyzer traverses the model layers and looks for the fusion pattern rules (listed in Section III-B4). If one of these patterns is found, then the corresponding fused operation’s latency is queried from the database and is used in the “lower-bound” computation (in either sequential or parallel model). If the benchmark is unavailable, or failed to run, then the latencies of the non-fused layers are used. The difference between the non-fused “lower-bound” latency and the fused “lower-bound” latency determines the profitability of layer fusion.

5) **Tensor Core Analysis**: The Analyzer determines if the target model execution utilizes Tensor Cores by looking at kernel names in the model execution profile. Kernel names that match the \_\[ish\]\d+ Regular-expression use Tensor Cores. By default, benchmarks targeting both float16 and float32 are generated. When benchmarks are run on systems with Tensor Core support, the difference between the “lower-bound” latency of float32 and float16 informs the profitability of using Tensor Cores and float16.

E. **Sustainability and Extensibility**

Sustainability of *Benanza* is ensured by providing an automated benchmark generation and analysis workflow design along with a continuously updated Performance Database. Benchmarking requires limited effort, as the micro-benchmarks are automatically generated, and the user only needs to compile and run the generated code on systems of interest. The Performance Database is continuously updated with new benchmark results. A big insight of the proposed design is that there is ample layer repeatability within and across models. This keeps the number of unique layers and thus the number of Performance Database entries in check over time. For new models, only the newly introduced unique layers are benchmarked.

For example, consider a scenario where all models in Table I except for ResNet*-v2 have already been benchmarked and the results are in the Performance Database. Using our design, benchmarking the ResNet*-v2 models requires measuring all the ResNet*-v2 layers that are not within the Performance Database. Evaluating this hypothetical scenario results in a 75% reduction (30 minutes) in benchmarking time on the Tesla_V100 system for batch size 32. The saving would be even larger on slower systems. By storing and reusing the micro-benchmark results in the Performance Database we minimize the time cost of running micro-benchmarks.

*Benanza* is extensible. As shown in Figure 2, *Benanza* is designed as a set of modular components. As new cuDNN functions are introduced, users update the *Benanza* runtime accordingly. For example, if a new cuDNN convolution algorithm is added, then the user can just add it to the list of algorithms to instantiate in the convolution benchmark implementation. If a new cuDNN/cuBLAS API or a fused API is added, then a user needs to add the benchmark implementation for the new API using the templates provided by *Benanza* as a basis. Users can also extend the Automatic Benchmark Generator to support other runtimes that target other software libraries or hardware, and leverage most of the other analysis components unmodified. These runtimes can target the frameworks’ Python or C++ API or other DL libraries (e.g. MIOpen [41] on AMD GPUs, or MKL-DNN [42] on CPUs). Through the novel benchmarking and analysis design, *Benanza* copes well with the fast evolving pace of DL innovations.

IV. **Evaluation**

We implemented *Benanza* and evaluated its design by answering Q1-6. We evaluated 30 ONNX models (listed in Table I) in the MXNet (v1.5.1), ONNX Runtime (v0.5.0), and PyTorch (v1.3) frameworks. Experiments were run on the 7 systems listed in Table III. All systems use Ubuntu 18.04.3 LTS, CUDA 10.1.243, cuDNN Version 7.6.3, and CUDA Driver 430.26. The micro-benchmarks were compiled with GCC 7.4.0. We first computed the float32 “lower-bound” latency in both sequential and parallel modes. Then we used the Analyzer to uncover and explore optimization opportunities — cuDNN heuristics, framework inefficiencies, layer fusion, and usage of Tensor Cores, and show their impact on the end-to-end latency. The reader is encouraged to explore further documentation and experimentation at benanza.mlmodelscope.org.

A. **End-to-end Latency vs. “Lower-Bound” Latency**

We measured the end-to-end latency of the 30 models using MXNet, ONNX Runtime, and PyTorch on the Tesla_V100 system. Figure 6 shows the latency difference across all models and Figure 7 compares the models across the frameworks. Due to the lack of support of some ONNX operators by ONNX Runtime [43] and PyTorch [44], not all models run within these
TABLE III: We used 7 GPU systems for evaluation. The systems cover the past GPU generations (from Kepler to the latest Turing). Amazon cloud (AWS) is used for 4 of the systems and the other 3 are local machines. The 4 Turing and Volta GPUs support Tensor Cores and their theoretical Tensor Core performance (Tensor TFLOPS) are listed.

| Name               | CPU                                   | GPU (Release Year) | GPU Architecture | GPU Memory Capacity, Bandwidth | Theoretical FP32 TFLOPS | Theoretical FP64 TFLOPS |
|--------------------|---------------------------------------|--------------------|------------------|--------------------------------|-------------------------|-------------------------|
| Tesla_K80 (AWS P2) | Intel Xeon CPU E5-2686 v4             | Tesla K80 (2014)   | Kepler           | 12 GB, 480 GB/s               | 5.6                     | ×                       |
| Tesla_M60 (AWS G3) | Intel Core i9-7900X CPU               | Tesla M60 (2015)   | Maxwell          | 7 GB, 1604 GB/s              | 4.8                     | ×                       |
| TITAN_Xp           | Intel Xeon CPU E5-2686 v4             | TITAN Xp (2017)    | Pascal           | 12 GB, 547.6 GB/s            | 12.2                    | ×                       |
| TITAN_V            | Intel Core i7-7820X CPU               | TITAN V (2017)     | Volta            | 16 GB, 672 GB/s              | 14.9                    | 110.0                   |
| Tesla_V100 (AWS P3)| Intel Xeon CPU E5-2686 v4             | Tesla V100 SXM2 (2018)| Volta    | 16 GB, 900 GB/s              | 15.7                    | 125.0                   |
| Quadro RTX         | Intel Xeon CPU E5-2630 v4             | Quadro RTX 6000 (2019)| Turing  | 24 GB, 624 GB/s              | 16.3                    | 130.5                   |
| Tesla_T4 (AWS G4)  | Intel Xeon Platinum 8250CL CPU        | Tesla T4 (2019)    | Turing           | 15 GB, 320 GB/s              | 8.1                     | 65.0                    |

**Fig. 6:** The end-to-end latency of all ONNX models using batch size 1 with MXNet backend on Tesla_V100 in Table III.

**Fig. 7:** The end-to-end latency of all ONNX models with MXNet, ONNX Runtime, and PyTorch backends (normalized to MXNet latency) using batch size 1 on Tesla_V100.

Frameworks. As MXNet is the fastest in general, subsequent sections of the paper (with the exception of Section IV-C) focus on informing optimizations in MXNet.

1) **Sequential Mode vs Parallel Mode:** The difference between the “lower-bound” latency and the end-to-end latency indicates the optimization opportunities in the framework and its use of the cuDNN and cuBLAS APIs. The “lower-bound” latency of a model normalized to the model’s end-to-end latency in a framework is referred to as normalized “lower-bound” latency. Figure 8 shows the normalized “lower-bound” latency using sequential and parallel modes in MXNet across all models using batch size 1 on Tesla_V100.

The sequential normalized “lower-bound” latencies across models have a geometric mean of 0.88, thus a potential latency speedup of $\frac{1.14}{0.88} = 1.32 \times$ can be made to MXNet. The parallel “lower-bound” latency across models has a geometric mean of 0.76, indicating a potential latency speedup of $\frac{1.0}{0.76} = 1.32 \times$. The difference between a model’s parallel and sequential “lower-bound” latency depends on the existence of parallel modules within the model and how compute-intensive the date-independent paths are. Models without parallel modules have the same parallel and sequential “lower-bound” latency. For models with compute-intensive parallel modules, such as the Inception models (ID=4, 9, 10), the potential speedup of the MXNet’s latency is $2.87 \times$, $2.69 \times$, and $2.45 \times$ respectively. The sequential and parallel normalized “lower-bound” latencies of LeNet (ID=11) are both low because LeNet is a simple model which has low latency ($0.33\text{ms}$ as shown in Figure 6). For LeNet, the MXNet overhead and other non-compute portion is high, thus the normalized “lower-bound” latency is low.

The sequential “lower-bound” latencies of the models with parallel modules (e.g. Inception and ResNet models) are closer to their end-to-end latency (when compared to the parallel “lower-bound”). This suggests that parallel modules are mostly executed sequentially in MXNet, even though the independent layers could be run in parallel. We verified the sequential execution behavior in MXNet by inspecting the model execution profile. We evaluated the benefits of the latter optimizations in terms of the serial “lower-bound” latency.

2) **Batch Sizes and Systems:** To demonstrate Benanza’s functions across batch sizes and systems, we evaluated the “lower-bound” latency of all models using different batch sizes from 1 to 32 on representative systems (shown in Table III).

**Fig. 8:** The sequential and parallel normalized “lower-bound” latency in MXNet using batch size 1 on Tesla_V100.

**Fig. 9:** The end-to-end latency of ResNet50_v1 in MXNet across batch sizes and systems.

**Fig. 10:** The cuDNN heuristic selects 8 non-optimal convolution layer algorithms for ResNet50_v1 using batch size 32 on Tesla_V100. Up to $2.75 \times$ speedup can be achieved if selection was ideal.
Inefficiencies in Frameworks

We select batch size 32, since some models cannot be run using batch sizes beyond 32 due to GPU memory limitations. Figure 9 shows the end-to-end latency of ResNet50-v1 on all systems in log scale. As expected, latencies are reversely correlated to the compute capability of the system (e.g. theoretical FP32 TFLOPS in Table III). ResNet50-v1 has a higher latency on Quadro_RTX when compared to Tesla_V100, since Quadro_RTX is a desktop-grade GPU with a memory bandwidth of 624 GB/s whereas Tesla_V100 is a server-grade GPU and a memory bandwidth of 900 GB/s.

Figure 12 shows the (sequential) normalized “lower-bound” latency of ResNet50-v1 across batch sizes and systems. The figure suggests that ResNet50-v1’s optimization opportunities are system and batch size dependent. Both Tesla_V100 and TITAN_V are highly optimized to run ResNet50-v1 across batch sizes, since their normalized latencies are high — ranging from 0.86 to 1.0. The normalized latencies for Tesla_T4 and Quadro_RTX are high for batch sizes 1 to 4 but drop beyond that. ResNet50-v1 is less optimized on the other systems and has low normalized “lower-bound” latencies.

The geometric mean of the normalized “lower-bound” latencies for all the models across systems and batch sizes is shown in Figure 13. Tesla_V100 and TITAN_V, are still observed to have high normalized latencies (0.76 – 0.88). A drop was still observed for Tesla_T4 and Quadro_RTX at batch size 4. Tesla_M60 and TITAN_Xp have normalized “lower-bound” latencies between 0.63 and 0.72. The oldest GPU generation, Tesla_K80, is the least optimized and has the lowest normalized “lower-bound” latency.

Overall, the current software stack (latest MXNet, cuDNN, and CUDA libraries used in the evaluation) is more optimized on the recent GPU generations (Turing and Volta) and for smaller batch sizes. Compared to Volta, the software stack is less optimized for Turing. This is possibly because Turing was newly released, and we expect optimizations that target Turing to increase. Moreover, the low normalized “lower-bound” latencies on the older GPUs suggest that vendors prioritize optimizations for newer GPU generations over older ones.

B. cuDNN Convolution Heuristics

Using the Benanza Analyzer, we observed that heuristics employed by cuDNN (and subsequently the frameworks) are not always optimal. For example, Figure 10 shows the convolution layer latencies using the algorithms informed by cuDNN heuristics (labeled as cuDNN Heuristic) normalized to using the optimal algorithm (labeled as Ideal Algorithm) for ResNet50_v1 using batch size 32 on Tesla_V100. The algorithm choices are listed in Section III-B2. Figure 14 shows the end-to-end latency speedup for ResNet50_v1 across batch sizes and systems by using the optimal convolution algorithm for all convolution layers. Figure 15 shows the geometric mean of the end-to-end latency speedup for all models by using the optimal algorithms. At batch size 32, the speedup ranges between 1.14 and 1.32× across GPUs. Both cutting edge and older GPU architectures can benefit from better cuDNN heuristics.

C. Inefficiencies in Frameworks

We used Benanza to identify the inefficiencies in MXNet and PyTorch. We then implemented the optimizations informed by Benanza and show the latency speedup after the framework modifications.

1) MXNet ONNX Model Loader: We observed through the Analyzer that there are layers in the model execution profile where the cuDNN API arguments deviate from what is expected. An inspection of the Analyzer’s parsed Nsight profile pointed to an image_2d_pad_constant_kernel GPU kernel function being invoked before every convolutional layer. Non-zero padding leads to the observed deviation between the expected and actual cuDNN API calls. We inspected the MXNet source code and found that padding layers are inserted during the loading of ONNX models in MXNet. ONNX supports specifying asymmetric padding as a parameter in convolution layers, whereas MXNet does not. Therefore, MXNet must insert padding layers before convolution layers where asymmetric padding is used when loading ONNX models. However, the MXNet ONNX model loader adds padding layers before every convolution layer (regardless of the use of asymmetric padding). A non-intrusive optimization is to only insert padding layers if asymmetric padding is used. With this simple one-line optimization, we observed up to 1.15× end-to-end latency speedup for ResNet50-v1 (shown in Figure 11).

2) PyTorch cuDNN Wrapper: Using Benanza we observed that there were excessive calls to cudaMemcpy/MemEvent between cuDNN API calls. Using the backtrace information from the Nsight profile, we identified the PyTorch source file that introduces these synchronizations. Upon further study of the source code, we found that all cuDNN functions are invoked by a cuDNN wrapper in PyTorch. The wrapper manages a pool of cuDNN handles and is designed to enable invoking cuDNN functions from different CPU threads. cuDNN functions managed by the same handle are synchronized and executed sequentially. In the current PyTorch (v1.3), however, a single handle is used for inference, and thus forced synchronization occurs before each cuDNN function call. The synchronizations cause 100μs stalls on average between cuDNN functions, thus the latency saved through this optimization is a function of the number of layers in a model. We modified PyTorch to elide the cuDNN wrapper and only synchronize before and after performing inference, and Figure 21 shows the speedup achieved using batch size 1. MobileNet-v2 achieves a 2.3× speedup, since it has low latency and a large number of layers.
implements NCHW convolutions in terms of NHWC with an implicit transposition. As compute dominates (i.e. larger batch sizes), the relative overhead of the transposition becomes small; hence, NCHW and NHWC have similar performance for larger batch sizes. Figure 19 shows the end-to-end latency speedup when using parallel execution of data-independent layers, optimal algorithm selection, layer fusion, and Tensor Cores (NHWC). Figure 20 shows the end-to-end latency speedup for ResNet50-v1 across batch sizes and systems. Up to a 1.95× and 1.8× speedup can be achieved by TITAN_V and Tesla_V100 respectively. We can surmise, from the previous analysis, that most of the profit for TITAN_V is attributed to its use of Tensor Cores. Quadro_RTX and Tesla_T4 achieve marginal speedup over the Tensor Core results.

V. RELATED WORK

DL Benchmarking: There has been no shortage of work on developing benchmarks to characterize DL models. These DL benchmarks either take a model as a black-box and measure the user-observable latency and throughput (end-to-end benchmarks) or delve deeper into models to characterize
the layer or kernel performance (micro-benchmarks). The end-to-end benchmarks [3], [4], [6] provide a corpus of models that are deemed to be of value to characterize for industry and research. Micro-benchmarks [4], [5], [45], [46] distill DL models into their layers or kernels, and are hand-curated. Micro-benchmarking enables easy measurements of layers within popular DL models and integrates easily with profiling tools. [7] has been recently proposed to enable fair comparison of DL techniques at different levels of granularity. At the operator level, [7] takes ONNX models and generates micro-benchmarks that target the framework’s Python API to measure the latency of each operator. Benanza also takes ONNX models as input, but generates lower-level cuDNN and cuBLAS micro-benchmarks to compute the “lower-bound” latency of the model, and perform analysis. The authors are unaware of previous work which generates micro-benchmarks from model layers and couples it with an analysis workflow to inform optimizations.

Performance Advising: There is past work on using profiling to inform users of possible optimizations. These optimizations are performed at the compiler level [47], or are plugins to code editors to inform proper usage of APIs [48], [49]. Low-level profile reports and some suggestions on how to address bottlenecks are provided by profilers and IDEs such as: NVIDIA’s Nvprof [8], Intel’s VTune [10], Oracle’s Solaris Studio [50], Microsoft’s Roslyn [51], and IBM’s XL [52]. To the author’s knowledge, there has been no work on applying or specializing the optimization advising to the DL domain.

VI. CONCLUSION

This paper presents Benanza, a sustainable and extensible DL benchmarking and analysis design that automatically generates layer-wise benchmarks for DL models to compute the “lower-bound” latency and inform optimizations on GPUs. We use Benanza to evaluate a set of 30 models using different frameworks on 7 GPUs, and pinpointed the optimizations in parallel layer execution, cuDNN algorithm selection, framework inefficiency, layer fusion, and Tensor Core usage. The results show that Benanza fills a significant gap within the characterization/optimization cycle and would boost the productivity of DL model, framework, and library developers.

ACKNOWLEDGMENTS

This work is supported by IBM-ILLINOIS Center for Cognitive Computing Systems Research (C3SR) - a research collaboration as part of the IBM Cognitive Horizon Network.

REFERENCES

[1] J. Dean, D. Patterson, and C. Young, “A new golden age in computer architecture: Empowering the machine-learning revolution,” IEEE Micro, vol. 38, no. 2, pp. 21–29, Mar. 2018. [Online]. Available: https://doi.org/10.1109/mm.2018.112130030

[2] K. Hatefwool, S. Bird, D. Brooks, S. Chintala, U. Diril, D. Dzbulgakov, M. Fawzy, B. Jia, Y. Jia, A. Kalro, J. Law, K. Lee, J. Lu, P. Noordhuis, M. Smelyanskiy, L. Xiong, and X. Wang, “Applied machine learning at facebook: A datacenter infrastructure perspective,” in 2018 IEEE International Symposium on High Performance Computer Architecture (HPCA), IEEE. IEEE, Feb. 2018, pp. 620–629. [Online]. Available: https://doi.org/10.1109/hpca.2018.00059

[3] “MLPerf,” github.com/mlperf/mlperf, 2019, accessed: 2019-10-04.

[4] “AI-Matrix,” github.com/alibaba/ai-matrix, 2019, accessed: 2019-10-04.

[5] Baidu, “DeepBench,” github.com/baidu-research/DeepBench, 2019.

[6] C. Coleman, M. Zaharia, D. Kang, D. Navayanan, L. Nardi, T. Zhao, J. Zhang, P. Bailis, K. Olukotun, and C. R. “Analysis of DAWNBenchmark, a time-to-accuracy machine learning performance benchmark,” SIGOPS Oper. Syst. Rev., vol. 53, no. 1, pp. 14–25, Jul. 2019. [Online]. Available: https://doi.org/10.1145/3352020.3352024

[7] T. Ben-Nun, M. Besta, S. Huber, A. N. Zlogas, D. Peter, and T. Hoefler, “A modular benchmarking infrastructure for high-performance and reproducible deep learning,” in 2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS). IEEE, May 2019, the 33rd IEEE International Parallel & Distributed Processing Symposium (IPDPS’19). [Online]. Available: https://doi.org/10.1109/ipdps.2019.00018

[8] “NVIDIA nvpprof,” docs.nvidia.com/cuda/profiler-users-guide/index.html, accessed: 2019-5-04.

[9] “NVIDIA Nsight System,” developer.nvidia.com/nsight-systems, accessed: 2019-5-04.

[10] “Intel VTune,” software.intel.com/en-us/vtune, accessed: 2019-5-04.

[11] “NVIDIA cuDNN,” developer.nvidia.com/cudnn, 2019, accessed: 2019-10-04.

[12] “NVIDIA cuBLAS,” developer.nvidia.com/cublas, accessed: 2019-10-04. [Online]. Available: developer.nvidia.com/cublas

[13] “ONNX: Open Neural Network Exchange,” onnx.ai, 2019, accessed: 2019-10-04.

[14] “ONNX Model Zoo,” github.com/onnx/onnx/models, 2019, accessed: 2019-10-04.

[15] “ONNX Model Zoo,” github.com/onnx/models, 2019, accessed: 2019-10-04.

[16] J. Deng, J. Guo, and S. Zafeiriou, “ArcFace: Additive angular margin loss for deep face recognition,” CoRR, vol. abs/1801.07698, 2018. [Online]. Available: arxiv.org/abs/1801.07698

[17] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” in Advances in Neural Information Processing Systems 25. Curran Associates, Inc., 2012.

[18] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “Going deeper with convolutions,” in 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, Jun. 2015, pp. 1–9. [Online]. Available: https://doi.org/10.1109/cvpr.2015.7298594

[19] R. B. Girshick, J. Donahue, T. Darrell, and J. Malik, “Rich feature hierarchies for accurate object detection and semantic segmentation,” CoRR, vol. abs/1311.2524, 2013. [Online]. Available: arxiv.org/abs/1311.2524

[20] G. Huang, Z. Liu, and K. Q. Weinberger, “Densely connected convolutional networks,” CoRR, vol. abs/1608.06993, 2016. [Online]. Available: arxiv.org/abs/1608.06993

[21] P. Wang, P. Chen, Y. Yuan, D. Liu, Z. Huang, X. Hou, and G. W. Cottrell, “Understanding convolution for semantic segmentation,” CoRR, vol./abs/1702.08502, 2017. [Online]. Available: arxiv.org/abs/1702.08502

[22] E. Barsoum, C. Zhang, C. Canton-Ferrer, and Z. Zhang, “Training deep networks for facial expression recognition with crowd-sourced label distribution,” CoRR, vol./abs/1608.01041, 2016. [Online]. Available: arxiv.org/abs/1608.01041

[23] S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” CoRR, vol./abs/1502.03167, 2015. [Online]. Available: arxiv.org/abs/1502.03167

[24] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, “Rethinking the inception architecture for computer vision,” CoRR, vol./abs/1512.00567, 2015. [Online]. Available: arxiv.org/abs/1512.00567

[25] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” Proc. IEEE, vol. 86, no. 11, pp. 2278–2324, 1998. [Online]. Available: https://doi.org/10.1109/5.726791

[26] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, “MobileNets: Efficient convolutional neural networks for mobile vision applications,” CoRR, vol./abs/1704.04861, 2017. [Online]. Available: arxiv.org/abs/1704.04861

[27] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” CoRR, vol./abs/1512.03385, 2015. [Online]. Available: arxiv.org/abs/1512.03385

[28] ——, “Identity mappings in deep residual networks,” in Computer Vision – ECCV 2016. B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham: Springer International Publishing, 2016, pp. 630–645.

[29] X. Zhang, X. Zhou, M. Lin, and J. Sun, “ShuffleNet: An extremely
