

MAPLE-Edge: A Runtime Latency Predictor for Edge Devices

Saeedjith Nair¹, Saad Abbasi¹, Alexander Wong^{1,2}, Mohammad Javad Shafiee^{1,2}

¹University of Waterloo, Waterloo, Canada

²DarwinAI, Waterloo, Canada

{smnair, srabbasi, a28wong, mjshafiee}@uwaterloo.ca

Abstract

Neural Architecture Search (NAS) has enabled automatic discovery of more efficient neural network architectures, especially for mobile and embedded vision applications. Although recent research has proposed ways of quickly estimating latency on unseen hardware devices with just a few samples, little focus has been given to the challenges of estimating latency on runtimes using optimized graphs, such as TensorRT and specifically for edge devices. As devices like NVIDIA’s Jetsons get more popular in embedded computing and robotics, we observe a pressing need to more accurately estimate inference latency of neural network architectures on diverse runtimes, including highly optimized ones. In this work, we propose MAPLE-Edge, an edge device-oriented extension of MAPLE, the state-of-the-art latency predictor for general purpose hardware, where we train a regression network on architecture-latency pairs in conjunction with a hardware-runtime descriptor to effectively estimate latency on a diverse pool of edge devices. Compared to MAPLE, MAPLE-Edge can describe the runtime and target device platform using a much smaller set of CPU performance counters that are widely available on all Linux kernels, while still achieving up to +49.6% accuracy gains against previous state-of-the-art baseline methods on optimized edge device runtimes, using just 10 measurements from an unseen target device. We also demonstrate that unlike MAPLE which performs best when trained on a pool of devices sharing a common runtime, MAPLE-Edge can effectively generalize across runtimes by applying a trick of normalizing performance counters by the operator latency, in the measured hardware-runtime descriptor. Lastly, we show that for runtimes exhibiting lower than desired accuracy, performance can be boosted by collecting additional samples from the target device, with an extra 90 samples translating to gains of nearly +40%.

1. Introduction

Despite the success of deep learning based approaches in computer vision tasks such as object detection [15, 16], clas-

sification [9, 18, 20], and segmentation [8, 17], it remains a challenge to manually design neural network architectures that are both fast, efficient and at the same time accurate. The field of Neural Architecture Search (NAS) [26] aims to address this problem by automating the architecture discovery process. However, most of the previous state-of-the-art techniques such as DARTS or ENAS [12, 14], formulate the optimization and search process with only the accuracy constraint as the main objective function in the search process.

The explosion of consumer applications for mobile and embedded devices necessitates the importance of multi-objective NAS methods that can discover models at the pareto-optimal frontier of accuracy, and constraints such as latency, memory, or power consumption. For embedded vision applications where extra focus gets paid to minimizing inference latency, NAS methods have relied on either measuring the model latency directly on-device [19] or estimating the inference latency using look-up-tables (LUTs) [2, 3, 5, 21, 23] or latency prediction modules (LPMs) [1, 7, 22, 24].

Due to the increasing size of NAS search spaces, as well as the diverse number of hardware devices and execution contexts (e.g. PyTorch, TensorFlow, TensorFlow Lite, TensorRT, OpenVino, etc.), performing measurements directly on-device is computationally intractable. Mechanisms like LUTs on the other hand are much faster at estimating latency in near constant time based on the assumption that the end-to-end latency of a network can be approximated by summing up individual latencies in a layer-wise manner. However as different studies such as Dudziak *et al.* showed [7] such LUT based approaches are not accurate and the layer-wise latencies can significantly deviate from the true latency; which means that the supposedly pareto-optimal models predicted by NAS methods may not actually lie on the pareto-optimal frontier. Instead, state-of-the-art approaches like HELP [10] use learning based approaches to predict the end-to-end model latency.

While these methods perform more accurately than LUT based estimators, the study presented in nn-Meter [25] showed that existing LPMs have difficulty estimating la-

tency when inference is executed on runtimes that optimize model graphs; we also further investigate this issue here in this manuscript. For example, frameworks like TensorRT (TRT) perform kernel auto-tuning and layer and tensor fusion to accelerate inference [4]. In other words, some layers and operations may get fused or executed in parallel, compared to the sequential execution order specified by the model designer. nn-Meter tackles this by creating a more fine-grained LUT approach which characterizes models at the kernel level, but their approach involves exhaustive testing and characterization of all sampled kernels which can take anywhere from 1 to 4.4 days based on the execution runtime.

Here we propose MAPLE-Edge, an edge device-oriented extension of the state-of-the-art latency predictor so-called MAPLE [1] for general purpose hardware, designed specifically for embedded devices. We show that by characterizing the hardware device and runtime at the operator level in just minutes (not days), the proposed MAPLE-Edge algorithm can accurately estimate the end-to-end latency of neural network architectures executed on optimized, previously unseen runtimes.

MAPLE-Edge is an LPM approach based on a hardware-aware regression model that can estimate the inference latency of a deep neural network architecture on an unseen embedded target device accurately. The proposed method specifically formulates the function $f(\alpha, \mathcal{S}; \theta) \rightarrow \hat{y}$ where α is the DNN architecture encoding, \mathcal{S} is a quantitative hardware descriptor, \hat{y} is the inference latency, and f is the regression model mapping architecture encoding and hardware descriptor to the inference latency [1].

We show that with some simple, but important modifications to the data processing and hardware characterization stages proposed in MAPLE [1], the proposed MAPLE-Edge becomes the most effective way of estimating latency on optimized edge runtimes, paving the way forward to more accurate NAS methods for embedded vision applications. As such our key contributions are as follows:

- A new latency estimator for optimized edge runtimes using an LPM algorithm is proposed. To the best of our knowledge, this is the first work exploring the idea of using an LPM algorithm to estimate neural network architecture latency on optimized runtimes such as TensorRT. It is very important to account for the effect of runtime in latency prediction for edge devices.
- A new runtime characterization on embedded devices using a hardware descriptor is proposed that can be easily constructed on all Linux devices. We show that by reducing the performance metric-based hardware descriptor proposed by MAPLE into a more dense form containing performance counters that are supported by Intel and ARM processors alike, embedded

device runtimes can be accurately characterized in just minutes.

- We propose two simple but very effective techniques to adapt MAPLE for diverse edge devices. We empirically show that our proposed techniques of targeted uniform sampling and performance counter normalization, have a significant impact on accuracy, resulting in average gains of up to +48.81% compared to MAPLE on certain runtimes.
- We provide comprehensive experimental results illustrating how using only 10 measurements from an unseen edge runtime, the proposed MAPLE-Edge can achieve state-of-the-art accuracy results outperforming the state-of-the-art methods including MAPLE [1] and HELP [10].

The rest of the manuscript is organized as follows. Section 2 describes the main methodology and the procedure of acquiring the training data from the edge devices. Section 3 dives into the experimental results and compares the proposed method and competing algorithms. Finally we conclude the paper in Section 4.

2. Method

In this section, we describe the framework for predicting the latency on edge devices including the search space, hardware descriptor, hardware cost collection pipeline, pre-processing, and data augmentation strategies.

2.1. MAPLE-Edge

The seminal MAPLE technique [1] was evaluated on general purpose hardware including Intel processors and server class GPUs. While the proposed algorithm outperformed other state-of-the-art techniques in estimating the latency, our experiments showed that MAPLE cannot always achieve strong results on ARM based embedded devices due to the added diversity in the device pool which needs further extension to address the issue. To this end, we propose a simple yet very effective trick to make the predictor generalize better: i) we take advantage of a targeted uniform sampling for selecting adaptation architectures, and ii) a new approach to represent hardware performance counters by the corresponding operator latency, to make the hardware descriptor adapt across diverse edge device runtimes.

2.1.1 Targeted Uniform Adaptation Sampling

The proposed MAPLE algorithm used a random selection approach to identify the adaptation set. While this approach makes the algorithm fairly easy and practical, it might result in instances where the selected samples may not be a good representation of the larger distribution.

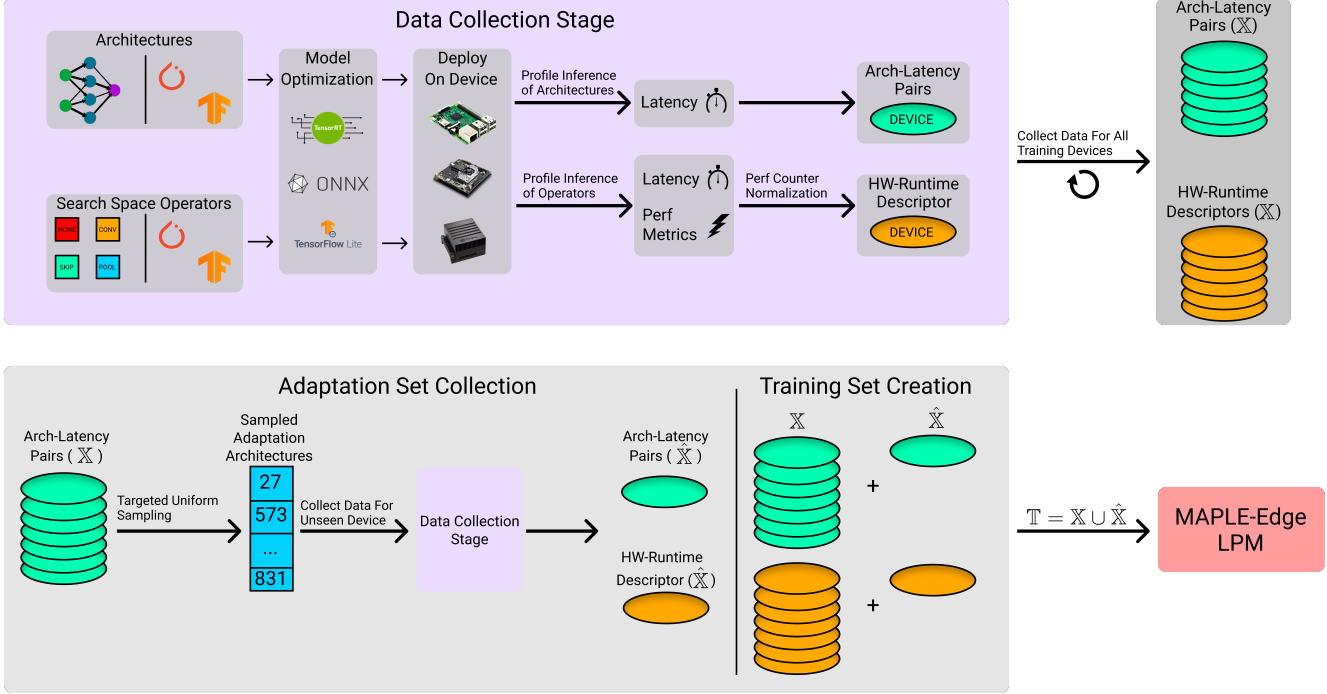


Figure 1. MAPLE-Edge overview. MAPLE-Edge is an edge device-oriented extension to MAPLE, designed specifically to estimate the latency of neural network architectures on unseen embedded devices. MAPLE-Edge trains a LPM-based hardware-aware regression model, that can effectively estimate architecture latency. To do this, it trains the LPM on a dense hardware descriptor made up of CPU performance counters, in conjunction with architecture-latency pairs. In the data collection stage, MAPLE-Edge uses an automated pipeline to convert models in NAS-Bench-201 to their optimized counterparts, deploy it to the corresponding target device, and profile inference. To adapt to previously unseen architectures, MAPLE-Edge creates a training set \mathcal{T} consisting of an initial set of architectures \mathcal{X} from a pool of known devices, as well as an adaptation set of architecture samples $\hat{\mathcal{X}}$ from the unseen device, which are selected through a targeted uniform sampling algorithm.

To minimize this risk, MAPLE-Edge leverages a targeted uniform sampling strategy where adaptation samples are selected to maximize the probability of falling across the latency space. Specifically, by creating N bins corresponding to the N adaptation samples and ranking architectures based on the measured end-to-end latency for each device in the training pool, we improve the probability of selecting a diverse set of samples, thereby representing the space more effectively. The ranked architectures for each device are then distributed uniformly across the N bins, yielding a total of $M \cdot \mathcal{X} / N$ architectures per bin, where M is the number of training devices, and \mathcal{X} is the number of architectures in the initial training set. One adaptation architecture is then randomly sampled per bin.

The intuition behind this is that although the absolute latency values are different across device runtimes, the relative orderings of architectures tend to be similar. Thus, by merging the latency distributions of all architectures in the training pool together and sampling from the combined pool, we create a more robust prior to sample the adaptation architectures from. Figure 2 demonstrates the effect of the

proposed Targeted Uniform sampling compared to random selection. As seen, the selected samples via the random selection exhibits similar latencies while the selected samples by the proposed sampling approach show a more diverse set spread across the distribution of training architectures.

2.1.2 Performance Counter Normalization

Due to the limited diversity of devices in the training pool, MAPLE was able to effectively characterize devices using the raw performance counters obtained directly from the *Perf* tool [13]. However on edge devices, there can be significant variance between the various runtimes and devices. Table 1 shows the mean latencies across the first 2700 neural network architectures in NAS-Bench-201 for an ImageNet style input of size 224×224 with batch size 1. The differences in latency between runtimes and even among devices of the same runtime imply that a hardware descriptor containing absolute performance counters measured over the entire duration of inference will be less suitable as a runtime descriptor, and as result the latency predictor will have to learn an additional mapping between an

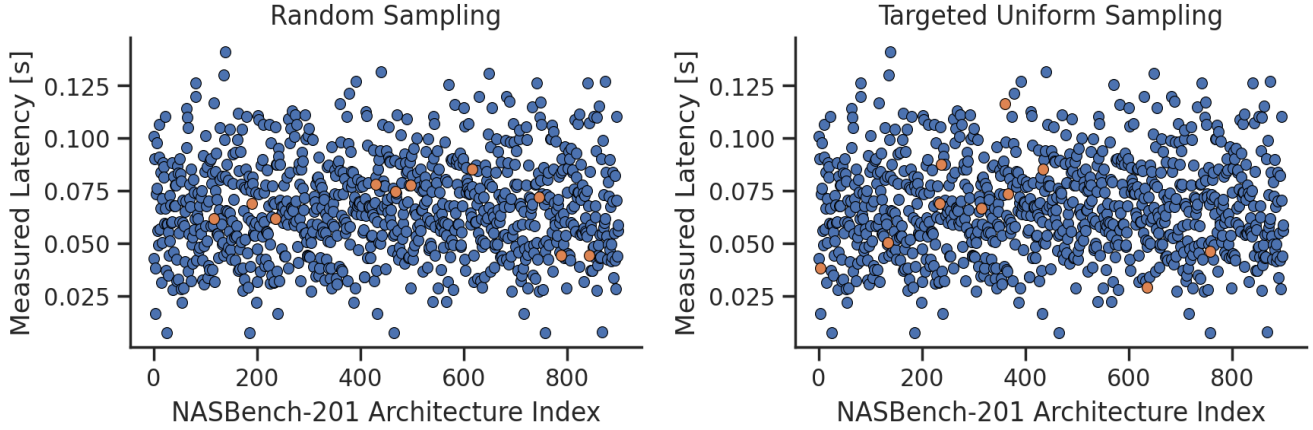


Figure 2. Comparison of the random adaptation sampling technique used by MAPLE (left) versus the targeted uniform sampling strategy proposed by MAPLE-Edge (right). The orange dots highlight ten adaptation architectures that were sampled using both methods from the first 900 architectures in NAS-Bench-201. The targeted uniform sampler intentionally selects architectures from across the latency distribution whereas the random sampler has no prior, and may result in the selected architectures being drawn from a narrow band of the latency space. By maximizing coverage of the latency space, we ensure that the LPM is more likely to be trained on samples that are representative of the test device.

Table 1. Average measured latency across architectures [0,2699] in NAS-Bench-201 (averaged over 50 trials each), using TensorFlow Lite and TensorRT runtimes. The large variance in latency between the different runtimes makes it difficult for the MAPLE hardware descriptor to effectively represent diverse device runtimes. We mitigate this issue by applying the performance counter normalization trick described in Section 2.1.2

Device	Runtime	
	TensorFlow Lite [s]	TensorRT [s]
Jetson TX1	1.0 ± 0.5	0.05 ± 0.01
Jetson TX2	0.9 ± 0.5	0.03 ± 0.01
Jetson Nano	1.1 ± 0.6	0.07 ± 0.02
Raspberry Pi 4B	0.6 ± 0.2	-

operator metric and its corresponding latency. Instead, to provide more effective representation and accelerate learning in the predictor, MAPLE-Edge normalizes all performance counters collected by *Perf* by the measured latency of the respective operator. By dividing each performance counter by the latency, this effectively transforms the hardware descriptor from an embedding of absolute values (e.g. cache-misses) to an embedding of relative rates (e.g. cache-misses per second).

2.2. Data Collection Pipeline

Embedded vision applications routinely operate under tight deadlines with low latency constraints. To maximize

performance in production, expert domain knowledge can be used to optimize deep learning models to more efficient formats (e.g. through the use of runtime specific compilers), as well as hardware modifications (e.g. overclocking, memory swapping, power usage maximization). Similar to HW-NAS-Bench [11], MAPLE-Edge leverages this domain knowledge to build an optimized, generic hardware-cost collection pipeline that automates the process of collecting latency measurements (as seen in Figure 1).

On the NVIDIA Jetson devices, MAPLE-Edge executes inference on the GPU using models compiled to serialized TensorRT (TRT) engine format, while CPU based inference is executed on all devices using the TensorFlow Lite (TFLite) runtime engine. Unlike HW-NAS-Bench however, MAPLE-Edge also records hardware performance counters for all 15 operations in NAS-Bench-201 [6], by executing the Linux *Perf* [13] tool while running inference.

2.3. Edge Dataset

Similar to MAPLE [1], BRP-NAS [7], HELP [10], and HW-NAS-Bench [11], MAPLE-Edge also uses the NAS-BENCH-201 [6] dataset for all experiments. NAS-BENCH-201 is a collection of 15,625 neural cell candidates with each architecture having a fixed cell topology with five possible operations in its search space including {none, skip-connection, conv1x1, conv3x3, avgpool3x3}. Each operation allows 16, 32, or 64 input and output channels, yielding a total of 15 possible variations for the operations that can be used in designing new architectures.

The key observation in MAPLE was that the latency

search space can be effectively characterized at the operator level, if the operator latency is augmented with a hardware description vector. Unlike methods like nn-Meter, MAPLE quickly builds a hardware descriptor for a device or runtime by executing the Linux tool *Perf* to measure 10 hardware performance counters while a model gets executed.

MAPLE-Edge follows a similar strategy, but builds a hardware descriptor using only 6 out of the original 10 performance counters that were found to be supported by default on all Linux kernels. These include performance counters for CPU-cycles, instructions, cache-references, cache-misses, level one (L1) data cache loads, and L1 data cache load misses. Even though we do not collect any metrics relating to LLC (last level cache) counters due to limited support on some ARM based processors, we show that the results are still competitive and achieve strong gains over all baselines. Each operation is characterized by running inference for $N = 1000$ runs, and the *Perf* tool is launched from a Python Subprocess to profile the execution of the inference occurring in the main thread.

To create our dataset, we measure the end-to-end latency of the first 2,700 architectures in NAS-Bench-201 on a variety of different edge devices and runtimes. These devices were selected due to their ubiquity in industrial embedded vision applications and include a i) 4GB 4xA57 core Jetson Nano with 128 NVIDIA Maxwell CUDA cores, ii) 4GB 4xA57 core Jetson TX1 with 256 NVIDIA Maxwell CUDA cores, iii) 8GB 6xA57 core Jetson TX2 with 256 NVIDIA Pascal CUDA cores, and iv) a 4GB 4xA72 core Raspberry Pi 4B running 64-bit Raspbian OS. Of these 2,700 architectures, we sample from the first 900 architectures for training our predictor, and report the test accuracy of all methods on architectures in range [1,800, 2,699]. This test strategy lets us evaluate how well the method performs on completely unseen architectures.

To train the regression model, we collect an initial set of architectures \mathbb{X} from all devices in the training pool, along with an adaptation set of few (i.e., typically 10) architectures $\hat{\mathbb{X}}$ from devices in the test pool. The initial set \mathbb{X} helps the model generalize to the search space while the adaptation samples $\hat{\mathbb{X}}$ help the predictor specialize to the test devices.

2.4. Data Augmentation

MAPLE-Edge also applies a simple yet effective data augmentation strategy where all selected adaptation samples from the test device are naively cloned $K - 1$ times to increase their representation in the training pool. Given the limited number of samples we collect, we empirically find that this approach helps the model prioritize the adaptation samples. However, we find that this data augmentation technique may result in worse performance if the adaptation samples are poorly chosen to begin with (such as with

a random selection strategy) instead of a more representative selection strategy such as targeted uniform sampling, which is more robust to outliers. Unless otherwise noted, all experiments in this paper use an augmentation factor of $K = 7$, meaning that for each adaptation sample, 6 additional clones are created.

We follow the same training strategy proposed in MAPLE [1], including the feedforward regression model, model training regime, loss function, and evaluation criteria. As such, training MAPLE-Edge remains fast, and takes less than 2 mins to train from scratch until completion (200 epochs, batch size 128) on a single NVIDIA GTX-1080Ti machine.

3. Results & Discussion

In this section we evaluate the performance of the proposed MAPLE-Edge approach and compared it with the state-of-the-art method in predicting the latency of neural network architectures on different edge devices. We also study the effect of each proposed techniques individually to illustrate the importance of the each step in improving the model accuracy.

3.1. Experimental Setup

To evaluate the performance of the proposed MAPLE-Edge, we compare it against look-up table (LUT) based method as the baseline and the state-of-the-art algorithms including HELP [10] and MAPLE [1]. For all experiments, HELP, MAPLE, and MAPLE-Edge were trained using architectures [0,899] from NAS-BENCH-201 across 10 trials, each with a unique adaptation set generated using our uniform targeted sampling technique. The 10 adaptation samples were kept consistent across all methods to ensure that the results are due to the technique’s ability to generalize and not because the adaptation samples are drawn from a more representative distribution. A more detailed study comparing the original MAPLE algorithm with each of our proposed improvements can be found in Section 3.3 and Table 4. Similar to BRP-NAS and MAPLE, we evaluate all methods using a $\pm 10\%$ error-bound accuracy metric, which describes the percentage of models with predicted latency within the corresponding error bound relative to the measured latency [7]. All methods were also evaluated on architectures [1800,2699] for consistency.

3.2. Results

Table 2 shows the comprehensive comparison analysis of the proposed method and competing algorithms. We can see how the LUT baseline performs well on the TensorFlow Lite runtimes but fails to generalize to any of the TensorRT runtimes. This is because the LUT approach estimates the total end-to-end latency by adding together the latency of each individual block and operator, without accounting for

Table 2. Summary of results comparing MAPLE-Edge to a LUT based estimator as well as HELP, and MAPLE baselines. The method column describes the method and training paradigm used. For example, *MAPLE-Edge TFLite* implies that each test device in that row was trained on MAPLE-Edge using a device pool containing only TensorFlow Lite runtimes. All methods were trained for ten trials with unique set adaptation samples (for each trial) used across all methods for consistency (selected using our targeted uniform sampling technique).

Method	Test Runtime - Mean $\pm 10\%$ Accuracy						
	Raspberry Pi TFLite	Jetson TX1 TFLite	Jetson TX2 TFLite	Jetson Nano TFLite	Jetson TX1 TRT	Jetson TX2 TRT	Jetson Nano TRT
LUT	55.18	80.96	80.22	79.96	12.07	0.00	7.85
HELP TFLite	42.77 \pm 5.07	54.90 \pm 9.62	58.82 \pm 12.00	58.67 \pm 11.99	-	-	-
MAPLE TFLite	45.04 \pm 7.69	99.22\pm0.31	96.81 \pm 1.75	98.53\pm0.83	-	-	-
MAPLE-Edge TFLite	54.00\pm11.43	98.89 \pm 0.35	97.4\pm1.59	96.18 \pm 1.85	-	-	-
HELP TRT	-	-	-	-	79.88 \pm 6.34	82.64 \pm 5.25	79.54 \pm 5.36
MAPLE TRT	-	-	-	-	96.52 \pm 1.79	96.64\pm2.05	94.37\pm2.76
MAPLE-Edge TRT	-	-	-	-	96.68\pm1.01	94.89 \pm 2.37	93.67 \pm 3.07
HELP TRT + TFLite	47.62 \pm 5.60	52.28 \pm 9.55	53.43 \pm 10.45	57.43 \pm 10.43	67.93 \pm 4.80	63.56 \pm 7.99	68.50 \pm 8.36
MAPLE TRT + TFLite	51.37 \pm 3.28	98.94 \pm 0.67	96.89 \pm 1.93	98.40\pm0.44	45.61 \pm 23.69	27.59 \pm 20.04	80.01 \pm 8.13
MAPLE-Edge TRT + TFLite	54.23\pm4.21	99.34\pm0.24	99.19\pm0.42	98.19 \pm 0.94	91.71\pm2.00	76.64\pm6.66	82.08\pm7.05

any runtime optimizations made by the execution engine. For example, NVIDIA’s TensorRT automatically eliminates layers with unused outputs and fuses together distinct layers to improve efficiency of running networks on the GPU [4]. Because of such optimizations, the resulting graph may not bear a lot of semblance to the original stack of layers, thus rendering the LUT based approach incapable of accurately estimating architecture latency. This can be easily seen in Figure 3 which shows how skewed the latency distribution looks for optimized edge runtimes, due to the LUT predictor overestimating architecture latency. On the other hand for un-optimized runtimes such as PyTorch under CPU based execution on an Intel processor, we can see how LUT based estimators achieve nearly perfect accuracy. This is why LUT based approaches have continued to be successful in a variety of non-embedded contexts, but this further underscores the need for more robust latency estimation methods for optimized edge runtimes.

The performance of all other methods were evaluated using two different types of training pools; runtime based pooling and combined pooling. In runtime based pooling, the initial set of samples are drawn from devices sharing the same runtime as the test device, excluding the test device itself. For the combined pooling setup, the initial set of samples are drawn from all available devices, excluding the test device. In almost all cases, we can see in Table 2 that MAPLE-Edge shows strong gains over the HELP baseline, especially in the case of runtime-based pooling where fewer devices are available. This is particularly significant in an embedded vision context as most applications would only

have the infrastructure to support a single runtime instead of all available runtimes. By being able to effectively generalize to an unseen device on the same runtime, MAPLE-Edge can propel significant advancements in the field of neural architecture search for embedded vision contexts.

The effectiveness of MAPLE-Edge is most evident in Table 3 where we attempt to predict the latency on a test device, using only a single device in the training pool. The reported results are obtained using leave-one-out cross-validation, where we rotate a device into the training pool to ensure that the results are representative of all available devices for that particular runtime. Here we can see that despite only being trained on 900 samples from a single device, MAPLE-Edge can effectively estimate the latency of architectures being executed on an unseen device. The strong gains over both HELP and MAPLE shows that the proposed LPM algorithm is very capable of estimating latency on optimized graphs, something that current state-of-the-art estimators have trouble with.

3.3. Improvements Over MAPLE

Table 4 provide a detailed glimpse into the impact of our proposed improvements on the MAPLE algorithm. The results labelled with a \star represent the performance of original MAPLE (i.e. with no performance counter normalization, no data augmentation, and using a random adaptation sample selection strategy). Although MAPLE shows strong performance across the board (90+% top $\pm 10\%$ accuracy), it appears to get particularly challenged when presented with a diverse group of devices such as the device pool with com-

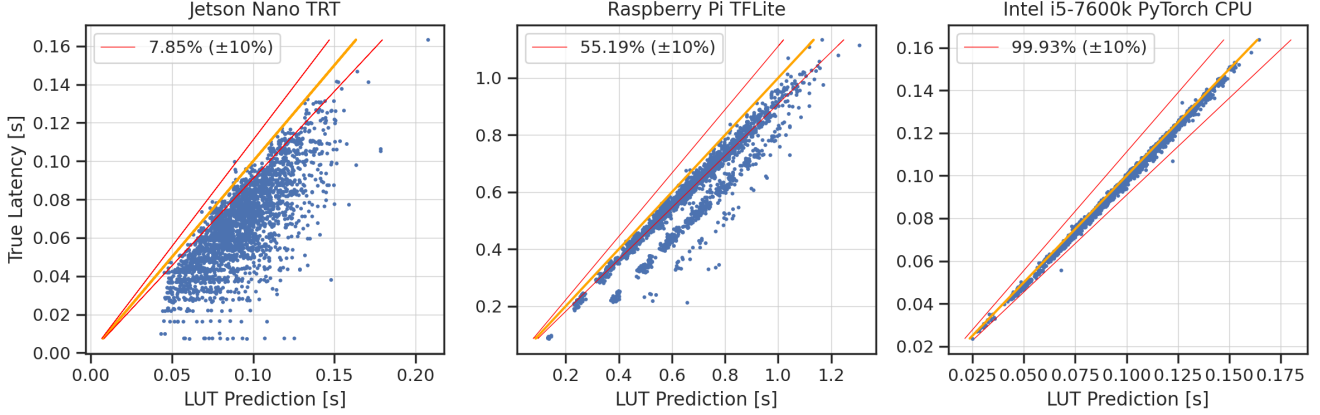


Figure 3. Look-up-table (LUT) latency predictions versus true latency for edge runtimes Jetson Nano TRT, and Raspberry Pi TFLite compared to non-edge runtime (PyTorch CPU inference on Intel i5-7600k). The red lines represent the $\pm 10\%$ error bound and the yellow line represents a perfect prediction. Note how LUT based estimation tends to overestimate the latency for inference executed on optimized graph, as it cannot take into account the runtime optimizations performed by the execution engine.

Table 3. Results from predicting the latency on each test device, using only one device in the training pool. Reported values are mean $\pm 10\%$ accuracy, obtained using leave-one-out cross validation across multiple training devices.

Method	Test Runtime - Mean $\pm 10\%$ Accuracy		
	Jetson TX1 TRT	Jetson TX2 TRT	Jetson Nano TRT
HELP TRT	78.95 \pm 5.02	76.35 \pm 7.13	76.44 \pm 6.54
MAPLE TRT	76.63 \pm 19.05	81.93 \pm 12.53	59.52 \pm 26.39
MAPLE-Edge TRT	95.23\pm3.27	96.59\pm2.09	91.04\pm4.567

bined TRT and TFLite runtimes. In such a scenario, the regression model appears unable to effectively generalize the hardware descriptor it has learned, to the adaptation samples presented from the test device. This is particularly visible in the Jetson TX1-TRT and Jetson TX2-TRT columns for MAPLE (TRT+TFLite) with baseline accuracies near 47% and 27% respectively. By applying the targeted uniform adaptation sampling strategy, the accuracy is seen to increase +1.99% on average. Stacking this with the latency normalization provides an average gain of +7.24% as well as a significant jump of +28.93% and +48.81% for the lowest performing TX1-TRT and TX2-TRT runtimes. Thus by only applying these small tricks, we show that the core ideas behind MAPLE can be successfully extended to diverse hardware pools as well, and that MAPLE-Edge can be used in Neural Architecture Search as an effective LPM tailored towards edge devices.

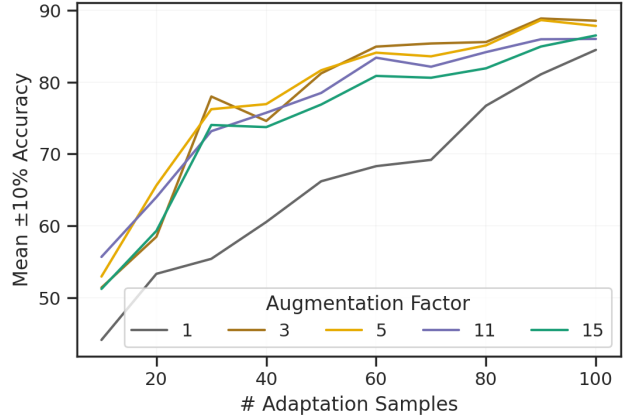


Figure 4. Impact of adaptation samples and augmentation factor on mean $\pm 10\%$ accuracy for Raspberry Pi TensorFlow Lite runtime. Each curve shows the result for using different number of augmentation sample K . Note, how increasing the number of adaptation samples from 10 to 100 provides a nearly 40% boost in prediction accuracy when no data augmentation is applied. On the other hand, we see that data augmentation proves most beneficial when adaptation samples are limited, and that beyond a certain point, additional data augmentation may in fact even harm accuracy. The training pool for this experiment consisted of all edge device runtimes excluding Raspberry Pi TFLite.

3.4. Effect of Adaptation Samples

Despite showing strong gains against baselines for all runtimes on the Jetson family of devices, MAPLE-Edge does not make any gains against the LUT approach for the Raspberry-Pi TensorFlow Lite runtime. To investigate which hyperparameters affect our method the most, we conduct an ablation study and experiment with varying the val-

Table 4. Detailed results showing how the additions we propose to MAPLE (targeted uniform adaptation sampling, performance counter normalization, and data augmentation) on average provide additive gains over the baseline method. The method column is broken down into 3 main sections: (top) These are the results from experiments where the device pool only consists of TensorFlow Lite runtimes. (middle) These are the results from experiments where the device pool only consists of TensorRT runtimes. (bottom) These are the results from experiments where the device pool consists of both TensorRT and TensorFlow Lite runtimes. The top row in each section is labelled with a \star signifying that these are the absolute accuracy values for the original MAPLE algorithm when evaluated on the corresponding test device runtime. A \dagger , \oplus , and \cup labels correspond respectively to targeted uniform adaptation sampling, latency based performance counter normalization, and data augmentation. These techniques are shown to augment the accuracy of MAPLE, with each technique providing additional additive gains over the baseline on average.

Method	Test Runtime - Mean $\pm 10\%$ Accuracy						
	Raspberry Pi TFLite	Jetson TX1 TFLite	Jetson TX2 TFLite	Jetson Nano TFLite	Jetson TX1 TRT	Jetson TX2 TRT	Jetson Nano TRT
MAPLE (TFLite) \star	41.70	98.53	94.39	95.33	-	-	-
MAPLE (TFLite) $\star\dagger$	+3.34	+0.69	+2.42	+3.20	-	-	-
MAPLE (TFLite) $\star\dagger\oplus$	+3.09	+0.83	+3.73	+2.18	-	-	-
MAPLE-Edge (TFLite) $\star\dagger\oplus\cup$	+12.30	+0.36	+3.01	+0.85	-	-	-
MAPLE (TRT) \star	-	-	-	-	94.10	94.66	91.91
MAPLE (TRT) $\star\dagger$	-	-	-	-	+2.42	+1.98	+2.46
MAPLE (TRT) $\star\dagger\oplus$	-	-	-	-	+3.32	-1.74	-0.81
MAPLE-Edge (TRT) $\star\dagger\oplus\cup$	-	-	-	-	+2.58	+0.23	+1.76
MAPLE (TRT+TFLite) \star	44.86	98.96	93.99	96.59	46.89	27.09	79.06
MAPLE (TRT+TFLite) $\star\dagger$	+6.51	-0.02	+2.90	+1.81	-1.28	+0.50	+0.95
MAPLE (TRT+TFLite) $\star\dagger\oplus$	+4.48	+0.42	+5.08	+1.18	+28.93	+48.81	+1.83
MAPLE-Edge (TRT+TFLite) $\star\dagger\oplus\cup$	+9.37	+0.38	+5.20	+1.60	+44.82	+49.55	+3.02

ues of the number of adaptation samples (previously fixed at 10) as well as the number of clones created per adaptation sample. The results of our experiments can be seen in Figure 4 which shows how additional adaptation samples and augmentation factor work together to significantly improve generalization, boosting the estimation accuracy on the Raspberry Pi TensorFlow Lite runtime from around 50% accuracy to nearly 90% accuracy. Although this requires almost 10x the number of adaptation samples, we believe that the additional time spent collecting samples (10 minutes vs. 100 minutes) is a tradeoff worth making in industrial scenarios, as it can significantly improve the latency estimation accuracy on embedded devices.

4. Conclusion

In this work, we propose MAPLE-Edge, a simple yet highly effective extension to MAPLE, enabling the most accurate latency estimation on embedded devices including optimized graphs. MAPLE-Edge leverages a reduced hardware descriptor that characterizes the device using just 6 performance counters that are available by default on all Linux kernels. We show that by measuring CPU based performance counters while executing inference using op-

timized graphs on embedded devices, we can more accurately estimate latency on TensorRT and TensorFlow Lite runtimes, compared to previous state-of-the-art approaches like HELP. We also observe additional gains over MAPLE by applying our proposed targeted uniform sampling and normalization strategies. We validate the proposed methods by conducting experiments with ten random trials each, using a unique set of adaptation samples in each trial. We find that on average, with ten adaptation samples and a targeted uniform adaptation sampling technique, MAPLE-Edge yields an improvement of 26.08% over HELP and 7.65% increase over MAPLE when there is more than one device in the training pool. MAPLE-Edge exhibits particularly strong gains when estimating the latency when trained on just 900 architectures from a single device, yielding 17.04% improvements over HELP and 21.59% improvement over MAPLE on average. By requiring significantly fewer adaptation samples than other techniques, MAPLE-Edge demonstrates that it is possible to effectively estimate latency on optimized edge runtimes, paving the way for more efficient architectures to be discovered for embedded vision using Neural Architecture Search.

References

- [1] Saad Abbasi, Alexander Wong, and Mohammad Javad Shafiee. Maple: Microprocessor a priori for latency estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022. 1, 2, 4, 5
- [2] Han Cai, Chuang Gan, Tianzhe Wang, Zhekai Zhang, and Song Han. Once-for-all: Train one network and specialize it for efficient deployment. *arXiv preprint arXiv:1908.09791*, 2019. 1
- [3] Han Cai, Ligeng Zhu, and Song Han. Proxylessnas: Direct neural architecture search on target task and hardware. *arXiv preprint arXiv:1812.00332*, 2018. 1
- [4] NVIDIA Corporation. Nvidia tensorrt, 2022. 2, 6
- [5] Xiaoliang Dai, Peizhao Zhang, Bichen Wu, Hongxu Yin, Fei Sun, Yanghan Wang, Marat Dukhan, Yunqing Hu, Yiming Wu, Yangqing Jia, et al. Chamnet: Towards efficient network design through platform-aware model adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11398–11407, 2019. 1
- [6] Xuanyi Dong and Yi Yang. Nas-bench-201: Extending the scope of reproducible neural architecture search. *arXiv preprint arXiv:2001.00326*, 2020. 4
- [7] Lukasz Dudziak, Thomas Chau, Mohamed Abdelfattah, Royson Lee, Hyeji Kim, and Nicholas Lane. Brp-nas: Prediction-based nas using gcns. *Advances in Neural Information Processing Systems*, 33:10480–10490, 2020. 1, 4, 5
- [8] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2961–2969, 2017. 1
- [9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 1
- [10] Hayeon Lee, Sewoong Lee, Song Chong, and Sung Ju Hwang. Hardware-adaptive efficient latency prediction for nas via meta-learning. *Advances in Neural Information Processing Systems*, 34, 2021. 1, 2, 4, 5
- [11] Chaojian Li, Zhongzhi Yu, Yonggan Fu, Yongan Zhang, Yang Zhao, Haoran You, Qixuan Yu, Yue Wang, and Yingyan Lin. Hw-nas-bench: Hardware-aware neural architecture search benchmark. *arXiv preprint arXiv:2103.10584*, 2021. 4
- [12] Hanxiao Liu, Karen Simonyan, and Yiming Yang. Darts: Differentiable architecture search. *arXiv preprint arXiv:1806.09055*, 2018. 1
- [13] The Linux Kernel Organization. perf: Linux profiling with performance counters, 2022. 3, 4
- [14] Hieu Pham, Melody Guan, Barret Zoph, Quoc Le, and Jeff Dean. Efficient neural architecture search via parameters sharing. In *International Conference on Machine Learning*, pages 4095–4104. PMLR, 2018. 1
- [15] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018. 1
- [16] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems*, 28, 2015. 1
- [17] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 1
- [18] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 1
- [19] Mingxing Tan, Bo Chen, Ruoming Pang, Vijay Vasudevan, Mark Sandler, Andrew Howard, and Quoc V Le. Mnasnet: Platform-aware neural architecture search for mobile. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2820–2828, 2019. 1
- [20] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pages 6105–6114. PMLR, 2019. 1
- [21] Alvin Wan, Xiaoliang Dai, Peizhao Zhang, Zijian He, Yuan-dong Tian, Saining Xie, Bichen Wu, Matthew Yu, Tao Xu, Kan Chen, et al. Fbnetv2: Differentiable neural architecture search for spatial and channel dimensions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12965–12974, 2020. 1
- [22] Hanrui Wang, Zhanghao Wu, Zhijian Liu, Han Cai, Ligeng Zhu, Chuang Gan, and Song Han. Hat: Hardware-aware transformers for efficient natural language processing. *arXiv preprint arXiv:2005.14187*, 2020. 1
- [23] Bichen Wu, Xiaoliang Dai, Peizhao Zhang, Yanghan Wang, Fei Sun, Yiming Wu, Yuandong Tian, Peter Vajda, Yangqing Jia, and Kurt Keutzer. Fbnet: Hardware-aware efficient convnet design via differentiable neural architecture search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10734–10742, 2019. 1
- [24] Yuhui Xu, Lingxi Xie, Xiaopeng Zhang, Xin Chen, Bowen Shi, Qi Tian, and Hongkai Xiong. Latency-aware differentiable neural architecture search. *arXiv preprint arXiv:2001.06392*, 2020. 1
- [25] Li Lyna Zhang, Shihao Han, Jianyu Wei, Ningxin Zheng, Ting Cao, Yuqing Yang, and Yunxin Liu. Nn-meter: Towards accurate latency prediction of deep-learning model inference on diverse edge devices. In *Proceedings of the 19th Annual International Conference on Mobile Systems, Applications, and Services*, pages 81–93, 2021. 1
- [26] Barret Zoph and Quoc V Le. Neural architecture search with reinforcement learning. *arXiv preprint arXiv:1611.01578*, 2016. 1