

# MobileTL: On-Device Transfer Learning with Inverted Residual Blocks

Hung-Yueh Chiang, Natalia Frumkin, Feng Liang, Diana Marculescu

The University of Texas at Austin  
Chandra Family Department of Electrical and Computer Engineering  
{hungyueh.chiang, nfrumkin, jeffliang, dianam}@utexas.edu

## Abstract

Transfer learning on edge is challenging due to on-device limited resources. Existing work addresses this issue by training a subset of parameters or adding model patches. Developed with inference in mind, Inverted Residual Blocks (IRBs) split a convolutional layer into depthwise and pointwise convolutions, leading to more stacking layers, *e.g.*, convolution, normalization, and activation layers. Though they are efficient for inference, IRBs require that additional activation maps are stored in memory for training weights for convolution layers and scales for normalization layers. As a result, their high memory cost prohibits training IRBs on resource-limited edge devices, and making them unsuitable in the context of transfer learning. To address this issue, we present MobileTL, a memory and computationally efficient on-device transfer learning method for models built with IRBs. MobileTL trains the shifts for internal normalization layers to avoid storing activation maps for the backward pass. Also, MobileTL approximates the backward computation of the activation layer (*e.g.*, Hard-Swish and ReLU6) as a signed function which enables storing a binary mask instead of activation maps for the backward pass. MobileTL fine-tunes a few top blocks (close to output) rather than propagating the gradient through the whole network to reduce the computation cost. Our method reduces memory usage by 46% and 53% for MobileNetV2 and V3 IRBs, respectively. For MobileNetV3, we observe a 36% reduction in floating-point operations (FLOPs) when fine-tuning 5 blocks, while only incurring a 0.6% accuracy reduction on CIFAR10. Extensive experiments on multiple datasets demonstrate that our method is Pareto-optimal (best accuracy under given hardware constraints) compared to prior work in transfer learning for edge devices.

## Introduction

With the plethora of mobile devices available for consumers today, there is a demand for fast, customized, and privacy-aware deep learning algorithms. To reach satisfactory performance, current deep learning trends promote models with billions of parameters (Dosovitskiy et al. 2021; Brown et al. 2020; Radford et al. 2021) whose on-device training is infeasible. One solution is to train the model with data in the cloud and rely on efficient communication to update the model on the mobile device (Samie et al. 2016). However,

Copyright © 2023, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

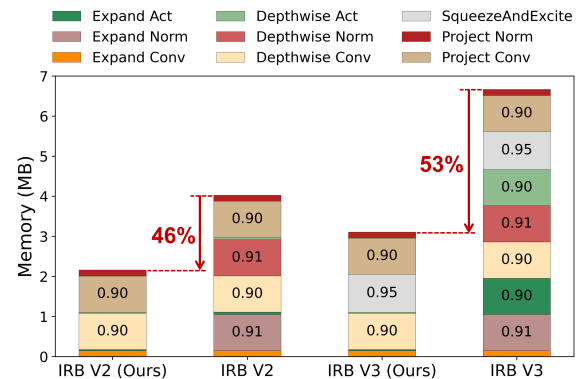


Figure 1: The figure shows stored activations of IRBs for the backward pass. We set the expansion ratio to 6. MobileTL reduces memory cost by 46.3% and 53.3% for MobileNetV2 and V3 blocks.

applications with tight data privacy constraints often make sending data from mobile devices to the cloud infeasible and suffer from degraded performance (Liu et al. 2021). In this setting, we consider *on-device transfer learning*, where a pre-trained model is downloaded to the mobile device and fine-tuned using local data. This way, the model is properly adapted to the target domain without sending potentially sensitive data to central servers.

On-device transfer learning is challenging due to limited computational resources, often prohibiting fine-tuning models. Inverted Residual Block (IRB) is one of the prevalent building blocks for models targeting mobile platforms. An IRB comprises one depthwise convolution, and two pointwise convolutions layers (Chollet 2017) with normalization and activation layers. IRBs expand the feature maps to higher dimensions using a pointwise convolution (with an expansion ratio of 3, 6, 8 *etc.*), apply a depthwise convolution in the higher dimensional space, and then project the features back into lower dimensions using a pointwise convolution. Though it is computationally efficient with fewer parameters, an IRB replaces one convolution with a stack of layers resulting in additional memory overhead for storing activation maps during training. Table 1 shows that MobileNetV2 and V3 blocks consume 0.913 and 1.362 MB

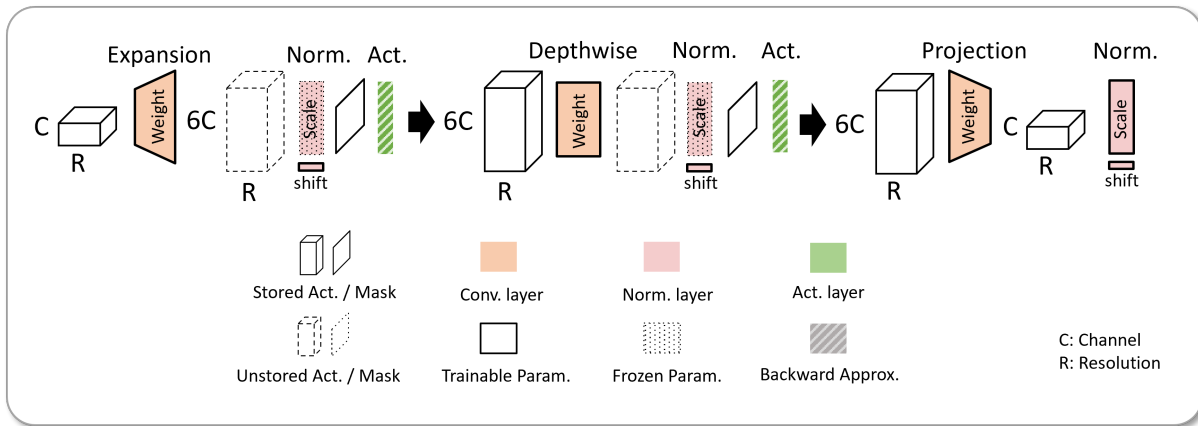


Figure 2: MobileTL is an efficient training scheme for IRBs. To avoid storing activation maps for two normalization layers after expansion and depthwise convolution, we only train shifts, and freeze scales and global statistics. The weights in convolutional layers are trained as usual. To adapt the distribution to the target dataset, we update the scale, shift, and global statistics for the last normalization layer in the block. MobileTL approximates the backward function of activation layers, *e.g.*, ReLU6 and Hard-Swish, by a signed function, so only a binary mask is stored for activation backward computing. Our method reduces the memory consumption by 46.3% and 53.3% for MobileNetV2 and V3 IRBs, as shown in Figure 1.

in training for storing activation maps, which is  $2.98\times$  and  $4.45\times$  more than for a vanilla convolution block.

To address this issue, we propose a memory-efficient back-propagation method for IRB-based models to enable fine-tuning on edge devices. As shown in Figure 2, for each intermediary normalization layer, we only update the shift but freeze the scale and global statistics (*i.e.*, mean and variance). This means we no longer have to store as many activation maps. The backward pass for memory-intensive activation layers, such as Hard-Swish (Howard et al. 2019) layers, are approximated as signed functions. The approximation allows us to store a binary mask when propagating the gradient. To reduce the memory footprint and FLOPs, we compute the gradient and update a few front-end blocks in floating point precision while the rest of the parameters are frozen and quantized during transfer learning. From our experiments, MobileTL reduces memory usage by 53% and 46% for MobileNetV3 and V2 blocks, respectively, and outperforms the baseline under the same memory constraint.

## Related Work

Low-cost, low-latency, and few-shot deep learning algorithms are key to budget-limited, customized, and data-sensitive use cases. A large body of research has been proposed to improve training, inference, and transfer learning efficiency.

**Efficient Model and Inference** Designing an efficient architecture with reduced parameters, memory footprint, and FLOPs has drawn much research attention. (Chollet 2017) decomposes an over-parameterized and computation-heavy convolution layer into separable convolution layers. (Iandola et al. 2016; Howard et al. 2017; Sandler et al. 2018; Zhang et al. 2018) handcraft efficient building blocks to build mod-

els for mobile platforms. The IRB comprised of depthwise and pointwise convolutions (Sandler et al. 2018) is now one of the most prevalent structures for mobile platforms. (Howard et al. 2019; Cai, Zhu, and Han 2019) search for model architectures for neural nets from the search space built with efficient building blocks. To reduce the memory footprint and to boost the latency in inference, pruning (Han, Mao, and Dally 2015) and quantization (Zhou et al. 2016; Courbariaux, Bengio, and David 2015; Dong et al. 2019) of model weights are prevalent methods. Our work addresses efficient *training* on devices and thus is different from the aforementioned work.

**Efficient Training** (Wu et al. 2018; Zhu et al. 2020) propose accelerating the training process with low-bitwidth training, thereby having a lower memory footprint. To save computational cost, (Jiacheng et al. 2021) skip the forward pass by caching the feature maps and only trains the last layer. Sparse training techniques (Xiaolong et al. 2021; Mostafa and Wang 2019; Dettmers and Zettlemoyer 2019) are proposed to update a subset of parameters under a constant resource constraint, but additional overhead for selecting trainable parameters is needed. (Cai et al. 2020; Mudrakarta et al. 2018) train lightweight operators and specific parameters, *e.g.*, scales and shifts, to achieve a lower memory footprint and transfer the pre-trained model to the target dataset. Our method studies the efficient training method for the existing blocks. We update all parameters in convolution layers in trainable blocks without adding patches or selecting training parameters.

**Transfer Learning** Our work is closely related to the transfer learning paradigm. (Mudrakarta et al. 2018) show that training scales and shifts in normalization layers effectively transfers the embedding to the target domain.

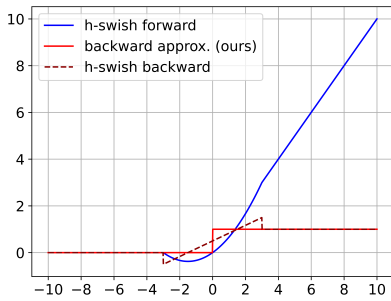


Figure 3: We approximate a Hard-Swish activation with a signed function. As a result, only a binary mask is involved in the backward computation for a Hard-Swish layer.

(Houlsby et al. 2019) propose adapter modules to transfer large transformer models to new tasks without global fine-tuning. (Cai et al. 2020) train bias to avoid storing activation maps and add memory-efficient lite-residual modules to recover the accuracy. Our work proposes an efficient transfer learning strategy for MobileNet-like models, which is orthogonal to previous work. In contrast, our method does not alter the model architecture and thus generalizes to any architecture.

## MobileTL Overview

### Efficient Transfer Learning

Fine-tuning all parameters can incur a huge computational cost. Figure 4 graphs FLOP counts on the left axis and memory on the right axis for fine-tuning MobileNetV3 Small (Howard et al. 2019). To perform global fine-tuning, the backward pass takes roughly twice the FLOPs of the forward pass (1934.92 vs. 1000.6 MFLOPs), and accumulated activation maps of each layer occupy 129.7 MB (19.9× model size) based on the chain-rule. To avoid global fine-tuning, we decompose a trained model  $f(x)$  into  $g$  and  $h$  functions such that  $f(x) = h(g(x))$ . We assume bottom blocks (close to input) learn primitive features, e.g., corners and edges, which can be shared across different tasks, while top blocks (close to output) recognize entire objects that are more task-specific (Zeiler and Fergus 2014). Based on this assumption, we freeze and quantize bottom blocks  $g(x)$  on 8-bit precision and only update top blocks  $h(x)$  for the target dataset during transfer learning. Therefore, our method does not propagate the gradient through the entire network and avoids storing activation maps for bottom blocks in  $g(x)$ .

### Efficient Transfer Learning Block

We argue that updating intermediary normalization layers in the IRB is ineffective in the transfer learning paradigm because it consumes a large amount of memory but without producing significant accuracy improvements. As a result, for transfer learning, MobileTL proposes to simplify the IRB by avoiding storing activation maps for normalization layers in between. More specifically, given a batch normalization

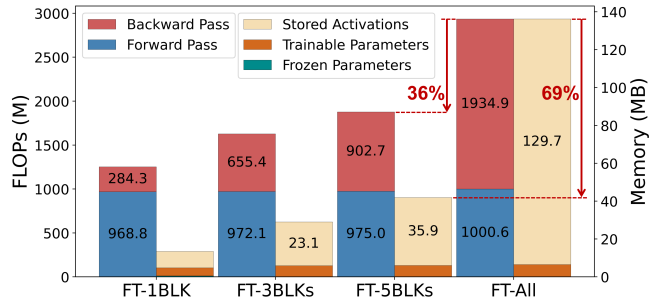


Figure 4: The FLOPs count and memory cost for training MobileNetV3 Small under different settings, e.g., fine-tune 3 blocks (FT-3BLKs), fine-tune all parameters (FT-All). In all settings, the classification layers and feature fusion layers are trained.

Block	# Param.	FLOPs (M)	Store Act. (MB)
Conv	230592	541.67	<b>0.306</b>
MBV2	<b>21408</b>	<b>51.56</b>	0.913
MBV3	26136	52.91	1.362

Table 1: We set the input and output sizes to be (8, 96, 7, 7), and the expansion ratio of IRBs to 1. The weight filter size for vanilla and depthwise convolution is  $5 \times 5$ . All convolution layers are followed by a normalization layer. IRBs reduce the number of parameters by  $10.7\times$  and FLOPs count by  $10.5\times$  while maintaining the same receptive field, but increase the training memory by  $2.9\times$  and  $4.4\times$ , respectively.

(BN) layer in evaluation mode, i.e.,  $\mu_{x_i}$  and  $\sigma_{x_i}$  are frozen.

$$\text{BN}(x_i, \gamma_i, \beta_i, \mu_{x_i}, \sigma_{x_i}) = \gamma_i \frac{x_i - \mu_i}{\sigma_{x_i}} + \beta_i = \bar{x}_i. \quad (1)$$

The gradient with respect to the scale  $\gamma_i$  and the shift  $\beta_i$  are

$$\frac{\partial L}{\partial \gamma_i} = \frac{\partial L}{\partial x_i} \frac{x_i - \mu_i}{\sigma_{x_i}}, \quad \frac{\partial L}{\partial \beta_i} = \frac{\partial L}{\partial \bar{x}_i}. \quad (2)$$

From Eq. 2, we have to store the activation  $x_i$  to compute the gradient for  $\gamma_i$ . Therefore, to avoid accumulating activation maps in memory, we freeze scales and global statistics for intermediary normalization layers when training IRBs, as Fig. 2 shows. Both normalization layers normalize inputs with pre-trained statistics. To recover the distribution difference between the pre-training dataset and target datasets, we update shifts in both intermediary normalization layers. We keep the global mean and variance in the final normalization layer updating during training, and both its scale and shift are trained for adapting to the distribution.

Though Hard-Swish (Howard et al. 2019) can boost the performance, it requires storing activation maps in memory during the backward pass, as shown in Table ???. We propose approximating its backward computation as a signed function. As a result, only a binary mask is stored in memory for later backward computation. Figure 3 depicts the

Activation	Forward	Backward	Memory
ReLU6	$\mathbf{a}_{i+1} = \min(\max(0, \mathbf{a}_i), 6)$	$\frac{\partial L}{\partial \mathbf{a}_i} = \frac{\partial L}{\partial \mathbf{a}_{i+1}} \circ \mathbf{1}_{0 \leq \mathbf{a}_i \leq 6}$	$2  \mathbf{a}_i $
Ours	$\mathbf{a}_{i+1} = \min(\max(0, \mathbf{a}_i), 6)$	$\frac{\partial L}{\partial \mathbf{a}_i} = \frac{\partial L}{\partial \mathbf{a}_{i+1}} \circ \mathbf{1}_{\mathbf{a}_i \geq 0}$	$ \mathbf{a}_i $
Hard-Swish	$\mathbf{a}_{i+1} = \mathbf{a}_i \circ \frac{\text{ReLU6}(\mathbf{a}_i+3)}{6}$	$\frac{\partial L}{\partial \mathbf{a}_i} = \frac{\partial L}{\partial \mathbf{a}_{i+1}} \circ \left( \frac{\text{ReLU6}(\mathbf{a}_i+3)}{6} + \mathbf{a}_i \circ \frac{\mathbf{1}_{-3 \leq \mathbf{a}_i \leq 3}}{6} \right)$	$32  \mathbf{a}_i $
Ours	$\mathbf{a}_{i+1} = \mathbf{a}_i \circ \frac{\text{ReLU6}(\mathbf{a}_i+3)}{6}$	$\frac{\partial L}{\partial \mathbf{a}_i} = \frac{\partial L}{\partial \mathbf{a}_{i+1}} \circ \mathbf{1}_{\mathbf{a}_i \geq 0}$	$ \mathbf{a}_i $

Table 2: We approximate the Hard-Swish and ReLU6 backward pass by a signed function. Instead of storing the full activation map, we store a compact binary mask for the backward pass, thereby reducing intermediary memory cost.  $|\cdot|$  denotes the number of elements. The memory is calculated in bits.

forward and backward mapping of our implementation for Hard-Swish.

Training  $L$  layers in a neural network with backward-approximated Hard-swish activation functions, we derive a theoretical bound with training steps ( $T$ ) and the number of approximated layers ( $L$ ). We use standard stochastic gradient descent with a learning rate  $\lambda$  and assume the magnitude of the gradient is bounded by  $G$ , and the number of elements in the output from each layer is bounded by  $N$ . We also assume that the network function  $F(\cdot)$  is Lipschitz continuous, *i.e.*,  $\forall \mathbf{x}, \mathbf{y} \in \text{dom}F, \|F(\mathbf{x}) - F(\mathbf{y})\|_2 \leq M\|\mathbf{x} - \mathbf{y}\|_2$ . We derive an error bound for the loss between the weights trained with MobileTL and the original weights. We let the weights in trainable  $L$  layers from MobileTL be  $\tilde{\mathbf{W}} = (\tilde{\mathbf{w}}_1, \tilde{\mathbf{w}}_2, \dots, \tilde{\mathbf{w}}_L)$ , and the original weights are  $\mathbf{W} = (\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_L)$ .

**Theorem 1.** *Given trainable  $L$  layers in a neural network with Hard-swish activation functions, whose backward calculation is approximated with a signed function. If we train the  $L$  layers for  $T$  steps, then the loss distance between  $\|F(\tilde{\mathbf{W}}^T) - F(\mathbf{W}^T)\|$  is bounded by  $\lambda MTG \left( \frac{\Psi(1-\Psi^L)}{1-\Psi} + \frac{\tilde{\Psi}(1-\tilde{\Psi}^L)}{1-\tilde{\Psi}} \right)$ , where  $\Psi = \frac{3}{2}\sqrt{NG}$ ,  $\tilde{\Psi} = \sqrt{NG}$ , and  $M$  is the constant from the Lipschitz continuous property of  $F(\cdot)$ .*

The proof of Theorem 1 can be found in the Supplementary Material. Theorem 1 shows that MobileTL transfers the pre-train weights to target datasets without incurring accuracy drop since on-device datasets are orders of magnitude smaller than pre-trained datasets (necessary updates  $T$  is small) and we approximate the Hard-swish layers only in trainable blocks ( $L$  is small). Figure 4 shows FT-3BLKs with MobileTL close to the knee point of the curves.

## Experimental Results

### Experiment Setup

**Model Profiling** We develop an analytical profiling tool to theoretically calculate the number of floating-point operations and memory footprint for both forward and backward passes. Floating-point operations can generalize to more operations such as normalization and activation layers and are not limited to multiplication-accumulate operations (*i.e.*, MACs). For simplicity, we approximate all operations as one FLOP, although not all floating-point operators consume

Methods	Mem. (MB)	FLOPs (M)	Train Param.
FT-All	382.7	16,205.0	2,927,612
FT-BN	189.9	11,066.1	162,596
FT-Bias	30.6	10,446.5	145,348
FT-3BLKs	40.5	7,728.0	1,695,972
FT-Last	29.0	5,325.9	128,100
TinyTL-B	31	10,446.5	145,348
TinyTL-L <sup>†</sup>	32	13,505.9	1,944,516
TinyTL-L-B <sup>†</sup>	37	14,087.6	1,959,748
<b>MobileTL-3BLKs*</b>	<b>33.7</b>	<b>7,699.0</b>	<b>1,691,364</b>

<sup>†</sup> uses model patches during fine-tuning. \* denotes Pareto-optimal

Table 3: We investigate different finetuning strategies with Proxyless Mobile on CIFAR100. All methods predicated with ‘‘FT’’ represent vanilla fine-tuning on the corresponding layer type. For example, FT-Bias represents fine-tuning only the bias term of each linear. MobileTL is Pareto-optimal among all methods across various datasets (*c.f.* Figure 5).

the same energy and number of clock cycles. For example, exponentiation and square root require many more cycles than addition. However, we still approximate them as a single FLOP as done by most analytical tools. To calculate the memory footprint, we consider all training components, including trainable and frozen parameters, accumulated activation maps, temporary memory for matrix multiplication, and residual connections. By considering all intermediary variables and operations, our profiling tool can better represent a complete training scheme, whereas many techniques only consider the model size.

**Dataset** We apply transfer learning on multiple image classification tasks. Similar to prior work (Kornblith, Shlens, and Le 2019; Houlsby et al. 2019; Cai et al. 2020), we begin with an ImageNet (Deng et al. 2009) pre-trained model, and transfer to eight downstream image classification datasets, including Cars (Krause et al. 2013), Flowers (Nilsback and Zisserman 2008), Aircraft (Maji et al. 2013), CUB-200 (Wah et al. 2011), Pets (Parkhi et al. 2012), Food (Bossard, Guillaumin, and Gool 2014), CIFAR10 (Krizhevsky, Hinton et al. 2009), and CIFAR100 (Krizhevsky, Hinton et al. 2009).

**Model Architecture** We apply MobileTL to a variety of IRB-based models, including MobileNetV2 (Sandler et al. 2018), MobileNetV3(Howard et al. 2019), and Proxyless Mobile (Cai, Zhu, and Han 2019). Though MobileTL targets models built with IRBs, we illustrate MobileTL’s flexibility in the analysis section, by extending our method to models built with conventional convolution blocks such as ResNet18 and ResNet50 (He et al. 2016).

**Training Details** For fair comparison, we follow the hyper-parameters and settings in TinyTL (Cai et al. 2020) where they train for 50 epochs with a batch size of 8 on a single GPU. We use the Adam optimizer and cosine annealing for all experiments, however, the initial learning rate is slightly tuned for each dataset and model. The classification layers are trained in all settings, and fusion layers are trained in block-wise fine-tuning. We ran our experiment using four random seeds, and average the results.

### Efficient Transfer Learning with MobileTL

Table 3 presents different fine-tuning strategies on Proxyless Mobile (Cai, Zhu, and Han 2019). TinyTL-B (Cai et al. 2020) only trains bias and avoids accumulating activation maps in memory. However, the gradient propagates to biases in the whole network, thereby requiring more FLOPs. In order to recover the accuracy, the lightweight patch in residual connection is proposed to train with the model (TinyTL-L-B). However, these patches introduce new parameters to the model. To best recover the performance for transfer learning, these patches need to be trained with the model on the large-scale pre-task dataset. We adopt a different approach by starting with fine-tuning a few IRBs, *e.g.*, 3 blocks of the model, and then applying MobileTL to trainable IRBs. MobileTL reduces memory usage by 16.7% when compared to vanilla fine-tuning of three IRBs and reduces FLOPs by 52% when compared to global fine-tuning.

Figure 5 depicts the accuracy versus memory footprint for transferring an ImageNet pre-trained Proxyless Mobile to eight downstream tasks. The radius of a circle represents the number of FLOPs, and therefore a smaller area means a smaller FLOP count. Compared with the baselines, MobileTL is Pareto-optimal under the same memory constraint for widely adopted datasets such as CIFAR10 (Krizhevsky, Hinton et al. 2009), Aircraft (Maji et al. 2013), CUB-200 (Wah et al. 2011), *etc.* For CIFAR100 (Krizhevsky, Hinton et al. 2009), MobileTL has comparable performance to TinyTL (Cai et al. 2020) but with lower FLOPs as well as lower latency on edge devices (*c.f.* Table 6). In our experiments, MobileTL outperforms the vanilla version in CIFAR10 and CIFAR100, illustrating that our method transfers the pre-trained model to the target dataset with lower memory costs. Our experiments also show that the pre-trained normalization statistics from the large pre-task dataset benefit downstream tasks. We can leverage this by only training the shift parameter and maintaining the original normalization statistics. MobileTL outperforms vanilla fine-tuning by 0.47% and 1.13% in accuracy and has lower memory costs when fine-tuning three blocks in CIFAR10 and CIFAR100 respectively.

IRB V3	Mem. (MB)	CIFAR10 (%)
Vanilla	47.4	95.2
remove-SE	43.6	94.3
ReLU	41.7	94.6
MobileTL	<b>35.8</b>	<b>95.0</b>

Table 4: We remove SE Layers (the second row) or replace H-Swish activation function with ReLU (the third row) for IRBs in MobileNetV3 Small. Although they reduce memory footprint, they lead to lower accuracy when transferring ImageNet to CIFAR10. MobileTL’s accuracy in the last row approaches the vanilla fine-tuning in the first row.

## Analysis

### Normalization Layers in IRBs

We study the effect of training different normalization layers for MobileTL and show the results in Figure 6. In this experiment, we adopt MobileNetV3 Small (Howard et al. 2019). For this ablation, the batch normalization layer following each expansion, depthwise, and pointwise convolution is trained in isolation (denoted in Figure 6 by -exp-bn, -dw-bn, and -proj-bn, respectively). For example, MobileTL-proj-bn fully updates the last normalization layer after the projection layer, while the other two normalization layers only update shifts. MobileTL-all-bn-shift only updates shifts and freezes scales and global statistics for all normalization layers in trainable IRBs. In contrast, MobileTL-all-bn fully updates scales, shifts, and global statistics for all normalization layers. In all MobileTL settings, Hard-Swish layers in trainable IRBs are approximated as a signed function in the backward pass.

In our experiments, we show that MobileTL-proj-bn (orange) is the most memory-efficient and Pareto-optimal. Training the last normalization layer involves less memory while adapting the pre-trained weights to the target dataset. MobileTL-all-bn-shift (red) has the least memory consumption but with degraded accuracy, which shows that *training shifts only in normalization layers fails to adapt the weights to the target domain*.

### Training with Efficient Operators

A natural way to reduce training memory is by substituting Hard-Swish with ReLU or removing Squeeze-and-Excitation layers whose backward update is very memory intensive (*c.f.* Fig. 1). However, pre-trained weights depend on the model architectures and are sensitive to operators in the network. In Table 4, we show that MobileTL is superior to these naive techniques. Furthermore, MobileTL does not alter the network structure and therefore avoids the performance drop when transferring pre-trained weights to the target dataset.

### Model Patches for Transfer Learning

MobileTL is orthogonal to previous work (Houlsby et al. 2019; Cai et al. 2020) that adds lightweight patches to the model and transfers the patches to the target dataset. In Table

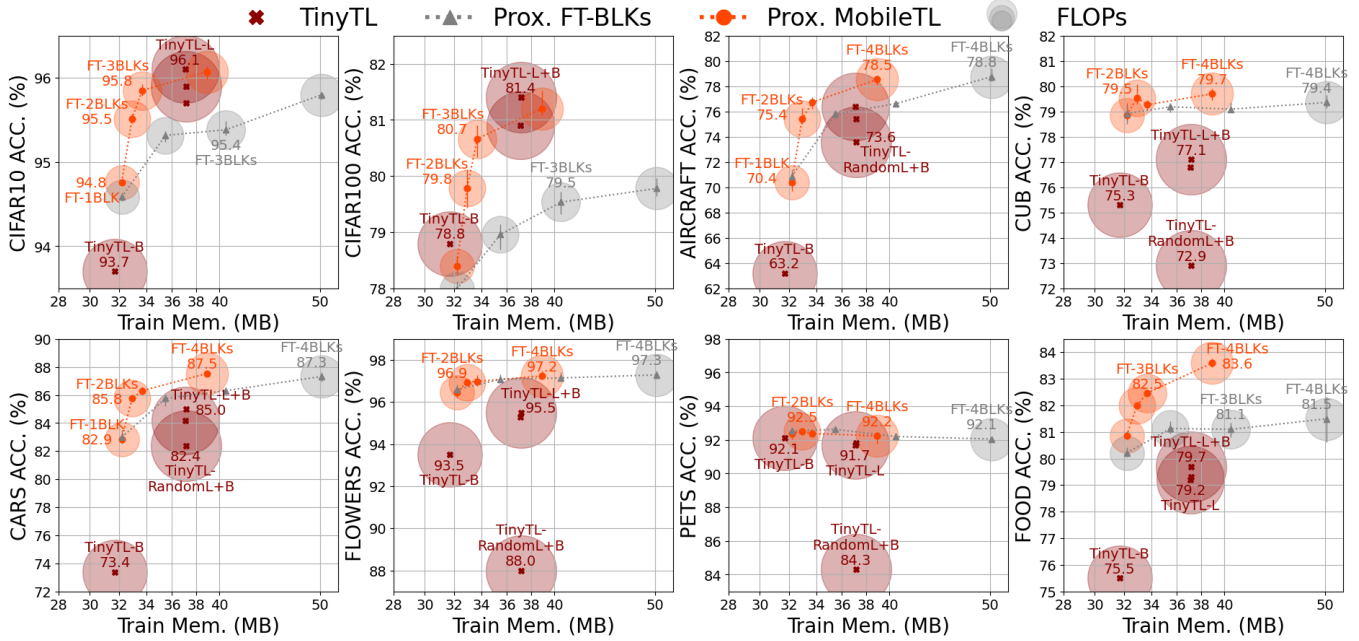


Figure 5: We experiment MobileTL with Proxyless Mobile (Cai, Zhu, and Han 2019) and transfer to eight downstream tasks. We compare our method (Prox. MobileTL in orange) with vanilla fine-tuning a few IRBs of the model (Prox. FT-BLKs in grey) and TinyTL (Cai et al. 2020). MobileTL maintains Pareto optimality for all datasets and improves accuracy over TinyTL in six out of eight datasets.

5, we transfer Proxyless Mobile from ImageNet-pre-trained weights to CIFAR10. We fine-tune the three top blocks (close to output) of the model. The first row is the vanilla block-wise fine-tuning approach. The following three rows correspond to MobileTL. We add model patches in the residual connections for the last two rows in the table. We adopt the lite-residual module as our experimental patches proposed in (Cai et al. 2020) with resolution down-sampling, group normalization layers (Wu and He 2018), and group convolutions. The patches are without pre-trained weights and are randomly initialized. The results show that patches present *additional 39.9% trainable parameters* while having marginal improvement training with main blocks. In contrast, MobileTL *reduces the memory footprint while keeping the accuracy without the need of adding additional modules or increasing parameters to the model*.

### Deployment on Edge Devices

To demonstrate practical feasibility, we deploy our method to edge devices. We experiment with Raspberry PI4 model B with Quad core ARM Cortex-A72 64-bit and 4 GB RAM, and NVIDIA JETSON NANO with 128-core GPU, Quad-core ARM Cortex-A57, and 2 GB RAM. We run our models with PyTorch framework on two devices, while on Raspberry PI is the CPU-only version. Batch size is set to 8 and input resolution is  $224 \times 224$  with output 10 classes. We measure the latency of forward and backward passes with random data. The latency is the average of 1000 training steps and is reported in seconds. The model is pre-loaded to the memory and run for a few warm-up steps before measuring.

Mobile TL	Main Blk	Res. Patch	Train Param.	Mem. (MB)	CIFAR10 (%)
	✓		1,580,682	40.1	95.4
✓	✓		<b>1,576,074</b>	<b>33.2</b>	<b>95.8</b>
✓	✓	✓	2,211,466	35.8	95.8
✓	frozen	✓	1,060,362	32.3	94.4

Table 5: We study the effectiveness of model patches. We transfer Proxyless Mobile from ImageNet to CIFAR10. The lightweight patches in the residual connection bring marginal improvement when training with main blocks. In contrast, MobileTL reduces the memory cost without incurring accuracy drop.

We report the result in Table 6, which shows that MobileTL reduces the latency by  $2.06 \times$  on Nano and  $1.82 \times$  on Raspberry PI when compared with global fine-tuning.

### Generalizing to Other Network Architectures

MobileTL does not alter the model architecture and therefore it generalizes to various models. As shown in Figure 7, we apply MobileTL to different architectures based on IRBs, such as MobileNetV2 (Sandler et al. 2018), MobileNetV3 Small and Large (Howard et al. 2019), Proxyless Mobile (Cai, Zhu, and Han 2019). Additionally, we extend MobileTL to models built with convolution blocks such as ResNet18 and ResNet50 (He et al. 2016) for comparison. Figure 7 depicts memory footprint versus accuracy of dif-

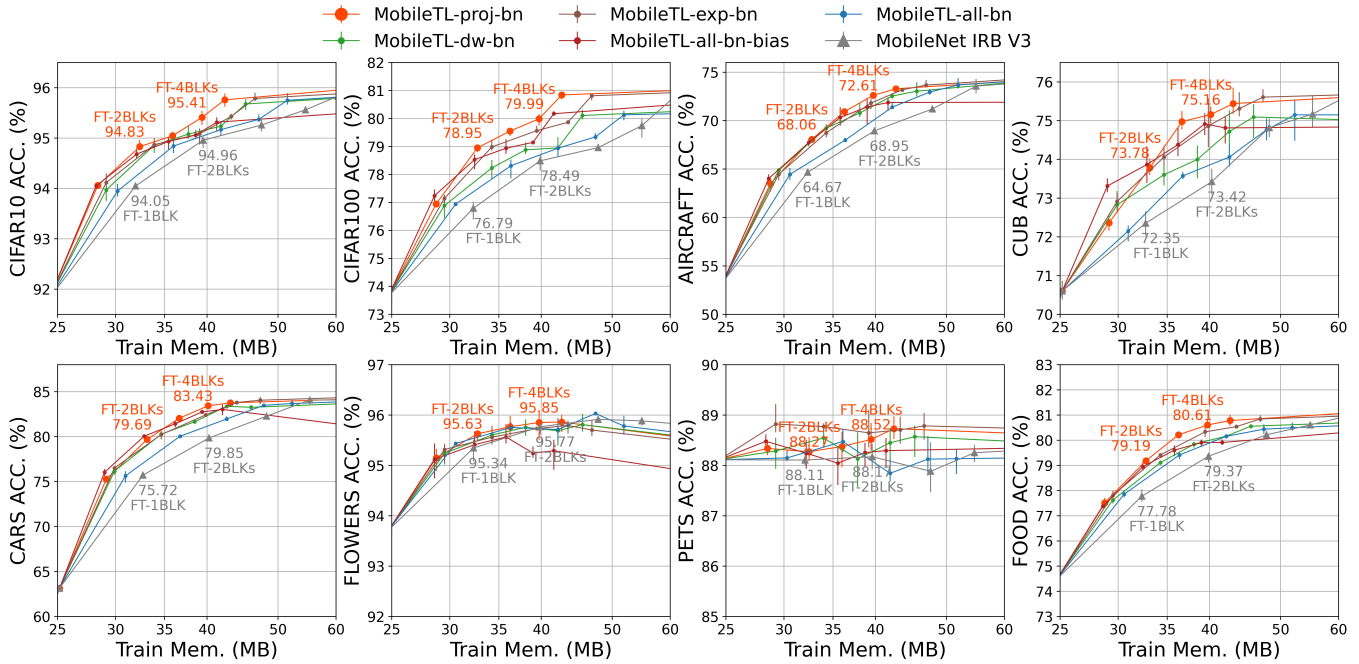


Figure 6: MobileTL-proj-bn is Pareto-optimal for MobileNetV3 Small across different settings. The last normalization layer following behind the projection layer takes less memory for training while adapting global statics to the target dataset. MobileTL-all-bn-shift (red) has the least memory consumption but with degraded accuracy. The results show that training shifts only in normalization layers fails to adapt the weights to the target domain.

Device	Method	Latency (s)
Nano	FT-All	0.235
	FT-BN	0.138
	FT-Bias	0.130
	<b>MobileTL-3BLKs</b>	<b>0.114</b>
RPI4	FT-All	2.465
	FT-BN	1.894
	FT-Bias	1.818
	<b>MobileTL-3BLKs</b>	<b>1.344</b>

Table 6: We deploy MobileTL on MobileNetV3 Small and measure the average latency for a training step (forward and backward pass) on a NVIDIA JETSON NANO, and a Raspberry PI 4. The latency is measured in seconds. Batch size is set to 8 and input size is  $224 \times 224$ . The experimental models do not have patches in all settings.

ferent models on CIFAR10 dataset. The radius corresponds to the FLOP count for fine-tuning. MobileTL pushes several models to the knee point, and Proxyless Mobile with MobileTL is Pareto-optimal.

## Conclusion

We present MobileTL, a new on-device transfer learning method that is memory- and computation-efficient. MobileTL reduces the memory footprint for separable convolution blocks by freezing the intermediary normalization layers and approximating activation layers in blocks during the

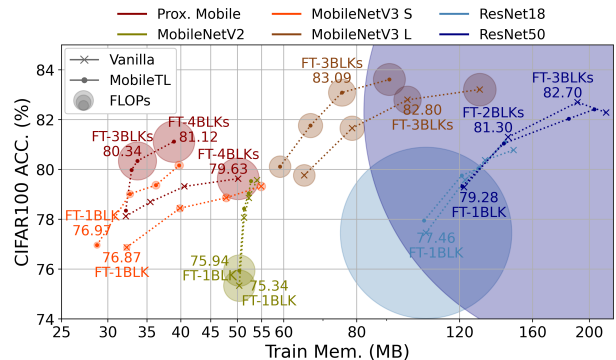


Figure 7: We generalize MobileTL to different architectures that built with IRBs, and models built with convolution blocks such as ResNet. The radius corresponds to FLOP count for fine-tuning. MobileTL pushes several models to the Pareto front.

backward pass. To reduce the FLOP counts, we only propagate the gradient through a few trainable top blocks (close to output) in the model to enable fine-tuning on small edge devices. We show the proposed method generalizes different architectures without re-training weights on the large pre-task dataset since it does not require adding patches to the model or altering the architecture. Our intensive ablation studies demonstrate the effectiveness and efficiency of MobileTL.

## Acknowledgements

This research was supported in part by NSF CCF Grant No. 2107085, NSF CSR Grant No. 1815780, and the UT Cockrell School of Engineering Doctoral Fellowship. Additionally, we thank Po-han Li for his help with the proof.

## References

- Bossard, L.; Guillaumin, M.; and Gool, L. V. 2014. Food-101—mining discriminative components with random forests. In *European Conference on Computer Vision*.
- Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*.
- Cai, H.; Gan, C.; Zhu, L.; and Han, S. 2020. Tinytl: Reduce memory, not parameters for efficient on-device learning. In *Advances in Neural Information Processing Systems*.
- Cai, H.; Zhu, L.; and Han, S. 2019. ProxylessNAS: Direct Neural Architecture Search on Target Task and Hardware. In *International Conference on Learning Representations*.
- Chollet, F. 2017. Xception: Deep learning with depthwise separable convolutions. In *Computer Vision and Pattern Recognition*.
- Courbariaux, M.; Bengio, Y.; and David, J.-P. 2015. Binaryconnect: Training deep neural networks with binary weights during propagations. In *Advances in Neural Information Processing Systems*.
- Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition*.
- Dettmers, T.; and Zettlemoyer, L. 2019. Sparse networks from scratch: Faster training without losing performance. In *arXiv preprint arXiv:1907.04840*.
- Dong, Z.; Yao, Z.; Gholami, A.; Mahoney, M. W.; and Keutzer, K. 2019. Hawq: Hessian aware quantization of neural networks with mixed-precision. In *International Conference on Computer Vision*.
- Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.
- Han, S.; Mao, H.; and Dally, W. J. 2015. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. In *International Conference on Learning Representations*.
- He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Computer Vision and Pattern Recognition*.
- Houlsby, N.; Giurgiu, A.; Jastrzebski, S.; Morrone, B.; De Laroussilhe, Q.; Gesmundo, A.; Attariyan, M.; and Gelly, S. 2019. Parameter-efficient transfer learning for NLP. In *International Conference on Machine Learning*.
- Howard, A.; Sandler, M.; Chu, G.; Chen, L.-C.; Chen, B.; Tan, M.; Wang, W.; Zhu, Y.; Pang, R.; Vasudevan, V.; et al. 2019. Searching for mobilenetv3. In *International Conference on Computer Vision*.
- Howard, A. G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; and Adam, H. 2017. Mobilenets: Efficient convolutional neural networks for mobile vision applications. In *arXiv preprint arXiv:1704.04861*.
- Iandola, F. N.; Han, S.; Moskewicz, M. W.; Ashraf, K.; Dally, W. J.; and Keutzer, K. 2016. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and 0.5 MB model size. In *arXiv preprint arXiv:1602.07360*.
- Jiacheng, Y.; James, G.; Mostafa, E.; and Gennady, P. 2021. MOIL: Enabling Efficient Incremental Training on Edge Device. In *International Conference on Learning Representations workshop*.
- Kornblith, S.; Shlens, J.; and Le, Q. V. 2019. Do better imagenet models transfer better? In *Computer Vision and Pattern Recognition*.
- Krause, J.; Stark, M.; Deng, J.; and Fei-Fei, L. 2013. 3d object representations for fine-grained categorization. In *International Conference on Computer Vision workshop*.
- Krizhevsky, A.; Hinton, G.; et al. 2009. Learning multiple layers of features from tiny images. Citeseer.
- Liu, B.; Ding, M.; Shaham, S.; Rahayu, W.; Farokhi, F.; and Lin, Z. 2021. When machine learning meets privacy: A survey and outlook. In *ACM Computing Surveys*.
- Maji, S.; Rahtu, E.; Kannala, J.; Blaschko, M.; and Vedaldi, A. 2013. Fine-grained visual classification of aircraft. In *arXiv preprint arXiv:1306.5151*.
- Mostafa, H.; and Wang, X. 2019. Parameter efficient training of deep convolutional neural networks by dynamic sparse reparameterization. In *International Conference on Machine Learning*.
- Mudrakarta, P. K.; Sandler, M.; Zhmoginov, A.; and Howard, A. 2018. K for the Price of 1: Parameter-efficient Multi-task and Transfer Learning. In *International Conference on Learning Representations*.
- Nilsback, M.-E.; and Zisserman, A. 2008. Automated flower classification over a large number of classes. In *Conference on Computer Vision, Graphics & Image Processing*.
- Parkhi, O. M.; Vedaldi, A.; Zisserman, A.; and Jawahar, C. 2012. Cats and dogs. In *Computer Vision and Pattern Recognition*.
- Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*.
- Samie, F.; Tsoutsouras, V.; Bauer, L.; Xydis, S.; Soudris, D.; and Henkel, J. 2016. Computation offloading and resource allocation for low-power IoT edge devices. In *2016 IEEE 3rd World Forum on Internet of Things (WF-IoT)*.
- Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; and Chen, L.-C. 2018. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Computer Vision and Pattern Recognition*.



- Wah, C.; Branson, S.; Welinder, P.; Perona, P.; and Belongie, S. 2011. The caltech-ucsd birds-200-2011 dataset. California Institute of Technology.
- Wu, S.; Li, G.; Chen, F.; and Shi, L. 2018. Training and inference with integers in deep neural networks. In *International Conference on Learning Representations*.
- Wu, Y.; and He, K. 2018. Group normalization. In *European Conference on Computer Vision*.
- Xiaolong, M.; Zhengang, L.; Geng, Y.; Wei, N.; Bin, R.; Yanzhi, W.; and Xue, L. 2021. Memory-Bounded Sparse Training on the Edge. In *International Conference on Learning Representations workshop*.
- Zeiler, M. D.; and Fergus, R. 2014. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision*.
- Zhang, X.; Zhou, X.; Lin, M.; and Sun, J. 2018. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *Computer Vision and Pattern Recognition*.
- Zhou, S.; Wu, Y.; Ni, Z.; Zhou, X.; Wen, H.; and Zou, Y. 2016. DoReFa-Net: Training Low Bitwidth Convolutional Neural Networks with Low Bitwidth Gradients. In *CoRR*.
- Zhu, F.; Gong, R.; Yu, F.; Liu, X.; Wang, Y.; Li, Z.; Yang, X.; and Yan, J. 2020. Towards unified int8 training for convolutional neural network. In *Computer Vision and Pattern Recognition*.