Supplementary : Dynamic Kernel Selection for Improved Generalization and Memory Efficiency in Meta-learning

Arnav Chavan*†‡, Rishabh Tiwari*†‡, Udbhav Bamba†‡ and Deepak K. Gupta†
†Transmute AI Lab (Texmin Hub), Indian Institute of Technology, ISM Dhanbad
{arnavchavan04,akchitra99,ubamba98,guptadeepak2806}@gmail.com

A. Algorithms

Algorithm 1: META LOSS

Input : Target output $\mathbf{y}$, Predicted output $\hat{\mathbf{y}_1}$, Task parameters $\boldsymbol{\phi}$, Meta parameters $\boldsymbol{\theta}$, Task masks $\boldsymbol{\zeta}$, Meta masks $\mathbf{z}$, Scalars λ_1, λ_2

Output : Loss $\mathcal{L}$

$\mathcal{L}_{ce} \leftarrow -\sum \mathbf{y} \log(\hat{\mathbf{y}})$

$\mathcal{L}_1 \leftarrow \frac{\lambda_1}{2} \|\boldsymbol{\theta} - \boldsymbol{\phi}\|^2$

$\mathcal{L}_2 \leftarrow \frac{\lambda_2}{2} \|\mathbf{z} - \boldsymbol{\zeta}\|^2$

$\mathcal{L} \leftarrow \mathcal{L}_{ce} + \mathcal{L}_1 + \mathcal{L}_2$

Algorithm 2: BUDGET LOSS

Input : Binary Masks: $\boldsymbol{\zeta}_b$, Target Budget: V_0

Output : Loss $\mathcal{L}$

$\mathcal{V} \leftarrow \frac{\text{Sum total of active kernels, } \sum \boldsymbol{\zeta}_b}{\text{Total number of kernels}}$

$\mathcal{L} \leftarrow (\mathcal{V} - \mathcal{V}_0)^2$

B. Experimental Details

We keep the hyperparameter same for pre-training and pruning experiments that were common in both. Hyperparameter values that were common across all experiments are - total iterations 60000, meta batch size of 4, $\lambda_1 = \lambda_2 = 0.5$, $\lambda_3 = 50$ and $\lambda_4 = 1e{-}04$. λ_3 increases linearly starting from 0 going upto 50 during the course of training. All other hyperparameters are reported in the Table 1. These hyperparameters were searched on the pre-training step and not on the pruning step

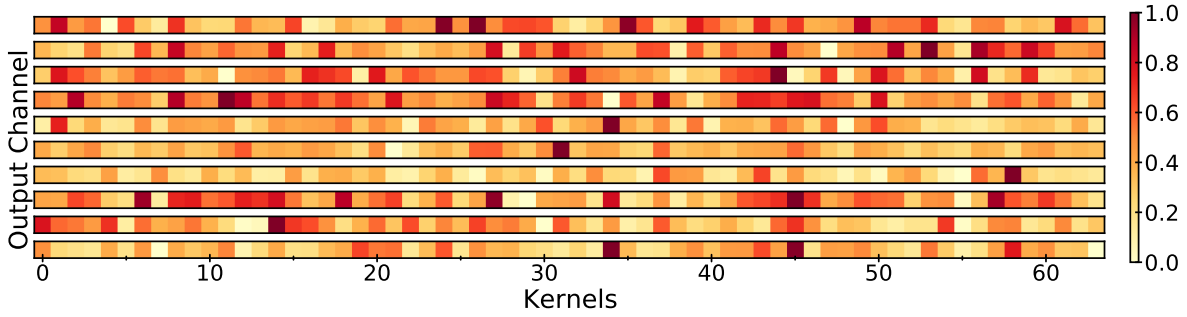
Dataset	Settings	Inner Learning Rate	Outer Learning Rate	Task Batch Size	Inner Steps
CIFAR-fs	5-way 5-shot	5.00E-03	5.00E-05	4	16
CIFAR-fs	5-way 1-shot	1.00E-02	5.00E-04	4	16
mini-ImageNet	5-way 5-shot	1.00E-02	5.00E-05	4	10
mini-ImageNet	5-way 1-shot	5.00E-03	5.00E-04	4	10

Table 1. Hyperparameters for all pre-training and pruning experiments

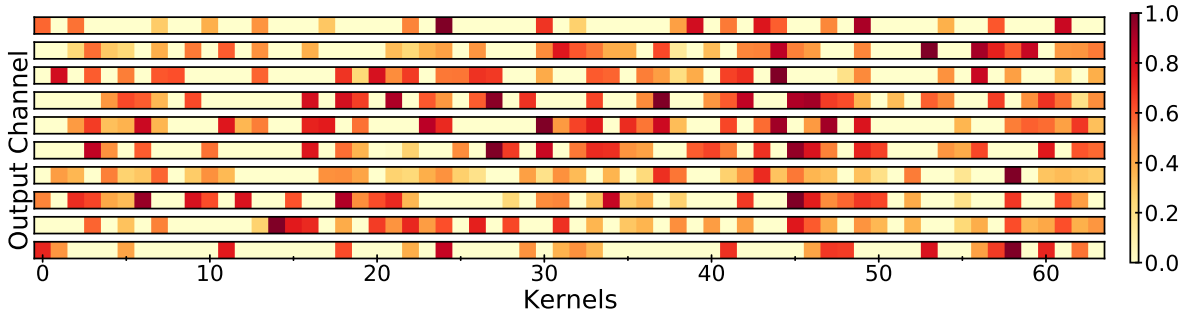
C. Kernel Visualizations

Figure 1 shows an example of all 64 input kernel usage for 10 random output channels from last layer of 4-conv model. For each output channel, we show the relative contribution of the all the kernels. From Figure 1a, it is seen that the contributions

*indicates equal contribution. Arnav Chavan is the corresponding author.



(a) Without budget constraint



(b) With budget constraint

Figure 1. Relative contribution of all 64, 3×3 kernels in mapping 10 different output channels for (a) no budget constraint, and (b) a budget constraint of 50% on the total number of kernels that are used per mapping.

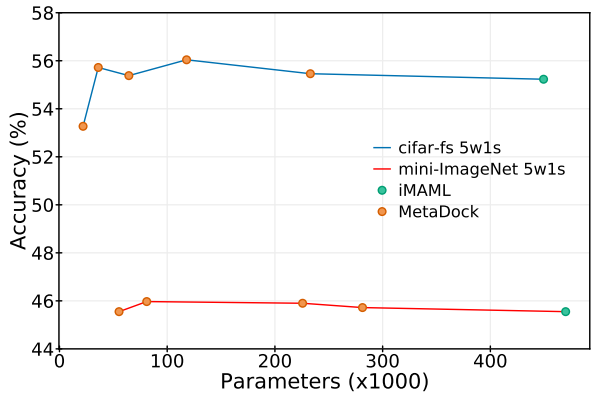


Figure 2. Accuracy vs Parameters for 4-Conv 128 channels pruned with METADOCK.

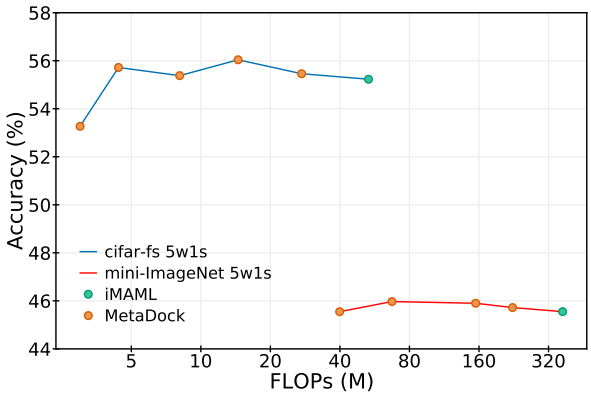


Figure 3. Accuracy vs FLOPs for 4-Conv 128 channels pruned with METADOCK.

of the different input kernels vary across different output channels with some kernels being more important than the others for an output channel. Figure 1b shows the same distribution but for a model pruned with METADOCK to a budget constraint of 50% on the total fraction of the kernels to be used.

D. Parameters and FLOPs

Figure 2 shows parameters vs. accuracy plot for CIFAR-fs and mini-ImageNet datasets on 5-way, 1-shot setting with 4-Conv-128 model and Figue 3 shows FLOPs vs. accuracy plot for CIFAR-fs and mini-ImageNet datasets on 5-way, 1-shot setting with 4-Conv-128 model