Peak Memory Usage (MobileNet-v2 on ImageNet)			
Pruning Method : DEPrune-BH			
Pruning Ratio	Pre-pruning	After-pruning	GAP
50%	7.22 MB	3.63 MB	3.59 MB

Table 1: Analysis of Peak Memory Usage (MB) with DEPrune-BH on ImageNet. 'GAP' means the after-pruning peak memory usage difference rate compared to pre-pruning. DEPrune-BH applies filter pruning using ℓ_2-norm to PW-conv.

MobileNet-v2 on CIFAR-10			
Method	Pruning ratio	Accuracy	diff.
MobileNet-v2	-	93.86%	-
NVIDIA n:m sparsity	50%	92.99%	-0.87%
DEPrune-B	50%	93.30%	-0.56%

Table 2: Comparison of accuracy (%) with DEPrune-B and NVIDIA n:m pruning on CIFAR-10. diff. means the accuracy difference rate compared to baseline. NVIDIA n:m pruning's n and m size are 2 and 4. DEPrune-B applies filter pruning using ℓ_2-norm to PW-conv.

ConvNeXt-Base on ImageNet						
Method	Pruning Ratio		Top-1 Accuracy			Inference Time Speed Up
	DW-conv	PW-conv	Baseline	Pruned	diff.	
ConvNeXt-Base	-	-	83.68%	-	-	1.00x
FP (ℓ_2-norm)	30%	30%	83.68%	82.00%	-1.68%	1.51x
DEPrune-BH	50%	30%	83.68%	82.46%	-1.22%	1.88x
DEPrune-BH	71%	41%	83.68%	82.22%	-1.46%	3.30x

Table 3: Comparison of inference time with DEPrune-BH and filter pruning (FP) on ConvNeXt-Base. diff. means the top-1 accuracy difference rate compared to baseline. DEPrune-BH applies filter pruning using ℓ_2-norm to PW-conv.

	Pre-processing Time (**offline**)	Inference Time (**online**)
Pruning Process (DEPrun-BH)	7minutes	-
Fine-tuning (65epoch)	19 hours 35 minutes	-
Pruned Model	-	450us

Table 4: Analysis of pre-processing time and inference time with DEPrune-BH in MobileNet-v3-small on ImageNet. Pre-processing time is offline time, which is a process that takes place before the model inference. We experiment of fine-tuning on A100 GPU-40GB.

Edge GPU	NVIDIA GPU Jetson Orin Nano 8GB Developer Kit		
Model	MobileNet-v2 on ImageNet		
Method	Baseline	GFS [50]	DEPrune-BH
Pruning Ratio	-	43%	DW-conv 75%, PW-conv 65%
Inference Time	27.41ms	17.90ms	11.04ms

Table 5: Analysis of inference time (ms) with DEPrune-BH and structured pruning on Edge GPU (NVIDIA Orin Nano 8GB). DEPrune-BH applies filter pruning using ℓ_2-norm to PW-conv.