A Supplementary Results for ReLUification

In this section, we display the ReLUification results which are not included in the main paper. In Section 3 and Section 4, we only presented the representative results for each model, using the best teacher model in Table 1.

A.1 ReLUification Results with and without KD

Table 9, 10 and 11 presents the ReLUification results with and without knowledge distillation using Swish, Mish and GeLU activation functions. The representative results are presented in the Table 4 of the main paper.

Table 9. ReLUification results of Swish models using high learning rate (Top-1 Accuracy in %). For each ReLUification method, the improvements of using a high learning rate compared to a low learning rate are also reported.

Model	Baseline		ReLUification w/ KD		ReLUification w/o KD	
	ReLU	Teacher	LR=0.1	LR=0.01	LR=0.1	LR=0.01
CIFAR100						
ResNet18	75.25	75.79	75.59 (▲0.34)	75.25	73.91 (▼1.09)	75.00
ResNet34	75.76	75.70	75.53 (▼0.01)	75.54	74.68 (▼0.42)	75.10
InceptionV3	74.51	73.89	72.74 (▲1.67)	71.07	72.71 (▼0.16)	72.87
ShuffleNetV1	69.04	70.97	70.49 (▲1.71)	68.78	69.43 (▲0.03)	69.40
ShuffleNetV2	67.16	68.60	67.49 (▲1.12)	66.37	67.51 (▲0.29)	67.22
MobileNetV1	67.35	69.39	67.04 (▲2.17)	64.87	44.15 (▼21.2)	65.35
ImageNet						
ResNet18	69.96	70.79	69.84 (▲0.32)	69.52	65.26 (▼3.34)	68.60
MobileNetV3	63.71	67.31	65.50 (▲1.52)	63.98	63.09 (▼1.52)	64.61

Table 10. ReLUification results of Mish models using high learning rate (Top-1 Accuracy in %).

Model	Baseline		ReLUification w/ KD		ReLUification w/o KD	
	ReLU	Teacher	LR=0.1	LR=0.01	LR=0.1	LR=0.01
CIFAR100						
ResNet18	75.25	75.53	75.16 (▲0.52)	74.64	73.92 (▼0.53)	74.45
ResNet34	75.76	75.94	75.79 (▲0.18)	75.61	74.68 (▼0.94)	75.62
InceptionV3	74.51	76.20	75.43 (▲2.10)	73.43	76.29 (▲1.47)	74.82
ShuffleNetV1	69.04	70.33	69.99 (▲1.22)	68.77	69.99 (▲0.66)	69.33
ShuffleNetV2	67.16	68.71	67.58 (▲1.18)	66.40	67.78 (▲0.76)	67.02
MobileNetV1	67.35	68.75	67.08 (▲2.31)	64.77	44.66 (▼20.5)	65.14
ImageNet						
ResNet18	69.96	70.83	69.83 (▲0.30)	69.53	62.48 (▼6.03)	68.51
MobileNetV3	63.71	67.04	64.66 (▲1.69)	62.97	62.58 (▼1.24)	63.82

Table 11. ReLUification results of GeLU models using high learning rate (Top-1 Accuracy in %).

Model	Baseline		ReLUification w/ KD		ReLUification w/o KD	
	ReLU	Teacher	LR=0.1	LR=0.01	LR=0.1	LR=0.01
CIFAR100						
ResNet18	75.25	75.68	75.78 (▲0.42)	75.36	74.21 (▼0.76)	74.97
ResNet34	75.76	75.53	75.39 (▲0.09)	75.30	74.11 (▼0.74)	74.85
InceptionV3	74.51	74.43	74.08 (▲1.13)	72.95	74.34 (▲0.90)	73.44
ShuffleNetV1	69.04	70.24	69.56 (▲1.28)	68.28	69.59 (▲1.32)	68.27
ShuffleNetV2	67.16	68.24	66.94 (▲0.74)	66.20	66.51 (▼0.40)	66.91
MobileNetV1	67.35	68.74	67.45 (▲1.58)	65.87	58.19 (▼7.72)	65.91
ImageNet						
ResNet18	69.96	70.83	70.05 (▲0.11)	69.94	65.54 (▼3.45)	68.99
MobileNetV3	63.71	67.45	65.23 (▲1.34)	63.89	63.52 (▼0.79)	64.31

A.2 ReLUification Results with Selective Exclusion

Table 12, 13 and 14 presents the ReLUification results with selective exclusion, using Swish, Mish and GeLU activation functions. The representative results are presented in the Table 5 of the main paper.

Table 12. ReLUification results of Swish models using selective exclusion (Top-1 Accuracy in %). All models are separated into three parts and during ReLUification, one of them remained as smooth function layer. Best ReLUification results are highlighted for each model.

Dataset	Model	Excluded Part			All
		Front	Middle	Rear	ReLU
CIFAR100	ResNet18	75.70	75.53	**75.96**	75.59
	ResNet34	75.63	75.47	**75.74**	75.53
	InceptionV3	73.00	73.05	**73.22**	72.74
	ShuffleNetV1	70.94	**70.95**	70.80	70.49
	ShuffleNetV2	67.68	**68.02**	67.89	67.49
	MobileNetV1	69.26	69.26	**69.39**	67.04
ImageNet	ResNet18	70.02	70.34	**70.34**	69.84
	MobileNetV3	66.21	66.37	**66.46**	65.50

Table 13. ReLUification results of Mish models using selective exclusion (Top-1 Accuracy in %).

Dataset	Model	Excluded Part			All
		Front	Middle	Rear	ReLU
CIFAR100	ResNet18	75.24	**75.30**	75.08	75.16
	ResNet34	**75.92**	75.62	75.73	75.79
	InceptionV3	75.24	**76.18**	75.34	75.43
	ShuffleNetV1	**70.63**	70.53	70.35	69.99
	ShuffleNetV2	67.78	68.06	**68.31**	67.58
	MobileNetV1	68.69	**68.84**	68.77	67.08
ImageNet	ResNet18	70.07	70.36	**70.43**	69.83
	MobileNetV3	65.45	65.53	**65.57**	64.66

Table 14. ReLUification results of GeLU models using selective exclusion (Top-1 Accuracy in %).

Dataset	Model	Excluded Part			All
		Front	Middle	Rear	ReLU
CIFAR100	ResNet18	**75.89**	75.70	75.59	75.78
	ResNet34	75.68	75.48	**75.68**	75.39
	InceptionV3	74.22	74.26	**74.36**	74.08
	ShuffleNetV1	**70.40**	70.38	70.03	69.56
	ShuffleNetV2	67.19	67.33	**67.43**	66.94
	MobileNetV1	**68.73**	68.56	68.60	67.45
ImageNet	ResNet18	70.12	**70.54**	70.51	70.05
	MobileNetV3	65.70	66.05	**66.22**	65.23

B Heuristics for Selective Exclusion

In this section, we elaborate more on the heuristics we use to divide each model into three groups for selective exclusion during ReLUification. Our image classification models reduce the width and height of the input feature map at every downsampling layer. On ImageNet, the models downsample the input images, whose width and height are 224, for five times. On CIFAR100, the models downsample the input images, whose width and height are 32, for three or four times. For both datasets, the last pooling layer is applied to shrink the width and height to 1, before feeding the produced feature map into the linear classifier.

Figure 8, 9, 10, 11, 12 and 13 visualizes the selective exclusion of the models. We defined the rear group as all the layers between the last trainable layer before linear layer and the last downsampling layer. The first and middle groups are defined with slight variations for each network architecture. ResNet18 and MobileNetV3 models are separated considering the number of layers included in each group. We designed middle group to have more layers than front group, while not leaving front group to be too thin. The other models on CIFAR100 are separated considering more about the position of downsampling layers. We pursued middle and rear group to have at least one downsampling layer and middle group to have equal or more downsampling layer than front group.

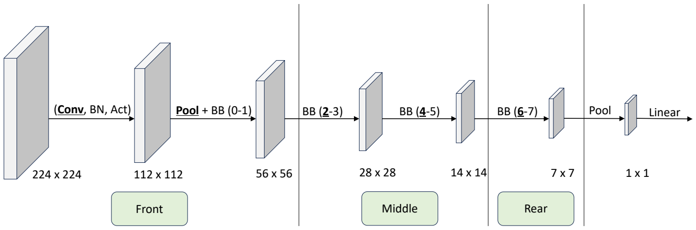


Fig. 8. Selective Exclusion (SE) of ResNet18 model on ImageNet. Downsampling layers are highlighted and the number of included blocks are presented for each group. BB stands for BasicBlock of ResNet18. ResNet18 and ResNet34 on CIFAR100 use the same separation.

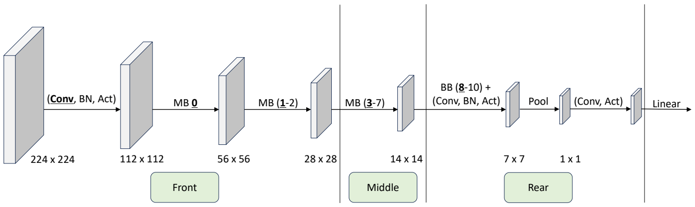


Fig. 9. Selective Exclusion (SE) of MobileNetV3 model on ImageNet. MB stands for Mobile Bottleneck block.

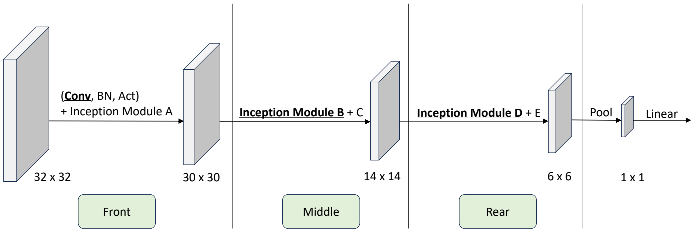


Fig. 10. Selective Exclusion (SE) of InceptionV3 model on CIFAR100.

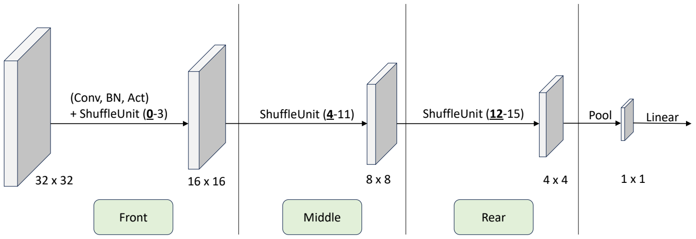


Fig. 11. Selective Exclusion (SE) of ShuffleNetV1 model on CIFAR100.

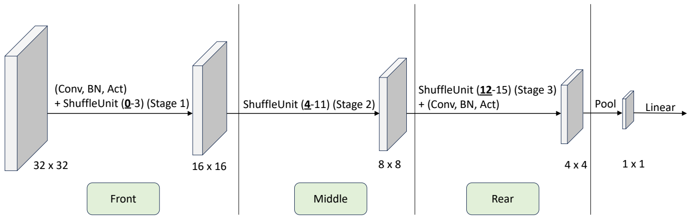

(Conv, BN, Act)
+ ShuffleUnit (0-3) (Stage 1)
ShuffleUnit (4-11) (Stage 2)
ShuffleUnit (12-15) (Stage 3)
+ (Conv, BN, Act)
Pool
Linear
32 x 32
16 x 16
8 x 8
4 x 4
1 x 1
Front
Middle
Rear

Fig. 12. Selective Exclusion (SE) of ShuffleNetV2 model on CIFAR100.

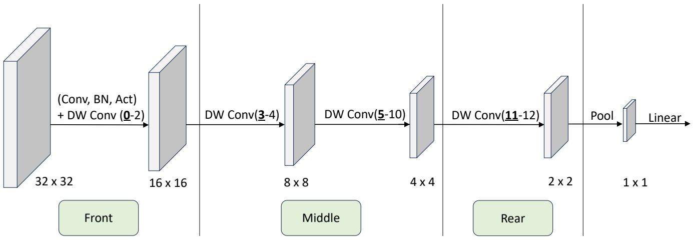


Fig. 13. Selective Exclusion (SE) of MobileNetV1 model on CIFAR100.

C Supplementary Results for Training Time and Accuracy Comparison

In Figure 7 of the main paper, we compared the training time against the model accuracy for our KD-based ReLUification methods with the approach of training ReLU models from scratch. In this section, we present the results for the other models not included in the main paper in Figure 14 and 15.

On CIFAR100, each model is trained from scratch for 50, 100, 200, and 400 epochs while ReLUification is conducted for 15, 30, 60, and 120 epochs. On ImageNet, each model is trained from scratch for 12, 24, 45, and 90 epochs while ReLUification is conducted for 6, 12, 24, and 48 epochs.

For ResNet18 and InceptionV3 models on CIFAR100, training from scratch with extended (=400) epochs results in improved baseline accuracy. Since our KD-based ReLUification results are tightly coupled with the teacher model's accuracy, it does not outperform this improved baseline. However, if such extra training cost is affordable, the improvement can also be applied to smooth function models and the ReLUification results of those models.

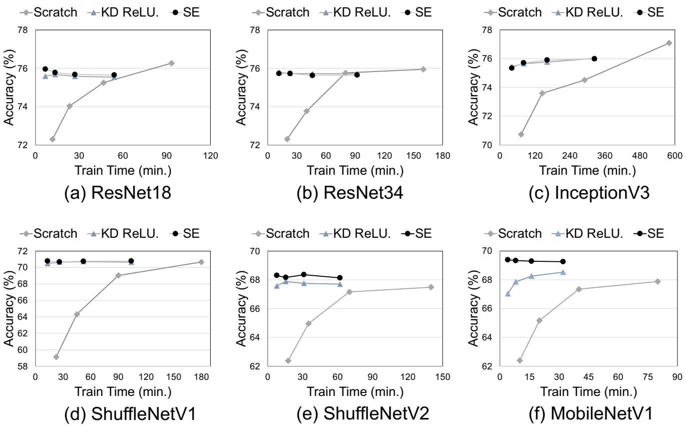
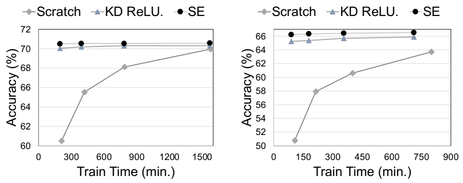


Fig. 14. Accuracy-training time comparison of 1) training from scratch 2) KD-based ReLUification with high LR and 3) with Selective Exclusion (SE). Accuracies are displayed with the total training time (min.) on CIFAR100.



(a) ResNet18 (b) MobileNetV3

Fig. 15. Accuracy-training time comparison of 1) training from scratch 2) KD-based ReLUification with high LR and 3) with Selective Exclusion (SE). Accuracies are displayed with the total training time (min.) on ImageNet.

D Hyperparameters for Experiments

Table 15 shows the default hyperparameters for training smooth function teacher models and ReLU baseline models. In Sec 5, we tested initial learning rate of {0.1, 0.01, 0.001, 0.0001, 0.00001} and adopted the best option respectively. As a result, Adam, AdamW and RMSProp used 0.001, AdaGrad used 0.01 and AdaDelta used 0.1. This is consistent with our observation with SGD, which achieved their best accuracy with their typical initial learning rate used for training from scratch. All other hyperparameters are identical to our main experiments in Sec 4.1.

Table 15. Default hyperparameter settings for training models.

Dataset	CIFAR100	ImageNet
Epochs	200	90
Batch Size	128	256
Optimizer	SGD with momentum 0.9	
Weight Decay	0.0001	
Learning Rate	Initial: 0.1 0.2× every 60 epochs	Initial: 0.1 0.1× every 30 epochs