PRUNE CHANNEL AND DISTILL: DISCRIMINATIVE KNOWLEDGE DISTILLATION FOR SEMANTIC SEGMENTATION - SUPPLEMENTARY MATERIAL -

Name of author

Korea University
Department of Electrical and Computer Engineering

1. TRAINING DETAILS

We train the student network with the Pytorch platform with an NVIDIA RTX TITAN GPUs.

Methods	Val mIoU (%)
T:DeepLabV3-RN101	78.07
S:DeepLabV3-RN18	74.21
+SKD	75.42
+PCD(ours)	77.55
T:DeepLabV3-RN101 + pruning	77.72
S:DeepLabV3-RN18	74.21
+SKD	75.88

Table 1. Comparison of different teacher networks for distillation on the validation set of Cityscapes.

1.1. Ablation Study

In this section, we will show the impact of the proposed method, and the performance differences according to the hyper-parameter are presented. We employ DeepLabV3-RN101 as the teacher and DeepLabV3-RN18 as the student by default. All experiments are conducted on validation of Cityscapes.

1.1.1. Pruning Ratios

Table 2 shows the mIoU variation according to different pruning ratios at each layer, r_l^d and r_l^m. The low l denotes semantically low feature maps and vice versa. (1) is a case of channel pruning without a discriminative score-based pruning process, and (2) is the opposite of (1). (1,2) still outperform student network and the results show that our PCD learning framework is effective in KD for semantic segmentation when we employ the proper channel pruning technique. (3) is the opposite case of (8), which is our hyper-parameter setting and has lower performance than (1,2) because discriminative score-based pruning is ineffective in semantically low feature maps. (4,5,6,7,8) show the results obtained by adding distillation sequentially from the lower layer. These results demonstrate that distillation for all semantic level features is effective.

	r_l^d (%)					r_l^m(%)					Val
l	1	2	3	4	5	1	2	3	4	5	mIoU (%)
S	0	0	0	0	0	0	0	0	0	0	74.21
(1)	0	0	0	0	0	75	75	75	75	50	76.80
(2)	75	75	75	75	50	0	0	0	0	0	77.05
(3)	75	75	50	25	0	0	0	25	50	50	75.82
(4)	0	0	0	0	0	75	0	0	0	0	75.36
(5)	0	0	0	0	0	75	75	0	0	0	76.90
(6)	0	0	25	0	0	75	75	50	0	0	76.61
(7)	0	0	25	50	0	75	75	50	25	0	77.08
(8)	0	0	25	50	50	75	75	50	25	0	77.55

Table 2. Evaluation of PCD with different pruning ratios on Cityscape validation set. S denotes the student model without distillation.

1.1.2. Feature Selection Methods

As shown in Table 3, we experimented with five feature selection methods: Random sort, which is distillation by randomly sorting the teacher and student features, Discriminative score sort that distillation is performed by sorting the features of the teacher and student based on discriminative score, Activation score sort that distillation is performed by sorting the features of the teacher and student based on activation score and Teacher matching is the opposite process of our Student matching. Random sort slightly improves the performance of the student network. Discriminative score sort and Activation score sort improved performance, but these are not the best models. Our student matching outperforms Teacher matching by 1.42%, and the result demonstrates that the distillation of the teacher's channel suitable for the student feature is effective.

2. EVALUATION OF PRUNED TEACHER

We train unstructured pruned teacher networks by applying [1] with a pruning ratio of 50% following [2]. The pruned teacher reduces the capacity gap when using SKD. However, this is still significantly lower compared to the performance of our proposed method. Since applying our SPD to pruned

feature selection	Val mIoU (%)
Random sort	74.26
Disciminative score sort	75.90
Activation score sort	76.87
Teacher matching	76.13
Student matching (ours)	77.55

Table 3. Comparison of different feature selection methods for distillation on the validation set of Cityscapes.

teachers does not fit our intention, we did not experiment.

3. ADDITIONAL QUALITATIVE RESULTS

Figure 1 shows additional qualitative results.

4. THE EXAMPLES OF ACTIVATION MAPS

Figure 2 is a visualization of the channel of the activation maps with low and high discriminative scores at layer $l = 4$. Figure 3 is a visualization of the channel of the activation maps with low and high discriminative scores at layer $l = 5$.

5. DISCUSSION AND LIMITATION

Our method can be applied to various CNN-based deep network architectures because our approach is based on channel pruning. Since our method directly transfers intermediate feature maps from the teacher network to the student, there is no significant performance improvement when heterogeneous architectures KD. In contrast, the student network achieves significant performance improvement when homogeneous architectures. However, if only high-level feature maps are considered, our proposed method can be sufficiently utilized in Transformer and CNN structures as depicted in the supplementary material.

6. WHY DO DISTILLATION INSTEAD OF PRUNING?

In the actual use of deep networks, there are situations in which a specific model architecture must be used, such as software, and hardware. However, structured pruning does not always provide the same architecture, and the generated model may be difficult to operate on the required hardware. Therefore, knowledge distillation is an effective way to satisfy this.

7. APPLICATION OF PCD TO TRANSFORMER

Due to the disparate structures of the intermediate feature maps in Transformer and CNN, our proposed method cannot be straightforwardly applied to Knowledge Distillation (KD) scenarios involving heterogeneous architectures. To address this, we can utilize the reshaping technique introduced by [3] for KD between Transformer and CNN. If the tokens of Transformer do not contain discriminative knowledge in intermediate feature representation, our PCD can be applied to the last layer of the transformer, which is expected to contain discriminative representations. When considering KD for the last layer, PCD can be employed irrespective of the model architecture differences between the teacher and student.

8. THE RANGE OF DISCRIMINATIVE SCORE

The discriminative score range depends on the sample's number of classes.

9. REFERENCES

[1] Song Han, Jeff Pool, John Tran, and William Dally, "Learning both weights and connections for efficient neural network," *Advances in neural information processing systems*, vol. 28, 2015.

[2] Heejo Kong, Gun-Hee Lee, Suneung Kim, and Seong-Whan Lee, "Pruning-guided curriculum learning for semi-supervised semantic segmentation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 5914–5923.

[3] Renjing Pei, Jianzhuang Liu, Weimian Li, Bin Shao, Songcen Xu, Peng Dai, Juwei Lu, and Youliang Yan, "Clipping: Distilling clip-based models with a student base for video-language retrieval," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 18983–18992.

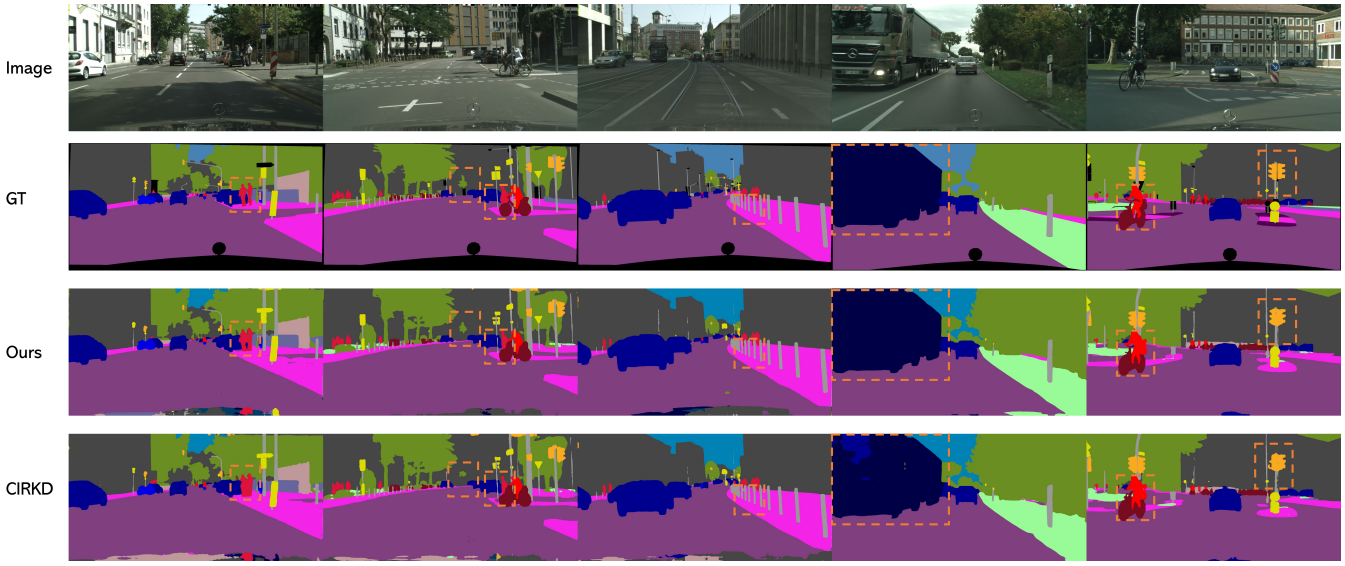

Fig. 1. Qualitative segmentation results using the DeepLabV3-RN18 on the validation set of Cityscapes

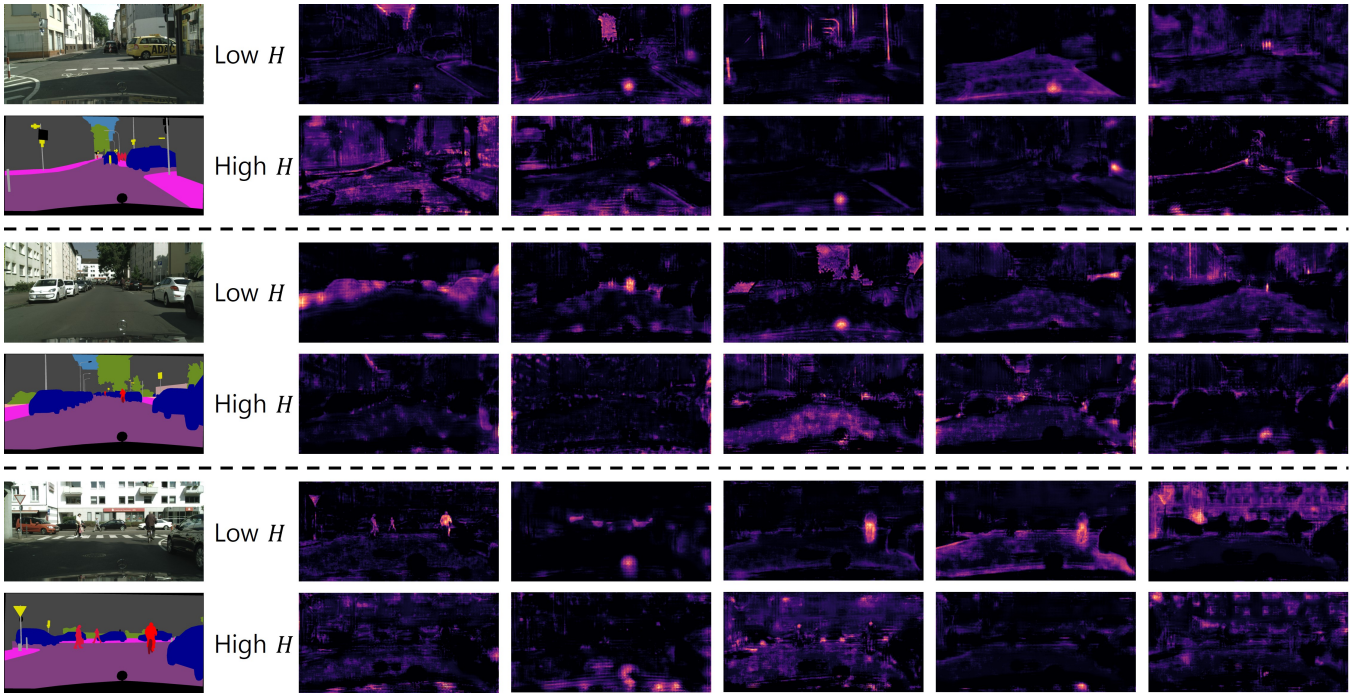

Fig. 2. The visualization of the channel of the activation maps with low and high H values. We extract feature maps $l = 4$ of teacher networks on the validation set of Cityscapes.

Fig. 3. The visualization of the channel of the activation maps with low and high H values. We extract feature maps $l = 5$ of teacher networks on the validation set of Cityscapes.