Class Attention Transfer for Semantic Segmentation

Yubin Cho
Department of Artificial Intelligence
Sogang University
Seoul, South Korea
ubinny77@naver.com

Sukju Kang
Department of Electronic Engineering
Sogang University
Seoul, South Korea
sjkang@sogang.ac.kr

***Abstract*—This paper proposes a knowledge transfer method using class attention maps (CAM) that are class-discriminative for training lightweight semantic segmentation networks. Since semantic segmentation classifies for each pixel, it is difficult to focus on the discriminative regions for each class in a single channel attention map. Thus, we generate attention maps for each class by using weights obtained from feature maps and class masks for squeezing the channels of the feature maps and then forcing a student network to generate the CAM that mimic the CAM of a teacher network. Our proposed method improves the state-of-the-art HRNetV2-W18+OCR by 4.78% in mIoU on the Cityscapes dataset.**

***Index Terms*—Knowledge Distillation, Semantic Segmentation, Attention Maps**

I. Introduction

Semantic segmentation is the task of classifying each pixel of an input image unlike typical classification that is a task of classifying an input image. In general, it requires complex networks with large capacity for accurate prediction. However, these networks require high computation, so it is difficult to be operated on edge devices with limited resources. Recently, the lightweight networks have been studied for computing on edge devices [1], [2]. As a result, lightweight models with satisfactory performance have drawn more attention. Methodologies such as pruning [1], [3], quantization [2], [4], knowledge distillation [5], [6] are mainly used to help train lightweight models.

Knowledge distillation is to transfer the knowledge from a large teacher network to a lightweight student network in the training stage. Since the student network has limitations in accepting all of the knowledge of teacher network, it is important to distill only useful knowledge from the teacher to the student network. The effectiveness of knowledge distillation has been proven in previous studies for classification tasks [2], [7], [8], [9]. Semantic segmentation, which classifies at the pixel-level, is a more challenging task than a classification task. However, previous studies [6] that directly applied knowledge distillation methods in classification [5] to semantic segmentation may not be effective.

In this paper, we propose a class attention transfer based on attention transfer [7], which is an effective knowledge distillation method in classification tasks, as shown in Fig. 1. Only the necessary information among the knowledge of teacher network can be transferred to the student network by using class attention maps (CAM) that activate the class-discriminative regions. We generate the CAM from the feature maps of each student and the teacher network, and then, the student network is guided to mimic the regions that need to be paid more attention. We show an example of the class attention maps in Fig. 2.

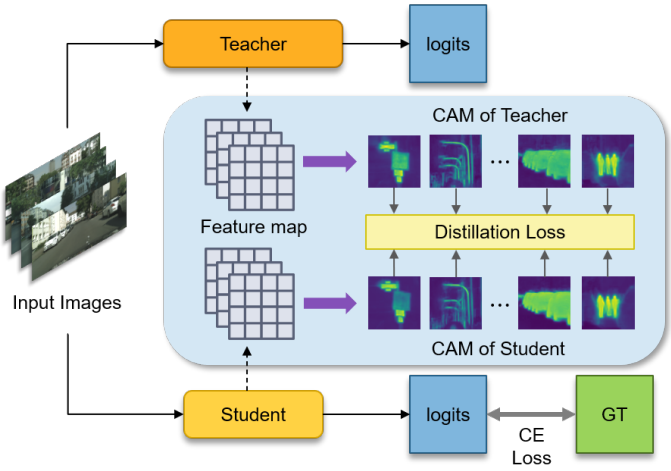


Fig. 1. The overall architecture of Class Attention Transfer. The feature map is used to generate Class Attention Maps. Only the student network is trained with the distillation loss and the task loss.

The main contributions of this paper are summarized as follows.

- We propose the attention transfer method using class attention maps generated by a novel method for semantic segmentation tasks.
- The proposed method shows better performance than applying attention transfer [7] directly to semantic segmentation.
- We implement our method on the latest state-of-the-art segmentation network, showing that the student network achieves similar performance to the teacher network.

II. Related Work

Most of the knowledge distillation methods are studied for classification tasks [2], [7], [8], [9], but we study the knowledge distillation method that focuses on efficient semantic segmentation. Prior to our work, there are several knowledge distillation methods for semantic segmentation tasks. In this

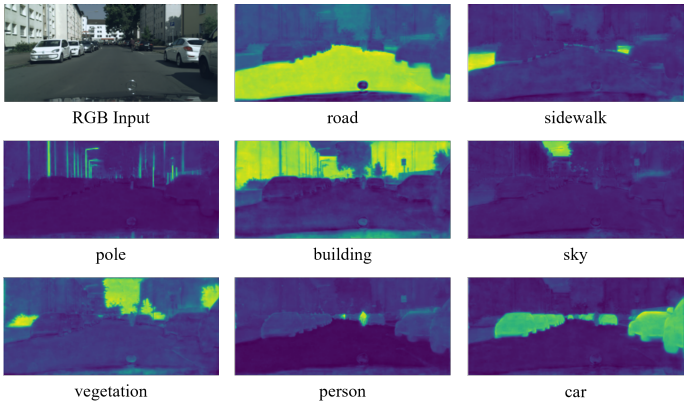


Fig. 2. Visualization Class Attention Maps generated by our method from the feature map of the teacher network. The CAM show activated regions corresponding to the categories.

section, we review some works related to the attention maps and knowledge distillation for semantic segmentation.

Attention Maps. Extracting attention maps from Convolutional Neural Networks (CNN) may be a method of informing how the hidden layers of CNN work. Visualizing the attention maps of CNN show where the networks focus in the images to perform tasks. [13] proposed deconvolutional networks to visualize what patterns caused activations in the feature maps. [14] proposed the guided backpropagation to visualize high level activations. However, they are non-class-discriminative. [15] proposed class activation maps using global average pooling as weights after the final convolution layer. The CAM was created by the weighted sum of the feature maps of the final convolution layer. It highlighted the class-discriminative image regions. [16] proposed Grad-CAM that combines CAM and the gradient of neurons. The gradient information in the final convolution layer of CNN was used as weights for the weighted sum of the feature maps.

Knowledge Distillation for Semantic Segmentation. [10] proposed local affinity to make the segmented boundary similar. It uses the local similarity map for each pixel, which is computed by the distance between itself and the 8-neighborhood pixels. L2 minimized the difference of the local similarity maps from the student and the teacher. [6] applied the pixelwise loss, the knowledge distillation method in classification [5], to semantic segmentation. It used KL divergence to align the class probability of each pixel from the student and teacher networks. [6] also proposed two structured distillation modules. The first module, pairwise distillation, transferred the similarity between pixel pairs using an affinity graph. The second module, holistic distillation, used the discriminator to align the high-order consistencies between output structures from the student and teacher networks. [11] was the first to propose CAM distillation. It created CAM by computing the gradient of the logits with respect to the feature maps. [12] proposed channel wise distillation to align the activations of each channel from the student and teacher networks. To use KL divergence, it converted the activations of each channel into a probability distribution for each channel by normalizing the activation maps in each channel.

III. Proposed Method

We propose the class attention transfer (CAT) in semantic segmentation, a task of performing classification for each pixel. In the classification task, the attention transfer [7] generates a single channel attention map by computing the summation of all channels at each pixel. This method is often used as a comparison method in several distillation works for semantic segmentation [6], [12]. However, the single attention map cannot focus on the discriminative regions for each class in semantic segmentation tasks. This section explains a new method to transfer attention information of each class from the teacher to the student by generating the class attention maps. We apply CAT to the feature map that is used as an input of the last convolution layer.

Let us denote the feature map as $F \in R^{C^{'} \times H \times W}$ which consists of $C^{'}$ feature channels with spatial dimensions of $H \times W$. We assume that the spatial resolution of the feature map in the teacher and student is the same. The feature map from the teacher and the student are F^T and F^S. To generate the class attention maps, we need to obtain weights using global average pooling (GAP). The feature map that focuses on the regions of the i^{th} class has a larger weight. The weights of the i^{th} class are calculated as:

$$w_i^{C^{'} \times 1 \times 1} = GAP\left(\frac{|F^T|^p}{\left\||F^T|^p\right\|_2} \otimes M_i\right), \tag{1}$$

where $i = 1, 2, ..., C$ indexes the class and $\otimes$ denotes element-wise multiplication. $|\cdot|^p$ denotes the element-wise power operator, and we use $p = 2$, which is experimentally optimal. M_i denotes the i^{th} class mask. $\left|F^T\right|^p$ is normalized in each channel using l_2-normalization. We obtain the class mask $M^{C \times H \times W}$ by applying a softmax function to the output logits of the teacher at each pixel and setting the pixel value in each channel to 1 if it is 0.5 or more. The i^{th} class attention map is created by the weighted sum of all channels of the feature map as below.

$$A_i^{1 \times H \times W} = \frac{1}{C} \sum_{k=1}^{C} w_i \otimes F. \tag{2}$$

We visualize the process of generating the i^{th} class attention map in Fig. 3. We repeat the processes of (1) and (2) C times to generate class attention maps corresponding to each class.

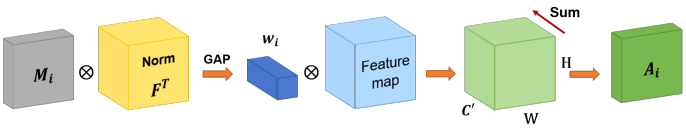


Fig. 3. The process of generating the i^{th} class attention map. The weights obtained from global average pooling are used to sum the channels of the feature map.

Let $A^{C\times H\times W}$ denote the concatenation of all class attention maps. Finally, the loss function of CAT is defined as follows.

$$Loss_{CAT} = \frac{1}{N}\sum_{j=1}^{N}\left(\frac{|A_j^S|^2}{\left\||A_j^S|^2\right\|_2} - \frac{|A_j^T|^2}{\left\||A_j^T|^2\right\|_2}\right)^2, \quad (3)$$

where $j = 1, 2, ..., N$ indexes the number of the batch. A^T is the concatenated CAM of the teacher and A^S is the concatenated CAM of the student. As shown in (3) , during class attention transfer, we use class attention maps that is l_2-normalized in each channel. The conventional cross entropy loss is also used for training. Finally, the total loss can be defined as follows.

$$Loss_{total} = Loss_{CE} + \lambda \cdot Loss_{CAT}, \quad (4)$$

where λ is a hyper-parameter to balance the loss terms.

IV. Experimental results

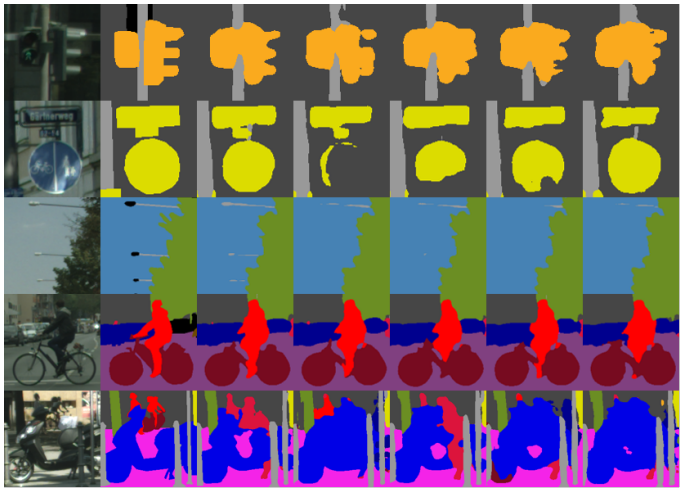
In this section, we explore the proposed Class attention transfer on semantic segmentation dataset. We compare our method with other distillation methods to verify the effectiveness of our method.

A. Experimental settings

1) Implementation Details: We used the semantic segmentation network of OCR [17] with HRNetV2-W48 [18] (HRNet-W48+OCR) as the teacher network for all experiments. We first used OCR with the backbone of HRNetV2-W18 [18] (HRNet-W18+OCR) as the student network to verify the effectiveness of our method. Then, we designed HRNetV2-W18 to be lighter by reducing the number of layer blocks from 4 to 2 per stage. We named it HRNet-W18-slim and used it as the backbone (HRNet-W18-slim+OCR). In training, networks used SGD with the momentum of 0.9 and the weight decay of 0.0005. The batch size was set to 4, and the number of iterations was 40K. The hyper-parameter λ was set to 0.4. The learning rate was initialized to be 0.01 and was multiplied by $\left(1 - \frac{iter}{max-iter}\right)^{0.9}$. We randomly cropped the images into 1024×512 pixels as the training input.

2) Datasets: The Cityscapes dataset [19] is collected for semantic urban scene understanding and contains 30 common classes with only 19 classes used for evaluation and testing. The dataset contains $5,000$ finely annotated images with $2,950/500/1,525$ images for training, validation, testing respectively. The image size is 2048×1024 pixels.

3) Evaluation Metrics: To evaluate the performance, we used the following metrics. The mean Intersection-over-Union (mIoU) is measured on the single scale and no flipping setting in all experiments. We also used the class IoU to evaluate the effectiveness of distillation on each class. The number of parameters of the network represented the model size. The floating-point operations per second (FLOPs) were computed with an input size of 512×1024 pixels to evaluate the complexity.



(a) Image (b) GT (c) Ours (d) AT (e) CD (f) PI (g) Student

Fig. 4. Qualitative segmentation results on validation set of Cityscapes produced from HRNetV2-W18+OCR. (a) input images, (b) ground truth, (c) class attention transfer, (d) attention transfer, (e) channel-wise distillation, (f) pixel-wise distillation, (g) original student model without KD.

TABLE I
Comparison of various distillation methods against our method on the validation set of Cityscapes.

Method	#Params (M)	FLOPs (G)	mIoU (%)
Teacher [17]	70.37	301.6	81.6
HRNet-W18+OCR	12.08	98.6	72.71
+ PI [6]	12.08	98.6	73.53
+ AT [7]	12.08	98.6	75.23
+ CD [12]	12.08	98.6	76.03
+ CAT (Ours)	12.08	98.6	**77.49**
HRNet-W18-slim+OCR	7.92	85.09	71.13
+ PI [6]	7.92	85.09	72.44
+ AT [7]	7.92	85.09	74.25
+ CD [12]	7.92	85.09	75.39
+ CAT (Ours)	7.92	85.09	**76.25**

B. Comparison with Knowledge Distillation Methods

We used state-of-the-art distillation methods [6], [7], [12] to compare with our method. We apply our method and attention transfer (AT) [7] to the inner feature map while pixel-wise distillation (PI) [6] and channel-wise distillation (CD) [12] are applied in the final logits map.

As reported in Table I, our method significantly improved mIoU of the student networks. This indicates that our method can be applied well to both a model with the reduced number of channels and a model with the reduced number of layers. Moreover, the proposed method outperformed all the other distillation methods on both student networks. Our method outperformed the best segmentation distillation method (CD) and the best spatial distillation method (AT) by 1.46% and by 2.27% on HRNet-W18+OCR, respectively.

The qualitative segmentation results intuitively demonstrated the effectiveness of our method in Fig. 4. Our method produced more accurate results compared to other methods.

TABLE II
CLASS IOU OF OUR PROPOSED METHOD COMPARED WITH OTHER TWO KNOWLEDGE TRANSFER METHODS ON THE VALIDATION SET OF CITYSCAPES, WHERE HRNETV2-W18+OCR IS USED AS THE STUDENT NETWORK.

Class	mIoU	road	sidewalk	building	wall	fence	pole	traffic light	traffic sign	vegetation
AT	75.23	97.97	83.95	91.47	51.84	58.08	63.02	69.67	76.13	91.98
CD	76.03	98.18	85.12	92.15	**60.66**	59.31	64.60	67.75	76.58	92.16
Ours	**77.49**	**98.24**	**85.85**	**92.46**	58.13	**60.56**	**66.37**	**71.72**	**78.48**	**92.43**
Class	terrian	sky	person	rider	car	truck	bus	train	motorcycle	bicycle
AT	63.89	93.86	78.67	57.29	93.76	73.99	83.19	70.97	55.77	73.78
CD	62.29	94.28	80.04	56.56	94.19	76.75	82.78	70.36	56.17	74.68
Ours	**65.85**	**94.58**	**81.53**	**59.31**	**94.53**	**78.32**	**83.88**	**73.57**	**60.19**	**76.25**

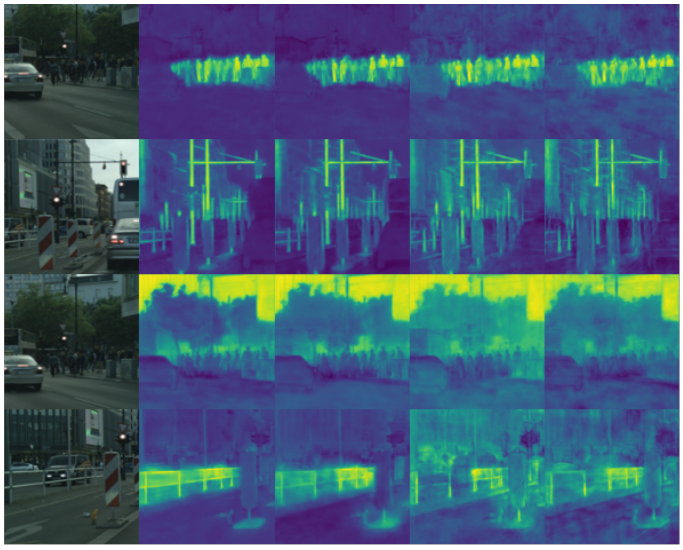


Fig. 5. The output logits of several classes on test set of Cityscapes, where HRNetV2-W18+OCR is used as the student. (a) input image, (b) teacher model, (c) class attention transfer, (d) attention transfer, (e) original student model without KD. From the top, each row shows the results for person, pole, building and fence.

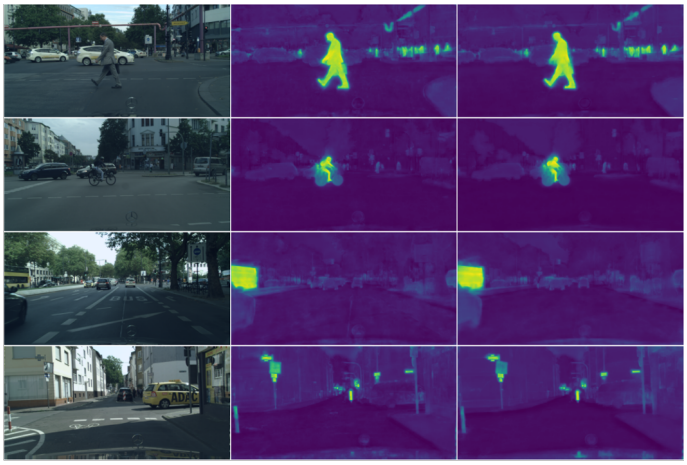


Fig. 6. Visualization CAM of several classes on test set of Cityscapes, where the CAM is generated by our method. (a) input image, (b) teacher model, (c) student model with class attention transfer. From the top, each row shows the results for person, rider, bus and traffic sign.

From Table II, we reported the detailed class IoU of our method and other two methods, attention transfer [7] and channel-wise distillation [12]. Our method improved the accuracy of almost all objects. This indicates that our class attention transfer can transfer the structural information well.

We visualized the output logits and CAM of several classes in Fig. 5 and Fig. 6, respectively. The logits of our method were significantly similar to the logits of the teacher unlike the logits of other models. The CAM of the student trained by our method hightlighted the class-discriminative regions. This indicates that the CAM generated by our method is useful for transferring the knowledge of the teacher to the student.

V. CONCLUSION

This paper proposed the class attention transfer with CAM generated by performing the weighted sum of the feature maps using GAP as weights. This enables to distill the information about the discriminative regions for each class. Experimental results showed that our method can significantly improve the performance of the student networks and outperforms other distillation methods on semantic segmentation. Furthermore, we hope that the proposed method can help train lightweight networks in many other tasks, such as instance segmentation, and panoptic segmentation.

ACKNOWLEDGMENT

This research was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.2021-0-02308, A Study on Edge Computer-based Deep Neural Network Technology for Solving Local Community and Living Problems in Multi-modal Environment)

REFERENCES

[1] D. Wofk, F. Ma, T.-J. Yang, S. Karaman, and V. Sze, "Fastdepth: Fast monocular depth estimation on embedded systems," *Int. Conf. on Robotics and Automation,* 2019.

[2] D. Zhang, J. Yang, D. Ye, and G. Hua, "Lq-nets: Learned quantization for highly accurate and compact deep neural networks," in *Proc. Eur. Conf. Comp. Vis.*, 2018.

[3] T. Zhang, S. Ye, K. Zhang, J. Tang, W. Wen, M. Fardad, and Y. Wang, "A systematic dnn weight pruning framework using alternating direction method of multipliers," in *Proc. Eur. Conf. Comp. Vis.*, 2018, pp. 184-199.

[4] H. V. Habi, R. H. Jennings, and A. Netzer, "Hmq: Hardware friendly mixed precision quantization block for cnns," in *Proc. Eur. Conf. Comp. Vis.*, 2020.

[5] G. E. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv: Comp. Res. Repository*, vol. abs/1503.02531, 2015.

[6] Y. Liu, K. Chen, C. Liu, Z. Qin, Z. Luo, and J. Wang, "Structured knowledge distillation for semantic segmentation," *IEEE Conf. Comp. Vis. Patt. Recog.*, 2019.

[7] S. Zagoruyko and N. Komodakis, "Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer," in *Proc. Int. Conf. Learn. Representations*, 2017.

[8] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, and Y. Bengio, "Fitnets: Hints for thin deep nets," in *Proc. Int. Conf. Learn. Representations*, 2015.

[9] S. I. Mirzadeh, M. Farajtabar, A. Li, N. Leveine, A.Marsukawa, and H. Ghasemzadeh, "Improved knowledge distillation via teacher assistance," in *Proc. AAAI Conf. Artificial Intell.*, 2020, pp. 5191-5198.

[10] J. Xie, B. Shuai, J.-F. Hu, J. Lin, and W.-S. Zheng, "Improving fast segmentation with teacher-student learning," in *Proc. British Machine Vis. Conf.*, 2018.

[11] N. K. Bavandpour and S. Kasaei, "Class attention map distillation for efficient semantic segmentation," *Int. Conf. Mach. Vis. Image Process.*, 2020.

[12] C. Shu, Y. Liu, J. Gao, Z. Yan, and C. Shen, "Channel-wise knowledge distillation for dense prediction," in *Proc. Int. Conf. Comp. Vis.*, 2021.

[13] M. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. Eur. Conf. Comp. Vis.*, 2014, pp. 818–833.

[14] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M.Riedmiller, "Striving for simplicity: The all convolutional net," *arXiv preprint arXiv: Comp. Res. Repository*, vol. abs/1412.6806, 2014.

[15] B. Zhou, A. Khosla, A. Lapedriza, A. Olivia, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2016, pp. 2921-2929.

[16] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comp. Vis.*, 2017, pp. 618–626.

[17] Y. Yuan, X. Chen, and J. Wang, "Segmentation transformer: Object-contextual representations for semantic segmentation," in *Proc. Eur. Conf. Comp. Vis.*, 2020.

[18] J. Wang, et al, "Deep high-resolution representation learning for visual recognition," *arXiv: Comp. Res. Repository*, vol. abs/1904.04514, 2019.

[19] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2016.