Class-wise Knowledge Distillation for Lightweight Segmentation Model

Ryota Ikedo[a], Kotaro Nagata[b] and Kazuhiro Hotta[c]
Meijo University, 1-501 Shiogamaguchi, Tempaku-ku, Nagoya 468-8502, Japan

Keywords: Knowledge Distillation, Class-wise, Semantic Segmentation.

Abstract: In recent years, we have been improving the accuracy of semantic segmentation by deepening segmentation models, but large amount of computational resources are required due to the increase in computational complexity. Therefore knowledge distillation has been studied as one of model compression methods. We propose a knowledge distillation method in which the output distribution of a teacher model learned for each class is used as a target of the student model for the purpose of memory compression and accuracy improvement. Experimental results demonstrate that the segmentation accuracy was improved without increasing the computational cost on two different datasets.

1 INTRODUCTION

Convolutional neural networks have shown good performance in various image recognition tasks such as image classification (He et al., 2016), object detection (Liu et al., 2016), pose estimation (Cao et al., 2018) and so on (Zhang et al., 2021). Semantic segmentation is a task that assigns class labels to all pixels in an input image. This technique has been applied to automatic driving (Cordts et al., 2016) and medical images (Zhao et al., 2020).

Semantic segmentation needs inference for all pixels in an input image, so relationship inter pixels and location of each class are important. Therefore, conventional methods have been proposed to improve the accuracy by enhancing the backbone network in order to enrich the extracted features (Zhao et al., 2017), and by introducing an attention mechanism that maintains the relationship inter-pixel (Vaswani et al., 2017) and inter-channel (Hu et al., 2018). However, those methods require a lot of computation cost because they require additional convolutional layers and other mechanisms to obtain informative features. Thus, there is generally a trade-off between computational complexity and accuracy in semantic segmentation.

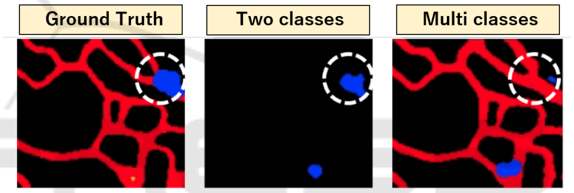


Figure 1: Comparison of two-class and multi-class segmentation. It shows that the segmentation accuracy is high in the case of only two classes, but not in the case of multiple classes.

In recent years, several methods have been proposed to solve the computational resource problem in various tasks, such as pruning (Han et al., 2015), quantization (Hubara et al., 2016) and knowledge distillation (Hinton et al., 2015). Knowledge distillation is an effective method among them, in which a computationally inefficient student model learns to the output distribution of a computationally expensive teacher model. As a result, high accuracy can be achieved with fewer computational resources.

Conventional knowledge distillation uses two kinds of loss functions; soft target loss and hard target loss. Soft target loss is the distillation loss that makes the output distribution of the student model closer to that of the teacher model using Mean Squared Error(MSE) loss or Kullback-Leibler(KL) divergence. Hard target loss is the use of cross entropy to make the output distribution of the student model closer to one hot vector representing the label. These two losses transfer knowledge from the teacher model to the student model. However, there is a difference in feature extraction capability between the teacher and student models. It is therefore highly difficult to exceed the performance of the teacher model by simply making

[a] https://orcid.org/0000-0002-8139-0623
[b] https://orcid.org/0000-0001-8256-2303
[c] https://orcid.org/0000-0002-5675-8713

Ikedo, R., Nagata, K. and Hotta, K.
Class-wise Knowledge Distillation for Lightweight Segmentation Model.
DOI: 10.5220/0011719900003414
In *Proceedings of the 16th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2023) - Volume 4: BIOSIGNALS*, pages 287-293
ISBN: 978-989-758-631-6; ISSN: 2184-4305

the output distribution of the student model closer to that of the teacher model.

In this paper, we propose a new knowledge distillation method for semantic segmentation. Semantic segmentation is usually more difficult to learn multiple classes in an input image, resulting in less accurate inference. Therefore, a model that inferences only for one class is more accurate for each class than a model that inferences for multiple classes as shown in Figure 1. We focus on this idea and propose a method to improve the accuracy by distilling the knowledge of multiple teacher models specific to one class into student models that inferences multiple classes.

We evaluate our proposed method using two different datasets; ssTEMD (Gerhard et al., 2013) and COVID-19 (Zhao et al., 2020). Experimental results show that the proposed method provide 0.96% and 1.30% improvements compared to the conventional knowledge distillation method which uses a single supervised model trained with multiple classes.

2 RELATED WORKS

2.1 Semantic Segmentation

Various segmentation methods have been proposed such as SegNet (Badrinarayanan et al., 2017), Deeplab (Chen et al., 2017), U-net (Ronneberger et al., 2015) so on. In recent studies, many methods have been proposed to utilize deep and large-scale networks such as ResNet (He et al., 2016) and EfficientNet (Tan and Le, 2019) as a backbone for encoders to improve the accuracy of feature extraction. For example, PSPNet (Zhao et al., 2017) adopted ResNet101 in the encode for feature extraction and introduced the Pyramid Pooling Module to extract features with high accuracy and handled both the global context of image and information on small parts of the image. The encoder for feature extraction is important for training these segmentation models. In this paper, low cost and high accuracy are achieved by transferring knowledge from a high accuracy model using an encoder with high computational cost to a model using a low cost encoder.

2.2 Knowledge Distillation

In recent years, the methods to provide high performance networks that are as lightweight as possible have been researched, e.g, pruning (Han et al., 2015) and quantization (Hubara et al., 2016). Knowledge distillation (Hinton et al., 2015) involves model compression by transferring knowledge from a large model called the teacher model, to a lightweight model called the student model. When distilling the knowledge of the teacher model into the student model, MSE loss and KL divergence are used to close the output distribution of the student model to that of the teacher model. Such distillation methods improved the accuracy and regularization of the student model with low computational cost. In the case of semantic segmentation, pixel-by-pixel knowledge distillation that makes the output distribution of each pixel in the student model closer to the output distribution of each pixel in the teacher model (Liu et al., 2019).

However, conventional knowledge distillation methods have the problem that teacher model's knowledge of small classes, i.e., classes that are difficult to recognize, is not well transferred to the student model. Thus, we propose class-wise knowledge distillation method, where knowledge is transferred from teacher models dedicated to each class to a student model. This derives more effective distillation than the case that knowledge of all classes were distilled from one teacher model or conventional knowledge distillation.

3 PROPOSED METHOD

We propose class-wise distillation method. The overview of our proposed method is shown in Figure 2. Our method provides C specific teacher models and one common teacher model with a large number of parameters. The specific teacher models are specialized for a particular class. The common teacher model is trained on all classes. Note that C is the number of classes. We also have one lightweight student model for knowledge distillation. When an image is fed into all models, the C specific teacher models output logits maps for two classes; specific class and the other class (background). The common teacher model and the student model output logits maps for all classes. The logits map for each class obtained by the student model is made closer to the output of teacher models.

Class-wise distillation method distils the knowledge of one class-specific teacher model into a student model. This allows class-specific knowledge to be obtained rather than the usual knowledge distillation of all classes from a single teacher model. The student model also learns to segment all classes and obtains relationships between classes that are not available in the teacher models specialized to only one class. The

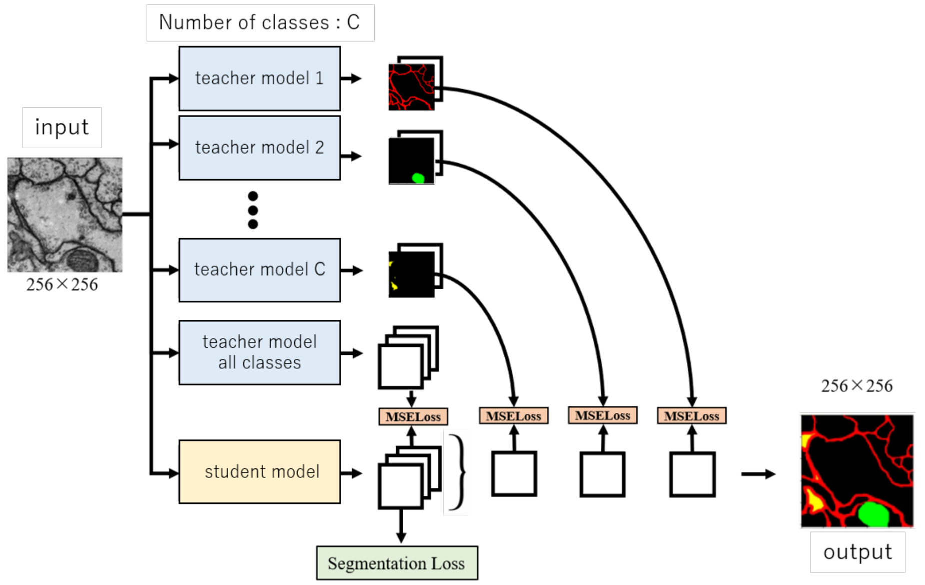


Figure 2: The overview of the proposed class-wise distillation. Teacher models are pre-trained models that are specialized to infer one particular class. These teacher models perform a two-class segmentation of specific classes and others (background). The student model is a lightweight network that performs segmentation for all classes. In class-wise distillation, the score for each class by the student model is learned to be close to the score by the teacher model specialized for that class.

student model is trained with the following loss:

$$\mathcal{L} = \mathcal{L}_{seg} + \sum_{n=1}^{c} \lambda_n \cdot \mathcal{L}_n + \lambda_{all} \cdot \mathcal{L}_{all} \quad (1)$$

where λ is a hyper-parameter representing the weight of each loss.

3.1 Segmentation Loss

Segmentation loss is the hard target loss that closes the score map between the label and the output of the student model. The image $x \in \mathbb{R}^{H\times W\times 3}$ is fed into the student model, and the student model outputs the logits map $s \in \mathbb{R}^{H\times W\times C}$ where H and W are the height and width of the input image and C is the number of classes. Segmentation loss is represented as softmax crossentropy as

$$\mathcal{L}_{seg} = \frac{1}{H\times W} \sum_{i=1}^{H\times W} \sum_{n=1}^{c} p_i^n \log q_i^n \quad (2)$$

where p_i^n and q_i^n represent the predicted probability and the target probability of class n at the i-th pixel.

3.2 Distillation Loss for all Classes

Distillation loss for all classes is a soft target loss that distills from the common teacher model trained on all classes to the student model. The common teacher model outputs the logits map $t \in \mathbb{R}^{H\times W\times C}$. All classes distillation loss computes MSE using the output of the teacher model and the student model.

$$\mathcal{L}_{all} = \frac{1}{H\times W} \sum_{i=1}^{H\times W} \sum_{n=1}^{C} (s_i^n - t_i^n)^2 \quad (3)$$

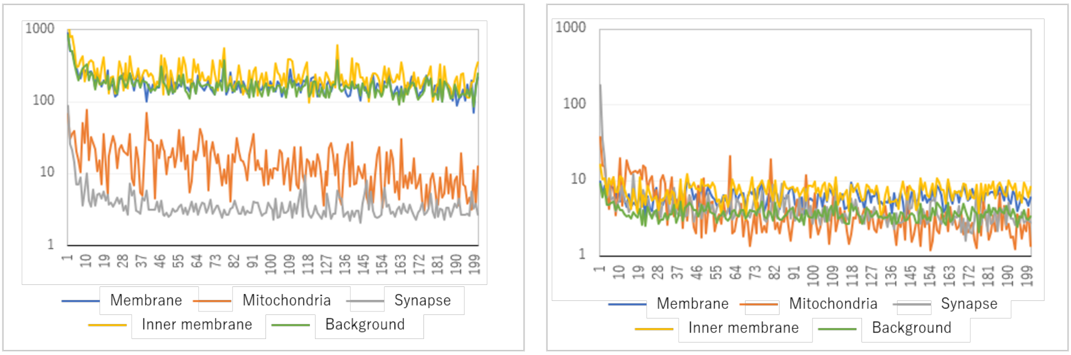
where s_i^n and t_i^n are the logits of class n at i-th pixel obtained by the student model and the common teacher model.

3.3 Class-wise Distillation Loss

Class-wise distillation loss is a soft target loss that distills from multiple teacher models specific to a particular class to one student model. The specific teacher models are pre-trained models that perform segmentation of two classes; a specific class and other class (background). They output a logits map $t \in \mathbb{R}^{H\times W\times 2}$. Class-wise distillation loss computes the MSE using the output of the specific teacher models and the student model.

$$\mathcal{L}_n = \frac{1}{H\times W} \sum_{i=1}^{H\times W} (s_i^n - t_i^n)^2 \quad (4)$$

where t_i^n is the logits of class n at i-th pixel obtained by the specific teacher models for the class n. Class-wise distillation loss allows the student model to mimic the logits maps of the teacher models specific to a particular class, and useful class-specific knowledge is transferred to the student model.



(a) Distillation losses without weights

(b) Distillation losses with weights

Figure 3: Transition of class-wise knowledge distillation loss.

4 EXPERIMENT

4.1 Experimental Setup

4.1.1 Dataset and Evaluation Metric

We conducted experiments using two dataset; ssTEM (Gerhard et al., 2013) and COVID-19 datasets (Zhao et al., 2020) The ssTEM dataset consists of Drosophila cell images with five classes; 'membrane', 'mitochondria', 'synapse', 'inner membrane' and 'background'. COVID-19 dataset has four classes of pneumonia; 'Background', 'Lungs other', 'Ground glass' and 'consolidations'. In both datasets, the input image is a monochrome image $x \in \mathbb{R}^{256\times256\times1}$. In both datasets, we also divided the annotated image data into training, validation and test for evaluation.

We used the standard evaluation protocol for semantic segmentation, Mean Intersection over Union (mIoU) averaged over all classes.

4.1.2 Training Setup

As described in the previous section, knowledge distillation is carried out using teacher models specialized for each class. These teacher models were pretrained two classes; a specific class and the other class (background). We used U-net for the student model and U-net with EffiicentNet-b7(U-net(EN-b7)) as encoder for the teacher model. The number of parameters in the teacher model is 65.75M and the number of parameters in the student model is 14.79M. That is, the student model has 1/4 parameters of the teacher model. We used softmax cross entropy as segmentation loss and MSE as class-wise distillation loss. We also used Adam as the optimizer.

4.1.3 Weight Parameters for Distillation Loss

We selected the weight parameter λ_n of class-wise distillation loss in equation (1) for optimal learning of our method. The loss of our method uses MSE Losses for each class. If those MSE losses are trained without weights, the loss per each class is shown in Figure 3(a). Figure shows that there is a large difference in loss between classes. In the case of such losses, the further the training focuses only on classes with large losses while it does not learns the classes with small losses. This may lead to learning bias in each class, and the data as a whole may not learn well. Thus, we adjusted the weight parameters so that the value of each distillation loss would be the same. The adjusted distillation losses are shown in Figure 3(b). The following experiment is performed when the distillation loss is corrected by weights.

4.2 Experimental Results

4.2.1 Results on ssTEM Dataset

Table 1 shows the results on the ssTEM dataset. First, we compare the performance of 2 classes U-net(EN-b7), teacher models specialised for each class, with 5 classes U-net(EN-b7), a teacher model trained on all classes. Teacher models specialized for each class achieved higher accuracy in almost classes than a model trained on all classes simultaneously.

We then compare the performance of 5 classes U-net(EN-b7) and 5 classes U-net which is a student model. 5 classes U-net(EN-b7) outperformed standard U-net by +2.96%. U-net trained by our proposed method outperformed standard U-net and U-net trained by the conventional knowledge distillation method by 3.52% and 0.96%. Furthermore, U-net trained by our proposed method improved the accu-

Table 1: Comparison our proposed method with baseline on ssTEM dataset. We denoted EfficientNet-b7 as EN-b7 and Knowledge distillation as KD. Standard U-net is used as the student model and U-net with EfficientNet-b7 backbone is used as the teacher models.

		IoU(%)					
	Method	membrane	mitochondria	synapse	Inner membrane	background	mIoU(%)
2 classes	U-net(EN-b7)	73.47	-	-	-	-	-
		-	84.35	-	-	-	-
		-	-	52.38	-	-	-
		-	-	-	69.06	-	-
		-	-	-	-	92.71	-
5 classes	U-net	69.82	78.12	48.45	64.54	91.37	70.46
	U-net(EN-b7)	72.26	83.87	49.26	69.67	92.04	73.42
	U-net + KD	71.55	81.83	47.90	71.79	92.01	73.02
	U-net + ours	72.14	83.93	50.48	70.79	92.54	73.98

Table 2: Comparison of different class-wise distillation weight parameters. We also use the common teacher model trained on all classes in this experiment. (a) is the result of learning without weights. (b) is the result of adjusting the weights so that the MSE loss for each class matches.

Method						mIoU(%)
student model : U-net						70.46
teacher model : U-net(EN-b7)						73.42
weight	λ_1	λ_2	λ_3	λ_4	λ_5	mIoU(%)
(a)	1	1.5	1	1	1	72.56
(b)	0.01	1.5	2.0	0.01	0.03	73.98

racy by 0.56% over 5 classes U-net(EN-b7) which is a teacher model.

We also confirmed that our proposed method is effective for classes which are difficult to inference, such as synapses. This results show that the proposed method is more effective than conventional knowledge distillation methods for classes that are difficult to infer.

Table 2 shows a comparison of the different weight parameters for class-wise distillation loss. Table 2 (a) shows the results of training without weights, and (b) shows the results of training with weights so that the MSE loss values for each class are about the same. (b) outperforms (a) by 1.42%. This is because the learning of classes with large losses interferes with the learning of classes with smaller losses. Therefore, for successful learning, it is necessary to select weights so that the distillation losses of all classes are about the same.

Figure 4 shows the qualitative segmentation results by each model. Figure shows that U-net, U-net enhanced with EfficientNet-b7, and U-net trained with vanilla knowledge distillation fail to correctly distinguish the mitochondrial area circled by the white dot lines. However, the proposed method allowed the student model to recognize mitochondria by transferring the knowledge from the teacher model specialized for it. These results on the ssTEM dataset show that, the proposed method transferred useful information from the teacher models specialized for each class to the student model, and achieved improved the accuracy with small model capacity described in Section 4.1.2.

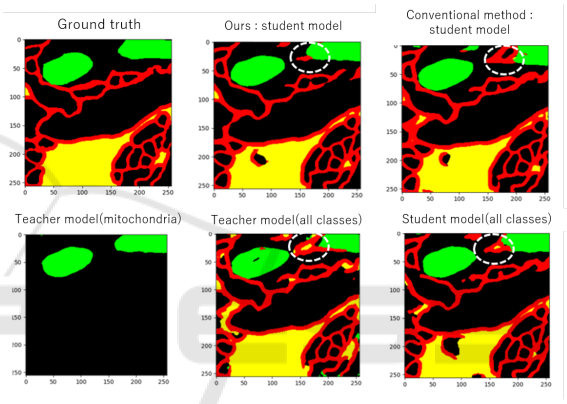


Figure 4: Comparison of our proposed method and baseline on ssTEM dataset. Student model trained with our proposed class-wise distillation method infer more accurately than those trained with conventional knowledge distillation methods.

4.2.2 Results on COVID-19 Dataset

Table 3 shows the results on the COVID-19 dataset. U-net trained with our distillation method outperformed the standard U-net and U-net trained with the conventional knowledge distillation by 5.16% and 1.30% in mIoU. Similarly with the experiment in previous section, U-net trained with our proposed method outperformed 4 classes U-net(EN-b7) which is a teacher model by 1.29% in mIoU. In ”consolidations” class, the most difficult class to distinguish,

Table 3: Comparison our proposed method and baseline on COVID-19 dataset.

		IoU(%)				
	Method	Background	Lungs other	Ground glass	Consolidations	mIoU(%)
2 classes	U-net(EN-b7)	93.14	-	-	-	-
		-	33.88	-	-	-
		-	-	49.22	-	-
		-	-	-	7.40	-
4 classes	U-net	92.41	30.18	41.23	2.44	41.57
	U-net(EN-b7)	95.04	34.37	46.45	5.86	45.43
	U-net + KD	96.11	37.83	47.72	0.00	45.42
	U-net + ours	95.38	37.06	48.13	6.30	46.72

the proposed method is able to inference with higher accuracy than the respective baselines. This results show that our proposed method is effective even for classes that are difficult to inference.

Experiments on two datasets demonstrated that our proposed class-wise distillation method is more effective in distilling knowledge for inference than conventional knowledge distillation methods.

5 CONCLUSIONS

In this paper, we proposed a new class-wise knowledge distillation method for multi-class semantic segmentation. Specifically, knowledge is transferred from teacher models specialized for each class to a student model. This enables better knowledge transmission than conventional knowledge distillation methods. By using this method, the student model achieved higher accuracy than the student model trained by conventional knowledge distillation methods.

In the future, we would like to make learning more effective without increasing computational resources.

ACKNOWLEDGEMENTS

This work is supported by SCAT Foundation and KAKENHI Grant Number 22H04735.

REFERENCES

Badrinarayanan, V., Kendall, A., and Cipolla, R. (2017). Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*.

Cao, Z., Hidalgo, G., Simon, T., Wei, S.-E., and Sheikh, Y. (2018). Openpose: realtime multi-person 2d pose estimation using part affinity fields. *arXiv preprint arXiv:1812.08008*.

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., and Yuille, A. L. (2017). Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*.

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., and Schiele, B. (2016). The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

Gerhard, S., Funke, J., Martel, J., Cardona, A., and Fetter, R. (2013). Segmented anisotropic sstem dataset of neural tissue. *figshare*.

Han, S., Mao, H., and Dally, W. J. (2015). Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. *arXiv preprint arXiv:1510.00149*.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

Hinton, G., Vinyals, O., Dean, J., et al. (2015). Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Hu, J., Shen, L., and Sun, G. (2018). Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

Hubara, I., Courbariaux, M., Soudry, D., El-Yaniv, R., and Bengio, Y. (2016). Binarized neural networks. *Advances in neural information processing systems*.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., and Berg, A. C. (2016). Ssd: Single shot multibox detector. In *European conference on computer vision*. Springer.

Liu, Y., Chen, K., Liu, C., Qin, Z., Luo, Z., and Wang, J. (2019). Structured knowledge distillation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical* *image computing and computer-assisted intervention*. Springer.

Tan, M. and Le, Q. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*. PMLR.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*.

Zhang, Y., Wang, C., Wang, X., Zeng, W., and Liu, W. (2021). Fairmot: On the fairness of detection and re-identification in multiple object tracking. *International Journal of Computer Vision*.

Zhao, H., Shi, J., Qi, X., Wang, X., and Jia, J. (2017). Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

Zhao, J., Zhang, Y., He, X., and Xie, P. (2020). Covid-ct-dataset: a ct scan dataset about covid-19. *arXiv preprint arXiv:2003.13865*.