# – Supplementary Material –
## Distilling Global and Local Logits with Densely Connected Relations

Youmin Kim[1,3*] Jinbae Park[1] YounHo Jang[1] Muhammad Ali[1] Tae-Hyun Oh[2] Sung-Ho Bae[1†]
[1]Kyung Hee University, [2]POSTECH, [3]Kakao Enterprise
{rladbals0733, qkrwlsqo94}@gmail.com, {2014104142, salmanali, shbae}@khu.ac.kr
taehyun@postech.ac.kr

## 1. Observations

We show the observation about the global and local logits and their relationships through the toy experiments shown in Figure 1 and 2 as suggested. The teacher and student are ResNet110 and ResNet20 [6], respectively. More detailed training settings are identical to Section 3.1.

Figure 1 shows the activation maps of the student with an image in CIFAR-100 [9] validation dataset where the distillation is performed on CIFAR-100 training dataset in three cases: (a) and (b) are trained with KL divergence for only global and local logits, respectively; (c) is trained with the relation distillation loss among the global and local logits. As shown in Figure 1, (a) only captures the global area around the wheel without the upper part of the vehicle separating the car and the truck. (b) captures the wheel and the upper part of the vehicle representing the truck, but also captures the top of the image, not related to the truck. (c) captures a broader coverage including the wheel and body regions together through their spatial relationship. Therefore, we use all of them to complement each other's shortcomings and fuse the strengths of each case for knowledge.

Figure 2 shows the class prediction of an image in CIFAR-100 validation dataset. We compare the class predictions of the network without distillation (baseline) and with our method using 4(2 x 2) local logits. The results of the class predictions show that ours has a higher chance to get more accurate local logits than the baseline as the bottom-right part is close to the turtle and the top-right/bottom-left part is closer to the shark due to the sharp object. Since our method transfers the spatial class information by the global and local logits and their relationships, it can be viewed that the distilled network with our method has an object localization ability even with only the classification task without object detection or segmentation tasks.

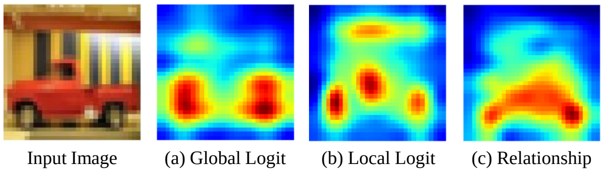
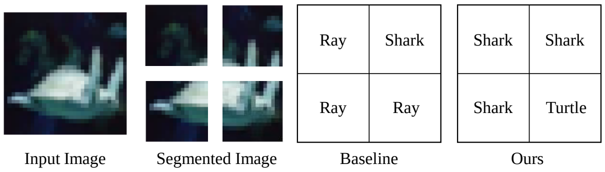


Figure 1. Activation map from the last feature of the student network after each distillation case. The map is computed by the max-pooling per each spatial position along with the channel-axis.



| Baseline | |
|---|---|
| Ray | Shark |
| Ray | Ray |

| Ours | |
|---|---|
| Shark | Shark |
| Shark | Turtle |

Figure 2. Class prediction by each local logit w.r.t a segmented part on a test sample of CIFAR-100 [9]. True label of the sample is Turtle.

## 2. Components and Computation Complexity

In this section, we verify the performance of each component of our method separately. The settings for teacher and student are the cases (a), (b) and (d) in Table 1 in the main paper. The training settings are identical to Section 3.1. Figure 3 shows the results, where the cases (a) and (b) show that the local logits play the most important role in increasing the performance. In case of (d), the combination of the global and local logits results in the highest performance among all the components. In addition, it shows that the combination of all the components (GLD) results in the highest performance. Through the results from the three cases, we verify the effectiveness of the local logits in our proposed method.

The computational complexity of our method is $O(B^2(L+1))$ where $L$ and $B$ are the number of local logits

[*] The work was done when Youmin Kim was at Kyung Hee University
[†] Corresponding author

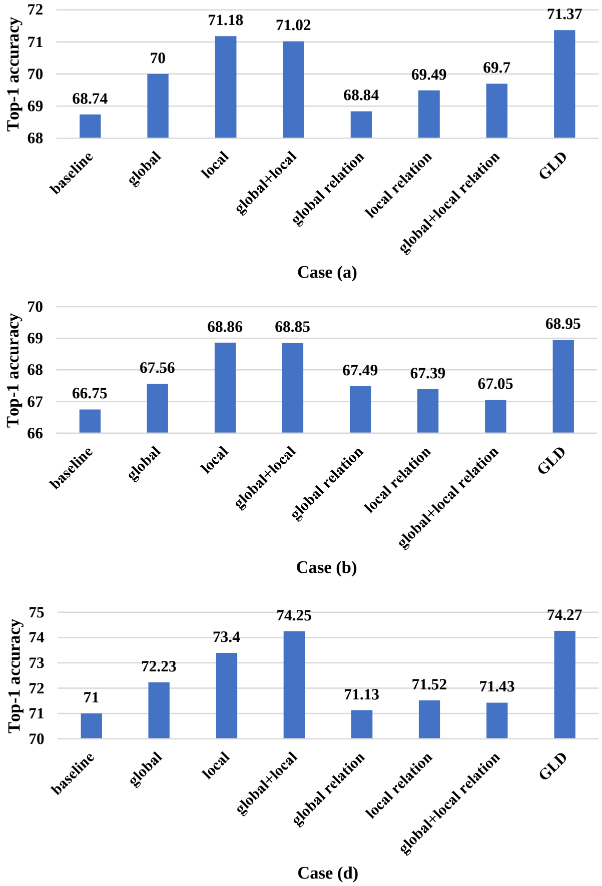


Figure 3. Top-1 accuracy (%) on CIFAR-100 [9] of each individual component. The results show that the local logits mostly contribute to the performance improvement, but the other components also have complementary information that can further improve.

and samples in a mini-batch, respectively. The relation- and non-relation-based distillation methods have $O(B^2)$ and $O(B)$, respectively.

## 3. Implementation details

In this section, we explain the training settings more specifically for the image classification, object detection and semantic segmentation tasks in the main paper. In CIFAR-100 [9], the settings for the teacher and student networks are identical to Table 1 in the main paper. In ImageNet [2] and the fine-grained datasets [12, 8, 7, 14], we choose ResNet34 and ResNet18 [6] as the teacher and student networks, respectively. In addition, the ResNet34 and ResNet18 are used as backbone networks for the object detection and semantic segmentation tasks.

### 3.1. Image Classification on Benchmark and Fine-grained Datasets

In the image classification experiments, we specify the training settings for the benckmark datasets (CIFAR-100 [9] and ImageNet [2]) and fine-grained datasets (Oxford 102 Flowers [12], Cars 196 [8], Standard Dogs [7] and CUB-200-2011 [14]).

CIFAR-100 is a 32 × 32 size RGB image dataset and consists of 100 classes. Each class has 500 training images and 100 test images. The training settings for the teacher and knowledge distillation experiments are as follows. We train all the networks for 200 epochs, with the batch size of 64. We use the SGD optimizer with the momentum of 0.9 and the weight decay of 0.0005. The learning rate starts from 0.1 and is decayed by a factor of 0.1 after each 100 and 150 epochs. ImageNet consists of 1.2 million RGB images for training and 50 thousand images for validation. Since the images in ImageNet have various sizes, both the training and validation images are cropped to $224 \times 224$ size. For the comparison with other state-of-the-art distillation methods, we follow the training settings suggested in [13].

In the fine-grained datasets, we fine-tune the teacher network which was pre-trained on the ImageNet dataset for each fine-grained dataset. In the distillation, the student network is trained with He initialization [5]. In the Oxford 102 Flowers and Cars 196, all weights in the teacher network are fine-tuned, while only the classifier in the teacher network is fine-tuned in the Stanford Dogs and CUB-200-2011. The training settings are as follows. We train all the networks for 200 epochs, with the batch size of 64. We use the SGD optimizer with the momentum of 0.9 and the weight decay of 0.0005. In the distillation, the learning rate starts from 0.05, and in the fine-tuning, it starts from 0.001, and is decayed by a factor of 0.1 after each 60, 120 and 180 epochs. The $\alpha$ and $\beta$ for the fine-grained datasets are identical to the setting of CIFAR-100 experiment in the main paper.

### 3.2. Object Detection and Sementic Segmentation

In the object detection experiments, we use Single Shot Detector (SSD) [11] as the detector with the size of input image as $300 \times 300$. We train the detector with the backbone for 250 epochs, with the batch size of 32. We use the SGD optimizer with the momentum of 0.9 and the weight decay of 0.0005. The learning rate warms up from 0 to 0.001 until 2 epochs and is decayed by a factor of 0.1 after each 150 and 200 epochs. We set the non-maximum suppression overlap and the confidence threshold values used in [11] as 0.45 and 0.01, respectively.

In the semantic segmentation experiments, we use DeepLabV3+[1] as the segmentation network with the size of input image as 513 × 513. We train the network with the backbone for 50 epochs in the PASCAL VOC2012 [3] and Semantic Boundaries dataset (SBD) [4], and 40 epochs in the COCO2017 segmentation [10] dataset. We use the batch size of 16 and the SGD optimizer with the momentum of 0.9 and the weight decay of 0.0005. The initial learning rate is 0.007 in the PASCAL VOC2012 and SBD, and 0.001 in COCO2017 segmentation dataset. We use the learning rate scheduler as the poly learning rate used in [15]. The power value in [15] is 0.9 in the PASCAL VOC2012, SBD and COCO2017 segmentation datasets.

## References

[1] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018.

[2] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[3] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, June 2010.

[4] Bharath Hariharan, Pablo Arbelaez, Lubomir Bourdev, Subhransu Maji, and Jitendra Malik. Semantic contours from inverse detectors. In *International Conference on Computer Vision (ICCV)*, 2011.

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[7] Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Li Fei-Fei. Novel dataset for fine-grained image categorization. In *First Workshop on Fine-Grained Visual Categorization, IEEE Conference on Computer Vision and Pattern Recognition*, Colorado Springs, CO, June 2011.

[8] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *4th International IEEE Workshop on 3D Representation and Recognition (3dRR-13)*, Sydney, Australia, 2013.

[9] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[10] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[11] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016.

[12] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Indian Conference on Computer Vision, Graphics and Image Processing*, Dec 2008.

[13] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. *arXiv preprint arXiv:1910.10699*, 2019.

[14] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

[15] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017.