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

Abstract

In prevalent knowledge distillation, logits in most image recognition models are computed by global average pooling, then used to learn to encode the high-level and task-relevant knowledge. In this work, we solve the limitation of this global logit transfer in this distillation context. We point out that it prevents the transfer of informative spatial information, which provides localized knowledge as well as rich relational information across contexts of an input scene. To exploit the rich spatial information, we propose a simple yet effective logit distillation approach. We add a local spatial pooling layer branch to the penultimate layer, thereby our method extends the standard logit distillation and enables learning of both finely-localized knowledge and holistic representation. Our proposed method shows favorable accuracy improvement against the state-of-the-art methods on several image classification datasets. We show that our distilled students trained on the image classification task can be successfully leveraged for object detection and semantic segmentation tasks; this result demonstrates our method's high transferability.

1. Introduction

Knowledge distillation is a method of transferring knowledge of a large network (i.e., teacher) to a smaller neural network (i.e., student). Unlike human-designed prior knowledge, the distillation is an optimization method that uses the representation of the network as prior knowledge. More specifically, the student is trained with respect to reducing a task-specific objective function, and the difference in knowledge from the teacher.

Due to the simplicity and effectiveness of targeting the teacher that has higher accuracy than the student, many researchers have used the distillation method to achieve the state-of-the-art accuracy on ImageNet dataset [11] of the network [53, 43]. For the same reason, the distillation has been combined with model-compression methods such as pruning [31, 37] and quantization [25, 4], or other optimization methods such as data augmentation [51] and ensemble [58]. In addition, the distillation is used not only for the image classification tasks, but also other vision tasks such as image super-resolution [17, 30], object detection [7, 49] and semantic segmentation [52, 36].

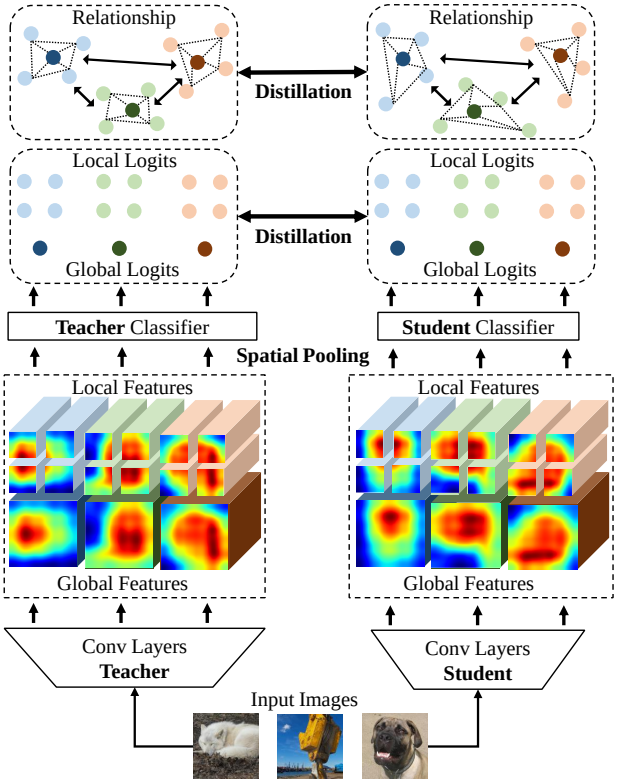

Figure 1. Overview of our method. The global and local logits and their densely connected relationships are used for the logit distillation. "Conv Layers" denote convolutional layers of teacher and student networks.

Depending on the representation levels of features to be transferred, knowledge distillation methods are divided into

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

two types: i) feature distillation that exploits the output features of intermediate convolutional layers [44, 55, 56, 23, 26, 19, 1, 39, 35, 18, 46, 45]; and ii) logit distillation that exploits the output logit in the final classifier [2, 20].

The logit is high-level and task-relevant knowledge with class information but loses spatial information due to the global average pooling, which spatially averages on the features of the last convolutional layer (i.e., penultimate layer). Most existing deep neural networks use global average pooling [32] for the object classification task because the pooling both significantly reduces the model parameters and prevents overfitting of the network while retaining the network's invariance to bounded spatial variants (e.g., translation, rotation, flipping) of the input data. However, in the logit distillation, the student only learns compressed knowledge, which includes no spatial information of the input data from the teacher.

Many studies have shown that spatial information [29, 15, 40, 5] and spatial relationships [21, 50, 22, 54] are essential factors for performance improvements in various computer vision tasks. To exploit both spatial information and spatial relationships for logit distillation, we propose a novel global and local logit distillation method (GLD) that transfers not only the global and local logits but also the relationships among the global and local logits of multiple input samples from the teacher to the student.

Figure 1 conceptualizes the global and local logits with their densely connected relationship from the teacher and student. Through the spatial pooling strategy in [15], we create the global logits from the features in the whole region (global features) and the local logits from local regions (local features) in the penultimate layer. Furthermore, the densely connected relationship consists of the global and local logits from all the input samples in a mini-batch. Specifically, our relationship is defined with global and local logits not only within the one input sample (intra-relationship) but also among all input samples (inter-relationship).

Therefore, the student can learn spatial class information composed of the global and local logits from the teacher. In addition, the student can learn not only the relationships among the global and local representations for one sample by the intra-relationship but also more detailed relationship among the all input samples in a mini-batch by the inter-relationship, densely connected through the global and local logits of each sample. The contributions of this work are summarized as follows:

- We propose a novel logit-distillation method that uses the global and local logits and their relationships within a single sample as well as among all samples in a mini-batch as knowledge.
- When using Kullback-Leibler (KL) divergence as knowledge distillation loss, we accommodate various distributions of both global and local logits by using the standard deviation of a logit as a softening factor.
- We validate the generalizability of our method on the image classification with benchmark datasets and its transferability to the object detection and semantic segmentation with various datasets.

2. Related Work

Knowledge distillation requires a method to convert the output feature or logit of the teacher to knowledge that the student can learn easily, considering the difference in capacity between the teacher and student. Various methods for this purpose have been studied; they can be categorized into two types: feature distillation and logit distillation.

In *feature distillation*, Romero *et al.* [44] improve the student network's performance by element-wise minimization of the difference between the respective output features of the teacher and the student through the regressor. Zagoruyko and Komodakis [56] use a spatial attention map of features to help the student learn where the teacher focuses on the input data. Unlike this method that uses the attention map to focus on the specific local regions in the whole region of the features, our method considers both the whole and local regions of the features regardless of the attention. Yim *et al.* [55] use the gram matrix for the input and output features of the intermediate layers in a whole network to consider the flow of solution procedure. Huang and Wang [23] use a kernel trick to reduce the difference of the feature distributions between the teacher and student. However, the gram matrix or the kernel trick is used to create knowledge only for the whole region of the features, whereas our method creates knowledge for the local regions as well as the whole region of the features. Kim *et al.* [26] encode the output features of the last convolutional layer using an auto-encoder and distill the core (encoded) information of the features. We use the logit encoded through the classifier for the features as core information.

Heo *et al.* [19] consider the activation boundary, i.e., the sign of the feature is used as knowledge rather than the value itself. Ahn *et al.* [1] formulate the knowledge distillation problem as mutual information maximization between the teacher and student. Heo *et al.* [18] perform distillation by filtering out redundant information, which can have adverse effects on the student. For this, they use a margin ReLU and a partial-L_2 distance loss for the output of all the batch normalization layers in the teacher. Park *et al.* [39] use euclidean distance and angle matrix among the global average pooled features for the inter-relationship. Liu *et al.* [35] use L_2 distance matrix between the relationship of the input features and the relationship of the output features in each layer of the network. Tung and Mori [46] use the covariance-based matrix among the one-dimensional flatten

feature for the inter-relationship. Peng *et al.* [42] use the kernel-based correlation among the embedded features for the inter-relationship. All of these methods [39, 35, 46, 42] use the relationship among the samples in the feature space regardless of the spatial information; our method differs from those in that it uses the global and local logits in the logit space to extract the spatial relationships for each sample, in addition to among the samples. Tian *et al.* [45] use mutual information to encourage that the teacher and student have the similar output features for the similar input samples (positive pair), while increasing the distance between the two features obtained from two different input samples (negative pair). Instead, we do not distinguish between positive and negative samples, and encourage the student to mimic all the logit behaviors of the teacher, including global and local logits and their relationships.

In *logit distillation*, Ba and Caruana [2] minimize the L_2 loss between the logits from the teacher and student. Hinton *et al.* [20] transfer the softened distribution of the softmax output as the knowledge to the student by dividing the logit values by a fixed *temperature* value; this approach allows the student to learn the true label class as well as the other classes. Our method uses both the global logits used in [20] and the local logits, so we normalize the logit to accommodate various distributions of global and local logits.

3. Method

Notation. We denote a convolutional layer of a network as $f = F(\cdot\,; W^F)$, and a classifier of a network as $z = C(\cdot\,; W^C)$, where f is the output feature of the last convolutional layer and z is a logit, and W is a set of trainable weights in a layer. The global average pooling [32] is denoted as $\texttt{GAP}(\cdot)$ and its receptive field size is the whole spatial region of the arbitrarily-sized input [15]. The input samples in a mini-batch is defined as $X = \{x_1, x_2, x_3, ..., x_n\}$ where the corresponding true label data is $Y = \{y_1, y_2, y_3, ..., y_n\}$ for a total of n samples in a mini-batch. Therefore, the last feature can be expressed as $f = F(x\,; W^F)$ and logit as $z = C(\texttt{GAP}(f)\,; W^C)$. Specifically, in the context of distillation, the teacher's last feature is $f_t = F_t(x\,; W_t^F)$ and logit is $z_t = C_t(\texttt{GAP}(f_t)\,; W_t^C)$, and the student's last feature is $f_s = F_s(x\,; W_s^F)$ and logit is $z_s = C_s(\texttt{GAP}(f_s)\,; W_s^C)$.

3.1. Global and Local Logits

We explain the global and local logits used in our method (Figure 1). The global logit is the final output of a classifier that takes the input as the global feature of the network. The global logit has been used in conventional neural networks [20]. During traditional logit distillation, only the global logit is transferred to the student; the spatial information is ignored due to the global average pooling [32].

To compensate for the loss of spatial information, we introduce "local" logits, which are the output of the classifier that takes the input of local features divided from the global feature. We share the same classifier to generate both local and global logits in a network.

The global and local logits are created by the classifier which takes $1 \times 1 \times c$ vectors spatially averaged from the global and local features where c is the number of input channels (Figure 1). The global and local features are denoted as f^0 and f^l ($l = 1, 2, ..., L$) where L is the number of local features, respectively. The local features are derived by dividing the width and height of f^0 by d where no overlapping among local features is applied, so $L = d^2$. The receptive field size of $\texttt{GAP}()$ for the global feature is $w \times h$, so the $\texttt{GAP}()$ for the local features takes the receptive field size of $\lfloor \frac{w}{d} \rfloor \times \lfloor \frac{h}{d} \rfloor$. Finally, the global logit z^0 and the l-th local logit z^l can be obtained as follows:

$$z^0 = C(\texttt{GAP}(f^0)\,; W^C), \;\; z^l = C(\texttt{GAP}(f^l)\,; W^C). \quad (1)$$

Hinton *et al.* [20] transfer the sufficient information for the true label class as well as the other classes from the teacher to the student via logits. The softmax function creates a peaky probability distribution, so a *temperature* parameter in [20] is used as a softening factor to produce a smooth probability distribution. After the softening process, distillation is performed by reducing the KL divergence between the two distributions of the teacher and student during the training phase. The distillation loss L_{hinton} [20] between all logits in a mini-batch from the teacher (Z_t) and student (Z_s) is defined as:

$$L_{hinton}(Z_t,\, Z_s) = \frac{1}{n}\sum_{i=1}^{n} \tau^2 KL(\psi(\frac{z_{t,i}}{\tau}),\, \psi(\frac{z_{s,i}}{\tau})), \quad (2)$$

$$\psi(z)_j = \frac{e^{z_j}}{\sum_{k=1}^{K} e^{z_k}},\; for \;\; j = 1, 2, ..., K.$$

where K is the number of classes and $\psi(\,\cdot\,)_j$ is a softmax function for j-th class, $z_{t,i}$ and $z_{s,i}$ are the output logits of the i-th input from the teacher and student, respectively.

The proposed logit distillation method exploits both global and local logits, which have different statistical characteristics, so we propose an adaptive distillation loss L_{ND}, which is derived from the KL divergence with normalized logits. L_{ND} is derived by replacing z/τ in Eq. 2 with nor-

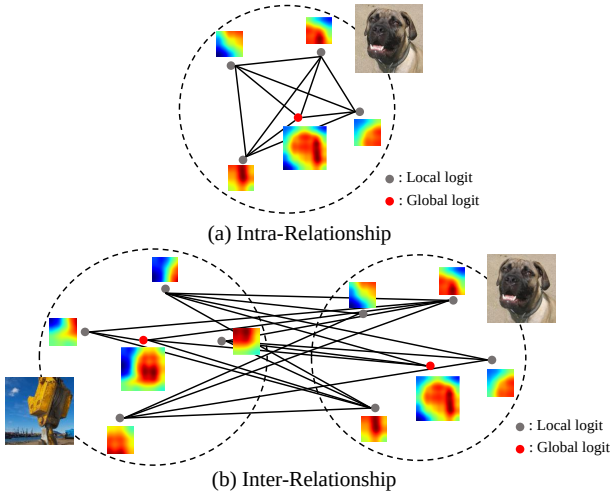

Figure 2. The examples of the proposed densely connected relationships. The bold-line in (a) denotes the intra-relationship within one sample and the bold-line in (b) denotes the inter-relationship between two samples.

malized logits as:

$$
\begin{aligned}
& L_{ND}(Z_t,\, Z_s) \qquad (3)\\
&= \frac{1}{n}\sum_{i=1}^{n} KL(\psi(\frac{z_{t,i}-\mu(z_{t,i})}{\sigma(z_{t,i})}),\, \psi(\frac{z_{s,i}-\mu(z_{s,i})}{\sigma(z_{s,i})}))\\
&= \frac{1}{n}\sum_{i=1}^{n} KL(\psi(\frac{z_{t,i}}{\sigma(z_{t,i})}-\frac{\mu(z_{t,i})}{\sigma(z_{t,i})}),\, \psi(\frac{z_{s,i}}{\sigma(z_{s,i})}-\frac{\mu(z_{s,i})}{\sigma(z_{s,i})}))\\
&= \frac{1}{n}\sum_{i=1}^{n} KL(\psi(\frac{z_{t,i}}{\sigma(z_{t,i})}),\, \psi(\frac{z_{s,i}}{\sigma(z_{s,i})})),
\end{aligned}
$$

where $\mu(\cdot)$ is the mean of the input logit and and $\sigma(\cdot)$ is its standard deviation. As a result, our L_{ND} improves the distillation performance by accommodating different statistical characteristics of the both global and local logits in the KL divergence. The effectiveness of L_{ND} in the proposed GLD is verified in Section 4.3.1.

Combining Eqs. 1 and 3 yields the global logit distillation loss L_{global} and local logit distillation loss L_{local} as:

$$
L_{global} = L_{ND}(Z_t^0,\, Z_s^0),\;\; L_{local} = \sum_{l=1}^{L} L_{ND}(Z_t^l,\, Z_s^l), \qquad (4)
$$

where Z^0 and Z^l are a set of the global logits and l-th local logits from all the samples in a mini-batch, respectively. By using L_{global} and L_{local}, the student can learn more spatial class information from the teacher than by using the previous logit distillation methods. Observations through toy experiments about the global and local logits and their relationships in the proposed method are given in the supplementary material.

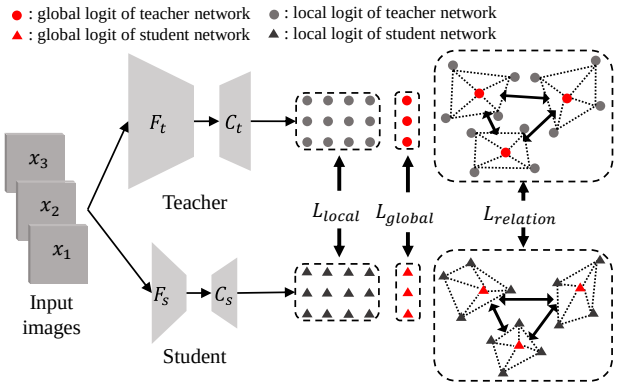

Figure 3. Loss components L_{local}, L_{global} and $L_{relation}$ for the proposed logit distillation method between teacher and student networks.

3.2. Densely Connected Relationship

Our method focuses on identifying what information should be contained in the relational knowledge to be transferred, unlike other methods [39, 46, 42], where they focus on how to form it. Specifically, our method transfers the spatial relationships through the simple L_2 distance matrix among the global and local logits for each sample (see Intra-relationship of Figure 2a) and across mini-batch samples (see Inter-relationship of Figure 2b). In practice, the intra- and inter-relationships are unified in the form of an $m \times m$ matrix where $m = n(L+1)$. By using these two types of relationship, our method can transfer more densely connected relational knowledge than other methods [39, 46, 42]; thus, our method can be regarded as a more general method than these methods, where they use only the inter-relationship among the features without the spatial information, i.e., special cases of ours.

The proposed intra-relationship between the global and local logits can be interpreted as capturing the co-activating behaviors, i.e., a local logit dominates the global statistic if the global and local logits are strongly related. This relationship also coincides with the local bias property of a convolutional neural network (CNN), i.e., a CNN is biased to specific local features that contribute to its final prediction [6, 3, 13, 48]. Similarly, the co-activating behaviors among the local logits can be captured as knowledge for distillation in the proposed method.

To measure the relationship, we use the L_2-based distance metric as:

$$
\begin{aligned}
& \tilde{D}(Z)_{i1,i2} = \|z_{i1} - z_{i2}\|^2,\;\; (z_{i1},\, z_{i2}) \in Z, \qquad (5)\\
& Z = Z^0 \cup Z^1 \cdots \cup Z^l \cdots \cup Z^L,
\end{aligned}
$$

Setup	Compression type	Teacher network	Student network	# of params teacher	# of params student	Compress ratio
(a)	Depth	ResNet 110	ResNet 20	1.7M	0.27M	15.9%
(b)	Channel	WideResNet 16-2	WideResNet 16-1	0.7M	0.18M	25.7%
(c)	Depth & channel	WideResNet 22-4	WideResNet 16-2	4.32M	0.7M	16.2%
(d)	Different architecture	WideResNet 16-4	ResNet 32	2.73M	0.46M	16.8%

Table 1. Information for settings on the teacher and student network architectures on CIFAR-100 dataset. The architecture is denoted as ResNet (depth) for ResNet [16] and WideResNet (depth)-(channel multiplication) for WideResNet [57].

where Z collects the global and local logits obtained from all the samples in a mini-batch, a union set of Z^0 and Z^l ($l = 1, 2, ..., L$). Collecting all the distance values measured with Eq. 5 for Z, we construct the distance matrix where L_2 normalization is performed for each row as:

$$D(Z)_{[r,c]} = \frac{\tilde{D}(Z)_{[r,c]}}{\|\tilde{D}(Z)_{[r,:]}\|_2}, \tag{6}$$

where $D(Z)_{[r,c]}$ indicates the $(r,c)^{th}$ element of the normalized distance matrix $D(Z)$. To transfer the relational knowledge from the teacher to student, we define a relation distillation loss $L_{relation}$ as:

$$L_{relation} = \frac{1}{m^2}\sum_{r=1}^{m}\sum_{c=1}^{m}\|D_t(Z_t)_{r,c} - D_s(Z_s)_{r,c}\|^2. \tag{7}$$

Loss components L_{global}, L_{local} and $L_{relation}$ for the proposed distillation method are calculated by comparing the logits of the teacher to the logits of the student (Figure 3). The final loss (L_{GLD}) of our method combines all the distillation losses with a task loss (L_{task}), which is the cross entropy between the network output and the true label if exists. Extending the conventional loss form of $L_{distill} = (1-\alpha)L_{task} + \alpha L_{global}$ in [20], we propose a new loss L_{GLD} by considering the local and relation losses as:

$$L_{GLD} = L_{distill} + L_{local} + \beta L_{relation}, \tag{8}$$

where α, β are balancing hyperparameters.

4. Experiments

Setup. We verify the effectiveness of our method through extensive experiments comparing with nine state-of-the-art knowledge distillation methods (i.e., KD [20], NST [23], AT [56], OD [18], RKD [39], IRG [35], CC [42], SP [46], CRD [45]) in the image classification task. The datasets used for the experiments are CIFAR-100 [28] and ImageNet [11]. We also show the transferability of the representation from the distilled student network. For verifying the transferability, CINIC-10 [10] and STL-10 [9] datasets are used as target datasets for the image classification task. For object detection and semantic segmentation, Pascal VOC2007 and VOC2012 [12], COCO dataset [33], and Semantic Boundaries dataset (SBD) [14] are used as target datasets. We also perform ablation studies to see the effectiveness of the proposed components. We use Pytorch library [41] for implementation. L is set to 4 for all experiments by default except for the setting of Table 8.

4.1. Image Classification

First, we compare our method with other knowledge distillation methods in the image classification task. The network trained for the image classification task is used as a backbone for other computer vision tasks such as object detection [34] and semantic segmentation [8], so image classification is considered a representative task. As in [20], the teacher is pre-trained and the student is randomly initialized before being used for the knowledge distillation.

4.1.1 CIFAR-100

We set four teacher-student cases using variants of ResNet [16] and WideResNet [57] to show the superiority of our method for various architectures of the teacher and student. Table 1 shows the experimental setups for four compression types with different combination of teacher/student network architectures. We set the α and β in Eq. 8 as 0.7 and 500, respectively. The hyperparameter values for other distillation methods are obtained from their respective papers. More detailed experimental settings are described in the supplementary material.

Table 2 shows the experimental results for all the cases in Table 1. Our method outperforms the state-of-the-art distillation methods for all the cases. This result indicates that the global and local logits with their densely connected relations play a crucial role in distilling the knowledge with high generalizability from the teacher to the student, even for classifying small image objects.

Setup	Teacher	Baseline	KD [20]	NST [23]	AT [56]	OD [18]	RKD [39]	IRG [35]	CC [42]	SP [46]	CRD [45]	GLD (ours)
(a)	72.53	68.75	70.27	69.13	69.73	69.16	69.36	69.87	68.85	70.29	71.10	**71.37**
(b)	73.04	66.76	68.06	66.70	67.49	67.40	67.41	68.03	66.82	67.61	67.90	**68.95**
(c)	77.67	73.04	74.75	73.40	74.36	76.07	73.34	75.17	72.96	74.23	75.55	**76.28**
(d)	76.67	71.01	73.03	70.51	72.23	73.68	71.11	73.26	70.99	72.13	73.65	**74.27**

Table 2. Top-1 accuracy (%) on CIFAR-100 validation dataset compared with various knowledge distillation methods. "Baseline" represents a result without distillation. The accuracy is averaged over three runs.

	Teacher	Baseline	AT [56]	KD [20]	OD* [18]	RKD* [39]	IRG* [35]	SP* [46]	CC [42]	Online KD [58]	CRD [45]	CRD+KD [45, 20]	GLD (ours)
Top-1	73.31	69.75	70.70	70.66	70.59	70.59	70.32	70.79	69.96	70.55	71.17	71.38	**71.63**
Top-5	91.42	89.07	90.00	89.88	89.81	89.68	89.99	89.80	89.17	89.59	90.13	90.49	**90.53**

Table 3. Top-1 and Top-5 accuracy (%) on ImageNet validation dataset compared with various knowledge distillation methods. The results of the other methods except "*" are quoted from [45]. Our method and "*" are performed with our implementation with the same training settings in [45]. "Baseline" denotes a result without distillation.

4.1.2 ImageNet

ImageNet classification is considered a difficult task because the network needs to classify each input image into one of 1000 classes. In this experiment, we choose ResNet34 and ResNet18 [16] as the teacher network and student network, respectively. For comparison with other state-of-the-art distillation methods, the training settings are same as [45]. We set α = 0.1 and β = 100. The hyperparameters of other methods follow their respective papers.

Table 3 shows the experimental results of the proposed method on ImageNet. Our method outperforms all other distillation methods, increasing the top-1 and top-5 accuracy for validation images by 1.88% and 1.46% over the baseline, respectively. This result supports that our method works well for the difficult image classification task with large image objects.

4.2. Transferability

To further verify the effectiveness of the proposed method, we show the transferability of our method in this section. Distilled networks tend to have higher generalization performance as compared to baseline in new datasets or tasks by learning more general knowledge through distillation methods [45]. To verify the transferability of our method for image classification tasks, we experiment with linear classification tasks in new datasets [10, 9] by using the distilled student on the original dataset used in the distillation process. Furthermore, for the object detection and semantic segmentation tasks, we replace backbone networks with the distilled network.

4.2.1 Other Datasets

In this section, we show the transferability of our method in experiments with CINIC-10 [10] and STL-10 [9] datasets, on which it had not been trained. CINIC-10 is 32×32 size RGB image dataset obtained from CIFAR-10 [28] and ImageNet [11]; it consists of 10 classes, and each class has 9,000 images for each training, validation, and test. CINIC-10 is a mixture of two heterogeneous datasets, so its data distribution is different from the source dataset (CIFAR-100 [28]). STL-10 consists of a labeled and an unlabeled image dataset constructed from the ImageNet dataset. We use the labeled dataset for transferability experiments; it is divided into ten classes with an image resolution of 96×96. Each class has 500 training images and 800 test images.

We evaluate the transferability of the distilled students for the case in which the teacher network is WideResNet16-2 and the student network is WideResNet16-1 (Table 1), respectively. We freeze the weights in the convolutional layers of the distilled student on CIFAR-100 and only train the classifier to adjust CINIC-10 and STL-10 image classification tasks as in [45]. The training settings are the same as in Section 4.1.1. Table 4 shows the experimental results on CINIC-10 and STL-10, the two target datasets. Our method shows higher transferability than other state-of-the-art distillation methods.

4.2.2 Object Detection and Semantic Segmentation

In this section, we show the transferability of the distilled student on object detection and semantic segmentation tasks. We use the distilled student (ResNet18) on ImageNet. For object detection and semantic segmentation, the student

	CIFAR-100 [28] → CINIC-10 [10]	CIFAR-100 [28] → STL-10 [9]
Teacher	62.15	70.65
Baseline	58.36	67.61
KD [20]	60.17	67.23
NST [23]	58.52	66.65
AT [56]	59.93	67.06
OD [18]	59.93	68.61
RKD [39]	60.09	67.58
IRG [35]	60.44	67.75
CC [42]	58.52	66.25
SP [46]	60.18	67.47
CRD [45]	60.90	68.28
GLD (ours)	**61.09**	**68.96**

Table 4. Top-1 accuracy (%) on CINIC-10 and STL-10 test datasets compared with various knowledge distillation methods. "Baseline" represents a result without distillation.

is used as the backbone network. More detailed experimental settings are described in the supplementary material.

In the object detection task, Single Shot Detector (SSD) [34] is used as the baseline detector. First, we measure the performance on the test dataset of Pascal VOC2007 [12] for the student trained with Pascal VOC2012 as the training dataset and VOC2007 as the validation dataset. Second, we measure the performance on the validation dataset of COCO2017 dataset detection [33] after training the detector with the training dataset of COCO2017.

In the semantic segmentation task, DeepLabV3+ (DLV3+) [8] is used as a segmentation network for two datasets. We measure the performance on the test dataset of Pascal VOC2007 segmentation dataset, with Pascal VOC2012 and Semantic Boundaries Dataset (SBD) [14] as training datasets. We also measure the performance on the validation dataset of the COCO2017 segmentation dataset by training the network with the training dataset in the same dataset. Table 5 and 6 show the experimental results for object detection and semantic segmentation, respectively. These results indicate that the proposed distillation can transfer knowledge with high generalizability in feature representation.

4.3. Ablation Study

In this section, we conduct experiments on each component of our proposed method and the number of the local logits with various hyperparameter settings. The training settings for CIFAR-100 [28] and ImageNet [11] are the same as in Section 4.1.1 and 4.1.2, respectively. Top-1 accuracy on CIFAR-100 is averaged over three runs.

Network (# of params)	Method	VOC2007 test set [12]	COCO2017 val set [33]
SSD-ResNet34 (31.68M)	Teacher	75.23	24.28
SSD-ResNet18 (21.57M)	Baseline	70.97	19.41
	OD [18]	71.55	19.72
	SP [46]	71.12	19.51
	GLD (Ours)	**71.83**	**19.83**

Table 5. Object detection results with mean Average Precision (mAP) on the test data of Pascal VOC2007 and COCO2017 detection dataset. "Baseline" represents a result without distillation.

Network (# of params)	Method	VOC2007 test set [12]	COCO2017 val set [33]
DLV3+-ResNet34 (26.72M)	Teacher	72.72	56.24
DLV3+-ResNet18 (16.61M)	Baseline	69.72	51.22
	OD [18]	70.14	53.68
	SP [46]	70.57	51.33
	GLD (Ours)	**70.77**	**53.75**

Table 6. Semantic segmentation results with mean Intersection of over Union (mIoU) on test data of Pascal VOC2007 and COCO2017 segmentation dataset. "Baseline" represents a result without distillation.

4.3.1 Components of GLD

To verify the performance of each component of our method, we experiment with each component of our method separately. The settings for teacher and student are identical to case (c) in Table 1: results of the case (a), (b) and (d) in Table 1 are described in the supplementary material. As shown in Figure 4, among all the components, the local logits have the strongest influence on the increase in the performance, and the combination of all the components (GLD) yields the highest performance.

To verify the effectiveness of using the standard deviation as a softening factor for the global and local logits in L_{ND} (Eq. 3), we perform experiments for all cases in Table 1, but with the softening factor either a fixed value (temperature) or the standard deviation for the conventional (global logit only) and our (global and local logits together) cases. In this experiment, $L_{relation}$ in our method is removed for the distillation. Table 7 shows the experimental results. For the conventional case, use of temperature achieve higher accuracy than use of standard deviation, whereas for our case, use of the standard deviation achieve higher accuracy than use of the temperature. These results indicate that standard

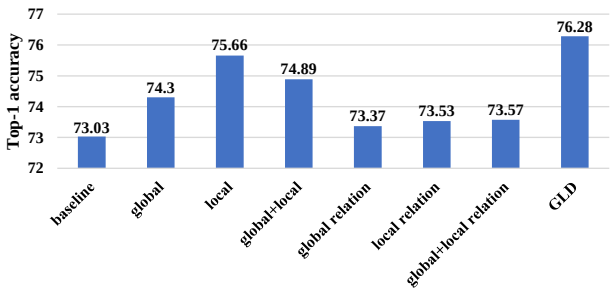

Figure 4. Top-1 accuracy (%) on CIFAR-100 [28] for each individual component. The local logits are the most effective for our method among the other separated components.

Setup		(a)	(b)	(c)	(d)
$\tau = 4$	global	**70.26**	**68.05**	**74.75**	**73.03**
σ	global	70.00	67.56	74.30	72.23
$\tau = 4$	global+local	69.77	67.44	**75.06**	72.50
σ	global+local	**71.02**	**68.85**	74.89	**74.25**

Table 7. Top-1 accuracy (%) on CIFAR-100 [28] of our method. The standard deviation (σ, ours) is more effective for the softening factor than the temperature (τ, [20]).

deviation can effectively accommodate different distributions from global and local logits.

4.3.2 Number of Local Logits and Hyperparameters

To determine the optimal number of local logits, we perform experiments according to the number of the local logits in our method, where we set to 4 (2×2), 9 (3×3) and 49 (7×7) for the fine-grained datasets (Oxford 102 Flowers [38], Cars 196 [27], Stanford Dogs [24] and CUB-200-2011 [47]) and ImageNet [11] and to 4 (2×2), 16 (4×4) and 64 (8×8) for CIFAR-100 [28], respectively. The detail settings are described in the supplementary material. Table 8 shows the tendency that increases in the importance of the locality of an object for classification (e.g., fine-grained classification) requires to increase in the number of the local logits/relations. These results suggest that classifying sub-categories of objects usually requires focusing more on local textures than global shapes of objects, where a large number of local logits of our method can extract more detailed local information.

We perform experiments to see the sensitivity of our method to hyperparameters α and β which control the conventional loss $L_{distill}$ and the relation loss $L_{relation}$, respectively. The settings for teacher and student are those of case (c) in Table 1. Table 9 shows the experimental results that our method is somewhat insensitive to the hyperparameters α and β.

	Teacher	Baseline	Local logits 2×2 (2×2)	3×3 (4×4)	7×7 (8×8)
102 Flowers	98.53	91.20	95.48	95.48	**96.82**
Cars 196	86.86	80.61	87.71	88.32	**88.77**
120 Dogs	86.86	65.69	79.07	79.28	**79.60**
CUB-200-2011	63.23	55.91	68.89	70.24	**70.81**
ImageNet	73.31	69.75	**71.63**	71.33	69.95
CIFAR-100 (a)	72.53	68.75	**71.37**	69.85	68.80
CIFAR-100 (b)	73.04	66.76	**68.95**	67.52	65.66
CIFAR-100 (c)	77.67	73.04	**76.28**	76.05	74.94
CIFAR-100 (d)	76.67	71.01	**74.27**	73.85	73.21

Table 8. Top-1 accuracy (%) on the fine-grained and benchmark datasets for the various number of the local logits. "Baseline" represents a result without distillation.

β \ α	0.1	0.3	0.5	0.7
100	76.00	76.05	76.02	76.27
500	76.02	76.03	76.19	**76.28**
1000	76.06	76.07	76.25	76.07

Table 9. Top-1 accuracy (%) on CIFAR-100 [28] for the α and β in our method. The results are stable in any case.

5. Conclusion

In this paper, we propose a novel logit distillation method, GLD, which uses both global and local logits and their densely connected relationship for distillation by exploiting the spatial information for the logits. We validated the effectiveness of our method in various experiments and achieved high performance in image classification, object detection and semantic segmentation. Since GLD can generate a variety of local class information depending on the spatial pooling strategy and provide how it is applied to the features of the penultimate layer, we plan to elaborate GLD by utilizing various spatial pooling strategies as future work.

Acknowledgement

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT & Future Planning (2018R1C1B3008159). Tae-Hyun Oh was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. NRF-2020R1F1A107517611). The authors are very grateful for Eunseop Shin for technical support.

References

[1] Sungsoo Ahn, Shell Xu Hu, Andreas Damianou, Neil D Lawrence, and Zhenwen Dai. Variational information distillation for knowledge transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9163–9171, 2019.

[2] Lei Jimmy Ba and Rich Caruana. Do deep nets really need to be deep? *arXiv preprint arXiv:1312.6184*, 2013.

[3] Nicholas Baker, Hongjing Lu, Gennady Erlikhman, and Philip J Kellman. Deep convolutional networks do not classify based on global object shape. *PLoS computational biology*, 14(12):e1006613, 2018.

[4] Yoonho Boo, Sungho Shin, Jungwook Choi, and Wonyong Sung. Stochastic precision ensemble: Self-knowledge distillation for quantized deep neural networks. *arXiv preprint arXiv:2009.14502*, 2020.

[5] Wieland Brendel and Matthias Bethge. Approximating cnns with bag-of-local-features models works surprisingly well on imagenet. *arXiv preprint arXiv:1904.00760*, 2019.

[6] Tom B Brown, Dandelion Mané, Aurko Roy, Martín Abadi, and Justin Gilmer. Adversarial patch. *arXiv preprint arXiv:1712.09665*, 2017.

[7] Guobin Chen, Wongun Choi, Xiang Yu, Tony Han, and Manmohan Chandraker. Learning efficient object detection models with knowledge distillation. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 742–751, 2017.

[8] 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.

[9] Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 215–223, 2011.

[10] Luke N Darlow, Elliot J Crowley, Antreas Antoniou, and Amos J Storkey. Cinic-10 is not imagenet or cifar-10. *arXiv preprint arXiv:1810.03505*, 2018.

[11] 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.

[12] 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.

[13] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*, 2018.

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

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 37(9):1904–1916, 2015.

[16] 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.

[17] Zibin He, Tao Dai, Jian Lu, Yong Jiang, and Shu-Tao Xia. Fakd: Feature-affinity based knowledge distillation for efficient image super-resolution. In *2020 IEEE International Conference on Image Processing (ICIP)*, pages 518–522. IEEE, 2020.

[18] Byeongho Heo, Jeesoo Kim, Sangdoo Yun, Hyojin Park, Nojun Kwak, and Jin Young Choi. A comprehensive overhaul of feature distillation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1921–1930, 2019.

[19] Byeongho Heo, Minsik Lee, Sangdoo Yun, and Jin Young Choi. Knowledge transfer via distillation of activation boundaries formed by hidden neurons. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3779–3787, 2019.

[20] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[21] Han Hu, Jiayuan Gu, Zheng Zhang, Jifeng Dai, and Yichen Wei. Relation networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3588–3597, 2018.

[22] Han Hu, Zheng Zhang, Zhenda Xie, and Stephen Lin. Local relation networks for image recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3464–3473, 2019.

[23] Zehao Huang and Naiyan Wang. Like what you like: Knowledge distill via neuron selectivity transfer. *arXiv preprint arXiv:1707.01219*, 2017.

[24] 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.

[25] Jangho Kim, Yash Bhalgat, Jinwon Lee, Chirag Patel, and Nojun Kwak. Qkd: Quantization-aware knowledge distillation. *arXiv preprint arXiv:1911.12491*, 2019.

[26] Jangho Kim, SeongUk Park, and Nojun Kwak. Paraphrasing complex network: Network compression via factor transfer. In *Advances in neural information processing systems*, pages 2760–2769, 2018.

[27] 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.

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

[29] Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *2006 IEEE Computer So-*

ciety Conference on Computer Vision and Pattern Recognition (CVPR'06), volume 2, pages 2169–2178. IEEE, 2006.

[30] Wonkyung Lee, Junghyup Lee, Dohyung Kim, and Bumsub Ham. Learning with privileged information for efficient image super-resolution. In *European Conference on Computer Vision*, pages 465–482. Springer, 2020.

[31] Yawei Li, Shuhang Gu, Christoph Mayer, Luc Van Gool, and Radu Timofte. Group sparsity: The hinge between filter pruning and decomposition for network compression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8018–8027, 2020.

[32] Min Lin, Qiang Chen, and Shuicheng Yan. Network in network. *arXiv preprint arXiv:1312.4400*, 2013.

[33] 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.

[34] 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.

[35] Yufan Liu, Jiajiong Cao, Bing Li, Chunfeng Yuan, Weiming Hu, Yangxi Li, and Yunqiang Duan. Knowledge distillation via instance relationship graph. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7096–7104, 2019.

[36] Yifan Liu, Ke Chen, Chris Liu, Zengchang Qin, Zhenbo Luo, and Jingdong Wang. Structured knowledge distillation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2604–2613, 2019.

[37] Jian-Hao Luo and Jianxin Wu. Neural network pruning with residual-connections and limited-data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1458–1467, 2020.

[38] 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.

[39] Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3967–3976, 2019.

[40] Nikolaos Passalis and Anastasios Tefas. Learning bag-of-features pooling for deep convolutional neural networks. In *Proceedings of the IEEE international conference on computer vision*, pages 5755–5763, 2017.

[41] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

[42] Baoyun Peng, Xiao Jin, Jiaheng Liu, Dongsheng Li, Yichao Wu, Yu Liu, Shunfeng Zhou, and Zhaoning Zhang. Correlation congruence for knowledge distillation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5007–5016, 2019.

[43] Hieu Pham, Qizhe Xie, Zihang Dai, and Quoc V Le. Meta pseudo labels. *arXiv preprint arXiv:2003.10580*, 2020.

[44] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*, 2014.

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

[46] Frederick Tung and Greg Mori. Similarity-preserving knowledge distillation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1365–1374, 2019.

[47] 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.

[48] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. In *Advances in Neural Information Processing Systems*, pages 10506–10518, 2019.

[49] Tao Wang, Li Yuan, Xiaopeng Zhang, and Jiashi Feng. Distilling object detectors with fine-grained feature imitation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4933–4942, 2019.

[50] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018.

[51] Longhui Wei, An Xiao, Lingxi Xie, Xin Chen, Xiaopeng Zhang, and Qi Tian. Circumventing outliers of autoaugment with knowledge distillation. *arXiv preprint arXiv:2003.11342*, 2(8), 2020.

[52] Jiafeng Xie, Bing Shuai, Jian-Fang Hu, Jingyang Lin, and Wei-Shi Zheng. Improving fast segmentation with teacher-student learning. *arXiv preprint arXiv:1810.08476*, 2018.

[53] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10687–10698, 2020.

[54] Hang Xu, Chenhan Jiang, Xiaodan Liang, and Zhenguo Li. Spatial-aware graph relation network for large-scale object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9298–9307, 2019.

[55] Junho Yim, Donggyu Joo, Jihoon Bae, and Junmo Kim. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4133–4141, 2017.

[56] Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *arXiv preprint arXiv:1612.03928*, 2016.

[57] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.

[58] Xiatian Zhu, Shaogang Gong, et al. Knowledge distillation by on-the-fly native ensemble. In *Advances in neural information processing systems*, pages 7517–7527, 2018.