Pseudo-spherical Knowledge Distillation

Kyungmin Lee* , **Hyeongkeun Lee**
Agency for Defense Development
{kyungmnlee, lhk528}@gmail.com

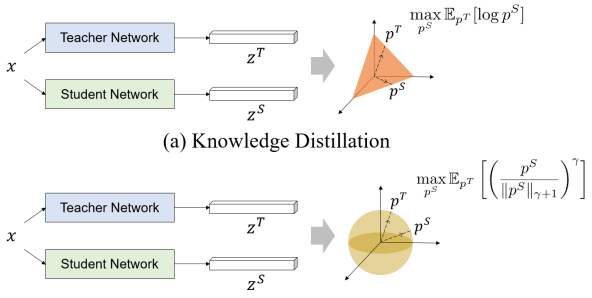
Abstract

Knowledge distillation aims to transfer the information by minimizing the cross-entropy between the probabilistic outputs of the teacher and student network. In this work, we propose an alternative distillation objective by maximizing the scoring rule, which quantitatively measures the agreement of a distribution to the reference distribution. We demonstrate that the proper and homogeneous scoring rule exhibits more preferable properties for distillation than the original cross entropy based approach. To that end, we present an efficient implementation of the distillation objective based on a pseudo-spherical scoring rule, which is a family of proper and homogeneous scoring rules. We refer to it as pseudo-spherical knowledge distillation. Through experiments on various model compression tasks, we validate the effectiveness of our method by showing its superiority over the original knowledge distillation. Moreover, together with structural distillation methods such as contrastive representation distillation, we achieve state of the art results in CIFAR100 benchmarks.

1 Introduction

The *knowledge distillation* (KD) aims to train a network by utilizing the information given by other networks. In particular, smaller networks benefit from knowledge distillation from a bigger network. The KD objective is consists of standard classification loss and a distillation loss, which maximizes the coincidence between the probabilistic outputs of student and teacher networks. Subsequently, many distillation methods proposed distilling auxiliary information such as intermediate features or attention maps. Yet, those methods do not marginally outperform KD. Many recent works focused on transferring the structural knowledge of teachers by using contrastive learning [Tian *et al.*, 2019; Xu *et al.*, 2020; Chen *et al.*, 2021]. However, they showed that attaching the KD objective to contrastive objectives significantly boosts the performance, avowing that the contrastive methods and KD are complementary to each other.

*Contact Author



(a) Knowledge Distillation

(b) Pseudo-spherical Knowledge Distillation

Figure 1: Overview on the comparison between (a) knowledge distillation that maximizes likelihood and proposed (b) pseudo-spherical knowledge distillation that maximizes pseudo-spherical scoring rule.

While the original KD uses cross-entropy loss for the distillation objective, we propose alternative distillation objectives that maximize the alignment between the probabilistic outputs of the teacher and student network. Our method is built on a scoring rule, which quantitatively measures the quality of a predictive distribution with respect to a reference distribution. Thus, we aim to optimize the student network by maximizing the scoring rule with respect to the probability of teacher network. Among various scoring rules, we focus on pseudo-spherical scoring rules, which is a representative family of *proper and homogeneous* scoring rules, that is suitable for optimization on softmax operator-based probabilistic outputs. To that end, we propose *pseudo-spherical knowledge distillation*, which utilizes pseudo-spherical scoring rule for distillation objective. We present an efficient implementation of the new distillation objectives and provide an in-depth analysis of the objectives. Through experiments on CIFAR-100 and ImageNet model compression benchmarks, we demonstrate the effectiveness of our new distillation methods. Especially, by combining contrastive distillation methods, we achieve state-of-the-art performance in CIFAR-100 benchmarks.

2 Related Work

Knowledge Distillation. Since the introduction of model compression [Zeng and Martinez, 2000; Buciluǎ *et al.*, 2006], [Hinton *et al.*, 2015] proposed knowledge distillation which effectively transfers the general and comprehensive knowledge

trained from a teacher network to a student network. It allows the student network to learn the knowledge of the teacher network by minimizing the cross-entropy loss between their probabilistic outputs. Later studies focused on transferring auxiliary information other than probabilistic outputs. For example, [Romero *et al.*, 2014] proposed FitNets, where the student learns from the feature maps of the intermediate layers. [Zagoruyko and Komodakis, 2016a] proposed attention transfer (AT), where the activated parts of the feature maps are transferred to the student. However, [Tian *et al.*, 2019] showed that those methods do not easily outperform knowledge distillation, even have inferior performance when the student and teacher network have different architectural styles.

Recent works focus on passing knowledge that captures the structural correlation between the representations by using contrastive learning objectives.

[Tian *et al.*, 2019] proposed contrastive representation distillation (CRD), where the contrastive objectives are used to maximize the mutual information between teacher and student representations. [Xu *et al.*, 2020] proposed SSKD, which applies self-supervised methods to refine richer representational knowledge of a teacher network and transfer it to a student network. Also, [Chen *et al.*, 2021] exploits Wasserstein distance to perform contrastive learning on both global and local scales.

Scoring Rules. In statistical decision theory [Dawid, 1998], the scoring rule measures the utility of a predictive distribution by assigning a numerical score on the events that materialize. [Gneiting and Raftery, 2007] elaborates the proper scoring rule, which necessitates the existence of the maximum. Thus, the proper scoring rules provide attractive loss and utility functions for estimation problems. Furthermore, the homogeneity of a scoring rule has been studied for unnormalized density estimation. [Takenouchi and Kanamori, 2017] proposed to learn unnormalized statistical models on discrete space by using the homogeneous scoring rule. Recently, [Yu *et al.*, 2021] proposed to train an energy-based model based on the pseudo-spherical scoring rule, with application to generative modeling on high dimensional data.

3 Preliminaries

3.1 Backgrounds on Knowledge Distillation

For a K-class classification problem, let $z^T, z^S \in \mathbb{R}^K$ be logits of teacher and student network. Then the standard KD train the student network by minimizing following objective:

$$\mathcal{L} = \alpha \mathcal{L}_{\text{cls}}(z^S, y) + \beta \mathcal{L}_{\text{dist}}(z^T, z^S), \tag{1}$$

where y is a one-hot label and $\alpha, \beta > 0$ are balancing weights. In addition to the conventional classification loss $\mathcal{L}_{\text{cls}}$, the distillation loss $\mathcal{L}_{\text{dist}}$ encourages the probability of student network to be similar to the probability of teacher network. The probability outputs are computed by a softened softmax operator, which is defined by the following:

$$[\sigma_\tau(z)]_k \coloneqq \frac{\exp(z_k/\tau)}{\sum_{j=1}^K \exp(z_j/\tau)}, \tag{2}$$

where $\tau > 0$ is a temperature that regulates the sharpness of the probability distribution. Let $p^T = \sigma_\tau(z^T)$ and $p^S = \sigma_\tau(z^S)$ be the probability outputs of each teacher and student network, then the distillation objective is a cross-entropy loss between them:

$$\mathcal{L}_{\text{dist}}(z^T, z^S) \coloneqq \tau^2 H(p^T, p^S), \tag{3}$$

where $H(p, q) = -\sum_k p_k \log q_k$ is a cross-entropy loss, and τ^2 is multiplied to match the magnitude of gradient [Hinton *et al.*, 2015]. However, we question the optimality of the cross-entropy objective in distilling probability outputs. To that end, We propose novel distillation objectives by using the scoring rule, an alternative measure that accounts for the probabilistic discrepancy of two distributions.

3.2 Proper and Homogeneous Scoring Rule

The scoring rule evaluates the quality of a probabilistic forecast by computing a score based on the predictive distribution and the events that substantiate. Formaly, let $\mathcal{P}_K$ be a space of all K-categorical distributions, i.e., for any $p \in \mathcal{P}_K$, $\sum_{k\in[K]} p_k = 1$. For a $q \in \mathcal{P}_K$, let us denote the scoring rule by $S(k, q)$ for each $k \in [K]$. Then given a reference distribution $p \in \mathcal{P}_K$, the expected score $S(p, q)$ is defined by

$$S(p, q) \coloneqq \mathbb{E}_p\left[S(k, q)\right] = \sum_{k\in[K]} p_k S(k, q), \tag{4}$$

Definition 1. *[Gneiting and Raftery, 2007] A scoring rule $S : [K] \times \mathcal{P}_K \to \mathbb{R}$ is proper if the corresponding expected score satisfies:*

$$S(p, q) \le S(p, p), \quad \forall p, q \in \mathcal{P}_K, \tag{5}$$

and it is strictly proper the equality holds if and only if $p = q$.

Therefore, the proper scoring rule encourages the forecaster to match their predictions to the true beliefs. Given a (strictly) proper scoring rule S, it induces generalized entropy

$$\mathcal{H}_S(p) \coloneqq \max_{q\in\mathcal{P}_K} S(p, q) = S(p, p). \tag{6}$$

Then one can show that $\mathcal{H}_S(p)$ is a (strictly) convex function of p, and it induces a Bregman divergence

$$\mathcal{D}_S(p, q) \coloneqq S(p, p) - S(p, q) = \mathcal{H}_S(p) - S(p, q). \tag{7}$$

Also, the reverse holds from the following proposition:

Proposition 1. *A scoring rule $S : [K] \times \mathcal{P}_K \to \mathbb{R}$ is (strictly) proper iff $S(p, p)$ is a (strictly) convex function of p.*

Since our goal is to match the probability of a student network to the probability of a teacher network, the proper scoring rule is an attractive distillation objective from the following:

$$\arg\max_{p^S\in\mathcal{P}_K} S(p^T, p^S) = \arg\max_{p^S\in\mathcal{P}_K} \sum_k p_k^T S(k, p^S) = p^T, \tag{8}$$

if S is a strictly proper scoring rule.

Moreover, we introduce homogeneous scoring rule which particularly suits for softmax based probability distributions.

Definition 2. *[Parry* et al.*, 2012] A scoring rule is homogeneous if*

$$S(k, q) = S(k, \lambda \cdot q), \quad \forall \lambda > 0, \forall k \in [K]. \tag{9}$$

Suppose a distribution is given by a softmax operator on a vector $z \in \mathbb{R}^K$, i.e., $q = \sigma_\tau(z)$. Then a (positive) scalar-multiplication on q does not change the inference of a classifier since

$$\arg\max_k q = \arg\max_k \lambda q, \tag{10}$$

for any $\lambda > 0$. Therefore, the homogeneity of a scoring rule reduces the search space for q, which makes the optimization more suitable. Moreover, if a scoring rule S is homogeneous, it suffices to compute the score with respect to the unnormalized distribution $\overline{q} = \exp(z)$. The following examples demonstrate the popular choices for proper scoring rules.

Example 1. ***(log scoring rule)*** *A* log scoring rule *is defined by* $S_{log}(k, q) \coloneqq \log q_k$. *Then the corresponding expected score with respect to* p *is given as* $S_{log}(p, q) = \sum_{k\in[K]} p_k \log q_k$, *where it is equivalent to the maximum likelihood estimation (MLE). Remark that the log scoring rule is strictly proper, but is not homogeneous.*

Example 2. ***(spherical scoring rule)*** *The classical spherical scoring rule [Friedman, 1983] is defined by*

$$S_2(k, q) \coloneqq \frac{q_k}{\|q\|_2}, \tag{11}$$

where $\|q\|_2^2 = \sum_k q_k^2$. *Then for any* $p, q \in \mathcal{P}_K$, *we have*

$$S(p, q) = \|p\|_2 \frac{\langle p, q\rangle}{\|p\|_2\|q\|_2} = \|p\|_2 \cos\angle(p, q), \tag{12}$$

where $\langle p, q\rangle$ *is a inner product and* $\angle(p, q)$ *is a angle between* p *and* q. *Therefore, the spherical scoring rule is strictly proper since it is maximized iff* $p = q$. *Also, since the scalar multiplication on* q *does not change the angle, it is homogeneous.*

3.3 Pseudo-spherical Scoring Rule

We introduce pseudo-spherical scoring rule [Gneiting and Raftery, 2007], which is a family of proper and homogeneous scoring rules that generalizes spherical scoring rule.

Definition 3. *Given* $\gamma > 0$ *and* $q \in \mathcal{P}_K$, *the pseudo-spherical scoring rule is defined as following:*

$$S_\gamma(k, q) \coloneqq \left(\frac{q_k}{\|q\|_{\gamma+1}}\right)^\gamma, \tag{13}$$

where $\|q\|_{\gamma+1} \coloneqq \left(\sum_k q_k^{\gamma+1}\right)^{\frac{1}{\gamma+1}}$.

Then the expected pseudo-spherical score with respect to reference distribution $p \in \mathcal{P}_K$ *is defined as*

$$S_\gamma(p, q) \coloneqq \mathbb{E}_p[S_\gamma(k, q)] = \frac{\sum_k p_k q_k^\gamma}{\|q\|_{\gamma+1}^\gamma}. \tag{14}$$

Alike spherical scoring rule in Example 2, the pseudo-spherical scoring rule is strictly proper and homogeneous.

Proposition 2. *[Gneiting and Raftery, 2007] The pseudo-spherical scoring rule* $S_\gamma(p, q)$ *is strictly proper and homogeneous for* $\gamma > 0$.

4 Pseudo-spherical Knowledge Distillation

We present *pseudo-spherical knowledge distillation* (PSKD), where the knowledge of a teacher is distilled to a student by maximizing the pseudo-spherical scoring rule.

4.1 Design of objective

Since $S_\gamma(p, q)$ is homogeneous, if we let $\overline{q} = q/Z$ be unnormalized distribution of q with some constant $Z > 0$, we have

$$S_\gamma(p, q) = S_\gamma(p, \overline{q}) = \frac{\sum_k p_k \overline{q}_k^\gamma}{\|\overline{q}\|_{\gamma+1}^\gamma}. \tag{15}$$

Moreover, if $q = \sigma_\tau(z)$ for some vector $z \in \mathbb{R}^K$, we have

$$S_\gamma(p, q) = \frac{\sum_k p_k e^{\gamma z_k/\tau}}{\left(\sum_k e^{(\gamma+1)z_k/\tau}\right)^{\frac{\gamma}{\gamma+1}}}, \tag{16}$$

where $\tau > 0$ is a temperature. However, maximizing $S_\gamma(p, q)$ contains fraction term, the gradient based optimization is intractable. Therefore, we define the PSKD objective $\mathcal{L}_\gamma(p, q)$ by taking a logarithm to make it feasible. We present two different approaches. First, by taking the $\frac{1}{\gamma}\log(\cdot)$ inside the expected score, we have

$$\begin{aligned}\mathcal{L}_\gamma^{in}(p, q) &= -\frac{1}{\gamma}\mathbb{E}_p[\log S_\gamma(k, q)] \\ &= -\sum_k p_k \log q_k + \log\|q\|_{\gamma+1},\end{aligned} \tag{17}$$

or by taking log outside of the expected score, we have

$$\begin{aligned}\mathcal{L}_\gamma^{out}(p, q) &= -\frac{1}{\gamma}\log\mathbb{E}_p[S_\gamma(k, q)] \\ &= -\frac{1}{\gamma}\log\left(\sum_k p_k q_k^\gamma\right) + \log\|q\|_{\gamma+1},\end{aligned} \tag{18}$$

where we negate the terms to fit in minimization problem. The negative of latter objective is called γ-score [Takenouchi and Kanamori, 2017], which is also a strictly proper and homogeneous scoring rule as it is the composition of strictly increasing log function and pseudo-spherical scoring rule. [Fujisawa and Eguchi, 2008] also called the latter objective as γ cross-entropy, as a generalization of cross-entropy loss. Indeed, following proposition reveal that the PSKD objectives are generalizations of cross-entropy loss.

Proposition 3. *[Yu* et al.*, 2021] When* $\gamma \to 0$, *we have*

$$\lim_{\gamma\to 0}\mathcal{L}_\gamma^{in}(p, q) = \lim_{\gamma\to 0}\mathcal{L}_\gamma^{out}(p, q) = H(p, q), \tag{19}$$

where $H(p, q) = -\mathbb{E}_p[\log q]$ *is a cross-entropy loss.*

Since the original KD uses cross-entropy loss for distillation, PSKD is a generalization of KD, which is a special case when $\gamma \to 0$.

Extension to negative orders. Even though the pseudo-spherical scoring rule is defined on $\gamma > 0$, we extend the PSKD objectives to $\gamma < 0$ (then the corresponding scoring rule might not be strictly proper).

Note that the following holds from the convexity of $-\log$ function and Jensen's inequality:

$$-\log\mathbb{E}_p[S_\gamma(k, q)] \le -\mathbb{E}_p[\log S_\gamma(k, q)], \tag{20}$$

thus we have

$$\mathcal{L}_\gamma^{out}(p, q) \le \mathcal{L}_\gamma^{in}(p, q), \quad \forall p, q \in \mathcal{P}_K, \forall\gamma > 0, \tag{21}$$

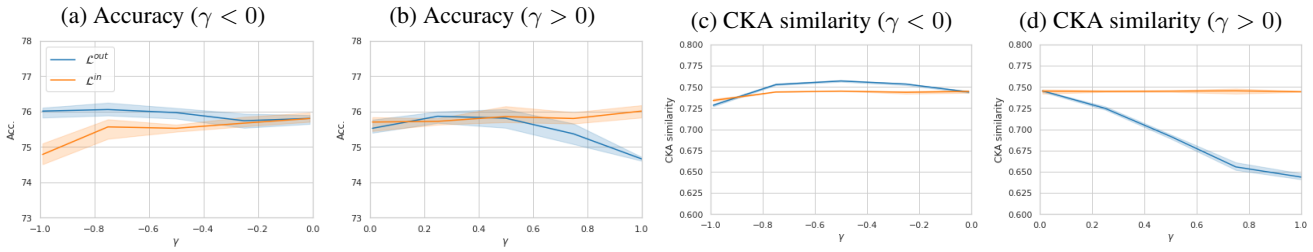


Figure 2: Ablation on different values of γ by comparing the accuracy of student network and the CKA similarity with respect to the teacher network. For $\gamma > 0$, using $\mathcal{L}^{\text{in}}$ is superior to $\mathcal{L}^{\text{out}}$ and for $\gamma < 0$, using $\mathcal{L}^{\text{out}}$ is superior to $\mathcal{L}^{\text{in}}$. The teacher network is WRN 40-2 and the student network is WRN-16-2. Experimented for 3 different seeds.

and it is reversed when $\gamma < 0$. One would expect $\mathcal{L}_\gamma^{in}$ to be superior since it upper bounds $\mathcal{L}_\gamma^{out}$ when $\gamma > 0$, and $\mathcal{L}_\gamma^{out}$ is superior when $\gamma < 0$. This is supported empirically by comparing the performance on model compression tasks (see section 5.1.

4.2 Implementation of PSKD

Let the logits of teacher and student network by z^T and z^S respectively, and let $p^T = \sigma_\tau(z^T)$ and $p^S = \sigma_\tau(z^S)$ be the probability output of each teacher and student network with temperature $\tau > 0$. Then from the homogeneity of pseudo-spherical scoring score and as $\overline{p}^T = \exp(z^T/\tau), \overline{p}^S = \exp(z^S/\tau)$, we have

$$\begin{aligned}\mathcal{L}_\gamma^{in}(p^T, p^S) &= -\sum_k p_k^T \log \overline{p}_k^S + \log \|\overline{p}^S\|_{\gamma+1} \\ &= -\sum_k p_k^T z_k^S/\tau + \frac{1}{\gamma+1}\log \sum_k e^{(\gamma+1)z_k^S/\tau},\end{aligned} \tag{22}$$

and

$$\begin{aligned}\mathcal{L}_\gamma^{\text{out}}(p^T, p^S) &= -\frac{1}{\gamma}\log \sum_k p_k^T (\overline{p}_k^S)^\gamma + \log \|\overline{p}^S\|_{\gamma+1} \\ &= -\frac{1}{\gamma}\log \sum_k p_k^T e^{\gamma z_k^S/\tau} + \frac{1}{\gamma+1}\log \sum_k e^{(\gamma+1)z_k^S/\tau}.\end{aligned} \tag{23}$$

Note that both objectives can be simply implemented by a simple log-sum-exp function, and requires no additional computation compared to the original cross-entropy.

Note that the gradient of $\mathcal{L}_\gamma^{out}$ is given by:

$$\frac{\partial \mathcal{L}_\gamma^{out}}{\partial z^S} = -\frac{1}{\tau}\left(\sigma_\tau\left(z^T + \gamma z^S\right) - \sigma_\tau\left((\gamma+1)z^S\right)\right). \tag{24}$$

Since $\exp(z/\tau) \approx 1 + z/\tau$ for sufficiently large τ and by assuming that the logits have zero-mean as in [Hinton *et al.*, 2015], we have $\sigma_\tau(z) \approx \frac{1+z/\tau}{K}$. Thus, it follows that

$$\begin{aligned}\frac{\partial \mathcal{L}_\gamma^{out}}{\partial z^S} &\approx -\frac{1}{K\tau^2}\left(z^T + \gamma z^S - (\gamma+1)z^S\right) \\ &= -\frac{1}{K\tau^2}\left(z^T - z^S\right).\end{aligned} \tag{25}$$

On the other hand, the gradient of $\mathcal{L}_\gamma^{in}$ is given by:

$$\frac{\partial \mathcal{L}_\gamma^{in}}{\partial z^S} = -\frac{1}{\tau}\left(\sigma_\tau\left(z^T\right) - \sigma_\tau\left((\gamma+1)z^S\right)\right), \tag{26}$$

and by using similar logic, we have

$$\frac{\partial \mathcal{L}_\gamma^{in}}{\partial z^S} \approx -\frac{1}{K\tau^2}\left(z^T - (\gamma+1)z^S\right). \tag{27}$$

Therefore for large τ, the PSKD objective conducts logit matching, and for small τ, it focuses on having the same label for student and teacher (i.e. label matching) similar to KD [Hinton *et al.*, 2015].

Finally, the complete objective is consists of standard classification loss and pseudo-spherical distillation loss as in eqn. (1):

$$\mathcal{L} = \alpha\mathcal{L}_{\text{cls}} + \beta\tau^2\mathcal{L}_\gamma^i, \quad i \in \{out, in\} \tag{28}$$

where $\alpha, \beta > 0$ are balancing weights and $\mathcal{L}_\gamma$ can be both $\mathcal{L}_\gamma^{in}$ and $\mathcal{L}_\gamma^{out}$. Note that we multiply τ^2 since the gradient scales as $1/\tau^2$.

5 Experiment

For the experiments, we demonstrate the effectiveness of PSKD on model compression tasks. First, we conduct ablation on PSKD objectives with respect to different values of γ. Then, we compare PSKD to cutting-edge distillation methods on CIFAR-100 and ImageNet benchmarks.

5.1 Ablation Study

Effect of γ. First, we conduct ablation studies on the performance of PSKD with different values of γ. Given a pre-trained wide ResNet 40-2 [Zagoruyko and Komodakis, 2016b] teacher network, we train wide ResNet 16-2 student network on CIFAR-100. For evaluation, we report top-1 test accuracy of student network and the centered kernel alignment (CKA) similarity [Kornblith *et al.*, 2019] of teacher and student network. The CKA similarity is a reliable measure of the similarity of two neural representations, which is invariant to orthogonal transformation and isotropic scaling. Thus, high CKA similarity indicates that the student learns similar representations to the teacher. We took the output of the penultimate layer (i.e. after average pooling) to calculate the CKA similarity.

Figure 2 demonstrates the effect of γ for both $\mathcal{L}_\gamma^{out}$ and $\mathcal{L}_\gamma^{in}$. We observe that the test accuracy of student networks distilled by $\mathcal{L}_\gamma^{in}$ increases as γ increases from -1 to 1, while the CKA similarity remains unchanged. On the other hand, the student networks distilled by $\mathcal{L}_\gamma^{out}$ achieve their maximum

Teacher Student	WRN-40-2 WRN-16-2	WRN-40-2 WRN-40-1	resnet56 resnet20	resnet110 resnet20	resnet110 resnet32	resnet32x4 resnet8x4	vgg13 vgg8
Teacher	75.61	75.61	72.34	74.31	74.31	79.42	74.64
Student	73.26	71.98	69.06	69.06	71.14	72.50	70.36
KD	74.92	73.54	70.66	70.67	73.08	73.33	72.98
FitNet	73.58	72.24	69.21	68.99	71.06	73.50	71.02
AT	74.08	72.77	70.55	70.22	72.31	73.44	71.43
SP	73.83	72.43	69.67	70.04	72.69	72.94	72.68
VID	74.11	73.30	70.38	70.16	72.61	73.09	71.23
RKD	73.35	72.22	69.61	69.25	71.82	71.90	71.48
PKT	74.54	73.45	70.34	70.25	72.61	73.64	72.88
AB	72.50	72.38	69.47	69.53	70.98	73.17	70.94
FT	73.25	71.59	69.84	70.22	72.37	72.86	70.58
CRD	75.48	74.14	71.16	71.46	73.48	75.51	73.94
CRD+KD	75.64	74.38	71.63	71.56	73.75	75.46	74.29
WCoRD	75.88	74.73	71.56	71.57	73.81	75.95	74.55
WCoRD+KD	76.11	74.72	**71.92**	**71.88**	74.20	76.15	74.72
SSKD	75.93	75.74	71.25	71.13	73.69	76.03	75.03
PSKD (Ours)	76.01	74.06	71.30	70.91	73.6	75.24	74.14
CRD+PSKD (Ours)	**76.53**	74.96	71.88	71.77	**74.19**	76.38	74.46
SSKD+PSKD (Ours)	76.09	**75.78**	71.32	71.38	74.03	**76.94**	**75.14**

Table 1: CIFAR-100 test accuracy (%) of student networks trained with various distillation methods where the teacher and student have similar architectural style. The citations and comparisons are in Appendix. Each results are provided by author, and for SSKD, we re-run the experiment to compare for same teacher network. Average over 5 runs.

around $\gamma = -0.5$ and it drops when γ becomes larger. Also, the CKA similarity significantly drops as γ increases. Remark that when distilling by $\mathcal{L}_\gamma^{out}$, the CKA similarity is strongly correlated with the test accuracy of student networks, yet it doesn't hold when distilling with $\mathcal{L}_\gamma^{in}$.

For the case when a student is trained by $\mathcal{L}_\gamma^{in}$, the CKA similarity doesn't change for different values of γ, showing that the value of γ on distilling representational knowledge doesn't vary. However, the test accuracy varies as γ becomes larger. We provide analysis by the following gradient. From eqn. (27), recall that the gradient of $\mathcal{L}^{in}$ can be approximated by a logit matching between z^T and $(\gamma+1)z^S$. Thus when the logit of a teacher network has a high value, the larger value of γ makes student logit easier to follow the teacher logit.

From eqn. (25), the approximation on the gradient of $\mathcal{L}_\gamma^{out}$ is a logit matching between z^T and z^S, which is independent to γ. But, from eqn (24), the gradient of $\mathcal{L}_\gamma^{out}$ is expressed by the difference between $\sigma_\tau(z^T + \gamma z^S)$ and $\sigma_\tau((\gamma+1)z^S)$. Then for $\gamma > 0$, the softmax output becomes smoother as γ increases, thus the gradient becomes smaller and reduces the effect of distillation. Therefore, $\mathcal{L}_\gamma^{out}$ has relatively better performance when $\gamma < 0$.

Comparison of two PSKD objectives. We further compare two objectives by conducting experiments on various teacher and student combinations. We choose $\gamma = -0.5$ and $\gamma = 1.0$ for baseline. Table 2 reports the test accuracy of student networks trained with different PSKD objectives and values of γ. Remark that $\mathcal{L}^{out}$ outperforms $\mathcal{L}^{in}$ when $\gamma = -0.5$ and $\mathcal{L}^{in}$ outperforms $\mathcal{L}^{out}$ when $\gamma = 1.0$, ascertaining the hypothesis asserted in section 4.1. While $\mathcal{L}^{in}$ with $\gamma = 1.0$ and $\mathcal{L}^{out}$ with $\gamma = -0.5$ achieve similar performances, the former achieves slightly better performance overall. Therefore, we use $\mathcal{L}^{out}$ with $\gamma = -0.5$ for the rest of the experiments.

γ	objective	WRN-40-2 WRN-16-2	resnet110 resnet32	resnet32x4 resnet8x4	vgg13 vgg8
-0.5	$\mathcal{L}^{\text{out}}$	**76.01**	**73.27**	75.49	**73.94**
	$\mathcal{L}^{\text{in}}$	75.69	73.26	75.35	73.88
1.0	$\mathcal{L}^{\text{out}}$	74.92	71.59	75.24	70.76
	$\mathcal{L}^{\text{in}}$	75.92	72.89	**75.58**	73.65

Table 2: Ablation on the distillation objective when γ is positive and negative. We report the test accuracy of student network. For each task, all the hyperparameters are same except the value of γ and distillation objective. Average over 3 runs.

5.2 Main Results

Setup. We follow model compression benchmarks that were proposed in CRD [Tian *et al.*, 2019]. For experiments on CIFAR-100, there are 13 teacher-student combinations where 7 are of similar architectures and 6 are of different architectures. Those architectures include resnet [He *et al.*, 2016], vgg [Simonyan and Zisserman, 2014], and ShuffleNet [Ma *et al.*, 2018; Zhang *et al.*, 2018]. We used pre-trained teacher models provided in the official CRD repository [1]. For the experiment on ImageNet, we used the official PyTorch pre-trained ResNet-34 for a teacher and ResNet-18 for a student.

Combination with other distillation methods. Recently, many distillation methods resort to transfer structural knowledge of teacher networks by using contrastive learning. It has been empirically verified that using the contrastive distillation method and KD simultaneously improves performance,

[1] https://github.com/HobbitLong/RepDistiller

Teacher Student	vgg13 MobileNetV2	ResNet50 MobileNetV2	ResNet50 vgg8	resnet32x4 ShuffleNetV1	resnet32x4 ShuffleNetV2	WRN-40-2 ShuffleNetV1
Teacher	74.64	79.34	79.34	79.42	79.42	75.61
Student	64.6	64.6	70.36	70.5	71.82	70.5
KD	67.37	67.35	73.81	74.07	74.45	74.83
FitNet	64.14	63.16	70.69	73.59	73.54	73.73
AT	59.40	58.58	71.84	71.73	72.73	73.32
SP	66.30	68.08	73.34	73.48	74.56	74.52
VID	65.56	67.57	70.30	73.38	73.40	73.61
RKD	64.52	64.43	71.50	72.28	73.21	72.21
PKT	67.13	66.52	73.01	74.10	74.69	73.89
AB	66.06	67.20	70.65	73.55	74.31	73.34
FT	61.78	60.99	70.29	71.75	72.50	72.03
CRD	69.73	69.11	74.30	75.11	75.65	76.05
CRD+KD	69.94	69.54	74.58	75.12	76.05	76.27
WCoRD	69.47	70.45	74.86	75.40	75.96	76.32
WCoRD+KD	70.02	70.12	74.68	75.77	76.48	76.68
SSKD	71.65	72.27	**76.20**	**78.10**	**78.82**	77.43
PSKD (Ours)	69.99	69.21	74.67	75.27	75.77	75.92
CRD+PSKD (Ours)	69.97	70.71	75.11	75.72	76.81	76.67
SSKD+PSKD (Ours)	**71.73**	**72.60**	**76.19**	78.03	**78.81**	**77.67**

Table 3: CIFAR-100 test accuracy (%) of student networks trained with various distillation methods where the teacher and student have different architectural style. The citations and comparisons are in Appendix. Each results are provided by author except for SSKD, which we re-run the experiment to compare for same teacher network. Average over 3 runs.

	Teacher	Student	AT	KD	SP	CC	CRD	WCoRD	SSKD	PSKD
Top-1	26.69	30.25	29.30	29.34	29.38	30.04	28.83	28.51	**28.38**	28.47
Top-5	8.58	10.93	10.00	10.12	10.20	10.83	9.87	9.84	**9.33**	9.51

Table 4: Top-1 and Top-5 error rates (%) of student network ResNet-18 on ImageNet validation set.

showing that they are complementary to each other. We show that using PSKD and contrastive distillation together further improves the performance. To that end, we add the contrastive distillation objective $\mathcal{L}_{\text{con}}$ to our objective:

$$\mathcal{L} = \alpha\mathcal{L}_{\text{cls}} + \beta\mathcal{L}_{\gamma} + \eta\mathcal{L}_{\text{con}}, \tag{29}$$

where $\alpha, \beta, \eta > 0$ are balancing weights. For contrastive methods, we experiment on both CRD [Tian *et al.*, 2019] and SSKD [Xu *et al.*, 2020].

Results on CIFAR-100 dataset. Table 1 compares the performance of various distillation methods when teacher and student share similar architectural styles. We observe that PSKD outperforms KD and other baselines except for the contrastive distillation methods. The contrastive methods such as CRD [Tian *et al.*, 2019], SSKD [Xu *et al.*, 2020] and WCoRD [Chen *et al.*, 2021] exhibit slightly better performance than PSKD as they transfer structural information of teacher network. We observe that using PSKD and contrastive methods together can further boost the performance. In addition to CRD and SSKD, model compression with PSKD achieves state-of-the-art performance in 5 out of 7 tasks, showing the effectiveness of PSKD.

Table 3 compares the performance when the teachers and students are of different architectures. Remark that the distillation methods using the information extracted from intermediate layers may perform worse when the architectures of teacher and student differ, but since PSKD only uses the output of a network, it doesn't rely on architecture-specific cues. Also, when using PSKD with contrastive methods, we achieve state-of-the-art results in 5 out of 6 tasks.

Results on ImageNet. Table 4 shows the results of model compression on ImageNet. Remark that our PSKD achieves lower error rates than KD and all other distillation methods except SSKD. Surprisingly, even though PSKD does not rely on transferring structural information, it achieves better performance than CRD and WCoRD. Meanwhile, PSKD enjoys simpler implementation and faster training.

6 Conclusion

We present pseudo-spherical knowledge distillation that generalizes knowledge distillation by using pseudo-spherical scoring rule. We propose two novel loss functions for pseudo-spherical knowledge distillation and provide an empirical and qualitative analysis. Our method achieves state-of-the-art results on CIFAR-100 and ImageNet supervised model compression, together with structural knowledge distillation methods.

Acknowledgments

This work was supported by the Agency for Defense Development by the Korean Government ("Adversarial AI based Deep Neural Nets Attack and Defense project")

References

[Buciluǎ *et al.*, 2006] Cristian Buciluǎ, Rich Caruana, and Alexandru Niculescu-Mizil. Model compression. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 535–541, 2006.

[Chen *et al.*, 2021] Liqun Chen, Dong Wang, Zhe Gan, Jingjing Liu, Ricardo Henao, and Lawrence Carin. Wasserstein contrastive representation distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16296–16305, 2021.

[Dawid, 1998] A Philip Dawid. Coherent measures of discrepancy, uncertainty and dependence, with applications to bayesian predictive experimental design. *Department of Statistical Science, University College London. http://www. ucl. ac. uk/Stats/research/abs94. html, Tech. Rep*, 139, 1998.

[Friedman, 1983] Daniel Friedman. Effective scoring rules for probabilistic forecasts. *Management Science*, 29(4):447–454, 1983.

[Fujisawa and Eguchi, 2008] Hironori Fujisawa and Shinto Eguchi. Robust parameter estimation with a small bias against heavy contamination. *Journal of Multivariate Analysis*, 99(9):2053–2081, 2008.

[Gneiting and Raftery, 2007] Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102(477):359–378, 2007.

[He *et al.*, 2016] 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.

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

[Kornblith *et al.*, 2019] Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International Conference on Machine Learning*, pages 3519–3529. PMLR, 2019.

[Ma *et al.*, 2018] Ningning Ma, Xiangyu Zhang, Hai-Tao Zheng, and Jian Sun. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In *Proceedings of the European conference on computer vision (ECCV)*, pages 116–131, 2018.

[Parry *et al.*, 2012] Matthew Parry, A Philip Dawid, and Steffen Lauritzen. Proper local scoring rules. *The Annals of Statistics*, 40(1):561–592, 2012.

[Romero *et al.*, 2014] 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.

[Simonyan and Zisserman, 2014] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[Takenouchi and Kanamori, 2017] Takashi Takenouchi and Takafumi Kanamori. Statistical inference with unnormalized discrete models and localized homogeneous divergences. *The Journal of Machine Learning Research*, 18(1):1804–1829, 2017.

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

[Xu *et al.*, 2020] Guodong Xu, Ziwei Liu, Xiaoxiao Li, and Chen Change Loy. Knowledge distillation meets self-supervision. In *European Conference on Computer Vision*, pages 588–604. Springer, 2020.

[Yu *et al.*, 2021] Lantao Yu, Jiaming Song, Yang Song, and Stefano Ermon. Pseudo-spherical contrastive divergence. *Advances in Neural Information Processing Systems*, 34, 2021.

[Zagoruyko and Komodakis, 2016a] 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.

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

[Zeng and Martinez, 2000] Xinchuan Zeng and Tony R. Martinez. Using a neural network to approximate an ensemble of classifiers. *Neural Processing Letters*, 12(3):225–237, 2000.

[Zhang *et al.*, 2018] Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6848–6856, 2018.