REDUCING THE CAPACITY GAP VIA SPHERICAL KNOWLEDGE DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Knowledge distillation aims to obtain a small and effective student model by learning the output from a large knowledgeable teacher model. However, when the student is distilled by an oversized teacher, a critical performance degradation problem is exposed. This paper revisits performance degradation problem from the perspective of model confidence. Specifically, we apply energy-based metrics to measure the confidence of models, and propose Spherical Knowledge Distillation (SKD): a more efficient knowledge distillation framework when distilling with larger teachers. A theoretical analysis is provided to show that SKD can effectively reduce the confidence gap between the teacher and student, thus alleviating the performance degradation problem. We demonstrate that SKD is easy to train, and can significantly outperform several strong baselines on various mainstream datasets, including CIFAR-100 and ImageNet.

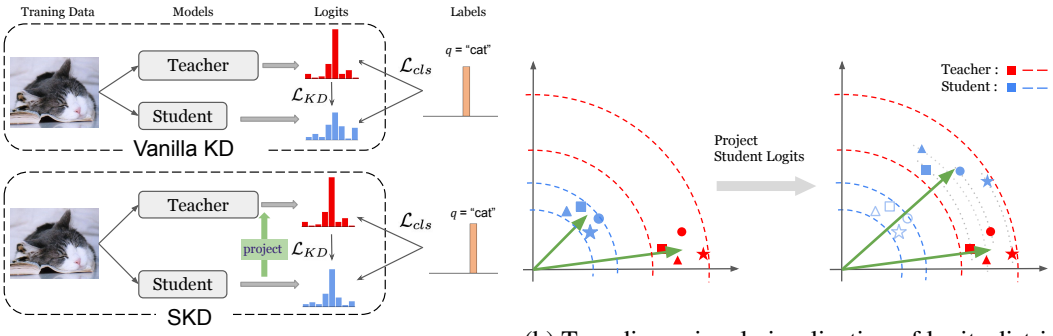
1 INTRODUCTION

Deep neural networks have achieved remarkable success in various fields such as computer vision (Deng et al., 2009; He et al., 2016; Zagoruyko & Komodakis, 2016), natural language processing (Vaswani et al., 2017; Kenton & Toutanova, 2019), and speech recognition (Ren et al., 2019). However, these state-of-the-art models have high costs in terms of storage, memory, and computation time, which hinders their deployment in practice. Therefore, model compression has attracted considerable research attention in recent years (Hinton et al., 2015; Polino et al., 2018; Choudhary et al., 2020; Kim et al., 2020; Ganesh et al., 2021; Xia et al., 2022).

One of the predominant approaches in model compression is knowledge distillation (KD) (Hinton et al., 2015; Tian et al., 2019; Guo et al., 2020; Shen et al., 2021; Chen et al., 2022), which trains a smaller model (i.e., the student) with the output of a larger and well-trained model (i.e., the teacher). Naturally, one would expect to train a better student with a larger and more accurate teacher. However, recent research has invalidated this hypothesis and found that knowledge distillation suffers from a mysterious performance degradation problem (Cho & Hariharan, 2019; Mirzadeh et al., 2019). Specifically, the student performance degrades with an oversized teacher, indicating that the knowledge of an oversized teacher cannot be effectively transferred to the student (Table 1).

This work proposes a Spherical Knowledge Distillation (SKD) framework, which examines the performance degradation problem from the perspective of confidence gap. We argue that the performance degradation problem is caused by the confidence gap between teachers and students. Specifically, a larger-capacity network is more likely to be overconfident (i.e., produces predicting probabilities that are close to one-hot distributions). This is because a larger-capacity network manages to further minimize the negative log likelihood by increasing its confidence, even when it can correctly classify almost all training samples.

To quantify the confidence of models, SKD applies two energy-based metrics, termed entropy and Helmholtz free energy. We show that these two energy-based metrics are effective to measure the confidence gap between the teacher and student. Further, considering the output logits space in a spherical coordination system, we show that student output logits are distributed in a smaller radius area with higher entropy and lower free energy, and vice versa for the teacher output logits (Figure 1). A theoretical analysis is provided to show that projecting the student output to the hyper sphere with teacher's radius can reduce the confidence gap, thus alleviating the performance degradation problem.



(a) Vanilla KD and SKD.

(b) Two-dimensional visualization of logits distribution of the teacher and student.

Figure 1: The illustration of Spherical Knowledge Distillation (SKD) framework. (a) Compared with Vanilla KD, SKD effectively reduces the confidence gap between the teacher and student. (b) In the spherical coordination system, the student's output logits have lower radius, where the teacher's is the contrary. SKD projects the student logits on the hyper sphere with the teacher's radius, thus reducing the confidence gap. Different color represents logits of different models, and different shape represents predictions of different samples.

To verify the effectiveness of our method, we present comprehensive experiments on CIFAR-100 (Krizhevsky et al., 2009) and ImageNet (Deng et al., 2009). The proposed SKD framework achieves excellent performance on these tasks. The experimental results show that SKD can mitigate the performance degradation problem and produce competitive students. For example, SKD distills the ResNet18 student with 73.01% accuracy on ImageNet, which achieves a new state-of-the-art result.

2 Methodology

2.1 Background

Vanilla Knowledge Distillation When training neural networks, we minimize the negative log-likelihood of the ground truth class to update model parameters. After the model is well-trained, the probability of the ground truth would be close to 1, while the probabilities of other labels are near 0. Hinton et al. (2015) noticed that the small wrong probabilities of large models are useful to unveil "dark knowledge". For example, given a picture of a "cat", the model is more likely to output a higher probability for class "dog" than class "airplane". These wrong probabilities imply the relationship between the two classes and unveil how a model tends to generalize. This observation inspired the usage of large models' outputs as soft targets to train efficient small models, which is the core idea of Knowledge Distillation. However, modern deep networks tend to produce peaky probabilities (Guo et al., 2017; Lee et al., 2018), i.e., the numbers of those wrong classes (near zero values) would be negligible compared to the ground truth (near one). Thus Hinton et al. (2015) proposed to raise the temperature of the last softmax layer to soften the output probabilities, which can be used as soft targets to train student networks.

For vanilla knowledge distillation, the KD loss can be defined as follows:

$$\begin{aligned} \mathcal{L}_{KD} &= -\sum_i q_i \log p_i \\ p_i = \frac{e^{z_i/\tau}}{\sum_j e^{z_j/\tau}}, q_i &= \frac{e^{v_i/\tau}}{\sum_j e^{v_j/\tau}} \end{aligned} \tag{1}$$

Where we denote logits of the teacher as v, logits of the student as z, student probability as p, teacher probability as q, temperature as τ, and the i-th and j-th value of logits (i.e. the i-th and j-th category of K classes) as i and j, respectively. The final loss for the student is then the weighted sum of the typical cross entropy loss $\mathcal{L}_{cls}$ and the knowledge distillation loss $\mathcal{L}_{KD}$:

$$\mathcal{L} = \lambda \mathcal{L}_{KD} + (1-\lambda)\mathcal{L}_{cls} \tag{2}$$

Table 1: The performance degradation problem.

Teacher	ResNet20	ResNet32	ResNet44	ResNet56	ResNet110
Teacher Acc	69.57	70.9	71.9	72.8	**73.8**
Student Acc	67.4	**68.2**	68	67.5	67.1
KD loss	**1.1**	1.7	2.1	2.5	3.3

The popular choice of the temperature τ is in $\{3, 4, 5\}$, and the weight λ is 0.9 (Hinton et al., 2015; Cho & Hariharan, 2019; Tian et al., 2020).

Performance Degradation Problem While knowledge distillation achieved success in many fields, a mysterious performance degradation problem was observed (Cho & Hariharan, 2019; Mirzadeh et al., 2019). Since the idea of knowledge distillation is transferring teacher knowledge to students, one natural hypothesis is that a larger and more accurate teacher would capture more knowledge and thus train better students. However previous studies invalidate this hypothesis by showing that student performance degenerates unexpectedly with larger teachers. For example, as shown in Table 1, applying a larger teacher model will increase the KD loss and decrease the accuracy of the student (with ResNet14 as the backbone). Cho & Hariharan (2019) hypothesize that the mismatch of capacity between the teacher and student causes the performance degradation. We further provide a new perspective about model confidence to explain and address the performance degradation problem, which we will discuss in the rest of this section.

2.2 Revisiting Performance Degradation Problem via Confidence Gap

There are two main theories explaining why knowledge distillation is effective. The first theory originated from vanilla Knowledge Distillation (Hinton et al., 2015), which argued that teachers are more accurate in capturing category similarities thus helping students to generalize better on unseen data. The second theory (Müller et al., 2019; Yuan et al., 2020) contends that the soft output of the teacher prevents the student network from overconfidence. However, when the teacher is larger, these two theories contradict each other. Such contradiction raises because larger models are usually more accurate and confident (i.e. with larger probability for ground truth) at the same time. While a more accurate teacher is beneficial to the student model, the output of a more confident teacher is closer to the one-hot distributions, thus degrading student performance. Therefore, the larger teacher capacity is a double-edged sword for distillation. We argue that it is crucial to investigate the confidence gap between the teacher and student, which previous studies have overlooked.

Recent works apply regularization techniques (e.g., label smoothing (Müller et al., 2019)) to prevent the teacher from overconfidence. From the prospective of confidence gap, such techniques are equivalent to reducing the confidence gap between the teacher and student on the teacher side. However, the teacher model trained by label smoothing could result in the loss of information in the logits about resemblances between instances of different classes Müller et al. (2019). Such information loss in the logits could degrade the student performance Chandrasegaran et al. (2022). Therefore, it is important to investigate how to reduce confidence gap without loss of information.

It should be noted that it is non-trivial for students with insufficient parameters to reduce the confidence gap (i.e., by multiplying a constant factor c ($c \geq 1$) to the student's logits; or by applying a lower constant temperature value for the student). This is because the predictions of the student and teacher model are inconsistent (i.e., the predictions of the teacher and student differ) for many samples. For those samples with inconsistent predictions, multiplying a factor to the student's output logits to make its confidence larger will increase the distillation loss. This could happen even if when students and teachers predict the same because the confidence gaps differ between separate samples. Consider a bi-classification example where the teacher's logits for two samples A and B are $(1.0, -1.0)$ and $(0.5, -0.5)$, respectively; while the student's logits are $(0.5, -0.5)$ and $(0.4, -0.4)$. At this time, multiplying the student model by the constant factor $c = 2$ will make the student's logits to be $(1.0, -1.0)$ and $(0.8, -0.8)$, respectively. In this case, although the loss for the first sample decreases, the loss for the second sample increases. Therefore, a more advanced method need to be explored to reduce the confidence gap in distillation.

Table 2: The gap of entropy and Helmholtz free energy using Vanilla KD and SKD.

	ResNet20	ResNet32	ResNet44	ResNet56	ResNet110
Entropy	0.74	0.45	0.35	0.22	0.09
$G_{entropy}$ - Vanilla KD	0.146	0.181	0.222	0.246	0.261
$G_{entropy}$ - SKD	**0.053**	**0.074**	**0.081**	**0.083**	**0.094**
Free Energy	13.11	13.84	14.45	15.37	16.13
G_f - Vanilla KD	0.042	0.053	0.063	0.068	0.074
G_f - SKD	**0.029**	**0.042**	**0.041**	**0.045**	**0.044**

In the following subsections, we first propose two energy-based metrics to measure the confidence gap between the teacher and student. Then we provide a theoretical analysis showing how to reduce the confidence gap between the teacher and student to alleviate the performance degradation problem.

2.3 Measuring Confidence Gap via Energy-based Metrics

Energy-based Metrics To investigate the relationship between model capacity and confidence, we apply two energy-based metrics termed *Entropy* and *Helmholtz Free Energy*, that are commonly used in quantifying thermal processes in thermodynamics. The concept of *Entropy* was also introduced into information science (Shannon, 1948) to measure the uncertainty (i.e., confidence) of a random variable (Kullback, 1997), and is defined as follows:

$$S = -\sum_i p_i \log(p_i) \tag{3}$$

According to its definition, high-confidence predictions would correspond to low entropy, and vise versa for low-confidence predictions. Thus entropy can be used to measure the model confidence. *Helmholtz Free Energy* is frequently used in energy-based model (LeCun et al., 2006). In thermodynamics, increasing the entropy of a system at a constant temperature will decrease the Helmholtz free energy (Lewis & Randall, 1963). Therefore, entropy and Helmholtz free energy has a close relationship with each other. Helmholtz free energy can be expressed as the log partition function as follows, where z_i is the i-th element of the logits, and τ is the temperature:

$$F(z) = \tau \log \sum_i e^{z_i/\tau} \tag{4}$$

It should be noted that when the temperature τ is set to 1, $F(z)$ is equivalent to the RealSoftMax $LSE(x) = \log \sum_i e^{x_i}$ (Nielsen & Sun, 2016). Therefore, $F(z)$ can be regarded as a smooth approximation to the maximum function, thus can also be used to measure the model confidence.

We measured the entropy and Helmholtz free energy for different models trained on the CIFAR-100 dataset with temperature $\tau = 1$ (with one-hot labels). As shown in Table 2, there exists a large gap in both entropy and free energy between networks with different sizes. And larger models generally produce lower entropy and higher free energy. In the following part of this subsection, we will first define *Entropy Gap* and *Helmholtz Free Energy Gap*, and then demonstrate how to reduce the gap.

Glossary

1. z_i: the i-th element of logits.
2. $||z||$: the norm of the logits, $||z|| = \sqrt{\sum_i z_i^2}$
3. p_i: the probability of the ith category. $p_i = -\frac{e^{z_i/\tau}}{\sum_j e^{z_j/\tau}}$
4. K: the dimension of the vector z and v (i.e. the number of label categories).

Entropy Gap Analysis We define the entropy gap as follows:

$$\begin{aligned} G_{entropy} &= S(p) - S(q) \\ &= \sum_i -p_i \log(p_i) + \sum_i q_i \log(q_i) \end{aligned} \tag{5}$$

Theorem 1. *Let p and q represent the student and the teacher output probability, z and v represent the student and teacher logits, and τ represents the temperature. The entropy gap $G_{entropy}$ approximate to $\frac{1}{K\tau^2}(||v||^2 - ||z||^2)$, under the assumption that z and v are zero meaned separately.*

Proof.

$$\begin{aligned} G_{entropy} &= S(p) - S(q) \\ &= \sum_i -p_i \log(p_i) + \sum_i q_i \log(q_i) \\ &= \sum_i (q_i v_i/\tau - p_i z_i/\tau) + \log \sum_j e^{z_j/\tau} - \log \sum_j e^{v_j/\tau} \\ &= \sum_i (\frac{e^{v_i/\tau} v_i/\tau}{\sum_j e^{v_j/\tau}} - \frac{e^{z_i/\tau} z_i/\tau}{\sum_j e^{z_j/\tau}}) + \log \sum_j e^{z_j/\tau} - \log \sum_j e^{v_j/\tau} \end{aligned} \tag{6}$$

With Taylor expansion $e^x \approx 1 + x$:

$$\begin{aligned} G_{entropy} \approx \sum_i (&\frac{(1 + v_i/\tau)v_i/\tau}{\sum_j (1 + v_j/\tau)} - \frac{(1 + z_i/\tau)z_i/\tau}{\sum_j (1 + z_j/\tau)}) \\ &+ \log \sum_j (1 + z_j/\tau) - \log \sum_j (1 + v_j/\tau) \end{aligned} \tag{7}$$

We follow the assumption from Hinton (Hinton et al., 2015), that the logits have been zero-meaned separately for each training example so that $\sum_j z_j = \sum_j v_j = 0$. We provide experimental validation to this assumption in Appendix A.2.

Given the above assumption, we can get:

$$\begin{aligned} G_{entropy} &\approx \sum_i (\frac{(v_i/\tau)^2}{K} - \frac{(z_i/\tau)^2}{K}) + \log(K) - \log(K) \\ &= \frac{1}{K\tau^2}(||v||^2 - ||z||^2) \end{aligned} \tag{8}$$

□

Helmholtz Free Energy Gap Analysis We define the Helmholtz free energy gap as follows:

$$\begin{aligned} G_f &= F(v) - F(z) \\ &= \tau \log \sum_i e^{v_i/\tau} - \tau \log \sum_i e^{z_i/\tau} \end{aligned} \tag{9}$$

Theorem 2. *Let z and v represent the student and the teacher output logits, and τ denotes the temperature. The Helmholtz free energy gap G_f approximate to $\frac{1}{2K\tau}(||v||^2 - ||z||^2)$, under the assumption that z and v are zero meaned separately and $2\tau^2 K$ is large compared with the square of logits norm.*

Proof.

$$\begin{aligned} G_f &= F(v) - F(z) \\ &= \tau \log \sum_i e^{v_i/\tau} - \tau \log \sum_i e^{z_i/\tau} \\ &\approx \tau \log(K + \sum_i v_i/\tau + \frac{1}{2}\sum_i v_i^2/\tau^2) \\ &\quad - \tau \log(K + \sum_i z_i/\tau + \frac{1}{2}\sum_i z_i^2/\tau^2) \end{aligned} \tag{10}$$

With Taylor expansion $e^x \approx 1 + x + \frac{1}{2}x^2$:

$$\begin{aligned} G_f &\approx \tau \log(K + \frac{1}{2}\sum_i v_i^2/\tau^2) - \tau \log(K + \frac{1}{2}\sum_i z_i^2/\tau^2) \\ &\approx \tau \log(1 + \frac{1}{2\tau^2 K}||v||^2) - \tau \log(1 + \frac{1}{2\tau^2 K}||z||^2) \end{aligned} \tag{11}$$

When $2\tau^2 K$ is large compared with the square of logits norm, with $\log(1+x) \approx x$ for $|x| < 1$:

$$G_f \approx \frac{1}{2K\tau}(||v||^2 - ||z||^2) \tag{12}$$

□

2.4 Spherical Knowledge Distillation (SKD)

In the previous subsection, we apply Entropy Gap and Helmholtz Free Energy Gap to measure the confidence gap between two models. We have shown that both Entropy Gap and Helmholtz Free Energy Gap can be represented as similar forms (i.e., $\frac{1}{\alpha K\tau}(||v||^2 - ||z||^2)$), whose value is controlled by the norm of the teacher and student logits. Consider logits in a spherical coordination system, the logits of the student are distributed in a smaller radius area with higher entropy and lower free energy, and vice versa for the teacher. Therefore, we propose Spherical Knowledge Distillation (SKD), which projects the student logits onto the hyper-sphere of the teacher. Specifically, the student logits for each sample would be transformed as follows:

$$\hat{z} = z * \frac{||v||}{||z||} \tag{13}$$

After the transformation, the student logits would have the same norm of the teacher logits, which significantly reduce the gap for both entropy and Helmholtz free energy: $G_f \approx \frac{1}{2K\tau}(||v||^2 - ||\hat{z}||^2) = 0$, and $G_{entropy} \approx \frac{1}{K\tau^2}(||v||^2 - ||z||^2) = 0$. In other words, SKD projects student logits from the area with high entropy and low free energy to the teacher's area, which is with low entropy and high free energy. The rest of SKD follows the standard distillation procedure:

$$\begin{aligned} &p_i = \frac{e^{\hat{z}_i/\tau}}{\sum_j e^{\hat{z}_j/\tau}}, q_i = \frac{e^{v_i/\tau}}{\sum_j e^{v_j/\tau}} \\ &\mathcal{L}_{SKD} = -\sum_i q_i \log p_i \\ &\mathcal{L} = \lambda \mathcal{L}_{SKD} + (1-\lambda)\mathcal{L}_{cls} \end{aligned} \tag{14}$$

3 Experiments

In this section, we show comprehensive experimental results to validate the effectiveness of SKD from several perspectives. Specifically, we first conducted experiments on two popular CV datasets to demonstrate the performance of SKD. Then we focused on evaluating whether SKD could alleviate the performance degradation problem.

Table 3: CIFAR-100 experiments.

Teacher Student	WRN-40-2 WRN-16-2	WRN-40-2 WRN-40-1	ResNet56 ResNet20	ResNet110 ResNet20	ResNet110 ResNet32	ResNet32*4 ResNet8*4	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
CC	73.56	72.21	69.63	69.48	71.48	72.97	70.71
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
FSP	72.91	-	69.95	70.11	71.89	72.62	70.23
NST	73.68	72.24	69.60	69.53	71.96	73.30	71.53
CRD	75.48	74.14	71.16	71.46	73.48	75.51	73.94
SKD	**75.75**	**75.06**	**72.08**	**72.12**	**74.09**	**76.40**	**74.17**

Table 4: ImageNet experiments with Top1 accuracy.

CE	KD	ES	SP	CC	CRD	AT	SKD
69.8	69.20	71.40	70.62	69.96	71.38	70.70	**72.80**

Dataset 1) *CIFAR-100* (Krizhevsky et al., 2009) is a relatively small data set and is widely used for testing various deep learning methods. CIFAR-100 contains 50,000 images in the training set and 10,000 images in the evaluation set, with 100 fine-grained categories. 2) *ImageNet* (Deng et al., 2009) is a much larger dataset than CIFAR-100. ImageNet contains 1.2M images for training and 50K for validation, with 1,000 fine-grained categories.

CIFAR Experimental settings We ran a total of 240 epochs for all methods. The learning rate was initialized as 0.05, then it decayed by 0.1 every 30 epochs after 150 epochs. For MobileNetV2, ShuffleNetV1 and ShuffleNetV2, we use a learning rate of 0.01 as this learning rate is optimal for these models in a grid search, while 0.05 is optimal for other models. For both vanilla KD and SKD, we set the temperature as 4, weight as 0.9, and cross-entropy as 0.1 for all settings.

ImageNet Experimental settings ResNet18 was used as the student for all methods. We applied the same training settings (e.g., learning rate, training epochs) as Heo et al. (2019). The teacher network had been trained in advance of the experiments and was fixed during training. The experiment requires 2 RTX 3090 GPU resources and takes around 40 hours.

3.1 Main Results

Baselines We selected various SOTA KD methods to evaluate the performances of SKD: 1) Knowledge defined from intermediate layers: FitNet (Romero et al., 2015), AT (Zagoruyko & Komodakis, 2017), SP (Tung & Mori, 2019), PKT (Passalis & Tefas, 2018), FT (Kim et al., 2020), and FSP (Yim et al., 2017); 2) Knowledge defined via mutual information: CC (Peng et al., 2019), VID (Ahn et al., 2019), CRD (Tian et al., 2020); 3) Structured Knowledge: RKD (Park et al., 2019); and 4) Knowledge from logits: KD (Hinton et al., 2015), NST (Huang & Wang, 2017), ES (Cho & Hariharan, 2019), and TA (Mirzadeh et al., 2019)

CIFAR-100 Table 3 and 9 shows that SKD always has higher accuracy than all other methods. In some situations (e.g. those where teacher/student is WRN-40-2/WRN-40-1 or ResNet110/ResNet32), the performances of SKD were even very close to those of the teacher.

ImageNet All experiments reported in Table 4 used ResNet34 as the teacher and ResNet18 as the student. Table 4 shows that SKD exceeds all of the previous SOTA by a large margin on ImageNet.

Table 5: Performance degradation problem on ImageNet.

Teacher	Method	Accuracy	Teacher	Method	Accuracy
ResNet34	KD	69.43	ResNet101	KD	68.91
	ES	70.98		**SKD**	**72.85**
	SKD	**72.80**			
ResNet50	KD	69.05	ResNet152	KD	68.84
	TA	70.65		TA	70.59
	ES	70.95		ES	70.74
	SKD	**73.01**		**SKD**	**72.70**

Table 6: Performance degradation problem on CIFAR-100. Student is ResNet14. SKD achieves lower training loss and higher accuracy. The entropy gap between the distilled student and teacher is also reduced significantly. Temperature is set to 4 in vanilla KD.

		ResNet20	ResNet32	ResNet44	ResNet56	ResNet110
Training loss	Vanilla KD	1.1	1.7	2.1	2.5	3.3
	SKD	**0.9**	**1.2**	**1.3**	**1.4**	**1.6**
Test acc	Vanilla KD	67.4	68.2	68	67.5	67.1
	SKD	**68.2**	**68.7**	**68.9**	**68.8**	**69.2**
$G_{entropy}$	Vanilla KD	0.146	0.181	0.222	0.246	0.261
	SKD	**0.053**	**0.074**	**0.081**	**0.083**	**0.094**
G_f	Vanilla KD	0.042	0.053	0.063	0.068	0.074
	SKD	**0.029**	**0.042**	**0.041**	**0.045**	**0.044**

Table 7: Performance degradation experiments with various teacher width

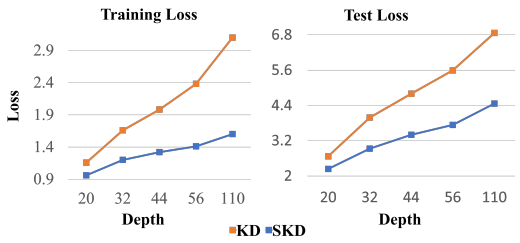
		WRN-16-2	WRN-16-3	WRN-16-4	WRN-16-5	WRN-16-6
Test acc	Vanilla KD	68.22	67.88	68.27	67.80	67.2
	SKD	**69.39**	**69.33**	**69.39**	**69.40**	**69.21**
$G_{entropy}$	Vanilla KD	0.063	0.082	0.095	0.115	0.134
	SKD	**0.045**	**0.058**	**0.079**	**0.083**	**0.09**
G_f	Vanilla KD	0.37	0.49	0.59	0.71	0.83
	SKD	**0.10**	**0.22**	**0.34**	**0.41**	**0.55**

Figure 2 shows the training process of vanilla KD and SKD. It is worth noting that SKD achieves comparable performance to KD's final performance after the first 30th epoch training.

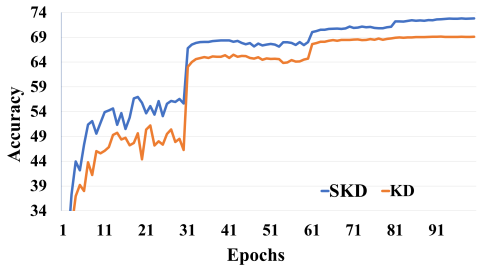
3.2 Performance Degradation Experiments

CFIFAR-100 We trained the ResNet14 with multiple teachers on the CIFAR-100 dataset. As shown in Table 6, the vanilla KD suffers from the performance degradation problem with oversized teachers (i.e., student accuracy continued decreasing when using teacher larger than ResNet32); while SKD continually improves the student performance as the teacher size is larger, which demonstrates that SKD can effectively alleviate the performance degradation problem. In addition, compared with vanilla KD, SKD significantly reduces both the entropy gap and free energy gap. We also added experiments where teacher models vary with width (Table 7, the student is WRN-16-1), which shows consistent results that the SKD outperformed the vanilla KD by a large margin.

ImageNet We compared SKD with two previous methods that aim to alleviate the degradation problem, Early Stop (Cho & Hariharan, 2019) (ES) and Teacher Assistant (Mirzadeh et al., 2019) (TA). Both of these two methods explicitly regularized the teacher capacity: 1) TA proposed to distill



(a) CIFAR-100 Training Loss.



(b) Training Process on ImageNet.

Figure 2: (a) As the teacher size grows, the loss of SKD increases slower than KD, which shows that SKD alleviates the performance degradation Problem. (b) When trained on ImageNet, SKD achieves comparable accuracy with KD on ImageNet at the 30th epoch.

the large teacher to an intermediate teacher and then distill to the student, so that each knowledge distillation step has a better match between student and teacher capacity; 2) ES methods use the early stopped teacher, the teacher capacity would be regularized by fewer training steps. Table 5 shows the degradation problem in ImageNet with ResNet18 as the student. We can see that SKD exceeds Early Stop and TA methods by a large margin in all teacher settings (Table 5). For example, when distilled by ResNet50 and ResNet152, the performance exceeded other methods by 2%. SKD achieves 73.01% accuracy, which is the best ResNet18 result that we know of.

4 Related Work

Buciluǎ et al. (2006) first proposed to compress a trained cumbersome model into a smaller model by matching the logits between them. Then Hinton et al. (2015) advanced this idea and formed a more widely used framework known as knowledge distillation (KD). Knowledge distillation tries to minimize the KL divergence between the soft output probabilities generated by the logits through softmax function. Different from Xu et al. (2020) that normalizes the features in the penultimate layer of the network to perform distillation, our methods perform normalization on the logits layer. Furthermore, knowledge distillation can also be regarded as a soft label training method. Specifically, previous studies have found that knowledge distillation helps to regularize the training of network. The relationship between KD and other regularization techniques (e.g., label smoothing) has been discussed in various works (Müller et al., 2019; Shen et al., 2021).

Although distillation has shown a great potential in many tasks, researchers found that larger teachers often unexpectedly harm the distillation performance, despite their more powerful ability (Cho & Hariharan, 2019; Mirzadeh et al., 2019). The performance degradation problem is particularly severe on ImageNet, resulting in poor performance of distilled student model. It was widely accepted that the capacity mismatch between teacher and student causes this problem Zhu & Wang (2021). Previous research proposed to regularize the teacher capacity to alleviate this problem heuristically. For example, Cho & Hariharan (2019) proposed to early stop the training of the teacher. Moreover, Mirzadeh et al. (2019) proposed to use a medium-size teacher assistant (TA) to perform a sort of sequence distillation. TA first learns from the teacher, then the student can learn from the TA. However, the accuracy of the early stopped teacher or TA is also lower than the original teacher.

5 Conclusion

The vanilla knowledge distillation overlooks the confidence gap between the student and the teacher, which may cause the performance degradation with oversized teachers. We presents the Spherical Knowledge Distillation framework, which address the performance degradation problem by reducing the confidence gap between the teacher and student. We validate the effectiveness of our method on CIFAR-100 and ImageNet. Experimental results show that SKD can effectively mitigate the performance degradation problem and produce competitive students.

REFERENCES

Sungsoo Ahn, Shell Xu Hu, Andreas C. Damianou, Neil D. Lawrence, and Zhenwen Dai. Variational information distillation for knowledge transfer. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9155–9163, 2019.

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*, KDD '06, pp. 535–541, New York, NY, USA, 2006.

Keshigeyan Chandrasegaran, Ngoc-Trung Tran, Yunqing Zhao, and Ngai-Man Cheung. Revisiting label smoothing and knowledge distillation compatibility: What was missing? In *International Conference on Machine Learning*, pp. 2890–2916. PMLR, 2022.

Defang Chen, Jian-Ping Mei, Hailin Zhang, Can Wang, Yan Feng, and Chun Chen. Knowledge distillation with the reused teacher classifier. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11933–11942, 2022.

Jang Hyun Cho and Bharath Hariharan. On the efficacy of knowledge distillation. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4793–4801, 2019.

Tejalal Choudhary, Vipul Mishra, Anurag Goswami, and Jagannathan Sarangapani. A comprehensive survey on model compression and acceleration. *Artificial Intelligence Review*, 53(7):5113–5155, 2020.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. Imagenet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009.

Prakhar Ganesh, Yao Chen, Xin Lou, Mohammad Ali Khan, Yin Yang, Hassan Sajjad, Preslav Nakov, Deming Chen, and Marianne Winslett. Compressing large-scale transformer-based models: A case study on bert. *Transactions of the Association for Computational Linguistics*, 9:1061–1080, 2021.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks, 2017.

Qiushan Guo, Xinjiang Wang, Yichao Wu, Zhipeng Yu, Ding Liang, Xiaolin Hu, and Ping Luo. Online knowledge distillation via collaborative learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11020–11029, 2020.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

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/CVF International Conference on Computer Vision*, pp. 1921–1930, 2019.

Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. *ArXiv*, abs/1503.02531, 2015.

Zehao Huang and Naiyan Wang. Like what you like: Knowledge distill via neuron selectivity transfer. *ArXiv*, abs/1707.01219, 2017.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pp. 4171–4186, 2019.

Jangho Kim, SeongUk Park, and Nojun Kwak. Paraphrasing complex network: Network compression via factor transfer, 2020.

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

Solomon Kullback. *Information theory and statistics*. Courier Corporation, 1997.

Yann LeCun, Sumit Chopra, Raia Hadsell, M Ranzato, and F Huang. A tutorial on energy-based learning. *Predicting structured data*, 1(0), 2006.

Kimin Lee, Honglak Lee, Kibok Lee, and Jinwoo Shin. Training confidence-calibrated classifiers for detecting out-of-distribution samples, 2018.

Gilbert Newton Lewis and Merle Randall. *Thermodynamics*. Number 44. Krishna Prakashan Media, 1963.

Seyed-Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, and Hassan Ghasemzadeh. Improved knowledge distillation via teacher assistant, 2019.

Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. When does label smoothing help? *Advances in neural information processing systems*, 32, 2019.

Frank Nielsen and Ke Sun. Guaranteed bounds on the kullback-leibler divergence of univariate mixtures using piecewise log-sum-exp inequalities. *arXiv preprint arXiv:1606.05850*, 2016.

Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3962–3971, 2019.

Nikolaos Passalis and Anastasios Tefas. Learning deep representations with probabilistic knowledge transfer. In *ECCV*, 2018.

Baoyun Peng, Xiao Jin, Jiaheng Liu, Shunfeng Zhou, Yichao Wu, Yu Liu, Dong sheng Li, and Zhaoning Zhang. Correlation congruence for knowledge distillation. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5006–5015, 2019.

Antonio Polino, Razvan Pascanu, and Dan Alistarh. Model compression via distillation and quantization. In *International Conference on Learning Representations*, 2018.

Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. Fastspeech: Fast, robust and controllable text to speech. *Advances in Neural Information Processing Systems*, 32, 2019.

Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. *CoRR*, abs/1412.6550, 2015.

Claude Elwood Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.

Zhiqiang Shen, Zechun Liu, Dejia Xu, Zitian Chen, Kwang-Ting Cheng, and Marios Savvides. Is label smoothing truly incompatible with knowledge distillation: An empirical study. *arXiv preprint arXiv:2104.00676*, 2021.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. In *International Conference on Learning Representations*, 2019.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. *ArXiv*, abs/1910.10699, 2020.

Frederick Tung and Greg Mori. Similarity-preserving knowledge distillation. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1365–1374, 2019.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Mengzhou Xia, Zexuan Zhong, and Danqi Chen. Structured pruning learns compact and accurate models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1513–1528, 2022.

Kunran Xu, Lai Rui, Yishi Li, and Lin Gu. Feature normalized knowledge distillation for image classification. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXV 16*, pp. 664–680. Springer, 2020.

Junho Yim, Donggyu Joo, Jihoon Bae, and Junmo Kim. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7130–7138, 2017.

Li Yuan, Francis EH Tay, Guilin Li, Tao Wang, and Jiashi Feng. Revisiting knowledge distillation via label smoothing regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3903–3911, 2020.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *ArXiv*, abs/1605.07146, 2016.

Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *ArXiv*, abs/1612.03928, 2017.

Yichen Zhu and Yi Wang. Student customized knowledge distillation: Bridging the gap between student and teacher. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5057–5066, October 2021.

A APPENDIX

A.1 THE SUM OF LOGITS

Hinton et al. (2015) presume that the $\sum_j z_j$ equals zero during training without further analysis. We conducted experiments on the CIFAR-100 dataset to verify this assumption. Table 8 shows that the logits are close to zero for all models used.

Table 8: The sum of logits

	ResNet20	ResNet32	ResNet44	ResNet56	ResNet110
Logits Sum	-5e-5	-4.7e-5	-5.7e-5	-7.9e-4	-6.1e-5
	WRN-16-1	WRN-16-2	WRN-16-3	WRN-16-4	VGG13
Logits Sum	-4.8e-5	-6.1e-6	-5.5e-5	-1.2e-5	-3.1e-5

A.2 MORE EXPERIMENTS ON CIFAR-100

Table 9: CIFAR-100 experiments when the teacher's architecture is significantly different.

Teacher Student	vgg13 MobileNetV2	ResNet50 MoblieNetV2	ResNet50 vgg8	ResNet32*4 ShuffleNetV1	ResNet32*4 ShuffleNetV2	WRN-40-2 ShuffleNetV1
Teacher	74.64	79.34	79.34	79.42	79.42	75.81
Student	64.60	64.60	70.36	70.50	71.82	70.50
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
CC	64.86	65.43	70.25	71.14	71.29	71.38
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
NST	58.16	64.96	71.28	74.12	74.68	74.89
CRD	**69.73**	69.11	74.30	**75.11**	75.65	76.05
SKD	68.62	**69.26**	**74.41**	75.08	**76.02**	**76.42**