Abstract

Knowledge distillation aims at obtaining a small but effective deep model by transferring knowledge from a much larger one. The previous approaches try to reach this goal by simply “logit-supervised” information transferring between the teacher and student, which somehow can be subsequently decomposed as the transfer of normalized logits and $l^2$ norm. We argue that the norm of logits is actually interference, which damages the efficiency in the transfer process. To address this problem, we propose Spherical Knowledge Distillation (SKD). Specifically, we project the teacher and the student’s logits into a unit sphere, and then we can efficiently perform knowledge distillation on the sphere. We verify our argument via theoretical analysis and ablation study. Extensive experiments have demonstrated the superiority and scalability of our method over the SOTAs.

1 Introduction

Recently, deep neural networks have achieved remarkable success in most of computer vision tasks. However, the increasing network depth in the current state-of-the-art models has resulted in a high computational burden and inference time. One direction to reduce the model size is knowledge distillation (KD) methods [8, 18, 19, 20, 23], which distills the knowledge from a large model (teacher) into a small model (student). The fundamental question of KD is “what knowledge is captured by the large model and can be transferred to the small model ?”. Two groundbreaking papers [2, 8] propose a knowledge transferring approach that uses the teacher’s logits (the output of the last linear layer) as the supervised information to train the students.

Similarity information between categories contained in teacher’s logits is generally believed the reason why distillation is effective[5, 8]. For instance, the image of cats is more similar to dogs than cars. Therefore, given a "cat" image, a trained network will assign a higher probability to the "dog" category than the "car" category. Knowledge distillation essentially transfers this similarity between teacher logits to the student.

The similarity of logits, as a vector, can be decomposed into the similarity of vector norm and the similarity of vector direction. We show that KD process tries to match the norm and direction in a mixed manner between the student and the teacher. Intuitively, it would be easier to match only the direction than match both. According to the success of the cosine similarity in the classification[13, 21], the direction of logits is sufficient to support the capacity of the model. In this paper, we show that the direction alone is capable of representing category similarity, and the norm will have a negative impact on knowledge distillation.

Therefore, we propose to project the logits of the teacher and the student into a spherical space by normalization. This makes the training of distillation optimized on a sphere, where $l^2$ norm is a
constant. The optimization constrained on the sphere becomes more smooth and simpler. Thus the model can concentrate on the real important learning part, the normalized logits, which captures the relative size between probability predictions for each category.

In summary, the main contribution of this paper has three-fold:

1. We reveal that the direction of logits is the key knowledge, and the norm of logits hinders the distilling process of this knowledge. We propose Spherical Knowledge Distillation (SKD) to transform logits to a unit sphere with the constant norm and perform KD on the sphere space.

2. We conduct extensive experiments to show the effectiveness of SKD method. We achieve better performance than SOTA methods in CIFAR100 and ImageNet with diverse student and teacher architectures.

3. Our method is simple to implement and can improve the performance of most existing KD methods. Our method is also robust to the temperature parameter and capacity gap between the student and teacher.

2 Related Work

Bucilua et al. [2] first proposed to compress a trained cumbersome model into a smaller model by matching the logits between them. Hinton et al. [8] advanced this idea and formed a more widely used framework known as knowledge distillation (KD). Hinton KD tries to minimize the KL divergence between the output probabilities generated by the logits through softmax. Hinton proposed to use a selected temperature applied on output logits to soften the distribution so that the loss can weigh more to the lower logits of the teacher.

In the later works, the main direction to improve the knowledge distillation is to transfer more knowledge from the teacher to the students, such as the intermediate representation [18] [23] [22] and relation among instances [17] [14] [20]. FitNets [18] uses the intermedia feature map of the teacher as the additional training hints for the student. The attention transfer method [23] makes the activation-based and gradient-based spatial attention map of the student consistent with the teacher. Peng et al. [17] transfers only the instance-level knowledge but also the correlation matrix between instances. Similarity-preserving KD [20] requires the student to mimic the pairwise similarity map between the instance features of the teacher.

Recent works [3] [10] also investigate the capacity mismatching between the teacher and the students, and small students fail to mimic much larger teachers. Cho et al. [3] propose stopping the training of the teacher to mitigate this problem. Inspired by curriculum learning, RCO [10] selects several checkpoints generated during the training of the teacher and makes the student gradually mimic those checkpoints of the teacher.

However, there are few works study and modify the origin knowledge distillation loss [2] [8]. We demonstrate that the training of the student can be more efficient by projecting the logits into a unit sphere. We also show that our method can alleviate the capacity mismatching problem.

3 Methodology

From the insight that direction, rather than the norm of logits, encodes the knowledge of classwise similarity, We propose to remove the influence of norm during distillation. In this section, we discuss the role of $l_2$ norm in distillation from a theoretical perspective. Specifically, We decompose the logits into two parts: the norm and the direction. The latter is expressed as a unit vector. We prove that the student learns the teacher from these two parts during distilling, and the learning of direction will be disturbed by the learning of the norm. The student can be trained much more efficiently on the unit sphere with a constant norm.

3.1 Hinton’s Distillation

Distillation is first proposed by Hinton [8]. Given two neural networks, a teacher network $T$, and a student network $S$, We denote the output logits of the teacher as $f^T(x)$, and output logits from the student as $f^S(x)$.
Knowledge distillation is an effective method to transfer knowledge into a small network. Generally, there are two losses used, MSE or KL divergence.

\[
\mathcal{L}_{MSE} = \sum_i (f^S_i(x) - f^T_i(x))^2
\]

(1)

\[
\mathcal{L}_{KL} = \sum_i p^T_i \log p^S_i, \text{ where } p_i = \frac{\exp(f_i(x)/\tau)}{\sum_j \exp(f_j(x)/\tau)}
\]

(2)

where \(\tau\) is the temperature parameter. Hinton points out that when the temperature is high, KL divergence will degenerate to MSE.

### 3.2 Analyzing Norm and Direction Loss

![Figure 1: The left elaborate the MSE loss of norm and normalized logits, where blue for norm loss and red for the latter, between the student and the teacher. The right shows the accuracy of the same student during distillation.](image1.png)

As the Figure 1 shows, the norm loss between student and teacher will change drastically during the training process. On the other hand, the MSE loss of normalized logits does not fluctuates that much. When the L2 norm loss stabilizes, the student’s accuracy begins to increase rapidly. This indicates that the norm part of distilling loss is the hard one to be optimized.

We conjecture the reason is that the norm is difficult for the student to learn because of the high variance in the norm of teacher among different data points. As shown in Figure 2 in binary classification case, the norm of the logits varies between the samples and the student have difficulty in mimicking this variance. However, after projected onto the sphere, it would be easier to match the logits between the teacher and the student. We valid this assumption in the ablation of the experiment.

### 3.3 Spherical Knowledge Distillation

Based on the above formula and experimental phenomena, we propose Spherical Knowledge Distillation (SKD). SKD aims to exclude the effects of the norm from distillation from two aspects. First, SKD normalizes the logits of the teacher. In this way, students do not need to learn the different norms of each training data of the teacher during distillation. Second, SKD also normalizes the logits of the student. With this setting, the norm will stay unchanged during distilling, so to stable the training process. Then SKD minimizes the KL divergence between the student and the teacher, after restoring the norm of teacher and student to the teacher’s norm space together. Finally, cross entropy loss with ground truth are added.
Table 1: CIFAR100 experiments

| Method | Teacher & Student | ResNet56 | ResNet20 | ResNet110 | ResNet32 | ResNet32*4 | vgg13 | vgg8 |
|--------|------------------|----------|----------|-----------|----------|-----------|-------|------|
| Teacher | 75.61  | 73.26    | 74.31    | 73.14    | 70.50    | 70.12    | 73.42 | 74.64 |
| Student | 74.11  | 72.50    | 72.61    | 73.09    | 71.23    | 71.90    | 71.43 | 71.48 |
| KD [8]  | 74.92  | 73.54    | 70.66    | 73.08    | 72.98    | 71.02    | 71.43 | 71.48 |
| FitNet [18] | 73.58 | 72.24    | 68.99    | 71.06    | 73.50    | 71.02    | 71.43 | 71.48 |
| SP [20] | 73.83  | 72.43    | 70.04    | 72.69    | 72.94    | 72.68    | 72.68 | 72.68 |
| CC [17] | 73.56  | 72.21    | 69.63    | 71.48    | 72.97    | 70.71    | 70.71 | 70.71 |
| VID [1] | 74.11  | 73.30    | 70.35    | 71.66    | 73.44    | 72.98    | 72.98 | 72.98 |
| RKD [15] | 73.35 | 72.22    | 69.61    | 71.82    | 71.90    | 71.48    | 71.48 | 71.48 |
| PKT [16] | 74.54 | 73.45    | 70.34    | 72.61    | 73.64    | 72.88    | 72.88 | 72.88 |
| AB [7]  | 72.50  | 72.38    | 69.47    | 70.98    | 73.17    | 70.94    | 70.94 | 70.94 |
| FT [11] | 72.91  | 71.59    | 69.84    | 72.37    | 72.86    | 70.58    | 70.58 | 70.58 |
| FSP [22] | 73.68 | 72.49    | 69.60    | 71.96    | 73.30    | 71.53    | 71.53 | 71.53 |
| NST [9] | 75.48  | 74.14    | 71.46    | 73.48    | 75.51    | 73.94    | 73.94 | 73.94 |
| SKD     | 75.75  | 75.06    | 71.89    | 73.09    | 76.40    | 74.17    | 74.17 | 74.17 |

Table 2: Imagenet experiments. All use ResNet18 as the student. The subscript denote the size of the teacher, and + denote training with 100 epochs and the reduced weight strategy.

To sum up, logits of teacher and student are all projected into the same constrained spherical space. SKD training convergence is faster than the regular knowledge distillation, and achieve much better performance than other methods.

The overall knowledge distillation loss for SKD:

\[ L_{SKD} = \sum_i \hat{p}_i^T \log \hat{p}_i^S, \text{ where } \hat{p}_i = \frac{\exp(f_i^T(x) * m/\tau)}{\sum_j \exp(f_j^T(x) * m/\tau)} \]  

\[ L = \lambda L_{CE}(\hat{p}_i^S, y) + (1 - \lambda) L_{SKD} \]  

where \( f_i^T(x) \) is the normalized logits and \( L_{CE} \) is the supervised loss, and \( \lambda \) is the trade-off parameter, and \( m \) is the average norm of the teacher.

In the experiments, our method is robust to the choice of the temperature \( \tau \), and less effort is needed to tune the hype-parameter.

4 Experiments

We conducted our experiments on two popular datasets, CIFAR100 and ImageNet. We prove the excellent performance of SKD through a broad range of teacher-student settings on CIFAR100 and ImageNet. In addition, we find our SKD can be used as a drop-in module to boost the performance of a range of knowledge distillation methods. Finally, we conduct two ablation experiments that help us to understand what happened to \( l^2 \) norm of students during the distillation.

Datasets (1) CIFAR100 [12] is a relatively small data set and is widely used for testing various of deep learning methods. CIFAR100 contains 50,000 images in the training set and 10,000 images in the dev set, divided into 100 fine-grained categories. (2) ImageNet [4] is a much harder task than CIFAR100. ImageNet contains 1.2M images for training and 50K for validation, that distributes in 1000 classes.

With all the experiment blow, we use \( \lambda \) as 0.9 and \( T \) as 4 without specifically mentioned.

4.1 Experiment on CIFAR100 & ImageNet

We achieve new STOAs in most of the experimental settings with our SKD. In order to test the effectiveness of our method, we conducted extensive experiments on CIFAR100 and ImageNet and
Table 3: Scalability Experiments on other methods.

|                | KD | FitNet | SP | CC | VID |
|----------------|----|--------|----|----|-----|
| ResNet56 w/o (S)KD | 69.06 | 69.21 | 69.67 | 69.63 | 70.38 |
| ResNet20 With KD | 70.66 | 71.06 | 71.58 | 71.56 | 71.24 |
| ResNet56 With SKD | 71.94 | 71.26 | 71.71 | 71.90 | 71.49 |
| ResNet20 w/o (S)KD | 70.34 | 69.47 | 69.95 | 69.60 | 71.16 |
| ResNet20 With KD | 71.52 | 71.28 | 70.97 | 71.02 | 71.63 |
| ResNet56 With SKD | 72.15 | 71.93 | 71.47 | 71.31 | 72.19 |

were successful in most experimental configurations. In particular, we exceeded the previous STOA on ImageNet by more than one point. We trained ResNet18 [6] to an acc of 73.01, close to the level of ResNet34, which is 73.3. To fully examine the performance of SKD, we compare with 13 other teacher-student framework methods.

**SKD on CIFAR100**

*Experimente settings* For the experiments on CIFAR100, we run total 240 epochs. The learning rate is initialized as 0.05, then decay by 0.1 every 30 epochs after the first 150 epochs. Temperature is set to 4, the weight of SKD or KD and cross-entropy is 0.9 and 0.1 for all the setting. As for the hyperparameters of other different distilling methods, we follow the setting of [19].

*Analysis* When the teacher and student share the same architecture, SKD has a stable improvement compared with all other methods.

**SKD on ImageNet**

*Experimente settings* We use ResNet18 as our backbone in ImageNet. We use 4 NVIDIA 1080ti with distributed learning to accelerate our experiment. For better performance, some experiments use 100 epochs, and others use 90 epochs. So the initial learning rate is set at 0.4, and decay by 0.1 in 30, 60, 80. A total 1024 batch size is used. Typically, our SKD with ResNet18 can finish within one day.

*Analysis* In the previous practice, the KD-based method often failed to bring too much improvement on ImageNet. The previous STOA is 71.34 accuracy with ResNet18, compared to 69.8 baseline trained with regular training with ground truth. In contrast, our method achieves an accuracy of 72.2 with 90 epochs, which is over two points higher than using cross-entropy. Furthermore, We found that until the end of the training, the accuracy of the model still has an upward trend. So to further push the limit, we add another 10 epochs and apply a strategy that gradually reduces the weight of distillation loss during the training process. This strategy is inspired by [3]. To be specific, we set λ to 0.9 before 60 epoch, and 0.5 before 80 epoch, and 0.1 with the rest of epochs.

Another thing worth noting is that the previous work [3] finds that KD often fails in distilling from regular ResNet50 or larger models. On the contrary, We find our SKD successfully distilled ResNet18 from teachers that up to ResNet101.

Finally, we achieve 73.01 accuracy, which is close to 73.3 of ResNet34.

### 4.2 Scalability Experiment

A lot of KD-based methods compared above are commonly used with a combination of three losses, which include their specially designed loss, plus KD loss and cross-entropy loss. Because our method is easy to implement, we are curious if using our method directly on other methods will improve these methods. We conducted experiments that replace the KD loss with our SKD loss, to test the scalability of SKD.

*Experiment settings* We use ResNet56 as the teacher and ResNet20 as the student to perform all these experiments. Hyperparameter is consistent with those used in CIFAR100. "The total loss of this experiment consists of three parts: cross-entropy, SKD loss, and their own specific loss. "

*Analysis* We find that SKD can improve performance on most of them. To be noticed, normalization on logits may change the representation consistency between the output logits and intermedia features,
Table 4: Ablation study on $l^2$ norm. "Norm loss" means student learned from the norm of the teacher. ‘*’ means KD with teacher’s logits normalized.

|       | CE   | Norm_loss | KD  | KD* | SKD |
|-------|------|-----------|-----|-----|-----|
| res20 | 69.50| 69.34     | 70.7| 71.2| 71.9|

thus harms the performance of some feature-learning based methods. The integration of SKD and feature-based KD may require further research. Still, SKD improves most of them. This result shows the high scalability and "easy-to-transfer ability" of SKD.

4.3 Ablation Experiment

To study the role of norm played during distillation, we conduct two experiments learning directly on the norm or exclude the teacher’s norm from distillation. Through these two experiments, we demonstrate that the norm of the teacher is hard to learn during the distilling process.

Directly Learn Norm of The Teacher Experiment

This experiment was meant to examine if the transferring of the norm will bring improvement to student performance.

Experiment settings We use resnet20 as the student and resnet56 as the teacher in this experiment. The student is trained to minimize the MSE loss of $l^2$ logits norm between the student and the teacher. The student is also trained by cross-entropy loss with ground truth. The weight of these two losses is set 0.05 and 1, respectively. The weight of the norm loss larger than 0.05 will harm more to the performance of the student.

Analysis "We found that the decline curve of norm in this experiment is similar to that in knowledge distillation". In the end, the accuracy of this experiment is not as good as that training with only ground truth. This indicates that directly transferring the norm from teacher to student is harmful to the student.

Learn Normalized Logits From Teacher Experiment

We further test if we only normalize the teacher, whether the performance is improved or harmed.

Experiment settings We conduct a distillation experiment that completely eliminates the teacher’s norm by normalizing the teacher’s logits and then multiply it by its average $l^2$ norm. Other settings are consistent with Hinton’s knowledge distillation. So all the teacher’s logits will be projected onto a sphere, and all information contained in the teacher’s $l^2$ norm will disappear. But the student still has to learn this constant norm as before.

Analysis In this way, students do not have to learn the complicated distribution of teacher’s $l^2$ norm among the training dataset. This method achieves much better results than regular KD, which demonstrates that the norms of teacher logits actually harm the distilling process. However, the student still has to adjust its norm to match the teacher norm, even now a constant. Thus the performance is not as good as SKD.

5 Analysis

Capacity gap and temperature are two aspects that affects the success of distillation. The former is mentioned in both [3] and [10], which states a phenomenon that when the size of the teacher becomes larger, the student performance obtained by distillation becomes worse. The latter, on the other hand, is used to soften the distribution of the teacher, so the student can learn more from the relative small numbers of logits, which is believed contains the knowledge of class-wise similarity. We study these two problems by decomposing the loss into MSE loss of norm and normalized logits like the former sectors. And we find Our SKD is robust to both capacity gap and temperature.

Robustness to Temperature

Temperature is used to soften the distribution, then the student can pay more attention to the small number of logits. But negative numbers in logits is not supervised too much during training, which could contains much noise [8]. So in practice, we need to balance these two aspects by selecting an appropriate temperature. We demonstrate that SKD is more robust with the choice of temperature.
Figure 3: Comparison between KD and SKD with respect to the change of capacity and temperature. The MSE loss of $l_2$ norm and normalized logits between teacher and student are presented.

**Experiment Settings** We conducted KD and SKD both on CIFAR100 dataset, with resnet20 as student and resnet56 as teacher. We logged the MSE loss between $l_2$ norm and normalized logits in both training process on validation dataset. The results are shown on Figure 3. We also plotted error rate to show the relevance.

**Analysis** With KD, we observed a high correlation between norm loss and error rate. When the norm loss is the smallest, the error rate is also very low. As the norm loss becomes larger, the error rate also starts to increase. In general, a good temperature will balance the norm loss and normalized logits loss, to get good performance.

On the other side, SKD is more robust to temperature, that achieves good performance even with temperature set at 80.

**Robustness to Capacity Gap**

Due to the limitations of representation ability, it is difficult for student to learn from a oversized teacher. We demonstrate that this gap can reflect in both the loss of norm and the loss of normalized logits between students and teachers. And SKD can alleviate this problem, thereby enjoying more performance improvements brought by larger teachers.

**Experiment settings** We conducted distillation with resnet14 as student. In terms of teachers, we choose a range of teachers, that including resnet20, resnet32, resnet44, resnet56 and resnet110. Hyperparameters are set as before in CIFAR100 experiments. We still log the MSE loss of $l_2$ norm and normalized logits between the student and the teachers.

**Analysis** From the figure 3, we can find that both norm and normalized logtis loss continue to increase as the teachers become larger. This can be the sign that capacity gap becomes larger. When the teacher is larger than resnet32, the damage caused by this gap overwhelms the advantages of the bigger teacher’s stronger performance, so the accuracy of student begins to decline.

At the same time, as the size of the teacher increases, the performance of SKD trained students become better. We think this is because SKD restricts learning to the sphere and does not need to learn teacher’s complicated norms, thus alleviating the capacity gap problem. We can also observe similar phenomena on Imagenet experiment, where we successfully distilled resnet101 to resnet18. This demonstrate that SKD can effectively learn more knowledge than KD from a larger network. While we need to select the most suitable teacher size in KD, we can choose to use a larger and therefore stronger network as the teacher in SKD.

6 Conclusion

We propose Spherical Knowledge Distillation (SKD), a very simple yet powerful framework that allows a small network to effectively approximate the performance of a much bigger one, showing a broad prospect in compressing networks. We demonstrate that the norm of the output logits will hinder the knowledge distillation. Therefore, we project the student and the teacher’s logits to the sphere space and only distill the direction of the logits. We conduct extensive experiments to show the effectiveness of SKD method, and our method can be easily plugged into most of the previous works.
Broader Impact

This work does not present any foreseeable societal consequence

References

[1] 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), pages 9155–9163, 2019.

[2] Cristian Buciluă, 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, page 535–541, New York, NY, USA, 2006.

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

[4] 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, pages 248–255, 2009.

[5] Tommaso Furlanello, Zachary C. Lipton, Michael Tschanne, Laurent Itti, and Anima Anandkumar. Born again neural networks, 2018.

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jun Sun. Deep residual learning for image recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 770–778, 2016.

[7] Byeongho Heo, Minsk Lee, Sangdoo Yun, and Jin Young Choi. Knowledge transfer via distillation of activation boundaries formed by hidden neurons. ArXiv, abs/1811.03233, 2019.

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

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

[10] Xiao Jin, Baoyun Peng, Yichao Wu, Yu Liu, Jiaheng Liu, Ding Liang, Junjie Yan, and Xiaolin Hu. Knowledge distillation via route constrained optimization. 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pages 1345–1354, 2019.

[11] Jangho Kim, Seonguk Park, and Nojun Kwak. Paraphrasing complex network: Network compression via factor transfer. In NeurIPS, 2018.

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

[13] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface: Deep hypersphere embedding for face recognition. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 6737–6746, 2017.

[14] Yufan Liu, Jiajiong Cao, Bing Li, Chunfeng Yuan, Weiming Hu, Yangxi Li, and Yunqiang Duan. Knowledge distillation via instance relationship graph. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 7089–7097, 2019.

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

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

[17] Baoyun Peng, Xiaojun Xie, Jiuhong 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), pages 5006–5015, 2019.

[18] 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.

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

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

[21] Hai Jun Wang, Yitong Wang, Zuo-Feng Zhou, Xing Ji, Zhifeng Li, Dihong Gong, Jingchao Zhou, and Wenwu Liu. Cosface: Large margin cosine loss for deep face recognition. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 5265–5274, 2018.
[22] 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)*, pages 7130–7138, 2017.

[23] 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.