Distilling Representational Similarity using Centered Kernel Alignment (CKA)

Aninda Saha[1,2]
a.saha@uqconnect.edu.au
Alina Bialkowski[1]
alina.bialkowski@uq.edu.au
Sara Khalifa[2]
sara.khalifa@data61.csiro.au

[1] School of ITEE
The University of Queensland
Brisbane, Australia

[2] Distributed Sensing Group
Data61, CSIRO
Brisbane, Australia

Abstract

Representation distillation has emerged as an effective knowledge distillation (KD) technique, which involves the transfer of an inter-example similarity matrix. However, existing methods use inadequate normalisation techniques combined with euclidean distance-based loss functions to distill inter-example similarity matrices. Such approaches are not invariant to uniform feature scaling, which is a key property for neural network similarity metrics. Therefore, we propose a novel loss function for representation distillation by adapting Centered Kernel Alignment (CKA), which computes the cosine similarity between the student and teacher's centered and normalised inter-example similarity matrices. We compare our proposed loss function against three popular representation distillation techniques, demonstrating CKA's outperformance on three benchmark image classification datasets. Our results reveal that distilling a centered and normalised distribution of the similarity matrix using the proposed CKA-based loss function is more effective than existing representation distillation techniques.

1 Introduction

Deep learning (DL) models play a key role in solving modern day computer vision problems, such as object detection and image classification. However, these models are very computationally expensive, making them difficult to deploy using off-the-shelf computing platforms. Therefore, model compression techniques are often used to reduce their compute requirements. Knowledge distillation (KD) has emerged as a popular end-to-end model compression technique whereby a compact student network learns to mimic a larger, more performante teacher network. KD is a popular choice due to its hardware and framework agnostic approach to model compression. Within KD, there are three main branches, including response-based, feature-based and similarity-based distillation [8]. Response-based distillation transfers knowledge from the output of the networks, feature-based distillation focuses on mimicking features from intermediate layers, whereas similarity-based distillation is about transferring inter-example or inter-layer similarities.

Response-based distillation involves distilling the output of the teacher model, however, it remains limited in the amount of knowledge that it can distill as it is restricted only to

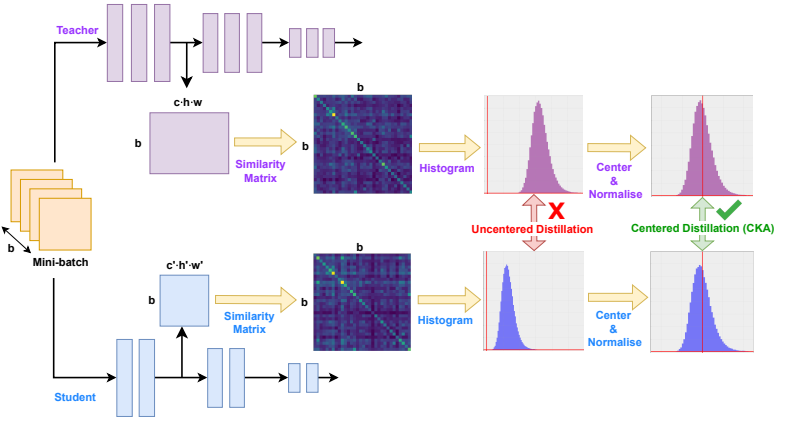

Figure 1: We propose a novel CKA-based loss function for representation distillation. Unlike existing methods which focuses on improving the quality of similarity matrices, our method shown above improves distillation by focusing on how they are distilled. CKA computes the cosine similarity between two centered and normalised similarity matrices, which outperforms currently used euclidean distance-based loss functions.

the output [8]. While feature-based distillation, also known as hint learning, theoretically provides access to knowledge from the teacher's latent space, how and where to effectively distill knowledge remains an open research question [8, 12]. One of the main obstacles is that due to the capacity gap between the student and teacher, directly mimicking the teacher's features is often infeasible for the student [16, 20]. Another significant drawback of feature matching is that it only supports distillation between equidimensional feature maps. This is considered a prohibitive issue since corresponding feature maps from different models can often have different shapes [8]. Several works [8, 21, 25] have addressed this by using adaptation layers, albeit at the expense of additional setup and compute requirements.

Upon close inspection, [24] and [20] discovered that semantically similar images elicited similar activations from different layers of a network. While this may sound intuitive, it led the authors to the powerful conclusion that representation distillation, where an inter-example similarity matrix is distilled, is a better knowledge formulation technique than direct feature matching, which performs point-to-point activation matching [20]. The work of [19] also demonstrated that capturing inter-example similarities provides richer context and information for the student to learn from. In addition to improving the quality of knowledge distilled, this approach also allows for the distillation between non-equidimensional feature maps as the resulting inter-example similarity matrices are equidimensional.

The work in [24] uses the Gram matrix as its inter-example similarity matrix, which calculates dot products between every pair of examples. [20] later proposed alternate kernel functions to compute the inter-example similarity matrix. However, both [24] and [20] distill their similarity matrices with the distance-based L2 norm loss, effectively forcing the student to mimic the scale of the teacher's features. This is at odds with some key properties, such as invariance to isotropic scaling, ideal for robust neural network similarity metrics [4, 13]. To address these properties, [13] proposed the use of Centered Kernel Alignment (CKA) as a robust similarity metric for neural networks.

Inspired by the success of CKA as a similarity metric, we propose a CKA-based loss function for distilling intermediate similarity matrices. As shown in Figure 1, CKA first centers and normalises the similarity matrices and then computes their cosine similarity. Due to the use of a similarity-based loss function as well as its invariance to isotropic scaling property [13], we demonstrate that CKA is a superior loss function compared to current representation distillation techniques, namely Similarity-Preserving (SP) KD [24], Correlational Congruence (CC) [20] and Relational Knowledge Distillation (RKD) [19], all of which make use of distance-based loss functions. As such, the main contributions of this paper are summarised as follows:

- We propose a novel loss function for representation distillation using CKA, which first centers and normalises the Gram matrix before distilling using the cosine similarity metric;
- We demonstrate the intuition behind the invariance to feature scaling properties, and hence our motivation for using CKA as a loss function; and
- We conduct a comprehensive study across three popular image classification datasets and compare against three popular techniques in representation distillation literature.

2 Background and Related Works

To motivate the use of CKA as a loss function, we first outline some key background works and expand the details of related works. In Subsection 2.1, we discuss some of the key works in KD literature and outline some of their key successes and challenges. In Subsection 2.2, we highlight the methodology employed by [24] in distilling the representational knowledge between networks. In Subsection 2.3, we discuss the mathematical intuition for CKA and how it employs a kernel function called Hilbert-Schmidt Independence Criterion (HSIC) [9] to improve representation distillation by centering and normalising the Gram matrices.

2.1 Knowledge Distillation (KD)

Taking inspiration from [2], KD research was sparked by [11] discovering that models can benefit from the knowledge of higher performing models, particularly if this knowledge is formulated and distilled appropriately. Compared to [2], which attempted to match the output logits of the student and teacher, [11]'s formulation of KD involved distilling softened class probabilities, which the authors expounded contains a more accurate representation of inter-class similarities.

Myriad of works followed that attempted to distil knowledge from different parts of the network to improve the performance of compact models. In [21], the idea of hint learning to match intermediate features was proposed, making use of adaptation layers to ensure shape compatibility. In [28] the idea of distilling an attention map was proposed, which eliminated hint learning's barrier of channel mismatch, but still requires matching in their spatial dimensions. [27] proposed the flow of solutions procedure (FSP) which computes the Gram matrix between the features of two different layers to measure the flow of knowledge between the two layers was proposed. Most of these methods, however, suffer from the challenge of dimension mismatch and require expensive workarounds.

Several works propose representation distillation [19, 20, 23, 24], whereby structural information between examples is transferred using inter-example similarity matrices rather than distilling features directly. [24] proposed Similarity-Preserving (SP) KD showing that rich contextual knowledge can be derived from an activation map by quantifying the similarity of each example with every other example in the mini-batch. SP uses the Gram matrix of the activation map, which produces a dot-product based similarity-matrix. Since the mini-batch has a fixed size b, the resulting Gram matrix always has the shape $b \times b$ for both student and teacher, making distillation much simpler. [20] has since proposed Correlational Congruence (CC) which uses alternate kernel functions to compute the inter-example similarities, including a Taylor-series approximation of the Gaussian Radial Basis Function (RBF). [19] have also proposed Relational Knowledge Distillation (RKD), which uses the distance and angle between each example pairs to derive their similarity matrices. However, these works distil their similarity matrices using euclidean distance-based loss functions such as the L2 norm loss or the Huber loss, neither of which are invariant to isotropic scaling, a property that improves the robustness of neural network similarities metrics [13].

2.2 Similarity Preserving (SP) KD

In this subsection we discuss the mathematical overview of SP in order to distinguish our CKA-based approach, using a consistent set of notations defined here. Representation distillation techniques, including SP, require access to one or more pairs of intermediate activation maps from the teacher and student networks. The activation map from layer l of the teacher is denoted as $A_T^{(l)} \in \mathbb{R}^{b \times c \times h \times w}$, whereas the activation map of layer l' of the student is denoted as $A_S^{(l')} \in \mathbb{R}^{b \times c' \times h' \times w'}$. The channel, height and width dimensions of the teacher are denoted by c, h and w respectively, whereas c', h' and w' denote that of the student. The mini-batch size is denoted by b.

The authors in [24] discovered that class distributions of differently initialised networks showed a similar pattern, leading the authors to propose inter-example similarities as an important source of knowledge for students to learn from. SP uses the outer product of the activation maps $A_T^{(l)}$ and $A_S^{(l')}$. This is calculated by first flattening the features into a matrix shaped $b \times chw$ and then multiplying the resulting matrix with its transpose, as shown in Equation 1.

$$Q^{(l)} = Flatten(A^{(l)}) \in \mathbb{R}^{b \times chw}; \widetilde{G}^{(l)} = Q^{(l)} \cdot Q^{(l)^T} \tag{1}$$

The resulting matrix $\widetilde{G}^{(l)}$ is a $b \times b$ Gram matrix, whereby each entry (i, j) is the dot product between the activation vectors of the ith and the jth examples in the mini-batch. However, since the dot product is a non-normalised measure of similarity between two vectors, the Gram matrix cannot be meaningfully distilled to the student without centering and normalising it first. The authors attempt to address the normalisation problem by applying row-wise normalisation, as shown in Equation 2, where $\|\cdot\|_2$ is the L2 norm.

$$G^{(l)} = \widetilde{G}^{(l)}_{[i,:]} / \|\widetilde{G}^{(l)}_{[i,:]}\|_2 \tag{2}$$

A Gram matrix is computed for the teacher and student, $G_T^{(l)}$ and $G_S^{(l)}$ respectively, which are used to calculate the SP loss by taking the Frobenius norm, $\|\cdot\|_F$, of the difference between the two matrices, as shown in Equation 3.

$$G^{(l)} = \frac{1}{b^2} \sum_{l,l' \in \mathcal{L}} \|G_T^{(l)} - G_S^{(l')}\|_F^2 \tag{3}$$

While this approach can transfer representational knowledge, due to an inadequate normalisation approach which does not normalise the columns of the similarity matrix, the student is penalised for differently scaled columns. Moreover, SP does not center the representational matrix, which again unnecessarily penalises the student for having a different scale, despite maintaining a similar distribution. [7] shows that the shape of the distribution strongly influences class separation. This limitation is addressed by our proposed CKA loss function, which distills a centered and normalised matrix using the cosine similarity metric.

2.3 Centered Kernel Alignment (CKA)

Centered Kernel Alignment (CKA) has been proposed in [13] as a robust way to measure representational similarity between neural networks. The authors present that three key properties must hold for a good neural network similarity metric: (i) invariance to isotropic scaling; (ii) invariance to orthogonal transformations; and (iii) non-invariance to invertible linear transformations [13]. Invariance to isotropic scaling simply means that uniformly scaling the input vectors does not affect the measure of similarity between them. This property is important because a model's discriminative capability is dependent on the distribution of its features rather than its scale, which is inconsequential for class separation [17, 18].

However, this key property is missing from previous works [19, 20, 24], which penalise students for not strictly mimicking the absolute values in the teacher's Gram matrix, even though the shape of distribution is more important. CKA can be used to address these challenges by applying a centering trick and normalisation, which ensures the focus of the optimisation is on the shape of the distribution, rather than the raw values in the Gram matrix. Furthermore, CKA computes the cosine similarity of the centered and normalised matrix, which produces a far more robust measure of similarity than euclidean distance-based metrics such as L2 or Huber losses [1]. CKA makes use of the Hilbert-Schmidt Independence Criterion (HSIC) proposed by [9] to estimate the similarity between the Gram matrices $G_S^{(l)}$ and $G_T^{(l)}$ by centering them first. The formula for HSIC is given by:

$$HSIC(G_S, G_T) = \frac{1}{(n-1)^2} tr(G_S H G_T H) = \frac{1}{(n-1)^2} \langle vec(G_S H)\ ,\ vec(G_T H) \rangle \tag{4}$$

where H is the centering matrix $H_n = I_n - \frac{1}{n}\mathbb{1}\mathbb{1}^T$. In contrast to SP's normalisation approach, HSIC's centering matrix performs row and column normalisation in tandem. While HSIC satisfies properties (ii) and (iii) above, it is not invariant to isotropic scaling on its own. To achieve this, the HSIC similarity between G_S and G_T is normalised as follows [4, 5]:

$$CKA(G_S, G_T) = \frac{\text{HSIC}(G_S, G_T)}{\sqrt{\text{HSIC}(G_S, G_S) \cdot \text{HSIC}(G_T, G_T)}} \tag{5}$$

In order to dissect the different components of CKA for our ablation study in Subsection 4.5, we identify that if we remove the centering process in Equation 4, CKA essentially boils down to the cosine similarity function. This equivalence is demonstrated in Equation 6, whereby *CKA_WC* represents "CKA without centering".

$$\begin{aligned} CKA\_WC(G_S, G_T) &= \frac{\frac{1}{(n-1)^2} vec(G_S) \cdot vec(G_T)}{\sqrt{\frac{1}{(n-1)^2} vec(G_S) \cdot vec(G_T)} \sqrt{\frac{1}{(n-1)^2} vec(G_S) \cdot vec(G_T)}} \\ &= \frac{vec(G_S) \cdot vec(G_T)}{\sqrt{vec(G_S) \cdot vec(G_T)} \sqrt{vec(G_S) \cdot vec(G_T)}} \\ &= cosine\_similarity(G_S, G_T) \end{aligned} \tag{6}$$

Due to its use of the cosine similarity, the resulting CKA metric is a value between 0 and 1 that essentially measures the similarity between the student-teacher pair's centered Gram matrices. [1] discusses the superiority of cosine similarity over distance-based losses. This motivates our proposal of CKA as a novel loss function to improve the discriminative capacity of the student, while allowing it to learn features at a different scale to the teacher.

3 Methodology

In this section, we outline the mathematical formulation of CKA as a novel loss function for distilling representational knowledge to students more effectively. To address the limitations of previous representation distillation techniques, such as SP, CC and RKD, we propose the use of CKA as a loss function, which ensures that the centered similarity matrix is distilled rather than forcing the student to mimic the teacher's similarity matrix with a different scale. This is done using the centering and normalisation strategies outlined in Equations 4 and 5.

Since CKA produces a similarity metric between 0 and 1, where 0 represents no similarity and 1 represents perfect similarity, we propose the loss function in Equation 7 which optimises towards a CKA value of 1.

$$\mathcal{L}_{CKA} = 1 - CKA(G_S, G_T) \tag{7}$$

The overall optimisation function, including the regular classification loss and the CKA loss is given by:

$$\mathcal{L}_{total} = \mathcal{L}_{CE} + \lambda_{CKA} \mathcal{L}_{CKA} \tag{8}$$

where $\mathcal{L}_{CE}$ is the cross-entropy loss applied to the classification vector and $\mathcal{L}_{CKA}$ is the CKA loss applied to the Gram matrices of the intermediate activation maps. λ_{CKA} is the weighting factor for the CKA loss, which we set to 1 for all of the experiments in this paper.

4 Results

4.1 Experimental Setup

Due to their pervasiveness in KD literature as well as computational efficiency, we conduct our experiments using the ResNet [10] and MobileNetV2 [22] architectures. To ensure reproducibility, we use a fixed seed of 2 on Python's PyTorch, NumPy and Random modules.

We first trained our baseline models ResNet18 (R18), ResNet34 (R34), ResNet50 (R50), ResNet101 (R101) and MobileNetV2 (MV2) on three popular benchmark image classification datasets: CIFAR100 [14], TinyImageNet [15] and ImageNet-1k [6]. In our KD experiments, we evaluate our proposed CKA-based technique and compared its performance to

three main state-of-the-art techniques SP, CC and RKD, treating R18 and MV2 as our two students and the remaining networks as their teachers.

The CIFAR100 and TinyImageNet experiments were run using a batch size of 128 images on a single Nvidia Tesla V100 GPU for 200 epochs. Meanwhile, the ImageNet-1k experiment were run in parallel across 2 V100 GPUs for 100 epochs with a batch size of 128 images in each GPU, which creates an effective batch size of 256. A learning rate scheduler was used to start with an initial learning rate of 0.1, which was reduced to 0.01 and 0.001, at $0.5 \times epochs$ and $0.75 \times epochs$ respectively.

We used the average pooling layer of both student and teacher networks as the distillation location due to their optimal performance in feature matching literature [8, 26]. For each mini-batch of 128 images, the activation map of the average pooling layer was flattened to produce a matrix of dimension $\mathbb{R}^{128 \times chw}$. This feature matrix was used to derive a similarity matrix using different metrics, which was then used to compute the SP, CC, RKD and CKA losses.

We also conducted an ablation study (see Subsection 4.5) whereby CKA loss was applied without centering the Gram matrices to empirically demonstrate the benefits of the centering component of CKA. While not explicitly a part of the ablation study, it is worth pointing out that the benefits of using cosine similarity over euclidean distance-based losses can be seen by comparing the CKA-WC experiment with the SP experiment, both of which apply their respective losses on the Gram matrices.

4.2 Results on CIFAR100

Table 1 shows the full set of experimental results for CIFAR100, which clearly shows that our proposed CKA-based loss function significantly outperforms state-of-the-art techniques SP, CC and RKD in distilling representational knowledge. The R18 student even manages to beat or perform close to its teacher across all experiments. Therefore, the results validate our hypothesis that it is better to teach students the shape of the distribution of the similarity matrices rather than their raw values. This is because the flexibility allows them to learn their own features within the constraints of their limited parameter space.

Architecture		**Model Accuracy (%)**					
Student	Teacher	Student	SP	CC	RKD	CKA	Teacher
R18	R18	77.98	78.22	78.54	78.63	**79.14**	77.98
R18	R34	77.98	78.41	78.94	78.83	**79.35**	78.97
R18	R50	77.98	78.15	78.46	78.32	**79.15**	79.19
R18	R101	77.98	78.27	78.69	78.47	**79.32**	79.68
MV2	R18	73.12	73.75	74.10	73.98	**75.30**	77.98
MV2	R34	73.12	73.47	74.04	73.89	**75.24**	78.97
MV2	R50	73.12	73.39	74.18	74.27	**75.19**	79.19
MV2	R101	73.12	73.32	74.14	74.00	**75.11**	79.68

Table 1: Results on CIFAR100 with two students R18 and MV2, distilled from a series of teachers using four main techniques: SP, CC, RKD and CKA (ours).

4.3 Results on TinyImageNet

Table 2 demonstrates a similar pattern of results as seen for CIFAR100 in Subsection 4.2. One notable difference is that the magnitude increase in CKA's accuracy compared to other methods is more pronounced for Tiny-ImageNet, indicating the effectiveness of CKA for small-scale natural image datasets.

Architecture		**Model Accuracy (%)**					
Student	Teacher	Student	SP	CC	RKD	CKA	Teacher
R18	R18	64.07	64.26	64.52	64.42	**65.26**	64.07
R18	R34	64.07	64.58	64.32	64.69	**66.14**	67.72
R18	R50	64.07	64.70	64.15	64.92	**66.24**	70.25
R18	R101	64.07	64.28	64.29	64.64	**65.87**	71.43
MV2	R18	62.27	62.46	62.58	62.50	**63.12**	64.07
MV2	R34	62.27	62.58	62.79	62.64	**63.35**	67.72
MV2	R50	62.27	62.67	62.85	62.45	**63.44**	70.25
MV2	R101	62.27	62.33	62.67	62.57	**63.27**	71.43

Table 2: Results on Tiny-ImageNet with two students R18 and MV2, distilled from a series of teachers using four main techniques: SP, CC, RKD and CKA (ours).

4.4 Results on ImageNet

ImageNet-1k is an important benchmark dataset due to its widespread usage in literature and its large-scale representation of one thousand classes across 1.2 million images. Table 3 shows our experimental results on ImageNet-1k, where each experiment took approximately 7-GPU days to train for 100 epochs. The results show that CKA once again outperforms its peers quite significantly, although the magnitude difference is not as high as smaller datasets, due to the difficulty of training ImageNet. Nonetheless, its outperformance on this important dataset proves the generalisability of CKA across different domains.

4.5 Ablation Study

Our ablation study is intended to demonstrate the benefit of centering the Gram matrix using HSIC before distillation. In this endeavour, we conduct an additional CKA experiment without centering the Gram matrices, as elaborated in Equation 6. Table 4 demonstrates our results and indicates that removing the centering matrix H generally degrades the performance of the model, except when performing self-distillation [29], where directly mimicking the similarity distribution is more feasible. It can be seen that the experiments with R18 student and R18 teacher perform at a somewhat similar level with and without centering. This suggests a potential link between model capacity and ability to mimic the uncentered similarity matrices. Moreover, by comparing our SP experiment which applies the L2 loss on the Gram matrices and our CKA-WC experiment which basically computes cosine similarity of the same uncentered Gram matrices, we see that cosine similarity outperforms euclidean distance-based loss functions for representation distillation.

Architecture		Model Accuracy (%)					
Student	Teacher	Student	SP	CC	RKD	CKA	Teacher
R18	R18	69.77	69.91	69.88	69.95	**70.21**	69.77
R18	R34	69.77	69.90	69.92	69.97	**70.32**	73.07
R18	R50	69.77	69.85	69.70	69.91	**70.08**	75.59
R18	R101	69.77	69.95	69.92	69.98	**70.23**	77.28
MV2	R18	71.23	71.35	71.41	71.39	**71.72**	73.07
MV2	R34	71.23	71.43	71.47	71.45	**71.89**	73.07
MV2	R50	71.23	71.46	71.53	71.58	**71.94**	75.59
MV2	R101	71.23	71.40	71.62	71.51	**72.02**	77.28

Table 3: Results on ImageNet with two students R18 and MV2, distilled from a series of teachers using four main techniques: SP, CC, RKD and CKA (ours).

Architecture		Model Accuracy (%)					
		CIFAR100		Tiny-ImageNet		ImageNet	
Student	Teacher	CKA-WC	CKA	CKA-WC	CKA	CKA-WC	CKA
R18	R18	79.04	79.14	65.13	65.26	70.04	70.21
R18	R34	78.81	79.35	65.04	66.14	69.97	70.32
R18	R50	78.59	79.15	65.18	66.24	67.82	70.08
R18	R101	78.79	79.32	64.78	65.87	69.89	70.23
MV2	R18	74.26	75.30	62.51	63.12	71.48	71.72
MV2	R34	74.41	75.24	62.95	63.35	71.52	71.89
MV2	R50	74.27	75.19	62.84	63.44	71.58	71.94
MV2	R101	74.36	75.11	62.92	63.27	71.69	72.02

Table 4: Ablation study indicating the importance of centering similarity matrices when performing representation distillation. Experiments conducted across three datasets CIFAR100, Tiny-ImageNet and ImageNet are shown. The CKA-WC columns refer to experiments applying CKA Without Centering, whereas the CKA columns apply centered distillation.

5 Conclusion

This paper proposes a novel loss function for representation distillation that uses CKA to distill a centered Gram matrix, outperforming state-of-the-art methods SP, CC and RKD. Our main intuition in proposing the CKA loss function is that representation distillation losses should be invariant to isotropic scaling, a transformation that does not affect class separation, and that cosine similarity is a more robust loss function than its euclidean distance-based counterparts. We provide the mathematical formulation for our CKA loss and demonstrate using an ablation study that centered distillation outperforms its uncentered counterpart. Our experiments conducted across three important image classification datasets, including ImageNet-1k, demonstrate that our method is generalisable to natural image datasets.

References

[1] Malik Boudiaf, Jérôme Rony, Imtiaz Masud Ziko, Eric Granger, Marco Pedersoli, Pablo Piantanida, and Ismail Ben Ayed. A unifying mutual information view of metric learning: cross-entropy vs. pairwise losses. In *European conference on computer vision*, pages 548–564. Springer, 2020.

[2] 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.

[3] Guobin Chen, Wongun Choi, Xiang Yu, Tony Han, and Manmohan Chandraker. Learning efficient object detection models with knowledge distillation. *Advances in neural information processing systems*, 30, 2017.

[4] Corinna Cortes, Mehryar Mohri, and Afshin Rostamizadeh. Algorithms for learning kernels based on centered alignment. *The Journal of Machine Learning Research*, 13: 795–828, 2012.

[5] Nello Cristianini, John Shawe-Taylor, Andre Elisseeff, and Jaz Kandola. On kernel-target alignment. *Advances in neural information processing systems*, 14, 2001.

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

[7] Xiang Deng and Zhongfei Zhang. Comprehensive knowledge distillation with causal intervention. *Advances in Neural Information Processing Systems*, 34:22158–22170, 2021.

[8] Jianping Gou, Baosheng Yu, Stephen John Maybank, and Dacheng Tao. Knowledge distillation: A survey. *International Journal of Computer Vision*, 2021.

[9] Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with hilbert-schmidt norms. In *International conference on algorithmic learning theory*, pages 63–77. Springer, 2005.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European conference on computer vision*, pages 630–645. Springer, 2016.

[11] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *NIPS 2014 Deep Learning Workshop*, 2015.

[12] Yunhun Jang, Hankook Lee, Sung Ju Hwang, and Jinwoo Shin. Learning what and where to transfer. In *International Conference on Machine Learning*, pages 3030–3039. PMLR, 2019.

[13] 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.

[14] Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. Cifar-100 (canadian institute for advanced research). URL http://www.cs.toronto.edu/~kriz/cifar.html.

[15] Ya Le and Xuan Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015.

[16] Seyed Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, and Hassan Ghasemzadeh. Improved knowledge distillation via teacher assistant. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 5191–5198, 2020.

[17] Thao Nguyen, Maithra Raghu, and Simon Kornblith. Do wide and deep networks learn the same things? uncovering how neural network representations vary with width and depth. *arXiv preprint arXiv:2010.15327*, 2020.

[18] A Emin Orhan and Xaq Pitkow. Skip connections eliminate singularities. *arXiv preprint arXiv:1701.09175*, 2017.

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

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

[21] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatte, and Yoshua Bengio. Fitnets: Hints for thin deep nets. *2015 International Conference on Learning Representations (ICLR)*, 2015.

[22] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018.

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

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

[25] Ji Wang, Weidong Bao, Lichao Sun, Xiaomin Zhu, Bokai Cao, and S Yu Philip. Private model compression via knowledge distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 1190–1197, 2019.

[26] Lin Wang and Kuk-Jin Yoon. Knowledge distillation and student-teacher learning for visual intelligence: A review and new outlooks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[27] Junho Yim, Donggyu Joo, Jihoon Bae, and Junmo Kim. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4133–4141, 2017.

[28] 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.

[29] Linfeng Zhang, Jiebo Song, Anni Gao, Jingwei Chen, Chenglong Bao, and Kaisheng Ma. Be your own teacher: Improve the performance of convolutional neural networks via self distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3713–3722, 2019.