Enhanced Knowledge Distillation via Parameter Re-definition

Ziquan Wang [1, a], Yulin Zhao [2, b], Lidong Cheng [2, c, *]

[1] Beihang University, Beijing, China

[2] Hefei University of Technology, Information Computing Sciences, Hefei, China

[a] zqwang@buaa.edu.cn, [b] 1544642195@qq.com, [c, *] orange253@163.com

* Corresponding Author Email: orange253@163.com

Abstract. Due to the high scalability of deep learning and its ability to manipulate large-scale hyperparameters, it has achieved great success in many fields. However, encoding such a large-scale data set is ultimately at the cost of expensive computing power and storage resources, which has also prompted model compression and model acceleration to become a hot topic in recent years. Model pruning, weight decomposition, reduction of model accuracy, weight sharing, etc. are all currently popular solutions, but they have a common problem that they cannot ensure that the compressed model is as good as the original model, and they are all based on the original model. to modify. This paper draws on the method based on knowledge distillation, introduces the concept of Renyi-divergence popularized by KL-divergence, and proposes a loss function that has been based on Renyi-divergence distance metric, and uses the rigor of the student network as a hyperparameter. A student network model that minimizes the loss function under rigor. We validated our results on ResNets using the cifar-10, cifar-100, and imagenet datasets. It improved the basic model by 0.6%, and the absolute gain of Top-l accuracy exceeded 1.6%.

Keywords: Knowledge Transfer; Renyi-divergence; Knowledge Distillation.

1. Introduction

In recent years, Artificial Intelligence (AI) has taken the opportunity of today's big data era to develop rapidly and penetrate people's lives. Such as intelligent voice recognition Siri, face recognition in airports and high-speed railway stations, and smart recommendation functions of e-commerce have brought great convenience to people's lives. The training of many well-performing network models is inseparable from massive data resources, expensive computing resources and storage resources. However, deploying these large-scale and cumbersome network models to the edge (such as smart bracelets, mobile phones, cameras, etc.) will be a huge challenge. To this end, scholars have developed a variety of model compression and model acceleration methods, which can be roughly divided into three categories, namely network pruning [LKD+17], network quantification [QKG+21] and knowledge transfer [PT18]. The basic logic of network pruning is to discard the weights that do not seriously affect the performance of the model. When we change the value of a certain parameter, it has little effect on the result, and then the parameter can be cut off. Network quantization means low precision, which is to convert the floating-point algorithm of the neural network into fixed-point. This can realize the real-time operation of the network on mobile phones, and it is also helpful for the deployment of cloud computing. However, these two methods are modified on the same model, which may lead to overfitting of the model. Good results cannot be obtained when testing on the test set, and their acceleration effect is still not as good as the model deployed on the GPU. The model compression method based on knowledge transfer overcomes the shortcomings of the above methods, and there are essential differences between them. Knowledge distillation [HVD15] is a common method of knowledge transfer, which is a process of one carrier (Student-Net) learning knowledge from another carrier (Teacher-Net). It needs to train a lightweight student model from the Teacher-Net, thereby greatly reducing the depth and width of the model Get a good compression effect. In 1992, Attewell's view that "due to knowledge is mostly inert, people can't simply accept it, but should reconstruct it" inspired Szulanski and led him to propose a Knowledge Transfer (KT) model, which is the earliest the work of Knowledge Transfer (KT). Later,

Gilbert and Cordey proposed a five-step Knowledge Transfer (KT) model, including knowledge acquisition, knowledge exchange, practical application, acceptance and assimilation. First of all, the learner should acquire knowledge and communicate with the knowledge imparter. Then, the learner should apply what they have learned to practice and summarize the results after application. The next step is to accept, and judge whether you should accept this kind of knowledge through the results of feedback in practice. Finally, is assimilation. Successful application of knowledge does not mean being assimilated, but at least it can show that the learner is not resistant. Until recently, Hinton et al. [HVD15] proposed a new Knowledge Transfer (KT) method based on Knowledge Distillation (KD).

There are two networks for Knowledge Distillation (KD), namely the Teacher-Net and the Student-Net. The basic idea of KD is to obtain the softened probability distribution (soft target or dark knowledge) by changing the temperature of the complex Teacher-Net, so that the Student-Net can acquire these dark knowledge and induce the training of the Student-Net. The higher the temperature gets, the smoother the distribution of distilled features becomes, and the stronger the creativity and imagination of the Student-Net. On the contrary, when at a lower temperature, the performance of the Student-Net is limited by the Teacher-Net. Through KD, the ability of the Student-Net should be at least the same as that of the Teacher-Net, so as to achieve the same or better prediction effect with less time complexity and computing resources.

Because in the neural network, the choice of loss function will affect the negative feedback process of the model, thereby affecting the performance of the model. In the work of this article, we present the effect of a loss function better for the training of the student network, that is, the distance metric based on Renyi-divergence, which extends KL-divergence. In order to demonstrate different choices, affect the knowledge of αdistillation results, we use a traceable Gaussian distribution q to generate an approximate complex distribution p on the basis of minimizing the loss function. As shown in Figure 1, it is the approximate value of α without normalization, as α changes, the coverage ranges of p and q are different. q is attracted to the mode with the greatest probability quality with α—>−∞, and when α—>+∞, q tends to cover all modes of p.

We use ResNet to verify our conjecture on various datasets, and the results show that the loss function of distance metric based on the Renyi-divergence in this paper significantly improves the ability of acquiring dark knowledge and generalization of the Student-Net. In summary, the contributions of our work are as follows:

We propose a new loss function in the Student-Net training based on knowledge distillation, based on the Renyi-divergence distance metric, and then we discusse the difference between it and KL-divergence, thus proves the superiority of the new method proposed in this work. We analyzed how to set the loss function to make the performance of the student network closer to or surpass that of the teacher network, and gave theoretical guarantees.

We will analyze the loss function in the negative feedback process of students' network, and give a theoretical guarantee.

We propose an algorithm for reasoning α, which will become a general reasoning method. We have evaluated in deep neural networks and proved that this method can achieve the best results currently on many existing open source data sets.

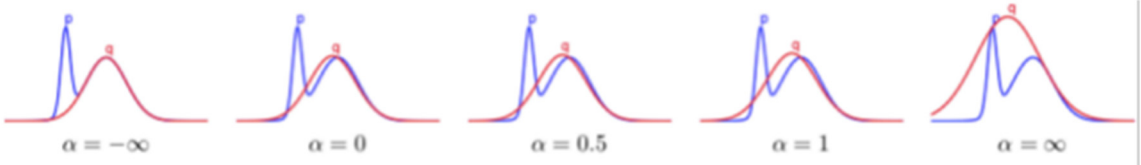


Figure 1. A mixture of two Gaussians q and p, as α changes, the coverage ranges of q and p are different

The traditional Knowledge Distillation (KD) model uses KL-divergence as a loss function to narrow the performance gap between the Teacher-Net and the Student-Net by minimizing the KL - divergence. In Murphy's work, he found that when KL-divergence fails to fully cover all the

characteristics of the Teacher-Net, the loss function will mistakenly over-punish the Student-Net. Therefore, the Student-Net may over-guess the behavior of the Teacher-Net based on the obtained dark knowledge, which may lead to incorrect predictions of certain results in the Teacher-Net.

In 1985, Amari suggested to replace the basic KL-divergence with a more generalized loss function based on Renyi-divergence, Specifically, we show that by adaptively controlling α in the proposed divergence metric, we can punish the bold and conservative predictions of the student network on the teacher network to encourage the student network to approach the teacher network more accurately. Meanwhile, in order to avoid Renyi-divergence from being affected by high gradient variance, we also propose a gradient clipping method to ensure the smooth progress of the training process. We show that the clipped gradient can still define an effective Renyi-divergence metric, thereby generating a suitable optimization target for the process of Knowledge Distillation (KD).

2. Related work

2.1 Compression and Acceleration of Network

In the early stage of Artificial Intelligence (AI) development, people just believed that it was feasible to shrink the model through model compression and acceleration. Later, some scholars proposed network pruning to pursue a balance between accuracy and efficiency. Han [LKD+17] et al. suggested to experiment in the deep network model. They believe that although the model is complex, factors with small weights can be discarded. However, this strategy requires relatively high environmental requirements. Network pruning is divided into weighted pruning, neuron pruning, connection pruning, filtering pruning, etc., which is to improve the efficiency of reasoning and is universally applicable to various scenarios. The weight selection of neurons in different application scenarios will be different. In addition to pruning, quantization, low-rank factorization, low-precision approximation and knowledge distillation are also popular compression models in recent years.

2.2 Knowledge Distillation

Knowledge distillation (KD) is one of the more mature solutions in knowledge transfer, and it is widely used. In Knowledge distillation (KD), the knowledge output by the Teacher-Net is defined as a soft target, also known as dark knowledge. Compared with a one-hot tag, soft targets provide more information than hard targets, and it reflects the potential connection between layers in the Teacher-Net. When the network uses one-hot (hard targets), it will lose part of the original data information, reducing the difficulty of the model to fit the data, making the model easier to fit, so overfitting may occur, resulting in a decline in the generalization ability of the model. When using soft targets, the model needs to learn more knowledge, such as the similarity and difference between two close probabilities, which brings challenges to the model's fitting ability and enhances the model's generalization ability. And as the entropy of the soft target's distribution is relatively high, the knowledge contained in the soft targets is richer.

Formally, the soft target of Teacher-Net (T) can be defined by $p = SoftMaxa/\tau$, where a is the logits vector of Teacher-Net (the activation value before SoftMax), and τ is the temperature. By increasing τ, by deviating the predicted value from 0 and 1, such inter-class similarity can be preserved. Then, the student network is trained by softening the combination of SoftMax and the original SoftMax. When the temperature rises, the weight of dark knowledge (negative label) will also increase, so that we can learn more dark knowledge instead of making it redundant in the network. In this way, we can allow a limited network to learn more knowledge for task processing. Higher temperatures will produce weaker probability distributions across categories. Specifically, when $\tau \to +\infty$, all categories have the same probability.

Subsequent work tried to supplement KD by transferring intermediate features. Recently, Romero et al. proposed FitNet to compress a wide and shallow network into a narrow and deep network. In order to learn the hidden features of the teacher network, FitNet allows students to imitate the teacher's

full feature map. However, this assumption is too strict, because the abilities of teachers and students may be very different. Under certain circumstances, FitNet may use its poor performance, or even overfitting. Recently, Zagoruyko et al. [Gre93] proposed to use the attention transfer mechanism (AT) to relax the assumption on FitNet: the attention map they transfer is the sum of all activations. NST [RHGS15] a new knowledge transfer method, which was treated as a division matching problem, matching the distribution of neuron selective patterns between teacher and student networks, and designed a new KT loss function with minimum Calculate the maximum average difference (mmd) between these distributions. Yim et al. [LWF+20] defined a new type of knowledge, the solution process flow (FSP) of knowledge transfer, which uses the Gram matrix of two different levels of features. They claim that this FSP matrix can reflect the problem-solving process of teachers. However, KD and these recent developments are independent of each other, and combining these efforts will achieve better results. As discussed later, their work can be regarded as a special case in our framework.

2.3 Domain Adaptation

Domain adaptation is a concept in transfer learning. In transfer learning, when the data distribution of the source domain and the target domain are different, but the two tasks are the same, the algorithm in the source domain is still available. The process from the source domain to the target domain is called domain adaptation. Since the distribution on the target domain is unknown, the effect of domain adaptation is related to the difference in the distribution of these two domains. At present, many similarity measures have been widely used in domain adaptation research, such as JS-divergence and α-divergence.

3. Our Methods

In this paper, we studied the concept of Renyi-divergence and discussed in detail the relationship between Renyi-divergence and Knowledge Distillation (KD). In the process of obtaining the Student-Net by knowledge distillation, we proposed a better objective loss function, that is, adding Renyi-divergence into the negative feedback of the Student-Net and minimizing it. By minimizing the distance between the Teacher-Net and Student-Net as the loss function, we can improve the difference based on different strict levels The Knowledge Distillation (KD) algorithm of the Student-Net. Compared with the traditional method of using information entropy as the loss function, the setting of the loss function significantly improves the learning ability and generalization ability of the Student-Net. We not only use multiple data sets to verify the accuracy of our method, but also compare the work of this paper with other existing methods, which demonstrated the superiority of the work of this paper.

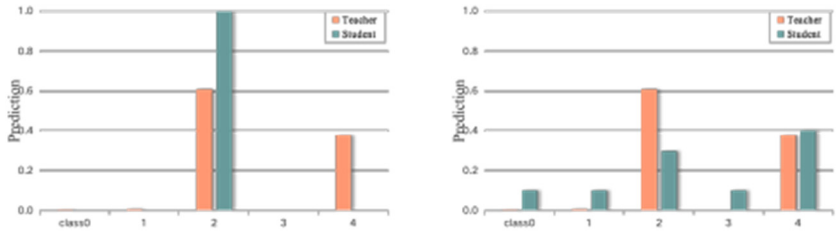
3.1 Motivation

In the Knowledge Distillation, the rigor of the student network has not been fully valued- how the student network views the supervision information provided by the teacher network.

The KL-divergence between the teacher and student models is calculated by $KL(p\|q)=E_p[\log p/q]$, Due to the zero-avoidance feature of KL-divergence, when $p > 0$, $q > 0$ must be satisfied. In contrast, when q > 0, only if under p > 0, it will not be punished. In figure 2, although the student-Net excessively reasoned about the dark knowledge of the Teacher-Net, and it produced a false prediction result (Class 4), The value of the objective loss function (KL) is still not large.

When other types of divergence are used, such as the reverse KL-divergence $KL(m//n)$, the above overestimation in Example 2 will be more penalized. For the reverse KL-divergence, if $n = 0$ and $m > 0$, $KL(m//n) = E_q[logm/n]$ is infinite. Therefore, if $n = 0$, we must ensure that $m = 0$, which is called zero-forcing property (Murphy, 2012). Therefore, minimizing the reverse KL-divergence encourages the student model m to avoid the low-probability mode of n, while paying attention to the

high-probability mode. Therefore, it may underestimate the uncertainty of the teacher model, as shown in Figure 1, 2.



(a) Example 1 (b) Example 2

Figure 2. Examples

Therefore, in the process of learning from the student network to the teacher network, there is an urgent problem that needs to be solved, that is, whether the loss function can be generalized to suppress the excessive useless reasoning of the student network while improving its ability to make reasonable guesses through dark knowledge.

The results in Figure 2 give us a new idea to control the boldness of the student-Net's reasoning. Firstly, we use the more flexible Renyi-divergence to summarize the typical KL divergence.

Considering $\alpha \in (0, 1) \cup (1, +\infty)$, Renyi-divergence is defined as

$$D_{\alpha}(p\| q) = \frac{1}{\alpha - 1} \sum_{i=1}^{m} p_i^{\alpha} q_i^{1-\alpha} \tag{1}$$

Where $q = [q_i]_{i=1}^{m}$ and $p = [p_i]_{i=1}^{m}$ are two discrete distributions on m categories. Renyi-divergence includes a wide range of classical divergence measures. In particular, the KL-divergence $\mathrm{KL}(p\| q)$ is the limit of $D_{\alpha}(p\| q)$ with $\alpha \to 1$, while the χ^2-divergence $\chi^2(P,Q)$ is the limit of $D_{\alpha(p\| q)}$ with $\alpha \to 2$

A key feature of Renyi-divergence is that we can decide to punish different types of differences (underestimated or overestimated) by choosing different α values. For example, as shown in Figure 2(c), when $\alpha \in (0,1)$, when q is more widely distributed than p (when q overestimates the uncertainty in p), $D_{\alpha}(p\| q)$ is large, when When q overestimates the uncertainty in $p, D_{\alpha(p\| q)}$ is small and q is more concentrated than p (when q underestimates the uncertainty in p). When $\alpha \in (1,+\infty)$, the trend is opposite: underestimation will be punished more severely than overestimation.

To alleviate the problems of overestimation and underestimation at the same time when training the super network, we consider $\alpha_1 \in (0,1)$ and $\alpha_2 \in (1,+\infty)$, and suggest using $D_{\alpha_1}(p\| q)$ and $D_{\alpha_2}(p\| q)$ in the KD loss function:

$$D_{\alpha_1,\alpha_2}(p\| q) = \max\left\{D_{\alpha_1}(p\| q), D_{\alpha 2}(p\| q)\right\} \tag{2}$$

Our KL loss now changes from equation to

$$\$_{KD}\left([\theta, s], \theta_t\right) = E_{x\sim D}\left[D_{\alpha_1,\alpha_2}(p(x;\theta)\| q(s;\theta,s))\right] \tag{3}$$

We denote this KD strategy that always selects the maximum value of D_{α_1} and D_{α_2} for optimization as Adaptive-KD.

3.2 Some Properties of Renyi-divergence

When α is in $[0,1]\cup(1,\infty)|\ D_\alpha<\infty, D_\alpha$ is continuous. For any $\alpha\in[0,\infty]$, there is

$$D_\alpha(P\|\ Q)\geq 0$$

For $\alpha>0$, if and only if $P=Q, D_\alpha(P\|\ Q)=0$.

For $\alpha=0$, if and only if $Q\ll P, D_\alpha(P\|\ Q)=0$.

For any $0<\alpha<1$, there is

$$D_\alpha(P\|\ Q)=\frac{\alpha}{1-\alpha}D_{1-\alpha}(Q\|\ P) \tag{4}$$

For any $\alpha\in(0,1)$, Renyi-divergence is a concave function.

4. Stabilizing α-divergence KD

Usually, α_1 and α_2 will be set to larger values at the same time to ensure that the Student-Net is sufficiently penalized when it underestimates or overestimates uncertainty the Teacher-Net. However, directly using large α to optimize Renyi-divergence is often difficult in actual problems. In Renyi-divergence, due to its gradient problem, we have the following formula

$$\nabla_\theta D_\alpha\left(p\|\ q_\theta\right)=-\frac{1}{\alpha}E_{q_\theta}\left[\left(\frac{p}{q_\theta}\right)^\alpha\nabla_\theta\log q_\theta\right] \tag{5}$$

If α is large, the power term $\left(p/q_\theta\right)^\alpha$ may be very important and cause the training process to be unstable. To enhance training stability, we clamp the maximum value of $\left(p/q_\theta\right)^\alpha$ to β, and obtain

$$\tilde{\nabla}_\theta D_\alpha\left(p\|\ q_\theta\right)\overset{\text{def}}{=}-\frac{1}{\alpha}\left[\text{Clip}_\beta\left(\frac{p}{q_\theta}\right)^\alpha\nabla_\theta\log q_\theta\right] \tag{6}$$

Equation (5) is a simple and effective heuristic approximation of $\nabla_\theta D_\alpha\left(p\|\ q_\theta\right)$. It is important to pay attention to the equation (5). Equal to the exact gradient of the special f divergence between p and q_θ. Therefore, our update is still equivalent to minimizing effective divergence. Note that the clipping function Clip beta $(\cdot)$ is only partially differentiable. So naively cut the $\left(p/q_\theta\right)^\alpha$ in the equation(3). It may prevent the gradient from propagating back to the density ratio term, thereby generating gradients that are not from the effective divergence.

To show that we are still optimizing the effective divergence with Eqn, please note that for the convex function $f:[0,+\infty)\to R$, the divergence of f between p and q_θ is defined as

$$D_f\left(p\|\ q_\theta\right)=E_{q_\theta}\left[f\left(\frac{p}{q_\theta}\right)-f(1)\right] \tag{7}$$

Its gradient w.r.t. θ is

$$\nabla_\theta D_f\left(p\|\ q_\theta\right)=-E_{q_\theta}\left[\rho_f\left(\frac{p}{q_\theta}\right)\nabla_\theta\log q_\theta\right] \tag{8}$$

Where $\rho_f(t)=f'(t)t-f(t)$ (Wang et al. (2018)). Note that α-divergence is a special case of f-divergence when $f(t)=t^\alpha/(\alpha(\alpha-1))$.

Proposition 4.1. There exists a convex function $f:(0,+\infty)\to R$ such that $\tilde{\nabla}_\theta D_\alpha\left(p\|\, q_\theta\right)$ in Eqn is the exact gradient of $D_f\left(p\|\, q_\theta\right)$, that is $\tilde{\nabla}_\theta D_\alpha\left(p\|\, q_\theta\right)=\nabla_\theta D_f\left(p\|\, q_\theta\right)$

Proof. Let $\rho_*(t)=\frac{1}{\alpha}\mathrm{Clip}_\beta(t)^\alpha$. We just need to find an f such that

$$\rho_f(t)=f'(t)t-f(t)=\rho_*(t) \tag{9}$$

Taking the derivation on both sides, we get $f''(t)t=\rho_*'(t)$. This gives $f''(t)=\rho_*'(t)/t$ and hence $f(t)=\iint \rho_*'(t)/t\,dt$, where $\iint$ denotes second-order antiderivative (or indefinite integral). Because $\rho_*(t)$ is nondecreasing, we have $\rho_*'(t)/t\geq 0$ for $t>0$, and hence f is convex on $(0,+\infty)$.

5. Experiments

We conceived the most complex data set for image classification in ImageNet [13] having approximately 1,2 million trainings and 505 thousand images, each belonging to 1000 categories. ResNets will also be experimented. The network is trained for 100 epochs and the learning rate begins at 0.1 and begins at every 30th epoch with 0.1 and the batch rate is 256, which is identical to conventional training without extra education skills. The performance of our ImageNet technique is reported by Table 1-2. It improves the basic models by 0.6% and absolute gains in Top-1 precision over 1.6%, which support the benefit of our method on the large-scale dataset.

Table 1. Results on the CIFAR-10 dataset. We report top-1 accuracies (%).

Method	KD	AT	KD+AT	alpha=0.1	alpha=0.2	alpha=0.3	alpha=0.4	alpha=0.5	alpha=0.6	alpha=0.7	alpha=0.8	alpha=0.9
acc(%)	94.42	94.51	94.75	94.22	94.17	94.26	94.55	94.74	94.31	94.44	94.49	94.22
40-1->16- 1				91.92	91.9	91.79	91.91	91.77	91.96	91.98	92.33	92.26
16-2->16-1				92.85	92.45	92.57	92.78	92.83	92.87	92.71	92.53	92.65

Table 2. Results on the CIFAR-100 dataset. We report top-1 accuracies (%)

Method	KD	AT	KD+AT	alpha=0.1	alpha=0.2	alpha=0.3	alpha=0.4	alpha=0.5	alpha=0.6	alpha=0.7	alpha=0.8	alpha=0.9	alpha=2
acc(%)	74.42	72.82	74.15	74.4	73.51	74.01	74.15	74.11	74.11	74.57	74.49	74.41	74.17

6. Conclusion

In this paper, we use the knowledge distillation method to compress the model. A new loss function is proposed to optimize the performance of the Student-Net, that is, the distance measurement based on Renyi-divergence, which is a generalization of the KL-divergence loss function. The experimental results show that, in the process of knowledge distillation, by modifying the value of α based on the Renyi-divergence based loss function, we realize Student-Net with different degrees of strictness. The Student-Net with a high degree of rigor will underestimate the uncertainty of the Teacher-Net's prediction, and vice versa, it will overestimate the uncertainty of the teacher network. We have tested on many open-source data sets. The experimental results show that compared with the traditional method based on information entropy, the improvement of this method significantly improves the learning ability and generalization ability of Student-Net.

References

[1] [AHD+19] Sungsoo Ahn, Shell Xu Hu, Andreas Damianou, Neil D. Lawrence, and Zhenwen Dai. Variational information distillation for knowledge transfer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2019.

[2] [CRCZ20] Xu Cheng, Zhefan Rao, Yilan Chen, and Quanshi Zhang. Explaining knowledge distillation by quantifying the knowledge. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2020.

[3] [DJW+21] Xing Dai, Zeren Jiang, Zhao Wu, Yiping Bao, Zhicheng Wang, Si Liu, and Erjin Zhou. General instance distillation for object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 7842–7851, June 2021.

[4] [Gre93] George D. Greenwade. The Comprehensive Tex Archive Network (CTAN). TUGBoat, 14(3): 342–351, 1993.

[5] [HVD15] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531, 2015.

[6] [LGG+17] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Oct 2017.

[7] [LKD+17] Hao Li, Asim Kadav, Igor Durdanovic, Hanan Samet, and Hans Peter Graf. Pruning filters for efficient convnets. In ICLR, 2017.

[8] [LWF+20] Xiaojie Li, Jianlong Wu, Hongyu Fang, Yue Liao, Fei Wang, and Chen Qian. Local correlation consistency for knowledge distillation. In European Conference on Computer Vision, pages 18–33. Springer, 2020.

[9] [PT18] Nikolaos Passalis and Anastasios Tefas. Learning deep representations with probabilistic knowledge transfer. In ECCV, 2018.

[10] [QKG+21] Lu Qi, Jason Kuen, Jiuxiang Gu, Zhe Lin, Yi Wang, Yukang Chen, Yanwei Li, and Ji- aya Jia. Multi-scale aligned distillation for low-resolution detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 14443–14453, June 2021.

[11] [RDS+15] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhi- heng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C Berg, and Fei-Fei Li. Imagenet large scale visual recognition challenge. International Journal of Computer Vision, 2015.

[12] [RHGS15] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In C. Cortes, N. D. Lawrence, D. D. Lee.

[13] M. Sugiyama, and R. Garnett, editors, Advances in Neural Information Processing Systems 28, pages 91–99. Curran Associates, Inc., 2015.

[14] [TM19] Frederick Tung and Greg Mori. Similarity-preserving knowledge distillation. In ICCV, 2019.