$V_k D$: Improving Knowledge Distillation using Orthogonal Projections

Roy Miles   Ismail Elezi   Jiankang Deng
Huawei Noah's Ark Lab

1. Supplementary Material

We provide more details on the datasets, architectures, and training pipelines. We also provide results on the much smaller CIFAR100 distillation benchmark, where we show competitive or improved performance over state-of-the-art. Finally, we include some qualitative results for the image generation and the complete derivations for orthogonality and whitening.

1.1. Datasets

We conduct experiments over a few widely adopted datasets including CIFAR [6], ImageNet [12], and COCO [7].

CIFAR classification consists of 60K 32×32 RGB images across either 10 or 100 classes with a 5:1 training/testing split. The models are each trained with 100%, 20% or 10% training images [3].

ImageNet classification uses 1.3 million images from 1000 different classes. In these experiment, we set the input size to 224×224, and follow the same training pipeline and augmentations provided by DeiT [14].

COCO includes a large-scale object detection benchmark, which we use to evaluate the ViDT model variants. It consists of 330k images with 80 different object categories.

1.2. Handcrafted projections

We compare our method to a handcrafted projection. For this projection we match the student features with a truncated SVD decomposition of the teacher features. In this way the student will align with the principle components of the teacher. The results are shown in figure 1 and although some good performance is achieved, it falls on achieving the top-end accuracy attained by an orthogonal projection. We expect this drop in performance is likely a consequence of the improved gradient flow when performing the loss in the larger teacher space and also that the smaller principle components are indeed contributing to the discriminative power of the learned representation. Furthermore, computing the SVD is much more computationally expensive due to the expensive decomposition required for each batch.

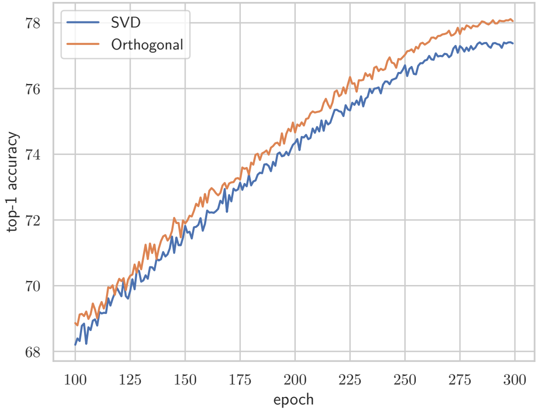

Figure 1. Comparison between using an orthogonal projection to using a truncated SVD of the teacher features. We observe consistent improvement in performance and convergence, while being much more computationally efficient.

1.3. Experiments on CIFAR100

CRD [13] provides an easy benchmark for most distillation methods. However, the results on this benchmark have become increasingly saturated, where many methods are even reporting better student performance than the teacher. This situation alone raises questions on whether the improvement is down to an improved knowledge distillation, or simply through the introduction of implicit model regularisation. Despite these limitations, we do provide some results on this CIFAR100 benchmark in table 1. Here we observe competitive performance to previous state-of-the-art on a few of the challenging cross-architecture settings.

Although ReviewKD [2] achieves very strong performance on this benchmark, its application is limited to the CNN → CNN settings. Furthermore, it requires many additional trainable parameters and has a much larger memory overhead since the intermediate representations are needed to compute the loss.

1.4. Implementation details

Patch token distillation is used to distill from or to a transformer based model and can be seen in Fig. 2.

Distillation	Teacher	vgg13	ResNet50	ResNet50	resnet32x4	resnet32x4	WRN-40-2
Mechanism	Student	MobileNetV2	MobileNetV2	vgg8	ShuffleNetV1	ShuffleNetV2	ShuffleNetV1
	Teacher	74.64	79.34	79.34	79.42	79.42	75.61
	Student	64.60	64.60	70.36	70.50	71.82	70.50
Logit	KD [4]	67.37	67.35	**73.81**	74.07	74.45	74.83
	DKD [17]	**69.71**	**70.35**	-	**76.45**	**77.07**	**76.70**
Intermediate	FitNet [11]	64.14	63.16	70.69	73.59	73.54	73.73
	AT [16]	59.40	58.58	71.84	71.73	72.73	73.32
	NST [5]	58.16	64.96	71.28	74.12	74.68	74.89
	SP [15]	66.30	68.08	**73.34**	73.48	74.56	74.52
	ReviewKD [2]	**70.37**	**69.89**	-	**77.45**	**77.78**	**77.14**
Representation	CC [10]	64.86	65.43	70.25	71.14	71.29	71.38
	RKD [8]	64.52	64.43	71.50	72.28	73.21	72.21
	PKT [9]	67.13	66.52	73.01	74.10	74.69	73.89
	CRD [13]	69.94	69.54	74.58	75.12	76.05	76.27
	WCoRD [1]	70.02	70.12	74.68	75.77	76.48	76.68
	V_kD	**70.11**	**70.65**	**74.95**	**77.05**	**77.51**	**77.19**
	Δ	+2.74	+3.30	+1.14	+2.98	+3.06	+2.36

Table 1. CIFAR-100 test *accuracy* (%) of student networks trained with a number of distillation methods. The best results for each distillation mechanism are highlighted in **bold**. Δ represents the performance improvement over classical KD. Representation is used here to describe the features directly before the final classifier.

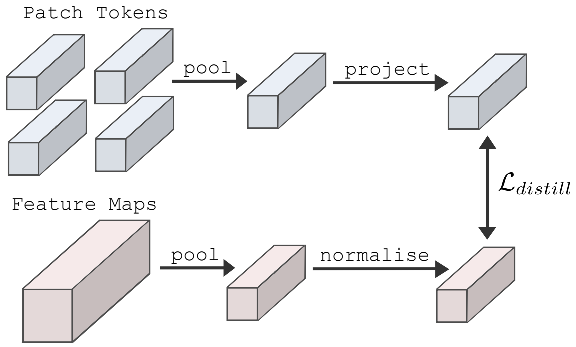

Figure 2. Patch token distillation between transformer and CNN models. We replace the projection with an orthogonal projection, while for the normalisation we either use layer norm or iterative whitening. For the pooling we adopt a simple global average.

This method was adopted since adding more distillation tokens [14] would introduce additional trainable parameters. Using a pooling strategy over the patch tokens proved to be very simple and effective.

2. Qualitative results for image generation

Fig. 3 shows some example images generated using the V_kD distilled BigGAN model trained on CIFAR100. The results demonstrate a very diverse generation of images, while preserving a lot of structural object information.

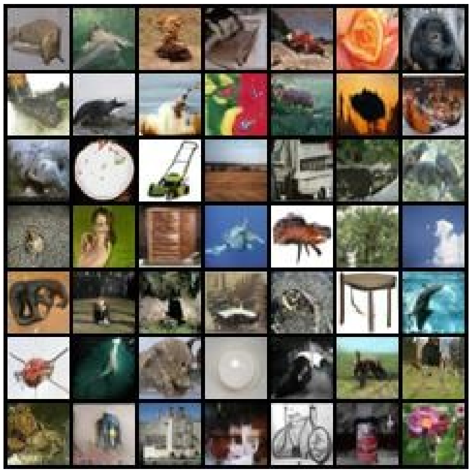

Figure 3. Example images generated using our V_kD distilled BigGAN model using the CIFAR-100 training dataset.

Method	Time (s)	Memory (GiB)
Linear	0.72 ± 0.01	17.67
Orthogonal	0.98 ± 0.02	17.82
RKD	0.76 ± 0.01	19.60
VID	0.93 ± 0.01	25.95
ICKD	0.78 ± 0.02	19.97
CRD	0.80 ± 0.02	21.46

Table 2. Timing and memory of different methods in ImageNet. We distill to a DeiT-S with an effective batch size of 1024 on ImageNet using 2 NVIDIA V100 GPUs.

2.1. Computational overhead

As shown in Tab. 2, during training, we only observe a small increase in latency but almost no increase in GPU memory. Most other methods come with a significant increase in memory, while also incurring additional training time overheads. There is also no inference overhead since we throw away the orthogonal projection after training.

Dataset	Metric	No Norm	Layer Norm	Whitening
C10	Precision	0.87	0.85	**0.88**
	Recall	0.72	0.69	**0.74**
	Density	**1.14**	1.00	1.12
	Coverage	0.95	0.95	**0.96**
C100	Precision	**0.88**	0.85	**0.88**
	Recall	0.04	0.00	**0.71**
	Density	1.07	0.86	**1.17**
	Coverage	0.56	0.15	**0.96**

Table 3. Evaluation for the fidelity and diversity of the distilled student models. Although we observe only a small improvement in recall when whitening the target teacher features for CIFAR10, we find it is critical in avoiding mode collapse for the more diverse CIFAR100 dataset.

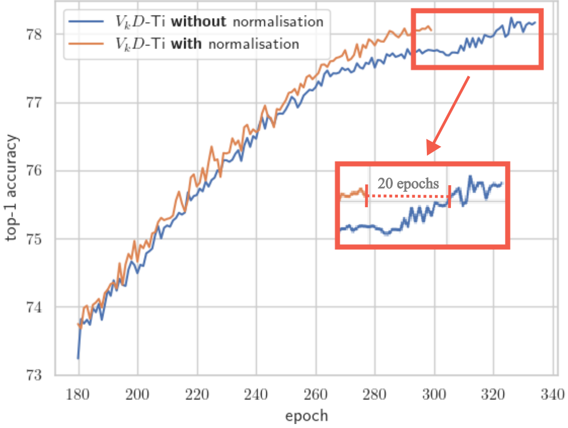

Figure 4. Normalisation improves convergence for discriminative tasks by improving the robustness of the loss to small/irrelevant perturbations in the input image.

2.2. Quantitative diversity metrics

We provide an evaluation metric on the feature diversity in Tab. 3. We generated 5k images from each model trained with or without our whitening and also with layer normalization. We see that whitening does encourage image diversity while also improves the FID and IS realism scores.

2.3. Normalisation improves convergence

Fig. 4 shows the evaluation results after each epoch of training. Here we confirm that the normalisation step does improve the model convergence. We find that simply extending the un-normalised training pipeline is enough to recover the drop in accuracy.

2.4. Further analysis

In this section we provide the complete derivations to supplement the illustrative analysis in the main manuscript.

Orthogonality of $exp(\mathbf{W})$. If $\mathbf{W}$ is a skew-symmetric matrix, then it admits the property $\mathbf{W}^T = -\mathbf{W}$. Its matrix-exponential is then given by:

$$\exp(\mathbf{W}) = \mathbf{I} + \mathbf{W} + \frac{\mathbf{W}^2}{2!} + \frac{\mathbf{W}^3}{3!} \dots \tag{1}$$

Since $\mathbf{W}$ is skew-symmetric, the transpose of this exponential is given as follows:

$$\begin{aligned} \exp(\mathbf{W})^T &= \mathbf{I} - \mathbf{W} + \frac{\mathbf{W}^2}{2!} - \frac{\mathbf{W}^3}{3!} \dots && (2) \\ &= \exp(-\mathbf{W}) && (3) \end{aligned}$$

Thus $\exp(\mathbf{W}) \exp(\mathbf{W})^T = \exp(\mathbf{W}) \exp(-\mathbf{W}) = \mathbf{I}$, which confirms that $\exp(\mathbf{W})$ is indeed orthogonal.

Whitening and feature diversity In this section we provide a more thorough investigation into the connection between the use of whitening and feature diversity. Our loss is simply the pair-wise distance between the student and teacher features. Through simple algebraic manipulation, we can re-express this loss as follows:

$$\begin{aligned} \mathcal{L}_{distill} &= \left\|\mathbf{Z}^s - \mathbf{Z}^t\right\|^2 \\ &= \sum_{i \neq j} |\mathbf{Z}^s_{:,j} - \mathbf{Z}^t_{:,i} - \mathbf{Z}^t_{:,j} + \mathbf{Z}^t_{:,i}|^2 \\ &= \sum_{i \neq j} |\mathbf{Z}^s_{:,j} - \mathbf{Z}^t_{:,i}|^2 + |\mathbf{Z}^t_{:,j} + \mathbf{Z}^t_{:,i}|^2 \\ &\qquad - 2\left\langle \mathbf{Z}^s_{:,j} - \mathbf{Z}^t_{:,i}, \mathbf{Z}^t_{:,j} + \mathbf{Z}^t_{:,i} \right\rangle, \end{aligned} \tag{4}$$

where $\mathbf{Z}^t$ is whitened such that $(\mathbf{Z}^t)^T(\mathbf{Z}^t) = \mathbf{I}$. This means that the magnitude of each feature will be equal to one and the dot product between the different features within a batch will be zero. We can express these two properties as follows:

$$\begin{aligned} &\textit{unit length :} && \left\|\mathbf{Z}^t_{:,i}\right\|^2 = 1, \\ &\textit{decorrelated :} && \langle \mathbf{Z}^t_{:,j}, \mathbf{Z}^t_{:,i} \rangle = 0 \end{aligned} \tag{5}$$

Substituting into the second term of equation 4 leads to the following simplification:

$$\mathcal{L}_{distill} = \sum_{i \neq j} \left\|\mathbf{Z}^s_{:,j} - \mathbf{Z}^t_{:,i}\right\|^2 + 2 - 2\left\langle \mathbf{Z}^s_{:,j} - \mathbf{Z}^t_{:,i}, \mathbf{Z}^t_{:,j} + \mathbf{Z}^t_{:,i} \right\rangle$$

Using the Cauchy-Schwartz inequality, we can find a lower bound on this loss.

$$
\begin{aligned}
\mathcal{L}_{distill} &\geq \sum_{i\neq j} \left\|\mathbf{Z}^s_{:,j} - \mathbf{Z}^t_{:,i}\right\|^2 + 2 - 2\sqrt{\left\|\mathbf{Z}^s_{:,j} - \mathbf{Z}^t_{:,i}\right\|^2 \left\|\mathbf{Z}^t_{:,j} + \mathbf{Z}^t_{:,i}\right\|^2} \\
&\geq \sum_{i\neq j} \left\|\mathbf{Z}^s_{:,j} - \mathbf{Z}^t_{:,i}\right\|^2 + 2 - 2\left\|\mathbf{Z}^s_{:,j} - \mathbf{Z}^t_{:,i}\right\|^2 \left\|\mathbf{Z}^t_{:,j} + \mathbf{Z}^t_{:,i}\right\|^2 \\
&= \sum_{i\neq j} \left\|\mathbf{Z}^s_{:,j} - \mathbf{Z}^t_{:,i}\right\|^2 + 2 - 4\left\|\mathbf{Z}^s_{:,j} - \mathbf{Z}^t_{:,i}\right\|^2 \\
&= \sum_{i\neq j} 2 - 3\left\|\mathbf{Z}^s_{:,j} - \mathbf{Z}^t_{:,i}\right\|^2 \\
&= \text{const} - 3\sum_{i\neq j} \underbrace{\left\|\mathbf{Z}^s_{:,j} - \mathbf{Z}^t_{:,i}\right\|^2}_{\mathbf{C}^2} \qquad (6)
\end{aligned}
$$

This bound is minimised when the distance between each $i \neq j$ student and teacher feature is maximised. Similarly, the $L2$ loss itself in equation 4 minimises the pair-wise distance between features. These two results show that the whitening operation jointly minimises the pairwise similarity, while also maximising an upper bound for the cross feature diversity.

References

[1] Liqun Chen, Dong Wang, Zhe Gan, Jingjing Liu, Ricardo Henao, and Lawrence Carin. Wasserstein Contrastive Representation Distillation. *CVPR*, 2020. 2

[2] Pengguang Chen, Shu Liu, Hengshuang Zhao, and Jiaya Jia. Distilling Knowledge via Knowledge Review. *CVPR*, 2021. 1, 2

[3] Kaiwen Cui, Yingchen Yu, Fangneng Zhan, Shengcai Liao, Shijian Lu1, and Eric Xing. Kd-dlgan: Data limited image generation via knowledge distillation. *CVPR*, 2023. 1

[4] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the Knowledge in a Neural Network. *NeurIPS*, 2015. 2

[5] Zehao Huang and Naiyan Wang. Like What You Like: Knowledge Distill via Neuron Selectivity Transfer. *arXiv preprint*, 2017. 2

[6] Alex Krizhevsky. Learning Multiple Layers of Features from Tiny Images. 2009. 1

[7] Tsung Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. *ECCV*, 2014. 1

[8] Wonpyo Park, Kakao Corp, Dongju Kim, and Yan Lu. Relational Knowledge Distillation. *CVPR*, 2019. 2

[9] Nikolaos Passalis and Anastasios Tefas. Learning Deep Representations with Probabilistic Knowledge Transfer. *ECCV*, 2018. 2

[10] Baoyun Peng, Xiao Jin, Dongsheng Li, Shunfeng Zhou, Yichao Wu, Jiaheng Liu, Zhaoning Zhang, and Yu Liu. Correlation congruence for knowledge distillation. *CVPR*, 2019. 2

[11] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. FitNets: Hints For Thin Deep Nets. *ICLR*, 2015. 2

[12] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *IJCV*, 2014. 1

[13] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. *ICLR*, 2019. 1, 2

[14] Hugo Touvron, Matthieu Cord, and Hervé Jégou. Deit iii: Revenge of the vit. In *ECCV*, 2022. 1, 2

[15] Fred Tung and Greg Mori. Similarity-preserving knowledge distillation. *ICCV*, 2019. 2

[16] Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In *ICLR*, 2019. 2

[17] Borui Zhao, Renjie Song, and Yiyu Qiu. Decoupled Knowledge Distillation. *CVPR*, 2022. 2