Cross-Architecture Distillation Made Simple with Redundancy Suppression

Supplementary Material

In this material, we provide additional details, analysis, and visualisations regarding our method.

6. Further Details

Datasets. We evaluate the proposed method on the CIFAR-100 [29] and ImageNet-1k [12] datasets for image classification, following OFA [22]. CIFAR-100 contains 50k 32 × 32 resolution images in 100 classes for training and 10k images for testing. ImageNet-1k includes 224×224 resolution images in 1,000 classes, of which 1.2 million images are for training and 50k for validation.

Implementation. For the RSD loss, since there are much more off-diagonal elements than diagonal elements, in practice we scale down the contribution of the decorrelation loss by a factor of κ. We set the default value of κ to 0.01, following Zbontar et al. [72], and tune it for ImageNet-1k experiments. For the MLP capacity in the AAD module, we adopt an expansion rate of 2 for CIFAR-100 experiments, and 4 for ImageNet-1k experiments by default. As for λ, we set it to 0.1 for CIFAR-100 and 0.01 for ImageNet-1k experiments, and tune it where necessary.

Training. We follow the same configurations as in OFA [22]. All images are resized to 224 × 224 as input. CNN students are trained using the SGD optimiser, whereas ViT and MLP students are trained using AdamW. We train all models for 300 epochs on CIFAR-100. We use a batch size of 128 for CNN students, and a batch size of 512 for ViT and MLP students. For ImageNet-1k, we train CNN students for 100 epochs with a batch size of 64 and ViT or MLP students for 300 epochs with a batch size of 128. All experiments are conducted using 8 NVIDIA A800 GPUs. All experimental results reported are the average top-1 accuracy of 3 independent runs. When taking the efficiency measurements, we use a local workstation with 20 Intel Core i9-10850K CPUs (10 cores) and an NVIDIA RTX 3090 GPU.

7. Further Analysis

Sensitivity to RSD loss weight λ. We explore how sensitive the proposed method is to different values of the loss weighting hyperparameter, λ. Compared to OFA and other generic KD methods, we find that our method is reasonably robust to varying λ, as demonstrated in Table 9. It consistently outperforms OFA with significant margin with different values of λ. While tuning this hyperparameter may lead to even better performance, as shown in Table 9, we stick to our default value of $\lambda = 0.1$ for CIFAR-100 experiments to save laborious tuning efforts and facilitate generalisation.

λ	0.05	0.075	0.1	0.15	0.2
Swin-T→ResNet18	83.36	83.80	83.92	**84.10**	83.75
Mixer-B/16→DeiT-T	78.16	78.53	78.50	**78.60**	78.48
ConvNeXt-T→ResMLP-S12	84.36	**84.41**	84.21	83.77	83.74

Table 9. **Sensitivity to varying RSD loss weight λ.**

γ	1/2	1	2	4	6
Swin-T→ResNet18	83.78	84.20	83.92	83.98	**84.21**
Mixer-B/16→DeiT-T	77.53	78.53	78.50	78.75	**79.30**
ConvNeXt-T→ResMLP-S12	82.92	83.33	**84.21**	84.06	84.18

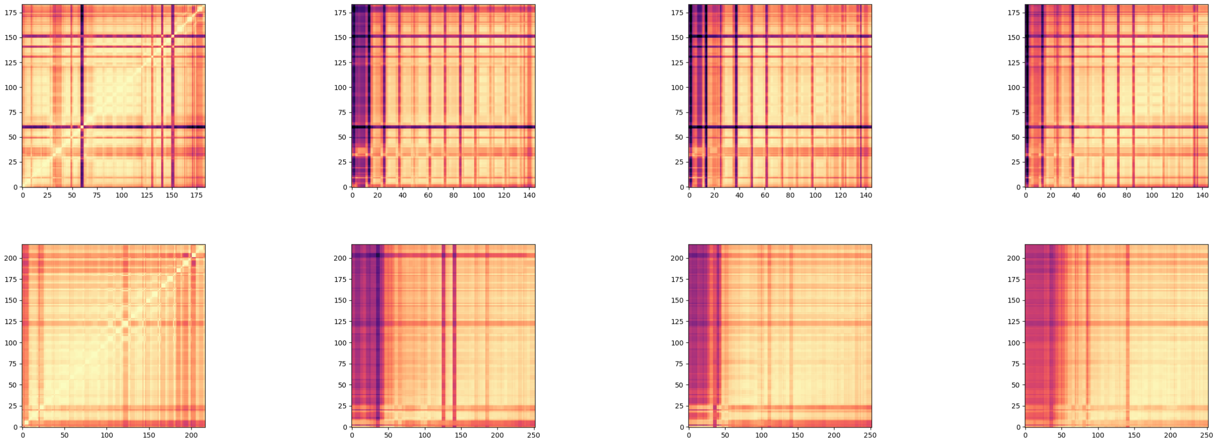
Table 10. **Sensitivity to varying AAD expansion rate γ.**

Design	Swin-T→ ResNet18	Mixer-B/16→ DeiT-T	ConvNeXt-T→ ResMLP-S12
OFA	80.54	73.90	81.22
w/o Norm.	82.90	75.87	81.73
Norm.	**83.92**	**78.50**	**84.21**
MSE	**83.92**	**78.50**	84.21
Huber	83.83	78.40	**84.30**
$\kappa = 0$	83.90	78.50	83.54
$\kappa = 0.001$	83.73	**78.78**	83.81
$\kappa = 0.01$	83.92	78.50	**84.21**
$\kappa = 0.02$	**84.01**	77.42	84.01

Table 11. **Effect of other design choices in RSD.**

Sensitivity to AAD expansion rate γ. We are interested to study how larger projectors with higher capability may impact the distillation performance of RSD. Let us denote the dimension expansion rate by γ. We experiment with different values of γ in Table 10. It can be observed that the performance of our method is relatively robust to varying capacities of the AAD module. A weak AAD module may become the performance bottleneck as it struggles to learn more useful representations according to the RSD criterion, but is still superior to OFA. A more capable AAD module tends to give better results, although it is not always the case. We postulate that when the AAD module is stronger, there is a greater decoupling effect. This may cause the student's internal representations to be insufficiently influenced by the RSD objective, thereby leading to degraded performance on the Mixer-B/16-to-DeiT-T set-up.

Other design considerations. In Table 11, we present further ablation studies to examine other design choices in-



(a) T. v.s. T.

(b) T. v.s. S. (w/o KD)

(c) T. v.s. S. (OFA)

(d) T. v.s. S. (RSD)

Figure 5. Visualisation of heterogeneous feature similarities via CKA across different teacher and student architectures. Top: ConvNeXt-T teacher and ResMLP-S12 student. Bottom: ViT-S teacher and MobileNetV2 student.

volved in our RSD formulation. We find that the normalisation operation in computing Pearson correlation coefficients helps improve learning, and removing it leads to performance degradation. RSD is also robust to the choice of the distance measure used in pulling the cross-architecture Pearson correlations towards the redundancy suppression target. κ modulates the importance of the decorrelation objective in RSD. When $\kappa = 0$, we completely give up decorrelating the feature dimensions. We observe that a very large κ degrades the performance, as the decorrelation objective overwhelms the invariance maximisation objective. Removing the decorrelation objective tends to hurt the performance.

Link to contrastive distillation. The proposed RSD objective shares some spirit with contrastive knowledge distillation, as both can be interpreted from an information bottleneck theory perspective [11, 17]. Essentially, one may also consider RSD as a form of *contrastive learning at the level of semantic units*. For RSD applied to the 1D representation embeddings, each value within the embedding is a feature unit that encodes distinct semantic knowledge. In this case, the batch-wise distributional pattern of each student semantic unit is maximally similar to its heterogeneous counterpart of the teacher, while those for other features units are maximally decorrelated. When RSD is applied over intermediate 2D feature maps, each semantic unit corresponds to the activation at a certain spatial location and a specific channel. Therefore, RSD effectively maximises the correlation between teacher and student activation distributions of the same spatial location and channel, while decorrelating those of different spatial locations and channels. In contrast, contrastive distillation methods are constrained to utilising sample-level contrastive learning, maintaining negative samples using a memory bank, and pulling apart negative samples, which makes them a vastly different group of methods than RSD. Additionally, we made attempts to make off-diagonal correlations in $\mathbf{P}$ anti-correlated rather than decorrelated, which should echo the idea of maximising dissimilarity between positive and negative samples, but obtained inferior performance.

CKA visualisation. We present higher-resolution visualisations for the CKA plots in the main text. Compared to OFA, our method can improve the heterogeneous feature similarity throughout the intermediate layers even though it does not directly access or process the intermediate features. Our method also effectively boosts feature similarity towards the final layers of the network.