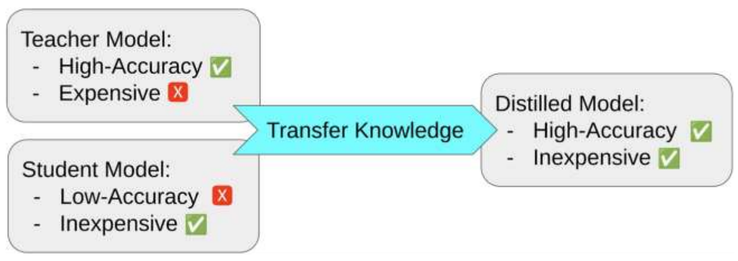
Logit-Based Losses Limit the Effectiveness of Feature Knowledge Distillation

Nicholas Cooper, Lijun Chen,
Sailesh Dwivedy, Danna Gurari

NEURAL INFORMATION PROCESSING SYSTEMS

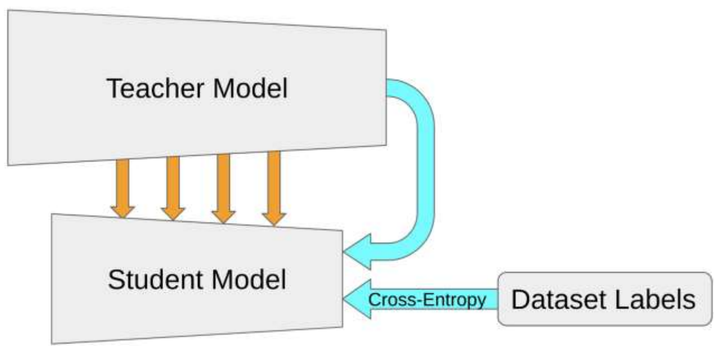


Knowledge distillation (KD) infuses the generalization ability of computationally expensive teacher models into lightweight student models.

Complete illustration of our feature knowledge distillation framework. The highest knowledge quality teacher layers are selected, and CE is only propagated through the classifier

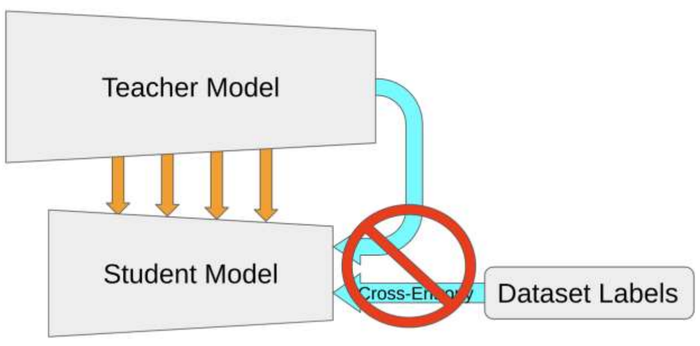
Teacher knowledge quality vs. hidden layer index. Separation shown in dark blue, information in light blue, efficiency in red, and total knowledge quality in gray. Best/standard layer selections indicated by black Xs/orange circles.

How to transfer knowledge?



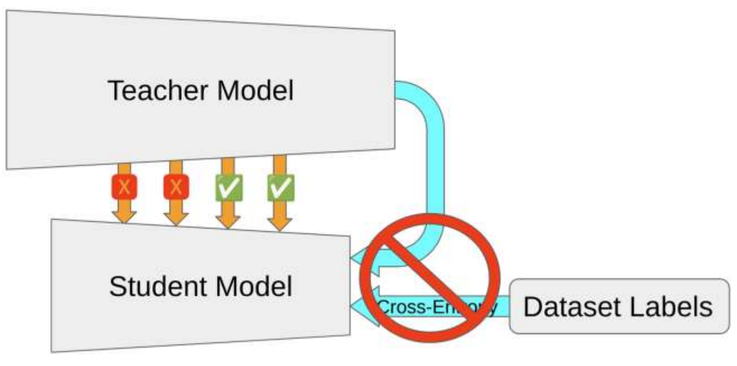
Logit-losses such as VKD [1] are computed at model output in *low-dimensional space*.
Feature-losses are computed at intermediate layers in *high-dimensional space*.
Multiple losses are combined in the final training recipe, including standard cross-entropy loss.

Training student backbones with *only* feature-based losses improves performance



We demonstrate that logit losses dilute the rich high-dimensional information transferred by feature losses. Removing these low-dimensional losses allows the student to maximally benefit from the rich high-dimensional feature losses.

What's the catch?



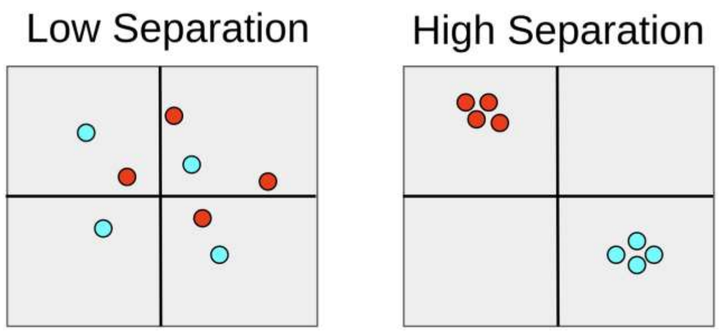
Naively removing all logit losses from the student training procedure can fail catastrophically if the student is not provided with enough information about the dataset labels. Care must be taken to select the "best" teacher layers. We address this by studying the geometric structure of teacher representations.

Quantifying Knowledge Quality

We compare teacher layers with three geometric measures: *separation, information,* and *efficiency*.

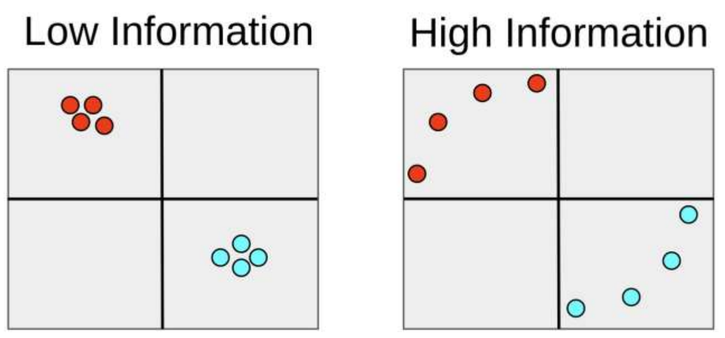
Separation is defined as the difference between the average within-class dot-product similarity and the average between-class dot-product similarity. It is a measure of how well the representations convey information about the ground truth labels.

$$\mathcal{S}(R) = avgDPW(R) - avgDPB(R)$$



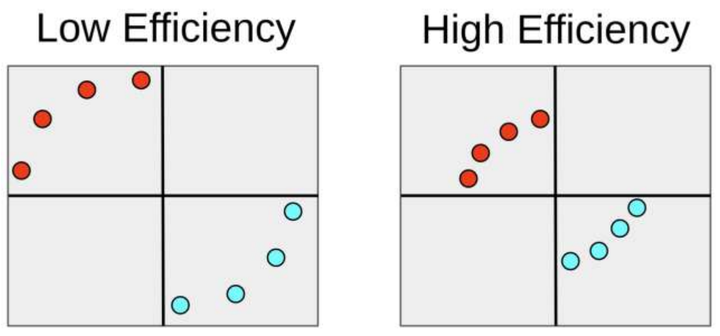
Information is defined as the product of the normalized SVD-entropy and the complement of the average intra-class similarity. It is a measure of the richness of the knowledge contained in the representations.

$$\mathcal{I}(R) = [1 - minDPW(R)]\, avgSVDE(R)$$



Efficiency is defined as the ratio of the "optimal" representation norm to the empirical norm. It is a measure of the size-efficiency of the features. The "optimal" norm is computed from a hypersphere packing problem based the representation intrinsic dimension.

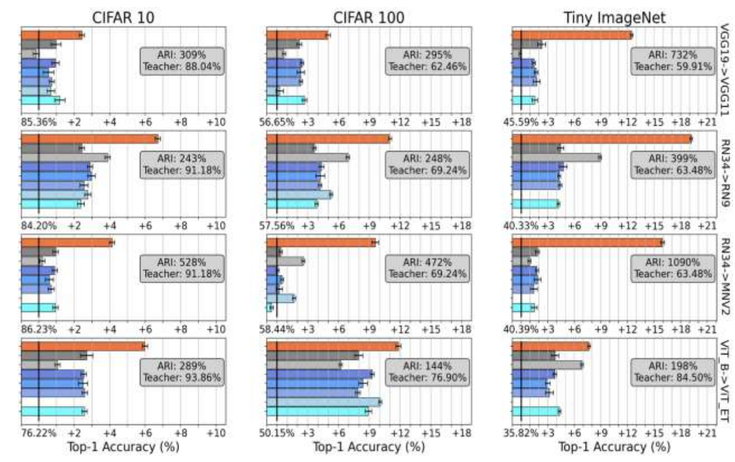
$$\mathcal{E}(R) = \frac{2K minDistB(R)}{avgNorm(R)} \quad K = \left(\frac{N}{\pi}\right)^{\frac{1}{D-1}}$$



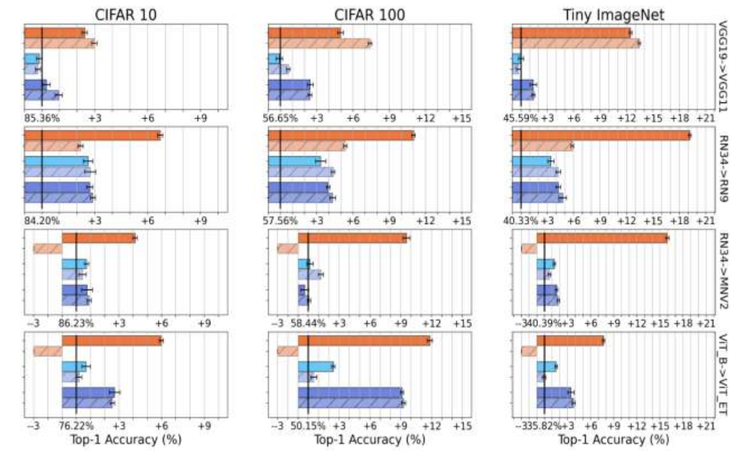
Knowledge Quality is constructed as follows:

$$\mathcal{Q}(R) := \mathcal{S}(R) + \sqrt{\mathcal{I}(R)\mathcal{E}(R)}$$

Experimental Results



Distillation results. Our method in orange, various baseline approaches in blues. Three image classification datasets of varied difficulty along the columns, and four popular teacher-student model pairs along the rows. ARI is the *average relative improvement* [2] of our method over baselines.



Fine-grained analysis. Solid/dashed bars denote ours/standard teacher layer selection; orange - no logit-losses, light blue - cross entropy, dark blue - cross entropy + vanilla KD [1].

We observe that representation knowledge quality closely follows the extraction and compression stages of deep neural networks [3, 4]. This is because the transition layers between stages exhibit high separation *and* high information *and* high efficiency. Selecting these "magic" layers ensures that the student receives sufficiently rich knowledge to converge without logit-losses.

Our proposed training recipe is highly effective at improving the generalization ability of the distilled students. On the standard KD benchmark of ResNet34->MobileNet V2 on CIFAR 100, our method boosts generalization performance by upwards of 5%. Detailed analysis shows that logit-losses act as "training wheels" which safeguard against poor knowledge quality features but limit the effectiveness of the distillation.

References

[1]: Hinton et al. "Distilling the knowledge in a neural network". 2015.

[2]: Krishnan et al. "Contrastive representation distillation". ICLR, 2021.

[3]: Brown et al. "Relating Regularization and Generalization through the Intrinsic Dimension of Activations". OPT, 2022.

[4]: Masarczyk et al. "The Tunnel Effect: Building Data Representations in Deep Neural Networks". NeurIPS, 2023.

Project GitHub