Cross-Layer Distillation with Semantic Calibration

Defang Chen,[1,2,3] **Jian-Ping Mei,**[4] **Yuan Zhang,**[1,3]
Can Wang,[1,2,3*] **Zhe Wang,**[1,3] **Yan Feng,**[1,3] **Chun Chen,**[1,3]

[1]College of Computer Science, Zhejiang University, China.
[2]Zhejiang Provincial Key Laboratory of Service Robot.
[3]Zhejiang University-LianlianPay Joint Research Center.
[4]College of Computer Science, Zhejiang University of Technology, China.
defchern@zju.edu.cn, jpmei@zjut.edu.cn, {yuan_zhang, wcan, fengyan, chenc}@zju.edu.cn

Abstract

Recently proposed knowledge distillation approaches based on feature-map transfer validate that intermediate layers of a teacher model can serve as effective targets for training a student model to obtain better generalization ability. Existing studies mainly focus on particular representation forms for knowledge transfer between manually specified pairs of teacher-student intermediate layers. However, semantics of intermediate layers may vary in different networks and manual association of layers might lead to negative regularization caused by semantic mismatch between certain teacher-student layer pairs. To address this problem, we propose Semantic Calibration for Cross-layer Knowledge Distillation (SemCKD), which automatically assigns proper target layers of the teacher model for each student layer with an attention mechanism. With a learned attention distribution, each student layer distills knowledge contained in multiple layers rather than a single fixed intermediate layer from the teacher model for appropriate cross-layer supervision in training. Consistent improvements over state-of-the-art approaches are observed in extensive experiments with various network architectures for teacher and student models, demonstrating the effectiveness and flexibility of the proposed attention based soft layer association mechanism for cross-layer distillation.

Introduction

The generalization ability of a lightweight model can be improved by training to match the prediction of a powerful model (Bucilua, Caruana, and Niculescu-Mizil 2006; Ba and Caruana 2014). This idea is popularized by knowledge distillation (KD) in which temperature scaling outputs from the teacher model are exploited to improve the performance of the student model (Hinton, Vinyals, and Dean 2015). Compared to discrete labels, soft targets predicted by the teacher model serve as an effective regularization to prevent the student model from being trapped in over-confident solutions during optimization (Pereyra et al. 2017; Müller, Kornblith, and Hinton 2019; Yuan et al. 2020).

In the vanilla KD framework, the knowledge learned by a classification model is represented only by the prediction of its final layer (Hinton, Vinyals, and Dean 2015). Although the relative probabilities assigned to different classes provide an intuitive understanding about how a model generalize, knowledge transfer in such a highly abstract form ignores a wealth of information contained in intermediate layers. Intending to further boost effectiveness of distillation, recent works proposed to align feature maps or their transformations of manually selected teacher-student layer pairs (Romero et al. 2015; Zagoruyko and Komodakis 2017; Ahn et al. 2019; Tung and Mori 2019; Passalis, Tzelepi, and Tefas 2020). An interpretation for the success of feature-map based distillation is that the multi-layer feature representations respect hierarchical concept learning process which may entail reasonable inductive bias (Bengio, Courville, and Vincent 2013).

Intermediate layers of teacher and student models with different capacity tend to have different levels of abstraction (Passalis, Tzelepi, and Tefas 2020). A peculiar challenge is thus to ensure appropriate layer associations in feature-map based distillation to achieve maximum performance improvement. However, existing efforts mainly focus on particular representations of feature maps to capture the enriched knowledge and enable knowledge transfer based on handcrafted layer assignments, such as random selection or one-to-one association (Romero et al. 2015; Zagoruyko and Komodakis 2017; Ahn et al. 2019; Tung and Mori 2019; Passalis, Tzelepi, and Tefas 2020). A naive allocation strategy may cause semantic mismatch between feature maps of candidate teacher-student layer pairs, leading to negative regularization effect in training of the student model. Since we have no access to prior knowledge of the semantic level of each intermediate layer, layer association becomes a non-trivial problem. Therefore, systematic approaches need to be developed for more effective and flexible knowledge transfer with feature maps.

In this paper, we propose **Sem**antic **C**alibration for **C**ross-layer **K**nowledge **D**istillation (SemCKD) to exploit intermediate knowledge by keeping the transfer in a matched semantic level. An attention mechanism is applied in our approach for automatic soft layer association, which effectively binds a student layer with those semantically similar target layers in the teacher model. Learning from multiple target layers with an attention allocation rather than turning to a fixed assignment can suppress over-regularization effect in training. To align the spatial dimension of each layer pair

*Corresponding author

for calculating the total loss, feature maps of each student layer are projected to the same dimension as those in the target layers. By taking advantage of semantic calibration and feature-map transfer across multiple layers, the student model can be effectively optimized with more appropriate guidance. The overall contributions of this paper are summarized as follows:

- We propose a novel technique to significantly improve effectiveness of feature-map transfer by semantic calibration via soft layer association. Our approach is readily applicable to heterogeneous settings where different architectures are used for the teacher and student models.
- Attention mechanism is used to achieve soft layer association for cross-layer distillation. Its capability to alleviate the semantic mismatch problem is supported by carefully designed experiments.
- Extensive experiments on CIFAR-100 and ImageNet datasets with a large variety of settings based on popular network architectures demonstrate that SemCKD consistently generalizes better than state-of-the-art approaches.

Related Work

Knowledge Distillation. KD serves as an effective recipe to improve the performance of a given student model by exploiting soft targets from a pre-trained teacher model (Hinton, Vinyals, and Dean 2015). Compared to discrete labels, fine-grained information among different categories provides extra supervision to optimize the student model better (Pereyra et al. 2017; Müller, Kornblith, and Hinton 2019). A new interpretation for the improvement is that soft targets act as a learned label smoothing regularization to keep the student model from producing over-confident predictions (Yuan et al. 2020). To save the expense of pre-training, some cost-effective online variants have been explored later (Anil et al. 2018; Chen et al. 2020).

Feature-Map Distillation. Rather than only formalizing knowledge in a highly abstract form like predictions, recent methods attempted to leverage information contained in intermediate layers by designing elaborate knowledge representations. A bunch of techniques have been developed for this purpose, such as aligning hidden layer responses called *hints* (Romero et al. 2015), mimicking spatial attention maps (Zagoruyko and Komodakis 2017), or maximizing the mutual information through variational principle (Ahn et al. 2019). The transferred knowledge can also be captured by crude pairwise activation similarities (Tung and Mori 2019) or hybrid kernel formulations built on them (Passalis, Tzelepi, and Tefas 2020). With the pre-defined representations, all of the above methods perform knowledge transfer with certain hand-crafted layer associations, such as random selection or one-to-one match. Unfortunately, as pointed in (Passalis, Tzelepi, and Tefas 2020), these hard associations would make the student model suffer from negative regularization, which limits the effectiveness of feature-map distillation. Based on the transfer learning framework, a newly solution is to learn association weights by a meta-network given only feature maps of the source network (Jang et al. 2019), while our proposed approach incorporates more information from teacher-student layer pairs.

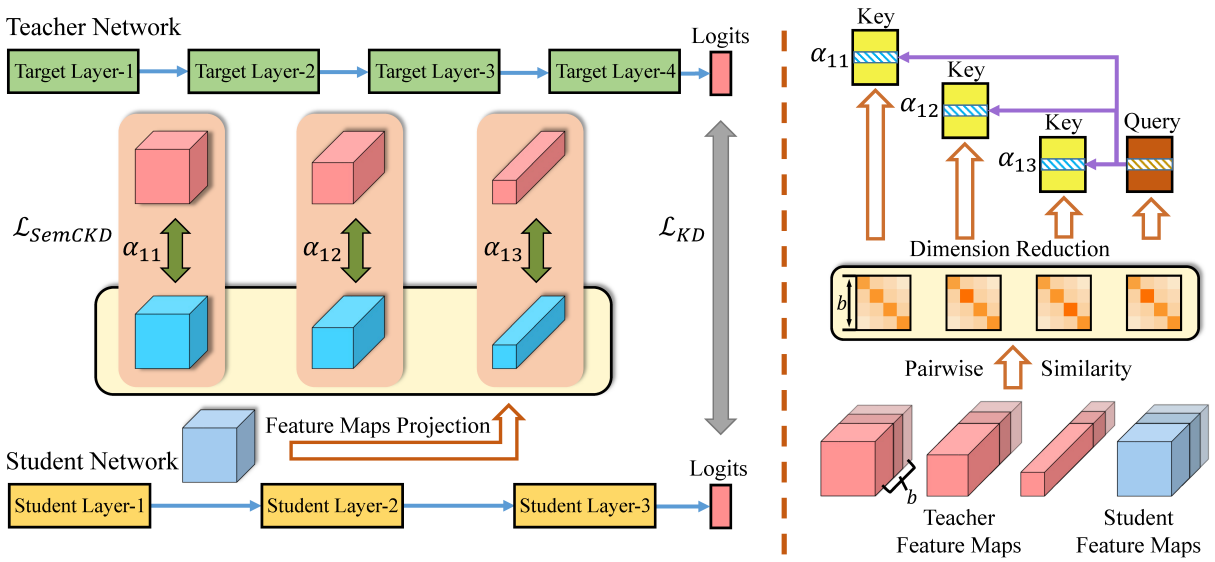
Feature-Embedding Distillation. Feature embedding is a good substitute for feature maps since low dimensional vectors are more tractable than high dimensional tensors. Meanwhile, feature embedding also preserves more structural information compared to the final predictions. Therefore, a variety of knowledge distillation approaches have been proposed based on feature embedding, especially the generated relational graphs where each node represents one instance (Passalis and Tefas 2018; Peng et al. 2019; Park et al. 2019; Liu et al. 2019). The main difference among these methods lies on how edge weights are constructed. Typical choices include cosine kernel (Passalis and Tefas 2018), truncated Gaussian RBF kernel (Peng et al. 2019), or combination of distance-wise as well as angle-wise potential functions (Park et al. 2019). In contrast to pairwise transfer, CRD formulates distillation as contrastive learning to capture higher-order dependencies in the representation space (Tian, Krishnan, and Isola 2020). Although our method mainly focuses on feature-map distillation, it is also compatible with the state-of-the-art feature-embedding distillation approach to further improve the performance.

Semantic Calibration for Distillation

Background and Notations

In this section, we briefly recap the basic concepts of classic knowledge distillation as well as provide necessary notations for the following illustration. Given a training dataset $\mathcal{D} = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^N$ consisting of N instances from K categories, and a powerful teacher model pre-trained on the dataset $\mathcal{D}$, our goal is reusing the same dataset to train another simple student model with cheaper computational and storage demand. For a mini-batch with size b, we denote the output of each target layer t_l and student layer s_l as $F^t_{t_l} \in \mathbb{R}^{b\times c_{t_l}\times h_{t_l}\times w_{t_l}}$ and $F^s_{s_l} \in \mathbb{R}^{b\times c_{s_l}\times h_{s_l}\times w_{s_l}}$, respectively, where c is the number of output channels, h and w are spatial dimensions, superscript t and s reflect the corresponding models. The value of candidate layers t_l and s_l range from 1 to t_L and s_L, respectively. Note that t_L and s_L may be different especially when the teacher and student architectures are different. The representations at the penultimate layer from the teacher and student models are denoted as $F^t_{t_L}$ and $F^s_{s_L}$, which are mainly used in feature-embedding distillation. Take the student model as an example, outputs of the last fully connected layer $g(\cdot)$ are known as logits $\boldsymbol{g}^s_i = g(F^s_{s_L}[i]) \in \mathbb{R}^K$ and the predicted probabilities are calculated with a softmax layer built on logits, i.e., $\boldsymbol{p}^s_i = \sigma(\boldsymbol{g}^s_i/T)$ with T usually equals to 1. The notation $F^s_{s_l}[i]$ denotes the output of student layer s_l for the i-th instance and is a shorthand for $F^s_{s_l}[i,:,:,:]$.

For classification tasks, in additional to regular cross entropy loss (CE) between the predicted probabilities $\boldsymbol{p}^s_i$ and the one-hot label $\boldsymbol{y}_i$ of each training sample, classic knowledge distillation (Hinton, Vinyals, and Dean 2015) incorporates another alignment loss to encourage the minimization of Kullback-Leibler (KL) divergence between $\boldsymbol{p}^s_i$ and soft



(a) Semantic Calibration Knowledge Distillation (b) Attention Allocation

Figure 1: An overview of the proposed Semantic Calibration for Knowledge Distillation (SemCKD). ($\boldsymbol{a}$) Feature maps for certain instance from the student layer-1 are projected into three individual forms to align with the spatial dimension of those from each target layer. The learned attention allocation adaptively helps the student model focus on the most semantic-related information for effective distillation. ($\boldsymbol{b}$) Pairwise similarities are first calculated between every stacked feature maps and then the attention weights are obtained by the proximities among generated *query* and *key* factors.

targets $\boldsymbol{p}_i^t$ of the teacher model

$$\mathcal{L}_{KD_i} = \mathcal{L}_{CE}(\boldsymbol{y}_i, \sigma(\boldsymbol{g}_i^s)) + T^2 \mathcal{L}_{KL}(\sigma(\boldsymbol{g}_i^t/T), \sigma(\boldsymbol{g}_i^s/T)), \tag{1}$$

where T is a hyper-parameter and a higher T leads to more considerable softening effect. We set T to 4 throughout this paper for fair comparison.

Feature-Map Distillation

As mentioned earlier, feature maps of a teacher model are valuable for helping a student model achieve better performance. Recently proposed feature-map distillation approaches can be summarized as adding the following loss term to Equation (1) for each mini-batch with size b

$$\mathcal{L}_{FMD} = \sum_{(s_l,t_l)\in\mathcal{C}} Dist\left(Trans^t\left(F_{t_l}^t\right), Trans^s\left(F_{s_l}^s\right)\right), \tag{2}$$

leading to the overall loss as

$$\mathcal{L}_{total} = \sum_{i=1}^{b} \mathcal{L}_{KD_i} + \beta \mathcal{L}_{FMD}, \tag{3}$$

where functions $Trans^t(\cdot)$ and $Trans^s(\cdot)$ in each method transform feature maps of candidate teacher-student layer pairs into a particular hand-designed representation, such as attention maps (Zagoruyko and Komodakis 2017) or pairwise similarity matrices (Tung and Mori 2019). The layer association sets $\mathcal{C}$ of existing methods are generated by random selection or one-to-one match. However, these simple association strategies may cause the loss of useful information. Take one-to-one match as an example, extra layers have to be discarded when the number of layers s_L and t_L are different, i.e., $\mathcal{C} = \{(1,1), ..., (min(s_L,t_L), min(s_L,t_L))\}$. With these associated layer pairs, the feature-map distillation loss is calculated by distance function $Dist(\cdot,\cdot)$. The hyper-parameter β in Equation (3) is used to balance two individual loss terms.

Rather than performing knowledge transfer based on fixed associations between candidate teacher-student layer pairs, our approach aims to learn associations for semantic calibrated cross-layer distillation.

Semantic Calibration Formulation

In our approach SemCKD, each student layer is automatically associated with those semantic-related target layers by attention allocation, as illustrated in Figure 1. Training with soft associations encourages the student model to collect and integrate multi-layer information to obtain a more suitable regularization. Moreover, SemCKD is readily applicable to the situation where the number of candidate layers from the teacher and student models differ.

The learned association set $\mathcal{C}$ in SemCKD is denoted as

$$\mathcal{C} = \{(s_l, t_l) \mid \forall\, s_l \in [1, ..., s_L],\ t_l \in [1, ..., t_L]\}, \tag{4}$$

with the corresponding weight satisfies $\sum_{t_l=1}^{t_L} \boldsymbol{\alpha}_{(s_l,t_l)} = \boldsymbol{1}, \forall\, s_l \in [1, ..., s_L]$. The weight $\boldsymbol{\alpha}_{(s_l,t_l)} \in \mathbb{R}^{b\times 1}$ represents the extent to which the target layer t_l is attended in deriving the semantic-aware guidance for the student layer s_l. We will elaborate on these attention-based weights later. All the feature maps of each student layer are projected into t_L individual forms to align with the spatial dimension of each of

target layers for the following distance calculation

$$F_{t_l}^{s'} = Proj\left(F_{s_l}^s \in \mathbb{R}^{b\times c_{s_l}\times h_{s_l}\times w_{s_l}}, t_l\right), t_l \in [1, ..., t_L], \tag{5}$$

with $F_{t_l}^{s'} \in \mathbb{R}^{b\times c_{t_l}\times h_{t_l}\times w_{t_l}}$. Each function $Proj(\cdot,\cdot)$ includes a stack of three layers with 1×1, 3×3 and 1×1 convolutions to meet the demand of capability for effective transformation[1].

Loss function. For a mini-batch with size b, the student model produces several feature maps across multiple layers, i.e., $F_{s_1}^s, ..., F_{s_L}^s$. After semantic layer associations and dimensional projections, the $\mathcal{L}_{FMD}$ loss of SemCKD is obtained by simply using Mean-Square-Error (MSE)

$$\begin{aligned}\mathcal{L}_{SemCKD} &= \sum_{(s_l,t_l)\in\mathcal{C}} \boldsymbol{\alpha}_{(s_l,t_l)} Dist\left(F_{t_l}^t, Proj\left(F_{s_l}^s, t_l\right)\right) \\ &= \sum_{s_l=1}^{s_L}\sum_{t_l=1}^{t_L}\sum_{i=1}^{b} \alpha_{(s_l,t_l)}^i \mathrm{MSE}\left(F_{t_l}^t[i], F_{t_l}^{s'}[i]\right),\end{aligned} \tag{6}$$

where feature maps from each student layer is transformed by a projection function $Trans^s(\cdot) = Proj(\cdot,\cdot)$ while those from the target layers remain unchanged by identity transformation $Trans^t(\cdot) = I(\cdot)$. The i-th element of vector $\boldsymbol{\alpha}_{(s_l,t_l)}$ is denoted as $\alpha_{(s_l,t_l)}^i$ for the corresponding instance. Equipped with the learned attention distributions, the total loss is aggregated by a weighted summation of each individual distance among the feature maps from candidate teacher-student layer pairs. Note that FitNet (Romero et al. 2015) is a special case of SemCKD by fixing $\alpha_{(s_l,t_l)}^i$ to 1 for certain (s_l, t_l) layer pair and 0 for the rest ones.

Attention Allocation. Feature representations contained in a trained neural network are progressively more abstract as the layer depth increases. Semantic level of those intermediates can vary among teacher and student architectures with different capacity. To further improve the performance of feature-map distillation, each student layer had better associate with the most semantic-related target layers to derive its own regularization. Random selection or forcing feature maps from the same layer depths to be aligned may not suffice due to negative effects from those mismatched layers.

Layer associations based on attention mechanism provides a potentially feasible solution to this problem. Since feature maps produced by similar instances probably become clustered at separate granularity in different layers, the proximity of pairwise similarity matrices can be regarded as a good measurement of the inherent semantic similarity (Tung and Mori 2019). These similarity matrices are calculated as

$$A_{s_l}^s = R(F_{s_l}^s)\cdot R(F_{s_l}^s)^T \quad A_{t_l}^t = R(F_{t_l}^t)\cdot R(F_{t_l}^t)^T, \tag{7}$$

where $R(\cdot) : \mathbb{R}^{b\times c\times h\times w} \mapsto \mathbb{R}^{b\times chw}$ is a reshaping operation, and therefore $A_{s_l}^s$ and $A_{t_l}^t$ are $b\times b$ matrices.

Based on the self-attention framework (Vaswani et al. 2017), we separately project the pairwise similarity matrices of each student layer and target layers into two subspaces by a Multi-Layer Perceptron (MLP) to alleviate the effect of noise and sparseness. For i-th instance

$$Q_{s_l}[i] = \mathrm{MLP}_Q(A_{s_l}^s[i]) \quad K_{t_l}[i] = \mathrm{MLP}_K(A_{t_l}^t[i]). \tag{8}$$

The parameters of $\mathrm{MLP}_Q(\cdot)$ and $\mathrm{MLP}_K(\cdot)$ are learned during training to generate *query* and *key* vectors and shared by all instances. Then, $\alpha_{(s_l,t_l)}^i$ is calculated as follows

$$\alpha_{(s_l,t_l)}^i = \frac{e^{Q_{s_l}[i]^T K_{t_l}[i]}}{\sum_j e^{Q_{s_l}[i]^T K_{t_j}[i]}}. \tag{9}$$

Attention-based allocation provides a possible way to suppress negative effects caused by layer mismatch and integrate positive guidance from multiple target layers, which is validated by Figure 2 and Table 3.

Although the proposed approach distills only the knowledge contained in intermediate layers, its performance can be further boosted by incorporating additional orthogonal techniques, e.g., feature-embedding transfer as shown in Table 5. The full training procedure with the proposed semantic calibration formulation is summarized in Algorithm 1.

Algorithm 1 Semantic Calibration for Distillation.

Input: Training dataset $\mathcal{D} = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^N$; A pre-trained teacher model with parameter θ^t; A student model with randomly initialized parameters θ^s;
Output: A well-trained student model;

```
while θ^s is not converged do
    Sample a mini-batch B with size b from D.
    Forward propagation B into θ^t and θ^s to obtain intermediate presentations F_{t_l}^t and F_{s_l}^s across layers.
    Construct pairwise similarity matrices A_{t_l}^t and A_{s_l}^s as Equation (7).
    Perform attention allocation as Equation (8-9).
    Align feature maps by projections as Equation (5).
    Update parameters θ^s by backward propagation the gradients of the loss in Equation (3) and Equation (6).
end while
```

Experiments

To demonstrate the effectiveness of the proposed semantic calibration strategy for cross-layer knowledge distillation, we conduct a series of classification tasks on the CIFAR-100 (Krizhevsky and Hinton 2009) and ImageNet datasets (Russakovsky et al. 2015). A large variety of teacher-student combinations based on popular network architectures are evaluated, including VGG (Simonyan and Zisserman 2015), ResNet (He et al. 2016), WRN (Zagoruyko and Komodakis 2016), MobileNet (Sandler et al. 2018) and ShuffleNet (Ma et al. 2018). In addition to comparing SemCKD with representative feature-map distillation approaches, we also provide results to support and explain the success of our semantic calibration strategy in helping student models obtain a proper regularization through three carefully designed experiments. Ablation studies on the attention mechanism

[1] In practice, we first use a pooling operation to align the height and weight dimensions of the $F_{s_l}^t$ and $F_{s_l}^s$ before projections to reduce computational consumption.

Student	VGG-8 70.46 ± 0.29	VGG-13 74.82 ± 0.22	ShuffleNetV2 72.60 ± 0.12	ShuffleNetV2 72.60 ± 0.12	MobileNetV2 65.43 ± 0.29	VGG-8 70.46 ± 0.29	ResNet-8x4 73.09 ± 0.30	ARI (%)
KD	72.73 ± 0.15	77.17 ± 0.11	75.60 ± 0.21	75.49 ± 0.24	68.70 ± 0.22	73.38 ± 0.05	74.42 ± 0.05	72.65 %
FitNet	72.91 ± 0.18	77.06 ± 0.14	75.44 ± 0.11	75.82 ± 0.22	68.64 ± 0.12	73.63 ± 0.11	74.32 ± 0.08	71.92 %
AT	71.90 ± 0.13	77.23 ± 0.19	75.41 ± 0.10	75.91 ± 0.14	68.79 ± 0.13	73.51 ± 0.08	75.07 ± 0.03	75.21 %
SP	73.12 ± 0.10	77.72 ± 0.33	75.54 ± 0.18	75.77 ± 0.08	68.48 ± 0.36	73.53 ± 0.23	74.29 ± 0.07	64.95 %
VID	73.19 ± 0.23	77.45 ± 0.13	75.22 ± 0.07	75.55 ± 0.18	68.37 ± 0.24	73.63 ± 0.07	74.55 ± 0.10	64.11 %
HKD	72.63 ± 0.12	76.76 ± 0.13	76.24 ± 0.09	76.64 ± 0.05	69.23 ± 0.16	73.06 ± 0.24	74.86 ± 0.21	61.23 %
SemCKD	**75.27 ± 0.13**	**79.43 ± 0.02**	**76.39 ± 0.12**	**77.62 ± 0.32**	**69.61 ± 0.05**	**74.43 ± 0.25**	**76.23 ± 0.04**	–
Teacher	ResNet-32x4 79.42	ResNet-32x4 79.42	VGG-13 74.64	ResNet-32x4 79.42	WRN-40-2 75.61	VGG-13 74.64	ResNet-32x4 79.42	Average 68.34 %

Table 1: Top-1 test accuracy of *feature-map distillation* approaches on CIFAR-100.

as well as dimensional projection are also conducted. Finally, we show that SemCKD is compatible with the feature-embedding distillation technique to achieve better results and analyze its sensitivity to the hyper-parameter β.

All evaluations are made in comparison to state-of-the-art approaches based on standard experimental settings and reported in means and standard deviations We regard the building blocks of teacher and student networks as *target layer* and *student layer* in practice for convenience. The detailed descriptions of computing infrastructure, network architectures, data processing, hyper-parameters in model optimization for reproducibility as well as more results are included in the technical appendix. The code is available at https://github.com/DefangChen/SemCKD.

Comparison of Feature-Map Distillation Approaches

Table 1 gives the Top-1 test accuracy (%) on CIFAR-100 based on seven different network combinations, which consist of two homogeneous settings, i.e. the teacher and student models share similar architectures (VGG-8/13, ResNet-8x4/32x4), and five heterogeneous settings. Each column apart from the first row includes the results of corresponding student models which are generated under the supervision of the same teacher model. The results of the vanilla KD are also included for comparison. According to Table 1, it is shown that SemCKD consistently achieves higher accuracy than state-of-the-art feature-map distillation approaches.

In order to obtain an intuitive sense about quantitative improvement, we adopt *Average Relative Improvement* (ARI) as the previous work (Tian, Krishnan, and Isola 2020)

$$ARI = \frac{1}{M}\sum_{i=1}^{M}\frac{Acc^i_{SemCKD} - Acc^i_{FMD}}{Acc^i_{FMD} - Acc^i_{STU}} \times 100\%, \quad (10)$$

where M is the number of different architecture combinations and Acc^i_{SemCKD}, Acc^i_{FMD}, Acc^i_{STU} refer to the accuracies of SemCKD, a certain feature-map distillation approach and a regularly trained student model in the i-th setting, respectively. This evaluation metric reflects the extent to which SemCKD further improves on the basis of existing approaches compared to improvements made by these approaches upon the baseline student models.

Student	ResNet-18 69.67	ShuffleV2x0.5 53.78	ResNet-18 69.67
KD	70.62	53.73	70.54
FitNet	70.31	51.46	70.42
AT	70.30	52.83	70.30
SP	69.99	51.73	70.12
VID	70.30	53.97	70.26
HKD	68.86	51.60	68.44
SemCKD	**70.87**	**53.99**	**70.66**
Teacher	ResNet-34 73.26	ResNet-34x4 73.54	ResNet-34x4 73.54

Table 2: Top-1 test accuracy of *feature-map distillation* approaches on ImageNet.

On average, SemCKD shows significantly relative improvement (68.34%) over all of the compared methods. Specifically, comparing with VID, which is the newest feature-map distillation approach under a single teacher-student training process, the relative improvement of SemCKD for each of the cases are 80.83%, 58.83%, 28.80%, 58.19%, 37.25%, 29.21%, respectively, leading to 64.11% in ARI. As for HKD, which relies on a costly teacher-auxiliary-student paradigm, the ARI becomes rather small on two settings (3.93% for "ShuffleNetV2 & ResNet-32x4", 9.98% for "MobileNetV2 & WRN-40-2"). But in general, SemCKD still relatively outperforms HKD for about 61.23%, showing that our approach can indeed make better use of intermediate information for effective distillation.

We also find that none of the compared methods can consistently beats the vanilla KD on CIFAR-100, which probably due to semantic mismatch among associated layer pairs. This problem becomes especially severe for random selection (FitNet method fails in 4/7 cases) and the situation where the number of candidate layers s_L is larger than t_L (4/5 of methods fail in the "ShuffleNetV2 & VGG-13" setting). Nevertheless, the semantic calibration formulation helps alleviate semantic mismatch to a great extent, leading to satisfied performance of SemCKD.

Table 2 shows the results on a large-scale image classification dataset and similar observations are obtained as above.

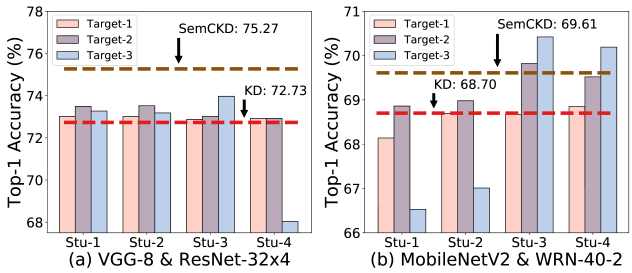


Figure 2: Negative regularization effect on CIFAR-100.

FitNet	AT	SP	VID	HKD	SemCKD
12.05	15.52	16.30	15.82	19.86	**11.27**

Table 3: Semantic Mismatch Score (log-scale) for VGG-8 & ResNet-32x4 on CIFAR-100.

Semantic Calibration Analysis

In this section, we experimentally study the negative regularization effect caused by manually specified layer associations and provide some explanations for the success of SemCKD by the proposed criterion and visual evidence.

Negative regularization effect occurs when feature-map distillation with certain layer association performs poorer than the vanilla KD. To reveal its existence, we conduct experiments by training the student model with only one specified teacher-student layer pair in "VGG-8 & ResNet-32x4" and "MobileNetV2 & WRN-40-2" settings. In both cases, the number of candidate target layers and student layers are 3 and 4, respectively. Figure 2 shows the results of student models with these 12 teacher-student layer combinations under the two settings on CIFAR-100. For better comparison, the results of the vanilla KD and SemCKD are plotted as dash horizontal lines with different colors.

As shown in Figure 2, the performance of a student model becomes extremely poor for some layer associations, which is probably caused by large semantic gaps. Typical results that suffer from negative regularization are "Student Layer-4 & Target Layer-3" in Figure 2a and "Student Layer-1, 2 & Target Layer-3" in Figure 2b. Another finding is that one-to-one layer matching is suboptimal since better results can be achieved by exploiting the information in a target layer with different depth, such as "Student Layer-1 & Target Layer-2" in Figure 2b. Although training with certain hand-craft layer association could outperform SemCKD in a few cases, such as "Student Layer-3 & Target Layer-3" in Figure 2b, SemCKD still performs reasonably well against a large selection of associations, especially the knowledge of the best layer association for each network combination is not available in advance. Nevertheless, those cases in which training with SemCKD are inferior to the best layer association indicates that there is room for refinement of our association strategy.

We then evaluate whether SemCKD actually leads to less semantic mismatch solutions compared with other approaches. A criterion called *semantic mismatch score* is proposed and measured by the average Euclidean distance between the similarity matrices generated by feature maps of each associated teacher-student layer pair, which hopefully represents the degree of difference between the captured pairwise similarity among instances in certain semantic level. As shown in Table 3, a lower semantic mismatch score is achieved by SemCKD thanks to our soft layer association mechanism. Detailed formulation as well as the calculation are provided in the technical appendix.

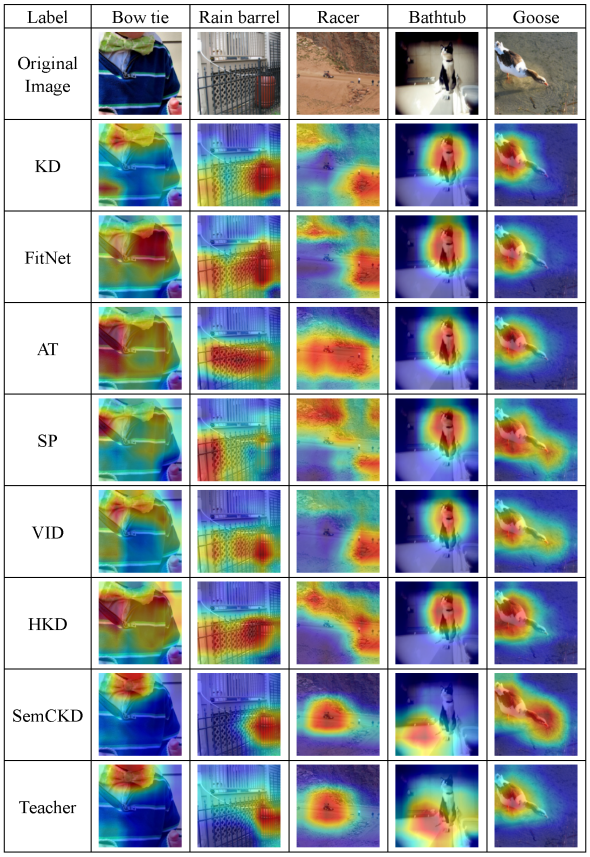


Figure 3: Grad-CAM visualization of *feature-map distillation* approaches on ImageNet. Region with a darker red is more important for the prediction. Best viewed in color.

To further provide visual explanations for the advantage of SemCKD, we randomly select several images from ImageNet labeled by "Bow tie", "Rain barrel", "Racer", "Bathhub" and "Goose", and use Grad-CAM (Selvaraju et al. 2017) to highlight the regions which are considered to be important for predicting the corresponding labels. As shown in Figure 3, the class-discriminative regions is centralized by SemCKD which is similar to the teacher model while being scatted around the surroundings by compared methods. As visualized in the fifth column, another failure mode of compared methods is that they sometimes regard the right regions as background while putting their attention on the spatial adjacency object. Moreover, SemCKD can capture more semantic-related information like highlighting the head and neck to identify a "Goose" in the image.

Equal Alloc	w/o Proj	w/o MLP	SemCKD
72.94 ± 0.87	72.51 ± 0.16	72.78 ± 0.29	75.27 ± 0.13

Table 4: Ablation study: Top-1 test accuracy for VGG-8 & ResNet-32x4 on CIFAR-100.

Ablation Study

Table 4 presents the evaluation of three SemCKD variants to further show the benefit of each individual component.

(1) Equal Allocation. In order to validate the effectiveness of allocating the attention of each student layer to multiple target layers, equal weight assignment is applied instead. This causes a lower accuracy by 2.33% (From 75.27% to 72.94%) and a considerably larger variance by 0.74%.

(2) w/o Projection. Rather than projecting the feature maps of each student layer to the same dimension as those in the target layers by Equation (5), we add a new $\mathrm{MLP}_V(\cdot)$ to project the pairwise similarity matrices of teacher-student layer pairs into another subspace to generate *value* vectors. Thus the Mean-Square-Error among feature maps in Equation (6) is replaced by these *value* vectors to calculate the overall loss, which reduces the performance by 2.76%.

(3) w/o MLP. A simple linear transformation is used to obtain *query* and *key* vectors in Equation (8) instead of the two-layer non-linear transformation, i.e., $\mathrm{MLP}(\cdot)$. The 2.49% performance drop indicates that the usefulness of $\mathrm{MLP}(\cdot)$ to alleviate the effect of noise and sparseness.

Extension to Feature-Embedding Distillation Approaches

Knowledge transfer based on feature embedding of the penultimate layer is another alternative to improve the generalization ability of student models. The results in Table 5 confirm that our approach holds a very satisfying property that it is highly compatible with the state-of-the-art feature-embedding distillation approach to achieve the better performance. We compare the performance of each student model trained with several newly proposed methods on three teacher-student network combinations. It is observed that by simply adding the loss term of CRD (Tian, Krishnan, and Isola 2020) into the original loss function of SemCKD without tuning any hyper-parameter, the performance has already been further boosted. Specifically, the ARI of SemCKD+CRD over CRD and SemCKD is about 40.13% and 13.90%, respectively.

Sensitivity Analysis

Finally, we evaluate the impact of hyper-parameter β on the performance of knowledge distillation. We compare three representative knowledge distillation approaches, including logits transfer (KD), feature-embedding transfer (CRD) and feature-map transfer (SemCKD). The range of hyper-parameter β for SemCKD is set as 100 to 1100 at equal interval of 100, while the hyper-parameter β for CRD ranges from 0.5 to 1.5 at equal interval of 0.1, adopting the same search space as the original paper (Tian, Krishnan, and Isola 2020). Note that the hyper-parameter β always equals to 0 for the vanilla KD, leading to a horizontal line in Figure 4.

Student	VGG-8	MobileNetV2	ResNet-8x4
	70.46 ± 0.29	65.43 ± 0.29	73.09 ± 0.30
PKT	73.11 ± 0.21	68.68 ± 0.29	74.61 ± 0.25
RKD	72.49 ± 0.08	68.71 ± 0.20	74.36 ± 0.23
IRG	72.57 ± 0.20	68.83 ± 0.18	74.67 ± 0.15
CC	72.63 ± 0.30	68.68 ± 0.14	74.50 ± 0.13
CRD	73.54 ± 0.19	69.98 ± 0.27	75.59 ± 0.07
SemCKD	75.27 ± 0.13	69.61 ± 0.05	76.23 ± 0.04
SemCKD+CRD	$\mathbf{75.52 \pm 0.09}$	$\mathbf{70.55 \pm 0.11}$	$\mathbf{76.68 \pm 0.19}$
Teacher	ResNet-32x4	WRN-40-2	ResNet-32x4
	79.42	75.61	79.42

Table 5: Top-1 test accuracy of *feature-embedding distillation* approaches on CIFAR-100.

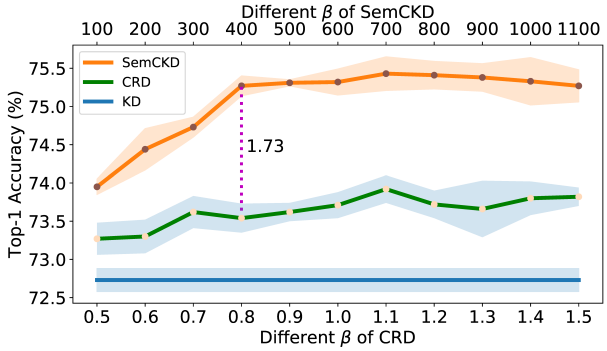


Figure 4: Impact of the hyper-parameter β for VGG-8 & ResNet-32x4 on CIFAR-100.

It is seen that SemCKD achieves the best results in all cases and outperforms CRD at about 1.73 absolute accuracy for the default hyper-parameter setting. Figure 4 also shows that the performance of SemCKD keeps very stable after the hyper-parameter β is greater than 400, which indicates our proposed method works reasonably well in a wide range of search space for the hyper-parameter β.

Conclusion

Feature maps produced by multiple intermediate layers of a powerful teacher model are valuable for improving knowledge transfer performance. A peculiar challenge for feature-map distillation is to ensure an appropriate association of teacher-student layer pairs. To alleviate negative regularization effect due to semantic mismatch between certain pairs of teacher-student intermediate layers, we propose semantic calibration via attention allocation for effective cross-layer distillation. Each student layer in our approach distills knowledge contained in multiple target layers with an automatically learned attention distribution to obtain proper supervision. Experimental results show that training with SemCKD leads to a relative low-level semantic mismatch score and its generalization ability outperforms the compared approaches. Visualization as well as detailed analysis provide some insights to the working principle of SemCKD.

Acknowledgments

This work is funded by National Key R&D Program of China (Grant No: 2018AAA0101505) and State Grid Corporation of China Scientific and Technology Project: Fundamental Theory of Human-in-the-loop Hybrid-Augmented Intelligence for Power Grid Dispatch and Control. The authors would like to thank the helpful comments from Ziying Guo and anonymous reviewers.

References

Ahn, S.; Hu, S. X.; Damianou, A. C.; Lawrence, N. D.; and Dai, Z. 2019. Variational Information Distillation for Knowledge Transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 9163–9171.

Anil, R.; Pereyra, G.; Passos, A.; Ormándi, R.; Dahl, G. E.; and Hinton, G. E. 2018. Large scale distributed neural network training through online distillation. In *International Conference on Learning Representations*.

Ba, J.; and Caruana, R. 2014. Do Deep Nets Really Need to be Deep? In *Advances in Neural Information Processing Systems*, 2654–2662.

Bengio, Y.; Courville, A. C.; and Vincent, P. 2013. Representation Learning: A Review and New Perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.* 35(8): 1798–1828.

Bucilua, C.; Caruana, R.; and Niculescu-Mizil, A. 2006. Model compression. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 535–541.

Chen, D.; Mei, J.-P.; Wang, C.; Feng, Y.; and Chen, C. 2020. Online Knowledge Distillation with Diverse Peers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 3430–3437.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.

Hinton, G. E.; Vinyals, O.; and Dean, J. 2015. Distilling the Knowledge in a Neural Network. *arXiv preprint arXiv:1503.02531* .

Jang, Y.; Lee, H.; Hwang, S. J.; and Shin, J. 2019. Learning What and Where to Transfer. In *International Conference on Machine Learning*.

Krizhevsky, A.; and Hinton, G. 2009. Learning multiple layers of features from tiny images. *Technical Report* .

Liu, Y.; Cao, J.; Li, B.; Yuan, C.; Hu, W.; Li, Y.; and Duan, Y. 2019. Knowledge Distillation via Instance Relationship Graph. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7096–7104.

Ma, N.; Zhang, X.; Zheng, H.; and Sun, J. 2018. ShuffleNet V2: Practical Guidelines for Efficient CNN Architecture Design. In *Proceedings of the European Conference on Computer Vision*, 122–138.

Müller, R.; Kornblith, S.; and Hinton, G. E. 2019. When Does Label Smoothing Help? In *Advances in Neural Information Processing Systems*.

Park, W.; Kim, D.; Lu, Y.; and Cho, M. 2019. Relational Knowledge Distillation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3967–3976.

Passalis, N.; and Tefas, A. 2018. Learning Deep Representations with Probabilistic Knowledge Transfer. In *European Conference on Computer Vision*, 283–299.

Passalis, N.; Tzelepi, M.; and Tefas, A. 2020. Heterogeneous Knowledge Distillation using Information Flow Modeling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Peng, B.; Jin, X.; Li, D.; Zhou, S.; Wu, Y.; Liu, J.; Zhang, Z.; and Liu, Y. 2019. Correlation Congruence for Knowledge Distillation. In *International Conference on Computer Vision*, 5006–5015.

Pereyra, G.; Tucker, G.; Chorowski, J.; Kaiser, Ł.; and Hinton, G. 2017. Regularizing neural networks by penalizing confident output distributions. *arXiv preprint arXiv:1701.06548* .

Romero, A.; Ballas, N.; Kahou, S. E.; Chassang, A.; Gatta, C.; and Bengio, Y. 2015. FitNets: Hints for thin deep nets. In *International Conference on Learning Representations*.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M. S.; Berg, A. C.; and Li, F. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision* 115(3): 211–252.

Sandler, M.; Howard, A. G.; Zhu, M.; Zhmoginov, A.; and Chen, L. 2018. MobileNetV2: Inverted Residuals and Linear Bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4510–4520.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In *International Conference on Computer Vision*, 618–626.

Simonyan, K.; and Zisserman, A. 2015. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*.

Tian, Y.; Krishnan, D.; and Isola, P. 2020. Contrastive Representation Distillation. In *International Conference on Learning Representations*.

Tung, F.; and Mori, G. 2019. Similarity-Preserving Knowledge Distillation. In *International Conference on Computer Vision*, 1365–1374.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, 5998–6008.

Yuan, L.; Tay, F. E.; Li, G.; Wang, T.; and Feng, J. 2020. Revisiting Knowledge Distillation via Label Smoothing Regularization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Zagoruyko, S.; and Komodakis, N. 2016. Wide Residual Networks. In *Proceedings of the British Machine Vision Conference*.

Zagoruyko, S.; and Komodakis, N. 2017. Paying more attention to attention: improving the performance of convolutional neural networks via attention transfer. In *International Conference on Learning Representations*.