Exploring Inter-Channel Correlation for Diversity-preserved Knowledge Distillation [2]

Li Liu[1], Qingle Huang[1], Sihao Lin[2], Hongwei Xie[1], Bing Wang[1], Xiaojun Chang[3*], Xiaodan Liang[4]

[1]Alibaba Group, [2]Monash University, [3]RMIT University, [4]Sun Yat-sen University

Motivation and Contribution

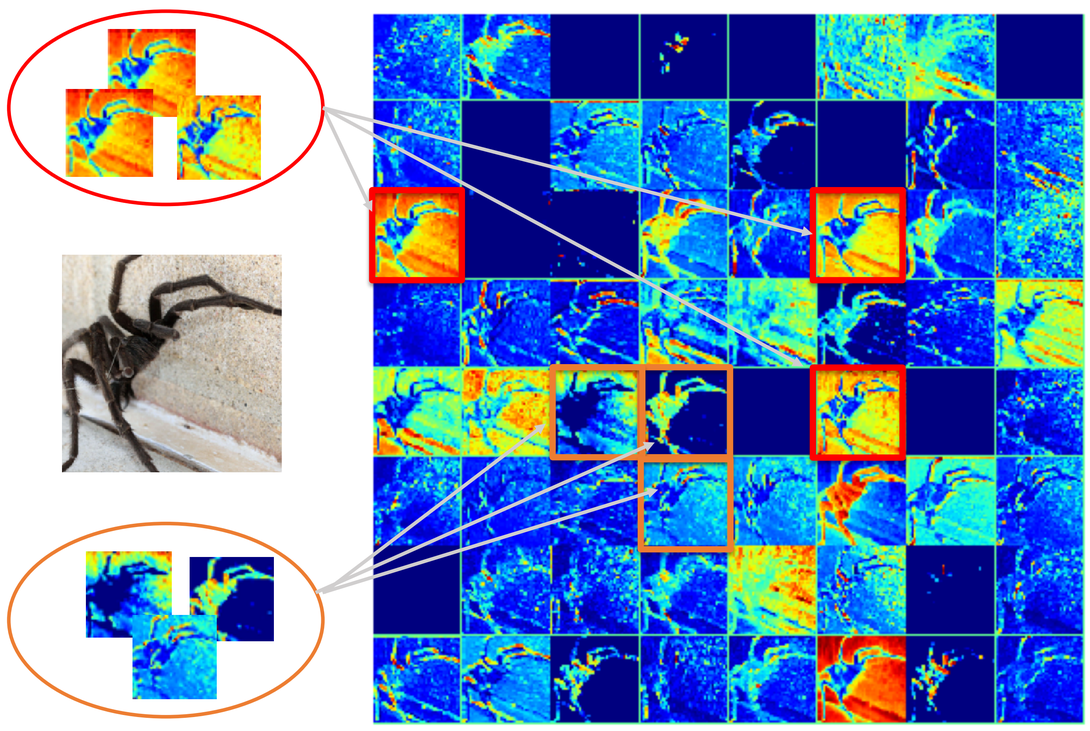


Illustration of inter-channel correlation. The channels orderly extracted from the second layer of ResNet18 have been visualized. The channels denoted by red boxes are homologous both perceptually and mathematically (*e.g.*, inner-product), while the channels denoted by orange boxes are diverse. We show the inter-channel correlation can effectively measure that each channel is homologous or diverse to others, which further reflects the *richness* of the feature spaces. Based on this insightful finding, our ICKD can enforce the student to *mimic* this property from the teacher.

We make the following contributions in this work.

- We introduce the inter-channel correlation, with the characteristic of being invariant to the spatial dimension, to explore and measure both the feature diversity and homology to help the student for better representation learning.
- We further introduce the grid-level inter-channel correlation to make our framework capable of dense prediction task, like semantic segmentation.
- To validate the effectiveness of the proposed framework, extensive experiments have been conducted on different (a) network architectures, (b) downstream tasks and (c) datasets. Our method consistently outperforms the state-of-the-arts methods by a large margin across a wide range of knowledge transfer tasks.

Approach

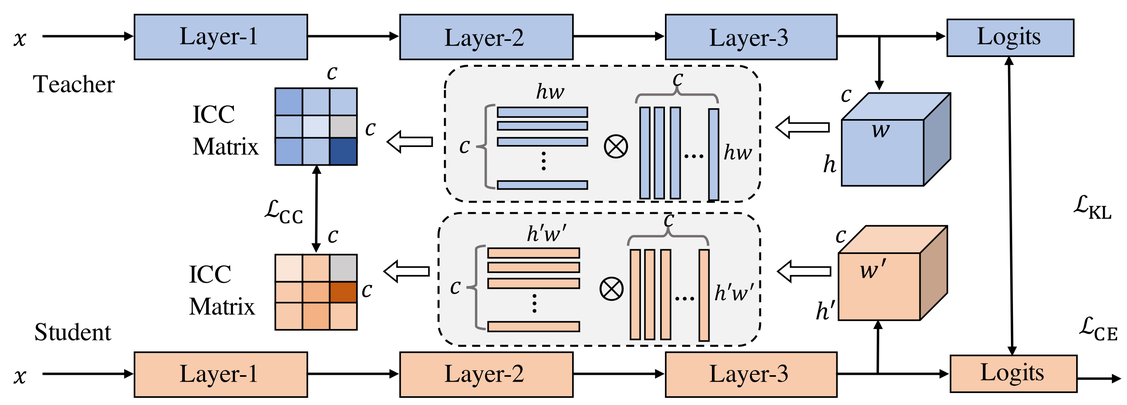
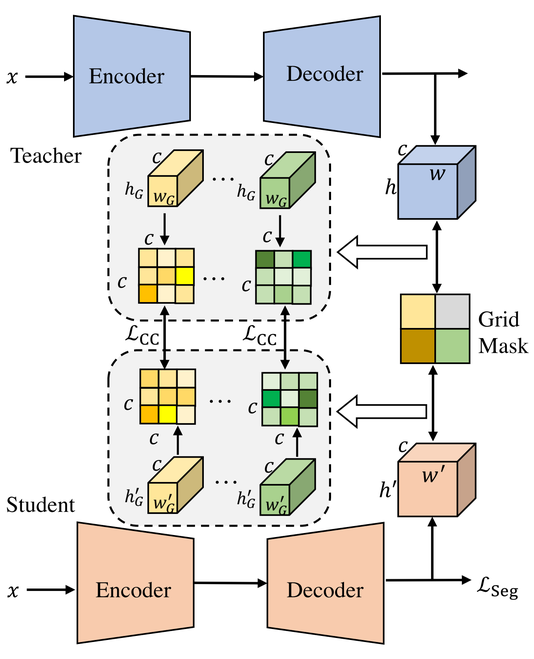


Illustration of the proposed ICKD. We measure the inter-channel correlation of the teacher feature and ask the student to share with the same property. The cubes represent the 3D feature tensors extracted from the teacher and student. They are flattened to the corresponding 2D matrices which are used to compute the ICC matrices. We minimize the MSE between the ICC matrices associated with the feature tensors. The student is also asked to minimize the KL-divergence between the logits of the teacher and student. Finally the cross-entropy loss is applied on the student.

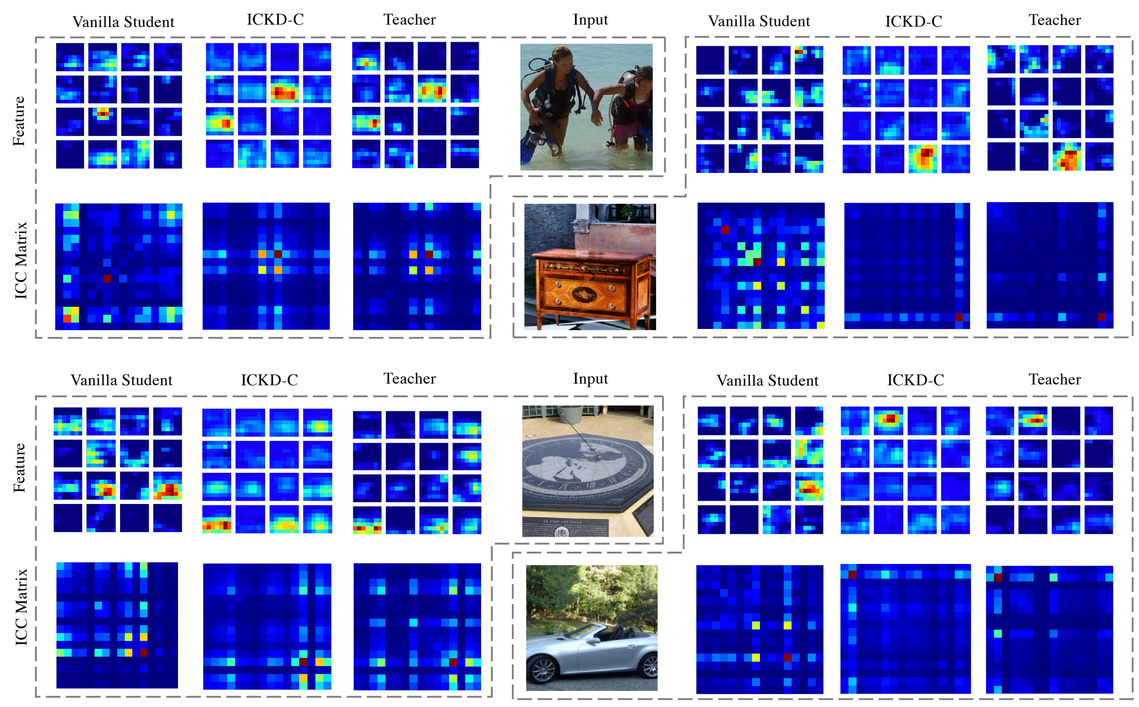
Grid-level Inter-Channel Correlation. When coming into semantic segmentation, the final feature map can be very large, *e.g.* 256×128×128. The correlation of two channels is generated by the inner-product of two vectors of length 16,384. Generally, these two vectors can be seen as sampled from an independent distribution, the correlation value will be of a very small order of magnitude, that means the correlation result is vulnerable to noise. In this situation, the inter-channel correlation of the student model in training process may be unstable. Motivated by the divide-and-conquer, we seek to split the feature map and then perform knowledge distillation individually. Based on this idea, we introduce the grid-level inter-channel correlation. We evenly divide the original feature into $n \times m$ parts and compute their ICC matrices individually. We then minimize the MSE on each paired ICC matrices. As depicted in the left figure, we use a Grid Mask to evenly divide the whole feature into different groups. Despite the change of spatial dimensions, the size of the resulting ICC matrix always depend on the numbers of channels



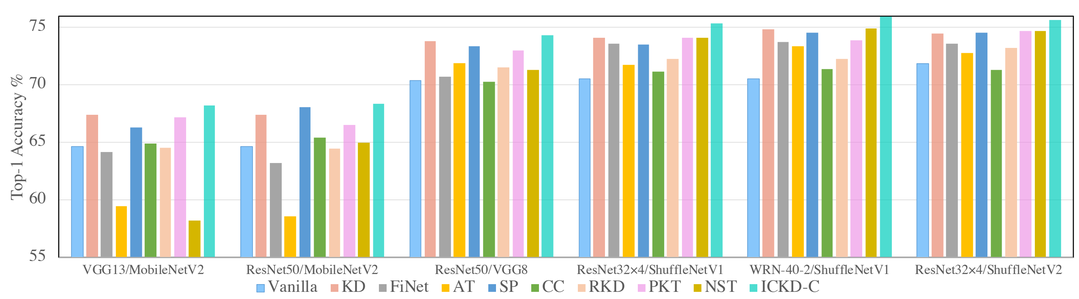
Experimental Results

Method	Network Architecture						
	WRN-40-2 WRN-16-2	WRN-40-2 WRN-40-1	ResNet56 ResNet20	ResNet110 ResNet20	ResNet110 ResNet32	ResNet32×4 ResNet8×4	VGG13 VGG8
Teacher	75.61	75.61	72.34	74.31	74.31	79.42	74.64
Vanilla	73.26	71.98	69.06	69.06	71.14	72.50	70.36
KD [12]	74.92	73.54	70.66	70.67	73.08	73.33	72.98
FitNet [22]	73.58 −1.34	72.24 −1.30	69.21 −1.45	68.99 −1.68	71.06 −2.02	73.50 +0.17	71.02 −1.96
AT [34]	74.08 −0.84	72.77 −0.77	70.55 −0.11	70.22 −0.45	72.31 −0.77	73.44 +0.11	71.43 −1.55
SP [28]	73.83 −1.09	72.43 −1.11	69.67 −0.99	70.04 −0.63	72.69 −0.39	72.94 −0.39	72.68 −0.20
CC [21]	73.56 −1.36	72.21 −1.33	69.63 −1.03	69.48 −1.19	71.48 −1.6	72.97 −0.36	70.71 −2.27
RKD [19]	73.35 −1.57	72.22 −1.32	69.61 −1.05	69.25 −1.42	71.82 −1.26	71.90 −1.43	71.48 −1.5
PKT [20]	74.54 −0.38	73.45 −0.09	70.34 −0.32	70.25 −0.42	72.61 −0.47	73.64 +0.31	72.88 −0.10
FSP [32]	72.91 −2.01	NA	69.95 −0.71	70.11 −0.56	71.89 −1.19	72.62 −0.71	70.20 −2.78
NST [13]	73.68 −1.24	72.24 −1.3	69.60 −1.06	69.53 −1.14	71.96 −1.12	73.30 −0.03	71.53 −1.45
ICKD-C (w/o $\mathcal{L}_{\mathrm{KL}}$)	**75.64** +0.72	74.33 +0.79	**71.76** +1.1	71.68 +1.01	73.89 +0.81	75.25 +1.92	73.42 +0.44
ICKD-C (Ours)	75.57 +0.65	**74.63** +1.09	71.69 +1.03	**71.91** +1.24	**74.11** +1.03	**75.48** +2.15	**73.88** +0.9

Top-1 accuracy (%) in Cifar-100 testing set. Methods are divided into two groups. The performance of each method against traditional KD [1] is reported. For fair comparison, we also report the performance of our method without $\mathcal{L}_{\mathrm{KL}}$. We find that our ICKD-C outperforms all the other methods.



Visualization of the features and the ICC matrices. We have visualized the feature maps and the corresponding ICC matrices of the vanilla student, our model (ICKD-C) and the teacher, respectively. The four input images are sampled from ImageNet testing set. The teacher architecture is ResNet34 and the student architecture is ResNet18. Without loss of generality, we orderly select 16 feature maps extracted from the 4-th block (*i.e.*, the distillation layer) of the network. The results show that our model possesses the similar feature diversity and pattern with the teacher, demonstrating that learning inter-channel correlation can effectively preserve feature diversity.



Knowledge distillation across different architectures on Cifar-100. Using teacher networks that completely different from that of students for knowledge distillation. Our method can enable the students to learn more general knowledge regardless of the specific architecture.

Conclusions

This work presents a method for knowledge distillation that explores the inter-channel correlation to mimic the feature diversity of the teacher network. In addition to image classification, we introduce the grid-level inter-channel correlation for semantic segmentation that most prior works do not pay attention to. We empirically demonstrate the effectiveness of the proposed method on a variety of network architectures and achieve the state-of-the-art in two vision tasks (image classification and semantic segmentation). Besides, the computation of the proposed ICC matrix is invariant to feature spatial dimensions and able to distill generic knowledge across different network architectures.

References

[1] Geoffrey E. Hinton, Oriol Vinyals, and J. Dean. “Distilling the Knowledge in a Neural Network”. In: *ArXiv* abs/1503.02531 (2015).

[2] Li Liu et al. “Exploring Inter-Channel Correlation For Diversity-Preserved Knowledge Distillation”. In: *ICCV*. 2021.

Acknowledgements

This work was supported by the funding of “Leading Innovation Team of the Zhejiang Province" (2018R01017) and Australian Research Council (ARC) Discovery Early Career Researcher Award (DECRA) under DE190100626.