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

Li Liu¹, Qingle Huang¹, Sihao Lin², Hongwei Xie³, Bing Wang⁴, Xiaojun Chang⁵, Xiaodan Liang⁴
¹Alibaba Group, ²Monash University, ³RMIT University, ⁴Sun 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 tasks, 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-art methods by a large margin across a wide range of knowledge transfer tasks.

Experimental Results

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 differ 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.

Acknowledgements

This work was supported by the funding of “Leading Innovation Team of the Zhejiang Province” (2018R010117) and Australian Research Council (ARC) Discovery Early Career Researcher Award (DECRA) under DE19010626.