Knowledge distillation on heterogeneous models

M. Gorpinich[1] O. Bakhteev[1,2] V. Strijov[1,2]

[1] Moscow Institute of Physics and Technology
[2] Dorodnicyn Computing Center RAS

Abstract. This paper investigates the deep learning knowledge distillation problem. Knowledge distillation is a model parameter optimization problem that allows transferring information contained in the model with high complexity, called teacher, to the simpler one, called student. In this paper we propose a cross-layer distillation method that can be applied to significantly heterogeneous models. The variational inference is applied to derive the loss function for metaparameter optimization. Metaparameters are the coefficients of the losses between each pair of layers. The proposed approach is evaluated in the computational experiment on the CIFAR-10 dataset.

Keywords: Machine learning · Knowledge distillation · Heterogeneous models.

1 Introduction

The paper investigates knowledge distillation problem on deep neural networks. Knowledge distillation or knowledge transfer is a technique that allows to transfer knowledge from a teacher model to a student model that is simpler comparing to the teacher model.

Many knowledge distillation methods require similarity of the architectures of the teacher and student model. An approach proposed by Hinton et al. [3] matches logits of the last softmax layer. There are approaches that also match intermediate layers of deep neural networks. In [1] the information-theoretic approach is used. In [5] the authors model information flows in the teacher network and teach student network to mimic them. In [6] intermediate-layer hints are used to guide the student network training process. In [8] the same problem is solved using attention transfer. In [7] the new type of loss function is used that preserves the similarity between different activation functions. In [4] the problem of heterogeneous models is solved using intermediate models called teacher assistant models. In [2] the feature map distillation with attention is used.

In this paper we propose an approach that allows to achieve high performance of a model trained with knowledge distillation even when the architectures of the teacher and student model are significantly heterogeneous.

Contributions of this paper are as follows:

–

2 Problem statement

The knowledge distillation problem is under consideration. In this paper we apply it to solve classification problem but it can be applied to other machine learning problems.

Given a dataset for K-classification problem:

$$\mathfrak{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{m},\ \mathbf{x}_i \in \mathbb{R}^n,\ y_i \in \mathbb{Y} = \{1, \ldots, K\},$$

Given a teacher model $\mathbf{f}$, which was trained on the dataset $\mathfrak{D}_{\text{train}}$. Optimize a student model $\mathbf{g}$ to transfer information obtained from the teacher.

Lets consider two significantly heterogeneous networks, T is a number of layers in a teacher network, S is a number of layers in a student network. Consider $T \cdot S$ pairs of layers $\{(t_i, s_j)\}_{i,j=1}^{T,S}$.

Solve the bi-level optimization problem:

$$\hat{\boldsymbol{\lambda}} = \arg \min_{\boldsymbol{\lambda} \in \mathbb{R}^{T \circ S}} \mathcal{L}_{\text{task}}(\hat{\mathbf{w}}, \boldsymbol{\lambda}) + \sum_{i,j=1}^{T,S} \lambda_{i,j} I(h_i^t, h_j^s),$$

$$\hat{\mathbf{w}} = \arg \min_{\mathbf{w} \in \mathbb{R}^m} \mathcal{L}_{\text{task}}(\hat{\mathbf{w}}, \boldsymbol{\lambda}),$$

where $\mathcal{L}_{\text{task}}$ is a cross-entropy loss for classification task, h_i^t and h_j^s are activations of the i-th layer of the teacher network and the j-th layer of the student network, $I(h_i^t, h_j^s)$ is the mutual information, $\lambda_{i,j}$ is a hyperparameter, $\sum_{i,j=1}^{T,S} \lambda_{i,j} = 1$.

3 Experiments

4 Conclusion

References

1. Ahn, S., Hu, S.X., Damianou, A.C., Lawrence, N.D., Dai, Z.: Variational information distillation for knowledge transfer. In: CVPR. pp. 9163–9171. Computer Vision Foundation / IEEE (2019), http://openaccess.thecvf.com/CVPR2019.py
2. Chen, D., Mei, J.P., Zhang, Y., 0001, C.W., Wang, Z., Feng, Y., 0001, C.C.: Cross-layer distillation with semantic calibration. CoRR **abs/2012.03236** (2020)
3. Hinton, G.E., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. CoRR **abs/1503.02531** (2015), http://arxiv.org/abs/1503.02531
4. Mirzadeh, S.I., Farajtabar, M., Li, A., 0001, H.G.: Improved knowledge distillation via teacher assistant: Bridging the gap between student and teacher. CoRR **abs/1902.03393** (2019), http://arxiv.org/abs/1902.03393
5. Passalis, N., Tzelepi, M., Tefas, A.: Heterogeneous knowledge distillation using information flow modeling. In: CVPR. pp. 2336–2345. IEEE (2020)
6. Romero, A., Ballas, N., Kahou, S.E., Chassang, A., Gatta, C., Bengio, Y.: Fitnets: Hints for thin deep nets. In: Bengio, Y., LeCun, Y. (eds.) 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings (2015), http://arxiv.org/abs/1412.6550

7. Tung, F., Mori, G.: Similarity-preserving knowledge distillation. In: ICCV. pp. 1365–1374. IEEE (2019)
8. Zagoruyko, S., Komodakis, N.: Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. CoRR **abs/1612.03928** (2016), http://arxiv.org/abs/1612.03928