Variational Mutual Information Distillation for Transfer Learning

Sungsoo Ahn *
Korea Advanced Institute of Science and Technology
Daejeon, Korea
sungsoo.ahn@kaist.ac.kr

Shell Xu Hu *
École des Ponts ParisTech
Champs-sur-Marne, France
hus@imagine.enpc.fr

Andreas Damianou
Amazon
Cambridge, United Kingdom
damianou@amazon.com

Neil D. Lawrence
Amazon
Cambridge, United Kingdom
lawrennd@amazon.com

Zhenwen Dai
Amazon
Cambridge, United Kingdom
zhenwend@amazon.com

Abstract

We consider the teacher-student framework for knowledge transfer, where the goal is to improve learning of a "student" neural network, given a "teacher" neural network pretrained on the same or a similar task. The majority of existing approaches for distilling knowledge from a teacher network to a student network rely on matching either activations or handcrafted features from the teacher network. Instead, in this paper we establish an information-theoretic framework for knowledge distillation which encourages high mutual information between two networks. Our framework can be applied to knowledge transfer between different tasks without any assumptions on the architectures of the teacher and the student network. We empirically validate our proposed framework by demonstrating its improvement over existing methods on various knowledge transfer tasks.

1 Introduction

Transfer learning for neural networks facilitates learning on a target task by leveraging knowledge gained from training on a source task. Hinton et al. [2015] introduce the teacher-student framework as a special case of transfer learning, where a "student" neural network learns a target task while efficiently using the knowledge present in a "teacher" neural network that was pretrained on a source task. Knowledge distillation methods have been proposed to solve this problem by matching either the activations [Romero et al., 2014, Hinton et al., 2015] or the handcrafted features [Zagoruyko and Komodakis, 2016a] of a specific layer of a student network to the ones of a teacher network. In this paper, we develop a new knowledge distillation method based on information content matching that maximizes the mutual information between student and teacher. Our argument is quite intuitive: for transfer learning to be effective, one has to maximize the amount of relevant knowledge being transferred from a teacher network to a student network; the mutual information between two networks provides quantification of such knowledge in a principled way. Despite its attractive properties, the mutual information is intractable to compute in general. Therefore, we replace the mutual information as a tractable analytic variational lower bound. Interestingly, we observe that maximization of such bound is equivalent to solving a density estimation task for activations in the teacher network. Our implementation based on a Gaussian observation model empirically outperforms state-of-the-art methods on various transfer learning tasks.

*Contributed during an internship at Amazon.

32nd Conference on Neural Information Processing Systems (NIPS 2018), Montréal, Canada.

2 Method

Consider training a student neural network on a particular target task, given another teacher neural network trained on a similar (or related) source task. From the perspective of information theory, knowledge distillation can be expressed as retaining high mutual information between the layers of the teacher and student networks while training the student network. More specifically, consider an input random variable $\boldsymbol{x}$ drawn from the target data distribution $p(\boldsymbol{x})$ and K pairs of representative layers $\mathcal{R} = \{(\mathcal{T}^{(k)}, \mathcal{S}^{(k)})\}_{k=1}^{K}$, where each pair $(\mathcal{T}^{(k)}, \mathcal{S}^{(k)})$ is selected from the teacher and student network respectively. Feedforwarding the input $\boldsymbol{x}$ through the neural networks induces K pairs of random variables $\{(\boldsymbol{t}^{(k)}, \boldsymbol{s}^{(k)})\}_{k=1}^{K}$ which indicate outputs from the corresponding layers as a function of input $\boldsymbol{x}$, e.g., $\boldsymbol{t}^{(k)} = \mathcal{T}^{(k)}(\boldsymbol{x})$. The mutual information between the pair of random variables $(\boldsymbol{t}, \boldsymbol{s})$ is defined by:

$$I(\boldsymbol{t}, \boldsymbol{s}) = H(\boldsymbol{t}) - H(\boldsymbol{t}|\boldsymbol{s}) = -\mathbb{E}_{\boldsymbol{t}\sim p(\boldsymbol{t})}[\log p(\boldsymbol{t})] + \mathbb{E}_{\boldsymbol{t},\boldsymbol{s}\sim p(\boldsymbol{t},\boldsymbol{s})}[\log p(\boldsymbol{t}|\boldsymbol{s})],$$

where the entropy $H(\boldsymbol{t})$ and the conditional entropy $H(\boldsymbol{t}|\boldsymbol{s})$ were derived from the joint distribution $p(\boldsymbol{t}, \boldsymbol{s})$. Note that the joint distribution $p(\boldsymbol{t}, \boldsymbol{s})$ is a result of aggregagation over the representative layers with input $\boldsymbol{x}$ sampled from the empirical distribution $p(\boldsymbol{x})$. We now define the following loss function which simultaneously trains the student network for the target task while encouraging high mutual information with the teacher network:

$$\mathcal{L} = \mathcal{L}_{\mathcal{S}} - \sum_{k=1}^{K} \lambda_k I(\boldsymbol{t}^{(k)}, \boldsymbol{s}^{(k)}), \tag{1}$$

where $\mathcal{L}_{\mathcal{S}}$ is the loss function for the target task and $\{\lambda_k\}_{k=1}^{K}$ is set of hyper-parameters introduced for regularizing the mutual information terms. Equation (1) needs to be minimized with respect to the student network's parameters. However, the minimization is hard since computation of the exact mutual information terms is intractable. We instead propose a variational lower bound for each mutual information term $I(\boldsymbol{t}, \boldsymbol{s})$, in which we define a variational distribution $q(\boldsymbol{t}|\boldsymbol{s})$ that approximates $p(\boldsymbol{t}|\boldsymbol{s})$:

$$\begin{aligned}
I(\boldsymbol{t}, \boldsymbol{s}) &= H(\boldsymbol{t}) - H(\boldsymbol{t}|\boldsymbol{s}) \\
&= H(\boldsymbol{t}) + \mathbb{E}_{\boldsymbol{t},\boldsymbol{s}\sim p(\boldsymbol{t},\boldsymbol{s})}[\log p(\boldsymbol{t}|\boldsymbol{s})] \\
&= H(\boldsymbol{t}) + \mathbb{E}_{\boldsymbol{s}\sim p(\boldsymbol{s})}[D_{\mathrm{KL}}(p(\boldsymbol{t}|\boldsymbol{s})||q(\boldsymbol{t}|\boldsymbol{s}))] + \mathbb{E}_{\boldsymbol{t},\boldsymbol{s}\sim p(\boldsymbol{t},\boldsymbol{s})}[\log q(\boldsymbol{t}|\boldsymbol{s})] \\
&\geq H(\boldsymbol{t}) + \mathbb{E}_{\boldsymbol{t},\boldsymbol{s}\sim p(\boldsymbol{t},\boldsymbol{s})}[\log q(\boldsymbol{t}|\boldsymbol{s})],
\end{aligned}$$

where the last inequality comes from the non-negativity of the Kullback-Leiber divergence $D_{\mathrm{KL}}(\cdot)$. Applying the above lower bound to (1), yields an upper bound of the original loss function:

$$\widetilde{\mathcal{L}} = \mathcal{L}_{\mathcal{S}} - \sum_{k=1}^{K} \lambda_k \mathbb{E}_{\boldsymbol{t}^{(k)},\boldsymbol{s}^{(k)}\sim p(\boldsymbol{t}^{(k)},\boldsymbol{s}^{(k)})}[\log q(\boldsymbol{t}^{(k)}|\boldsymbol{s}^{(k)})]. \tag{2}$$

This objective is jointly minimized over the parameters of the student network and the auxiliary distribution q. Note that the entropy term $H(\boldsymbol{t})$ has been removed from the equation since it is constant with respect to the parameters to be optimized. Such a technique of maximizing the lower bound of mutual information is known as variational information maximization [Agakov and Felix, 2004].

Implementation. We describe a specific instance of our framework by choosing a form made for the variational distribution $q(\boldsymbol{t}|\boldsymbol{s})$, which depends on the type of layers used for $\boldsymbol{t}$ and $\boldsymbol{s}$. In general, we use a Gaussian distribution with heteroscedastic mean $\boldsymbol{\mu}(\cdot)$ and homoscedastic variance $\boldsymbol{\sigma}$ as the auxiliary distribution $q(\boldsymbol{t}|\boldsymbol{s})$, i.e., the mean $\boldsymbol{\mu}(\cdot)$ is a function of $\boldsymbol{s}$ and the standard deviation $\boldsymbol{\sigma}$ is not. When the corresponding representative layer of the teacher network is one-dimensional, i.e., $\boldsymbol{t} = \mathcal{T}(\boldsymbol{x}) \in \mathbb{R}^D$, the variational distribution is expressed as follows:

$$-\log q(\boldsymbol{t}|\boldsymbol{s}) = \sum_{d=1}^{D} \log \sigma_d + \frac{(t_d - \mu_d(\boldsymbol{s}))^2}{2\sigma_d^2} + \text{constant}, \tag{3}$$

where t_d indicates the d-th entry of the vector $\boldsymbol{t}$ and μ_d represents the output of a single unit of $\boldsymbol{\mu}$. Furthermore, representative layer of the teacher network can also contain spatial dimensions

N	Full (5000)	1000	500	100
Teacher	94.46	-	-	-
Student	90.71	85.21	79.82	59.59
KD	91.34	85.39	81.88	65.94
FitNet	90.97	86.45	83.71	76.07
AT	91.51	87.28	84.46	75.18
MI	**91.76**	**88.71**	**86.49**	**78.14**
KD + AT	**91.81**	87.34	85.01	76.29
KD + MI	91.7	**88.59**	**86.53**	**78.48**

(a) CIFAR-10 with varying size of the dataset

N_ℓ	1000	500	100
KD	**91.33**	**91.30**	**91.29**
FitNet	85.25	82.03	76.85
AT	89.30	88.72	86.34
MI	90.33	90.23	89.03
KD + AT	91.56	**91.81**	**91.34**
KD + MI	**91.57**	91.44	**91.34**

(b) CIFAR-10 with varying number of the labels

Table 1: Experimental results (accuracy) of knowledge distillation (model compression) on CIFAR-10 and CIFAR-100 dataset from WRN-40-2 (teacher network) to WRN-16-1 (student network) with varying (a) number of data points per class (denoted by N) and (b) number of labels per class (denoted by N_ℓ) provided for training of student network.

corresponding to channel, height and length, i.e., $\boldsymbol{t} \in \mathbb{R}^{C\times H\times L}$. For this case, our choice of variational distribution is expressed as follows:

$$-\log q(\boldsymbol{t}|\boldsymbol{s}) = \sum_{c=1}^{C}\sum_{h=1}^{H}\sum_{\ell=1}^{L} \log \sigma_c + \frac{(t_{c,h,\ell} - \mu_{c,h,\ell}(\boldsymbol{s}))^2}{2\sigma_c^2} + \text{constant}, \tag{4}$$

where $t_{c,h,\ell}$ denote scalar components of $\boldsymbol{t}$ indexed by (c, h, ℓ). Further, μ_c represents the output of a single unit from the neural network $\boldsymbol{\mu}$ consisting of convolutional layers.

3 Experiments

In this section, we demonstrate the effectiveness of the proposed method in knowledge transfer for various tasks. We consider using various deep convolutional neural networks such as residual networks [He et al., 2016], wide residual networks [Zagoruyko and Komodakis, 2016b] and VGG networks [Simonyan and Zisserman, 2014]. We also consider various vision datasets: CIFAR-10 [Krizhevsky, 2009], ImageNet [Russakovsky et al., 2015], CUB-200-2011 [Welinder et al., 2010] and MIT-67 [Quattoni and Torralba, 2009]. Throughout the experiments, 20% of the dataset provided for training the student network was used for validation, i.e., choosing the best set of hyper-parameters for each method. Every result was computed as the mean of accuracies over 3 runs.

3.1 Knowledge transfer between same datasets

We first consider knowledge transfer task where the student network is being trained on the same dataset that the teacher network has already been trained on. A wide residual network with 40 layers of depth (WRN-40-2) was pretrained on the same dataset as the teacher network. The goal is to transfer knowledge from the teacher network for the training of a smaller network with 16 layers of depth (WRN-16-1). Here, we implement our mutual information loss (MI) by choosing ends of each residual blocks in the teacher and student network as the representative layers. The mean function $\mu(\cdot)$ for auxiliary distribution in equation (4) was parameterized by two layers of 1×1 convolutional layers with batch normalization and rectified linear unit. For comparison, we consider three candidates for loss function: traditional knowledge distillation loss (KD) proposed by Hinton et al. [2015], attention transfer loss (AT) proposed by Zagoruyko and Komodakis [2016a] and hint based loss (FitNet) [Romero et al., 2014]. Additionally, we provide the baseline results from training on the provided dataset without any transfer learning applied to the teacher network (Teacher) and student network (Student). In the first set of experiments (Table 1a), we train the student network on the subset of the given dataset with a varying number of data points. For the next set of experiments (Table 1b), labels were provided only for the subset of the original dataset, which results in a mixture of labeled and unlabeled dataset. In Table 1a, we observe that the proposed MI loss outperform baselines accross most regimes. Especially, one can observe that improvement from MI compared to other methods gets larger with a smaller size of the number of data points. In Table 1b, we observe that the KD loss outperforms other algorithms. This is as expected since KD can be seen as providing"soft" labels to the student network which have been removed in the unlabeled training data. Still, we observe that the MI loss outperforms other methods for transferring knowledge from intermediate layers of the teacher network, i.e., FitNet and AT.

N	FULL (<127)	50	25	10
Student	48.46	37.49	25.02	16.32
Finetuned	71.89	66.72	57.74	47.36
LwF	64.05	57.66	43.43	27.84
FitNet	70.15	65.40	56.37	40.82
AT	58.56	49.50	40.65	24.80
MI-L	68.11	60.27	48.68	30.45
MI-I	**72.09**	**68.13**	**60.25**	**49.83**
LwF + FitNet	70.75	65.87	56.02	40.07
MI-L + MI-I	**71.27**	**67.59**	**61.69**	**50.07**

(a) MIT-67, ResNet-34 to ResNet-18

N	FULL (<127)	50	25	10
Student	54.15	43.31	28.11	15.45
Finetuned	67.39	61.77	52.11	39.33
LwF	64.50	60.12	50.55	34.20
FitNet	**71.00**	**65.10**	54.83	41.27
AT	59.10	52.11	40.75	25.45
MI-L	68.78	62.89	52.69	37.69
MI-I	69.38	64.98	**57.41**	**43.51**
LwF + FitNet	70.67	65.10	54.95	40.92
MI-L + MI-I	**71.04**	**66.89**	**58.21**	**48.26**

(b) MIT-67, ResNet-34 to VGG-9

N	FULL (<35)	20	10	5
Student	43.15	25.59	10.31	5.11
Finetuned	74.76	69.10	53.29	33.15
LwF	62.64	53.46	38.49	21.84
FitNet	67.85	62.58	51.15	36.51
AT	59.94	47.75	28.20	17.39
MI-L	66.11	59.80	44.20	27.13
MI-I	**71.03**	**66.08**	**53.86**	**41.44**
LwF + FitNet	68.53	62.63	51.08	34.28
MI-L + MI-I	**70.97**	**64.73**	**53.87**	**40.58**

(c) CUB-200-2011, ResNet-34 to ResNet-18

N	FULL (<35)	20	10	5
Student	49.33	33.28	14.44	7.41
Finetuned	66.94	63.57	50.79	29.56
LwF	66.12	55.72	41.63	24.89
FitNet	69.06	62.83	45.81	29.96
AT	59.31	47.66	29.13	15.07
MI-L	68.08	**63.13**	**47.83**	28.73
MI-I	**69.84**	62.63	45.91	**31.01**
LwF + FitNet	**70.56**	62.44	47.36	30.52
MI-L + MI-I	70.00	**65.14**	**53.78**	**38.76**

(d) CUB-200-2011, ResNet-34 to VGG-9

Table 2: Experimental results (accuracy) of transfer from ResNet-34 (teacher network) to ResNet-18/VGG-9 (student network) for the MIT-67/CUB-200-2011 dataset with varying number of data points per class (denoted by N).

3.2 Knowledge transfer between different datasets

Next, we consider transfer learning between heterogeneous tasks, where the goal is to enhance the training of student network for classifying a dataset that has never been provided for training of the teacher network. The teacher network is trained on the ImageNet dataset and is then used for transferring knowledge to a student network which is trained on the MIT-67 or the CUB-200-2011 dataset. We consider the residual network (ResNet-34) as the teacher network and either one of the residual network (ResNet-18) or VGG network (VGG-9) as the student network. This time, we conduct each set of experiments by varying the number of data points in the dataset provided to the student network. We consider two candidates for the MI loss. The first type of loss is based on mutual information between logit layer of the teacher network and the penultimate layer of the student network (MI-L). The other kind of loss considered corresponds to choosing representative layers as the intermediate layers with spatial dimensions (MI-I). Specifically, ends of residual blocks and max pooling layers were chosen from the residual and VGG networks respectively. The auxiliary distributions for MI-L and MI-I are parameterized as in equation (3) and (4) with mean function $\boldsymbol{\mu}(\cdot)$ as linear transformation and two layers of 1×1 convolutional layers with batch normalization and rectified linear units respectively. We compare with the same choice of knowledge transfer loss as considered in Section 3.1 except for the KD loss, which was replaced by its extension to transfer learning between heterogeneous tasks, i.e., the learning without forgetting loss (LwF) [Li and Hoiem, 2017]. For baselines, we additionally provide the results from training the student network without transfer learning (Student). We also provide the results of finetuning the student network with initialization provided from training on the ImageNet dataset (Finetuned). In Table 2, our method outperforms others in most regions of comparison. Especially, we observe that our algorithm shows similar performance even with the finetuning method, which requires the student network to be pretrained on the source task.

4 Conclusion

We proposed a new framework for maximizing the mutual information between two neural networks for efficient knowledge transfer. We also presented an accompanying tractable variational formulation equipped with a recognition model for efficient optimization. The implementation of our algorithm is based on Gaussian observation models and is empirically shown to outperform other benchmarks in the distillation and transfer learning tasks.

References

David Agakov and Barber Felix. The IM algorithm: a variational approach to information maximization. 2004.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009.

Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

Ariadna Quattoni and Antonio Torralba. Recognizing indoor scenes. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 413–420. IEEE, 2009.

Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*, 2014.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona. Caltech-UCSD Birds 200. Technical Report CNS-TR-2010-001, California Institute of Technology, 2010.

Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *arXiv preprint arXiv:1612.03928*, 2016a.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016b.