Supplementary Material:

Variational Information Distillation for Knowledge Transfer

A. Implementation details

A.1. Network architectures

For the WRNs and ResNets used throughout the experiments, we use the same architectures as originally described by Zagoruyko *et al.*, [32] and He *et al.*, [11] respectively. For the VGG-9 network used in transfer learning, *i.e.*, Section 3.2, we make a slight modification from the VGG-11 network [25] without deviating from the VGG design philosophy. It is conducted by first stacking eight 3×3 convolutional layers with $64, 128, 256, 256, 512, 512, 512, 512$ channels in order with batch normalization and rectified linear unit (ReLU) after every convolutional layers. Furthermore, additional max-pooling layers are inserted after the $\{1, 2, 4, 6, 8\}$-th ReLUs. Then the final max-pooling layer is followed by global average pooling layer and a linear layer leading up to the prediction of the labels. For the MLP used in knowledge transfer from CNN to MLP, *i.e.*, Section 3.3, we sequentially stack one linear layer, three bottleneck linear layers and one linear layer leading to the prediction of labels, where dropout with drop rate of 0.2, batch normalization and ReLU was inserted between each layers. Here, the bottleneck layer indicates a composition of two linear layers without non-linearity that is introduced to speed up learning by reducing the number of parameters. All of the hidden layers have the same h number of units and the bottleneck linear layer is composed of two linear layers with a size of $h \times \frac{h}{4}$ and $\frac{h}{4} \times h$.

A.2. Parameterization of VID

In the knowledge distillation experiments, *i.e.*, Section 3.1, we parameterize the mean function $\boldsymbol{\mu}(\cdot)$ in equation (5) for VID-I by three 1×1 convolutional layers with batch normalization and ReLU between each layers. The hidden channel sizes were chosen to be twice of the output channel size. For the transfer learning experiments, *i.e.*, Section 3.2, we first parameterize the mean function $\boldsymbol{\mu}(\cdot)$ in equation (5) for VID-I by two 1×1 convolutional layers with batch normalization and ReLU between the layers. For this case, the hidden channel sizes were chosen to be half of the output channel size. Furthermore, VID-LP was parameterized as in equation (6) with mean function $\boldsymbol{\mu}(\cdot)$ being a single linear layer, *i.e.*, a linear transformation. Finally, we consider the knowledge transfer from CNN to MLP, *i.e.*, Section 3.3, based on VID-I with equation (5). For this case, the mean function maps the one-dimensional input $\boldsymbol{s}$ from intermediate layer of the student network (MLP) into a three-dimensional output $\boldsymbol{t}$ corresponding to intermediate layer of the teacher network (CNN), *i.e.*, $\boldsymbol{\mu} : \mathbb{R}^N \to \mathbb{R}^{C \times H \times W}$. To this end, the input is first treated as a three-dimensional tensor with with unit spatial dimensions, *i.e.*, $\boldsymbol{s} \in \mathbb{R}^{N \times 1 \times 1}$. Then the input goes through a single transposed convolutional layer with a 4×4 kernel, unit stride and zero padding followed by multiple transposed convolutional layers with a 4×4 kernel, two strides and unit padding. The number of transposed convolutional layers were varied for corresponding layer of the teacher network, in order to match the spatial dimension.

A.3. Loss function and training scheme

In the experiments, the loss function for VID takes the following form:

$$\widehat{\mathcal{L}} = \lambda_1 \mathcal{L}_{\mathcal{S}} - \sum_{k=1}^{K} \frac{\lambda_2}{N_k} \mathbb{E}_{\boldsymbol{t}^{(k)}, \boldsymbol{s}^{(k)}} [\log q(\boldsymbol{t}^{(k)} | \boldsymbol{s}^{(k)}))], \tag{8}$$

where $\mathcal{L}_{\mathcal{S}}$ is the task-specific loss function for the target task, $\lambda_1, \lambda_2 > 0$ are hyper-parameters introduced for balancing between the cross-entropy and the regularization terms, and N_k denotes the total number of dimensions for each layer selected from the teacher network for knowledge transfer, i.e., $\boldsymbol{t}^{(k)} \in \mathbb{R}^{N_k}$ or $N_k = C_k H_k W_k$ when $\boldsymbol{t}^{(k)} \in \mathbb{R}^{C_k \times H_k \times W_k}$. For all of the experiments and both VID-I and VID-LP, we select λ_1 and λ_2 from $\{0.1, 1\}$ and $\{10, 100\}$ respectively, based on the performance evaluated on the validation set. For other knowledge transfer methods, we also choose the scaling of the cross-entropy term, *i.e.*, $\lambda_1 > 0$, from $\{0.1, 1\}$. Furthermore, the corresponding regularization terms are scaled by $\{1, 10\}, \{10, 100\}, \{100, 1000\}, \{5, 50\}$ for KD, FitNet, AT and NST respectively, based on the performance on the validation set. Additionally, KD was implemented with temperature scaling parameter set to $T = 4$.

Finally, we describe the training scheme used for the experiments. Due to unstable gradients in some cases, we clipped the norm of gradients by 100 throughout the experiments. Additionally, the homoscedastic variance for the variational distribution in equation (5) and (6) was initialized with value of 5.0. In the knowledge distillation experiments, *i.e.*, Section 3.1, when training on the full dataset, we used stochastic gradient descent (SGD) for 200 epochs with batch size of 128 and weight decay of 0.0005. Initial learning rate of 0.1 is decayed 0.2 times at $\{60, 120, 160\}$-th

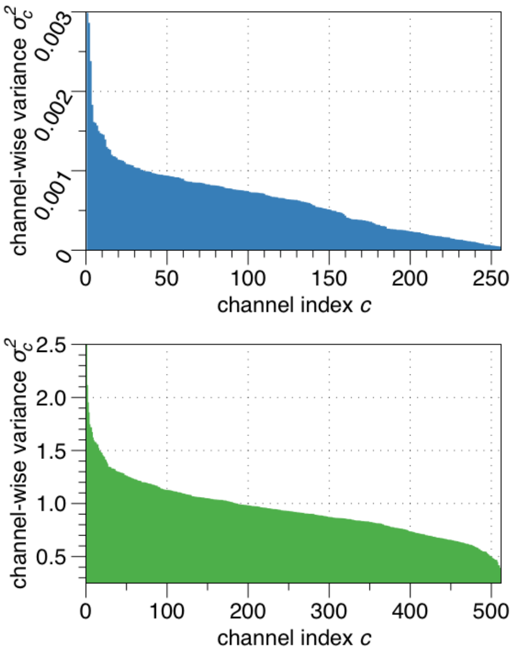


Figure 3: *Channel-wise variance $\sigma_n^2 = \sigma_c^2$ (sorted) learned by VID-I in transfer learning from ResNet34 trained on ImageNet to ResNet18 trained on CUB-200, corresponding to the ends of third (top) and fourth (bottom) residual blocks respectively.*

epoch. When training on subset of the dataset, the numbers are appropriately scaled to have similar number of updates for parameters. In the transfer learning experiments, *i.e.*, Section 3.2, when training on the full dataset for ResNet-34 and ResNet-18, we use SGD for 250 epochs with batch size 128 and weight decay of 0.0005. Initial learning rate of 0.05 is decayed by 0.2 times at $\{150, 200\}$-th epoch. For the case of VGG-9, we use SGD for 250 epochs with batch size 12 without weight decay. Initial learning rate of 0.01 is decayed by 0.2 times at 150 and 200-th epoch. Again, the numbers are appropriately scaled to match the number of updates for parameters when training on subset of the full dataset. In the knowledge transfer from CNN to MLP experiments, *i.e.*, Section 3.3, we used SGD for 700 epochs with batch size of 128 and weight decay of 0.0005. Initial learning rate of 0.001 was decayed by 0.2 times at 500 and 600-th epoch.

B. Additional Experiments

B.1. Visualization of learned parameters

In order to examine the behavior of the learned variance parameters σ_n in VID, we plot its channel-wise value for different layers in Figure 3. Here, one can observe that the learned variance parameters σ_n are diverse, especially accross different layers. Hence, modeling of homoscedastic variance is necessary for obtaining a tighter lower bound of the mutual information in the equation (3).

Student	KD	FitNet	AT	NST	VID-L	VID-I
93.34	93.58	93.43	92.89	94.01	93.88	**94.17**

Table 5: *Experimental results (test accuracy) of transfer learning from grayscale-SVHN to MNIST with* 200 *samples per class for LeNet-based architectures.*

Teacher	Student	KD	FitNet	ANC	VID-I
92.36 / <u>93.43</u>	91.69 / <u>91.42</u>	91.12	<u>91.61</u>	<u>91.92</u>	**92.15**

Table 6: *Experimental results (validation accuracy) in comparison to Adversarial Network Compression (ANC), for knowledge distillation from ResNet-164 to ResNet-20 on CIFAR-10 dataset. Underlined numbers are results reported by Belagiannis* et al. *[3].*

B.2. Transfer learning from SVHN to MNIST

We also provide additional experimental results for transfer learning from SVHN to MNIST in Table 5. To this end, the teacher network is trained on the full SVHN dataset that was converted to grayscale, then the student network is trained on MNIST with 200 data points per class. We employ LeNet-like architectures for both networks. Again, one observes that VID outperforms over other methods.

B.3. Comparison with adversarial network compression

We additionally compare with the recently proposed adversarial network compression [3]. by repeating the knowledge distillation experiment on CIFAR-10 between ResNets presented by [3]. The corresponding results are reported in Table 6. One observes that our methods outperforms the ANC with a small margin.

B.4. Experimental results with standard deviation

In Table 7, 8, 9, 10, 11 and 12, we provide full experimental results corresponding to the Table 1, 2, 3a, 3b, 3d and 4 respectively with additional standard deviations for the three repeated runs.

B.5. Additional heat maps for VID training

In Figure 4, we provide additional visualization results of the knowledge transfer based on VID that was plotted in the same way as in Figure 2.

M	5000	1000	500	100
Teacher	94.36 (± 0.27)	-	-	-
Student	90.82 (± 0.17))	84.64 (± 0.05)	79.64 (± 0.05)	55.03 (± 6.59)
KD	91.66 (± 0.13)	85.52 (± 0.02)	81.48 (± 0.24)	55.03 (± 0.05)
FitNet	90.79 (± 0.31)	84.84 (± 0.35)	80.82 (± 0.19)	68.57 (± 0.84)
AT	91.54 (± 0.10)	87.43 (± 0.35)	84.78 (± 0.27)	73.96 (± 0.96)
NST	91.11 (± 0.12)	86.76 (± 0.37)	82.68 (± 0.13)	64.76 (± 0.45)
VID-I	91.94 (± 0.31)	**89.76** (± 0.07)	**88.33** (± 0.43)	**82.03** (± 1.13)
KD + AT	91.39 (± 0.26)	87.11 (± 0.03)	84.54 (± 0.01)	75.11 (± 0.83)
KD + VID-I	**92.31** (± 0.31)	89.33 (± 0.21)	87.34 (± 0.19)	81.80 (± 0.01)

Table 7: Experimental results (test accuracy) of knowledge distillation on the CIFAR-10 dataset from teacher network (WRN-40-2) to student network (WRN-16-1) with varying number of data points per class (denoted by M).

(d, w)	(40,2)	(16, 2)	(40, 1)	(16, 1)
Teacher	74.16 (± 0.33)	-	-	-
Student	74.34 (± 0.46)	70.42 (± 0.63)	68.79 (± 0.19)	65.46 (± 0.13)
KD	75.54 (± 0.25)	72.94 (± 0.38)	71.34 (± 0.19)	66.97 (± 0.46)
FitNet	74.29 (± 0.17)	70.89 (± 0.61)	68.66 (± 0.27)	65.38 (± 0.05)
AT	74.76 (± 0.36)	71.06 (± 0.07)	69.85 (± 0.51)	65.31 (± 0.51)
NST	74.81 (± 0.19)	71.19 (± 0.54)	68.00 (± 0.20)	64.95 (± 0.33)
VID-I	75.25 (± 0.37)	73.31 (± 0.30)	71.51 (± 0.15)	66.32 (± 0.52)
KD + AT	**75.90** (± 0.40)	73.16 (± 0.15)	71.48 (± 0.15)	66.48 (± 0.67)
KD + VID-I	**75.90** (± 0.26)	**73.50** (± 0.06)	**72.47** (± 0.21)	**66.91** (± 0.06)

Table 8: Experimental results (test accuracy) of knowledge distillation on the CIFAR-100 dataset from the teacher network (WRN-40-2) to the student networks (WRN-d-w) with varying factor of depth d and width w.

M	≈80	50	25	10
Student	48.78 (± 0.72)	37.46 (± 0.88)	25.52 (± 1.37)	14.68 (± 0.41)
Finetuned	71.22 (± 0.85)	65.30 (± 0.83)	58.56 (± 0.55)	48.86 (± 0.87)
LwF	61.34 (± 0.54)	50.07 (± 0.22)	38.76 (± 0.34)	22.09 (± 0.58)
FitNet	70.37 (± 0.97)	61.34 (± 0.94)	54.60 (± 1.31)	36.54 (± 0.34)
AT	57.99 (± 0.39)	48.66 (± 0.67)	42.51 (± 1.09)	25.90 (± 1.29)
NST	56.79 (± 1.20)	46.92 (± 0.80)	34.38 (± 1.19)	20.70 (± 0.22)
VID-LP	67.54 (± 0.42)	59.18 (± 0.76)	47.89 (± 0.75)	31.22 (± 1.12)
VID-I	**72.04** (± 0.62)	66.42 (± 0.45)	60.77 (± 0.91)	**50.60** (± 1.06)
LwF + FitNet	70.32 (± 0.69)	61.19 (± 0.45)	53.83 (± 0.91)	36.67 (± 0.88)
VID-LP + VID-I	71.69 (± 0.37)	**66.87** (± 0.59)	**61.29** (± 0.04)	49.65 (± 0.97)

Table 9: Experimental results (test accuracy) of transfer learning from the teacher network (ResNet-34) to the student network (ResNet-18) for the MIT-67 dataset with varying number of data points per class (denoted by M).

M	≈80	50	25	10
Student	54.13 (± 0.50)	44.13 (± 0.30)	29.05 (± 0.72)	15.92 (± 0.67)
Finetuned	66.39 (± 0.41)	58.51 (± 0.45)	51.97 (± 0.31)	39.93 (± 0.58)
LwF	58.18 (± 0.53)	49.68 (± 2.09)	38.08 (± 3.33)	26.09 (± 1.08)
FitNet	71.00 (± 0.60)	64.05 (± 0.63)	55.30 (± 1.42)	40.67 (± 0.13)
AT	60.57 (± 0.30)	53.11 (± 0.83)	42.64 (± 0.57)	26.12 (± 0.52)
NST	55.40 (± 0.34)	47.29 (± 1.23)	34.03 (± 1.19)	21.27 (± 0.71)
VID-LP	68.21 (± 0.59)	61.77 (± 0.57)	50.75 (± 0.49)	39.23 (± 0.11)
VID-I	**71.99** (± 0.19)	66.62 (± 0.75)	**59.00** (± 0.38)	46.24 (± 0.31)
LwF + FitNet	70.75 (± 0.47)	64.38 (± 1.13)	55.60 (± 0.13)	41.34 (± 0.33)
VID-LP + VID-I	71.44 (± 1.21)	**66.67** (± 0.50)	57.59 (± 0.23)	**46.42** (± 1.01)

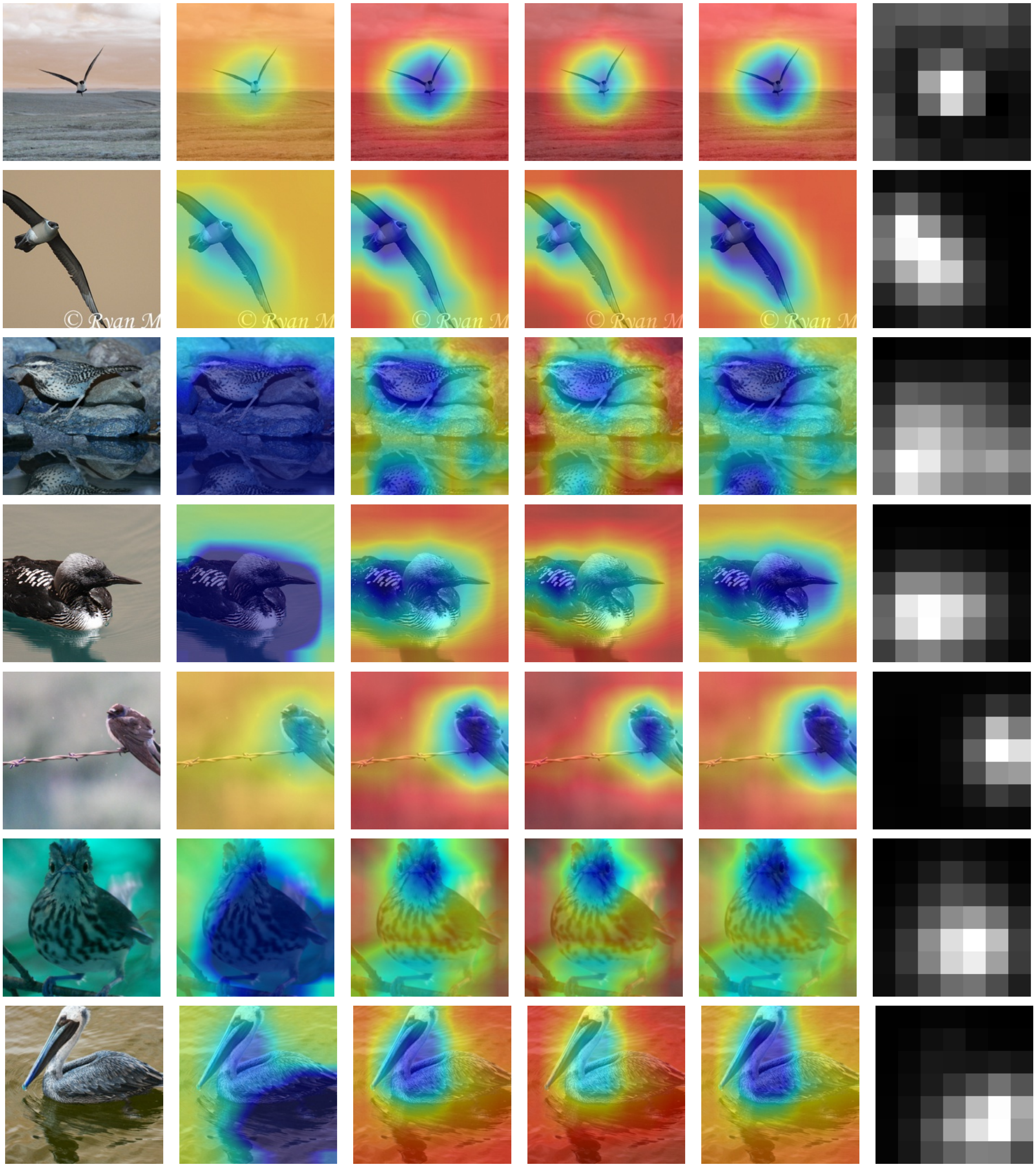
Table 10: Experimental results (test accuracy) of transfer learning from the teacher network (ResNet-34) to the student network (VGG-9) for the MIT-67 dataset with varying number of data points per class (denoted by M).

M	≈29.95	20	10	5
Student	44.59 (± 1.93)	32.10 (± 0.65)	15.69 (± 0.27)	9.66 (± 0.22)
Finetuned	60.96 (± 1.88)	51.86 (± 0.99)	46.88 (± 0.92)	39.98 (± 0.33)
LwF	52.54 (± 0.12)	36.38 (± 0.14)	22.79 (± 0.35)	11.52 (± 0.15)
FitNet	68.96 (± 0.45)	61.52 (± 0.80)	48.04 (± 0.64)	32.89 (± 1.95)
AT	56.28 (± 1.75)	43.96 (± 0.80)	28.33 (± 0.17)	13.98 (± 1.01)
NST	56.55 (± 2.05)	44.95 (± 0.36)	28.43 (± 0.35)	14.66 (± 2.48)
VID-LP	66.82 (± 0.41)	55.94 (± 0.27)	38.10 (± 0.83)	30.47 (± 0.31)
VID-I	**71.51** (± 1.48)	**65.69** (± 0.68)	**53.29** (± 1.20)	**38.09** (± 1.05)
LwF + FitNet	68.40 (± 0.50)	61.40 (± 0.40)	45.57 (± 0.04)	28.41 (± 0.24)
VID-LP + VID-I	70.03 (± 0.05)	63.46 (± 0.40)	48.79 (± 0.04)	32.35 (± 0.24)

Table 11: Experimental results (test accuracy) of transfer learning from the teacher network (ResNet-34) to the student network (VGG-9) for the CUB-200 dataset with varying number of data points per class (denoted by M).

Network	MLP-4096	MLP-2048	MLP-1024
Student	70.60 (± 0.26)	70.78 (± 0.45)	70.90 (± 0.13)
KD	70.42 (± 0.26)	70.53 (± 0.18)	70.79 (± 0.35)
FitNet	76.02 (± 0.26)	74.08 (± 0.18)	72.91 (± 0.35)
VID-I	**85.18** (± 0.20)	**83.47** (± 0.29)	**78.57** (± 0.11)

Table 12: Experimental result (test accuracy) of distillation on CIFAR-10 from the convolutional teacher network (WRN-40-2) to the fully connected student network (MLP-h) with varying size of hidden dimensions h.



(a) input (b) 0-th epoch (c) 40-th epoch (d) 160-th epoch (e) no transfer (f) magnitude of $\boldsymbol{t}_{h,w}$

Figure 4: Plots for the heat maps corresponding to the variational distribution evaluated for spatial dimensions of the intermediate layer in the teacher network, *i.e.*, $\log q(\boldsymbol{t}_{h,w}|\boldsymbol{s}) = \sum_c \log q(t_{c,h,w}|\boldsymbol{s})$. Each figure corresponds to (a) original input image, (b, c, d) log-likelihood $\log q(\boldsymbol{t}_{h,w}|\boldsymbol{s})$ that was normalized and interpolated to fit the spatial dimension of the input image (red pixels correspond to high probability), (d) log-likelihood of variational distribution optimized for the student network trained without any knowledge transfer applied and (f) magnitude of the layer $\boldsymbol{t}$ averaged for each spatial dimensions.