

Paper review for Variational Information Distillation for Knowledge Transfer

Paper author : Sungsoo Ahn et. al.

Korea Advanced Institute of Science and Technology, Daejeon, Korea

Abstract. 본 논문에서는 지금까지 제안된 knowledge distillation 기법들의 문제점으로 그들에게 아직 common agreed theory가 없으며 단순히 neural network의 feature를 통해 hand-crafted feature를 정의한다는 점을 지적했다. 따라서 이를 해결하기 위한 방법으로 mutual information을 기반으로 한 knowledge 정의 방식을 제안하였다. 이를 통해 기존 기법에 비해 더 common agreed theory에 가까울뿐만 아니라 더 높은 성능을 보인다고 주장하고 있다. 또한 실험 결과를 통해 convolutional neural network(CNN)을 이용한 small dataset training, transfer learning은 물론 CNN의 knowledge를 multilayer perceptron(MLP)에 transfer하는 것도 가능하다고 주장하고 있다.

Keywords: Knowledge distillation, mutual information, transfer learning

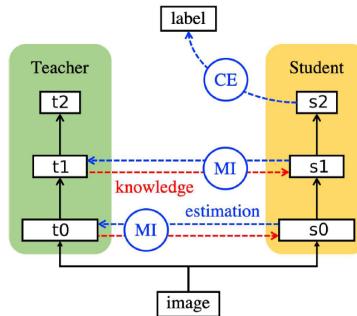
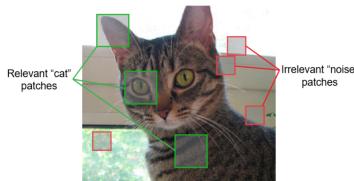
1 제안 기법

제안 기법의 개념도는 1과 같다. 제안 기법은 teacher network와 student network의 같은 level의 feature maps가 가진 mutual information을 maximize하도록 학습한다. 하지만 mutual information은 intractable하기 때문에 variational lower bound로 이를 대체한다. 이 과정은 다른 mutual information을 학습하는 알고리즘에서 자주 사용되는 trivial한 방법이기 때문에 생략한다. 저자들은 variational lower bound를 variational information이라 부르며 이에 따라 제안 기법명을 variational information distillation (VID)로 정의하였다. Knowledge를 transfer하기 위해 target task와 VID를 multi-task learning으로 학습하며 target task를 위한 loss function \mathcal{L}_S 를 포함한 loss function은 아래 수식과 같다.

$$\tilde{\mathcal{L}} = \mathcal{L}_S - \sum_{k=1}^K \lambda_k \mathbb{E}_{\mathbf{t}^{(k)}, \mathbf{s}^{(k)}} \left[\log q \left(\mathbf{t}^{(k)} | \mathbf{s}^{(k)} \right) \right] \quad (1)$$

여기서 λ_k 는 regularization scale이며 $\mathbf{t}^{(k)}, \mathbf{s}^{(k)}$ 는 각각 teacher와 student의 random variables로 sensed feature map을 통해 정의된다.

저자들은 1을 기반으로 두 가지 loss function을 제안하였다. 하지만 수식적으로, 기법적으로 차이가 없기 때문에 feature map을 target으로 한 VID-I에 대해 설명하며 이는 아래 수식과 같다.

**Fig. 1.** 제안 기법의 전체적인 개념도**Fig. 2.** student network와 VID를 통해 학습한 student network의 attention map 비교.

$$\begin{aligned} -\log q(\mathbf{t}|\mathbf{s}) &= -\sum_{c=1}^C \sum_{h=1}^H \sum_{w=1}^W \log q(t_{c,h,w}|\mathbf{s}) \\ &= \sum_{c=1}^C \sum_{h=1}^H \sum_{w=1}^W \log \sigma_c + \frac{((t_{c,h,w} - \mu_{c,h,w}(\mathbf{s})))^2}{2\sigma_n^2} \end{aligned} \quad (2)$$

여기서 μ 는 mean parameter로 student feature map에 세 개의 convolution layer를 적용하여 구해진다. σ 는 student feature map과 독립적인 trainable parameter이다.

저자들을 이를 통해 학습할 경우 VID는 기존의 다른 task와 연결될 수 있다고 주장한다. 먼저 infomax problem과 관련이 있다고 주장한다. 2에서 배경은 information이 적고 고양이는 information이 많다고 할 수 있다. 따라서 image의 information을 잘 학습하려면 고양이 부분에 attention을 줘야한다. 실제로 VID를 통해 학습한 결과 1과 같이 student network만 학습할 때보다 중요한 부분에 집중적으로 attention map이 형성되는 경향을 보인다. 또한 기존 knowledge distillation 기법들이 많이 사용하는 L_2 -distance 기반의 feature matching의 generalization version이 될 수 있다고 주장한다.

2 실험 결과

사용한 dataset은 cifar10, cifar100 그리고 CUB-200-2011, MIT-67이다. 실험 결과는 아래 table1, 2, 3, 4과 같다. 모든 실험 결과에서 제안 기법이 다른 기법에 비해 우위에 있는 것을 알 수 있다. 특히 transfer learning에서 outperform하는 것을 알 수 있다.

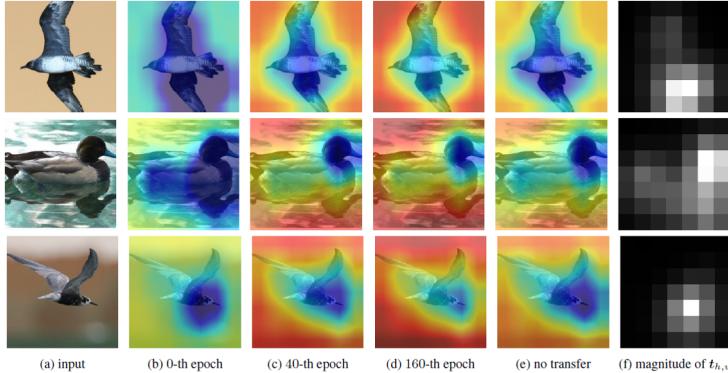


Figure 2: Plots for the heat maps corresponding to the variational distribution evaluated for spatial dimensions of the intermediate layer in the teacher network, i.e., $\log q(t_{h,w}|s) = \sum_c \log q(t_{c,h,w}|s)$. Each figure corresponds to (a) original input image, (b, c, d) log-likelihood $\log q(t_{h,w}|s)$ 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 t averaged for each spatial dimensions.

| M | 5000 | 1000 | 500 | 100 |
|------------|--------------|--------------|--------------|--------------|
| Teacher | 94.26 | - | - | - |
| Student | 90.72 | 84.67 | 79.63 | 58.84 |
| KD | 91.27 | 86.11 | 82.23 | 64.24 |
| FitNet | 90.64 | 84.78 | 80.73 | 68.90 |
| AT | 91.60 | 87.26 | 84.94 | 73.40 |
| NST | 91.16 | 86.55 | 82.61 | 64.53 |
| VID-I | 91.85 | 89.73 | 88.09 | 81.59 |
| KD + AT | 91.81 | 87.34 | 85.01 | 76.29 |
| KD + VID-I | 91.7 | 88.59 | 86.53 | 78.48 |

Table 1: 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 | - | - | - |
| Student | 74.34 | 70.42 | 68.79 | 65.46 |
| KD | 75.80 | 72.87 | 70.99 | 66.03 |
| FitNet | 74.29 | 70.89 | 68.66 | 65.38 |
| AT | 74.76 | 71.06 | 69.85 | 65.31 |
| NST | 74.81 | 71.19 | 68.00 | 64.95 |
| VID-I | 75.25 | 73.31 | 71.51 | 66.32 |
| KD + AT | 75.86 | 73.13 | 71.4 | 67.07 |
| KD + VID-I | 76.11 | 73.69 | 72.16 | 67.19 |

Table 2: 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 .

3 결론 및 개인적인 견해

저자들은 variational lower bound를 학습하는 방법으로 기존의 기법에 비해 더 유연하고 강력한 알고리즘을 설계했다고 주장하고 있다. 하지만 제안 기법이 다른 기법과 다른 점은 그저 포장지일 뿐으로 보인다. 저자들이 처음 지적한 문제 즉 ”기존 기법들은 common agreed theory가 없으며 hand-crafted feature를 사용한다.”는 문제와 제안 기법과 관련성이 없어보인다. 저자들은 variational information을 구하기 위해 많은 가정과 trainable layers를 통한 hand-crafted feature를 사용했으며 이를 teacher와 비교하고 있다. 또한 Mutual information은 두 random variable의 차이를 측정하는 도구일 뿐이다. FitNet, AT, AB 모두 student와 teacher feature map의 크기를 맞추기 위해 몇 개의 convolutional layer를 사용하며, 유일한 차이점은 두 feature map의 차이를 비교하는 방법이기 때문에 결국 이들과 대등조이하다. 특히 저자들이 제안 기법의 효과를 보이기 위한 1의 경우 AT에서 주장하는 것과 거의 같아보인다. 그렇다면 실제로 infomax problem을 직접적으로 풀고있는 AT와 비교를 하는 것이 맞다고 보여진다.

| M | ≈80 | 50 | 25 | 10 |
|----------------|--------------|--------------|--------------|--------------|
| Student | 48.13 | 37.69 | 27.01 | 14.25 |
| fine-tuning | 70.97 | 66.04 | 58.13 | 47.91 |
| LwF | 63.43 | 51.79 | 41.04 | 22.76 |
| FitNet | 71.34 | 60.45 | 54.78 | 36.94 |
| AT | 58.21 | 48.66 | 43.66 | 27.01 |
| NST | 55.52 | 46.34 | 33.21 | 20.82 |
| VID-LP | 67.91 | 58.51 | 47.09 | 31.94 |
| VID-I | 71.34 | 63.66 | 60.07 | 50.97 |
| LwF + FitNet | 70.97 | 60.37 | 54.48 | 38.73 |
| VID-LP + VID-I | 71.87 | 65.75 | 61.79 | 50.37 |

| (a) MIT-67, ResNet-34 to ResNet-18 | | | | |
|------------------------------------|--------------|--------------|--------------|--------------|
| M | ≈29.95 | 20 | 10 | 5 |
| Student | 37.22 | 24.33 | 12.00 | 7.09 |
| fine-tuning | 76.69 | 71.00 | 59.25 | 44.07 |
| LwF | 55.18 | 42.13 | 26.23 | 14.27 |
| FitNet | 66.63 | 56.63 | 46.68 | 31.04 |
| AT | 54.62 | 41.44 | 28.90 | 16.55 |
| NST | 55.01 | 41.87 | 23.76 | 15.63 |
| VID-LP | 65.59 | 54.12 | 39.20 | 27.86 |
| VID-I | 73.25 | 67.20 | 56.86 | 46.21 |
| LwF + FitNet | 68.69 | 58.81 | 48.86 | 31.30 |
| VID-LP + VID-I | 69.71 | 63.94 | 52.87 | 41.12 |

| (c) CUB-200-2011, ResNet-34 to ResNet-18 | | | | |
|--|--------------|--------------|--------------|--------------|
| M | ≈29.95 | 20 | 10 | 5 |
| Student | 44.59 | 32.10 | 15.69 | 9.66 |
| fine-tuning | 60.96 | 51.86 | 46.88 | 39.98 |
| LwF | 52.18 | 38.05 | 25.57 | 13.93 |
| FitNet | 68.96 | 61.52 | 48.04 | 32.89 |
| AT | 56.28 | 43.96 | 28.33 | 13.98 |
| NST | 56.55 | 44.95 | 28.43 | 14.66 |
| VID-LP | 66.82 | 55.94 | 38.10 | 30.47 |
| VID-I | 71.51 | 65.69 | 53.29 | 38.09 |
| LwF + FitNet | 70.56 | 62.44 | 47.36 | 30.52 |
| VID-LP + VID-I | 70.00 | 65.14 | 53.78 | 38.76 |

Table 3: Experimental results (test accuracy) of transfer learning from the teacher network (ResNet-34) to the student network (ResNet-18/VGG-9) for the MIT-67/CUB-200-2011 dataset with varying number of data points per class (denoted by M). We use $M \approx M_{\text{avg}}$ to denote the setting where the number of data points per class is non-uniform and M_{avg} in average. Fine-tuning gives good results on transfer learning, but is not directly comparable as it is not a knowledge transfer method.

[t]