Learn from Concepts: Towards the Purified Memory for Few-shot Learning

1 Purified Memory

Let's start by the definition of information bottleneck:

$$\max I(v;y) - \beta I(v;z), \tag{1}$$

where $I(.;.)$ denotes the mutual information, y represents the label, z is the feature, v denotes our purified concept and β is the Lagrange coefficient. However, directly optimizing Eq. (1) would be intractable. Instead, we propose an another objective to ensure:

$$I(z;y) = I(v;y), \tag{2}$$

Based on the definition of mutual information, we have:

$$I(v;z) := H(v) - H(v|z), \tag{3}$$

where $H(v)$ denotes Shannon entropy, and $H(v|z)$ is the conditional entropy of z given v. Based on the symmetry of mutual information, we have:

$$I(v;z) = I(z;v). \tag{4}$$

Hence we have the following results:

$$\begin{aligned}
&I(v;y) = I(z;y) \\
&\iff I(y;v) = I(y;z) \\
&\iff H(y) - H(y|v) = H(y) - H(y|z) \\
&\iff H(y|v) = H(y|z).
\end{aligned} \tag{5}$$

Based on the definition of conditional entropy, for any continuous variables v, y and z, we have:

$$\begin{aligned}
&H(y|v) - H(y|z) = \\
&-\int p(v)dv \int p(y|v)\log p(y|v)dy \\
&+\int p(z)dz \int p(y|z)\log p(y|z)dy \ = \\
&-\iint p(v)p(y|v)\log\left[\frac{p(y|v)}{p(y|z)}p(y|z)\right]dvdy \\
&+\iint p(z)p(y|z)\log\left[\frac{p(y|z)}{p(y|v)}p(y|v)\right]dzdy.
\end{aligned} \tag{6}$$

By factorizing the double integrals in Eq. (6) into another two components, we show the following:

$$\begin{aligned}
&\iint p(v)p(y|v)\log\left[\frac{p(y|v)}{p(y|z)}p(y|z)\right]dvdy = \\
&\iint \underbrace{p(v)p(y|v)\log\frac{p(y|v)}{p(y|z)}dvdy}_{\text{term } V_1} + \\
&\iint \underbrace{p(v)p(y|v)\log p(y|z)dvdy}_{\text{term } V_2}.
\end{aligned} \tag{7}$$

Conduct similar factorization for the second term in Eq.(6), we have:

$$\begin{aligned}
&\iint p(z)p(y|z)\log\left[\frac{p(y|z)}{p(y|v)}p(y|v)\right]dzdy = \\
&\iint \underbrace{p(z)p(y|z)\log\frac{p(y|z)}{p(y|v)}dzdy}_{\text{term } Z_1} + \\
&\iint \underbrace{p(z)p(y|z)\log p(y|v)dzdy}_{\text{term } Z_2}.
\end{aligned} \tag{8}$$

Integrate term V_1 and term Z_1 over y:

$$V_1 = \int p(v)D_{KL}[p(y|v)\|p(y|z)]dv, \tag{9}$$

$$Z_1 = \int p(z)D_{KL}[p(y|z)\|p(y|v)]dz, \tag{10}$$

where $D_{KL}[.||.]$ denotes KL-divergence. Integrate term V_2 and term Z_2 over v and z respectively, we have:

$$V_2 = \int p(y)\log p(y|z)dy \tag{11}$$

$$Z_2 = \int p(y)\log p(y|v)dy. \tag{12}$$

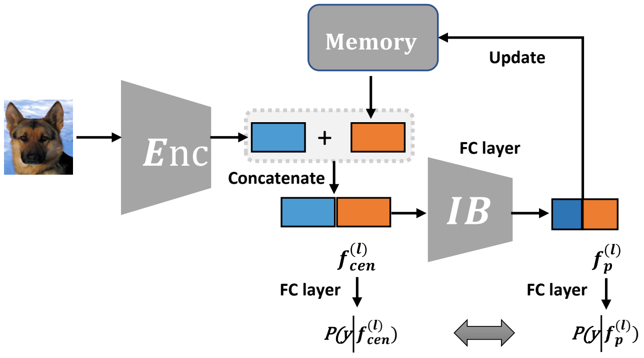

Figure 1: Purified Memory Updating Scheme.

In the view of above, we have the following:

$$I(v;y)-I(z;y)=H(y|v)-H(y|z)=$$
$$\int p(z)D_{KL}[p(y|z)\|p(y|v)]dz+\int p(y)\log\left[\frac{p(y|v)}{p(y|z)}\right]dy$$
$$-\int p(v)D_{KL}[p(y|v)\|p(y|z)]dv \tag{13}$$

Based on the non-negativity of KL-divergence, Eq. (13) is upper bounded by:

$$\int p(z)D_{KL}[p(y|z)\|p(y|v)]dz+\int p(y)\log\left[\frac{p(y|v)}{p(y|z)}\right]dy. \tag{14}$$

Equivalently, we have the upper bound as:

$$\mathbb{E}_{v\sim E_\theta(v|x)}\mathbb{E}_{z\sim E_\phi(z|v)}[D_{KL}[p(y|z)\|p(y|v)]]$$
$$+\mathbb{E}_{v\sim E_\theta(v|x)}\mathbb{E}_{z\sim E_\phi(z|v)}\left[\log\left[\frac{p(y|v)}{p(y|z)}\right]\right], \tag{15}$$

where θ,ϕ denote the parameters of the encoder and the information bottleneck. In this sense, we omit the second term since it is proportional to $D_{KL}[p(y|z)\|p(y|v)]$. And finally draw our loss.

Also there is a clear physical interpretation behind our loss. That is the predicted information is unchanged after the purification process. From this perspective, our learned concept is progressively stable and consistent by compressing the representation. In practice, we implement memory updating as shown in Fig. 1.

2 Experiment

Datasets. We evaluate MA-GNN on four few-shot learning benchmarks followed by [Yang *et al.*, 2020]: miniImageNet [Vinyals *et al.*, 2016], tieredImageNet [Ren *et al.*, 2018], CUB-200-2011 [wah,] and CIFAR-FS [Bertinetto *et al.*, 2018]. Among them, miniImageNet and tieredImageNet are collected from ImageNet, and CIFAR-FS is a subset from CIFAR-100. Unlike theses datasets, CUB-200-2011 is a fine-grained bird classification dataset. In Table 1, we summarize the statistics of datasets including number of images, number of categories, train/val/test splits and image resolution.

Dataset	Images	Classes	Train-val-test	Resolution
miniImageNet	60000	100	64/16/20	84 × 84
tieredImageNet	779165	608	351/97/160	84 × 84
CUB-200-2011	11788	200	100/50/50	84 × 84
CIFAR-FS	60000	100	64/16/20	32 × 32

Table 1: Statistics for Few-Shot Learning Datasets.

	miniImageNet 5-way		tieredImageNet 5-way	
Model	**1-shot**	**5-shot**	**1-shot**	**5-shot**
B	60.27	77.50	64.58	81.66
Non	63.16		68.38	
Naive	64.27		69.98	
PB	64.84	80.89	70.33	82.50
Ours	**65.87** $^{\uparrow 5.6}$	**82.23** $^{\uparrow 4.83}$	**71.69** $^{\uparrow 7.11}$	**84.4** $^{\uparrow 2.77}$

Table 2: Quantitative results when using different memory mechanism. ”B”: our baseline (EGNN); “Non-Mem”: meta-knowledge nodes are implemented by the class center from the current episode; “Naive-Mem”: a memory storing the entire features; “PB-Mem”: prototype-based memory.

3 Ablation Study

Here we also present the quantitative results when using different memory mechanism. The results are shown in Table 2. Note that compared with the baseline method, our new memory mechanism could bring a nearly 5% improvement on average. The results are consistent with our observations in the main paper.

References

[Bertinetto *et al.*, 2018] Luca Bertinetto, Joao F Henriques, Philip HS Torr, and Andrea Vedaldi. Meta-learning with differentiable closed-form solvers. *arXiv preprint arXiv:1805.08136*, 2018.

[Ren *et al.*, 2018] Mengye Ren, Eleni Triantafillou, Sachin Ravi, Jake Snell, Kevin Swersky, Joshua B Tenenbaum, Hugo Larochelle, and Richard S Zemel. Meta-learning for semi-supervised few-shot classification. *arXiv preprint arXiv:1803.00676*, 2018.

[Vinyals *et al.*, 2016] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. In *Advances in neural information processing systems*, pages 3630–3638, 2016.

[wah,]

[Yang *et al.*, 2020] Ling Yang, Liangliang Li, Zilun Zhang, Xinyu Zhou, Erjin Zhou, and Yu Liu. Dpgn: Distribution propagation graph network for few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13390–13399, 2020.