Instance and Category Supervision are Alternate Learners for Continual Learning

1. Appendix

In this section, we introduce and prove the theorems mentioned in the main text of this paper.

A. Proof to Theorem 1

Given sample x, we have the data observations v^A, v^B of its augmented viewpoints. On this basis, the embeddings z^A, z^B are obtained from the projector f_ψ.

To maximally preserve sample information that is invariant *w.r.t.* distortions, an initial objective can be formulated as follows:

$$\max_{\psi} I(z^A; v^B), \tag{1}$$

where $I(z^A; v^B)$ denotes the mutual information between z^A and v^B. Based on the information processing inequality, we show:

$$I(z^A; v^B) \leq I(v^A; v^B), \tag{2}$$

which suggests another solution to Eq. (1), *i.e.*, approximating $I(z^A; v^B)$ to its upper-bound. Hence, a refined objective can be given as:

$$\min_{\psi} I(v^A; v^B) - I(z^A; v^B). \tag{3}$$

According to the definition of mutual information [1],

$$I(z; v) := H(v) - H(v|z), \tag{4}$$

where $H(v)$ denotes Shannon entropy, and $H(v|z)$ is the conditional entropy of z given v [1]. On this basis, we rearrange Eq. (3) by:

$$\begin{aligned} &I(v^A; v^B) - I(z^A; v^B) \\ &= H(v^B) - H(v^B|v^A) - H(v^B) + H(v^B|z^A) \\ &= H(v^B|z^A) - H(v^B|v^A). \end{aligned} \tag{5}$$

Therefore, Eq. (3) is equivalent to:

$$\min H(v^B|z^A) - H(v^B|v^A). \tag{6}$$

Recall the definition of conditional entropy, for continual variables v^A, z^A, and v^B, we have:

$$\begin{aligned} &I(v^A; v^B) - I(z^A; v^B) = H(v^B|z^A) - H(v^B|v^A) = \\ &- \int p(z^A)dz^A \int p(v^B|z^A) \log p(v^B|z^A)dv^B \\ &+ \int p(v^A)dv^A \int p(v^B|v^A) \log p(v^B|v^A)dv^B \ = \\ &- \iint p(z^A)p(v^B|z^A) \log \left[\frac{p(v^B|z^A)}{p(v^B|v^A)} p(v^B|v^A)\right] dz^A dv^B \\ &+ \iint p(v^A)p(v^B|v^A) \log \left[\frac{p(v^B|v^A)}{p(v^B|z^A)} p(v^B|z^A)\right] dv^A dv^B. \end{aligned} \tag{7}$$

By factorizing the double integrals in Eq. (7) into another two components, we show the following:

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

Conduct similar factorization for the second term in Eq. (7), we have:

$$\begin{aligned} &\iint p(v^A)p(v^B|v^A) \log \left[\frac{p(v^B|v^A)}{p(v^B|z^A)} p(v^B|z^A)\right] dv^A dv^B \\ &= \underbrace{\iint p(v^A)p(v^B|v^A) \log \frac{p(v^B|v^A)}{p(v^B|z^A)} dv^A dv^B}_{\text{term } V_1} \\ &+ \underbrace{\iint p(v^A)p(v^B|v^A) \log p(v^B|z^A) dv^A dv^B}_{\text{term } V_2}. \end{aligned} \tag{9}$$

Integrate term Z_1 and term V_1 over v^B:

$$Z_1 = \int p(z^A) D_{KL}[p(v^B|z^A) \| p(v^B|v^A)] dz^A, \quad (10)$$

$$V_1 = \int p(v^A) D_{KL}[p(v^B|v^A) \| p(v^B|z^A)] dv^A, \quad (11)$$

where D_{KL} denotes KL-divergence. By integrating term Z_2 and term V_2 over z^A and v^A respectively, we have:

$$Z_2 = \int p(v^B) \log p(v^B|v^A) dv^B, \quad (12)$$

$$V_2 = \int p(v^B) \log p(v^B|z^A) dv^B. \quad (13)$$

In the view of above, we have the following:

$$\begin{aligned} &I(v;y) - I(z;y) = H(y|z) - H(y|v) \\ &= \int p(v^A) D_{KL}[p(v^B|v^A) \| p(v^B|z^A)] dv^A \\ &+ \int p(v^B) \log \left[\frac{p(v^B|z^A)}{p(v^B|v^A)} \right] dv^B \\ &- \int p(z^A) D_{KL}[p(v^B|z^A) \| p(v^B|v^A)] dz^A. \end{aligned} \quad (14)$$

Based on the non-negativity of KL-divergence, Eq. (14) is upper-bounded by:

$$\begin{aligned} &\int p(v^A) D_{KL}[p(v^B|v^A) \| p(v^B|z^A)] dv^A \\ &+ \int p(v^B) \log \left[\frac{p(v^B|z^A)}{p(v^B|v^A)} \right] dv^B. \end{aligned} \quad (15)$$

Equivalently, we have the upper-bound as:

$$\begin{aligned} &\mathbb{E}_{v^A, v^B \sim f(x;\theta)} \mathbb{E}_{z^A \sim f(v;\psi)} [D_{KL}[p(v^B|v^A) \| p(v^B|z^A)]] \\ &+ \mathbb{E}_{v^A, v^B \sim f(x;\theta)} \mathbb{E}_{z^A \sim f(v;\psi)} \left[\log \left[\frac{p(v^B|z^A)}{p(v^B|v^A)} \right] \right], \end{aligned} \quad (16)$$

where θ, ψ parameterize the encoder and projector, respectively. Therefore, the objective of maximizing the invariant sample information in z^A can be formulated as:

$$\min_{\theta,\psi} \mathbb{E}_{v \sim f(x;\theta)} \mathbb{E}_{z \sim f(v;\psi)} \left[D_{KL}[\mathbb{P}_v^A || \mathbb{P}_z^A] + \log \left[\frac{\mathbb{P}_z^A}{\mathbb{P}_v^A} \right] \right], \quad (17)$$

in which $\mathbb{P}_z^A = p(v^B|z^A)$ and $\mathbb{P}_v^A = p(v^B|v^A)$ denote the predicted distributions of the representation and observation, respectively.

Clearly, the objective of preserving sample information is equivalent to minimizing the discrepancy between the predicted distributions of v^A and z^A. Notice that this can be achieved by minimizing $D_{KL}(\mathbb{P}_v^A || \mathbb{P}_z^A)$, which explicitly approximates $p(v^B|z^A)$ to $p(v^B|v^A)$ and implicitly reduce the second term in Eq.(17) in the same time. Ideally, the representation z^A retrieves all sample information shared by the other viewpoint v^B when $\mathbb{P}_z^A$ coincides with $\mathbb{P}_v^A$, *i.e.*,:

$$D_{KL}[\mathbb{P}_v^A || \mathbb{P}_z^A] = 0 \Rightarrow \mathbb{P}_z^A = \mathbb{P}_v^A \quad (18)$$

Based on Eq. (14) , we show the following:

$$\begin{aligned} &D_{KL}[p(v^B|v^A) \| p(v^B|z^A)] dv^A = 0 \\ &\Rightarrow H(v^B|z^A) - H(v^B|v^A) = I(v^A; v^B) - I(z^A; v^B) = 0, \end{aligned} \quad (19)$$

which reveals that minimizing $D_{KL}[\mathbb{P}_v^A || \mathbb{P}_z^A]$ is consistent with the objective of preserving the invariant sample information for z^A (symmetric to z^B). Thus Theorem 1 holds.

Implementation. To avoid intractability in $\mathcal{L}_{SSL}^s$, we assume a variational distribution $\mathcal{N}\left(v^B | R\left(z^A\right), \sigma \mathbf{I}\right)$ with R as a deterministic mapping to reconstruct v^B from z^A (more details could be found in [4]). On this basis, we have $\mathcal{L}_{SSL}^s$ defined as:

$$\min_{z^A = f(v^A;\psi), R} D_{KL}\left[p(v^B|v^A) || p(v^B|R(z^A))\right], \quad (20)$$

where $p(v^B|R(z^A))$ (similar for $p(v^B|v^A)$) is approximated with $\frac{1}{n}\sum_{i=1}^{n} \log \frac{e^{\langle R(z_i^A), v_i^B \rangle}}{\frac{1}{n}\sum_{j=1}^{n} e^{\langle R(z_i^A), v_j^B \rangle}}$, a common formulation in self-supervised learning.

B. Proof to Theorem 2

Given v^A, v^B as the observations of sample x from different augmentations, and z^A, z^B are the corresponding representations. As analyzed in [5, 4, 2], removing variation caused by distortions plays a crucial role in unleashing instance-level discrimination. To this end, we provide the following analytical solutions to explicitly eliminating $I(v^A; z^A|v^B)$ and $I(v^B; z^B|v^A)$ without diminishing the sample information.

Considering that $z^A, z^B \sim f(v;\psi)$, $I(v_A; z_A|v_B)$ can be expressed as:

$$\begin{aligned} &I(v^A; z^A|v^B) = \mathbb{E}_{v^A, v^B \sim f(x;\theta)} \mathbb{E}_{z^A, z^B \sim f(v;\psi)} \left[\log \frac{p(z^A|v^A)}{p(z^A|v^B)} \right] \\ &= \mathbb{E}_{v^A, v^B \sim f(x;\theta)} \mathbb{E}_{z^A, z^B \sim f(v;\psi)} \left[\log \frac{p(z^A|v^A) p(z^B|v^B)}{p(z^B|v^B) p(z^A|v^B)} \right] \\ &= D_{KL}[p(z^A|v^A) || p(z^B|v^B)] - D_{KL}[p(z^B|v^A) || p(z^B|v^B)] \\ &\leq D_{KL}[p(z^A|v^A) || p(z^B|v^B)]. \end{aligned} \quad (21)$$

Notice this bound is tight whenever z^A and z^B preserve sufficient sample information [2], which can be assured by the Theorem 1 (proved in appendix.A). On this basis, we

Table 1: Evaluation of our alternate leaner using different memory size. All experiments are conducted on CIFAR-100 under task-incremental protocol. Average time of our baseline with the same setting is adopted as 1.0x.

Setting	0.5k	1k	2k	4k
Avg. Acc	73.49	77.03	80.04	81.93
Avg. Time	1.14x	1.31x	1.42x	1.79x

formulate the following objective to explicitly removing the variation $I(v_A; z_A|v_B)$:

$$\mathbb{E}_{v^A,v^B\sim f(x;\theta)}\mathbb{E}_{z^A,z^B\sim f(v;\psi)}\left[D_{KL}[\mathbb{P}^A_{z|v}||\mathbb{P}^B_{z|v}]\right], \quad (22)$$

in which $\mathbb{P}^A_{z|v} = p(z^A|v^A)$ and $\mathbb{P}^B_{z|v} = p(z^B|v^B)$ denote the predicted distributions. Similarly, we introduce the following objective to minimize $I(v^B; z^B|v^A)$.

$$\mathbb{E}_{v^A,v^B\sim f(x;\theta)}\mathbb{E}_{z^A,z^B\sim f(v;\psi)}\left[D_{KL}[\mathbb{P}^B_{z|v}||\mathbb{P}^A_{z|v}]\right], \quad (23)$$

For simplicity, we apply Eq. (24) to eliminate the variation for both views.

$$\mathbb{E}_{v^A,v^B\sim f(x;\theta)}\mathbb{E}_{z^A,z^B\sim f(v;\psi)}\left[D_{JS}[\mathbb{P}^A_{z|v}||\mathbb{P}^B_{z|v}]\right], \quad (24)$$

where D_{JS} denotes the Jensen-Shannon divergence. In the view of above, Theorem 2 holds.

Implementation. Suggested by [2], the projector ψ is modeled by Normal distribution parametrized with a neural network $(\mu_\psi, \sigma^2_\psi)$, which defines $p(z^A|v^A)$ as $\mathcal{N}\left(z^A|\mu_\psi(v^A), \sigma^2_\psi(v_A)\right)$ (symmetric for $p(z^B|v^B)$). On this basis, the density of ψ can be evaluated, and thus the JS-divergence between $p(z^A|v^A)$ and $p(z^B|v^B)$ can be directly computed.

C. Additional experiments

In this section, we evaluate our approach with different memory sizes and demonstrate the generalization to other vision task (*i.e.*, image classification).

According to Tab. 1, the memory size acts a critical role to the performance, which provides significant improvement with linearly increased cost. On the other hand, our method outperforms all competitors with the most popular memory budget (*i.e.*, 2k), and achieves competitive accuracies with much less storage (*i.e.*, 0.5k and 1k).

Furthermore, we apply our SSL strategy to image classification and exhibit the comparison with other techniques in Tab. 2, where the superior performance demonstrates the generalization ability of our theories.

Table 2: Evaluation of different self-supervised strategies on image classification. All experiments obtained by following the baseline in [3] with the ResNet-18 architecture and a KNN classifier.

Setting	CIFAR-10	CIFAR-100	Tiny-ImageNet
SimSiam	95.76	86.31	82.89
BarlowTwins	95.48	87.16	82.42
Ours	96.09	87.32	83.31

References

[1] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm. Mutual information neural estimation. In *Proceedings of the 35th International Conference on Machine Learning*, pages 531–540, 2018.

[2] Marco Federici, Anjan Dutta, Patrick Forré, Nate Kushman, and Zeynep Akata. Learning robust representations via multi-view information bottleneck. In *Proceedings of the International Conference on Learning Representations*, 2020.

[3] Divyam Madaan, Jaehong Yoon, Yuanchun Li, Yunxin Liu, and Sung Ju Hwang. Representational continuity for unsupervised continual learning. In *International Conference on Learning Representations, ICLR*, 2022.

[4] Yao-Hung Hubert Tsai, Yue Wu, Ruslan Salakhutdinov, and Louis-Philippe Morency. Self-supervised learning from a multi-view perspective. In *International Conference on Learning Representations*, 2021.

[5] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning, (ICML)*, pages 12310–12320, 2021.