Mutual Conditional Probability for Self-Supervised Learning

Takumi Kobayashi
National Institute of AIST, Japan
University of Tsukuba

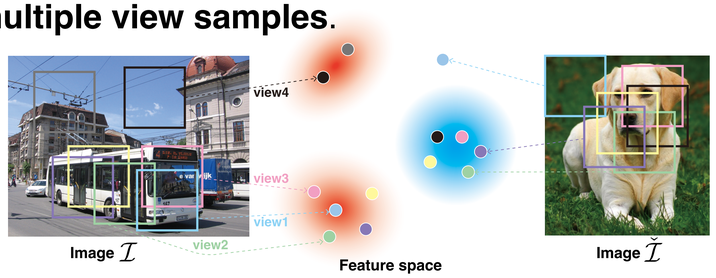
BMVC 2022

Contributions

- ✓ We propose a novel **self-supervised learning (SSL)** method to exploit **multiple view samples**.
- ✓ **Mutual conditional probability** embeds compact and discriminative feature representation into SSL.
- ✓ Our SSL produces favorable performance even on only 100-epoch ImageNet training.

Probabilistic Model

Multiple data augmentations produce **multiple view samples**.



Multiple view samples are probabilistically modeled by means of kernel density estimation (**KDE**).

Generative prob. (GenPro) $\mathrm{p}(\boldsymbol{z}|\mathcal{I}) = \frac{1}{MC_\sigma}\sum_{i=1}^{M}\exp\left(-\frac{1}{2\sigma^2}\|\boldsymbol{z}-\boldsymbol{x}_i\|_2^2\right)$

loss $\ell_{gen}(\boldsymbol{z}) = -\sigma^2\log\mathrm{p}(\boldsymbol{z}|\mathcal{I}) \Rightarrow \nabla_{\boldsymbol{z}}\ell_{gen} =$ MeanShift-vec.

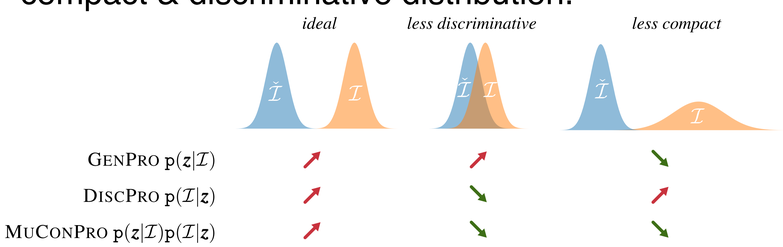
Discriminative prob. (DiscPro) $\mathrm{p}(\mathcal{I}|\boldsymbol{z}) = \frac{\mathrm{p}(\boldsymbol{z}|\mathcal{I})}{\mathrm{p}(\boldsymbol{z})} \propto \frac{\sum_{i|y_i=\mathcal{I}}\exp(-\frac{1}{2\sigma^2}\|\boldsymbol{z}-\boldsymbol{x}_i\|_2^2)}{\sum_{i=1}^{N}\exp(-\frac{1}{2\sigma^2}\|\boldsymbol{z}-\boldsymbol{x}_i\|_2^2)}$

loss $\ell_{disc}(\boldsymbol{z}) = -\sigma^2\log\mathrm{p}(\mathcal{I}|\boldsymbol{z})$

Mutual conditional prob. (MuConPro) $\mathrm{p}(\boldsymbol{z}|\mathcal{I})\mathrm{p}(\mathcal{I}|\boldsymbol{z})$

loss $\ell(\boldsymbol{z}) = -\sigma^2\log[\mathrm{p}(\boldsymbol{z}|\mathcal{I})\mathrm{p}(\mathcal{I}|\boldsymbol{z})] = \ell_{gen}(\boldsymbol{z}) + \ell_{disc}(\boldsymbol{z})$

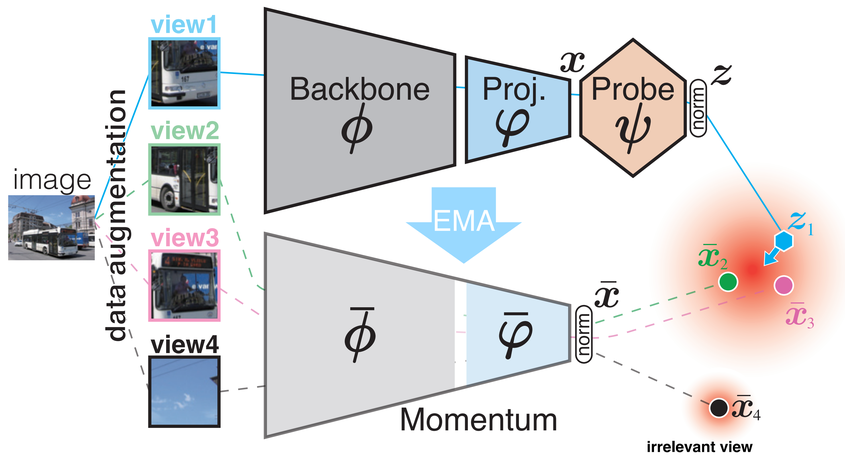
MuConPro effectively describes compact & discriminative distribution.



Proposed SSL

Our SSL trains a model based on **MuConPro** loss.

Feature representation is updated by maximizing MuConPro toward **compact and discriminative** ones, while **disregarding irrelevant view samples** via KDE.



$$\ell_{SSL}(\boldsymbol{z}_{bi}) = -2T\log\sum_{j=1|j\neq i}^{M}\exp\left(\boldsymbol{z}_{bi}^\top\bar{\boldsymbol{x}}_{bj}/T\right) + T\log\sum_{c=1,j=1|c\neq b\vee j\neq i}^{B,M}\exp\left(\boldsymbol{z}_{bi}^\top\bar{\boldsymbol{x}}_{cj}/T\right)$$

Compactness within image — Discrimination across images

Multiple view samples are effectively aggregated in a MuConPro manner which naturally induces compact and discriminative feature representation.

Temperature T controls the bandwidth in KDE.

Analysis by loss gradients.

- $T\to\infty$ results in BYOL [Grill+20].

$$\frac{\partial\ell_{gen}(\boldsymbol{z}_{bi})}{\partial\boldsymbol{z}_{bi}} = -\frac{\sum_{j\neq i}\exp(\boldsymbol{z}_{bi}^\top\bar{\boldsymbol{x}}_{bj}/T)\bar{\boldsymbol{x}}_{bj}}{\sum_{j\neq i}\exp(\boldsymbol{z}_{bi}^\top\bar{\boldsymbol{x}}_{bj}/T)} \xrightarrow{T\to\infty} -\frac{1}{M-1}\sum_{j\neq i}\bar{\boldsymbol{x}}_{bj} = \frac{\partial}{\partial\boldsymbol{z}}\mathop{\mathbb{E}}_{j\neq i}[-\boldsymbol{z}_{bi}^\top\bar{\boldsymbol{x}}_{bj}]$$

- $T\to 0$ results in too sparse updating.

$$\frac{\partial\ell_{gen}(\boldsymbol{z}_{bi})}{\partial\boldsymbol{z}_{bi}} = -\frac{\sum_{j\neq i}\exp(\boldsymbol{z}_{bi}^\top\bar{\boldsymbol{x}}_{bj}/T)\bar{\boldsymbol{x}}_{bj}}{\sum_{j\neq i}\exp(\boldsymbol{z}_{bi}^\top\bar{\boldsymbol{x}}_{bj}/T)} \xrightarrow{T\to 0} -\arg\max_{\{\bar{\boldsymbol{x}}_{bj}\}_{j\neq i}}\boldsymbol{z}_{bi}^\top\bar{\boldsymbol{x}}_{bj}$$

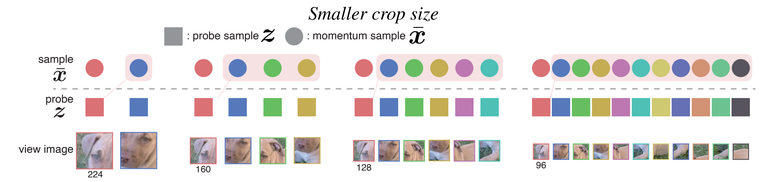
We simply apply $T=1$.

Experimental Results

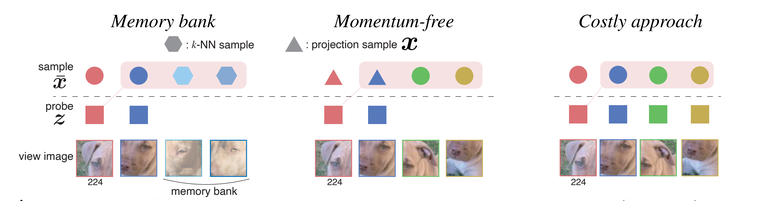
Ablation Study on ImageNet-100 with ResNet18.

Multiple samples are generated in various ways in a constant computation budget.

- Cropping small-sized image patches.
- Drawing from external memory bank.
- Processing large-sized patch via momentum-free mode.



views	$2\cdot(224\times224)$	$4\cdot(160\times160)$	$6\cdot(128\times128)$	$11\cdot(96\times96)$
linear	77.94	**81.64**	81.30	79.20
k-NN	68.90	**74.06**	73.82	71.54



views	2-NN samples	2 projection samples	$4\cdot(224\times224)$
linear	78.04	78.68	81.12
k-NN	69.94	69.86	73.54

Probabilistic models are compared with various T.

Model	(a) Linear evaluation $T=0$	1	∞	(b) k-NN evaluation $T=0$	1	∞
GenPro $\mathrm{p}(\boldsymbol{z}\|\mathcal{I})$	79.24	81.38	80.90	71.50	73.38	72.98
DiscPro $\mathrm{p}(\mathcal{I}\|\boldsymbol{z})$	17.86	81.56	81.08	16.58	73.46	73.38
MuConPro $\mathrm{p}(\boldsymbol{z}\|\mathcal{I})\mathrm{p}(\mathcal{I}\|\boldsymbol{z})$	78.38	**81.64**	81.50	70.16	**74.06**	74.00

Performance comparison with various SSL on 100-training epochs.

MuConPro with $4\cdot(160\times160)$ views and $T=1$.

Model	Acc. (%)	Model	Acc. (%)
SimCLR [5]	66.5	*Supervised [16]*	*76.2*
BYOL [14]	66.5	SwAV [3]	72.1
MoCo-v2 [7]	67.4	MS-SSL [22]	72.4
MoCo-v3 [8]	68.9	GenPro $\mathrm{p}(\boldsymbol{z}\|\mathcal{I})$	70.9
SimSiam [6]	68.1	DiscPro $\mathrm{p}(\mathcal{I}\|\boldsymbol{z})$	72.1
W-MSE [12]	69.4	MuConPro $\mathrm{p}(\boldsymbol{z}\|\mathcal{I})\mathrm{p}(\mathcal{I}\|\boldsymbol{z})$	**72.9**