RényiCL : Contrastive Representation Learning with Skew Rényi Divergence

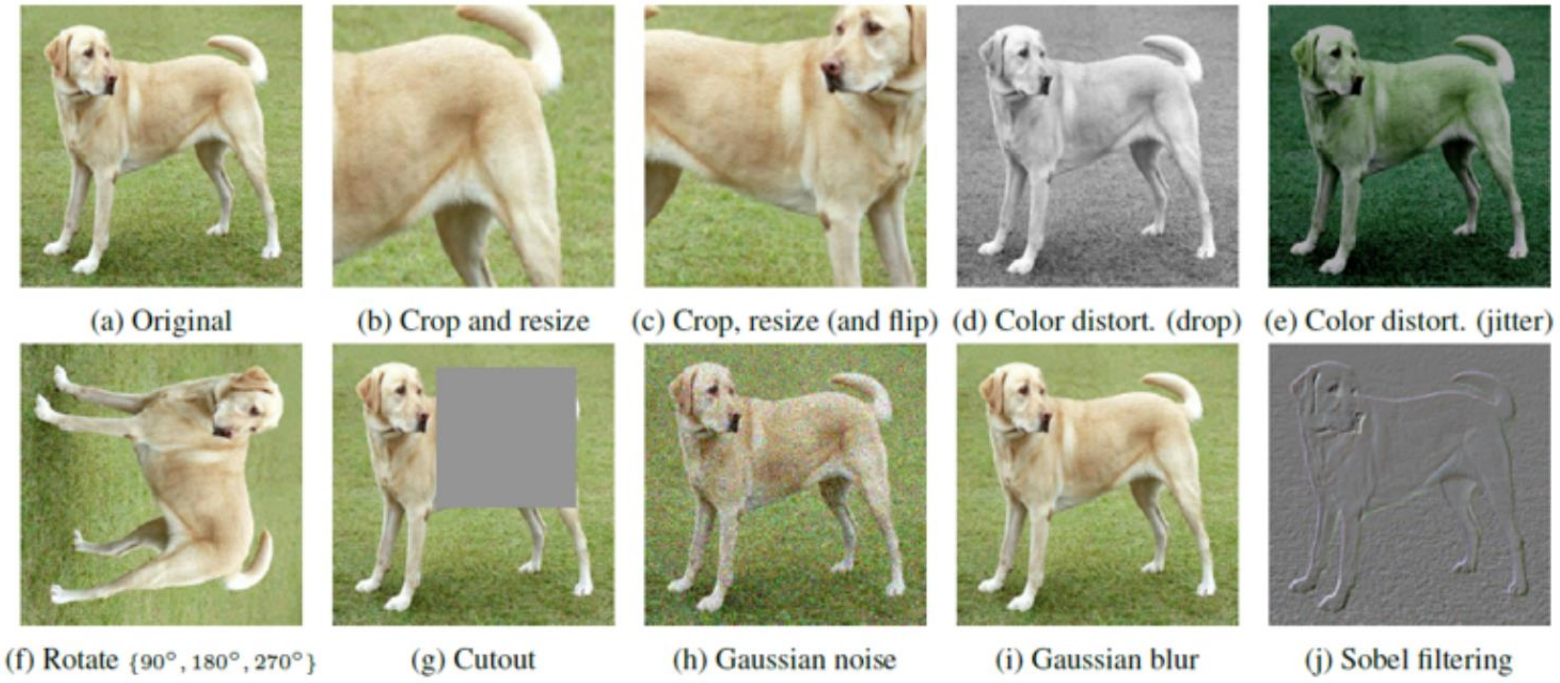
-Summary-

Introduction

- Two key components of contrastive learning:

1. Data augmentation [starting from SimCLR, 2020]

   : Generate different views for positive pairs where the views share relevant information



Introduction

2. Contrastive objective

: Enforces the representation to capture the shared information between two positive pairs and pushing the negative pairs

1) Donsker-Varadhan (DV) objective [Belghazi et al., 2018] :

$$D_{KL}(P \parallel Q) = \sup_{f \in \mathcal{F}} I_{DV}(f), \qquad where\ \ I_{DV} \coloneqq \mathbb{E}_P[f] - \log \mathbb{E}_Q\left[e^f\right]$$

Hence, $I(X;Y) = D_{KL}(P_{XY} \parallel P_X P_Y) = \sup_{f \in \mathcal{F}} \mathbb{E}_{P_{XY}}[f(x,y)] - \log \mathbb{E}_{P_X P_Y}\left[e^{f(x,y)}\right]$

: I_{DV} for mutual information

Problem :

a. DV objective may have large variance unless one uses large number of samples

b. Empirically, Contrastive learning with DV objective suffers from training instability

https://www.ee.bgu.ac.il/~haimp/ml/lectures/lec9_MINE/MINE_MUTUAL_INFORMATION.pdf (Donsker-Varadhan representation lecture note)

Introduction

2) Contrastive predictive coding (CPC) objective [Oord et al., 2018] :

$$I_{CPC}(f) \coloneqq \mathbb{E}\left[\frac{1}{B}\sum_{i=1}^{B}\log\frac{(K+1)\cdot e^{f(x_i,y_i^+)}}{e^{f(x_i,y_i^+)}+\sum_{j=1}^{K} e^{f\left(x_i,y_{ij}^-\right)}}\right]$$

: Address issues from DV-objective => Popular choice for contrastive objective

B : batch size of samples

Problem :

a. It is well known $I_{CPC}(f) \leq \min\{\, I(X;Y), \log(K+1)\,\}$

(When $I(X;Y) \gg \log\,(K+1)$, it suffers high bias problem)

Introduction

3) α-CPC objective [Poole et al., 2019] :

$$I_{CPC}^{(\alpha)}(f) \coloneqq \mathbb{E}\left[\frac{1}{B}\sum_{i=1}^{B}\log\frac{e^{f(x_i,y_i^+)}}{\alpha\cdot e^{f(x_i,y_i^+)}+\frac{1-\alpha}{K}\cdot\sum_{j=1}^{K}e^{f\left(x_i,y_{ij}^-\right)}}\right]$$

: Can achieves smaller bias ($\because I_{CPC}^{(\alpha)}(f) \leq \log\left(\frac{K+1}{\alpha}\right)$) using trade-off between bias & variance

***B* : batch size of samples**

***K* : size of negative samples**

Problem :

a. It is not guaranteed that $I_{CPC}^{(\alpha)}(f) \leq I(X;Y)$ (exists counter-example)

: Smaller α can reduce the bias, but also induce higher estimate above $I(X;Y)$

Introduction

4) **α-MLCPC objective** [Song et al., 2020] :

$$I_{MLCPC}^{(\alpha)}(f) \coloneqq \mathbb{E}\left[\frac{1}{B}\sum_{i=1}^{B}\log\frac{e^{f(x_i, y_i^+)}}{\frac{\alpha}{B}\sum_{i=1}^{B} e^{f(x_i, y_i^+)} + \frac{1-\alpha}{BK}\cdot\sum_{i=1}^{B}\sum_{j=1}^{K} e^{f\left(x_i, y_{ij}^-\right)}}\right]$$

: Can achieves smaller bias ($\because I_{MLCPC}^{(\alpha)}(f) \leq \log\left(\frac{K+1}{\alpha}\right)$) using trade-off between bias & variance. Also, It is guaranteed that $I_{MLCPC}^{(\alpha)}(f) \leq I(X;Y)$.

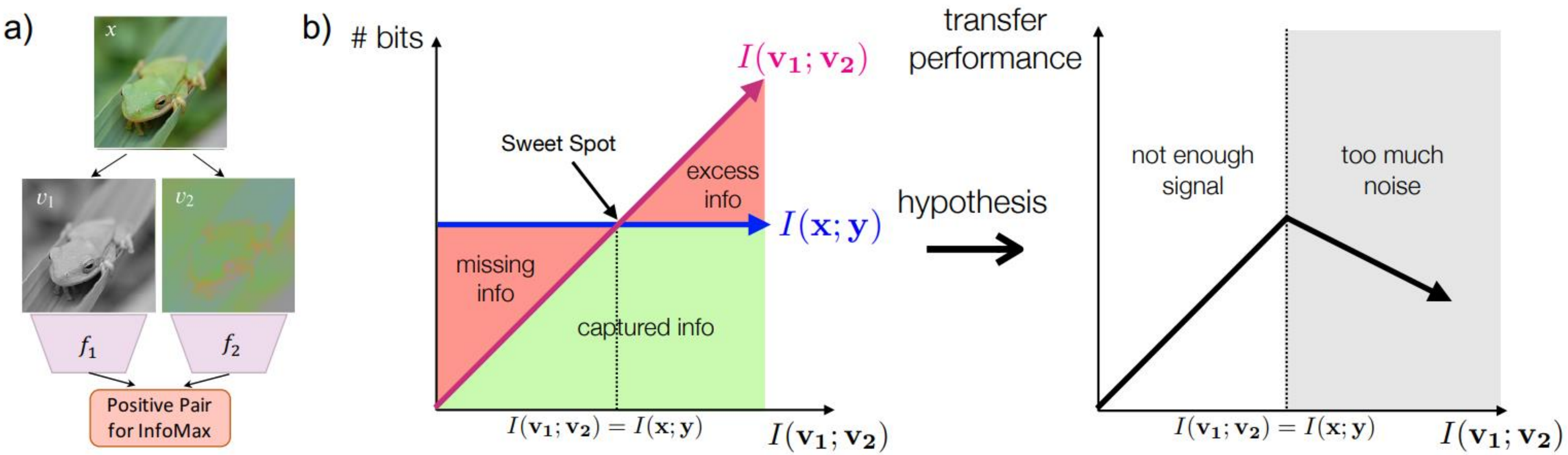
B **: batch size of samples**
K **: size of negative samples**

Introduction

What makes for good views for Contrastive Learning? [Tian et al., 2020]

- One problem on data augmentation
  1. Augmented views can share insufficient information.
     => The representation of them is hard to be learned with sufficient features.

  2. Augmented views can share too much information.
     => the views may share nuisance features that degrade the generalization.



Introduction (Rényi divergence)

- How about adopting Rényi divergence instead of KL-divergence for MI estimation?

- Rényi divergence of order $\gamma \in (0,1) \cup (1, \infty)$: (when $P \ll Q$)

$$R_\gamma(P \parallel Q) := \frac{1}{\gamma(\gamma-1)} \log \mathbb{E}_P \left[\left(\frac{dP}{dQ} \right)^{\gamma-1} \right]$$

- **[Background]** Rényi entropy and Rényi divergence (continuous ver) [from Erven, 2007] :

  1. Rényi entropy : most general way to quantify information while preserving some axioms to satisfy :

$$H_\gamma(X) = \frac{1}{1-\gamma} \log \mathbb{E}_P[(dP)^{\gamma-1}] = \frac{1}{1-\gamma} \log \int P^\gamma d\mu$$

Renyi divergence basics paper (https://arxiv.org/pdf/1206.2459.pdf)

Introduction (Rényi divergence)

- **[Background]** Rényi entropy and Rényi divergence :

  2. Rényi divergence :

$$R_r(P \parallel Q) \coloneqq \frac{1}{\gamma(\gamma-1)} \log \mathbb{E}_P\left[\left(\frac{dP}{dQ}\right)^{\gamma-1}\right]$$

Property :

① [Positivity] : For any $\gamma \in (0,1) \cup (1,\infty)$:

$$R_\gamma(P \parallel Q) \geq 0 \quad \text{and equality holds iff } P = Q$$

② [Extended orders] :

$$\lim_{\gamma \to 1} R_r(P \parallel Q) = D_{KL}(P \parallel Q)$$

$$R_2(P \parallel Q) = \frac{1}{2}\log(1 + \mathcal{X}^2(P,Q)), \text{where } \mathcal{X}^2(P,Q) \coloneqq \int \left(\frac{dP}{dQ} - 1\right)^2 dQ$$

Rényi divergence

- Using similar DV objective approach to Rényi divergence :

- **[Lem]** Rényi divergence of order γ admits following variational form [Birrell et al., 2021]:

$$R_r(P \parallel Q) = \sup_{f \in \mathcal{F}} I^{(\gamma)}_{Renyi}(f) \quad \text{for} \quad I^{(\gamma)}_{Renyi}(f) \coloneqq \frac{1}{\gamma - 1} \log \mathbb{E}_P\left[e^{(\gamma-1)f}\right] - \frac{1}{\gamma} \log \mathbb{E}_Q\left[e^{\gamma f}\right]$$

where optimal $f^* = \log(\frac{dP}{dQ})$

- Using this lemma, $R_r(P_{XY} \parallel P_X P_Y) = \sup_{f \in \mathcal{F}} I^{(\gamma)}_{Renyi}(f)$

where $I^{(\gamma)}_{Renyi}(f) \coloneqq \frac{1}{\gamma-1} \log \mathbb{E}_{P_{XY}}\left[e^{(\gamma-1)f}\right] - \frac{1}{\gamma} \log \mathbb{E}_{P_X P_y}\left[e^{\gamma f}\right]$

=> **Can we estimate MI by maximizing $\boldsymbol{I^{(\gamma)}_{Renyi}(f)}$?** [No, it suffer from high variance]

Rényi divergence

- Theoretically, We can show as below :

- **[Thm]** Let P_m and Q_n be the empirical distributions of m i.i.d samples from P and Q, then

$$\lim_{n\to\infty} n \cdot Var_{P,Q}[\,\hat{I}^{(\gamma)}_{Renyi}(f^*)] \geq \frac{e^{\gamma^2 D_{KL}(P\|Q)} - \gamma^2}{e^{\gamma(\gamma-1)R_\gamma(P\|Q)}}$$

where $\hat{I}^{(\gamma)}_{Renyi}(f) \coloneqq \frac{1}{\gamma-1}\log \mathbb{E}_{P_m}\left[e^{(\gamma-1)f}\right] - \frac{1}{\gamma}\log \mathbb{E}_{Q_n}\left[e^{\gamma f}\right]$

- Interpretation :

Even if we achieve optimal f^*, the variance of $\hat{I}^{(\gamma)}_{Renyi}(f)$ can explodes as we suggested previous slide.

Interpretation of CPC and MLCPC

- Recall that generalized form of α-CPC and α-MLCPC:
  1. α-CPC :
  $$I_{CPC}^{(\alpha)}(f) = \mathbb{E}_{P_{XY}}[f(x,y)] - \mathbb{E}_{P_X}\left[\log\left(\alpha\mathbb{E}_{P_{Y|X}}\left[e^{f(x,y)}\right] + (1-\alpha)\mathbb{E}_{P_Y}\left[e^{f(x,y)}\right]\right)\right]$$
  2. α-MLCPC :
  $$I_{MLCPC}^{(\alpha)}(f) = \mathbb{E}_{P_{XY}}[f(x,y)] - \log\left(\alpha\mathbb{E}_{P_{XY}}\left[e^{f(x,y)}\right] + (1-\alpha)\mathbb{E}_{P_X P_Y}\left[e^{f(x,y)}\right]\right)$$

- If we define ***α*-skew KL divergence**, then we can connect above things : ($\alpha \in [0,1]$)
$$D_{KL}^{(\alpha)}(P \parallel Q) := D_{KL}(P \parallel \alpha P + (1-\alpha)Q)$$

Interpretation of CPC and MLCPC

- First observe that we can write DV objective of α-skew KL divergence as follows :

$$D_{KL}^{(\alpha)}(P \parallel Q) = \sup_{f \in \mathcal{F}} \mathbb{E}_P[f] - \log\left(\alpha \mathbb{E}_P\left[e^f\right] + (1-\alpha)\mathbb{E}_Q\left[e^f\right]\right)$$

- Sketch : (Put $T^*(X) = \log \frac{P(X)}{Q\,(x)}$ and replace $Q(x) \Rightarrow \alpha P(x) + (1-\alpha)Q(x)$)

$$\begin{aligned}
\mathbb{E}_P[T^*(X)] - \log(\mathbb{E}_Q[e^{T^*(X)}]) &\overset{(a)}{=} \mathbb{E}_P\left[\log \frac{P(X)}{Q(X)}\right] - \log\left(\mathbb{E}_Q\left[e^{\log \frac{P(X)}{Q(X)}}\right]\right) \\
&\overset{(b)}{=} D_{KL}(P||Q) - \log\left(\mathbb{E}_Q\left[\frac{P(X)}{Q(X)}\right]\right) \\
&\overset{(c)}{=} D_{KL}(P||Q) - \log\left(\int_{\mathcal{X}} Q(x)\frac{P(x)}{Q(x)}dx\right) \\
&= D_{KL}(P||Q) - \log\left(\int_{\mathcal{X}} P(x)dx\right) \\
&\overset{(d)}{=} D_{KL}(P||Q) - \log(1) \\
&= D_{KL}(P||Q).
\end{aligned}$$

Optimal $f^* = log\,\left(\frac{P}{\alpha P + (1-\alpha)Q}\right)$

Interpretation of CPC and MLCPC

- Then, we can figure out the followings :

1. $\sup_{f \in \mathcal{F}} I_{CPC}^{(\alpha)}(f) = \mathbb{E}_{P_x}\left[D_{KL}^{(\alpha)}\left(P_{Y|X} \parallel P_Y\right)\right]$ where $f^* = \log\left(\frac{P_{Y|X}}{\alpha P_{Y|X}+(1-\alpha)P_Y}\right)$
2. $\sup_{f \in \mathcal{F}} I_{MLCPC}^{(\alpha)}(f) = D_{KL}^{(\alpha)}(P_{XY} \parallel P_X P_Y)$ where $f^* = \log\left(\frac{P_{XY}}{\alpha P_{XY}+(1-\alpha)P_X P_Y}\right)$

(By comparing above formula to DV objective of α-skew KL divergence)

Bounded variance property for α-skew KL divergence

- Using α-skew KL divergence concept, we can show that the variance of CPC and MLCPC become low-variance estimator (having bounded variance) of the MI :

  (Similarly, we can show bounded variance property for $R_2^{(\alpha)}$)

- **[Thm]** Let P_m and Q_n be the empirical distributions of m i.i.d samples from P and Q, and assume there is $\hat{f} \in \mathcal{F}$ such that $\left| I_{KL}^{(\alpha)}(\hat{f}) - D_{KL}^{(\alpha)}(P_m \parallel Q_n) \right| < \epsilon_f$ for some $\epsilon_f > 0$. Then, for any $\alpha < \frac{1}{8}$, the variance of estimator satisfies :

$$Var_{P,Q}\left[\hat{I}_{KL}^{(\alpha)}(\hat{f})\right] \leq c_1 \epsilon_f + \frac{c_2(\alpha)}{\min\{n, m\}} + \frac{\boxed{c_3 \log^2(\alpha m)}}{m} + \frac{c_4 \log^2(c_5 n)}{\boxed{\alpha^2 n}}$$

where $c_2(\alpha) = \min\left\{\frac{1}{\alpha}, \frac{\mathcal{X}^2(P\|Q)}{1-\alpha}\right\}$

Small α can leads to loose upper bound

https://arxiv.org/pdf/2106.07898.pdf (Proof outline material)

Variational estimation of skew Rényi divergence

- Again, similarly define **α-skew Rényi divergence of order γ** :

$$R_\gamma^{(\alpha)}(P \parallel Q) \coloneqq R_\gamma(P \parallel \alpha P + (1-\alpha)Q)$$

- Also, define **(α, γ)- Rényi MLCPC (RMLCPC)** objective using DV-objective :

$$I_{RMLCPC}^{(\alpha,\gamma)}(f) \coloneqq \frac{1}{\gamma - 1}\log \mathbb{E}_{P_{XY}}\left[e^{(\gamma-1)f(x,y)}\right] - \frac{1}{\gamma}\log\left(\alpha \mathbb{E}_{P_{XY}}\left[e^{\gamma f(x,y)}\right] + (1-\alpha)\mathbb{E}_{P_X P_Y}\left[e^{\gamma f(x,y)}\right]\right)$$

such that $\sup_{f\in\mathcal{F}} I_{RMLCPC}^{(\alpha,\gamma)}(f) = R_\gamma^{(\alpha)}(P_{XY} \parallel P_X P_Y)$

- Now, we can use $I_{RMLCPC}^{(\alpha,\gamma)}(f)$ to estimate $R_\gamma^{(\alpha)}(P_{XY} \parallel P_X P_Y)$ by maximizing it.
  [InfoMax principle]

Rényi Contrastive Representation Learning

- Inspired by the fact that "Rényi divergence penalizes more when two distributions differ",
  We can expect RényiCL can learn more discriminative representation.
  (∵Harder data augmentation in CL results in more dissimilar augmented positive pairs)

- Adopting InfoMax principle :
  Our goal : find g which discriminates positive / negative pairs as much as possible

$$\sup_{g:\mathcal{X}\to\mathbb{R}^d} D_{KL}\left(P_{g(V)g(V')} \parallel P_{g(V)}P_{g(V')}\right) = \sup_{g:\mathcal{X}\to\mathbb{R}^d} I\big(g(V); g(V')\big) \leq I(V; V')$$

Rényi Contrastive Representation Learning

- If we adopt α-MLCPC or (α,γ)-RMLCPC objective, our goal becomes as follows :

1. α-MLCPC objective :

$$\sup_{g:\mathcal{X}\to\mathbb{R}^d, f\in\mathcal{F}} I^{(\alpha)}_{MLCPC}(f,g) = \sup_{g:\mathcal{X}\to\mathbb{R}^d} D^{(\alpha)}_{KL}\left(P_{g(V)g(V')} \parallel P_{g(V)}P_{g(V')}\right)$$

2. (α,γ)-RMLCPC objective :

$$\sup_{g:\mathcal{X}\to\mathbb{R}^d, f\in\mathcal{F}} I^{(\alpha,\gamma)}_{RMLCPC}(f,g) = \sup_{g:\mathcal{X}\to\mathbb{R}^d} R^{(\alpha)}_{\gamma}\left(P_{g(V)g(V')} \parallel P_{g(V)}P_{g(V')}\right)$$

Gradient analysis for RényiCL

- We claimed that "Rényi divergence penalizes more when two distributions differ", then How does $I_{RMLCPC}^{(\alpha,\gamma)}(f)$ induces different updating rule during GD update?

  Recall : $\boxed{\mathcal{I}_{\mathrm{MLCPC}}^{(\alpha)}(f_\theta) = \mathbb{E}_{v,v^+}[f_\theta(v,v^+)] - \log\left(\alpha\mathbb{E}_{v,v^+}[e^{f_\theta(v,v^+)}] + (1-\alpha)\mathbb{E}_{v,v^-}[e^{f_\theta(v,v^-)}]\right)}$

- For the baseline case ($I_{MLCPC}^{(\alpha)}$) :

$$
\begin{aligned}
&\nabla_\theta \mathcal{I}_{\mathrm{MLCPC}}^{(\alpha)}(f_\theta) \\
&= \mathbb{E}_{v,v^+}[\nabla_\theta f_\theta(v,v^+)] - \frac{\alpha\mathbb{E}_{v,v^+}[e^{f_\theta(v,v^+)}\nabla_\theta f_\theta(v,v^+)] + (1-\alpha)\mathbb{E}_{v,v^-}[e^{f_\theta(v,v^-)}\nabla_\theta f_\theta(v,v^-)]}{\alpha\mathbb{E}_{v,v^+}[e^{f_\theta(v,v^+)}] + (1-\alpha)\mathbb{E}_{v,v^-}[e^{f_\theta(v,v^-)}]} \\
&= \mathbb{E}_{v,v^+}[\nabla_\theta f_\theta(v,v^+)] - \left(\mathbb{E}_{\mathtt{sg}(q_\theta(v,v^+))}[\nabla_\theta f_\theta(v,v^+)] + \mathbb{E}_{\mathtt{sg}(q_\theta(v,v^-))}[\nabla_\theta f_\theta(v,v^-)]\right),
\end{aligned}
$$

where

$$
q_\theta(v,v^+) \propto \frac{\alpha p(v,v^+)e^{f_\theta(v,v^+)}}{\alpha\mathbb{E}_{v,v^+}[e^{f_\theta(v,v^+)}] + (1-\alpha)\mathbb{E}_{v,v^-}[e^{f_\theta(v,v^-)}]}
$$

$$
q_\theta(v,v^-) \propto \frac{(1-\alpha)p(v)p(v^-)e^{f_\theta(v,v^-)}}{\alpha\mathbb{E}_{v,v^+}[e^{f_\theta(v,v^+)}] + (1-\alpha)\mathbb{E}_{v,v^-}[e^{f_\theta(v,v^-)}]},
$$

Gradient analysis for RényiCL

- Recall : $\boxed{\mathcal{I}^{(\alpha,\gamma)}_{\mathrm{RMLCPC}}(f_\theta) = \frac{1}{\gamma-1}\log \mathbb{E}_{v,v^+}[e^{(\gamma-1)f_\theta(v,v^+)}] - \frac{1}{\gamma}\log\left(\alpha\mathbb{E}_{v,v^+}[e^{\gamma f_\theta(v,v^+)}] + (1-\alpha)\mathbb{E}_{v,v^-}[e^{\gamma f_\theta(v,v^-)}]\right)}$
- For the RMLCPC case ($I^{(\alpha,\gamma)}_{RMLCPC}$) :

$$
\begin{aligned}
&\nabla_\theta \mathcal{I}^{(\alpha,\gamma)}_{\mathrm{RMLCPC}}(f_\theta) \\
&= \mathbb{E}_{q_\theta^{(1)}(v,v^+)}[\nabla_\theta f_\theta(v,v^+)] - \frac{\alpha\mathbb{E}_{v,v^+}[e^{\gamma f_\theta(v,v^+)}\nabla_\theta f_\theta(v,v^+)] + (1-\alpha)\mathbb{E}_{v,v^-}[e^{\gamma f_\theta(v,v^-)}\nabla_\theta f_\theta(v,v^-)]}{\alpha\mathbb{E}_{v,v^+}[e^{\gamma f_\theta(v,v^+)}] + (1-\alpha)\mathbb{E}_{v,v^-}[e^{\gamma f_\theta(v,v^-)}]} \\
&= \mathbb{E}_{\mathtt{sg}(q_\theta^{(1)}(v,v^+))}[\nabla_\theta f_\theta(v,v^+)] - \left(\mathbb{E}_{\mathtt{sg}(q_\theta^{(2)}(v,v^+))}[\nabla_\theta f_\theta(v,v^+)] + \mathbb{E}_{\mathtt{sg}(q_\theta^{(2)}(v,v^-))}[\nabla_\theta f_\theta(v,v^-)]\right),
\end{aligned}
$$

where

$$
\begin{aligned}
q_\theta^{(1)}(v,v^+) &\propto p(v,v^+)\underline{e^{(\gamma-1)f_\theta(v,v^+)}} \\
q_\theta^{(2)}(v,v^+) &\propto \frac{\alpha p(v,v^+)\underline{e^{\gamma f_\theta(v,v^+)}}}{\alpha\mathbb{E}_{v,v^+}[e^{\gamma f_\theta(v,v^+)}] + (1-\alpha)\mathbb{E}_{v,v^-}[e^{\gamma f_\theta(v,v^-)}]} \\
q_\theta^{(2)}(v,v^-) &\propto \frac{(1-\alpha)p(v)p(v^-)\underline{e^{\gamma f_\theta(v,v^-)}}}{\alpha\mathbb{E}_{v,v^+}[e^{\gamma f_\theta(v,v^+)}] + (1-\alpha)\mathbb{E}_{v,v^-}[e^{\gamma f_\theta(v,v^-)}]},
\end{aligned}
$$

Higher $\nabla_\theta f_\theta$ -> larger importance weight q_θ

Gradient analysis for RényiCL

- In conclusion , RMLCPC has following properties :

1. **Hard negative sampling :**
   the gradient weights more on harder negatives (v, v^-) with high values of $f_\theta(v, v^-)$
   as γ increases.

2. **Easy positive sampling :**
   the gradient weights more on easier positives (v, v^+) with high value of $f_\theta(v, v^+)$
   as $\gamma \in (1, \infty)$ increases.

- **Effect of α in RényiCL :** Using small α (practical case to reduce bias) leads to the training largely affected by easy positive and hard negative (small effect on 2nd term)
  => helps to learn discriminative representation

Gradient analysis for RényiCL

- In conclusion , RMLCPC has following properties :

- Effect of α in RényiCL : Using small α (practical case to reduce bias) leads to the training largely affected by easy positive and hard negative (small effect on 2nd term)
  => helps to learn discriminative representation

α^{-1}	1024	4096	16384	65536
Base Aug.	79.0	79.3	78.6	78.4
Hard Aug.	81.1	81.3	**81.6**	81.1
Gap	+2.1	+2.0	**+3.0**	+2.7

Experiments (Linear evaluation / Semi-supervised learning)

Method	Epochs	Top-1
SimCLR [3]	800	70.4
Barlow Twins [24]	800	73.2
BYOL [44]	800	74.3
MoCo v3 [6]	800	74.6
SwAV [41]	800	75.3
DINO [41]	800	75.3
NNCLR [19]	1000	75.6
C-BYOL [45]	1000	75.6
RényiCL	**200**	**75.3**
RényiCL	**300**	**76.2**

	1% ImageNet		10% ImageNet	
Method	Top-1	Top-5	Top-1	Top-5
Supervised [3]	25.4	48.4	56.4	80.4
SimCLR [3]	48.3	75.5	65.6	87.8
BYOL [44]	53.2	78.4	68.8	89.0
SwAV [41]	53.9	78.5	70.2	89.9
Barlow Twins [24]	55.0	79.2	69.7	89.3
NNCLR [19]	56.4	80.7	69.8	89.3
C-BYOL [45]	**60.6**	**83.4**	70.5	90.0
RényiCL	56.4	80.6	**71.2**	**90.3**

Linear evaluation (Top-1 acc) on the ImageNet validation set **[different methods for CL] (left)**

Semi-supervised top-1 / top-5 acc by fine-tuning a pre-trained ResNet-50 **[different methods for CL] (right)**

- **Linear evaluation protocol** :
  self-supervised learning (Pre-training) → feature extractor (freeze) + Linear classifier training → linear evaluation using validation set.

Experiments (Linear evaluation with different objectives)

	ImageNet-100		CIFAR-100		CIFAR-10		CovType		Higgs-100K	
Method	Base	Hard	Base	Hard	Base	Hard	Base	Hard	Base	Hard
CPC	79.7	81.1(+1.4)	65.4	67.1(+1.7)	91.7	91.9(+0.2)	71.6	74.3(+2.7)	64.7	71.3(+6.6)
MLCPC	79.5	81.2(+1.7)	65.6	66.6(+1.0)	91.9	92.1(+0.2)	71.7	74.1(+2.4)	64.9	71.5(+6.6)
RMLCPC	78.9	**81.6(+2.7)**	64.5	**68.5(+4.0)**	90.7	**92.5(+1.8)**	72.1	**74.9(+2.8)**	64.5	**72.4(+7.9)**

Top-1 linear evaluation acc of unsupervised representation learning on image dataset **[Contrastive objective]**

Experiments (Mutual information estimation)

- Recall **the convexity of KL divergence** :

$$D_{KL}(\lambda P_1 + (1-\lambda)P_2 \parallel \lambda Q_1 + (1-\lambda)Q_2] \leq \lambda D_{KL}(P_1 \parallel Q_1) + (1-\lambda)D_{KL}(P_2 \parallel Q_2)$$

- By plugging $P_1 = P_2 = Q_1 = P,\ Q_2 = Q$, We get following inequality : ($\alpha \in (0,1)$)

$$D_{KL}^{(\alpha)}(P \parallel Q) \leq (1-\alpha)D_{KL}(P \parallel Q) < D_{KL}(P \parallel Q)$$

- Similarly, **the convexity of Rényi divergence** is proved : ($\gamma \in [0,\infty]$)

$$R_\gamma(\lambda P_1 + (1-\lambda)P_2 \parallel \lambda Q_1 + (1-\lambda)Q_2) \leq \lambda R_\gamma(P_1 \parallel Q_1) + (1-\lambda)R_\gamma(P_2 \parallel Q_2)$$

- Hence, We get similar inequality : ($\alpha \in (0,1)$)

$$R_\gamma^{(\alpha)}(P \parallel Q) \leq (1-\alpha)R_\gamma(P \parallel Q) < R_\gamma\,(P \parallel Q)$$

Experiments (Mutual information estimation)

- And recall the **optimal critic** f^* for α-CPC / α-MLCPC / (α, γ)-RMLCPC :

$$f^* = \log\left(\frac{Z \cdot P_{XY}(x,y)}{\alpha P_{XY}(x,y) + (1-\alpha)P_X(x)P_Y(y)}\right) = \log\left(\frac{Z \cdot r(x,y)}{\alpha r(x,y) + 1 - \alpha}\right)$$

where $r(x,y) = \frac{P_{XY}}{P_X P_Y}$ (true density ration), Z = log-normalization constant

- By modifying above equation we get following :

$$\int_{\mathcal{X}} \int_{\mathcal{Y}} e^{f^*(x,y)} \big(\alpha P_{XY}(x,y) + (1-\alpha)P_X(x)P_Y(y)\big) dx dy = Z$$

- Here, we can estimate Z using **MC approximation** on batch environment of size B:

$$\hat{Z} = \frac{\alpha}{B} \sum_{i=1}^{B} e^{f^*(x_i, y_i)} + \frac{1-\alpha}{B(B-1)} \sum_{i=1}^{B} \sum_{j \neq i} e^{f^*(x_i, y_j)}$$

Experiments (Mutual information estimation)

- Then, we can approximate **true density ratio $\hat{\boldsymbol{r}}$** as follows :

$$\hat{r}(x,y) = \frac{(1-\alpha)e^{f^*(x,y)}}{\hat{Z} - \alpha e^{f^*(x,y)}}$$

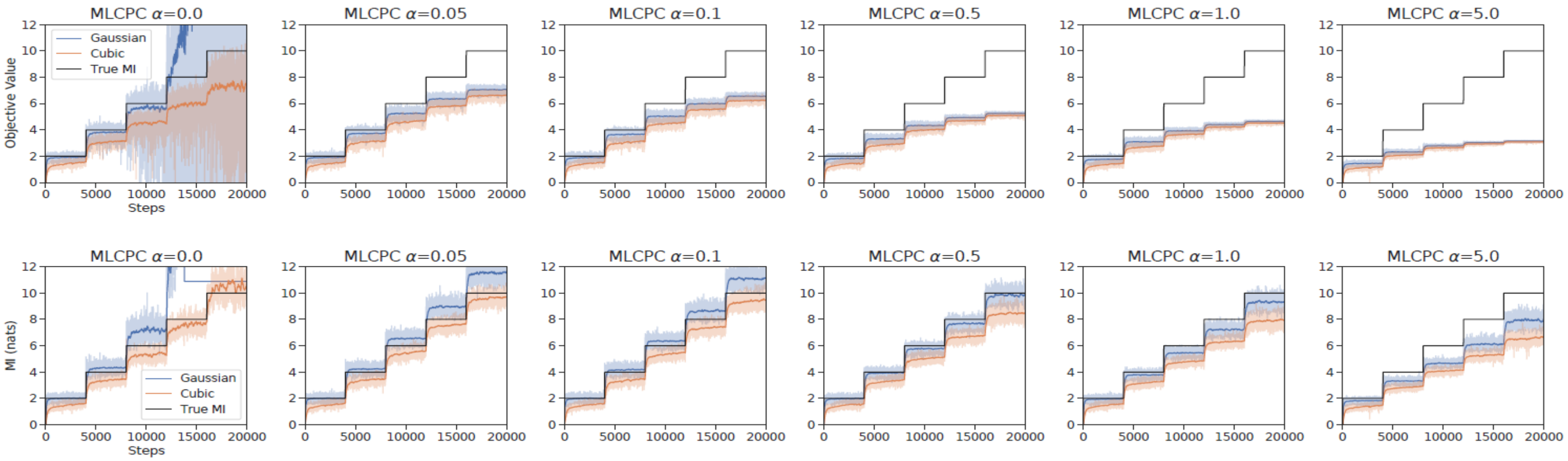
- Finally, using estimated $\hat{r}$, and formula that $I(X;Y) = \mathbb{E}_{P_{XY}}\left[\log \frac{P_{XY}}{P_X P_Y}\right] = \mathbb{E}_{P_{XY}}\ [\log r]$:

$$\hat{I}(X;Y) = \frac{1}{B}\sum_{i=1}^{B} \log \hat{r}(x_i, y_i)$$

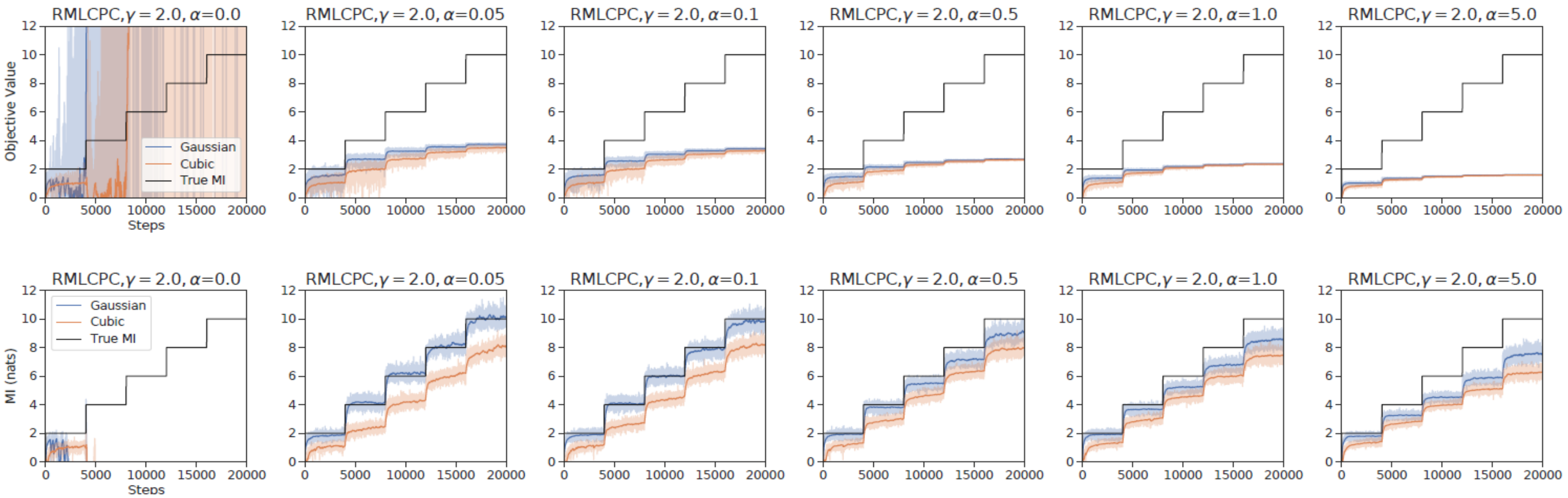
- Obviously, larger batch size results in much more accurate estimate for $I(X;Y)$

- Under **standard correlated gaussian experiments** using joint critic, we get following results

Experiments (Mutual information estimation)



Top : α-MLCPC objective \ bottom : MI estimation from α-MLCPC



Top : (α, γ)-RMLCPC objective \ bottom : MI estimation from (α, γ)-RMLCPC