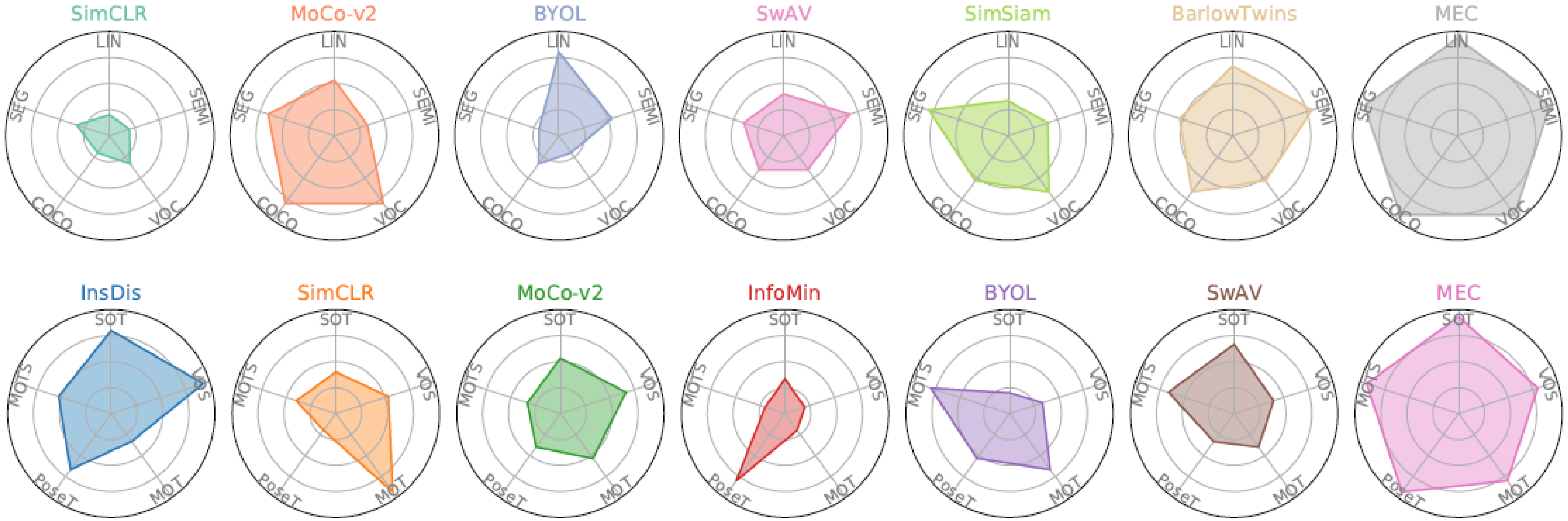
Self-Supervised Learning via Maximum Entropy Coding

-Summary-

Introduction

- Current problem of SSL :
  - Famous SSL methods (SimCLR, SwAV, MoCo-v2, …) suffers from biases into the learned representation.
    - Ex : Learned representations with image-level task (image classification) show not good performance in patch or pixel-level tasks (object detection, semantic segmentation)



Comparison of transfer learning performance on 5 image-based tasks (top) and 5-video based tasks (bottom)

Maximum Entropy Coding

- Q1 : what makes for generalizable representations?
- Q2 : what is the optimization + criterion that directly measures the structure of representations with the aim of minimizing the biases brought by the pretext task?
- → Answer on paper : **Maximum Entropy Coding**

- [Background for rate distortion function] (d = feature dimension, m = # of samples)
  - Given a set of sample $Z = [z^1, \ldots, z^m] \in \mathbb{R}^{d \times m}$, **the minimum # of bits needed to encode Z subject to a distortion upper bound ϵ** : (when $\mathbb{E}_{p(z)}[\|z - \hat{z}\|_2] \leq \epsilon$)

$$L = \left(\frac{m+d}{2}\right) \log\det\left(I_m + \frac{d}{m\epsilon^2} Z^T Z\right) = \left(\frac{m+d}{2}\right) \log\det\left(I_d + \frac{d}{m\epsilon^2} Z Z^T\right)$$

  - *(In fact, this is the approximated upper bound of rate distortion function* ***when $\boldsymbol{Z \sim MN(0, \Sigma)}$*** *and especially m is sufficiently large.)*

Segmentation of Multivariate Mixed Data via Lossy Data Coding and Compression[Ma et al.,] : rate distortion function theory

Maximum Entropy Coding

- But, calculating det on high-dimensional matrix $Z^T Z$ is intractable, especially from ill-condition property of $Z^T Z$.

- By rewriting L as $L = \mu \log \det(I_m + \lambda Z^T Z)$, where $\mu = \frac{m+d}{2}$, $\lambda = \frac{d}{m\epsilon^2}$ and using some identical equation $[\det(\exp(A)) = \exp\big(Tr(A)\big)]$, we can apply **Taylor expansion to get** **approximated L** :

$$L = Tr\left(\mu \sum_{k=1}^{\infty} \frac{(-1)^{k+1}}{k} \left(\lambda Z^T Z\right)^k\right), \qquad \text{convergence condition}: \left\|\lambda Z^T Z\right\|_2 < 1$$

- By maximizing L, we can achieve well-generalized representation in terms of required # of bits to encode Z within upper bound of error ϵ.

Maximum Entropy Coding

- However, this optimization will lead to trivial solutions such as uniform distribution if we do not impose the model to have certain level of accuracy (or performance).

- One method is to exploit the data augmentation (often adopted in current SSL method).
  - Authors suggest to replace empirical covariance matrix ZZ^T to correlation matrix $Z_1 Z_2^T$ to enforce contrastive learning effects <ins>*(motivation seems not clear in theoretical view)*</ins> (where Z_1, Z_2 are two views from Z) :

$$\mathcal{L}_{MEC} = -\mu \log\det\left(\boldsymbol{I_m} + \lambda \boldsymbol{Z}_1^\top \boldsymbol{Z}_2\right) \approx -\operatorname{Tr}\left(\mu \sum_{k=1}^{n} \frac{(-1)^{k+1}}{k} \left(\lambda \boldsymbol{Z}_1^\top \boldsymbol{Z}_2\right)^k\right)$$

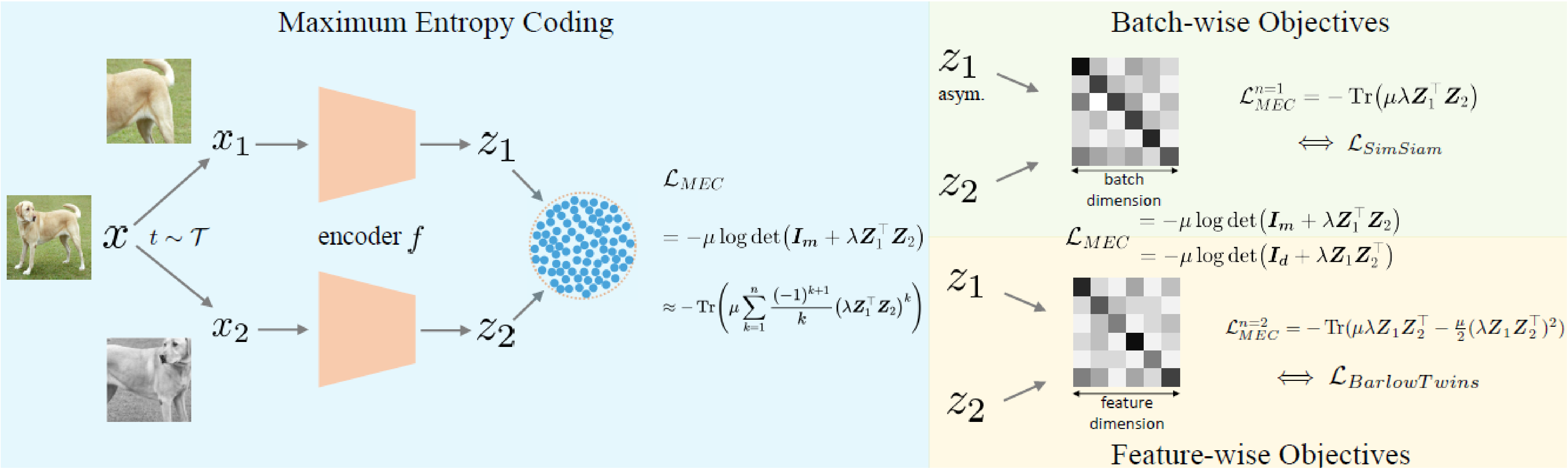
  - However, we can observe view consistency effect on L_{MEC} loss.
    <ins>***(through $Z_1^T Z_2$ terms in $Tr(\cdot)$)***</ins>

Maximum Entropy Coding

- Note that we can easily show two version of L_{MEC} (batch-wise / feature-wise):

$$\mathcal{L}_{MEC} = \underbrace{-\mu \log\det\left(\boldsymbol{I_m} + \lambda \boldsymbol{Z}_1^\top \boldsymbol{Z}_2\right)}_{\text{batch-wise}} = \underbrace{-\mu \log\det\left(\boldsymbol{I_d} + \lambda \boldsymbol{Z}_1 \boldsymbol{Z}_2^\top\right)}_{\text{feature-wise}}$$

- Illustration of MEC method :



Maximum Entropy Coding

- In fact, by adopting this replacement ($ZZ^T \rightarrow Z_1Z_2^T$), they can achieve similar objectives such as SimSiam, Barlow Twins :
  (Note : InfoNCE is also similar, but not exact due to $\log \exp(\cdot)$ in InfoNCE term)

1. SimSiam (similar to 1st order Taylor approximation of L_{MEC} (batch-wise) :

$$\mathcal{L}_{SimSiam} = -\sum_{i=1}^{m} z_1^i \cdot z_2^i \quad \Longleftrightarrow \quad \mathcal{L}_{MEC}^{n=1} = -\operatorname{Tr}\left(\mu\lambda \boldsymbol{Z}_1^\top \boldsymbol{Z}_2\right) = -\mu\lambda \sum_{i=1}^{m} z_1^i \cdot z_2^i$$

   where μ and λ can be absorbed into learning rate.

2. Barlow (similar to 2nd order Taylor approximation of L_{MEC} (feature-wise) :

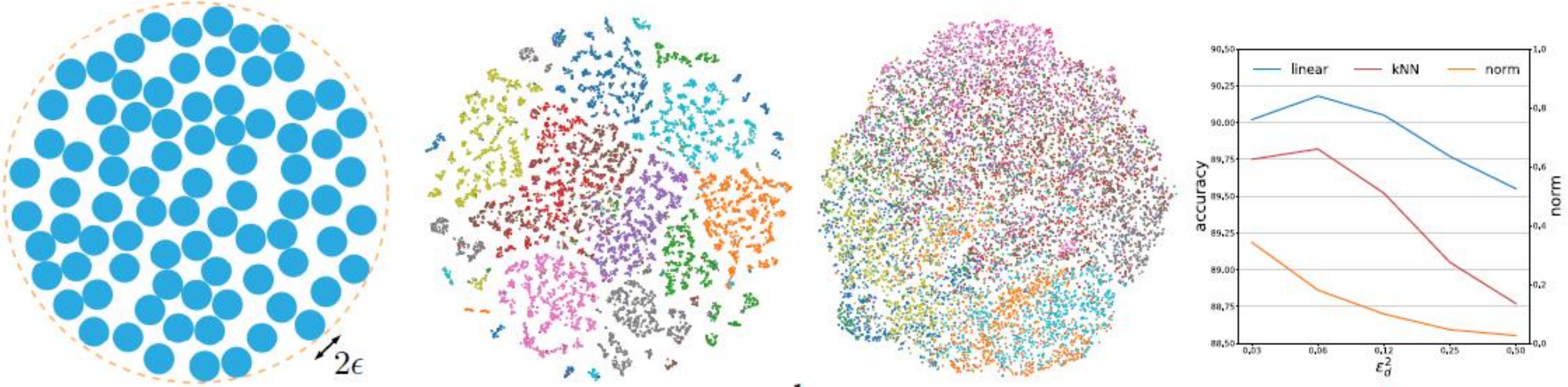
$$\mathcal{L}_{Barlow} = \sum_{i=1}^{d} (1 - \boldsymbol{C}_{ii})^2 + \lambda_{barlow} \sum_{i=1}^{d} \sum_{j \neq i}^{d} \boldsymbol{C}_{ij}{}^2 \quad \Longleftrightarrow \quad \mathcal{L}_{MEC}^{n=2} = -\operatorname{Tr}\left(\mu\lambda \boldsymbol{Z}_1 \boldsymbol{Z}_2^\top - \frac{\mu}{2}\left(\lambda \boldsymbol{Z}_1 \boldsymbol{Z}_2^\top\right)^2\right)$$

$$= \mu \sum_{i=1}^{d} \left(-\boldsymbol{C}_{ii} + \frac{1}{2}\boldsymbol{C}_{ii}{}^2\right) + \frac{\mu}{2} \sum_{i=1}^{d} \sum_{j \neq i}^{d} \boldsymbol{C}_{ij}{}^2$$

   where $C = \lambda Z_1 Z_2^T$

Maximum Entropy Coding – Supplement

- What is the meaning of ϵ in representation learning?:
  - The upper-bound of distances to distinguish different samples well (in l_2 distance)
  - Higher ϵ will result in representation learning robust to gaussian noise
  - Smaller ϵ will result in representation learning based on train samples.



a b c

Effect of the distortion measure ϵ on MEC

(a) : Encoding the representation is akin to packing ϵ-ball in representation space.

(b) T-SNE of the learned representation (left : large ϵ, right : small ϵ)

(c) Linear and kNN accuracy and the spectral norm w.r.t the degree of distortion ϵ (norm analysis for convergence of Taylor expansion)

Maximum Entropy Coding – Thoughts

- What we get from maximizing L_{MCE}?
  - Under satisfying accuracy, we achieve representation having (almost) largest entropy.
  - Similar approaches in some papers :
    - EX : To remove simplicity bias (such as classify only using color information), some paper used models which have largest data size (in bytes) by zipping the model parameters.
  - Maximizing entropy of learned representation may be beneficial for removing simplicity bias, which results in good overall downstream performance suggested on this paper.

- One idea : According to [Ma et al., 2007], we can further well estimate rate distortion using mixture of gaussian model assumption :

Can we use identity + Taylor approximation here?

$$L = \sum_{j=1}^{k} \frac{tr(\Pi_j)}{2m} \log_2 \det\left(I + \frac{n}{\epsilon^2 tr(\Pi_j)} Z \Pi_j Z^T\right)$$

$k = \#\ of\ classes$
$m = \#\ of\ samples$
$\Pi_j = membership\ martix\ of\ class\ j$

Maximum Entropy Coding – Thoughts

- The crucial assumption that $z \sim MN(0, \Sigma)$ seems not satisfactory.
  - We don't have guarantee that $z = f_\theta(x) \sim MN(0, \Sigma)$.
  - We usually assume that input data will follow mixture of multivariate normal distribution, this will collapse on complex data (CIFAR-10, ImageNet)

- How about adding regularization term to enforce distribution on representation approximately to follow MN?
  - Such as $KL(p(z) \| mixture\ of\ MN(0, \Sigma))$
  - Although there is no closed form for mixture of multivariate normal, we can approximate them using lower / upper bound of it. (not checked tractability …) [JL Durrieu et al., 2012], [JR Hershey., 2007]