Squeezing Water from a Stone: Improving Pre-Trained SSL Embeddings Through Effective Entropy Maximization Criterion (E2MC)

Deep Chakraborty[1]
Yann LeCun[2,3]
Tim G. J. Rudner[2]
Erik Learned-Miller[1]

[1] UMassAmherst
Manning College of Information & Computer Sciences
dchakraborty@umass.edu

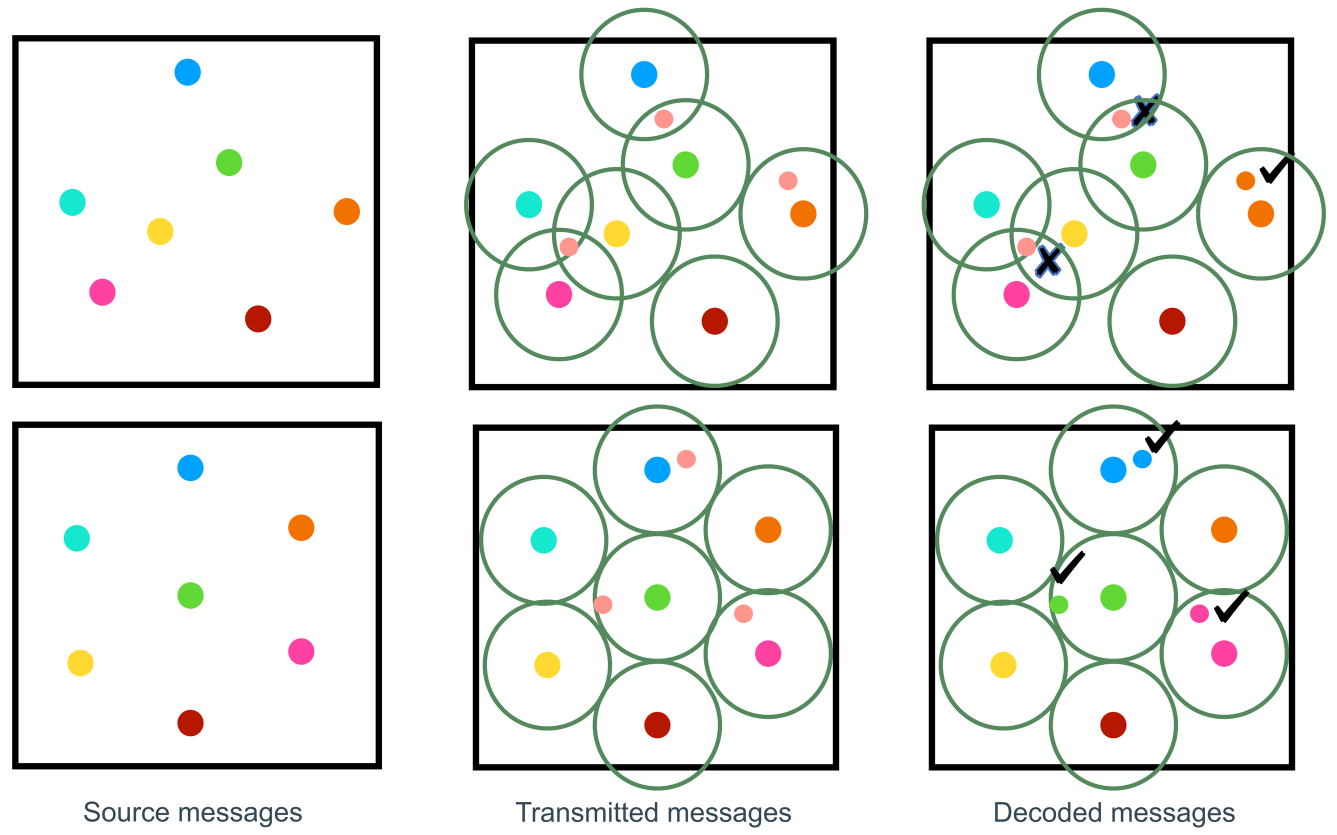
[2] NYU

[3] Meta

arXiv

What Constitutes *Good* Embeddings?

- Embeddings with **maximum entropy** preserve the most amount of information about the inputs.
- By maximizing the amount of information retained, we can hope to do well on future discrimination tasks when they are *unknown*.
- An information-theoretic viewpoint:



So What's the Problem?

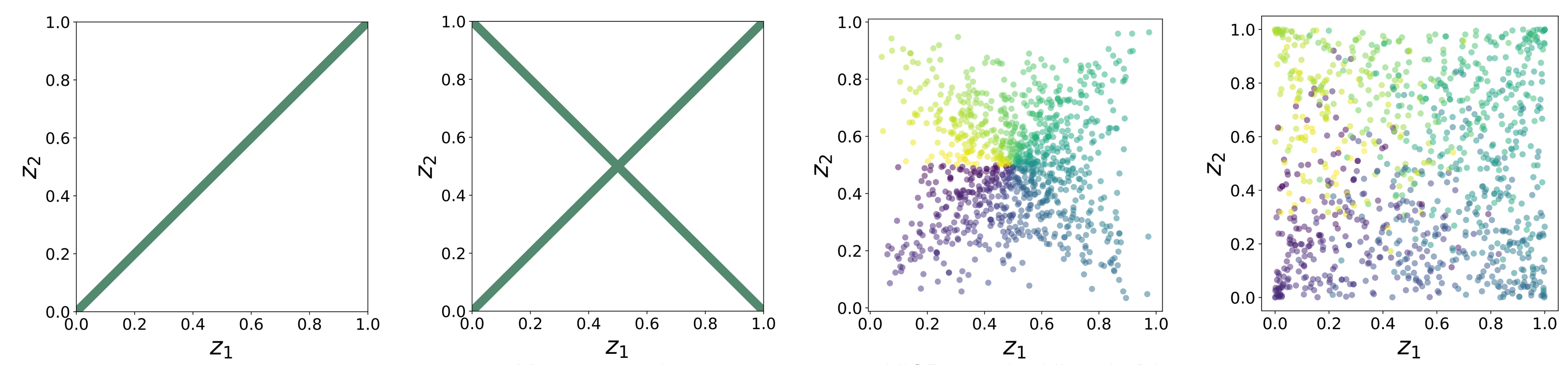
- No direct access to the embedding distribution p(z), so we must use finite amount of samples for entropy estimation, which grows exponentially with number of dimensions.
  - Can we find constraints for which we **have sufficient data?**
- SSL models are already highly optimized and their performance is close to saturation, so it is challenging to improve them further!
  - Can we find a **model agnostic criterion** which can be used to improve pre-trained models using a handful of epochs?

What Have Others Tried?

- Alignment and Uniformity on the Hypersphere (**AUH**) [1]
  - Distribute points uniformly on the hypersphere by minimizing the energy configuration of points using pairwise potentials.
  - **Limitation**: Operates on samples of the high dimensional joint distribution!
- Approximate Log-Determinant Maximization (**CorInfoMax** [2], **VICReg** [3])
  - Maximize the spread of the latent vectors in embedding space by using the log determinant of the covariance matrix as an approximation of the mutual information between input views.
  - **Limitation**: Gaussian distribution assumption!

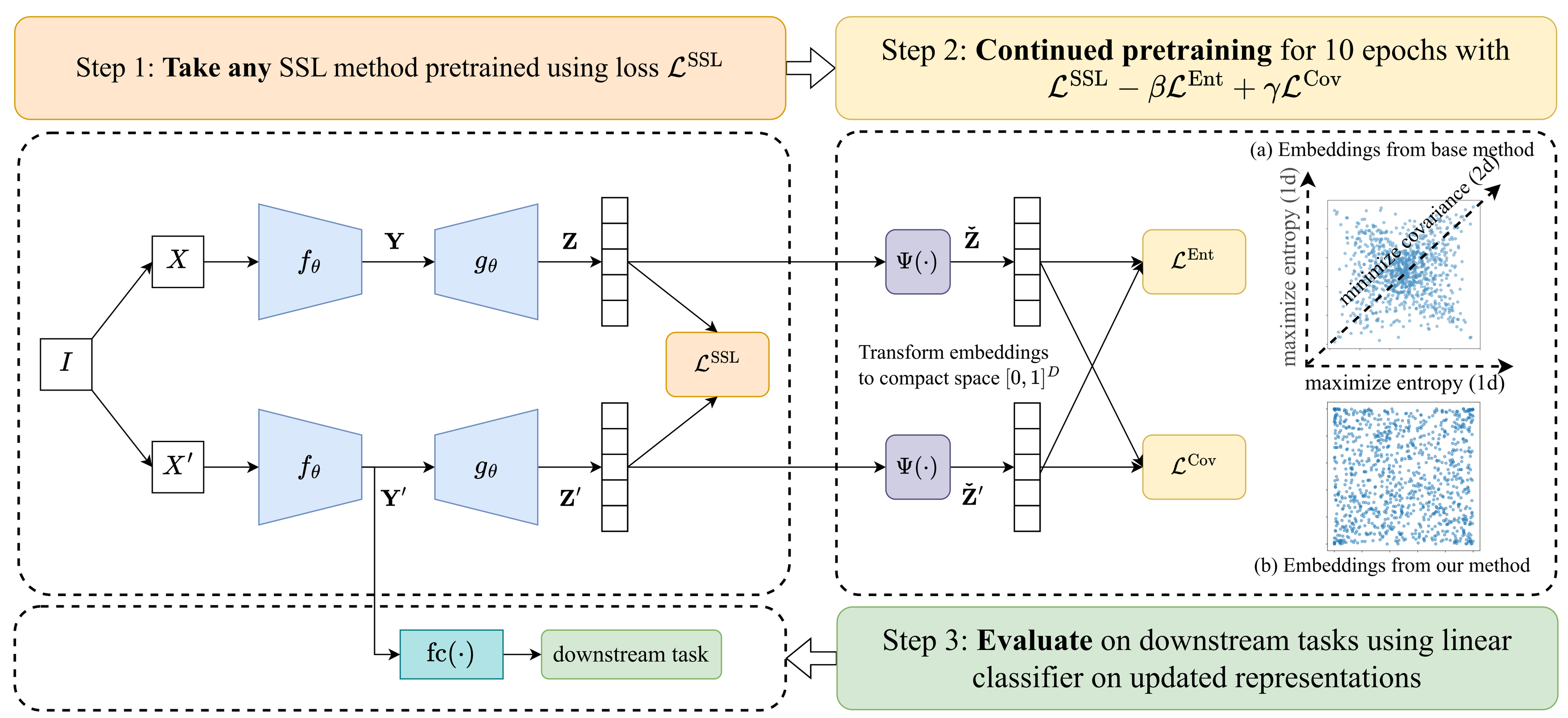
Low-Dimensional Statistics: The Hero That SSL Needs, but Not the One It Deserves

1. The **one-dimensional entropy** of each marginal component of our embeddings.
2. The **covariance of all pairs** of marginals.



Max-ent marginals, Perfect correlation — Max-ent marginals, No correlation (really?) — VICReg embeddings in 2d (Baseline) — VICReg + E2MC in 2d (Ours)

Talk is Cheap, Show Me the Model!



The m-spacings estimate of differential entropy [7]

$$\mathcal{L}^{\text{Entropy}}(\theta) = \frac{1}{d}\sum_{j=1}^{d}\left(\widehat{\mathcal{H}}_j(\check{Z}_\theta^1, \cdots, \check{Z}_\theta^n) + \widehat{\mathcal{H}}_j(\check{Z}_\theta'^1, \cdots, \check{Z}_\theta'^n)\right)$$

where

$$\widehat{\mathcal{H}}_j(\check{Z}_\theta^1, ..., \check{Z}_\theta^n) = \frac{1}{(n-m)}\sum_{i=1}^{n-m}\log\left(\frac{n+1}{m}\left(\check{Z}_{\theta j}^{(i+m)} - \check{Z}_{\theta j}^{(i)}\right)\right)$$

and $\check{Z}_{\theta j}^{(1)} \leq \check{Z}_{\theta j}^{(2)} \leq \cdots \leq \check{Z}_{\theta j}^{(n)}$

Sample covariance estimator [3]

$$\mathcal{L}^{\text{Covariance}}(\theta) = \frac{1}{nd}\left(\|(K_\theta - \text{diag}(K_\theta))\|_F^2 + \|(K'_\theta - \text{diag}(K'_\theta))\|_F^2\right)$$

where $K_\theta = \check{\check{Z}}_\theta^\top \check{\check{Z}}_\theta$ and $\check{\check{Z}}_\theta = \check{Z}_\theta - \frac{1}{n}\sum_{i=1}^{n}\check{Z}_\theta^i$

Does It *Really* Help?

Short answer: Yes. Long Answer: It *really* does.

Table 1: Evaluation of self-supervised embeddings. Top-1-Accuracy / mAP under different paradigms on the base (ImageNet) and other datasets.

	Linear Evaluation			Semi-supervised Learning		Transfer Learning	
Method (checkpoint)	1% labels	10% labels	100% labels	1% labels	10% labels	iNat18	VOC07
VICReg base [3] (1000 ep)	53.50 ±0.11	66.57 ±0.02	73.20†	54.53* ±0.12	67.97* ±0.03	47.00†	86.60†
VICReg continued (1010 ep)	53.51 ±0.07	66.57 ±0.06	73.16 ±0.02	–	–	–	–
VICReg+ E2MC **[ours]** (1010 ep)	**54.54** ±0.05	**66.82** ±0.05	**73.45** ±0.07	**55.05** ±0.08	**68.12** ±0.04	**47.18** ±0.11	**86.80**
SwAV base [4] (400 ep)	52.34 ±0.07	67.61 ±0.02	74.30†	52.57 ±0.15	69.25 ±0.05	46.00	**88.38**
SwAV continued (410 ep)	52.31 ±0.07	67.56 ±0.05	74.31 ±0.02	–	–	–	–
SwAV+ E2MC **[ours]** (410 ep)	**53.40** ±0.01	**67.73** ±0.03	**74.44** ±0.03	52.70 ±0.54	69.24 ±0.02	**46.71** ±0.17	88.24
SwAV base [4] (800 ep)	53.70 ±0.05	68.86 ±0.03	75.30†	53.89† ±0.13	70.22† ±0.05	49.08*	88.56*
SwAV continued (810 ep)	53.69 ±0.05	68.87 ±0.04	75.32 ±0.01	–	–	–	–
SwAV+ E2MC **[ours]** (810 ep)	**55.27** ±0.07	**68.98** ±0.02	**75.41** ±0.02	53.94 ±0.30	70.32 ±0.05	**49.72** ±0.20	**88.69**
SimSiam base [5] (100 ep)	43.71 ±0.04	60.15 ±0.02	68.37*	–	–	38.75	**84.62**
SimSiam continued (110 ep)	43.78 ±0.05	60.23 ±0.08	68.45 ±0.08	–	–	–	–
SimSiam+ E2MC **[ours]** (110 ep)	43.78 ±0.06	60.23 ±0.07	**68.52** ±0.05	–	–	**38.99** ±0.20	84.54

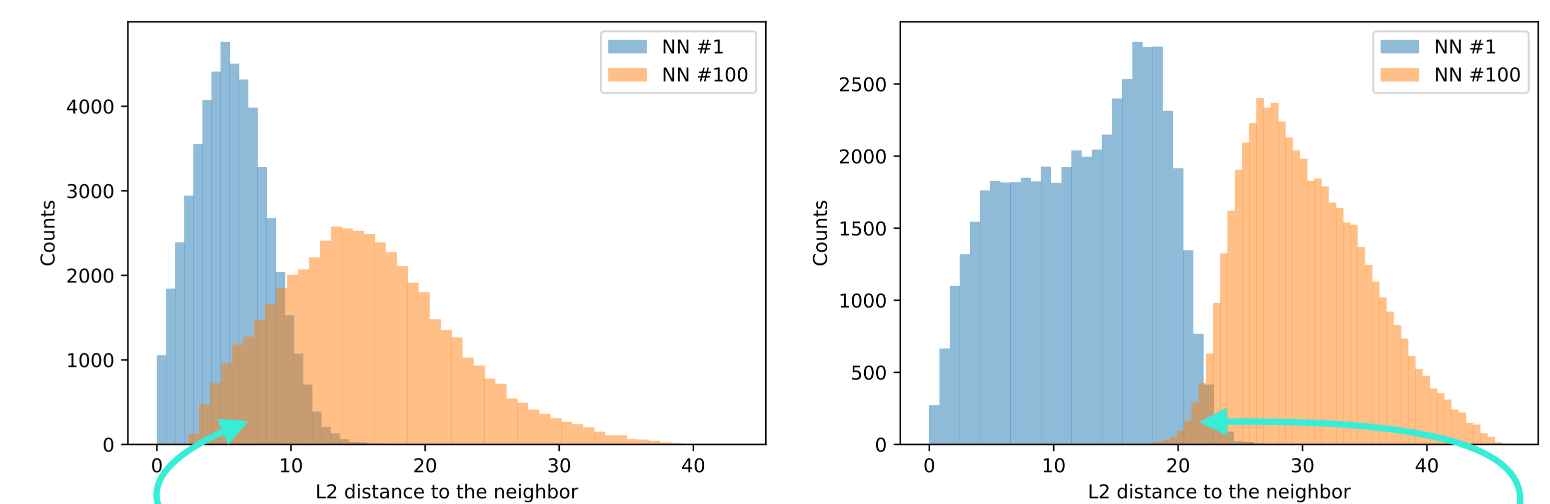
But What About That *Other* Method?

Table 2: Top-1-Accuracy of linear classifier trained on ImageNet

Method	1% labels	10% labels	100% labels
SwAV base [4]	53.70 ±0.05	68.86 ±0.03	75.30†
SwAV continued	53.69 ±0.05	68.87 ±0.04	75.32 ±0.01
SwAV + VCReg [3]	54.02 ±0.05	68.88 ±0.03	75.36 ±0.02
SwAV + MMCR [6]	53.30 ±0.02	68.77 ±0.04	75.27 ±0.01
SwAV + AUH [1]	53.84 ±0.07	68.90 ±0.04	75.33 ±0.01
SwAV + E2MC **[ours]**	**55.27** ±0.07	**68.98** ±0.02	**75.41** ±0.02

What Else Can We Show?

Embedding separability under different criteria



VICReg embeddings show high overlap signifying that the 100th nearest neighbor of a point is often closer than the 1st nearest neighbor of another point.

After continued pre-training with E2MC criterion, the overlap is significantly reduced, showing better separability for downstream tasks.

Why Should *You* Care?

- **Squeeze** the most out of your SSL model!
  - With our model agnostic plug-and-play criterion, you could get improved performance from your model, especially if you have limited data, or you care about generalization to unseen tasks.
  - Help your SSL model converge faster using our criterion.
- Continued pre-training is *relatively* **inexpensive**!
  - You can adapt off-the-shelf models with ResNet-50 backbone using our criterion in **under 10 hrs** using 2 RTX-8000 GPUs.
- **Fundamental research** into properties of large-scale SSL models.
  - Do these methods work for LLMs? You can find out!

References

[1] Wang & Isola (ICML 2020). "Understanding contrastive representation learning through alignment and uniformity on the hypersphere".
[2] Ozsoy et al. (NeurIPS 2022). "Self-supervised learning with an information maximization criterion".
[3] Bardes et al. (ICLR 2022). "VICReg: variance-invariance-covariance regularization for self-supervised learning".
[4] Caron et al. (NeurIPS 2020). "Unsupervised learning of visual features by contrasting cluster assignments".
[5] Chen & He. (CVPR 2021). "Exploring simple Siamese representation learning".
[6] Yerxa et al. (NeurIPS 2023). "Learning efficient coding of natural images with maximum manifold capacity representations".
[7] Learned-Miller & Fisher (JMLR 2003). "ICA using spacings estimates of entropy".