Assessing Neural Network Representations During Training Using Noise-Resilient Diffusion Spectral Entropy

Presenter: Chen Liu
Nov 2023

Outline

- ❑ Motivation
- ❑ Background
- ❑ Methods
  - ❑ Definition of Diffusion Spectral Entropy (DSE)
  - ❑ Definition of Diffusion Spectral Mutual Information (DSMI)
  - ❑ Propositions and Properties
- ❑ Experiments & Results
  - ❑ Toy test cases for DSE and DSMI
  - ❑ DSMI at very high dimension
  - ❑ Computational Efficiency
  - ❑ Evolution along neural network training
  - ❑ Utility Study: Network Initialization Experiment for DSE
  - ❑ Utility Study: ImageNet cross-model correlation

Outline

Motivation

- ***Entropy*** and ***mutual information*** in neural networks provide rich information on the learning process.
- But they are historically **difficult to compute when the dimension is high** due to curse of dimensionality.
- We leverage diffusion geometry to access the underlying manifold and reliably compute these information-theoretic measures.

Outline

Krishnaswamy Lab
Yale University

Background

- ***Entropy***

Shannon $$H(X) = \mathbb{E}[-\log p(X)] = -\sum_{x \in X} p(x) \log p(x)$$

von Neumann $$H(\rho) = -tr(\rho \log \rho) = -\sum_i \eta_i \log \eta_i$$

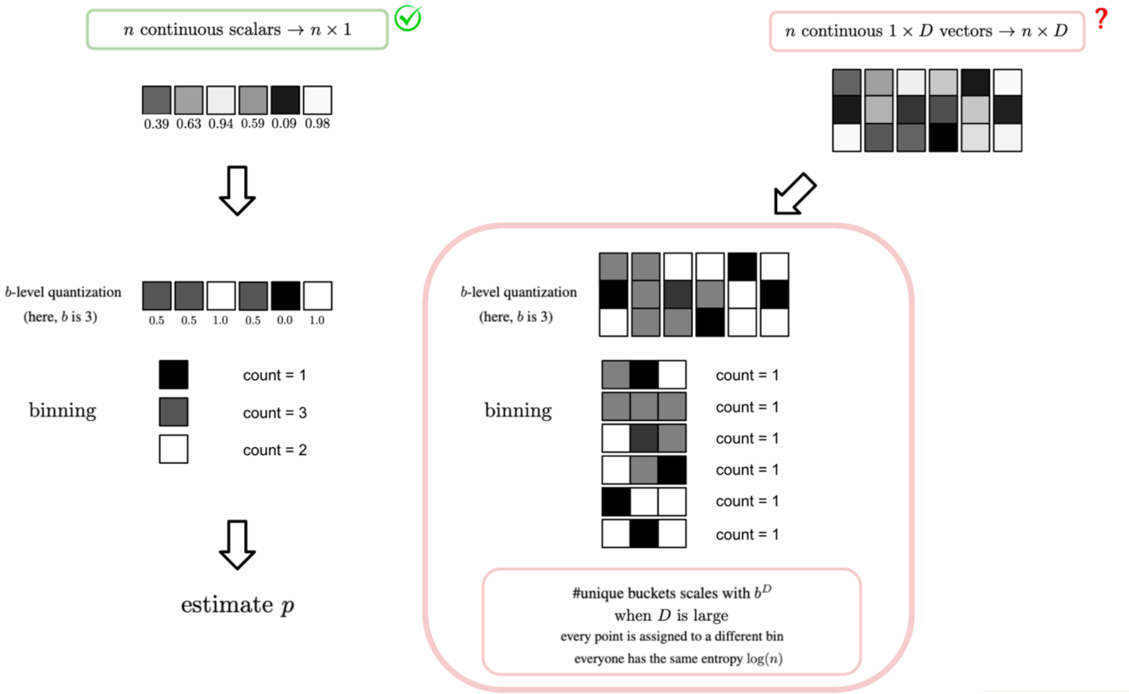
- ***Mutual Information***

$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) \\ &= H(X) - \sum_i p(Y = y_i) H(X|Y = y_i) \end{aligned}$$

Background

- Classic method is binning + quantization.

entropy & mutual information: some form of $-\sum p \log p \rightarrow$ how to estimate p?



Background

- Diffusion geometry

Diffusion Map

$$\mathcal{K}(z_1, z_2) = \frac{\mathcal{G}(z_1, z_2)}{\|\mathcal{G}(z_1, \cdot)\|_1^\alpha \|\mathcal{G}(z_2, \cdot)\|_1^\alpha}, \text{ where}$$

$$\mathcal{G}(z_1, z_2) = e^{-\frac{\|z_1 - z_2\|^2}{\sigma}}$$

$$\mathbf{P}_{i,j} = p(z_i, z_j) = \frac{\mathcal{K}(z_1, z_2)}{\|\mathcal{K}(z_1, \cdot)\|_1}$$

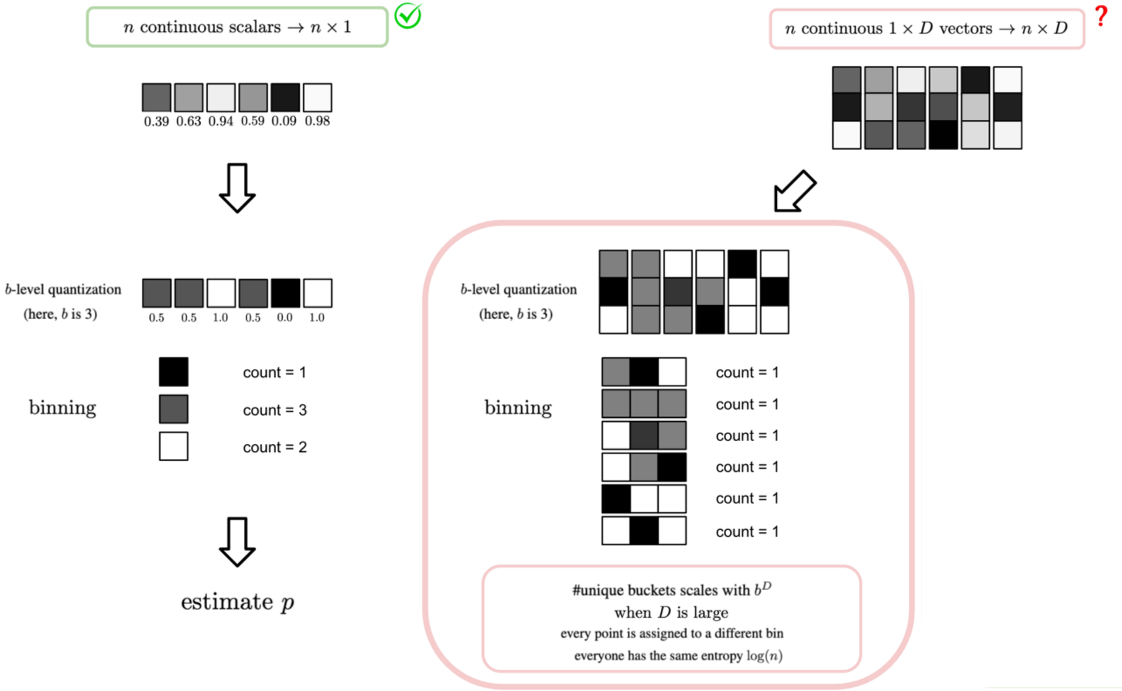
Outline

- Motivation
- Background
- Methods
  - Definition of Diffusion Spectral Entropy (DSE)
  - Definition of Diffusion Spectral Mutual Information (DSMI)
  - Propositions and Properties
- Experiments & Results
  - Toy test cases for DSE and DSMI
  - DSMI at very high dimension
  - Computational Efficiency
  - Evolution along neural network training
  - Utility Study: Network Initialization Experiment for DSE
  - Utility Study: ImageNet cross-model correlation

Methods

- Classic method is binning + quantization.

entropy & mutual information: some form of $-\sum p \log p$ $\rightarrow$ how to estimate p?

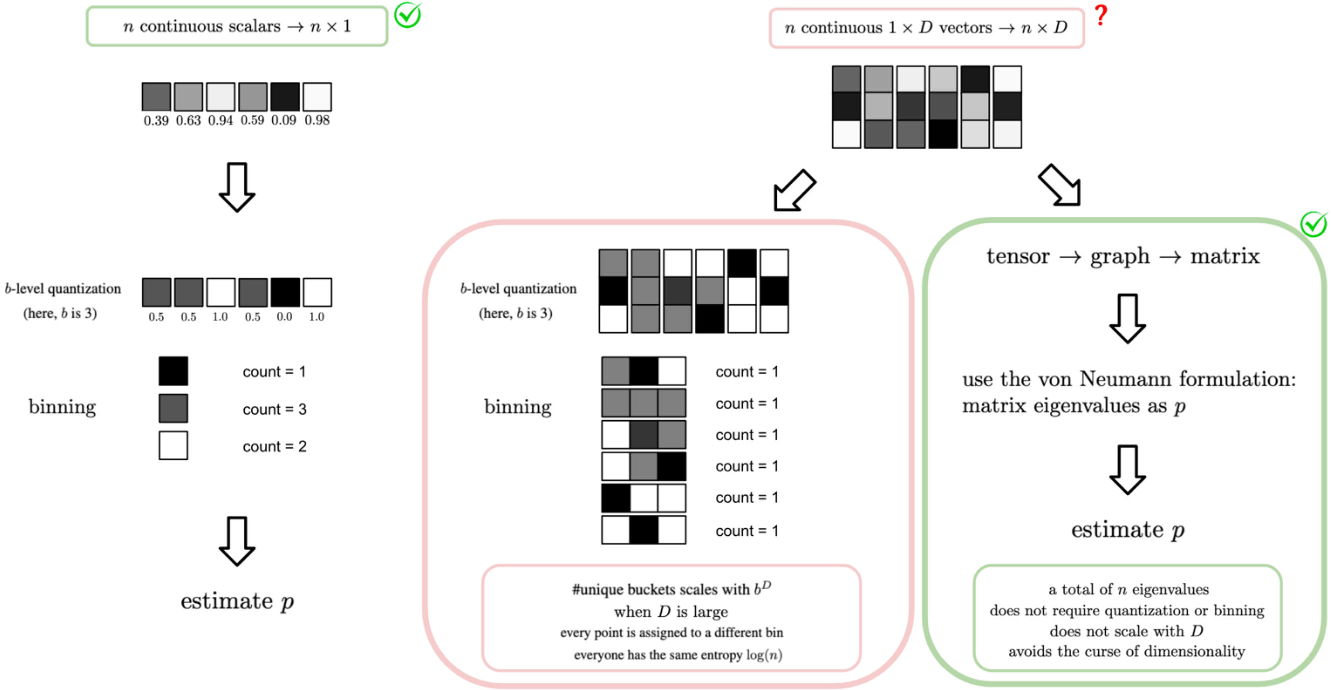


Methods

Krishnaswamy Lab
Yale University

- Our method: use diffusion geometry.

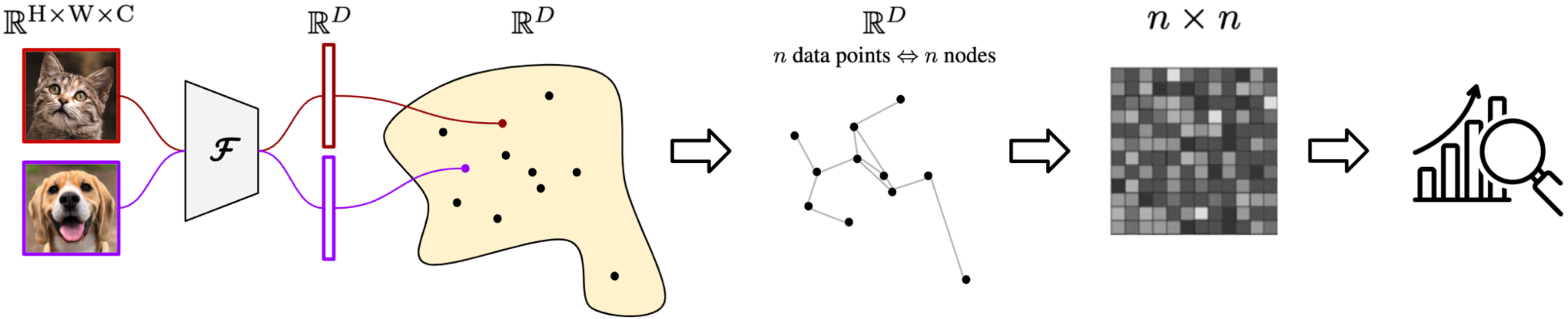
entropy & mutual information: some form of $-\sum p \log p$ $\rightarrow$ how to estimate p?



Methods

Krishnaswamy Lab
Yale University

- Our method: use diffusion geometry.



Methods (definitions)

Definition 4.1. *We define* ***Diffusion Spectral Entropy (DSE)*** *as an entropy of the eigenvalues of the diffusion operator* $\mathbf{P}_X$ *computed on a dataset* X *where* $x \in X$ *is a multidimensional vector* $[x_1, x_2 \dots x_d]^T$*:*

$$S_D(\mathbf{P}_X, t) := -\sum_i \alpha_{i,t} \log(\alpha_{i,t}) \qquad (7)$$

where $\alpha_{i,t} := \frac{|\lambda_i^t|}{\sum_j |\lambda_j^t|}$*, and* $\{\lambda_i\}$ *are the eigenvalues of the diffusion matrix* $\mathbf{P}_X$.

Methods (definitions)

Krishnaswamy Lab
Yale University

Definition 4.2. *We define* ***Diffusion Spectral Mutual Information (DSMI)*** *as the difference between conditional and unconditional diffusion spectral entropy*

$$I_D(X;Y) = S_D(\mathbf{P}_X, t) - \sum_{y_i \in Y} p(Y = y_i) S_D(\mathbf{P}_{X|Y=y_i}, t) \quad (8)$$

Methods (propositions)

- First, we provide the lower bound and the upper bound of DSE when t -> inf, and we explain the conditions when they are reached.

Proposition 4.1. *S_D achieves a minimal entropy of 0 when the diffusion operator defines an ergodic Markov chain, and is in steady state (as $t \to \infty$).*

Proposition 4.2. *As $t \to \infty$, $S_D(\boldsymbol{P}_X, t)$ on data with k well-separated clusters is $\log(k)$.*

- 4.1 implies that if all data points are very similar, i.e., have the same probability of transitioning to any other point, then it has minimal entropy.
- 4.2 shows that DSE will reach its maximum value when the points are spread out very far apart.

Methods (propositions)

Krishnaswamy Lab
Yale University

- Next, we examine the expected value of DSE.

Proposition 4.3. *Let $X \in \mathbb{R}^{n \times d}$ be a dataset of n independent and identically distributed multivariate Gaussian vectors in $\mathbb{R}^d$, where $x_i \sim \mathcal{N}(0, I_d)$. Then, using K as defined in Eqn 1 with $\alpha = 1/2$,*

$$\mathbb{E}[S_D(\mathbf{P}_X, t=1)] \lessapprox \log(\frac{n}{1-\beta}) - \left(\frac{1}{n} + \left(\frac{n-1}{n}\right)\beta\right)\log\left(1 + \frac{\beta n}{1-\beta}\right)$$

$$\text{where } \beta = \left(1 + \frac{4}{\sigma}\right)^{-\frac{d}{2}}$$

$$\mathcal{K}(z_1, z_2) = \frac{\mathcal{G}(z_1, z_2)}{\|\mathcal{G}(z_1, \cdot)\|_1^{\alpha} \|\mathcal{G}(z_2, \cdot)\|_1^{\alpha}}, \text{ where}$$

$$\mathcal{G}(z_1, z_2) = e^{-\frac{\|z_1 - z_2\|^2}{\sigma}}$$

This establishes a theoretical upper bound on the DSE at any given layer.

Also reinforces that for large d, beta is close to 0, so DSE <= log(n).

Methods (propositions)

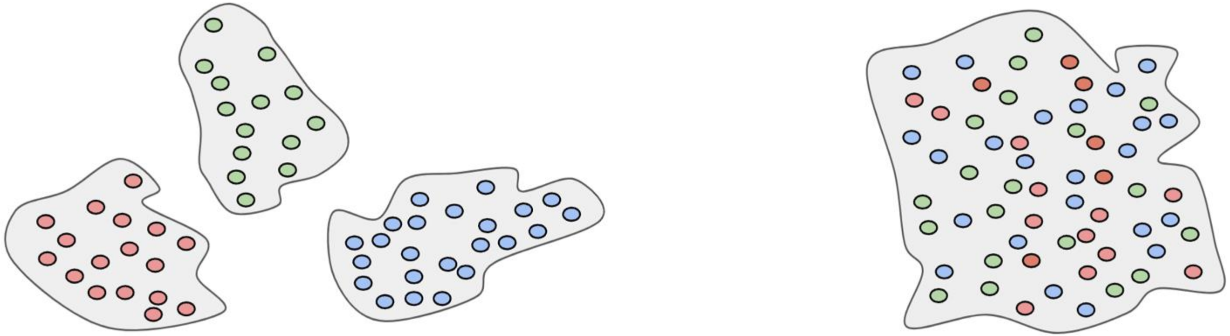
- Finally, we investigate the entropy progression in neural network training.

Proposition 4.4. *Take n to be arbitrarily large. Let $X \in \mathbb{R}^{n \times d}$ be a matrix of i.i.d. random values $x_{ij} \sim f$. Let $Y \in \mathbb{R}^{n \times d}$ be a matrix of i.i.d. random values $y_{ij} \sim f$, but in $k \in \mathbb{N}$ distinct clusters such that when the anisotropic probability matrix is computed for $\alpha = 1/2$, the probability of diffusion between points of different clusters is arbitrarily small. Then, using β as defined in Proposition 4.3, the approximate upper bound on DSE increases by $\beta \log(k)$.*

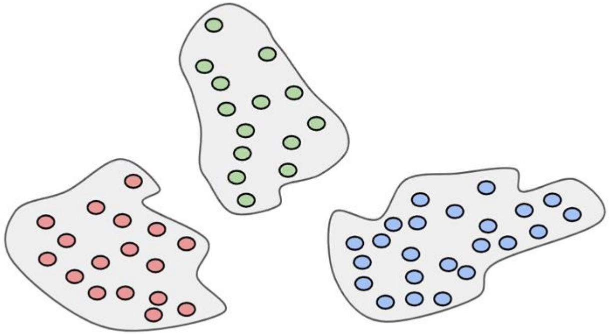
Recall the training process of a classification neural network. During training, the embeddings will spread out into different clusters. This proposition suggests that the upper bound of DSE will increase along the training process.

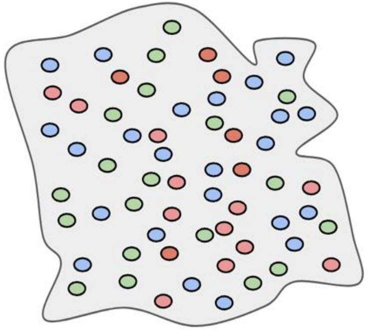
Outline

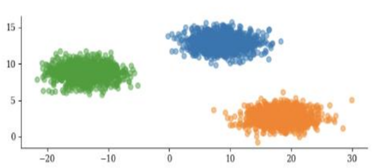
Results (Intuition)

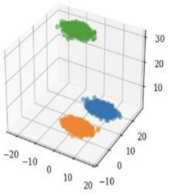


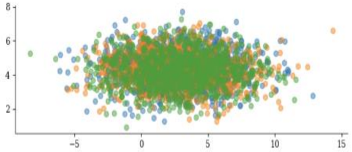
Results (Intuition)

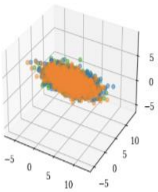






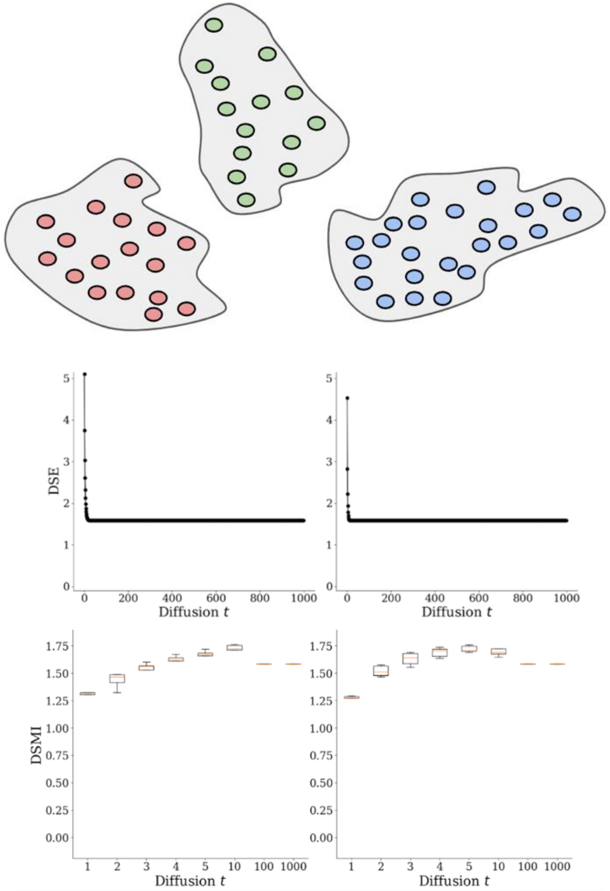


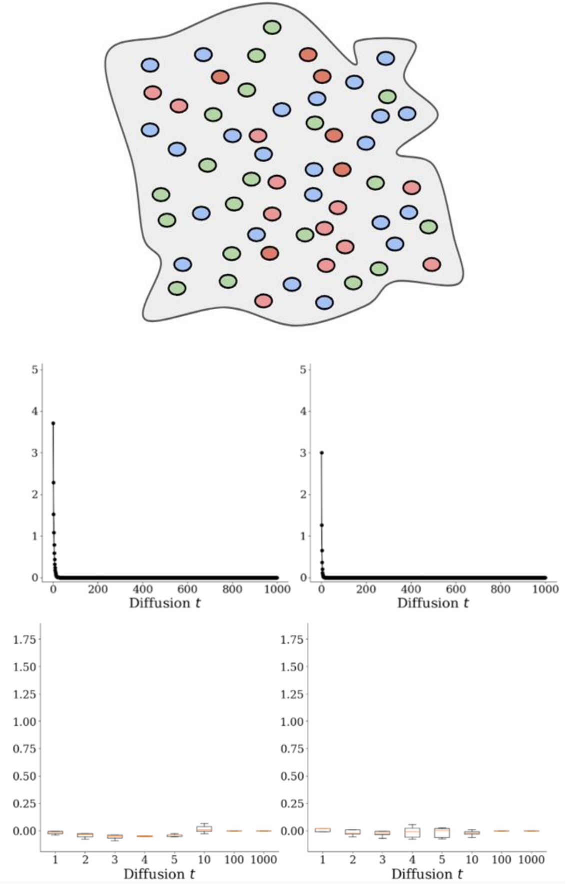




Results (Intuition)

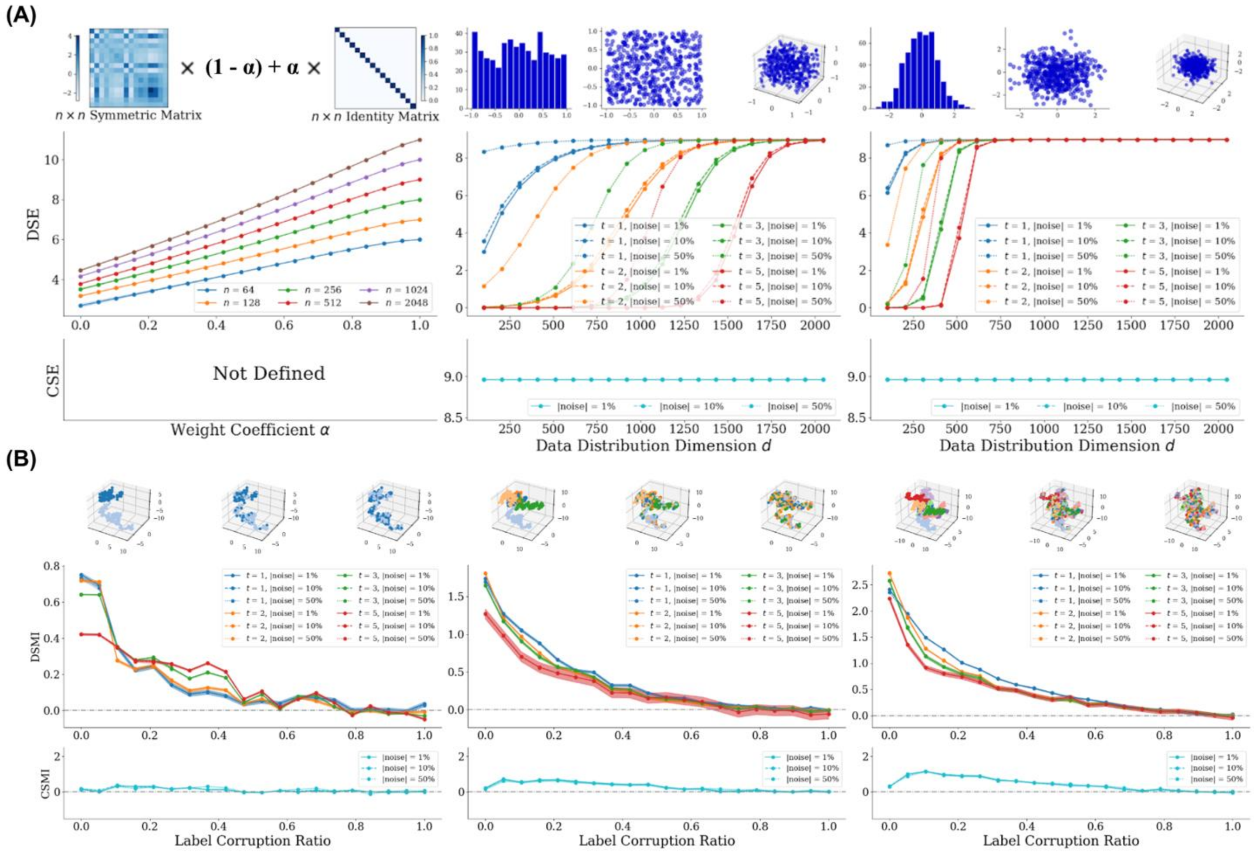
Krishnaswamy Lab
Yale University





Results (Verify Trending)

Krishnaswamy Lab
Yale University



- (A) DSE increases as intrinsic dimension grows, while CSE does not capture this trend due to curse of dimensionality.
- (B) When two random variables are dependent, DSMI negatively correlates with the level of data corruption, while CSMI does not capture this trend. DSMI I D(Z; Y) and CSMI are computed on synthetic, 20-dimensional trees with { 2, 5, 10 } branches (Left, Mid, Right).

Results (DSMI at very high dimension)

Krishnaswamy Lab
Yale University

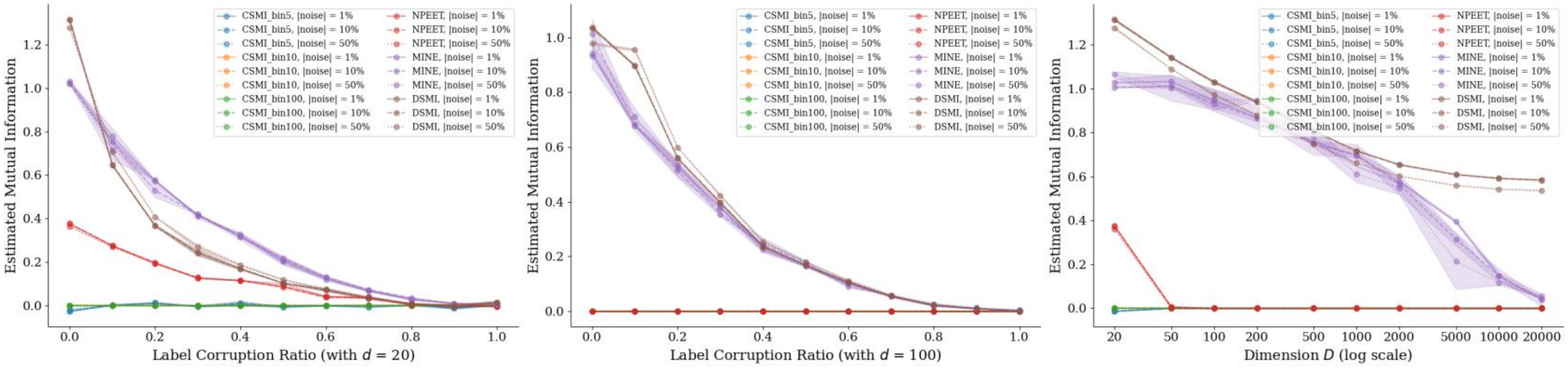


Figure: Mutual information estimation on toy Gaussian blobs

- All methods generally obey the expected behavior
- (3rd panel) CSMI, NPEET and MINE fail as the dimension increases beyond 10,000, while DSMI still remains significant.

Results (Computational Efficiency)

Krishnaswamy Lab
Yale University

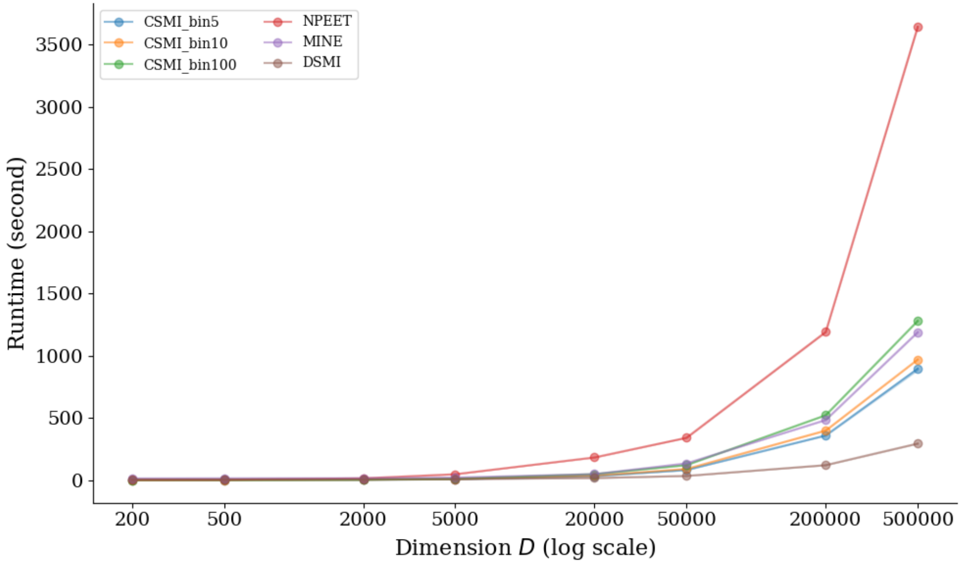


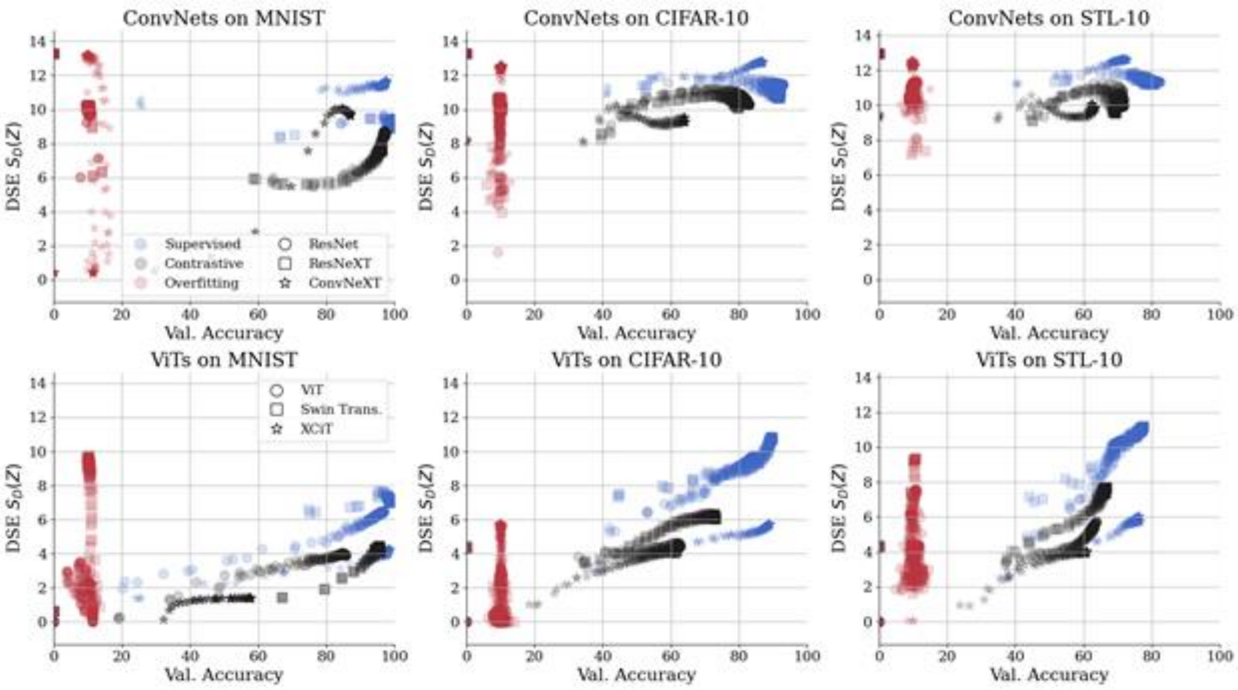
Figure: DSMI scales better than other methods at high dimensions.

Results (Evolution along neural network training)

Experimented with

- **6 models**: 3 ConvNets, 3 Vision Transformers
- **3 learning settings**: supervised, self-supervised and nonsense overfitting.
- **3 datasets**: MNIST, CIFAR-10, and STL-10.
- **3 random seeds**

DSE of embedding vectors



- DSE(Z) generally increases as models perform better in proper learning.

Figure: Diffusion Spectral Entropy DSE(Z) of embedding vector Z.

DSMI with output

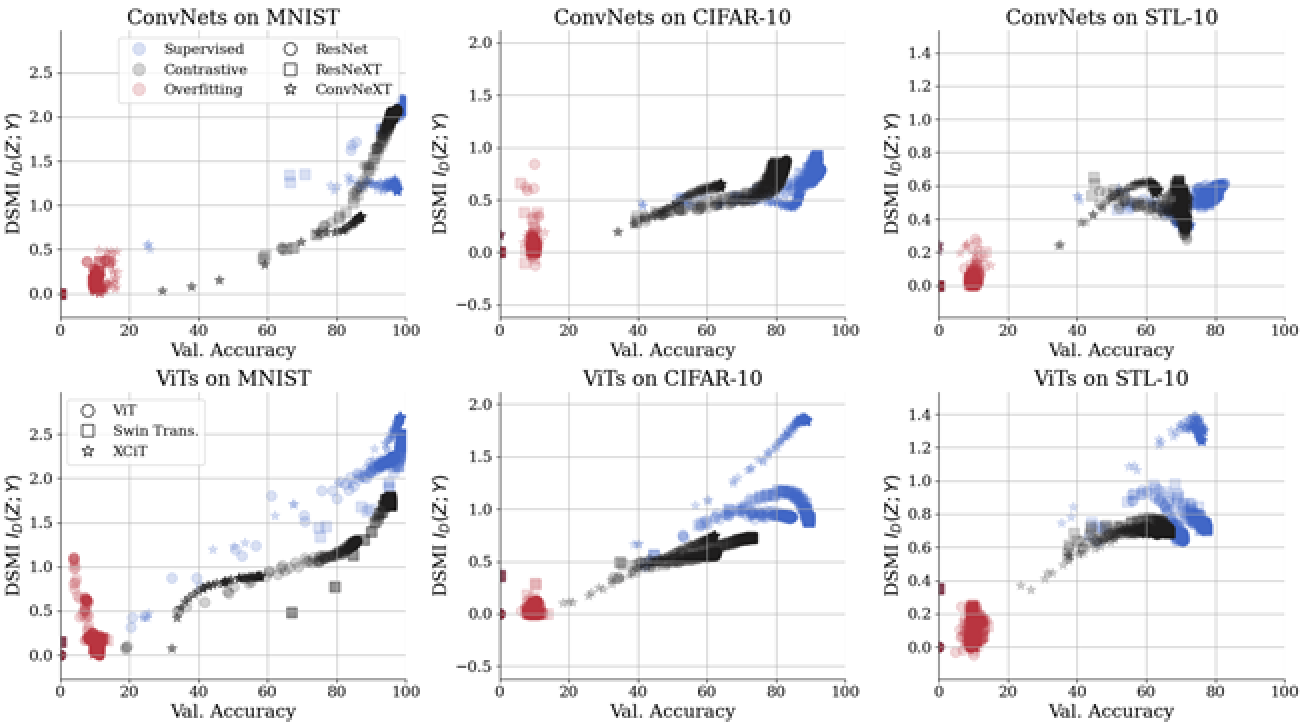
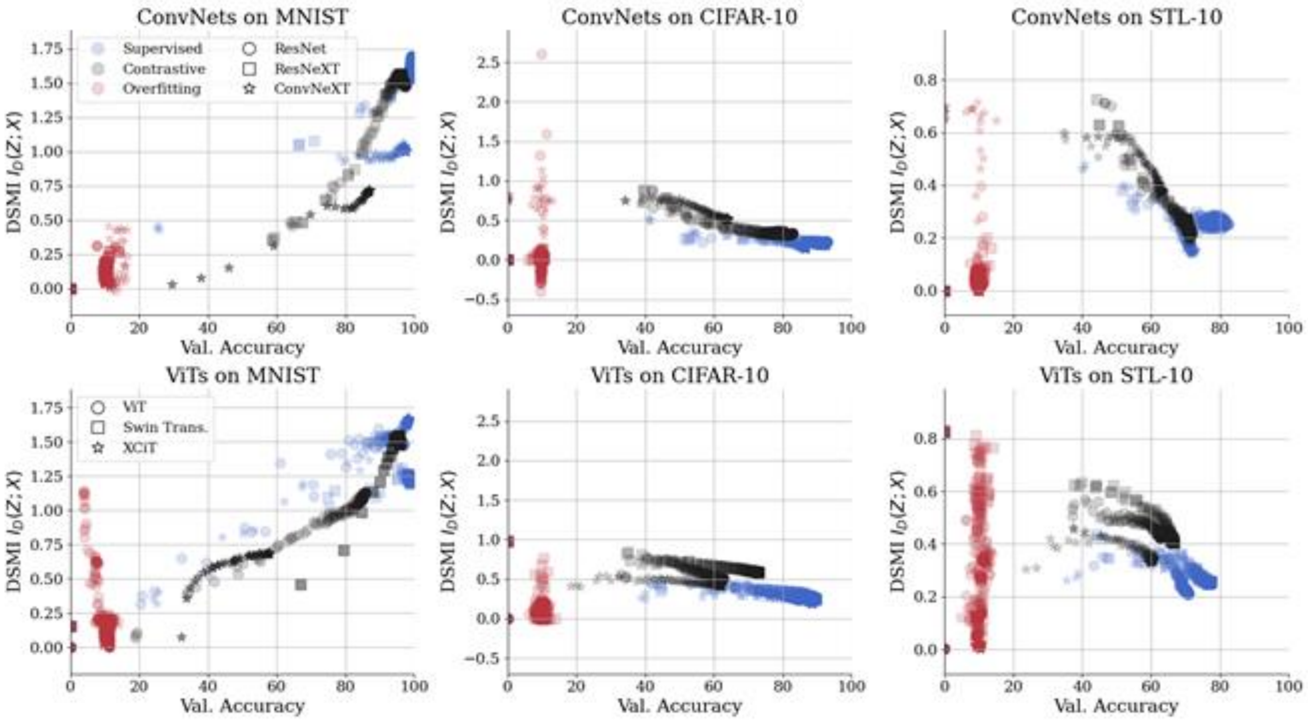


Figure: Diffusion Mutual Information DSMI(Z; Y) between embedding vector Z and class label Y.

- DSMI(Z;Y) consistently increases in proper learning.
- DSMI climbs more slowly in contrastive learning compared to supervised learning and ends up at a lower terminal value.
- In nonsense memorization, DSMI quickly converges to around zero.

DSMI with input

Krishnaswamy Lab
Yale University



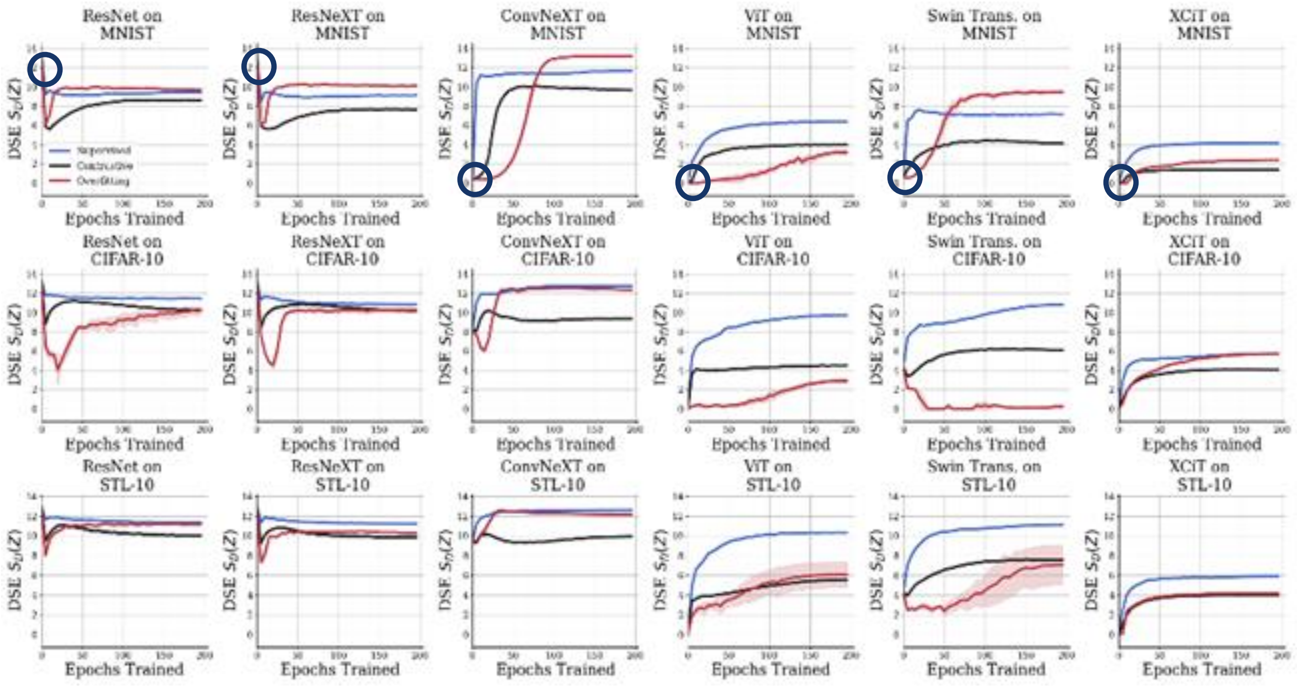
- DSMI(Z; X) keeps increasing during learning on the MNIST dataset.
- DSMI(Z; X) mostly decreasing on the CIFAR-10 and STL-10 datasets.
- In nonsense memorization, DSMI(Z; X) rises to a significant level in most cases
- In contrast to information bottleneck theory

Figure: Diffusion Spectral Mutual Information DSMI(Z; X) between embedding vector Z and input X.

Outline

Network Initialization Experiment for DSE

Krishnaswamy Lab
Yale University



- **Observation 1**: Even under the same initialization code, some networks starts with low DSE while others starts with high DSE.
- **Observation 2**: If starting at low DSE, DSE will increase monotonically. If starting at high DSE, DSE will decrease first and then increase.

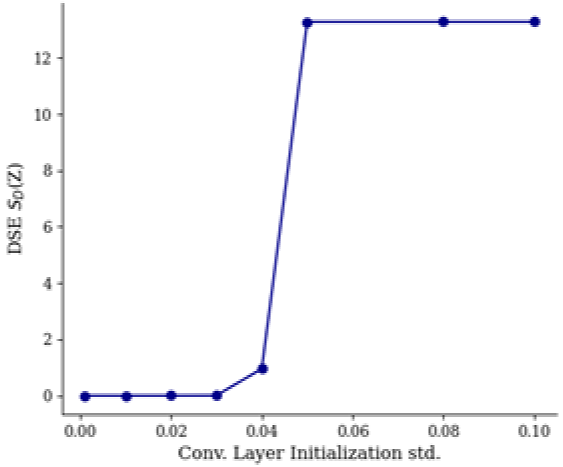
- **Question**: Will initial DSE affect the training dynamics?

Network Initialization Experiment for DSE

Krishnaswamy Lab
Yale University

Will initializing the network at a high DSE vs. a low DSE affect the learning process?

- Initializing convolutional layers with a normal distribution with a mean of 0 and a tunable standard deviation.

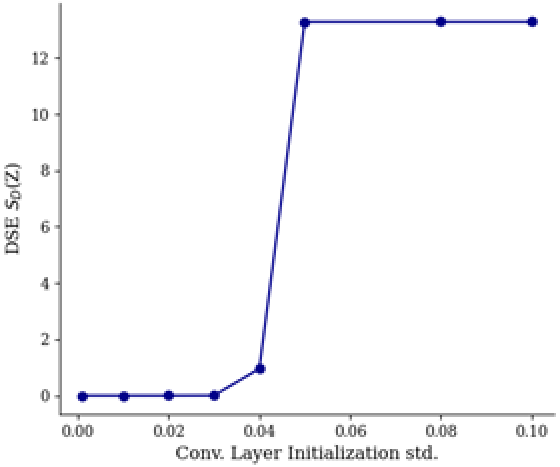


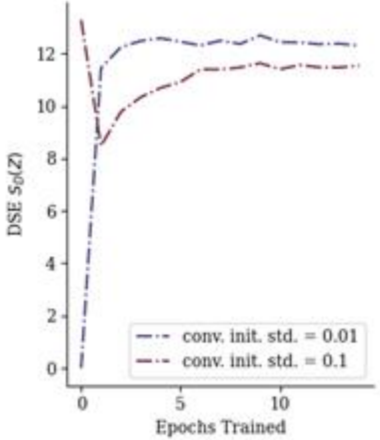
Network Initialization Experiment for DSE

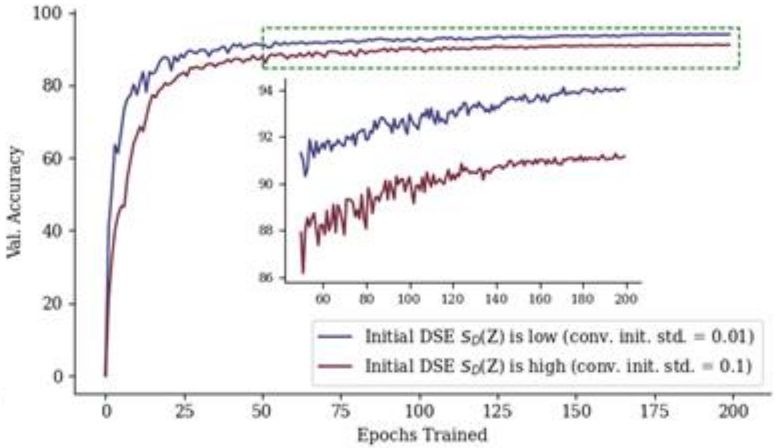
Krishnaswamy Lab
Yale University

Will initializing the network at a high DSE vs. a low DSE affect the learning process?

- ❑ Initializing convolutional layers with a normal distribution with a mean of 0 and a tunable standard deviation.
- ❑ **Initializing the network at a low DSE allows faster convergence and better final performance.**





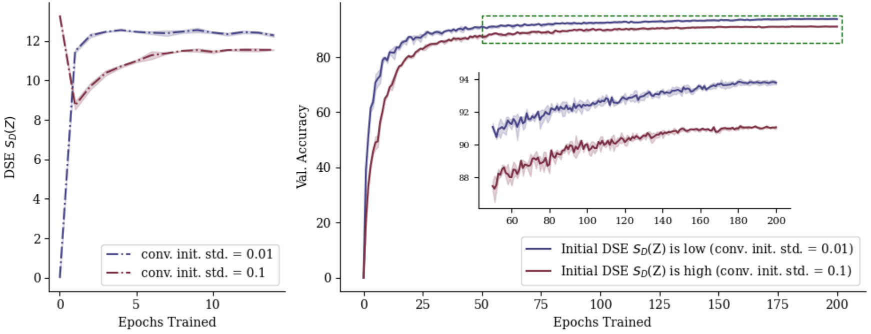


Network Initialization Experiment for DSE

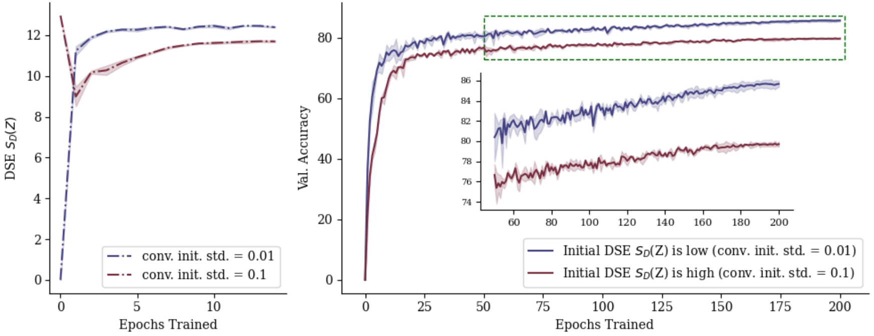
Krishnaswamy Lab
Yale University

Will initializing the network at a high DSE vs. a low DSE affect the learning process?

- ❑ Initializing convolutional layers with a normal distribution with a mean of 0 and a tunable standard deviation.
- ❑ **Initializing the network at a low DSE allows faster convergence and better final performance.**



CIFAR-10



STL-10

Network Initialization Experiment for DSE

Krishnaswamy Lab
Yale University

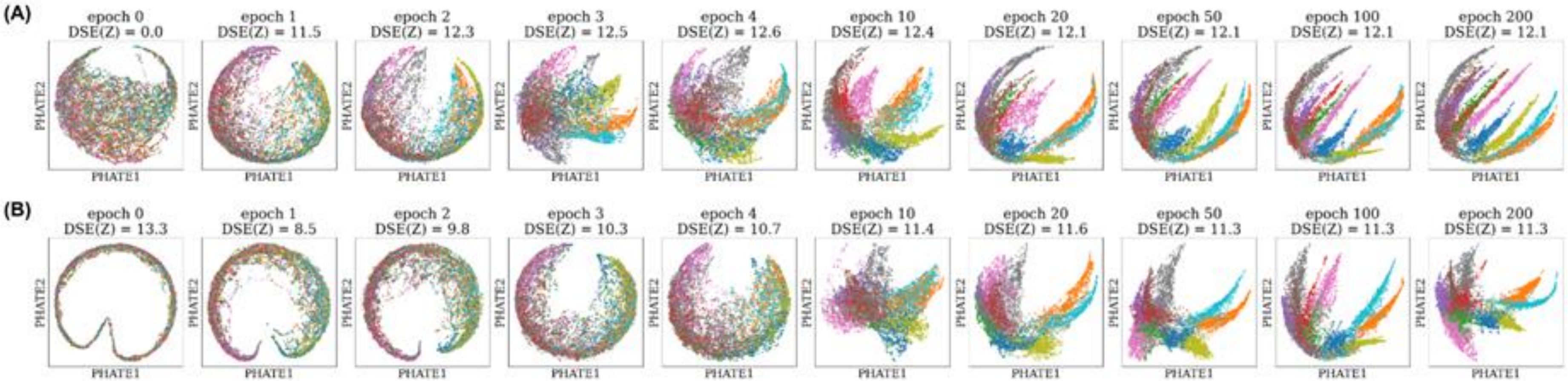
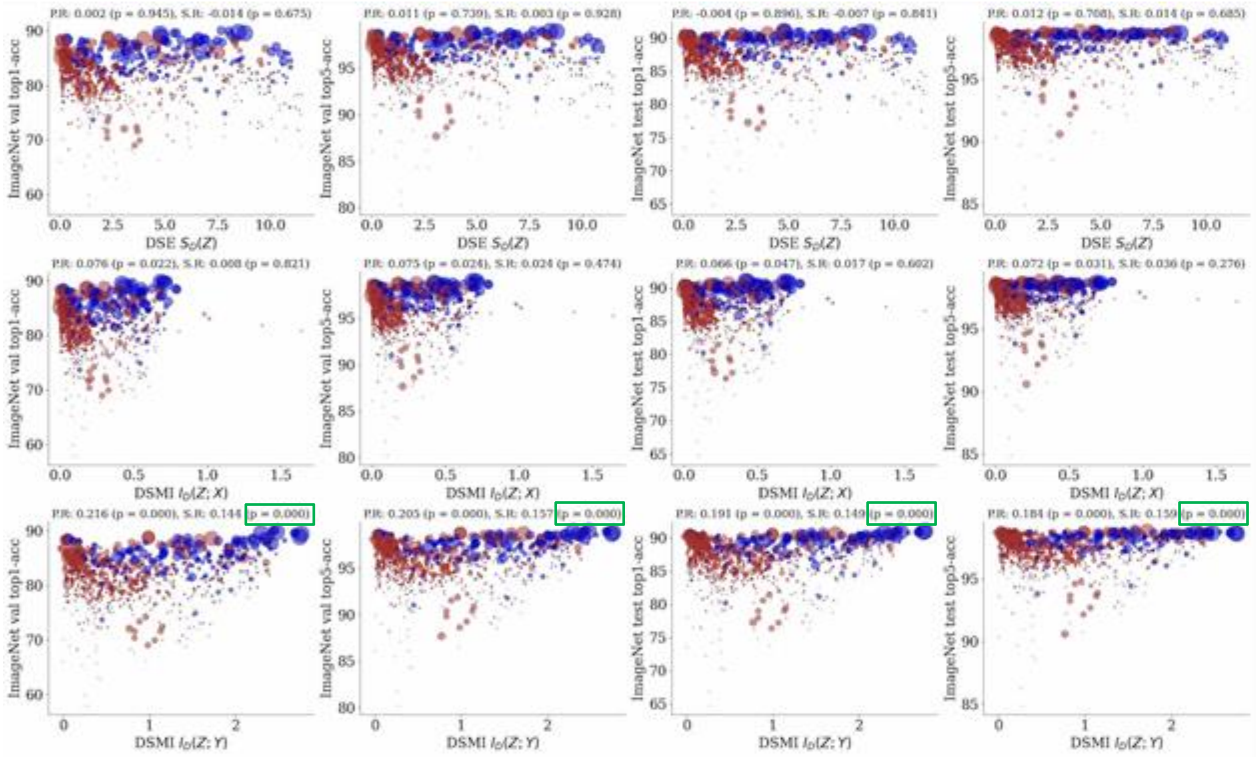


Figure: PHATE representation of the embedding spaces during training for low (panel A) and high (panel B) initial DSE. Colors represent ground truth class labels.

Results (ImageNet cross-model correlation)



- ❑ Correlation analysis between DSE(Z), DSMI(Z; X), DSMI (Z; Y) and ImageNet accuracy evaluated on 962 pretrained models.
- ❑ Red circles are ConvNets and blue circles are ViTs. Circle sizes indicate model sizes.
- ❑ DSMI (Z; Y) (last row) shows a strong positive correlation ($p < 0.001$).

Future works

- ❑ Further investigate the effect of network initialization
- ❑ Explore DSE and/or DSMI as regularizations for supervised learning
- ❑ Use DSE and/or DSMI to regularize self-supervised learning
- ❑ Can further extend this framework to data from other systems, in addition to neural networks, to understand how neural networks such as brain networks.

Author List

Danqi Liao*, Yale University
Chen Liu*, Yale University
Benjamin W. Christensen, Yale University

Alexander Tong, Université de Montréal & Mila -- Quebec AI Institute
Guillaume Huguet, Université de Montréal & Mila -- Quebec AI Institute
Guy Wolf, Université de Montréal & Mila -- Quebec AI Institute

Maximilian Nickel, Meta AI Research (FAIR)
Ian Adelstein, Yale University

Smita Krishnaswamy, Yale University & Meta AI Research (FAIR)