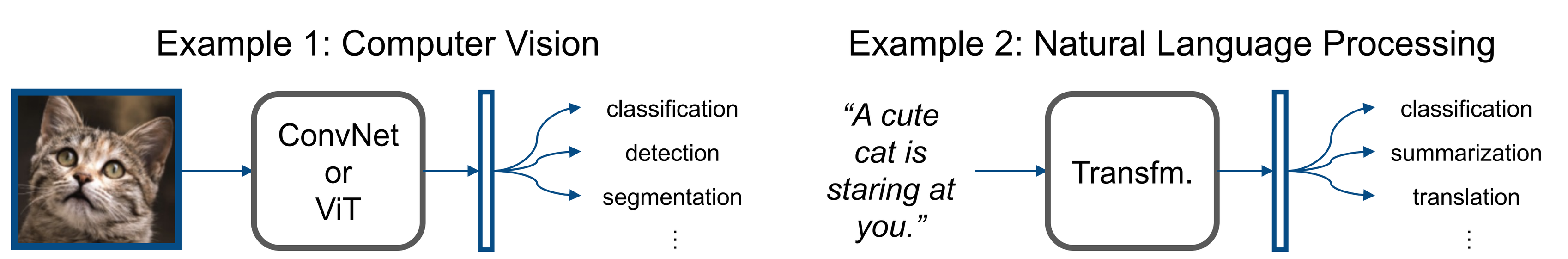
Assessing Neural Network Representations During Training Using Data Diffusion Spectra

Danqi Liao*[1], Chen Liu*[1], Alexander Tong[2], Guillaume Huguet[2], Guy Wolf[2], Maximilian Nickel[3], Ian Adelstein[1], Smita Krishnaswamy[1,3]

*Equal contribution. Order of co-first authors determined by coin toss.
[1] Yale University, [2] The Quebec AI Institute (Mila) and Université de Montréal, [3] The FAIR Team, Meta AI

1. Motivations

- Deep neural networks are powerful as they **learn meaningful representations** of data.



- While the representations vectors reside in high dimensional spaces, they in fact lie on a lower dimensional manifold. **Assessing the properties of this embedding manifold** therefore is key to better understanding the neural network.
- Bonus: this is true for neural representations in intermediate layers as well.

2. Main Innovations

We hereby introduce **two measurements** to quantify properties of neural manifolds.

1. **Diffusion Spectral Entropy (DSE)**
   - Spectral entropy of the diffusion operator (to be explained).
   - A robust quantifier of the intrinsic information measure of data representation despite the presence of noise.
2. **Diffusion Spectral Mutual Information (DSMI)**
   - A mutual information metric derived using DSE.
   - Quantifies information an embedding manifold has on the output labels or the input data of the dataset. Can be extended to other targets.

3. Existing Solutions

The famous ***Deep Learning and the Information Bottleneck Principle*** [1] as well as many other prior work used a simple method for quantifying information content during neural network training.

They **binned** the vectors along each feature dimension to form a probability distribution and computed the **Shannon entropy and mutual information**. This method is known to suffer from the **curse of dimensionality**, which **limits it to toy models** (e.g., a network with 12 neurons at the widest layer).

A follow-up work [2] used kernel density estimation and Kraskov estimator to remedy this issue, yet both methods **require specific assumptions** on the distributions of hidden layer activation.

[1] Tishby & Zaslavsky (2015) [2] On the information bottleneck theory of deep learning, ICLR 2018

4. Background

1. Diffusion geometry

Diffusion geometry seeks to describe data points based on random-walk probabilities to one another. This has been seen to be a noise-tolerant and adaptive way of representing data.
From a pair of points z_1 and z_2, we can compute an anisotropic kernel:

$$\mathcal{K}(z_1, z_2) = \frac{\mathcal{G}(z_1, z_2)}{\|\mathcal{G}(z_1, \cdot)\|_1^\alpha \|\mathcal{G}(z_2, \cdot)\|_1^\alpha}, \quad \mathcal{G}(z_1, z_2) = e^{-\frac{\|z_1 - z_2\|^2}{\sigma}}$$

and the diffusion matrix $\boldsymbol{P}$ is given by:

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

2. Entropy and mutual information

Entropy, a basic quantity in information theory, quantifies the amount of uncertainty or "surprise" when given the value of a random variable.

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

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

Mutual Information: $I(X;Y) = H(X) - H(X|Y) = H(X) - \sum_i p(Y = y_i) H(X|Y = y_i)$

5. Methods

1. Definitions

DSE:

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

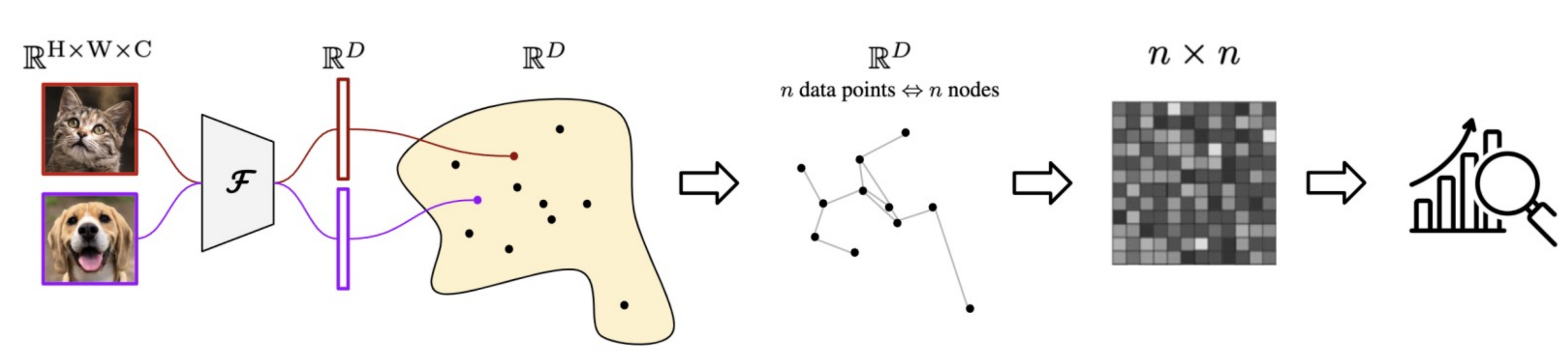
where $\alpha_{i,t} := \frac{|\lambda_i^t|}{\sum_j |\lambda_j^t|}$, *and* $\{\lambda_i\}$ *are the eigenvalues of the diffusion matrix* $\mathbf{P}_X$

DSMI:

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

2. You are talking about "graph diffusion" all the time, but where does the *graph* come from?

Data Graph! We build a data graph from the data samples!

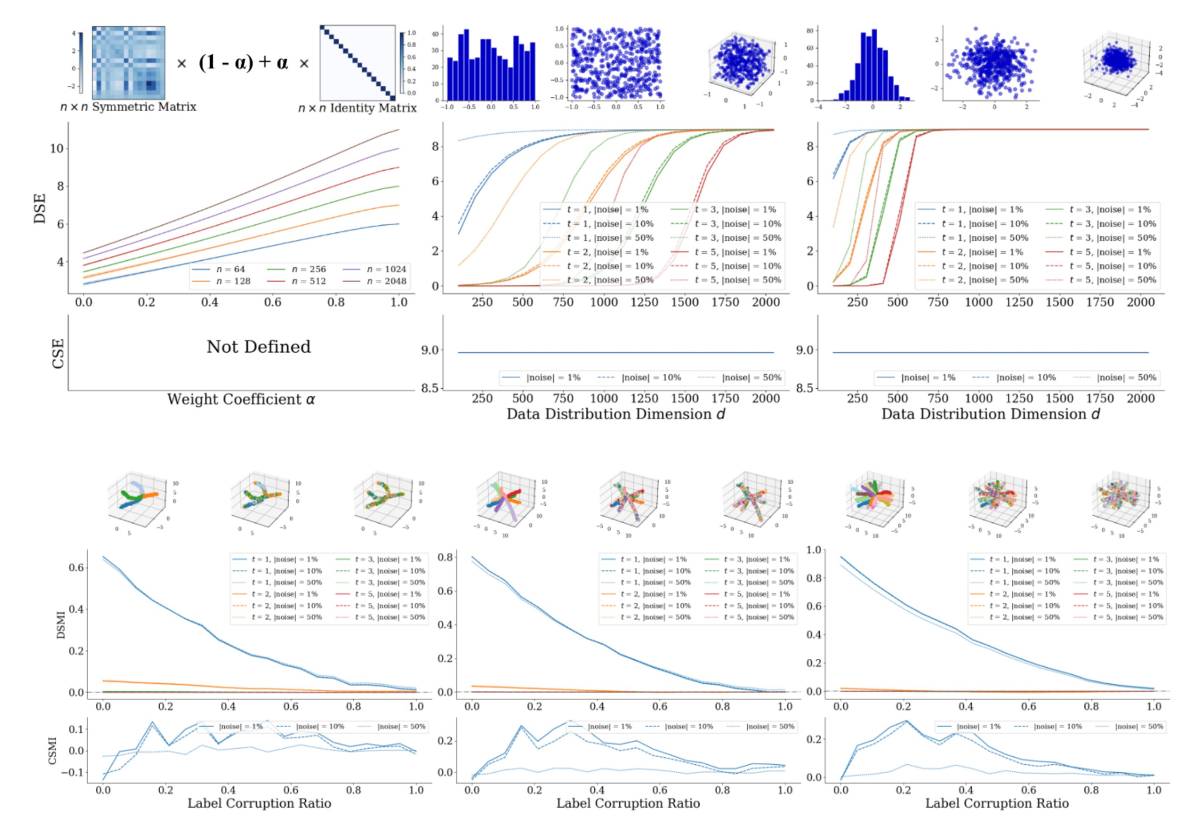


6. Results

1. Toy examples

DSE vs. CSE

DSMI vs. CSMI

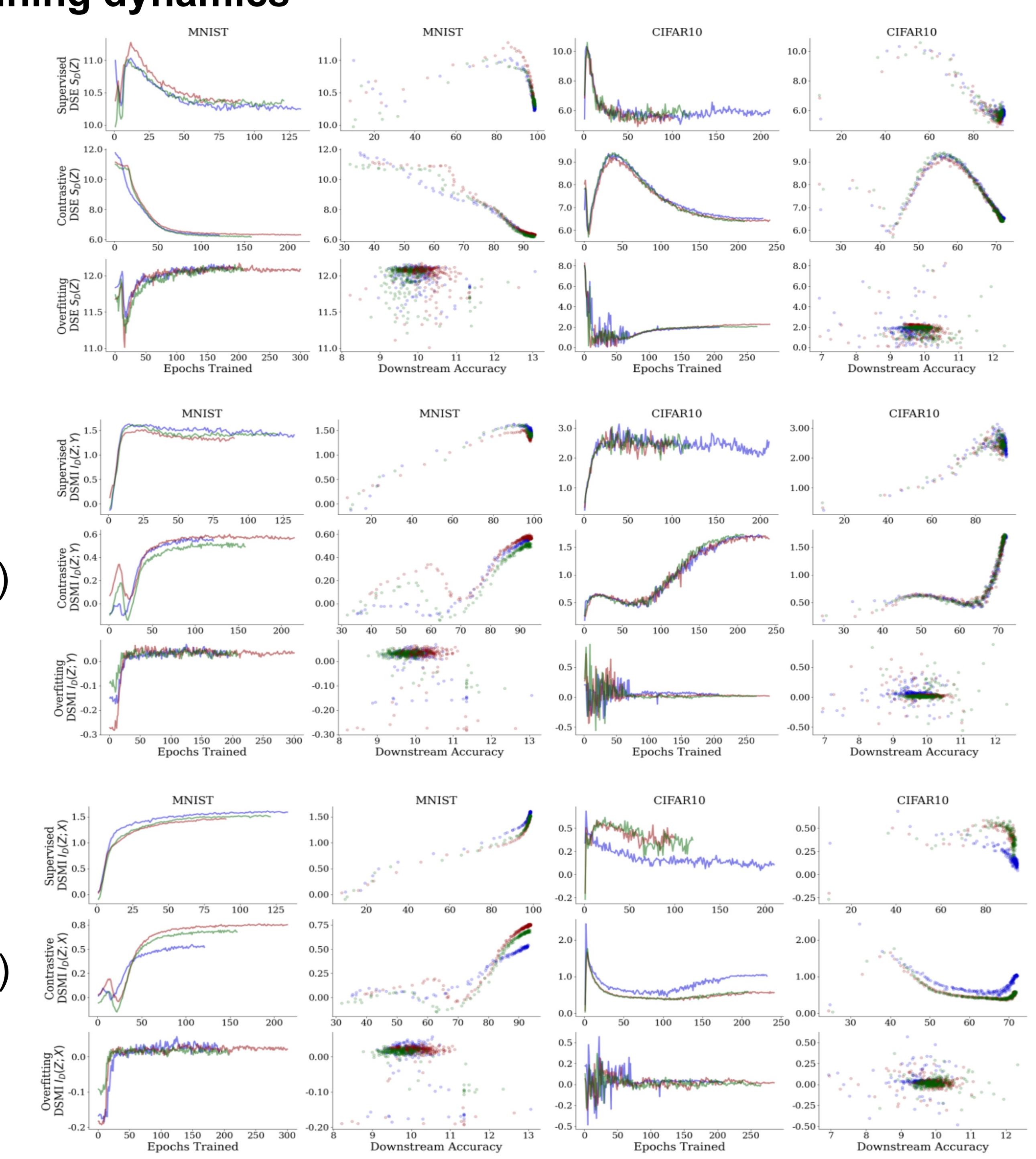


2. Real training dynamics

DSE(Z)

DSMI(Z; Y)

DSMI(Z; X)



Paper PDF

GitHub