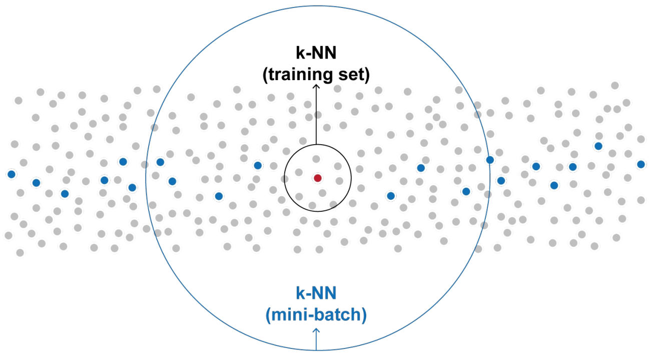
Efficient Local Intrinsic Dimensionality Estimation in Evolving Deep Representations

Problem

Poor-locality in mini-batch LID estimation for efficiency

LID estimation using k-NN distances with respect to a sample set:

$$\widehat{\mathrm{LID}} = -\left(\frac{1}{k}\sum_{j=1}^{k} \ln \frac{r_j}{r_k}\right)^{-1}$$



- reference point
- mini-batch (b=20, k=8)
- training set (n=300, k=8)

k/n (high locality) << k/b (poor locality)

Main Idea

Enhancing locality of LID estimation by reusing historical distances in deep network training

Sliding window of mini-batch distances (in latent space) for point i :

$$O_i^{t-\tau+1} \quad O_i^{t-1} \quad O_i^{t} = \{r_1, r_2, \cdots, r_{b-1}\}$$

$t-\tau+1 \quad t-1 \quad t$ epoch

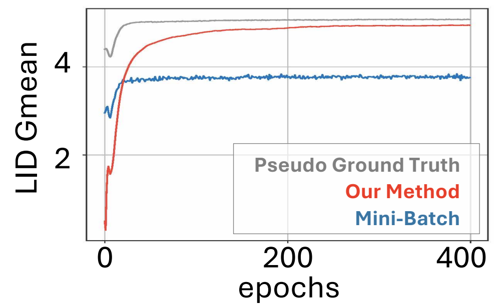
LID estimation using	Spatial locality
O_i^t	$\frac{k}{b-1}$
$^*W_i^t = \cup_{j=t-\tau+1}^{t} O_i^j$	$\frac{k}{\tau \cdot (b-1)}$

*With additional strategies to handle distributional drift over time and reduce memory requirements.

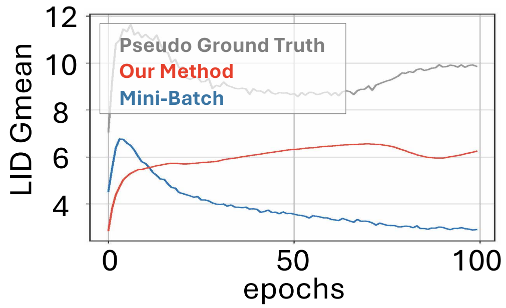
Results

Better approximation of pseudo ground truth than mini-batch estimation

Autoencoder + Synthetic data



Resnet + CIFAR10



NJIT New Jersey Institute of Technology

Michael E. Houle, Vincent Oria, Hao Xu