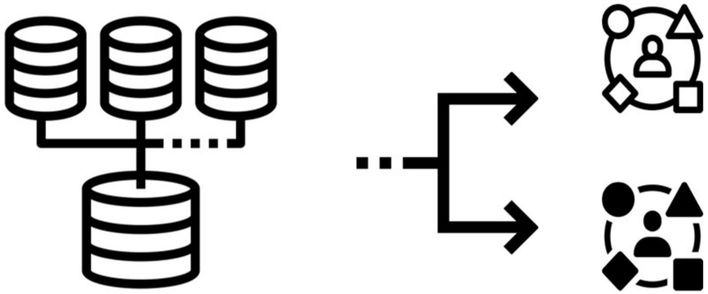
NEURAL INFORMATION PROCESSING SYSTEMS

On the Trade-off of Intra-/Inter-class Diversity for Supervised Pre-training

Jieyu Zhang*, Bohan Wang*, Zhengyu Hu,
Pang Wei Koh, Alexander Ratner

Diversities in Supervised Pre-training

Two kind of diversity for a supervised pre-training dataset



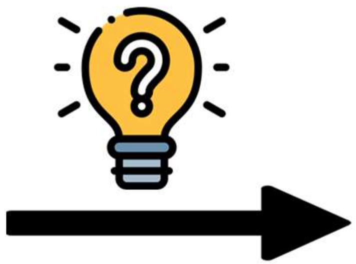
Intra-class diversity:

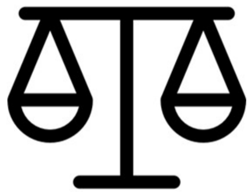
Number of different samples within each pre-training class.

Inter-class diversity:

Number of different pre-training classes.

Trade-off Between Diversities





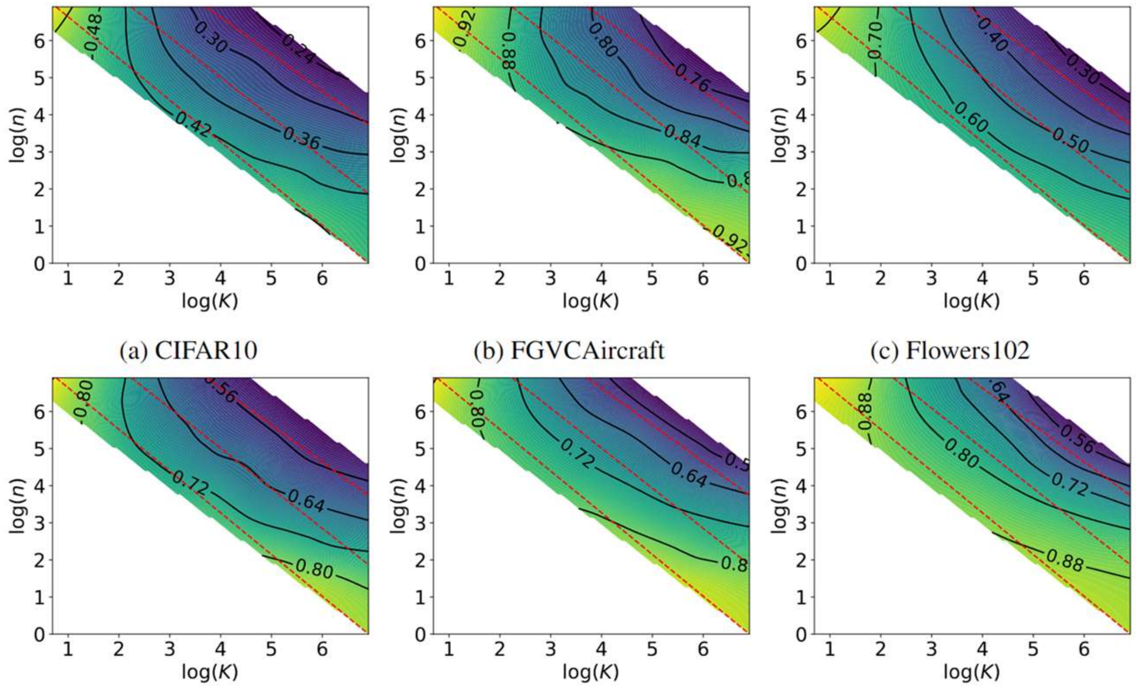
With a fixed Dataset budget(size)

Intra-class diversity VS Inter-class diversity

Empirical Observations on Intra-/Inter-Class Diversity

***Both* intra-/inter-class diversity are beneficial for downstream tasks.**

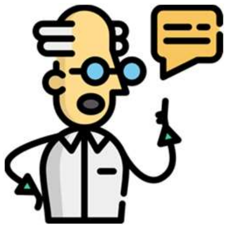
***A trade-off* of intra-/inter-class diversity on downstream task performance.**



(a) CIFAR10 (b) FGVCAircraft (c) Flowers102

(d) MIT67 (e) Stanford40 (f) StanfordDogs

Theoretical Understanding: Impact of Intra-/Inter-Class Diversity Trade-off



Theorem 3.1. *Let Assumptions 1 and 2 hold. Then, with probability over the sampling of the datasets at least* $1-\delta$*, we have*

$$\mathcal{E}_d(f_{S^d}\circ h_{S_p}, \tilde{\mathcal{P}}) \le \left(\nu_1^{\tilde{\mathcal{P}}}(\mathcal{D}) + M_1\sqrt{\frac{\log\frac{4}{\delta}}{2K}} + \frac{C_1}{\sqrt{K}}\right)\left(5M_\ell\sqrt{\frac{\log\frac{6}{\delta}}{2n}} + \frac{2G\sqrt{2}}{\sqrt{n}}\right) + \nu_0^{\tilde{\mathcal{P}}}(\mathcal{D}) + M_0\sqrt{\frac{\log\frac{6}{\delta}}{2K}} + \frac{C_0}{\sqrt{K}} + 5M_\ell\sqrt{\frac{\log\frac{6}{\delta}}{2\tilde{N}}} + 2\sqrt{2}G\frac{1}{\sqrt{\tilde{N}}}. \tag{1}$$

A simplified version

$$U = \frac{A}{\sqrt{n}} + \frac{B}{\sqrt{K}} + \frac{C}{\sqrt{N}} + D$$

Put in: N = n × K

$$U(K) = \frac{A\sqrt{K}}{\sqrt{N}} + \frac{B}{\sqrt{K}} + \frac{C}{\sqrt{N}} + D$$

Theoretical Understanding:
Optimal Class-to-Sample Ratio

When N is fixed, by leveraging the fact that N = n × K, we can express U as

$$U = \frac{1}{N^{\frac{1}{4}}}\left(Ax^{\frac{1}{4}} + B\frac{1}{x^{\frac{1}{4}}}\right) + c$$

Optimal class-to-sample ratio: $\bar{x} = \frac{B^2}{A^2}$ ***invariant to N !!!***

Predicting the optimal number of pre-training classes

Extrapolation:

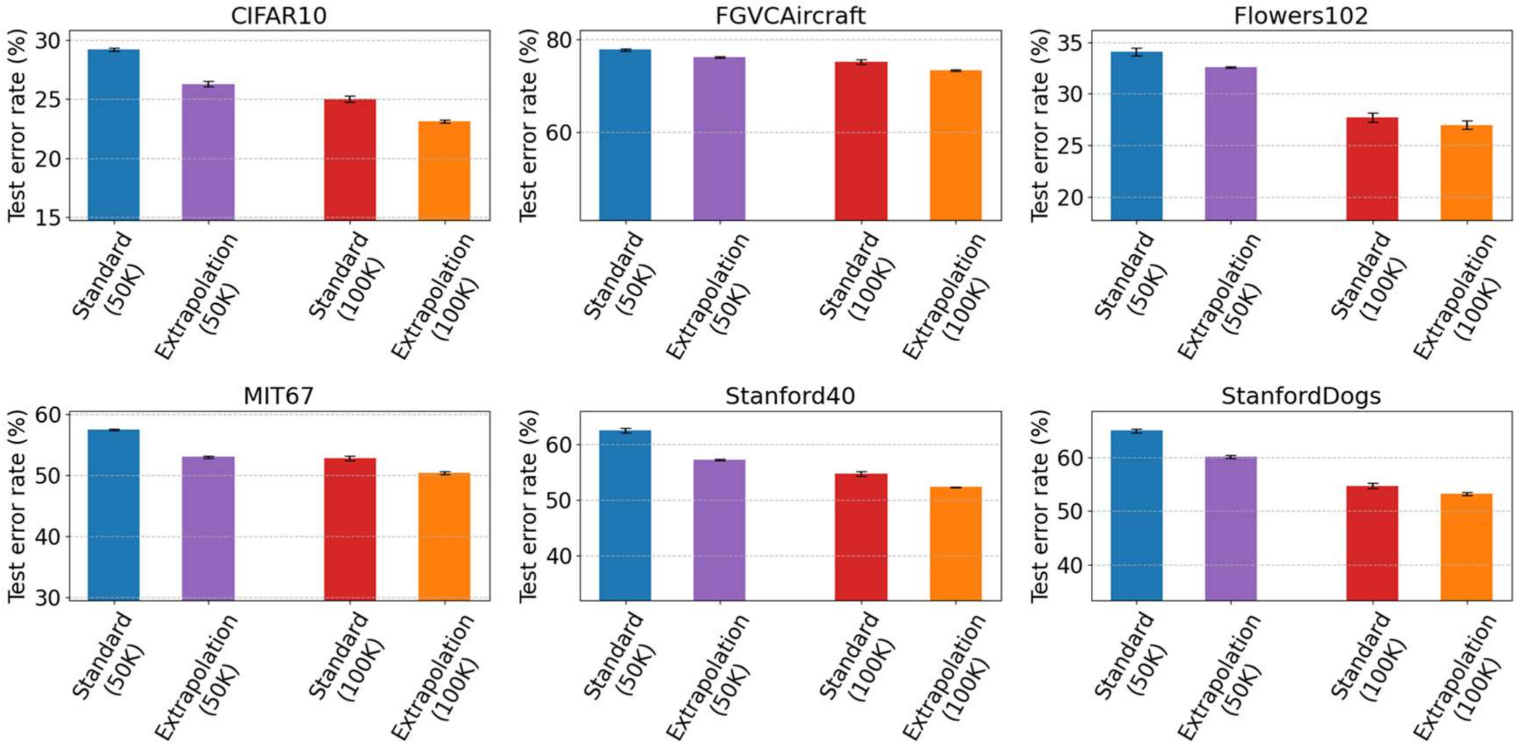
Small dataset —*calculate*→ $\bar{x} = \frac{B^2}{A^2}$ —*estimate*→ $\bar{K} = \sqrt{\bar{x}N}$ —*construct*→ Large dataset

Small dataset | Optimal class-to-sample ratio | The number of classes | Large dataset

Standard:

The number of classes equals 1000 as the standard design choice of ImageNet

Predicting the optimal number of pre-training classes



The number of classes K Extrapolation finds are all superior to the Standard.

Thank you for your listening!