ParNeC 模式识别与神经计算研究组
PAttern Recognition and NEural Computing

When do Curricula Work?

Xiaoxia Wu*
UChicago and TTIC
xwu@ttic.edu

Ethan Dyer
Blueshift, Alphabet
edyer@google.com

Behnam Neyshabur
Blueshift, Alphabet
neyshabur@google.com

ICLR 2021

Curriculum Learning (CL)

Curriculum: the **learning order** of the examples (e.g., from easy to hard)

human learning

machine learning

Contributions

- **Train over 25,000 models over 4 datasets (i.e.,** CIFAR10/100, FOOD101, and FOOD101N**) to better understand CL**

  ✓ **Implicit Curricula**: examples are learned is consistent across runs, similar training methods, and similar architectures.

  ✓ **Effectiveness**:

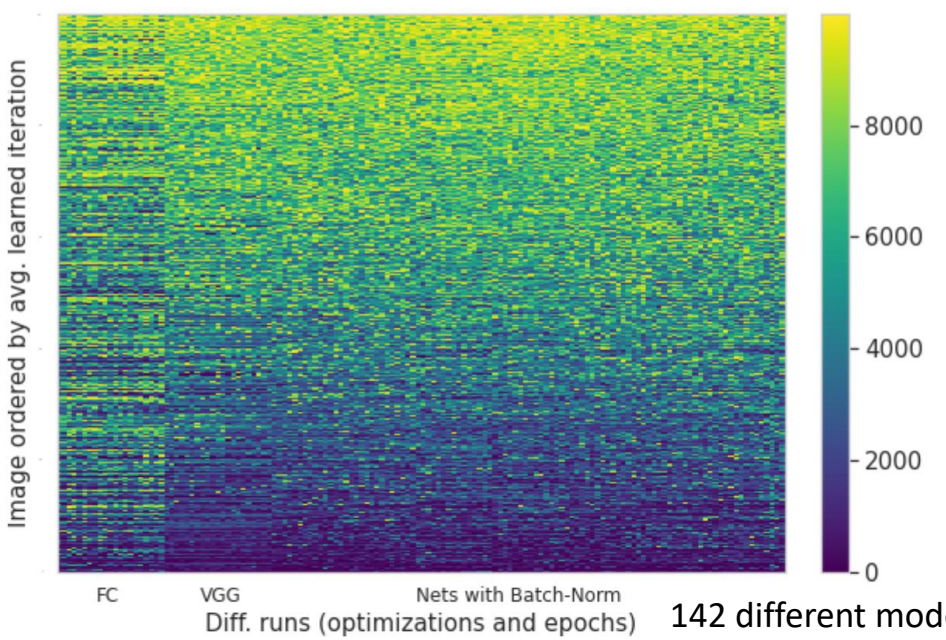
  - Almost no improvement in the standard setting.
  - Effective with limited training time.
  - Effective under the noisy regime.

Implicit Curricula

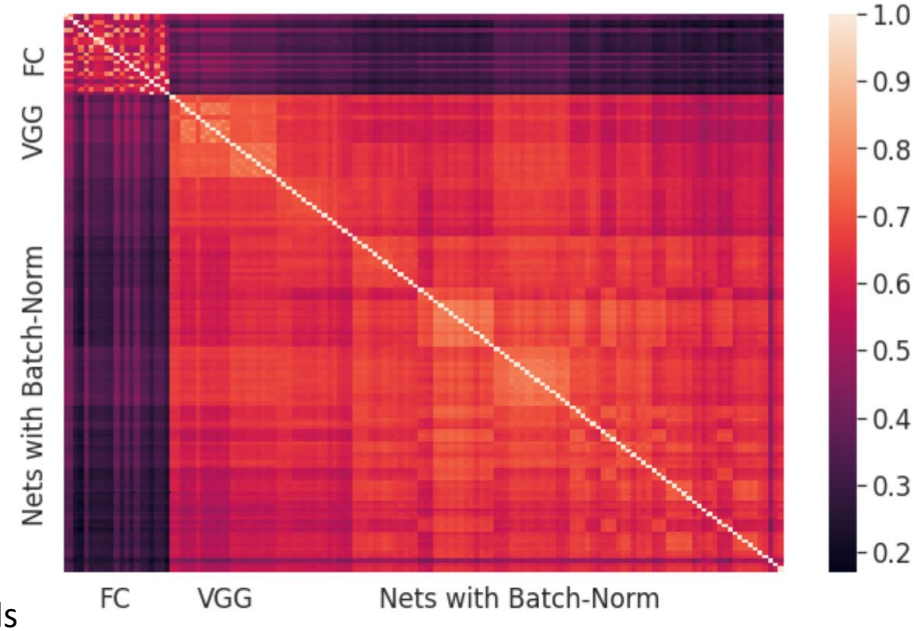
- ***Learned iteration:*** the epoch for which the model correctly predicts the sample for that and all subsequent epochs.

$$\min_{t^*}\{t^* | \hat{y}_{\mathbf{w}}(t)_i = y_i, \forall t^* \leq t \leq T\}$$

Examples in CIFAR-10



142 different models



Images are learned in a similar order for similar architectures and training methods, which implies **the difficulty of a given image is less model-dependent**.

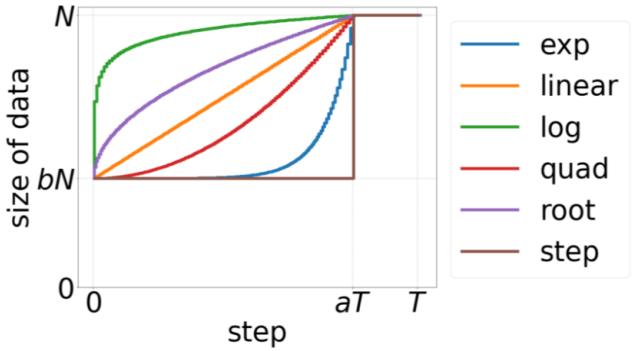
Effectiveness

✓ Scoring Functions

- Loss function.
- Learned epoch/iteration.
- Estimated c-score. $-\mathbb{E}_{D\overset{n}{\sim}\hat{\mathcal{D}}\setminus\{(\boldsymbol{x}_i,y_i)\}}\left[\mathbb{P}(\hat{y}_{\mathbf{w},i}=y_i|D)\right]$

✓ Pacing Functions

Name	Expression $g_{(a,b)}(t)$
log	$Nb+N(1-b)\left(1+.1\log\left(\frac{t}{aT}+e^{-10}\right)\right)$
exp	$Nb+\frac{N(1-b)}{e^{10}-1}\left(\exp\left(\frac{10t}{aT}\right)-1\right)$
step	$Nb+N\left\lceil\frac{x}{aT}\right\rceil$
polynomial	$Nb+N\frac{(1-b)}{(aT)^p}t^p \quad - \; p=1/2,1,2$

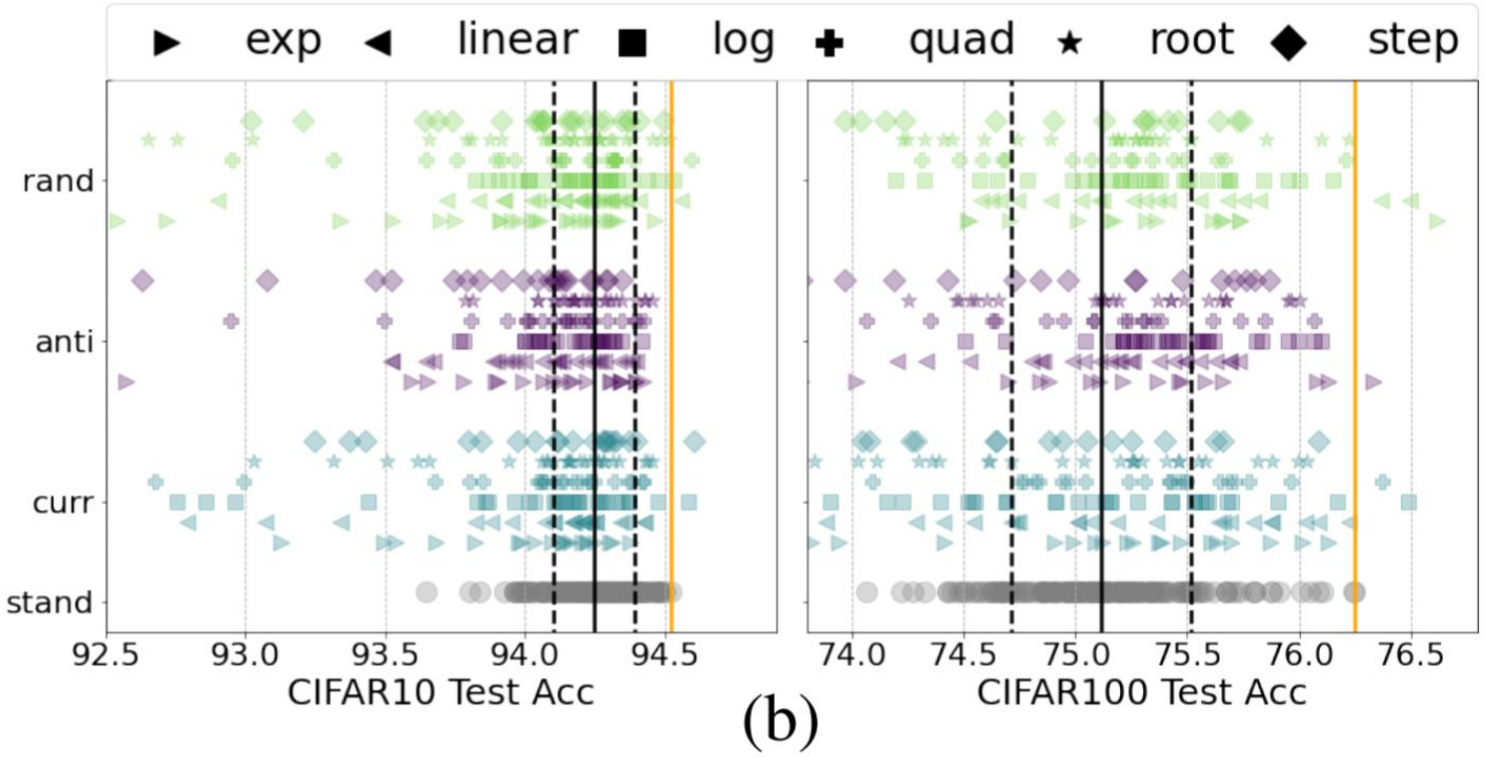


✓ The Order:

- Curriculum or anti-curriculum or random

Standard setting

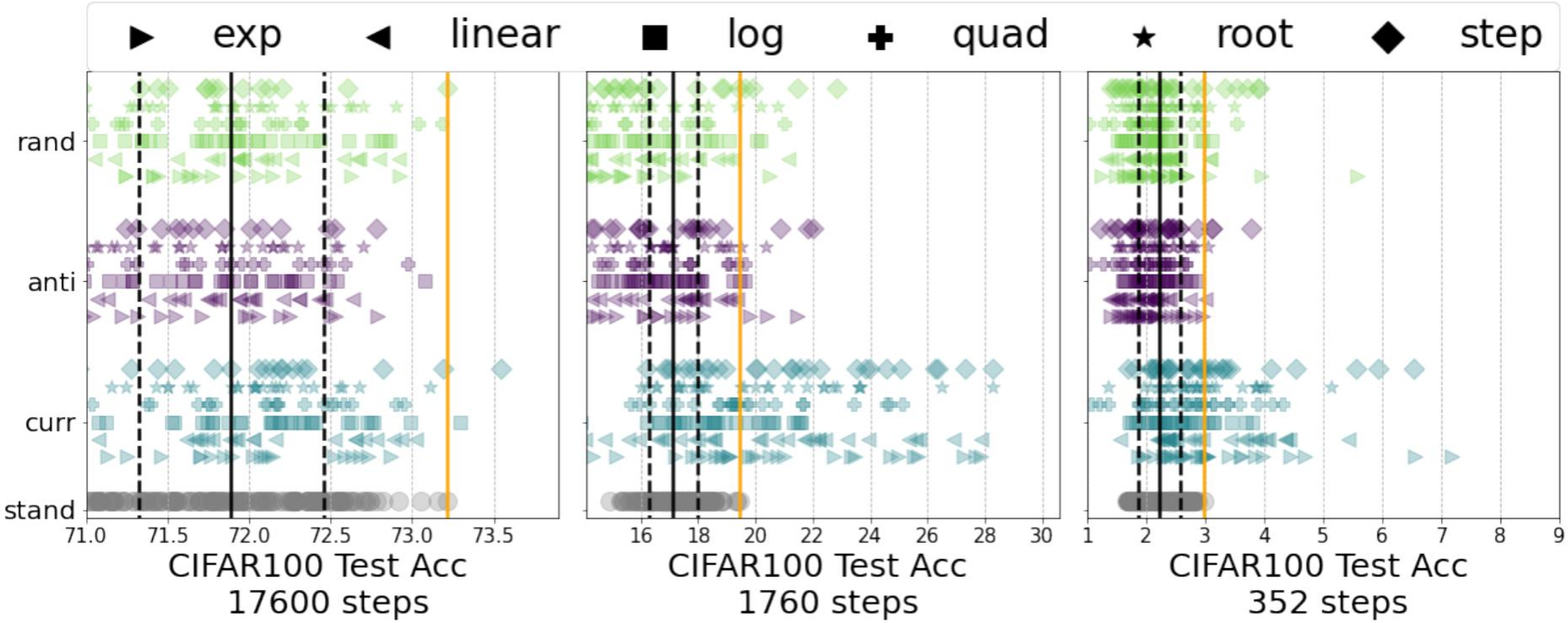
- Train a ResNet-50 model for 100 epoch (expected to be converged)
- 540 configs (180 different pacing func. x 3 orders), each one is repeated for 3 times with different seeds.



PACING FUNCTIONS GIVE MARGINAL BENEFIT, CURRICULA GIVE NONE

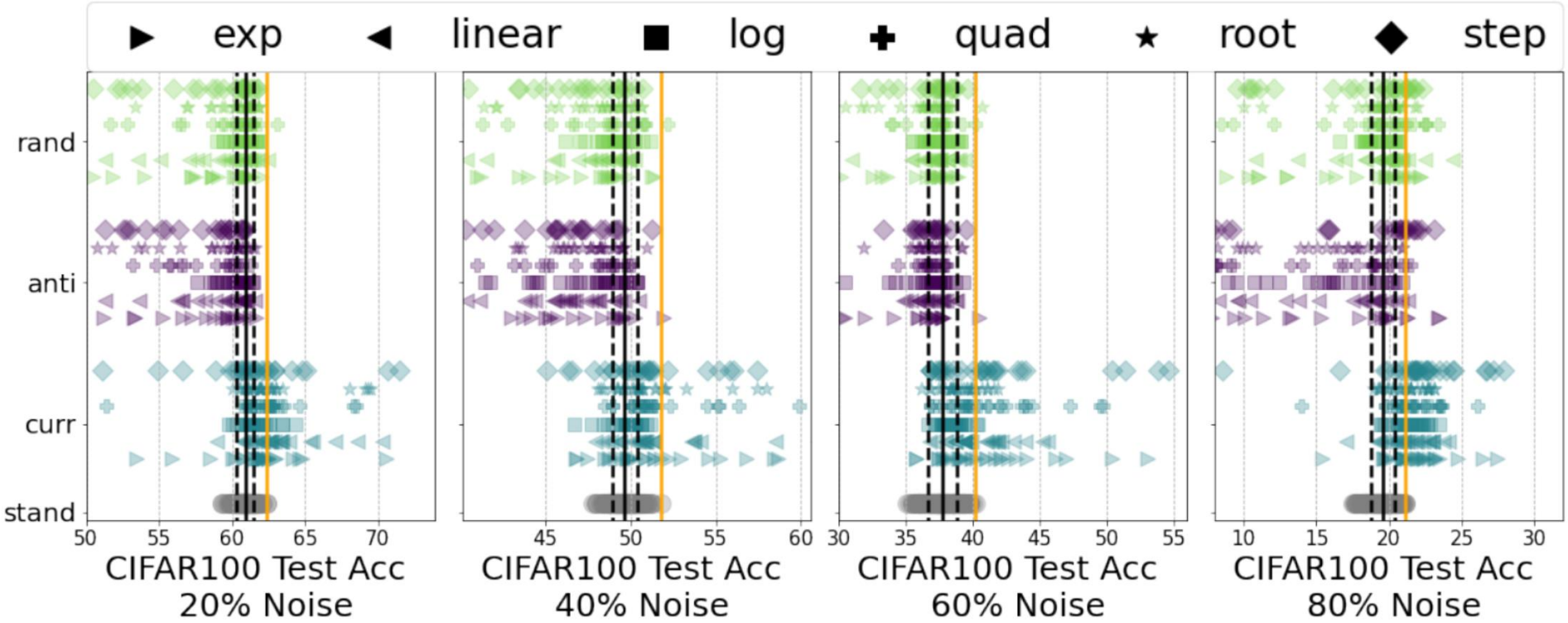
Limited time setting

- Train a ResNet-50 model for 1/5/50 epochs



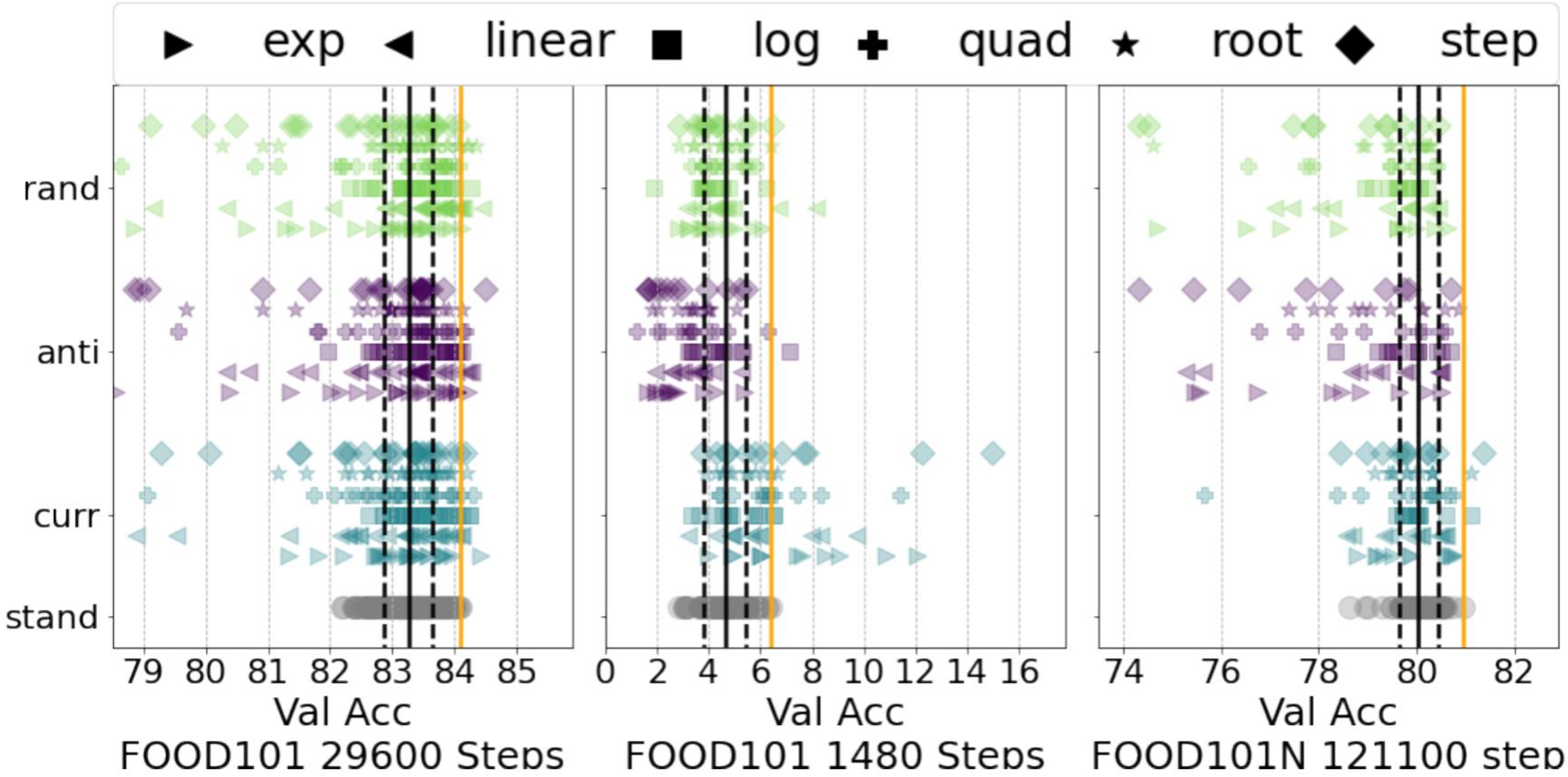
Noisy setting

- Train a ResNet-50 model for 100 epochs



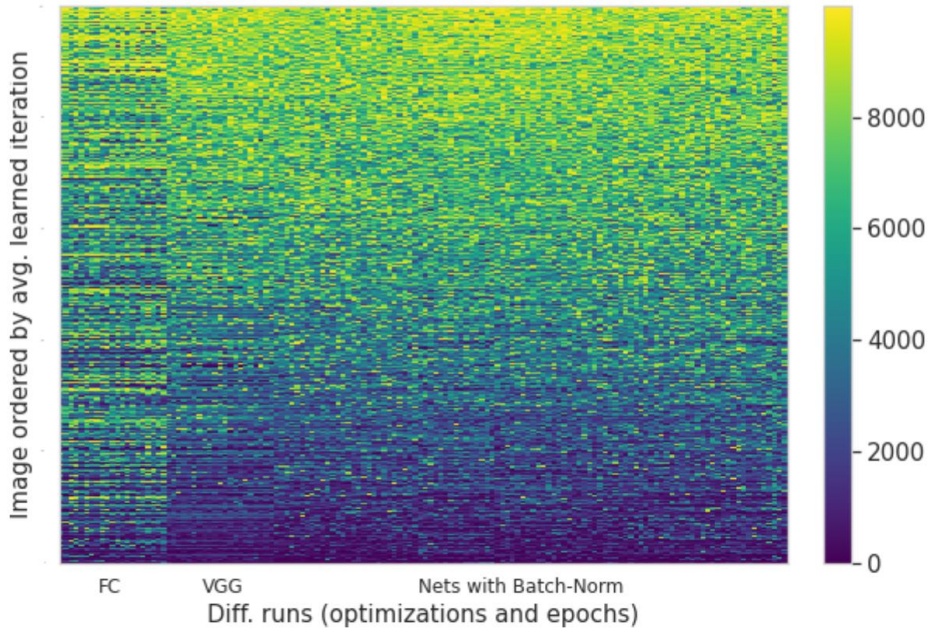
Large datasets

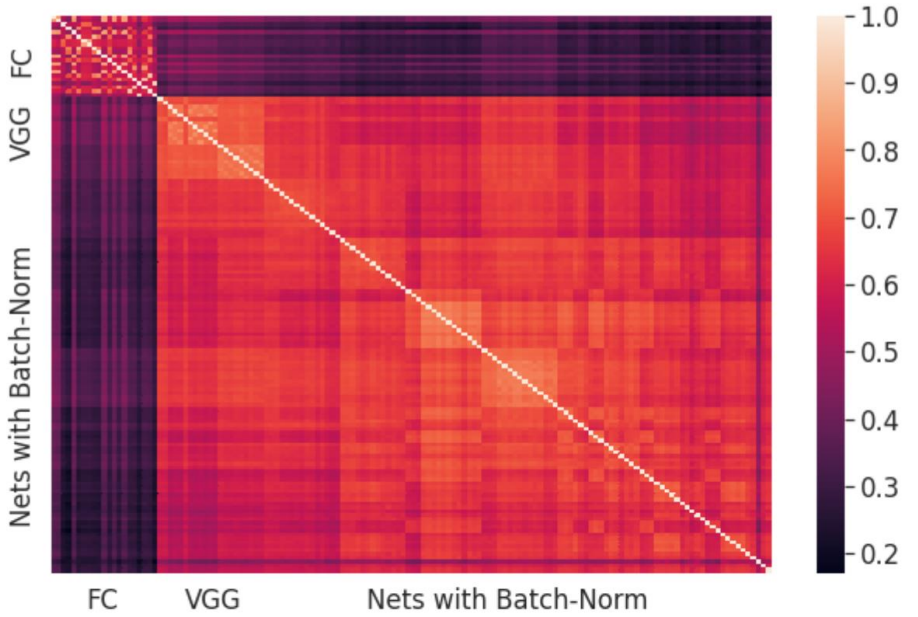
- FOOD101 and FOOD101N datasets, which contain 75,000 and 310,000 examples, respectively.



✓ **Similar phenomena are observed**

Discussion





Images are learned in a similar order for similar architectures and training methods, which implies **the difficulty of a given image is less model-dependent**.

implies

Although the training instances are fed to the model with different orders, they may be learned by the model in a consistent order if there are enough training epochs.

ParNeC 模式识别与神经计算研究组
PAttern Recognition and NEural Computing

THANKS