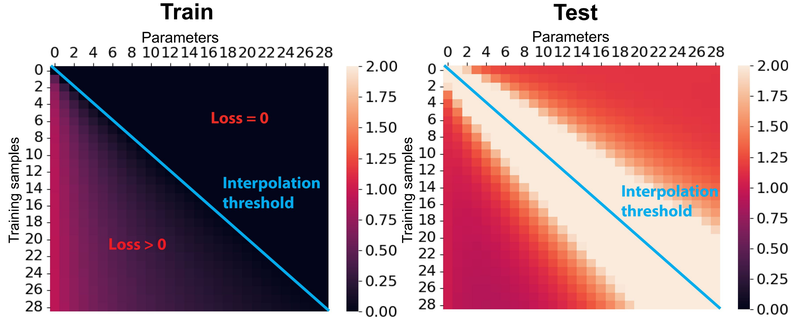
Breaking Neural Network Scaling Laws with Modularity

Akhilan Boopathy, Sunshine Jiang, William Yue, Jaedong Hwang, Abhiram Iyer, Ila Fiete

Massachusetts Institute of Technology

The Exponential Cost of High-Dimensionality

- Linearization analysis accurately predicts generalization behavior of neural networks

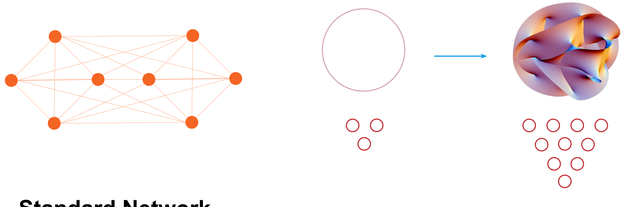


- Linearization analysis also predicts the "curse of dimensionality"

Theorem 1 (informal): Standard Networks

$$\mathbb{E}[|y(x) - \hat{y}(x)|^2] = O(B^m)$$

True label · Predicted label · Task input dimensionality



Standard Network

Increased input dimension
Exponential increase in required data

Breaking the Curse: Modular Networks Generalize Efficiently

Thetheorem 2 (informal): Modular Networks

$$\mathbb{E}[|y(x) - \hat{y}(x)|^2] = O(1)$$

True label · Predicted label

Modular Network

Increased input dimension
No increase in required data!

Modular Networks Require Specialized Optimization

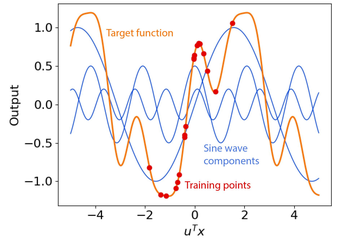
$$\min_{\hat{U}} Y^T K^{-1} Y \qquad K(x_1, x_2) = \sum_i \kappa(x_1, x_2; \hat{U}_i)$$

Module pretraining objective · True labels on all training data · Kernel · Task inputs · Module input parameters

- We pretrain module input parameters before training the full network
- Module input parameters are trained to align label similarity with module input similarity

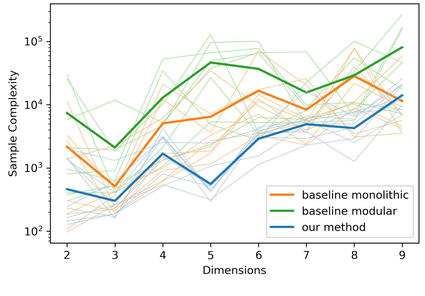
Modular Initialization Rule Improves Generalization

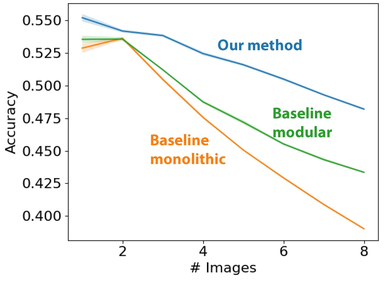
Sine wave regression



Compositional CIFAR-10

frog-horse-truck-airplane





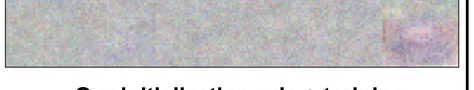
- Our modular initialization rule outperforms baselines

Module feature visualization

Untrained network

After standard training

Our initialization rule



Our initialization rule + training

- Our rule specializes modules even before training