The Pitfalls of Simplicity Bias in Neural Networks

Harshay Shah, Kaustav Tamuly, Aditi Raghunathan, Prateek Jain, and Praneeth Netrapalli
{harshay.rshah, ktamuly2, aditir1994, pjain9, praneethn}@gmail.com
Microsoft

Simplicity Bias (SB)

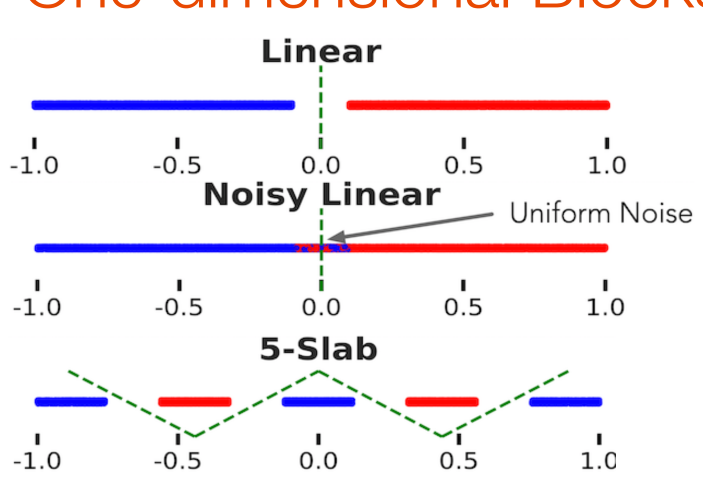
Simplicity Bias, the tendency of standard training methods like SGD to find simple models, is often used to justify why neural networks (NNs) generalize well.

However, existing works on SB vis-a-vis generalization lack a precise notion of simplicity and do not shed light on why neural networks lack robustness in practice.

Our goal is to better understand (a) the effect of *Simplicity Bias* (SB) on **feature learning** and (b) its implications on **robustness** and **generalization**.

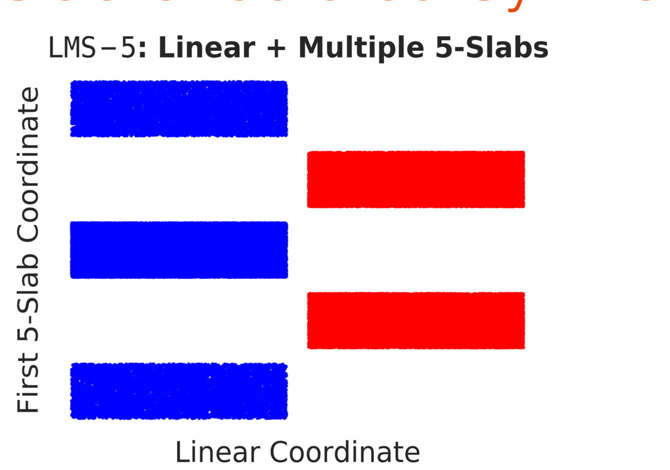
Datasets

One-dimensional Blocks and Simplicity

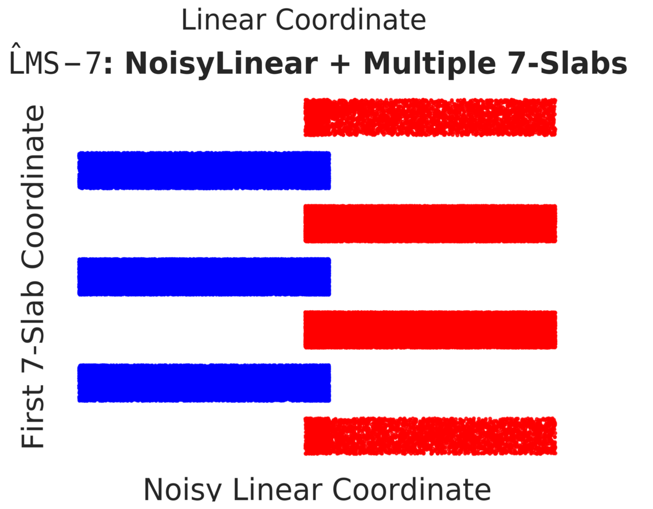


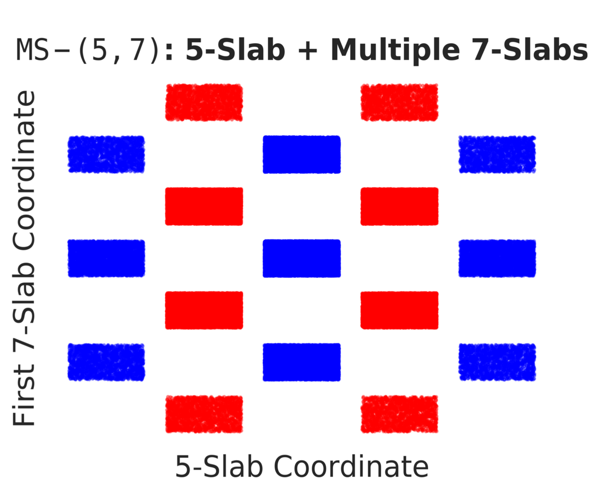
The blocks have a natural notion of feature simplicity: *minimum number of pieces required by a piecewise linear classifier to attain Bayes optimal accuracy.*

Slab-structured Synthetic Datasets



The datasets comprise features of varying simplicity and predictive power, as each coordinate maps to a one-dimensional block.





MNIST-CIFAR Dataset

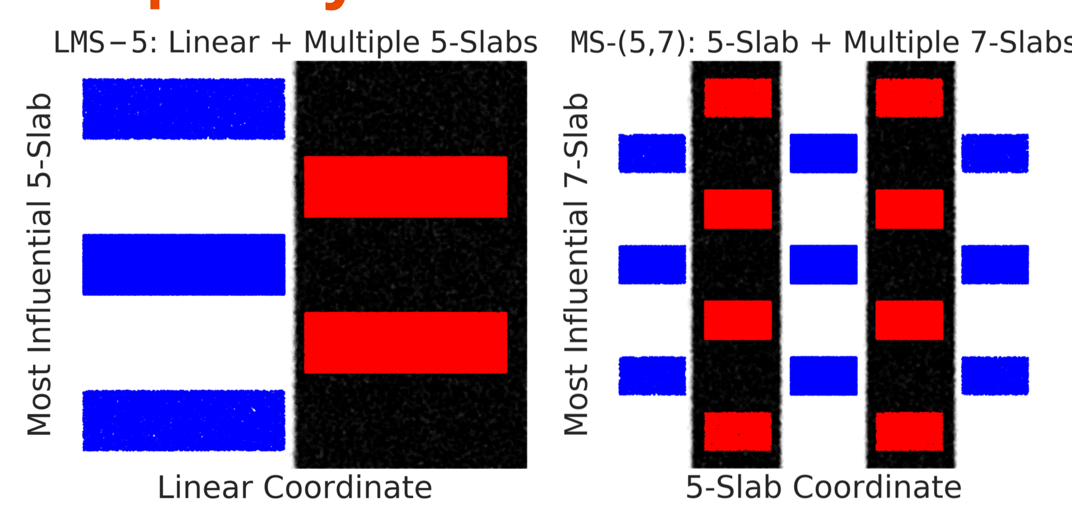
The data consists of two classes. Images are **vertical concatenations** of MNIST and CIFAR-10 images.

Our Findings

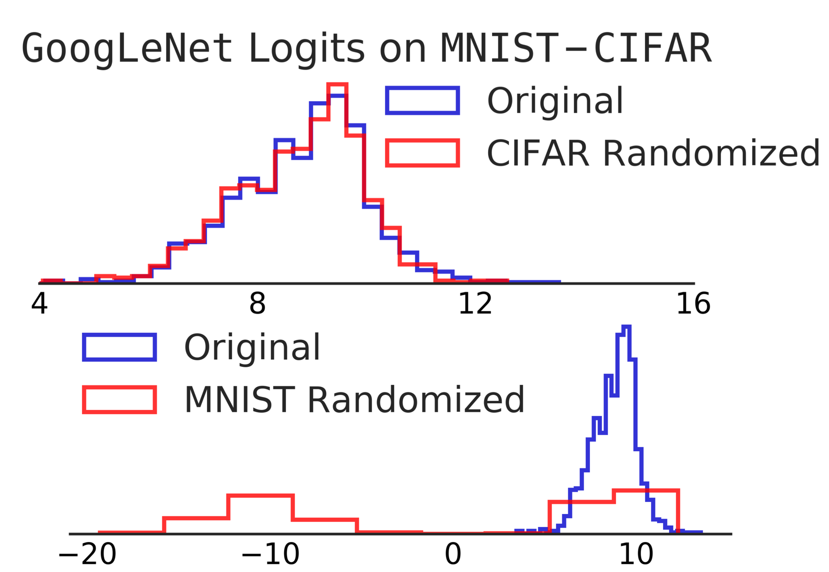
Theory and empirics on piecewise-linear and semi-real image datasets support three findings about of Simplicity Bias in neural networks:

1. SB is extreme: SGD-trained NNs **exclusively rely on the simplest feature** and **remain invariant to all complex features**, even if they've equal predictive power.
2. Extreme SB jointly shed light on why seemingly **benign distribution shifts** and **universal adversarial attacks** can drastically degrade model performance.
3. Contrary to conventional wisdom, **SB can hurt standard generalization** as well.

Simplicity Bias is Extreme in Practice



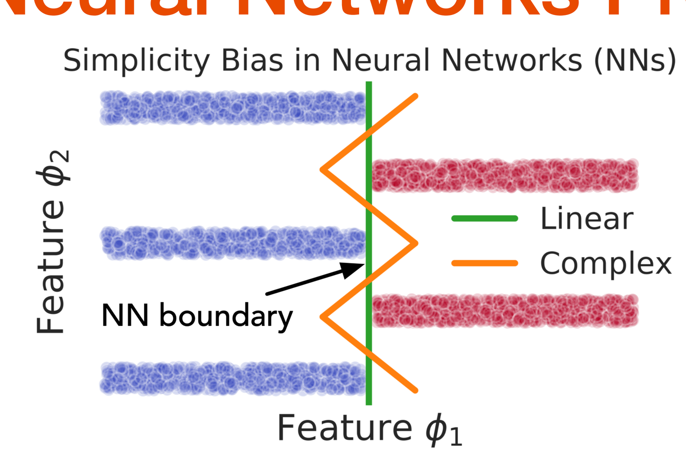
SGD-trained NNs *exclusively* rely on the simplest feature **S** and remain invariant to all complex features **S^c**, even though **S** and **S^c** have equal predictive power.



Randomizing the simplest feature **S** nullifies model performance and **randomly shuffles the logits** across classes

However, randomizing *all* complex features **S^c** has no effect on model performance; **logits remain unchanged**.

Neural Networks Provably Exhibit Simplicity Bias (SB)



One-hidden-layer ReLU NNs trained on LSN data learn small-margin classifiers that **only rely on the linear coordinate** instead of large-margin classifiers that **rely on linear & slab coordinates.**

Theorem 1. *Let $f(x) = \sum_{j=1}^{k} v_j \cdot ReLU(\sum_{i=1}^{d} w_{i,j} x_i)$ denote a one-hidden-layer neural network with k hidden units and ReLU activations. Set $v_j = \pm 1/\sqrt{k}$ w.p. $1/2\ \forall j \in [k]$. Let $\{(x^i, y^i)\}_{i=1}^{m}$ denote i.i.d. samples from* LSN *where $m \in [cd^2, d^\alpha/c]$ for some $\alpha > 2$. Then, given $d > \Omega(\sqrt{k}\log k)$ and initial $w_{ij} \sim \mathcal{N}(0, \frac{1}{dk\log^4 d})$, after $O(1)$ iterations, mini-batch gradient descent (over w) with hinge loss, constant step size, mini-batch size $\Theta(m)$, satisfies:*

- *Test error is at most $1/poly(d)$*
- *The learned weights of hidden units w_{ij} satisfy:*

$$\underbrace{|w_{1j}| = \frac{2}{\sqrt{k}}\left(1 - \frac{c}{\sqrt{\log d}}\right) + O\left(\frac{1}{\sqrt{dk}\log d}\right)}_{\textit{Linear Coordinate}}, \underbrace{|w_{2,j}| = O\left(\frac{1}{\sqrt{dk}\log d}\right)}_{\textit{3-Slab Coordinate}}, \underbrace{\|w_{3:d,j}\| = O\left(\frac{1}{\sqrt{k}\log d}\right)}_{d-2\textit{ Noise Coordinates}}$$

with probability greater than $1 - \frac{1}{poly(d)}$. Note that c is a universal constant.

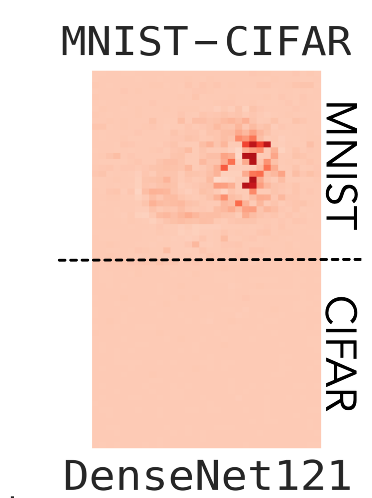
Pitfalls of Extreme Simplicity Bias

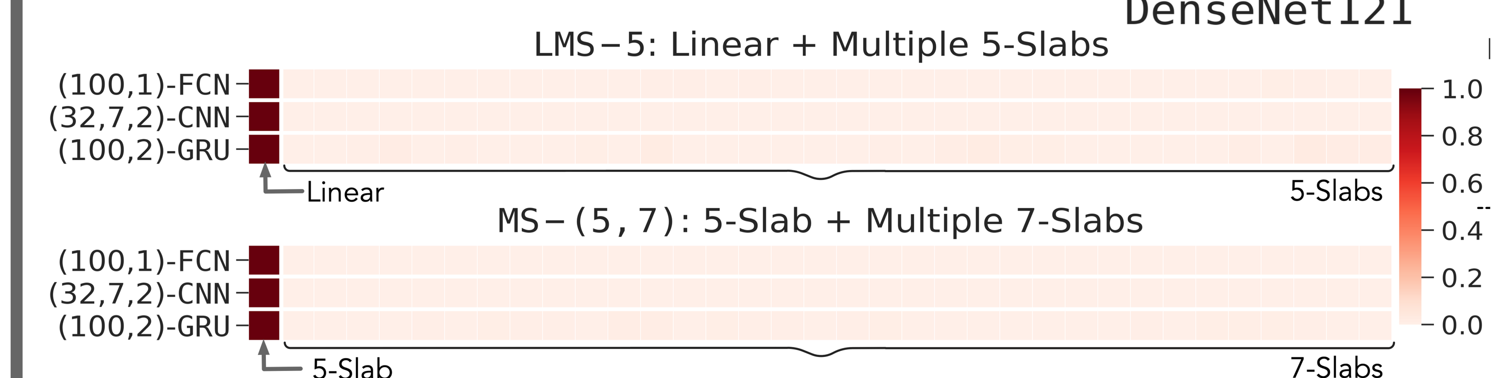
Unreliable Out-of-Distribution Performance

Sensitivity to simple feature(s) **S** and invariance to complex features **S^c** result in NNs that exhibit **unreliably high confidence estimates**, even when **S^c** contradicts **S**.

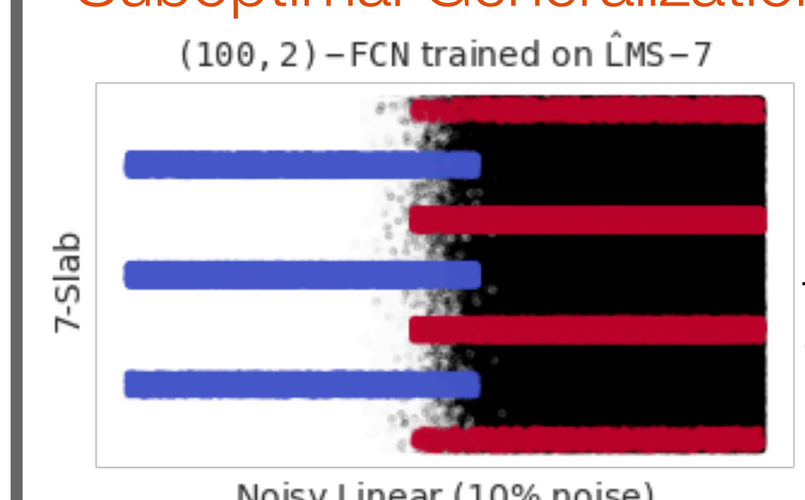
Adversarial Vulnerability

Extreme SB results in small-margin classifiers that are vulnerable to "universal" data-agnostic as well as model-agnostic adversarial attacks that only perturb the simple feature(s) **S**.





Suboptimal Generalization



FCNs trained on noisy LMS-7 data with SGD latch on to the noisy linear feature **S** even though all 7-slabs **S^c** have 100% predictive power.

Accuracy	(100,1)-FCN	(200,1)-FCN	(300,1)-FCN
Training Data	0.984 ± 0.003	0.998 ± 0.000	0.995 ± 0.000
Test Data	0.940 ± 0.002	0.949 ± 0.003	0.948 ± 0.002
S^c-Randomized	0.941 ± 0.001	0.946 ± 0.001	0.946 ± 0.001
S-Randomized	0.498 ± 0.001	0.498 ± 0.000	0.497 ± 0.001

Mitigating the pitfalls of Simplicity Bias

Adversarial training and **ensembles of independently trained models** do not mitigate the pitfalls of SB in the proposed datasets (see Appendix E).

Our datasets and metrics collectively motivate the need for *new* algorithmic approaches to mitigate the pitfalls of SB.