





#### Can Neural Nets Learn the Same Model Twice?

### **Investigating Reproducibility and Double Descent from the**

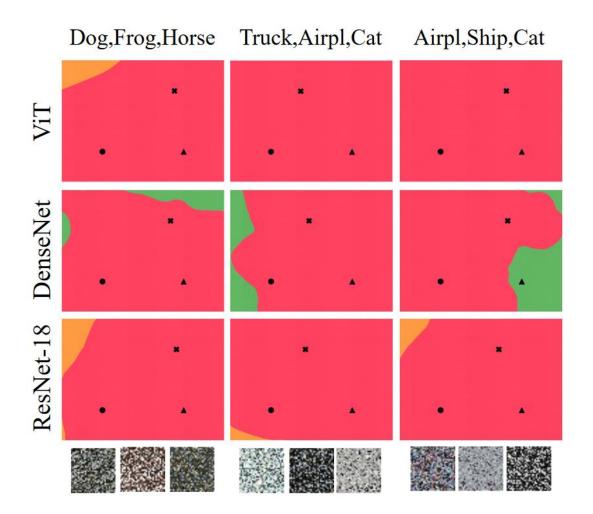
#### **Decision Boundary Perspective**

Gowthami Somepalli<sup>1</sup>, Liam Fowl<sup>1</sup>, Arpit Bansal<sup>1</sup>, Ping Yeh-Chiang<sup>1</sup>, Yehuda Dar<sup>2</sup>, Richard Baraniuk<sup>2</sup>, Micah Goldblum<sup>3</sup>, Tom Goldstein<sup>1</sup> <sup>1</sup>University of Maryland, College Park <sup>2</sup>Rice University <sup>3</sup>New York University

CVPR 2022



# Preliminary



**Lemma 2.1** Let  $f : [0,1]^n \rightarrow [0,1]$  be a neural network satisfying  $|f(x) - f(y)| \le \frac{L}{\sqrt{n}} ||x - y||$ . Let  $\overline{f}$  denote the median value of f on the unit hypercube. Then, for an image  $x \in [0,1]^n$  of uniform random pixels, we have  $|f(x) - \overline{f}| \le t$ with probability at least

$$1-\frac{Le^{-2\pi nt^2/L^2}}{\pi t\sqrt{n}}.$$

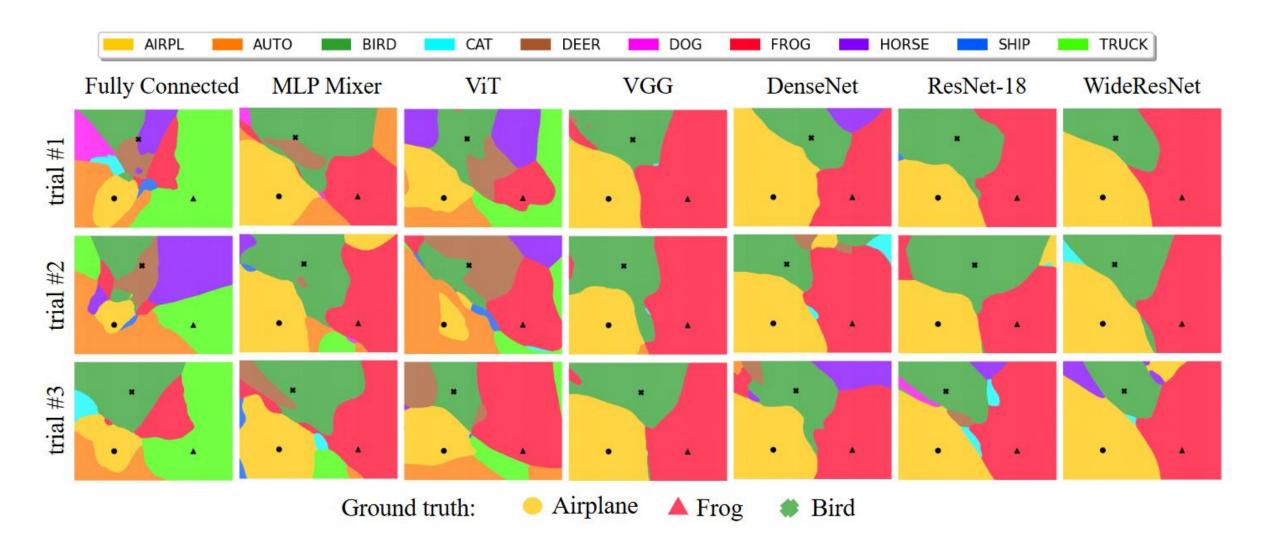
2



# **Decision Boundary**

We take a page from the mixup playbook and plot decision boundaries along the convex hull between data samples. We first sample a triplet  $(x_1, x_2, x_3) \sim D^3$  of *i*. *i*. *d*. images from the distribution D. Then, we construct the plane spanned by the vectors  $\overrightarrow{v_1} = x_2 - x_1$ ,  $\overrightarrow{v_2} = x_3 - x_1$  and plot the decision boundaries in this plane. To be precise, we sample inputs to the network with coordinates  $\alpha \cdot max(\overrightarrow{v_1} \cdot \overrightarrow{v_2}, |proj_{\overrightarrow{v_1}} \cdot \overrightarrow{v_2}|) \overrightarrow{v_1} + \beta(\overrightarrow{v_2} - proj_{\overrightarrow{v_1}} \overrightarrow{v_2})$ for  $-0.1 \leq \alpha, \beta \leq 1.1$ 





4



## **Reproducibility Score**

$R(\theta_1, \theta_2) = \mathbb{E}_{T_i \sim \mathcal{D}}$	$\left[ ( f(S_i, \theta_1) \cap f(S_i, \theta_2) ) /  S_i  \right]$
---	---

			-						-	
WideRN30	0.87	0.85	0.85	0.82	0.81	0.78	0.63	0.61	0.45	-0.85
WideRN20	0.85	0.86	0.85	0.82	0.81	0.78	0.63	0.6	0.44	-0.80
WideRN10	0.85	0.85	0.86	0.81	0.81	0.78	0.63	0.6	0.44	-0.75 e.
ResNet18	0.82	0.82	0.81	0.83	0.81	0.78	0.63	0.61	0.45	- 0.70 <sup>S</sup>
DenseNet	0.81	0.81	0.81	0.81	0.82	0.77	0.64	0.61	0.44	-0.65 cipili
VGG	0.78	0.78	0.78	0.78	0.77	0.79	0.63	0.6	0.44	Reproducibility
ViT	0.63	0.63	0.63	0.63	0.64	0.63	0.75	0.64	0.47	-0.55
MLPMixer	0.61	0.6	0.6	0.61	0.61	0.6	0.64	0.67	0.46	-0.50
FullyCon	0.45	0.44	0.44	0.45	0.44	0.44	0.47	0.46	0.69	-0.45
2	Wide RLAS	WideRX2	WideR 11	Restant	Densetlet	100	Sit .	MPNIXE	FullyCon	



Distillation					Vanilla Training					_
0.87	0.87	0.86	0.85	0.77	0.82	0.82	0.81	0.78	0.75	-0.85 ല
0.86	0.86	0.86	0.82	0.75	0.82	0.84	0.82	0.78	0.76	ty Score
0.86	0.86	0.87	0.82	0.75	0.81	0.82	0.83	0.78	0.75	- 0.75 - 0.70 - 0.70
0.81	0.81	0.81	0.82	0.73	0.78	0.78	0.78	0.8	0.73	- 0.70
0.76	0.76	0.76	0.75	0.72	0.75	0.76	0.75	0.73	*	
Distillation Student Vanilla Trained Model										
	0.86 0.86 0.81 0.76 Restvet	0.87 0.87 0.86 0.86 0.86 0.86 0.81 0.81 0.76 0.76	0.87 0.87 0.86   0.86 0.86 0.86   0.86 0.86 0.87   0.81 0.81 0.81   0.76 0.76 0.76	0.870.860.850.860.870.820.860.860.870.820.810.810.810.820.760.760.750.75	0.870.860.850.770.860.860.860.820.750.860.860.870.820.750.810.810.810.820.730.760.760.760.750.72	0.87 0.87 0.86 0.85 0.77 0.82   0.86 0.86 0.82 0.75 0.82   0.86 0.86 0.82 0.75 0.82   0.86 0.86 0.87 0.82 0.75 0.82   0.86 0.86 0.87 0.82 0.75 0.81   0.81 0.81 0.82 0.75 0.78 0.78   0.76 0.76 0.76 0.75 0.72 0.75   Restrict Denservet VGG Vitter   Vitter Denservet VGG Vitter	0.87 0.87 0.86 0.85 0.77 0.82 0.82   0.86 0.86 0.82 0.75 0.82 0.84   0.86 0.86 0.82 0.75 0.82 0.84   0.86 0.86 0.87 0.82 0.75 0.81 0.82   0.81 0.81 0.82 0.75 0.81 0.82   0.76 0.76 0.76 0.75 0.72 0.75 0.76   Restrict to the total t	0.87 $0.87$ $0.86$ $0.85$ $0.77$ $0.82$ $0.82$ $0.81$ $0.86$ $0.86$ $0.86$ $0.82$ $0.75$ $0.82$ $0.84$ $0.82$ $0.86$ $0.86$ $0.87$ $0.82$ $0.75$ $0.81$ $0.82$ $0.83$ $0.81$ $0.81$ $0.81$ $0.82$ $0.73$ $0.78$ $0.78$ $0.78$ $0.76$ $0.76$ $0.76$ $0.75$ $0.72$ $0.75$ $0.76$ $0.75$ Restlict to the part of the	0.87 0.87 0.86 0.85 0.77 0.82 0.82 0.81 0.78   0.86 0.86 0.86 0.82 0.75 0.82 0.84 0.82 0.78   0.86 0.86 0.87 0.82 0.75 0.82 0.84 0.82 0.78   0.86 0.86 0.87 0.82 0.75 0.81 0.82 0.78   0.86 0.86 0.87 0.82 0.75 0.81 0.82 0.83 0.78   0.81 0.81 0.81 0.82 0.73 0.78 0.78 0.78 0.81   0.76 0.76 0.76 0.75 0.72 0.75 0.76 0.75 0.73   Restrict of the persentet VGG   Wither persentet VGG	$0.87$ $0.87$ $0.86$ $0.85$ $0.77$ $0.82$ $0.82$ $0.81$ $0.78$ $0.75$ $0.86$ $0.86$ $0.86$ $0.82$ $0.75$ $0.82$ $0.84$ $0.82$ $0.78$ $0.76$ $0.86$ $0.86$ $0.87$ $0.82$ $0.75$ $0.81$ $0.82$ $0.84$ $0.82$ $0.78$ $0.76$ $0.86$ $0.86$ $0.87$ $0.82$ $0.75$ $0.81$ $0.82$ $0.83$ $0.78$ $0.75$ $0.81$ $0.81$ $0.81$ $0.82$ $0.75$ $0.78$ $0.78$ $0.78$ $0.78$ $0.75$ $0.81$ $0.81$ $0.81$ $0.82$ $0.75$ $0.78$ $0.78$ $0.78$ $0.78$ $0.78$ $0.76$ $0.76$ $0.76$ $0.75$ $0.72$ $0.75$ $0.76$ $0.75$ $0.73$ $\bigstar$ Restrict Viet Viet Viet Viet Viet Viet Viet Vie



#### Reproducibility

	1	~				
	Adam	SGD	SGD + SAM			
ResNet-18	79.81%	83.74%	87.22%			
VGG	81.19%	80.92%	84.21%			
MLPMixer	67.80%	66.51%	68.06%			
VIT	69.55%	75.13%	75.19%			
Test Accuracy						

Test Accuracy								
	Adam	SGD	SGD + SAM					
ResNet-18	93.04	95.30	95.68					
VGG	92.87	93.13	93.90					
MLPMixer	82.22	82.04	82.18					
VIT	70.89	75.49	74.72					



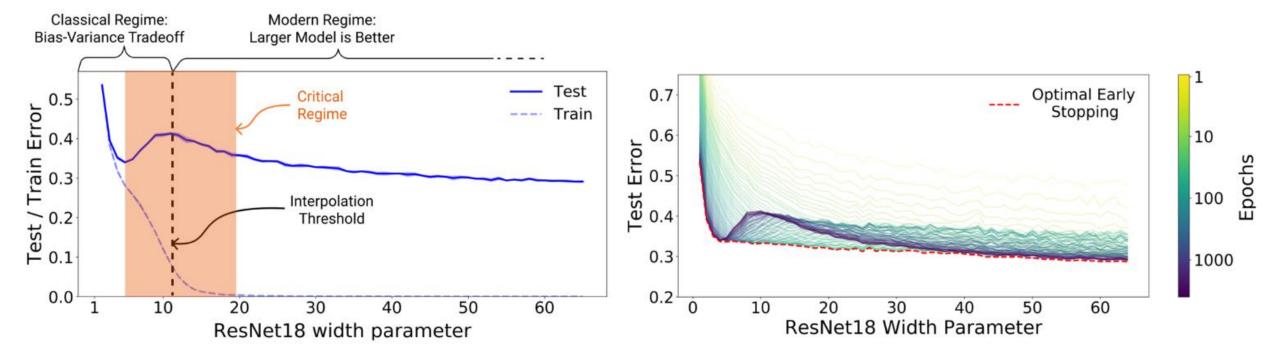
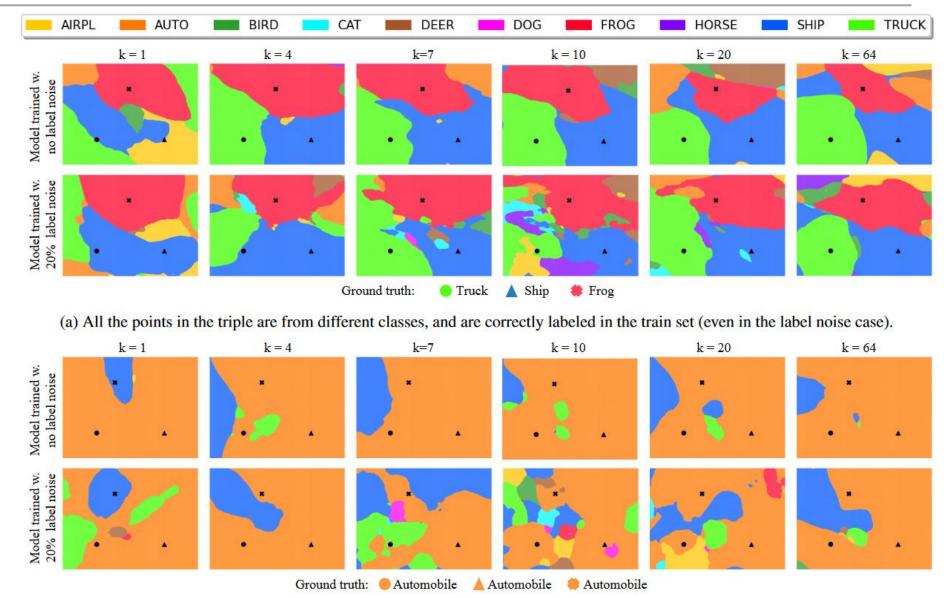


Figure 1: Left: Train and test error as a function of model size, for ResNet18s of varying width on CIFAR-10 with 15% label noise. **Right:** Test error, shown for varying train epochs. All models trained using Adam for 4K epochs. The largest model (width 64) corresponds to standard ResNet18.



### **Double Descent**



(b) All points in the triple are from the same class, Automobile, and are correctly labeled in the train set (even in the label noise case).



#### k=10 for all the plots

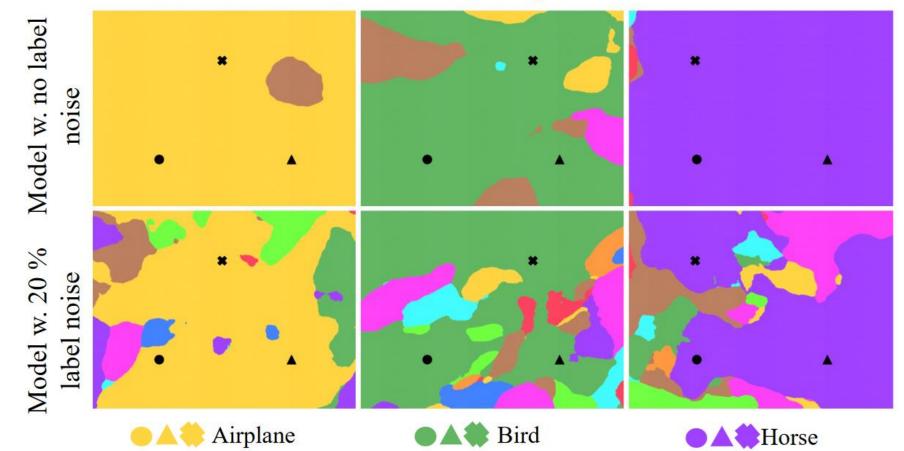


Figure 8. Decision boundaries of 3 correctly labeled points at k = 10 on models with and without label noise.

# **Double Descent**



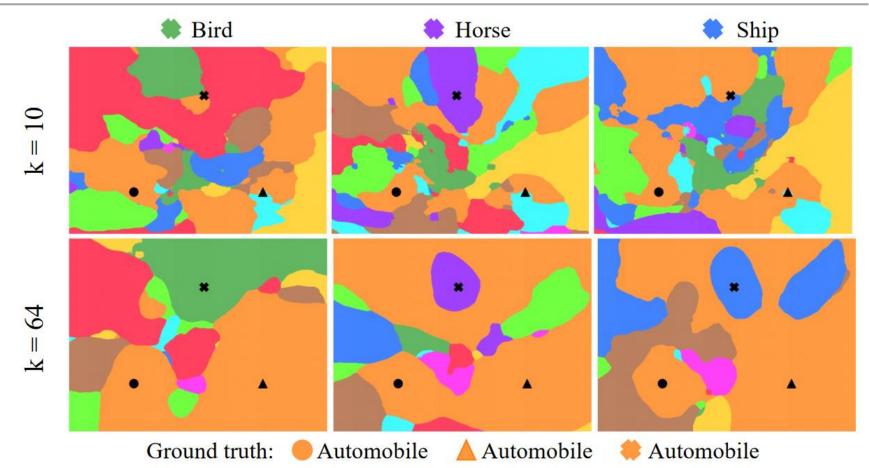


Figure 9. Decision boundaries with 1 mislabeled automobile and 2 correctly labeled automobiles. Each column represents a different image triplet. The mislabeled point is marked by x.

# **Fragmentation Score**



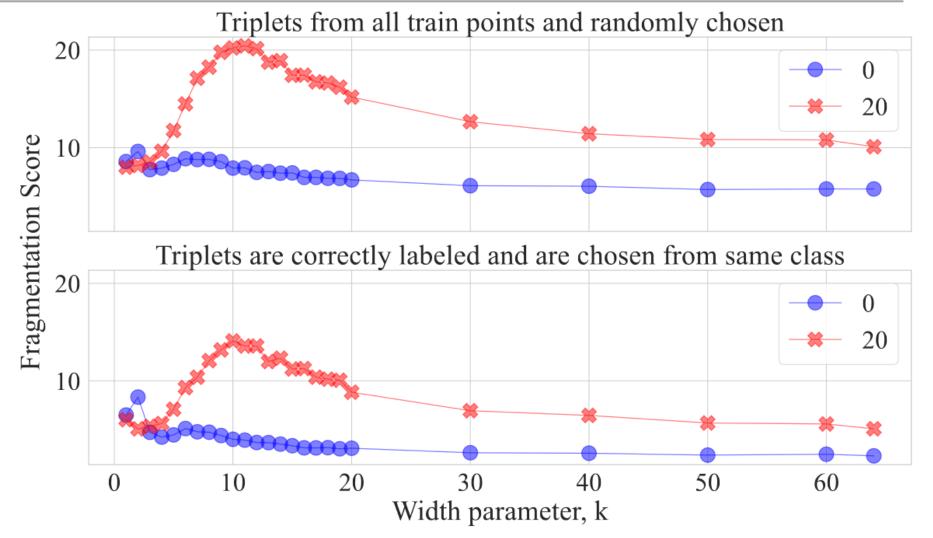


Figure 10. Fragmentation scores as a function of model width for models trained with and without label noise.



# **Reproducibility Score**

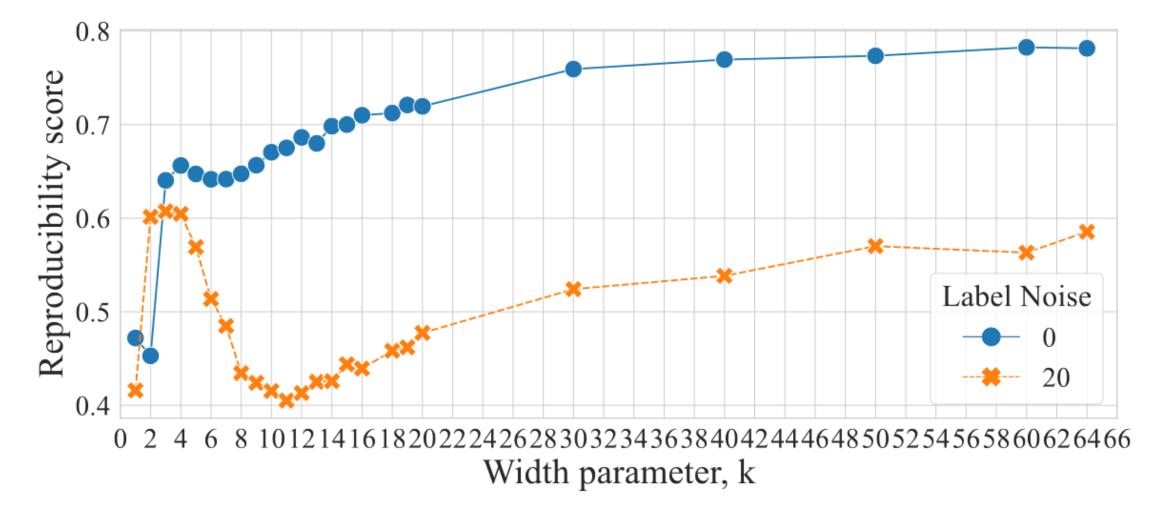


Figure 11. Reproducibility with respect to random initialization for models of different widths.

# **Median Margins**



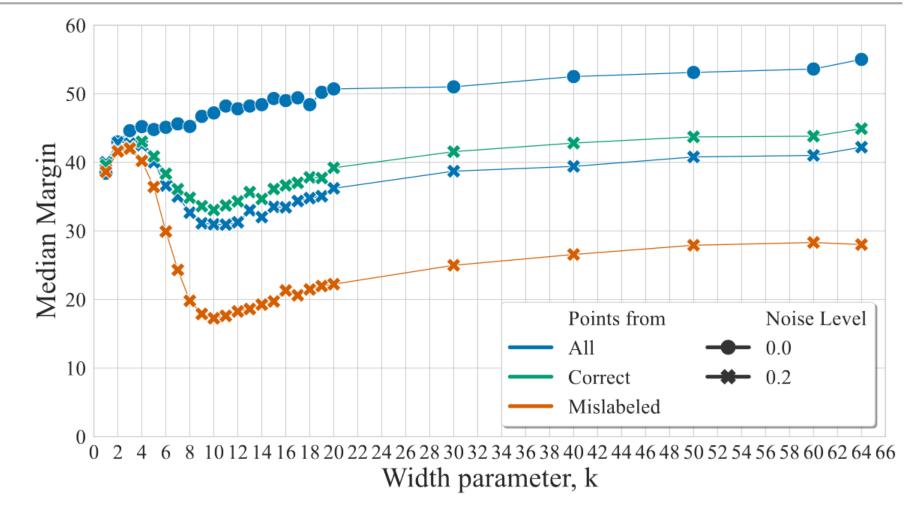


Figure 12. Median Margins - models with and without label noise. Y-axis reflects the average perturbation size needed to reach decision boundary in a random direction.