Can Neural Nets Learn the Same Model Twice? Investigating Reproducibility and Double Descent from the Decision Boundary Perspective

Supplementary Material

A. Proof of Lemma 2.1

For clarity, we restate the lemma here.

Lemma 2.1 *Let $f : [0,1]^n \to [0,1]$ be a neural network satisfying $|f(x) - f(y)| \leq \frac{L}{\sqrt{n}}\|x - y\|$. Let $\bar{f}$ 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 $|f(x) - \bar{f}| \leq t$ with probability at least*

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

Consider a set $\mathcal{A} \subset [0,1]^n$, and let d denote the ℓ_2 distance metric. We define the ϵ-expansion of the set $\mathcal{A}$ as $\mathcal{A}(\epsilon) = \{x \in [0,1]^n \mid d(x, \mathcal{A}) \leq \epsilon\}$. In plain words, $\mathcal{A}(\epsilon)$ is the set of all points lying within ϵ units of the set $\mathcal{A}$.

Our proof will make use of the isoperimetric inequality first presented by Ledoux [24]. We use the following variant with tighter constants proved by Shafahi et al. in [31].

Lemma A.1 (Isoperimetric inequality on the unit cube) *Consider a measurable subset of the cube $\mathcal{A} \subset [0,1]^n$, and a 2-norm distance metric $d(x,y) = \|x - y\|_2$. Let $\Phi(z) = (2\pi)^{-\frac{1}{2}} \int_{-\infty}^{z} e^{-t^2/2} dt$, and let α be the scalar that satisfies $\Phi(\alpha) = vol[\mathcal{A}]$. Then*

$$vol[\mathcal{A}(\epsilon)] \geq \Phi\left(\alpha + \epsilon\sqrt{2\pi}\right). \tag{3}$$

In particular, if $vol(\mathcal{A}) \geq 1/2$, then we simply have

$$vol[\mathcal{A}(\epsilon)] \geq 1 - \frac{e^{-2\pi\epsilon^2}}{2\pi\epsilon}. \tag{4}$$

To prove Lemma 2.1, we start by choosing $\mathcal{A} = \{x | f(x) \leq \bar{f}\}$. Now, consider any $x \in \mathcal{A}\left(t\frac{\sqrt{n}}{L}\right)$. From the Lipschitz bound on f we have

$$|f(x) - f(y)| \leq \frac{L}{\sqrt{n}}\|x - y\|,$$

for any y. If we choose $y = \arg\min_{z \in \mathcal{A}} \|x - z\|$ to be the closest point to x in the set $\mathcal{A}$, we have that $\|z - y\| \leq t\frac{\sqrt{n}}{L}$, and so

$$|f(x) - f(y)| \leq t.$$

But $f(y) \leq \bar{f}$ because $y \in \mathcal{A}$. From this, we see that for any choice of $x \in \mathcal{A}\left(t\frac{\sqrt{n}}{L}\right)$ we have

$$f(x) - \bar{f} \leq t. \tag{5}$$

Recall that $\bar{f}$ is the median value of f on the unit cube, and so we have that

$$\text{vol}[\mathcal{A}] \geq \frac{1}{2}.$$

We can then apply Lemma A.1 with $\epsilon = t\frac{\sqrt{n}}{L}$, and we see that

$$\text{vol}\left[\mathcal{A}\left(t\frac{\sqrt{n}}{L}\right)\right] \geq 1 - \frac{Le^{-2\pi t^2 n/L^2}}{2\pi t\sqrt{n}}.$$

We conclude that a *randomly* chosen $x \in [0,1]^n$ will lie in $\mathcal{A}\left(\frac{t\sqrt{n}}{L}\right)$, and therefore satisfy (5) with probability at least $1 - \frac{Le^{-2\pi t^2 n/L^2}}{2\pi t\sqrt{n}}$.

An analogous argument with $\mathcal{A} = \{x | f(x) \geq \bar{f}\}$ shows that a randomly chosen $x \in [0,1]^n$ will satisfy

$$\bar{f} - f(x) \leq t. \tag{6}$$

with the same probability. Applying a union bound, we see that a randomly chosen x will satisfy (5) and (6) simultaneously with probability at least $1 - \frac{Le^{-\pi t^2 n/L^2}}{\pi t\sqrt{n}}$.

B. Decision regions

Off-manifold decision regions We present a few off-manifold decision boundaries in this section. In Fig. 11, we show decision regions of multiple off manifold images where all the pixels are uniformly sampled in the image space. Each row is a model, and each column is a randomly sampled triplet. We observe that the decision regions assigned to such off-manifold images are quite uniform for a given model. For example, in DenseNet, all such images are assigned to *Bird* class, while in ViT, they are assigned to *Frog* or *Automobile*. In Fig. 12, we show decision regions for a multiple triplets of shuffled images. (Expanded version of Fig.2). Even in this type of off-manifold images, we see a similar pattern that the models are assigning the samplings to a certain set of classes. This emphasises that the decision regions are more structured close to the image manifold and are rather uniform farther away from the manifold.

C. Additional Reproducibility results

Region similarity score ablation We performed an additional ablation to check if the region similarity score is correlated with prediction similarity (Intersection over union of predictions on a given set of points) on augmented test data across models trained on different seeds. So we repeated

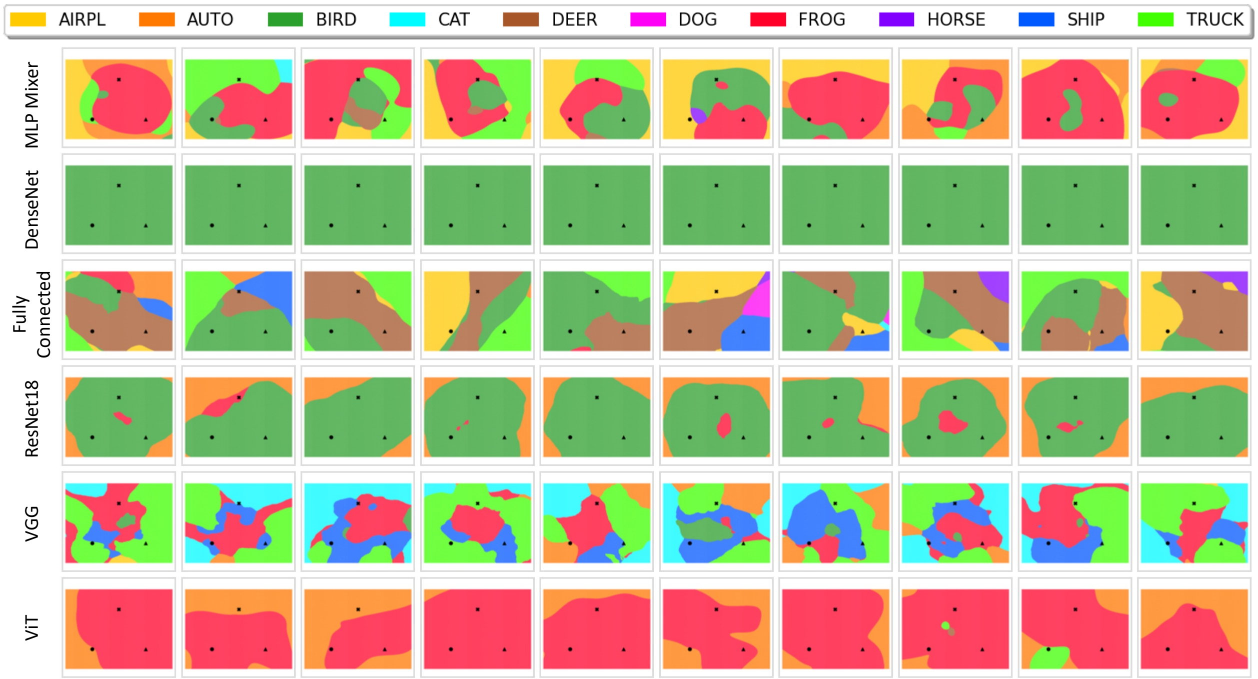


Figure 11. Decision regions when all the images are uniformly sampled. Each row corresponds to a model,while each column is a new sampling of the triplet

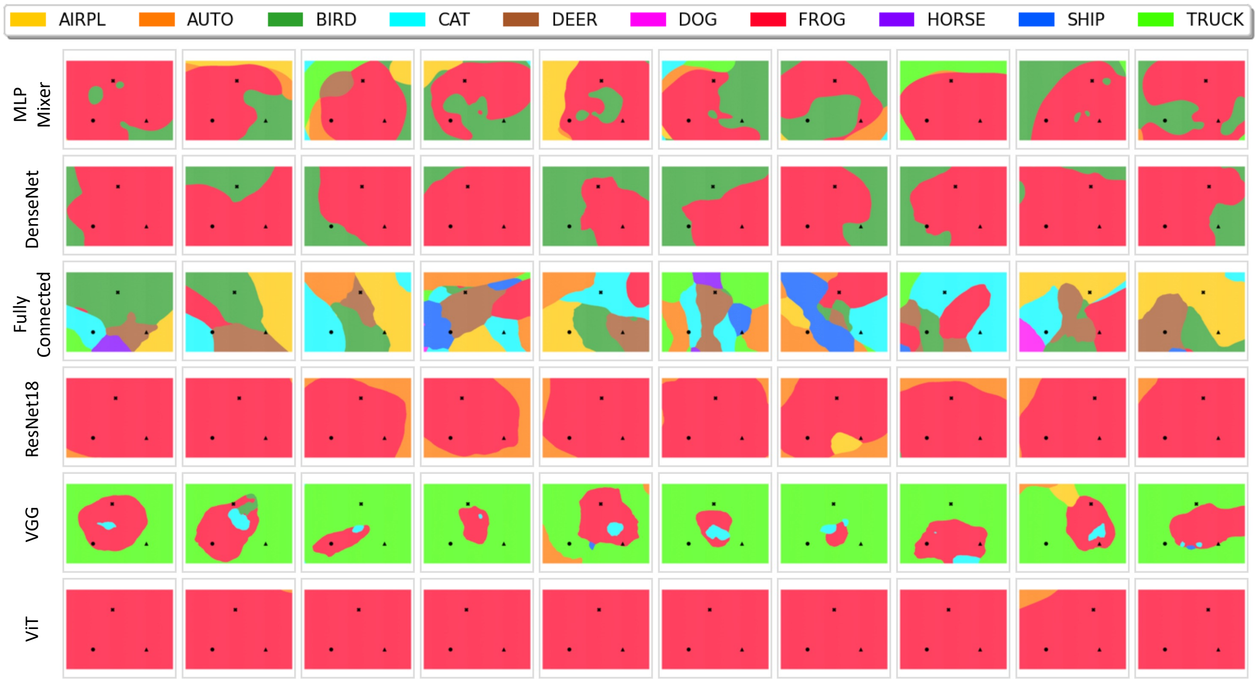


Figure 12. Decision regions when the pixels are randomly shuffled. Each row corresponds to a model,while each column is a new sampling of the triplet. Extended version of Fig 2.

Model	Test data	Aug. test data	Original
WideResNet-30	96.24	84.79	86.61
ResNet-18	95.96	84.02	83.74
ViT	86.94	78.44	75.13
MLP-Mixer	82.55	69.89	66.51

Table 2. Scores are computed across 3 trained models (with different init seeds and SGD) for each architecture.

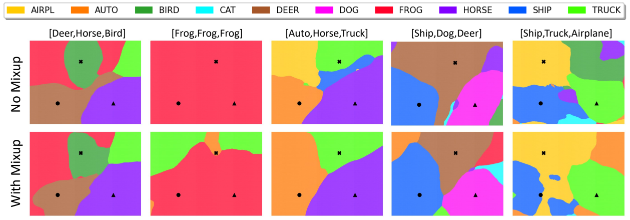


Figure 13. We present decision regions for random triplets sampled from the training set for ResNet18. We see the decision regions are almost same with and without mixup.

our analysis, but without interpolated planes, on 4 models with test images and randomly augmented test images (one augmentation per image - 100k images). Augmentations include flips, crops, perspectives, rotations, Gaussian blur, color-jitter, and random contrast. Results are in Table 2 with 2 scores, one is IoU on the test data (10k images) predictions, and the second score is IoU calculated on augmented data (100k images). The relative ordering of the networks is preserved exactly by all three metrics.

With and without Mixup in training In order to understand how having mixup in the training affects the decision boundaries, we examined 2 cases, ResNet18 and Vision Transformer. In Fig. 13, we show 5 randomly sampled triplets and their decision regions produced by ResNet18 trained with and without mixup. We can see there is a slight difference, but not quite significant. We quantified how “similar” the decision surfaces are with region similarity score introduced in Section 3.2. The score for Resnet18 is 0.774, and for ViT is 0.808.

D. Additional Double Descent results

Additional plots at $k = 10$ As referenced in main, in Figure 14, we plotted decision regions of points sampled from the same class and are given correct labels (even in label noise scenario). We can see that when there is no label noise there are a few fragmented class regions and the regions explode when the model is trained with label noise.

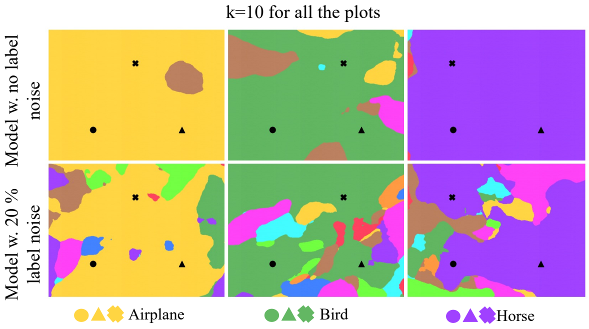


Figure 14. Decision boundaries of 3 correctly labeled points at $k = 10$ on models with and without label noise.

Margin plots We computed median margins for 1000 train data points (averaged over 5 random directions for each point) as shown in Figure 15. We see that the overall margins drop by the introduction of label noise in training. In label noise case, we also show how margins differ for correctly labeled and mislabeled points. We see mislabeled points have significatly lower margins than correctly labeled points, however the margins increase for both types of points as we increase the model capacity (i.e. k in this case).

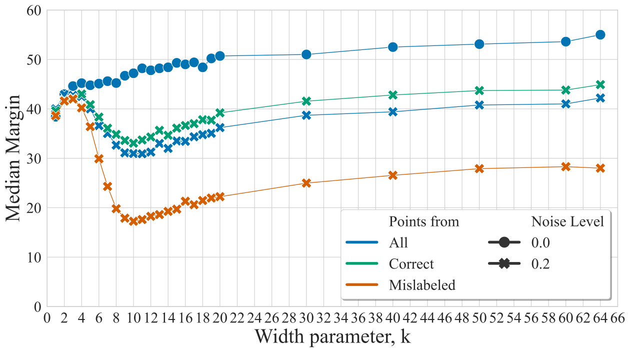


Figure 15. Median Margins - models with and without label noise. Y-axis reflects the average perturbation size needed to reach decision boundary in a random direction.

Additional error plots In Fig. 6, we have seen how the test errors change as we progressively increase the model capacity. Figure 16 shows how training errors change in addition to test errors. We can see that the train error reaches 0 at much higher k with label noise than without. In model without label noise, the interpolation begins at $k = 10$ which is the true interpolation threshold when there is no label noise. We further examine how correctly labeled and mislabeled points are behaving in Figure 17. The green lines represent the overall train error, while orange shows the error on correctly labeled points. The mislabeled points are shown in grey, and the error is computed as incorrect

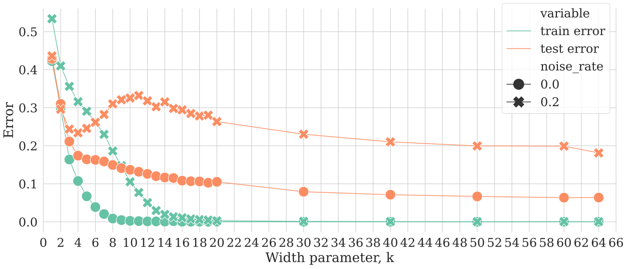


Figure 16. In this figure we show the train and test errors with and without label noise.

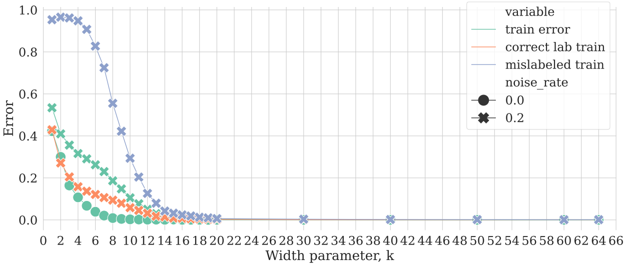


Figure 17. In this figure, we show the train data errors for with and without label noise cases in green color. We also investigate how the errors are changing for correctly labeled points (orange curve) and in mislabeled points (grey curve).

predictions with respect to assigned class. We see that till $k = 4$ the mislabeled points are not fit to their assigned class which partially explains the low test error of test data. However at $k = 10$ most of the correctly labeled points are fit while some of the mislabeled points are still not fit to their assigned class. This trend diverges from what is seen in simple model families where the second peak of test error coincides with the model capacity with 0 training error. This shows that double descent in more complicated in neural-network architectures than what is seen in simple linear models.

Region similarity scores from random data sampling In Figure 10, we have seen how decision boundaries change when we compare two runs of the same model architecture with different initializations. In Figure 18, we show how the ordering of the data changes the decision boundaries. We see that the region similarity or reproducibility across training runs is high in the under- and over-parametrized regimes, but it drops drastically closer to the interpolation threshold. This is the exact same behaviour observed in Figure 10. This shows that $k = 10$ is a quite unstable with respect to different types of variations in the model training.

Additional plots across varying model capacities and noise In Figures 19 and 20 , we show how the decision regions change with and without label noise and with varying model capacities across different samplings of triplets.

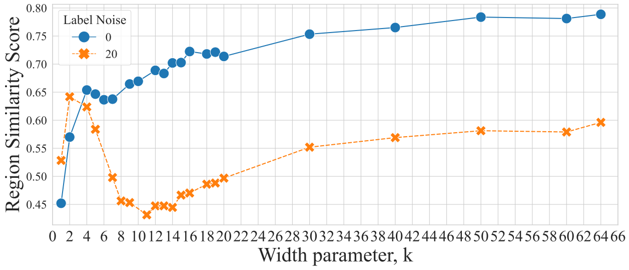
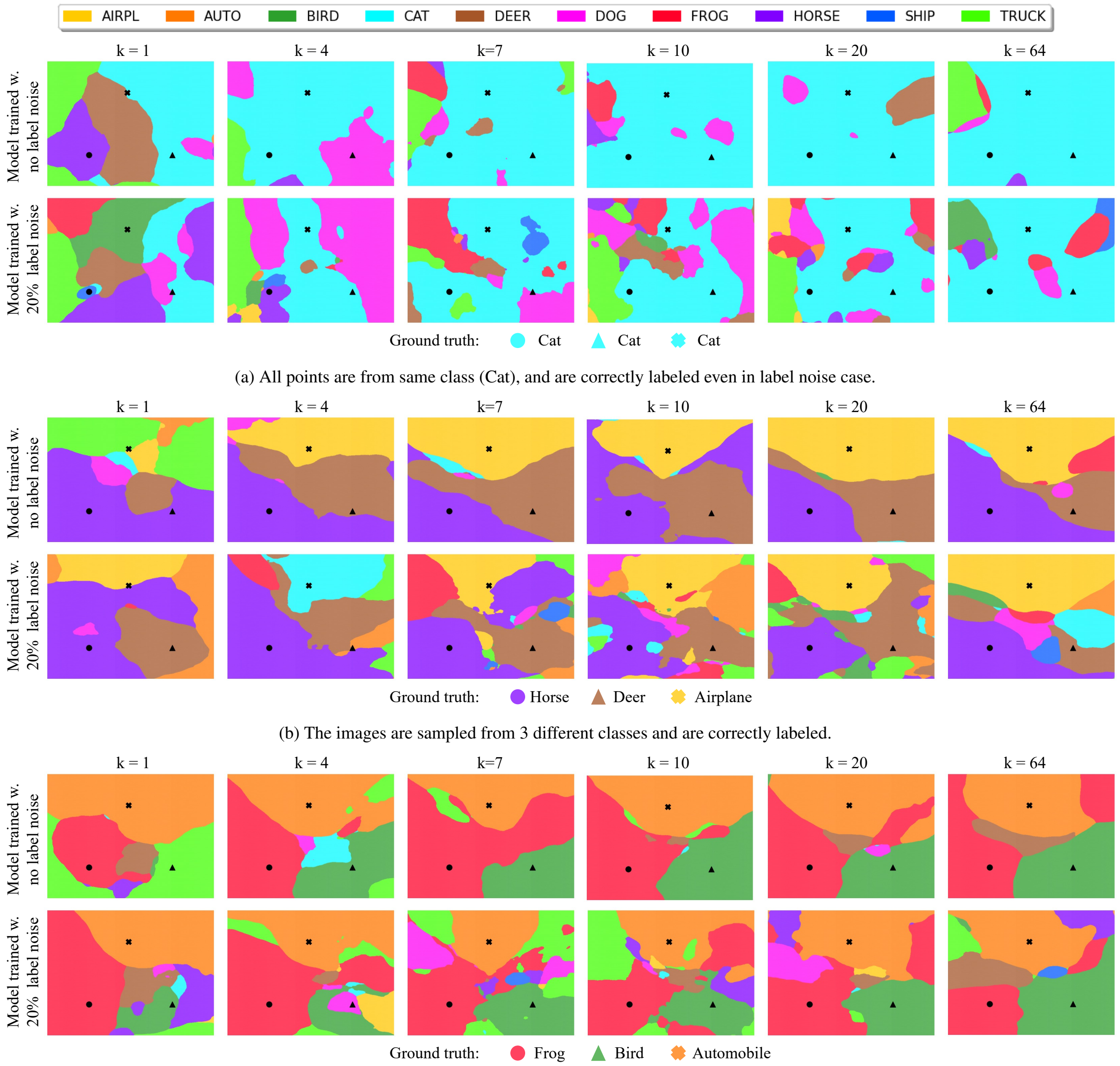


Figure 18. Region similarity with respect to random data samplings for models of different widths

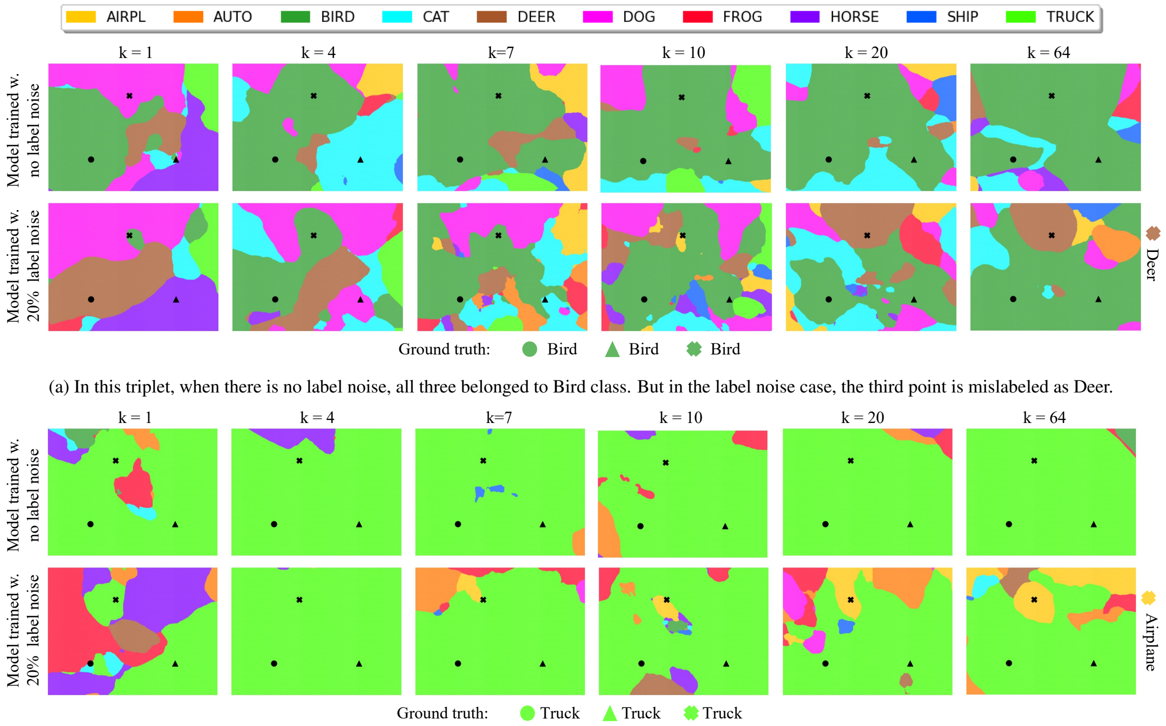


(a) All points are from same class (Cat), and are correctly labeled even in label noise case.

(b) The images are sampled from 3 different classes and are correctly labeled.

(c) The images are sampled from 3 different classes and are correctly labeled. Additional case

Figure 19. **Decision boundaries for models of varying width.** We show additional decision surfaces with different types of triplets here. All points are correctly labeled



(a) In this triplet, when there is no label noise, all three belonged to Bird class. But in the label noise case, the third point is mislabeled as Deer.

(b) In this triplet, when there is no label noise, all three belonged to Truck class. But in the label noise case, the third point is mislabeled as Airplane.

Figure 20. **Decision boundaries for models of varying width.** We show decision surfaces with different types of triplets here. One point is mislabeled.