A Additional Experimental Results

A.1 Ablation Studies

Sensitivity to *p* We first evaluate noisy ImageNet classification with varying p. A higher p includes more clean examples at the cost of involving more noisy examples. From Figure 8, ODD is not very sensitive to p, and empirically $p = 10$ represents the best trade-off.

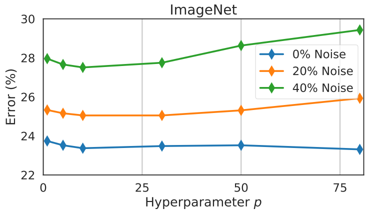

Fig. 8. Ablation studies over the hyperparameter p on ImageNet under different levels of mislabeled examples.

Sensitivity to *E* We evaluate the validation error of ODD on CIFAR10 with 20% and 40% input-agnoistic label noise where $E \in \{25, 50, 75, 100, 150, 200\}$ ($E = 200$ is equivalent to ERM). The results in Figure 9 suggest that our method is able to separate noisy and clean examples if E is relatively small where the learning rate is high, but is unable to perform well when the learning rate decreases.

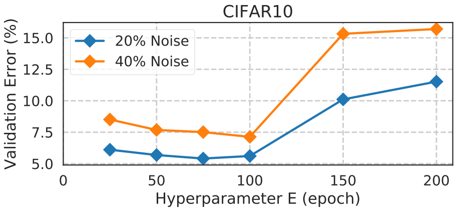

Fig. 9. Validation errors of ODD on CIFAR10 with different values of E.

Sensitivity to the amount of noise Finally, we evaluate the training error of ODD on CIFAR10 under input-agnostic label noise of $\{1\%, 5\%, 10\%, 20\%, 30\%, 40\%\}$ with $p = 5$, $E = 50$ or 75. This reflects how much examples exceed the threshold and are identified as noise at epoch E. From Figure 10, we observe that the training error is almost exactly the amount of noise in the dataset, which demonstrates that the loss distribution of noise can be characterized by our threshold regardless of the percentage of noise in the dataset.

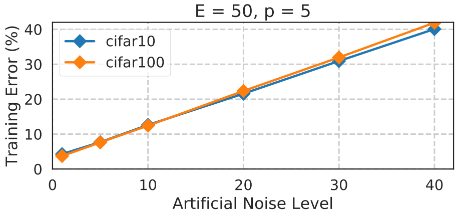

Fig. 10. Training errors of ODD on CIFAR10 with different amount of uniform noise.

Precision and recall for classifying noise We evaluate precision and recall for examples classified as noise on CIFAR10 and CIFAR100 for different noise levels (1, 5, 10, 20, 30, 40) in Figure 11. The recall values are around 0.84 to 0.88 where as the precision values range from 0.88 to 0.92. This demonstrates that ODD is able to achieve good precision/recall with default hyperparameters even at different noise levels.

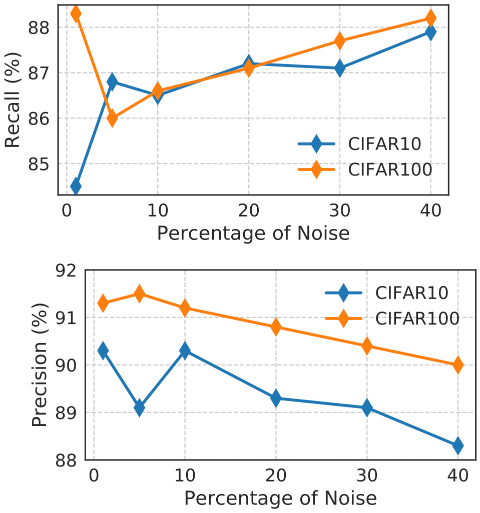

Fig. 11. Recall and precision for ODD on CIFAR10 and CIFAR100 with different levels of uniform random noise.

Percentage of samples discared by ODD We show the percentage of examples discarded by Noise Classifier in Table 6; the percentage of discarded examples by $p = 10$ is very close to the actual noise level, suggesting that it is a reasonable setting.

Table 6. Percentage of example discraded by ODD on ImageNet-2012.

% Mislabeled	Hyperparameter p					Network
	1	10	30	50	80	
0%	6.2	2.6	1.1	0.7	0.5	ResNet-50
20%	24.4	21.1	19.3	17.6	11.5	
40%	44.8	40.3	36.2	28.1	7.8	

Ablation studies on WebVision We include additional ablation on p for WebVision (Table 7), where we consider different levels of p values.

Table 7. Additional results on WebVision with varying p.

	Webvision		ImageNet	
p	Top1	Top5	Top1	Top 5
1	71.97	88.55	65.32	84.95
10	72.55	88.93	65.53	85.20
30	72.51	89.21	65.52	85.25
50	72.41	89.23	65.60	85.19
80	72.47	89.18	65.73	85.19

A.2 Images in CIFAR-100 Classified as Noise

We display the examples in CIFAR-100 training set for which our ODD methods identify as noise across 3 random seeds. One of the most common label such examples have is “leopard”; in fact, 21 of 50 “leopard” examples in the training set are perceived as hard, and we show some of them in Figure 12. It turns out that a lot of the “leopard” examples contains images that clearly contains tigers and black panthers (CIFAR-100 has a label corresponding to “tiger”).

We also demonstrate random examples from the CIFAR-100 that are identified as noise in Figure 13 and those that are not identified as noise in Figure 14. The examples identified as noise often contains multiple objects, and those not identified as noise often contains only one object that is less ambiguous in terms of identity. Therefore, when the datasets contains all “clean” examples, ODD would tend to discard examples that are hard to learn well.

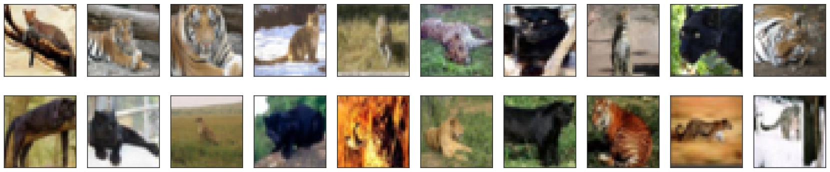

Fig. 12. Examples with label “leopard” that are classified as noise.

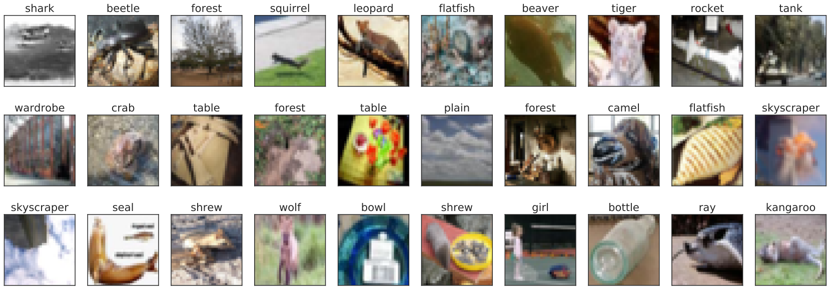

Fig. 13. Random CIFAR-100 examples that are classified as noise.

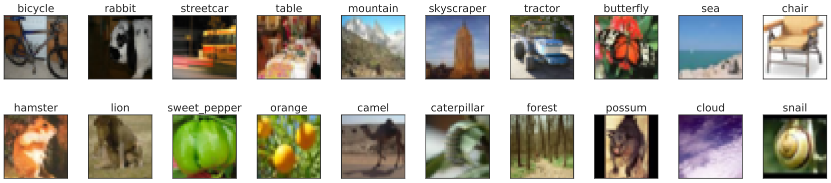

Fig. 14. Random CIFAR-100 examples that are not classified as noise.