Using Pre-Training Can Improve Model Robustness and Uncertainty Supplementary Material

Dan Hendrycks [1] **Kimin Lee** [2] **Mantas Mazeika** [3]

1. CIFAR-10-Related Classes Excluded from Downsampled ImageNet

The ImageNet-1K classes that are related to CIFAR-10 are as follows:

n03345487, n03417042, n04461696, n04467665, n02965783, n02974003, n01514668, n01514859, n01518878, n01530575, n01531178, n01532829, n01534433, n01537544, n01558993, n01560419, n01580077, n01582220, n01592084, n01601694, n01608432, n01614925, n01616318, n01622779, n02123045, n02123159, n02123394, n02123597, n02124075, n02085620, n02085782, n02085936, n02086079, n02086240, n02086646, n02086910, n02087046, n02087394, n02088094, n02088238, n02088364, n02088466, n02088632, n02089078, n02089867, n02089973, n02090379, n02090622, n02090721, n02091032, n02091134, n02091244, n02091467, n02091635, n02091831, n02092002, n02092339, n02093256, n02093428, n02093647, n02093754, n02093859, n02093991, n02094114, n02094258, n02094433, n02095314, n02095570, n02095889, n02096051, n02096177, n02096294, n02096437, n02096585, n02097047, n02097130, n02097209, n02097298, n02097474, n02097658, n02098105, n02098286, n02098413, n02099267, n02099429, n02099601, n02099712, n02099849, n02100236, n02100583, n02100735, n02100877, n02101006, n02101388, n02101556, n02102040, n02102177, n02102318, n02102480, n02102973, n02104029, n02104365, n02105056, n02105162, n02105251, n02105412, n02105505, n02105641, n02105855, n02106030, n02106166, n02106382, n02106550, n02106662, n02107142, n02107312, n02107574, n02107683, n02107908, n02108000, n02108089, n02108422, n02108551, n02108915, n02109047, n02109525, n02109961, n02110063, n02110185, n02110341, n02110627, n02110806, n02110958, n02111129, n02111277, n02111500, n02111889, n02112018, n02112137, n02112350, n02112706, n02113023, n02113186, n02113624, n02113712, n02113799, n02113978, n01641577, n01644373, n01644900, n03538406, n03095699, n03947888.

We chose these by using the WordNet hierarchy, though different class selections are conceivable.

2. Evaluating Adversarial Robustness with Random Restarts

PGD attacks with multiple random restarts and more iterations make for stronger adversaries. However, at the time of writing it is currently not standard to evaluate models with random restarts, even though restarts have been shown to reduce accuracy of adversarially trained models (Mosbach et al., 2018). Other models completely break against adversaries using 100 steps. For completeness we include results with random restarts and more steps. An external evaluation by Maksym Andriushchenko of our adversarially pre-trained CIFAR-10 model found that using 100 steps and 1,000 random restarts reduces accuracy to 52.9%. Compared to the baseline of normal training evaluated with a weaker adversary, adversarial pre-training remains 7.1% higher in absolute accuracy. While it is not standard to evaluate with multiple random restarts, it is currently standard to evaluate with adversaries which take many steps. Adversaries with 100 steps hardly affect the model's accuracy, in that accuracy changes by only 0.2% from 57.4% to 57.2%.

3. Full Out-of-Distribution Detection Results

We use the problem setup from Hendrycks et al. (2019). We use various datasets such as Gaussian Noise, Rademacher Noise, etc. Note that the ImageNet-21K dataset is the ImageNet-22K dataset with the ImageNet-1K classes held out. Results are in Table 1.

Table 1: Out-of-distribution example detection for the maximum softmax probability baseline detector and the MSP detector after pre-training. All results are percentages and an average of 5 runs.

		AUROC ↑		AUPR ↑	
$\mathcal{D}_{\text{in}}$	$\mathcal{D}_{\text{out}}^{\text{test}}$	Normal	Pre-Training	Normal	Pre-Training
Tiny ImageNet	Gaussian	49.4	67.4	15.2	21.1
	Rademacher	70.7	75.0	23.0	25.5
	Blobs	76.2	69.5	28.2	23.1
	Textures	68.7	72.4	29.5	31.8
	SVHN	86.6	89.1	53.2	58.8
	Places365	76.8	74.6	36.8	31.8
	LSUN	73.2	71.6	30.4	27.4
	ImageNet-21K	72.7	71.7	29.9	28.5
	Mean	71.8	**73.9**	30.8	**31.0**
CIFAR-10	Gaussian	96.2	96.7	70.5	73.1
	Rademacher	97.5	97.6	79.4	78.4
	Blobs	94.7	97.2	69.0	83.5
	Textures	88.3	93.7	56.6	70.4
	SVHN	91.8	95.7	63.7	76.9
	Places365	87.4	91.0	56.1	67.7
	LSUN	88.7	93.7	57.4	72.4
	CIFAR-100	87.1	90.7	54.1	65.4
	Mean	91.5	**94.5**	63.4	**73.5**
CIFAR-100	Gaussian	48.8	96.5	14.6	82.7
	Rademacher	52.3	98.8	15.7	92.5
	Blobs	85.9	89.6	44.9	56.4
	Textures	73.5	79.7	33.1	44.1
	SVHN	74.5	79.6	32.0	48.5
	Places365	74.1	74.6	34.0	34.2
	LSUN	70.5	70.9	28.7	27.7
	CIFAR-10	75.5	75.3	34.5	35.8
	Mean	69.4	**83.1**	29.7	**52.7**

References

Mosbach, M., Andriushchenko, M., Trost, T., Hein, M., and Klakow, D. Logit pairing methods can fool gradient-based attacks. *arXiv preprint arXiv:1810.12042*, 2018.