Table 1: We consider pretrained CIFAR-100 model and compare transfer learning to each of the three downstream tasks with ($\lambda = 0.01$) and without ($\lambda = 0.$) weight matrix regularization. The presented metrics are the row-wise maximal L2 norm of the weight matrix, clean and robust cross-entropy (CE) loss, the absolute difference between both losses, and LHS & RHS (rob. score) of Lemma 1.

	λ	max L2 norm	clean /robust CE	dif.	LHS	RHS
CIFAR-10	0	4.4	0.81 / 1.55	0.74	0.084	0.98
CIFAR-10	0.01	2.1	0.98 / 1.44	0.46	0.107	0.98
FashionMNIST	0	4.6	0.45 / 1.25	0.8	0.087	2.11
FashionMNIST	0.01	2.3	0.62 / 1.13	0.51	0.113	2.11
Intel Image	0	3.1	0.61 / 1.08	0.47	0.077	1.05
Intel Image	0.01	2.2	0.67 / 1.05	0.38	0.087	1.05

Table 2: Comparison of transfer learning metrics when pre-training on a small-scale (CIFAR-10) or large-scale (CIFAR-100) dataset, where the downstream task remains the same (CIFAR-100). The metrics are the same as in Table 1.

	max L2 norm	clean /robust CE	dif.	LHS	RHS
CIFAR-10 $\rightarrow$ CIFAR-100	3.44	3.75 / 4.07	0.32	0.047	0.43
CIFAR-100 $\rightarrow$ CIFAR-100	1.73	2.71 / 3.27	0.56	0.157	1.0

Table 3: Two adversarial pre-training approaches for the CIFAR-100 model evaluated on transfer learning to CIFAR-10. The first row is the version used in our paper and the second approach is denoted by its robustbench model name (Sehwag et al., ICLR, 2022).

pre-training	max L2 norm	clean /robust CE	dif.	LHS	RHS
Paper	4.4	0.81 / 1.55	0.74	0.084	0.98
Sehwag2021Proxy	3.8	0.81 / 2.08	1.27	0.167	2.11

Table 4: Transfer Llarning on an ImageNet pre-trained model with differently resized CIFAR-10 images. The table contains clean and robust CE loss and accuracy, proportion of images fulfilling the theoretical bound (1), LHS and RHS of Lemma 1.

input image	clean /robust CE	clean /robust acc	prop. fulfilled	LHS	RHS
256x256	0.22 / 4.36	92% / 16%	0%	1.14	15.7
128x128	0.41 / 1.31	86% / 57%	0.1%	0.28	12.7
64x64	1.04 / 1.49	64% / 48%	0%	0.14	8.2

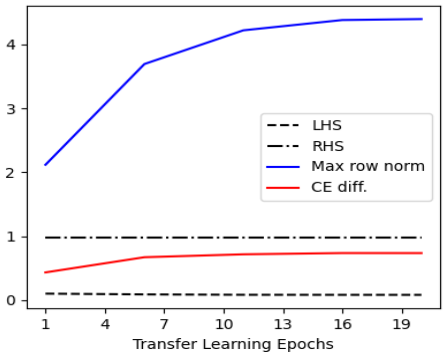

Figure 1: This graphic illustrates the changing robustness level during transfer learning. The blue curve is the row-wise maximal L2 norm of the linear probe's weight matrix. The red curve is the difference between clean and robust cross-entropy (CE) loss, which increases mainly due to the decreasing clean CE. The black lines show the LHS and RHS (robustness score) of Lemma 1.