Supplementary Material: Learning ODIN

Amir Jevnisek
amirjevn@mail.tau.ac.il
Shai Avidan
avidan@eng.tau.ac.il

School of Electrical Engineering
Tel-Aviv University
Tel-Aviv, Israel

1 Models Capacity

We show the number of parameters for various models in Table 1. The first three models are the models demonstrated in our paper for GQ training. The last three models are those mentioned in [7] and [1]. As can be seen, the size of the network ranges from fairly small ones (less than 0.5M parameters) all the way up to networks with more than 20M parameters.

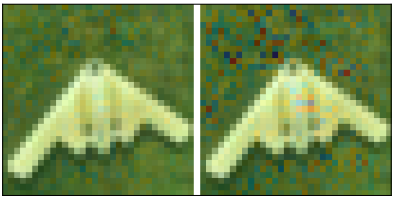
Model	CIFAR-10	CIFAR-100
Madry's ResNet	$467K$	$472K$
ResNet18	$11.17M$	$11.22M$
ResNet34	$21.28M$	$21.32M$
DenseNet-BC-100	$769K$	$800K$
MSDNet	$2.8M$	$2.8M$
WideResNet	$8.9M$	$8.9M$

Table 1: **Models: Number of Parameters** Architectures for CIFAR-10 and CIFAR-100 classification and their respective number of network parameters. We train the top three architectures (Madry's ResNet, ResNet18 and ResNet34) with our novel Gradient Quotient regularization loss. We compare against methods using the rest of the architectures listed on this table.

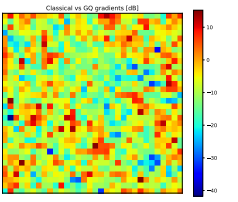
2 $|S_{\hat{y}}(\tilde{x}) - S_{\hat{y}}(x)|$ overlaid on an image

We visualize $|S_{\hat{y}}(\tilde{x}) - S_{\hat{y}}(x)|$ for a particular image (Figure 1(a)). Figure 1(b) shows $\nabla_x S_{\hat{y}}$ overlaid as a heat map on top of the original image. The left side of figure 1(b) is for a ResNet-18 trained with CE-loss on CIFAR-10. The right side of figure 1(b) is for ResNet-18 trained with the CE-loss and $\mathcal{L}_{GQ}$ term. It is clear that when training with $\mathcal{L}_{GQ}$, the gradients $\nabla_x S_{\hat{y}}$ are stronger, as expected. Figure 1(c) shows per-pixel increase of $\nabla_x S_{\hat{y}}$ (in dB) of our method w.r.t to the base CE-loss. As can be seen, most values our positive, indicating that GQ does indeed strengthen $\nabla_x S_{\hat{y}}$.

(a) Original Image



(b) $\nabla_x S_{\hat{y}}$ overlaid on the image. left is training with CE-loss, right training with CE-loss+GQ,



(c) Gradient Amplification per pixel [dB].

Figure 1: **Gradient Visualization** ResNet18 backbone trained on CIFAR-10 once classicaly (solely with CE-loss) and once with CE-loss and GQ. (a) shows the original image, (b) shows the gradient of the estimated logit w.r.t image pixels, overlaid on the original image, (c) shows our per-pixel improvement in [dB].

3 Log-Softmax Gradient w.r.t Image-Pixels Distribution

We show a zoom in on the distributions of $||\nabla_x \log S_{\hat{y}}(x;T)||_1$ in Figure 2. Dashed lines represent distributions for networks trained classically (that is, with CE-loss). Solid lines represent distributions for models trained with CE-loss and our loss term as a regularizer. When our loss term is added, the distribution of $||\nabla_x \log S_{\hat{y}}(x;T=1000)||_1$ for In-Distribution samples is wider than the distribution of the same quantity for classically trained networks. This figure enables the observer to notice that all OOD datasets but one did not have significant support growth. This enables the detection of OOD samples: their $||\nabla_x \log S_{\hat{y}}(x;T=1000)||_1$ is measured, ID samples will have higher scores on average and OOD samples will have lower scores.

4 10 Datasets Benchmark Results

We follow [7] to evaluate AUROC and FPR95 scores on the full 10 datasets evaluation: MNIST [6], K-MNIST [3], fashion-MNIST [9], LSUN (crop) [11], SVHN [8], Textures [2], STL10 [4], Places365 [12], iSUN [10] and LSUN (resize) [11]. We evaluate these scores for all combinations of In-Distribution datasets and models: we consider CIFAR-10 and CIFAR-100 datasets as In-Distribution and Madry's ResNet, ResNet-18 and ResNet-34 models. Scores are summarized in Table 2. The table shows AUROC and FPR95 scores for 10 different OOD datasets and 3 different models.

5 Is GQ sensitive to the tuning parameter λ?

We test the sensitivity of the GQ regularizer to the trade-off parameter λ. Specifically in this case, the experiment is performed on the logit-based flavour of our method (instead of $\mathcal{L}_{GQ}$ evaluated for softmax values, we evaluate it for the logits). As a test point, we take Madry's ResNet on trained on CIFAR-10 with $\mathcal{L}_{GQ}$ as a regularizer. Since this is a high-capacity

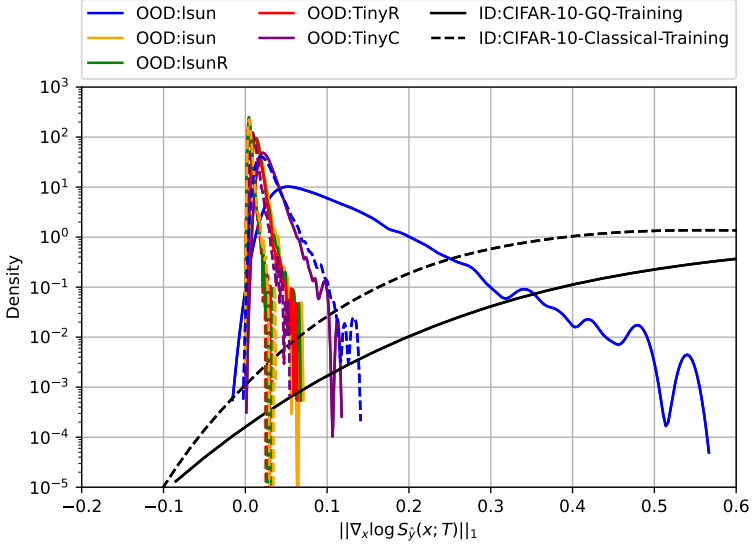


Figure 2: **Log-Softmax Gradient w.r.t Image-Pixels Distribution**. We show the Log-Softmax Gradient w.r.t Image-Pixels Distribution ($||\nabla_x \log S_{\hat{y}}(x;T)||_1$) for two models trained on CIFAR-10: one trained solely with CE-loss (dashed lines) and one trained with our $\mathcal{L}_{GQ}$ loss term (solid lines). The figure shows a Zoom-In on the distribution of $||\nabla_x \log S_{\hat{y}}(x;T = 1000)||_1$. When our loss term is added, the distribution of $||\nabla_x \log S_{\hat{y}}(x;T = 1000)||_1$ for In-Distribution samples is wider than the distribution of $||\nabla_x \log S_{\hat{y}}(x;T = 1000)||_1$ for the In-Distribution samples of the classically trained network. One can observe that all OOD datasets but one did not have significant support growth.

OOD-Dataset $\mathcal{D}_{OOD}^{test}$	Architecture	ID-Dataset $\mathcal{D}_{ID}^{train}$	AUROC ↑	FPR95 ↓	ID-Accuracy ↑ (%)
MNIST	Madry's ResNet	CIFAR-10	99.07	3.31	93.29
		CIFAR-100	94.84	30.56	71.17
	ResNet18	CIFAR-10	99.12	2.69	94.26
		CIFAR-100	88.88	64.92	77.72
	ResNet34	CIFAR-10	99.31	2.06	95.56
		CIFAR-100	78.38	67.31	78.27
K-MNIST	Madry's ResNet	CIFAR-10	98.83	4.68	93.29
		CIFAR-100	95.29	25.16	71.17
	ResNet18	CIFAR-10	99.06	2.58	94.26
		CIFAR-100	92.24	45.79	77.72
	ResNet34	CIFAR-10	99.01	3.71	95.56
		CIFAR-100	81.35	69.53	78.27
fashion-MNIST	Madry's ResNet	CIFAR-10	98.53	7.16	93.29
		CIFAR-100	97.80	11.68	71.17
	ResNet18	CIFAR-10	99.29	1.86	94.26
		CIFAR-100	95.63	25.43	77.72
	ResNet34	CIFAR-10	98.95	4.42	95.56
		CIFAR-100	96.33	18.74	78.27
LSUN (crop)	Madry's ResNet	CIFAR-10	98.81	5.18	93.29
		CIFAR-100	92.94	41.30	71.17
	ResNet18	CIFAR-10	98.99	4.67	94.26
		CIFAR-100	83.37	78.68	77.72
	ResNet34	CIFAR-10	98.56	5.27	95.56
		CIFAR-100	88.07	60.60	78.27
SVHN	Madry's ResNet	CIFAR-10	89.59	5.18	93.29
		CIFAR-100	70.21	94.22	71.17
	ResNet18	CIFAR-10	93.32	38.76	94.26
		CIFAR-100	84.84	72.89	77.72
	ResNet34	CIFAR-10	87.69	39.44	95.56
		CIFAR-100	76.91	87.35	78.27
Textures	Madry's ResNet	CIFAR-10	88.88	48.37	93.29
		CIFAR-100	69.05	89.77	71.17
	ResNet18	CIFAR-10	92.06	39.40	94.26
		CIFAR-100	81.47	74.34	77.47
	ResNet34	CIFAR-10	80.02	47.62	95.56
		CIFAR-100	77.38	73.55	78.27
STL-10	Madry's ResNet	CIFAR-10	66.32	83.91	93.29
		CIFAR-100	73.95	84.54	71.17
	ResNet18	CIFAR-10	69.07	82.50	94.26
		CIFAR-100	79.80	79.95	77.72
	ResNet34	CIFAR-10	59.07	83.91	95.56
		CIFAR-100	76.55	80.46	78.27
Places365	Madry's ResNet	CIFAR-10	88.59	48.13	93.29
		CIFAR-100	75.19	82.86	71.17
	ResNet18	CIFAR-10	91.08	43.08	94.26
		CIFAR-100	77.16	81.20	77.72
	ResNet34	CIFAR-10	81.71	46.93	95.56
		CIFAR-100	74.05	81.56	78.27
iSUN	Madry's ResNet	CIFAR-10	95.81	21.43	93.29
		CIFAR-100	81.87	72.75	71.17
	ResNet18	CIFAR-10	96.35	43.08	94.26
		CIFAR-100	92.28	41.68	77.72
	ResNet34	CIFAR-10	92.44	23.62	95.56
		CIFAR-100	89.55	47.42	78.27
LSUN (resize)	Madry's ResNet	CIFAR-10	96.92	16.39	93.29
		CIFAR-100	83.45	68.51	71.17
	ResNet18	CIFAR-10	97.20	15.84	94.26
		CIFAR-100	93.10	28.23	77.72
	ResNet34	CIFAR-10	94.02	19.31	95.56
		CIFAR-100	89.94	45.94	78.27

Table 2: Per dataset performance: Area Under the Receiver Operating Curve, FPR95 and In-Distribution classifier accuracy.

λ	10^{-6}	10^{-5}	10^{-4}	10^{-3}	10^{-2}	10^{-1}
AUROC	91.94	92.29	91.93	92.05	92.44	92.60
FPR95	30.19	28.15	30.51	29.47	28.26	32.15
ID Accuracy	93.29	93.06	92.80	92.75	92.59	92.60

Table 3: **Sensitivity of GQ to λ** We evaluate the sensitivity of GQ to λ when CIFAR-10 is the in-distribution dataset (using the setting discussed in Table 1 in the main paper). Metrics stay stable - under 0.5% peak-to-peak for AUROC and $\sim 2.5\%$ in FPR95. The latter metric is inherently noisier since it is induced from a single point in the ROC curve. In-Distribution accuracy falls by 0.7% for λs up to 10^{-1}.

ID	OOD	Madry's Resnet		ResNet18		ResNet34	
		Ours ↑(%)	Ours+pNML ↑(%)	Ours ↑(%)	Ours+pNML ↑(%)	Ours ↑(%)	Ours+pNML ↑(%)
CIFAR-10	iSUN	**96.65**	88.30	**97.61**	94.47	92.82	**96.45**
	LSUN (R)	**97.46**	87.73	**98.10**	95.48	94.27	**97.40**
	LSUN (C)	**98.28**	93.96	98.26	**99.30**	97.07	**97.55**
	Tiny-Imagenet-Resize	**94.64**	86.44	**96.50**	93.00	89.53	**95.13**
	Tiny-Imagenet-Crop	**97.66**	88.44	98.13	**98.36**	95.77	**97.22**
	Uniform	99.31	**99.99**	99.19	**99.98**	99.91	**100**
	Gaussian	99.39	**100**	99.16	**100**	99.94	**100**
	SVHN	**97.24**	95.89	96.39	**96.86**	95.15	**96.76**
CIFAR-100	iSUN	**83.22**	64.05	**94.39**	79.52	91.49	**94.12**
	LSUN (R)	**84.22**	62.98	**94.80**	79.39	91.56	**95.48**
	LSUN (C)	**93.98**	81.82	**92.23**	56.34	88.99	**90.14**
	Tiny-Imagenet-Resize	**84.50**	63.97	**93.85**	81.23	91.51	**96.56**
	Tiny-Imagenet-Crop	**92.02**	70.13	**93.38**	62.20	92.38	**94.37**
	Uniform	94.22	**99.97**	**99.97**	99.20	98.63	**100**
	Gaussian	86.92	**99.99**	**99.98**	99.90	97.67	**100**
	SVHN	**88.95**	76.86	**96.11**	80.47	94.86	**98.13**

Table 4: **AUROC OOD detection scores for GQ and pNML on top of GQ** We show AUROC scores for OOD detection on 8 OOD datasets and 2 ID datasets. We evaluate the scores for GQ trained networks and GQ trained networks with pNML on top of it. This evaluation is done with the pNML [1] code base which scans for the optimal perturbation magnitude ε. **Bold** depicts highest (best) AUROC scores, for the same setting of OOD dataset and network architecture.

experiment, we let the networks train for 100 epochs instead of the 200 epochs we evaluate for CIFAR-10 and CIFAR-100 in the paper.

AUROC results are summarized in Table 3. Substantial degradation is not apparent even after λ changes by 5 orders of magnitude. We therefore conclude that GQ is insensitive to the weight λ, and that its contribution is mainly due to the loss term itself.

6 pNML applied over GQ trained models

We follow [1] and evaluate AUROC and FPR95 on a benchmark of 8 datasets: iSUN [10], LSUN(R), LSUN (C) [11], Tiny-Imagenet Resize and Crop [5], Uniform and Gaussian noises and SVHN [8]. We evaluate these scores for the case of models trained with our method, GQ, and for the case of pNML applied on it. AUROC scores are summarized in Table 4. FPR95 scores are summarized in Table 5. For pNML we use the penultimate layer to extract features in all models. One can observe that as the network complexity is higher (that is with more parameters) - pNML applied on GQ trained models, achieve better results than GQ. We therefore conclude that the two tables highlight that pNML is sensitive to the

ID	OOD	Madry's Resnet		ResNet18		ResNet34	
		Ours ↓ (%)	Ours+pNML ↓ (%)	Ours ↓ (%)	Ours+pNML ↓ (%)	Ours ↓ (%)	Ours+pNML ↓ (%)
CIFAR-10	iSUN	**18.25**	48.06	**14.49**	28.40	24.37	**18.63**
	LSUN (R)	**14.11**	48.49	**10.78**	25.73	20.63	**12.62**
	LSUN (C)	**9.43**	25.08	9.96	**3.18**	15.17	**12.65**
	Tiny-Imagenet-Resize	**28.23**	51.03	**21.77**	36.52	31.74	**24.93**
	Tiny-Imagenet-Crop	**12.90**	46.24	10.50	**8.17**	18.16	**16.02**
	Uniform	1.11	**0**	0.06	**0**	0.01	**0**
	Gaussian	1.07	**0**	0.02	**0**	**0**	**0**
	SVHN	**13.42**	21.25	18.39	**14.25**	22.36	**16.86**
CIFAR-100	iSUN	**70.48**	89.71	**31.41**	58.71	40.41	**22.95**
	LSUN (R)	**66.69**	92.19	**29.37**	61.71	40.10	**20.13**
	LSUN (C)	**31.73**	60.07	**40.80**	85.85	49.72	**41.03**
	Tiny-Imagenet-Resize	**64.27**	85.67	**34.29**	55.51	38.85	**15.45**
	Tiny-Imagenet-Crop	**42.83**	81.33	**37.22**	80.12	36.53	**25.53**
	Uniform	42.35	**0**	**0**	2.34	1.99	**0**
	Gaussian	83.03	**0**	**0**	0.02	9.12	**0**
	SVHN	**52.75**	80.39	**24.57**	53.52	30.24	**9.32**

Table 5: **FPR95 OOD detection scores for GQ and pNML on top of GQ** We show FPR95 scores for OOD detection on 8 OOD datasets and 2 ID datasets. We evaluate the scores for GQ trained networks and GQ trained networks with pNML on top of it. This evaluation is done with the pNML [1] code base which scans for the optimal perturbation magnitude ε. **Bold** depicts lowest (best) FPR95 scores, for the same setting of OOD dataset and network architecture.

features extracted from the backbone to compute the regret.

References

[1] Koby Bibas, Meir Feder, and Tal Hassner. Single layer predictive normalized maximum likelihood for out-of-distribution detection. *Advances in Neural Information Processing Systems*, 34, 2021.

[2] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3606–3613, 2014.

[3] Tarin Clanuwat, Mikel Bober-Irizar, Asanobu Kitamoto, Alex Lamb, Kazuaki Yamamoto, and David Ha. Deep learning for classical japanese literature. *arXiv preprint arXiv:1812.01718*, 2018.

[4] Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 215–223. JMLR Workshop and Conference Proceedings, 2011.

[5] Ya Le and Xuan S. Yang. Tiny imagenet visual recognition challenge. 2015.

[6] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[7] Ziqian Lin, Sreya Dutta Roy, and Yixuan Li. Mood: Multi-level out-of-distribution detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15313–15323, 2021.

[8] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. 2011.

[9] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

[10] Pingmei Xu, Krista A Ehinger, Yinda Zhang, Adam Finkelstein, Sanjeev R Kulkarni, and Jianxiong Xiao. Turkergaze: Crowdsourcing saliency with webcam based eye tracking. *arXiv preprint arXiv:1504.06755*, 2015.

[11] Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015.

[12] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017.