Shifting Transformation Learning for Robust Out-of-Distribution Detection

Authors: Sina Mohseni, Arash Vahdat, JBS Yadawa

Introduction

The real-world deployment of Deep Neural Network (DNN) algorithms in safety-critical applications needs to address DNNs vulnerabilities to Out-of-distribution (OOD) samples. We propose a new technique relying on self-supervision for generalizable out-of-distribution:

- ✓ It does not require additional training samples nor need to pre-know the D^{out}_{target} for tuning.
- ✓ It incurs no extra computation and memory overheads compared methods like DNN ensembles and MC-dropout.

Overview

- ✓ The role of self-supervised learning in OOD detection is poorly explored and limited to one-class classification problems with rather simple geometric transformations tasks:
  - ✓ Intuitively, we simultaneously train the base encoder on multiple shifted distributions of the training data using auxiliary self-supervised objectives and the main classification objective.
  - ✓ We present a simple framework to select effective transformations and module their effect to maximize OOD detection performance only using the source training data without any OOD samples.

Multi-task Transformation Learning

➢ *1) Multi-Task Transformation Learning:* our technique trains a multi-tasked network using self-supervised or fully-supervised training objectives. We define auxiliary transformations learning tasks $T = \{(T_n, \lambda_n)\}_{n=1}^{N}$ based on geometric (translation, rotation) and non-geometric (blurring, sharpening, color jittering, Gaussian noise, cutout) domain-invariant image transformation.

We use auxiliary self-supervised tasks to learn multiple shifted distributions of the training set $S = \{(x_i, y_i)\}_{i=1}^{M}$

$$\mathcal{L}_{main} = \sum_{(x_i,y_i) \in S} \ell(f(T_n(x_i)), \hat{y}_i)$$

$$\mathcal{L}_{aux} = \frac{1}{N}\sum_{n=1}^{N} \lambda_n \sum_{(T_n(x_i),\hat{y}_i) \in S_{T_n}} \ell(f(T_n(x_i)), \hat{y}_i)$$

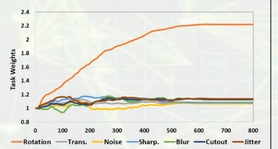
$$\mathcal{L}_{total} = \lambda_0 \mathcal{L}_{main}(\theta) + \lambda_n \mathcal{L}_{aux}(\theta)$$

Each auxiliary task is selected and optimized only using the in domain training set.

➢ *2) Selecting Transformations Set T:* we use Bayesian optimization to identify effective transformations for each in-domain training from a pool of available transformations. In our experiments we observed that gradient-based optimization is not able to capture the effects of transformation itself on the training.

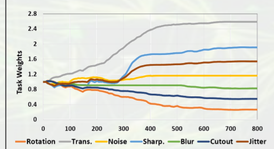
➢ *3) Optimizing Transformations Weights λ:* we optimize training coefficients (λ) for selected transformations to efficiently modulate the impact of each transformation in the training loss. We use differentiable hyperparameter optimization for λ weights as meta-parameter in the outer loop by backpropagating through the gradient's updates of the network parameters θ.

	Rot.	Trans.	Noise	Jitter	Sharp.	Cutout	Blur
Rot.	92.18	92.55	92.25	90.65	91.47	91.43	92.48
Trans.	92.55	90.16	90.65	90.62	90.54	89.95	90.57
Noise	92.25	90.65	90.08	90.24	90.41	89.99	90.35
Jitter	90.65	90.62	90.24	89.63	89.84	90.07	90.23
Sharp.	91.47	90.54	90.41	89.84	89.58	89.40	90.14
Cutout	91.43	89.95	89.99	90.07	89.40	89.93	89.92
Blur	92.48	90.57	90.35	90.23	90.14	89.92	90.50



(a) CIFAR-10

	Rot.	Trans.	Noise	Jitter	Sharp.	Cutout	Blur
Rot.	76.04	75.78	75.49	74.84	74.69	76.00	75.51
Trans.	75.78	78.81	76.58	78.96	78.20	76.77	77.47
Noise	75.49	76.58	78.96	76.70	76.40	75.26	76.70
Jitter	74.84	78.96	76.70	77.77	77.09	77.20	76.09
Sharp.	74.69	78.20	76.40	77.09	78.69	76.59	75.95
Cutout	76.00	76.77	75.26	77.20	76.59	77.99	76.08
Blur	75.51	77.47	76.70	76.09	75.95	76.08	77.86



(b) CIFAR-100

➢ *4) Detection Score:* we calculate an ensemble detection score by combining prediction entropy from the main classification task and all auxiliary transformation heads to compute the OOD detection score.

Experiments and Results

➢ *OOD Detection Performance:* we evaluate our method in comparison to self-supervised and contrastive learning based works. Our results indicate the efficiency of multi-task learning for training of a wide range of invariances to in-domain representations.

✓ *Table 1 presents evaluation results (AUROC%) in comparison to the supervised training (Baseline) and six state-of-the-art techniques all trained on ResNet-18 backbone.*

✓ Our results are reported with both self-supervised only ($\mathcal{L}_{ssl}$) and combination of it with labeled supervised ($\mathcal{L}_{ssl} + \mathcal{L}_{sup}$) for auxiliary transformation learning loss.

✓ *Our technique outperforms other state-of-the-art by learning a diverse set of transformations. Notably, unlike contrastive learning, our framework modulates the impact of different transformations via trainable λ weights instead of explicitly dividing transformations into positive and negative sets.*

✓ *We present evaluation results trained on CIFAR-10, CIFAR-100, and ImageNet-30 datasets. The outlier test sets are diverse set of disjoint datasets to*

Table 1: OOD detection results (AUROC%) on multi-class classification setup

D^{in}_{train}	D^{out}_{test}	Supervised Baseline	Geometric SSL [21]	SupSimCRL [24]	CSI-ens [48]	SSD [46]	Our ($\mathcal{L}_{\text{SSL}}$)	Our ($\mathcal{L}_{\text{SSL}} + \mathcal{L}_{\text{Sup}}$)
CIFAR10	SVHN	92.89	97.96	97.22	97.38	93.8	99.92	96.6
	Texture	87.69	96.25	94.21	97.18	94.05	97.61	96.91
	Places365	88.34	92.57	91.11	93.11	91.77	93.72	98.73
	ImageNet	87.44	92.06	92.1	92.49	90.28	92.99	93.57
	LSUN	89.87	93.57	92.13	94.02	94.4	95.03	94.12
	CIFAR-100	87.62	91.91	88.36	92.06	90.4	93.24	94.07
	Average	88.98	94.05	92.52	94.37	92.45	95.42	95.67
CIFAR100	SVHN	79.18	83.62	81.55	87.38	83.6	87.11	90.64
	Texture	75.28	82.39	76.83	78.31	81.35	85.47	77.99
	Places365	76.07	74.57	75.37	78.1	79.16	77.87	92.62
	ImageNet	78.53	77.56	80.77	82.41	76.29	80.66	79.25
	LSUN	73.73	71.86	73.5	75.22	63.77	74.32	74.01
	CIFAR-10	78.26	74.73	73.28	78.44	73.94	79.25	91.56
	Average	76.84	77.46	76.88	79.98	76.35	80.78	84.35
ImageNet-30	Places365	89.6	92.17	90.81	94.28	95.47	92.97	97.54
	Flowers 101	87.7	92.13	93.81	96.18	96.47	94.19	97.18
	CUB-200	85.26	90.58	89.19	94.15	96.57	93.34	96.44
	Dogs	90.3	93.25	95.16	97.64	95.23	93.63	97.07
	Food	78.93	85.09	83.61	89.04	85.48	82.51	96.49
	Pets	92.88	95.28	96.38	98.49	96.24	94.82	96.37
	Texture	86.98	92.16	98.7	98.54	94.86	93.99	96.56
	Average	87.01	91.42	92.81	95.67	94.14	92.08	96.69

Out-of-Distribution Generalizability and Robustness

✓ *We propose four criteria for an ideal OOD detection technique, including i) zero-shot OOD training, ii) no hyperparameter dependency, iii) generalization to various unseen OOD distributions, and iv) robustness against test-time perturbations. Table 2 situates our proposed technique against a diverse range of state-of-the-art OOD detection techniques trained on the same network with the same training budget.*

✓ *Hyperparameters Dependency: hyperparameter disentanglement is a crucial property for OOD detection and an ideal detector should not be sensitive to hyperparameters tied to the target outlier distribution.*

✓ *Zero-shot Training: despite the previous trend in using samples from D^{out}_{target} for tuning or D^{out}_{train} for training, having access to target outlier samples is not a realistic assumption in many setups.*

✓ *Detection Robustness: we investigate the effect of natural perturbations and corruptions on OOD detection performance. We measure averaged OOD detection results for 15 image distortions tests with 5 levels of intensity presented in [18].*

✓ *Generalizable OOD Detection: we* recognize the necessity of diverse D^{out}_{test} sets to evaluate the generalizability of OOD detection techniques w.r.t. the distance between the D^{in}_{train} set and D^{out}_{test} sets.

Table 2: OOD detection generalizability and robustness (AUROC%) in different criteria

Detection Technique	OOD Detection Criteria			Averaged Detection Performance	Generalizability Tests		
	Hyp.-Para. Dependency	Generalizable	Zero-shot		Mixed Distribution	Far-OOD	Near-OOD
ODIN [28]	High	--	--	91.15	88.1	96.7	85.8
Mahalanobis [29]	High	--	--	95.35	92.24	99.1	88.51
Outlier Exposure [20]	Low	✓	--	96.24	96.88	98.76	93.41
Geometric SSL [21]	Low	✓	✓	94.05	94.29	97.96	91.91
CSI-ens [48]	Low	✓	✓	94.37	94.1	97.38	90.6
Gram Matrices [45]	No	--	✓	94.17	95.08	99.5	79.01
SSD [46]	No	✓	✓	92.47	92.7	93.8	90.4
Ours	No	✓	✓	95.67	95.55	96.6	94.07

References: Please see the paper for references