Published as a conference paper at ICLR 2019

DEEP ANOMALY DETECTION WITH OUTLIER EXPOSURE

Dan Hendrycks
University of California, Berkeley
hendrycks@berkeley.edu

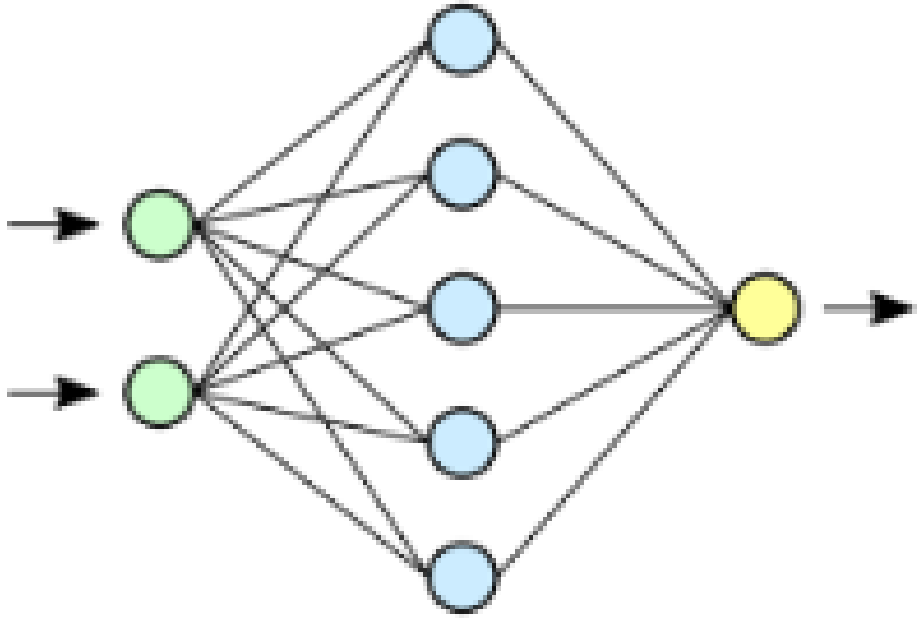
Mantas Mazeika
University of Chicago
mantas@ttic.edu

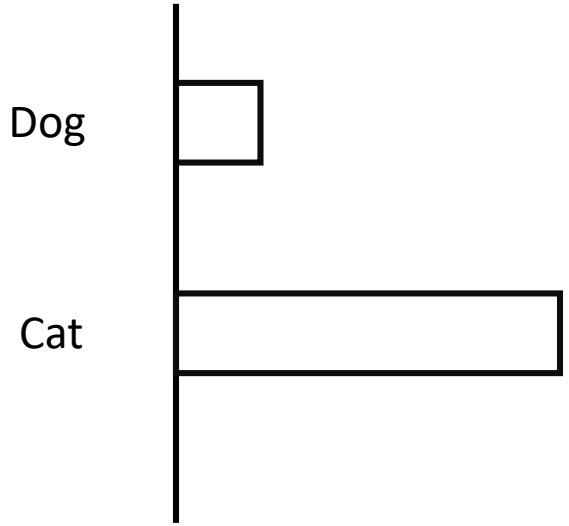
Thomas Dietterich
Oregon State University
tgd@oregonstate.edu

TAEWON KIM

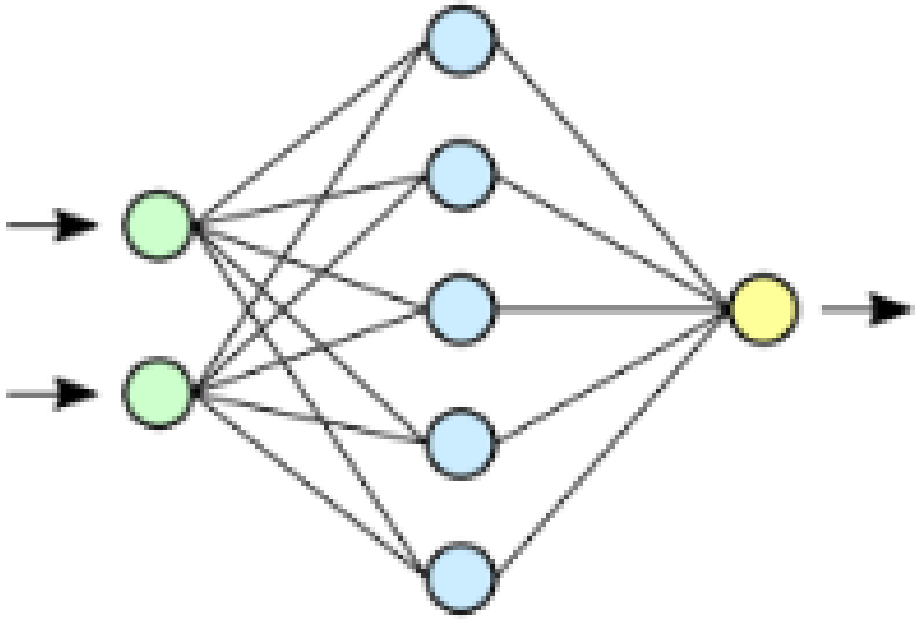
0. Contents

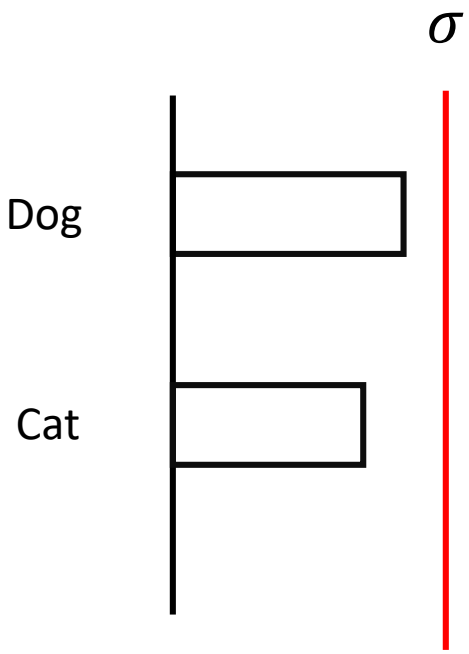
1. OOD detection



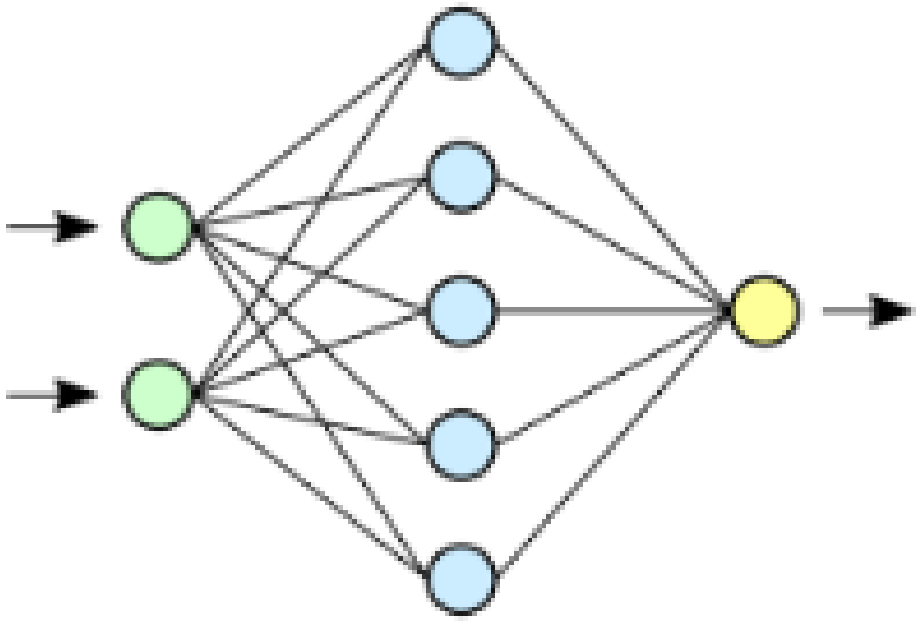


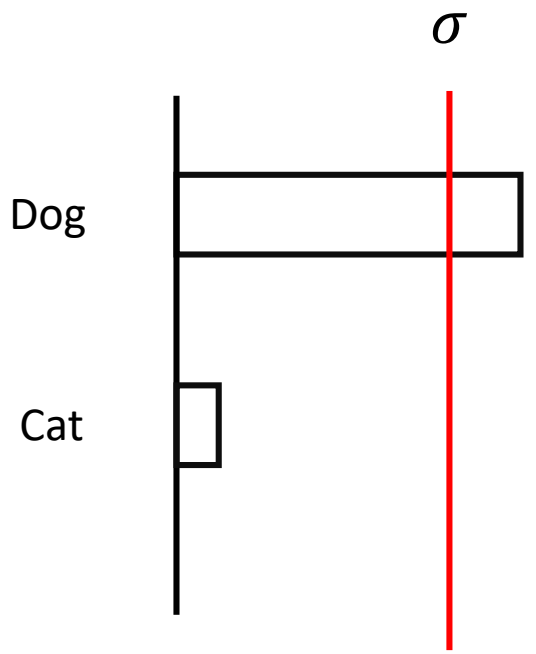
1. OOD detection





1. OOD detection





1. Outlier Exposure

Outlier Exposure

- 본 논문은 기존의 OOD detection 방법론에 간단하게 추가할 수 있는 새로운 방법론을 제시함

- OE(Outlier Exposure) 방법론은 문자 그대로 OOD에 해당하는 데이터를 학습에 사용하여 각 방법론에서 OOD detection의 성능을 높임

Related works

Published as a conference paper at ICLR 2017

Baseline: A Baseline for Detecting Misclassified and OOD Data in NN

- OOD detection이라는 task를 처음으로 정의하고 단순 soft-max 값을 OOD score로 사용하는 방식을 향후 연구의 baseline 모델로 제시함
- 아래는 이에 대한 실험 결과 중의 하나로 단순 soft-max 값을 OOD score를 사용하는 경우에도 어느정도 좋은 성능을 보이는 것을 확인함

In-Distribution / Out-of-Distribution	AUROC /Base	AUPR In /Base	AUPR Out/Base	Pred. Prob (mean)
CIFAR-10/SUN	95/50	89/33	97/67	72
CIFAR-10/Gaussian	97/50	98/49	95/51	77
CIFAR-10/All	96/50	88/24	98/76	74
CIFAR-100/SUN	91/50	83/27	96/73	56
CIFAR-100/Gaussian	88/50	92/43	80/57	77
CIFAR-100/All	90/50	81/21	96/79	63
MNIST/Omniglot	96/50	97/52	96/48	86
MNIST/notMNIST	85/50	86/50	88/50	92
MNIST/CIFAR-10bw	95/50	95/50	95/50	87
MNIST/Gaussian	90/50	90/50	91/50	91
MNIST/Uniform	99/50	99/50	98/50	83
MNIST/All	91/50	76/20	98/80	89

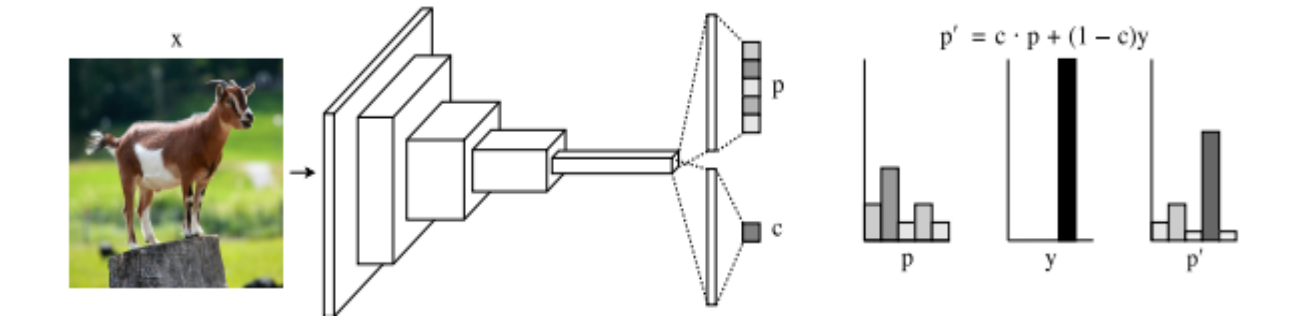
Table 2: Distinguishing in- and out-of-distribution test set data for image classification. CIFAR-10/All is the same as CIFAR-10/(SUN, Gaussian). All values are percentages.

Related works

arXiv:1802.04865v1 [stat.ML] 13 Feb 2018

Learning Confidence for Out-of-Distribution Detection in Neural Networks

Terrance DeVries [1 2] Graham W. Taylor [1 2]



- 그림과 같이 네트워크 구조와 loss function을 변형하여 모델이 confidence score을 산출할 수 있도록 함
- P는 기존의 softmax score를 의미
- C는 [0,1] 사이의 값을 갖는 confidence score를 산출
- 최종적으로 산출하는 output은 p' = c * p + (1 - c) * y 이고 y는 실제 정답 label
- L = -Σ(log(p') * y - log(c))

Related works

Published as a conference paper at ICLR 2018

TRAINING CONFIDENCE-CALIBRATED CLASSIFIERS FOR DETECTING OUT-OF-DISTRIBUTION SAMPLES

Kimin Lee* **Honglak Lee**[§,†] **Kibok Lee**[†] **Jinwoo Shin***
*Korea Advanced Institute of Science and Technology, Daejeon, Korea
†University of Michigan, Ann Arbor, MI 48109
§Google Brain, Mountain View, CA 94043

Cross Entropy loss

KL-Div. Term for Out of-dist. data

$$\min_G \max_D \min_\theta \underbrace{\mathbb{E}_{P_{in}(\hat{\mathbf{x}},\hat{y})}\left[-\log P_\theta\left(y=\hat{y}|\hat{\mathbf{x}}\right)\right]}_{(c)} + \beta \underbrace{\mathbb{E}_{P_G(\mathbf{x})}\left[KL\left(\mathcal{U}\left(y\right) \| P_\theta\left(y|\mathbf{x}\right)\right)\right]}_{(d)}$$

$$+\underbrace{\mathbb{E}_{P_{in}(\hat{\mathbf{x}})}\left[\log D\left(\hat{\mathbf{x}}\right)\right] + \mathbb{E}_{P_G(\mathbf{x})}\left[\log\left(1-D\left(\mathbf{x}\right)\right)\right]}_{(e)}.$$

Modified GAN objective function

- 해당 논문에서는 GAN을 이용하여 out of distribution 데이터를 생성 후 이를 이용하여 모델을 재 학습하여 OOD를 탐지하는 방식 제안
- GAN loss를 변형하여 GAN의 generator가 out of distribution sample을 생성하도록함
- 여기서 u는 균등 분포를 의미하고 Pθ 는 Classifier를 의미 각 class에 대해 균등 분포를 산출 하도록 유도
- Cross Entropy loss 같은 경우는 In-distribution data가 올바른 클래스를 예측하는 데 얼마나 잘 수행되는지
- KL-Div. Term 은 out-of-distribution data가 uniform한 분포를 가지도록 유도하는 과정
- Modified GAN objective function
  in-distribution data 대해 올바르게 분류하고 생성자가 생성한 가짜 데이터가 실제와 구분 안되게 학습

울산대학교 UNIVERSITY OF ULSAN 서울아산병원 Asan Medical Center

Related works

Published as a conference paper at ICLR 2018

TRAINING CONFIDENCE-CALIBRATED CLASSIFIERS FOR DETECTING OUT-OF-DISTRIBUTION SAMPLES

Kimin Lee* **Honglak Lee**[§,†] **Kibok Lee**[†] **Jinwoo Shin***
*Korea Advanced Institute of Science and Technology, Daejeon, Korea
†University of Michigan, Ann Arbor, MI 48109
§Google Brain, Mountain View, CA 94043

Cross Entropy loss

KL-Div. Term for Out of-dist. data

$$\min_G \max_D \min_\theta \underbrace{\mathbb{E}_{P_{in}(\hat{\mathbf{x}},\hat{y})}\left[-\log P_\theta\left(y=\hat{y}|\hat{\mathbf{x}}\right)\right]}_{(c)} + \beta \underbrace{\mathbb{E}_{P_G(\mathbf{x})}\left[KL\left(\mathcal{U}\left(y\right) \| P_\theta\left(y|\mathbf{x}\right)\right)\right]}_{(d)}$$

$$+\underbrace{\mathbb{E}_{P_{in}(\hat{\mathbf{x}})}\left[\log D\left(\hat{\mathbf{x}}\right)\right] + \mathbb{E}_{P_G(\mathbf{x})}\left[\log\left(1-D\left(\mathbf{x}\right)\right)\right]}_{(e)}.$$

Modified GAN objective function

(c)

In distribution data

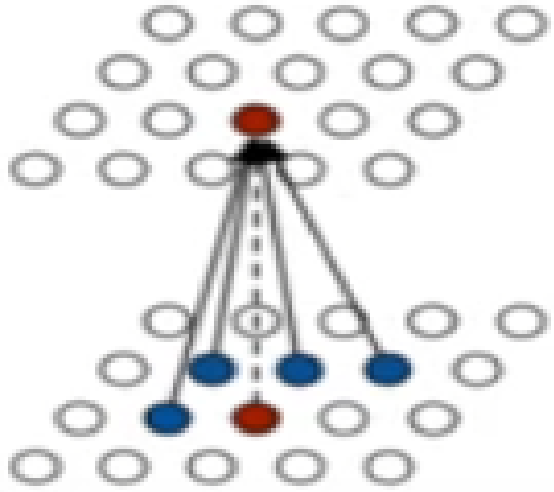
(d)

GAN data(out of distribution data)

Related works

Bits Per Pixel(BPP)

- BPP는 Pixel CNN++ 모델을 사용하여 OOD를 탐지
- Pixel CNN ++ 는 CNN기반의 autoregressive 모델
- 주변의 context를 통해 해당 pixel의 등장 확률을 계산하는 모델로 학습 시에는 모든 pixel이 1이 되도록 학습을 진행



PixelCNN

Related works

Bits Per Pixel(BPP)

- 따라서 Pixel CNN++를 사용하면 image 데이터에서 주변 pixel을 통해 각 pixel의 등장 확률을 계산할 수 있고, 이를 활용하여 OOD score인 BPP를 구할 수 있음
- $BPP = \frac{\sum NLL(x)}{Num\_Pixel}$

BPP는 이미지를 압축할 때 각 픽셀을 표현하는 데 필요한 평균 비트 수를 나타내는 지표로, 이미지의 압축률을 측정하는 데 사용됩니다. BPP를 계산하는 공식은 다음과 같습니다:

Negative Log Likelihood 의 합계가 높을수록 모델이 데이터를 더 잘 예측하고 있다는 것을 의미하며, 이는 BPP가 낮아짐을 의미합니다.

Pixel CNN++를 통해 예측된 픽셀 값의 확률이 낮다면, 해당 이미지는 OOD로 분류될 가능성이 높고, 이는 BPP 값이 높아질 수 있음을 나타냅니다

Method

Outlier Exposure

- 기존의 OOD detection 방법론에 간단하게 추가할 수 있는 새로운 방법론을 제시
- OE(Outlier Exposure) 방법론은 문자 그대로 OOD에 해당하는 데이터를 학습에 사용하여 각 방법론에서 OOD detection의 성능을 높임
- [D_in,D_out-oe,D_out-test] 같이 데이터 셋이 구성되어 있고 D_out-oe와 D_out-test는 완전히 다른 데이터로 구성

Method

Outlier Exposure: Data Set

- [D_{in}]: SVHN, CIFAR-10&100, Tiny ImageNet, Places365, 20Newsgroups, TREC, SST
- [D_{out-oe}]: 80 Million Tiny Images(SVHN, CIFAR), ImageNet-22K(Tiny ImageNet, Places365), WikiText-2(나머지)
- [$D_{out-test}$]: SVHN, CIFAR-10&100, Tiny ImageNet, Places365, 20Newsgroups, TREC, SST
- D_{out-oe}, $D_{out-test}$ 이 두개의 데이터 셋은 겹치면 안되므로 겹치는 데이터의 경우 다 제거해주었다고 함

Method

Outlier Exposure : Experiment on Maximum Softmax Probability (MSP).

- MSP의 경우 out-of-distribution 데이터의 정답 label을 uniform 하여 classifier를 학습한 후 진행함
- $\mathbf{E}_{(x,y)\sim D_{in}}[-\log f_y(x)] + \beta \mathbf{E}_{x\sim D_{out}}[H(u; f(x))]$ $(\beta = 1.0)$

	FPR95 ↓		AUROC ↑		AUPR ↑	
$\mathcal{D}_{in}$	MSP	+OE	MSP	+OE	MSP	+OE
SVHN	6.3	0.1	98.0	100.0	91.1	99.9
CIFAR-10	34.9	9.5	89.3	97.8	59.2	90.5
CIFAR-100	62.7	38.5	73.1	87.9	30.1	58.2
Tiny ImageNet	66.3	14.0	64.9	92.2	27.2	79.3
Places365	63.5	28.2	66.5	90.6	33.1	71.0

Table 1: Out-of-distribution image detection for the maximum softmax probability (MSP) baseline detector and the MSP detector after fine-tuning with Outlier Exposure (OE). Results are percentages and also an average of 10 runs. Expanded results are in Appendix A.

Method

Outlier Exposure : Experiment on Maximum Softmax Probability (MSP).

- MSP의 경우 out-of-distribution 데이터의 정답 label을 uniform 하여 classifier를 학습한 후 진행함
- $\mathbf{E}_{(x,y)\sim \mathrm{In}}[-\log f_y(x)] + \beta \mathbf{E}_{x\sim \mathrm{OE}}[H(u; f(x))]$ $(\beta = 1.0)$

	FPR90 ↓		AUROC ↑		AUPR ↑	
$\mathcal{D}_{\text{in}}$	MSP	+OE	MSP	+OE	MSP	+OE
20 Newsgroups	42.4	4.9	82.7	97.7	49.9	91.9
TREC	43.5	0.8	82.1	99.3	52.2	97.6
SST	74.9	27.3	61.6	89.3	22.9	59.4

Table 2: Comparisons between the MSP baseline and the MSP of the natural language classifier fine-tuned with OE. Results are percentages and averaged over 10 runs.

Method

Outlier Exposure : Experiment on Confidence Branch

- Confidence branch 방법론의 경우 out-of-distribution 데이터에 대해서는 p는 학습하지 않고 c만 다음과 같은 loss를 추가하여 학습
- $L_{In} = -\sum_{i=1}^{M} \log(p')\, y - \alpha \log(c)$ & $L_{OE} = \alpha \log(c)$ (α=0.5)

	FPR95 ↓			AUROC ↑			AUPR ↑		
$\mathcal{D}_{in}$	MSP	Branch	+OE	MSP	Branch	+OE	MSP	Branch	+OE
CIFAR-10	49.3	38.7	20.8	84.4	86.9	93.7	51.9	48.6	66.6
CIFAR-100	55.6	47.9	42.0	77.6	81.2	85.5	36.5	44.4	54.7
Tiny ImageNet	64.3	66.9	20.1	65.3	63.4	90.6	30.3	25.7	75.2

Table 3: Comparison among the maximum softmax probability, Confidence Branch, and Confidence Branch + OE OOD detectors. The same network architecture is used for all three detectors. All results are percentages, and averaged across all $\mathcal{D}_{out}^{test}$ datasets.

Method

Outlier Exposure : Experiment on GAN

- MSP 의 GAN 이미지를 추가한 경우이므로 단순하게 out-of-distribution을 또 새롭게 추가하여 학습

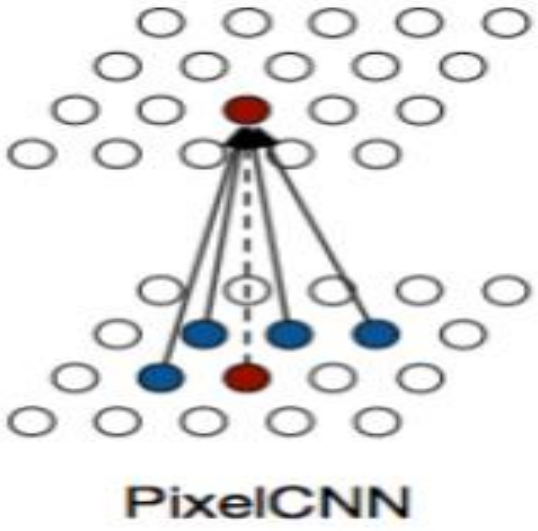
	FPR95 ↓			AUROC ↑			AUPR ↑		
$\mathcal{D}_{\text{in}}$	MSP	+GAN	+OE	MSP	+GAN	+OE	MSP	+GAN	+OE
CIFAR-10	32.3	37.3	11.8	88.1	89.6	97.2	51.1	59.0	88.5
CIFAR-100	66.6	66.2	49.0	67.2	69.3	77.9	27.4	33.0	44.7

Table 4: Comparison among the maximum softmax probability (MSP), MSP + GAN, and MSP + GAN + OE OOD detectors. The same network architecture is used for all three detectors. All results are percentages and averaged across all $\mathcal{D}_{\text{out}}^{\text{test}}$ datasets.

Method

Outlier Exposure : Experiment on BPP

- 이 경우 in-distribution 으로 먼저 pixel CNN ++ 을 학습한 후 D_out-oe 을 활용하여 Pixel CNN++ 를 재 학습 하여 사용함
- Pixel CNN++은 모든 pixel의 등장 확률이 0이 되도록 학습을 진행함



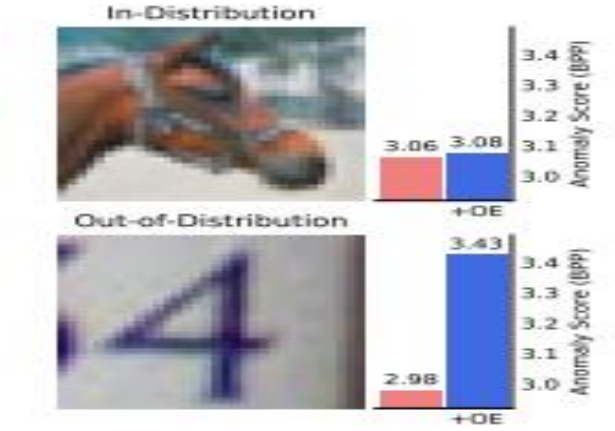


Figure 2: OOD scores from PixelCNN++ on images from CIFAR-10 and SVHN.

Method

Outlier Exposure : Experiment on BPP

- 이 경우 in-distribution 으로 먼저 pixel CNN ++ 을 학습한 후 D_out-oe 을 활용하여 Pixel CNN++ 를 재 학습 하여 사용함
- Pixel CNN++은 모든 pixel의 등장 확률이 0이 되도록 학습을 진행함(D_out-oe)

$\mathcal{D}_{in}$	$\mathcal{D}_{out}^{test}$	FPR95 ↓		AUROC ↑		AUPR ↑	
		BPP	+OE	BPP	+OE	BPP	+OE
CIFAR-10	Gaussian	0.0	0.0	100.0	100.0	100.0	99.6
	Rademacher	61.4	50.3	44.2	56.5	14.2	17.3
	Blobs	17.2	1.3	93.2	99.5	60.0	96.2
	Textures	96.8	48.9	69.4	88.8	40.9	70.0
	SVHN	98.8	86.9	15.8	75.8	9.7	60.0
	Places365	86.1	50.3	74.8	89.3	38.6	70.4
	LSUN	76.9	43.2	76.4	90.9	36.5	72.4
	CIFAR-100	96.1	89.8	52.4	68.5	19.0	41.9
Mean		66.6	**46.4**	65.8	**83.7**	39.9	**66.0**

Table 5: OOD detection results with a PixelCNN++ density estimator, and the same estimator after applying OE. The model's bits per pixel (BPP) scores each sample. All results are percentages. Test distributions $\mathcal{D}_{out}^{test}$ are described in Appendix A.

Discussion

Outlier Exposure :Discussion

- OOD detection 성능을 높이기 위한 D_out-oe 데이터의 특징에 대해서 실험적으로 밝힘
- 첫 번째 D_out-oe 데이터 셋의 다양성
- CIFAR10을 in distribution 으로 한 실험에서 CIFAR 100의 10개의 class를 D_out-oe로 활용했을 때와 30개의 class를 D_out-oe로 활용하였을 때 그 성능 차이가 약 7%가 있음
- 반면 50개의 class를 사용한 경우는 30개의 class를 사용한 경우와 성능에 큰 차이가 없었고, 다양성이 확보된 경우 데이터 셋의 개수는 큰 상관이 없다
- D_out-oe, D_out-test 그리고 D_in 간의 유사성
- 실험 결과 D_out-oe와 D_out-test는 유사성이 떨어져도 전혀 상관이 없음, SVHN을 out of distribution으로 한 실험에서 SVHN과 비슷한 digit 이미지 데이터를 D_out-oe로 사용한 경우가 오히려 자연 이미지를 D_out-oe로 사용한 경우에 비해 성능이 안좋음
- 반대로 D_out-oe와 D_in은 유사성이 있는 경우 성능이 좋았음
- 결론적으로 D_out-oe는 충분한 다양성이 확보된 D_in과 유사성이 높은 (즉 쉽게 구분하기 어려운) 데이터 셋을 사용하는게 좋음

Discussion

Outlier Exposure :Discussion

- 본 논문에서는 OE를 사용한 경우 모델의 output의 calibration이 좋아지는 것을 확인함
- Calibration은 모델의 결과를 확률 값으로 보는 관점, 만약 MSP의 값이 0.9인 경우 calibration이 완벽하다면 이는 90% 확률로 분류 결과가 올바른 것을 의미

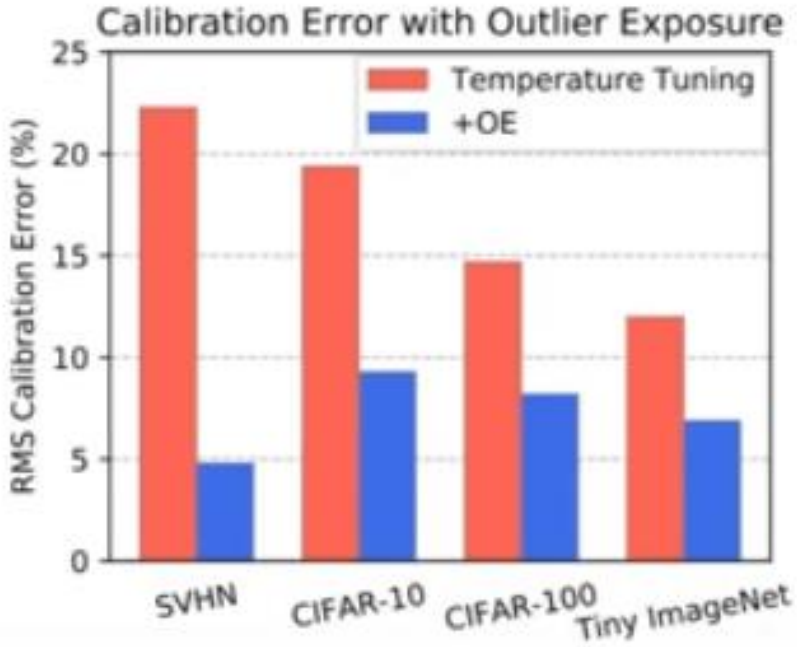


Figure 3: Root Mean Square Calibration Error values with temperature tuning and temperature tuning + OE across various datasets.

감사합니다