Published as a conference paper at ICLR 2017

A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks

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

Kevin Gimpel
Toyota Technological Institute at Chicago
kgimpel@ttic.edu

- Motivation
- Contribution
- Background
- Concept
- Experiment
- Conclusion

Intro.

기존 Discrimitive Model(classifier)의 문제 - Overconfidence

https://mnist-demo.herokuapp.com/

Intro.

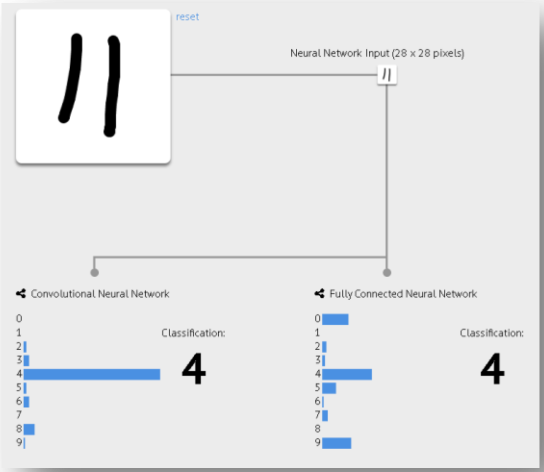
◆ 기존 DL based Discriminative Model(classifier)의 문제

✓ Overconfidence

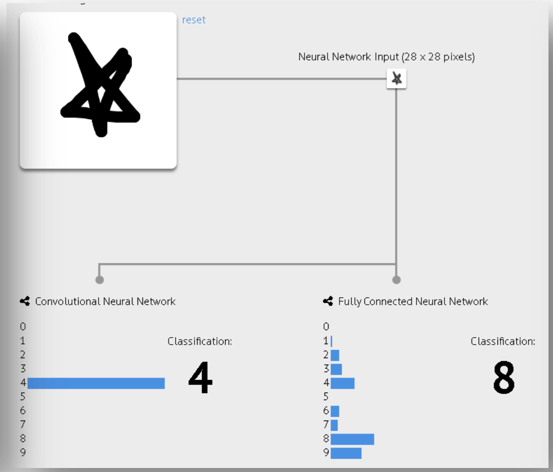
◆ Out-of-Distribution(Abnormal) Sample Inference

Miss-classification
= error

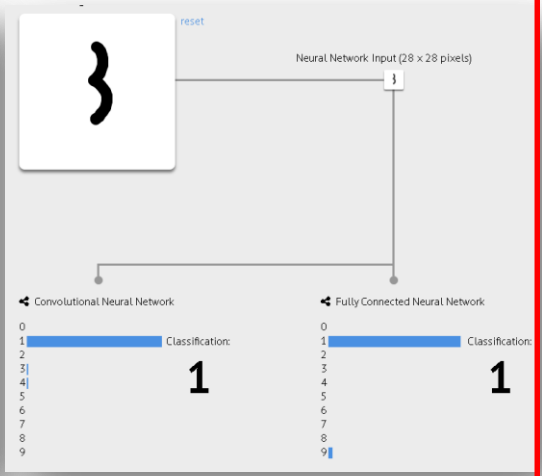
Outlier
(신규 class)



Outlier
(오염, Deployee된 App을 비정상 이용)



Novelty



overconfidence *overconfidence* *overconfidence* *overconfidence* *overconfidence*

Papers(Anomaly Detection)

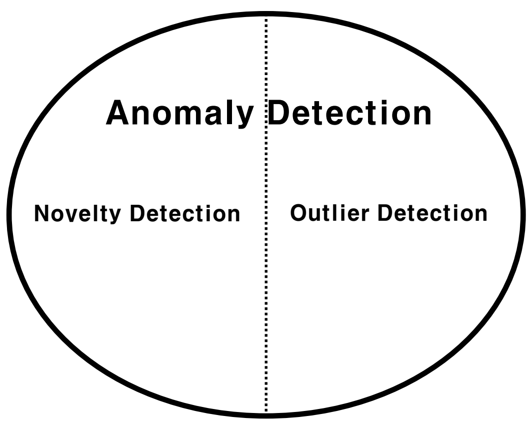
- A Baseline For Detecting Misclassified and Out-of-Distribution Examples in Neural Networks (Hendrycks et. al., ICLR 2017)
- Enhancing The Reliability of Out-of-Distribution Image Detection in Neural Networks (Liang et. al., ICLR 2018)
- Training Confidence-Calibrated Classifiers for Detecting Out-of-Distribution Samples (Lee et. al., ICLR 2018)
- A Simple Unified Framework for Detecting Out-of-Distribution Samples and Adversarial Attacks (Lee el. al., NeurIPS 2018)
- Learning Confidence for Out-of-Distribution Detection in Neural Networks (DeVries et. al., arXiv 2018)
- Deep Anomaly Detection with Outlier Exposure (Hendrycks et. al., ICLR 2019)

Background

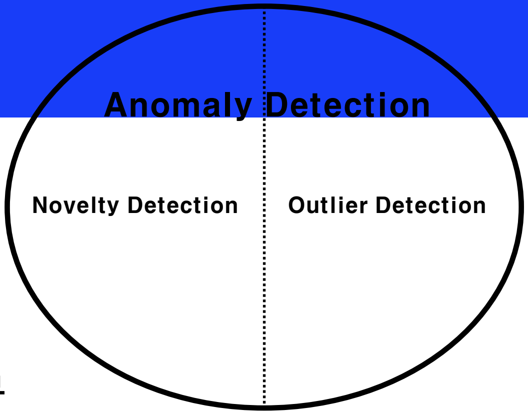
◆ Anomaly Detection 용어 구분

✓ **Normal Sample Class 개수와 Abnormal Sample 성격**

<table>
<tr><td rowspan="3"></td><td colspan="2">Anomaly Detection</td></tr>
<tr><td colspan="2">Goal: Test-time Abnormal-sample 찾기</td></tr>
<tr><td colspan="2">Abnormal 성격
(=Unknown=Unseen)</td></tr>
<tr><td>보유한 학습세트에
Normal Sample(In-distribution sample) 개수
(Normal Class = 1개)</td><td>Open-set에서 충분히 등장
-> Novelty Detection 문제
(Novel class=Normal class)</td><td>Open-set에서 등장 가능성 X
-> Outlier Detection 문제
(Outlier class=Abnormal class)</td></tr>
<tr><td>보유한 학습세트에
Normal Sample(In-distribution sample) 개수
(Normal Class > 1개)</td><td colspan="2">OoD(Out-of-Distribution) 문제</td></tr>
</table>



Background



◆ Anomaly Detection 용어 구분

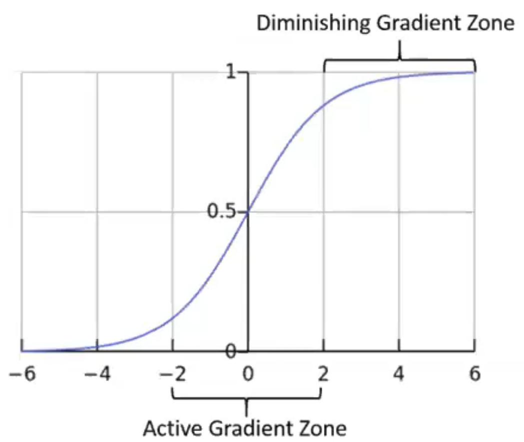
✓ **학습데이터의 레이블링 유무와 Normal/Abnormal Sample 학습 시 사용 유무**

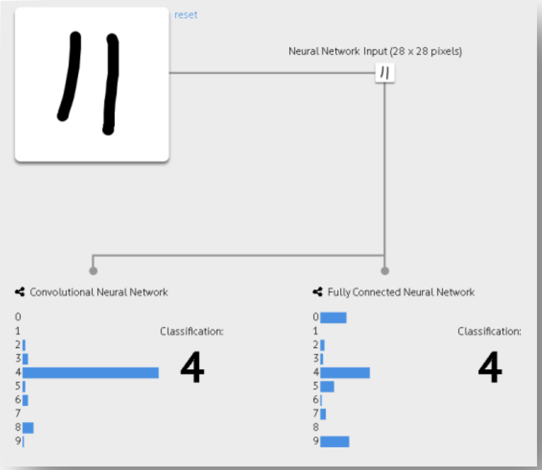
	학습데이터에 레이블링	Normal Sample (In-distribution)	Abnormal Sample (Out-of-distribution)	장점	단점
Supervised Anomaly Detection([1])	O	학습 사용 O	학습 사용 O	Acc 높음	수집시 Cost 발생, Class-imbalance 문제
Semi-supervised Anomaly Detection([2]) = One-Class Anomaly Detection	O (필터링)	학습 사용 O	학습 사용 X	정상이미지만 가지고 학습하므로 불량이미지 수집 비용 X	[1] 대비 Acc 낮다, 여전히 정상이미지에 대한 label 작업이 필요하다(필터링)
Un-supervised Anomaly Detection([3])	X	대다수 사용 O **(대다수 Normal 가정)**	극소수 사용 O	데이터에 대한 레이블링 작업이 필요없다.	[1] 대비 Acc 낮다, [1] 에 비해 하이퍼파리미터에 의한 모델 성능에 대한 일관성이 없다(성능에 영향을 주는 요소 많다)

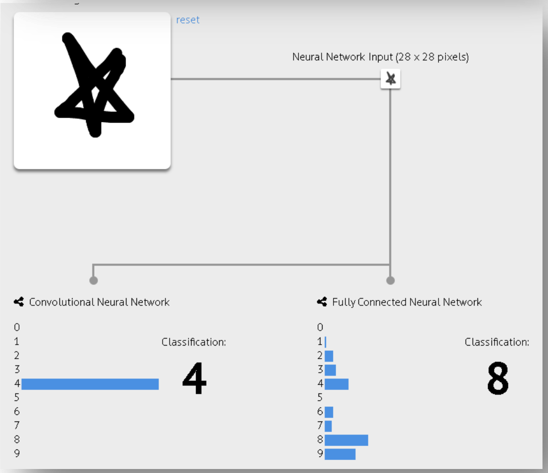
[1]
Discrimitive Model(Traditional Softmax based Classifier)
[2]
ML-based : Energy-based Generative, Model based(GMM), One-Class SVM
DL-based : Generative Model(GAN), Deep-SVDD
[3]
ML-based : PCA
DL-based : AE

Background – Overconfidence in DL

$$A = \frac{1}{1+e^{-x}}$$







Logit

: Exponential -> Output is Sensitivity !
-> Over confidence in NN

Background – Overconfidence in DL

EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks

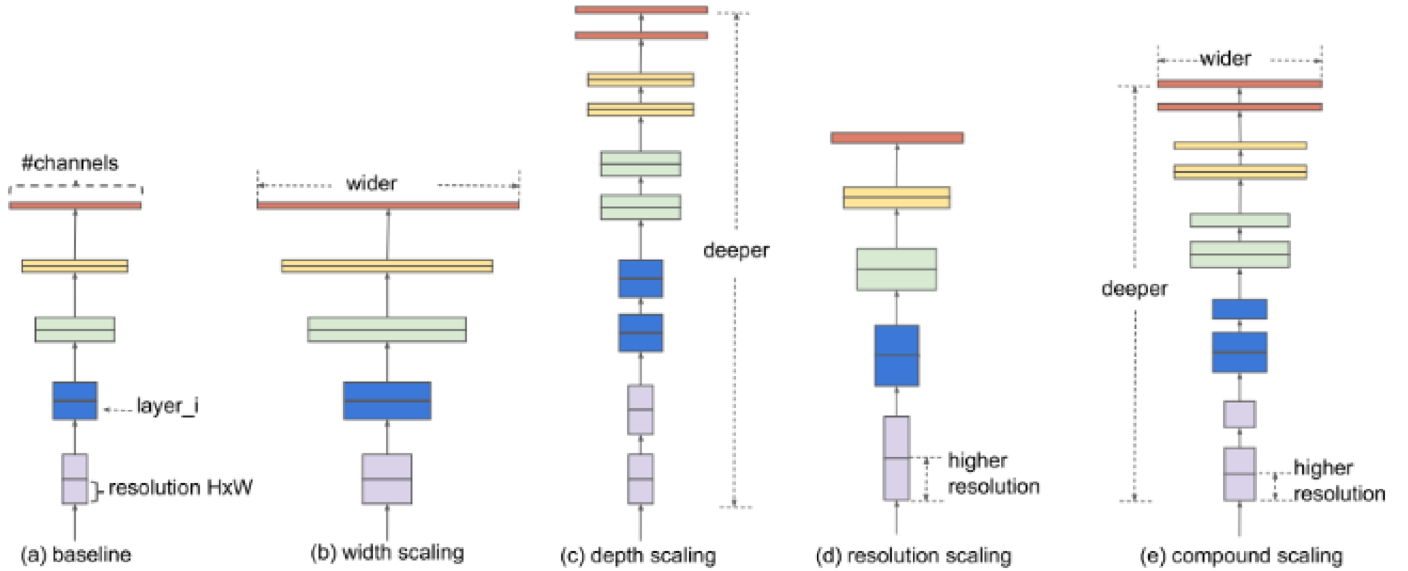
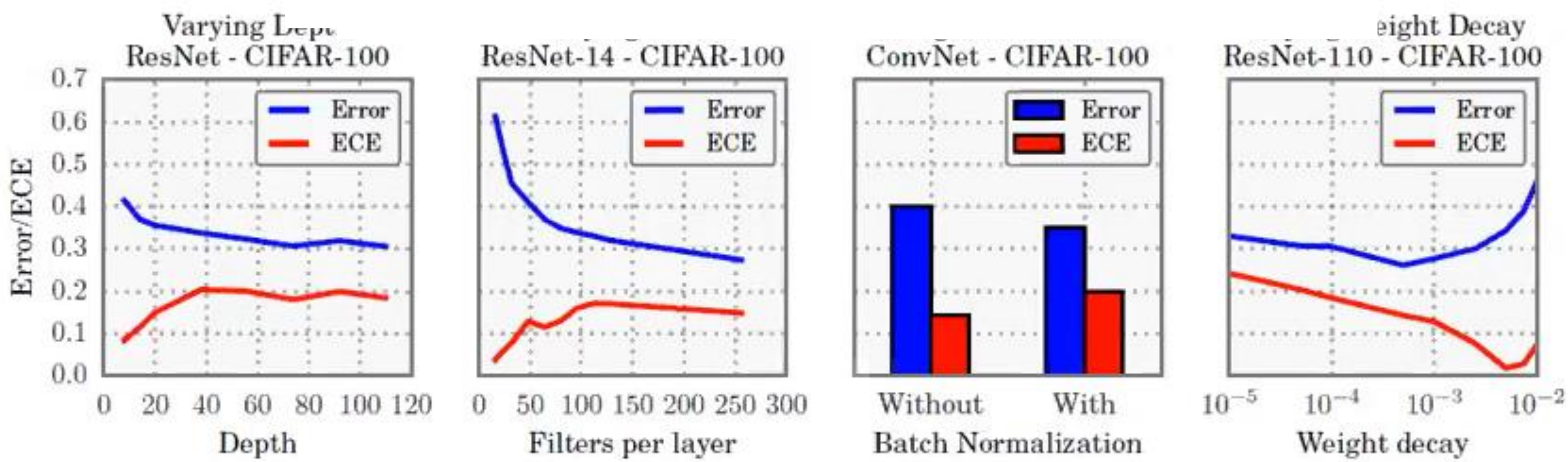


Figure 2. **Model Scaling.** (a) is a baseline network example; (b)-(d) are conventional scaling that only increases one dimension of network width, depth, or resolution. (e) is our proposed compound scaling method that uniformly scales all three dimensions with a fixed ratio.

Background – Overconfidence in DL

On Calibration of Modern Neural Networks

* ECE = Expected Calibration Error



① Depth ↑
② Filters ↑
③ Batch Normalization 有
④ Weight Decay ↓

It remains future work to understand why these trends affect calibration while improving accuracy.

Contribution

이 논문은 Anomaly Detection 태스크의 최초의 논문

"OOD Detection" 문제를 해결하기 위한 Baseline 논문
-> Anomaly detection 최초 논문
-> CNN based
-> Classifier 기반에서 해결하고자 함 (인퍼런스타임에 집중=이미 학습이 끝난 logit을 재활용) -> loss 재설계 맥락 X
-> 기존 Discrimitive Model"에서 사용 가능

기존 Classifier가 발생시키는 Over-confidence 문제를 별다른 가정 없이(기존 모델 변경 없이) Anomaly Detection 을 하는 Solution을 Baseline으로써 제시

이 논문 이후의 Anomaly Detection 연구에 Scheme 표준(모델평가과 실험설계)을 제시

논문에서 제시한 Baseline method는 다양한 태스크NLP, Vision, Speech Recognition)에서 효과가 있었음

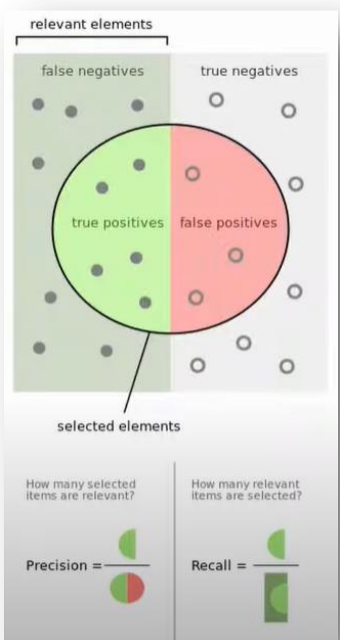
Prediction$_{\text{OK SAMPLE}}$ > Prediction$_{\text{NG SAMPLE}}$, Prediction$_{\text{OOD SAMPLE}}$

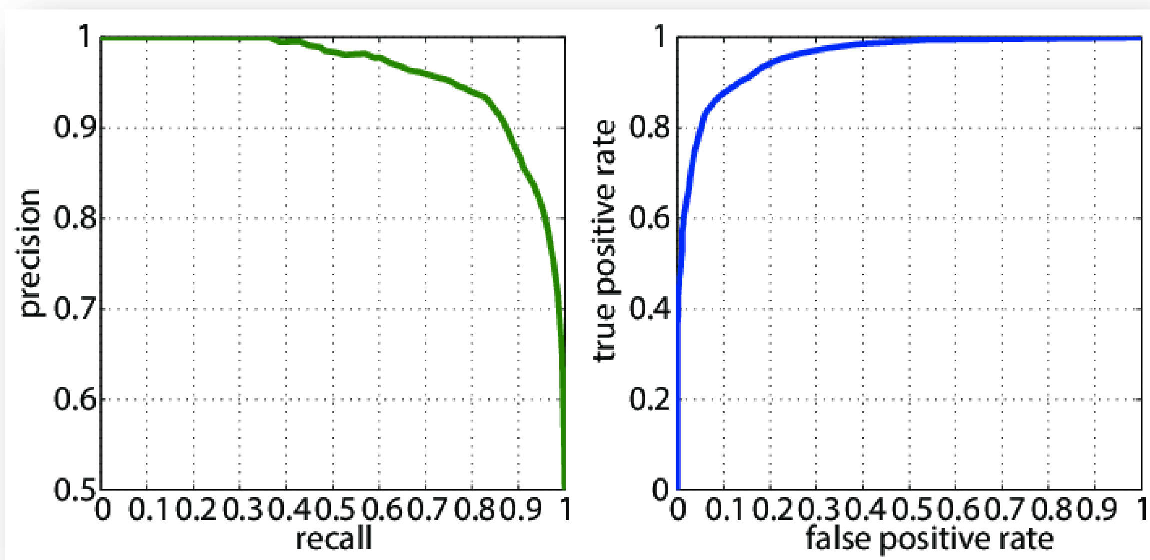
-> 모두 overconfident. Overconfidence 정도가 차이가 있는 대체적인 경향이 있음을 확인. 따라서 Maximum Softmax Probability를 이용한 Anomaly Sample 판단은 가능함.

Background

- Anomaly Detection 용어
  - Anomaly Detection, Novelty Detection, Out-of-Detection(OOD)

- Anomaly Detection의 평가지표
  - AUROC, AUPR, Pred. Prob(mean)
    -> TH 무관하게 measure.
    -> Open-set-dataset is imbalanced sample.





Background

Confusion Matrix with imbalanced data distribution e.g. binary classification

		GT	
		NG	OK
Pred	NG	0	0
	OK	1	99

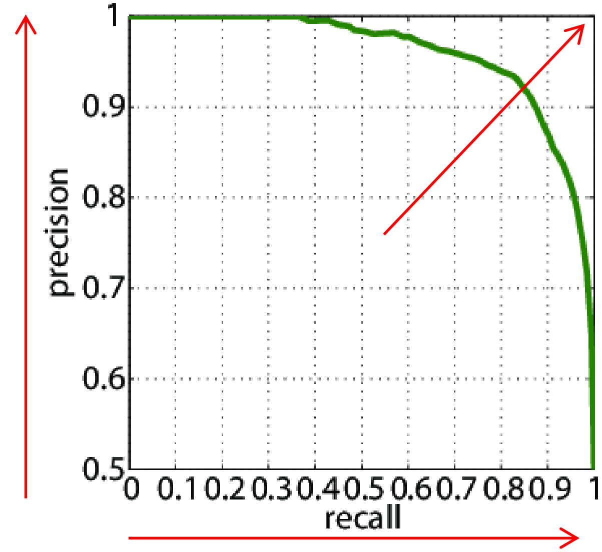
- OK(In-distribution) 입장에서 NG는 Out-of-distribution
- 모델은 realworld특성상 학습셋은 Imbalaced
- OOD sample(Unseen)에 대해 잘 예측 못함.

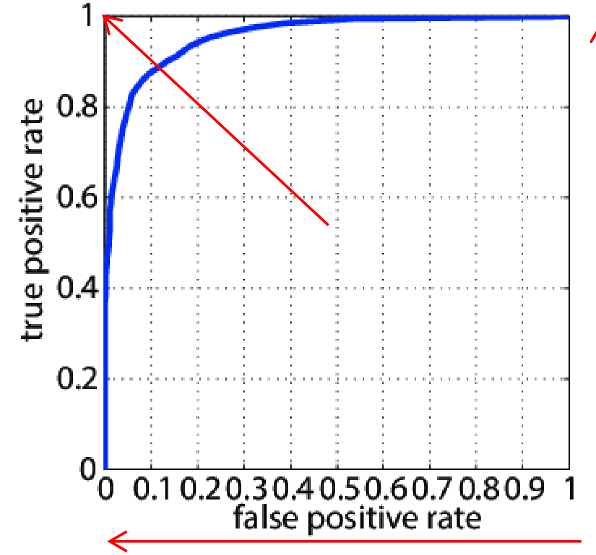
- 모델성능평가지표
  -> If Accuracy = 99%
- 실데이터는 Imbalaced 되었다
  -> Anormaly 태스크에 적합X

Background

		GT	
		NG	OK
Pred	NG	5 TP	10 FP
	OK	5 FN	80 TN

ACC = 85/100 = 0.85





5 TP	10 FP
5 FN	80 TN

5 TP	10 FP
5 FN	80 TN

precision = 5/15 = 0.3

5 TP	10 FP
5 FN	80 TN

5 TP	10 FP
5 FN	80 TN

recall = 5/10 = 0.5

5 TP	10 FP
5 FN	80 TN

5 TP	10 FP
5 FN	80 TN

FP-R = 10/90 = 0.1

5 TP	10 FP
5 FN	80 TN

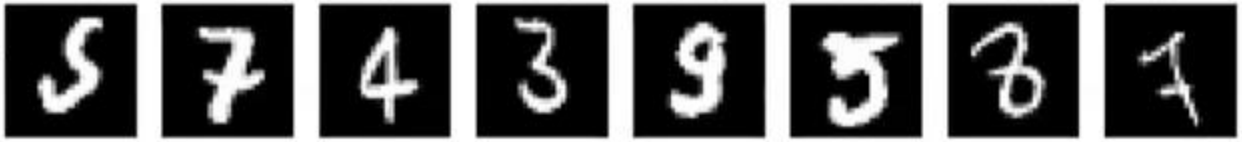
5 TP	10 FP
5 FN	80 TN

TP-R = 5/10 = 0.5

EXP-1

Miss-classified

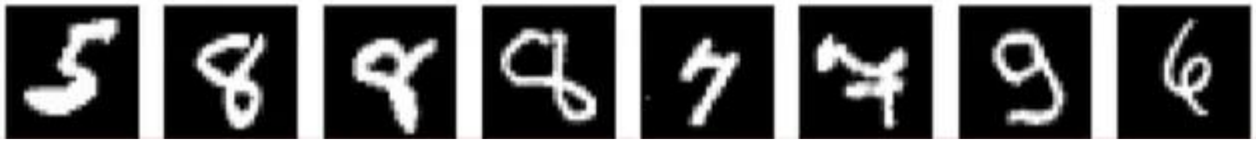
Mean : 0.86



0.81, 6(5)　0.91, 3(7)　0.84, 6(4)　0.91, 6(3)　0.86, 3(9)　0.75, 3(5)　0.90, 3(8)　0.88, 2(7)

Correct

Mean : 0.91

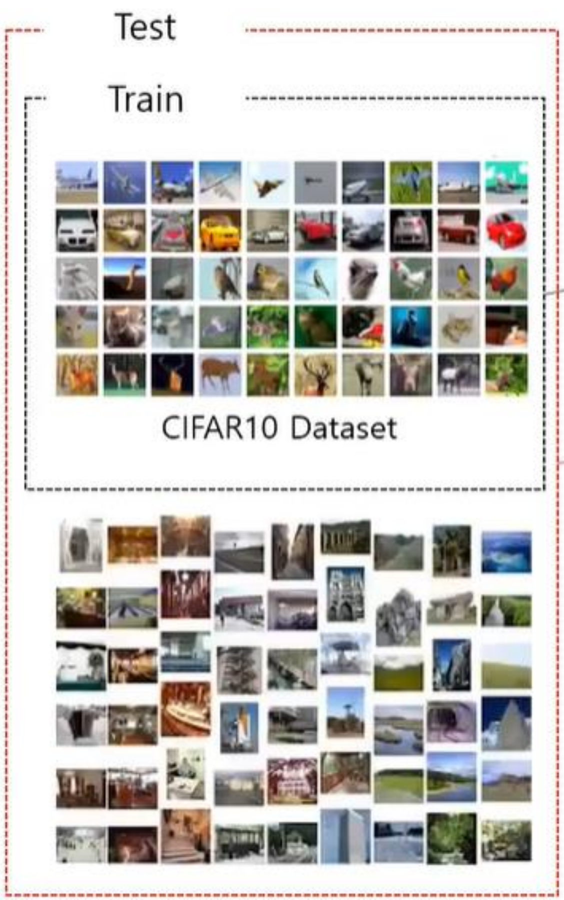


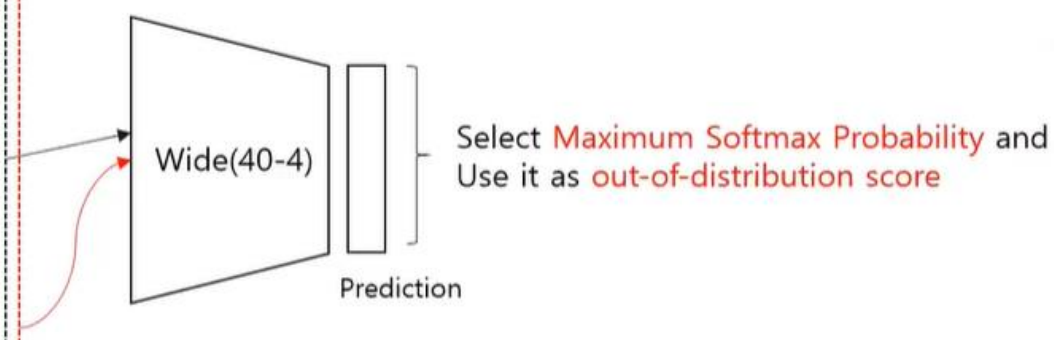
0.90　0.95　0.85　0.95　0.92　0.88　0.95　0.86

Dataset	AUROC /Base	AUPR Succ/Base	AUPR Err/Base	Pred. Prob Wrong(mean)	Test Set Error
MNIST	97/50	100/98	48/1.7	86	1.69
CIFAR-10	93/50	100/95	43/5	80	4.96
CIFAR-100	87/50	96/79	62/21	66	20.7

-> Very Over-confidence !

EXP-2





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

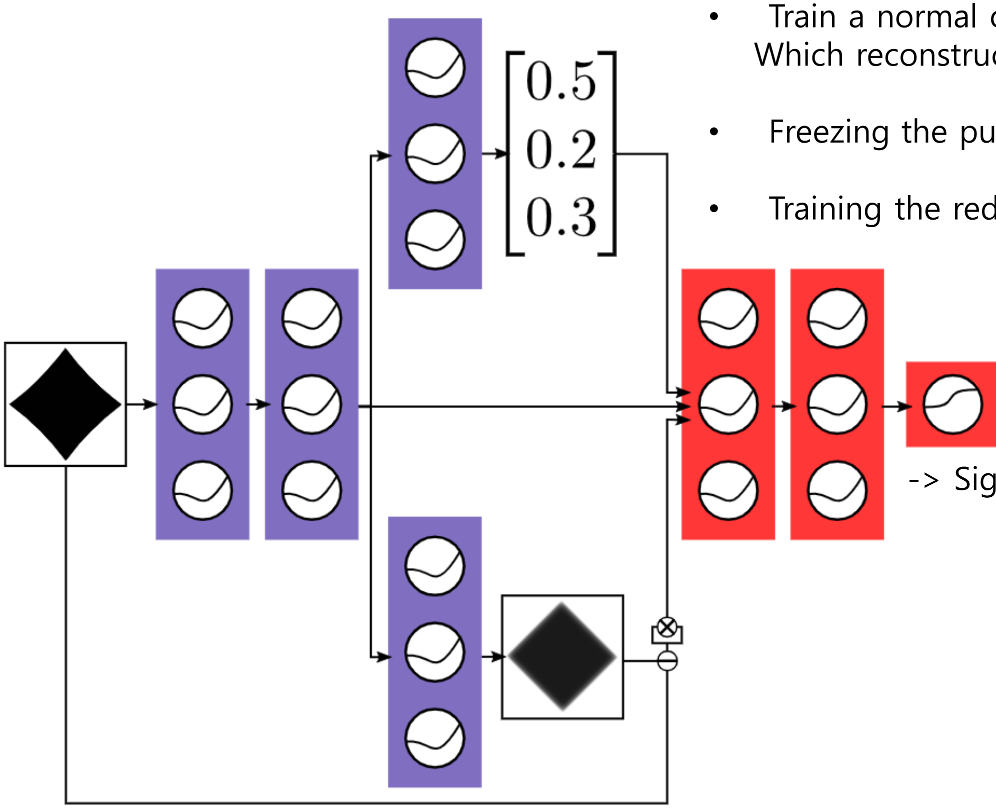
Table 2: Distinguishing in- and out-of-distribution test set data for image classification. CIFAR-10/All is the same as CIFAR-10/(SUN, Gaussian). All values are percentages.

-> Maximum Softmax Prob as OOD score ! (Various task)

~ Vision, NLP(Sentiment Classification, Text Categorization, Autoimatic Speech Recognition)

Pipeline

Abnormality Module



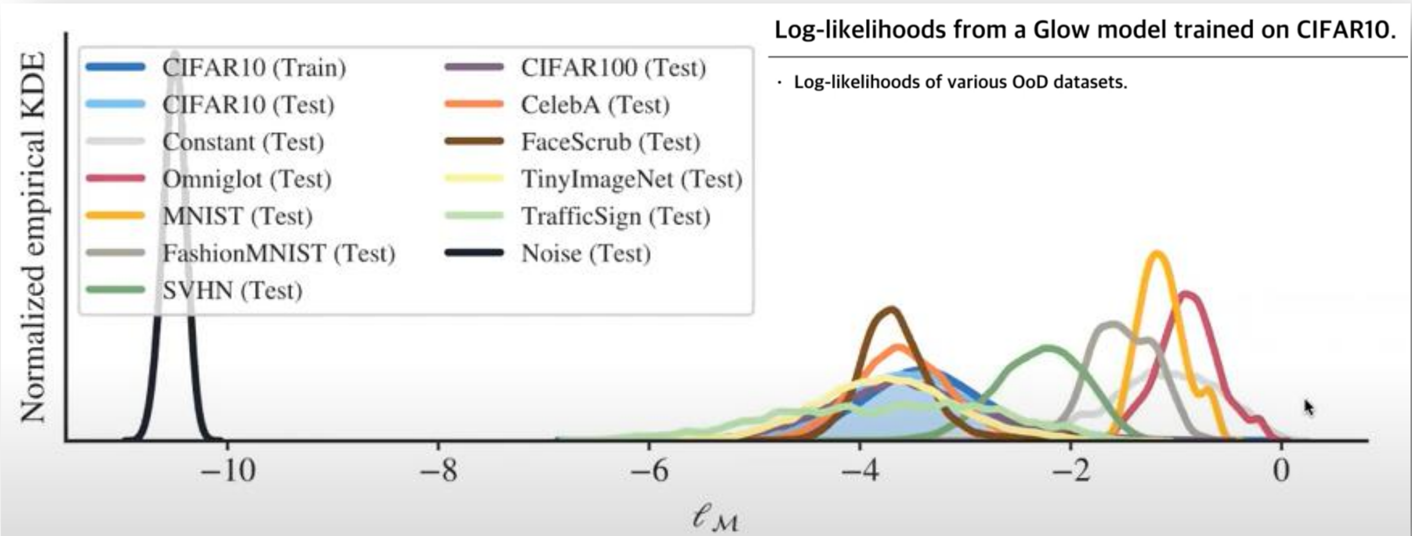
- Train a normal classifier and append an auxiliary decoder Which reconstructs the input with in-distribution samples.
- Freezing the purple layer
- Training the red layrers on Normal Samples, Noised-Normal Samples

-> Sigmoid output of the red layers is OOD Scores we can use

Experiments

In-Distribution / Out-of-Distribution	AUROC /Base Softmax	AUROC /Base AbMod	AUPR In/Base Softmax	AUPR In/Base AbMod	AUPR Out/Base Softmax	AUPR Out/Base AbMod
MNIST/Omniglot	95/50	100/50	95/52	100/52	95/48	100/48
MNIST/notMNIST	87/50	100/50	88/50	100/50	90/50	100/50
MNIST/CIFAR-10bw	98/50	100/50	98/50	100/50	98/50	100/50
MNIST/Gaussian	88/50	100/50	88/50	100/50	90/50	100/50
MNIST/Uniform	99/50	100/50	99/50	100/50	99/50	100/50
Average	93	100	94	100	94	100

Table 11: Improved detection using the abnormality module. All values are percentages.



Conclusion and Follow-ups

- Demonstrated a softmax prediction probability baseline for error, out-of-distribution detect
- Presented the abnormality module (+ gain)
- Presented Evaluation Metric in OOD task(property)

Deep Anomaly Detection with Outlier Exposure, 2019 ICLR

$\mathcal{D}_{\text{in}}$	FPR95 ↓ MSP	FPR95 ↓ +OE	AUROC ↑ MSP	AUROC ↑ +OE	AUPR ↑ MSP	AUPR ↑ +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.

$\mathcal{D}_{\text{in}}$	FPR95 ↓ MSP	FPR95 ↓ +GAN	FPR95 ↓ +OE	AUROC ↑ MSP	AUROC ↑ +GAN	AUROC ↑ +OE	AUPR ↑ MSP	AUPR ↑ +GAN	AUPR ↑ +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.

- Outlier Exposure는 기존 방법들에 독립적으로 추가가 가능한 아이디어
- 기존 detector들에 Outlier Exposure를 추가하였을 때 얼마나 성능이 향상되는지를 논문에서 결과로 제시
- 다만 Outlier Exposure로 **어떤 데이터 셋**을 사용하는지에 따라 성능이 크게 달라질 수 있다는 점이 풀어야 할 문제(Future work)
- Gaussian noise나 GAN으로 생성한 sample 등을 활용하는 것은 크게 효과적이지 않음
- 반면, Outlier Exposure로 사용하는 데이터 셋을 최대한 realistic 하면서 size도 크고, 다양하게 구축하는 것이 좋은 성능을 달성하는 데 도움을 준다고 가이드를 제시해주고 있습니다.
- 기존에 존재하던 Out-of-distribution Detection 알고리즘들에 추가로 적용이 가능하면서도 손쉽게 구현이 가능한 방법론을 제안하였고, 실제로 효과적인 성능 향상다.

감사합니다