DMQA Open Seminar

Score-Based OOD Detection for Image Classification: Part1

2024. 01. 26

고려대학교 산업경영공학과

Data Mining & Quality Analytics Lab.

임 새 린

발표자 소개

❖ 임새린 (Saerin Lim)

- 고려대학교 산업경영공학과 Data Mining & Quality Analytics Lab.
- Ph.D. Student (2021.03 ~ Present)
- 지도 교수 : 김성범 교수님

❖ Research Interest

- Self-supervised learning & Semi-supervised learning

❖ Contact

- E-mail : momo_om@korea.ac.kr

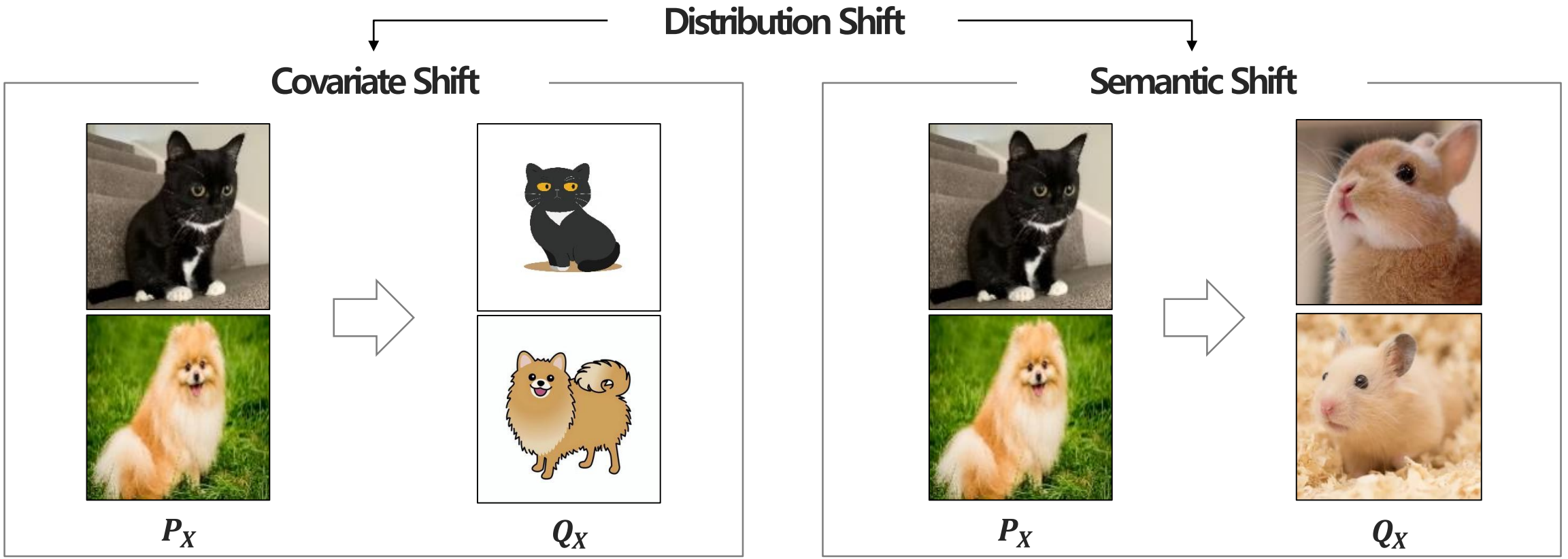
목차

Contents

OOD Detection

Out-of-Distribution: OOD

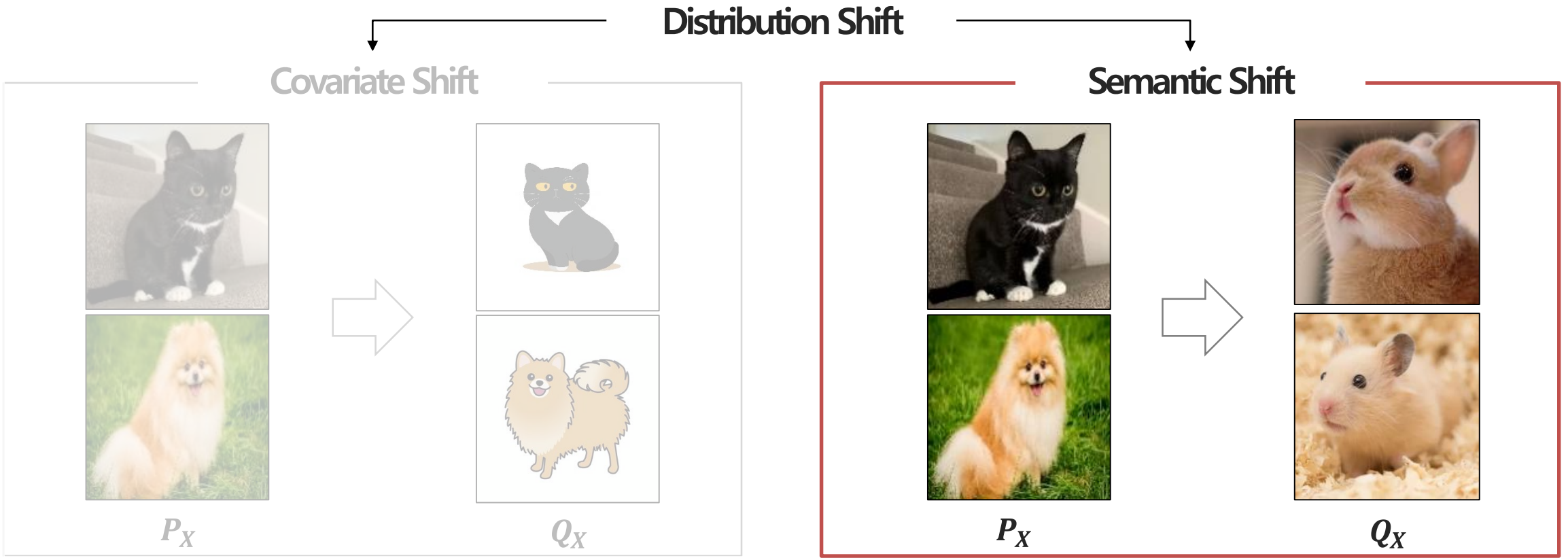
- ❖ OOD 데이터란 학습에 활용된 입력 데이터 분포 P_X와 상당히 다른 분포 Q_X에서 샘플링된 입력 데이터
- ❖ 일반적으로 이미지 분류 문제에서는 학습 데이터에 존재하지 않는 클래스를 가진 이미지를 의미



OOD Detection

Out-of-Distribution: OOD

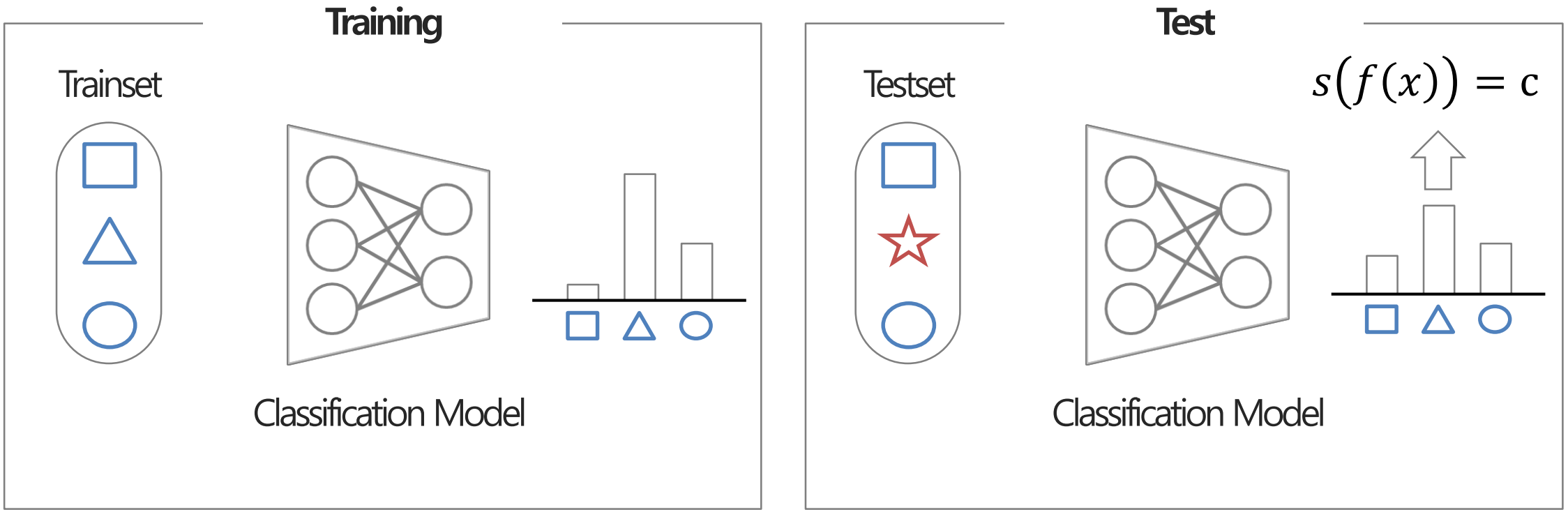
- ❖ OOD 데이터란 학습에 활용된 입력 데이터 분포 P_X와 상당히 다른 분포 Q_X에서 샘플링된 입력 데이터
- ❖ 일반적으로 이미지 분류 문제에서는 학습 데이터에 존재하지 않는 클래스를 가진 이미지를 의미



OOD Detection

Out-of-Distribution Detection

- ❖ OOD detection은 학습이 완료된 모델이 새로운 샘플을 입력 받았을 때 ID인지 OOD인지를 구분하는 태스크
- ❖ 학습할 때는 OOD탐지가 아닌 이미지 분류 모델로 학습이 됨



OOD Detection

Out-of-Distribution Detection

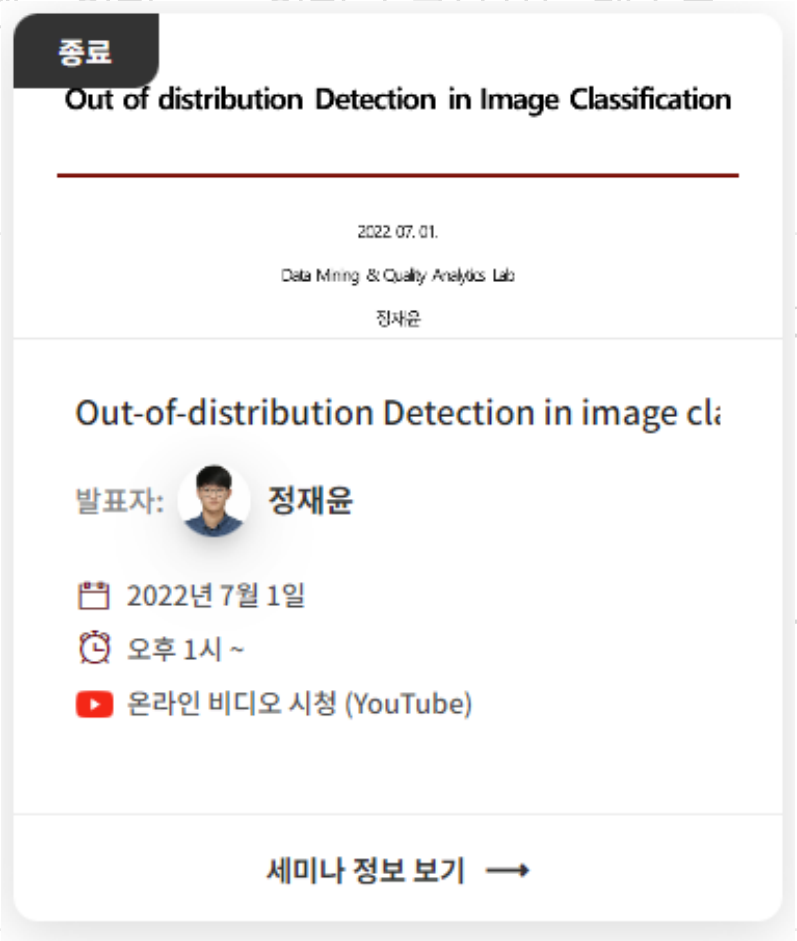
- OOD detection은 학습이 완료된 모델이 새로운 샘플을 입력 받았을 때
- 학습할 때는 OOD탐지가 아닌 이미지 분류 모델로 학습이 됨

OOD Detection

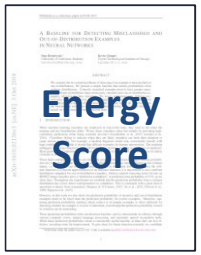
Open-Set Recognition

Anomaly Detection

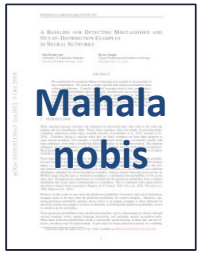
Novelty Detection

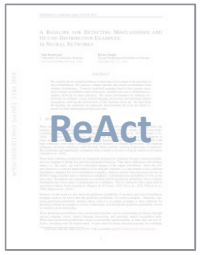


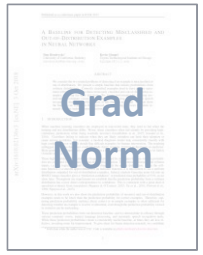
Timeline



2017 2018 2019 2020 2021 2022







A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks (2017, ICLR)

Baseline

Introduction

- ❖ 이미지 분류, 자연어 처리, 음성인식 분야에서 OOD탐지 **문제를 체계적으로 정의하고 평가 프로토콜을 제안**
- ❖ 실험적 발견을 기반으로 간단한 OOD score를 정의하고 **OOD 탐지 문제의 baseline 제공**

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

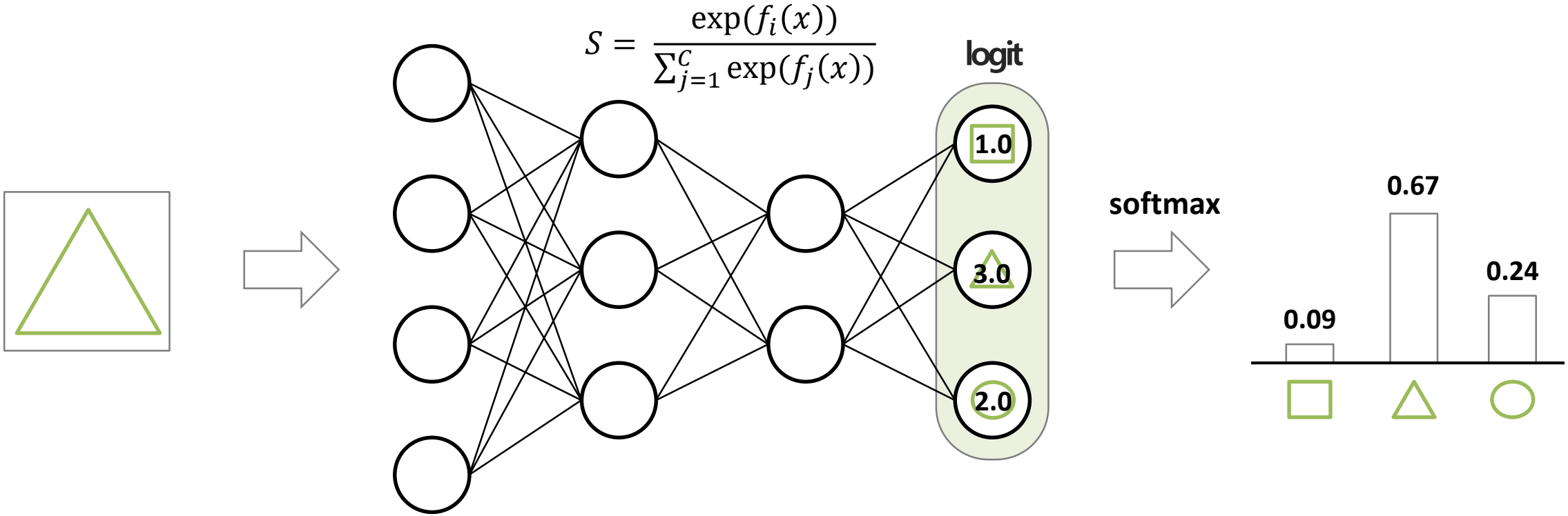
ABSTRACT

We consider the two related problems of detecting if an example is misclassified or out-of-distribution. We present a simple baseline that utilizes probabilities from softmax distributions. Correctly classified examples tend to have greater maximum softmax probabilities than erroneously classified and out-of-distribution examples, allowing for their detection. We assess performance by defining several tasks in computer vision, natural language processing, and automatic speech recognition, showing the effectiveness of this baseline across all. We then show the baseline can sometimes be surpassed, demonstrating the room for future research on these underexplored detection tasks.

Baseline

Motivations

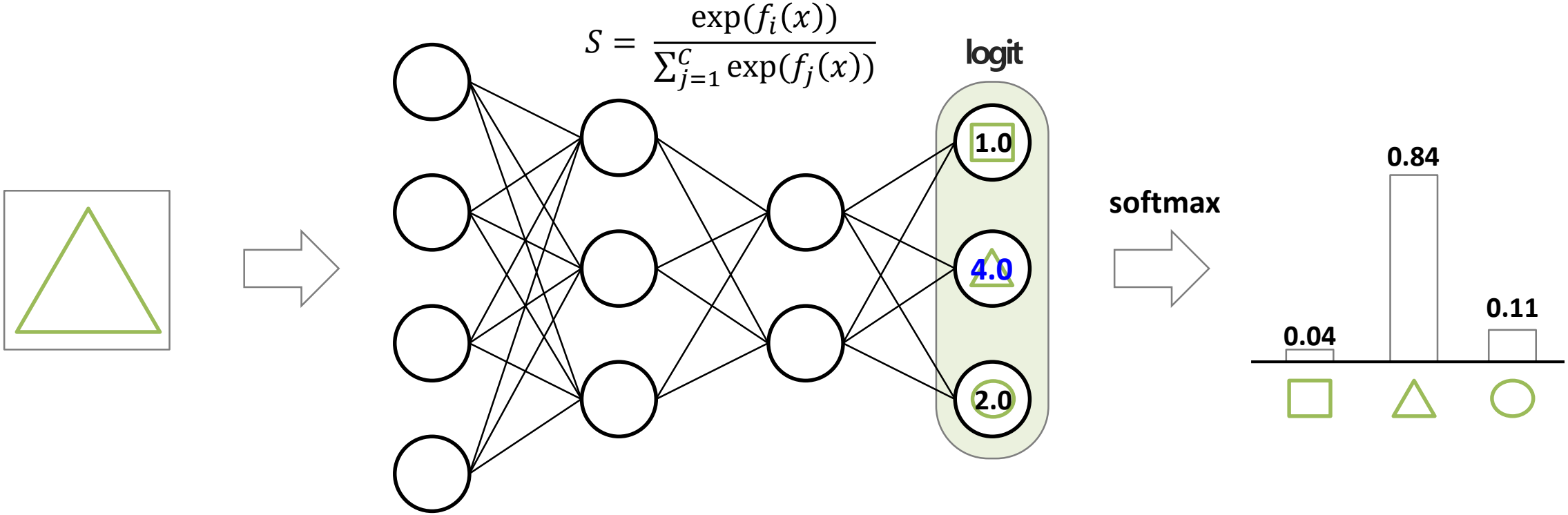
- 모델에서 출력된 logit은 일반적으로 softmax 함수를 거쳐서 확률 분포로 변환
- 이 때 softmax 함수는 지수 함수를 사용하기 때문에 logit의 **작은 변화에도 확률에 큰 변화**를 일으킴



Baseline

Motivations

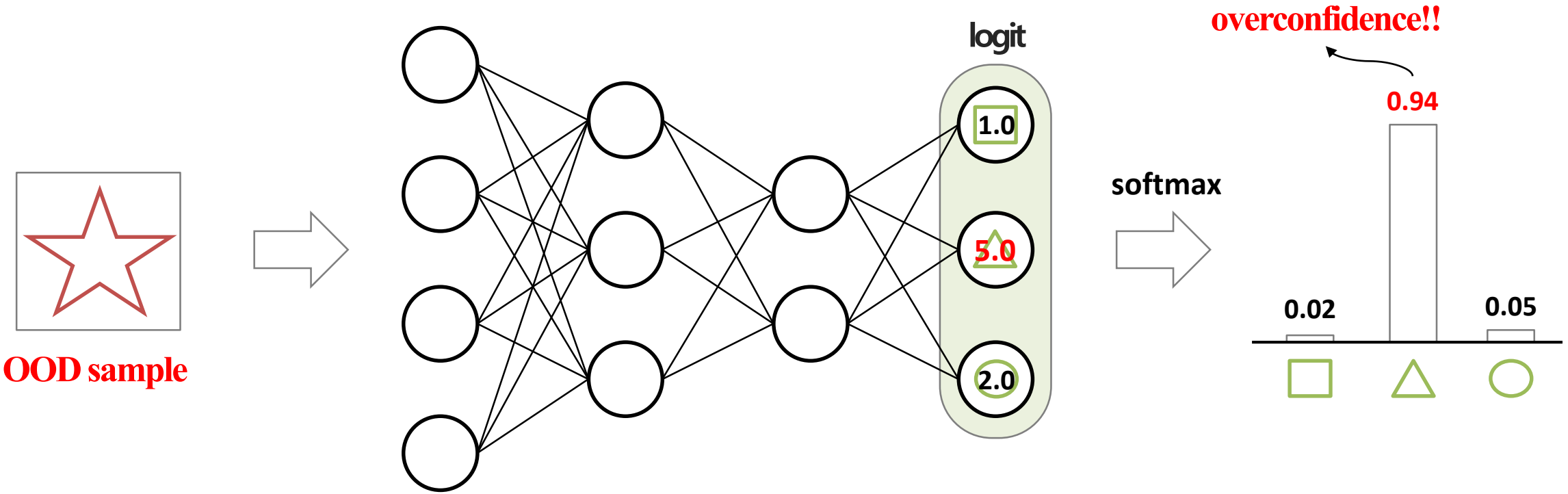
- 모델에서 출력된 logit은 일반적으로 softmax 함수를 거쳐서 확률 분포로 변환
- 이 때 softmax 함수는 지수 함수를 사용하기 때문에 logit의 **작은 변화에도 확률에 큰 변화**를 일으킴



Baseline

Motivations

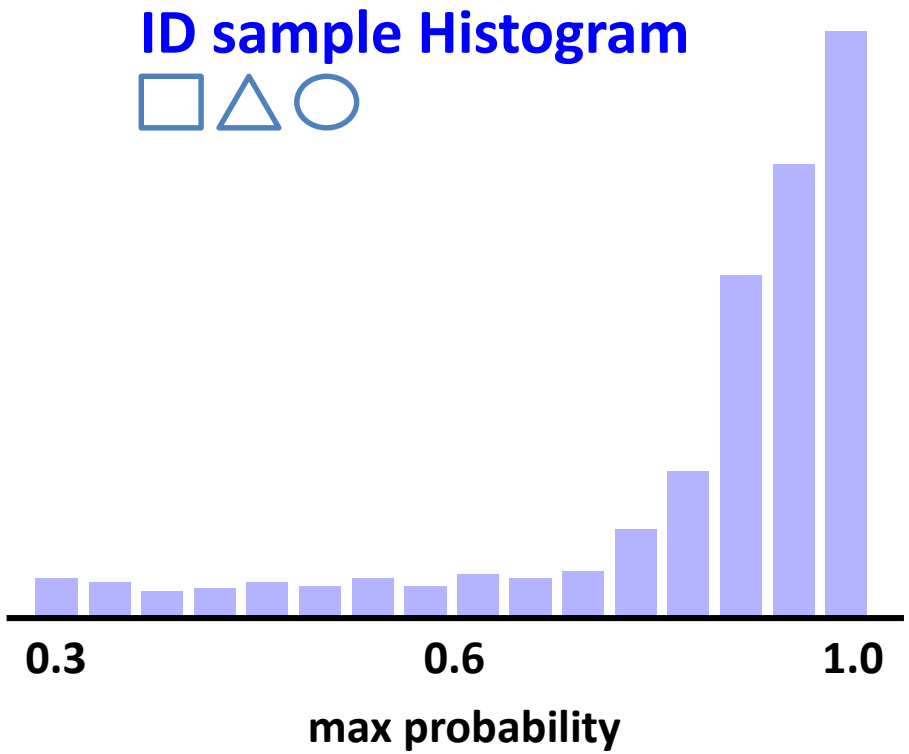
- ❖ 이러한 특성에 의해서 잘못 예측된 샘플이나 OOD 샘플에 대해서도 큰 확률값을 출력 (overconfidence problem)
- ❖ 하지만 저자들은 overconfidence problem에도 불구하고 **OOD 샘플의 최대 확률값보다 ID 샘플이 도 높다**는 것을 발견

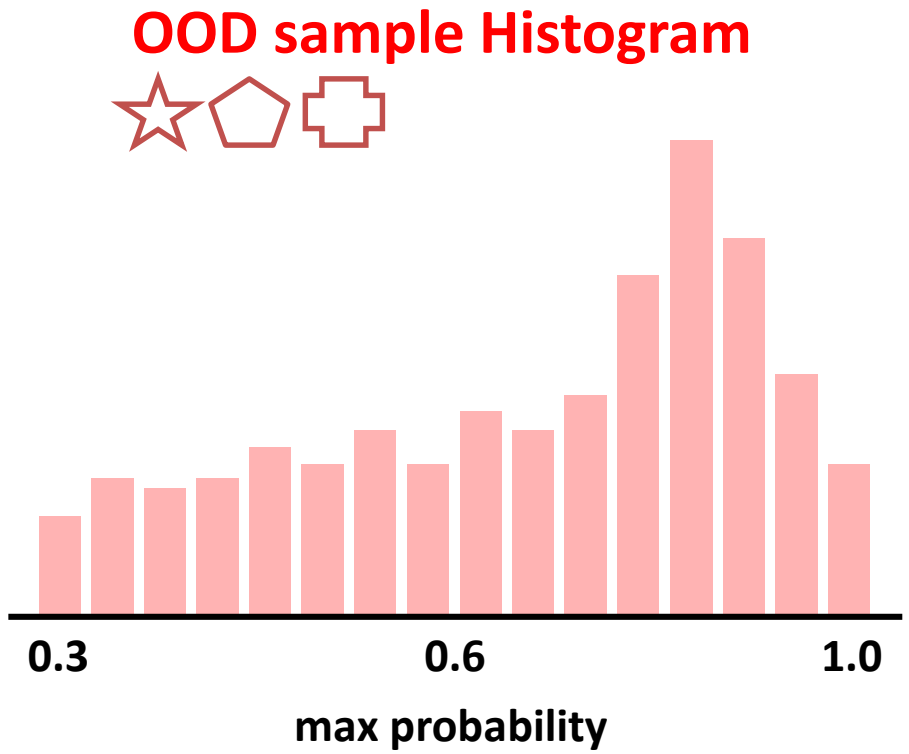


Baseline

Motivations

- ❖ 이러한 특성에 의해서 잘못 예측된 샘플이나 OOD 샘플에 대해서도 큰 확률값을 출력 (overconfidence problem)
- ❖ 하지만 저자들은 overconfidence problem에도 불구하고 **OOD 샘플의 최대 확률값보다 ID 샘플이 도 높다**는 것을 발견

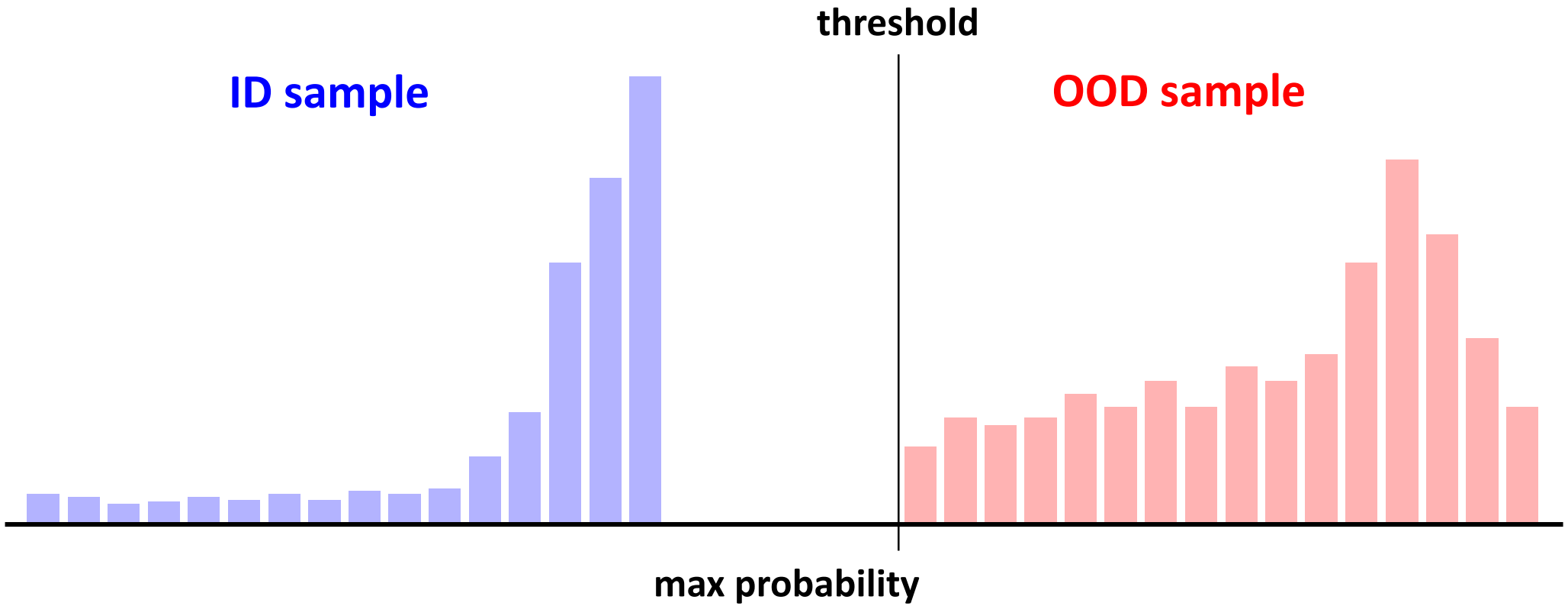




Baseline

OOD score

- ❖ 이러한 발견을 기반으로 확률 최대값을 OOD score로 정의하고 특정 threshold를 통해 ID와 OOD를 분류
- ❖ 제안 OOD score를 랜덤하게 ID와 OOD 둘 중 하나를 선택하는 방법과 비교하여 더 좋은 성능을 낸다는 것을 보임

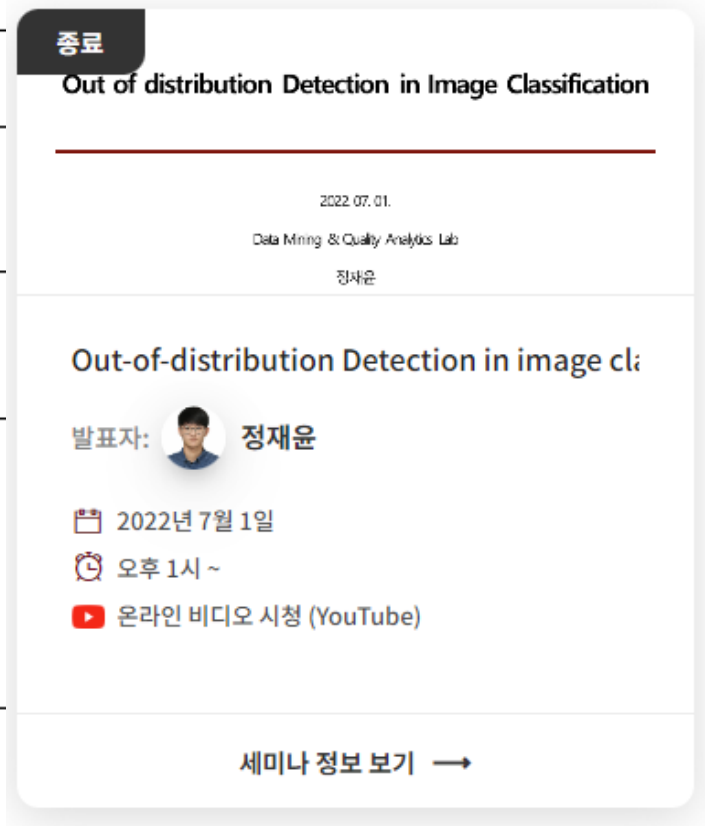


Baseline

OOD score

- ❖ 이러한 발견을 기반으로 확률 최대값을 OOD score로 정의하고 특정 threshold를 통해 ID와 OOD를 분류
- ❖ 제안 OOD score를 랜덤하게 ID와 OOD 둘 중 하나를 선택하는 방법과 비교하여 더 좋은 성능을 낸다는 것을 보임

In-Distribution / Out-of-Distribution	AUPR ut/Base	Pred. Prob (mean)
CIFAR-10/SUN	97/67	72
CIFAR-10/Gaussian	95/51	77
CIFAR-10/All	98/76	74
CIFAR-100/SUN	96/73	56
CIFAR-100/Gaussian	30/57	77
CIFAR-100/All	96/79	63
MNIST/Omniglot	96/48	86
MNIST/notMNIST	38/50	92
MNIST/CIFAR-10bw	95/50	87
MNIST/Gaussian	91/50	91
MNIST/Uniform	98/50	83
MNIST/All	98/80	89



Enhancing the Reliability of Out-of-Distribution Image Detection in Neural Networks (2018, ICLR)

ODIN

Introduction

- ❖ 앞서 활용된 max probability의 ID와 OOD 차이를 극대화시키기 위해서 **temperature scaling**과 **input preprocessing**을 제안
- ❖ 추가적인 학습이 없는 간단한 방법에도 불구하고 baseline 대비 높은 수준의 성능 향상을 보여줌

Published as a conference paper at ICLR 2018

ENHANCING THE RELIABILITY OF OUT-OF-DISTRIBUTION IMAGE DETECTION IN NEURAL NETWORKS

Shiyu Liang
Coordinated Science Lab, Department of ECE
University of Illinois at Urbana-Champaign
sliang26@illinois.edu

Yixuan Li
University of Wisconsin-Madison*
sharonli@cs.wisc.edu

R. Srikant
Coordinated Science Lab, Department of ECE
University of Illinois at Urbana-Champaign
rsrikant@illinois.edu

ABSTRACT

We consider the problem of detecting *out-of-distribution* images in neural networks. We propose *ODIN*, a simple and effective method that does not require any change to a pre-trained neural network. Our method is based on the observation that using temperature scaling and adding small perturbations to the input can separate the softmax score distributions between in- and out-of-distribution images, allowing for more effective detection. We show in a series of experiments that *ODIN* is compatible with diverse network architectures and datasets. It consistently outperforms the baseline approach (Hendrycks & Gimpel, 2017) by a large margin, establishing a new state-of-the-art performance on this task. For example, *ODIN* reduces the false positive rate from the baseline 34.7% to 4.3% on the DenseNet (applied to CIFAR-10 and Tiny-ImageNet) when the true positive rate is 95%.

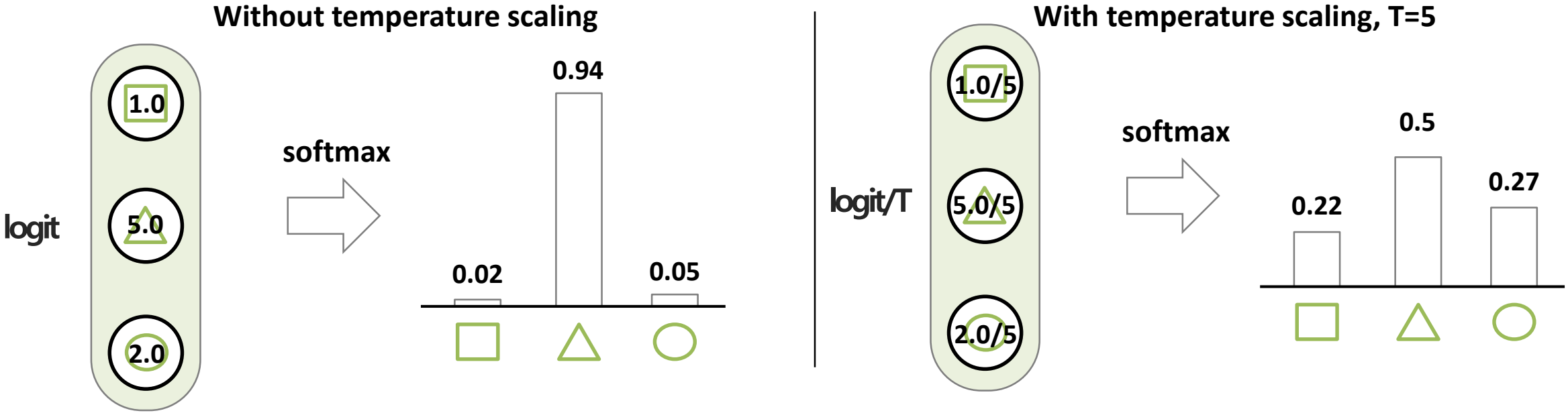
	Out-of-distribution dataset	FPR (95% TPR) ↓	Detection Error ↓	AUROC ↑	AUPR In ↑	AUPR Out ↑
		Baseline (Hendrycks & Gimpel, 2017) / ODIN				
Dense-BC CIFAR-10	TinyImageNet (crop)	34.7/**4.3**	10.0/**4.7**	95.3/**99.1**	96.4/**99.1**	93.8/**99.1**
	TinyImageNet (resize)	40.8/**7.5**	11.5/**6.1**	94.1/**98.5**	95.1/**98.6**	92.4/**98.5**
	LSUN (crop)	39.3/**11.4**	10.2/**7.2**	94.8/**97.9**	96.0/**98.0**	93.1/**97.9**
	LSUN (resize)	33.6/**3.8**	9.8/**4.4**	95.4/**99.2**	96.4/**99.3**	94.0/**99.2**
	Uniform	23.5/**0.0**	5.3/**0.5**	96.5/**99.0**	97.8/**100.0**	93.0/**99.0**
	Gaussian	12.3/**0.0**	4.7/**0.2**	97.5/**100.0**	98.3/**100.0**	95.9/**100.0**
Dense-BC CIFAR-100	TinyImageNet (crop)	67.8/**26.9**	36.4/**12.9**	83.0/**94.5**	85.3/**94.7**	80.8/**94.5**
	TinyImageNet (resize)	82.2/**57.0**	43.6/**22.7**	70.4/**85.5**	71.4/**86.0**	68.6/**84.8**
	LSUN (crop)	69.4/**18.6**	37.2/**9.7**	83.7/**96.6**	86.2/**96.8**	80.9/**96.5**
	LSUN (resize)	83.3/**58.0**	44.1/**22.3**	70.6/**86.0**	72.5/**87.1**	68.0/**84.8**
	Uniform	100.0/**100.0**	35.86/**17.9**	43.1/**99.5**	63.2/**87.5**	41.9/**65.1**
	Gaussian	100.0/**100.0**	41.2/**38.0**	30.6/**40.5**	53.4/**60.5**	37.6/**40.9**

ODIN

Temperature Scaling

- Temperature는 확률 분포의 sharpness를 조절하는 인자로 knowledge distillation연구에서 제안
- 1보다 큰 temperature는 확률 분포를 더 soft하게 만듦으로써 더 많은 정보량을 가지게 됨

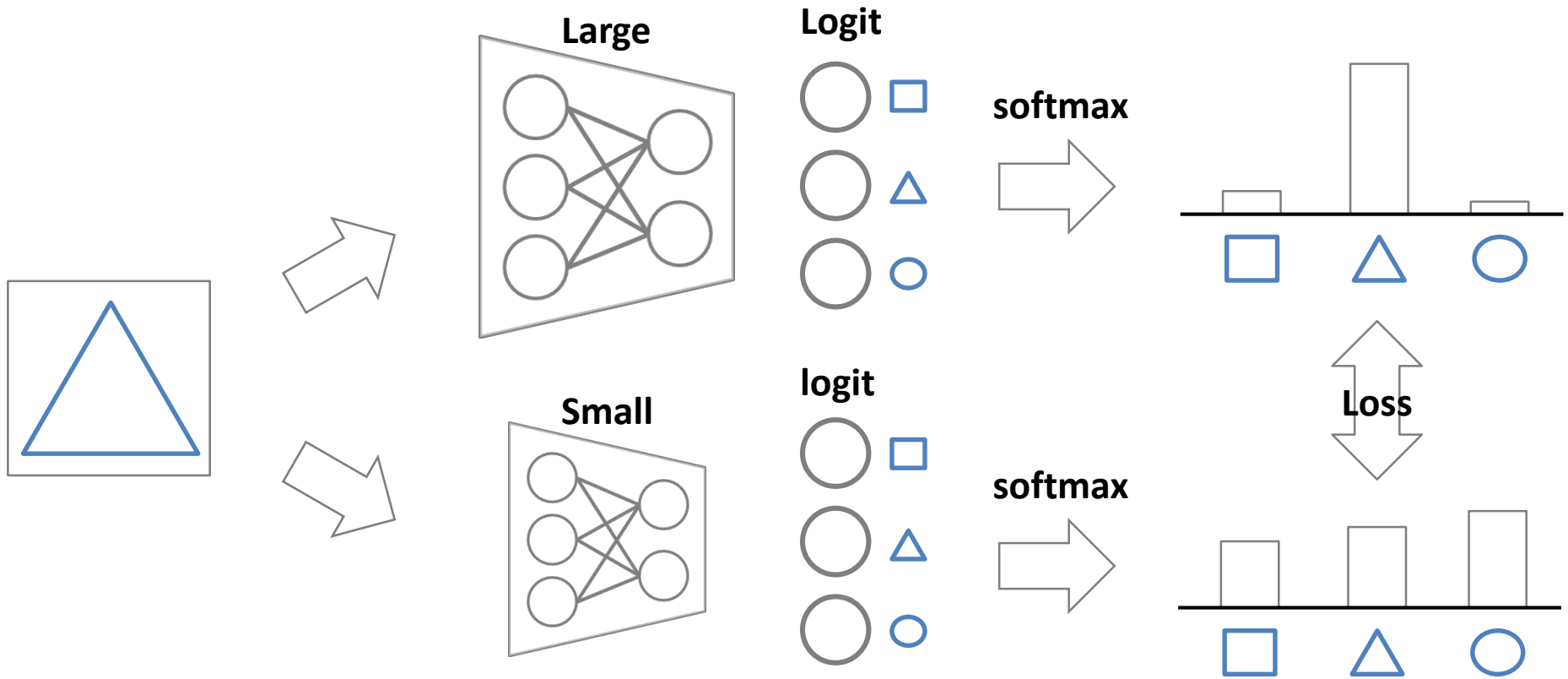
$$S = \frac{\exp(f_i(x))}{\sum_{j=1}^{C} \exp(f_j(x))} \Rightarrow S = \frac{\exp(f_i(x)/\boldsymbol{T})}{\sum_{j=1}^{C} \exp(f_j(x)/\boldsymbol{T})}$$



ODIN

Temperature Scaling

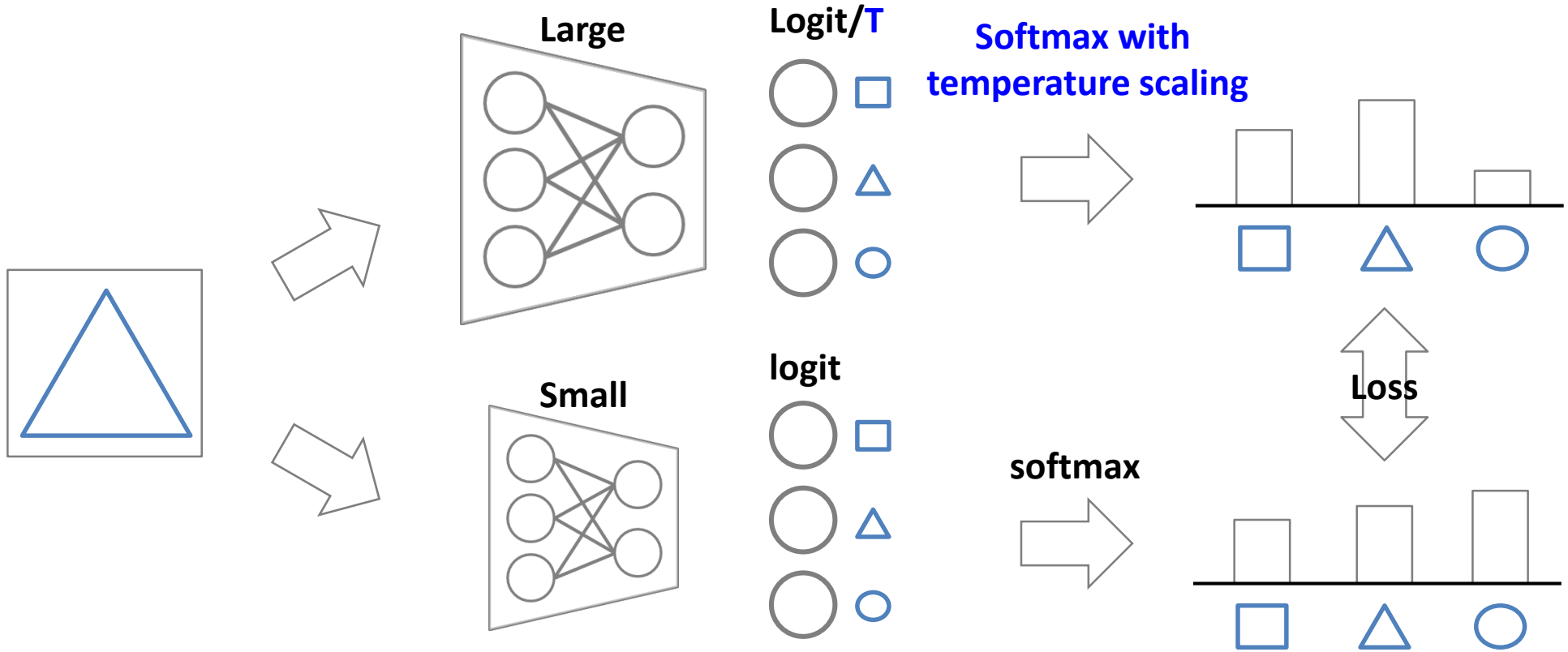
- Temperature는 확률 분포의 sharpness를 조절하는 인자로 knowledge distillation연구에서 제안
- 1보다 큰 temperature는 확률 분포를 더 soft하게 만듦으로써 더 많은 정보량을 가지게 됨



ODIN

Temperature Scaling

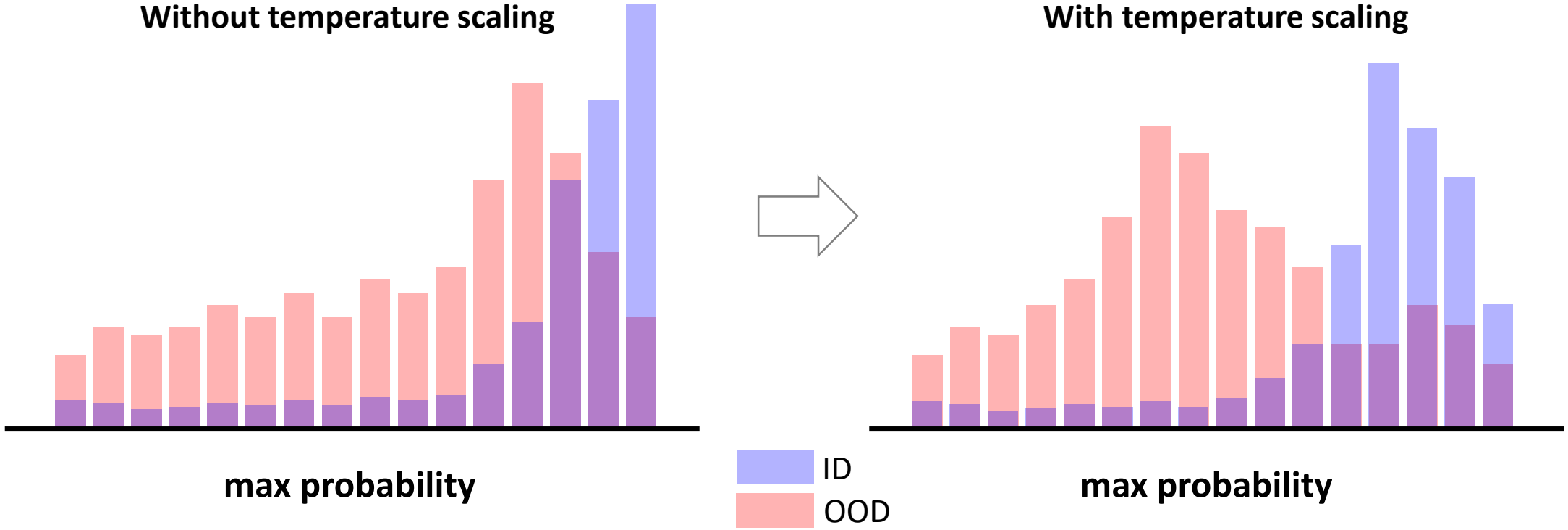
- Temperature는 확률 분포의 sharpness를 조절하는 인자로 knowledge distillation연구에서 제안
- 1보다 큰 temperature는 확률 분포를 더 soft하게 만듦으로써 더 많은 정보량을 가지게 됨



ODIN

Temperature Scaling

- ❖ 저자들은 temperature를 큰 값으로 설정하면 ID와 OOD의 max probability 차이가 커진다는 것을 발견
- ❖ 또한, softmax 함수와 temperature 사이의 관계를 수식화하여 큰 값의 temperature가 OOD 탐지에 효과적임을 보임



ODIN

Temperature Scaling

- 저자들은 temperature를 큰 값으로 설정하면 ID와 OOD의 max probability 차이가 커진다는 것을 발견
- 또한, softmax 함수와 temperature 사이의 관계를 수식화하여 큰 값의 temperature가 OOD 탐지에 효과적임을 보임

$$S_{\hat{y}} = \frac{\exp(f_{\hat{y}}(x)/T)}{\sum_{j=1}^{C}\exp(f_j(x)/T)} \approx \frac{1}{C - \frac{1}{T}\sum_i\left[f_{\hat{y}}(x) - f_i(x)\right] + \frac{1}{2T^2}\sum_i\left[f_{\hat{y}}(x) - f_i(x)\right]^2}$$

$$U_1(x) = \frac{1}{C-1}\sum_{i\neq\hat{y}}\left[f_{\hat{y}}(x) - f_i(x)\right]$$

$$U_2(x) = \frac{1}{C-1}\sum_{i\neq\hat{y}}\left[f_{\hat{y}}(x) - f_i(x)\right]^2$$

$$let\, f_{\hat{y}}(x) - f_i(x) = \delta, U_2(x) = \underbrace{\frac{1}{C-1}\sum_{i\neq\hat{y}}\left[\delta_i - \bar{\delta}\right]^2}_{\text{분산}} + \underbrace{\bar{\delta}^2}_{\text{평균}}$$

ODIN

Temperature Scaling

- 저자들은 temperature를 큰 값으로 설정하면 ID와 OOD의 max probability 차이가 커진다는 것을 발견
- 또한, softmax 함수와 temperature 사이의 관계를 수식화하여 큰 값의 temperature가 OOD 탐지에 효과적임을 보임

$$S_{\hat{y}} \propto \frac{1}{T}\left(U_1 - \frac{U_2}{2T}\right)$$

$$U_1(x) = \frac{1}{C-1}\sum_{i \neq \hat{y}}\left[f_{\hat{y}}(x) - f_i(x)\right]$$

예측 클래스와 다른 클래스의 logit 차이 정도

$$U_2(x) = \frac{1}{C-1}\sum_{i \neq \hat{y}}\left[f_{\hat{y}}(x) - f_i(x)\right]^2$$

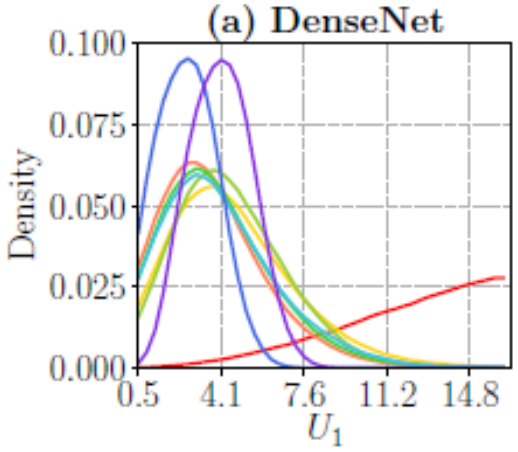
다른 클래스들 사이에 logit 차이 정도

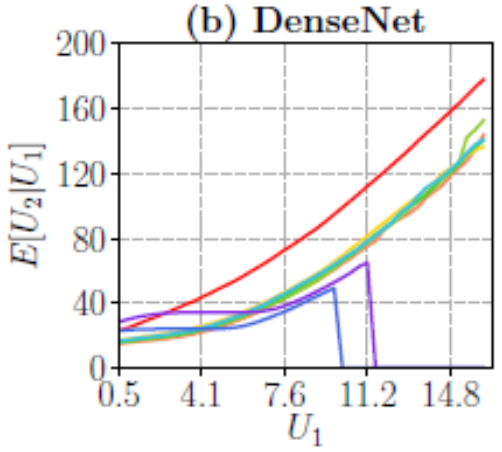
ODIN

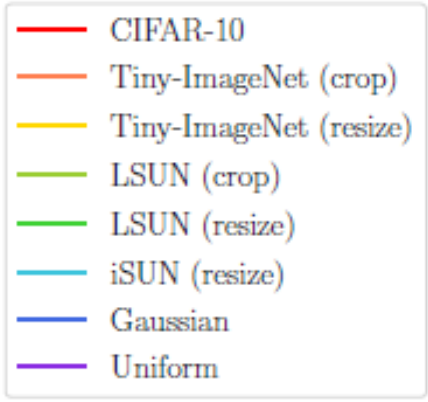
Temperature Scaling

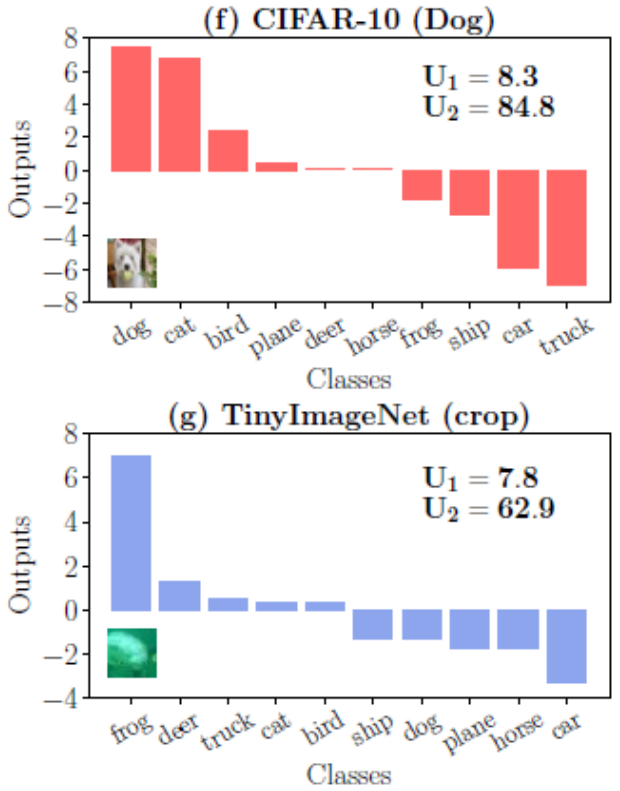
- 저자들은 temperature를 큰 값으로 설정하면 ID와 OOD의 max probability 차이가 커진다는 것을 발견
- 또한, softmax 함수와 temperature 사이의 관계를 수식화하여 큰 값의 temperature가 OOD 탐지에 효과적임을 보임

$$S_{\hat{y}} \propto \frac{1}{T}\left(U_1 - \frac{U_2}{2T}\right)$$







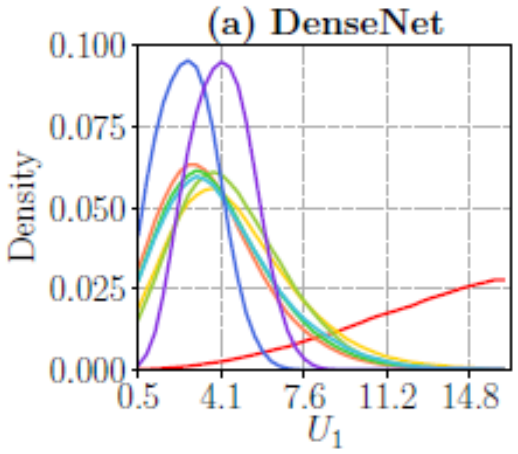


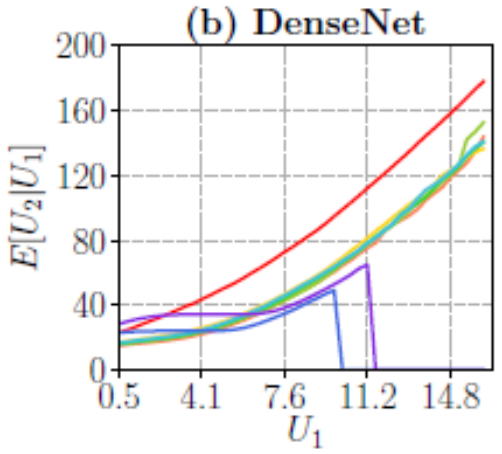
ODIN

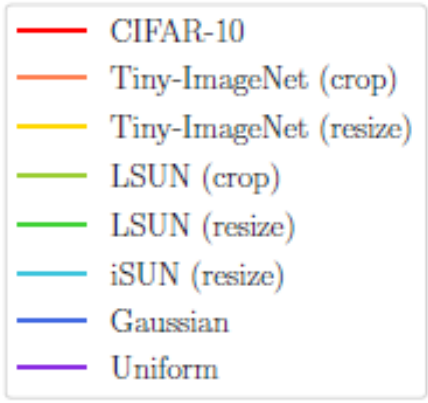
Temperature Scaling

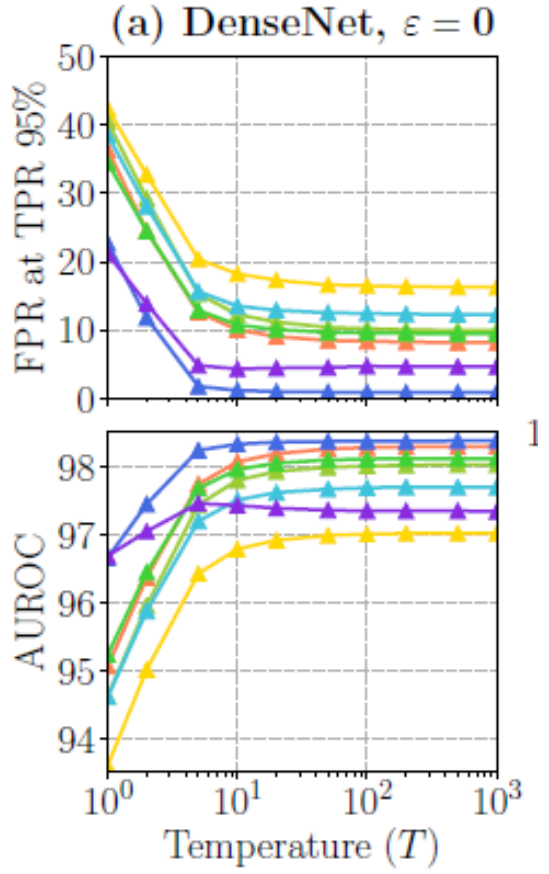
- 저자들은 temperature를 큰 값으로 설정하면 ID와 OOD의 max probability 차이가 커진다는 것을 발견
- 또한, softmax 함수와 temperature 사이의 관계를 수식화하여 큰 값의 temperature가 OOD 탐지에 효과적임을 보임

$$S_{\hat{y}} \propto \frac{1}{T}\left(U_1 - \frac{U_2}{2T}\right)$$







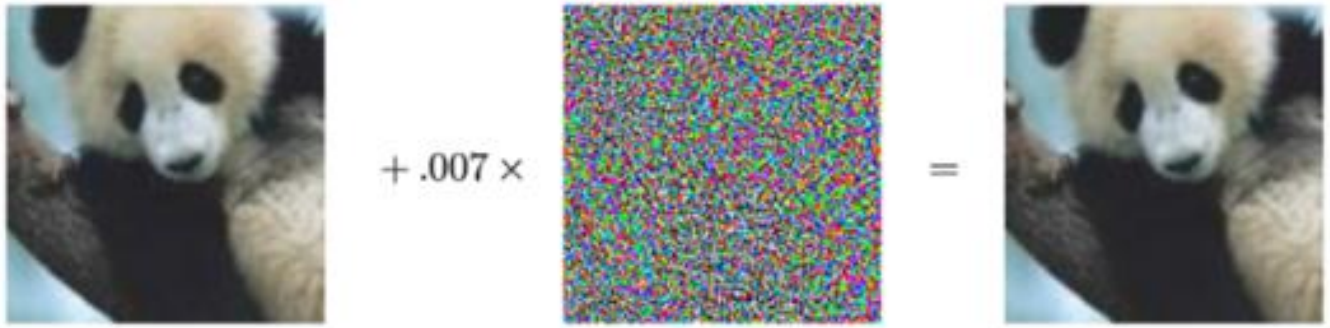


ODIN

Input Preprocessing

- Adversarial noise는 입력 변수의 출력층 활성화값을 증폭시켜 모델의 max probability를 낮춤
- Fast Gradient Sign Method (FGSM)은 loss를 증가시키는 방향의 gradient를 계산하여 adversarial noise를 쉽게 얻는 방법

$$\tilde{x} = x + \underbrace{\epsilon sign(\nabla_x J(\theta, x, y))}$$



ODIN

Input Preprocessing

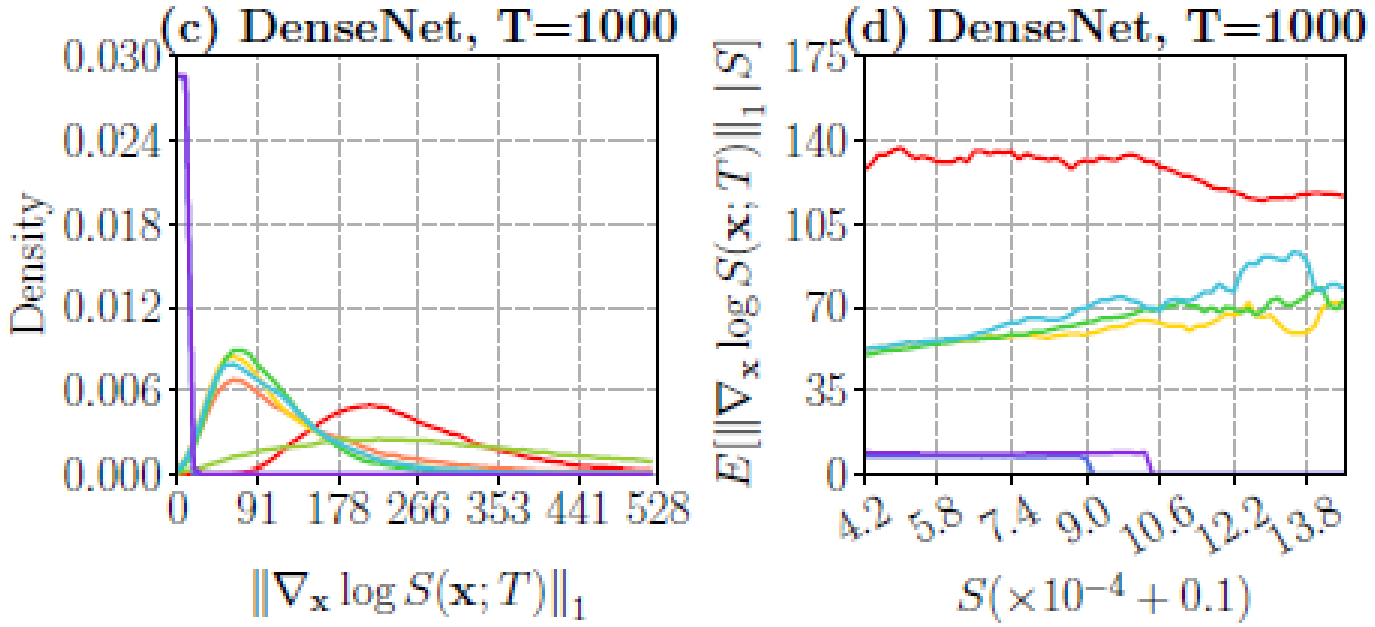
- 저자들은 이와 반대로 입력 이미지에 max probability를 낮추는 noise를 제거하여 max probability를 높임
- 또한, ID가 OOD보다 max probability값 증가폭이 높다는 것을 실험적으로 증명하며 OOD 탐지에 효과적임을 보임

Loss를 커지게 만드는 노이즈 추가

$$\tilde{x} = x + \epsilon sign(\nabla_x J(\theta, x, y))$$

Max probability를 낮추는 노이즈 제거

$$\tilde{x} = x - \epsilon sign(-\nabla_x S_{\hat{y}}(x))$$



A Simple Unified Framework for Detecting Out-of-Distribution Samples and Adversarial Attacks (2018, NeurIPS)

Mahalanobis-Based Score

Introduction

- ❖ 모델의 인코더 부분에서 출력되는 feature의 분포가 클래스별로 **multivariate Gaussian distribution**을 따른다는 가정
- ❖ 테스트 샘플의 feature vector와 Gaussian distribution 평균 vector의 **Mahalanobis distance**를 활용한 OOD score 제안

A Simple Unified Framework for Detecting Out-of-Distribution Samples and Adversarial Attacks

Kimin Lee[1], Kibok Lee[2], Honglak Lee[3,2], Jinwoo Shin[1,4]
[1]Korea Advanced Institute of Science and Technology (KAIST)
[2]University of Michigan
[3]Google Brain
[4]AItrics

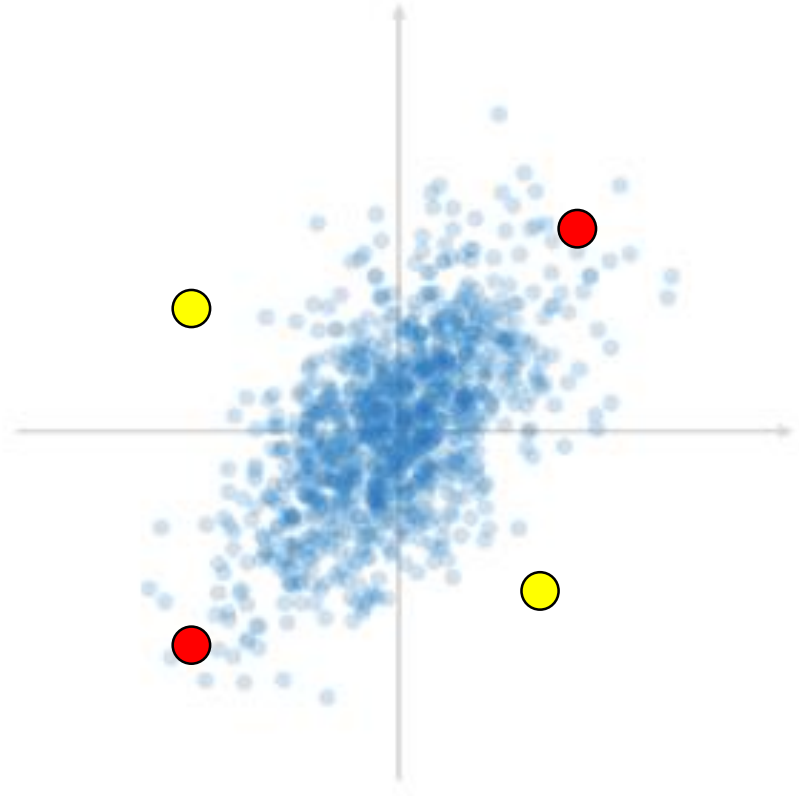
Abstract

Detecting test samples drawn sufficiently far away from the training distribution statistically or adversarially is a fundamental requirement for deploying a good classifier in many real-world machine learning applications. However, deep neural networks with the softmax classifier are known to produce highly overconfident posterior distributions even for such abnormal samples. In this paper, we propose a simple yet effective method for detecting any abnormal samples, which is applicable to any pre-trained softmax neural classifier. We obtain the class conditional Gaussian distributions with respect to (low- and upper-level) features of the deep models under Gaussian discriminant analysis, which result in a confidence score based on the Mahalanobis distance. While most prior methods have been evaluated for detecting either out-of-distribution or adversarial samples, but not both, the proposed method achieves the state-of-the-art performances for both cases in our experiments. Moreover, we found that our proposed method is more robust in harsh cases, e.g., when the training dataset has noisy labels or small number of samples. Finally, we show that the proposed method enjoys broader usage by applying it to class-incremental learning: whenever out-of-distribution samples are detected, our classification rule can incorporate new classes well without further training deep models.

Mahalanobis-Based Score

Mahalanobis Distance

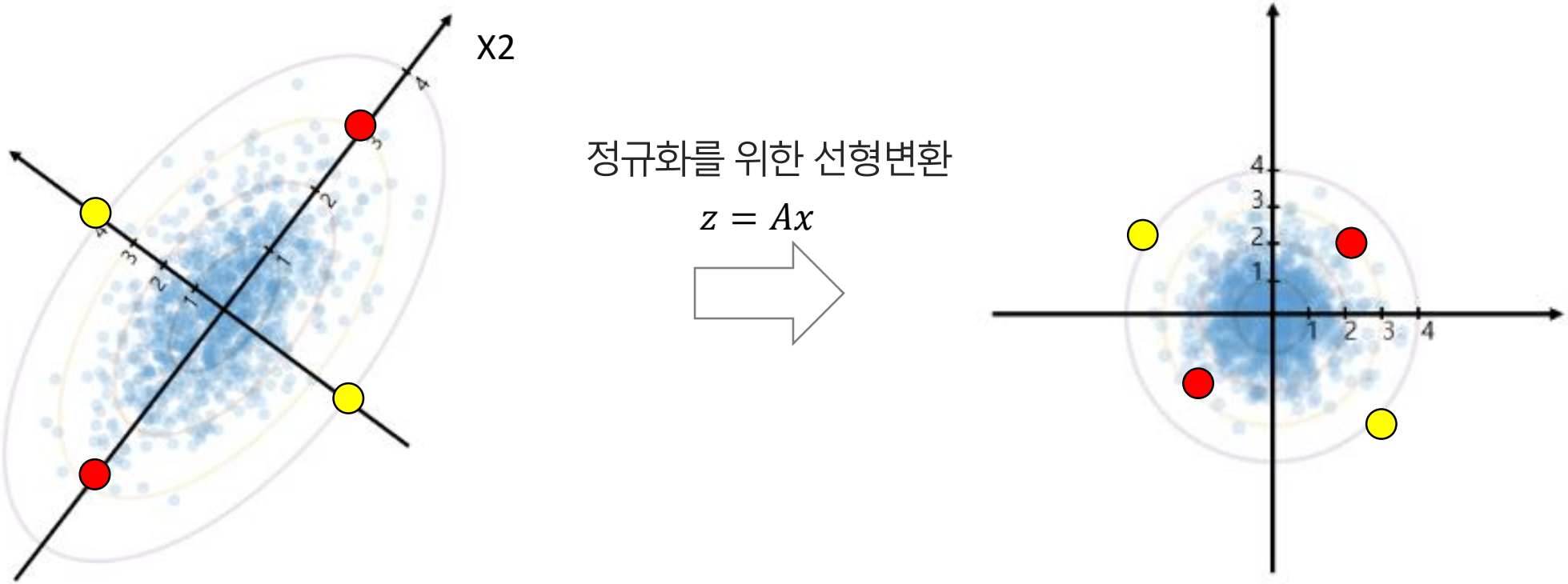
- Mahalanobis distance는 한 샘플이 확률분포상에서 평균과의 거리가 표준편차의 몇 배인지를 나타내는 값
- 다음과 같은 확률 분포가 주어졌을 때 노란점 사이의 거리와 빨간점 사이의 거리 중 어떤 것이 더 멀까?



Mahalanobis-Based Score

Mahalanobis Distance

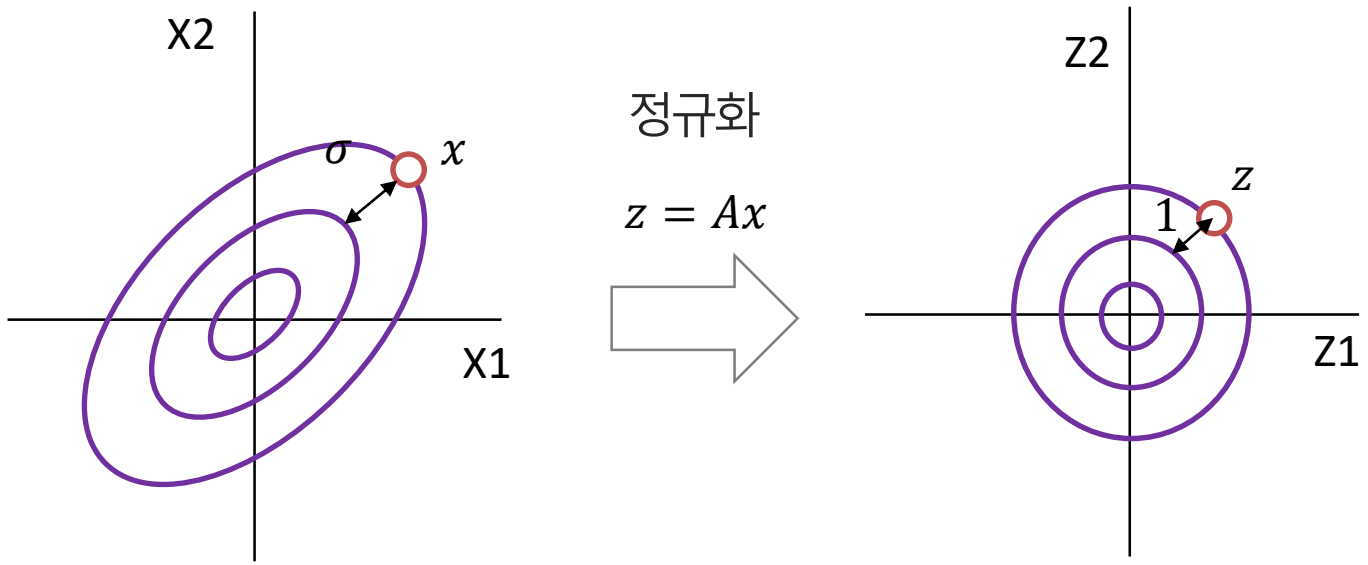
- Mahalanobis distance는 한 샘플이 확률분포상에서 평균과의 거리가 표준편차의 몇 배인지를 나타내는 값
- 다음과 같은 확률 분포가 주어졌을 때 노란점 사이의 거리와 빨간점 사이의 거리 중 어떤 것이 더 멀까?



Mahalanobis-Based Score

Mahalanobis Distance

- Mahalanobis distance는 한 샘플이 확률분포상에서 평균과의 거리가 표준편차의 몇 배인지를 나타내는 값
- Mahalanobis distance는 정규화된 분포로 선형변환된 샘플의 Euclidean distance



$$Euclidean\ distance$$

$$d_E = \sqrt{x^T x}$$

$$Mahalanobis\ distance$$

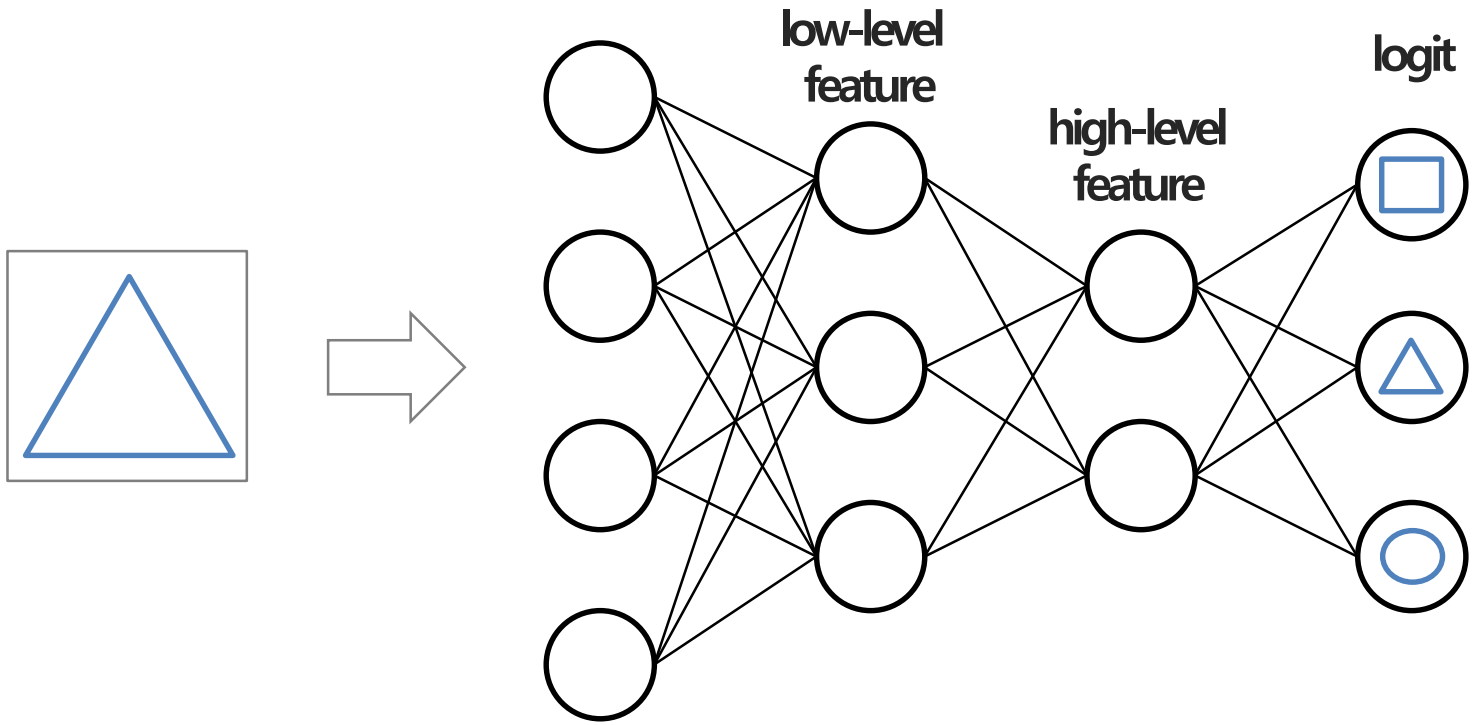
$$d_M = \sqrt{z^T z} = \sqrt{(Ax)^T (Ax)}$$

$$= \sqrt{x^T A^T A x} = \sqrt{x^T \sum^{-1} x}$$

Mahalanobis-Based Score

Mahalanobis Distance-Based Confidence Score

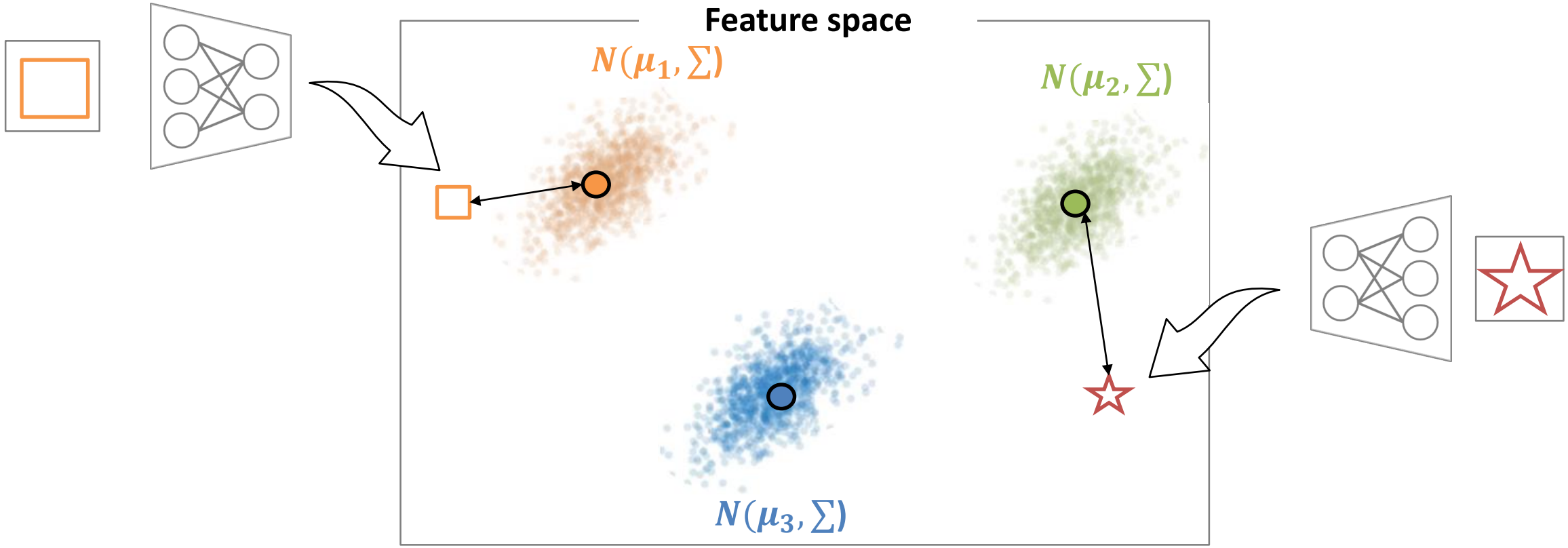
❖ 저자들은 logit을 활용하기 보다는 더 많은 정보가 담겨 있는 feature vector를 활용한 방법론을 제안



Mahalanobis-Based Score

Mahalanobis Distance-Based Confidence Score

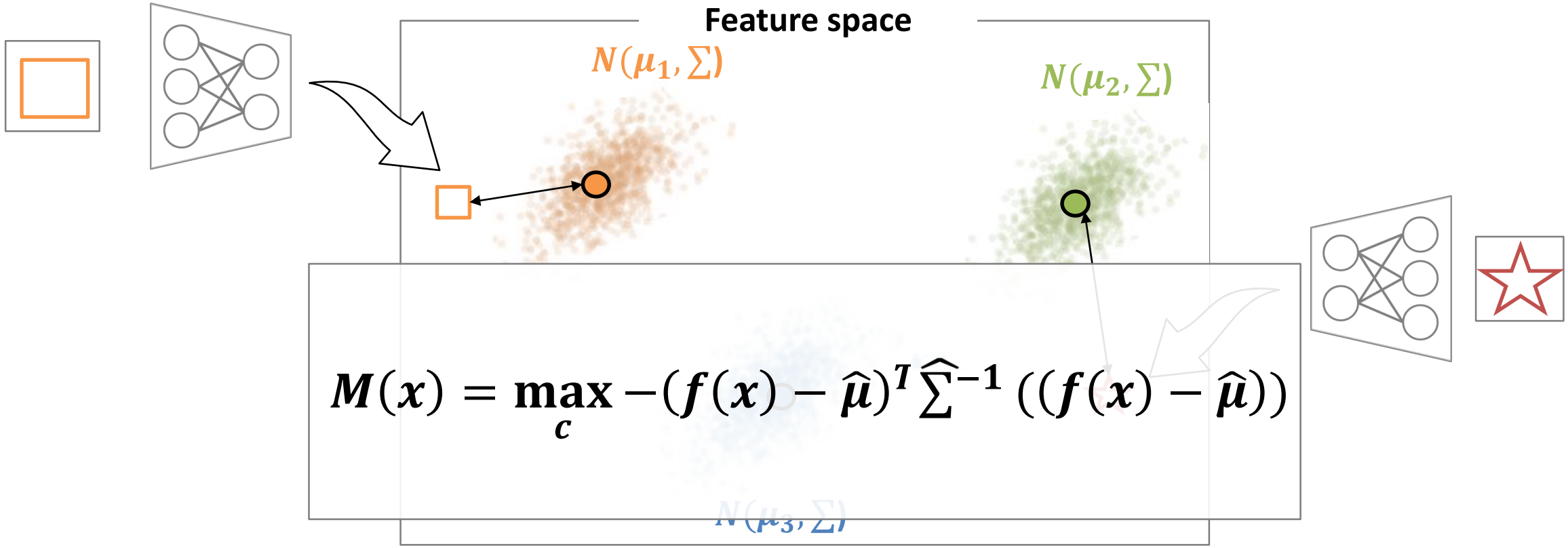
- 모델에서 출력되는 특징 벡터들이 특징 공간 상에서 class-dependent multivariate Gaussian distribution을 따른다고 가정
- 새로운 샘플이 입력되었을 때, 해당 특징 벡터와 각 클래스별 분포 사이의 Mahalanobis distance를 통해 OOD 구별



Mahalanobis-Based Score

Mahalanobis Distance-Based Confidence Score

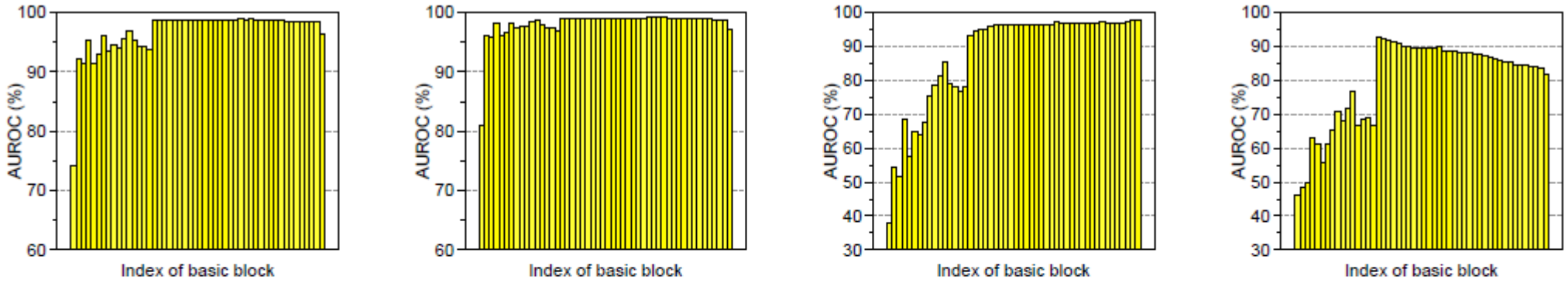
- 모델에서 출력되는 특징 벡터들이 특징 공간 상에서 class-dependent multivariate Gaussian distribution을 따른다고 가정
- 새로운 샘플이 입력되었을 때, 해당 특징 벡터와 각 클래스별 분포 사이의 Mahalanobis distance를 통해 OOD 구별



Mahalanobis-Based Score

Feature Ensemble

- ❖ 여기에 저자들은 각 layer별로 다른 수준의 특징이 출력된다는 것을 기반으로 feature ensemble 방법 제안
- ❖ Esemble을 통해서 calibration 효과를 얻을 수 있으며, 특히 데이터셋의 복잡도마다 OOD 탐지에 효과적인 특징 활용 가능



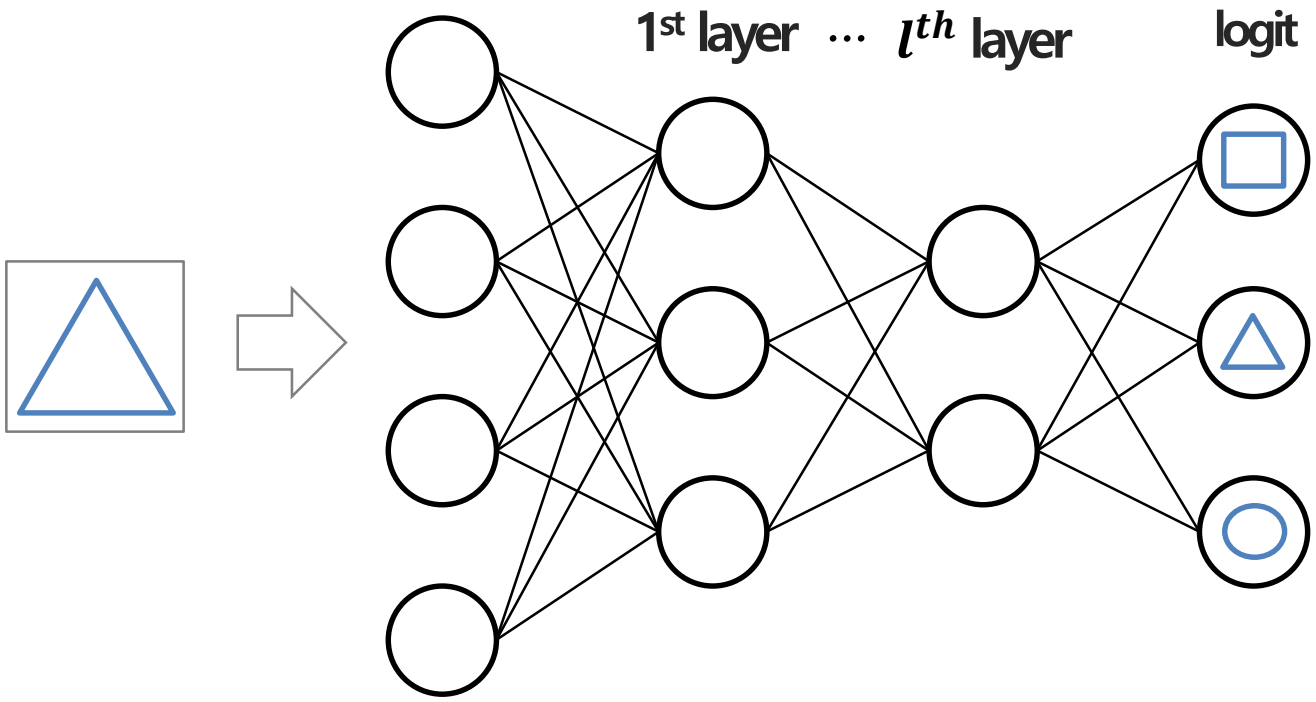
(a) TinyImageNet (b) LSUN (c) SVHN (d) DeepFool

Figure 2: AUROC (%) of threshold-based detector using the confidence score in (2) computed at different basic blocks of DenseNet trained on CIFAR-10 dataset. We measure the detection performance using (a) TinyImageNet, (b) LSUN, (c) SVHN and (d) adversarial (DeepFool) samples.

Mahalanobis-Based Score

Feature Ensemble

- ❖ 여기에 저자들은 각 layer별로 다른 수준의 특징이 출력된다는 것을 기반으로 feature ensemble 방법 제안
- ❖ Esemble을 통해서 calibration 효과를 얻을 수 있으며, 특히 데이터셋의 복잡도마다 OOD 탐지에 효과적인 특징 활용 가능



$$M_1(x) = \max_c -(f_1(x) - \hat{\mu}_1)^T \sum_1^{-1} ((f_1(x) - \hat{\mu}_1))$$

$$M_2(x) = \max_c -(f_2(x) - \hat{\mu}_2)^T \sum_2^{-1} ((f_2(x) - \hat{\mu}_2))$$

$$\vdots$$

$$M_l(x) = \max_c -(f_l(x) - \hat{\mu}_l)^T \sum_l^{-1} ((f_l(x) - \hat{\mu}_l))$$

$$M(x) = \sum_l \alpha_l M_l$$

Mahalanobis-Based Score

Input Preprocessing

❖ 제안한 Mahalanobis distance-based confidence score가 커지는 방향의 노이즈를 추가하여 탐지 능력 향상

Adversarial attack

Loss를 커지게 만드는 노이즈 추가

$$\tilde{x} = x + \epsilon sign(\nabla_x J(\theta, x, y))$$

ODIN

Max probability를 낮추는 노이즈 제거

$$\tilde{x} = x - \epsilon sign(-\nabla_x S_{\hat{y}}(x))$$

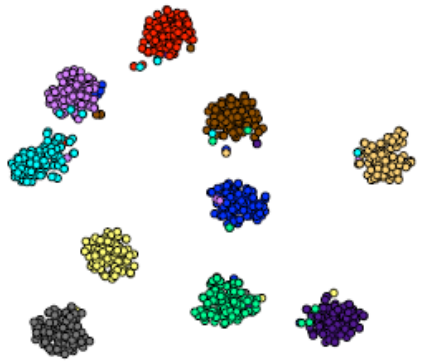
Mahalanobis

Confidence score를 높이는 노이즈 추가

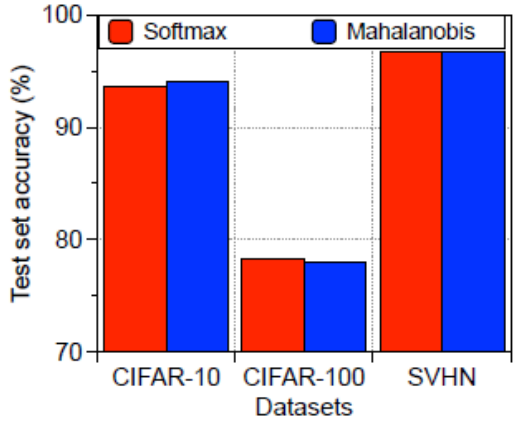
$$\tilde{x} = x + \epsilon sign(+\nabla_x M(x))$$

Mahalanobis-Based Score

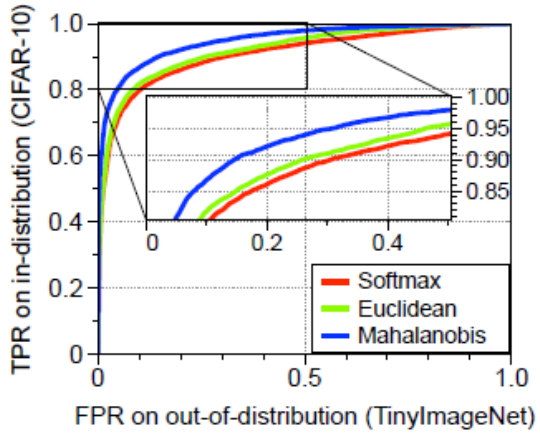
Experimental Results



(a) Visualization by t-SNE



(b) Classification accuracy



(c) ROC curve

Method	Feature ensemble	Input pre-processing	TNR at TPR 95%	AUROC	Detection accuracy	AUPR in	AUPR out
Baseline [13]	-	-	32.47	89.88	85.06	85.40	93.96
ODIN [21]	-	-	86.55	96.65	91.08	92.54	98.52
Mahalanobis (ours)	-	-	54.51	93.92	89.13	91.56	95.95
	-	✓	92.26	98.30	93.72	96.01	99.28
	✓	-	91.45	98.37	93.55	96.43	99.35
	✓	✓	**96.42**	**99.14**	**95.75**	**98.26**	**99.60**

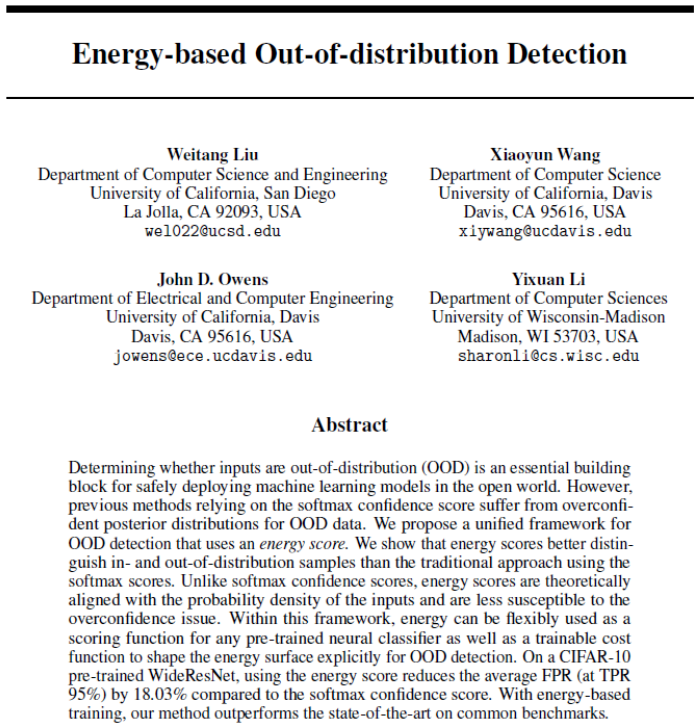
Energy-based Out-of-Distribution Detection

(2020, NeurIPS)

Energy-Based Score

Introduction

- ❖ 이전에 연구된 softmax score 기반 방법론들과는 다르게 입력 샘플의 확률 밀도와 관련이 있는 energy score를 제안
- ❖ Energy score가 OOD detection에 중요한 정보를 가지고 있다는 것을 이론적으로 증명하며 우수한 성능을 보임
- ❖ 또한, energy score가 discriminative classification model에서 쉽게 계산될 수 있음을 보임



Energy-based Out-of-distribution Detection

Weitang Liu
Department of Computer Science and Engineering
University of California, San Diego
La Jolla, CA 92093, USA
wel022@ucsd.edu

Xiaoyun Wang
Department of Computer Science
University of California, Davis
Davis, CA 95616, USA
xiywang@ucdavis.edu

John D. Owens
Department of Electrical and Computer Engineering
University of California, Davis
Davis, CA 95616, USA
jowens@ece.ucdavis.edu

Yixuan Li
Department of Computer Sciences
University of Wisconsin-Madison
Madison, WI 53703, USA
sharonli@cs.wisc.edu

Abstract

Determining whether inputs are out-of-distribution (OOD) is an essential building block for safely deploying machine learning models in the open world. However, previous methods relying on the softmax confidence score suffer from overconfident posterior distributions for OOD data. We propose a unified framework for OOD detection that uses an *energy score*. We show that energy scores better distinguish in- and out-of-distribution samples than the traditional approach using the softmax scores. Unlike softmax confidence scores, energy scores are theoretically aligned with the probability density of the inputs and are less susceptible to the overconfidence issue. Within this framework, energy can be flexibly used as a scoring function for any pre-trained neural classifier as well as a trainable cost function to shape the energy surface explicitly for OOD detection. On a CIFAR-10 pre-trained WideResNet, using the energy score reduces the average FPR (at TPR 95%) by 18.03% compared to the softmax confidence score. With energy-based training, our method outperforms the state-of-the-art on common benchmarks.

Energy-Based Score

Energy-Based Model

- ❖ 통계물리학에서 에너지란 한 시스템 내에서 입자들이 존재할 확률을 표현하는 물리적인 양
- ❖ 이 개념을 활용해서 (x,y) 사이의 dependency를 에너지를 통해 인코딩하는 Energy-Based Models (EMBs) 제안
- ❖ (x,y)가 올바른 페어라면 낮은 에너지를 가지고, 틀린 페어라면 높은 에너지를 가지도록 하는 energy surface 학습

Boltzmann Distribution

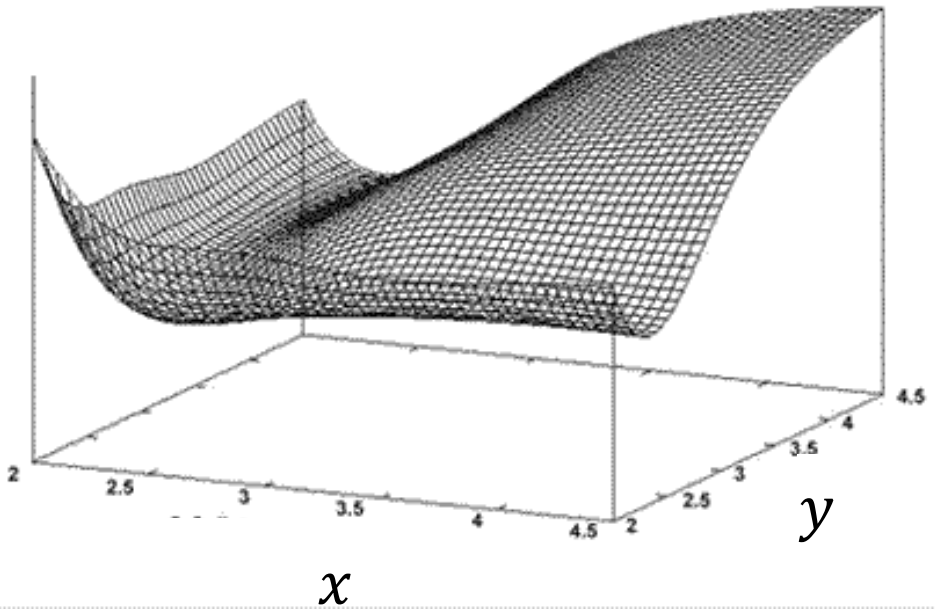
$$p(y|x) = \frac{e^{-E(x,y)/T}}{\int_{y'} e^{-E(x,y')/T}} = \frac{e^{-E(x,y)/T}}{e^{-E(x)/T}}$$

Partition function

Helmholtz free energy

$$E(x) = -T * \log \int_{y'} e^{-E(x,y')/T}$$

Surface of Energy $E(x,y)$

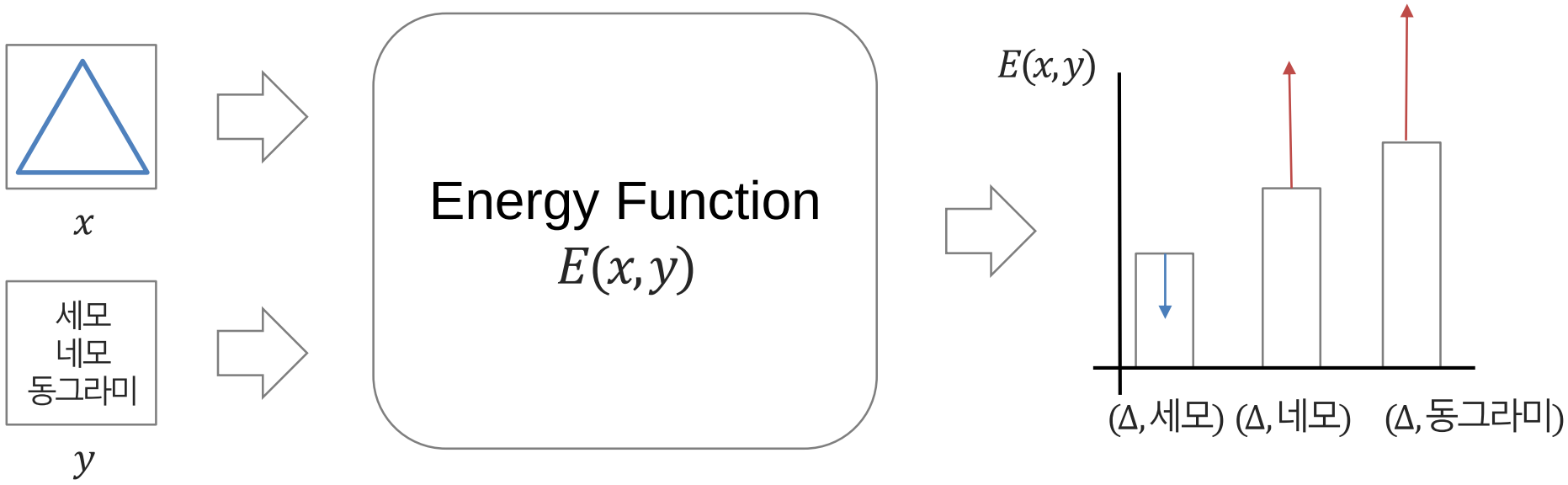


LeCun, Yann, et al. "A tutorial on energy-based learning." Predicting structured data 1.0 (2006).

Energy-Based Score

Energy-Based Model

- 통계물리학에서 에너지란 한 시스템 내에서 입자들이 존재할 확률을 표현하는 물리적인 양
- 이 개념을 활용해서 (x, y) 사이의 dependency를 에너지를 통해 인코딩하는 Energy-Based Models (EMBs) 제안
- (x, y)가 올바른 페어라면 낮은 에너지를 가지고, 틀린 페어라면 높은 에너지를 가지도록 하는 energy surface 학습



LeCun, Yann, et al. "A tutorial on energy-based learning." Predicting structured data 1.0 (2006).

Energy-Based Score

Classification Model as EBMs

❖ Classification model을 학습하는 것은 EBM을 학습하는 것과 같으며 logit을 통해서 energy를 알 수 있음

Classification model

$$p(y|x) = S_y(x) = \frac{e^{f_y(x)/T}}{\sum_{j=1}^{K} e^{f_j(x)/T}}$$

Energy-Based Model

$$p(y|x) = \frac{e^{-E(x,y)/T}}{\int_{y'} e^{-E(x,y')/T}} = \frac{e^{-E(x,y)/T}}{e^{-E(x)/T}}$$

$$E(x,y) = -f_y(x)$$

$$E(x) = -T * \log \int_{y'} e^{-E(x,y')/T}$$

$$= -T * \log \sum_{j=1}^{K} e^{f_j(x)/T}$$

$$g(x;\tau,f) = \begin{cases} 0 \; if \; -E(x;f) \leq \tau, \\ 1 \; if \; -E(x;f) > \tau \end{cases}$$

Energy-Based Score

Classification Model as EBMs

❖ Softmax score를 통해서 negative log likelihood를 최소화하는 과정은 energy surface를 학습하는 과정과 동일

$$\mathrm{L}_{nll} = -\log p(y|x) = \mathbb{E}_{(x,y)\sim P_{in}}\left[-\log \frac{e^{f_y(x)/T}}{\sum_{j=1}^{K} \mathrm{e}^{f_j(x)/T}}\right]$$

다른 페어는 에너지를 높게

$$= \mathbb{E}_{(x,y)\sim P_{in}}\left[\frac{1}{T}E(x,y) + \log \sum_{j=1}^{K} e^{-E(x,j)/T}\right]$$

정답 페어는 에너지를 낮게

Classification model

||

Energy-based model

Energy-Based Score

Softmax Score vs Energy Score

- Softmax score는 Energy score에서 maximum logit value에 의해 bias된 특별한 경우
- 이 bias에 의해서 OOD detection에 유용한 정보를 잃게 됨

$$\max_y p(y|x) = \max_y \frac{e^{f_y(x)}}{\sum_i e^{f_i(x)}} = \frac{e^{f^{max}(x)}}{\sum_i e^{f_i(x)}} = \frac{1}{\sum_i e^{f_i(x) - f^{max}(x)}}$$

$$\Rightarrow \quad \log \max_y p(y|x) = E(x; f(x) - f^{max}(x))$$

$$E(x) = \log \sum_{j=1}^{K} e^{f_j(x)}, where\ T = 1$$

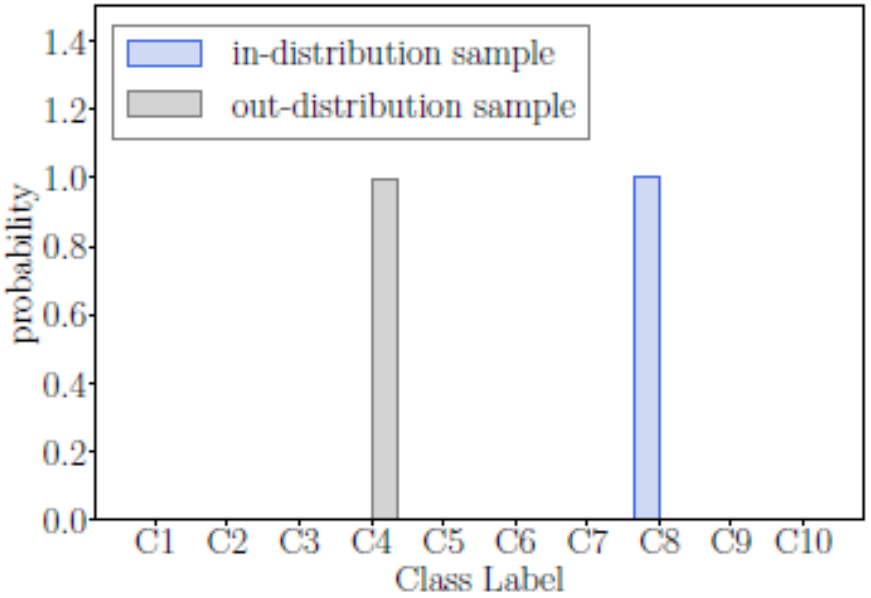
$$= E(x; f) + \overbrace{f^{max}(x)}^{Bias}$$

$\downarrow for\ In-dist\ x$ $\uparrow for\ In-dist\ x$

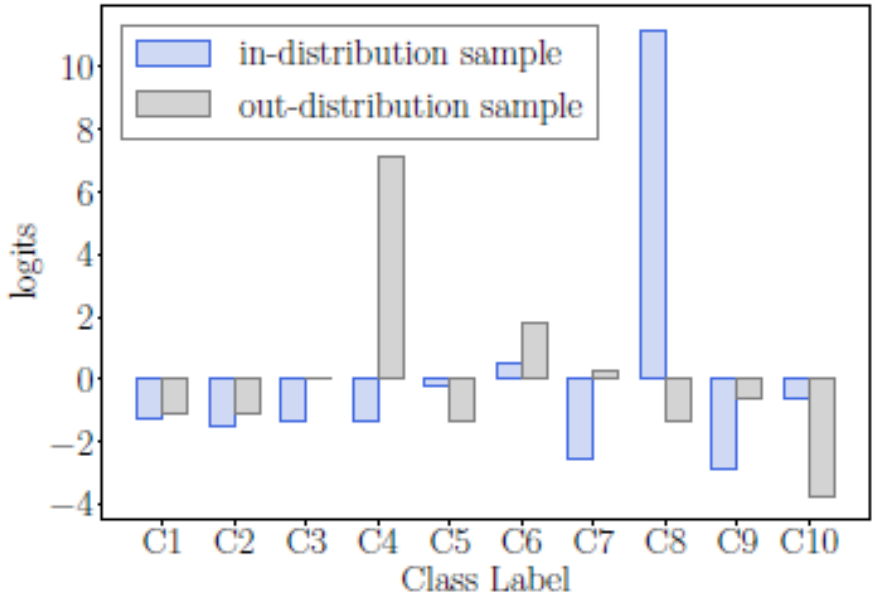
Energy-Based Score

Softmax Score vs Energy Score

- Softmax score는 Energy score에서 maximum logit value에 의해 bias된 특별한 경우
- 이 bias에 의해서 OOD detection에 유용한 정보를 잃게 됨



(a) softmax scores 1.0 vs. 0.99



(b) negative energy scores: 11.19 vs. 7.11

Energy-Based Score

Energy-bounded Learning for OOD Detection

❖ 사용 가능한 OOD 데이터셋을 통해서 OOD 샘플에 대한 에너지를 명시적으로 낮추는 손실 함수 제안

$$\min_{\theta} Loss = L_{nll} + \lambda * L_{energy}$$

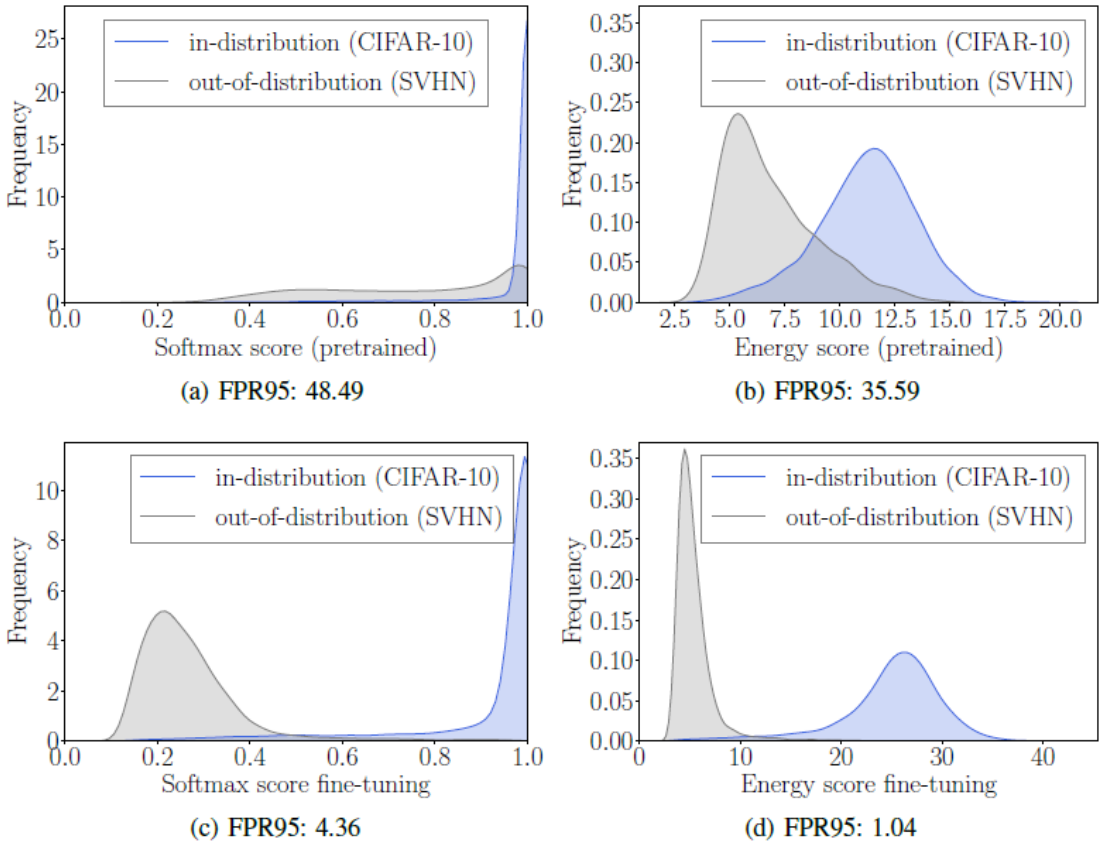
$$where, L_{energy} = \mathbb{E}_{(x_{in},y)\sim D_{in}^{train}}[\max(0, E(x_{in}) - m_{in})]^2$$

$$+\mathbb{E}_{(x_{out})\sim D_{out}^{train}}[\max(0, m_{out} - E(x_{out}))]^2$$

Energy-Based Score

Reuslts

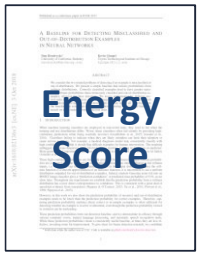
$\mathcal{D}_{\text{in}}^{\text{test}}$	Method	FPR95 ↓	AUROC ↑	AUPR ↑	In-dist Test Error ↓
CIFAR-10 (WideResNet)	Softmax score [13]	51.04	90.90	97.92	5.16
	Energy score (ours)	33.01	91.88	97.83	5.16
	ODIN [24]	35.71	91.09	97.62	5.16
	Mahalanobis [23]	37.08	93.27	98.49	5.16
	OE [14]	8.53	98.30	99.63	5.32
	Energy fine-tuning (ours)	**3.32**	**98.92**	**99.75**	**4.87**
CIFAR-100 (WideResNet)	Softmax score [13]	80.41	75.53	93.93	**24.04**
	Energy score (ours)	73.60	79.56	94.87	24.04
	ODIN [24]	74.64	77.43	94.23	24.04
	Mahalanobis [23]	54.04	84.12	95.88	24.04
	OE [14]	58.10	85.19	96.40	24.30
	Energy fine-tuning (ours)	**47.55**	**88.46**	**97.10**	24.58



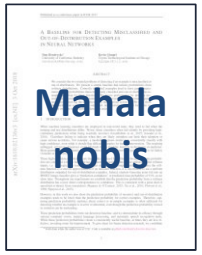
(a) FPR95: 48.49 (b) FPR95: 35.59

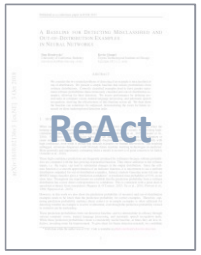
(c) FPR95: 4.36 (d) FPR95: 1.04

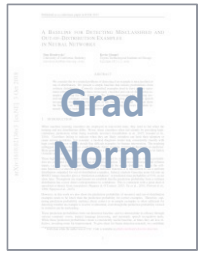
Timeline



2017 2018 2019 2020 2021 2022







Reference

Review Papers

1. Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In International Conference on Learning Representations, 2017.
2. Shiyu Liang, Yixuan Li, and R Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In International Conference on Learning Representations, 2018.
3. Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. Advances in Neural Information Processing Systems, 31, 2018.
4. Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. Advances in Neural Information Processing Systems, 33:21464–21475, 2020.

Related Papers

1. Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531, 2015.
2. I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in Proc. Int. Conf. Learn. Representations, 2015.
3. Y. LeCun, S. Chopra, R. Hadsell, M. Ranzato, and F. J. Huang, "A tutorial on energy-based learning," in Predicting Structured Data, Cambridge, MA, USA: MIT Press, 2006.

Others

1. http://dmqm.korea.ac.kr/activity/seminar/369
2. https://angeloyeo.github.io/2022/09/28/Mahalanobis_distance.html

감사합니다