Open DMQA Seminar

Out of distribution Detection in Image Classification

2022. 07. 01.

Data Mining & Quality Analytics Lab

정재윤

고려대학교 KOREA UNIVERSITY

Data Mining Quality Analytics

발표자 소개

❖ 정재윤(Jaeyoon Jeong)

- 고려대학교 산업경영공학과 전공
- Data Mining & Quality Analytics Lab.(김성범 교수님)
- 석사과정 재학중(2021.09.~ Present)

❖ Research Area

- Machine learning / Deep learning Algorithms
- Anomaly Detection, Out-of-distribution Detection

❖ Contact

- jj950310@korea.ac.kr

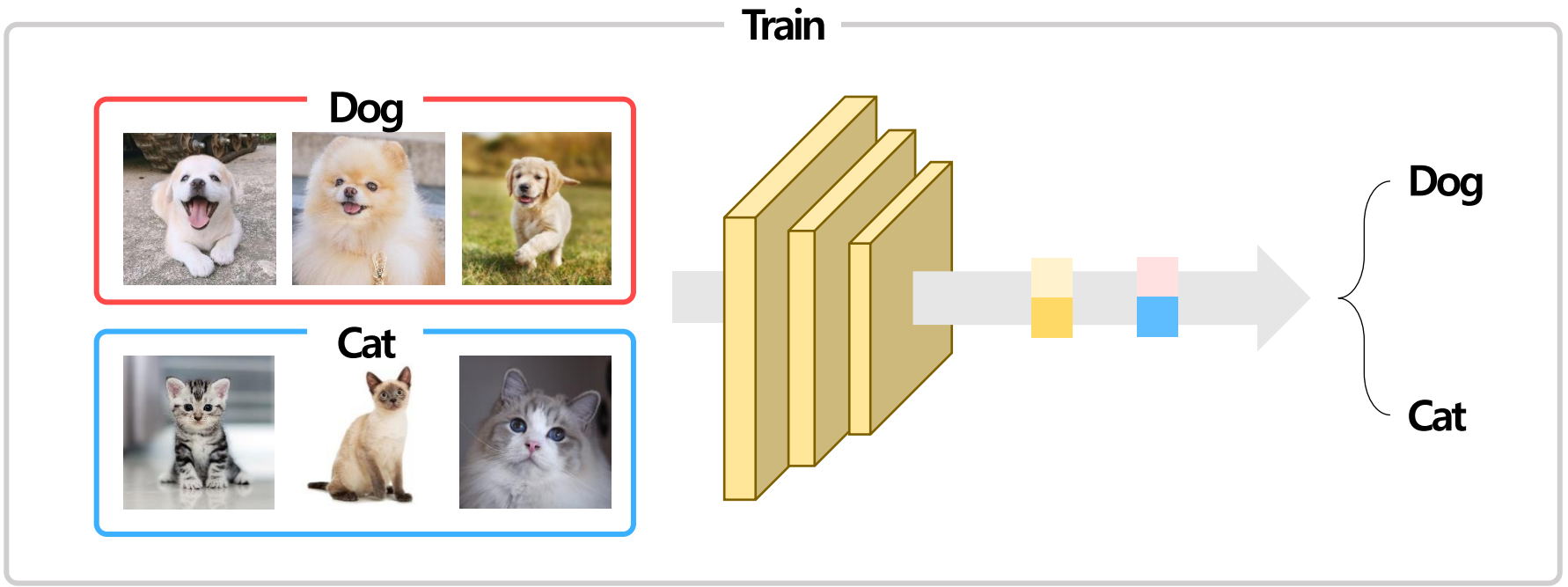
목차

Introduction

Motivation

❖ Image Classification

- 모델은 Image와 label을 Input으로 받아, 설정한 클래스 중 하나로 분류하도록 학습
- 최종적으로 Image만을 넣어, 정답 클래스 예측

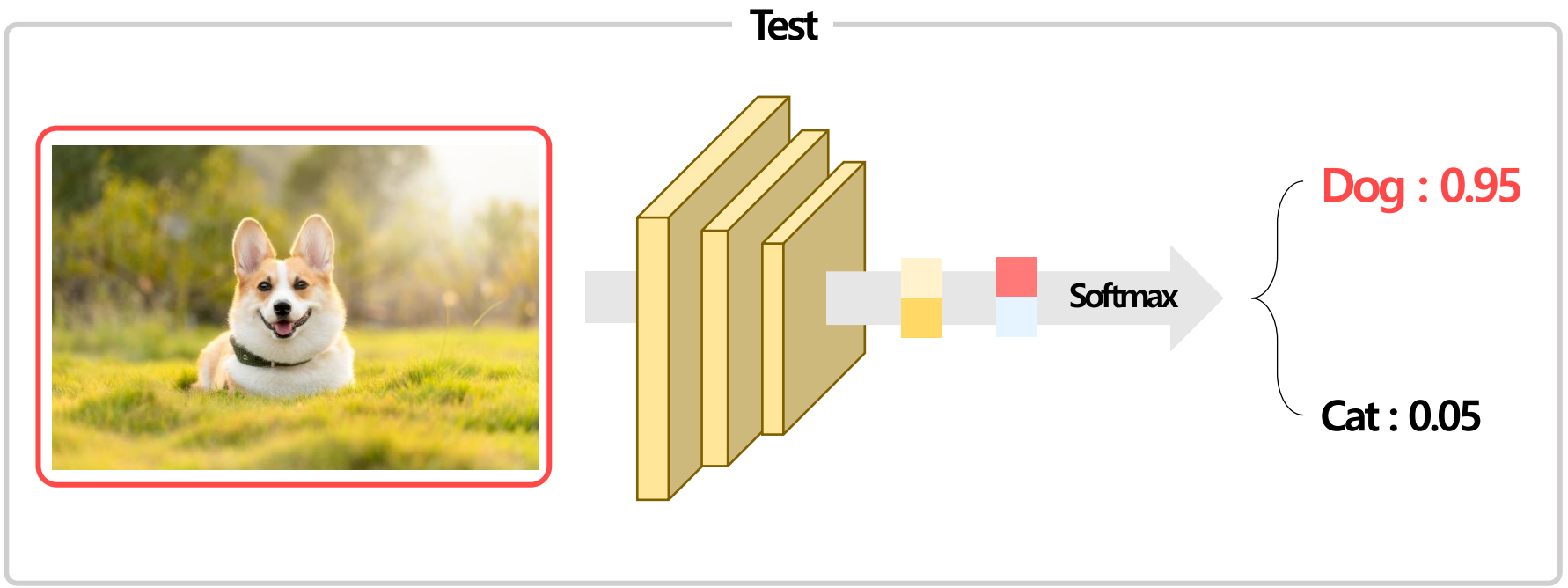


Introduction

Motivation

❖ Image Classification

- 모델은 Image와 label을 Input으로 받아, 설정한 클래스 중 하나로 분류하도록 학습
- 최종적으로 Image만을 넣어, 정답 클래스 예측

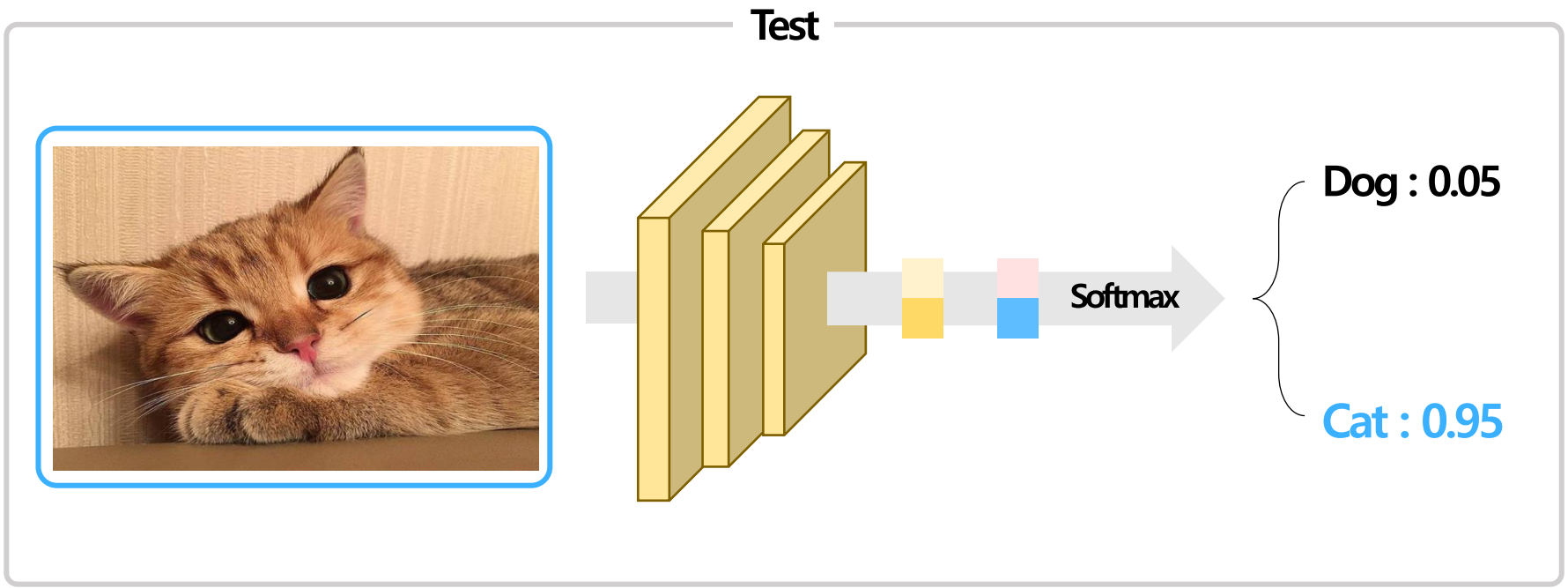


Introduction

Motivation

❖ Image Classification

- 모델은 Image와 label을 Input으로 받아, 설정한 클래스 중 하나로 분류하도록 학습
- 최종적으로 Image만을 넣어, 정답 클래스 예측

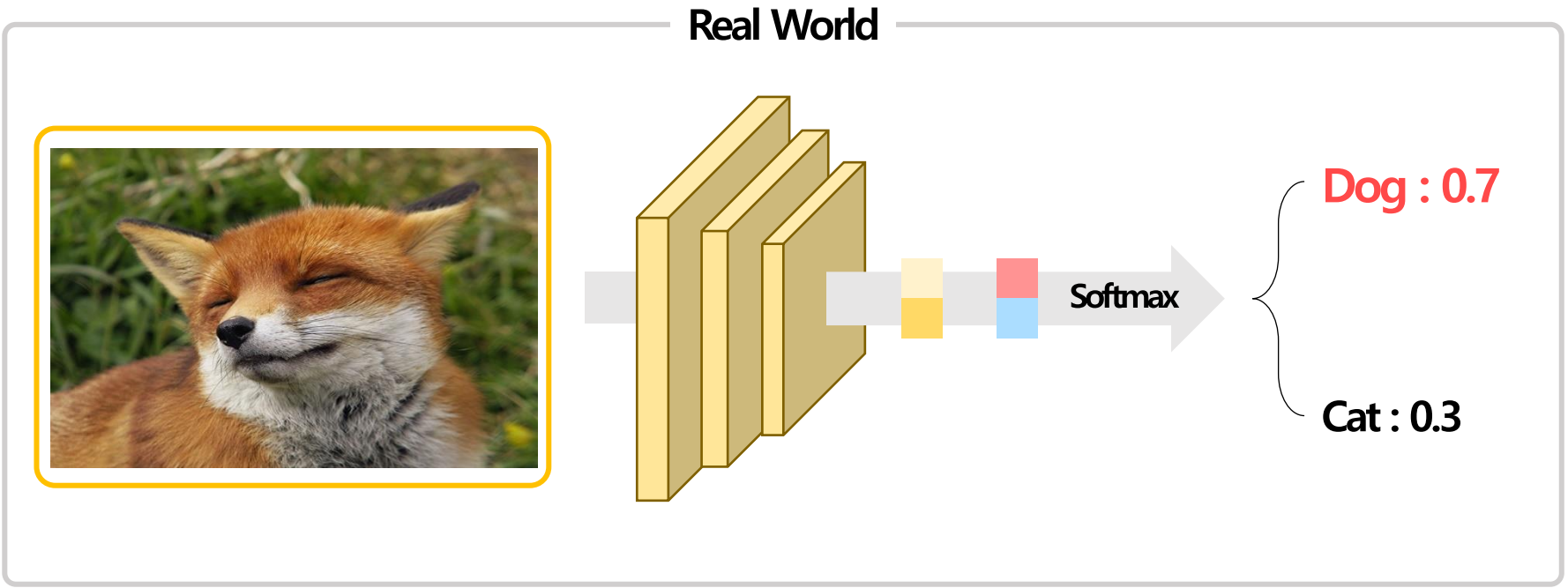


Introduction

Motivation

❖ Limitation

- 학습 완료 후, 모델의 전반적인 성능은 높음
- 그러나 학습되지 않은 class의 데이터를 넣으면, 기존의 class로 잘못 분류



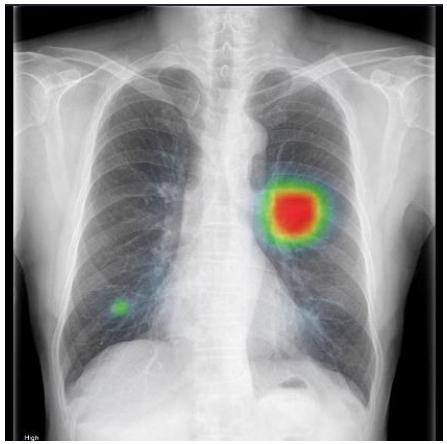
Introduction

Motivation

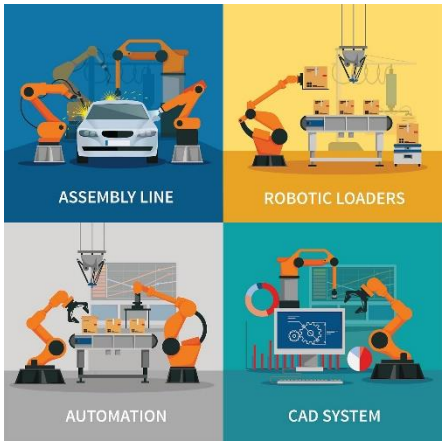
❖ Limitation

- 현실의 특정 도메인에서 이러한 문제가 발생하면 위험한 결과를 도출할 가능성 존재

Real World



의료 진단



스마트 팩토리

자율 주행

Introduction

Motivation

❖ Limitation

- 현실의 특정 도메인에서 이러한 문제가 발생하면 위험한 결과를 도출할 가능성 존재

Real World

자율 주행

스마트 팩토리

의료 진단

모델이 학습한 데이터셋은 정확하게 분류하되,

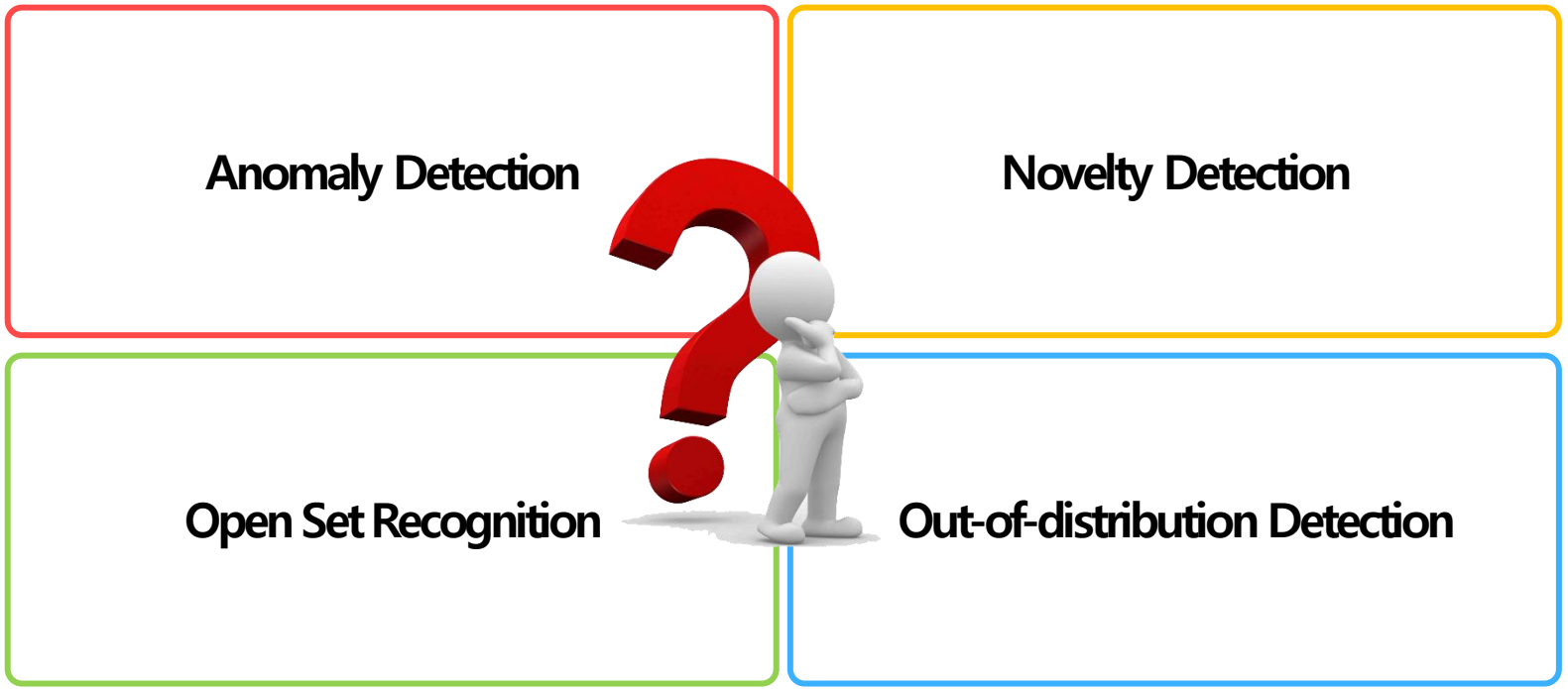
학습하지 않은 데이터셋은 걸러내는 것이 목표

Introduction

What is out-of-distribution Detection?

❖ **Limitation**

- 해당 목표를 가지고 발전한 여러 연구분야 존재
- 개념들이 유사하여 혼동되기 쉬움



Introduction

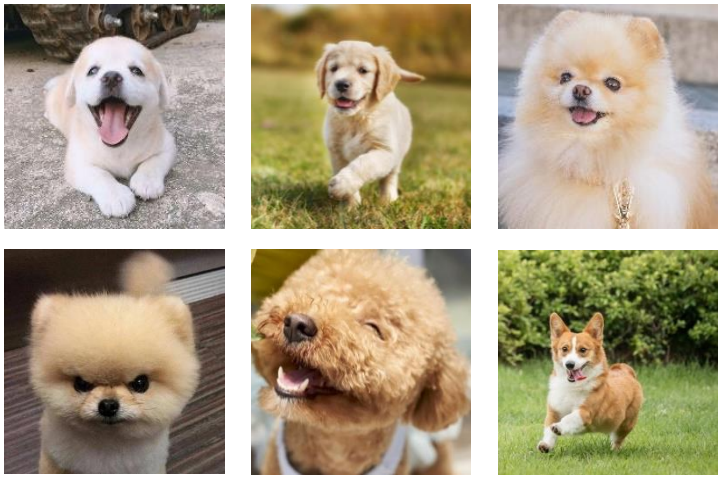
What is out-of-distribution Detection?

❖ AD vs ND

- 보통 모델이 클래스(Normal)가 하나인 데이터셋에 대하여 학습
- 정상과 정상이 아닌 것을 분류하는 Binary Classification

Space

Normal Dataset

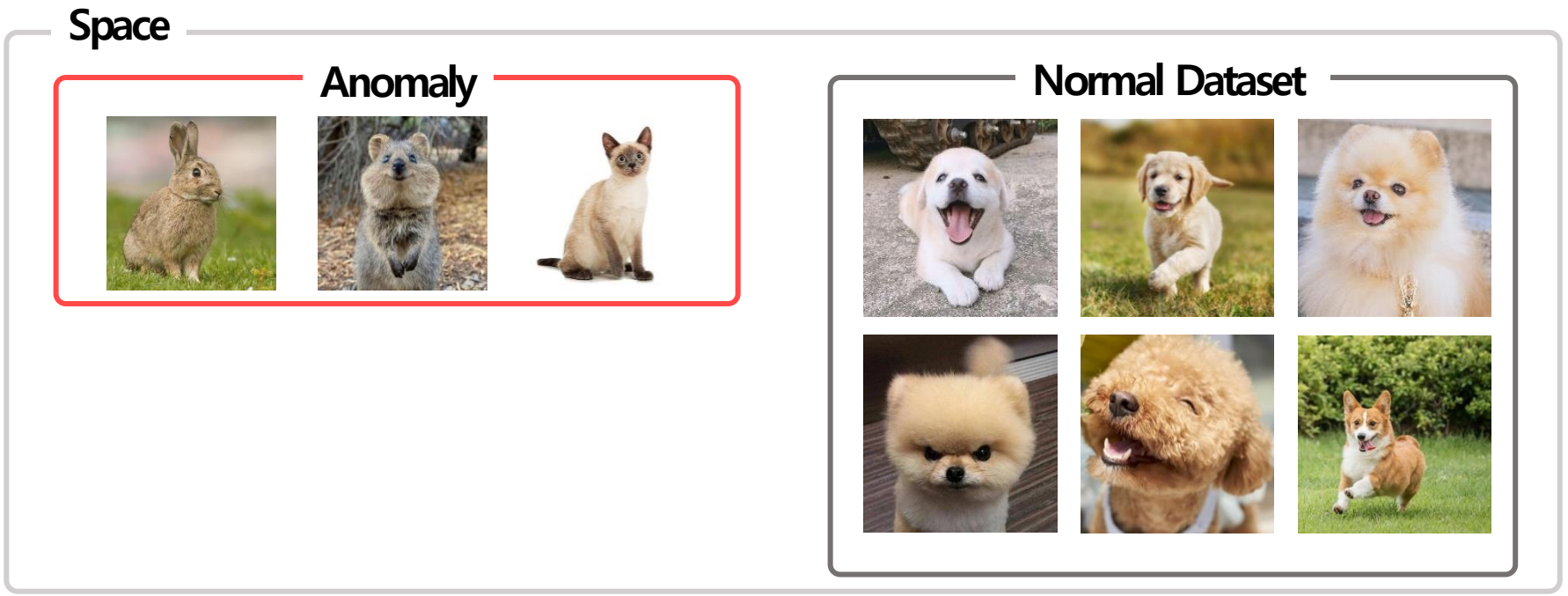


Introduction

What is out-of-distribution Detection?

❖ AD vs ND

- AD : 학습한 데이터와 본질적으로 특징이 다른 Anomaly를 찾는 Task
- ND : 학습한 데이터와 본질적인 특징은 같으나, 새로운 형태의 Novelty를 찾는 Task

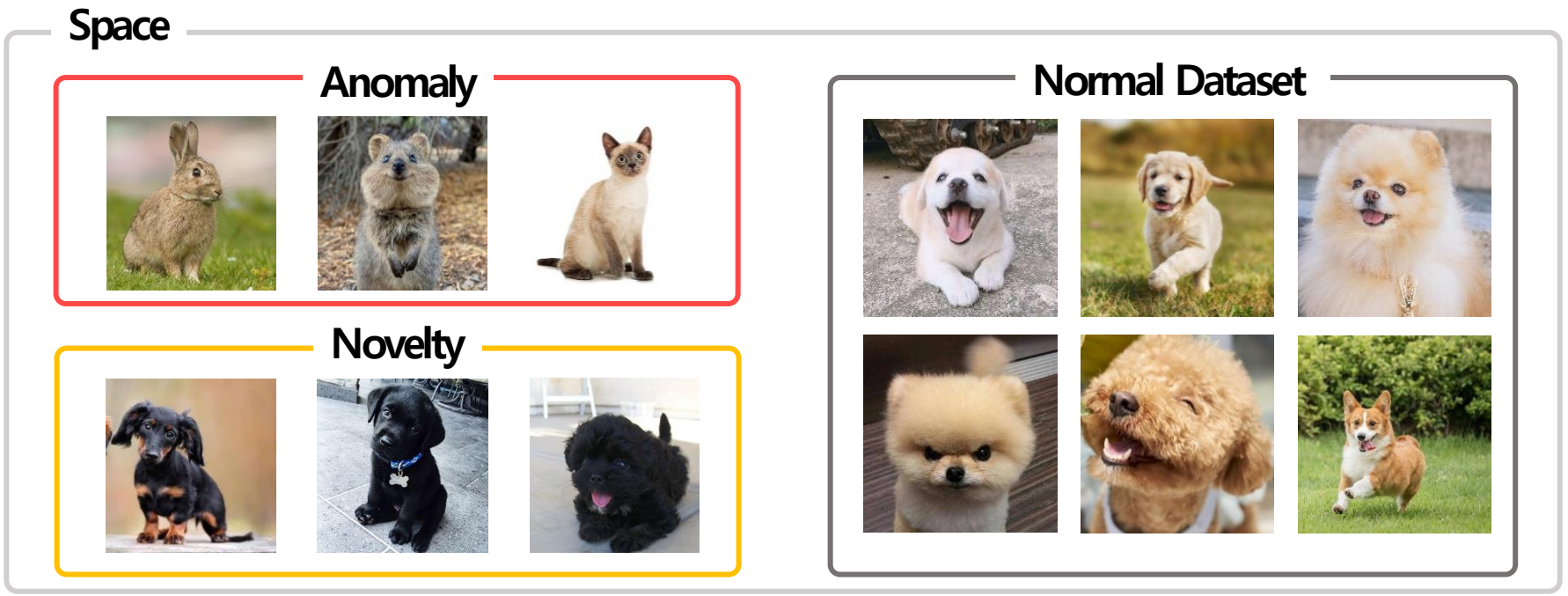


Introduction

What is out-of-distribution Detection?

❖ AD vs ND

- AD : 학습한 데이터와 본질적으로 특징이 다른 Anomaly를 찾는 Task
- ND : 학습한 데이터와 본질적인 특징은 같으나, 새로운 형태의 Novelty를 찾는 Task



Introduction

What is out-of-distribution Detection?

❖ OSR vs OOD Detection

- 모델이 클래스가 여러 개인 데이터셋에 대하여 학습
- Multi-class classification과 학습하지 않은 것을 찾는 Task 동시 진행

Space

Normal Dataset

Introduction

What is out-of-distribution Detection?

❖ OSR vs OOD Detection

- OSR : Multi-class classification & 학습하지 않은 클래스를 찾음
- OOD Detection : Multi-class classification & 전혀 다른 범주의 학습하지 않은 데이터셋을 찾음

Space

Open-set

Normal Dataset

Introduction

What is out-of-distribution Detection?

❖ OSR vs OOD Detection

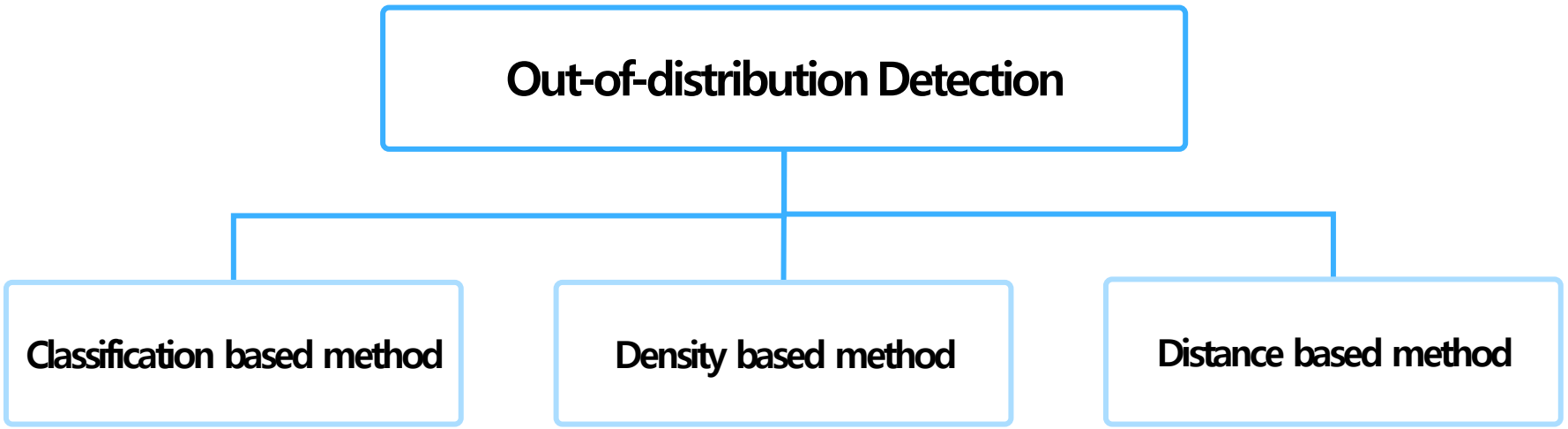
- OSR : Multi-class classification & 학습하지 않은 클래스를 찾음
- OOD Detection : Multi-class classification & 전혀 다른 범주의 학습하지 않은 데이터셋을 찾음

Introduction

What is out-of-distribution Detection?

❖ OOD Detection

- OOD Detection은 크게 Classification based method, Density based method, Distance based method로 구분

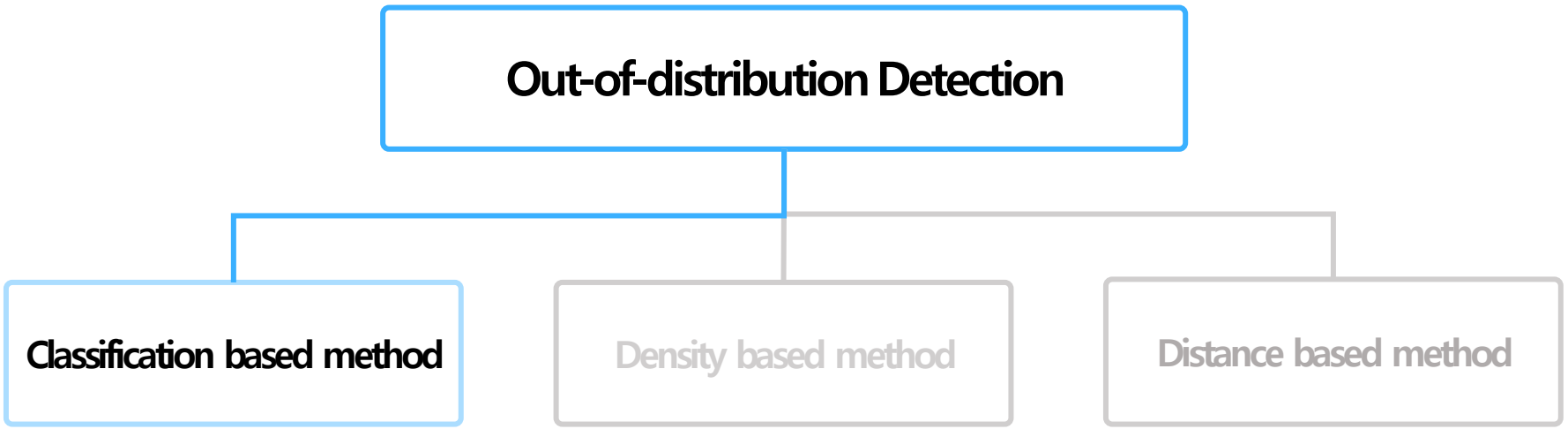


Introduction

What is out-of-distribution Detection?

❖ OOD Detection

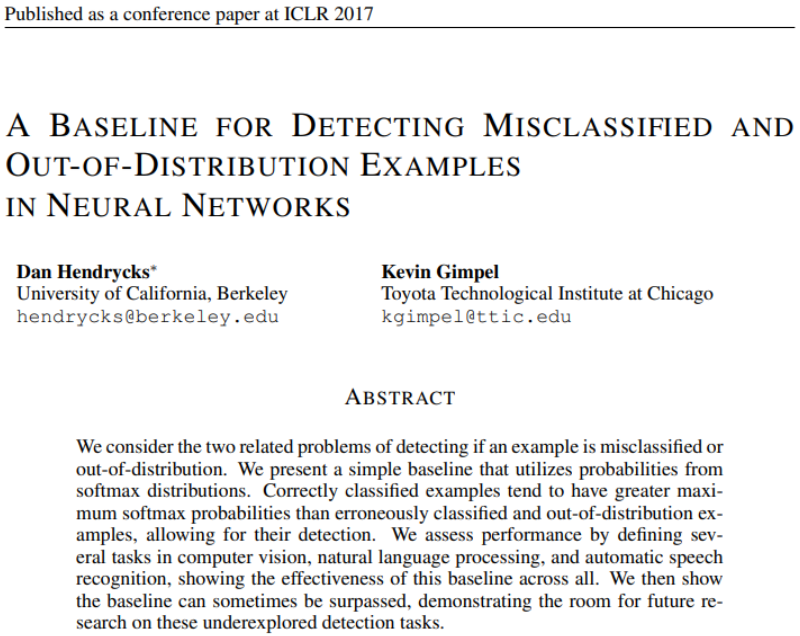
- OOD Detection은 크게 Classification based method, Density based method, Distance based method로 구분



Baseline method

❖ **A Baseline for Detecting Misclassified and Out-of-distribution Examples in Neural Networks**

- 2017년 ICLR에 발표된 논문으로, 2022년 6월 28일 기준으로 총 1431회 인용
- OOD Detection 문제 정의와 실험 세팅 및 평가 방법을 제시



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

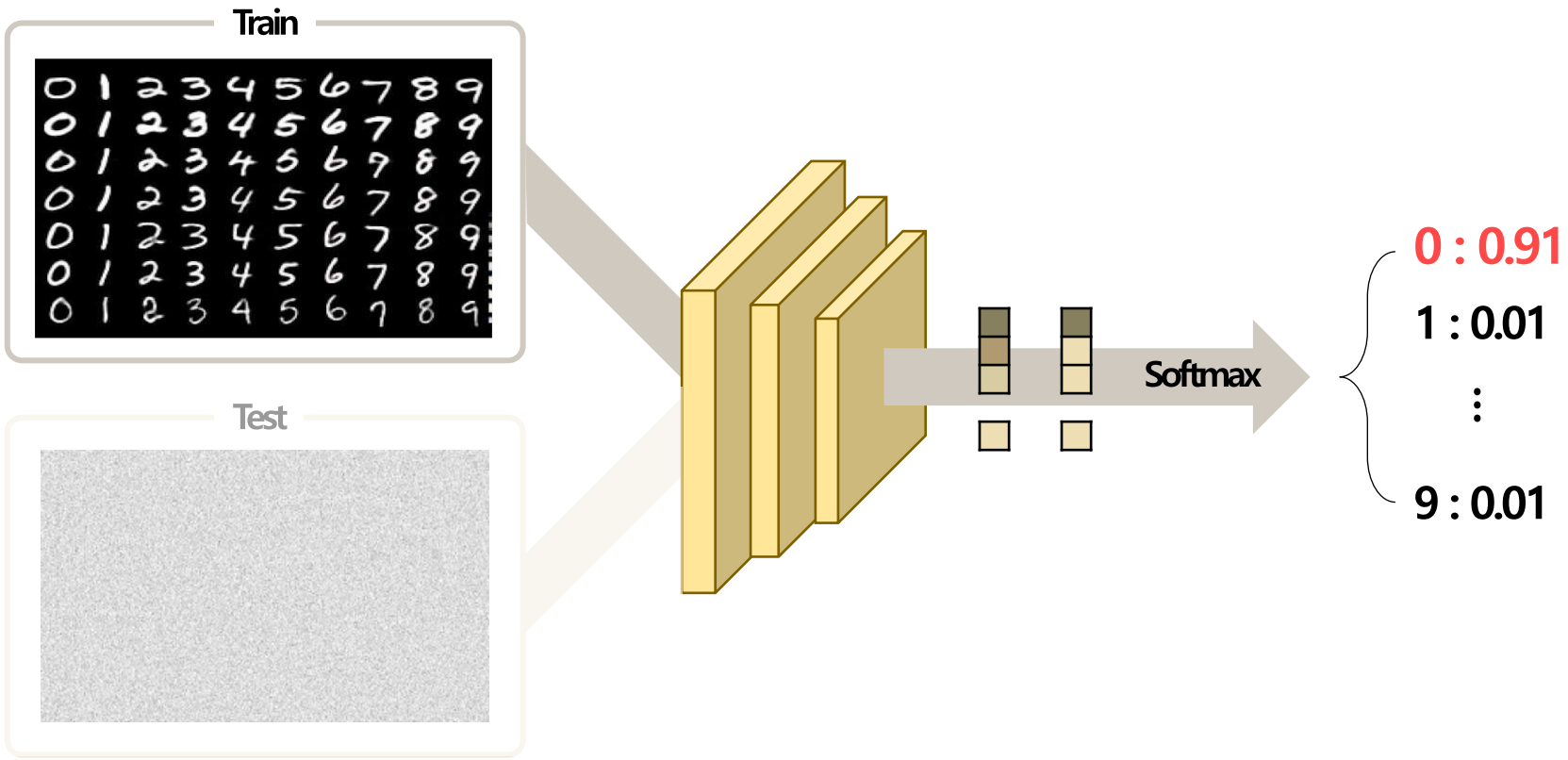
ABSTRACT

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 method

❖ **Misclassification problem**

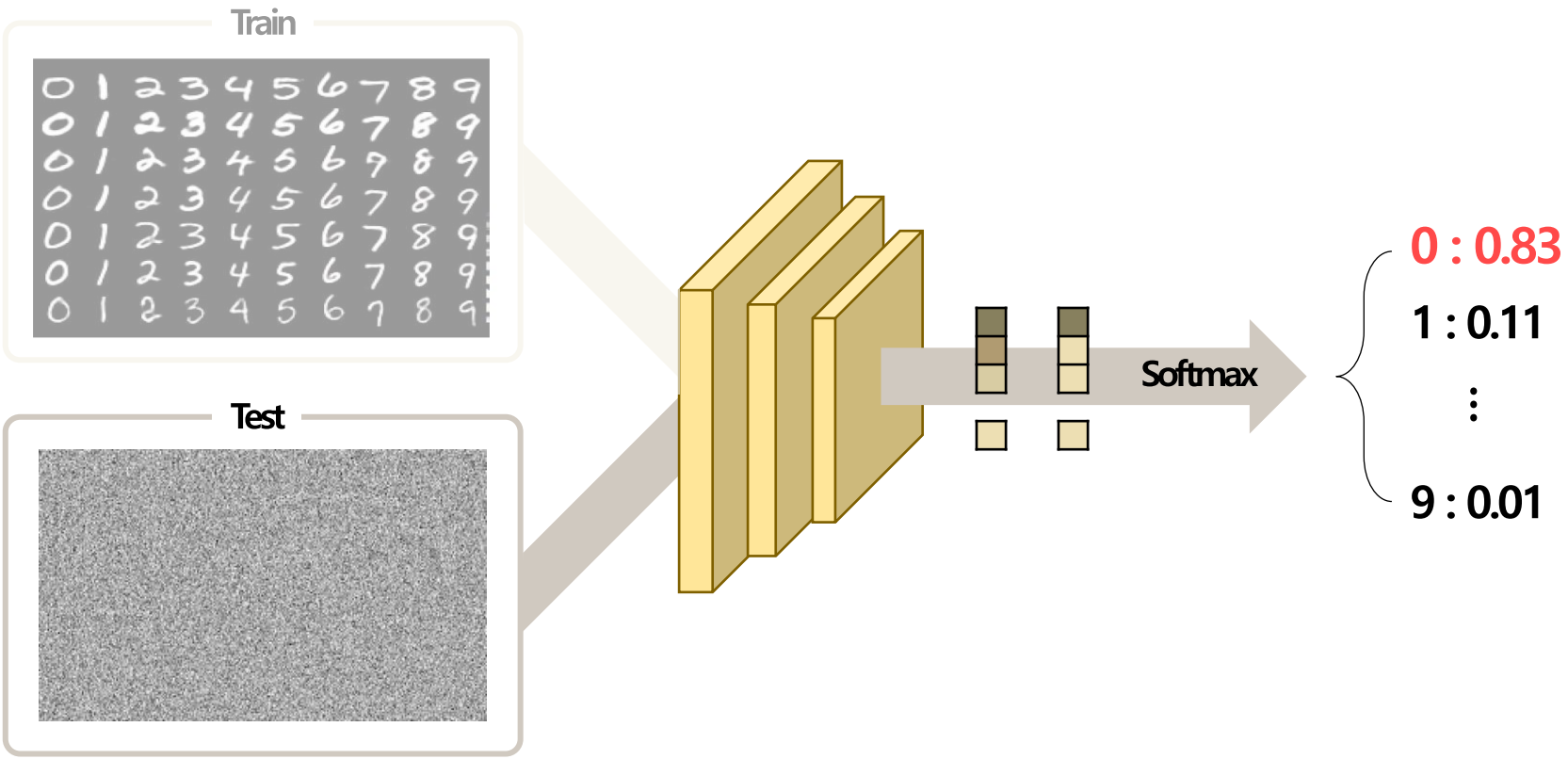
- MNIST로 학습시킨 분류기에 Gaussian noise를 통과시키면 높은 확률로 오분류
- 이러한 문제는 모델의 마지막 단이 Softmax에 의해서 발생함을 제시



Baseline method

❖ Misclassification problem

- MNIST로 학습시킨 분류기에 Gaussian noise를 통과시키면 높은 확률로 오분류
- 이러한 문제는 모델의 마지막 단이 Softmax에 의해서 발생함을 제시

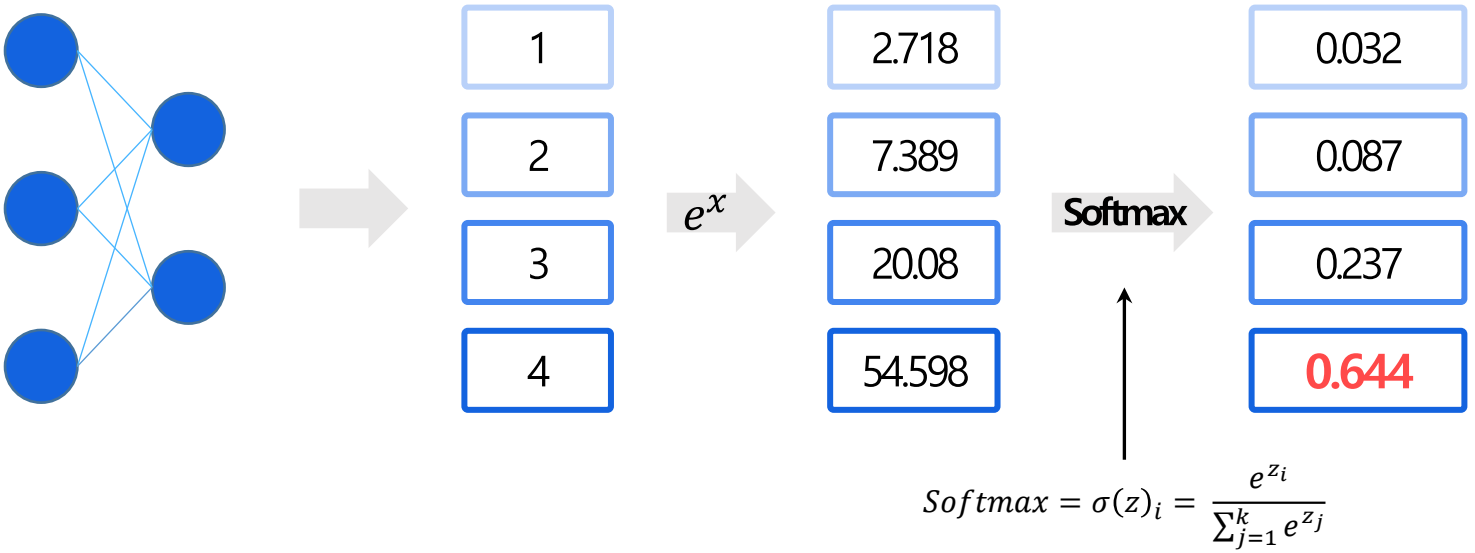


Baseline method

❖ Softmax function

- Multi-class classification에 출력층에서 사용되는 활성화 함수
- 함수에서 사용하는 e^x로 인해, 작은 값의 차이도 확연히 구분

  → 이 때문에 모델은 분류 시, Misclassification problem을 야기

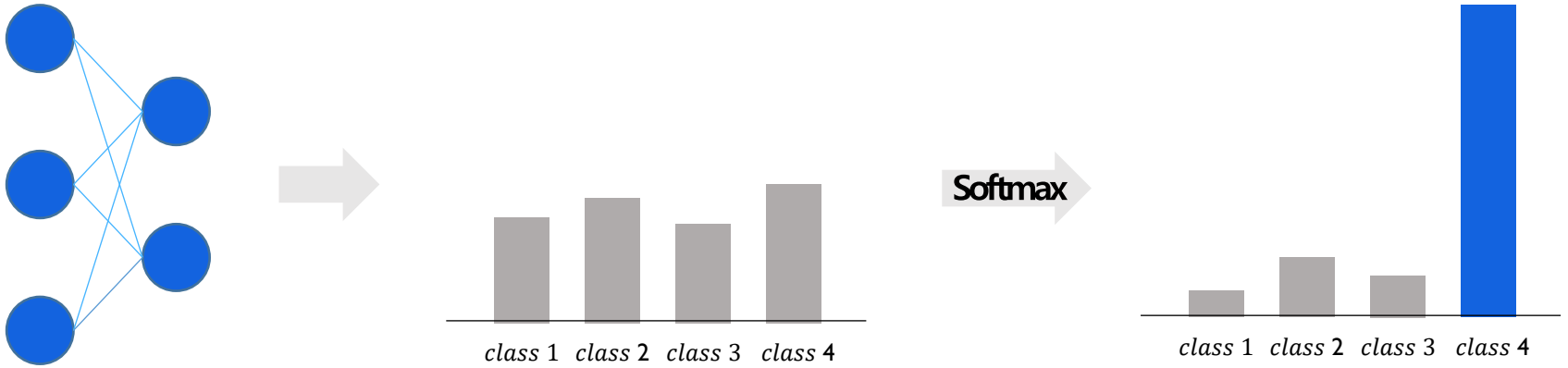


Baseline method

❖ Softmax function

- Multi-class classification에 출력층에서 사용되는 활성화 함수
- 함수에서 사용하는 e^x로 인해, 작은 값의 차이도 확연히 구분

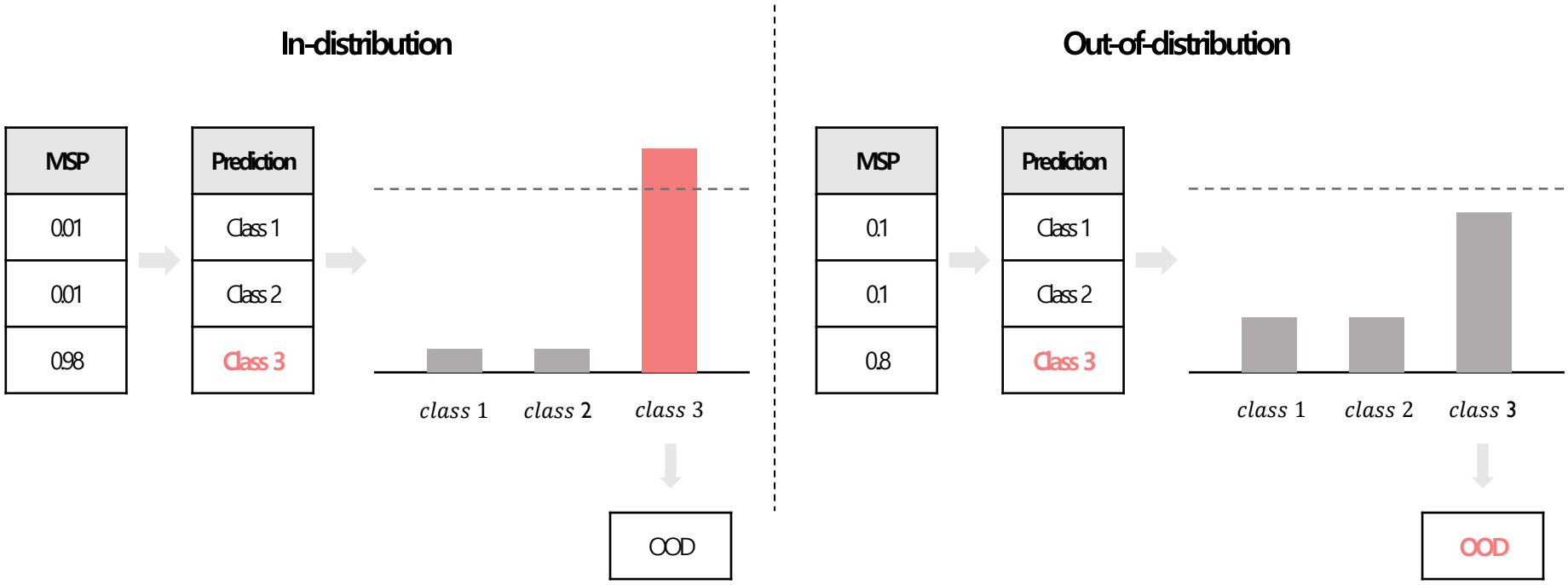
  → 이 때문에 모델은 분류 시, Misclassification problem을 야기



Baseline method

❖ Maximum Softmax Probability

- Test 시, Out-of-distribution이 들어오면 Softmax Probability가 비교적 낮음을 확인
- Softmax Probability 중 가장 큰 값과 임의의 Threshold를 비교
  - ✓ 확률값이 Threshold보다 높으면 In-distribution, 낮으면 Out-of-distribution



Baseline method

❖ Evaluation Metrics

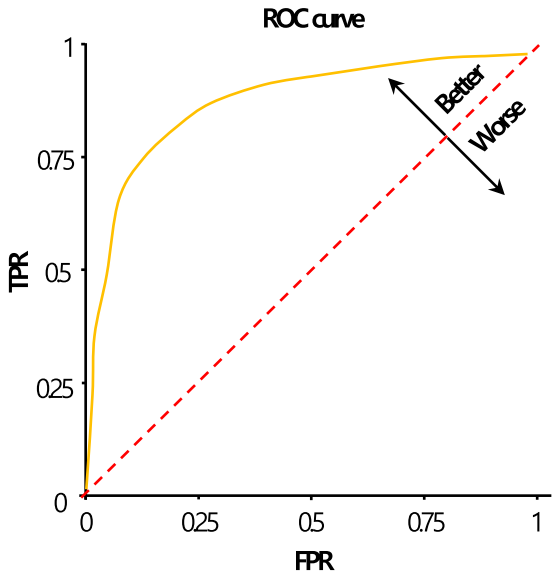
- AUROC : FPR과 TPR을 축으로 그린 ROC 곡선 아래의 넓이
- AUPR : Recall과 Precision을 축으로 그린 곡선 아래의 넓이

Confusion Matrix

	Actually Positive (1)	Actually Negative (0)
Predicted Positive (1)	True Positives	False Positives
Predicted Negative (0)	False Negatives	True Negatives

$$TPR = \frac{TP}{all\ positive} = \frac{TP}{TP + FN}$$

$$FPR = \frac{FP}{all\ negative} = \frac{FP}{FP + TN}$$



Baseline method

❖ Evaluation Metrics

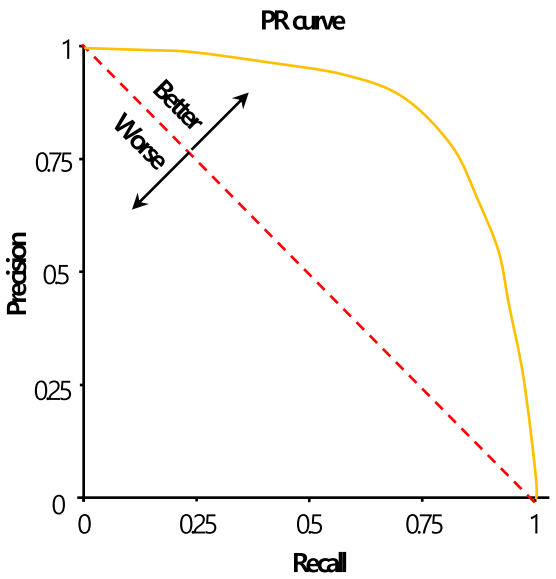
- AUROC : FPR과 TPR을 축으로 그린 ROC 곡선 아래의 넓이
- AUPR : Recall과 Precision을 축으로 그린 곡선 아래의 넓이

Confusion Matrix

	Actually Positive (1)	Actually Negative (0)
Predicted Positive (1)	True Positives	False Positives
Predicted Negative (0)	False Negatives	True Negatives

$$TPR = \frac{TP}{all\ positive} = \frac{TP}{TP + FN} = Recall$$

$$Precision = \frac{TP}{TP + FP}$$



Advanced methods

OOD data generation method

❖ Training Confidence-calibrated Classifiers for Detecting Out-of-distribution Samples

- 2018년 ICLR에 발표된 논문으로, 2022년 6월 28일 기준으로 총 533회 인용
- GAN과 Classifier를 활용한 학습방법 제안

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

ABSTRACT

The problem of detecting whether a test sample is from in-distribution (i.e., training distribution by a classifier) or out-of-distribution sufficiently different from it arises in many real-world machine learning applications. However, the state-of-art deep neural networks are known to be highly overconfident in their predictions, i.e., do not distinguish in- and out-of-distributions. Recently, to handle this issue, several threshold-based detectors have been proposed given pre-trained neural classifiers. However, the performance of prior works highly depends on how to train the classifiers since they only focus on improving inference procedures. In this paper, we develop a novel training method for classifiers so that such inference algorithms can work better. In particular, we suggest two additional terms added to the original loss (e.g., cross entropy). The first one forces samples from out-of-distribution less confident by the classifier and the second one is for (implicitly) generating most effective training samples for the first one. In essence, our method jointly trains both classification and generative neural networks for out-of-distribution. We demonstrate its effectiveness using deep convolutional neural networks on various popular image datasets.

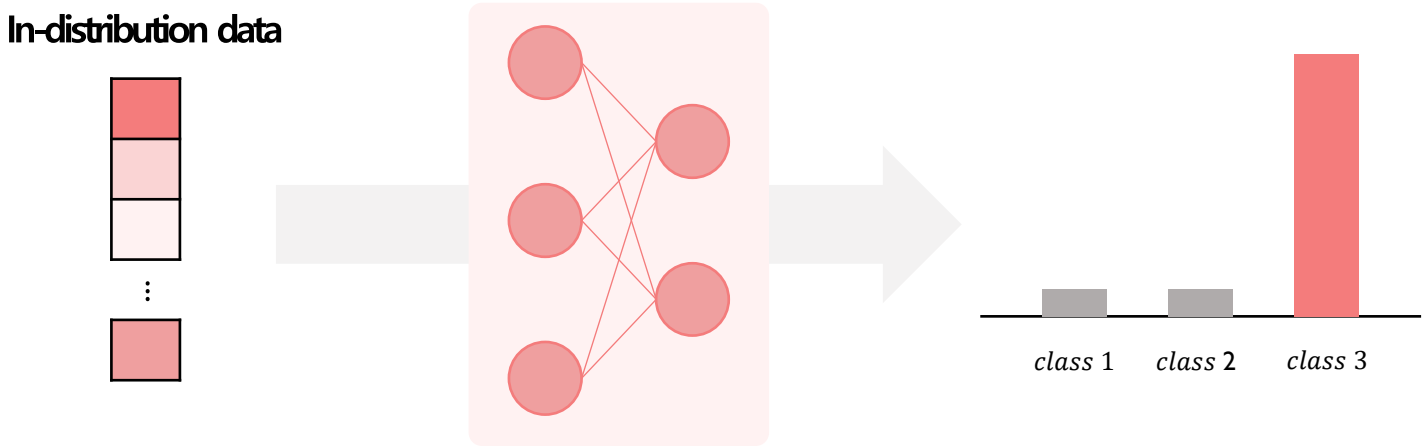
Advanced methods

OOD data generation method

❖ Confidence loss

- In-distribution data는 Classification task에 주로 사용되는 Cross Entropy loss
- Out-of-distribution data는 모델의 예측 확률이 Uniform distribution과 같아지도록 KL-divergence를 사용

$$\min_{\theta} \underbrace{E_{P_{in}(\hat{x},\hat{y})}[-\log P_{\theta}\ (y = \hat{y} \mid \hat{x})]}_{\text{(a) term}} + \beta E_{P_{out}(x)}[KL\ (U(y)\ ||\ P_{\theta}(y|x)]$$



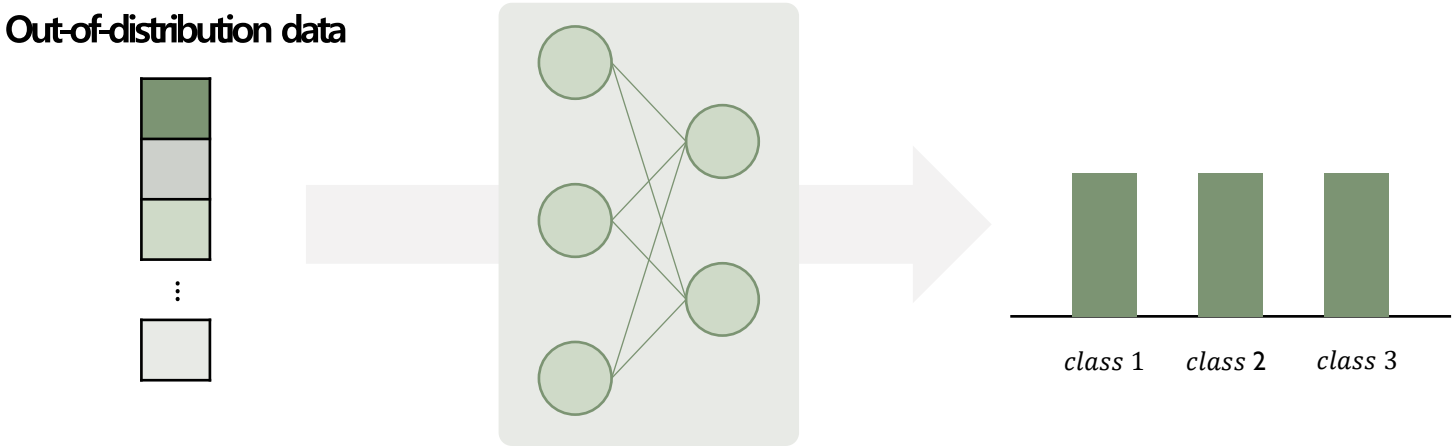
Advanced methods

OOD data generation method

❖ Confidence loss

- In-distribution data는 Classification task에 주로 사용되는 Cross Entropy loss
- Out-of-distribution data는 모델의 예측 확률이 Uniform distribution과 같아지도록 KL-divergence를 사용

$$\min_{\theta} E_{P_{in}(\hat{x},\hat{y})}[-\log P_{\theta}\ (y = \hat{y} \mid \hat{x})] + \beta \underbrace{E_{P_{out}(x)}[KL\ (U(y)\ ||\ P_{\theta}(y|x)]}_{\text{(b) term}}$$

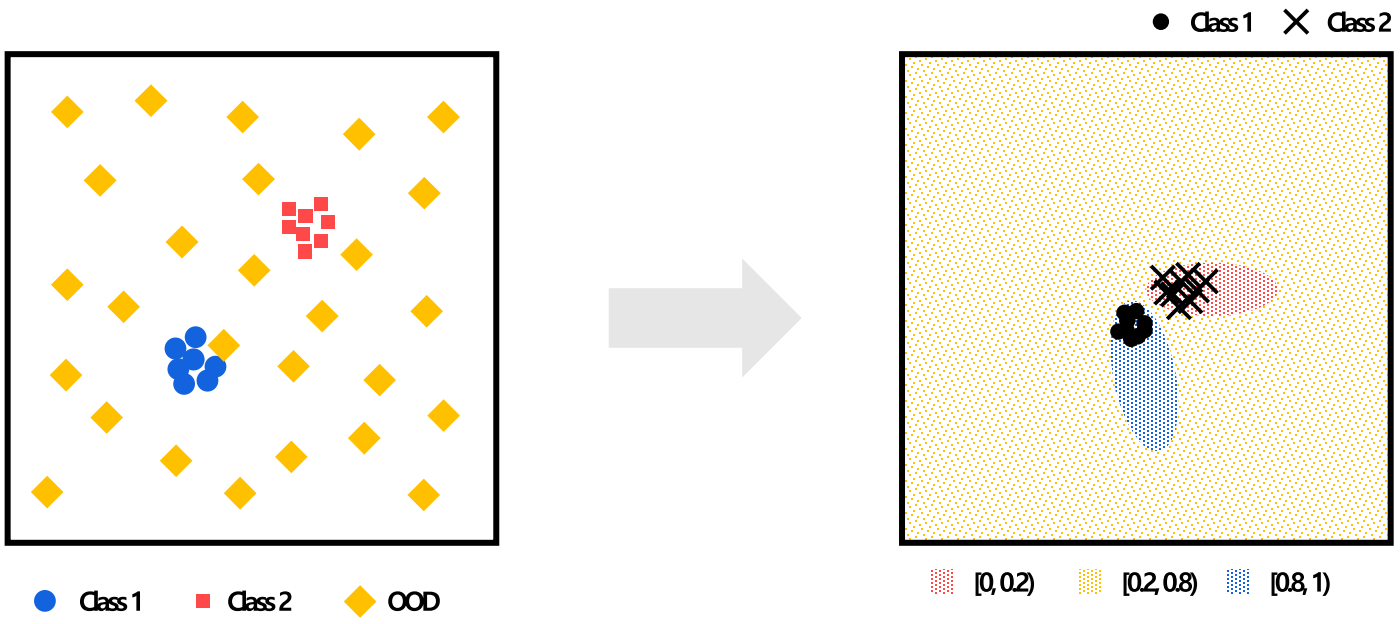


Advanced methods

OOD data generation method

❖ Confidence loss

- 해당 loss를 사용한 Classifier를 사용하여 간단한 실험 진행
  - ✓ 전체 Space에서 OOD를 추출해 학습 진행 시, 비교적 넓은 Boundary 형성
  - ✓ Class 주변의 OOD를 추출해 학습 진행 시, 이전보다 더 좁은 Boundary 형성

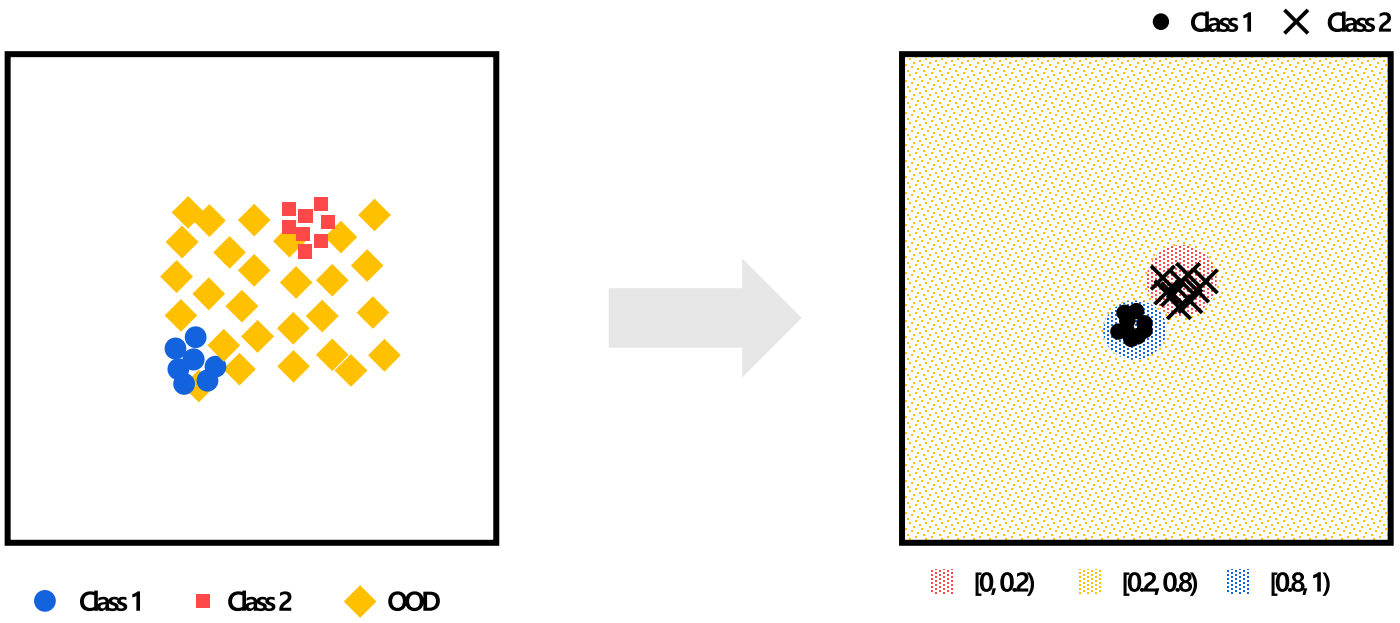


Advanced methods

OOD data generation method

❖ Confidence loss

- 해당 loss를 사용한 Classifier를 사용하여 간단한 실험 진행
  - ✓ 전체 Space에서 OOD를 추출해 학습 진행 시, 비교적 넓은 Boundary 형성
  - ✓ Class 주변의 OOD를 추출해 학습 진행 시, 이전보다 더 좁은 Boundary 형성



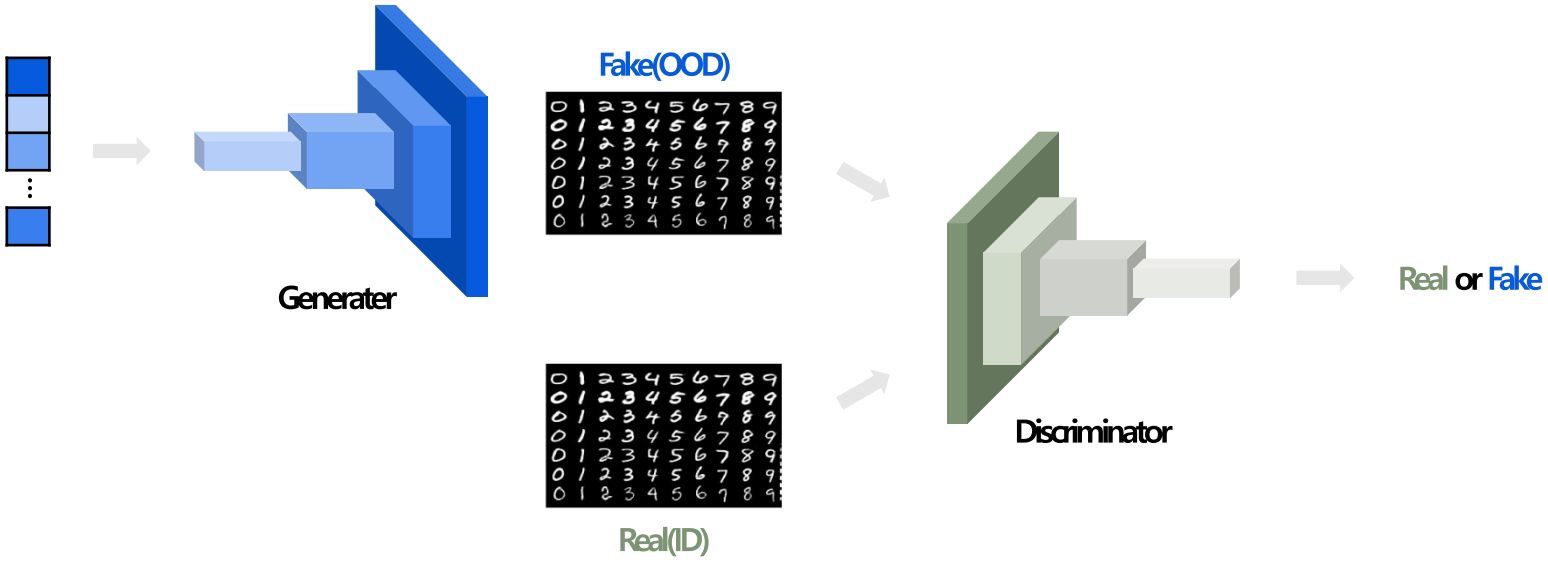
Advanced methods

OOD data generation method

❖ GAN loss

- 전체 Space에 대해 많은 Out-of-distribution sample을 수집하기는 어려움
- In-distribution 주변의 Out-of-distribution sample을 수집하는 것이 성능 향상에 효율적
  - ✓ GAN을 통해 In-distribution 주변의 Out-of-distribution sample 생성

$$\min_G \max_D \beta E_{P_{G(x)}}[KL(U(y) \,||\, P_\theta(y|x)] + \underbrace{E_{P_{in}(x)}[\log D(x)] + E_{P_{G(x)}}[\log(1 - D(x))]}_{\text{(b) term}}$$



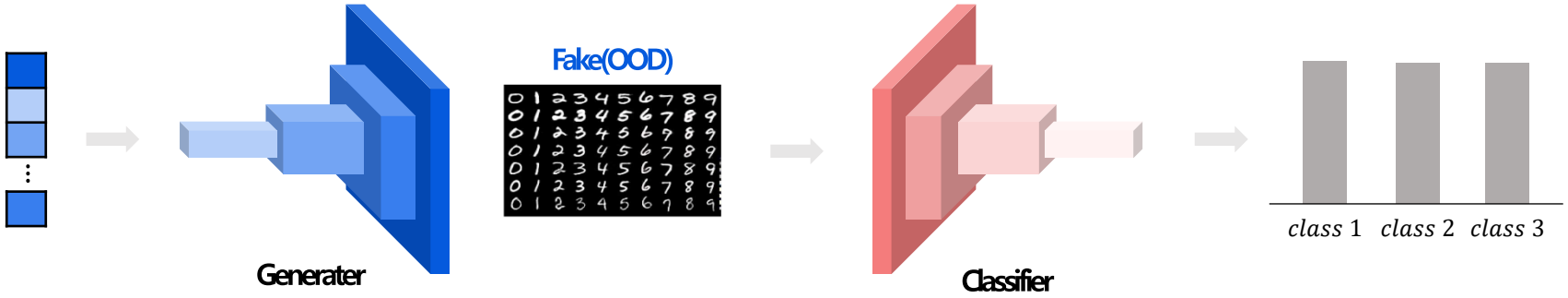
Advanced methods

OOD data generation method

❖ GAN loss

- 전체 Space에 대해 많은 Out-of-distribution sample을 수집하기는 어려움
- In-distribution 주변의 Out-of-distribution sample을 수집하는 것이 성능 향상에 효율적
  - ✓ GAN을 통해 In-distribution 주변의 Out-of-distribution sample 생성

$$\min_G \max_D \beta \underbrace{E_{P_{G(x)}}[KL(U(y) \,||\, P_\theta(y|x)]}_{\text{(a) term}} + E_{P_{in}(x)}[\log D(x)] + E_{P_{G(x)}}[\log(1 - D(x))]$$



Advanced methods

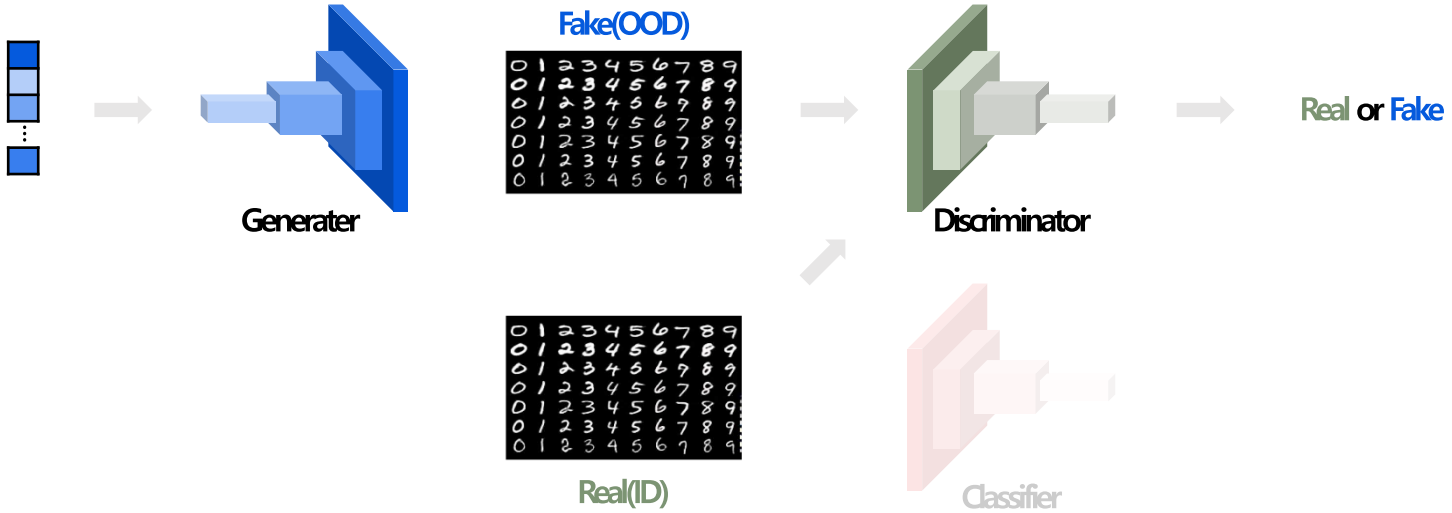
OOD data generation method

❖ Joint Training Method

- Confidence loss와 GAN loss를 합하여 새로운 Training 방법 제시
- 상호 보완적 학습으로 서로의 성능을 향상 시키게 됨

$$\min_G \max_D \min_\theta \; E_{P_{in}(\hat{x},\hat{y})}[-\log P_\theta(y=\hat{y}|\hat{x})] + \beta E_{P_{G(x)}}[KL\big(U(y)\big)||P_\theta(y|x)]$$

$$+E_{P_{in}(\hat{x})}[\log D(\hat{x})] + E_{P_{G(x)}}[\log(1-D(x))]$$



Advanced methods

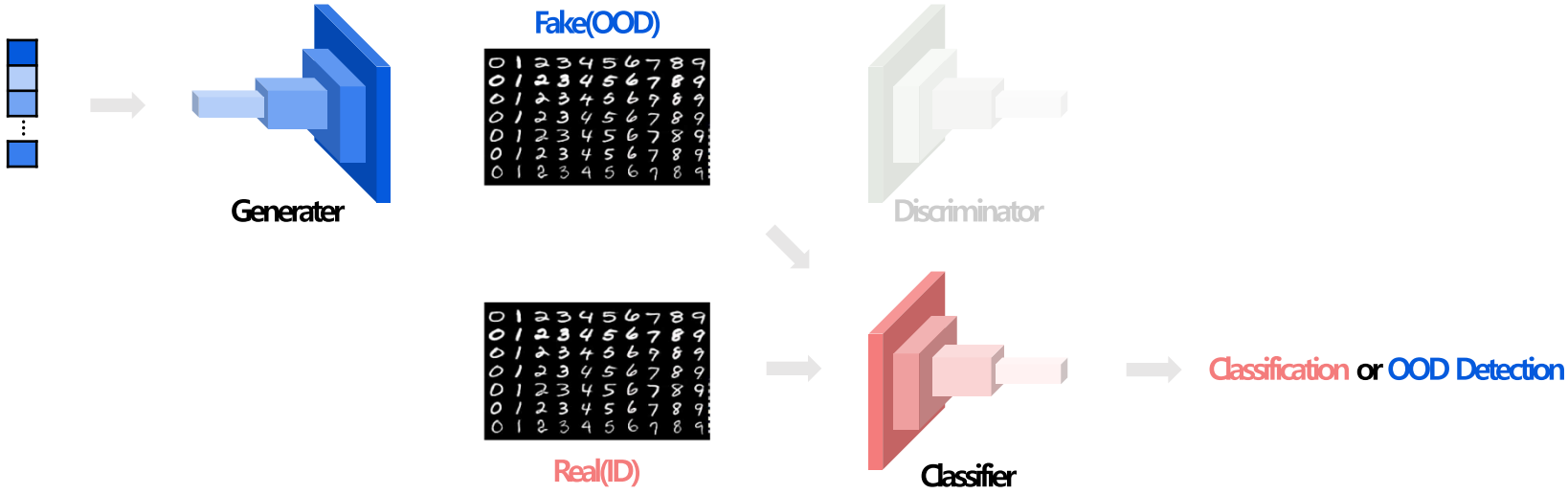
OOD data generation method

❖ Joint Training Method

- Confidence loss와 GAN loss를 합하여 새로운 Training 방법 제시
- 상호 보완적 학습으로 서로의 성능을 향상 시키게 됨

$$\min_G \max_D \min_\theta \; E_{P_{in}(\hat{x},\hat{y})}[-\log P_\theta(y=\hat{y}|\hat{x})] + \beta E_{P_{G(x)}}[KL\big(U(y)\big)||P_\theta(y|x)]$$

$$+E_{P_{in}(\hat{x})}[\log D(\hat{x})] + E_{P_{G(x)}}[\log(1-D(x))]$$

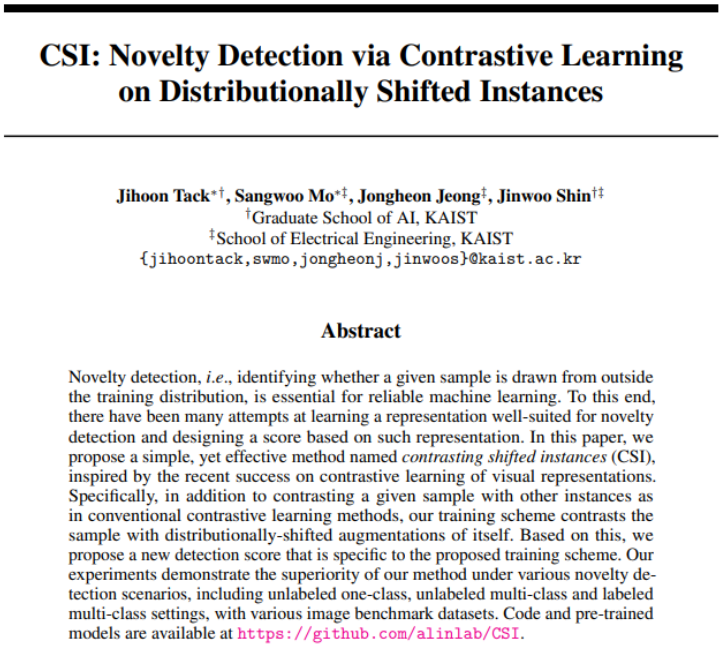


Advanced methods

Self-supervised learning method

❖ CSI: Novelty Detection via Contrastive Learning on Distributionally Shifted Instances

- 2020년 NeurIPS에 발표된 논문으로, 2022년 6월 28일 기준으로 총 182회 인용
- Contrastive Learning을 OOD Detection에 적용한 논문



CSI: Novelty Detection via Contrastive Learning on Distributionally Shifted Instances

Jihoon Tack[*†], **Sangwoo Mo**[*‡], **Jongheon Jeong**[‡], **Jinwoo Shin**[†‡]
[†]Graduate School of AI, KAIST
[‡]School of Electrical Engineering, KAIST
{jihoontack,swmo,jongheonj,jinwoos}@kaist.ac.kr

Abstract

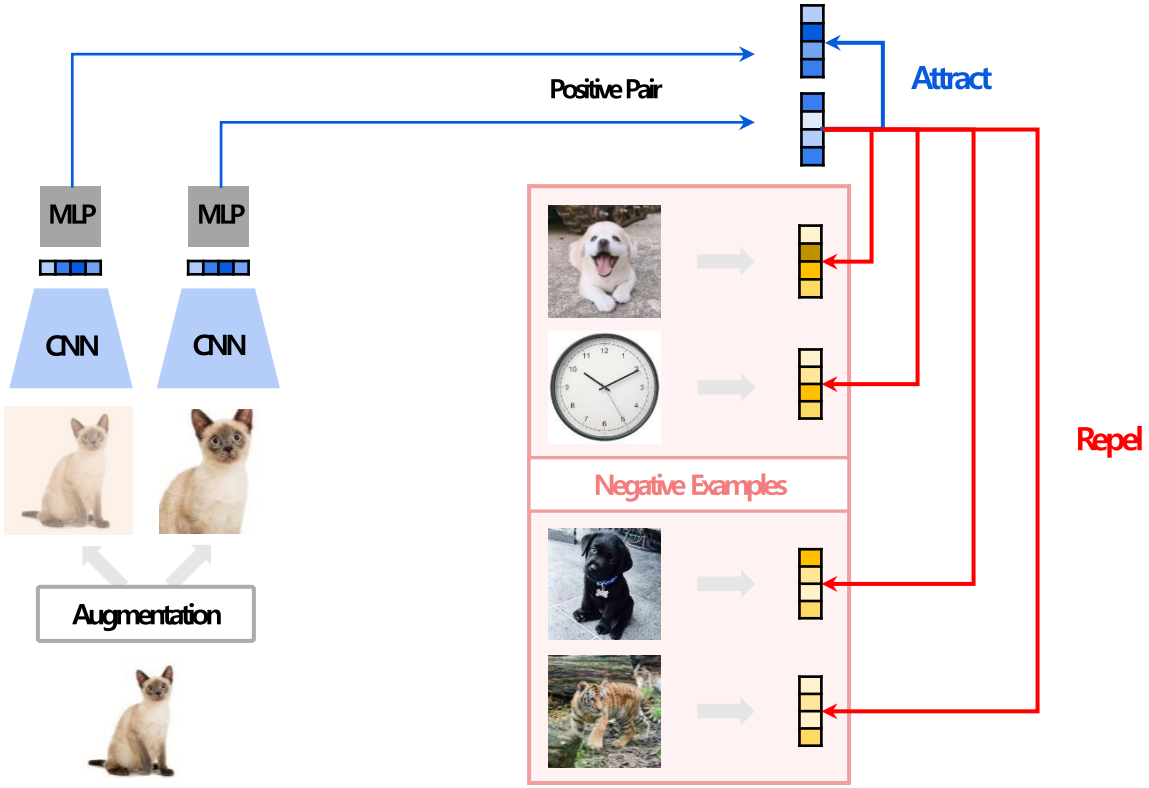
Novelty detection, *i.e.*, identifying whether a given sample is drawn from outside the training distribution, is essential for reliable machine learning. To this end, there have been many attempts at learning a representation well-suited for novelty detection and designing a score based on such representation. In this paper, we propose a simple, yet effective method named *contrasting shifted instances* (CSI), inspired by the recent success on contrastive learning of visual representations. Specifically, in addition to contrasting a given sample with other instances as in conventional contrastive learning methods, our training scheme contrasts the sample with distributionally-shifted augmentations of itself. Based on this, we propose a new detection score that is specific to the proposed training scheme. Our experiments demonstrate the superiority of our method under various novelty detection scenarios, including unlabeled one-class, unlabeled multi-class and labeled multi-class settings, with various image benchmark datasets. Code and pre-trained models are available at https://github.com/alinlab/CSI.

Advanced methods

Self-supervised learning method

❖ SimCLR

- 배치 내의 이미지를 추출하여 2개의 data augmentation을 적용
  - ✓ 동일 이미지에 적용된 augmentation은 positive
  - ✓ 다른 이미지에 적용된 augmentation은 negative



Advanced methods

Self-supervised learning method

❖ Shifting Transformation

- Rotation 등의 일부 Augmentation이 SimCLR의 성능 하락 야기
- 이러한 Augmentation을 동일 이미지에 적용 시 negative sample로 활용

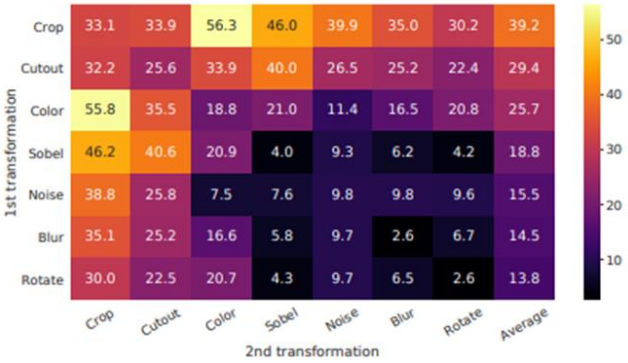
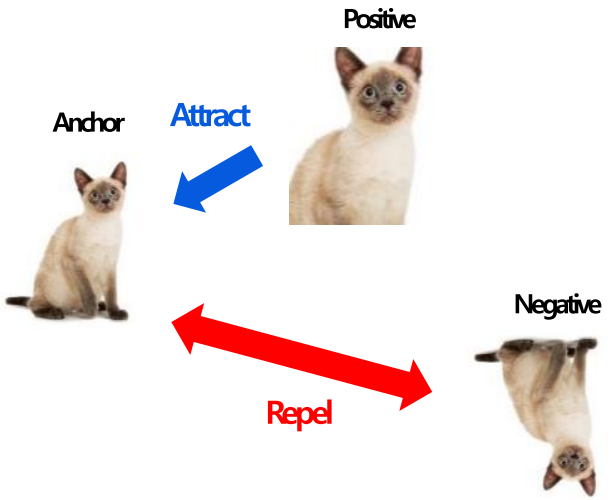


Figure 5. Linear evaluation (ImageNet top-1 accuracy) under individual or composition of data augmentations, applied only to one branch. For all columns but the last, diagonal entries correspond to single transformation, and off-diagonals correspond to composition of two transformations (applied sequentially). The last column reflects the average over the row.



Chen, Ting, et al. "A simple framework for contrastive learning of visual representations." International conference on machine learning. PMLR, 2020.

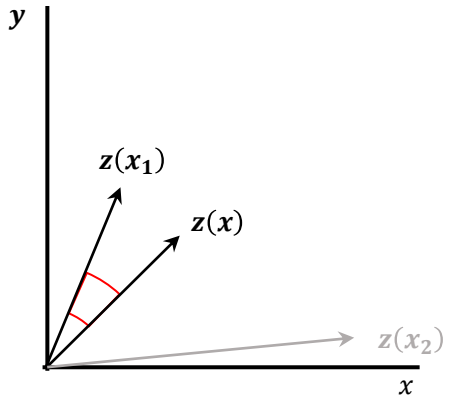
Advanced methods

Self-supervised learning method

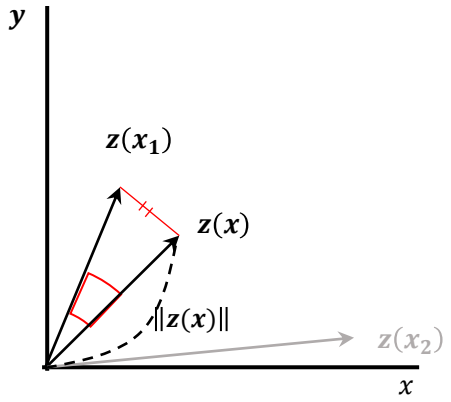
❖ OOD-ness

- 각 Augmentation 기법의 OOD-ness를 측정하여 Shifting transformation을 선택
- OOD-ness : OOD detection score를 통해 In-distribution과 Transformed sample간의 AUROC 값
  - ✓ OOD detection score : 가장 가까운 sample의 코사인 유사도 · Representation vector의 Norm

$$s_{con}(x;\{x_m\}) := \max_m sim\big(z(x_m), z(x)\big) \cdot \|z(x)\|$$



$$sim\big(z(x_m), z(x)\big) = \frac{z(x_m) \cdot z(x)}{\|z(x_m)\| \, \|z(x)\|}$$



$$\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p\right)^{1/p}$$

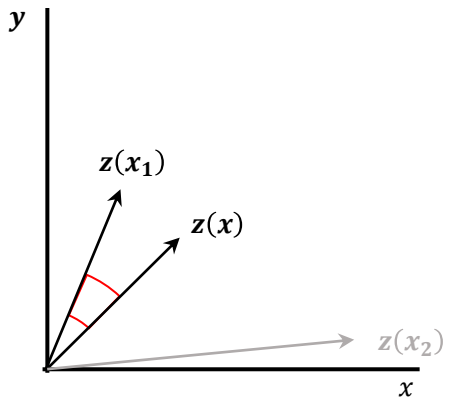
Advanced methods

Self-supervised learning method

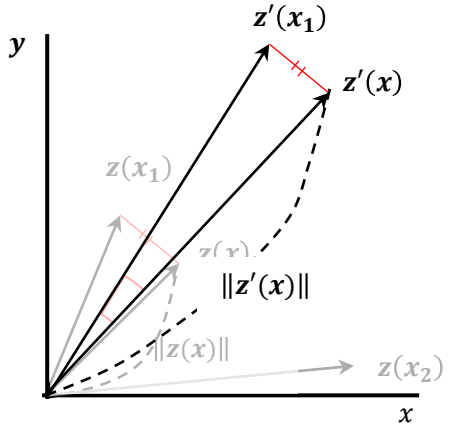
❖ OOD-ness

- 각 Augmentation 기법의 OOD-ness를 측정하여 Shifting transformation을 선택
- OOD-ness : OOD detection score를 통해 In-distribution과 Transformed sample간의 AUROC 값
  - ✓ OOD detection score : 가장 가까운 sample의 코사인 유사도 · Representation vector의 Norm

$$s_{con}(x;\{x_m\}) := \max_m sim\big(z(x_m), z(x)\big) \cdot \|z(x)\|$$



$$sim\big(z(x_m), z(x)\big) = \frac{z(x_m) \cdot z(x)}{\|z(x_m)\| \, \|z(x)\|}$$



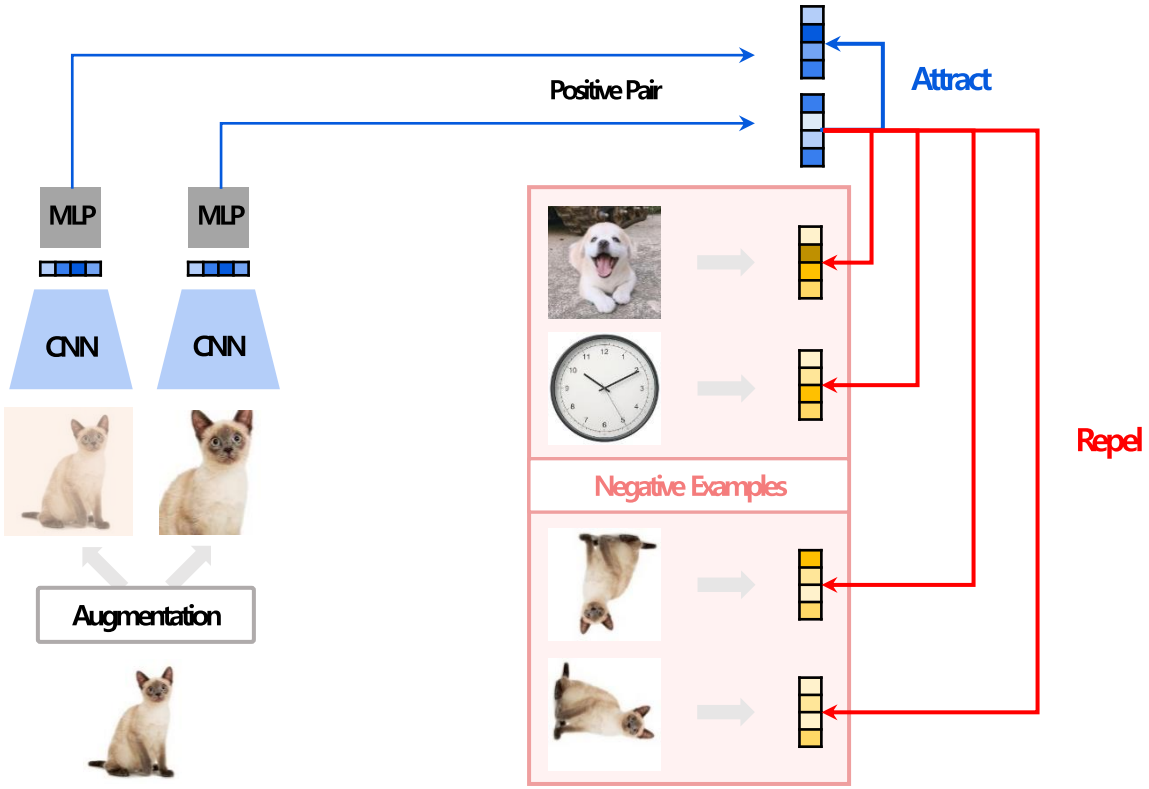
$$\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p\right)^{1/p}$$

Advanced methods

Self-supervised learning method

❖ CSI: Contrasting Shifted Instances

- 기존의 SimCLR와 동일하나, Shifting transformation을 적용한 이미지를 Negative sample로 분류
- Contrastive Learning을 적용하여, Negative samples는 밀어내고 Positive pair는 끌어당기게 학습

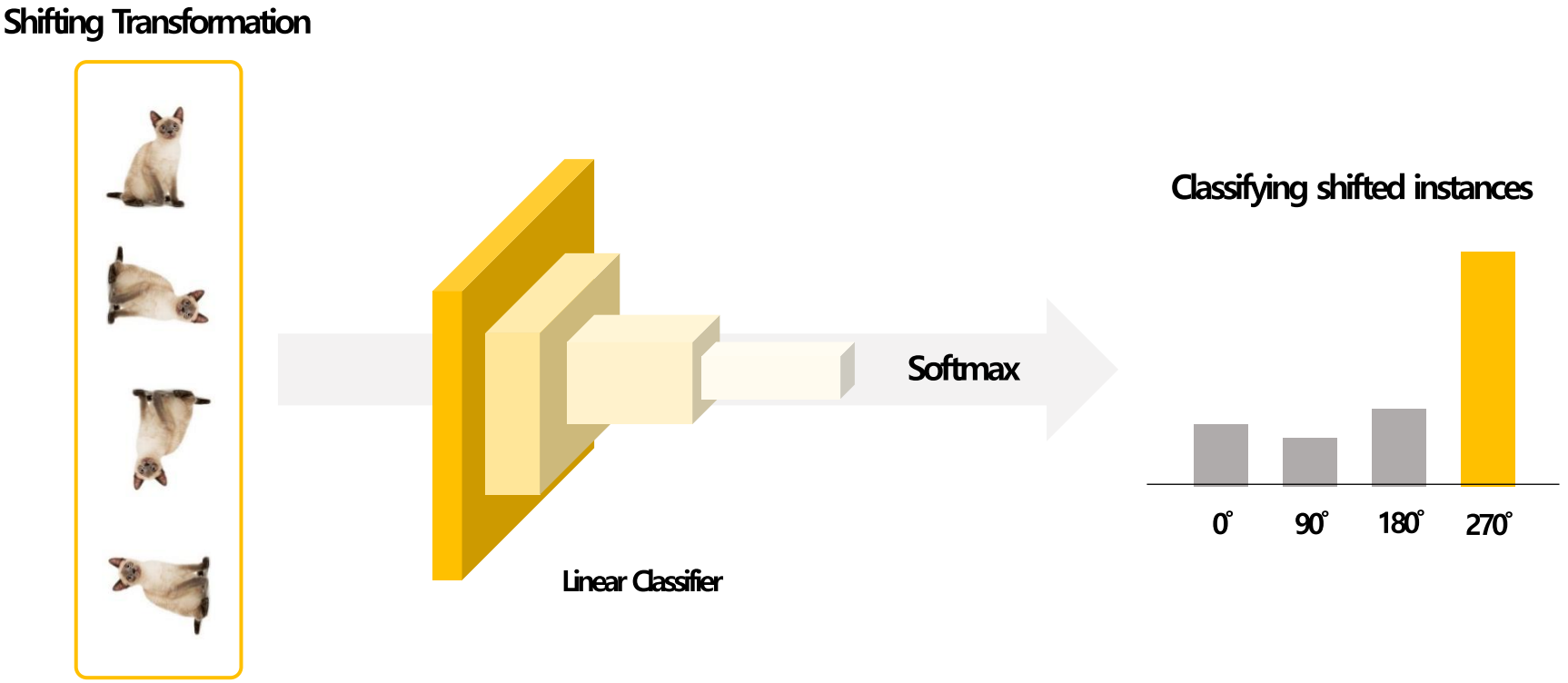


Advanced methods

Self-supervised learning method

❖ CSI: Classifying Shifted Instances

- OOD-ness가 가장 높은 Rotation을 Shifting Transformation으로 설정
- 추가적인 Representation을 위해, Auxiliary task로 회전한 각도를 Classification하는 작업을 수행

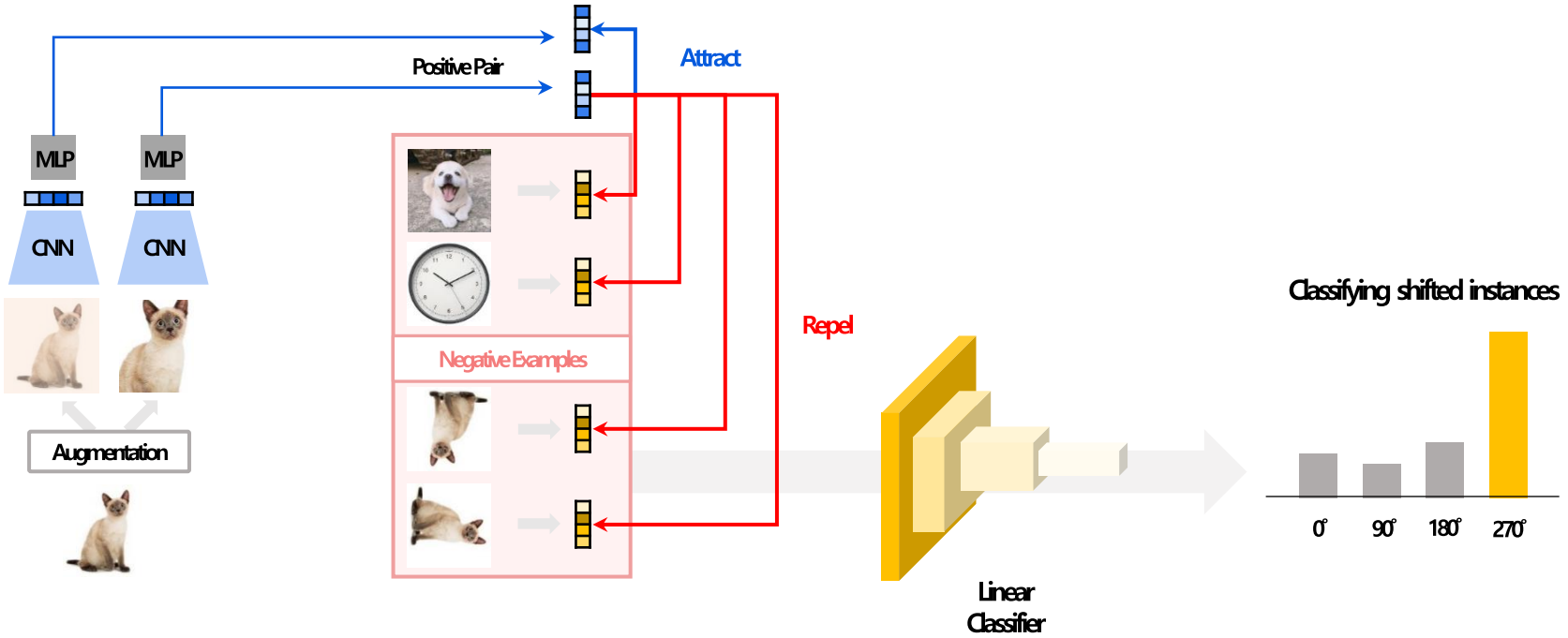


Advanced methods

Self-supervised learning method

❖ CSI

- 최종적으로 Contrastive Learning과 Classification을 동시에 학습 진행



Conclusion

❖ Conclusion

- Out-of-distribution Detection이란 Multi-class Classification 상황에서 Classification 성능을 유지함과 동시에 학습하지 않은 데이터셋도 찾아내는 연구분야

- 본 세미나에서는 OOD Detection에 대한 정의와 아래의 세가지 방법론을 간략히 소개
  - ✓ Baseline method : MSP
  - ✓ OOD data generation method : Confidence loss + GAN loss
  - ✓ Self-supervised learning method : CSI

Thank you

E-mail : jj950310@korea.ac.kr

References

- ✓ Yang, J., Zhou, K., Li, Y., & Liu, Z. (2021). Generalized out-of-distribution detection: A survey. arXiv preprint arXiv:2110.11334.
- ✓ Hendrycks, D., & Gimpel, K. (2016). A baseline for detecting misclassified and out-of-distribution examples in neural networks. arXiv preprint arXiv:1610.02136.
- ✓ Lee, K., Lee, H., Lee, K., & Shin, J. (2017). Training confidence-calibrated classifiers for detecting out-of-distribution samples. arXiv preprint arXiv:1711.09325.
- ✓ Tack, J., Mo, S., Jeong, J., & Shin, J. (2020). Csi: Novelty detection via contrastive learning on distributionally shifted instances. Advances in neural information processing systems, 33, 11839-11852.

- ✓ https://hoya012.github.io/blog/anomaly-detection-overview-1/
- ✓ https://hoya012.github.io/blog/anomaly-detection-overview-2/
- ✓ https://www.youtube.com/watch?v=NOzDB2Rpbi0
- ✓ https://bo-10000.tistory.com/125