**2023 동계 세미나**

# Long-Tail Distribution Learning

***Sogang University***
*Vision & Display Systems Lab, Dept. of Electronic Engineering*

***Presented By***
이창헌

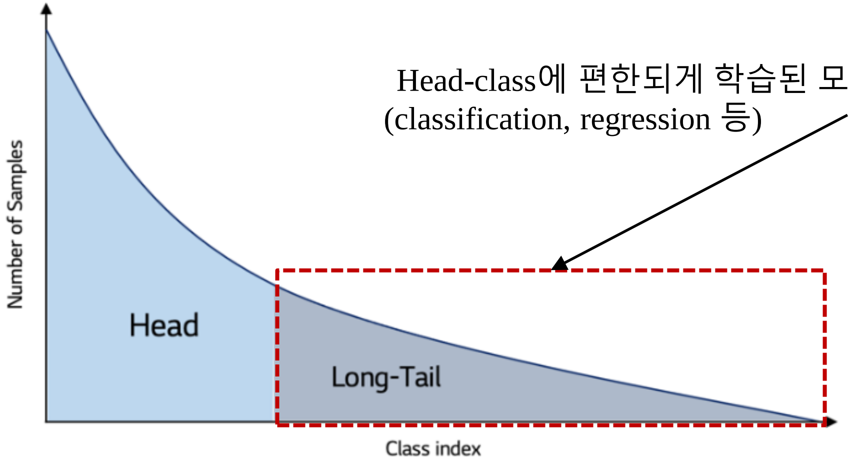
# Outline

- Background
  - Long-tail distribution learning
    - Re-sampling
      - Under-sampling
      - Over-sampling
    - Cost-Sensitive Learning
    - Decoupled Learning
- Recent paper
  - Partial and asymmetric contrastive learning for out-of-distribution detection in long-tailed recognition

# Background

- Long-tail distribution
  - Long-tail distribution을 가진 데이터셋이란?
    - Class imbalance의 불균형 비율이 매우 큰 특수한 경우
  - Head-class
    - 데이터의 수가 많은 dominant class
  - Long-tail class
    - 데이터의 수가 적은 scarce class

Head-class에 편한되게 학습된 모델은 tail-class에 대해 성능이 떨어짐
(classification, regression 등)

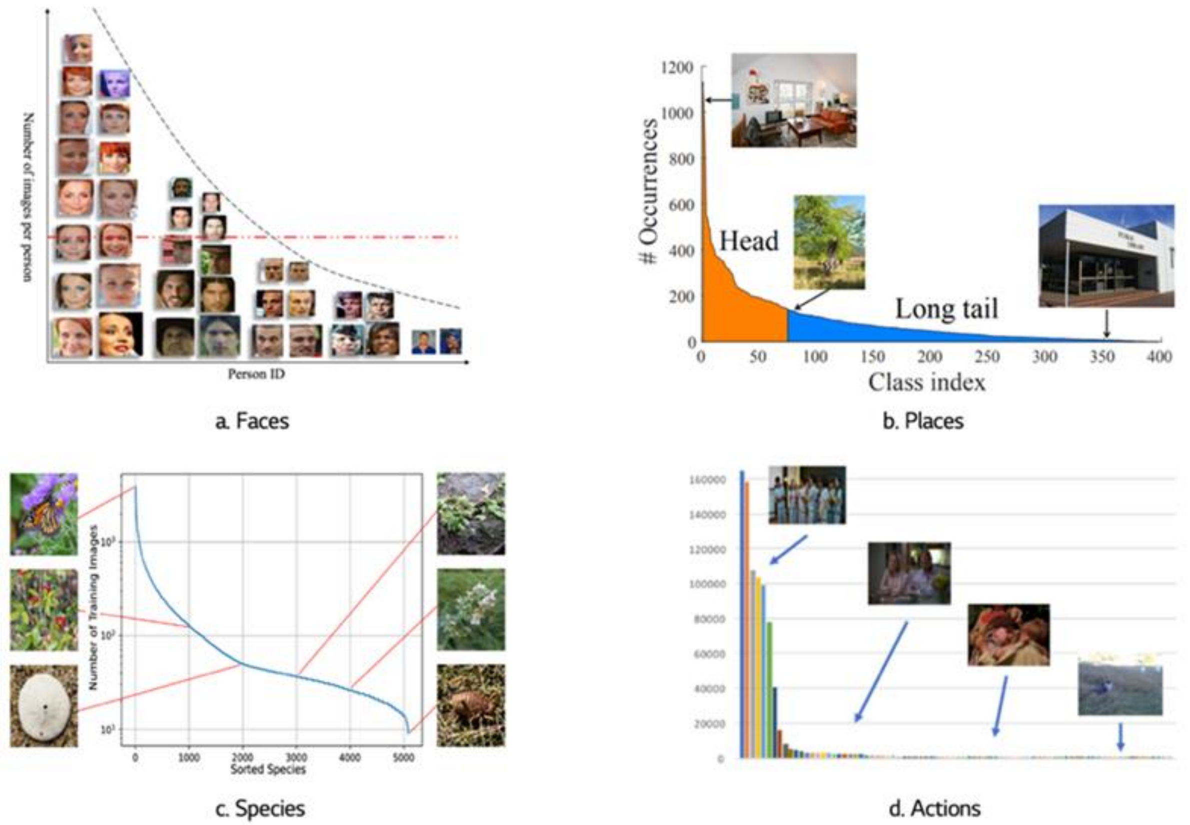



<Long-tail distribution 예시>

1) Ho, Jonathan et al. "Denoising diffusion probabilistic models." *Advances in Neural Information Processing Systems* (2020).

# Background

- Long-tail distribution
  - Real-world에서 수집된 large-scale dataset에서는 피해갈 수 없는 문제




<다양한 domain에서 long-tail distribution 예시>

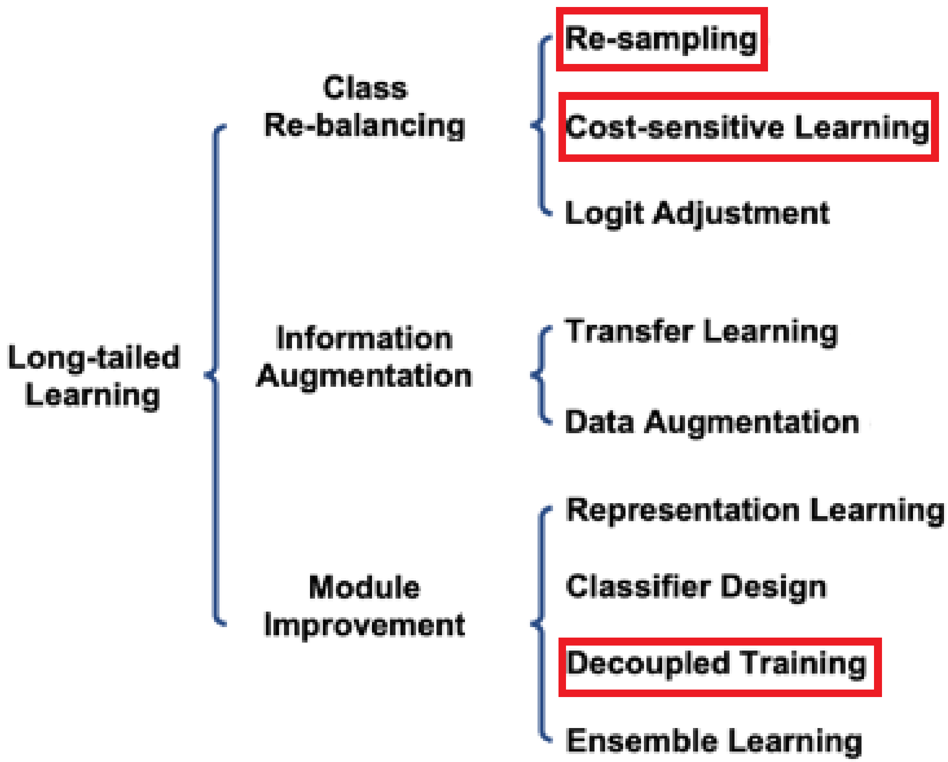
서강대학교
SOGANG UNIVERSITY



VDS LAB

# Background

- Long-tail distribution
  - How do we solve long-tail distributions?




<Long-tail distribution learning 분류>

1) Ho, Jonathan et al. "Denoising diffusion probabilistic models." *Advances in Neural Information Processing Systems* 33 (2020).

# Background

- Long-tail distribution [1]
  - Re-Sampling
    - Under-sampling
      - Head-class를 줄여서 tail-class와 데이터의 갯수를 맞추는 방법
    - Over-sampling
      - Tail-class를 늘려서 head-class와 데이터의 갯수를 맞추는 방법


<Under-sampling> <Over-sampling>

<Re-sampling 방법 예시>

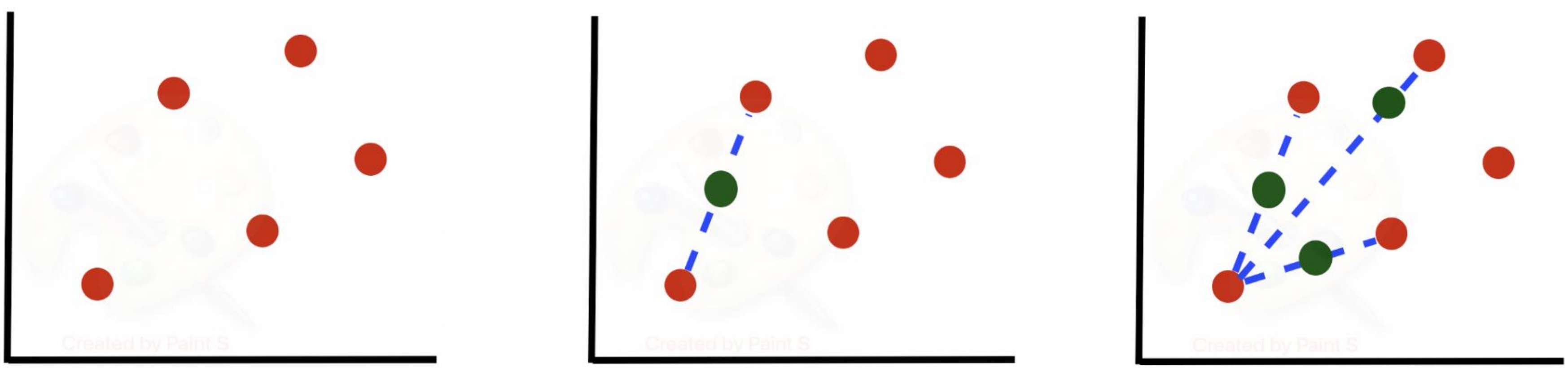
1) Chawla, Nitesh V., et al. "SMOTE: synthetic minority over-sampling technique." *Journal of artificial intelligence research* 16 (2002): 321-357.

# Background

- Long-tail distribution [1]
  - Re-Sampling
    - Over-sampling
      - SMOTE[1)]

1) 특정 데이터에 대해 nearest neighbor와의 distance를 구함
2) Distance에 0~1 사이의 랜덤 상수 $c$를 곱함
3) 기존 데이터에 $c$ * Distance를 더해 feature space 상에서 신규 데이터를 생성함
4) 1)~3)의 과정을 $n$번 반복함

: 생성된 신규 데이터
: 기존 데이터



<SMOTE 알고리즘을 활용한 데이터 생성 예시>

# Background

- Long-tail distribution [2]
  - Cost-Sensitive Learning
    - 각 class 별 loss 값을 다르게 주어 re-balance하는 algorithm적인 접근법
      - Softmax
        - ✓ $S(z_i) = \frac{e^{z_i}}{\sum_j e^{z_j}}$
      - Negative log-likelihood loss
        - ✓ $L_{nl} = -\log(\frac{e^{z_i}}{\sum_j e^{z_j}}) = -\log(p_i)$
      - Weighted softmax loss
        - ✓ $L_{wnl} = -\frac{1}{n_i}\log(p_i)$
      - Focal loss[1)]
        - ✓ $L_{fl} = -(1-p_i)^{\gamma}\log(p_i)$

$z$ : predicted logits

$p$ : softmax probability

$n_i$ : i class의 데이터 수

$\gamma$ : 상수 parameter

$i$ : class

# Background

- Long-tail distribution [2]
  - Cost-Sensitive Learning
    - 각 class 별 loss 값을 다르게 주어 re-balance하는 algorithm적인 접근법
      - Negative log-likelihood loss
        - ✓ $L_{nl} = -\log(p_i)$
      - Focal loss[1]
        - ✓ $L_{fl} = -(1 - p_i)^{\gamma}\log(p_i)$

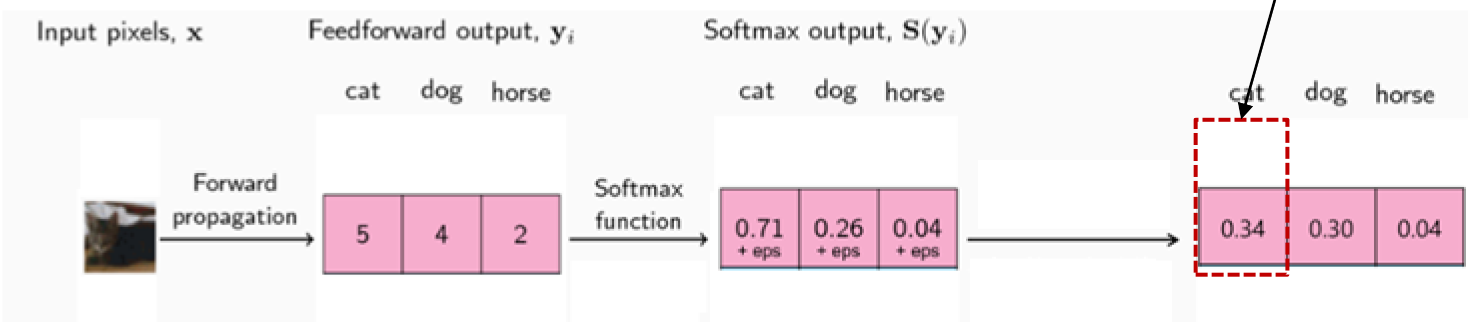
    - Case 1 : $p_i = 0.1$인 경우 (uncertainty가 높은 tail-class)
      - Negative log-likelihood loss = $-log$ (0.1) = 2.3
      - Focal loss = $-(1 - 0.1)log$ (0.1) = 2.1
    - Case 2 : $p_i = 0.9$인 경우 (uncertainty가 낮은 head-class)
      - Negative log-likelihood loss = $-log$ (0.9) = 0.1
      - Focal loss = $-(1 - 0.9)log$ (0.9) = 0.01

# Background

- Long-tail distribution [2]
  - Cost-Sensitive Learning
    - Balanced softmax loss[1)] 예시

$$L_{bs} = -\log\left(\frac{n_i e^{z_i}}{\sum_j n_j e^{z_j}}\right)$$

$-\log\left(\frac{e^5}{e^5+e^4+e^2}\right) = 0.34$




- Case 1 : cat이 head-class인 경우 cat = 10, dog = 5, horse = 2

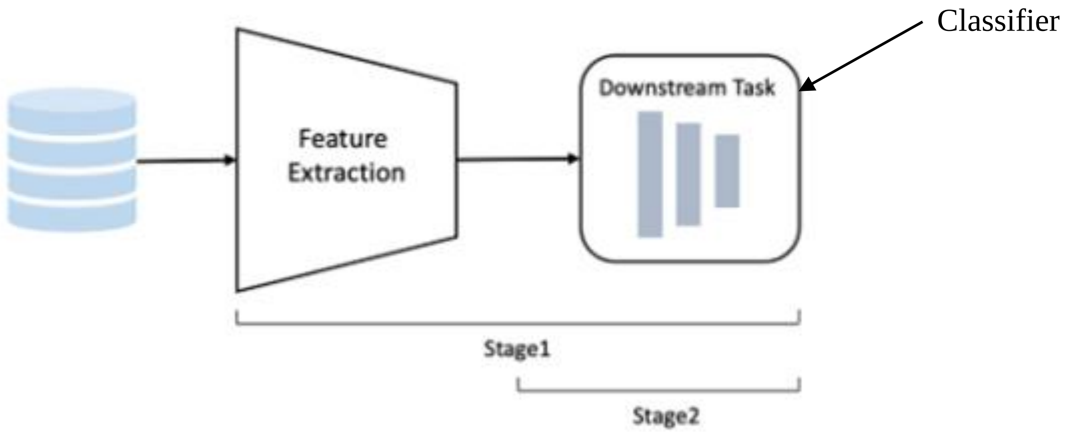
$$-log\left(\frac{10e^5}{10e^5+5e^4+2e^2}\right) = 0.07$$

- Case 2 : cat이 tail-class인 경우 cat = 2, dog = 5, horse = 10

$$-log\left(\frac{2e^5}{2e^5+5e^4+10e^2}\right) = 0.33$$

# Background

- Long-tail distribution [3]
  - Decoupled Learning[1)]
    - Stage 1 : Representation learning (class-imbalance를 고려하지 않음)
      - 기존의 end-to-end 방식으로 feature extractor와 classifier를 함께 학습함
    - Stage 2 : Classifier finetuning (class-imbalance를 고려함)
      - Feature extractor는 freeze된 상태에서 classifier를 재학습함

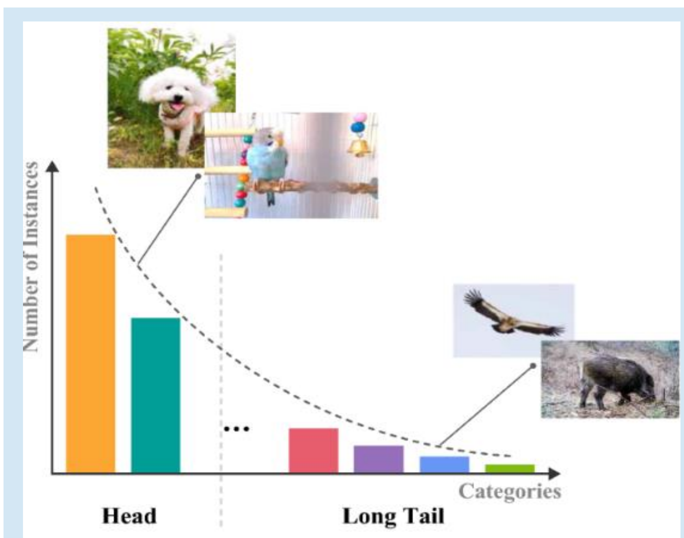



<Decoupling representation and classifier 학습 방법>

1) Wang, Haotao, et al. "Partial and asymmetric contrastive learning for out-of-distribution detection in long-tailed recognition." International Conference on Machine Learning. PMLR, 2022.

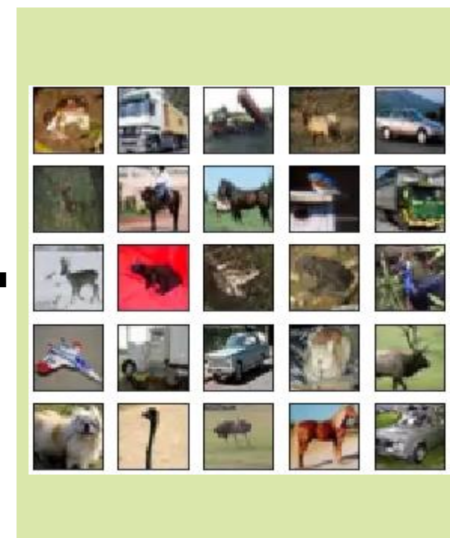
# OOD Detection in Long-tailed recognition

- Partial and asymmetric contrastive learning for out-of-distribution detection in long-tailed recognition[1)]
  - Out-of-distribution detection
    - In-distribution 데이터와 unlabeled out-of-distribution dataset을 사용해 학습 후 inference 시에 새로운 sample이 in-distribution일 경우에는 정확하게 분류
    - Uncertainty가 높아서 out-of-distribution 데이터로 판단될 경우 outlier로 걸러냄
  - OOD detection + Long-tailed recognition 문제가 동시에 존재하는 데이터셋은?




<Long-tailed in-distribution 데이터셋 예시>

+



<무작위 데이터로 구성된 OOD 데이터셋>

# OOD Detection in Long-tailed recognition

- Partial and asymmetric contrastive learning for out-of-distribution detection in long-tailed recognition[1)]
  - OOD detection + Long-tailed recognition 문제가 동시에 존재하는 데이터셋은?

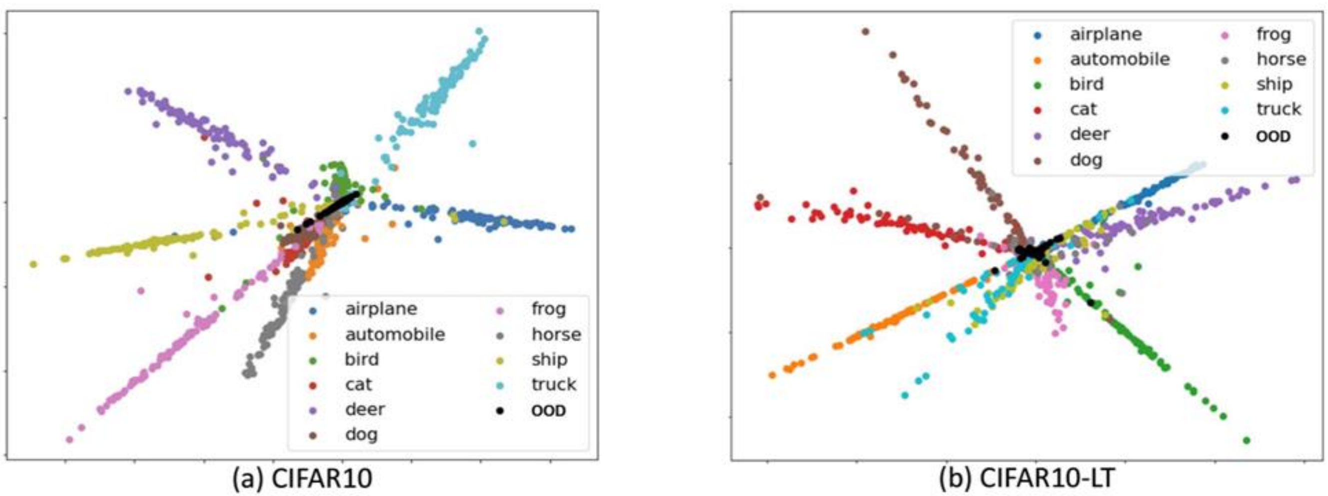
| Method | Dataset | AUROC (↑) | AUPR (↑) | FPR95 (↓) | ACC (↑) |
|---|---|---|---|---|---|
| NT (MSP) | CIFAR10 | 85.86 | 84.37 | 52.52 | 93.45 |
| | CIFAR10-LT | 72.28 (-13.58) | 70.27 (-14.10) | 66.07 (+13.55) | 72.34 (-21.11) |
| OE | CIFAR10 | 96.68 | 96.29 | 14.59 | 92.81 |
| | CIFAR10-LT | 89.92 (-6.75) | 87.71 (-8.58) | 34.80 (+20.21) | 73.30 (-19.51) |
| EnergyOE | CIFAR10 | 96.59 | 96.37 | 14.80 | 93.07 |
| | CIFAR10-LT | 89.31 (-7.27) | 88.92 (-7.45) | 40.88 (+26.08) | 74.68 (-18.39) |
| SOFL | CIFAR10 | 96.74 | 96.60 | 14.57 | 89.13 |
| | CIFAR10-LT | 91.13 (-5.61) | 90.49 (-6.10) | 34.98 (+20.41) | 54.42 (-34.71) |
| OECC | CIFAR10 | 96.27 | 95.41 | 14.77 | 91.95 |
| | CIFAR10-LT | 87.28 (-8.99) | 86.29 (-9.12) | 45.24 (+30.47) | 60.16 (-31.79) |
| NTOM | CIFAR10 | 96.92 | 96.95 | 14.95 | 91.44 |
| | CIFAR10-LT | 92.89 (-4.03) | 92.31 (-4.65) | 29.03 (+14.09) | 66.41 (-25.03) |

<OOD class 추가 이후 일반 데이터셋과 long-tailed 데이터셋 성능 저하 비교>

1) Wang, Haotao, et al. "Partial and asymmetric contrastive learning for out-of-distribution detection in long-tailed recognition." International Conference on Machine Learning. PMLR, 2022.

# OOD Detection in Long-tailed recognition

- Partial and asymmetric contrastive learning for out-of-distribution detection in long-tailed recognition[1)]
  - Tail-class로 인한 out-of-distribution detection 성능 저하
    - 정확한 OOD-detection을 위해서는 OOD-sample의 uncertainty가 높아야함
    - Tail-class의 낮은 데이터량, variance로 인해 학습 시 under-represent 됨
      - OOD-detection 모델이 OOD sample과의 decision boundary를 잘 찾지 못하여 OOD sample에 대해서 over-confident한 prediction을 내게 됨




<일반 데이터셋과 long-tailed 데이터셋의 OOD-detection 성능 시각화>

서강대학교
SOGANG UNIVERSITY

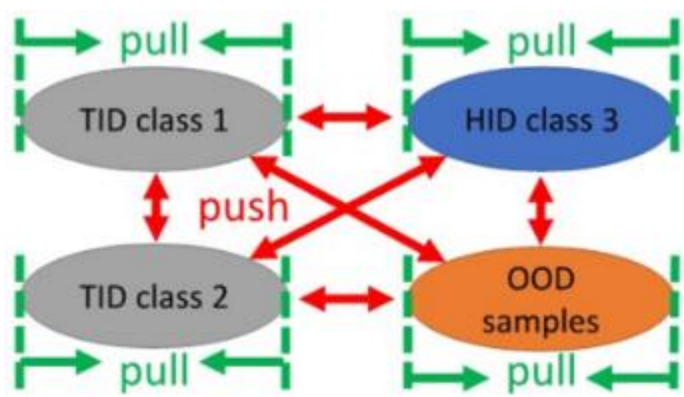


VDS LAB

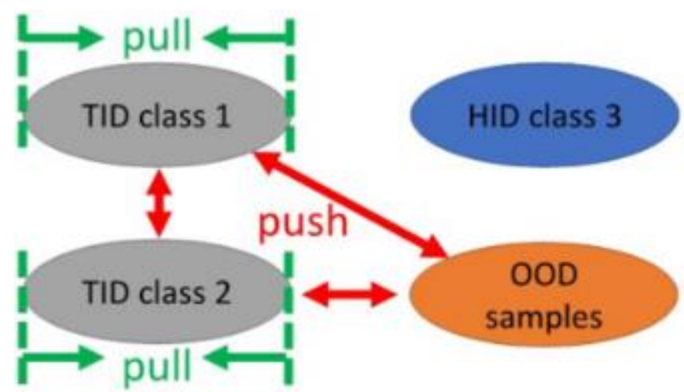
1) Wang, Haotao, et al. "Partial and asymmetric contrastive learning for out-of-distribution detection in long-tailed recognition." International Conference on Machine Learning. PMLR, 2022.

# OOD Detection in Long-tailed recognition

- Partial and asymmetric contrastive learning for out-of-distribution detection in long-tailed recognition[1)]
  - Contrastive learning
    - Feature space 상에서 같은 class의 데이터는 가깝게, 다른 class의 데이터는 멀어지게 모델을 학습함



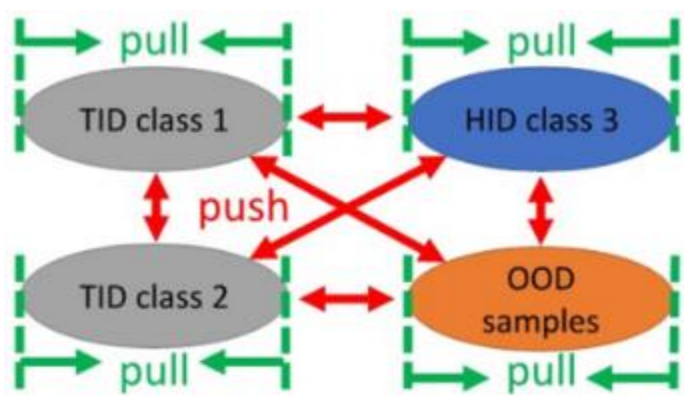

<Naive한 contrastive learning>



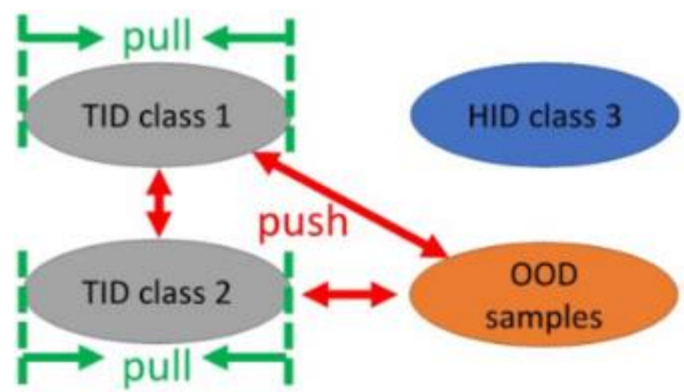

<제안된 contrastive learning>

# OOD Detection in Long-tailed recognition

- Partial and asymmetric contrastive learning for out-of-distribution detection in long-tailed recognition[1)]
  - Partial and asymmetric supervised contrastive learning(PASCL)
    - Partiality
      - OOD sample과 tail-class sample에만 부분적으로 contrastive Learning을 적용함
    - Asymmetry
      - OOD sample은 feature space에서 서로 가까운 공간에 위치하도록 pull 하지 않음




<Naive한 contrastive learning>




<제안된 contrastive learning>

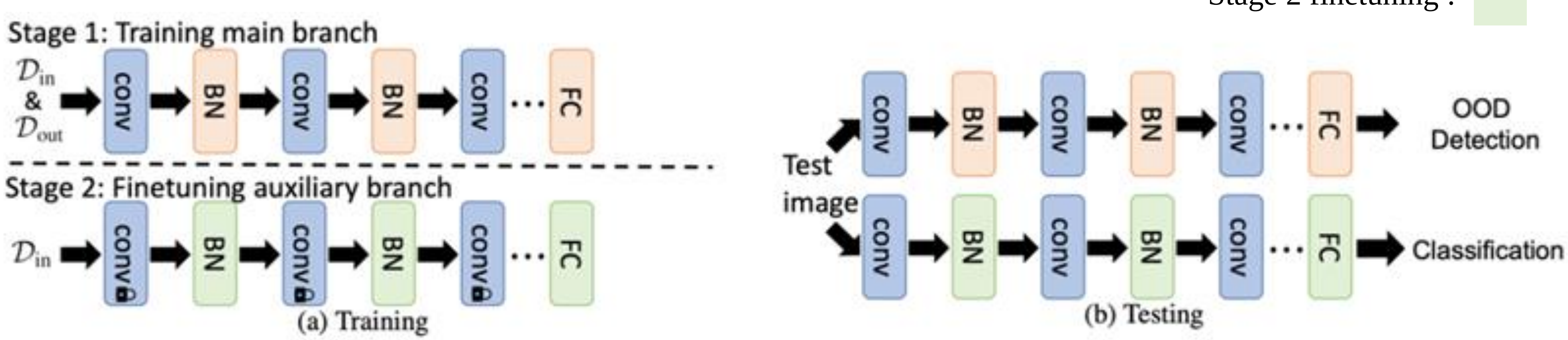
1) Wang, Haotao, et al. "Partial and asymmetric contrastive learning for out-of-distribution detection in long-tailed recognition." International Conference on Machine Learning. PMLR, 2022.

# OOD Detection in Long-tailed recognition

- Partial and asymmetric contrastive learning for out-of-distribution detection in long-tailed recognition[1)]
  - Auxiliary Branch Fine-Tuning (ABF) - Decoupled learning
    - Stage 1에서 OOD + in-distribution 데이터를 사용해서 end-to-end 방식으로 모든 네트워크를 학습함
    - Stage 2에서는 in-distribution 데이터만 사용해서 convolutional layer를 freeze한 상태에서 classifier를 finetuning함
      - 학습 데이터가 바뀌었음으로 batch normalization layer에 대해서도 finetuning 진행함

Stage 1 training :

Stage 2 finetuning :




<제안된 decoupled learning>

# OOD Detection in Long-tailed recognition

- Partial and asymmetric contrastive learning for out-of-distribution detection in long-tailed recognition[1)]
  - Experiment results

(a) OOD detection results and in-distribution classification results in terms of ACC95.

| $\mathcal{D}_{out}^{test}$ | Method | AUROC (↑) | AUPR (↑) | FPR95 (↓) | ACC95 (↑) |
|---|---|---|---|---|---|
| Texture | OE | **76.71** ± 1.20 | **58.79** ± 1.39 | 68.28 ± 1.53 | 71.43 ± 1.58 |
| | Ours | 76.01 ± 0.66 | 58.12 ± 1.06 | **67.43** ± 1.93 | **73.11** ± 1.55 |
| SVHN | OE | 77.61 ± 3.26 | 86.82 ± 2.50 | 58.04 ± 4.82 | 64.27 ± 3.26 |
| | Ours | **80.19** ± 2.19 | **88.49** ± 1.59 | **53.45** ± 3.60 | **64.50** ± 1.87 |
| CIFAR10 | OE | 62.23 ± 0.30 | **57.57** ± 0.34 | 80.64 ± 0.98 | **82.67** ± 0.99 |
| | Ours | **62.33** ± 0.38 | 57.14 ± 0.20 | **79.55** ± 0.84 | 82.30 ± 1.07 |
| Tiny ImageNet | OE | 68.04 ± 0.37 | **51.66** ± 0.51 | 76.66 ± 0.47 | 76.22 ± 0.61 |
| | Ours | **68.20** ± 0.37 | 51.53 ± 0.42 | **76.11** ± 0.80 | **77.56** ± 1.15 |
| LSUN | OE | 77.10 ± 0.64 | **61.42** ± 0.99 | 63.98 ± 1.38 | 65.64 ± 1.03 |
| | Ours | **77.19** ± 0.44 | 61.27 ± 0.72 | **63.31** ± 0.87 | **68.05** ± 1.24 |
| Places365 | OE | 75.80 ± 0.45 | **86.68** ± 0.38 | 65.72 ± 0.92 | 67.04 ± 0.49 |
| | Ours | **76.02** ± 0.21 | 86.52 ± 0.29 | **64.81** ± 0.27 | **69.04** ± 0.90 |
| Average | OE | 72.91 ± 0.68 | 67.16 ± 0.57 | 68.89 ± 1.07 | 71.21 ± 0.84 |
| | Ours | **73.32** ± 0.32 | **67.18** ± 0.10 | **67.44** ± 0.58 | **72.43** ± 0.66 |

(b) in-distribution classification results in terms of ACC@FPR$n$.

| Method | ACC@FPRn (↑) 0 | 0.001 | 0.01 | 0.1 |
|---|---|---|---|---|
| OE | 39.04 ± 0.37 | 39.07 ± 0.38 | 39.38 ± 0.38 | 42.40 ± 0.44 |
| Ours | **43.10** ± 0.47 | **43.12** ± 0.47 | **43.39** ± 0.48 | **46.14** ± 0.38 |

(c) Comparison with other methods.

| $\mathcal{D}_{out}^{test}$ | Method | AUROC (↑) | AUPR (↑) | FPR95 (↓) | ACC (↑) |
|---|---|---|---|---|---|
| Average | ST (MSP) | 61.00 | 57.54 | 82.01 | <u>40.97</u> |
| | OECC | 70.38 | 66.87 | 73.15 | 32.93 |
| | EnergyOE | 71.10 | <u>67.23</u> | 71.78 | 39.05 |
| | OE | <u>72.91</u> ± 0.68 | 67.16 ± 0.57 | <u>68.89</u> ± 1.07 | 39.04 ± 0.37 |
| | Ours | **73.32** ± 0.32 | **67.18** ± 0.10 | **67.44** ± 0.58 | **43.10** ± 0.47 |

<논문 실험 결과>

서강대학교 SOGANG UNIVERSITY

VDS LAB

# Experiment

- Partial and Asymmetric Contrastive Learning for Out-of-Distribution Detection in Long-tailed Recognition
  - Contrastive learning

| $\mathcal{D}_{\text{in}}$ | Asymmetry | Partiality | ABF | **AUROC** (↑) | **AUPR** (↑) | **FPR95** (↓) | **ACC95** (↑) | **ACC@FPR$n$** (↑) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | 0 | 0.001 | 0.01 | 0.1 |
| CIFAR10-LT | No contrastive loss (OE) | | | 95.10 ± 1.01 | 97.14 ± 0.81 | 16.15 ± 1.52 | 81.33 ± 0.81 | 73.84 ± 0.77 | 73.90 ± 0.77 | 74.46 ± 0.81 | 78.88 ± 0.66 |
| | ✗ | ✗ | ✗ | 95.34 ± 1.58 | 97.30 ± 1.20 | 15.12 ± 3.07 | 81.94 ± 1.28 | 75.03 ± 1.46 | 75.09 ± 1.45 | 75.60 ± 1.44 | 80.02 ± 1.10 |
| | ✗ | ✓ | ✗ | 95.01 ± 1.25 | 96.74 ± 0.78 | 15.31 ± 4.35 | 82.34 ± 1.56 | 74.46 ± 1.80 | 74.52 ± 1.80 | 75.04 ± 1.76 | 80.21 ± 0.99 |
| | ✓ | ✗ | ✗ | 94.91 ± 1.43 | 96.86 ± 1.47 | 15.57 ± 1.19 | 82.08 ± 0.47 | 75.24 ± 0.99 | 75.29 ± 0.98 | 75.77 ± 0.98 | 79.85 ± 0.77 |
| | ✓ | ✓ | ✗ | **96.63** ± 0.90 | **98.06** ± 0.56 | **12.18** ± 3.33 | 81.70 ± 1.21 | 76.20 ± 0.79 | 76.26 ± 0.79 | 76.85 ± 0.81 | 81.07 ± 0.58 |
| | ✓ | ✓ | ✓ | **96.63** ± 0.90 | **98.06** ± 0.56 | **12.18** ± 3.33 | **82.72** ± 1.51 | **77.08** ± 1.01 | **77.13** ± 1.02 | **77.64** ± 0.99 | **81.96** ± 0.85 |
| CIFAR100-LT | No contrastive loss (OE) | | | 77.61 ± 3.26 | 86.82 ± 2.50 | 58.04 ± 4.82 | 64.27 ± 3.26 | 39.04 ± 0.37 | 39.07 ± 0.38 | 39.38 ± 0.38 | 42.40 ± 0.44 |
| | ✗ | ✗ | ✗ | 78.05 ± 2.12 | 87.18 ± 0.87 | 59.10 ± 5.03 | **66.44** ± 3.90 | 40.21 ± 0.43 | 40.25 ± 0.43 | 40.56 ± 0.45 | 43.71 ± 0.42 |
| | ✗ | ✓ | ✗ | 79.46 ± 1.83 | 88.01 ± 1.90 | 54.59 ± 3.34 | 63.86 ± 2.52 | 40.24 ± 0.53 | 40.28 ± 0.53 | 40.60 ± 0.55 | 43.93 ± 0.57 |
| | ✓ | ✗ | ✗ | 79.54 ± 2.38 | 87.68 ± 1.51 | 54.27 ± 3.69 | 63.33 ± 2.87 | 40.00 ± 0.42 | 40.04 ± 0.41 | 40.36 ± 0.42 | 43.60 ± 0.42 |
| | ✓ | ✓ | ✗ | **80.19** ± 2.19 | **88.49** ± 1.59 | **53.45** ± 3.60 | 63.10 ± 1.87 | 40.33 ± 0.20 | 40.36 ± 0.20 | 40.66 ± 0.18 | 43.79 ± 0.22 |
| | ✓ | ✓ | ✓ | **80.19** ± 2.19 | **88.49** ± 1.59 | **53.45** ± 3.60 | 64.50 ± 1.87 | **43.10** ± 0.47 | **43.12** ± 0.47 | **43.39** ± 0.48 | **46.14** ± 0.38 |

<논문 ablation study>

# Conclusion

- Long-tailed distribution 특성의 dataset에서 OOD detection task 성능을 높일 수 있는 방법을 제안함
  - 그러나 real-life application과 다르게 학습 시 in-distribution sample과 OOD sample이 이미 다른 distribution이라는 정보가 존재함
    - 위의 문제들을 극복하기 위해서는 unsupervised 환경에서 tail-class sample과 OOD sample을 구분할 수 있는 새로운 방법이 필요함
- Long-tail distribution의 특성과 이를 개선하기 위한 다양한 방법론이 존재하나 task, domain specific한 경우가 있음
  - Class imbalance가 존재하는 데이터셋의 근본적인 문제를 정확히 파악하고 특성에 따라서 적합한 알고리즘을 선정하는 능력이 필요함

# Thank you!