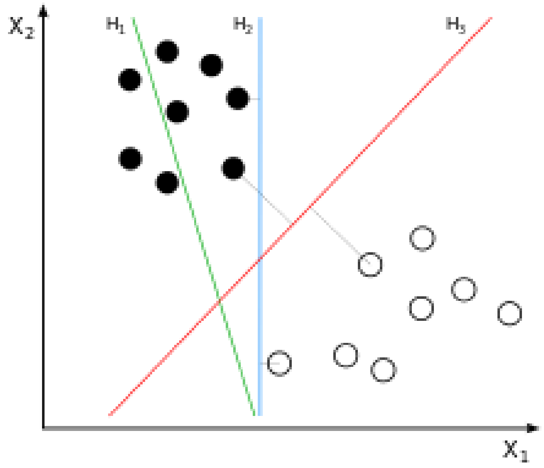
Maximum Class Separation as Inductive Bias in One Matrix

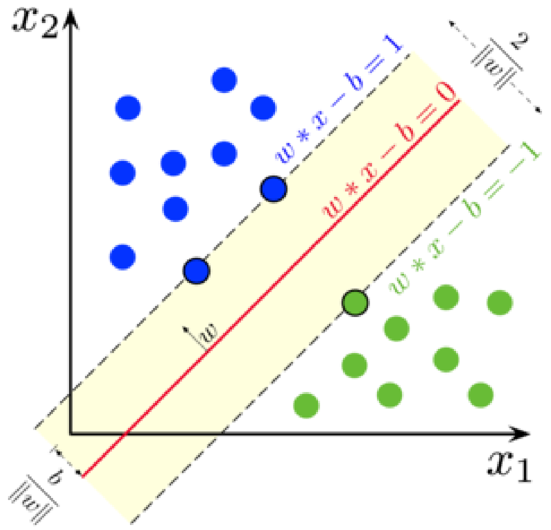
Tejaswi Kasarla, Gertjan J. Burghouts, Max van Spengler, Elise van der Pol , Rita Cucchiara, Pascal Mettes

전북대학교 IT정보공학과
202018392
박나현

Inductive bias

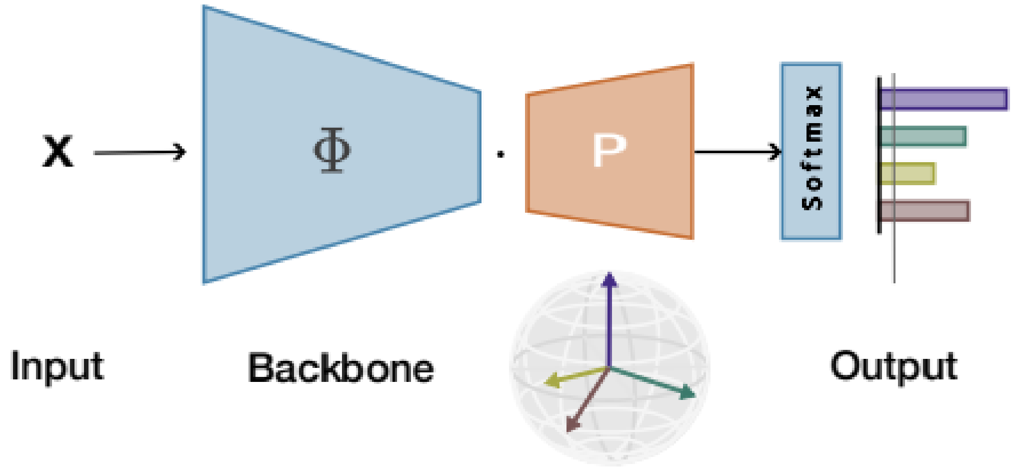
- Inductive bias
  - 문제를 더 잘 풀기위해 설계한 가정
  - Separation
    - SVM (margin 최대화), 알고리즘
    - 딥러닝에서 optimization
    - class 분리 위해 설계에 통합되어 unseen data 에 대한 분류 세팅 일반화 향상





Closed-Form Maximum Separation between Classes

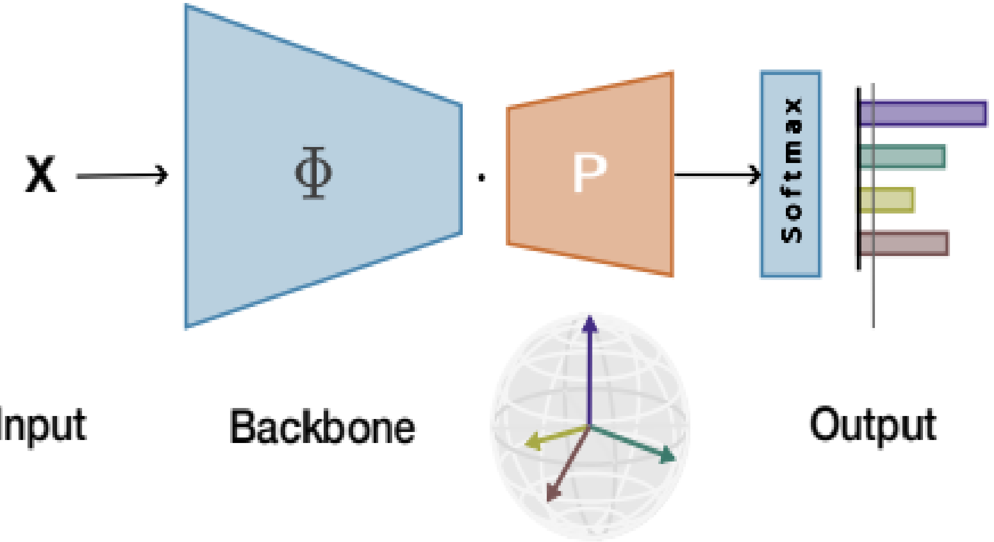
- simple alternative
  - **Encoding maximum separation as an inductive bias in the network by adding one fixed matrix multiplication before computing the softmax activation**
  - Optimization 대신 closed-form으로 해결
    - Closed-form : an equation which can be solved in terms of functions and mathematical operations.



Closed-Form Maximum Separation between Classes

Definition 1. (Maximally separated matrix) *For $k+1$ classes, let $P_k = [p_0, \ldots, p_k]$ denote a matrix of $k+1$ column vectors $\{p_i\}_{i=0}^{k} \subset \mathbb{S}^{k-1}$, such that $\forall_{i,j,k,i\neq j,j\neq k} \langle p_i, p_j \rangle = \langle p_i, p_k \rangle$ and* $\sum_{i=0}^{k} p_i = 0$.

- **두가지 조건** 전제 하에 모든 class vector 들이 서로 멀리 떨어짐 보장
  - **어떤 두 class vector 사이의 각도는 동일**
  - **모든 class vector 의 평균은 원점**



Closed-Form Maximum Separation between Classes

- Definition 1
  - k output 차원 사용 시, k+1 class가 최적으로 분리
  - Embedding 구성위해 pair-wise angular similarity 개념 필요

Cosine similarity

Lemma 1. *For maximally separated matrix P_k, $\forall_{i,j,i \neq j} \langle p_i, p_j \rangle = -\frac{1}{k}$*

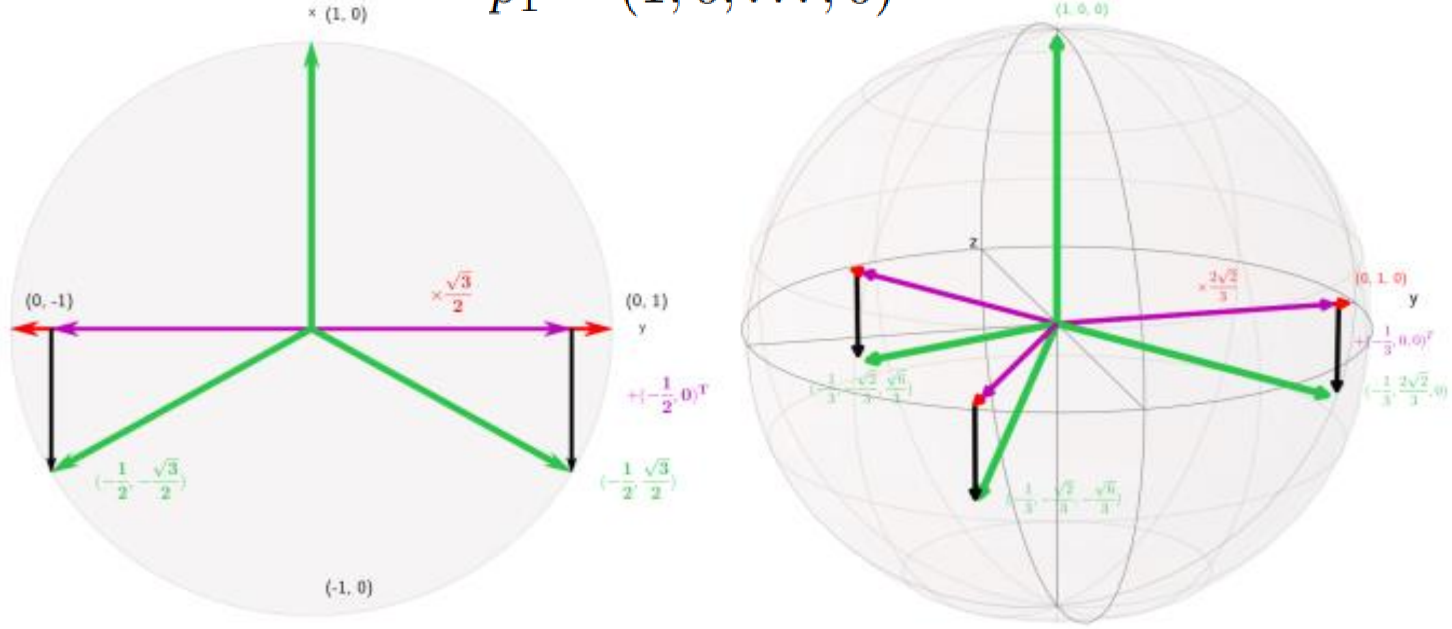
- k+1 class 의 최적 separation은 **전체 공간**을 사용한다고 생각 가능
  - 직교 분리만을 제공하는 one-hot encoding 과는 다르게 직교성을 넘어 클래스가 분리
- **재귀적으로 $\boldsymbol{P_k}$ 구성 가능**

Closed-Form Maximum Separation between Classes

$$P_1 = \begin{pmatrix} 1 & -1 \end{pmatrix} \in \mathbb{R}^{1\times 2}$$

$$P_k = \begin{pmatrix} 1 & -\frac{1}{k}\mathbf{1}^T \\ \mathbf{0} & \sqrt{1-\frac{1}{k^2}}\, P_{k-1} \end{pmatrix} \in \mathbb{R}^{k\times(k+1)}$$

$$p_1 = (1, 0, \ldots, 0)$$



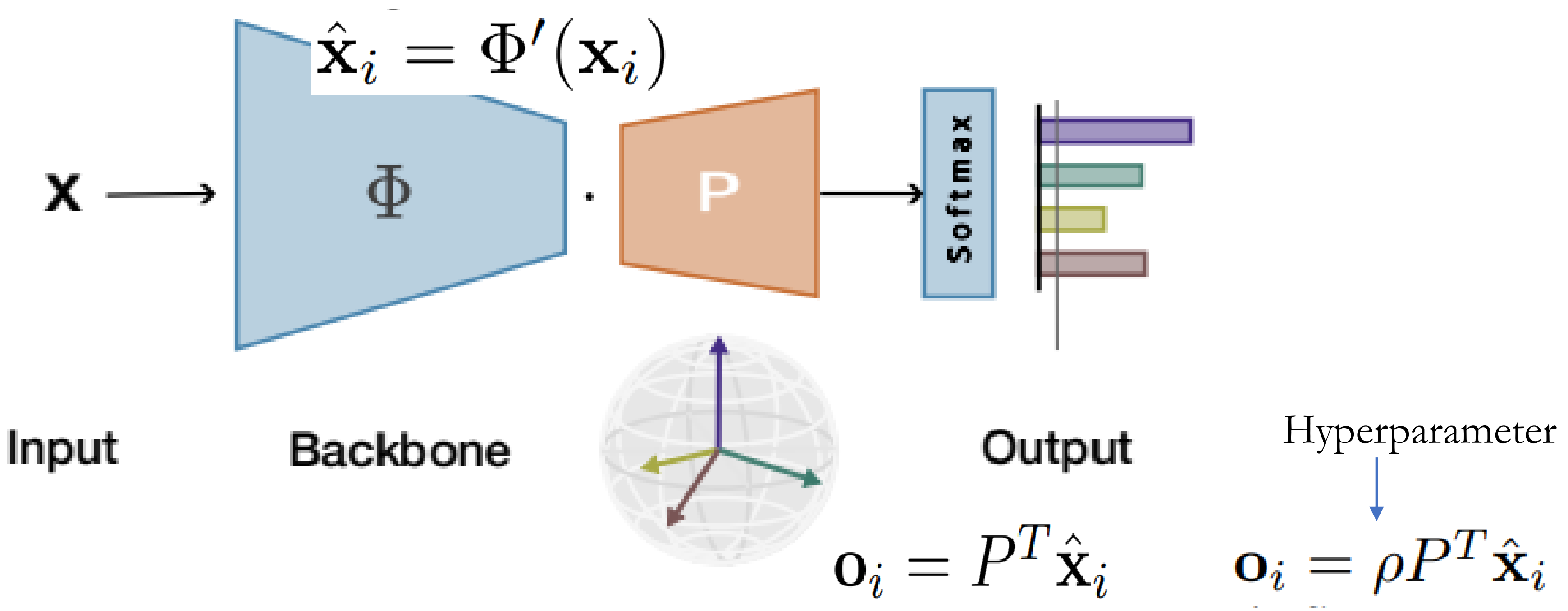
(a) Recursive update from 2 to 3 classes.

(b) Recursive update from 3 to 4 classes.

Theorem 1. *For any $k \geq 1$, P_k is a maximally separated matrix.*

Closed-Form Maximum Separation between Classes

- 단일 행렬 곱셈으로 network 출력 증가
- Maximum separation 위한 학습 불필요



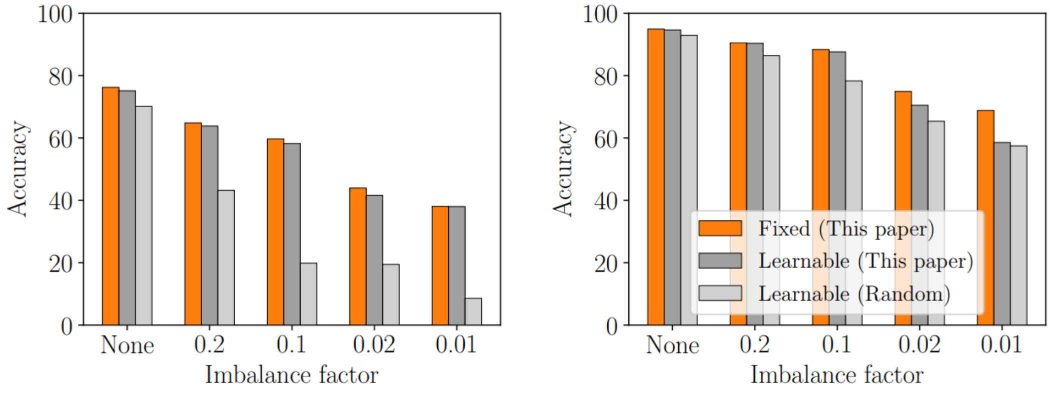
$$\hat{\mathbf{x}}_i = \Phi'(\mathbf{x}_i)$$

$$\mathbf{o}_i = P^T \hat{\mathbf{x}}_i \qquad \mathbf{o}_i = \rho P^T \hat{\mathbf{x}}_i$$

Experiments

- Classification and long-tailed recognition
  - Classification with maximum separation

	CIFAR-100					CIFAR-10				
	-	0.2	0.1	0.02	0.01	-	0.2	0.1	0.02	0.01
ConvNet	56.70	45.97	40.34	27.35	16.59	86.68	79.47	73.90	51.40	43.67
+ This paper	**57.05**	**46.59**	**40.44**	**28.27**	**18.40**	**86.76**	**79.63**	**75.88**	**55.25**	**48.05**
	+0.35	+0.62	+0.10	+0.92	+1.81	+0.08	+0.16	+1.98	+3.85	+4.38
ResNet-32	75.77	65.74	58.98	42.71	35.02	94.63	88.17	83.10	68.64	56.98
+ This paper	**76.54**	**66.01**	**60.54**	**45.12**	**38.85**	**95.09**	**91.42**	**88.16**	**77.02**	**69.70**
	+0.77	+0.27	+1.56	+2.41	+3.83	+0.46	+3.25	+5.06	+8.38	+12.72

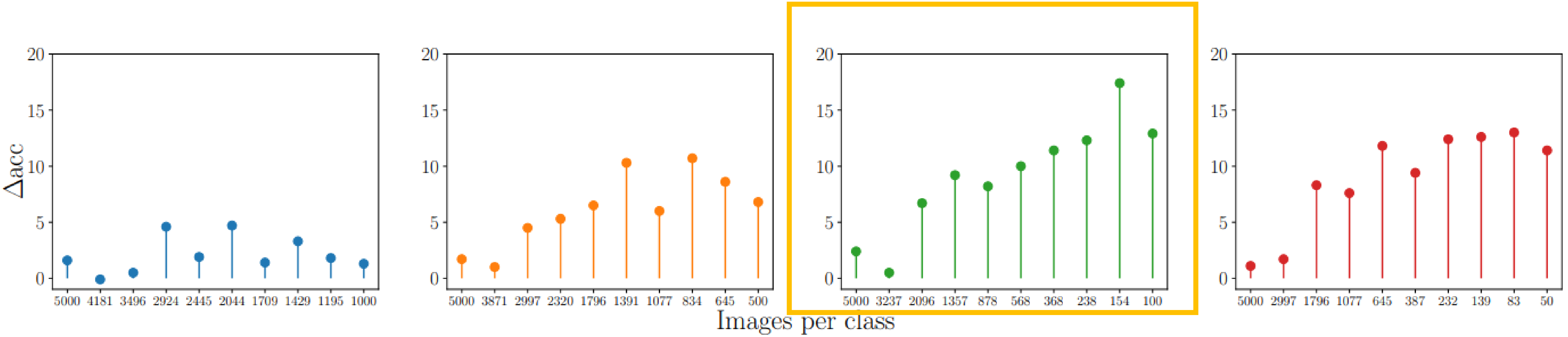


(a) CIFAR-100.
(b) CIFAR-10.

Accuracy 향상과 train sample frequency 사이 관계

Experiments

- Classification and long-tailed recognition
  - Should maximum separation be a fixed inductive bias?

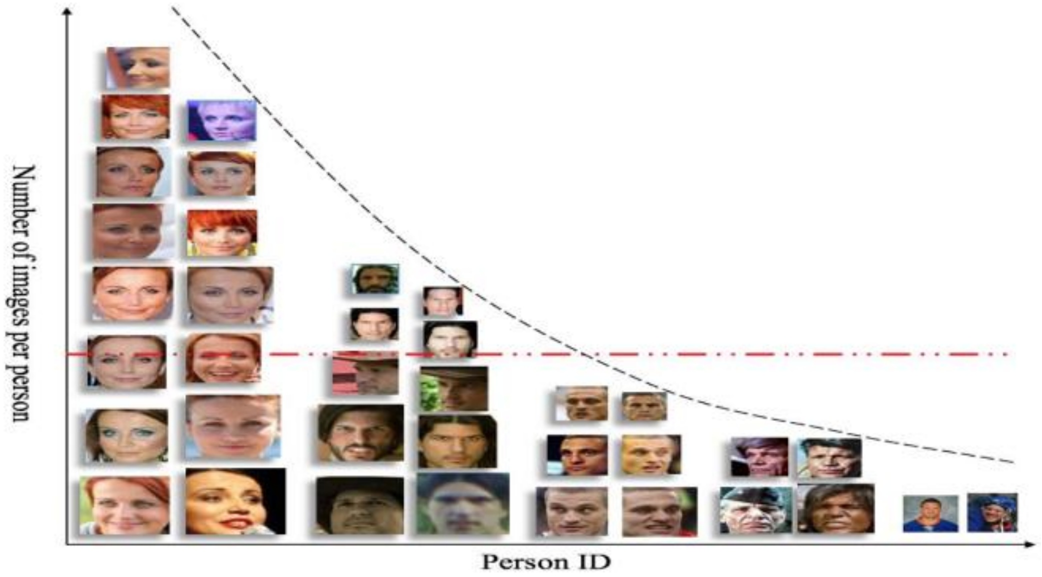


Learnable matrix

무작위 초기화

Fixed matrix

Experiments



- Classification and long-tailed recognition
  - Enriching long-tailed recognition approaches
    - LDAM : 기존의 cross-entropy loss 대신하는 label-distribution-aware margin loss를 사용한 방식
    - MiSLAS : label-aware smoothing과 shifted batch nomalization을 이용한 방식

	Imbalance factor 0.1	0.02	0.01
LDAM-SGD	55.05	43.85	39.87
+ This paper	**57.72**	**45.14**	**42.02**
	+2.67	+1.29	+2.20
LDAM-DRW	57.45	47.56	42.37
+ This paper	**58.37**	**48.02**	**43.19**
	+0.92	+0.46	+0.82

	Imbalance factor 0.1	0.02	0.01
MiSLAS (stage 1)	58.36	44.69	40.29
+ This paper	**59.63**	**45.65**	**40.56**
	+1.27	+0.96	+0.27
MiSLAS (stage 2)	61.93	52.53	48.00
+ This paper	**63.52**	**53.36**	**48.42**
	+1.59	+0.83	+0.42

Experiments

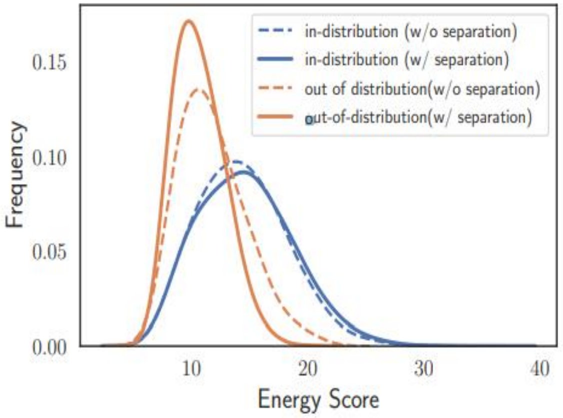
- Classification and long-tailed recognition
  - ImageNet experiments
    - ImageNet: large-scale dataset (1000개 class, 백만 개 넘는 데이터)

	Resnet-50				Resnet-152			
	top 1		top 5		top 1		top 5	
	Base	+ Ours	Base	+ Ours	Base	+ Ours	Base	+ Ours
Imagenet	73.2	**74.8**	92.4	**94.9**	77.9	**78.5**	94.3	**95.1**
Imagenet-LT	43.8	**47.3**	70.4	**73.6**	48.3	**49.7**	73.9	**74.8**

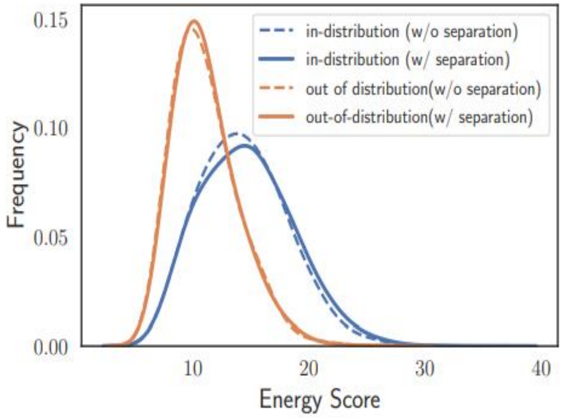
Experiments

- Out-of-distribution detection and open-set-recognition
  - Out-of-distribution with maximum separation
    - out-of-distribution : 학습 데이터의 분포와는 다른 분포를 갖는 데이터, 학습데이터에 포함되지 않은 class를 가진 데이터 의미

	Metric	w/o maximum separation			w/ maximum separation		
		FPR95 ↓	AUROC ↑	AUPR ↑	FPR95 ↓	AUROC ↑	AUPR ↑
SVHN	Softmax Score	84.00	71.40	92.87	**83.04**	**75.58**	**94.59**
	Energy Score	85.76	73.94	93.91	**78.86**	**85.42**	**96.92**
	Mahalanobis	44.02	90.48	97.83	**35.88**	**91.45**	**97.93**
Places 365	Softmax Score	**82.86**	73.46	93.14	83.08	**73.53**	**93.54**
	Energy Score	**80.87**	75.17	93.40	81.36	**76.01**	**93.64**
	Mahalanobis	**88.83**	67.87	90.71	89.16	**69.33**	**91.49**



(a) SVHN as OOD.



(b) Places365 as OOD.

Experiments

- Out-of-distribution detection and open-set-recognition
  - Open-set recognition with maximum separation
    - Open-set recognition : 모델이 학습할 때 없었던 클래스가 test시 나타나 학습 클래스 중 하나가 아닌 unknown 으로 분류가 가능한 알고리즘

	SVHN	CIFAR10	CIFAR + 10	CIFAR + 50
MSP+	95.94	90.10	94.48	93.58
+ This paper	**96.22**	**91.05**	**95.57**	**94.31**
	+0.28	+0.95	+1.09	+0.73
MLS	97.10	93.82	97.94	96.48
+ This paper	**97.58**	**95.30**	**98.33**	**96.74**
	+0.48	+1.48	+0.39	+0.26

Conclusions

- Deep network 에 class 최대로 분리하는 inductive bias 포함하도록 노력
  - Separation 은 optimization 문제가 아니고 closed-form 으로 해결 가능
- Fixed matrix 가 분류에서 효과적이고 광범위하게 적용가능하고 사용 편리
- Limitation
  - Supervised setting에 적용되며 class 벡터 간의 관계 정보를 요구하는 설정에는 일반화 불가

Thank you