21.09.10

Deal with Contrastive Learning

고은성

Korea University

Data **M**ining & **Q**uality **A**nalytics Lab.

KOREA UNIVERSITY Data Mining Quality Analytics

발표자 소개

❖ 고은성

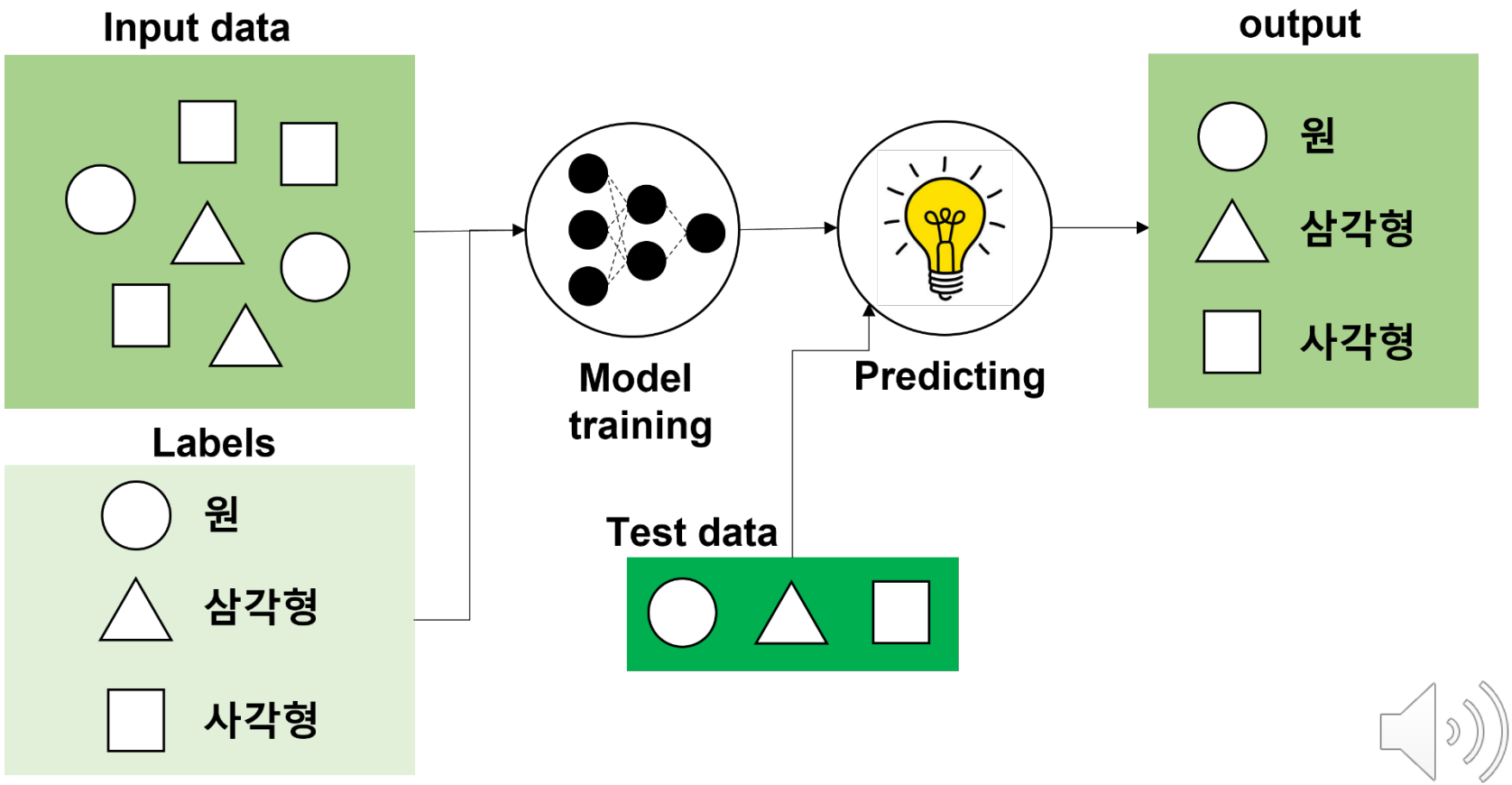
- 고려대학교 산업경영공학과
- Data Mining & Quality Analytics Lab. (김성범 교수님)
- M.S Student (2021.03~)
- Research Interest
  - Deep Learning
  - Self-Supervised Learning
- E-mail
  - esko328@korea.ac.kr

목차

Introduction

❖ Self-Supervised Learning 등장 배경

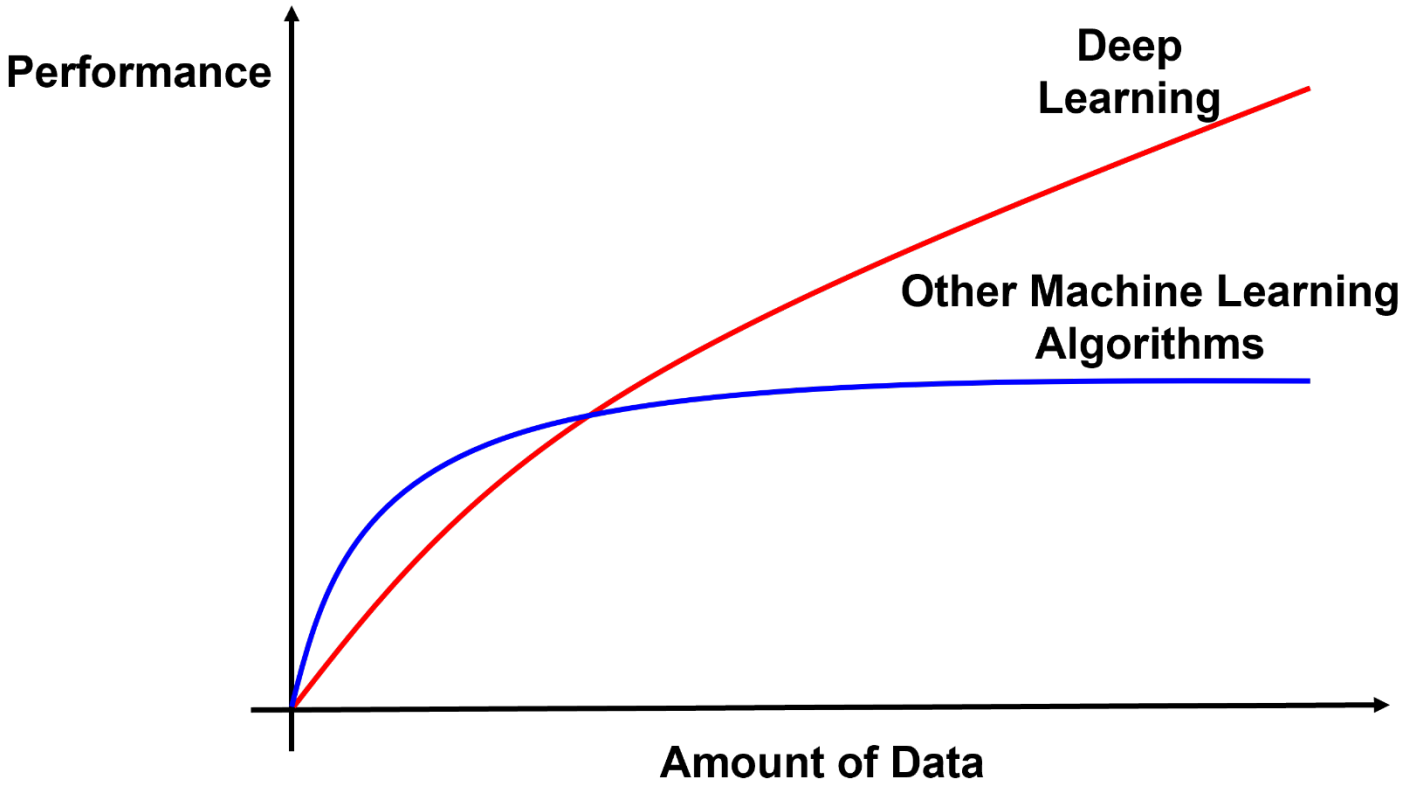
- Supervised Learning
  - ➢ 주어진 데이터와 라벨로부터 모델을 학습시켜 새로운 데이터의 라벨을 예측



Introduction

❖ Self-Supervised Learning 등장 배경

- 딥러닝 모델
  - ➢ 모델 사이즈가 증가함에 따라 정확도 향상 -> 대량의 데이터 필요

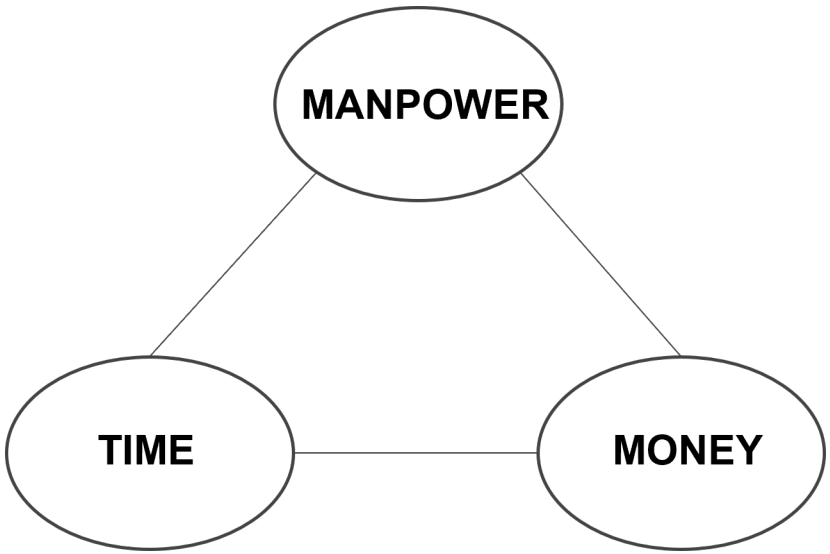


Introduction

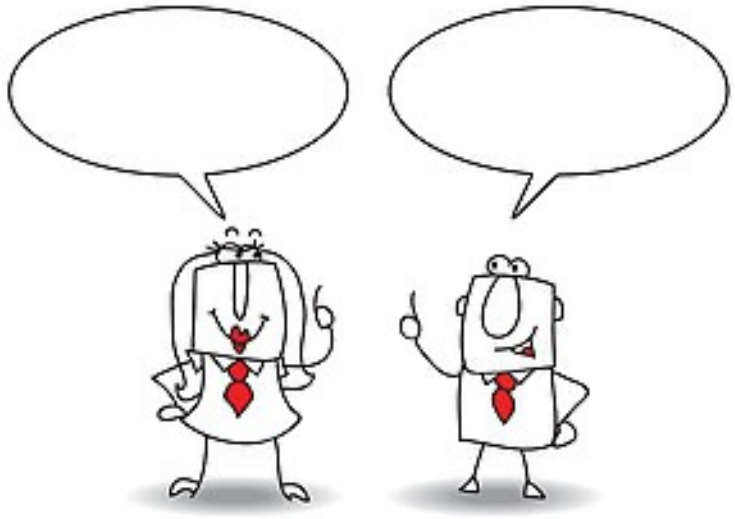
❖ Self-Supervised Learning 등장 배경

- Supervised Learning
  - ➢ 많은 양의 양질의 데이터가 있을 때 좋은 성능을 냄
  - ➢ ***레이블링된 데이터가 없다면 사용 불가능!***
  - ➢ ***레이블링에 많은 시간과 비용 소요, 작업자의 편향된 지식***

COST



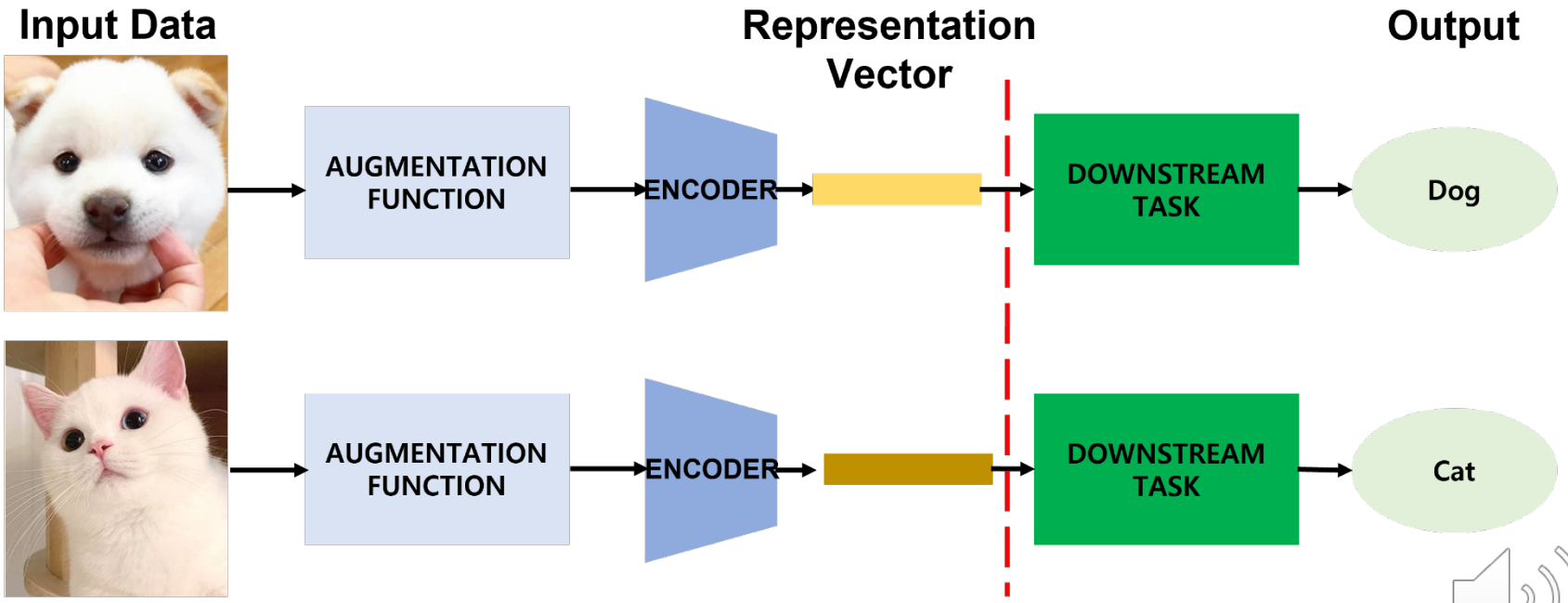
BIAS



Introduction

❖ Self-Supervised Learning 등장 배경

- Self-Supervised Learning
  - ➢ 라벨링 되어있지 않은 데이터를 기반으로 하여 특징을 학습하고, 이를 다운스트림 테스크에 활용
  - ➢ ***Representation learning model + Downstream task model***
  - ➢ ***데이터의 레이블이 소량인 경우에도 학습 성능을 극대화***

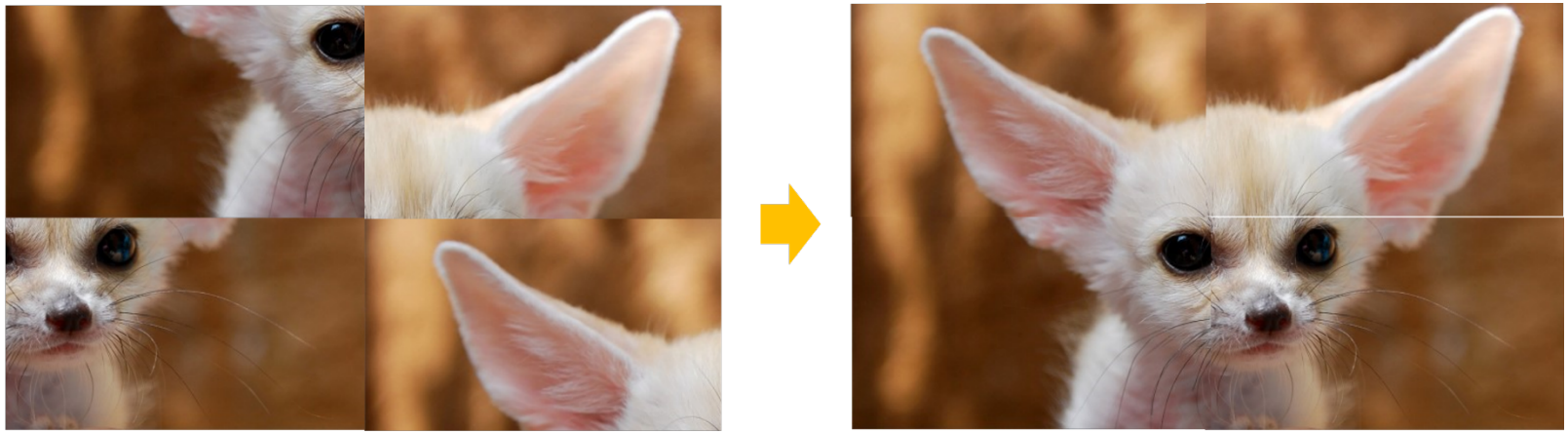


Introduction

❖ Self-Supervised Representation Learning Methods

- Pretext task
  - ***<u>특징 추출을 위해 사전에 문제를 정의하고 label을 직접 생성하여 모델을 학습</u>***
  - 구조 유추, 변환 예측, 재건, 시간적 관계 유추 task 등

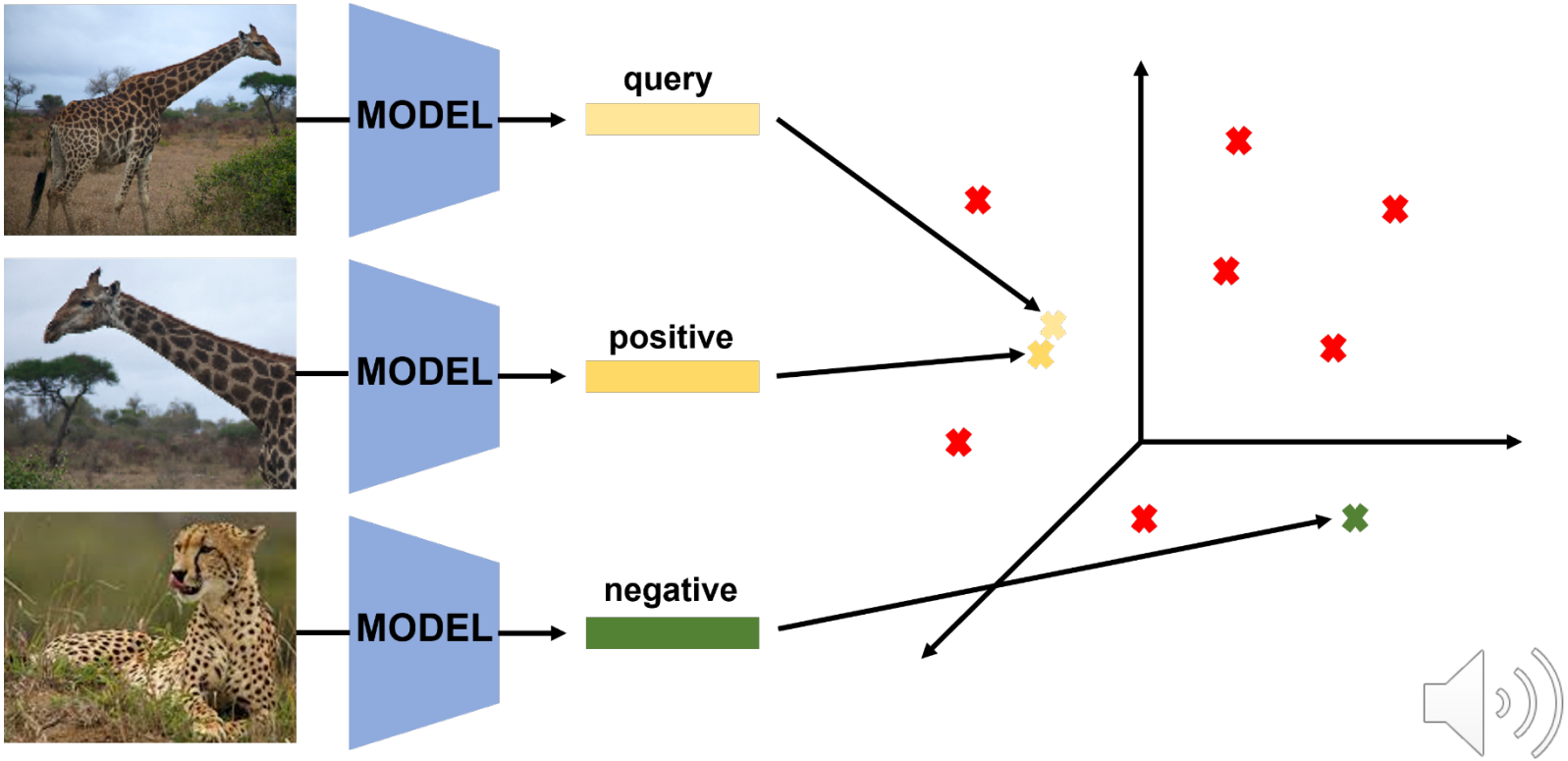
<순서 예측>



Introduction

❖ Self-Supervised Representation Learning Methods

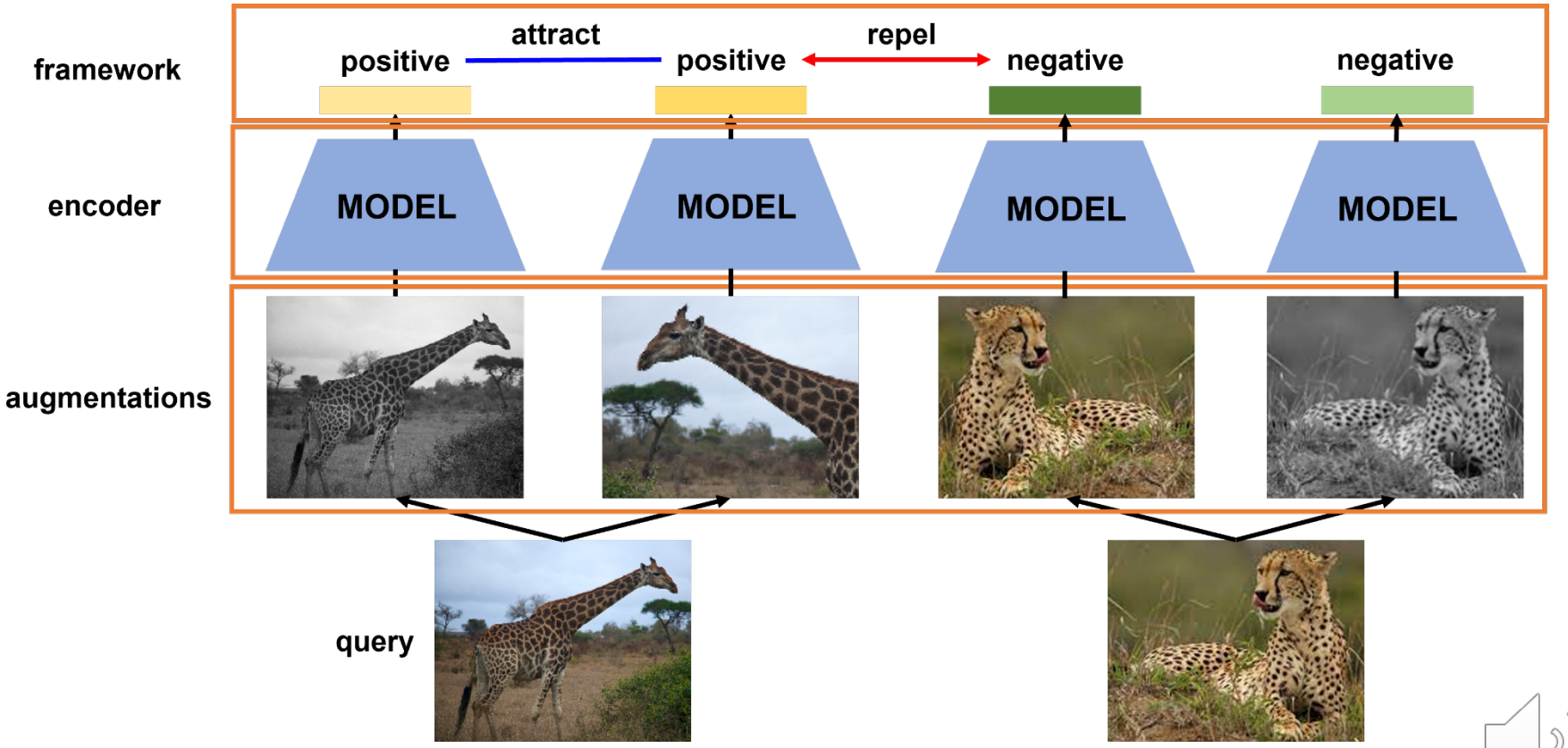
- Contrastive Learning
  - ***Representation space에서 positive pair는 가까이, negative pair는 멀리 있도록 학습하여 특징벡터를 추출***
  - ***Downstream task model의 성능으로 contrastive learning 성능을 확인***



Introduction

❖ Self-Supervised Representation Learning Methods

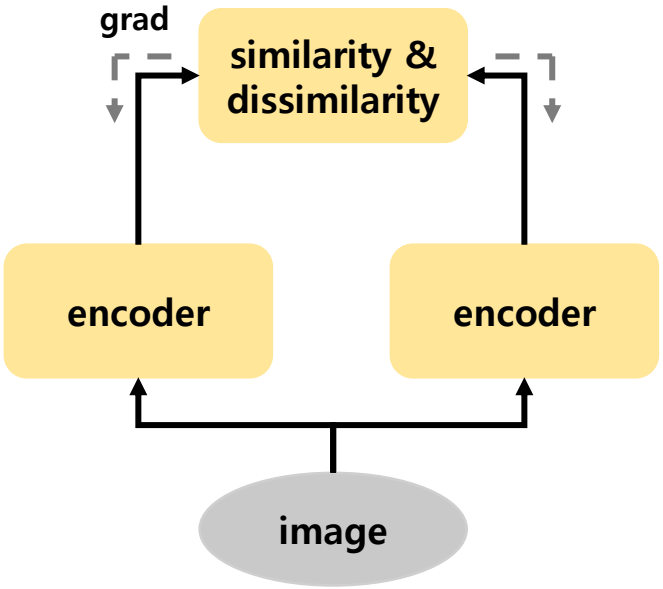
- Contrastive Learning 3요소
  - ***Data augmentation, Encoder, Framework (Positive & Negative Pairs, Loss Function)***



Methods

❖ Contrastive Learning Methods

- SimCLR



$$\textbf{Loss Function} \rightarrow NT-Xent$$

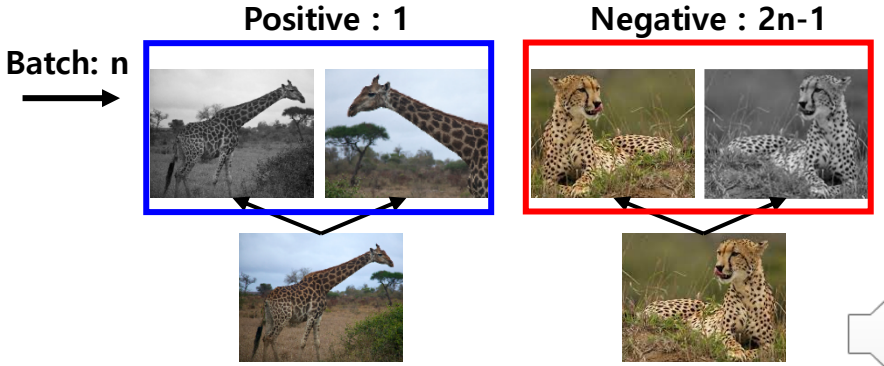
$$-\log \frac{\exp\left(\frac{sim(z_i, z_j)}{\tau}\right)}{\sum_{k=1(k\neq i)}^{2N} \exp\left(\frac{sim(z_i, z_k)}{\tau}\right)}$$

positive

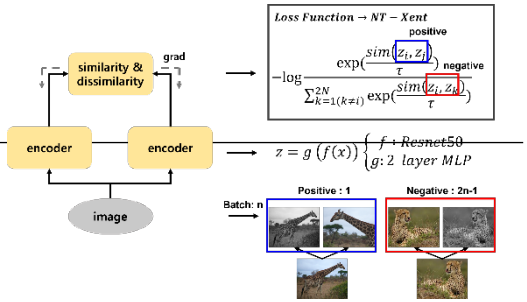
negative

$$z = g(f(x)) \quad \begin{cases} f : Resnet \\ g: 2\ layer\ MLP \end{cases}$$

실제 학습에 사용

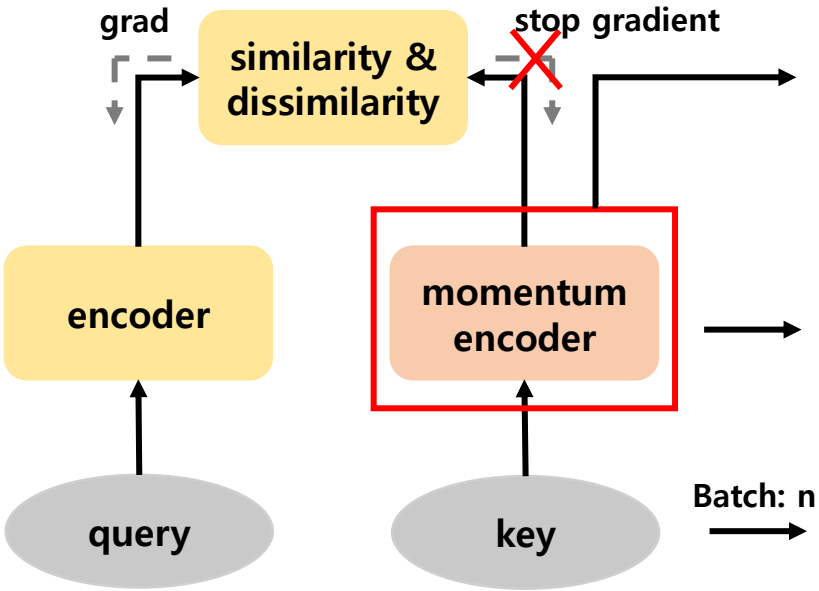


Methods



❖ Contrastive Learning Methods

- MoCo



momentum encoder : key encoder

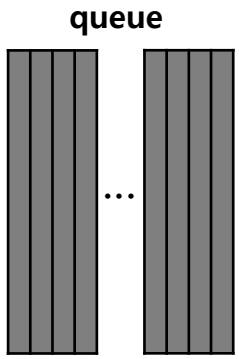
-일관성 유지가 중요!

-query encoder parameter의 변화에 따라 점진적으로 key encoder parmeter 변화

-$\theta_k = m\theta_k + (1-m)\theta_q$

-모든 negative sample에 대해 gradient 전파 어려움

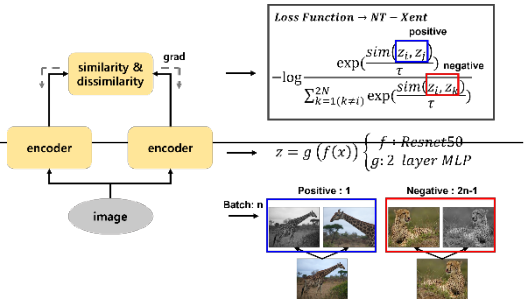
$$z = g(f(x)) \begin{cases} f : Resnet \\ g : 2\ layer\ MLP \end{cases}$$



batch size보다 큰 queue 구조의 dictionary에 negative key 미리 구성

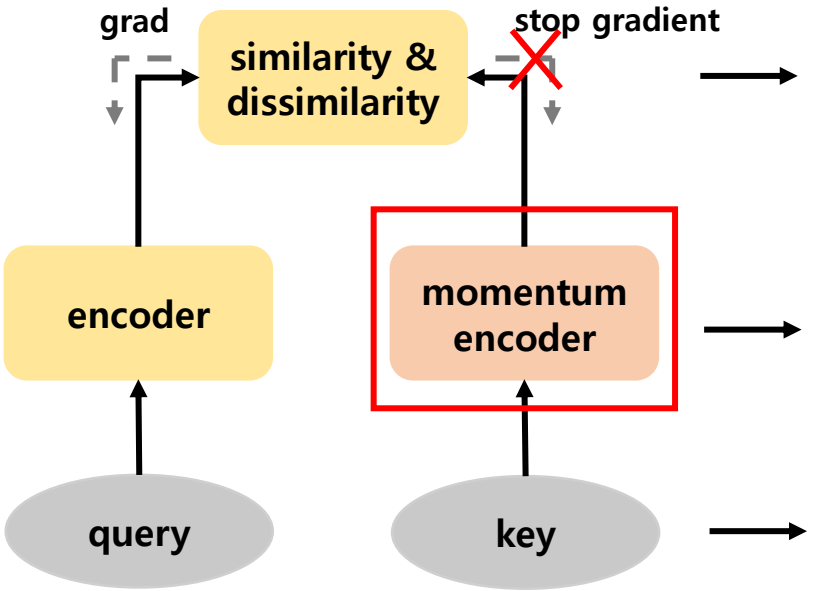
-> simCLR보다 batch size에 영향을 덜 받음

Methods



❖ Contrastive Learning Methods

- MoCo v2



SimCLR 장점 반영

-Cosine learning rate scheduling 방법 이용

$$z = g(f(x)) \begin{cases} f: Resnet \\ g: 2\ layer\ MLP \end{cases}$$

실제 학습에 사용

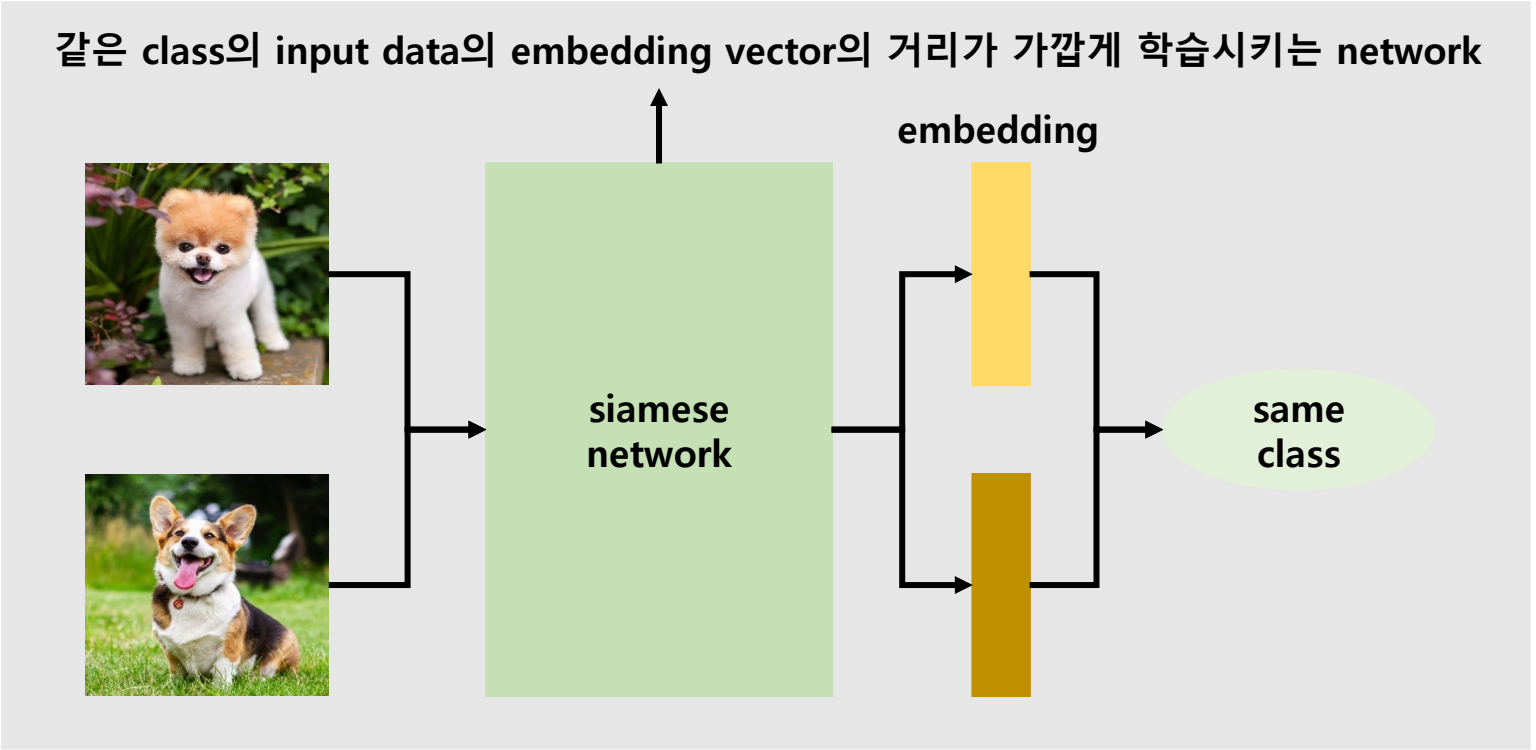
Data augmentation

-Gaussian blurring 추가

Methods

❖ Self-Supervised Representation Learning Methods

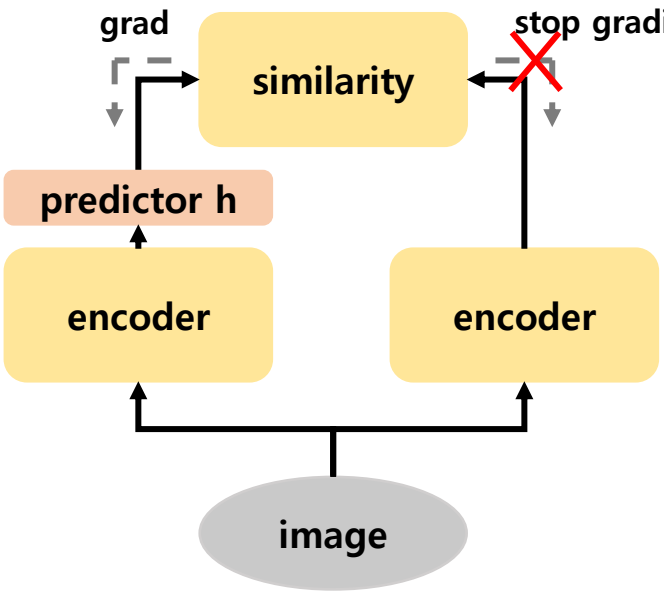
- Collapsing
  - ➢ ***Siamese network구조로 representation vector학습시킬 때, representation vector가 constant vector로 고정되는 현상***
  - ➢ SimCLR, MoCo : negative pair, momentum encoder로 collapsing 방지



Methods

❖ Contrastive Learning Methods

- Simsiam

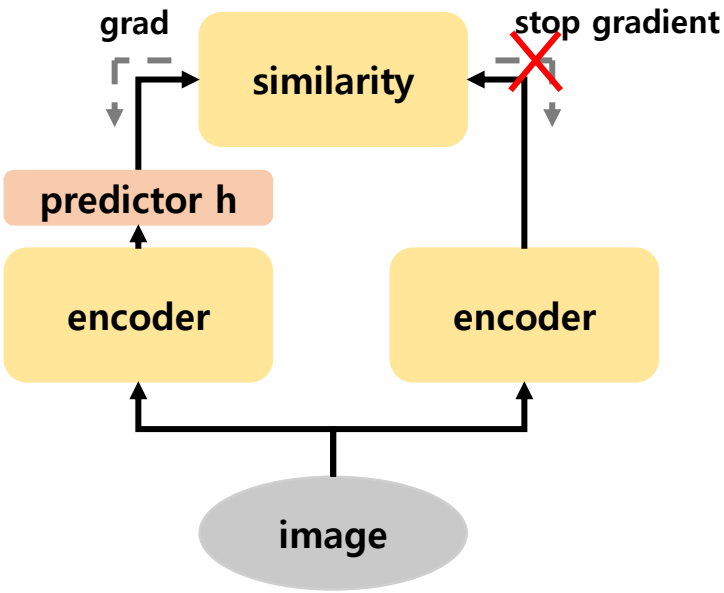


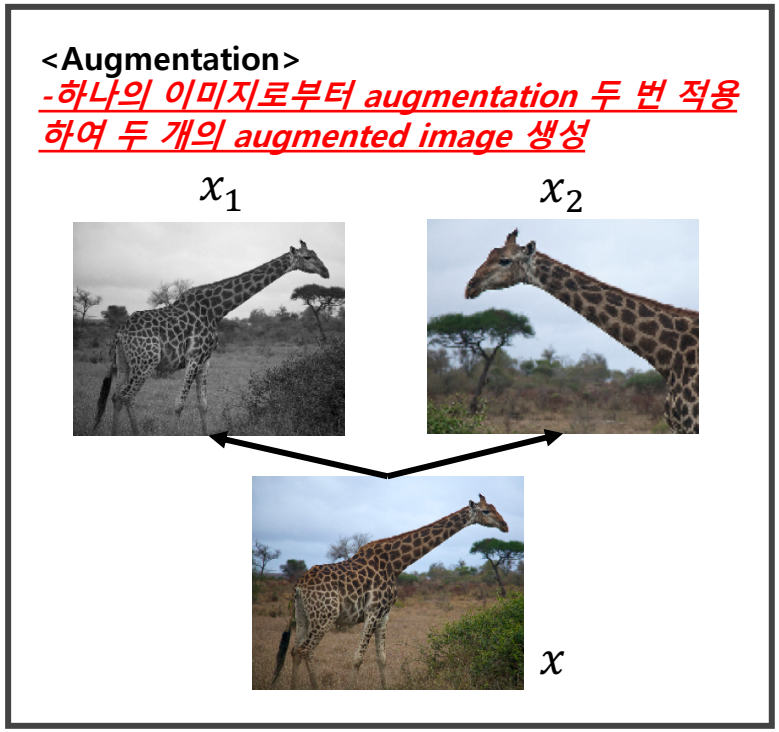
<Simple Siamese>
-negative pair, momentum encoder 없이
Stop gradient 방식을 통해 Siamese network의
Collapsing 해결

Methods

❖ Contrastive Learning Methods

- Simsiam

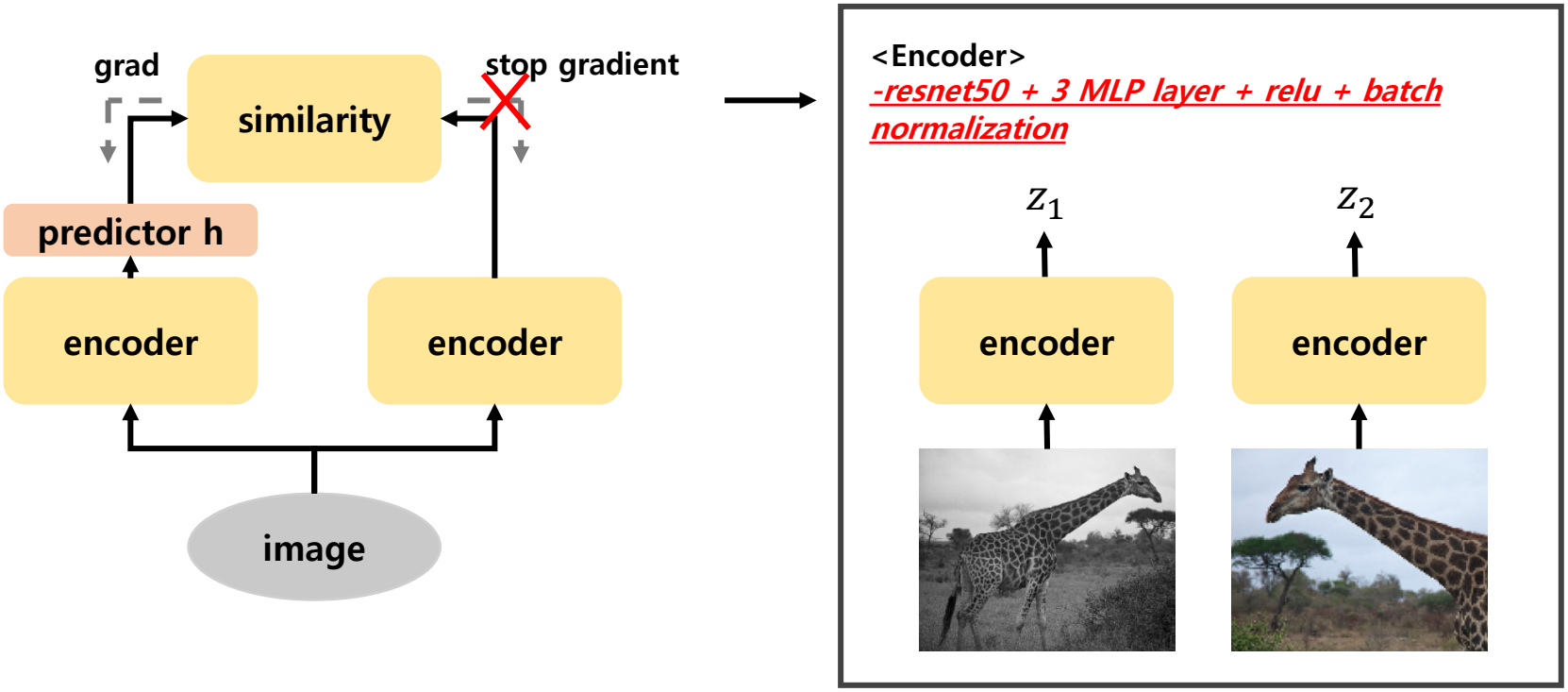




Methods

❖ Contrastive Learning Methods

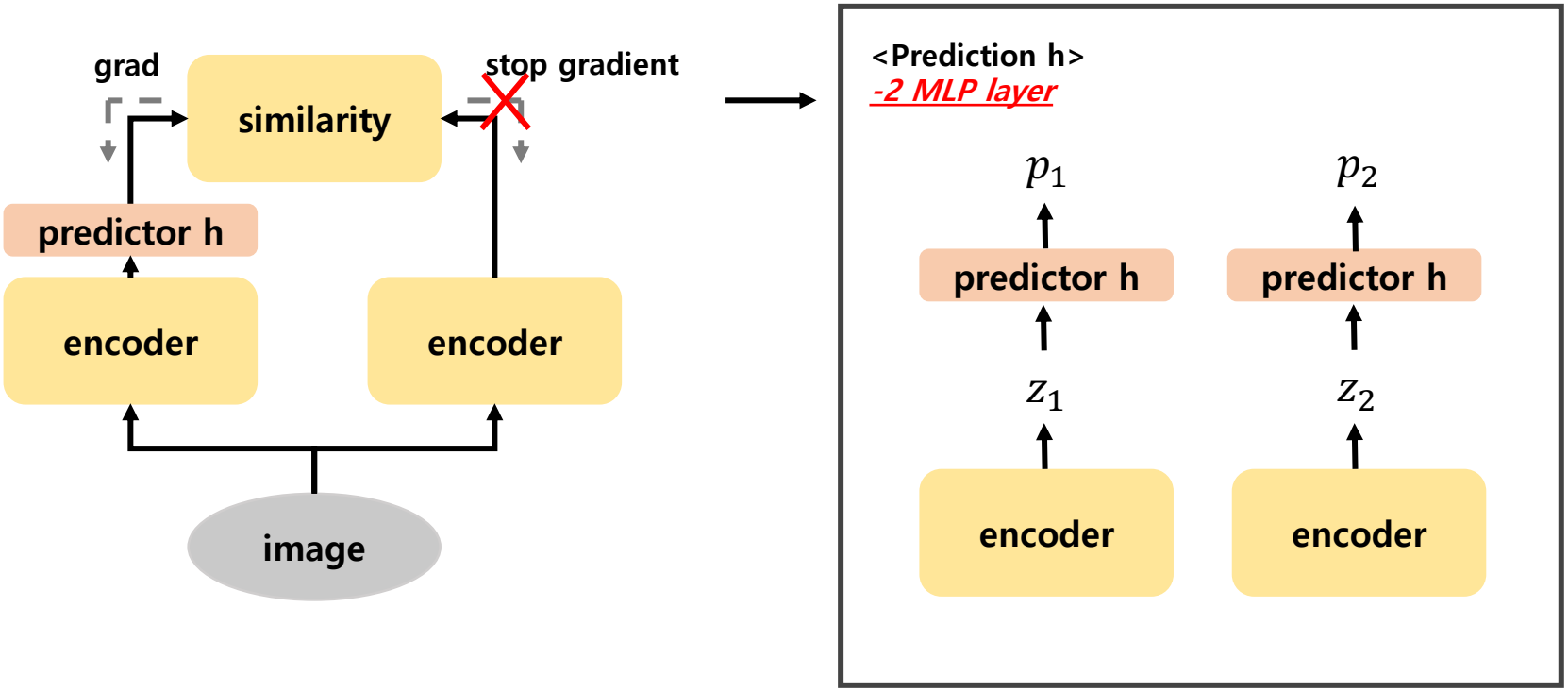
- Simsiam



Methods

❖ Contrastive Learning Methods

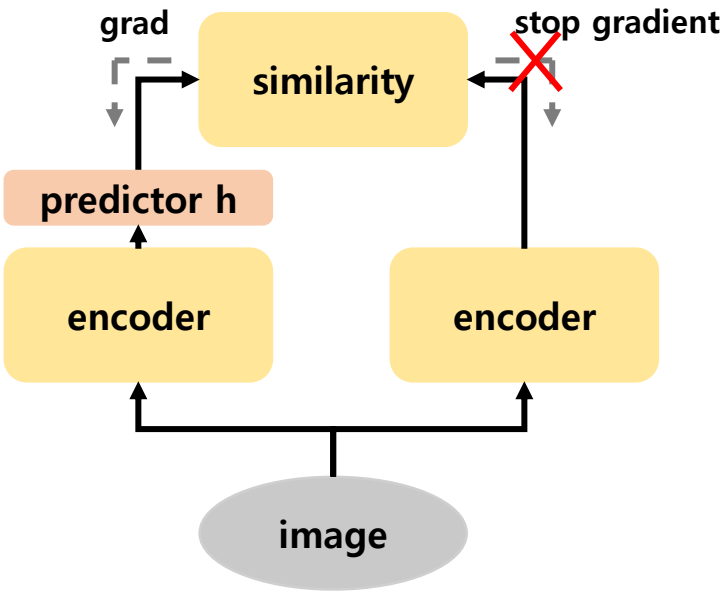
- Simsiam



Methods

❖ Contrastive Learning Methods

- Simsiam



<Negative cosine similarity>

$$D(p_1, z_2) = -\frac{p_1}{\|p_1\|_2} \cdot \frac{z_2}{\|z_2\|_2}$$

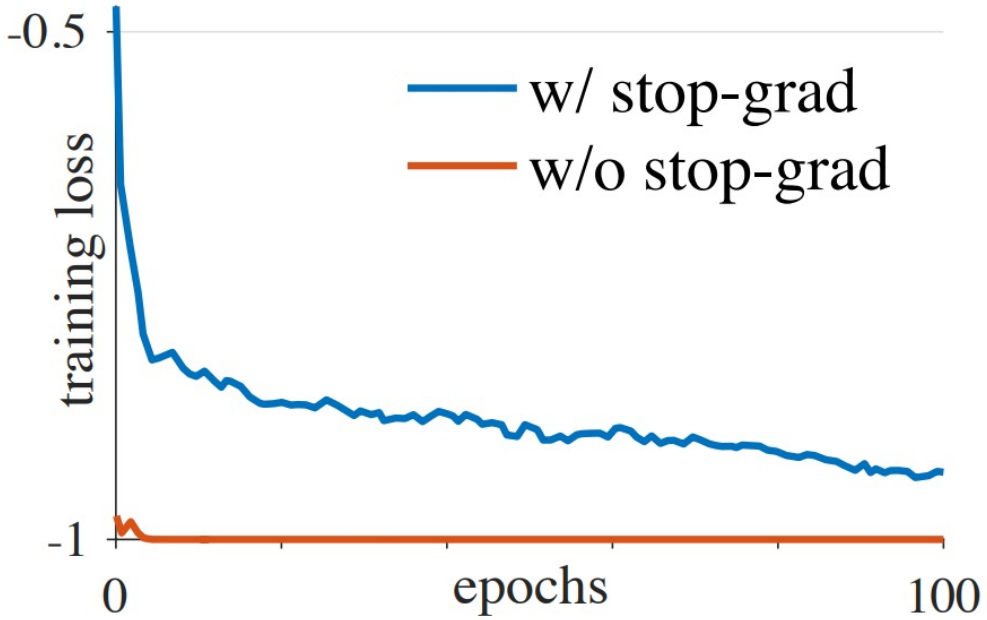
<Loss function>

$$L = D(p_1, stopgrad(z_2))/2 + D(p_2, stopgrad(z_1))/2$$

Methods

❖ Contrastive Learning Methods

- Simsiam
  - ***Stop-grad가 없이 representation 학습시키면 loss가 빠르게 -1로 수렴***
  - ***Stop-grad 없는 경우 accuracy 0.1%로 전혀 학습되지 않음***
  - ***Architecture design만으로 collapsing 방지할 수 없음***



	acc. (%)
w/ stop-grad	67.7±0.1
w/o stop-grad	0.1

Methods

❖ 기존 Contrastive Learning Methods 단점

- augmentation으로 유도된 inductive bias
- downstream task 진행 시, 사용된 augmentation들에 대한 불변성 가정이 필요
  - ➢ ***augmentation 관련 정보 손실***
  - ➢ 꽃 분류를 위해 색 정보가 필요!

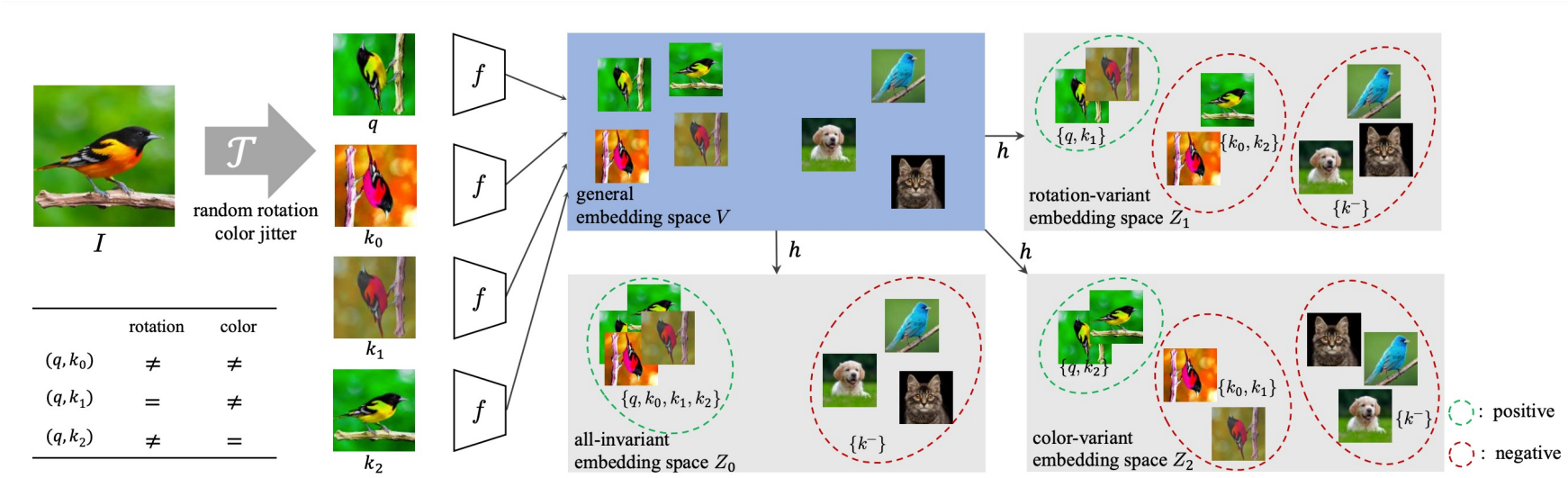
<엉겅퀴>

<방가지똥>

Methods

❖ Contrastive Learning Methods

- *LOOC* (Leave-One-Out Contrastive Learning)
  - ➢ 기존 contrastive learning과 달리 augmentation으로 인한 변동 요인도 포함하는 representation 학습

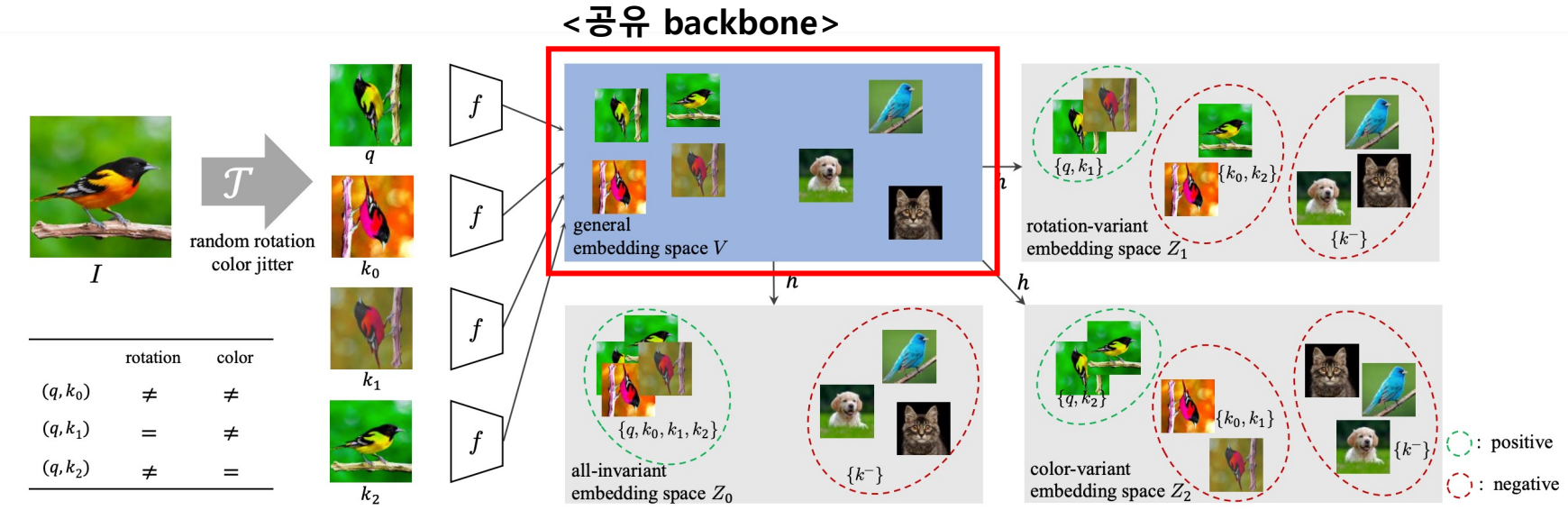


	rotation	color
(q, k_0)	$\neq$	$\neq$
(q, k_1)	$=$	$\neq$
(q, k_2)	$\neq$	$=$

Methods

❖ Contrastive Learning Methods

- *LOOC* (Leave-One-Out Contrastive Learning)
  - ➢ 공유 backbone을 가지는 multi-head architecture를 사용해 구현
  - ➢ 특정 augmentation에 민감하고 나머지 augmentation들에는 그렇지 않은 부분공간들을 건설

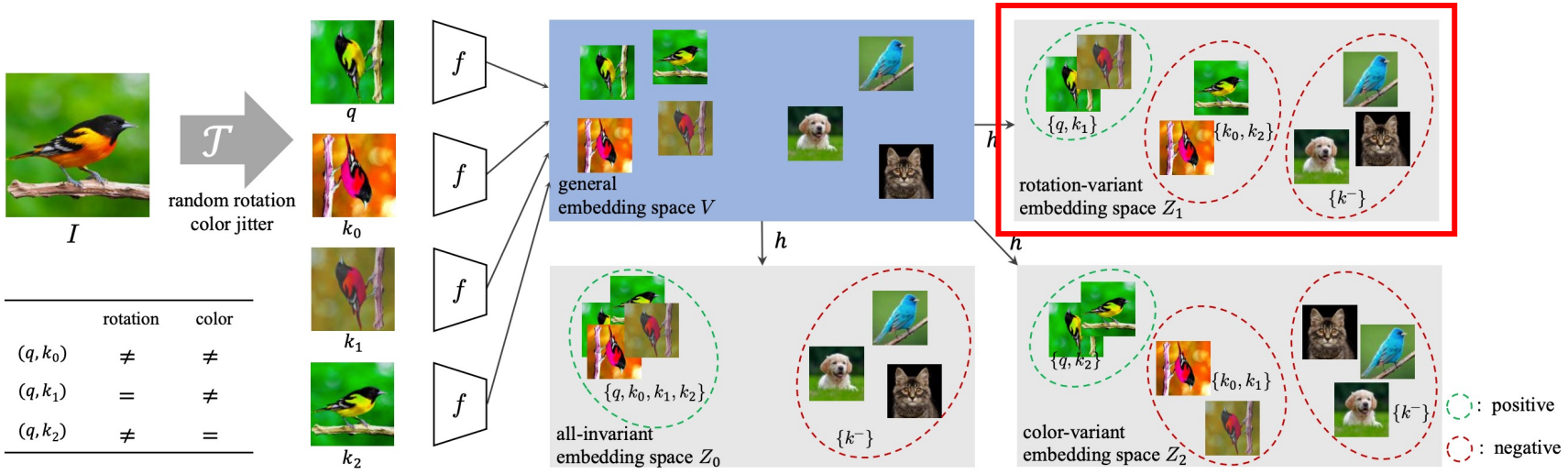


Methods

❖ Contrastive Learning Methods

- *LOOC* (Leave-One-Out Contrastive Learning)
  - ➢ 공유 backbone을 가지는 multi-head architecture를 사용해 구현
  - ➢ 특정 augmentation에 민감하고 나머지 augmentation들에는 그렇지 않은 부분공간들을 건설

같은 data sample인데,
rotation augmentation이 다르게 적용되면
negative sample로 정의한 부분 공간



	rotation	color
(q, k_0)	$\neq$	$\neq$
(q, k_1)	$=$	$\neq$
(q, k_2)	$\neq$	$=$

Methods

❖ Contrastive Learning Methods

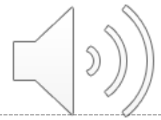
- *LOOC* (Leave-One-Out Contrastive Learning)
  - ➢ Loss Function

k^+(동일 reference image augmentation)중, i번째 augmentation operator의 parameter가 같은 positive pair similarity

$$\mathcal{L}_q = -\frac{1}{n+1}\left(\log\frac{E^+_{0,0}}{E^+_{0,0}+\sum_{k-}E^-_{0,0}} + \sum_{i=1}^{n}\log\frac{E^+_{i,i}}{\sum_{j=0}^{n}E^+_{i,j}+\sum_{k-}E^-_{i,i}}\right)$$

k^+(동일 reference image augmentation)중, i번째 augmentation operator의 parameter가 다른 negative pair similarity

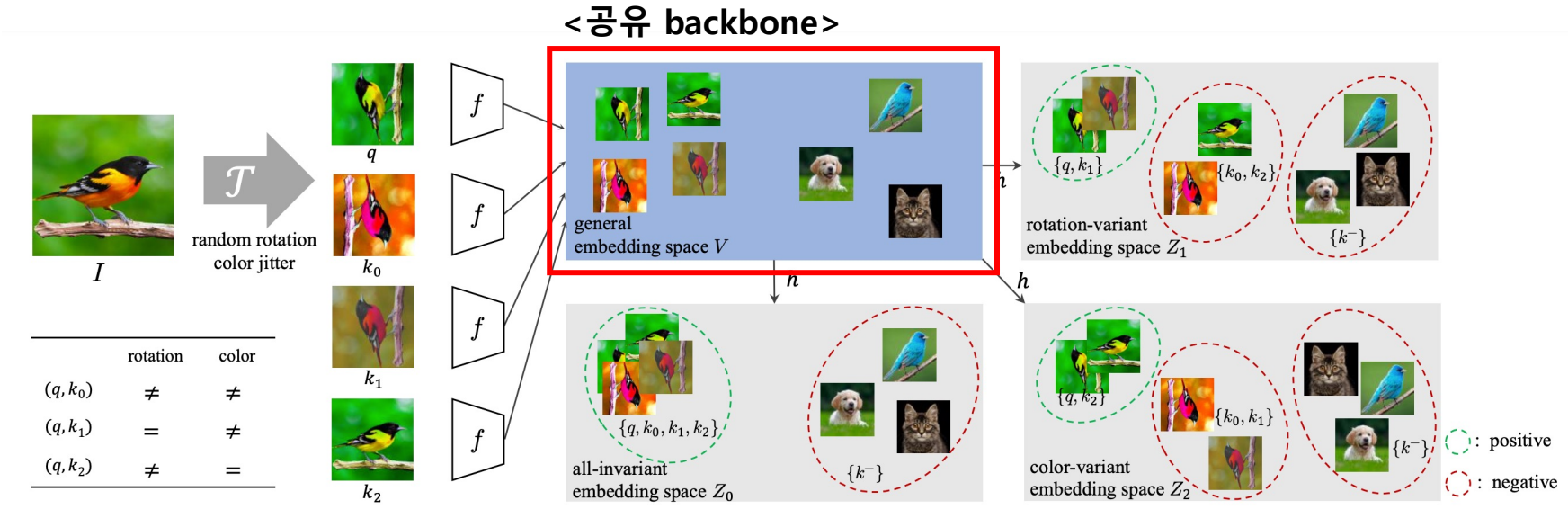
k^-(다른 reference image augmentation)중, i번째 augmentation operator의 parameter가 같은 negative pair similarity



Methods

❖ Contrastive Learning Methods

- *LOOC* (Leave-One-Out Contrastive Learning)
  - ➢ General representation v가 augmentation으로 인해 불변하는 요소와 변하는 요소를 모두 담음
  - ➢ General representation v를 downstream task에 활용
  - ➢ 부분 공간의 representation을 concatenate하여 사용 -> looc++



	rotation	color
(q, k_0)	$\neq$	$\neq$
(q, k_1)	$=$	$\neq$
(q, k_2)	$\neq$	$=$

Conclusion

❖ Conclusion

- *Simsiam & MoCo v2*

method	batch size	negative pairs	momentum encoder	100 ep	200 ep	400 ep	800 ep
SimCLR (repro.+)	4096	✓		66.5	68.3	69.8	70.4
MoCo v2 (repro.+)	**256**	✓	✓	67.4	69.9	71.0	72.2
BYOL (repro.)	4096		✓	66.5	**70.6**	**73.2**	**74.3**
SwAV (repro.+)	4096			66.5	69.1	70.7	71.8
SimSiam	**256**			**68.1**	70.0	70.8	71.3

	VOC 07 detection			VOC 07+12 detection			COCO detection			COCO instance seg.		
pre-train	AP_{50}	AP	AP_{75}	AP_{50}	AP	AP_{75}	AP_{50}	AP	AP_{75}	AP^{mask}_{50}	AP^{mask}	AP^{mask}_{75}
scratch	35.9	16.8	13.0	60.2	33.8	33.1	44.0	26.4	27.8	46.9	29.3	30.8
ImageNet supervised	74.4	42.4	42.7	81.3	53.5	58.8	58.2	38.2	41.2	54.7	33.3	35.2
SimCLR (repro.+)	75.9	46.8	50.1	81.8	55.5	61.4	57.7	37.9	40.9	54.6	33.3	35.3
MoCo v2 (repro.+)	**77.1**	**48.5**	**52.5**	**82.3**	**57.0**	**63.3**	**58.8**	**39.2**	**42.5**	**55.5**	**34.3**	**36.6**
BYOL (repro.)	**77.1**	47.0	49.9	81.4	55.3	61.1	57.8	37.9	40.9	54.3	33.2	35.0
SwAV (repro.+)	75.5	46.5	49.6	81.5	55.4	61.4	57.6	37.6	40.3	54.2	33.1	35.1
SimSiam, base	75.5	47.0	50.2	**82.0**	56.4	62.8	57.5	37.9	40.9	54.2	33.2	35.2
SimSiam, optimal	**77.3**	**48.5**	**52.5**	**82.4**	**57.0**	**63.7**	**59.3**	**39.2**	**42.1**	**56.0**	**34.4**	**36.7**

Conclusion

❖ Conclusion

- *LOOC* (Leave-One-Out Contrastive Learning)

Table 1: **Classification accuracy on 4-class rotation and IN-100** under linear evaluation protocol. Adding rotation augmentation into baseline MoCo significantly reduces its capacity to classify rotation angles while downgrades its performance on IN-100. In contrast, our method better leverages the information gain of the new augmentation.

회전도에 대한 4범주 분류

model	Rotation Acc.	IN-100 top-1	IN-100 top-5
Supervised	72.3	83.7	95.7
MoCo	61.1	81.0	95.2
MoCo + Rotation	43.3	79.4	94.1
MoCo + Rotation (same for q and k)	45.5	78.1	94.3
LooC + Rotation [ours]	65.2	80.2	95.5

-> 반영된 augmentation관련 정보가 기존 contrastive learning의 경우 손실되는 것을 확인

-> 반면 LooC는 augmentation에 따라 변하는 정보도 포함하고 있는 것을 확인

Conclusion

❖ Conclusion

- *LOOC* (Leave-One-Out Contrastive Learning)

Table 3: **Evaluation on datasets of real-world corruptions.** Rotation augmentation is beneficial for ON-13, and texture augmentation if beneficial for IN-C-100.

model	Aug.		ON-13		IN-C-100 (top-1)						IN-100	
	Rot.	Tex.	top-1	top-5	Noise	Blur	Weather	Digital	All	$d \geq 3$	top-1	top-5
Supervised			30.9	54.8	28.4	47.1	44.9	58.5	47.2	36.5	83.7	95.7
MoCo			29.2	54.2	37.9	38.5	47.7	60.1	48.2	37.2	81.0	95.2
LooC	✓		34.2	59.6	31.3	33.1	42.4	54.9	42.7	31.8	80.2	95.5
		✓	30.1	54.1	42.4	39.6	54.0	61.9	51.3	41.9	81.0	94.7
	✓	✓	33.3	59.2	37.0	35.2	50.2	56.9	46.5	37.2	79.4	94.3
LooC++	✓	✓	32.6	57.3	38.3	37.6	52.0	60.0	48.8	38.9	82.1	95.1

-> 논문에서 제안하는 방법론이 기존 방법론보다 noise에 대한 robustness가 좋음을 확인

Conclusion

❖ Conclusion

- *LOOC* (Leave-One-Out Contrastive Learning)

Table 5: **Comparisons of concatenating features from different embedding spaces in LooC++** jointly trained on color, rotation and texture augmentations. Different downstream tasks show non-identical preferences for augmentation-dependent or invariant representations.

Model	Variance Head			IN-100		iNat-1k		Flowers-102		IN-C-100
	Col.	Rot.	Tex.	top-1	top-5	top-1	top-5	5-shot	10-shot	all-top-1
LooC++				78.5	94.3	38.5	64.7	68.6 (± 0.6)	77.6 (± 0.1)	48.0
	✓			79.7	94.4	42.9	68.7	69.1 (± 0.7)	79.5 (± 0.2)	47.1
		✓		81.5	94.9	41.4	67.4	70.5 (± 0.6)	80.0 (± 0.2)	52.6
			✓	80.3	94.9	43.0	68.6	70.4 (± 0.5)	80.5 (± 0.2)	44.1
	✓	✓	✓	82.2	95.3	45.9	71.4	71.0 (± 0.7)	81.9 (± 0.3)	48.0

-> LooC++에서 ablation실험을 통해 모든 부분 공간의 head를 concatenate 시킨 feature를 downstream task에 사용했을 때 가장 성능이 좋음을 확인

Conclusion

❖ Summary

- 라벨링 문제로 Self-Supervised Learning의 장점이 부각되면서 많은 연구가 이루어지고 있음.
- Contrastive learning의 다양한 Method가 제안되고 있음. (SimCLR, MoCo(v1&v2), Simsiam, Looc)
- Downstream task에 맞는 특징 추출 방법을 사용하는 것이 앞으로 더욱 중요하게 연구되어야 할 것.

감사합니다.