Augmentation and Generalization in Contrastive Self-Supervised Learning

Towards the Generalization of Contrastive Self-Supervised Learning

Weiran Huang[1*] Mingyang Yi[2,3*] Xuyang Zhao[4*]

[1]Huawei Noah's Ark Lab
[2]University of Chinese Academy of Sciences
[3]Academy of Mathematics and Systems Science, Chinese Academy of Sciences
[4]Peking University

ICLR2023

Background

Contrastive Learning

- **Alignment**: Similar samples have similar features.
- **Uniformity**: Preserve maximal information.

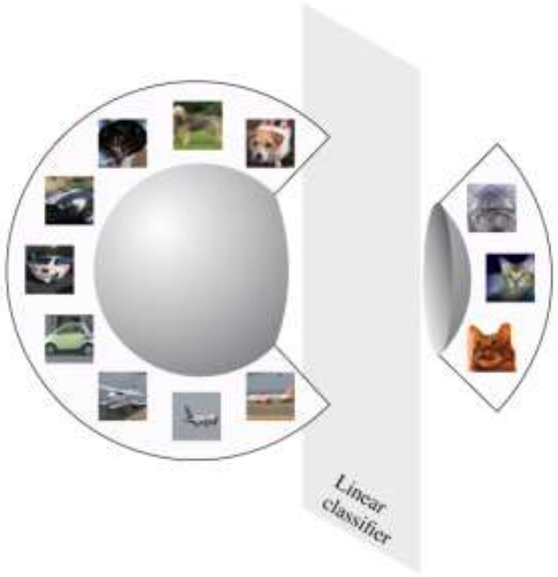
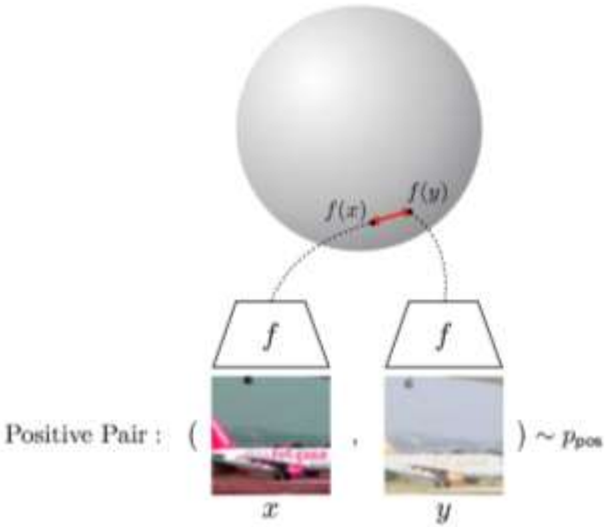


Figure 2: **Hypersphere:** When classes are well-clustered (forming spherical caps), they are linearly separable. The same does not hold for Euclidean spaces.



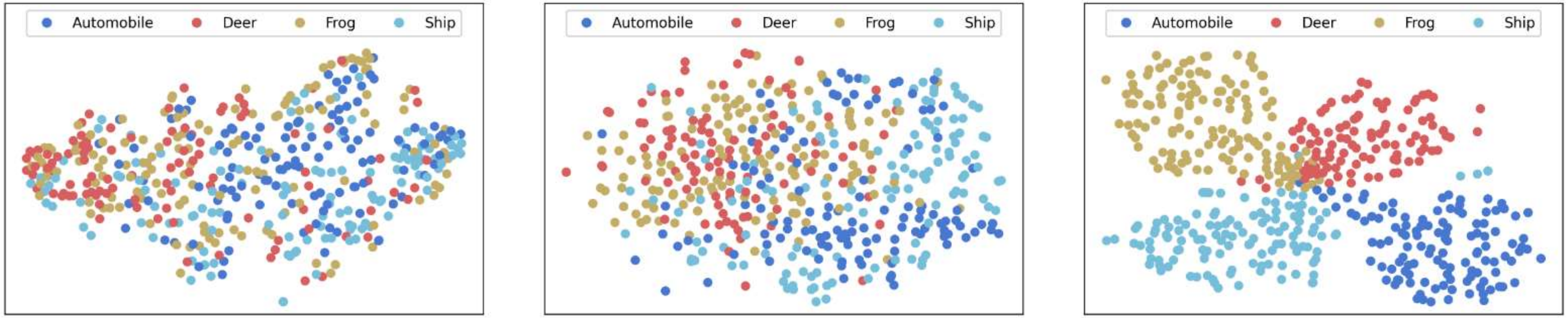
Alignment: Similar samples have similar features.
(Figure inspired by Tian et al. (2019).)

Uniformity: Preserve maximal information.

Background

The generalization of contrastive SSL is related to three key factors

- **Alignment** of positive samples
- **Divergence** of class centers
- **Concentration** of augmented data



(a) Initial (b) Only color distortion (c) Multiple transformations

Figure 1: SimCLR's embedding space with different richnesses of data augmentations on CIFAR-10.

Background

The generalization of contrastive SSL is related to three key factors

- **Alignment** of positive samples
- **Divergence** of class centers
- **Concentration** of augmented data

The first two factors are properties of learned representations.
The third one is determined by pre-defined data augmentation.

Experiments

Table 1: Downstream performance under different richness of augmentations.

Dataset	Transformations (a)	(b)	(c)	(d)	(e)	Accuracy SimCLR	Barlow Twins	MoCo	SimSiam
CIFAR-10	✓	✓	✓	✓	✓	**89.76 ± 0.12**	**86.91 ± 0.09**	**90.12 ± 0.12**	**90.59 ± 0.11**
	✓	✓	✓	✓		88.48 ± 0.22	85.38 ± 0.37	89.69 ± 0.11	89.34 ± 0.09
	✓	✓	✓			83.50 ± 0.14	82.00 ± 0.59	86.78 ± 0.07	85.38 ± 0.09
	✓	✓				63.23 ± 0.05	67.83 ± 0.94	75.12 ± 0.28	63.27 ± 0.30
	✓					62.74 ± 0.18	67.77 ± 0.69	74.94 ± 0.22	61.47 ± 0.74
CIFAR-100	✓	✓	✓	✓	✓	**57.74 ± 0.12**	**57.99 ± 0.29**	**64.19 ± 0.14**	**63.48 ± 0.16**
	✓	✓	✓	✓		55.43 ± 0.10	55.22 ± 0.25	62.50 ± 0.28	60.31 ± 0.41
	✓	✓	✓			45.10 ± 0.25	50.40 ± 0.64	57.04 ± 0.21	51.42 ± 0.14
	✓	✓				28.01 ± 0.18	34.11 ± 0.59	40.18 ± 0.04	26.26 ± 0.30
	✓					27.95 ± 0.09	34.05 ± 1.13	39.63 ± 0.31	25.90 ± 0.83

Table 2: Downstream performance under different strength of augmentations.

Dataset	Color Distortion Strength	Accuracy SimCLR	Barlow Twins	MoCo	SimSiam
CIFAR-10	1	**82.75 ± 0.24**	**82.58 ± 0.25**	**86.68 ± 0.05**	**82.50 ± 1.05**
	1/2	78.76 ± 0.18	81.88 ± 0.25	84.30 ± 0.14	81.80 ± 0.15
	1/4	76.37 ± 0.11	79.64 ± 0.34	82.76 ± 0.09	78.80 ± 0.17
	1/8	74.23 ± 0.16	77.96 ± 0.16	81.20 ± 0.12	76.09 ± 0.50
CIFAR-100	1	**46.67 ± 0.42**	**50.39 ± 1.09**	**58.50 ± 0.51**	**49.94 ± 2.01**
	1/2	40.21 ± 0.05	48.76 ± 0.25	55.08 ± 0.09	46.27 ± 0.46
	1/4	36.67 ± 0.08	46.22 ± 0.71	52.09 ± 0.18	42.02 ± 0.34
	1/8	34.75 ± 0.20	44.72 ± 0.26	49.43 ± 0.16	36.26 ± 0.34

(a) random cropping
(b) random Gaussian blur
(c) color dropping
(d) color distortion
(e) random horizontal flipping

ARCL: ENHANCING CONTRASTIVE LEARNING WITH AUGMENTATION-ROBUST REPRESENTATIONS

Xuyang Zhao[1,2*] **Tianqi Du**[1,3*] **Yisen Wang**[3,4] **Jun Yao**[5] **Weiran Huang**[2†]

[1] School of Mathematical Sciences, Peking University
[2] Qing Yuan Research Institute, Shanghai Jiao Tong University
[3] National Key Lab of General Artificial Intelligence, School of Intelligence Science and Technology, Peking University
[4] Institute for Artificial Intelligence, Peking University [5] Huawei Noah's Ark Lab

ICLR2023

Formulation

Augmentation Set $\mathcal{A}$
Transformation-induced domain D_A
Training domain set $\{D_A\}_{A\in\mathcal{A}}$

The goal of contrastive learning is equivalent to align different D_A

We naturally expect that the features it learn are domain invariant
The learned representation is not domain-invariant

Formulation

The population loss of InfoNCE (Chen et al., 2020a; He et al., 2020) is well known as:

$$\mathcal{L}_{\text{InfoNCE}} = -\mathop{\mathbb{E}}_{\boldsymbol{x},\boldsymbol{x}'} \mathop{\mathbb{E}}_{\substack{\boldsymbol{x}_1,\boldsymbol{x}_2 \in A(\boldsymbol{x}) \\ \boldsymbol{x}^- \in A(\boldsymbol{x}')}} \log \frac{e^{f(\boldsymbol{x}_1)^\top f(\boldsymbol{x}_2)}}{e^{f(\boldsymbol{x}_1)^\top f(\boldsymbol{x}_2)} + e^{f(\boldsymbol{x}_1)^\top f(\boldsymbol{x}^-)}},$$

where encoder f is normalized by $\|f\| = 1$. It can be divided into two parts:

$$\begin{aligned}\mathcal{L}_{\text{InfoNCE}} &= \mathop{\mathbb{E}}_{\boldsymbol{x},\boldsymbol{x}'} \mathop{\mathbb{E}}_{\substack{\boldsymbol{x}_1,\boldsymbol{x}_2 \in A(\boldsymbol{x}) \\ \boldsymbol{x}^- \in A(\boldsymbol{x}')}} \left[-f(\boldsymbol{x}_1)^\top f(\boldsymbol{x}_2) + \log\left(e^{f(\boldsymbol{x}_1)^\top f(\boldsymbol{x}_2)} + e^{f(\boldsymbol{x}_1)^\top f(\boldsymbol{x}^-)}\right)\right] \qquad (5) \\ &= \underbrace{\frac{1}{2}\mathop{\mathbb{E}}_{\boldsymbol{x}} \mathop{\mathbb{E}}_{\boldsymbol{x}_1,\boldsymbol{x}_2 \in A(\boldsymbol{x})} [\|f(\boldsymbol{x}_1) - f(\boldsymbol{x}_2)\|^2] - 1}_{=:\mathcal{L}_1^{\text{InfoNCE}}(f)} + \underbrace{\mathop{\mathbb{E}}_{\boldsymbol{x},\boldsymbol{x}'} \mathop{\mathbb{E}}_{\substack{\boldsymbol{x}_1,\boldsymbol{x}_2 \in A(\boldsymbol{x}) \\ \boldsymbol{x}^- \in A(\boldsymbol{x}')}} \left[\log\left(e^{f(\boldsymbol{x}_1)^\top f(\boldsymbol{x}_2)} + e^{f(\boldsymbol{x}_1)^\top f(\boldsymbol{x}^-)}\right)\right]}_{=:\mathcal{L}_2^{\text{InfoNCE}}(f)}.\end{aligned}$$

Methods

The original align loss:

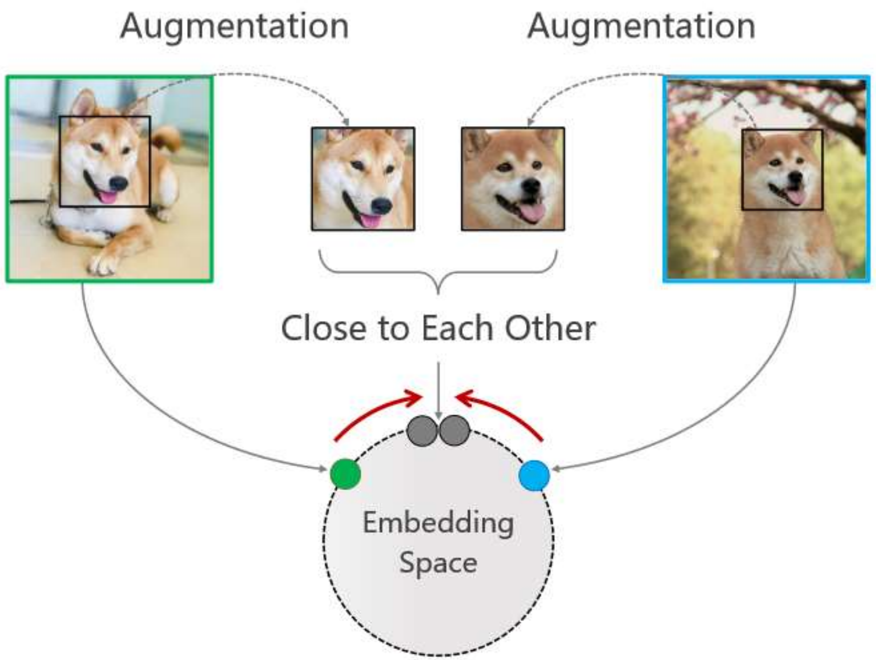
$$\mathcal{L}_{\text{align}}(f;\mathcal{D},\pi) := \mathbb{E}_{X\sim\mathcal{D}}\mathbb{E}_{(A_1,A_2)\sim\pi^2}\|f(A_1(X)) - f(A_2(X))\|^2$$

Define the augmentation-robust loss as

$$\mathcal{L}_{\text{AR}}(f;\mathcal{D}) := \mathop{\mathbb{E}}_{X\in\mathcal{D}} \sup_{A,A'\in\mathcal{A}} \|f(A(X)) - f(A'(X))\|^2$$

An approximation:

$$\widehat{\mathcal{L}}_{\text{AR}}(f) := \frac{1}{n}\sum_{i=1}^{n} \sup_{A_1,A_2\in\widehat{\mathcal{A}}_m(X_i)} \|f(A_1(X_i)) - f(A_2(X_i))\|_2^2.$$



Methods

Algorithm 1: SimCLR + ArCL

input : Batch size N, temperature τ, augmentation π, number of views m, epoch T, encoder f, projector g.

for $t = 1, \dots, T$ **do**
- sample minibatch $\{X_i\}_{i=1}^{N}$;
- **for** $i = 1, \dots, N$ **do**
  - draw m augmentations $\widehat{\mathcal{A}} = \{A_1, \dots, A_m\} \sim \pi$;
  - $z_{i,j} = g(f(A_j X_i))$ for $j \in [m]$;
  - *# select the worst positive samples*;
  - $s_i^{+} = \min_{j,k \in [m]} \{ z_{i,j}^{\top} z_{i,k} / (\|z_{i,j}\| \|z_{i,k}\|) \}$;
  - *# select the negative samples*;
  - **for** $j = 1, \dots, N$ **do**
    - $s_{i,j}^{-} = z_{i,1}^{\top} z_{j,1} / (\|z_{i,1}\| \|z_{j,1}\|)$;
    - $s_{i,j+N}^{-} = z_{i,1}^{\top} z_{j,2} / (\|z_{i,1}\| \|z_{j,2}\|)$;
- compute $L = -\frac{1}{N} \sum_{i=1}^{N} \log \frac{\exp(s_i^{+}/\tau)}{\sum_{j=1, j \neq i}^{2N} \exp(s_{i,j}^{-}/\tau)}$;
- update f and g to minimize L;

return f

Experiments

Table 1: 5 different augmentations.

	Grayscale	RandomCrop	HorizontalFlip	ColorJitter
Aug 1	✓	–	–	–
Aug 2	–	✓	–	–
Aug 3	–	–	✓	–
Aug 4	–	–	–	✓
Aug 5	✓	–	–	✓

Table 2: Linear evaluation results (%) of pretrained CIFAR10 models on CIFAR10, CIFAR100 and their modified versions.

	Method	Batch Size	Aug 1	Aug 2	Aug 3	Aug 4	Aug 5	Original
CIFAR10	SimCLR	256	86.36	83.21	86.93	86.42	86.13	86.76
	SimCLR + ArCL (views=4)	256	88.68	86.77	89.01	88.70	88.31	88.95
	SimCLR + ArCL (views=6)	256	**88.95**	**87.18**	**89.54**	**88.92**	**88.61**	**89.11**
	SimCLR	512	88.62	86.27	88.96	88.56	88.37	88.81
	SimCLR + ArCL (views=4)	512	89.97	88.06	90.48	89.91	89.59	90.20
	SimCLR + ArCL (views=6)	512	90.24	89.54	90.69	90.43	90.07	90.69
	SimCLR + ArCL (views=8)	512	**90.44**	**88.96**	**90.98**	**90.63**	**90.31**	**90.84**
CIFAR100	SimCLR	256	51.65	47.55	53.17	52.05	51.36	52.75
	SimCLR+ArCL(views=4)	256	53.76	49.80	55.68	54.19	52.96	54.83
	SimCLR+ArCL(views=6)	256	**54.13**	**50.74**	**55.74**	**54.75**	**53.46**	**55.29**
	SimCLR	512	52.28	48.09	53.45	52.58	51.53	53.12
	SimCLR+ArCL(views=4)	512	53.40	50.16	54.92	53.77	52.61	54.20
	SimCLR+ArCL(views=6)	512	54.00	50.57	**56.24**	**55.04**	**53.77**	55.60
	SimCLR+ArCL(views=8)	512	**54.59**	**50.85**	55.74	54.62	53.21	**55.96**

Experiments

		Epochs	ImageNet	Aircraft	Caltech101	Cars	CIFAR10	CIFAR100	DTD	Flowers	Food	Pets	Avg
Linear	MoCo	800	70.68	41.79	87.92	39.31	92.28	74.90	73.88	90.07	68.95	83.30	72.49
	MoCo	800+50	70.64	41.00	87.63	39.01	92.27	75.14	74.31	88.31	68.57	83.69	72.21
	MoCo + ArCL(views=2)	800+50	69.70	44.29	89.79	42.15	93.07	76.70	74.20	90.40	70.94	83.68	73.91
	MoCo + ArCL(views=3)	800+50	69.80	44.57	89.48	42.11	**93.29**	**77.33**	74.63	91.13	71.16	**84.23**	74.21
	MoCo + ArCL(views=4)	800+50	69.80	**44.62**	**89.66**	**42.88**	93.22	76.83	**75.00**	**91.59**	**71.35**	83.99	**74.35**
	MoCo	800+100	70.64	41.65	87.64	39.31	92.12	75.03	73.94	89.53	68.31	83.55	72.34
	MoCo + ArCL(views=2)	800+100	70.26	41.86	89.52	40.21	92.64	75.73	**74.04**	88.97	70.06	84.31	73.04
	MoCo + ArCL(views=3)	800+100	70.92	43.87	89.36	**42.37**	93.30	76.93	72.93	90.50	**71.14**	84.03	73.83
	MoCo + ArCL(views=4)	800+100	69.42	**45.55**	**89.61**	41.91	**93.55**	**77.05**	**74.04**	**91.17**	70.93	**85.48**	**74.37**
	MoCo	200	67.72	40.02	86.59	37.41	90.90	72.43	73.88	87.97	66.97	80.00	70.69
	MoCo + ArCL(views=2)	200	68.04	42.34	87.92	36.45	92.29	71.71	74.68	89.00	67.87	81.42	71.85
	MoCo + ArCL(views=3)	200	68.74	**43.21**	88.26	38.27	92.49	75.02	74.68	**89.61**	68.31	**81.64**	72.39
	MoCo + ArCL(views=4)	200	68.92	41.65	**88.42**	**38.77**	**92.70**	**75.73**	**75.43**	88.95	**68.63**	81.60	**72.43**
	Supervised*		77.20	43.59	90.18	44.92	91.42	73.90	72.23	89.93	69.49	91.45	74.12
Finetune	MoCo	800		83.56	82.54	85.09	95.89	71.81	69.95	95.26	76.81	88.83	83.30
	MoCo	800+50		83.15	84.50	85.90	96.13	72.58	70.16	94.44	79.34	86.12	83.59
	MoCo + ArCL(views=2)	800+50		**86.05**	87.38	**87.28**	96.33	79.39	72.18	95.89	81.36	89.03	86.10
	MoCo + ArCL(views=3)	800+50		84.03	87.64	86.34	**96.88**	80.98	72.87	**96.14**	**81.90**	89.20	86.22
	MoCo + ArCL(views=4)	800+50		84.19	**88.42**	86.67	96.68	**81.17**	**73.09**	95.90	81.70	**89.52**	**86.37**
	MoCo	800+100		83.18	84.50	84.27	96.01	72.14	70.27	95.53	78.23	88.73	83.65
	MoCo + ArCL(views=2)	800+100		84.45	86.84	87.20	96.40	78.40	71.91	95.93	80.54	88.56	85.58
	MoCo + ArCL(views=3)	800+100		**85.94**	86.85	**87.34**	96.36	79.75	71.44	96.00	81.48	88.26	85.94
	MoCo + ArCL(views=4)	800+100		85.65	**88.50**	86.39	**96.91**	**81.29**	**73.35**	**96.17**	**81.82**	**89.30**	**86.60**
	MoCo	200		83.18	82.66	84.47	95.51	72.54	70.43	94.99	77.39	86.12	83.03
	MoCo + ArCL(views=2)	200		81.09	83.93	**86.54**	95.88	76.18	**70.69**	94.44	76.78	86.98	83.61
	MoCo + ArCL(views=3)	200		84.79	85.61	85.39	**96.56**	**78.81**	70.59	**95.84**	**80.71**	87.91	85.13
	MoCo + ArCL(views=4)	200		**84.88**	**86.19**	85.90	96.35	78.62	**70.69**	95.77	80.46	**88.00**	**85.21**
	Supervised*			83.50	91.01	82.61	96.39	82.91	73.30	95.50	84.60	92.42	86.92

Experiments

Table 7: Linear evaluation results of pretrained models using SimCLR with two different alignment losses on MNIST-CIFAR dataset. The average results under three diffrent random seeds are given.

Methods	Accuracy(%)	Methods	Accuracy(%)
SimCLR	85.6		
SimCLR + ArCL(views=3)	86.0	SimCLR+AAL(views=3)	85.4
SimCLR + ArCL(views=4)	87.2	SimCLR+AAL(views=4)	86.1
SimCLR + ArCL(views=5)	87.3	SimCLR+AAL(views=5)	86.3
SimCLR + ArCL(views=6)	88.4	SimCLR+AAL(views=6)	85.8

WHAT SHOULD NOT BE CONTRASTIVE IN CONTRASTIVE LEARNING

Tete Xiao
UC Berkeley

Xiaolong Wang
UC San Diego

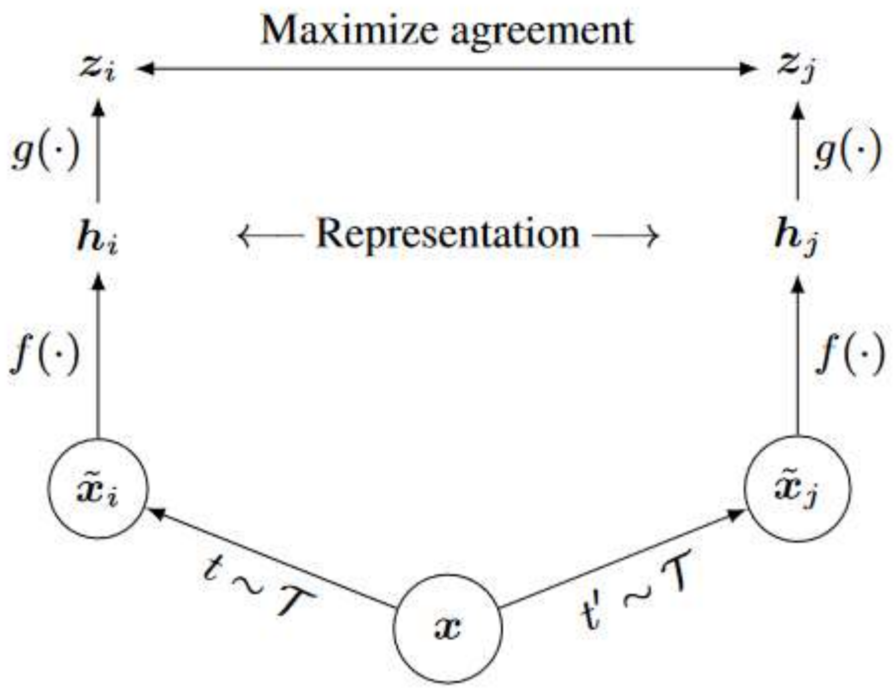
Alexei A. Efros
UC Berkeley

Trevor Darrell
UC Berkeley

ICLR2021

Background

Is the classical contrastive learning framework capable of achieving optimal in generalization?

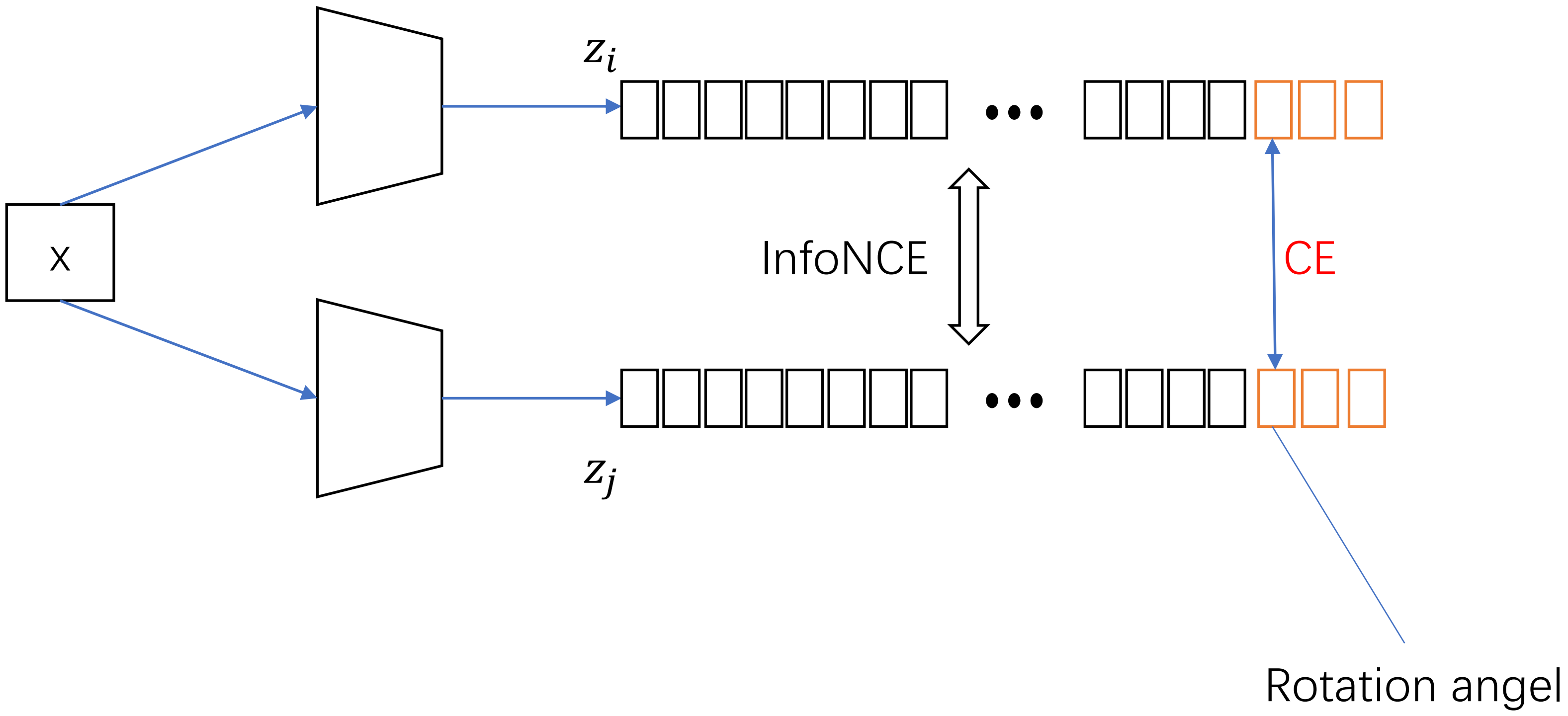


Consider the extreme situation:

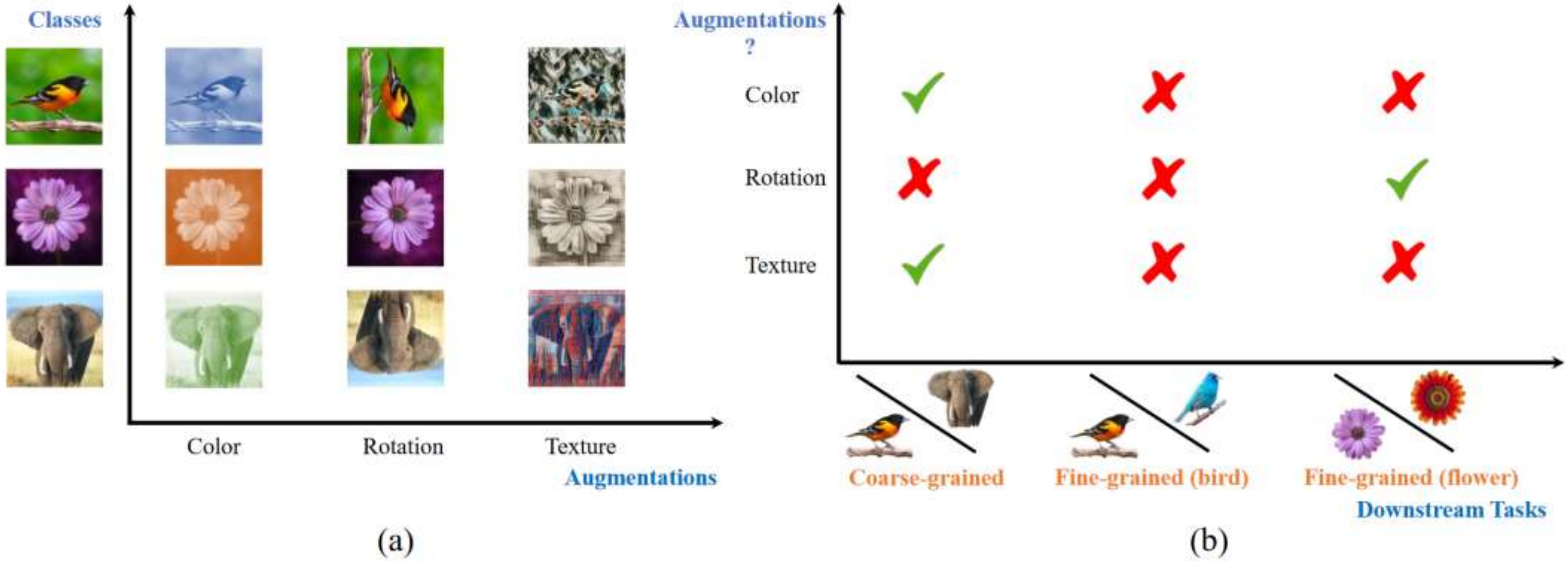
$$z_i = z_j, \forall x \in D_x$$

If T contains Rotation, then the features contains non information to justify the rotation angel in some rotation-sensitive tasks.

Background

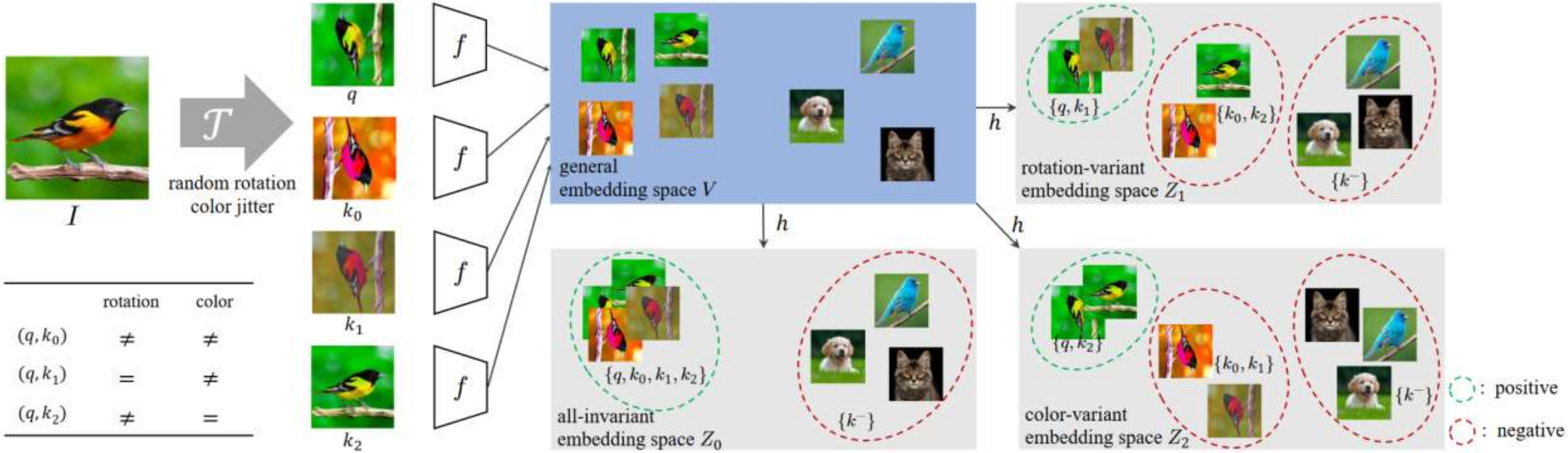


Motivation



Classes
Color
Rotation
Texture
Augmentations
(a)
Augmentations
?
Color
Rotation
Texture
Coarse-grained
Fine-grained (bird)
Fine-grained (flower)
Downstream Tasks
(b)

Method



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

$$E_{i,j}^{\{+,-\}} = \exp\left(\boldsymbol{z}_i^q \cdot \boldsymbol{z}_i^{k_j^{\{+,-\}}} / \tau\right)$$

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

Experiments

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.

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

Table 2: **Evaluation on multiple downstream tasks**. Our method demonstrates superior generalizability and transferability with increasing number of augmentations.

model	Augmentation Color	Augmentation Rotation	iNat-1k top-1	iNat-1k top-5	CUB-200 top-1	CUB-200 top-5	Flowers-102 5-shot	Flowers-102 10-shot	IN-100 top-1	IN-100 top-5
MoCo	✓		36.2	62.0	36.7	64.7	67.9 (± 0.5)	77.3 (± 0.1)	81.0	95.2
LooC	✓		41.2	67.0	40.1	69.7	68.2 (± 0.6)	77.6 (± 0.1)	81.1	95.3
		✓	40.0	65.4	38.8	67.0	70.1 (± 0.4)	79.3 (± 0.1)	80.2	95.5
	✓	✓	44.0	69.3	39.6	69.2	70.9 (± 0.3)	80.8 (± 0.2)	79.2	94.7
LooC++	✓	✓	46.1	71.5	39.3	69.3	68.1 (± 0.4)	78.8 (± 0.2)	81.2	95.2

Experiments

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 ($\pm$ 0.6)	77.6 ($\pm$ 0.1)	48.0
	✓			79.7	94.4	42.9	68.7	69.1 ($\pm$ 0.7)	79.5 ($\pm$ 0.2)	47.1
		✓		81.5	94.9	41.4	67.4	70.5 ($\pm$ 0.6)	80.0 ($\pm$ 0.2)	52.6
			✓	80.3	94.9	43.0	68.6	70.4 ($\pm$ 0.5)	80.5 ($\pm$ 0.2)	44.1
	✓	✓	✓	82.2	95.3	45.9	71.4	71.0 ($\pm$ 0.7)	81.9 ($\pm$ 0.3)	48.0

Experiments

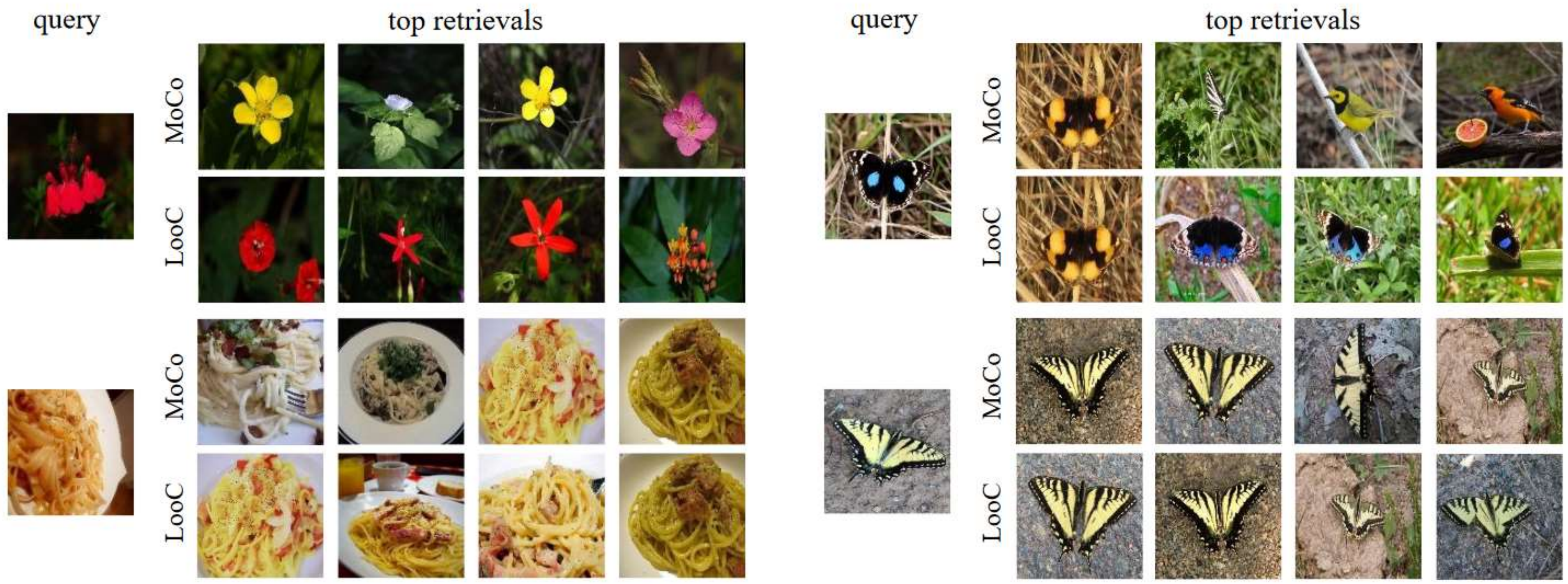


Figure 3: **Top nearest-neighbor retrieval results** of LooC vs. corresponding invariant MoCo baseline with color (left) and rotation (right) augmentations on IN-100 and iNat-1k. The results show that our model can better preserve information dependent on color and rotation despite being trained with those augmentations.

Thanks

Augmented Distance

$$d_A(\boldsymbol{x}_1, \boldsymbol{x}_2) = \min_{\boldsymbol{x}_1' \in A(\boldsymbol{x}_1), \boldsymbol{x}_2' \in A(\boldsymbol{x}_2)} \|\boldsymbol{x}_1' - \boldsymbol{x}_2'\| .$$

Definition 1 ((σ, δ)-Augmentation). The augmentation set A is called a (σ, δ)-augmentation, if for each class C_k, there exists a subset $C_k^0 \subseteq C_k$ (called a main part of C_k), such that both $\mathbb{P}[\boldsymbol{x} \in C_k^0] \geq \sigma\, \mathbb{P}[\boldsymbol{x} \in C_k]$ where $\sigma \in (0, 1]$ and $\sup_{\boldsymbol{x}_1, \boldsymbol{x}_2 \in C_k^0} d_A(\boldsymbol{x}_1, \boldsymbol{x}_2) \leq \delta$ hold.

Larger σ and smaller δ indicate the sharper concentration of augmented data.

For any $A' \supseteq A$, $\quad d_{A'}(x_1, x_2) \leq d_A(x_1, x_2)$

Theories

南京航空航天大學
NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

Theorem 1. *Given a (σ, δ)-augmentation used in contrastive SSL, if*

$$\mu_\ell^\top \mu_k < r^2 \left(1 - \rho_{max}(\sigma, \delta, \varepsilon) - \sqrt{2\rho_{max}(\sigma, \delta, \varepsilon)} - \frac{\Delta_\mu}{2}\right) \tag{2}$$

holds for any pair of (ℓ, k) with $\ell \neq k$, then the downstream error rate of NN classifier G_f

$$\mathrm{Err}(G_f) \leq (1 - \sigma) + R_\varepsilon, \tag{3}$$

where $\rho_{max}(\sigma, \delta, \varepsilon) = 2(1 - \sigma) + \frac{R_\varepsilon}{\min_\ell p_\ell} + \sigma \left(\frac{L\delta}{r} + \frac{2\varepsilon}{r}\right)$ *and* $\Delta_\mu = 1 - \min_{k \in [K]} \|\mu_k\|^2 / r^2$.

Proof of Proposition 4.1. For any $\varepsilon > 0$, let $t = \sqrt{\varepsilon}/2$ and $f(x_1, x_2) = x_1 + tx_2$. Then, the alignment loss of f satisfies

$$\mathcal{L}_{\text{align}}(f; \mathcal{D}, \pi) = t^2\, \mathbb{E}\, X_2^2 \mathop{\mathbb{E}}_{(\theta_1,\theta_2)\sim\mathcal{N}(0,1)^2} (\theta_1 - \theta_2)^2 = 2t^2 < \varepsilon.$$

Let $c = 0$ and $c' = 1/t$. Then obviously

$$\mathcal{R}(f; \mathcal{D}_c) = 0,$$

but

$$\mathcal{R}(f; \mathcal{D}_{c'}) = P(X_1 < 0, X_1 + X_2 \geq 0) + P(X_1 \geq 0, X_1 + X_2 \leq 0) = \frac{1}{4}.$$

Formulation

南京航空航天大學
NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

$$\mathcal{R}(f;\mathcal{D}^{\text{tar}}) := \min_{h\in\mathbb{R}^{K\times m}} \mathop{\mathbb{E}}_{X\sim\mathcal{D}^{\text{tar}}} \ell(h\circ f(X),Y), \tag{2}$$

$$\mathcal{L}_{\text{align}}(f;\mathcal{D},\pi) := \mathbb{E}_{X\sim\mathcal{D}}\mathbb{E}_{(A_1,A_2)\sim\pi^2}\|f(A_1(X)) - f(A_2(X))\|^2$$

Proposition 4.1. *Consider a two-dimensional classification problem with data* $(X_1,X_2)\sim\mathcal{N}(0,I_2)$. *The label* Y *satisfies* $Y=1(X_1\geq 0)$, *and the data augmentation is to multiply* X_2 *by standard normal noise, i.e.,*

$$A_\theta(X) = (X_1, \theta\cdot X_2),$$
$$\theta\sim\mathcal{N}(0,1).$$

The corresponding transformation-induced domain set is $\mathcal{P}=\{\mathcal{D}_c:\mathcal{D}_c=(X_1,c\cdot X_2)\ for\ c\in\mathbb{R}\}$. *We consider the 0-1 loss in equation 2. Then for every* $\varepsilon>0$, *there exists representation* f *and two domains* $\mathcal{D}_c$ *and* $\mathcal{D}_{c'}$ *such that*

$$\mathcal{L}_{\text{align}}(f;\mathcal{D},\pi)<\varepsilon,$$

but

$$|\mathcal{R}(f;\mathcal{D}_c)-\mathcal{R}(f;\mathcal{D}_{c'})|\geq\frac{1}{4}$$