中国人民大学
RENMIN UNIVERSITY OF CHINA
高瓴人工智能学院
Gaoling School of Artificial Intelligence

What to contrast?

Bing Su
Joint work with
Wenwen Qiang, Jiangmeng Li, Hui Xiong, Ji-Rong Wen

Contrastive Learning

- CL: A long story …
- CL has achieved empirical progress in *Self-supervised learning*

Contrastive Learning

- Contrastive loss guides the learned features to bring positive pairs together and push negative pairs farther apart.

$$\mathcal{L} = - \mathop{\mathbb{E}}_{X_S} \left[\log \frac{d\left(\{z^+\}\right)}{d\left(\{z^+\}\right) + \sum_{k=1}^{K} d\left(\{z^-\}_k\right)} \right]$$

  - X_S : a set of pairs randomly sampled from X
  - $\{z^+\}$: a positive pair
  - $\{z^-\}_k$: negative pairs, $k \in \{1, ..., K\}$
  - $d(\cdot)$: a discriminating function

Contrastive Learning

- Key points
  - Positive sample
  - Negative samples

What to contrast is crucial !

Contents

Motivation

- **Contrastive learning heavily relies on informative features, or "hard" (positive or negative) features**
  - Early works include informative features by applying complex data augmentations or adopting large batch size or memory bank
  - Recent works design elaborate sampling approaches to explore informative features
- **Learning anti-collapsed feature augmentation**

Challenge

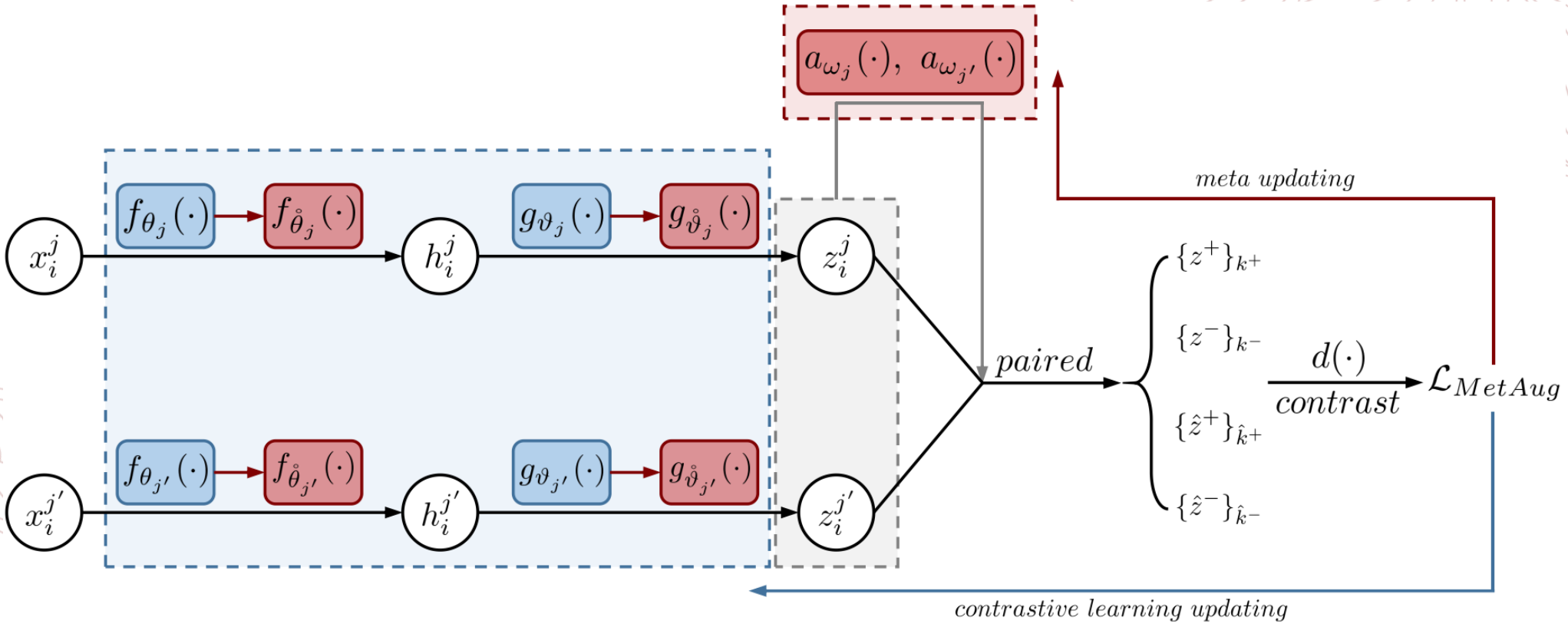
- **It is desirable to *learn* informative feature augmentations**
  - alleviate the need of strong augmentations on data
  - from a restricted amount of images (small batch size)
  - anti-collapsed

MetAug

- Meta Feature Augmentation (MetAug)
- Tackle augmentations on features
- Learn view-specific encoders (with projection heads) and auxiliary meta feature augmentation generators (MAGs) by ***margin-injected meta feature augmentation*** and ***optimization-driven unified contrast***.

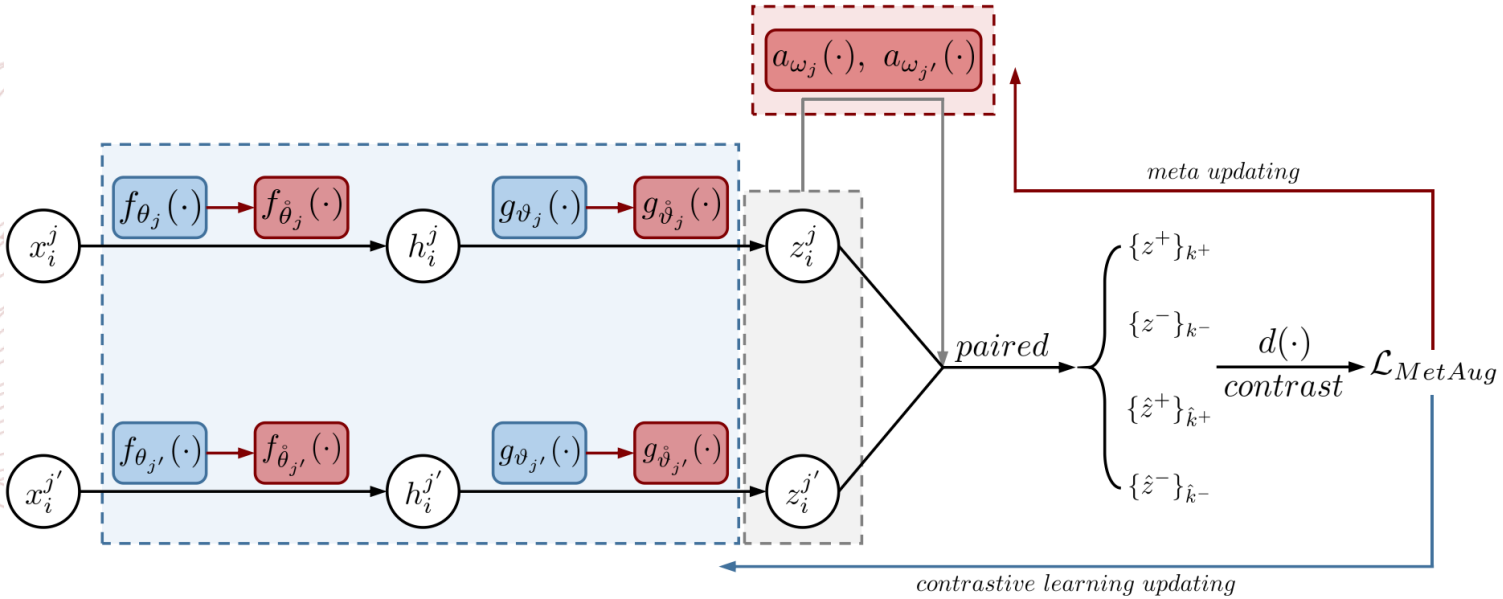
MetAug

- Overview

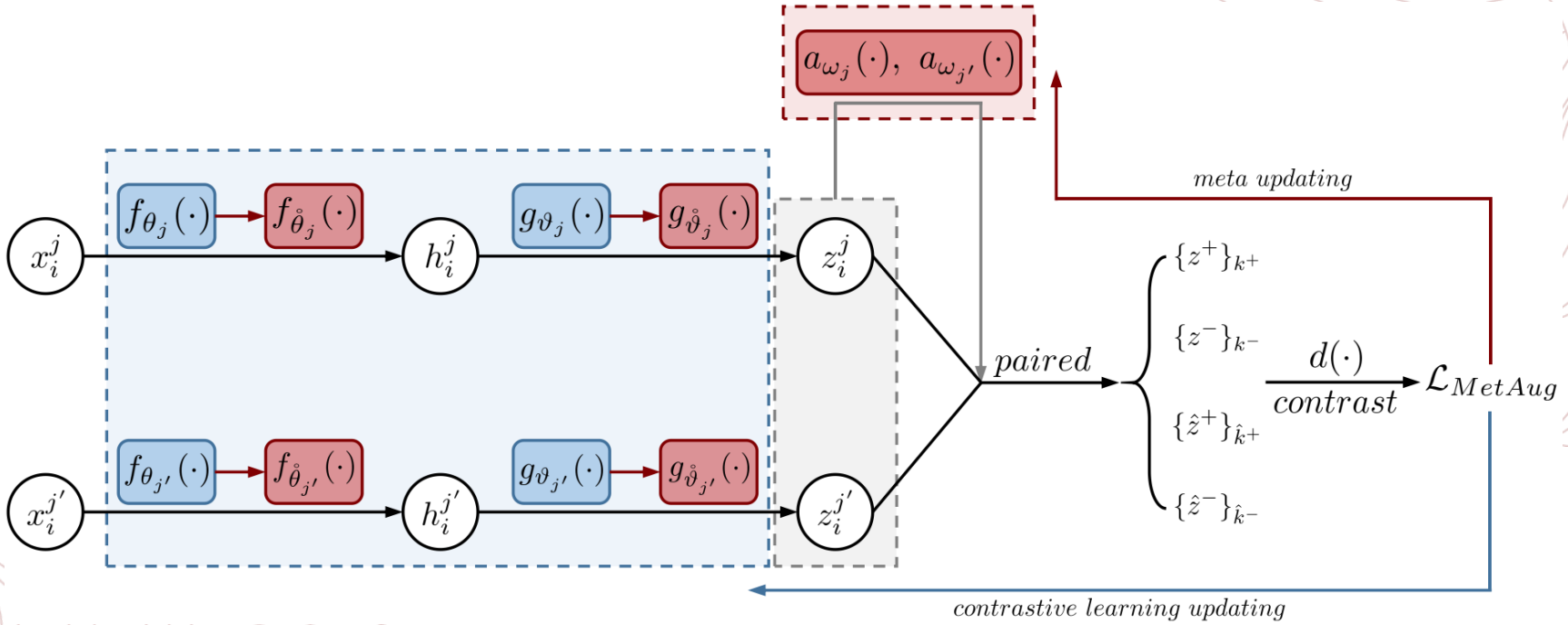


Training

- Updating the encoders and projectors
- Updating MAGs in a meta learning manner: leverages second-derivative technique to update the parameters with respect to the improvement of the contrastive learning



Updating MAGs



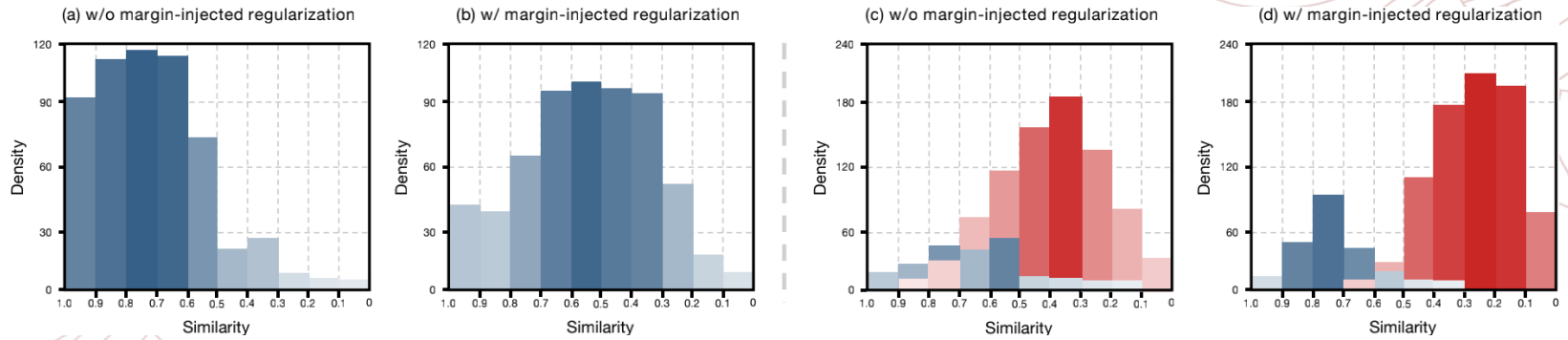
$$\mathring{\theta} = \theta - \ell \cdot \nabla_\theta \mathcal{L}\left(\left\{g_\vartheta\left(f_\theta(\widetilde{X})\right), a_\omega\left(g_\vartheta\left(f_\theta(\widetilde{X})\right)\right)\right\}\right)$$

$$\mathring{\vartheta} = \vartheta - \ell \cdot \nabla_\vartheta \mathcal{L}\left(\left\{g_\vartheta\left(f_\theta(\widetilde{X})\right), a_\omega\left(g_\vartheta\left(f_\theta(\widetilde{X})\right)\right)\right\}\right)$$

$$\arg\min_\omega \mathcal{L}\left(\left\{g_{\mathring{\vartheta}}\left(f_{\mathring{\theta}}(\widetilde{X})\right), a_\omega\left(g_{\mathring{\vartheta}}\left(f_{\mathring{\theta}}(\widetilde{X})\right)\right)\right\}\right)$$

Margin-injected regularization

- Injects a margin to encourage MAGs to generate anti-collapsed augmented features



$$\sigma^{+} = \min\left[\min\left(\{d(\{z^{+}\}_{k^{+}})\}\right), \max\left(\{d(\{z^{-}\}_{k^{-}})\}\right)\right]$$

$$\sigma^{-} = \max\left[\min\left(\{d(\{z^{+}\}_{k^{+}})\}\right), \max\left(\{d(\{z^{-}\}_{k^{-}})\}\right)\right]$$

$$\omega \leftarrow \omega - \ell' \cdot \nabla_{\omega}\mathcal{L}\Big(\Big\{g_{\mathring{\vartheta}}(f_{\mathring{\theta}}(\widetilde{X})), a_{\omega}\Big(g_{\mathring{\vartheta}}(f_{\mathring{\theta}}(\widetilde{X}))\Big)\Big\}\Big) + \alpha \cdot \mathcal{R}_{\sigma}$$

$$\mathcal{R}_{\sigma} = \frac{1}{\hat{K}^{+}} \sum_{\hat{k}^{+}=1}^{\hat{K}^{+}} \left[d(\{\hat{z}^{+}\}_{\hat{k}^{+}}) - \sigma^{+}\right]_{+} + \frac{1}{\hat{K}^{-}} \sum_{\hat{k}^{-}=1}^{\hat{K}^{-}} \left[\sigma^{-} - d(\{\hat{z}^{-}\}_{\hat{k}^{-}})\right]_{+}$$

Optimization-Driven Unified Contrast

- Jointly contrasts all features in one gradient back-propagation step
- Emphasizes the weight to the similarity that deviates from the optimum and decrases the weight to the similarity having close proximity with the optimum

$$\mathcal{L}_{OUCL} = \left[\sum_{k^-=1}^{K^-} d(\{z^-\}_{k^-}) - \sum_{k^+=1}^{K^+} d(\{z^+\}_{k^+}) + \lambda \right]_+$$

$$\mathcal{L}_{OUCL} = \frac{1}{\beta} log\left\{ 1 + \sum_{k^-=1}^{K^-} \sum_{k^+=1}^{K^+} exp\Big[\beta\Big((d(\{z^+\}_{k^+}) - 1)^2 + (d(\{z^-\}_{k^-}))^2 - 2\gamma^2 \Big) \Big] \right\}$$

$$\mathcal{L}_{OUCL} = \frac{1}{\beta} log\left\{ 1 + \sum_{k^-=1}^{K^-} \sum_{k^+=1}^{K^+} exp\Big[\beta\Big(\Gamma^- (d(\{z^-\}_{k^-}) - \gamma^-) - \Gamma^+ (d(\{z^+\}_{k^+}) - \gamma^+) \Big) \Big] \right\}$$

Evaluation

- Comparison with self-supervised learning methods

Model	Tiny ImageNet		STL-10		CIFAR10		CIFAR100	
	conv	fc	conv	fc	conv	fc	conv	fc
Fully supervised	36.60		68.70		75.39		42.27	
BiGAN	24.38	20.21	71.53	67.18	62.57	62.74	37.59	33.34
NAT	13.70	11.62	64.32	61.43	56.19	51.29	29.18	24.57
DIM	33.54	36.88	72.86	70.85	73.25	73.62	48.13	45.92
SplitBrain‡	32.95	33.24	71.55	63.05	77.56	76.80	51.74	47.02
SwAV	39.56 ± 0.2	38.87 ± 0.3	70.32 ± 0.4	71.40 ± 0.3	68.32 ± 0.2	65.20 ± 0.3	44.37 ± 0.3	40.85 ± 0.3
SimCLR	36.24 ± 0.2	39.83 ± 0.1	75.57 ± 0.3	77.15 ± 0.3	80.58 ± 0.2	80.07 ± 0.2	50.03 ± 0.2	49.82 ± 0.3
CMC‡	41.58 ± 0.1	40.11 ± 0.2	83.03	85.06	81.31 ± 0.2	83.28 ± 0.2	58.13 ± 0.2	56.72 ± 0.3
MoCo	35.90 ± 0.2	41.37 ± 0.2	77.50 ± 0.2	79.73 ± 0.3	76.37 ± 0.3	79.30 ± 0.2	51.04 ± 0.2	52.31 ± 0.2
BYOL	41.59 ± 0.2	41.90 ± 0.1	81.73 ± 0.3	81.57 ± 0.2	77.18 ± 0.2	80.01 ± 0.2	53.64 ± 0.2	53.78 ± 0.2
Barlow Twins	39.81 ± 0.3	40.34 ± 0.2	80.97 ± 0.3	81.43 ± 0.3	76.63 ± 0.3	78.49 ± 0.2	52.80 ± 0.2	52.95 ± 0.2
DACL	40.61 ± 0.2	41.26 ± 0.1	80.34 ± 0.2	80.01 ± 0.3	81.92 ± 0.2	80.87 ± 0.2	52.66 ± 0.2	52.08 ± 0.3
LooC	42.04 ± 0.1	41.93 ± 0.2	81.92 ± 0.2	82.60 ± 0.2	83.79 ± 0.2	82.05 ± 0.2	54.25 ± 0.2	54.09 ± 0.2
SimCLR + Debiased	38.79 ± 0.2	40.26 ± 0.2	77.09 ± 0.3	78.39 ± 0.2	80.89 ± 0.2	80.93 ± 0.2	51.38 ± 0.2	51.09 ± 0.2
SimCLR + Hard	40.05 ± 0.3	41.23 ± 0.2	79.86 ± 0.2	80.20 ± 0.2	82.13 ± 0.2	82.76 ± 0.1	52.69 ± 0.2	53.13 ± 0.2
CMC‡ + Debiased	41.64 ± 0.2	41.36 ± 0.1	83.79 ± 0.3	84.20 ± 0.2	82.17 ± 0.2	83.72 ± 0.2	58.48 ± 0.2	57.16 ± 0.2
CMC‡ + Hard	42.89 ± 0.2	42.01 ± 0.2	83.16 ± 0.3	85.15 ± 0.2	83.04 ± 0.2	86.22 ± 0.2	58.97 ± 0.3	59.13 ± 0.2
MetAug (only OUCL)‡	42.02 ± 0.1	42.14 ± 0.2	84.09 ± 0.2	84.72 ± 0.3	85.98 ± 0.2	87.13 ± 0.2	59.21 ± 0.2	58.73 ± 0.2
MetAug‡	**44.51 ± 0.2**	**45.36 ± 0.2**	**85.41 ± 0.3**	**85.62 ± 0.2**	**87.87 ± 0.2**	**88.12 ± 0.2**	**59.97 ± 0.3**	**61.06 ± 0.2**

Evaluation

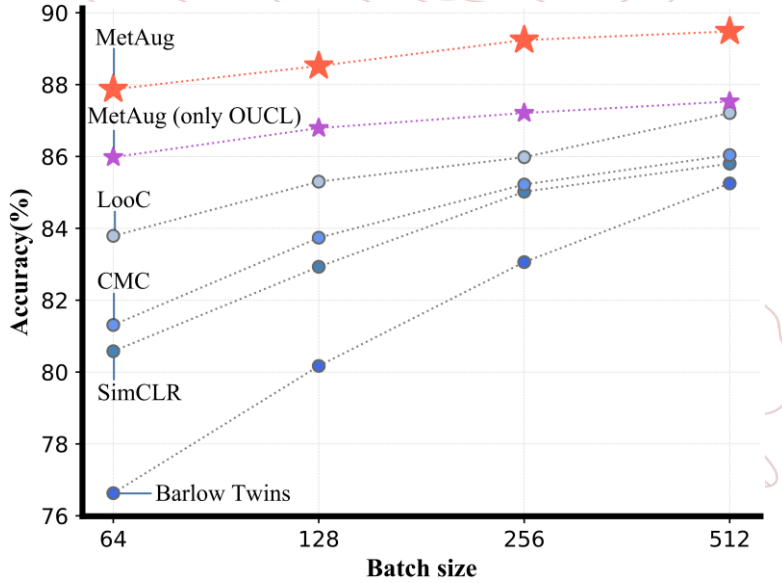
- Comparison with self-supervised learning methods

Model	CIFAR10	STL-10	Average
SwAV	83.15	82.93	83.04
SimCLR	84.63	83.75	84.19
CMC	86.10	86.83	86.47
BYOL	87.14	87.56	87.35
Barlow Twins	85.84	86.02	85.93
DACL	86.93	88.11	87.52
LooC	87.80	88.62	88.21
SwAV + Hard	83.99	84.51	84.25
SimCLR + Hard	86.91	85.48	86.20
CMC + Hard	88.25	87.79	88.02
MetAug (*only OUCL*)	88.79	88.31	88.55
MetAug	**91.09**	**90.26**	**90.68**

Model	ImageNet		
	conv	ResNet-50	
	top 1		top 5
Fully supervised	50.5	-	-
SplitBrain	32.8	-	-
CPC v2	-	63.8	85.3
SwAV	38.0 ± 0.3	71.8	-
SimCLR	37.7 ± 0.2	71.7	-
CMC	42.6	-	-
MoCo	39.4 ± 0.2	71.1	-
SimSiam	-	71.3	-
InfoMin Aug.	-	73.0	91.1
BYOL	41.1 ± 0.2	74.3	91.6
Barlow Twins	39.6 ± 0.2	-	-
NNCLR	-	75.4	92.3
DACL	41.8 ± 0.2	-	-
LooC	43.2 ± 0.2	-	-
SimCLR + Debiased	38.9 ± 0.3	-	-
SimCLR + Hard	41.5 ± 0.2	-	-
MetAug	$\mathbf{45.1 \pm 0.2}$	-	-
MetAug*	-	**76.0**	**93.2**

Evaluation

- Comparison under multiple batch sizes



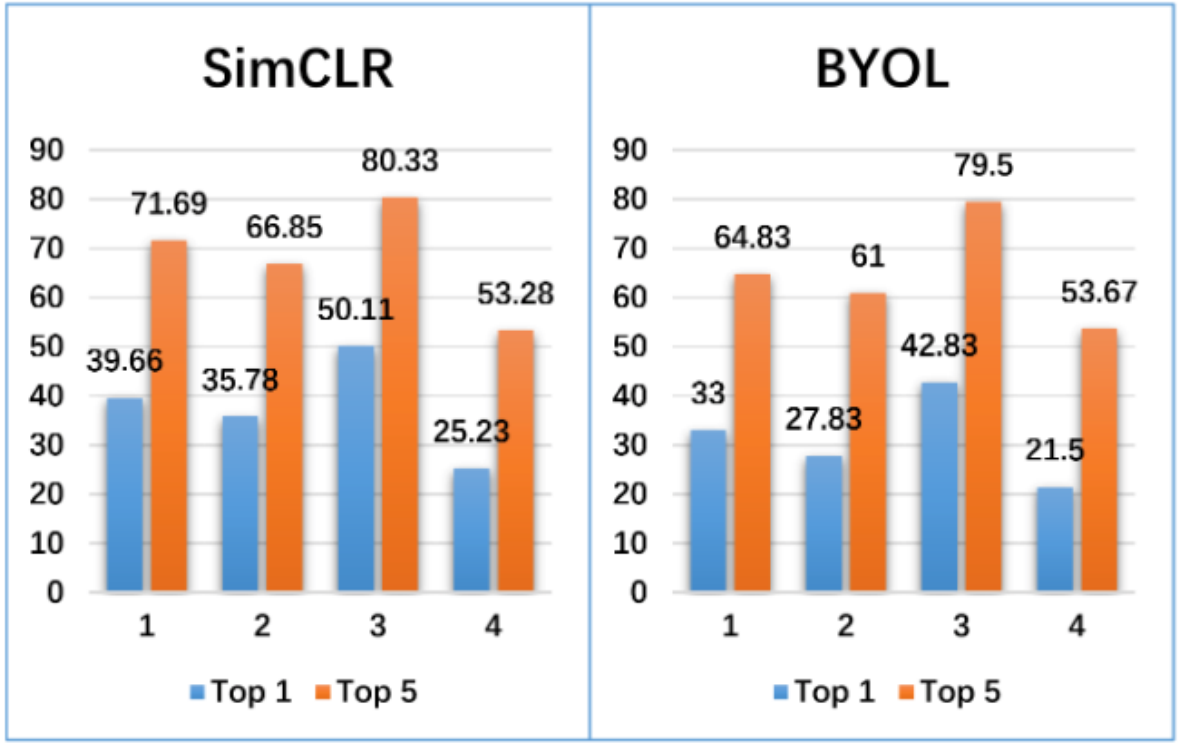
- Comparisons with different data augmentations

ıset. ID	Data augmentations						Methods		
	horizontal flip	rotate	random crop	random grey	color jitter	mixup	DACL	LooC	MetAug
1	✓	✓					-	80.73	**87.05**
2			✓				-	81.16	**87.53**
3				✓			-	80.70	**86.81**
4					✓		-	81.64	**87.79**
5	✓		✓				-	82.05	**88.12**
6		✓			✓		-	82.16	**88.01**
7	✓		✓			✓	80.87	82.21	**88.22**
8	✓	✓	✓	✓	✓	✓	82.09	83.17	**88.65**

Contents

Motivation

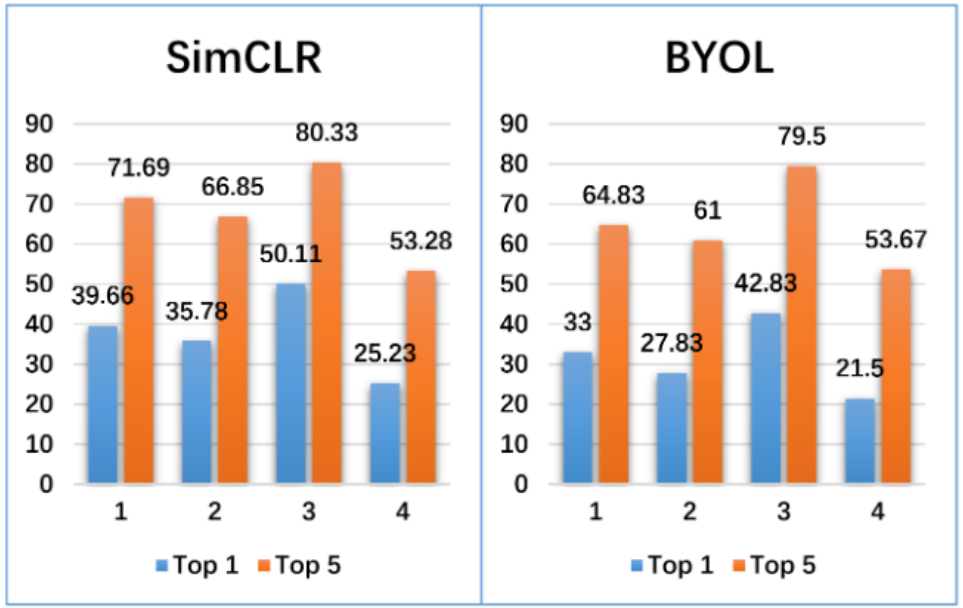
- Often-Overlooked Characteristic of Current Contrastive Learning Methods



- 1: training and testing on full images
- 2: training on full images and testing on foreground images
- 3: training and testing on foreground images
- 4: training on foreground images and testing on full images

Motivation

- Often-Overlooked Characteristic of Current Contrastive Learning Methods



Observation: background-related information degrades the performance of the CL models.

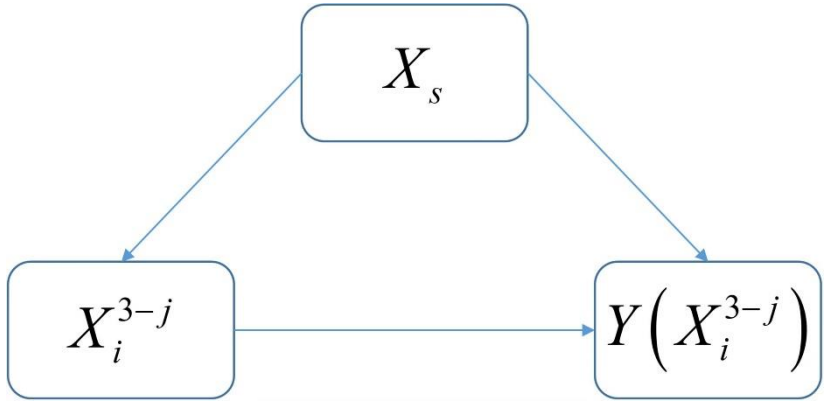
Explanation: the feature extractor trained on full images so that it extracts background-dependent semantic features. But contrastive learning strives to be adaptable to a variety of downstream tasks. Only foreground-related semantic information can ensure the robustness of the learned features to various tasks.

Intuition

1. To capture the causal links between semantic information, positive sample, and anchor, we establish a Structural Causal Model (SCM).
2. We propose a new method by implementing backdoor adjustments to the planned SCM.

Problem Formulation

- **Structural Causal Model**
  - The nodes in SCM represent the abstract data variables and the directed edges represent the (functional) causality



- X_s : semantic information
- X_i^{3-j} : positive sample
- $Y(X_i^{3-j})$: anchor (or label)

Problem Formulation

- **Causal Intervention via Backdoor Adjustment**
  - The backdoor adjustment assumes that we can observe and stratify the confounder

$$P(Y(X_i^{3-j})|do(X_i^{3-j})) = \sum_{i=1}^{n} P(Y(X_i^{3-j})|X_i^{3-j}, Z_s^i)P(Z_s^i)$$

- Z_s^i : a stratification of semantic feature
- $P(Y(X_i^{3-j})|do(X_i^{3-j}))$: the true causality between $Y(X_i^{3-j})$ and X_i^{3-j} .

Meta Semantic Regularizer

- The implementation of the backdoor adjustment during the training phase

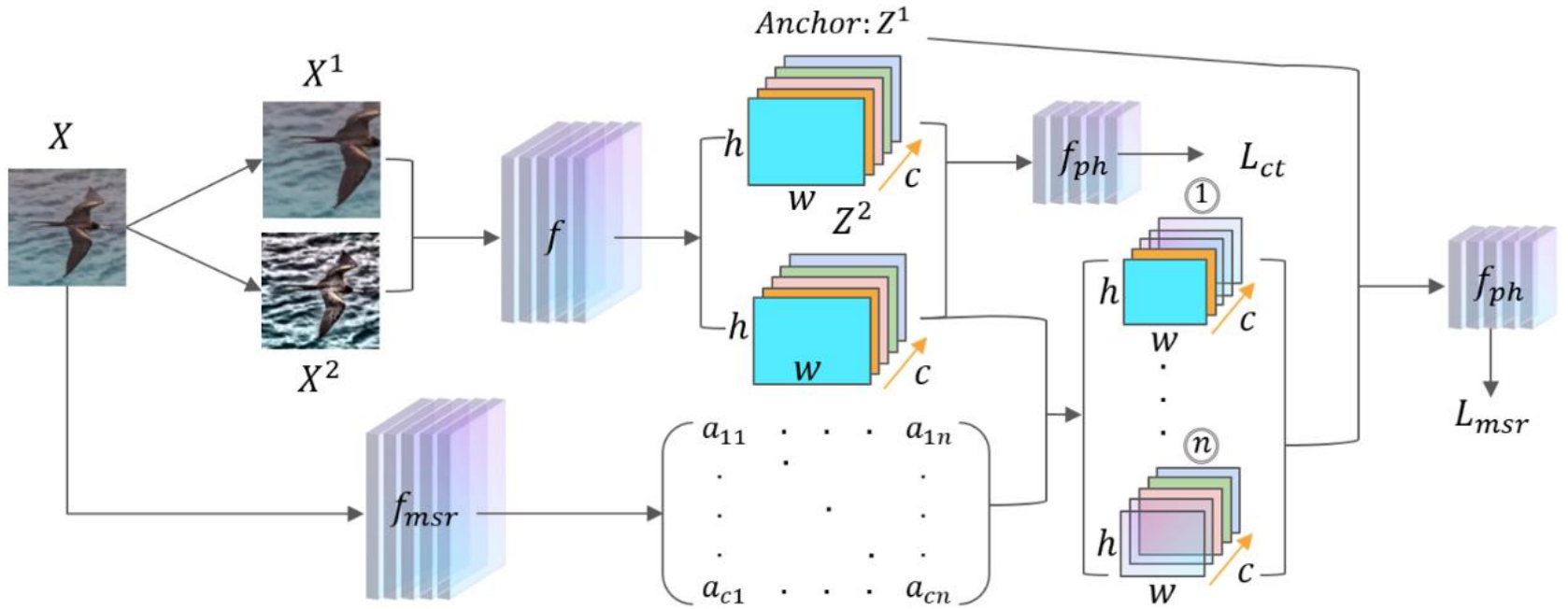
$$Z_s^t = a_t \qquad a_t = [a_{1,t}, ..., a_{c,t}]^{\mathrm{T}} \qquad P(Z_s^t) = 1/n$$

$$P(Y(X_i^{3-j})|X_i^{3-j}, Z_s^t) = \frac{\exp\left(\frac{\mathrm{sim}\left(Z_i^j, a_t \odot Z_i^{3-j}\right)}{\tau}\right)}{\exp\left(\frac{\mathrm{sim}\left(Z_i^j, a_t \odot Z_i^{3-j}\right)}{\tau}\right) + \sum_{\substack{k=1,\\ k\neq i}}^{N} \sum_{\substack{l=1,\\ l\neq j}}^{2} \exp\left(\frac{\mathrm{sim}\left(Z_i^j, Z_k^l\right)}{\tau}\right)}$$

$$P(Y(X_i^{3-j})|do(X_i^{3-j})) = \sum_{t=1}^{n} \frac{\exp\left(\frac{\mathrm{sim}\left(Z_i^j, a_t \odot Z_i^{3-j}\right)}{\tau}\right) \times \frac{1}{n}}{\exp\left(\frac{\mathrm{sim}\left(Z_i^j, a_t \odot Z_i^{3-j}\right)}{\tau}\right) + \sum_{\substack{k=1,\\ k\neq i}}^{N} \sum_{\substack{l=1,\\ l\neq j}}^{2} \exp\left(\frac{\mathrm{sim}\left(Z_i^j, Z_k^l\right)}{\tau}\right)}$$

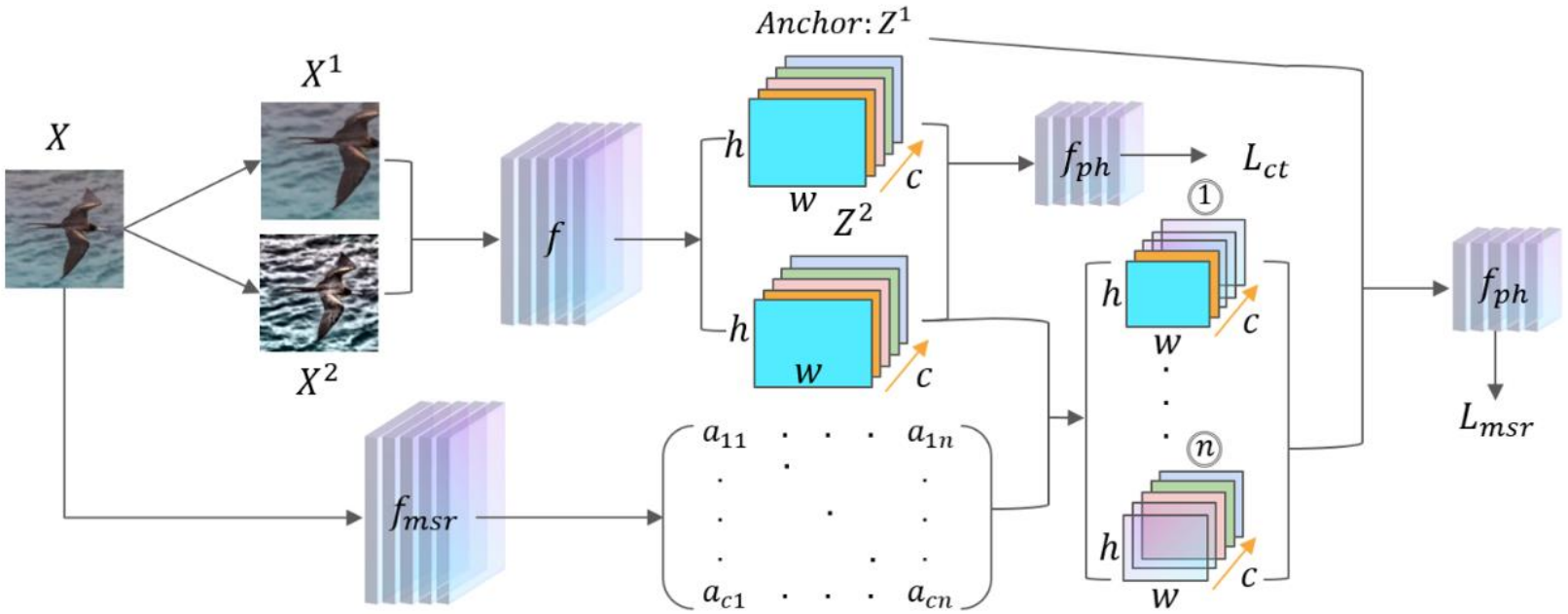
Meta Semantic Regularizer

- The meta semantic regularizer is trained alongside the feature extractor, with two stages per epoch



- In the first stage, f and f_{ph} are learned using the two augmented training set X_{tr}^{aug}, and the semantically relevant weight matrix A_s. In the second stage, f_{msr} is updated by computing its gradients with respect to the contrastive loss.

Meta Semantic Regularizer

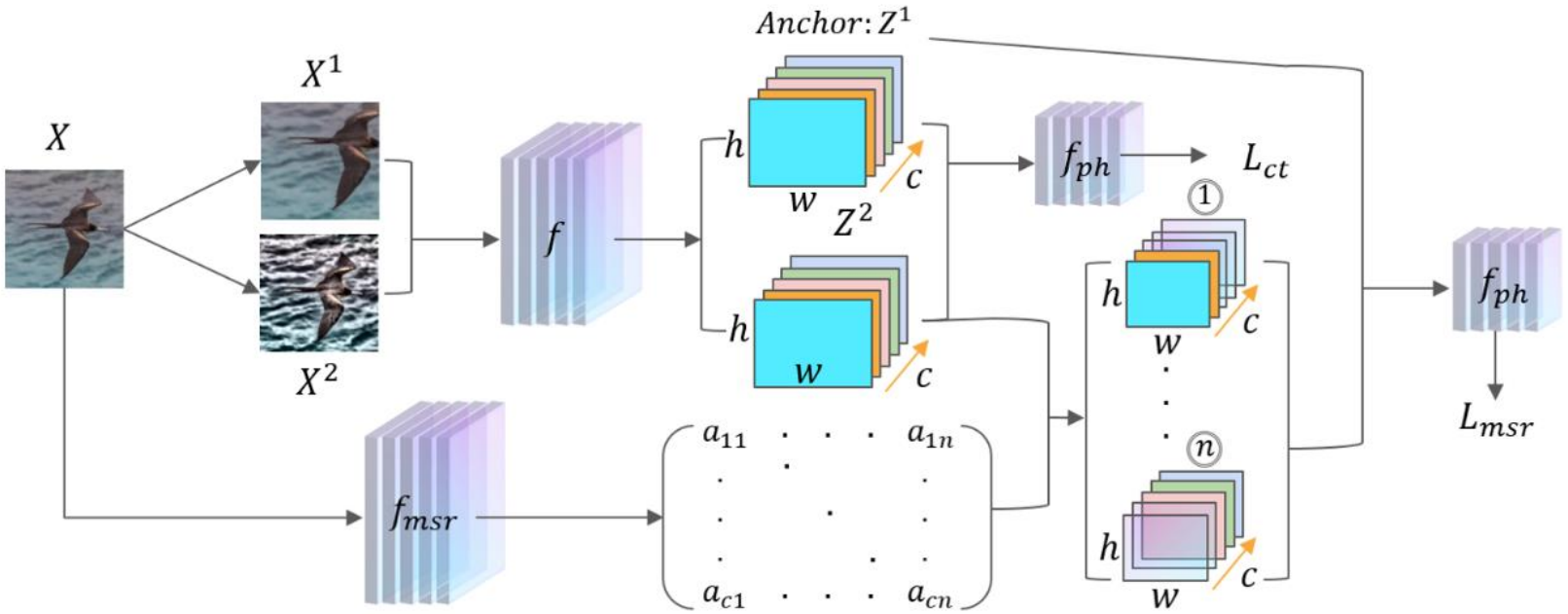


$$\min_{f, f_{ph}} L_{to} = L_{ct} + \lambda L_{msr}$$

$$L_{ct} = \sum_{i=1}^{N} \sum_{j=1}^{2} -\log \frac{\exp\left(\frac{\mathrm{sim}\left(Z_i^j, Z_i^{3-j}\right)}{\tau}\right)}{\sum_{k=1,}^{N} \sum_{l=1, l \neq j}^{2} \exp\left(\frac{\mathrm{sim}\left(Z_i^j, Z_k^l\right)}{\tau}\right)}$$

$$L_{msr} = \sum_{i=1}^{N} \sum_{j=1}^{2} -\log P\left(Y\left(X_i^{3-j}\right) | do\left(X_i^{3-j}\right)\right)$$

Meta Semantic Regularizer



$$f^1 = f - \alpha \nabla_f L_{to},$$
$$f_{ph}^1 = f_{ph} - \alpha \nabla_{f_{ph}} L_{to}$$

$$\min_{f_{msr}} L_{ct}\left(f^1, f_{ph}^1\right) + \gamma L_{uni}$$

$$L_{uni} = \log \sum_{a_i, a_j \in A_s} G_t\left(a_i, a_j, t\right)$$

$$G_t\left(a_i, a_j, t\right) \triangleq \exp\left(2t \cdot {a_i}^{\mathrm{T}} a_j - 2t\right)$$

Error Bound

- Downstream classification task
- Linear classifier; fine-tuning

Theorem 5.1. *Let* $f^* \in \arg\min_f L_{cl} + \lambda L_{msr}$. *Then with probability at least* $1-\delta$, *we have that*

$$\left|L_{SM}^{T}\left(f^*\right) - L_{cl}\left(f^*\right)\right| \leq O\left(\frac{Q_1 \mathscr{R}_H\left(\lambda\right)}{M} + \sqrt{\frac{Q_2}{M}}\right) \tag{10}$$

where M *is the total number of training samples,* N *is the size the mini-batch, and* $Q_1 = \sqrt{1+1/N}$, $\mathscr{R}_H\left(\lambda\right)$ *is the rademacher complexity. Also,* $\mathscr{R}_H\left(\lambda\right)$ *is monotonically decreasing w.r.t.* λ.

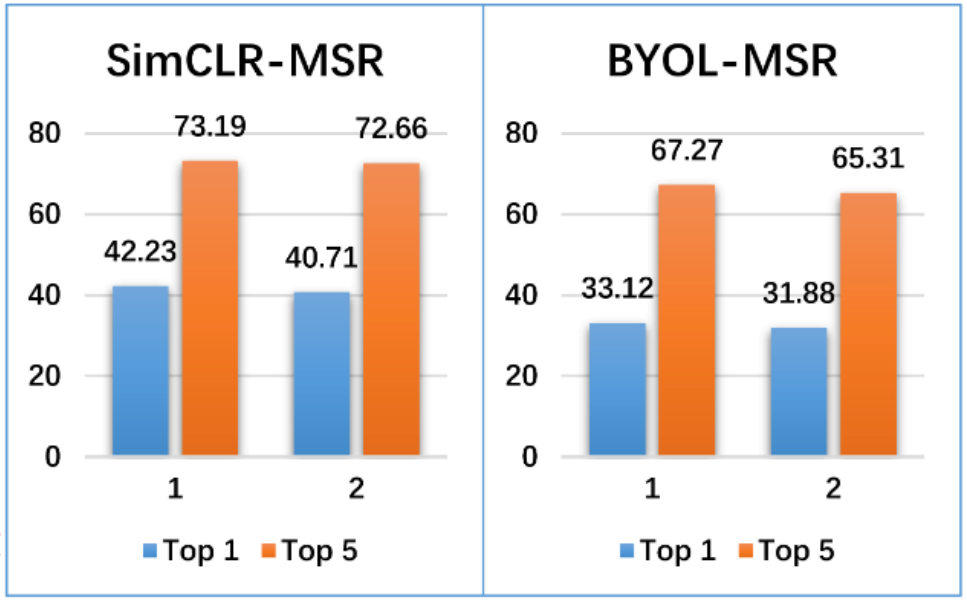
Evaluation

- Comparison with self-supervised learning methods

Methods	CIFAR-10		CIFAR-100		STL-10		Tiny ImageNet	
	linear	5-nn	linear	5-nn	linear	5-nn	linear	5-nn
SimCLR (Chen et al., 2020a)	91.80	88.42	66.83	56.56	90.51	85.68	48.84	32.86
BYOL (Grill et al., 2020)	91.73	89.45	66.60	56.82	91.99	88.64	51.00	36.24
W-MSE (Ermolov et al., 2021)	91.99	89.87	67.64	56.45	91.75	88.59	49.22	35.44
ReSSL (Zheng et al., 2021)	90.20	88.26	63.79	53.72	88.25	86.33	46.60	32.39
LMCL (Chen et al., 2021a)	91.91	88.52	67.01	56.86	90.87	85.91	49.24	32.88
SSL-HSIC (Li et al., 2021)	91.95	89.99	67.23	57.01	92.09	88.91	51.37	36.03
RELIC (Mitrovic et al., 2021)	91.96	89.35	67.24	56.88	91.15	86.21	49.17	32.97
ICL-MSR(SimCLR + MSR)	92.34	89.47	67.59	57.64	92.03	86.94	50.12	32.88
ICL-MSR(BYOL + MSR)	92.26	**90.12**	66.97	**57.97**	**93.22**	**89.36**	52.54	**37.54**
ICL-MSR(LMCL + MSR)	**92.45**	89.38	**67.99**	57.71	91.56	87.73	**52.61**	32.35
ICL-MSR(ReSSL + MSR)	91.77	89.06	65.12	55.07	89.91	88.06	47.17	33.03

Evaluation

- The experimental results for two kinds of ICL-MSR models



- 1: training and testing on full images
- 2: training on full images, and testing on foreground images

Contents

Summary

- What to contrast is important
- Learning informative samples to contrast via Meta Feature Augmentation
- Learning foreground to contrast via Interventional Meta Semantic Regularizer

Future/On-going work

- What to contrast?
- Hard/false negative/positive mining
- Uncertainty/distribution

- Contrasting structured data

参考文献

1. Wenwen Qiang, Jiangmeng Li, Changwen Zheng, Bing Su, and Hui Xiong. **“Interventional Contrastive Learning with Meta Semantic Regularizer”**, International Conference on Machine Learning (ICML), 2022.
2. Jiangmeng Li, Wenwen Qiang, Changwen Zheng, Bing Su, and Hui Xiong. **“MetAug: Contrastive Learning via Meta Feature Augmentation”**, International Conference on Machine Learning (ICML), 2022.
3. Bing Su and Ji-Rong Wen, **“Temporal Alignment Prediction for Supervised Representation Learning and Few-Shot Sequence Classification”**, International Conference on Learning Representations (ICLR), 2022.

Thanks!