How does Labeling Error Impact Contrastive Learning? A Perspective from Data Dimensionality Reduction

Jun Chen

Huazhong Agricultural University, Wuhan, China

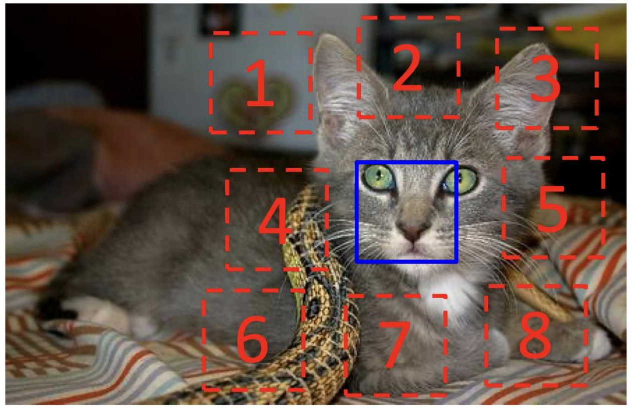
cj850487243@163.com

Jun. 2025

This work is jointed with Hong Chen, Yonghua Yu, and Yiming Ying.

Self-supervised Learning

- By context[1]



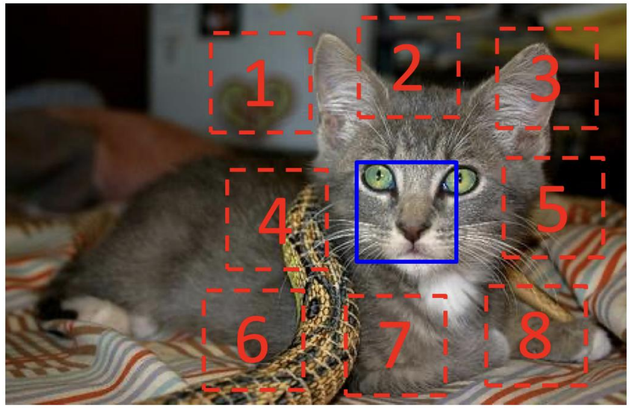
X = ( , ); Y = 3

[1] C. Doersch, A. Gupta, A. Efros. Unsupervised visual representation learning by context prediction. IEEE International Conference on Computer Vision (ICCV), 2015.

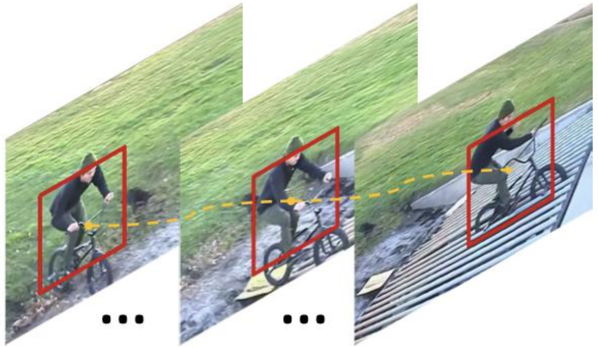
Backgrounds

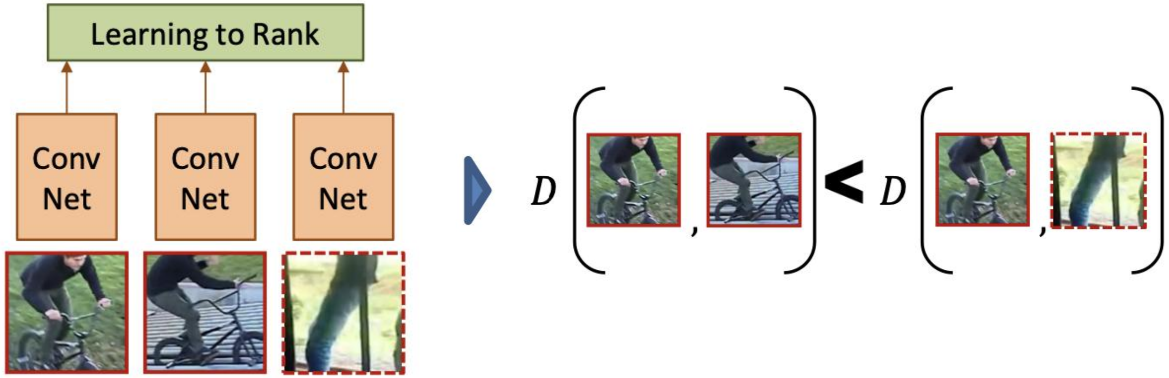
- **Self-supervised Learning**

  - By context[1]



  - By time series[2]



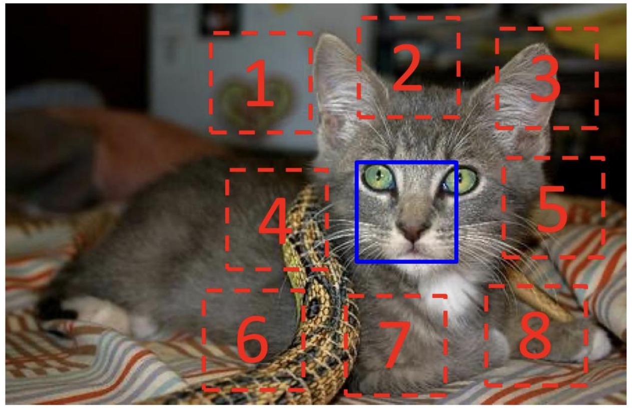


[1] C. Doersch, A. Gupta, A. Efros. Unsupervised visual representation learning by context prediction. IEEE International Conference on Computer Vision (ICCV), 2015.
[2] X. Wang, G. Abhinav. Unsupervised learning of visual representations using videos. IEEE International Conference on Computer Vision (ICCV), 2015: 2794-2802.

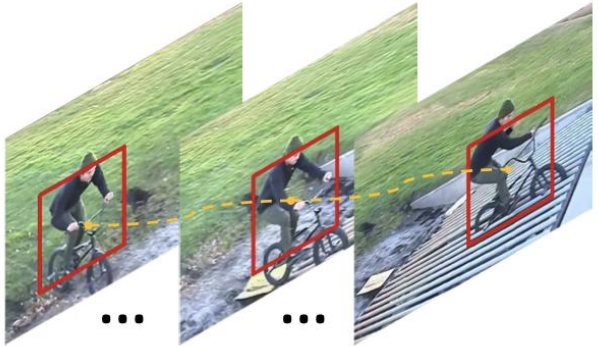
Backgrounds

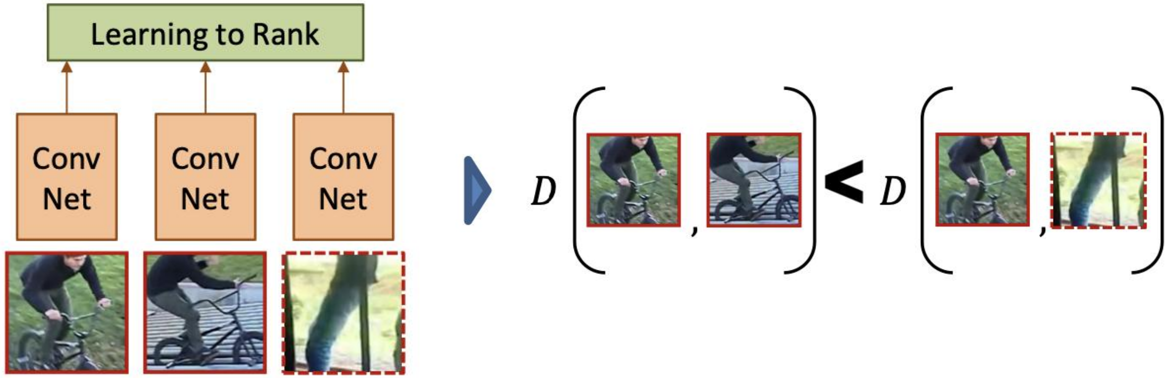
- **Self-supervised Learning**

  - By context[1]



  - By time series[2]



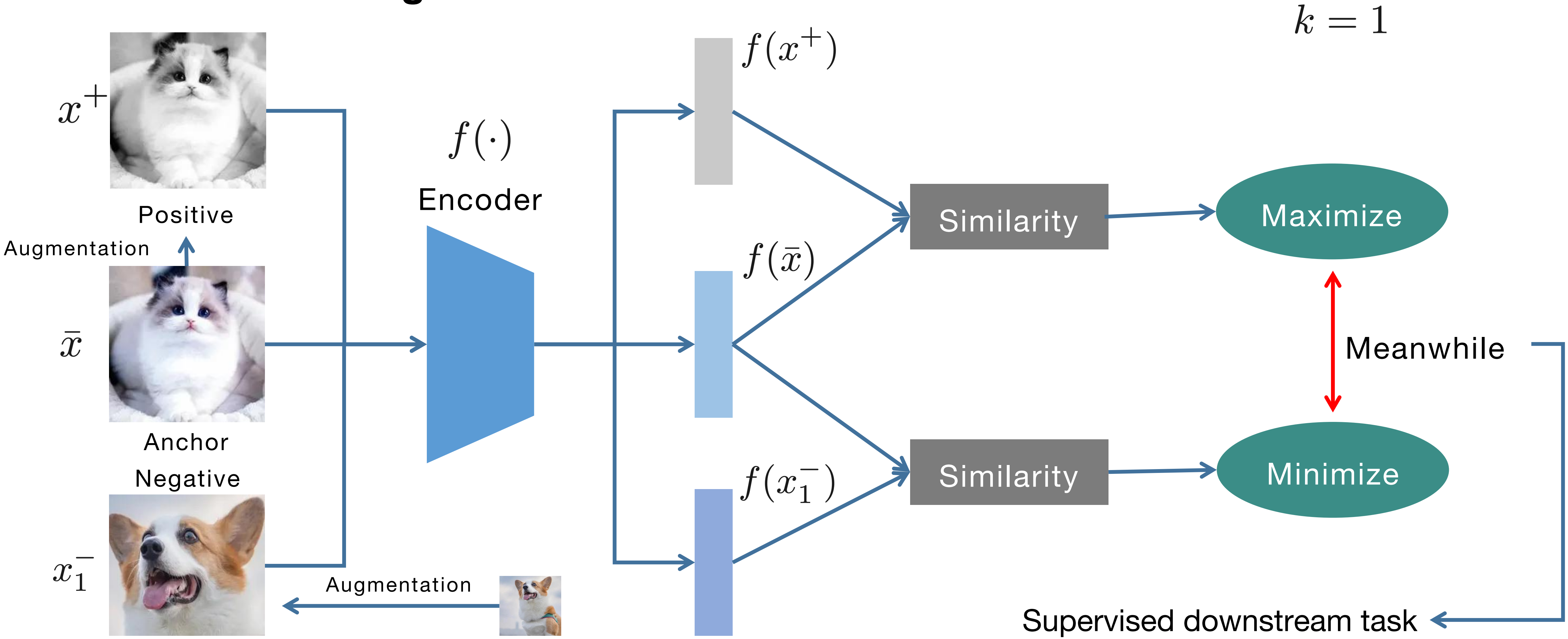


  - By contrastive[3]

[1] C. Doersch, A. Gupta, A. Efros. Unsupervised visual representation learning by context prediction. IEEE International Conference on Computer Vision (ICCV), 2015.
[2] X. Wang, G. Abhinav. Unsupervised learning of visual representations using videos. IEEE International Conference on Computer Vision (ICCV), 2015: 2794-2802.
[3] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations. ICML, 2020.

Backgrounds

- **Contrastive Learning**

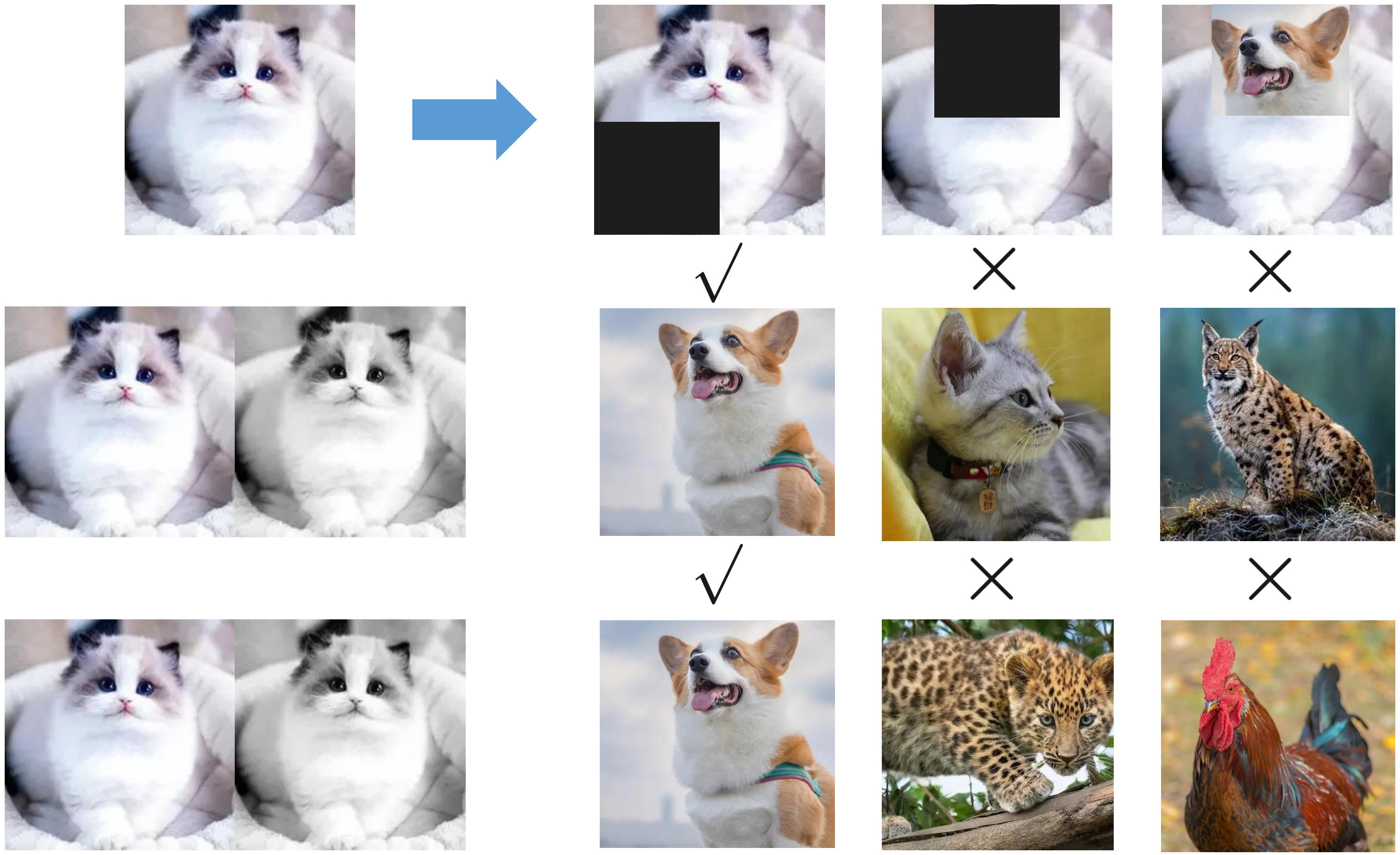


[3] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations. ICML, 2020.

Backgrounds

● Untrustworthy Phenomena

- False Positive Samples[4]
- False Negative Samples[5]
- Soft Negative Samples Mining[6]



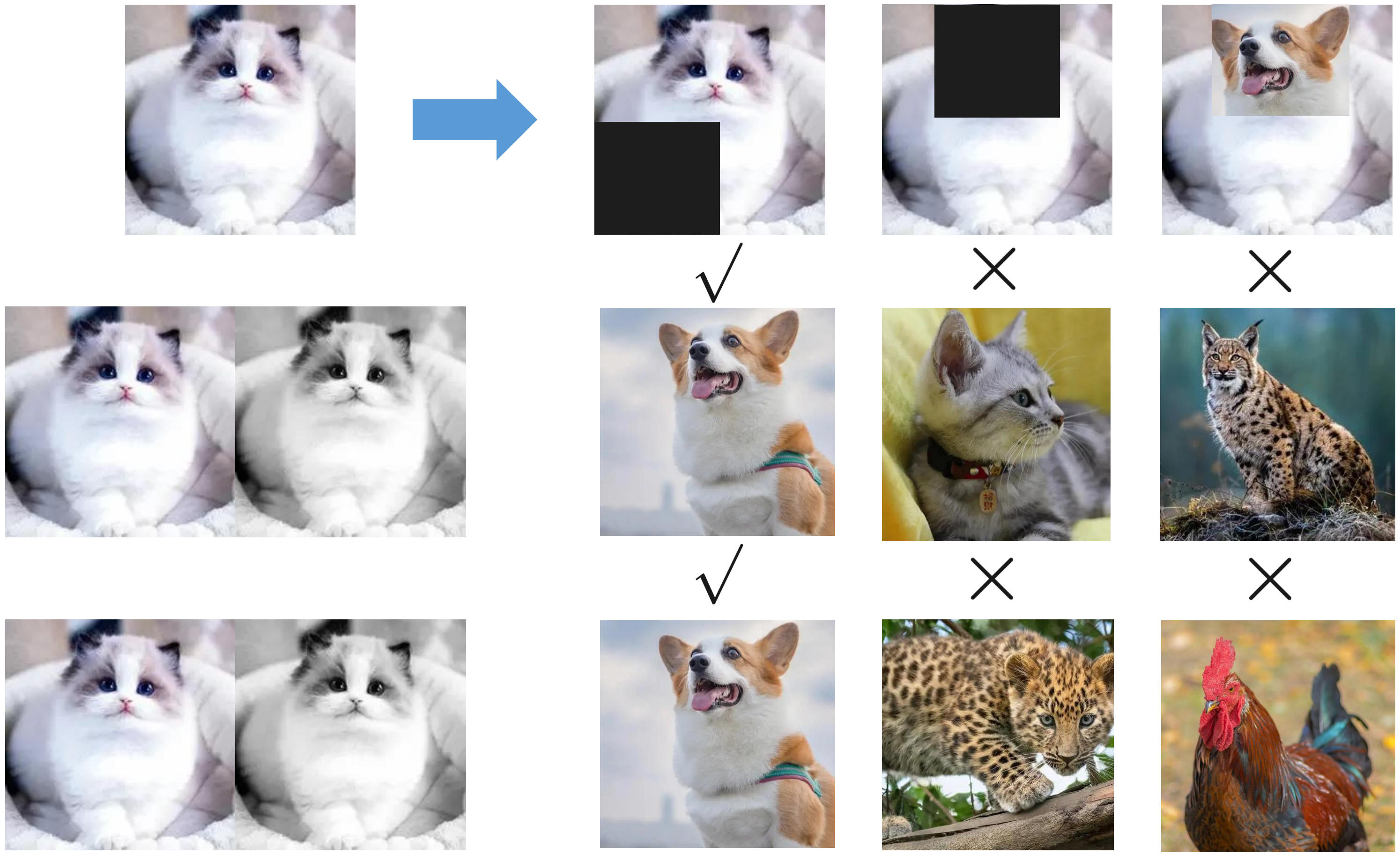
Positive Sample Pair   Negative Samples

[4] J. HaoChen, C. Wei, A. Gaidon, and T. Ma. Provable guarantees for self-supervised deep learning with spectral contrastive loss. NeurIPS, 2021.
[5] S. Arora, H. Khandeparkar, M. Khodak, O. Plevrakis, and N. Saunshi. A theoretical analysis of contrastive unsupervised representation learning. ICML, 2019.
[6] S. Lee, T. Park, and K. Lee. Soft contrastive learning for time series. ICLR, 2024.

Backgrounds

● Untrustworthy Phenomena

- **False Positive Samples[4]**
- False Negative Samples[5]
- Soft Negative Samples Mining[6]



Positive Sample Pair Negative Samples

[4] J. HaoChen, C. Wei, A. Gaidon, and T. Ma. Provable guarantees for self-supervised deep learning with spectral contrastive loss. NeurIPS, 2021.
[5] S. Arora, H. Khandeparkar, M. Khodak, O. Plevrakis, and N. Saunshi. A theoretical analysis of contrastive unsupervised representation learning. ICML, 2019.
[6] S. Lee, T. Park, and K. Lee. Soft contrastive learning for time series. ICLR, 2024.

Theoretical Impact of Labeling Error

False Positive Samples

- Augmentation overlap[7]

Intra-class overlap

Definition 1 (Augmentation Overlap)

Given a collection of augmentation strategies $\mathcal{T}$, we say that two original samples $\bar{x}, \bar{x}' \in \bar{\mathcal{D}}$ are $\mathcal{T}$-augmentation overlapped if they have overlapped views, i.e., $\exists t, t' \in \mathcal{T}$ such that $t(\bar{x}) = t'(\bar{x}')$.

Assumption (Label Consistency)[7]

For any $x, x^+ \sim p(x, x^+)$, we assume the labels are deterministic (one-hot) and consistent: $p(y|x) = p(y|x^+)$.

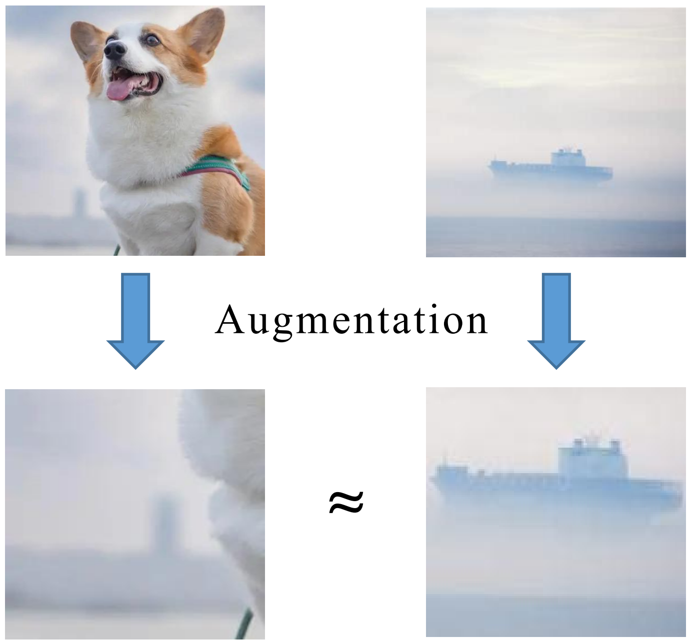
Without false positive samples

[7] Y. Wang, Q. Zhang, Y. Wang, J. Yang, and Z. Lin. Chaos is a ladder: A new theoretical understanding of contrastive learning via augmentation overlap. In International Conference on Learning Representations (ICLR), 2022.

Theoretical Impact of Labeling Error

False Positive Samples

- Augmentation overlap[7]



Inter-class overlap
(caused by false positive samples)

Assumption 1 (Labeling Error)

For any $\bar{x} = \bar{\mathcal{D}}$, its its latent label $y_{\bar{x}}$, and its augmented sample $x \sim p(\cdot|\bar{x})$, we assume that the true label of x is not consistent with $y_{\bar{x}}$ with the probability $\alpha \in (0, 1)$. That is,

$$\mathbb{E}_{\bar{x} \in \bar{\mathcal{D}}, x \sim p(\cdot|\bar{x})} \left[\mathbb{I} \left[y_x \neq y_{\bar{x}} \right] \right] = \alpha.$$

[7] Y. Wang, Q. Zhang, Y. Wang, J. Yang, and Z. Lin. Chaos is a ladder: A new theoretical understanding of contrastive learning via augmentation overlap. In International Conference on Learning Representations (ICLR), 2022.

Theoretical Impact of Labeling Error

- **Bound of Classification Risk**

Theorem 1 (Bounds of Mean Classification Risk)

Let the labeling error assumption hold. For any $f \in \mathcal{F}_1, g \in \mathcal{F}_2$, the gap between the mean downstream classification risk and the contrastive risk $\mathcal{L}_{CE}(g_{f,\mu}) + \log\left(\frac{M}{K}\right) - \mathcal{L}_{InfoNCE}(f)$ can be upper bounded by

$$\mathbb{E}_{p(x,y_{\bar{x}}^-)}\left[f(x)^\top \mu_{y_{\bar{x}}}\right] + \sqrt{V_{y_{\bar{x}}^-}(f(x)|y_{\bar{x}})} + \sqrt{V(f(x)|y_{\bar{x}})} + \mathcal{O}\left(M^{-\frac{1}{2}}\right)$$

and lower bounded by

$$\mathbb{E}_{p(x,x^+,y_{\bar{x}}^-)}\left[f(x)^\top f(x^+)\right] - \sqrt{V(f(x)|y_{\bar{x}})} - \frac{1}{2}V(f(x)|y_{\bar{x}}) - \frac{1}{2}V(f(x^-)|y^-) - \mathcal{O}\left(M^{-\frac{1}{2}}\right),$$

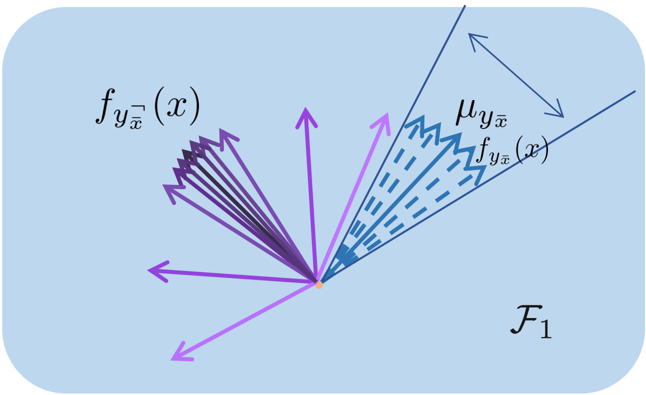
where $V_{y_{\bar{x}}^-}(f(x)|y_{\bar{x}}) = \mathbb{E}_{p(x,y_{\bar{x}}^-)}\left[\|f(x) - \mu_{y_{\bar{x}}}\|^2\right]$, $V(f(x)|y_{\bar{x}}) = \mathbb{E}_{p(x,y_{\bar{x}})}\left[\|f(x) - \mu_{y_{\bar{x}}}\|^2\right]$, $V(f(x^-)|y^-) =$ $\mathbb{E}_{p(x,y^-)}\left[\|f(x) - \mu_{y^-}\|^2\right]$ are the conditional intra-class variances of the representations of false positive, true positive and negative augmented samples, respectively.

Theoretical Impact of Labeling Error

- **Result Analysis**

$$\mathbb{E}_{p(x,y_{\bar{x}}^-)}\left[f(x)^\top \mu_{y_{\bar{x}}}\right]$$

$$\mathbb{E}_{p(x,x^+,y_{\bar{x}}^-)}\left[f(x)^\top f(x^+)\right]$$

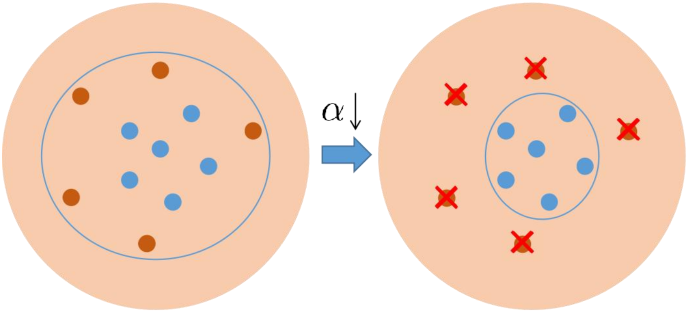


Relationship among f(x) and μ

$$V_{y_{\bar{x}}^-}(f(x)|y_{\bar{x}}) = \mathbb{E}_{p(x,y_{\bar{x}}^-)}\left[\|f(x) - \mu_{y_{\bar{x}}}\|^2\right]$$

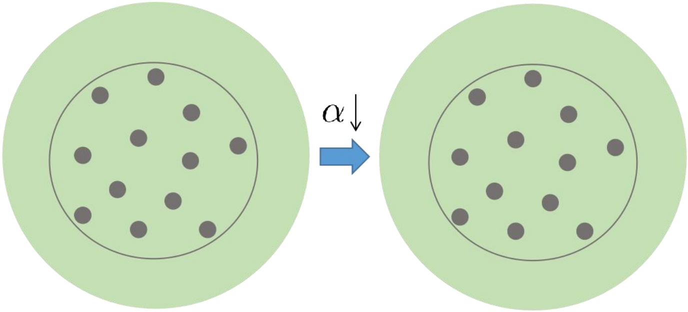
↓

$$V(f(x)|y)$$



Positive Augmented Samples

$$V(f(x^-)|y^-) = \mathbb{E}_{p(x,y^-)}\left[\|f(x) - \mu_{y^-}\|^2\right]$$



Negative Augmented Samples

Dimensionality Reduction as A New Perspective

- **Data Dimensionality Reduction (SVD)**

Definition 2 (Singular Value Decomposition)

For a matrix $X \in \mathbb{R}^{m \times m'}$ (without of loss generality, we let $m \leq m'$), its SVD equation is $X = USV^\top$, where $U = [u_1, ..., u_m] \in \mathbb{R}^{m \times m}$($V = [v_1, ..., v_{m'}] \in \mathbb{R}^{m' \times m'}$) is the left (right) singular matrix with $m(m')$ orthonormal column vectors (eigen vectors of $XX^\top(X^\top X)$) , $S = [diag(s_1, ..., s_m), \mathbf{0}]$ is composed of a diagonal matrix $diag(s_1, ..., s_m) \in \mathbb{R}^{m \times m}$ and a zero matrix $\mathbf{0}$ with size $m \times (m' - m)$, s_i denotes the i-th largest singular value, $s_1 \geq s_2 \geq ... \geq s_m \geq 0$.

[8] C. Eckart and G. Young. The approximation of one matrix by another of lower rank. Psychometrika, 1:211–218, 1936.

Dimensionality Reduction as A New Perspective

Data Dimensionality Reduction (SVD)

Lemma 1 (Eckart-Young Theorem[8])

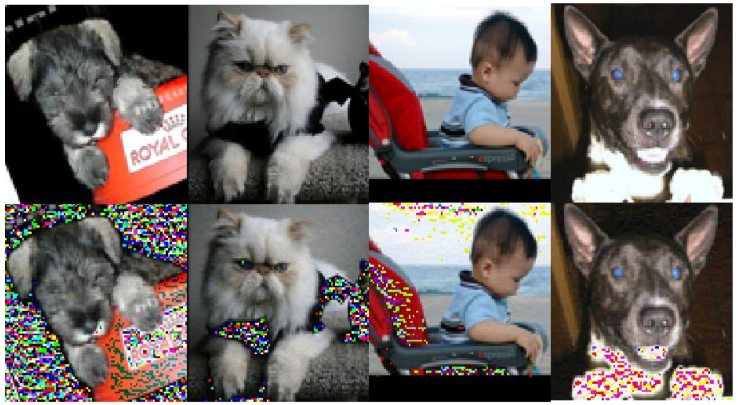
Let X be a $m \times m'$ matrix of rank $r \in [m]$ which has complex elements. Let P_q be the set of all $m \times m'$ matrices with rank $q \in [r]$. Then for all matrices B in P_q, there holds $\left\|X - \hat{X}_q\right\|_F \leq \|X - B\|_F$.

- Eckart-Young Theorem implies that the majority of the informational content is captured by the dominant singular subspace[9].
- We assume by default that there is a positive correlation between the amount of information and the semantical relevance of information.

[8] C. Eckart and G. Young. The approximation of one matrix by another of lower rank. Psychometrika, 1:211–218, 1936.
[9] M. Kilmer, L. Horesh, H. Avron, and E. Newman. Tensor-tensor algebra for optimal representation and compression of multiway data. Proceedings of the National Academy of Sciences, 118, 2021.

Dimensionality Reduction as A New Perspective

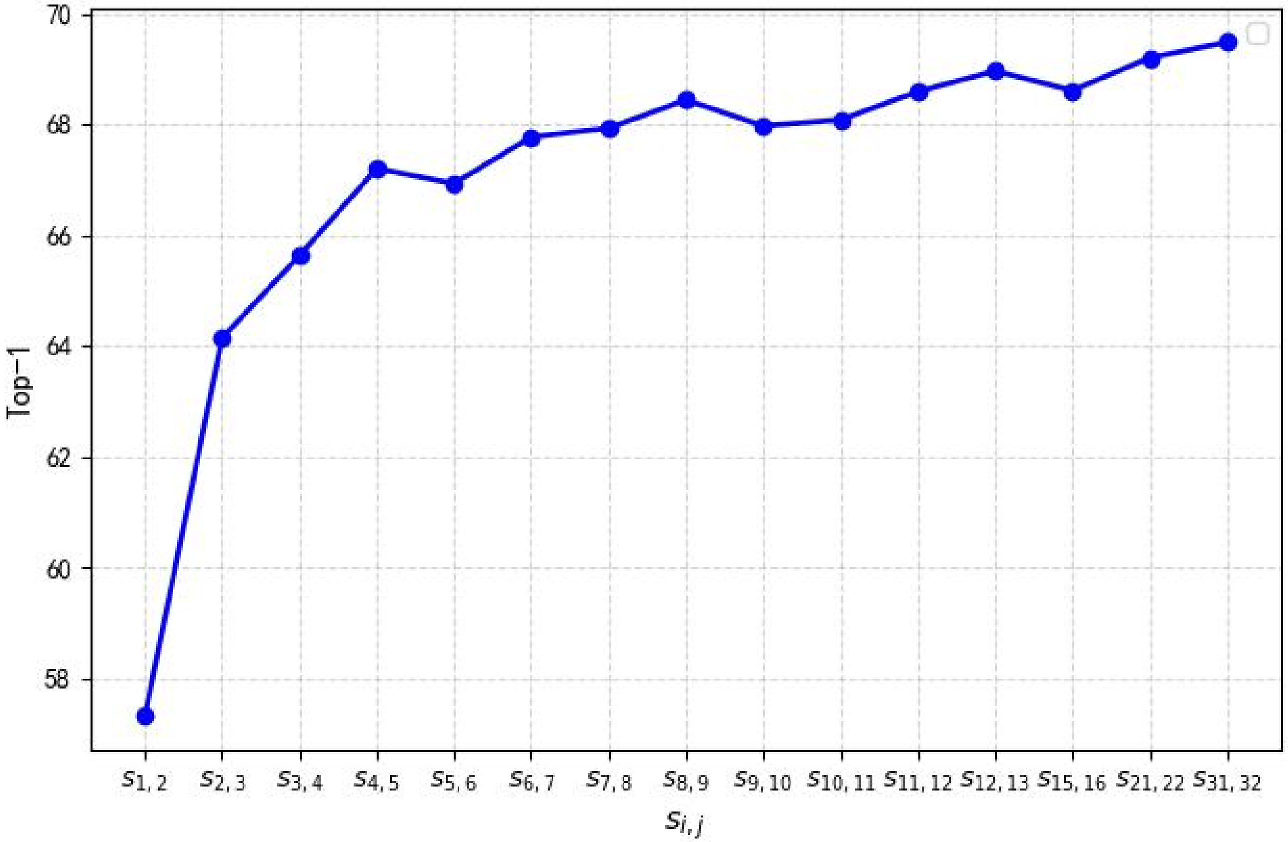
- **Data Dimensionality Reduction (SVD)**
  - STL-10



Raw Images

after taking SVD

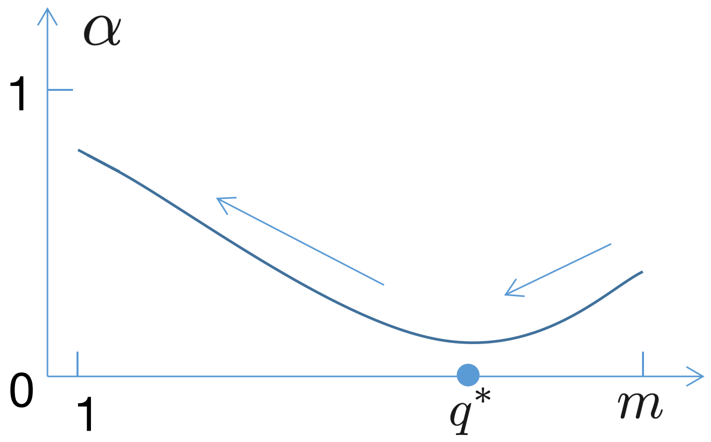
  - CIFAR-10



Dimensionality Reduction as A New Perspective

Proposition 2

Let a sample and the corresponding sample after SVD be represented as the matrices $X, \hat{X}_q \in \mathbb{R}^{m \times m'}$. Assume that there are q^* singular values regrading the semantics-related information. When $q \geq q^*$, under the assumption of labeling error and the augmentation collection $\mathcal{T}$, the true label of the augmented sample of $\hat{X}_q$ is not consistent with the latent label of X with the probability $\alpha_q \leq \alpha$ .When $q < q^*$, the corresponding probability $\alpha_q > \alpha_{q^*}$.



[9] M. Kilmer, L. Horesh, H. Avron, and E. Newman. Tensor-tensor algebra for optimal representation and compression of multiway data. Proceedings of the National Academy of Sciences, 118, 2021.

Dimensionality Reduction as A New Perspective

Assumption 2

Let the assumption of labeling error hold. When performing SVD with the truncated value q the encoder f with the empirical InfoNCE loss $\hat{\mathcal{L}}_{InfoNCE}(f)$ can align any positive sample pair $(x, x^+) \sim p(x, x^+, y_{\bar{x}}^{\neg})$ such that their distance in the embedding space lies within $[\epsilon(\alpha_{q^*}), \epsilon(\alpha_q)]$. For simplicity, let $\epsilon_{q^*} = \epsilon(\alpha_{q^*}), \epsilon_q = \epsilon(\alpha_q)$. Consequently, the alignment satisfies $\epsilon_{q^*} \leq \|f(x) - f(x^+)\| \leq \epsilon_q$.

Theorem 3

Given the condition of Theorem 1 and Assumption 2, after taking the optimal truncated SVD on the original dataset $\bar{\mathcal{D}}$, the mean downstream classification risk $\mathcal{L}_{CE}(g_{f,\mu})$ with the empirical optimal encoder f can be upper bounded by $\mathcal{L}_{InfoNCE}(f) + \epsilon_{q^*} + \epsilon_q - \frac{1}{2}\epsilon_{q^*}^2 + \mathcal{O}\left(M^{-\frac{1}{2}}\right) - \log\left(\frac{M}{eK}\right)$ and lower bouned by

$$\mathcal{L}_{InfoNCE}(f) - \epsilon_{q^*} - \epsilon_{q^*}^2 - \frac{1}{2}\epsilon_q^2 - \mathcal{O}\left(M^{-\frac{1}{2}}\right) - \log\left(\frac{M+1}{K}\right).$$

Dimensionality Reduction as A New Perspective

- **Experimental Results**

Table 2. Downstream classification top-1 accuracies (%) of SimCLR ($\mathcal{L}_{InfoNCE}$) using the truncated SVD with different truncated parameter q.

$\mathcal{T}$	Encoder	Dataset	w/o SVD	$q = 30$	$q = 25$	$q = 20$	$q = 15$	$q = 10$
$\mathcal{T}_1$	Resnet-18	CIFAR-10	68.82	69.48	69.75	**69.87**	69.01	68.26
$\mathcal{T}_1$	Resnet-50	CIFAR-10	63.20	63.36	**63.96**	62.23	60.97	60.06
RRC	Resnet-18	CIFAR-10	58.56	**58.83**	58.67	58.61	58.54	58.32
$\mathcal{T}_1$	Resnet-18	CIFAR-100	38.48	38.81	**40.10**	39.05	38.98	38.10
$\mathcal{T}$	Encoder	Dataset	w/o SVD	$q = 90$	$q = 70$	$q = 50$	$q = 30$	$q = 10$
$\mathcal{T}_1$	Resnet-18	STL-10	71.54	**73.12**	72.29	71.10	70.04	67.52

Table 4. Downstream classification top-1 accuracies (%) of SimCLR ($\mathcal{L}_{InfoNCE}$) on CIFAR-10 using the truncated SVD with different augmentations ($\mathcal{T}_2 = \{\mathcal{T}_1 +$ Cutout$\}$; $\mathcal{T}_3 = \{$RRC, Cutout, Hide patch$\}$; $\mathcal{T}_4 = \{$RRC, Cutout, Color jitter$\}$; $\mathcal{T}_5 = \{$RRC, Cutout$\}$; $\mathcal{T}_6 = \{$RRC(0.08, 0.5), Cutout$\}$; $\mathcal{T}_7 = \{$RRC(0.08, 0.5), Cutout(0.5, 1.0)$\}$).

SVD	Encoder	$\mathcal{T}_2$	$\mathcal{T}_3$	$\mathcal{T}_4$	$\mathcal{T}_5$	$\mathcal{T}_6$	$\mathcal{T}_7$	RRC(0.08,0.5)
w.o. SVD	Resnet-18	62.90	50.53	60.00	56.67	54.97	54.09	57.11
$q = 30$	Resnet-18	**64.86**	**51.00**	**61.57**	**57.85**	**55.69**	**54.75**	**58.10**

Further Understanding of Labeling Error

Definition 3 (Augmentation Graph [4])

Given an original dataset $\bar{\mathcal{D}}$ and an augmentation collection $\mathcal{T}$, there exist n augmented samples that form the augmentation dataset $\mathcal{D}_{aug} = \{x | x = t(\bar{x}), \bar{x} \in \bar{\mathcal{D}}, t \in \mathcal{T}\}$. An augmentation graph $\mathcal{G}$ is obtained by taking the n augmented samples as the graph vertices and assuming there exists an edge between two vertices $x, x' \in \mathcal{D}_{aug}$ (if they can be generated from a random original sample $\bar{x} \in \bar{\mathcal{D}}$.)

According to spectral graph theory, we define $A \in \mathbb{R}^{n \times n}$ as the adjacency matrix of the augmentation graph $\mathcal{G}$. For two augmented samples $x, x' \in \mathcal{D}_{aug}$, the element $A(x, x')$ denotes the marginal probability of generating x, x' from a random original sample $\bar{x} \in \bar{\mathcal{D}}$. Formally, $A(x, x') = \mathbb{E}_{\bar{x} \in \bar{\mathcal{D}}} [p(x|\bar{x}) p(x'|\bar{x})]$.
The corresponding normalized graph Laplacian matrix is $L = I - D^{-\frac{1}{2}} A D^{-\frac{1}{2}}$, where D represents a diagonal degree matrix with the diagonal element $D_{x,x} = \sum_{x' \in \mathcal{D}_{aug}} A(x, x')$. The eigenvalues of L are denoted as $\{\lambda_i\}_{i=1}^{n}$, where $0 = \lambda_1 \leq ... \leq \lambda_n \leq 2$.

[4] J. HaoChen, C. Wei, A. Gaidon, and T. Ma. Provable guarantees for self-supervised deep learning with spectral contrastive loss. NeurIPS, 2021.

Further Understanding of Labeling Error

Definition 3 (Augmentation Graph)

Given an original dataset $\bar{\mathcal{D}}$ and an augmentation collection $\mathcal{T}$, there exist n augmented samples that form the augmentation dataset $\mathcal{D}_{aug} = \{x | x = t(\bar{x}), \bar{x} \in \bar{\mathcal{D}}, t \in \mathcal{T}\}$. An augmentation graph $\mathcal{G}$ is obtained by taking the n augmented samples as the graph vertices and assuming there exists an edge between two vertices $x, x' \in \mathcal{D}_{aug}$ (if they can be generated from a random original sample $\bar{x} \in \bar{\mathcal{D}}$.)

Theorem 4

Let the assumption of labeling error hold. For the empirical optimal encoder f^*, after taking the truncated SVD with hyper-parameter q on the original dataset $\bar{\mathcal{D}}$, there exists a linear head W with norm $\|W^*\|_F \leq 1/(1 - \lambda_{k,q})$ such that

$$\mathcal{E}(f^*, W^*) \leq \frac{4\alpha_q \downarrow}{\lambda_{k+1,q} \downarrow} + 8\alpha_q \downarrow$$

Maybe $\lambda_{k+1,q} \leq \lambda_{k+1}$

where k denotes the dimension of embedding space and $\lambda_{k+1,q}$ denotes the $k+1$-th eigenvalues of L.

Further Understanding of Labeling Error

- **Augmentation Suggestion**

  - Wang et al,.[10] suggested：Weak augmentation + Data inflation
  - We suggest：Weak augmentation + Data inflation + SVD

Table 5. Downstream classification top-1 accuracies (%) of SimCLR ($\mathcal{L}_{spe}$) on CIFAR-10 using the truncated SVD with different q or the data inflation strategy under the weak data augmentation adopted by Wang et al. (2024) ($\mathcal{T}_8 = \{$RRC(0.2, 1.0), Color jitter(0.5, 0.4), Random horizontal flip, Random grayscale, Gaussian blur$\}$).

$\mathcal{T}$	Encoder	Inflation	w/o SVD	$q = 30$	$q = 25$	$q = 20$	$q = 15$	$q = 10$
$\mathcal{T}_8$	Resnet-18	71.54	71.21	71.64	**71.65**	71.11	70.41	67.83

$\mathcal{T}$	Encoder	Inflation	Inflation + ($q = 30$)	Inflation + ($q = 25$)	Inflation + ($q = 20$)
$\mathcal{T}_8$	Resnet-18	71.54	71.64	**72.55**	71.19

[10] Y. Wang, J. Zhang, and Y. Wang. Do generated data always help contrastive learning? In International Conference on Learning Representations (ICLR), 2024.

Further Understanding of Labeling Error

Augmentation Suggestion

- Wang et al,.[10] suggested：Weak augmentation + Data inflation
- We suggest：Weak augmentation + Data inflation + SVD + moderate embedding dimension

Table 7. Downstream classification top-1 accuracies (%) of SimCLR ($\mathcal{L}_{spe}$) using the truncated SVD ($q = 30$ for CIFAR-10 and CIFAR-100, $q = 90$ for STL-10) with different embedding dimension k.

$\mathcal{T}$	Encoder	Dataset	Embedding Dimension				
			$k = 128$	$k = 256$	$k = 512$	$k = 1024$	$k = 2048$
$\mathcal{T}_1$	Resnet-18	CIFAR-10	67.71	68.51	68.54	**69.09**	68.65
$\mathcal{T}_1$	Resnet-50	CIFAR-10	**67.43**	65.99	66.50	66.83	66.22
$\mathcal{T}_1$	Resnet-18	CIFAR-100	35.00	36.68	36.78	**37.78**	37.18
$\mathcal{T}_1$	Resnet-50	CIFAR-100	35.46	35.42	35.39	**35.59**	35.53
$\mathcal{T}_1$	Resnet-18	STL-10	72.35	72.42	73.12	**73.88**	73.47
$\mathcal{T}_1$	Resnet-50	STL-10	74.68	74.94	75.01	**76.26**	75.57

[10] Y. Wang, J. Zhang, and Y. Wang. Do generated data always help contrastive learning? In International Conference on Learning Representations (ICLR), 2024.

Thanks

Jun Chen

Huazhong Agricultural University, Wuhan, China

cj850487243@163.com

Jun. 2025