Learning the Unlearned: Mitigating Feature Suppression in Contrastive Learning

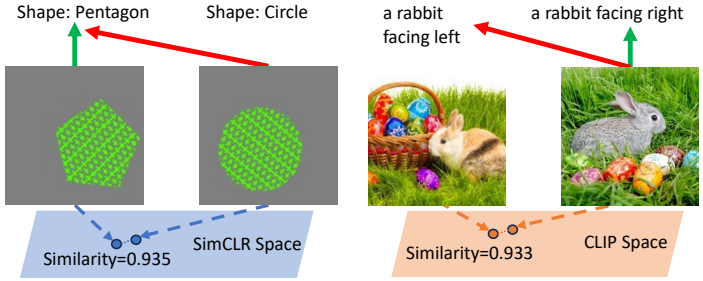
Jihai Zhang *,1, Xiang Lan *,2, Xiaoye Qu[3], Yu Cheng[1], Mengling Feng[2], Bryan Hooi[2]

[1]The Chinese University of Hong Kong [2]National University of Singapore [3]Shanghai AI Laboratory

*Equal Contribution

Paper Code

1. Feature suppression harms contrastive learning



Feature suppression in unimodal and multimodal contrastive learning.

Feature	SimCLR	MoCo-v2
C-M(CIFAR)	0.10	0.77
C-M(MNIST)	0.99	0.98
Trifeature(Shape)	0.44	0.85
Trifeature(Texture)	0.92	0.99
Trifeature(Color)	1.00	1.00

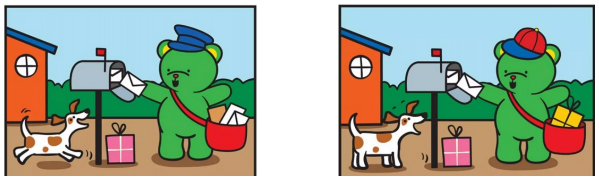
Attribute	Accuracy	Attribute	Accuracy
O&D	26.7	**C&A**	40.0
PSF	13.3	**S&P**	26.7
S&C	26.7	**Texts**	13.3
Q&C	6.7	**V&P**	20.0
P&R	6.7	Average	20.0

Unimodal: Image from Trifeature with different shapes and textures have high similarity in the SimCLR space. In Trifeature, SimCLR significantly ignores the shape information, and in CIFAR-MNIST, it almost completely neglects the CIFAR information.

Multimodal: Images from MMVP with different semantics have high similarity in the CLIP space. The performance of CLIP on MMVP benchmark is low on most of the attributes.

🤔 The trained model captures only a limited portion of the information from the input data while overlooking other potentially valuable content.

2. Concept of our solution

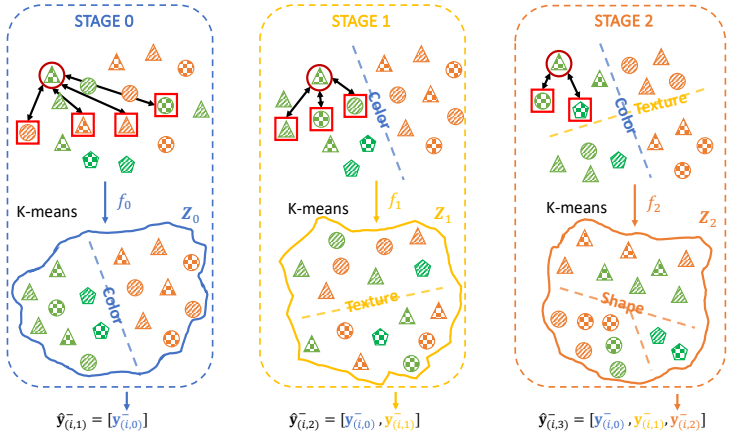


Can you find any difference of the two images.

EYE: The hats on the bear!
EYE: The contents in the bucket!
EYE: The gestures of the dog!
BRAIN: Any difference **EXCEPT FOR** the hat?
BRAIN: Any difference **EXCEPT FOR** the hat and the bucket?
BRAIN: ...

💡 We "learn" for multiple stages: at each stage, we force the model to learn the previously unlearned information.

3. Multistage Contrastive Learning (MCL)



Overview of MCL:

1. An initial model is trained, and its output representations are then clustered.
2. At each subsequent stage, cluster assignments from the prior stages are used to create a pseudo label.
3. During training at each stage, negative samples are selected from those sharing the same pseudo label as the anchor.
4. The final representations are formed by concatenating outputs from each training stage.

Advantages of MCL:

- Facilitates the learning of new features.
- Simple, effective, and efficient.
- Easily adaptable across various frameworks.
- Offers benefits in both uni/multimodal contexts.

4. Experiments across unimodal and multimodal contrastive learning

Performance of the MCL trained SimCLR and the baselines.

	Trifeature			CelebA			CIFAR-MNIST		
	Shape	Texture	Color	Hair	Smiling	Gender	CIFAR	MNIST	Average
IFM	0.99	0.99	**1.00**	0.82	0.70	0.93	0.11	**0.99**	0.82
TS	0.93	**1.00**	**1.00**	0.62	0.65	0.87	0.09	**0.99**	0.77
FD	0.75	0.73	0.78	0.81	**0.89**	0.94	0.78	0.86	0.82
SimCLR	0.81	0.99	**1.00**	0.84	0.75	0.94	0.29	0.98	0.83
MCL	**1.00**	**1.00**	**1.00**	**0.85**	0.79	**0.95**	**0.87**	**0.99**	**0.93**

Performance of the MCL tuned CLIP models (MMVP benchmark).

	O&D	PSF	S&C	Q&C	P&R	C&A	S&P	Texts	V&P	Average
ResNet STAGE0	6.7	6.7	**46.7**	0.0	**13.3**	46.7	**33.3**	6.7	13.3	19.3
ResNet STAGE1	**13.3**	**13.3**	40.0	**20.0**	0.0	33.3	33.3	6.7	26.7	20.7
ResNet STAGE2	6.7	6.7	33.3	6.7	6.7	**73.3**	13.3	13.3	**33.3**	21.5
ResNet STAGE3	0.0	13.3	33.3	20.0	0.0	60.0	33.3	**26.7**	20.0	23.0
ResNet MCL	0.0	13.3	40.0	6.67	6.67	73.3	40.0	6.7	33.3	<u>24.4</u>
ViT STAGE0	**26.7**	13.3	26.7	6.7	6.7	40.0	26.7	**13.3**	20.0	20.0
ViT STAGE1	13.3	**33.3**	**66.7**	26.7	**13.3**	53.3	20.0	13.3	**26.7**	29.6
ViT STAGE2	0.0	13.3	46.7	**40.0**	6.7	53.3	20.0	13.3	20.0	23.7
ViT STAGE3	13.3	6.67	46.7	13.3	13.3	**66.7**	**40.0**	13.3	20.0	25.9
ViT MCL	6.67	20.0	73.3	13.3	13.3	80.0	46.7	13.3	26.7	<u>32.6</u>

Performance of the MCL tuned CLIP Model Soup on retrieval.

	COCO						Flickr30k					
	Image-to-Text			Text-to-Image			Image-to-Text			Text-to-Image		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
OpenAI	56.1	79.5	86.8	35.4	60.1	70.2	48.5	72.6	80.8	28.0	49.3	58.7
MCL MS	60.4	82.4	89.5	42.2	66.9	76.2	52.7	76.6	84.4	36.5	59.0	68.1

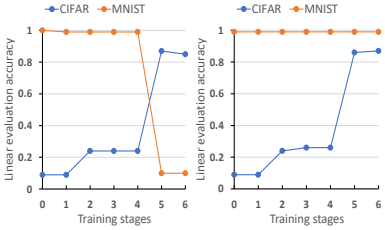
AMI scores between the K-means clusters in different MCL stages.

AMI	Stage 0	Stage 1	Stage 2
Stage 0	1.00	0.33	0.25
Stage 1	0.33	1.00	0.24
Stage 2	0.25	0.24	1.00

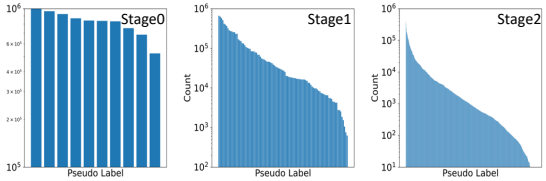
Linear evaluation accuracy of MCL trained SimCLR on ImageNet.

	Stage0	Stage0+1	Stage0+1+2
MCL	0.456	0.477	0.482

Focus shifting in the MCL learning process.



Distribution of pseudo labels across different MCL stages.



Visualization of features learned at different MCL stages (showing the three most similar samples to the anchor).