Paper Reading Session
Contrastive Learning meets Masked Modeling [1]

Lei Wang[1,2]

[1]Australian National University
[2]Data61/CSIRO

June 2, 2023

[1]Inspired by Dr. Liang's TPAMI'18 paper 'SIFT Meets CNN: A Decade Survey of Instance Retrieval'.

Table of Contents[2]

Contents

[2]The materials presented in this paper reading session are based on papers published in top venues *e.g.*, CVPR, ICLR, JMLR with google citations > 1000.

Widely used self-supervised learning methods

Contrastive Learning (CL)

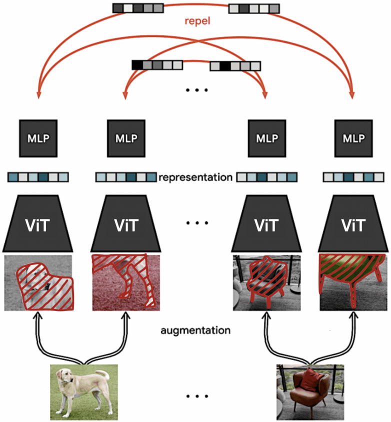


Figure 1: SimCLR[a].

[a] Chen *et al.* "A simple framework for contrastive learning of visual representations." ICLR'20.

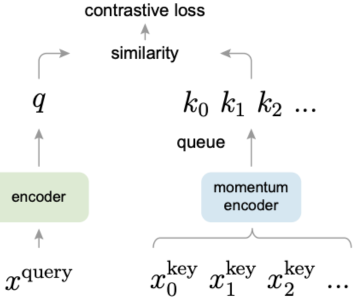


Figure 2: MoCo[a].

Image-level approach:

- learn invariant semantics of two random views (explore global repre. to contrast)
- make globally projected repre. sim./dissim. for pos./neg. samples

[a] He *et al.* "Momentum contrast for unsupervised visual representation learning." CVPR'20.

Masked Modeling (MM)

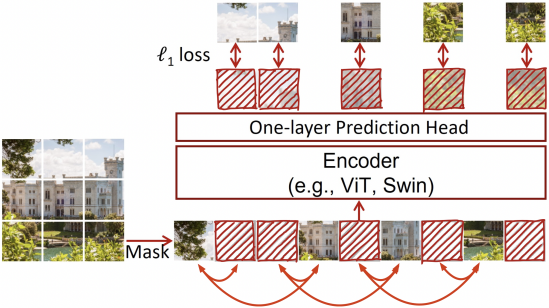


Figure 3: SimMIM[3].

Deviating from **CL**, token-level approach:

- a strong competitor / impressive performances of downstream tasks
- *e.g.*, Masked Image Modeling (MIM/**MM**)
    - reconstruct the correct semantics of masked input patches
    - learn the semantics of patch tokens, unlike **CL**
    - outperform **CL** in finetuning acc./a more effective pretraining method than **CL**

[3]Xie *et al.* "Simmim: A simple framework for masked image modeling." CVPR'22.

MM (cont.)

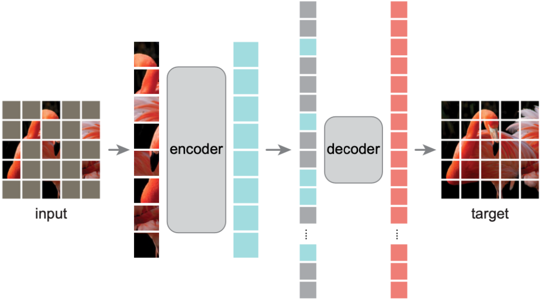


Figure 4: MAE architecture[4].

Token-level approach, *e.g.*, masked autoencoders (MAE):

- a large random subset of patches is masked out
- encoder is applied to the small subset of visible patches
- masked tokens are introduced after the encoder
- the full set of encoded patches & masked tokens are processed by a decoder
- reconstruct the original image in pixels (loss only on masked patches)

[4] He *et al.* "Masked autoencoders are scalable vision learners." CVPR'22.

CL *vs.* MM

Which method, **CL** or **MM**, for self-supervised learning of ViTs[5]?

- Observations/little is known about what they learn:
  - To better understand self-superv. & can potentially affect future improv.)
  - Both methods are widely used
  - **MM** outperforms **CL** in **finetuning**/dense prediction tasks[6] with **large models**
  - **CL** works well for **linear probing**[7]/classification tasks with **small models**

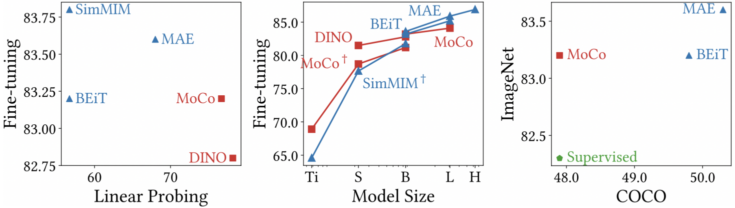


Figure 5: **CL** *vs.* **MM** (outperform/underperform & superior scalability / downstream dense pred. *e.g.*, OD with Mask R-CNN on COCO)[8].

[5] Dosovitskiy *et al.* "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale." ICLR'21

[6] Learn a mapping from input images to complex output structures *e.g.*, SS, DE, OD, PL, *etc.*

[7] Linear classifiers, a probe uses the hidden units of a given intermed. layer as feat., these probes cannot affect the training phase of model & generally added after training

[8] Park *et al.* "What Do Self-Supervised Vision Transformers Learn?" ICLR'23.

CL *vs.* MM (cont.)

CL and **MM** have advantages over different tasks, key components different?

- architecture (early layer → low-level info., later layer → high-level info.)
- self-attention (global / local relationships)

Figure 6: Perth Lights[9].
Image-level (global rep.) *vs.* token-level (patch semantics)

- representation (shape-/texture-oriented, low-/high-frequency, different levels of detail, token-level info. preserved?)

(a) Low-freq. (shapes)

(b) High-freq. (texture)

[9] This photo was captured by Lei Wang on 21/07/2019 in Perth CBD.

Comparisons & Discussions

Architecture: early or later layers

- Early layers: low-level features, *e.g.*,
  - local patterns, texture info. & high frequency signals
- Later layers:
  - global patterns, shape info. & low frequency signals

- Which component matters?
  - measure linear probing acc. using intermediate repre.
  - **CL** & **MM** exploit global & local patterns
  - Later layer of **CL** & early layer of **MM**?
    - linear probing acc. of **MM** > **CL** at the beginning
    - **CL** outperforms **MM** at the end of the model
    - acc of **CL** ↑ with depth ↑
    - acc of **MM** ↓ at the end of model (later layers are not helpful in separating repre.)
    - Later layer of **CL** & early layer of **MM** play an important role in making linearly separable repre.
    - shallow pred. head impairs performance / explicit decoder (*e.g.*, reconstruct masked tokens) helps ViTs

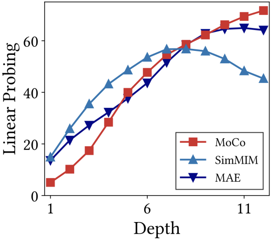


Figure 8: Linear probing acc. of rep. of intermediate layers.

Self-attention: attention maps

Visualizations of attention maps:

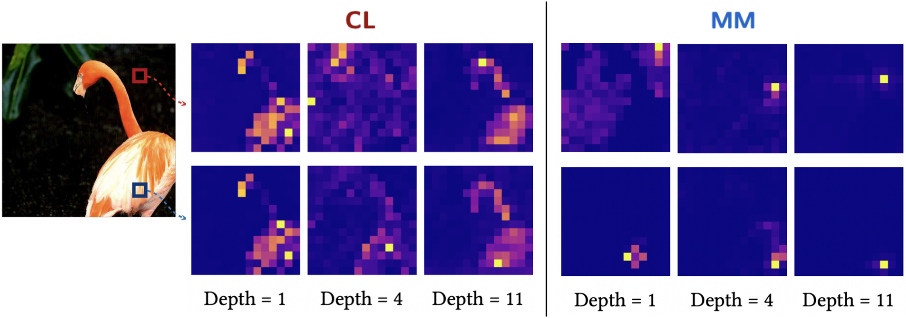


Figure 9: Self-attentions of **CL** (MoCo v3) *vs.* **MM** (SimMIM) for selected depths/layers.

- ViT-B/16 pretrained on ImageNet-1k
- select 2 different tokens in different layers, *e.g.*, 1, 4 & 11
- using ImageNet val image:
  - **CL**: global pat., shape of obj., all attns capture the same pat.; reg. of tokens
  - **MM**: capture local pat., correlated with tokens
  - self-attn heads show almost consistent results

Self-attention: attention distance

Attn dist.[10]: the avg. dist. between Q and K tokens w.r.t. self-attn weights
$\approx$ receptive field size of CNNs

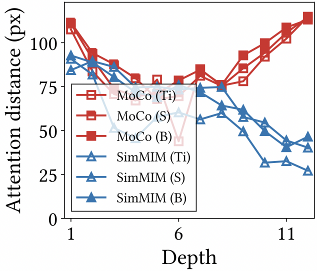


Figure 10: Recep. fields of **CL** *vs.* **MM**.

- AD of **CL** > **MM**, *e.g.*, later layers, implies
  - rep. of **CL** contains global pat. & shape info.
  - **CL** helps ViTs classify between obj. of imgs.
  - **MM** mainly captures local relationships
  - **MM** may have difficulty recognizing whole obj & shapes
- '*An attn collapse into homogeneity*'[a]
  - self-attn of **CL** indicates different spatial tokens have *e.g.*, identical obj. shapes
  - 'Homogeneity' of **CL** is observed across all heads & tokens

[a] Attn collapse reduces rep. diversity, which may lead to homogeneous token rep.

[10] Dosovitskiy *et al.* "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale." ICLR'21

Self-attention: attention collapse

Normalized mutual information (NMI)[11]:

- measure the attn collapse
- low mutual info. values $\rightarrow$ attn maps less dependent on the tokens
- high mutual info. $\rightarrow$ attn maps strongly depend on the tokens

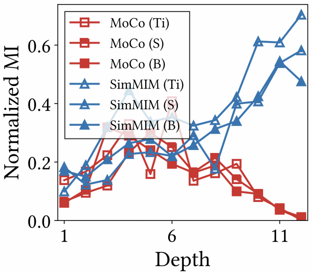


Figure 11: Degree of attn collapse w.r.t. NMI of **CL** *vs.* **MM**.

- MI of **CL** $\ll$ **MM** (later layers)
- self-attn of **CL** have little to do with tokens
- self-attn of **CL** tends to collapse into homog. distr.

[11] Strehl & Ghosh. "Cluster Ensembles — A Knowledge Reuse Framework for Combining Multiple Partitions." JMLR'03.

Self-attention: diversity of representations

Measure representations of self-attn using cosine similarity:

- different self-attn **heads** (*left fig.*)
- between the before & after self-attn layers (**depths**, *middle fig.*)
- between different **tokens**/spatial locations (*right fig.*)

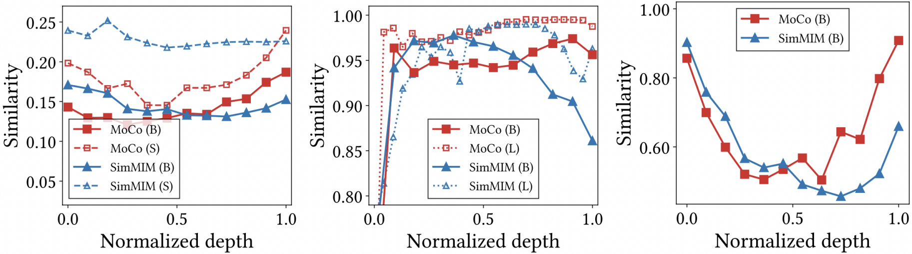


Figure 12: Cosine sim. of rep. in self-attn of **CL** *vs.* **MM** w.r.t. heads, depths and tokens.

- rep. sim. of **CL** > **MM** in later layers ('homogenity')
- ↑ heads (ViT-S to -B)/depths (ViT-B to -L) of **CL** → not effective in ↑ diversity; ViT-S to -B (*left*) ↑ rep. diversity of **MM**
- **CL** lacks rep. diversity in later layers → not suitable for dense pred. (token feat. are homo w.r.t. spatial coord.)

Representation: feature space

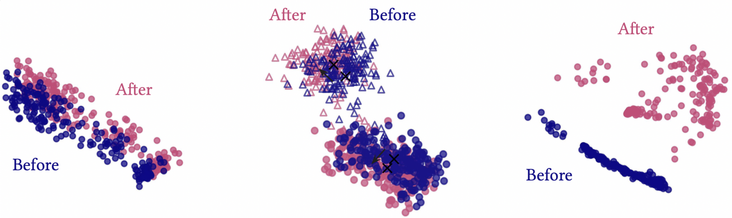


Figure 13: 'all tokens in unison' of **CL** *vs.* 'diff. transf. of individual tokens' of **MM**[12]

- Disp./Visual. rep. in crucial layers *e.g.*, the first layer & the last layer: *left*: **CL** (1 image), *middle*: **CL** (2 images), *right*: **MM** (1 image)
- 'unison' of **CL**: self-attn maps are homo. w.r.t. spatial loc. of tokens
- modules add near-constant to all token rep. → inter-rep. dis. & volume of rep. do not ↑ → **CL** cares less about individ. tokens
- self-attns helps discriminative power of **CL**, *e.g.*, *middle*, moving centers of rep. distr. away from each other: **CL** makes imgs linearly separable even though it losses the ability to distinguish tokens
- different self-attn are assigned to individual spatial tokens of **MM** (dis., vol.)

[12] Park *et al.* "What Do Self-Supervised Vision Transformers Learn?" ICLR'23.

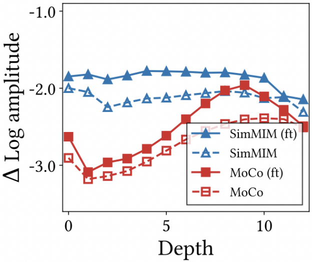
Representations: low-/high-frequency info.

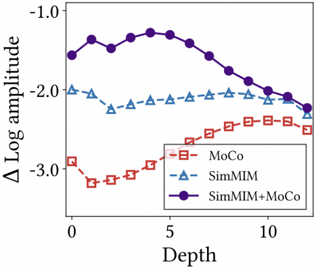
CL captures low-frequency info. & **MM** captures high-frequency info.?

- **CL**: provides image-level self-supervision / global patterns
- **MM**: provides token-level self-supervision / local patterns

Fourier analysis[13]:

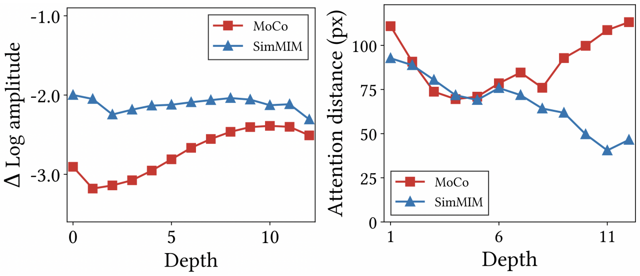
- show relative log amplitude of Fourier-transformed rep.
- by computing the amplitude difference between the highest & lowest frequencies of rep.





[13] Park & Kim. "How do vision transformers work?" ICLR'22

Representation: low-/high-frequency info. (cont.)



(a) low-/high-freq. of **CL** & **MM**   (b) Recep. fields of **CL** & **MM**

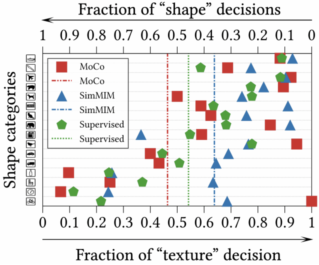
CL exploits low-frequencies & **MM** exploits high-frequencies:

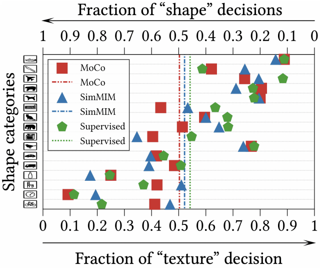
- high-freq. ampl. of **CL** $\ll$ **MM**:
  - **CL** uses low-freq. *e.g.*, global structures/shapes;
  - **MM** uses high-freq. spatial info. *e.g.*, narrow structures/fine textures
- Recall Fig. 8:
  - **CL** help linearly separate images in their repre. spaces
  - self-supervised models trained with **CL** & **MM** learn repre. in different levels of details

Representation: shape-/texture-biased

CL & **MM** each has a bias towards shapes & texture?

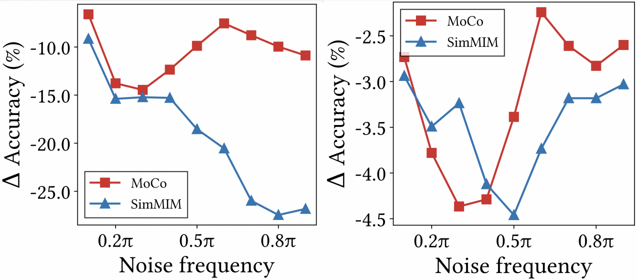
- using a texture-altered dataset: Stylized ImageNet[14]
- reporting the results of linear probing to evaluate the shape & texture biases of pretrained *left* & finetuned *right* models (ViT on ImageNet-1K of **superv.**)
- **CL** is more shape-biased > **MM** > **supervised**
- **CL** depends more on shape & **MM** depends on texture to classify imgs
- **CL** is robust to texture changes & **MM** is vulnerable to them





[14] Geirhos *et al*. "ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness". ICLR'19.

Representation: Robustness



Robustness for noise frequency (*left* pretrained & *right* finetuned):

- measure the decrease in acc on ImageNet with frequency-based random noise
- frequency window size of the noise is 0.1π
- **CL** is robust to high-freq. noises, **MM** is more vulnerable to them
- Why?
  - high-freq. noises harm the fine details of imgs
  - **CL** is more shape-biased, **MM** is texture-biased
  - Explained 'the robustness of **CL** against adversarial perturbations[15]

[15] Bordes *et al.* "High fidelity visualization of what your self-supervised representation knows about." TMLR'22.

Conclusion

Conclusion

Conclusion:

	CL (img-level invariants)	**MM** (token-level similarities)
Behaviour	linear probing & small model	finetuning & large model
Architecture	later layers	early layers
Self-attention	capture globalities & shapes	capture localities & textures
Representation	distinguish images	distinguish tokens

Future work:

- Complementary to each other? A simple way: linearly combining 2 losses *e.g.*, $\mathcal{L}=(1-\lambda)\mathcal{L}_{\textbf{MM}}+\lambda\mathcal{L}_{\textbf{CL}}$: Page 16 right fig.: hybrid models > **MM** ($\lambda=0$) > **CL** ($\lambda=1$)
- Enhance individual properties of **CL** & **MM** w.r.t. learning shapes / texture, may improve?
- Restricted receptive fields/locally restricted self-attentions of **CL**
- Apply **CL** in the later layers & **MM** in the early layers

Thank you!