Exploring Self-supervised Learned Representations

Representation Learning

Vladimir Zaigrajew

2025-03-25

Vladimir Zaigrajew - vladimir.zaigrajew.dokt@pw.edu.pl

Tymoteusz Kwieciński - tymoteuszkwiecinski@gmail.com

You can find us in Room 316, MINI, PW

Remember every information you can find on our Github Repo:

Figure 1: QR code to course Github Repo

Figure 2: QR code to our Github Repo

Recap from last lecture

- Both contrastive and unmasking task are derived from how humans learn
- Contrastive task is based on the idea of learning by comparing:
  - Discriminative learning: Anchor sample with its augmentation as a positive sample to be closer compared to other samples (negative samples)
  - Postive samples only: Samples are changed by augmentations and the model learns to have similar representations to them
- Unmasking task is based on the idea of learning by filling in the blanks:
  - Masked language modeling: The model learns to predict the missing words in a sentence
  - Image inpainting: The model learns to predict the missing pixels in an image
- Both tasks are used to pre-train models on large datasets, allowing them to learn useful representations that can be fine-tuned for specific tasks
- Mode collapse and Representation collapse are two common problems self supervised learning where the model fails to learn meaningful representations due to the lack of diversity in the training data or the model's hacking of the task

Recap from last lecture

- Contrastive learning requires a large amount of negative samples to be effective, while unmasking tasks can be more efficient with fewer samples per batch
- Contrastive learning is more suitable for tasks where the goal is to learn a similarity metric between samples, while unmasking tasks are more suitable for tasks where the goal is to learn a more general representation of the data
- The current best models from self-supervised learning are based on unmasking tasks, such as BERT or MAE but CLIP is also very popular from contrastive learning

When to use supervised transfer learning?

Small number of labeled data

Big amount of unlabeled data

Few-shot classification (fast adaptation)

Small domain shift to the training data distribution

5. Discussion

We have conducted the first thorough and up-to-date empirical evaluation of state of the art SSL performance when applied to diverse downstream tasks, a comparison that has been missing in the literature until now. Our evaluation showed that: (1) The best self-supervised methods today can usually outperform supervised pre-training as a source of knowledge transfer, an exciting milestone for the field that has long been speculated on, but now clearly confirmed. (2) Performance of self-supervised representations on ImageNet is reassuringly broadly representative of downstream performance on natural image recognition tasks, confirming the relevance of this metric for research. (3) However, ImageNet performance is not reliably representative of downstream performance on unstructured image recognition, or other spatially sensitive tasks such as detection, surface normal prediction and semantic segmentation. Thus the vision of a 'universal' pre-trained feature with best performance on diverse downstream tasks is yet to be realised. Furthermore, SSL researchers should adopt a wider range of benchmarks to better impact the broader computer vision community.

[0]Ericsson, Linus, Henry Gouk, and Timothy M. Hospedales. "How well do self-supervised models transfer?." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021.

When to use supervised transfer learning?

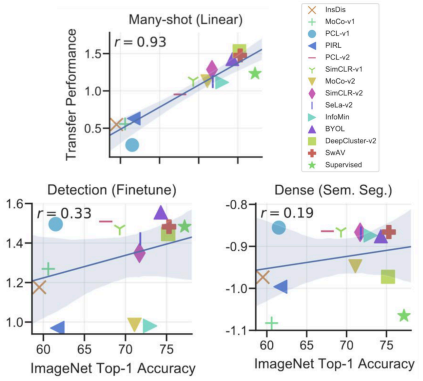
Do Imagenet self-supervised CNNs perform well on diverse downstream datasets and tasks?

Yes to highly correlated datasets such as CIFAR-10, CIFAR-100, STL-10 and others

However it has low correlation for different tasks such as detection and segmentation, why?

Is there a best SSL representation overall?

Is universal pre-training for several downstream tasks possible?



[0]Ericsson, Linus, Henry Gouk, and Timothy M. Hospedales. "How well do self-supervised models transfer?." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021.

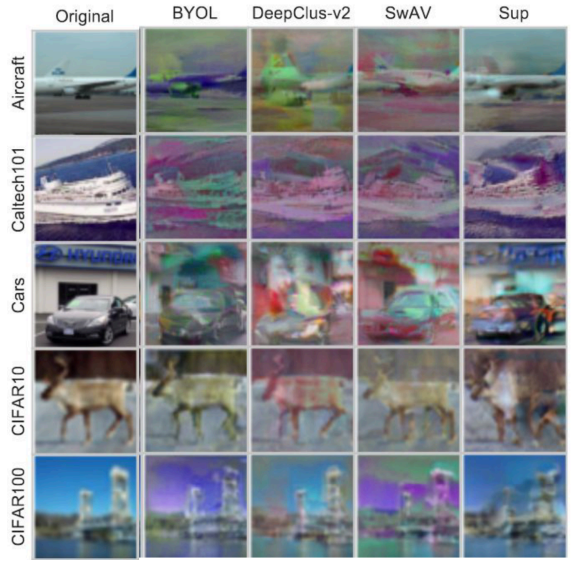
Do self-supervised and supervised features represent the same information?

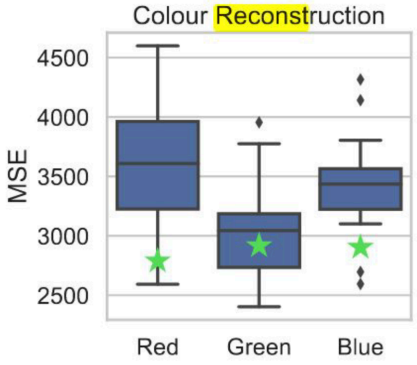
No

- self-supervised features seem to discard colour information (next slide)
- improved uncertainty calibration
- Complementary learned features

[0]Ericsson, Linus, Henry Gouk, and Timothy M. Hospedales. "How well do self-supervised models transfer?." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021.

Reconstruct RGB images from the features





Ericsson, Linus, Henry Gouk, and Timothy M. Hospedales. "How well do self-supervised models transfer?." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021.

Linear probing VS Fine-tuning on the same domain

		ImageNet	Cars	Pets	Cars	Pets	CIFAR100
Linear	InsDis	59.50	28.98	68.78	61.56	76.22	68.26
	MoCo-v1	60.60	27.99	69.84	65.02	76.96	71.52
	PCL-v1	61.50	12.93	75.34	73.24	86.98	79.62
	PIRL	61.70	28.72	71.36	61.02	76.26	66.48
	PCL-v2	67.60	30.51	82.79	71.68	85.39	80.26
	SimCLR-v1	69.30	43.73	83.33	83.78	84.10	84.53
	MoCo-v2	71.10	39.31	83.30	75.20	79.80	71.33
	SimCLR-v2	71.70	50.37	84.72	79.84	83.20	79.05
	SeLa-v2	71.80	36.86	83.22	85.62	88.55	84.37
	InfoMin	73.00	41.04	86.24	78.76	85.28	71.15
	BYOL	74.30	56.40	89.10	84.60	89.62	83.95
	DeepCluster-v2	75.20	**58.60**	89.36	**87.27**	89.43	**85.15**
	SwAV	75.30	54.06	87.60	86.76	89.05	84.37
	Supervised	**77.20**	44.92	**91.45**	82.61	**92.42**	82.91

Finetune

[0]Ericsson, Linus, Henry Gouk, and Timothy M. Hospedales. "How well do self-supervised models transfer?." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021.

Out-of-domain Few-shot transfer (20-shot) of pre-trained CNNs

	CropDiseases	EuroSAT	ISIC	ChestX
InsDis	91.95 ± 0.44	86.52 ± 0.51	52.19 ± 0.53	29.13 ± 0.44
MoCo-v1	92.04 ± 0.43	86.55 ± 0.51	$\mathbf{53.79 \pm 0.54}$	30.00 ± 0.43
PCL-v1	80.74 ± 0.57	75.19 ± 0.67	38.01 ± 0.44	25.54 ± 0.43
PIRL	91.19 ± 0.49	87.06 ± 0.50	53.24 ± 0.56	29.48 ± 0.45
PCL-v2	92.58 ± 0.44	87.94 ± 0.40	44.40 ± 0.52	28.28 ± 0.42
SimCLR-v1	94.03 ± 0.37	89.38 ± 0.40	53.00 ± 0.54	30.82 ± 0.43
MoCo-v2	92.12 ± 0.46	88.92 ± 0.41	52.39 ± 0.49	29.43 ± 0.45
SimCLR-v2	94.92 ± 0.34	91.05 ± 0.36	53.15 ± 0.53	30.90 ± 0.44
SeLa-v2	94.75 ± 0.37	88.34 ± 0.57	48.43 ± 0.54	30.43 ± 0.46
InfoMin	92.34 ± 0.44	86.76 ± 0.47	48.21 ± 0.54	29.48 ± 0.44
BYOL	96.07 ± 0.33	89.62 ± 0.39	53.76 ± 0.55	30.71 ± 0.47
DeepCluster-v2	$\mathbf{96.63 \pm 0.29}$	$\mathbf{92.02 \pm 0.37}$	49.91 ± 0.53	$\mathbf{31.51 \pm 0.45}$
SwAV	96.15 ± 0.31	91.99 ± 0.36	47.08 ± 0.50	30.91 ± 0.45
Supervised	93.09 ± 0.43	88.36 ± 0.43	48.79 ± 0.53	29.26 ± 0.44

[0]Ericsson, Linus, Henry Gouk, and Timothy M. Hospedales. "How well do self-supervised models transfer?." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021.

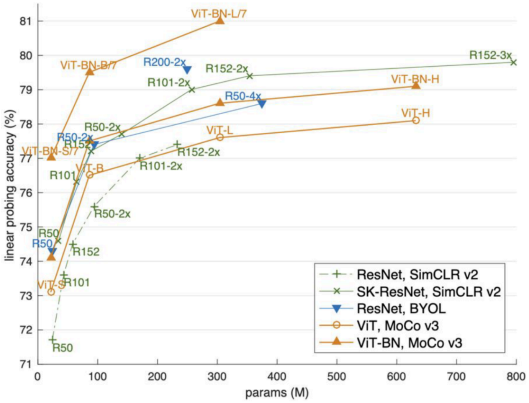
What do vision transformers (ViTs) learn “on their own”? MoCov3

Outperform Resnet50 on ImageNet (supervised)

Better generalization than supervised

model	MoCo v3	SimCLR	BYOL	SwAV
R-50, 800-ep	73.8	70.4	**74.3**	71.8
ViT-S, 300-ep	**72.5**	69.0	71.0	67.1
ViT-B, 300-ep	**76.5**	73.9	73.9	71.6

Table 4. **ViT-S/16 and ViT-B/16 in different self-supervised learning frameworks** (ImageNet, linear probing). R-50 results of other frameworks are from the improved implementation in



pre-train	CIFAR-10 [26] ViT-B	ViT-L	ViT-H	CIFAR-100 [26] ViT-B	ViT-L	ViT-H	Oxford Flowers-102 [33] ViT-B	ViT-L	ViT-H
random init.	77.8	77.1	75.9	48.5	48.3	48.0	54.4	54.3	52.8
ImNet supervised [16]	98.1	97.9	n/a	87.1	86.4	n/a	89.5	89.7	n/a
ImNet self-sup., MoCo v3	98.9 ↑0.8	99.1 ↑1.2	99.1	90.5 ↑3.4	91.1 ↑4.7	91.2	97.7 ↑8.2	98.6 ↑8.9	98.8

[0]Chen, Xinlei, Saining Xie, and Kaiming He. ”An empirical study of training self-supervised vision transformers.” Proceedings of the IEEE/CVF International conference on computer vision.

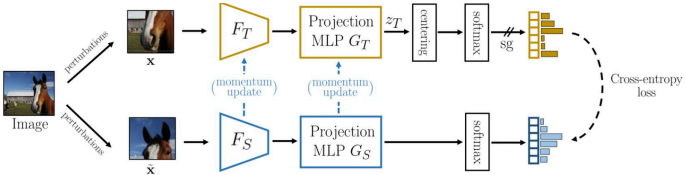
What do vision transformers (ViTs) learn “on their own”? DINO

Excellent k-NN classifiers

Features are organized in an interpretable way

Connects categories based on visual characteristics

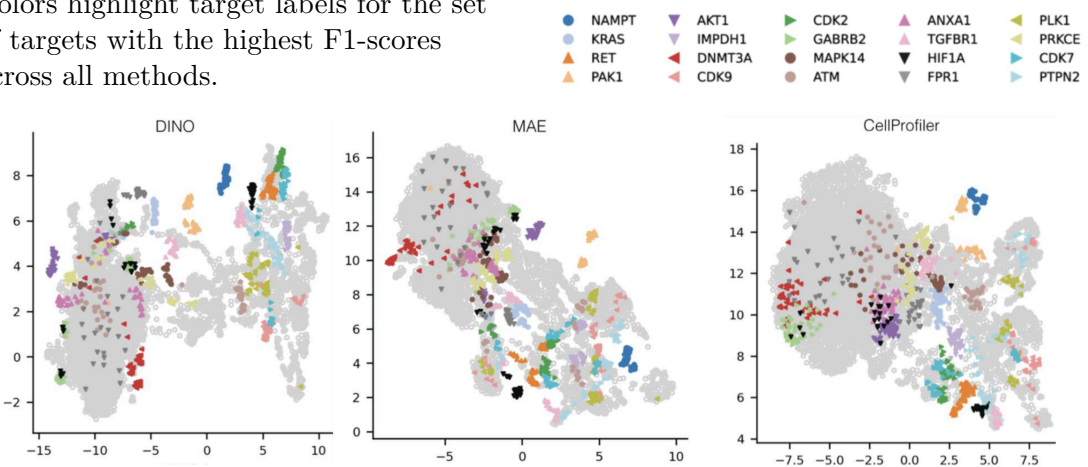
https://ai.meta.com/blog/dino-paws-computer-vision-with-self-supervised-transformers-and-10x-more-efficient-training/



Method	Arch.	Param.	im/s	Linear	k-NN
SwAV [10]	RN50w2	93	384	77.3	67.3
BYOL [30]	RN50w2	93	384	77.4	–
DINO	ViT-B/16	85	312	78.2	76.1
SwAV [10]	RN50w5	586	76	78.5	67.1
BYOL [30]	RN50w4	375	117	78.6	–
BYOL [30]	RN200w2	250	123	79.6	73.9
DINO	ViT-S/8	21	180	79.7	**78.3**
SCLRv2 [13]	RN152w3+SK	794	46	79.8	73.1
DINO	ViT-B/8	85	63	**80.1**	77.4

[0]Caron, Mathilde, et al. ”Emerging properties in self-supervised vision transformers.” Proceedings of the IEEE/CVF international conference on computer vision. 2021.

Colors highlight target labels for the set of targets with the highest F1-scores across all methods.



[0]Kim, Vladislav, et al. "Self-supervision advances morphological profiling by unlocking powerful image representations. bioRxiv." (2023): 6.

Visualization - ViT Attention maps

Class-specific features lead to unsupervised segmentation masks

Correlate with the shape of semantic objects in the images

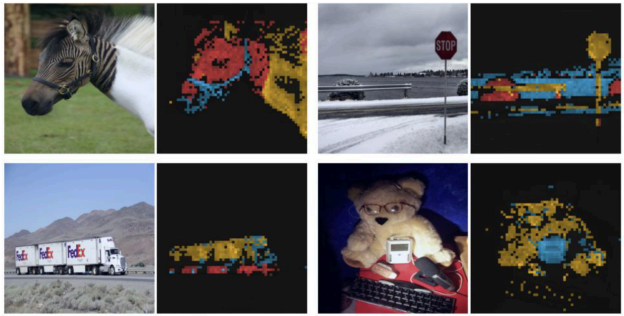


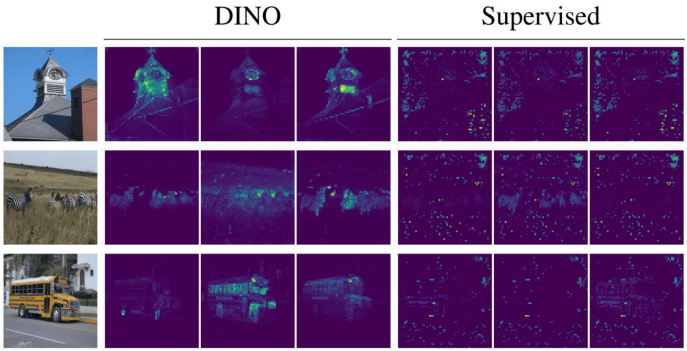
Figure 3: **Attention maps from multiple heads.** We consider the heads from the last layer of a ViT-S/8 trained with DINO and display the self-attention for `[CLS]` token query. Different heads, materialized by different colors, focus on different locations that represents different objects or parts (more examples in Appendix).

[0]Caron, Mathilde, et al. "Emerging properties in self-supervised vision transformers." Proceedings of the IEEE/CVF international conference on computer vision. 2021.

Visualization - ViT Attention maps

Attention maps capture explicit semantic information

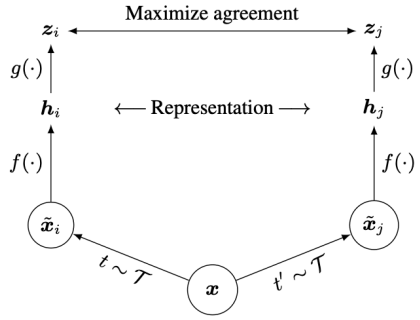
Does not emerge as clearly with supervised ViTs, nor with convnets!



[0]Caron, Mathilde, et al. "Emerging properties in self-supervised vision transformers." Proceedings of the IEEE/CVF international conference on computer vision. 2021.

"Image-level" VS "token-level" self-supervised learning

SimCLR/MoCO



Chen, Ting, et al. "A simple framework for contrastive learning of visual representations." International conference on machine learning. PmLR, 2020.

SimMIM

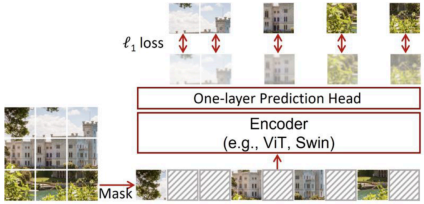


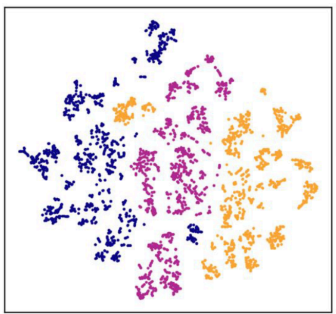
Figure 1. An illustration of our simple framework for masked language modeling, named *SimMIM*. It predicts raw pixel values of the randomly masked patches by a lightweight one-layer head, and performs learning using a simple ℓ_1 loss.

Xie, Zhenda, et al. "Simmim: A simple framework for masked image modeling." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022.

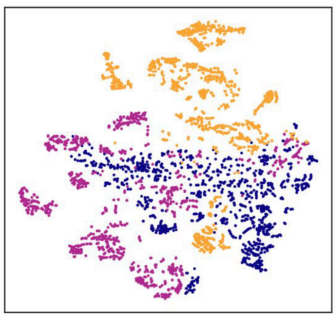
The tokens of MoCo form a cluster for each image

Contrastive -based features are more linearly separable (3 classes)

3528 tokens
(196 tokens x 18 images)



(a) MoCo

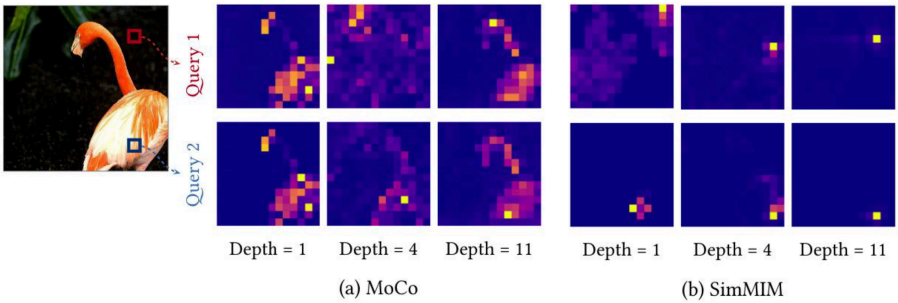


(b) SimMIM

[0]Park, Namuk, et al. "What do self-supervised vision transformers learn?." arXiv preprint arXiv:2305.00729 (2023).

Contrastive learning → global information

Masked Image Modeling (MIM) → local areas and similar tokens



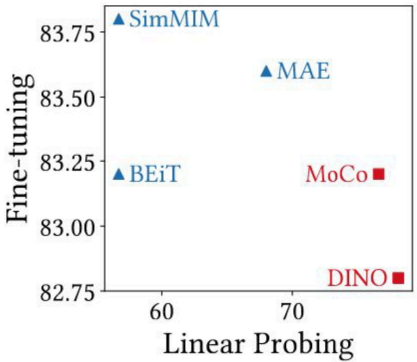
collapses into homogeneous maps for all queries and heads

[0]Park, Namuk, et al. "What do self-supervised vision transformers learn?." arXiv preprint arXiv:2305.00729 (2023).

MoCo, DINO outperform MIM methods in linear probing and small model regimes.

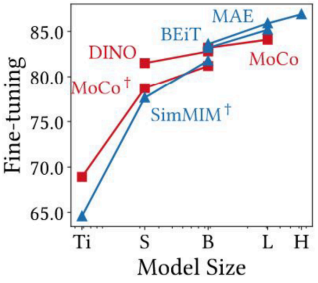
MIM excels in fine-tuning, large model regimes, and dense prediction.

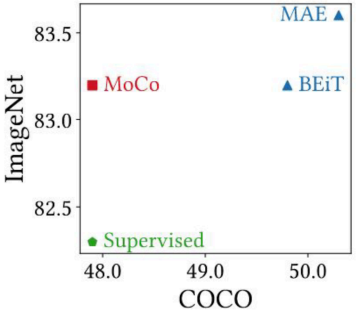
DINO, BEiT, MAE have consistent properties



[0]Park, Namuk, et al. "What do self-supervised vision transformers learn?." arXiv preprint arXiv:2305.00729 (2023).

MIM shows superior scalability in large model regimes





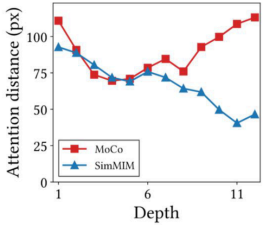


Figure 3: **Effective receptive fields of CL are global, but those of MIM are local.** This is particularly evident in the later layers.

[0]Park, Namuk, et al. "What do self-supervised vision transformers learn?." arXiv preprint arXiv:2305.00729 (2023).

None :)

For more, read this lecture from the Lab of HHU Dusseldorf (clickable link).

Also these papers which focus on the analysis of self-supervised representation and comparing it to supervised ones are also good places to go:

- How well do self-supervised models transfer?
- An Empirical Study of Training Self-Supervised Vision Transformers
- Emerging Properties in Self-Supervised Vision Transformers
- What Do Self-Supervised Vision Transformers Learn?