Vision Transformers Needs Registers

Presenter: Niccolo Avogaro

18.03.2025

Background

Self-Supervised Learning in Computer Vision

- Using the vision sense is a natural and powerful way to gain perception of our world.
- Traditional computer vision pipelines require extremely expensive labeling processes.
- Learning via image content without any labels has proven to be extremely hard.

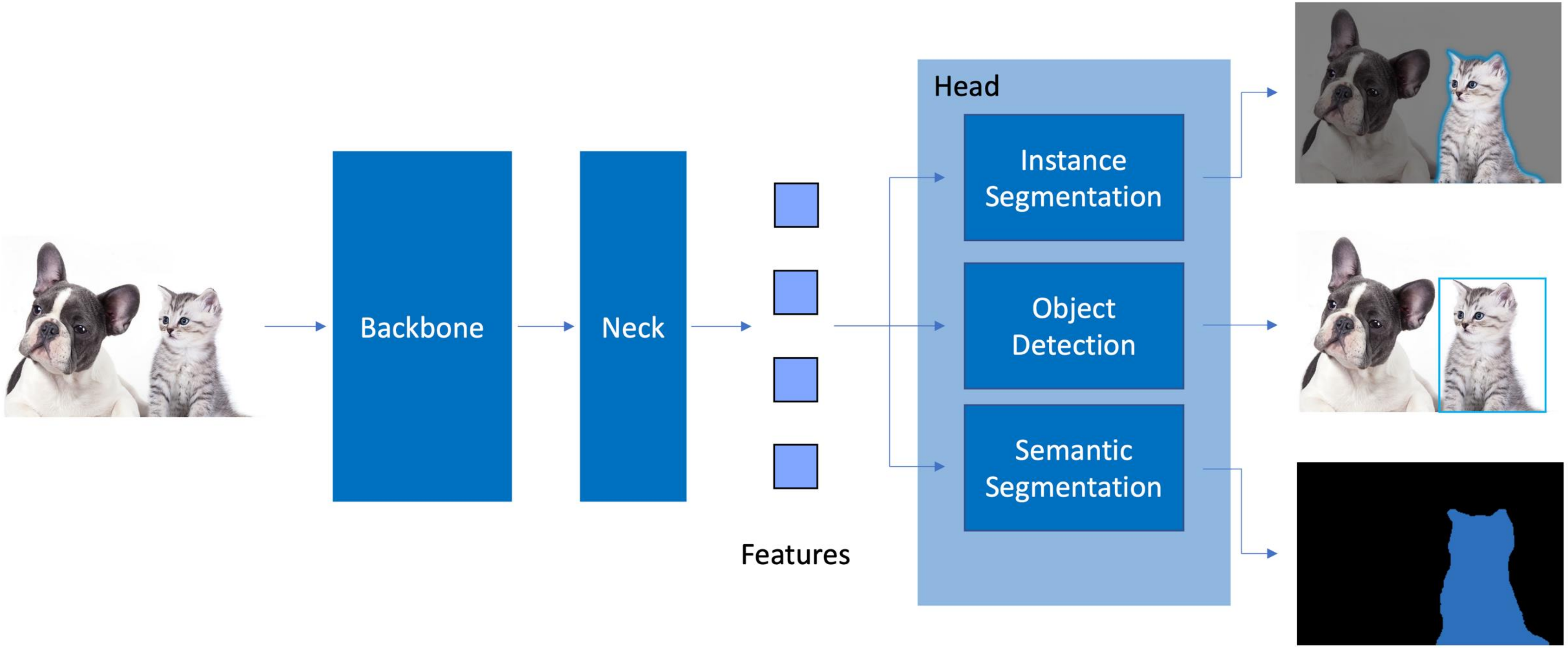
Yann LeCun
@ylecun

* Language is low bandwidth: less than 12 bytes/second. A person can read 270 words/minutes, or 4.5 words/second, which is 12 bytes/s (assuming 2 bytes per token and 0.75 words per token). A modern LLM is typically trained with 1x10^13 two-byte tokens, which is 2x10^13 bytes. This would take about 100,000 years for a person to read (at 12 hours a day).

* Vision is much higher bandwidth: about 20MB/s. Each of the two optical nerves has 1 million nerve fibers, each carrying about 10 bytes per second. A 4 year-old child has been awake a total 16,000 hours, which translates into 1x10^15 bytes.

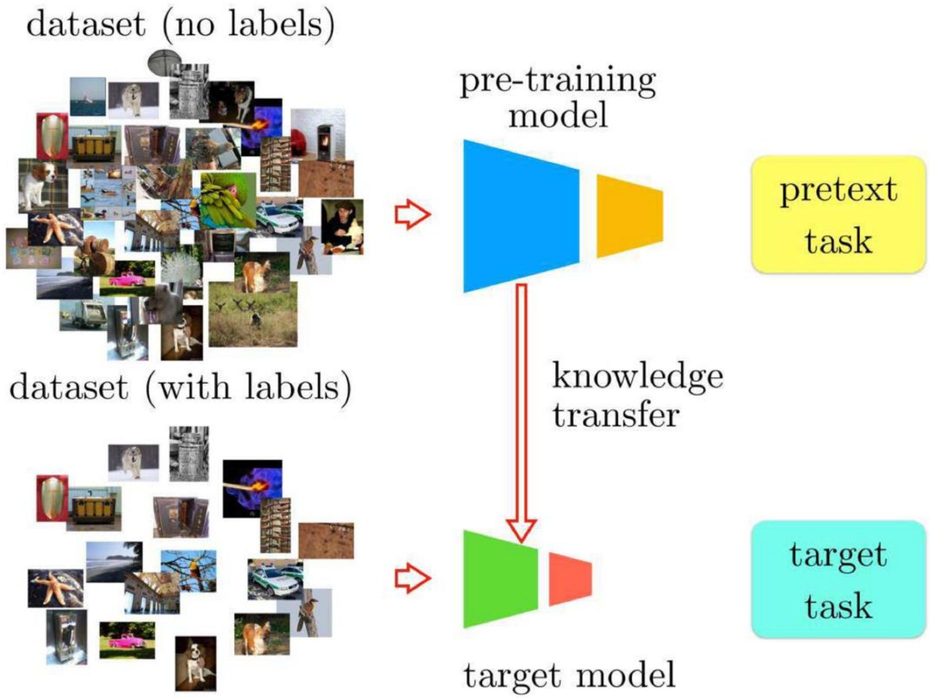
In other words:
- The data bandwidth of visual perception is roughly 16 million times higher than the data bandwidth of written (or spoken) language.
- In a mere 4 years, a child has seen 50 times more data than the biggest LLMs trained on all the text publicly available on the internet.

Classic Computer Vision Pipeline



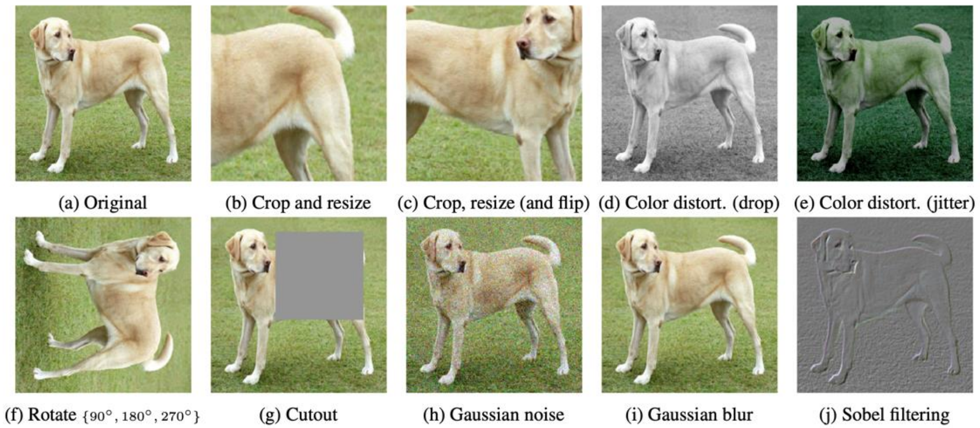
Self-Supervised Pipeline

- Self-Supervised Learning aims at creating a strong representation via training on a pre-text task.



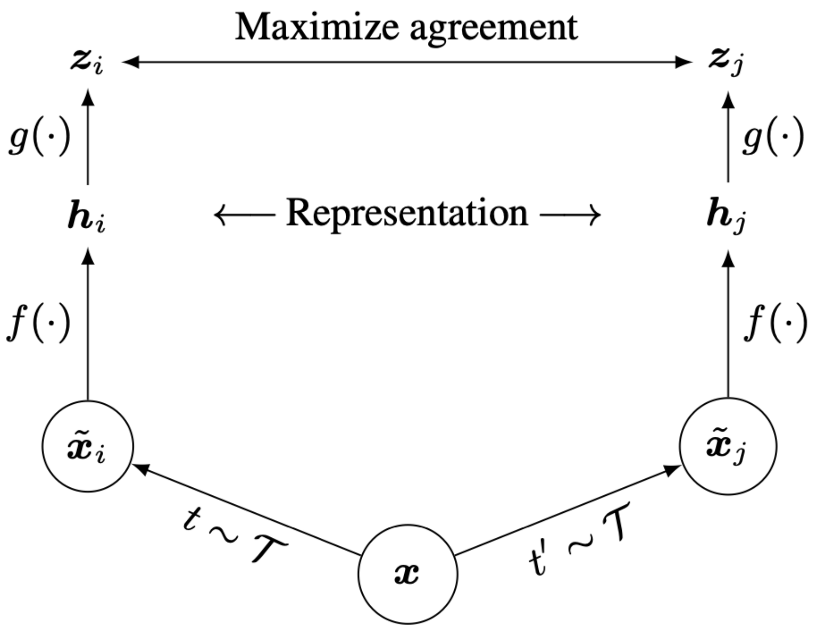
Self-Supervised Pipeline

- Self-Supervised Learning aims at creating a strong representation via training on a pre-text task.
- We want to enforce two different looking images of the same object to be mapped closely in the feature space

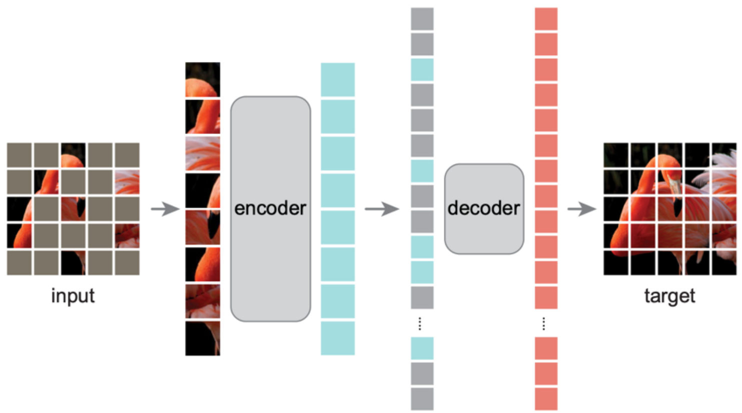


Self-Supervised Pipeline

- Self-Supervised Learning aims at creating a strong representation via training on a pre-text task.
- We want to enforce two different looking images of the same object to be mapped closely in the feature space
- The model is usually trained maximizing the agreement of two augmentations

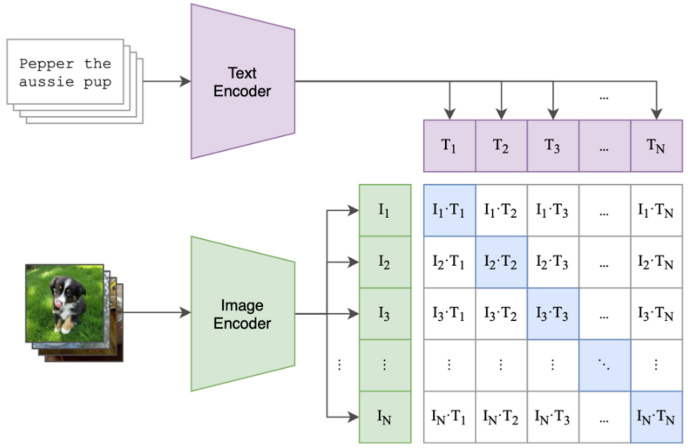


Some of the Greatest



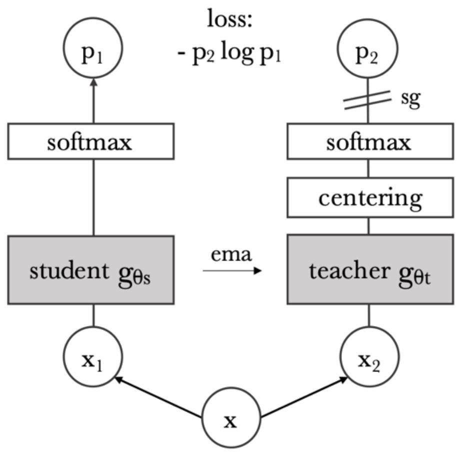
MAE

He et al, Masked Autoencoders Are Scalable Vision Learners, 2021



CLIP (Again)

Radford et al, Learning Transferable Visual Models From Natural Language Supervision, 2021



DINO

Caron et al, Emerging Properties in Self-Supervised Vision Transformers, 2021

Emerging Properties in Self-Supervised Vision Transformers

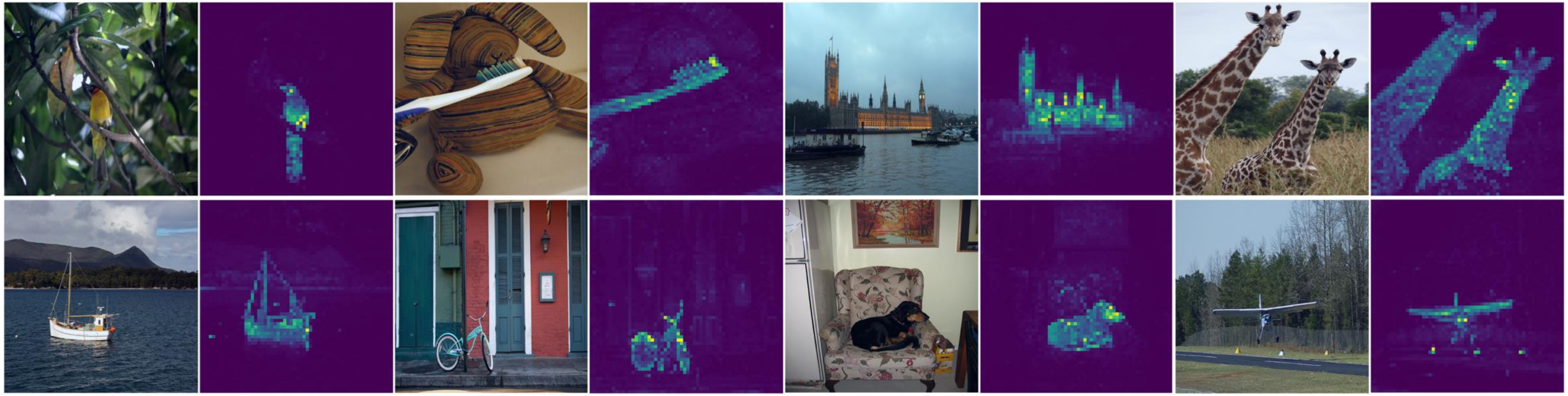
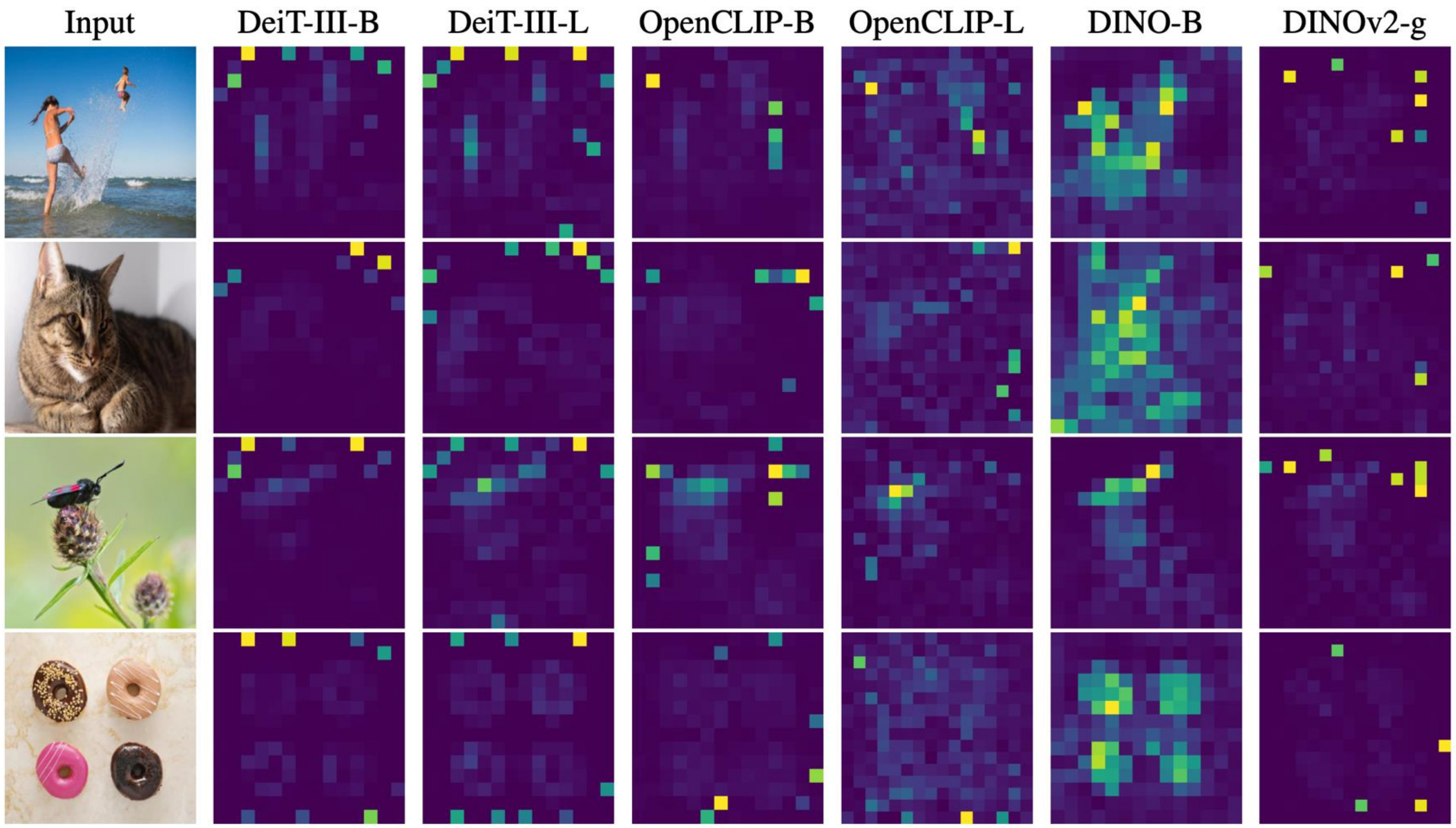


Figure 1: **Self-attention from a Vision Transformer with 8×8 patches trained with no supervision.** We look at the self-attention of the [CLS] token on the heads of the last layer. This token is not attached to any label nor supervision. These maps show that the model automatically learns class-specific features leading to unsupervised object segmentations.

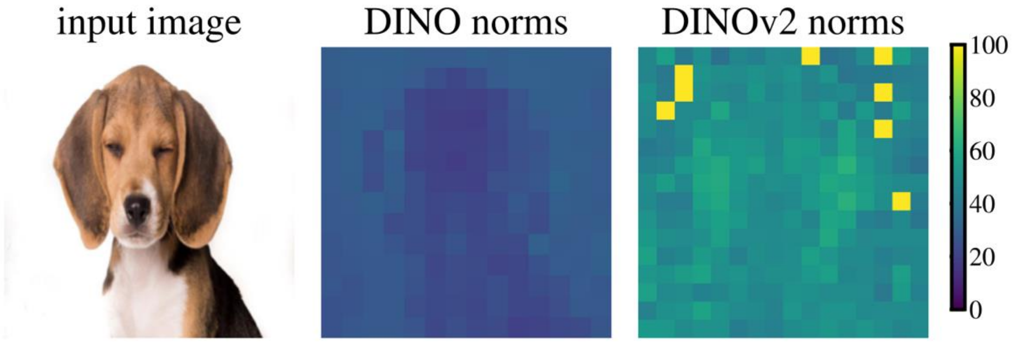
Vision Transformers Need Registers

Problem



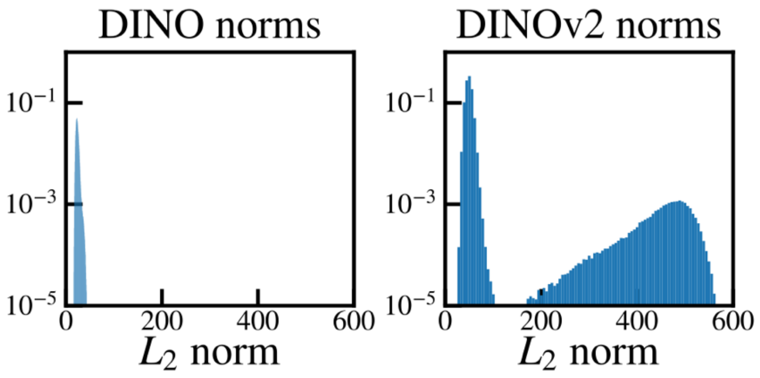
- The emergence of clean attention maps in inference is a behavior not seen in modern SSL methods.

Distribution of artifacts



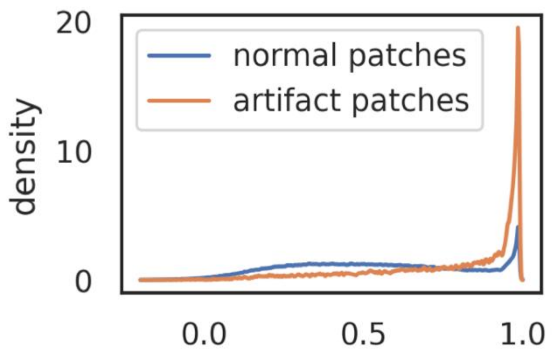
- Looking at the distribution of the norms of the attention values, DINOv2 has a few outlier patches, whereas DINO does not present these artifacts.

Distribution of artifacts



- Looking at the distribution of the norms of the attention values, DINOv2 has a few outlier patches, whereas DINO does not present these artifacts.
- Seems like a bimodal distribution of values.

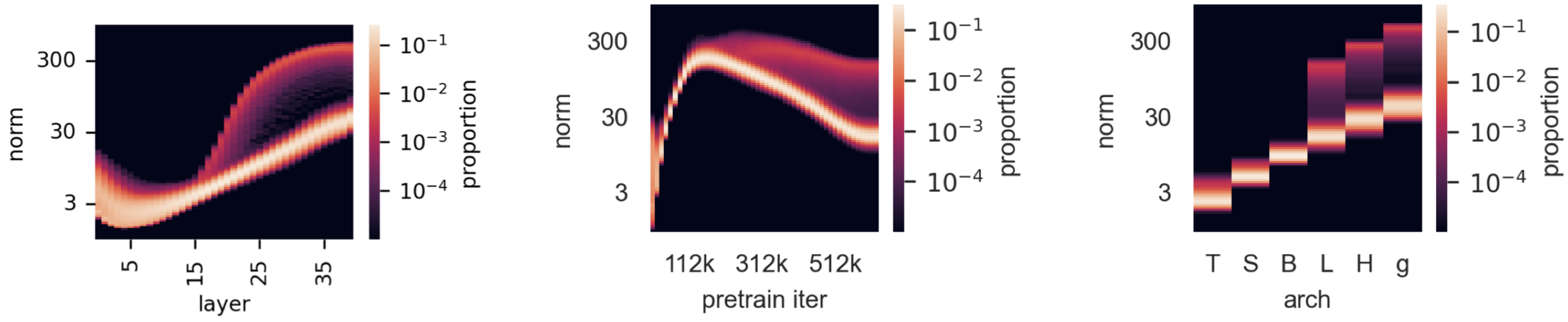
Distribution of artifacts



(a) Cosine similarity to neighbors.

- Looking at the distribution of the norms of the attention values, DINOv2 has a few outlier patches, whereas DINO does not present these artifacts.
- Seems like a bimodal distribution of values.
- The outlier patches are very dissimilar in the feature space from their neighbors

Distribution of artifacts - More in detail



(a) Norms along layers. (b) Norms along iterations. (c) Norms across model size.

Figure 4: Illustration of several properties of outlier tokens in the 40-layer DINOv2 ViT-g model. **(a)**: Distribution of output token norms along layers. **(b)**: Distribution of norms along training iterations. **(c)**: Distribution of norms for different model sizes. The outliers appear around the middle of the model during training; they appear with models larger than and including ViT-Large.

Other interesting observations

- The behavior is per-se not bad, as the models having artifacts still carry the most of performance
- On the other hand, the model discards local patch information
- Linear probing of the representation (CLS, normal and outliers) shows that outliers contain global information

	position prediction		reconstruction
	top-1 acc	avg. distance ↓	L2 error ↓
normal	**41.7**	**0.79**	**18.38**
outlier	22.8	5.09	25.23

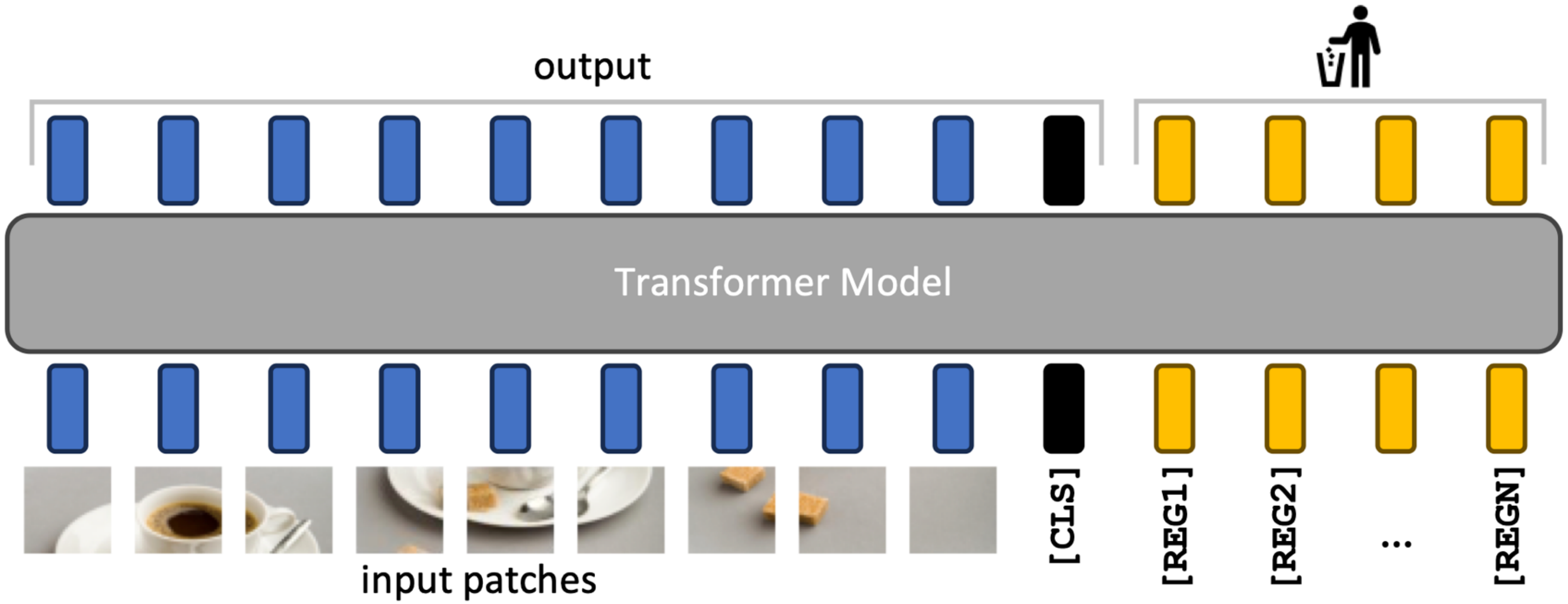
	IN1k	P205	Airc.	CF10	CF100	CUB	Cal101	Cars	DTD	Flow.	Food	Pets	SUN	VOC
[CLS]	**86.0**	**66.4**	**87.3**	**99.4**	**94.5**	**91.3**	96.9	**91.5**	**85.2**	**99.7**	**94.7**	**96.9**	**78.6**	89.1
normal	65.8	53.1	17.1	97.1	81.3	18.6	73.2	10.8	63.1	59.5	74.2	47.8	37.7	70.8
outlier	69.0	55.1	79.1	99.3	93.7	84.9	**97.6**	85.2	84.9	99.6	93.5	94.1	78.5	**89.7**

Hypothesys

Large, *sufficiently trained* models learn to recognize *redundant* tokens, and to use them as places to *store*, *process* and *retrieve* global information.

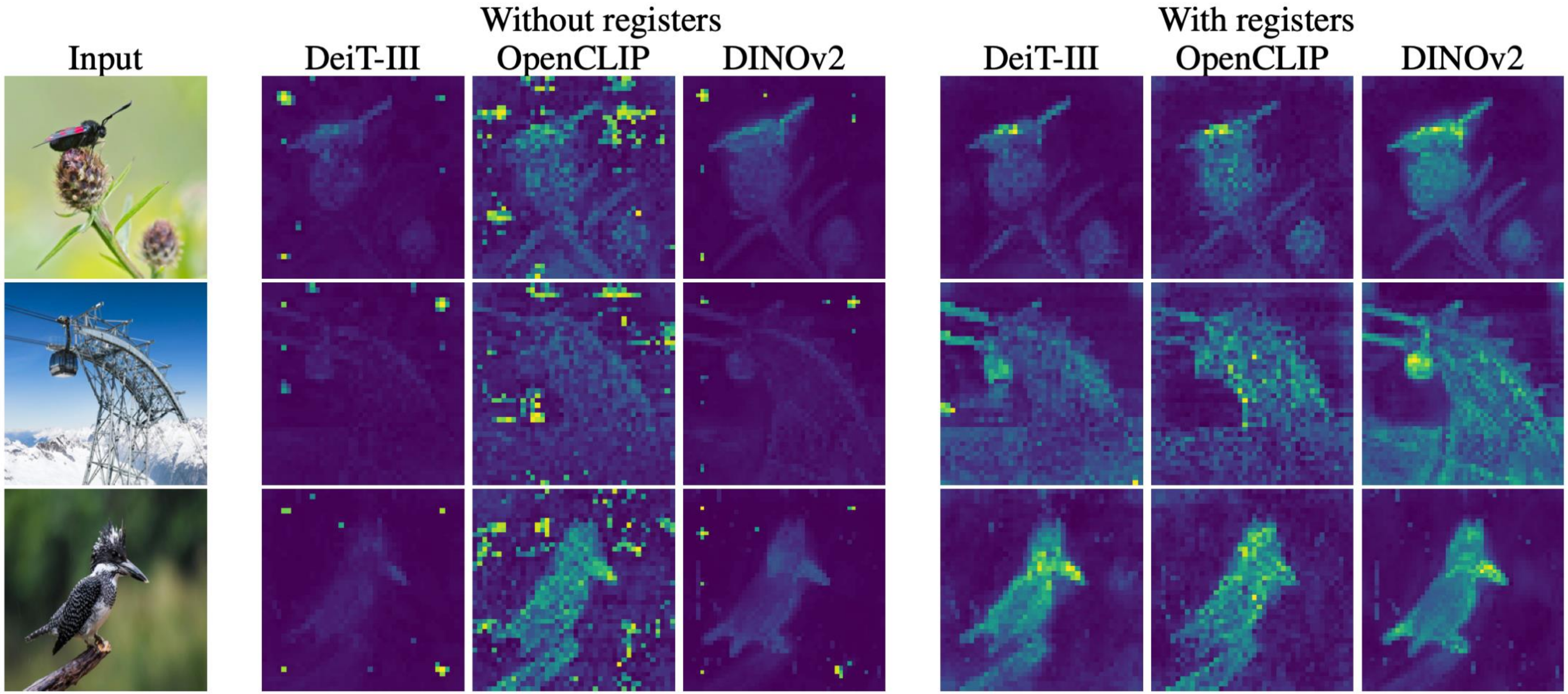
Solution

Registers



Adding new tokens not used in downstream tasks empowers the model to store and process additional information while reducing artifacts.

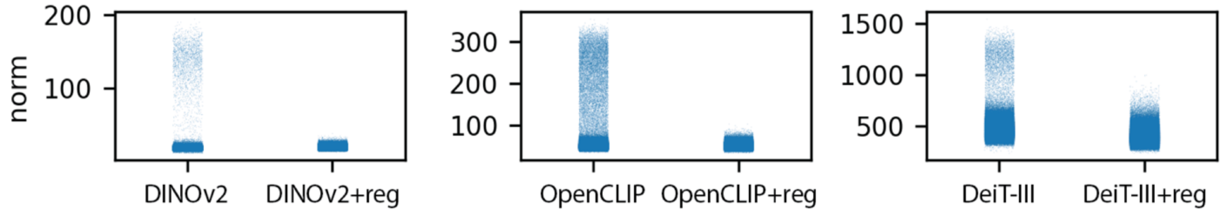
Results



Without registers
With registers
Input
DeiT-III
OpenCLIP
DINOv2
DeiT-III
OpenCLIP
DINOv2

Results

- The distribution of norms becomes unimodal, with way less outliers.
- The performance is similar or slightly better in downstream tasks



	ImageNet Top-1	ADE20k mIoU	NYUd rmse ↓
DeiT-III	84.7	38.9	0.511
DeiT-III+reg	84.7	39.1	0.512
OpenCLIP	78.2	26.6	0.702
OpenCLIP+reg	78.1	26.7	0.661
DINOv2	84.3	46.6	0.378
DINOv2+reg	84.8	47.9	0.366

(a) Linear evaluation with frozen features.

	ImageNet Top-1
OpenCLIP	59.9
OpenCLIP+reg	60.1

(b) Zero-shot classification.

Results

- The performance improves dramatically for unsupervised object discovery tasks

	VOC 2007	VOC 2012	COCO 20k
DeiT-III	11.7	13.1	10.7
DeiT-III+reg	27.1	32.7	25.1
OpenCLIP	38.8	44.3	31.0
OpenCLIP+reg	37.1	42.0	27.9
DINOv2	35.3	40.2	26.9
DINOv2+reg	55.4	60.0	42.0

Emergent Property (?)

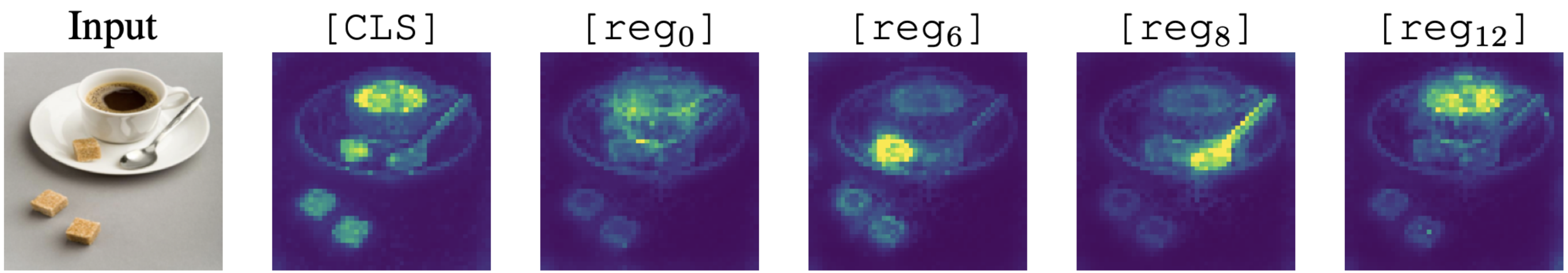


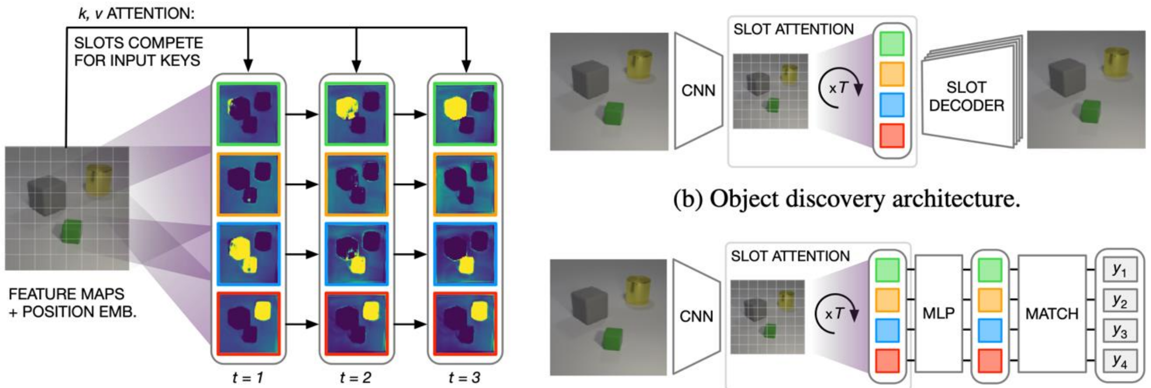
Figure 9: Comparison of the attention maps of the [CLS] and register tokens. Register tokens sometimes attend to different parts of the feature map, similarly to slot attention (Locatello et al., 2020). This behaviour was never required from the model, and emerged naturally from training.

Conlusions and Limitations

- Darcet et al. finds that the attention maps of modern transformer-based models is corrupted.
- They introduce a registers to clean these maps, resulting in clearer visualizations.
- Great explainability work.
- While attention maps improve, the downstream performance is left unchanged, with unsupervised segmentation models being far from SOTA.
- Self-contained.

Future Work

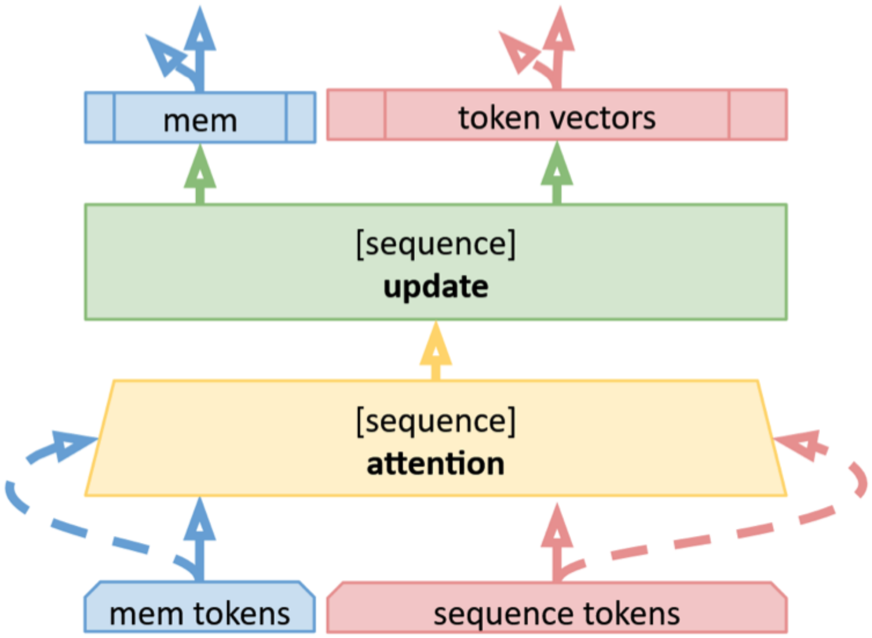
- Emergent object-centric behavior is interesting.
- Solving object centric self-supervised representation learning would mean solve the tokenization problem in CV



(a) Slot Attention module.
(b) Object discovery architecture.
(c) Set prediction architecture.

Future Work

- Adding registers to allow for better computation of results is very interesting
- Input-level computation is a field sometimes overlooked



Burtsev et al, Memory Transformer, 2020