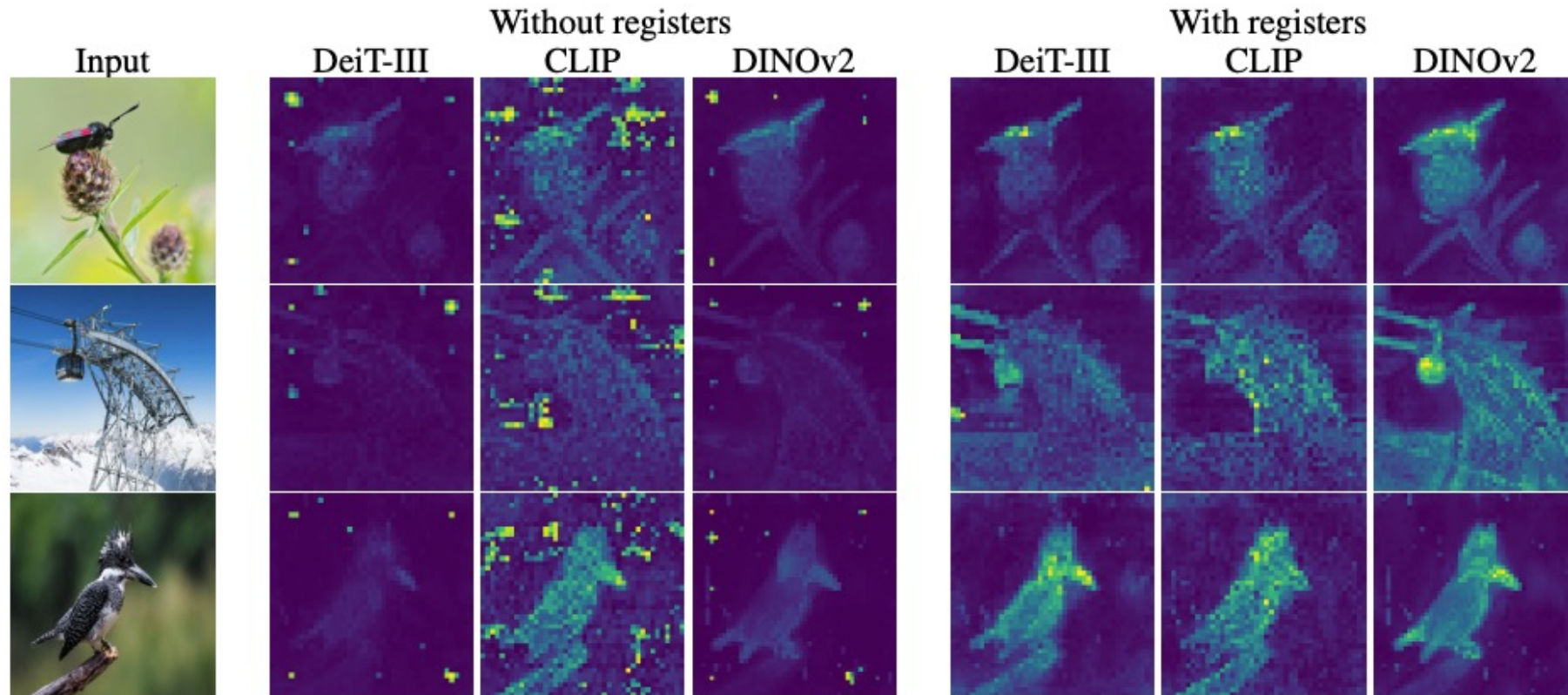


“Vision Transformers Need Registers”

Tom Davies



Vision Transformers - Recap

- Apply transformers to vision with as few modifications to the architecture as possible.
- Previous attempts struggled with attention over pixels.

Published as a conference paper at ICLR 2021

AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE

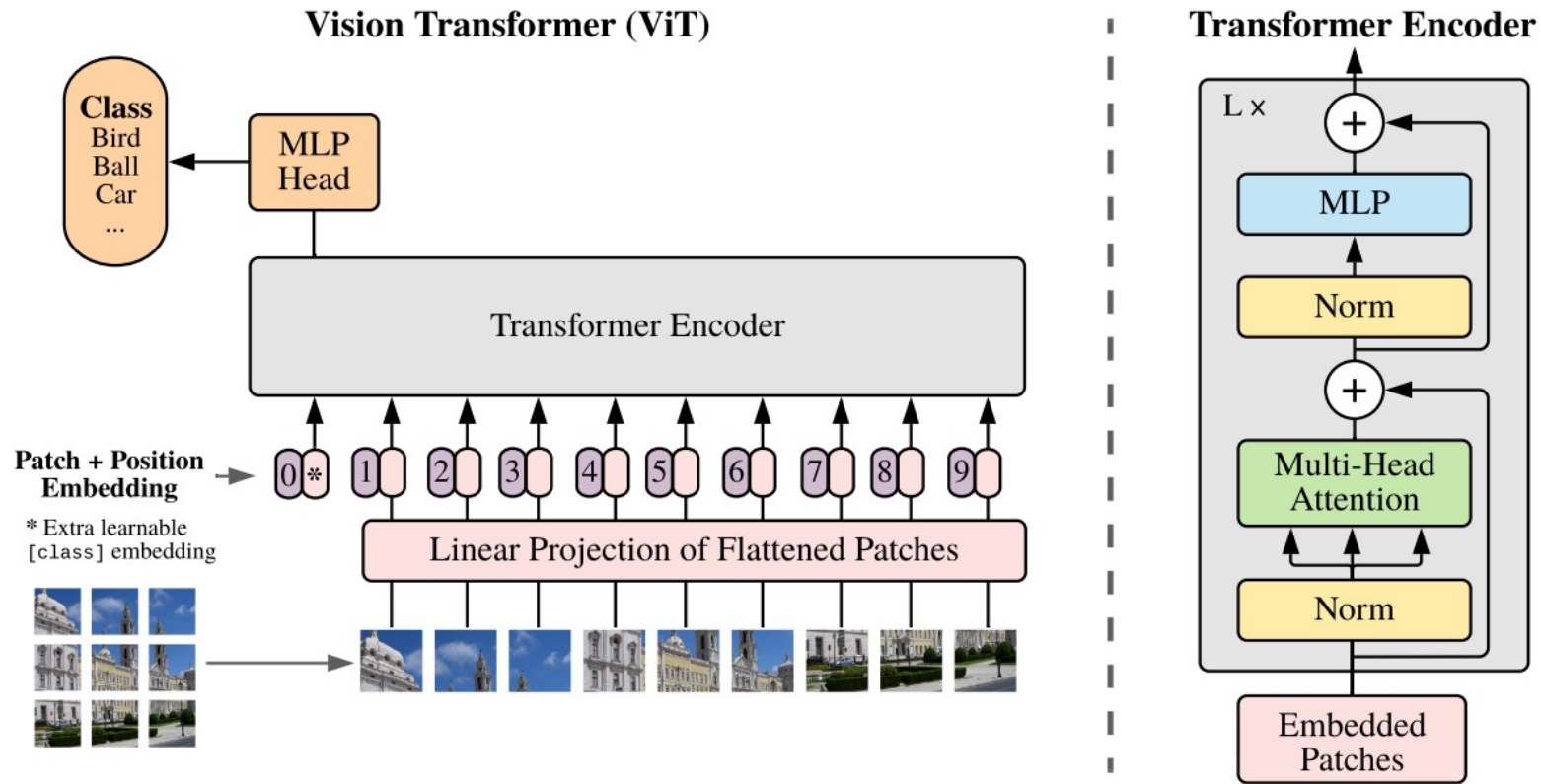
Alexey Dosovitskiy^{*,†}, Lucas Beyer^{*}, Alexander Kolesnikov^{*}, Dirk Weissenborn^{*},
Xiaohua Zhai^{*}, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer,
Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby^{*,†}

^{*}equal technical contribution, [†]equal advising

Google Research, Brain Team

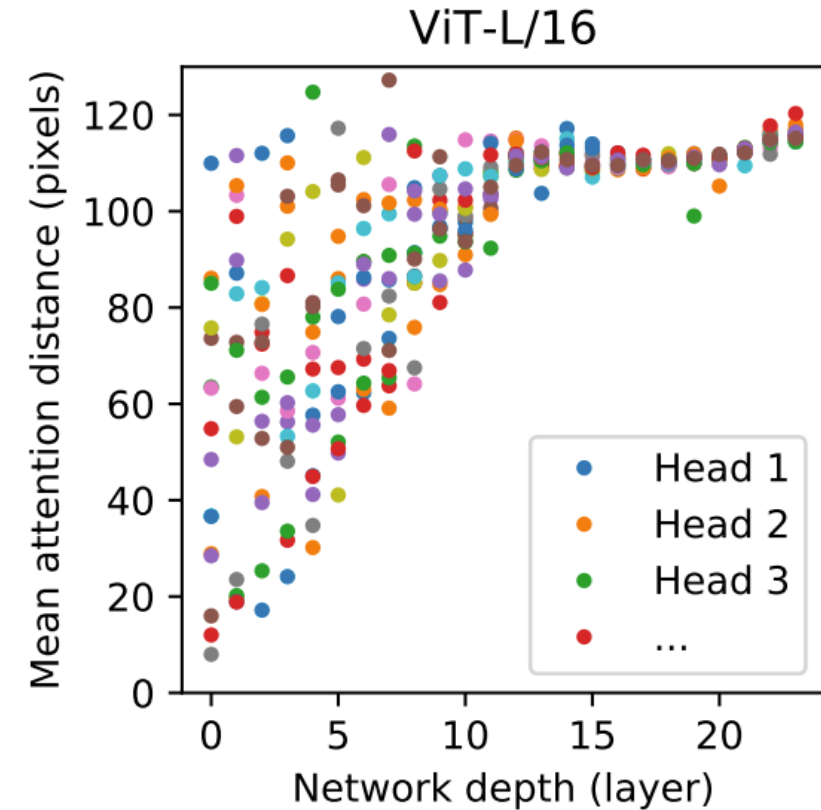
{adosovitskiy, neilhoulby}@google.com

Vision Transformers - Recap

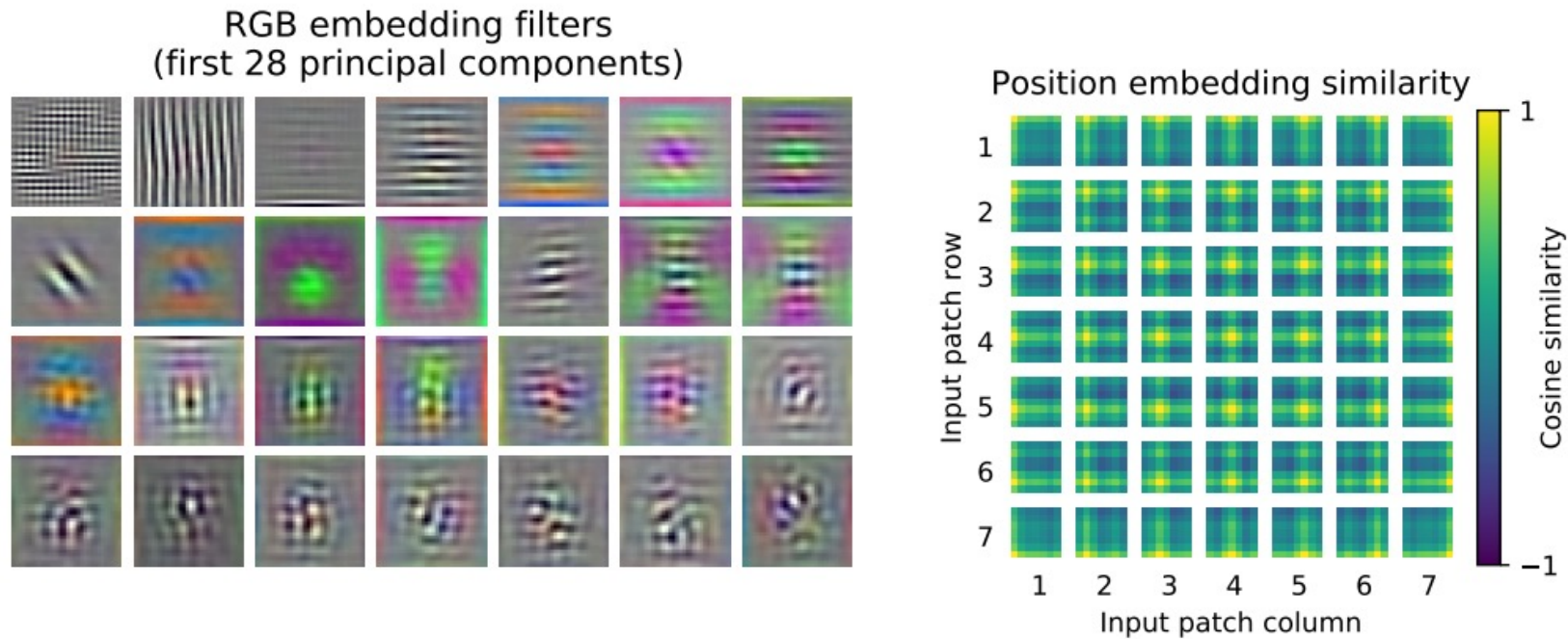


Vision Transformers - Recap

- Performs worse than CNNs on ImageNet-1k.
- CNNs have inductive biases which are difficult to learn.
- With more data (e.g., ImageNet-21k), they were SOTA.



Vision Transformers - Recap



Vision Transformers - Recap

	Ours-JFT (ViT-H/14)	Ours-JFT (ViT-L/16)	Ours-I21k (ViT-L/16)	BiT-L (ResNet152x4)	Noisy Student (EfficientNet-L2)
ImageNet	88.55 ± 0.04	87.76 ± 0.03	85.30 ± 0.02	87.54 ± 0.02	88.4/88.5*
ImageNet ReaL	90.72 ± 0.05	90.54 ± 0.03	88.62 ± 0.05	90.54	90.55
CIFAR-10	99.50 ± 0.06	99.42 ± 0.03	99.15 ± 0.03	99.37 ± 0.06	—
CIFAR-100	94.55 ± 0.04	93.90 ± 0.05	93.25 ± 0.05	93.51 ± 0.08	—
Oxford-IIIT Pets	97.56 ± 0.03	97.32 ± 0.11	94.67 ± 0.15	96.62 ± 0.23	—
Oxford Flowers-102	99.68 ± 0.02	99.74 ± 0.00	99.61 ± 0.02	99.63 ± 0.03	—
VTAB (19 tasks)	77.63 ± 0.23	76.28 ± 0.46	72.72 ± 0.21	76.29 ± 1.70	—
TPUv3-core-days	2.5k	0.68k	0.23k	9.9k	12.3k

ViT-Huge – 632M parameters. ViT-Large – 307M parameters.

JFT-300M – internal Google dataset. (Now JFT-3B!)

Other ViTs

- Three models used in the registers paper:
 - DEiT-III (supervised w/ labels)
 - OpenCLIP (supervised w/ text)
 - DinoV2 (self-supervised)

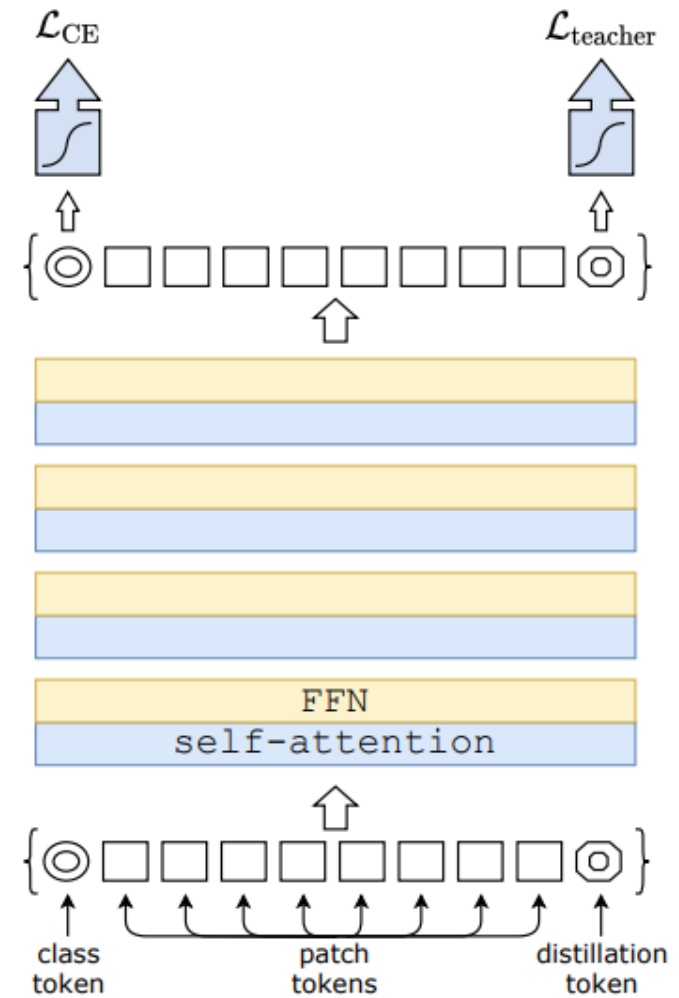
DeiT III: Revenge of the ViT

Hugo Touvron^{*,†} Matthieu Cord[†] Hervé Jégou^{*}

^{*}Meta AI [†]Sorbonne University

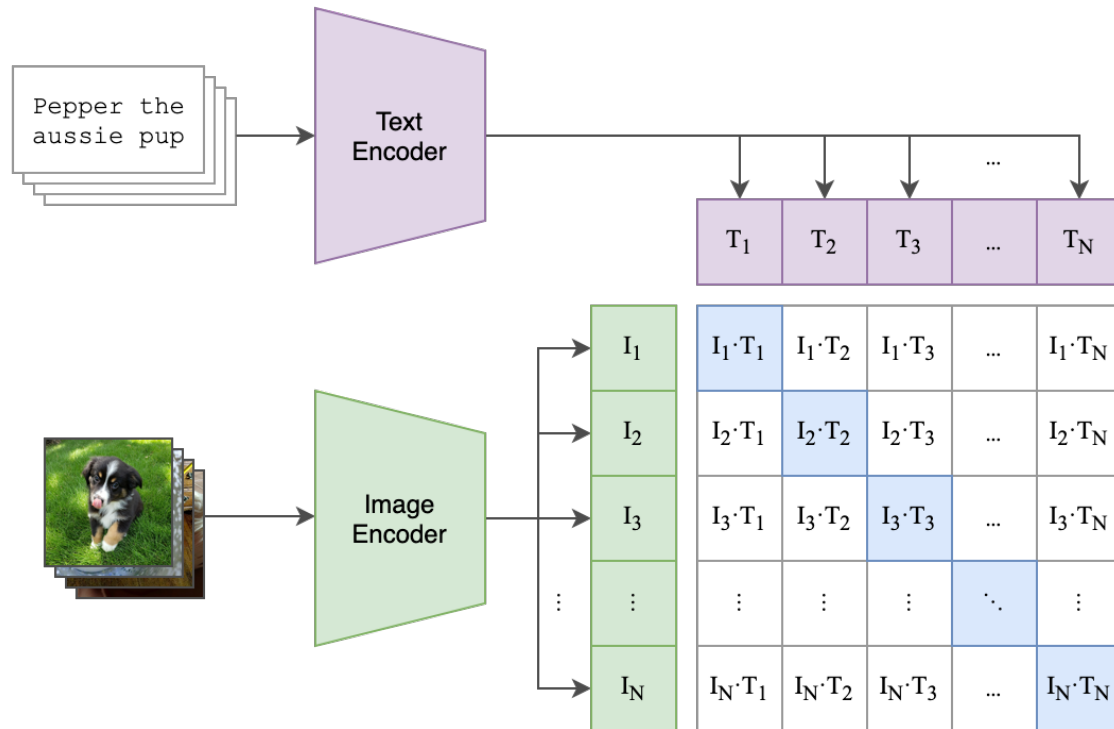
Other ViTs – DEIT-III

- DEIT – Data Efficient Image Transformers.
- DEIT-I used a *distillation token*, DEIT-III does not.

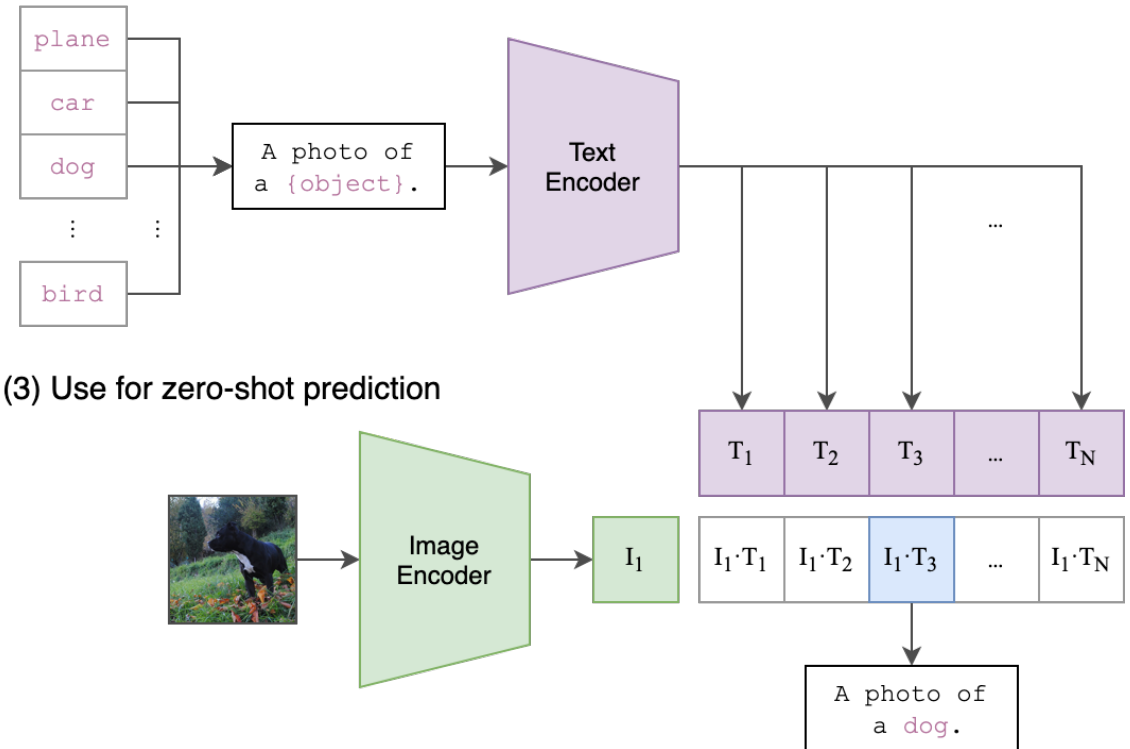


Other ViTs – OpenCLIP

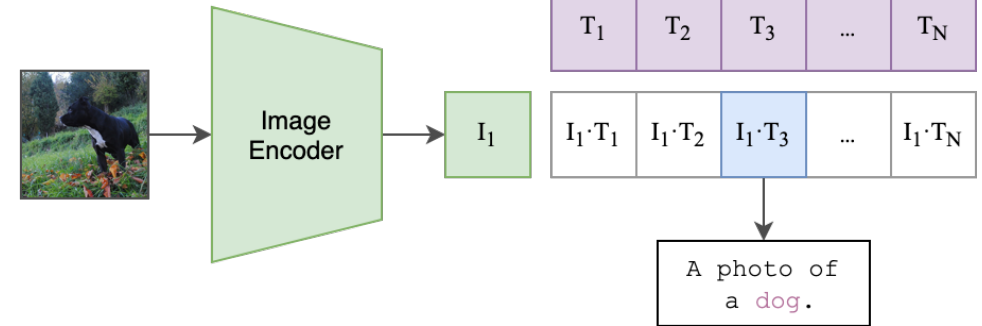
(1) Contrastive pre-training



(2) Create dataset classifier from label text

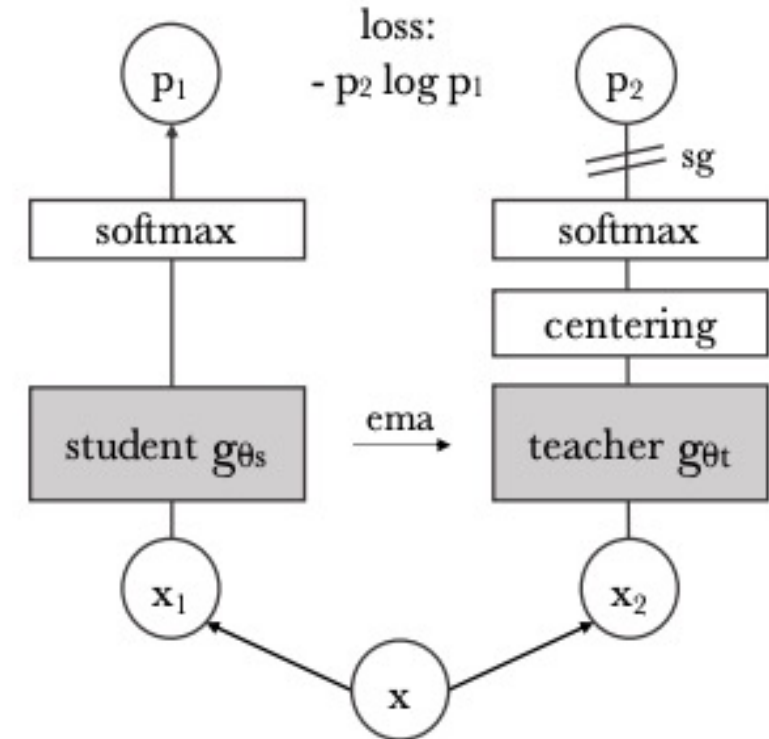


(3) Use for zero-shot prediction



Other ViTs – DINO

- A collection of self-supervised models.
- ‘Self-distillation with no labels’
 - Pass different views on the same image to student/teacher models
 - Contrastive loss



Other ViTs – DINO

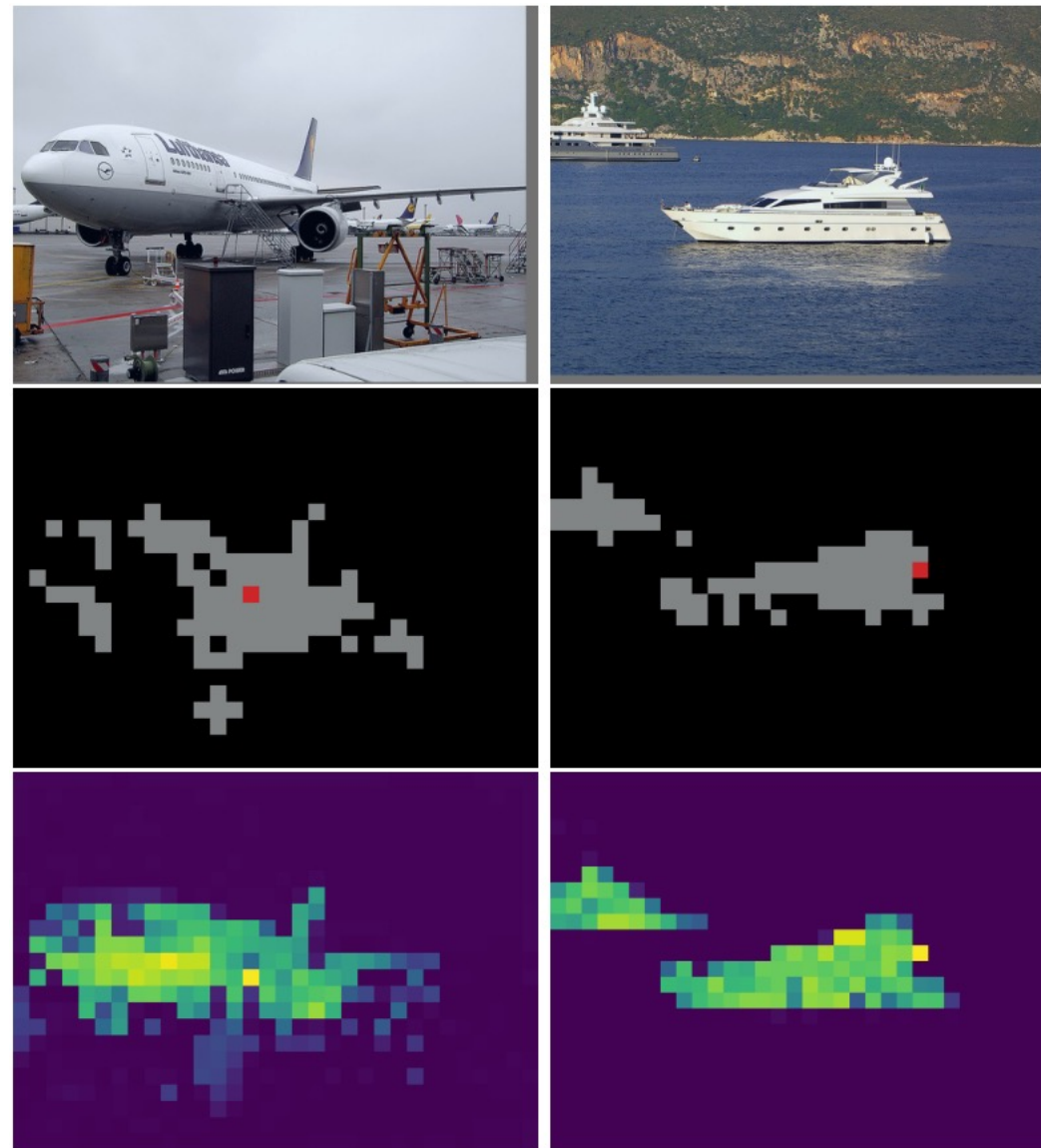
- Importantly, they claim that:
 - They produce features that can be used without finetuning.
 - The attention on the final layer uniquely learns to segment images (based on object discovery algo LOST).



DINO - LOST

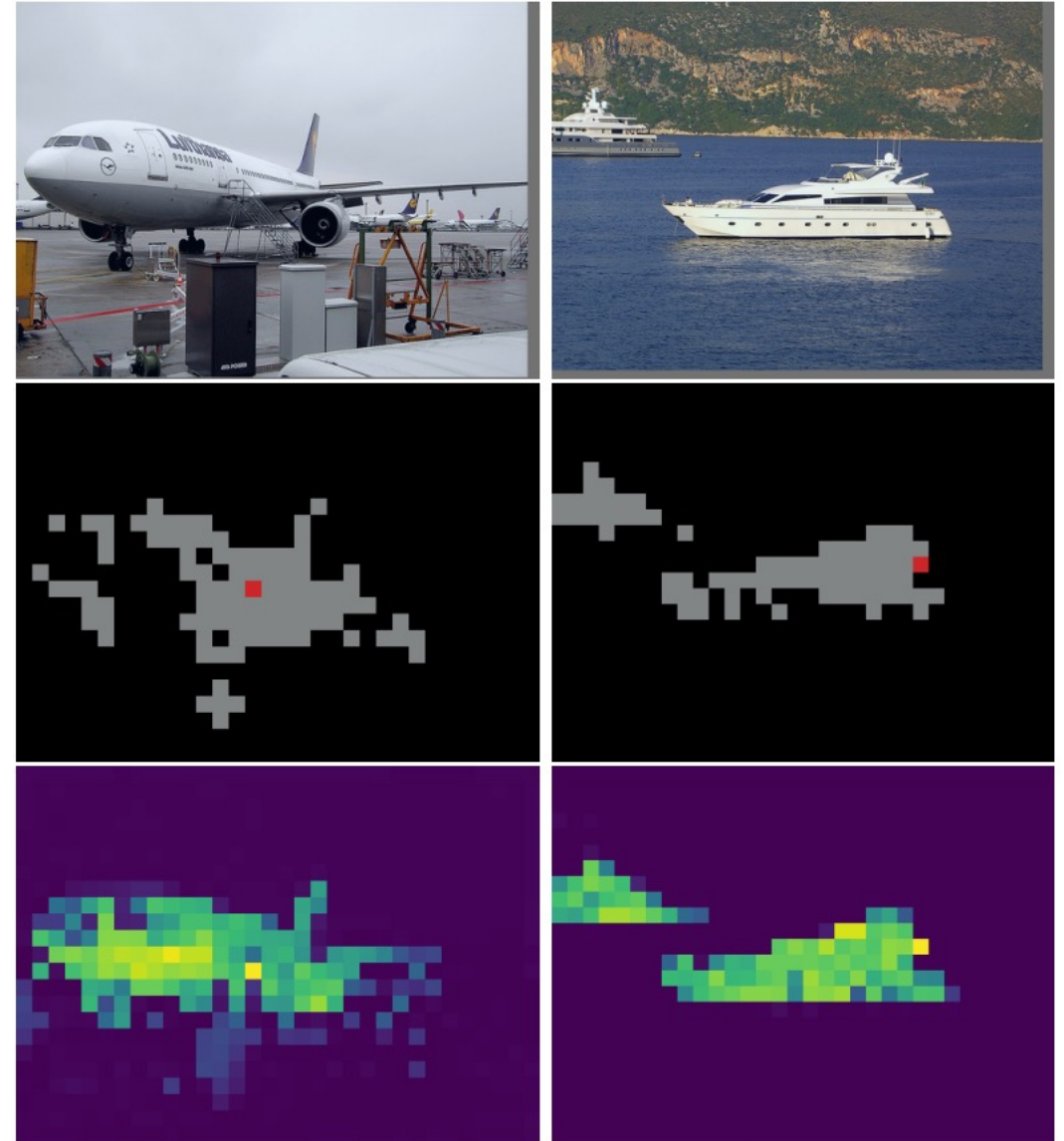
- Compute attention map similarity matrix & correlation graph.
 - Take lowest degree node in patch similarity graph as seed.

$$a_{pq} = \begin{cases} 1 & \text{if } \mathbf{f}_p^\top \mathbf{f}_q \geq 0, \\ 0 & \text{otherwise.} \end{cases}$$



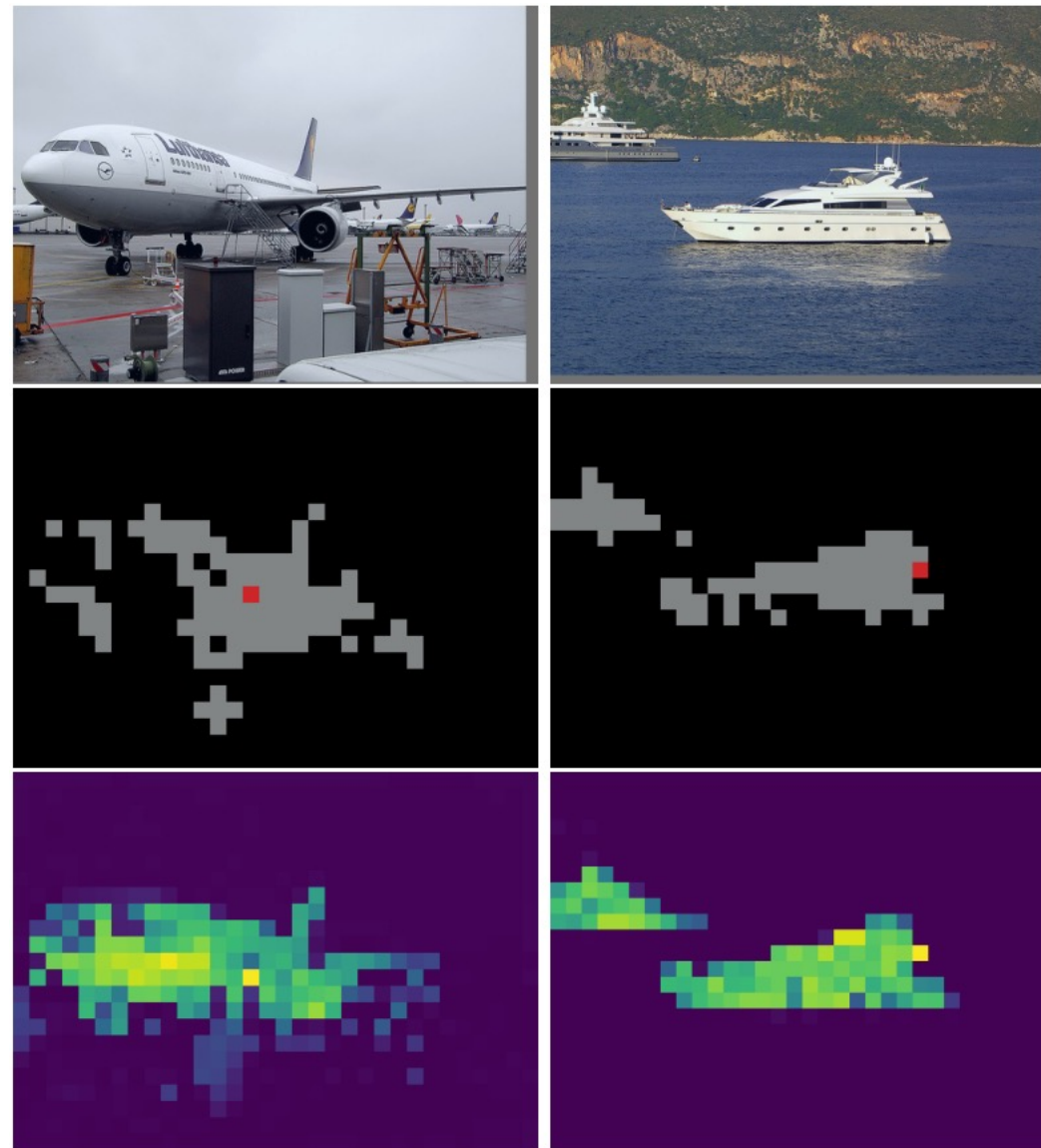
DINO - LOST

- Compute attention map similarity matrix & correlation graph.
 - Take lowest degree node in patch similarity graph as seed.
- Expand to patches which have low global correlation but are positively correlated with the seed.



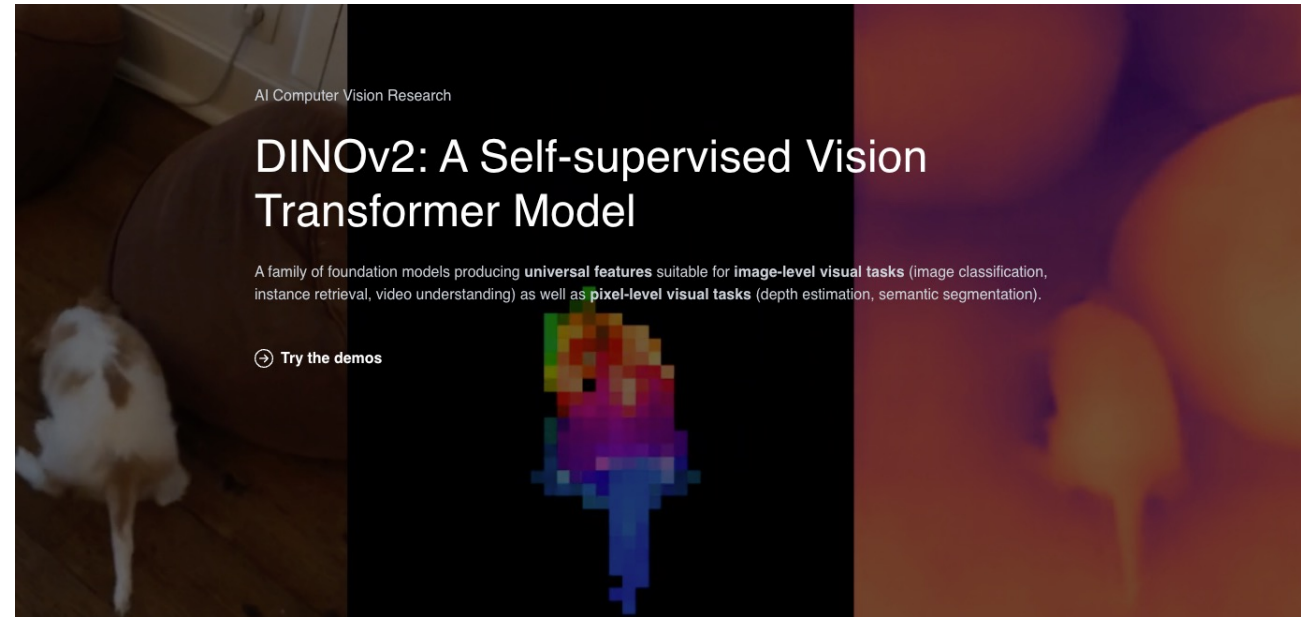
DINO - LOST

- Compute attention map similarity matrix & correlation graph.
 - Take lowest degree node in patch similarity graph as seed.
- Expand to patches which have low global correlation but are positively correlated with the seed.
- Select connected component containing seed, output bounding box of patches.



Other ViTs – DINOv2

- Mainly engineering improvements for scale.
- However, their attention maps are worse for object detection.
 - Their LOST performance is.. lost.



Vision Transformers Need Registers

- Meta's work to understand + fix the degraded attention maps.
- Accepted at ICLR 2024.

VISION TRANSFORMERS NEED REGISTERS

Timothée Darcet^{1,2}, Maxime Oquab¹, Julien Mairal² & Piotr Bojanowski¹

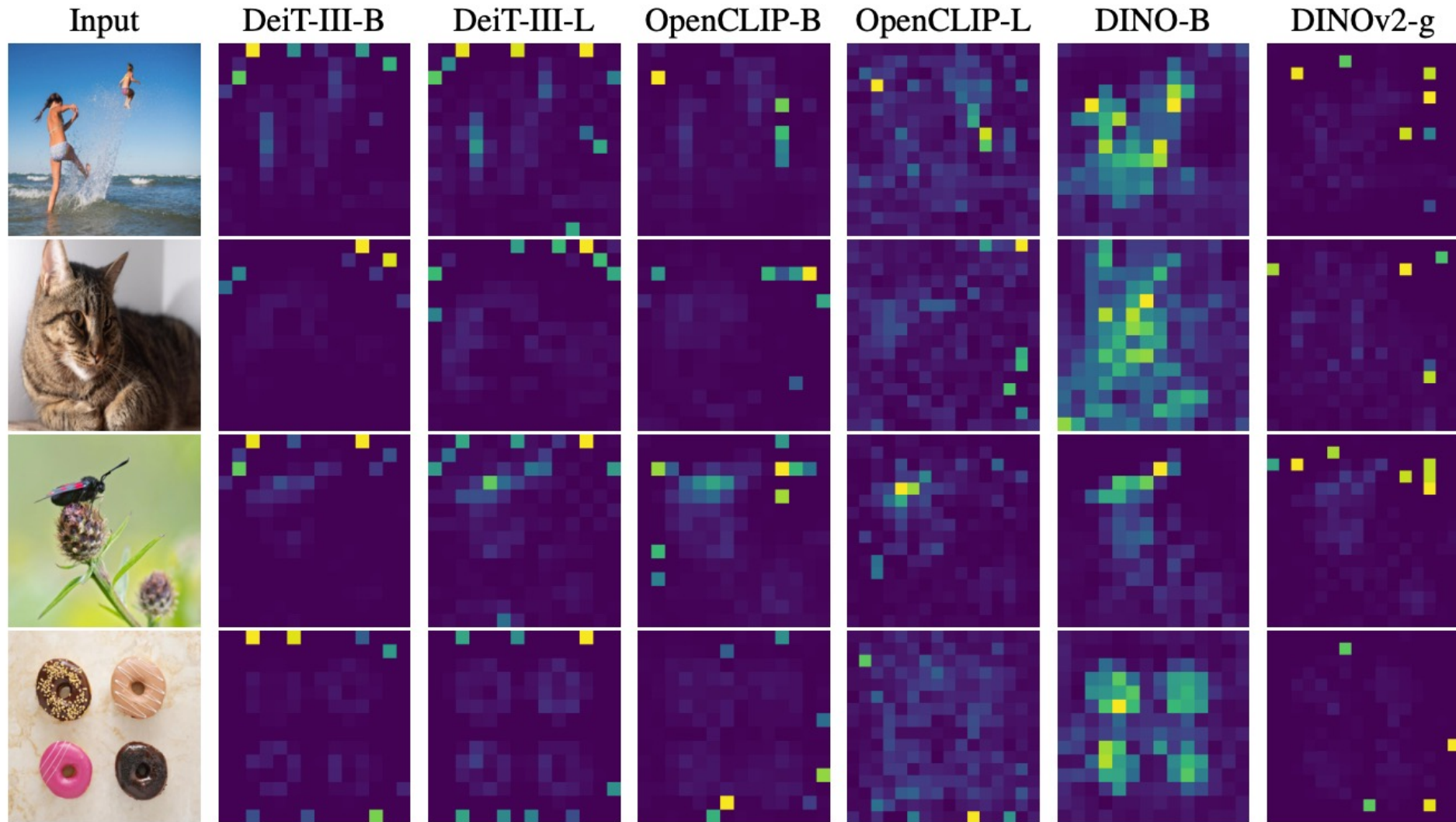
¹ FAIR, Meta

² INRIA

{timdarcet,qas,bojanowski}@meta.com

julien.mairal@inria.fr

Artifacts in attention maps



Investigating the artifacts

- Artifacts come from paying attention to patches with high norm features.

Investigating the artifacts

- Artifacts come from paying attention to patches with high norm features.
- High-norm patch embeddings correspond to redundant patches.

Investigating the artifacts

- Artifacts come from paying attention to patches with high norm features.
- High-norm patch embeddings correspond to redundant patches.
- High-norm patch embeddings hold little local information.

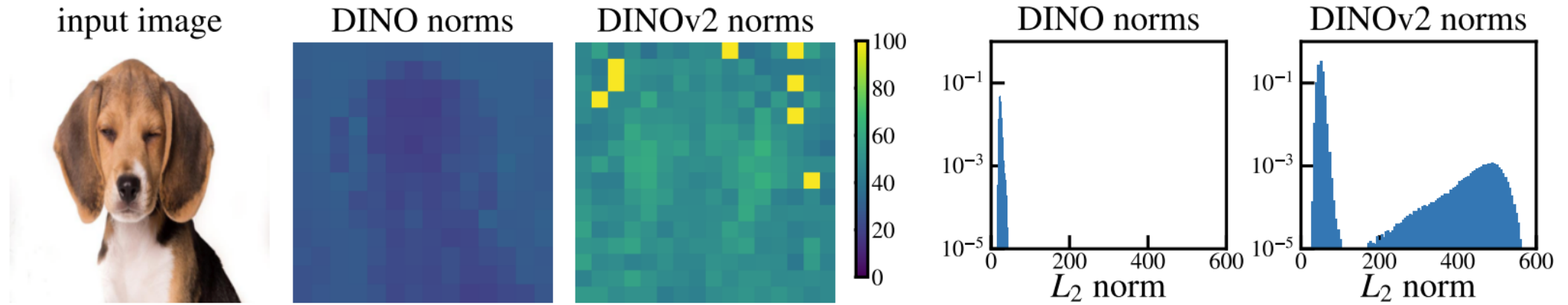
Investigating the artifacts

- Artifacts come from paying attention to patches with high norm features.
- High-norm patch embeddings correspond to redundant patches.
- High-norm patch embeddings hold little local information.
- High-norm patch embeddings hold global information.

Investigating the artifacts

- Artifacts come from paying attention to patches with high norm features.
- High-norm patch embeddings correspond to redundant patches.
- High-norm patch embeddings hold little local information.
- High-norm patch embeddings hold global information.
- Conclusion: background patches are being used as ‘processing space’ to aggregate global information.
 - Good for model performance.
 - Bad for meaningful attention maps.

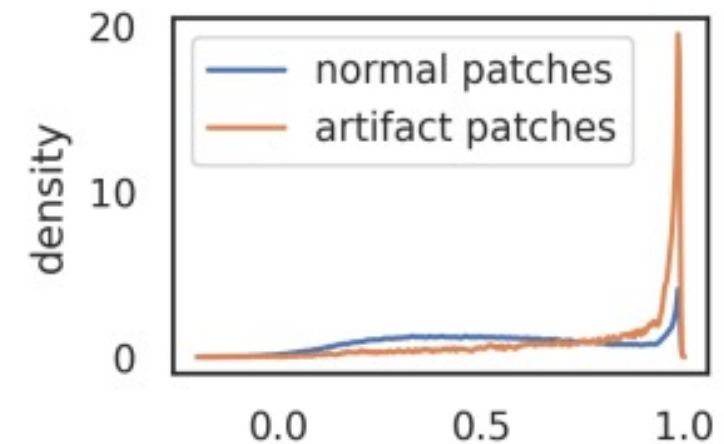
Artifacts come from high norm patch embeddings



- Artifacts in attention maps correspond to patches with high-norms.
- They investigate artifacts further by denoting any patch with norm > 150 as an artifact.

High norms correspond to redundant patches

- Measure the cosine similarity of (initial) patch embeddings to 4 neighbouring patches.
- High norm tokens are very similar to their neighbours.



(a) Cosine similarity to neighbors.

High norm tokens hold little local 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

(b) Linear probing for local information.

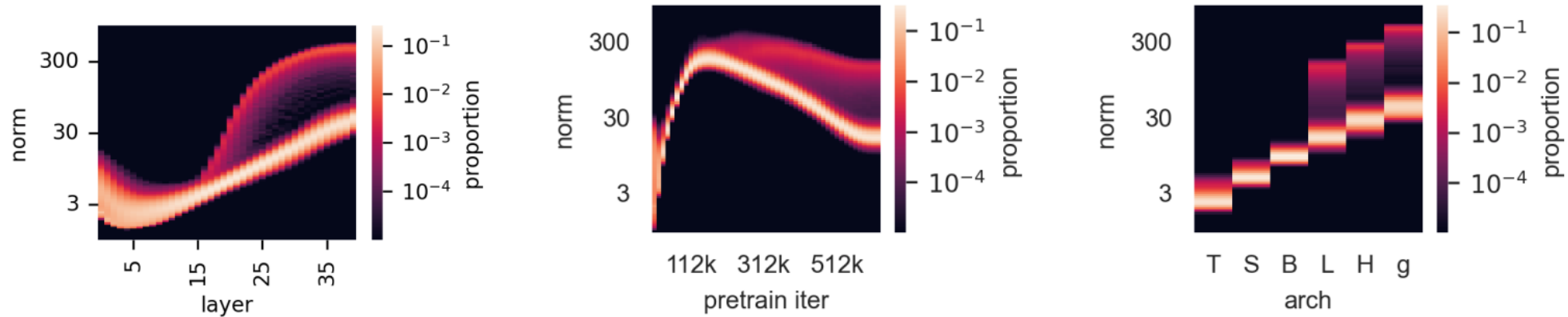
- Train a linear model on (final) patch embeddings for:
 - Position prediction. Can the position embedding be recovered?
 - Pixel reconstruction. Is there still information about the patch?

High norm tokens hold global information

	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	<u>96.9</u>	91.5	85.2	99.7	94.7	96.9	78.6	<u>89.1</u>
normal	65.8	53.1	17.1	97.1	81.3	18.6	<u>73.2</u>	10.8	63.1	59.5	74.2	47.8	37.7	<u>70.8</u>
outlier	<u>69.0</u>	<u>55.1</u>	<u>79.1</u>	<u>99.3</u>	<u>93.7</u>	<u>84.9</u>	97.6	<u>85.2</u>	<u>84.9</u>	<u>99.6</u>	<u>93.5</u>	<u>94.1</u>	<u>78.5</u>	89.7

- Train a model to classify image using one token embedding.
- Outlier token outperforms normal at global classification.

Why DINOv2? -- When do artifacts happen?



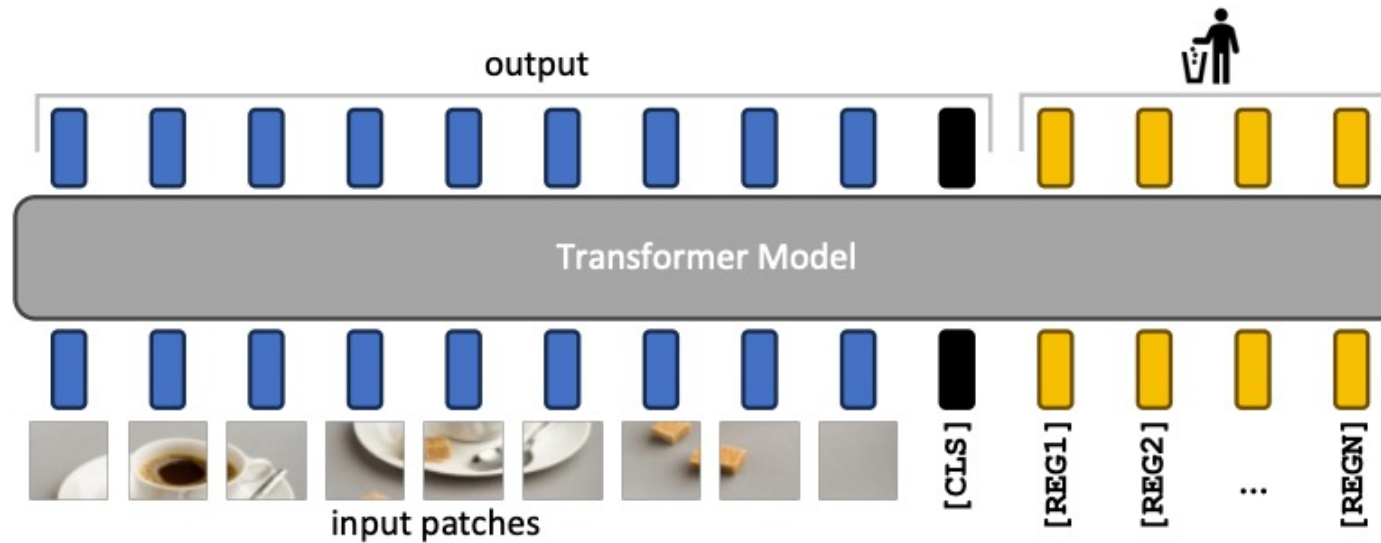
(a) Norms along layers.

(b) Norms along iterations.

(c) Norms across model size.

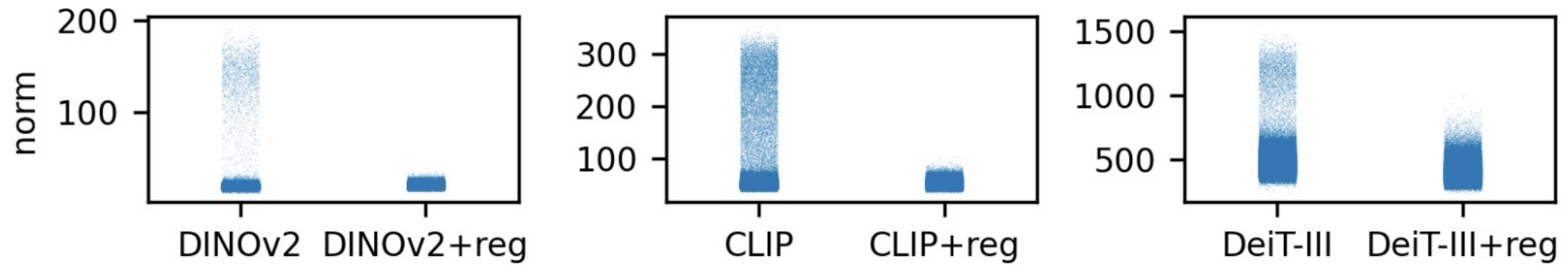
- They typically appear in middle layers.
- They typically appear only after sufficient training.
- These artifacts only appear in sufficiently large models.

Solution: Registers



- Additional tokens that are discarded.
- First proposed in NLP for Memory Transformers (Burtsev et al., 2020).

Results - Norms



- Norms immediately drop with registers.

Results - Performance

	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.

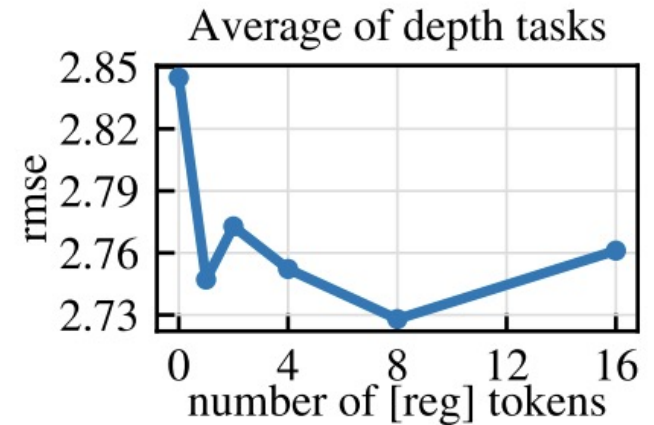
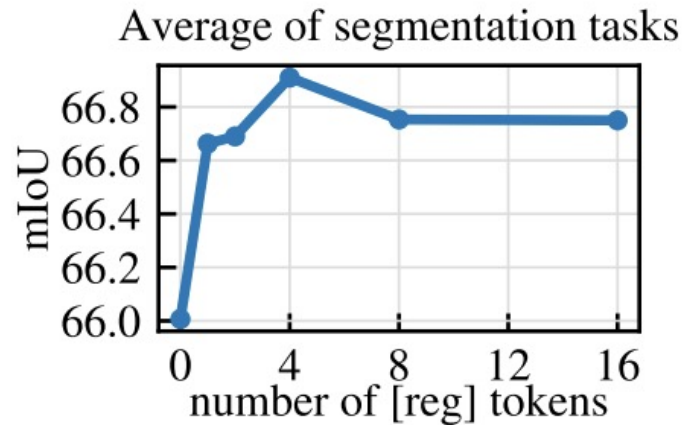
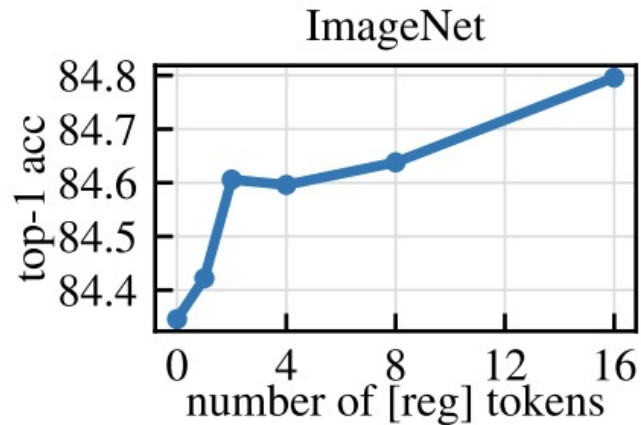
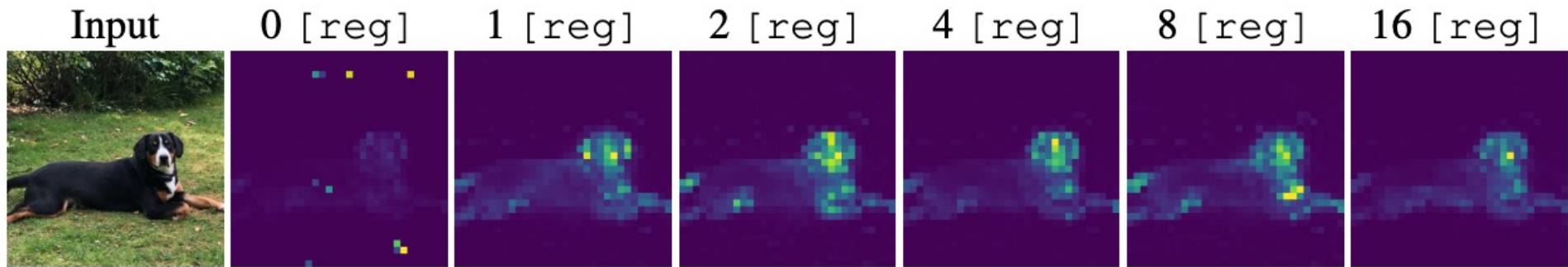
- Task performance is generally the same or better.

Results - LOST

	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

Table 3: Unsupervised Object Discovery using LOST (Siméoni et al., 2021) on models with and without registers. We evaluated three types of models trained with various amounts of supervision on VOC 2007, 2012 and COCO. We measure performance using corloc. We observe that adding register tokens makes all models significantly more viable for usage in object discovery.

Ablation – Number of Registers



Examining registers

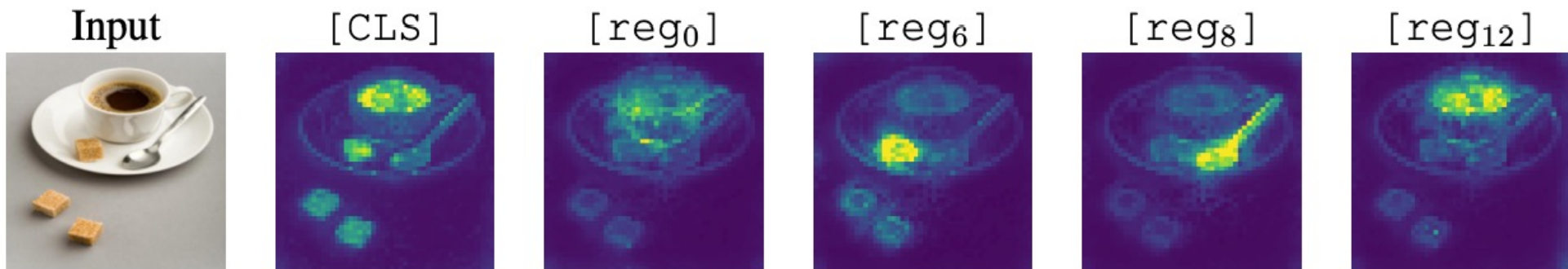




Figure 9: Comparison of the attention maps of the `[CLS]` and register tokens. Register tokens sometimes attend to different parts of the feature map, in a way similar to slot attention (Locatello et al., 2020). Note that this behaviour was never required from the model, and emerged naturally from training.


Models

 `timm/vit_small_patch14_reg4_dinov2.lvd142m`


 Feature Extraction • Updated Nov 22, 2023 •  10.1k

 `timm/vit_base_patch14_reg4_dinov2.lvd142m`

 Feature Extraction • Updated Nov 22, 2023 •  3.72k

 `timm/vit_large_patch14_reg4_dinov2.lvd142m`

 Feature Extraction • Updated Nov 22, 2023 •  1.72k •  1

 `timm/vit_giant_patch14_reg4_dinov2.lvd142m`

 Feature Extraction • Updated Nov 22, 2023 •  552