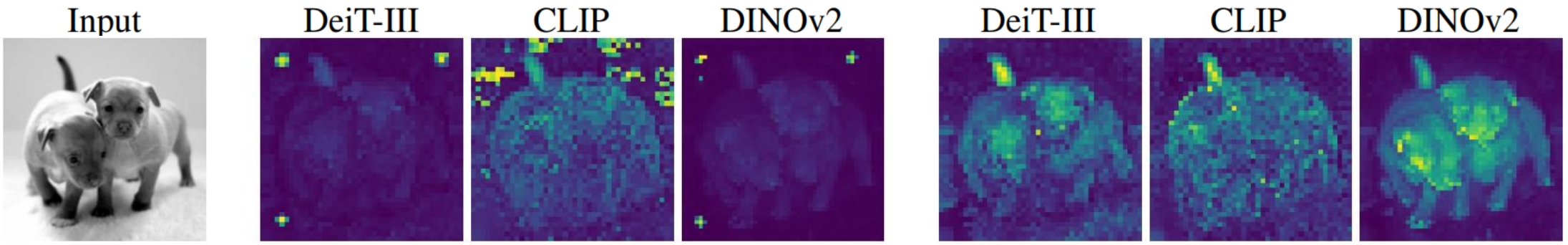
Vision Transformers Need Registers

(28 Sep 2023, FAIR, Meta, INRIA)



Denis Shepelev, Nov 2023
https://github.com/denzist/presentations

Plan

- Problem
- Investigation pipeline
- Solution = additional tokens (registers)
- Results
- Conclusion

Problem

- Background outliers on the attention maps
- Outliers have 10x **higher norm**, 2% of total sequence
- Outliers are in random positions
- This is typical for both supervised and self-supervised ViTs

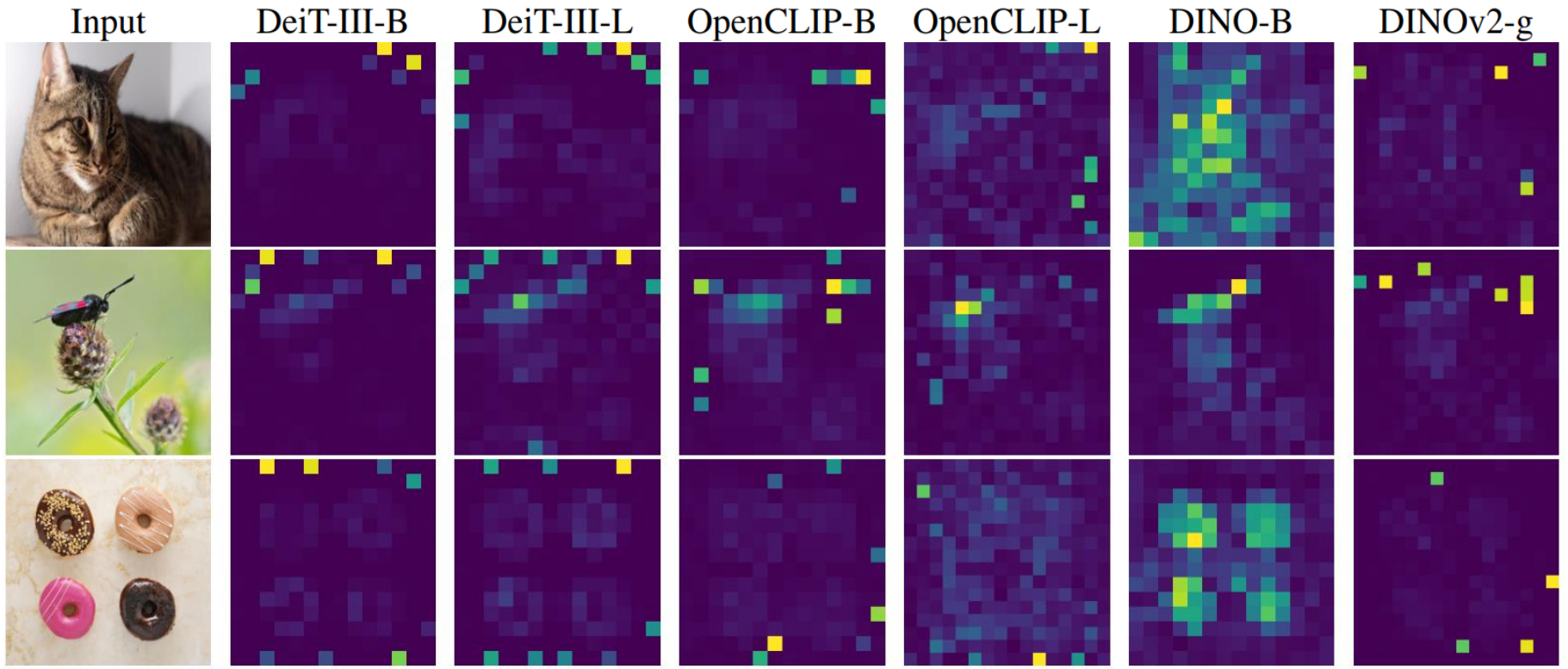


Figure 2: Illustration of artifacts observed in the attention maps of modern vision transformers. We consider ViTs trained with label supervision (DeiT-III), text-supervision (OpenCLIP) or self-supervision (DINO and DINOv2). Interestingly, all models but DINO exhibit peaky outlier values in the attention maps. The goal of this work is to understand and mitigate this phenomenon.

DINO vs DINOv2

- Outliers: norm > 150
- DINOv2 exhibits artifacts
- But DINO is exception

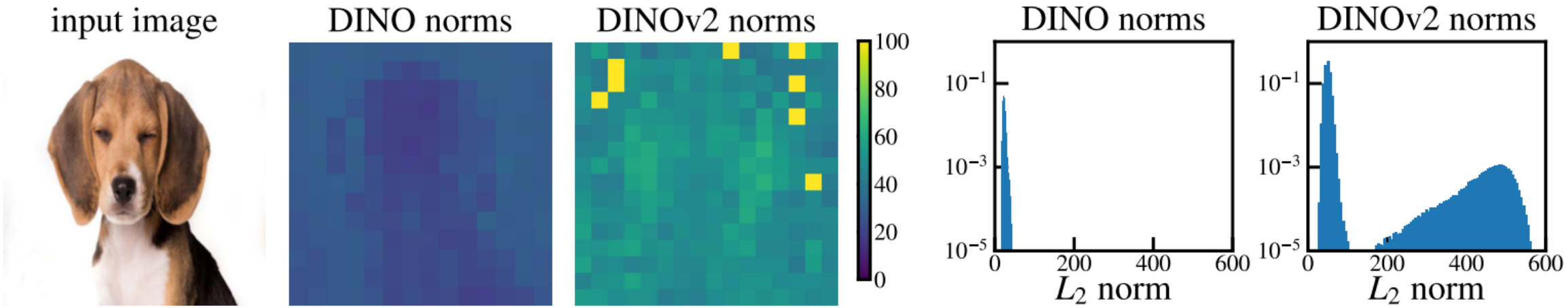
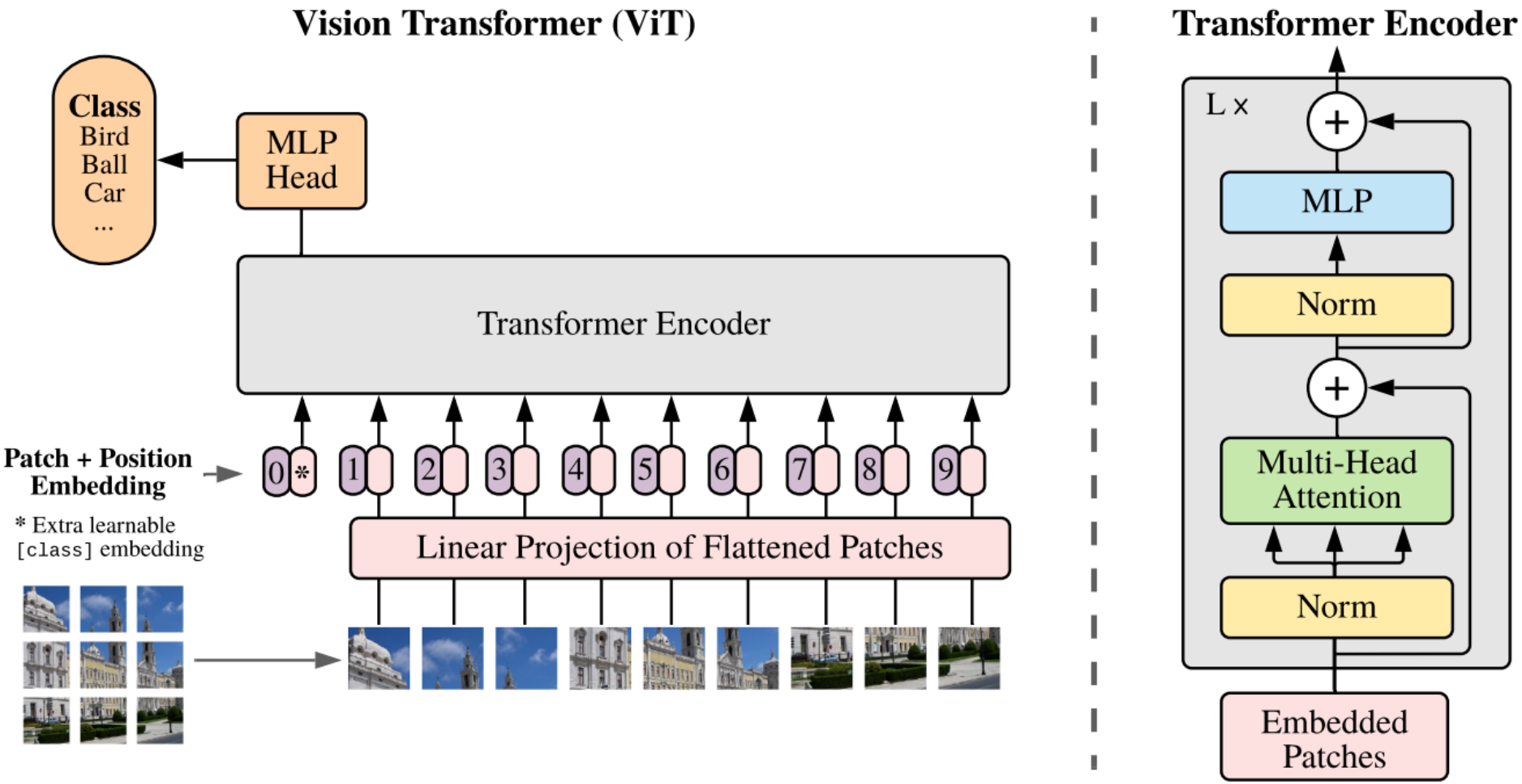
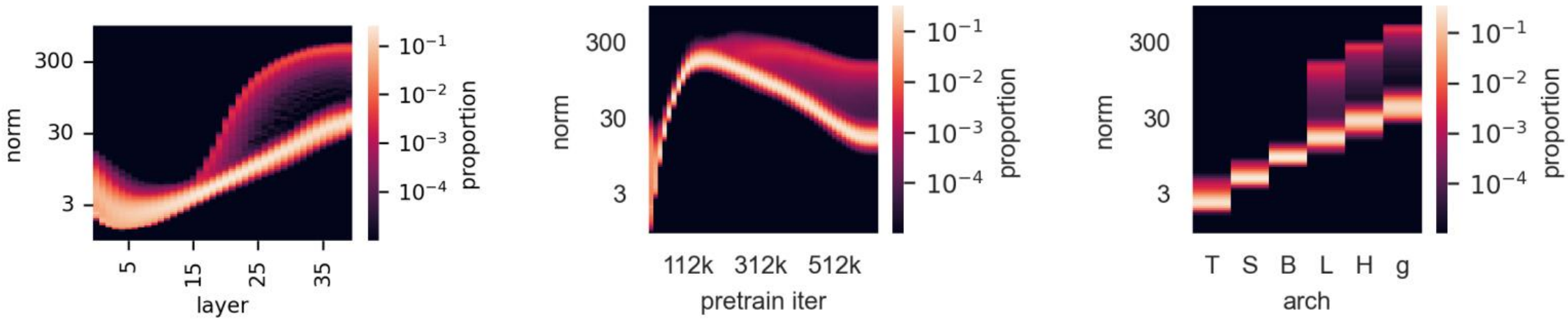


Figure 3: Comparison of local feature norms for DINO ViT-B/16 and DINOv2 ViT-g/14. We observe that DINOv2 has a few outlier patches, whereas DINO does not present these artifacts. For DINOv2, although most patch tokens have a norm between 0 and 100, a small proportion of tokens have a very high norm. We measure the proportion of tokens with norm larger than 150 at 2.37%.

ViT


Vision Transformer (ViT)
Class
Bird
Ball
Car
...
MLP Head
Transformer Encoder
Patch + Position Embedding
* Extra learnable [class] embedding
0 * 1 2 3 4 5 6 7 8 9
Linear Projection of Flattened Patches
Transformer Encoder
L x
MLP
Norm
Multi-Head Attention
Norm
Embedded Patches

Tokens norms distributions

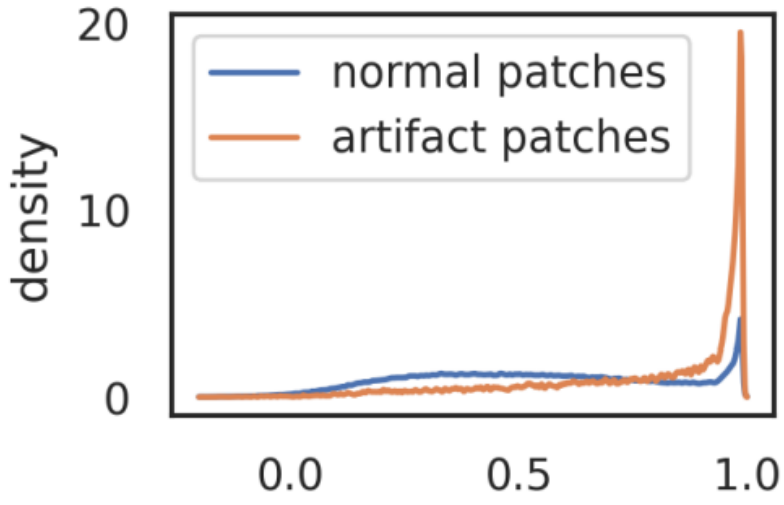


(a) Norms along layers. (b) Norms along iterations. (c) Norms across model size.

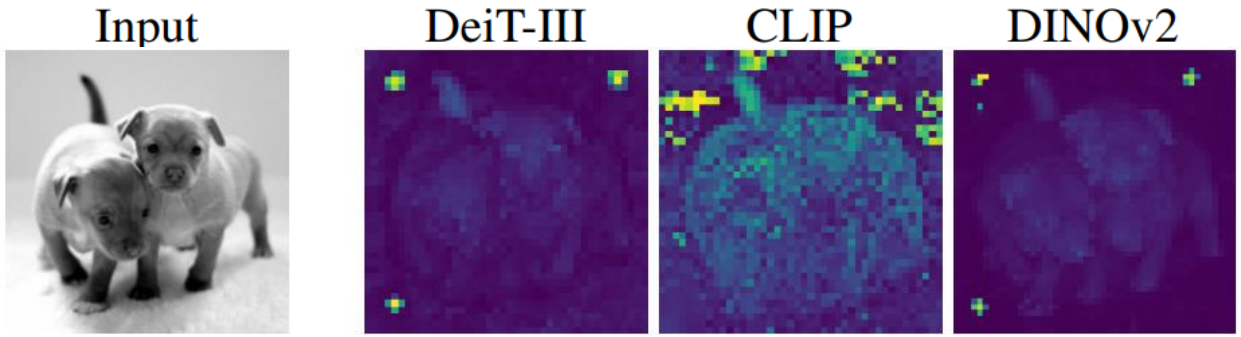
Figure 4: Illustration of several properties of outlier tokens in the 40-layer DINOv2 ViT-g model. **(a)**: Distribution of output token norms along layers. **(b)**: Distribution of norms along training iterations. **(c)**: Distribution of norms for different model sizes. The outliers appear around the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

Outliers appear where patch information is redundant

- To verify this, the cosine similarity between high-norm tokens and their 4 neighbors right after the patch embedding layer (at the beginning of the vision transformer) is measured
- High-norm tokens appear on patches that are very similar to their neighbors
- This suggests that these patches contain redundant information and that the model could discard their information without hurting the quality of the image representation
- This matches qualitative observations that they often appear in uniform, background areas



(a) Cosine similarity to neighbors.



Outliers hold little local information

- Position prediction:
  - Note that position information was injected in the tokens before the first ViT layer
  - Outliers have much lower accuracy than the other tokens, suggesting **they contain less information about their position in the image**
- Pixel reconstruction:
  - Outliers achieve much lower accuracy than other tokens
  - This suggests that **high-norm tokens contain less information to reconstruct the image than the others**

	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.

Outliers hold global information

- Forward each image in a classification dataset through DINOv2-g and extract the patch embeddings
- Choose a single token at random, either high-norm or normal
- This token is then considered as the image representation
- Train a logistic regression classifier to predict the image class from this representation and measure the accuracy. **Outliers have a much higher accuracy than the other tokens**
- This suggests that outlier tokens contain more global information than other patch tokens

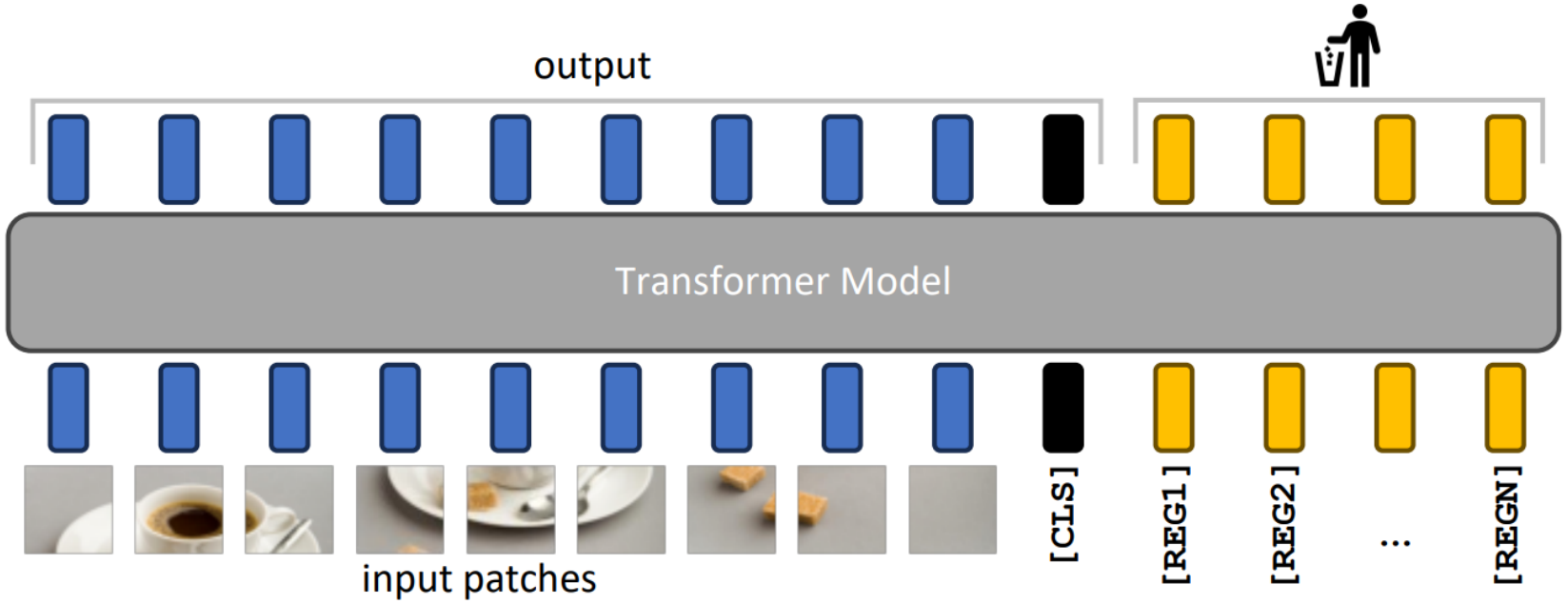
	IN1k	P205	Airc.	CF10	CF100	CUB	Cal101	Cars	DTD	Flow.	Food	Pets	SUN	VOC
[CLS]	**86.0**	**66.4**	**87.3**	**99.4**	**94.5**	**91.3**	96.9	**91.5**	**85.2**	**99.7**	**94.7**	**96.9**	**78.6**	89.1
normal	65.8	53.1	17.1	97.1	81.3	18.6	73.2	10.8	63.1	59.5	74.2	47.8	37.7	70.8
outlier	69.0	55.1	79.1	99.3	93.7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

Summing up problem causes

- Large, sufficiently trained models learn to recognize redundant tokens, and to use them as places to store, process and retrieve global information
  - This is not a bad behavior
  - But it leads possibly to decreased performance on dense prediction tasks
- Fix: explicitly add new tokens to the sequence, that the model can learn to use as registers:
  - We add these tokens after the patch embedding layer, with a learnable value, similarly to the [CLS] token
  - At the end of the vision transformer, these tokens are discarded, and the [CLS] token and patch tokens are used as image representations, as usual
  - This mechanism was first proposed in Memory Transformers (Burtsev et al., 2020), improving translation tasks in NLP
- Why DINO is ok, but DINOv2 is not?
  - Not been able to fully investigated
  - Possibly scaling the model size beyond ViT-L, and longer training length may be possible causes

Fix

- Explicitly add new tokens to the sequence, that the model can learn to use as registers
- Add these tokens after the patch embedding layer, with a learnable value, similarly to the [CLS] token
- At the end of the vision transformer, these tokens are discarded, and the [CLS] token and patch tokens are used as image representations, as usual
- This mechanism was first proposed in Memory Transformers (Burtsev et al., 2020), improving translation tasks in NLP



Experimental setup

- DEIT-III (Touvron et al., 2022):
  - A simple and robust supervised training recipe for classification with ViTs on ImageNet-1k and ImageNet-22k
  - Train this method on the ImageNet-22k dataset, using the ViT-B settings
- OpenCLIP (Ilharco et al., 2021):
  - A method for producing text-image aligned models, following the original CLIP work
  - We run the OpenCLIP method on a text-image-aligned corpus based on Shutterstock that includes only licensed image and text data
  - We use a ViT-B/16 image encoder
- DINOv2 (Oquab et al., 2023):
  - A self-supervised method for learning visual features, following the DINO work
  - We run this method on ImageNet-22k with the ViT-L configuration

Ablations

- In all our experiments, we kept 4 register tokens

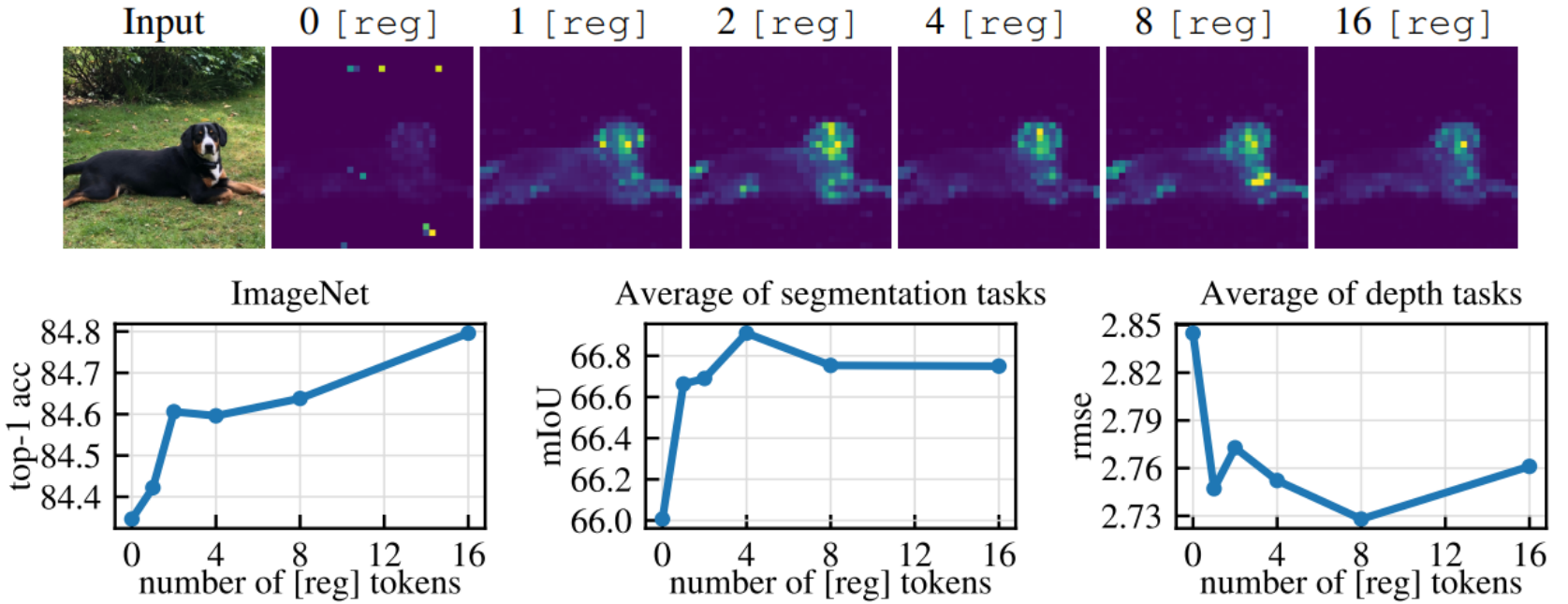


Figure 8: Ablation of the the number of register tokens used with a DINOv2 model. **(top)**: qualitative visualization of artifacts appearing as a function of number of registers. **(bottom)**: performance on three tasks (ImageNet, ADE-20k and NYUd) as a function of number of registers used. While one register is sufficient to remove artefacts, using more leads to improved downstream performance.

Registers fix tokens norms

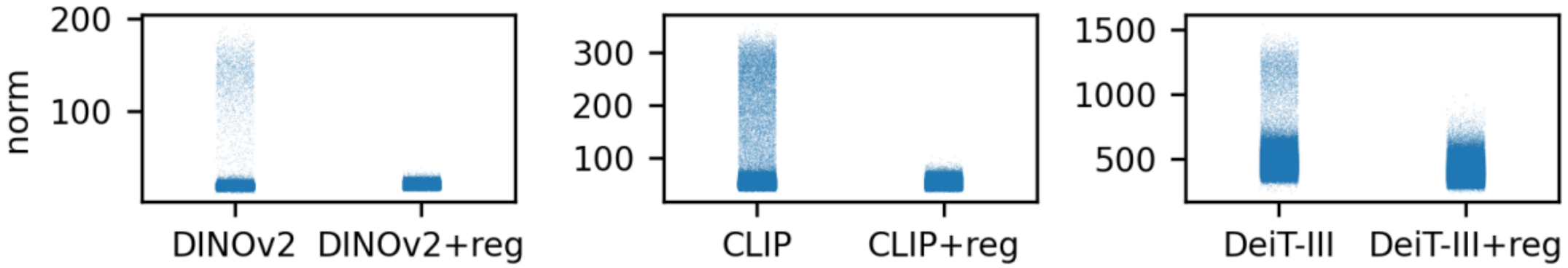


Figure 7: Effect of register tokens on the distribution of output norms on DINOv2, CLIP and DeiT-III. Using register tokens effectively removes the norm outliers that were present previously.

Registers attention maps

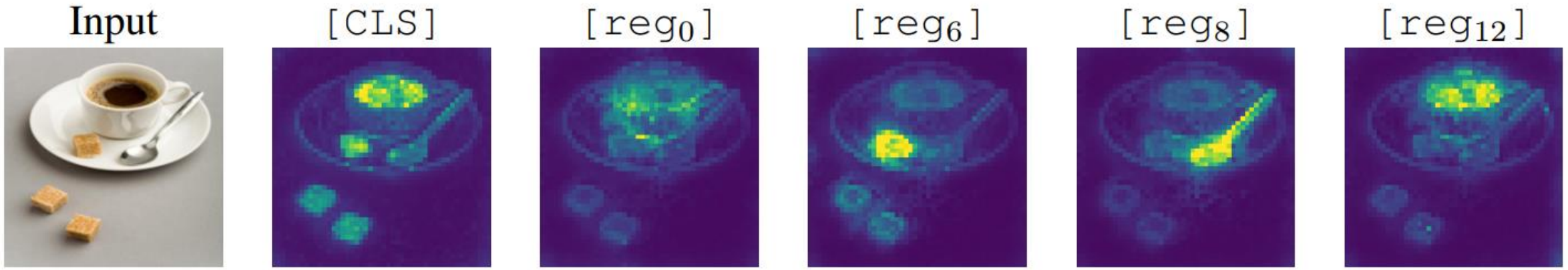


Figure 9: Comparison of the attention maps of the [CLS] and register tokens. Register tokens sometimes attend to different parts of the feature map, in a way similar to slot attention (Locatello et al., 2020). Note that this behaviour was never required from the model, and emerged naturally from training.

Registers do not lose performance (1)

	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.

Table 2: Evaluation of downstream performance of the models that we trained, with and without registers. We consider linear probing of frozen features for all three models, and zero-shot evaluation for the OpenCLIP model. We see that using register not only does not degrade performance, but even improves it by a slight margin in some cases.

Registers do not lose performance (2)

model	# of params	with registers	ImageNet k-NN	ImageNet linear	download
ViT-S/14 distilled	21 M	❌	79.0%	81.1%	backbone only
ViT-S/14 distilled	21 M	✅	79.1%	80.9%	backbone only
ViT-B/14 distilled	86 M	❌	82.1%	84.5%	backbone only
ViT-B/14 distilled	86 M	✅	82.0%	84.6%	backbone only
ViT-L/14 distilled	300 M	❌	83.5%	86.3%	backbone only
ViT-L/14 distilled	300 M	✅	83.8%	86.7%	backbone only
ViT-g/14	1,100 M	❌	83.5%	86.5%	backbone only
ViT-g/14	1,100 M	✅	83.7%	87.1%	backbone only

Conclusions

- ViTs model can learn to store and retrieve information during the forward pass
- This decreases slightly performance on dense tasks
- Fix – additional registers

- **Pros:**
  - Simple fix
  - Fixed DINOv2 is available
  - <2% FLOPS added
- **Coins:**
  - Still FLOPS added
  - Why DINOv1 ok is unclear
  - **Retrain is needed**

Literature

- Paper
- DINOv2