VISION TRANSFORMERS NEED REGISTERS

Timothée Darcet[1,2], Maxime Oquab[1], Julien Mairal[2] & Piotr Bojanowski[1]
[1] FAIR, Meta
[2] INRIA
{timdarcet,qas,bojanowski}@meta.com
julien.mairal@inria.fr

Seungwoo

TL;DR (on X)

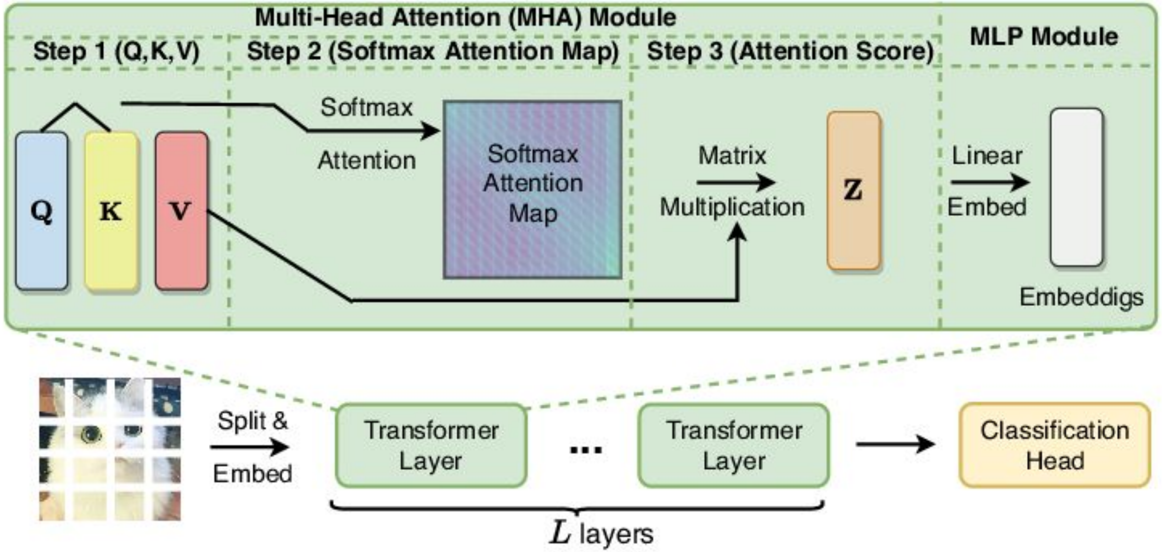
The Vision Transformer recognizes useless patches, discards the info in them, and uses them as *aggregators of global information*.

> **Timothee Darcet**
> @TimDarcet
>
> Vision transformers need registers!
> Or at least, it seems they *want* some…
> ViTs have artifacts in attention maps. It’s due to the model using these patches as “registers”.
> Just add new tokens (“[reg]”):
> - no artifacts
> - interpretable attention maps 🦖
> - improved performances!

Image Source: https://arxiv.org/abs/2211.05109

Preliminaries: Vision Transformer Attention Map

Splitting the image in to patches and use it as a token like NLP



Preliminaries: Vision Transformer Attention Map

Reshape [CLS] token’s attention to (nxn)

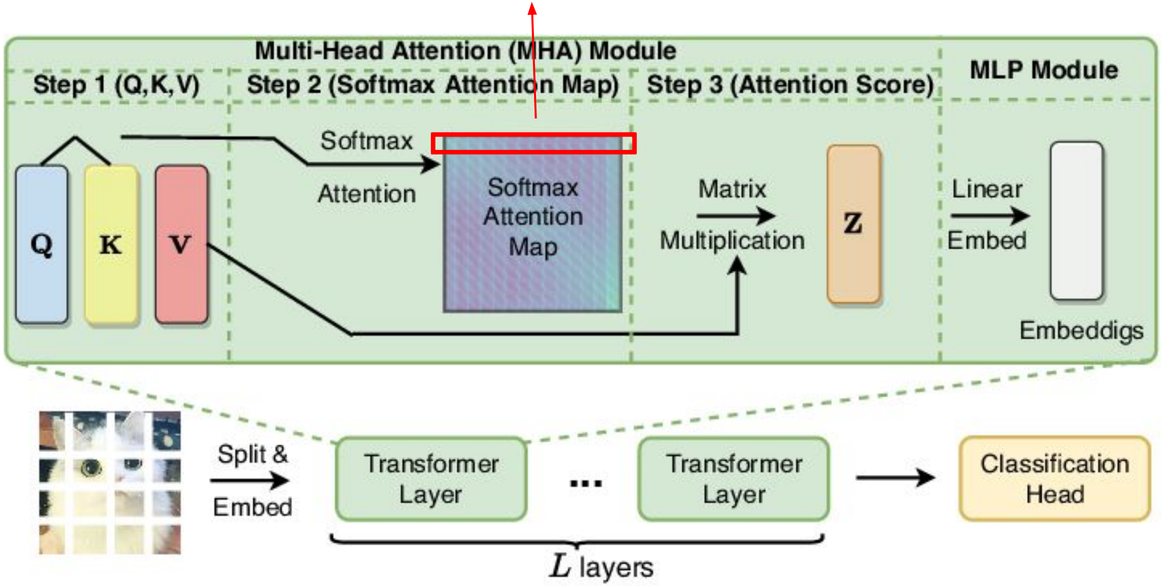
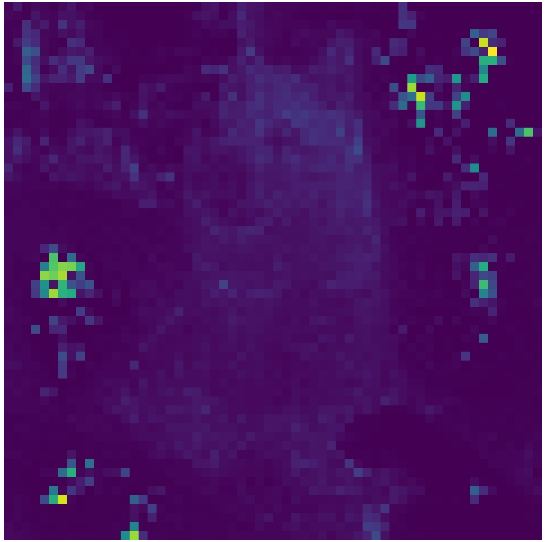


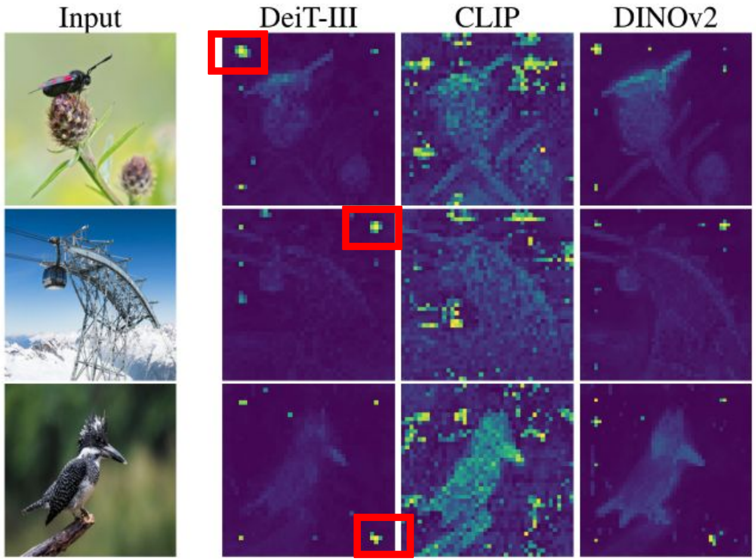
Image Source: Googling (Shiba dog)

Preliminaries: Vision Transformer Attention Map



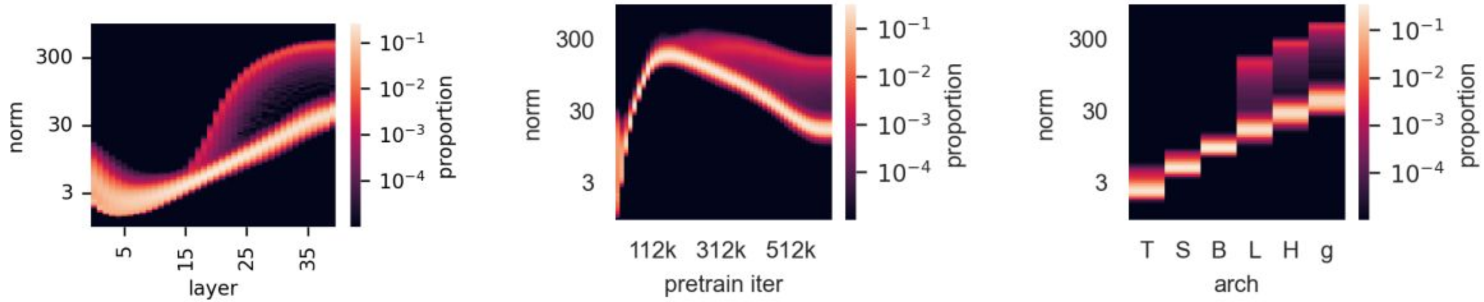
Motivation

most modern vision transformers exhibit artifacts in the attention maps



Features of high-norm outlier tokens

At the output of the model, the norm of artifact patches is much higher than the norm of other patches.



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

Features of high-norm outlier tokens

High-norm tokens hold little local information but hold 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

(b) Linear probing for local information.

Features of high-norm outlier tokens

High-norm tokens hold little local information but 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**	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**

Table 1: Image classification via linear probing on normal and outlier patch tokens. We also report the accuracy of classifiers learnt on the class token. We see that outlier tokens have a much higher accuracy than regular ones, suggesting they are effectively storing global image information.

Hypothesis

High-norm tokens hold little local information but hold global information.

The Vision Transformer **recognizes useless patches**, discards the info in them, and uses them as ***aggregators of global information***.

[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	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	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**

Table 1: Image classification via linear probing on normal and outlier patch tokens. We also report the accuracy of classifiers learnt on the class token. We see that outlier tokens have a much higher accuracy than regular ones, suggesting they are effectively storing global information.

ViT Needs Registers

Adding the [reg] tokens only during training, and discard them during inference

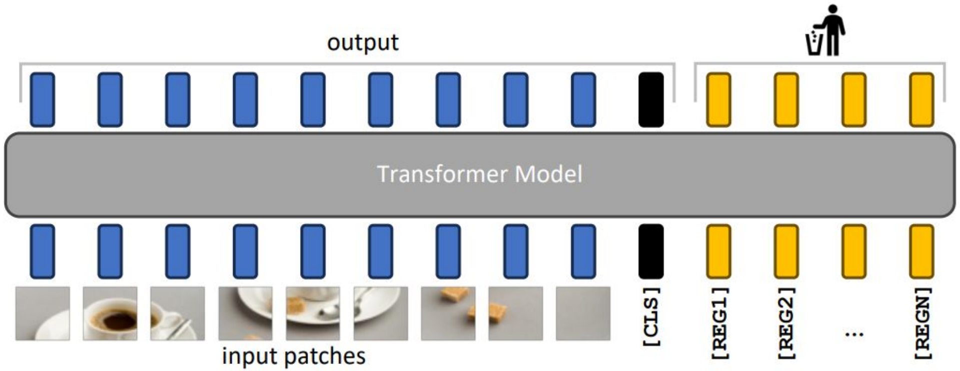
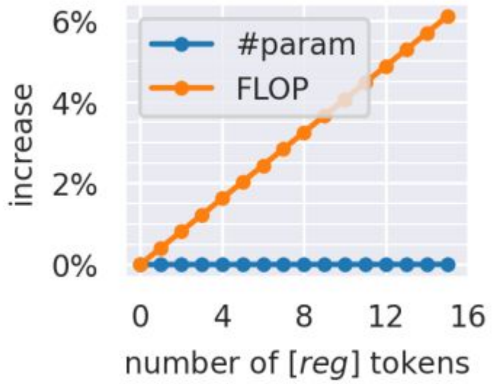
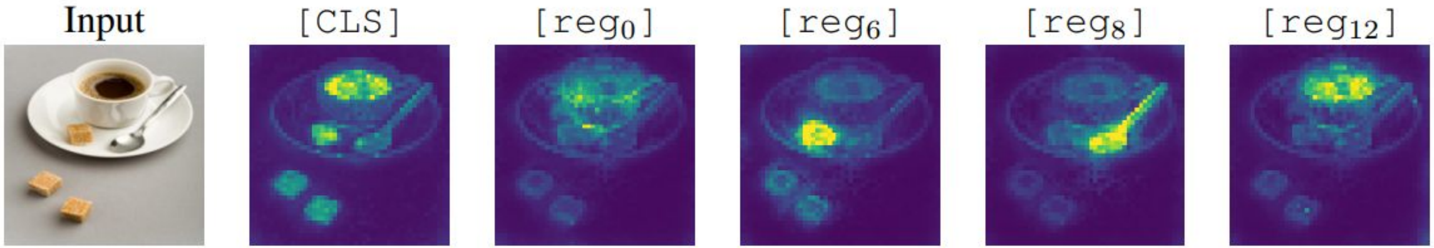


Figure 6: Illustration of the proposed remediation and resulting model. We add N additional learnable input tokens (depicted in yellow), that the model can use as *registers*. At the output of the model, only the patch tokens and CLS tokens are used, both during training and inference.

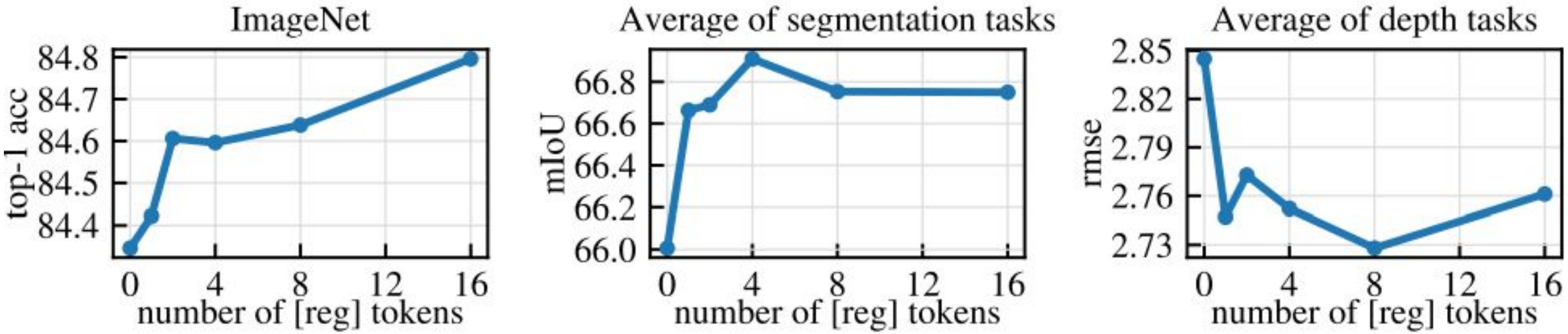


ViT Needs Registers

Each [reg] token is responsible for focusing on different regions of the image.



ViT Needs Registers



Related Works of Additional Tokens In Transformers

for classification: [cls] tokens in BERT and ViT

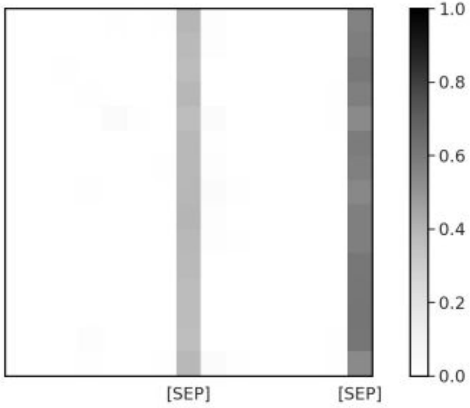
for generative learning: [Mask] tokens in BERT and BEiT

for detection: [detection] tokens in YOLOS, and ViDT

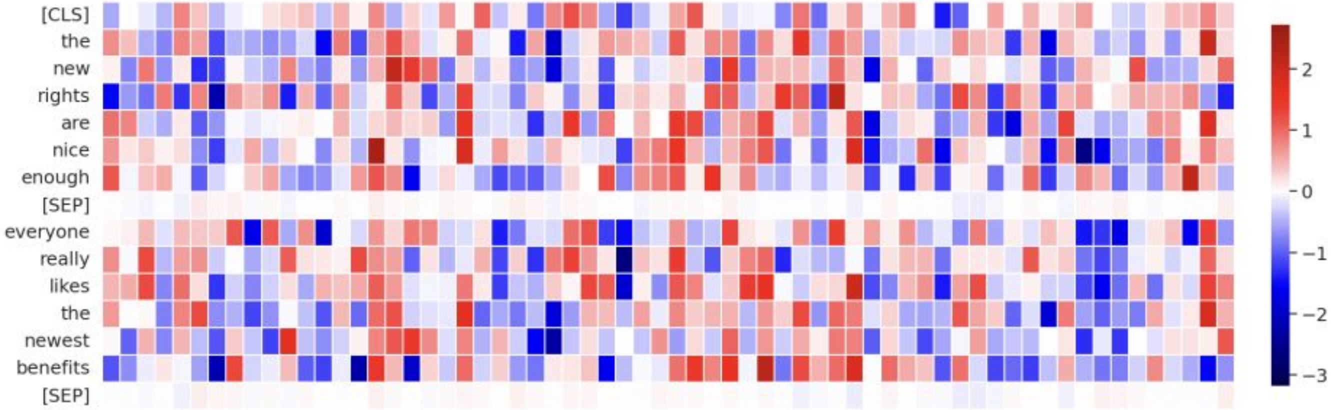
Different to these works, the tokens they add to the sequence add <u>no information</u>, and <u>their output value is not used for any purpose</u>.

Attention Sink in BERT

Meaningless tokens (e.g., [SEP] token) take much attention in BERT.



Attention



Value

Attention Sink in BERT

Attention sinks need very large QK and this gives rise to **big outlier channels** (arguably).

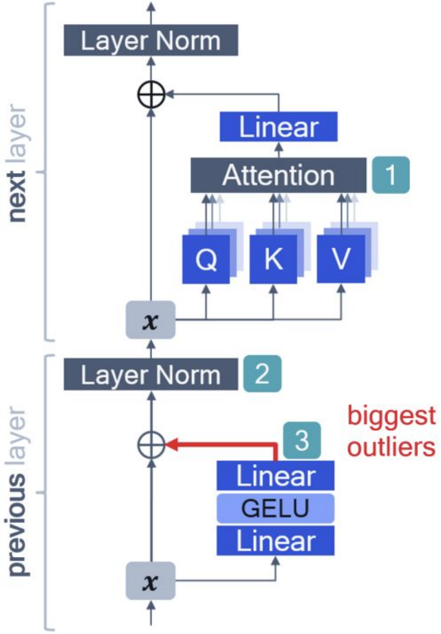


Figure 4: A schematic illustration of the attention layer in BERT. Hidden activation tensor is denoted by **x**. ⊕ is an element-wise addition. A problematic output of the FFN that generates largest in magnitude outliers is highlighted in red. Notice how those outliers in the *previous layer* influence the behavior in the attention mechanism in the *next layer*.

Attention Sink in LLMs

Attention sinks occur in LLM, for **first N tokens.**

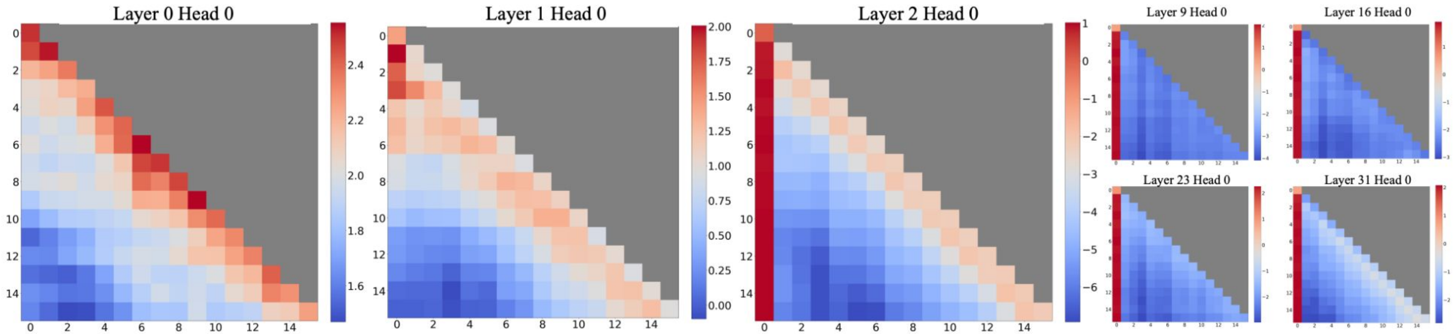


Figure 2: Visualization of the *average* attention logits in Llama-2-7B over 256 sentences, each with a length of 16. Observations include: (1) The attention maps in the first two layers (layers 0 and 1) exhibit the "local" pattern, with recent tokens receiving more attention. (2) Beyond the bottom two layers, the model heavily attends to the initial token across all layers and heads.

Thank you and Questions?