ICLR 2024 oral

汇报人：程泽轩

Vision Transformers Need Registers

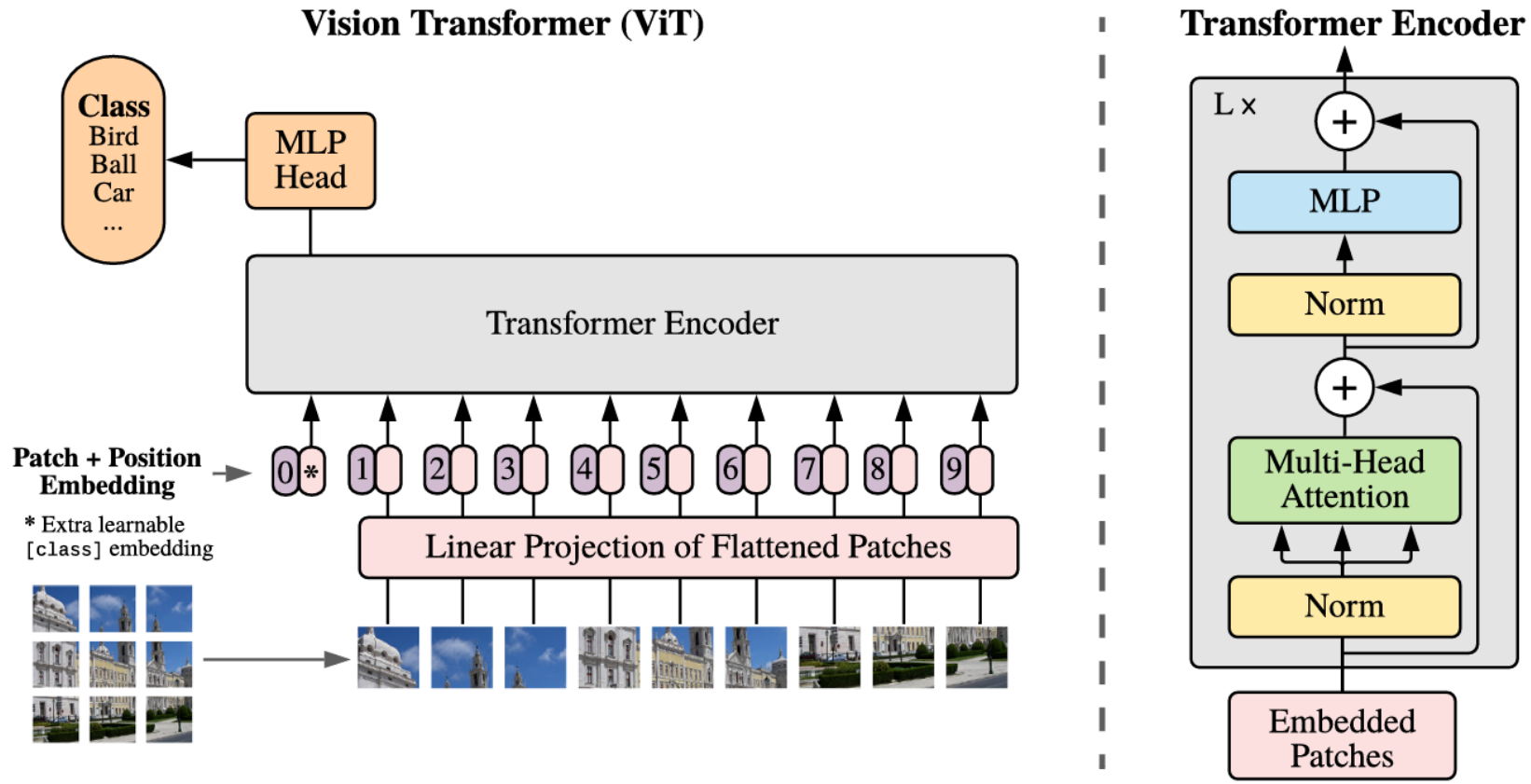
Timothée Darcet, Maxime Oquab, Julien Mairal, Piotr Bojanowski

FAIR, Meta
INRIA

研究问题

➢问题背景

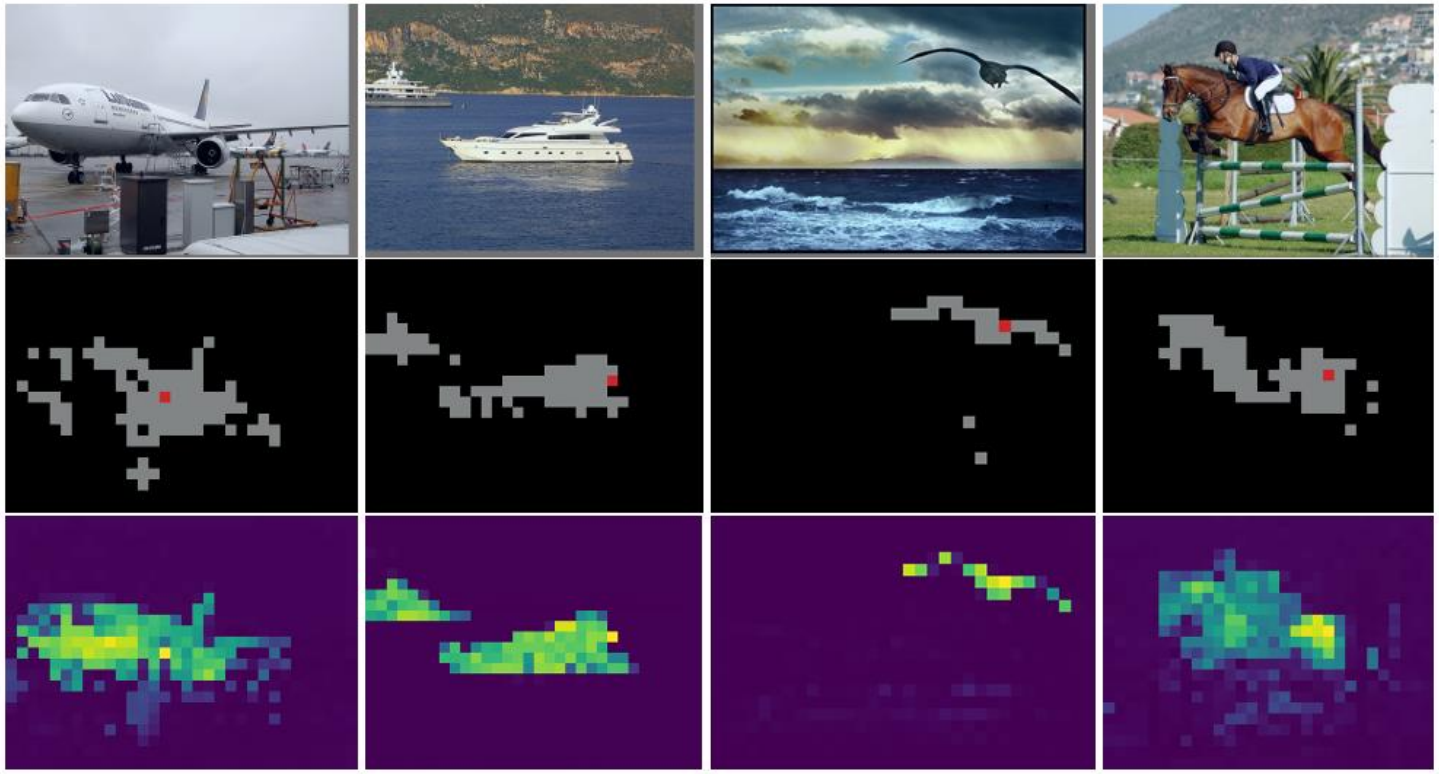
- ➢Vision Transformer, ViT（ICLR 2021）
- ➢Image -> Patches



研究问题

➢问题发现

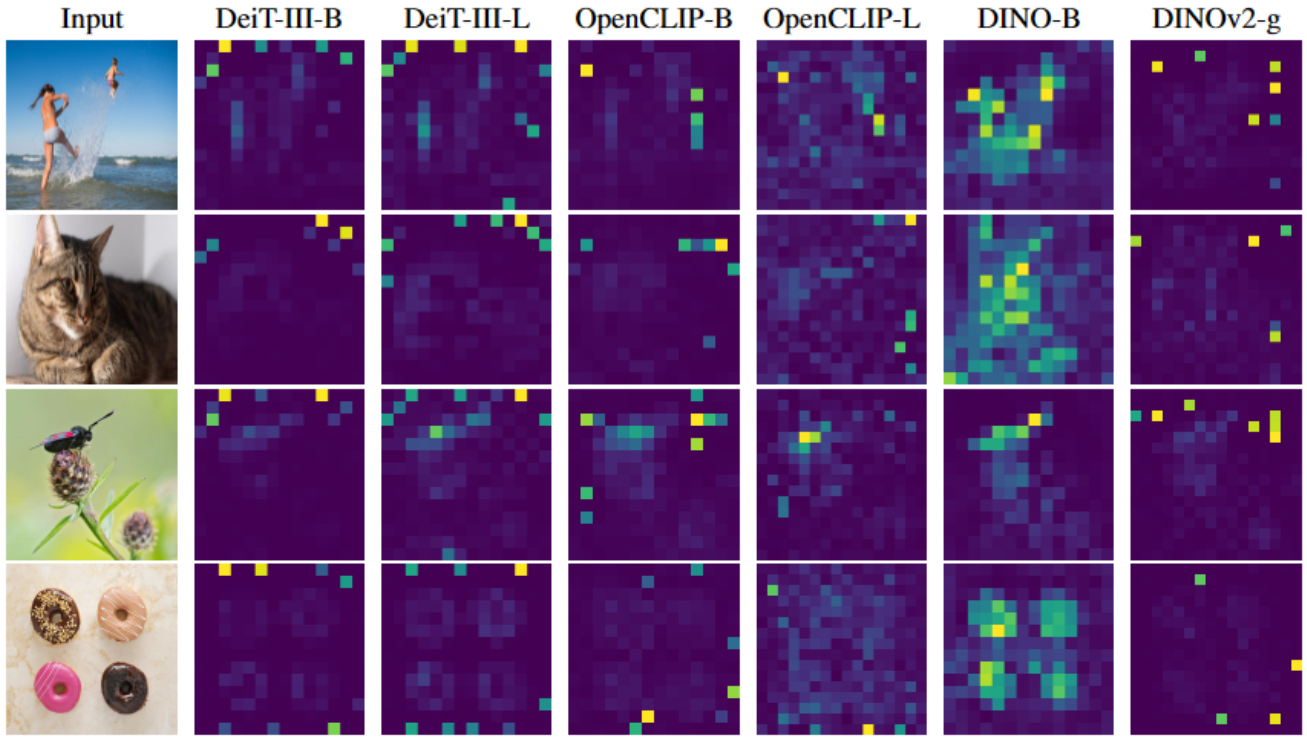
- ➢DINO：自监督训练的Vision Transformer基座（ICCV2021）
- ➢LOST：在DINO等Vision Transformer基座的特征图上完成视觉任务
- ➢特征图即Patches的特征，也即在前向过程中Transformer的q、k、v



研究问题

➢问题发现

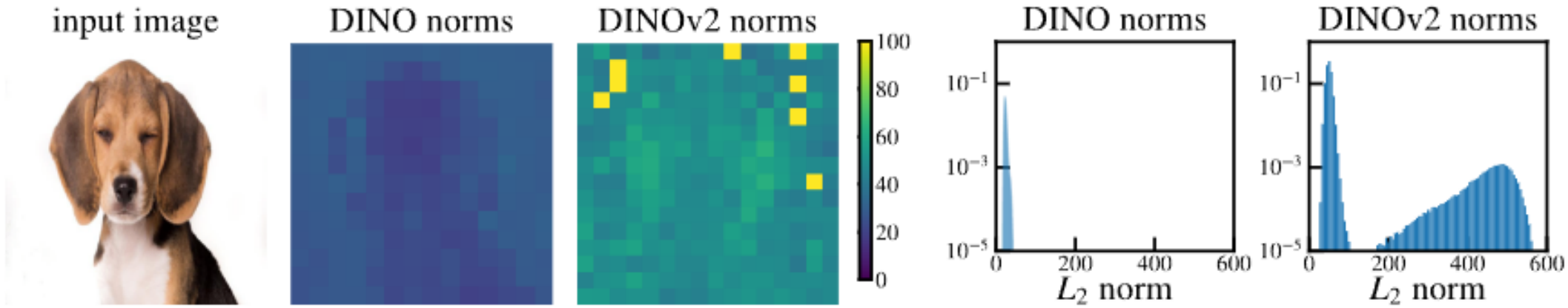
- ➢LOST方法在使用DINO时效果出色，而Vision Transformer基座替换为更大更强的DINOv2时反而变差
- ➢作者检查发现，在DeiT、OpenCLIP、DINOv2等更强基座的特征图上出现了异常点，而DINO没有。本文的目标是尝试寻找产生该现象的原因



研究问题

➢问题深入

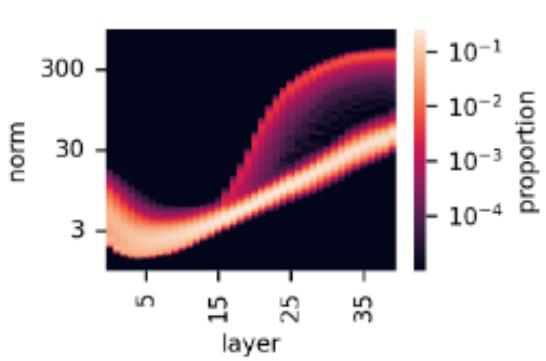
- ➢作者分析了DINO与DINOv2在前向过程中，Patch（也即Transformer的q、k、v）的L2范数
- ➢作者发现，相比于DINO，DINOv2有一小部分的的Patch有明显更高的L2范数。根据统计，有2.37%的Patch具有高于150的L2范数



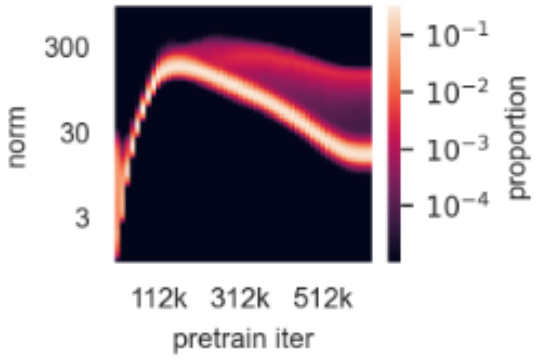
研究问题

➢问题深入

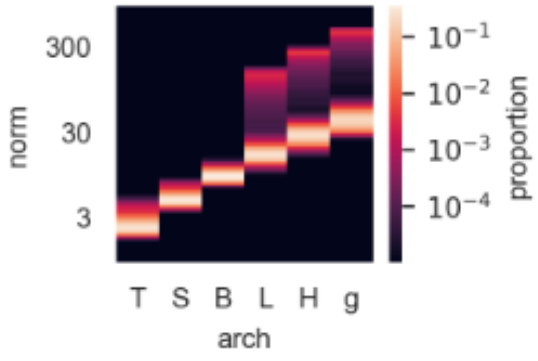
- ➢针对出现这一异常现象的Vision Transformer基座之一的DINOv2，作者进行了详细分析
- ➢作者选择40层的DINOv2 ViT-g模型。图a展示了每一层中Patch的L2范数的分布情况；图b展示了随着模型预训练迭代次数的增加，Patch的L2范数的变化趋势
- ➢图c展示了不同模型大小下，Patch的L2范数的分布情况，可以看到只有在L以上的模型中，出现了异常的L2范数分布



(a) Norms along layers.



(b) Norms along iterations.

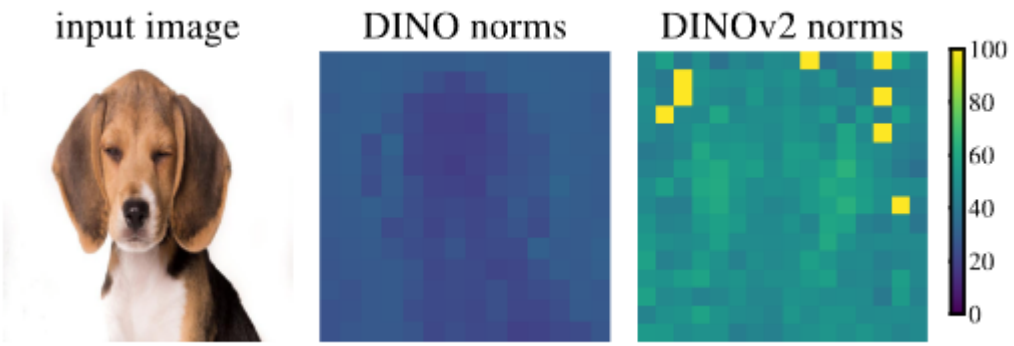


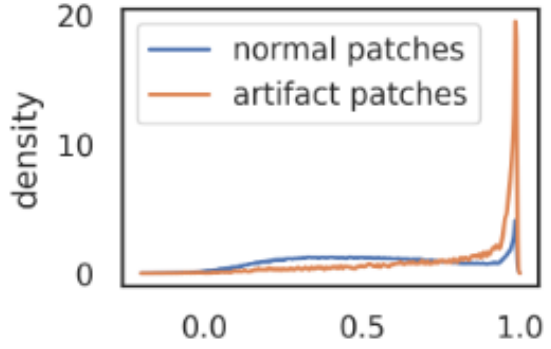
(c) Norms across model size.

研究问题

➢问题深入

- ➢观察出现L2范数异常的Patch，发现这些Patch大都出现在图像中没有关键信息的部分，例如下图中异常Patch均出现在图像中的空白区域
- ➢作者提出猜想：异常Patch出现在图像中具有多余信息的部分
- ➢猜想验证：作者计算了所有异常Patch与周围相邻Patch的相似度，与正常Patch的计算结果对比，发现异常Patch与周围Patch的相似度明显更高





(a) Cosine similarity to neighbors.

研究问题

➢问题深入

- ➢作者思考：这些Patch是否携带了其他特殊信息呢?
- ➢作者将Patch携带的信息分为局部信息和全局信息。
- ➢针对局部信息，作者设计了实验，训练了额外一个线性层，使用Patch的Embedding来完成位置预测和像素重建的任务
- ➢根据实验结果，异常Patch在位置预测和像素重建任务上的表现明显差于正常Patch，这表明这些Patch携带更少的局部信息

	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.

研究问题

➢问题深入

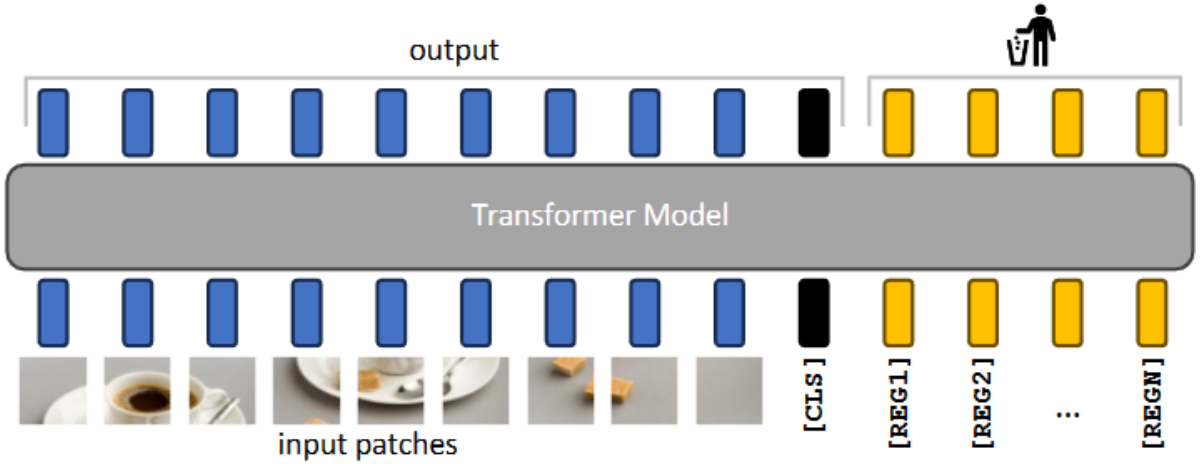
- ➢对于全局信息，作者设计了实验，使用不同的Patch来预测图像分类结果
- ➢作者得到了如图所示的实验结果。CLS表示使用额外补充的分类token来预测，而normal和outlier表示使用正常Patch和异常Patch来进行预测
- ➢可以看到，异常Patch比正常Patch在预测图像分类这种全局信息上反而效果更好，这证明异常Patch携带更多的全局信息。

	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**

研究问题

➢思考与假设

- ➢在综合以上现象后，作者给出了假设：大型的Vision Transformer基座模型已经学习到识别局部冗余信息，并将携带局部冗余信息的Patch用作存储、处理和而检索全局信息的载体
- ➢作者认为这种现象会导致模型丢弃Patch本身的局部信息，可以尝试改进
- ➢作者给出方案：与其让模型将局部冗余的Patch作为全局信息的载体，不如直接补充额外的“Patch”以实现冗余信息的处理，即Register；在进行预测任务时，这些Register被直接丢弃



研究问题

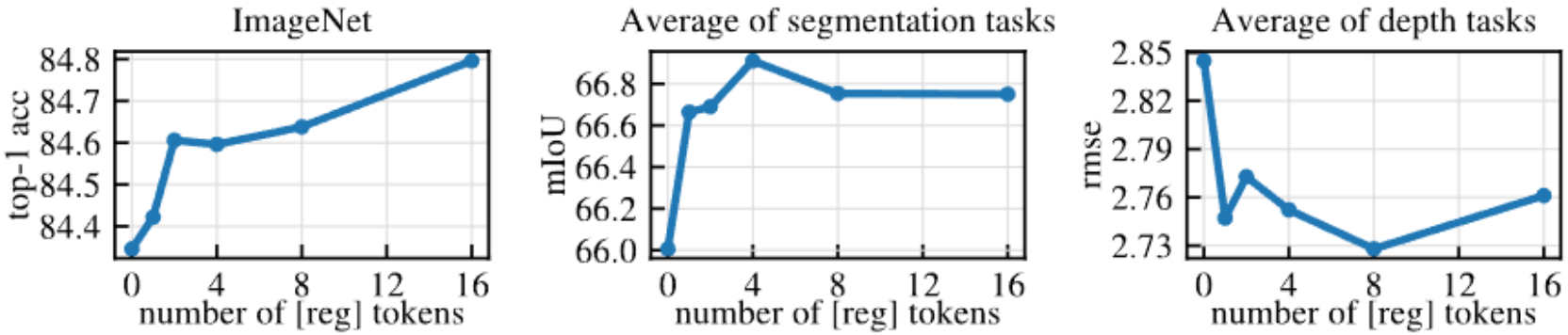
➢定量实验

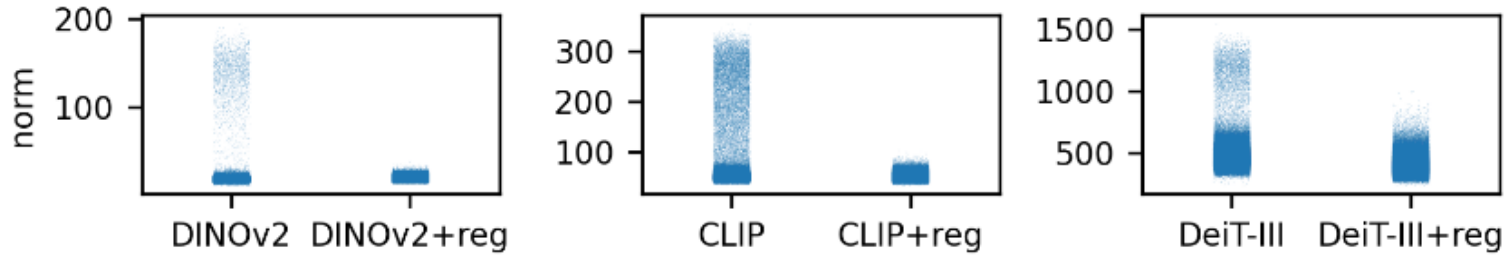
	ImageNet Top-1	ADE20k mIoU	NYUd rmse ↓
DeiT-III	84.7	38.9	0.511
DeiT-III+reg	84.7	39.1	0.512
OpenCLIP	78.2	26.6	0.702
OpenCLIP+reg	78.1	26.7	0.661
DINOv2	84.3	46.6	0.378
DINOv2+reg	84.8	47.9	0.366

(a) Linear evaluation with frozen features.

	ImageNet Top-1
OpenCLIP	59.9
OpenCLIP+reg	60.1

(b) Zero-shot classification.





研究问题

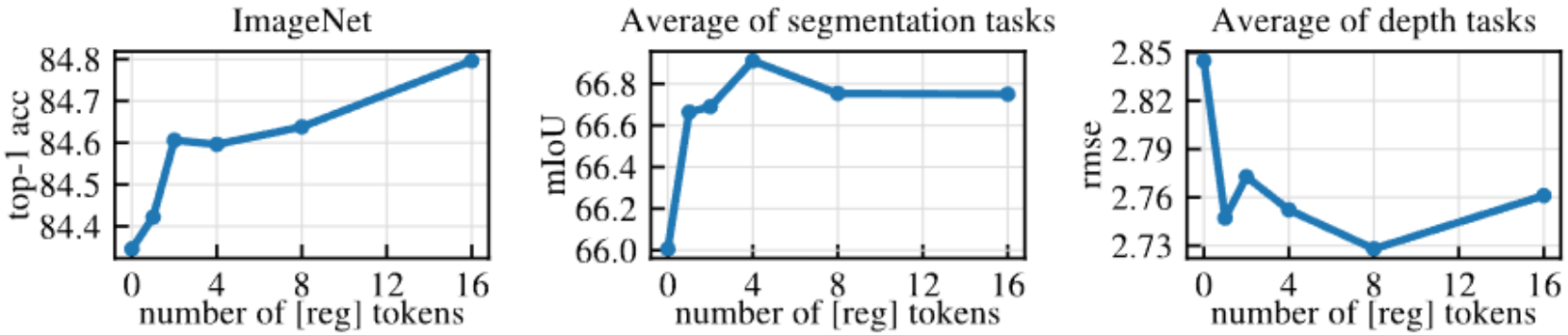
➢定量实验

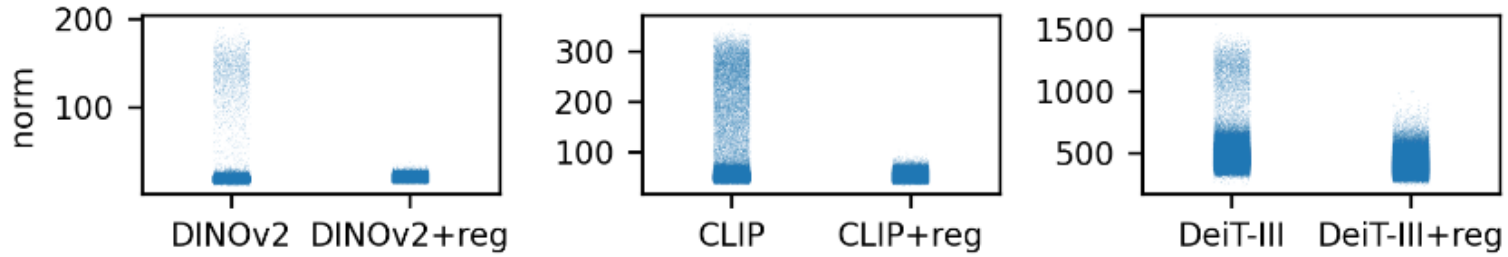
	ImageNet Top-1	ADE20k mIoU	NYUd rmse ↓
DeiT-III	84.7	38.9	0.511
DeiT-III+reg	84.7	39.1	0.512
OpenCLIP	78.2	26.6	0.702
OpenCLIP+reg	78.1	26.7	0.661
DINOv2	84.3	46.6	0.378
DINOv2+reg	84.8	47.9	0.366

(a) Linear evaluation with frozen features.

	ImageNet Top-1
OpenCLIP	59.9
OpenCLIP+reg	60.1

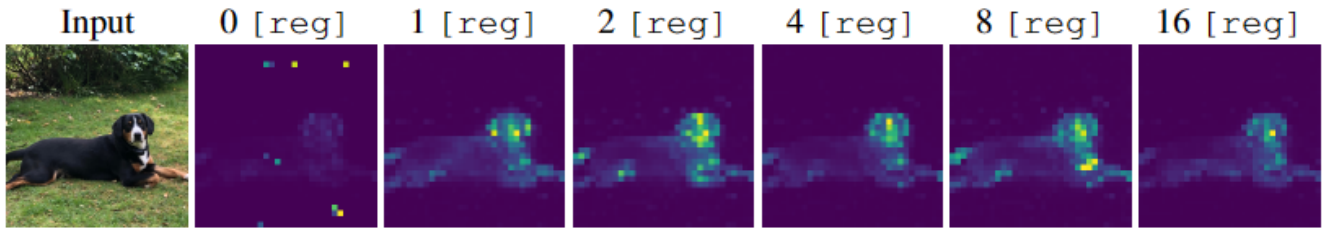
(b) Zero-shot classification.

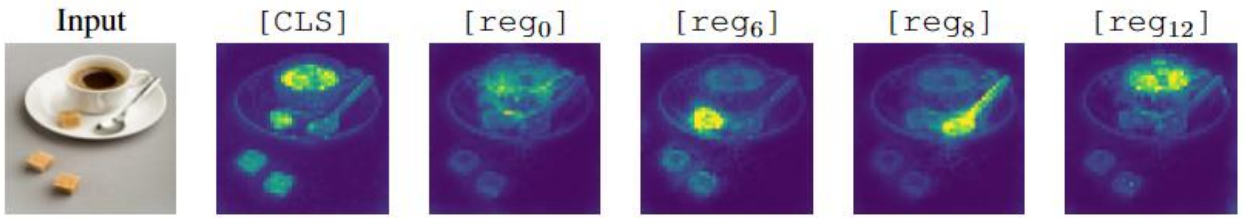


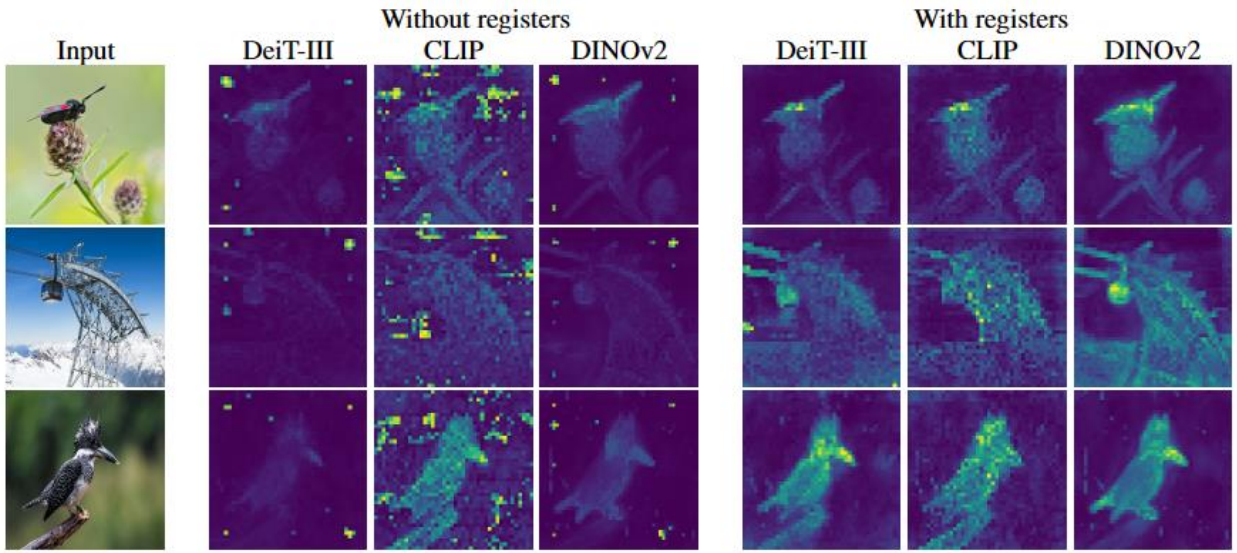


研究问题

➢定性实验


Input
0 [reg]
1 [reg]
2 [reg]
4 [reg]
8 [reg]
16 [reg]


Input
[CLS]
[reg0]
[reg6]
[reg8]
[reg12]


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

研究问题

➢总结

- ➢作者从异常Patch这一现象出发，寻找原因
- ➢运用多种可视化手段，从多种因素的角度，刻画出异常现象的变化趋势，并逐渐发现异常Patch的固有特性
- ➢设计针对性实验，探究异常Patch在局部信息、全局信息等方面的深层次性质
- ➢总结现象，给出思考和假设，以及尝试性的解决方案
- ➢完成定量和定性的分析实验，均确认了解决方案的有效性，以及思考和假设的合理性

ICLR 2024 oral

汇报人：程泽轩

谢谢大家！

Timothée Darcet, Maxime Oquab, Julien Mairal, Piotr Bojanowski

FAIR, Meta
INRIA