DINO
Team

Meta

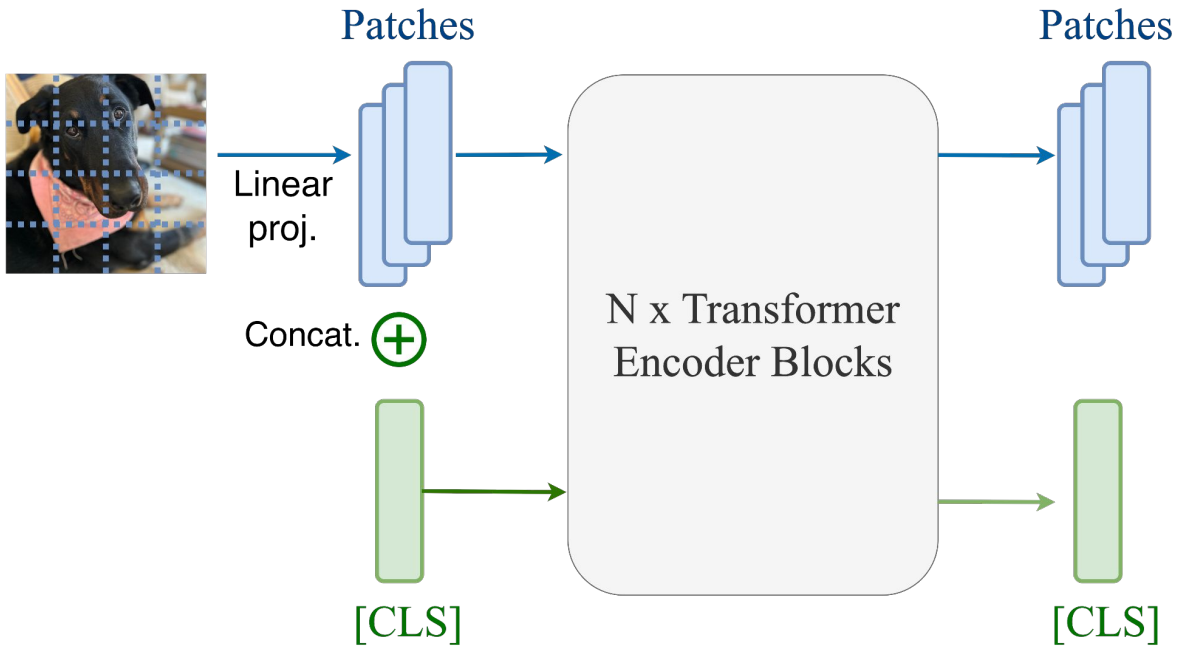
Revisiting [CLS] and Patch Token Interaction in Vision Transformers

ICLR 26

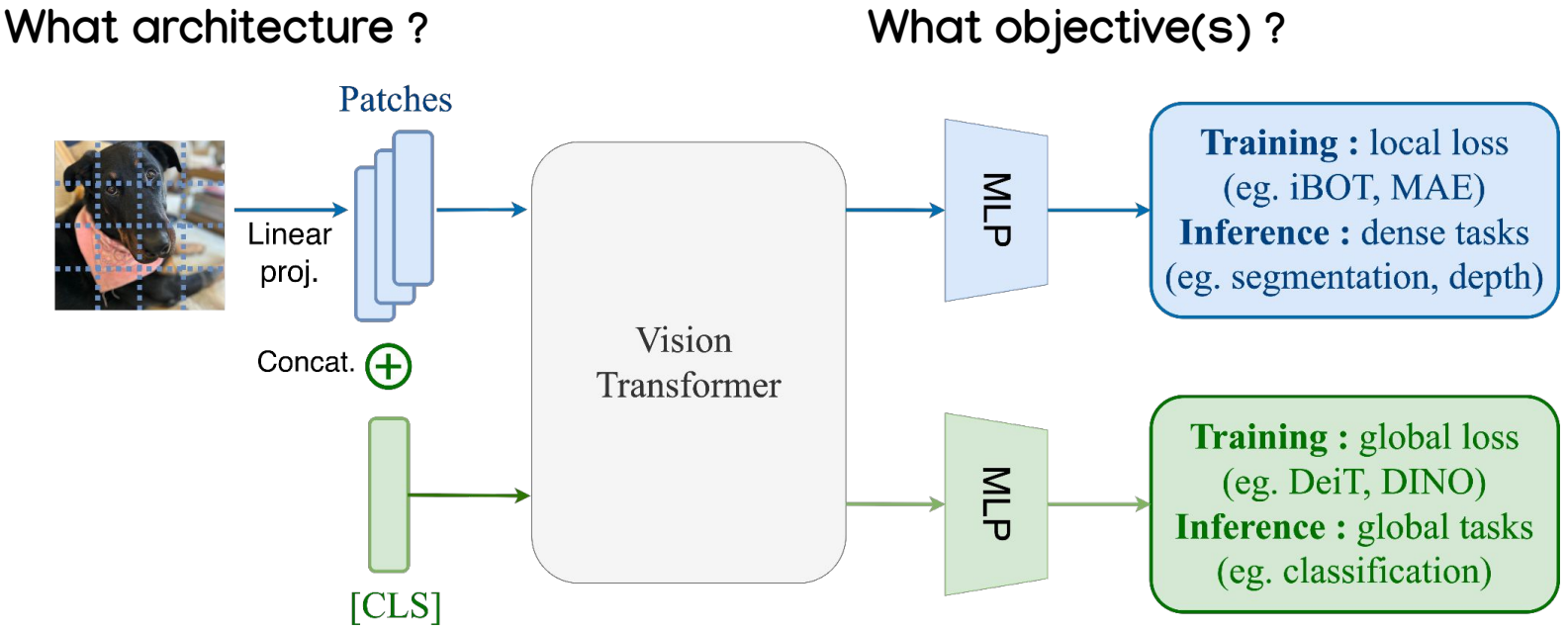
Alexis Marouani | Oriane Siméoni | Hervé Jégou | Piotr Bojanowski | Huy V. Vo

Introduction : local and global features in SSL

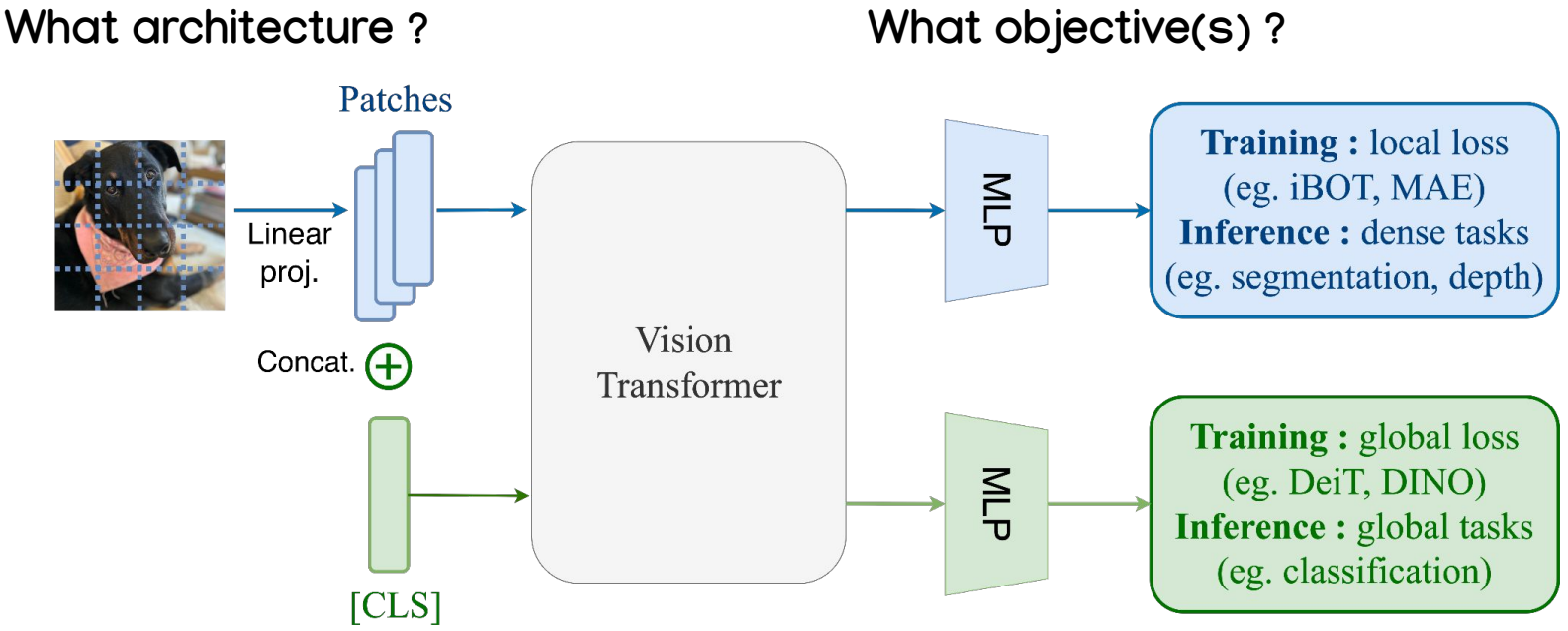
What architecture ?



Introduction : local and global features in SSL



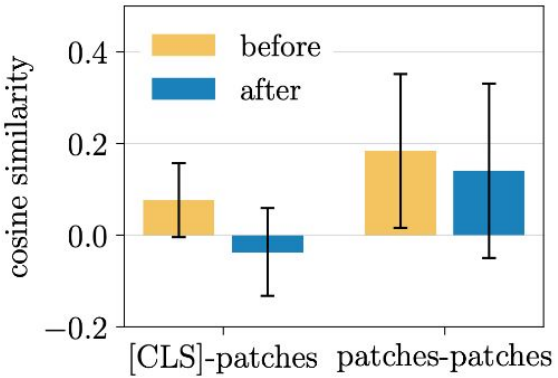
Introduction : local and global features in SSL



Same computational pipeline => Is it a problem ? Is there room for improvement ?

Frictions between CLS and patches

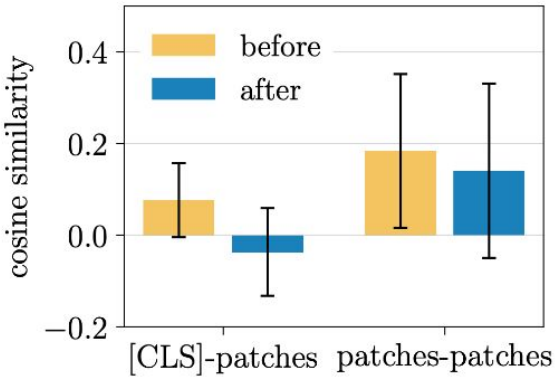
before / after LayerNorm



ViTs **implicitly distinguish** [CLS] and patches with LayerNorm (here pre-attention)

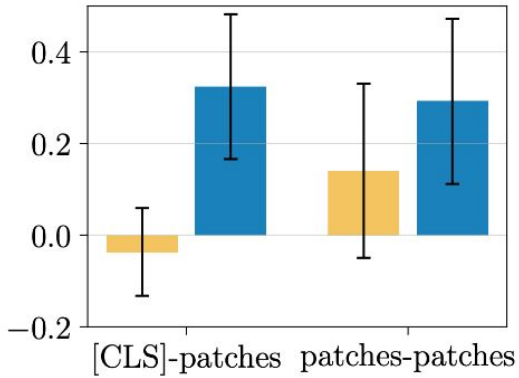
Frictions between CLS and patches

before / after LayerNorm



ViTs **implicitly distinguish** [CLS] and patches with LayerNorm (here pre-attention)

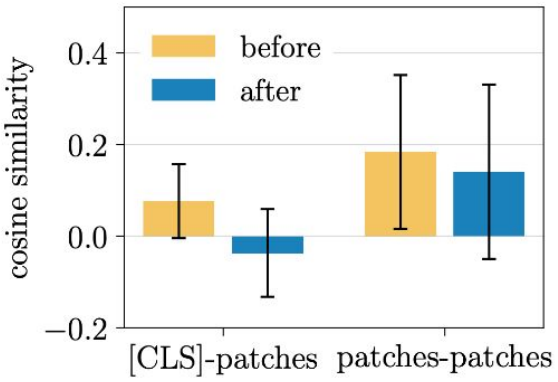
before / after Attention



Attention aligns [CLS] and patches representations

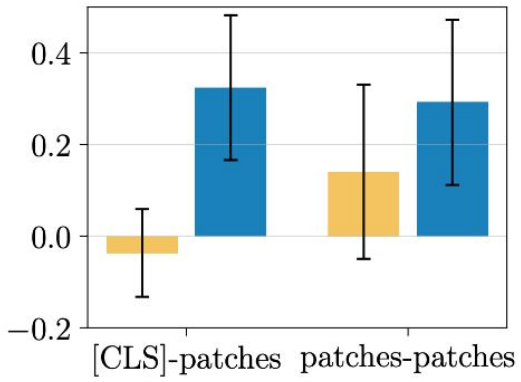
Frictions between CLS and patches

before / after LayerNorm

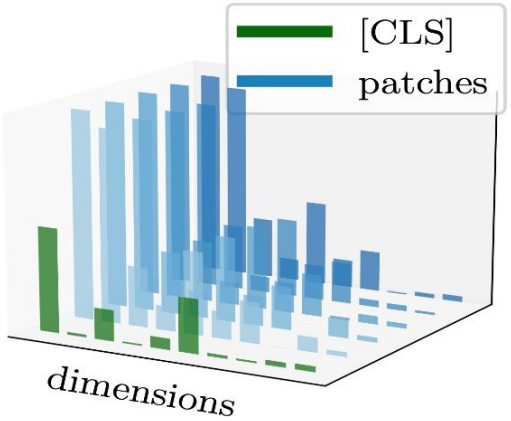


ViTs **implicitly distinguish** [CLS] and patches with LayerNorm (here pre-attention)

before / after Attention



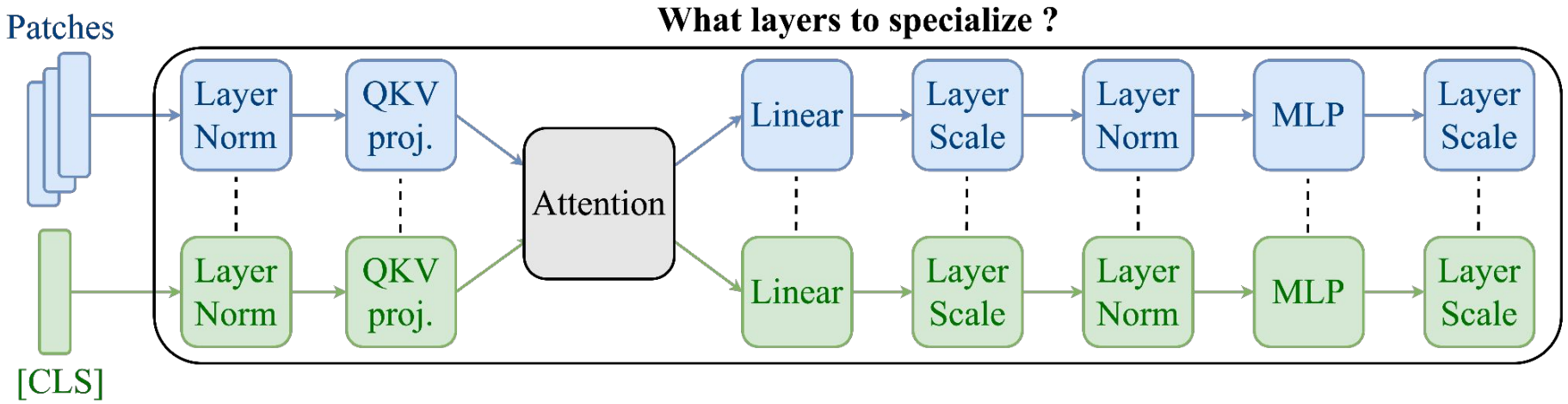
Attention aligns [CLS] and patches representations



[CLS] and patches occupy different dimensions → **LayerNorm can apply different operations** on both

A new architecture : specializing the layers

Our proposal: Specialize the weights while keeping interaction in the attention



Framework for main results

Training framework:

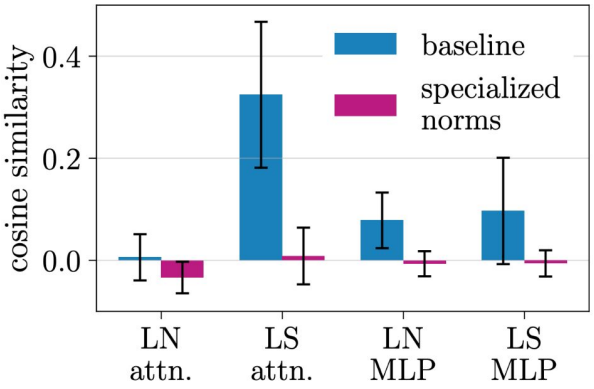
- DINOv2 – ViT L – 625k its on ImNet22k

Evaluation framework:

- Global task : Classification on ImNet1k
- Segmentation : ADE20k, CityScapes, VOC
- Depth : KITTI, NYU V2, SUN RGB-D

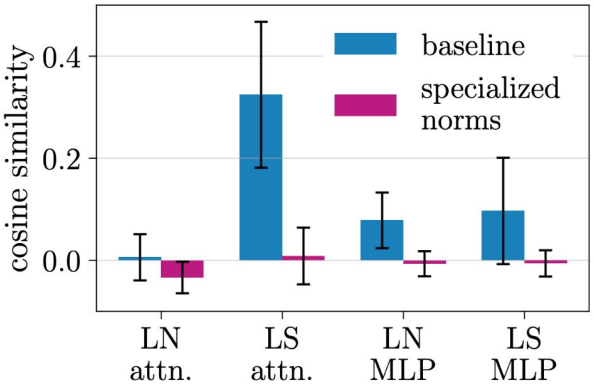
The effect of normalization specialization

Cosine sim.
[CLS]-patches

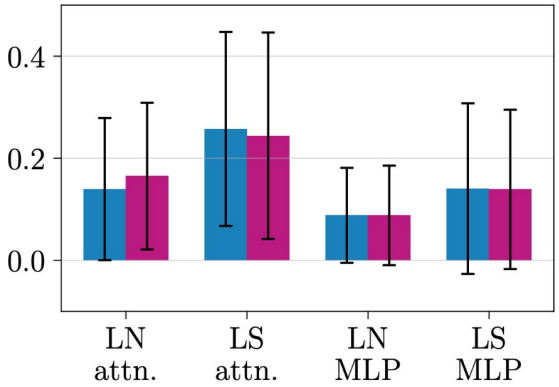


The effect of normalization specialization

Cosine sim. [CLS]-patches

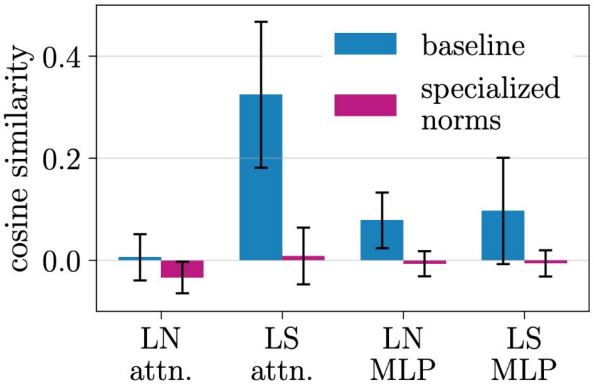


Cosine sim. patches-patches

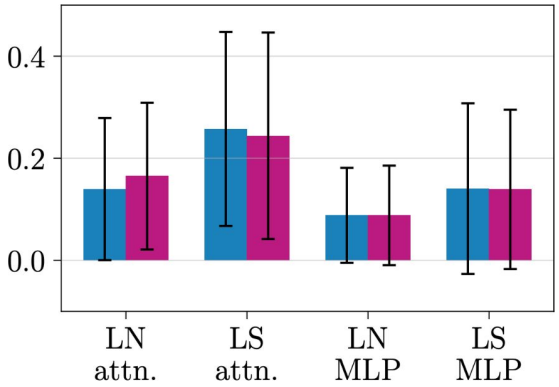


The effect of normalization specialization

Cosine sim. [CLS]-patches



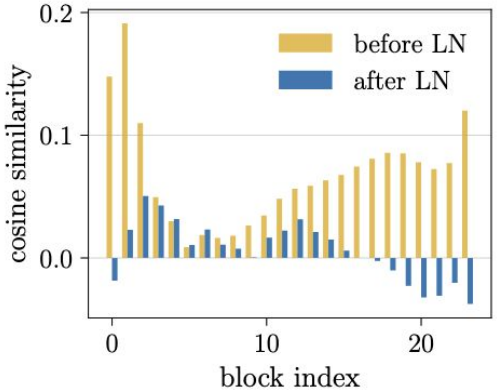
Cosine sim. patches-patches



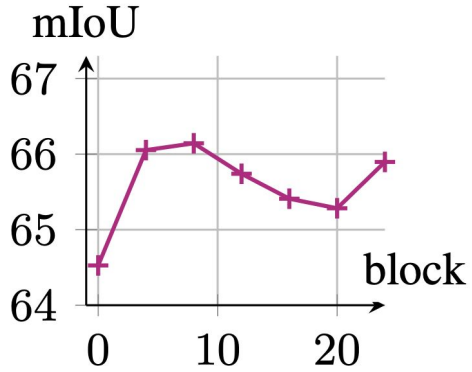
Dense performance boosted

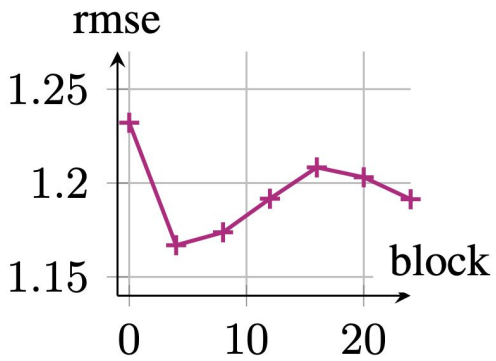
Spec.	Linear Acc.	Avg Seg.	Avg Depth ↓
–	**85.4**	64.5	1.232
norms	85.1	**65.6**	**1.178**

Where to specialize ?



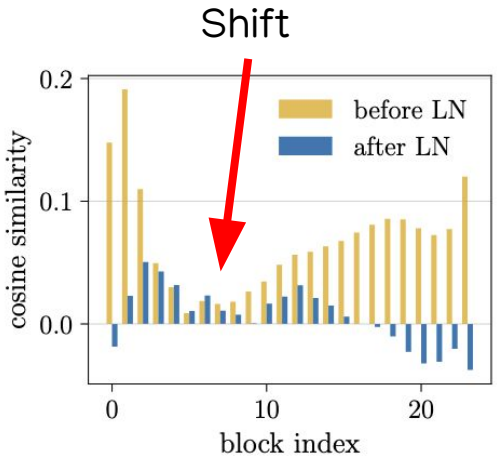
Mean cosine similarity between [CLS] and all patches (no specialization)



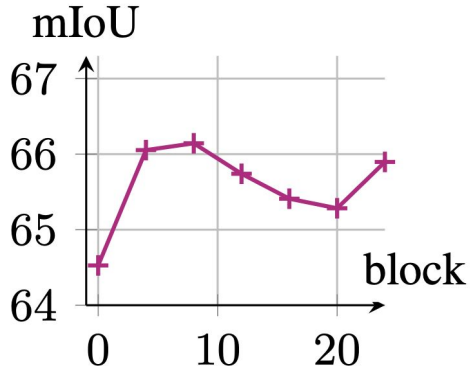


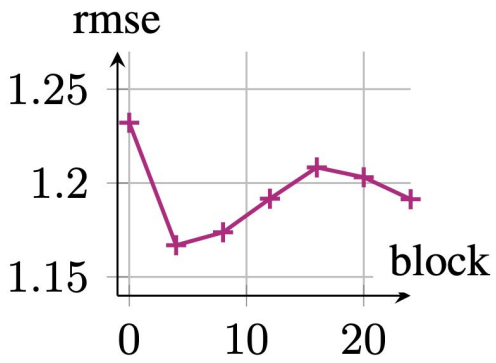
Average segmentation (left) scores and average depth rmse (↓) vs number of specialized blocks at the beginning of the model. Normalization layers are specialized in all blocks.

Where to specialize ?



Mean cosine similarity between [CLS] and all patches (no specialization)

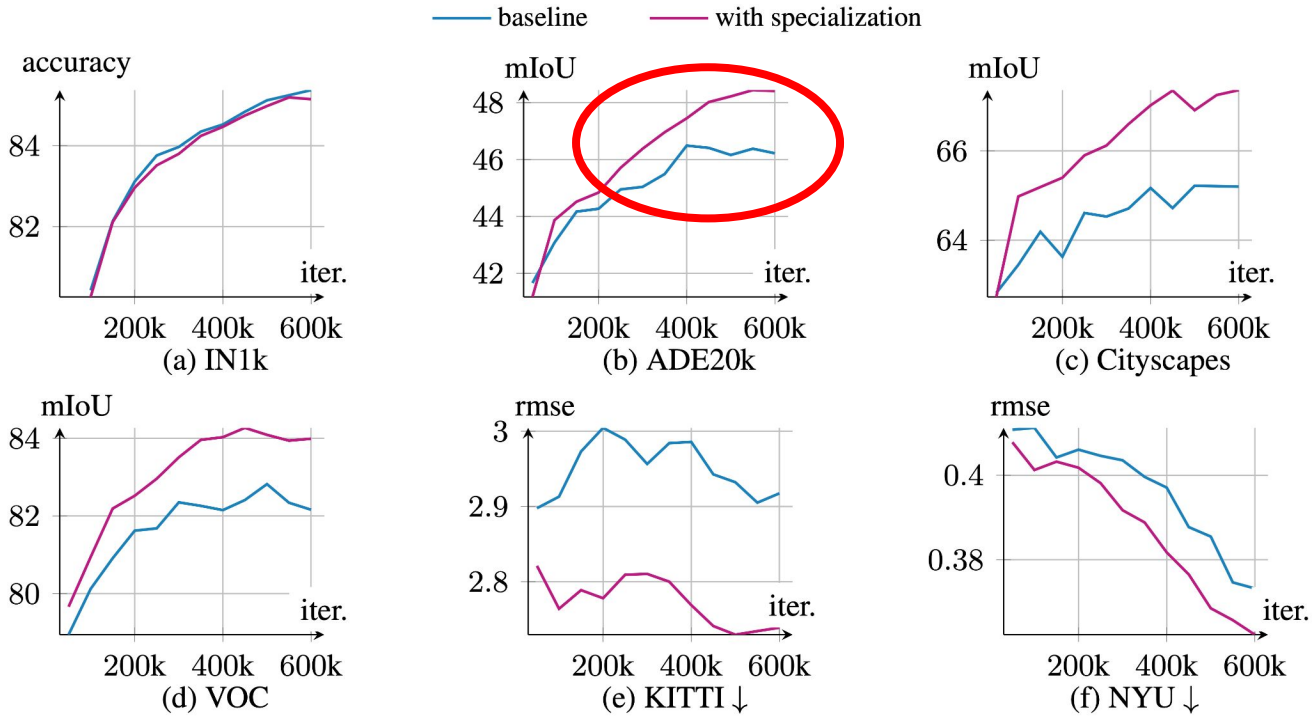




Average segmentation (left) scores and average depth rmse (↓) vs number of specialized blocks at the beginning of the model. Normalization layers are specialized in all blocks.

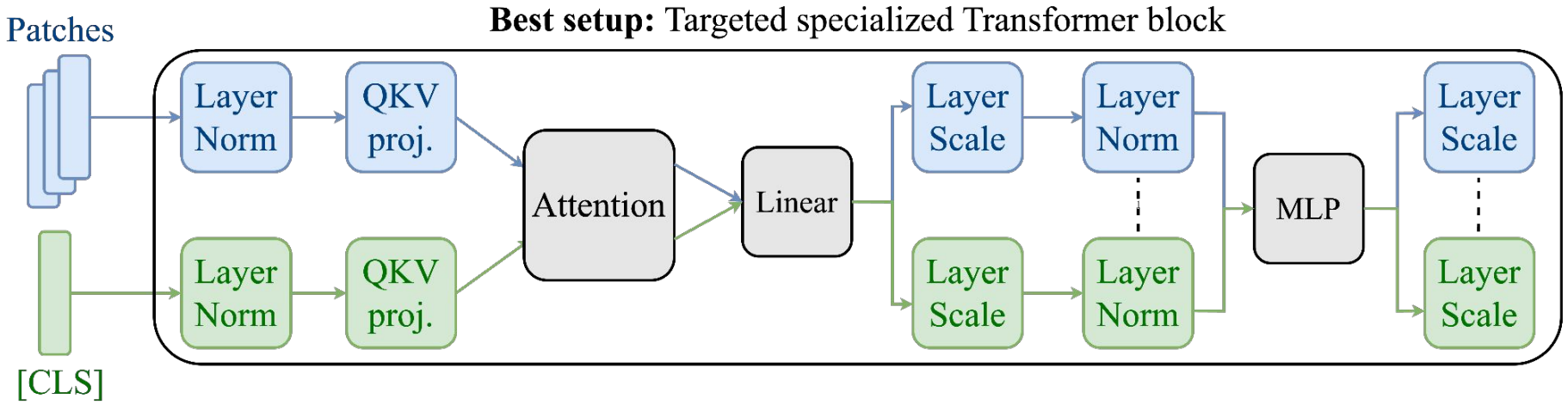
What to specialize ?

Specializing QKV projection during ⅓ gives best results => 8% increase



What to specialize ?

Specializing QKV projection during first ⅓ layers gives best results => 8% increase



Remark : Sora LoRA mechanisms are possible to get gains at low parameter cost

Generalizability

- Different model sizes

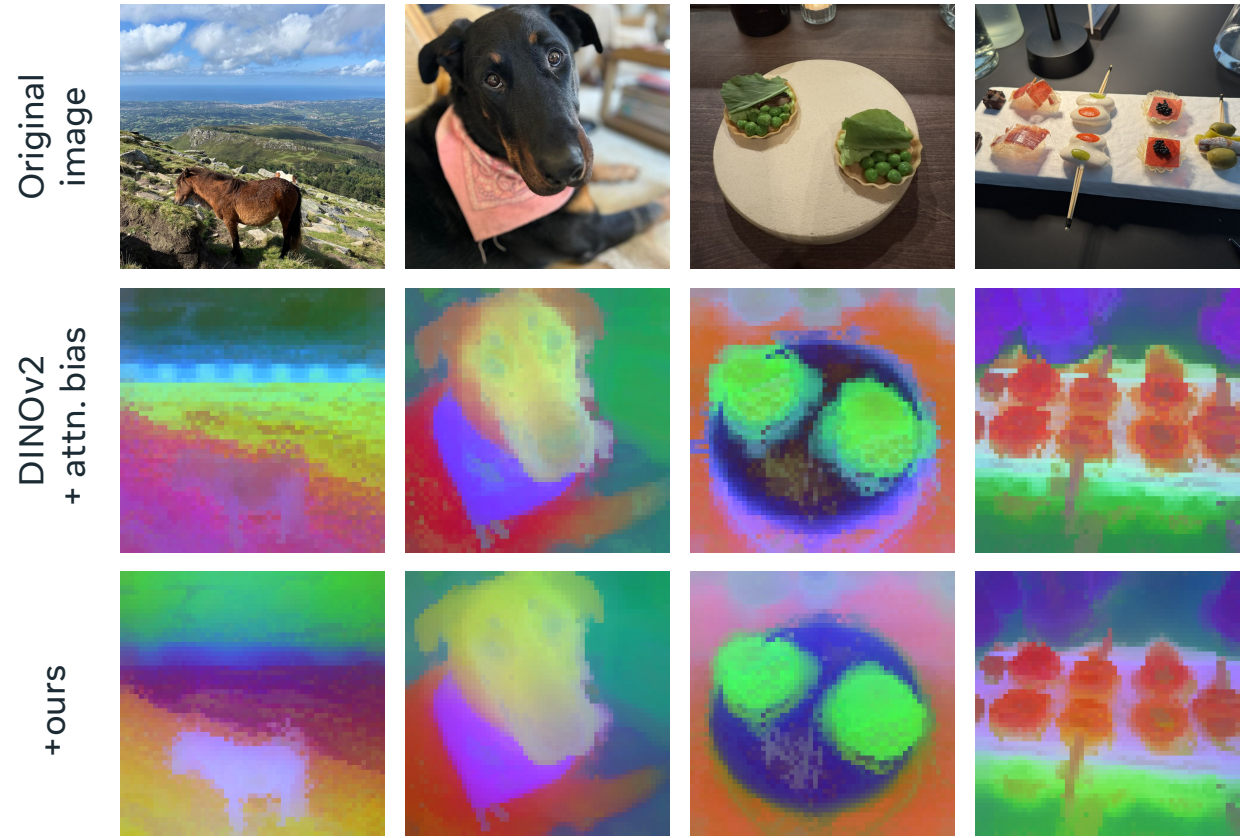
Method	Size	Classif.	Segmentation			Depth ↓		
		ImNet	ADE	City	VOC	KITTI	NYU	SUN
DINOv2 - With high-norm handling strategies								
∅	L	85.3	45.7	64.2	82.1	2.868	0.389	0.410
+ours		85.3+0.0%	47.3+3.5%	66.6+3.7%	83.7+1.9%	2.787-2.8%	0.369-5.1%	0.390-4.9%
4 registers	L	85.3	45.6	64.9	82.2	2.893	0.372	0.411
+ours		85.3+0.0%	47.5+4.2%	65.9+1.5%	83.6+1.7%	2.906+0.4%	0.367-1.3%	0.395-3.9%
Attn. bias	L	85.4	46.2	65.2	82.2	2.917	0.373	0.406
+ours		85.2-0.2%	48.4+4.8%	67.4+3.4%	84.0+2.2%	2.739-6.1%	0.362-2.9%	0.393-3.2%
DINOv2 - Other sizes								
Attn. bias	B	80.4	38.3	58.4	76.6	3.250	0.462	0.464
+ours		80.6+0.2%	38.5+0.5%	60.3+3.3%	76.5-0.1%	3.236-0.4%	0.448-3.0%	0.470+1.3%
Attn. bias	H	86.2	48.1	67.0	83.1	2.717	0.359	0.387
+ours		86.1-0.1%	49.2+2.3%	67.1+0.1%	83.5+0.5%	2.752+1.3%	0.344-4.2%	0.386-0.3%
DeiT-III								
Attn. bias	B	81.8	25.4	61.7	48.9	5.040	0.747	0.823
+ours		81.7-0.1%	26.3+3.5%	62.7+1.6%	50.7+3.7%	4.900-2.8%	0.732-2.0%	0.809-1.7%

Generalizability

- Different model sizes
- Different training objective

Method	Size	Classif.	Segmentation			Depth ↓		
		ImNet	ADE	City	VOC	KITTI	NYU	SUN
DINOv2 - With high-norm handling strategies								
∅	L	85.3	45.7	64.2	82.1	2.868	0.389	0.410
+ours		85.3+0.0%	47.3+3.5%	66.6+3.7%	83.7+1.9%	2.787-2.8%	0.369-5.1%	0.390-4.9%
4 registers	L	85.3	45.6	64.9	82.2	2.893	0.372	0.411
+ours		85.3+0.0%	47.5+4.2%	65.9+1.5%	83.6+1.7%	2.906+0.4%	0.367-1.3%	0.395-3.9%
Attn. bias	L	85.4	46.2	65.2	82.2	2.917	0.373	0.406
+ours		85.2-0.2%	48.4+4.8%	67.4+3.4%	84.0+2.2%	2.739-6.1%	0.362-2.9%	0.393-3.2%
DINOv2 - Other sizes								
Attn. bias	B	80.4	38.3	58.4	76.6	3.250	0.462	0.464
+ours		80.6+0.2%	38.5+0.5%	60.3+3.3%	76.5-0.1%	3.236-0.4%	0.448-3.0%	0.470+1.3%
Attn. bias	H	86.2	48.1	67.0	83.1	2.717	0.359	0.387
+ours		86.1-0.1%	49.2+2.3%	67.1+0.1%	83.5+0.5%	2.752+1.3%	0.344-4.2%	0.386-0.3%
DeiT-III								
Attn. bias	B	81.8	25.4	61.7	48.9	5.040	0.747	0.823
+ours		81.7-0.1%	26.3+3.5%	62.7+1.6%	50.7+3.7%	4.900-2.8%	0.732-2.0%	0.809-1.7%

Qualitative results



Conclusion

- ViTs has 2 types of tokens (CLS and patches) with same computational pipeline
- But, already in vanilla architecture there are mechanisms to disentangle their features (via the normalizations)
- By specializing specific weights → we obtain better separation, better local features & better perfs on dense taks

Thank you !