
ConViT: Improving Vision Transformers with Soft Convolutional Inductive Biases

Stéphane d'Ascoli^{1 2} Hugo Touvron² Matthew L. Leavitt² Ari S. Morcos² Giulio Biroli^{1 2} Levent Sagun²

2021. 11. 24

HyunJung Kim

¹Department of Physics, Ecole Normale Supérieure, Paris, France ²Facebook AI Research, Paris, France. Correspondence to: Stéphane d'Ascoli <stephane.dascoli@ens.fr>.

Proceedings of the 38th International Conference on Machine Learning, PMLR 139, 2021. Copyright 2021 by the author(s).

Contents

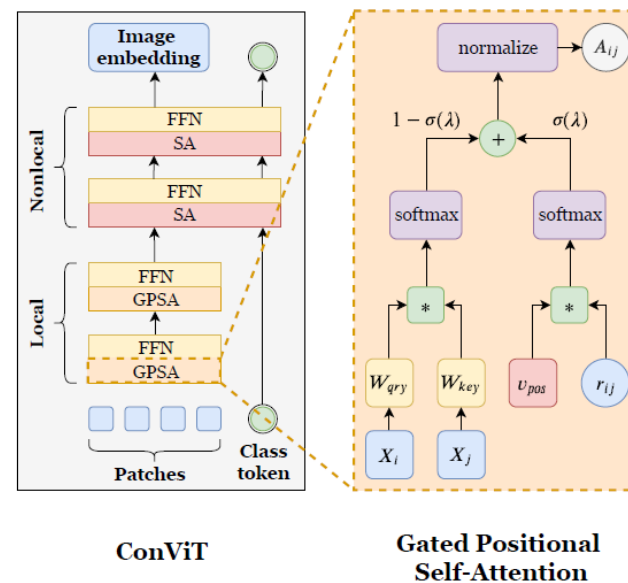
1. Background

2. Approach

3. Performance

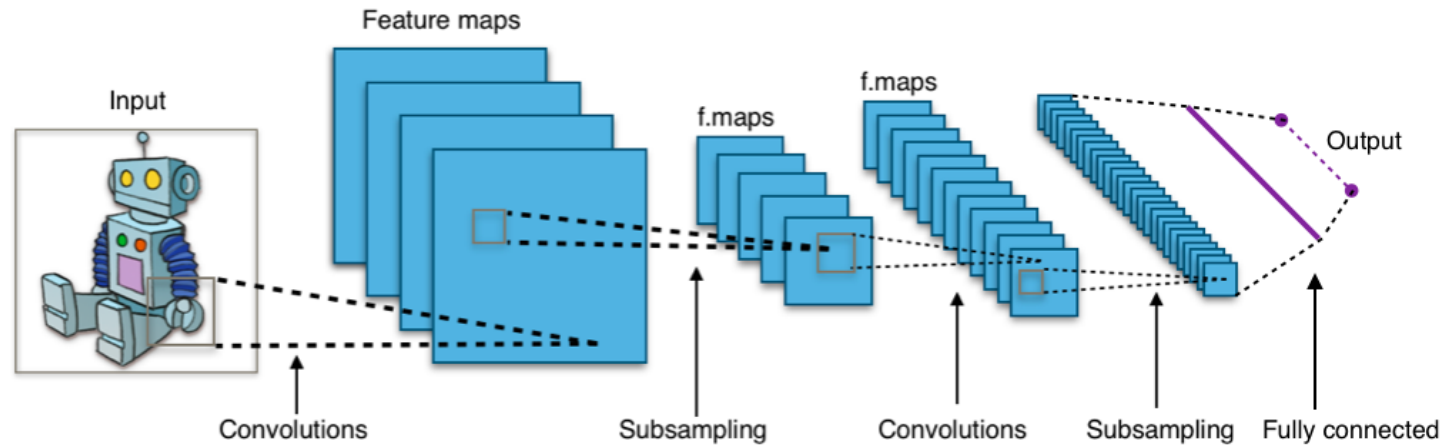
4. Ablation study

ConViT: Improving Vision Transformers with Soft Convolutional Inductive Biases



- **ConViT = CNN (Convolutional Neural Networks) + Transformer**

- **CNN (Convolutional Neural Networks)**



- **Transformer**

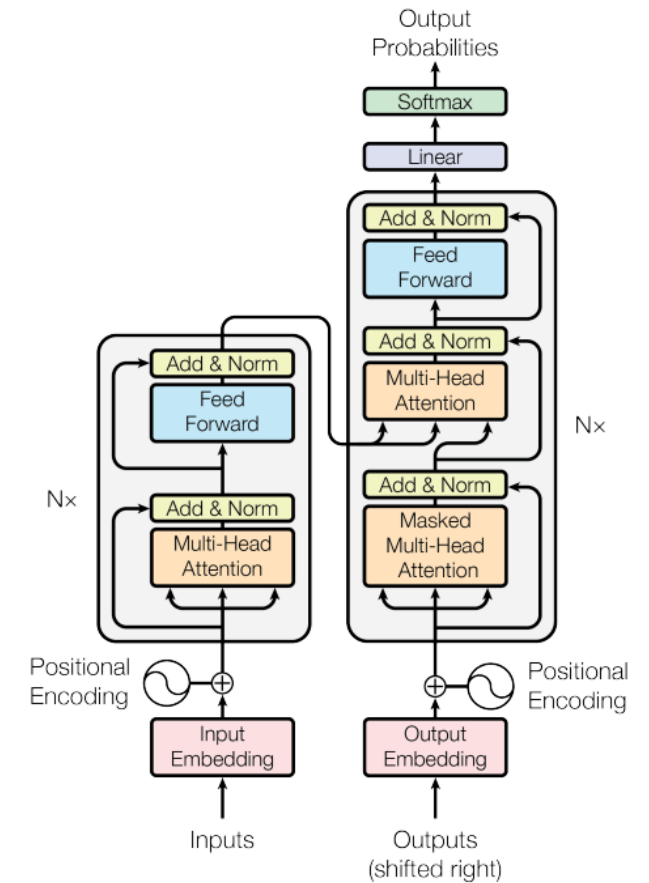
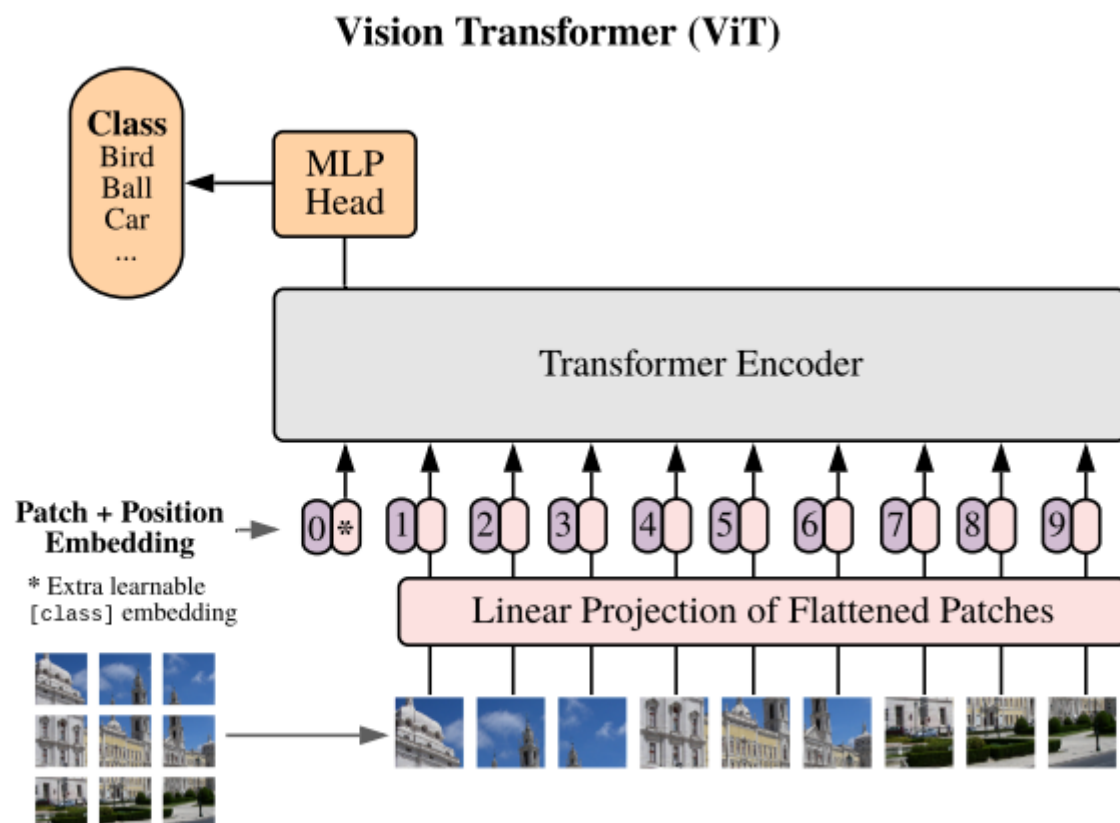
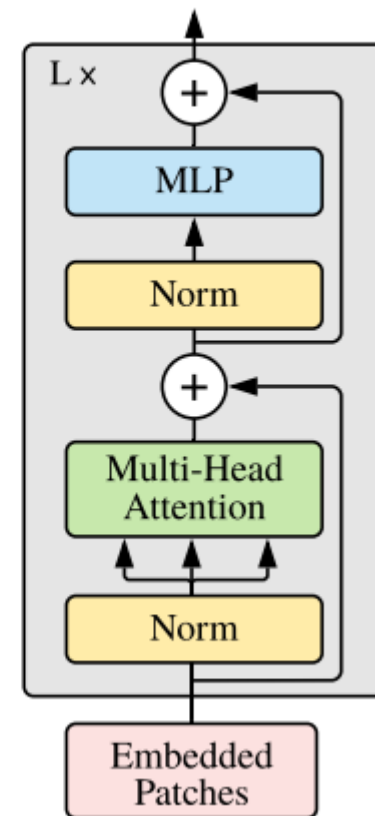


Figure 1: The Transformer - model architecture.

- ViT: Vision Transformer



Transformer Encoder



Original Transformer

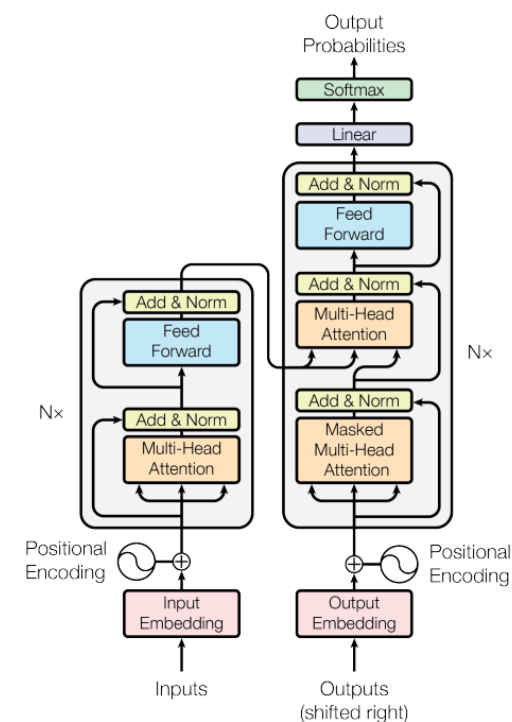
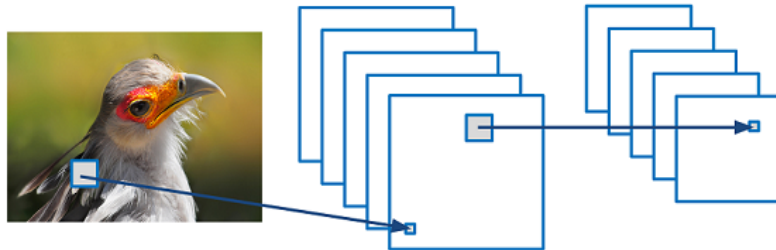


Figure 1: The Transformer - model architecture.

- Inductive Bias

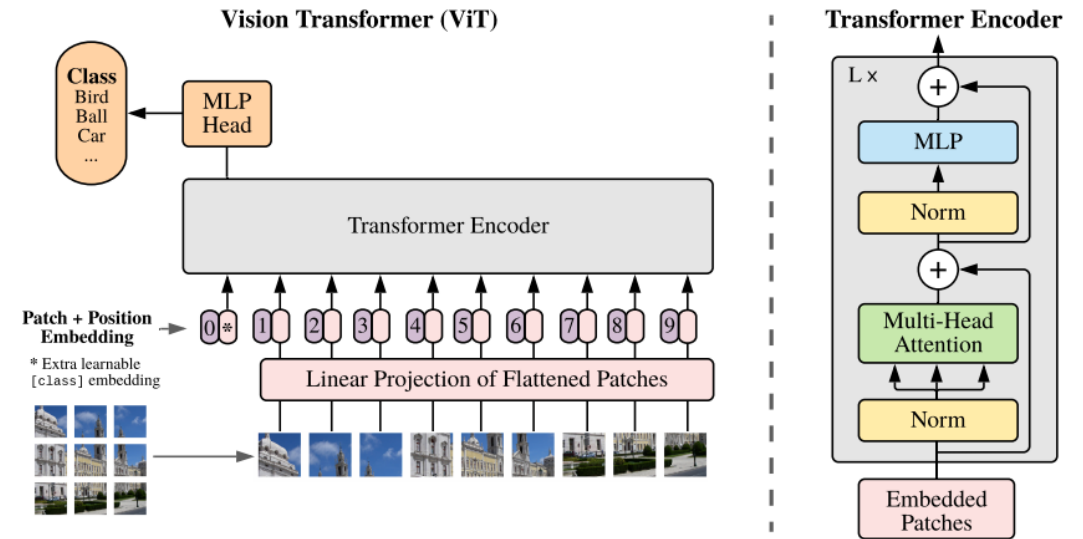
- CNN

- ✓ Locally connected (Local)
 - ✓ Same weight for all inputs

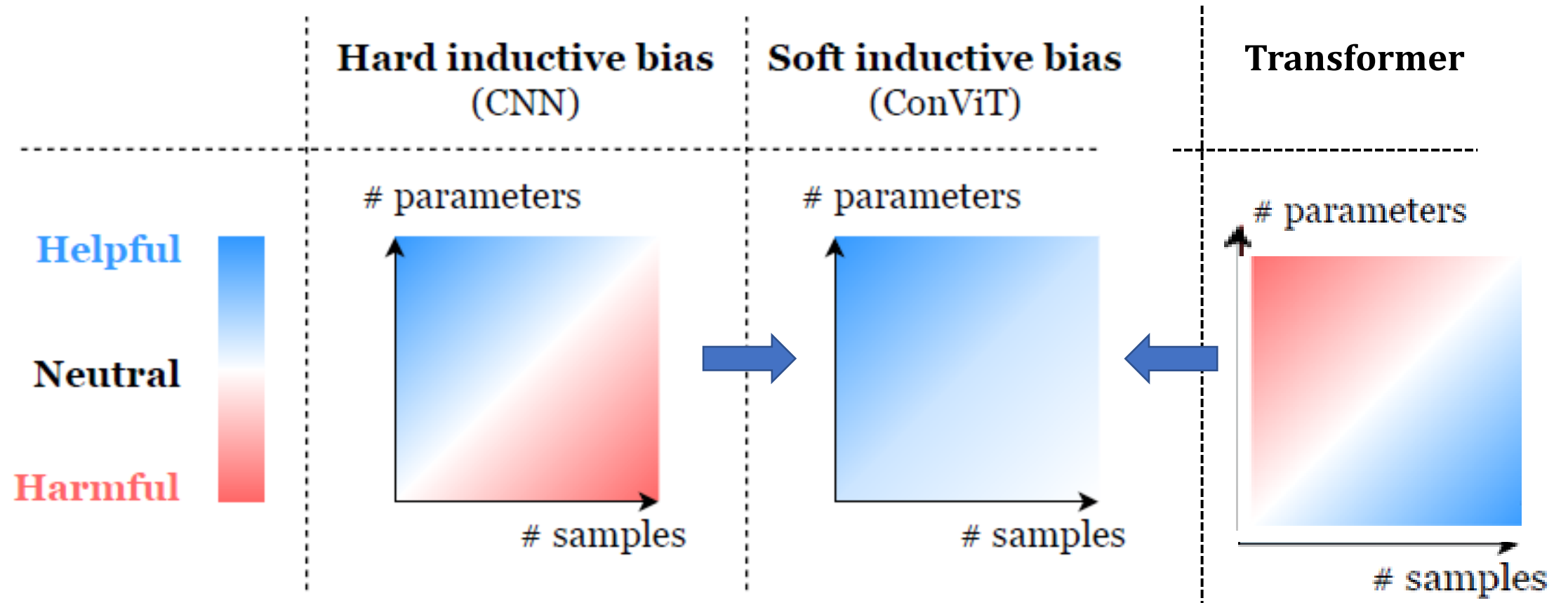


- Transformer

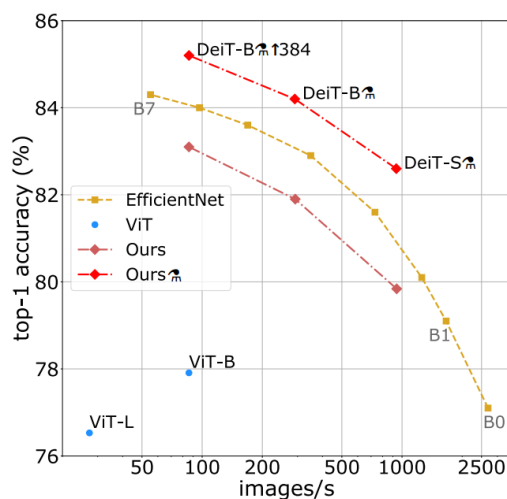
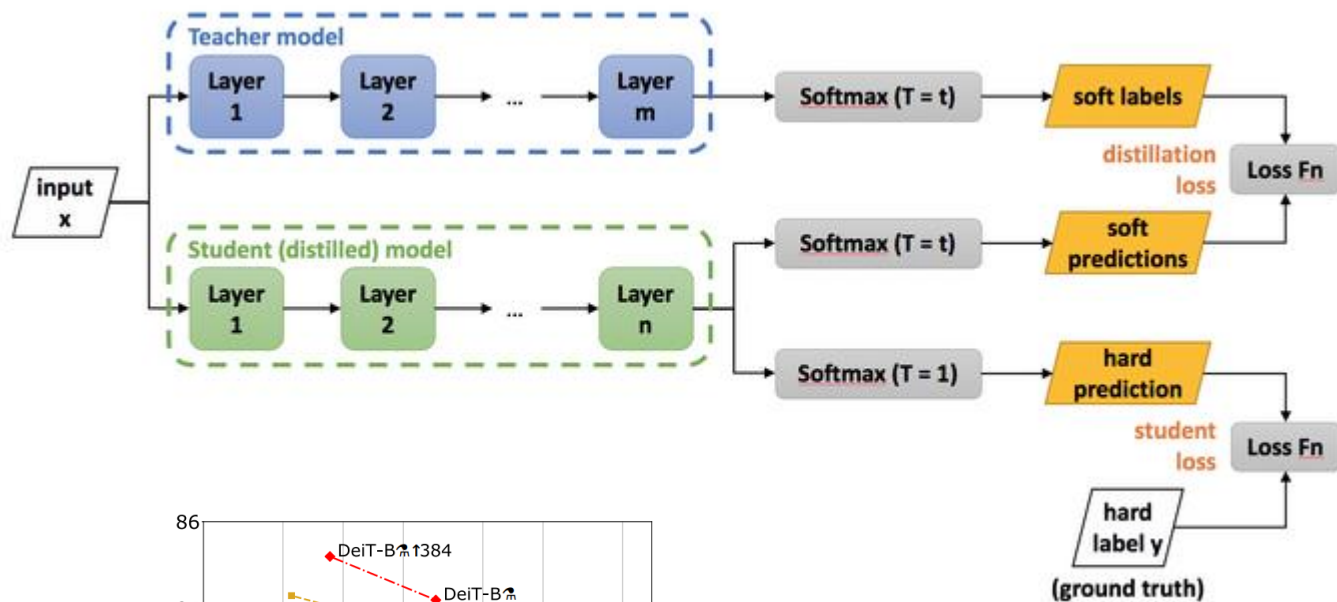
- ✓ Fully connected (Global)
 - ✓ Different weight for all inputs



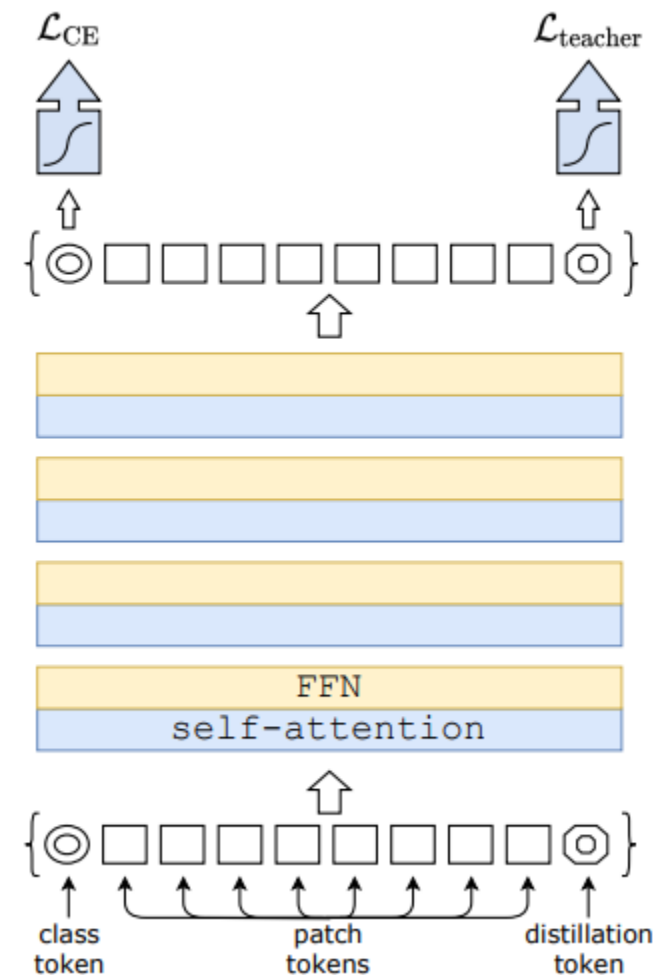
- ConViT: Soft Inductive Biases



- Knowledge Distillation



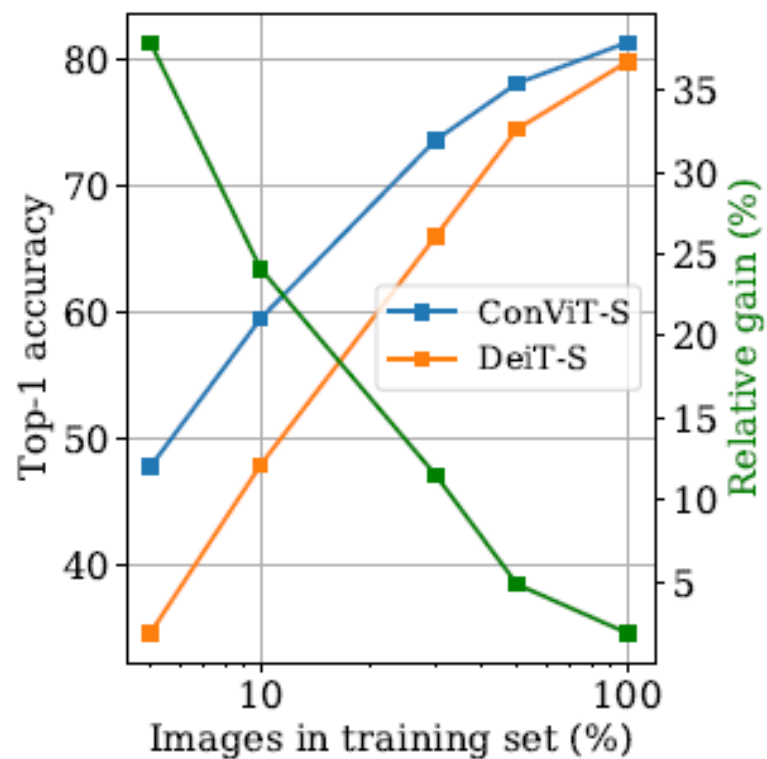
- Distillation



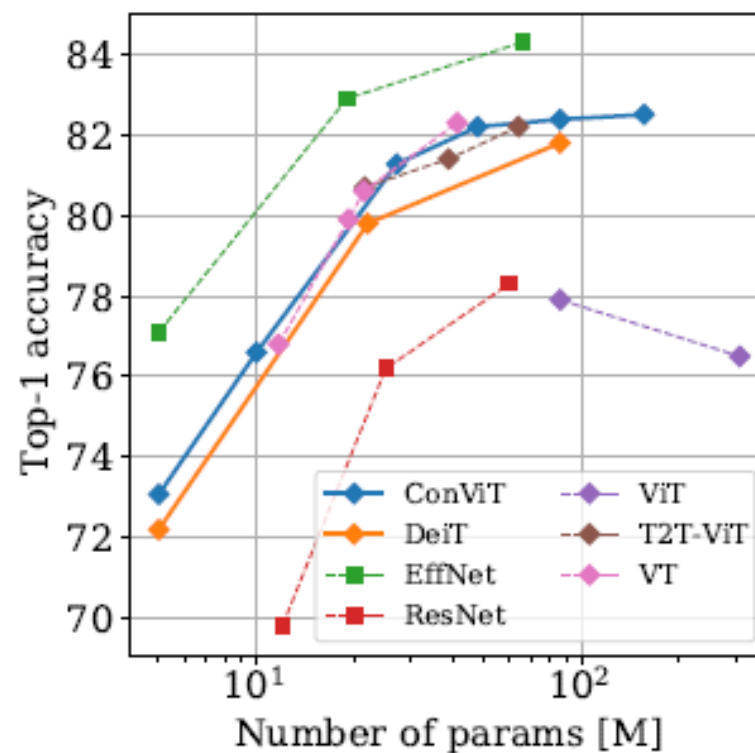
- **Contributions**

1. **A new form of SA layer, named gated positional self-attention (GPSA),
Can initialize as a convolutional layer.**
2. **The resulting Convolutional Vision Transformer (ConViT) outperforms the DeiT while boasting a much improved sample-efficiency (Fig. 2).**
3. **Ablation study: Benefits from the convolution initialization.**

- Sample & Parameter Efficiency



(a) Sample efficiency




(b) Parameter efficiency

- **Transformer Notion (Multi-head self-attention)** >> associated memory with (key, query) vector pairs

$$(1) \quad \mathbf{A} = \text{softmax} \left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{D_h}} \right) \in \mathbb{R}^{L_1 \times L_2},$$

$$(2) \quad \text{MSA}(\mathbf{X}) := \text{concat}_{h \in [N_h]} [\text{SA}_h(\mathbf{X})] \mathbf{W}_{out} + \mathbf{b}_{out},$$

$$(3) \quad \boxed{\text{SA}_h(\mathbf{X}) := \mathbf{A}^h \mathbf{X} \mathbf{W}_{val}^h},$$




- Attention
- Multi-Head Self Attention
by concatenating each head's self-attention
- Each head's self-attention

- Cf. Absolute / Relative Attention

- Absolute Attention (ViT)

- Relative Attention (ConViT)

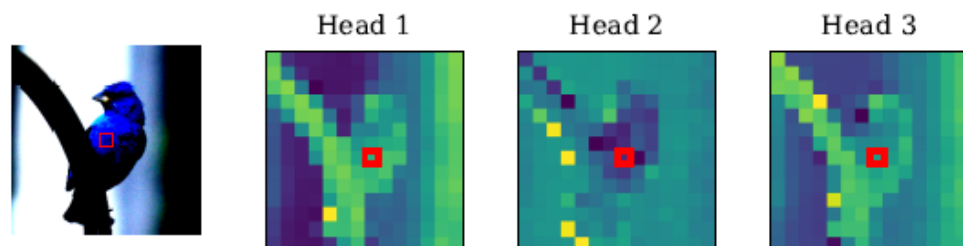
$$A_{ij}^h := \text{softmax} (Q_i^h K_j^{h\top} + v_{pos}^{h\top} r_{ij}) \quad (4)$$

 Attention Parameter
 Patch's Relative position

- Convolution Initialization

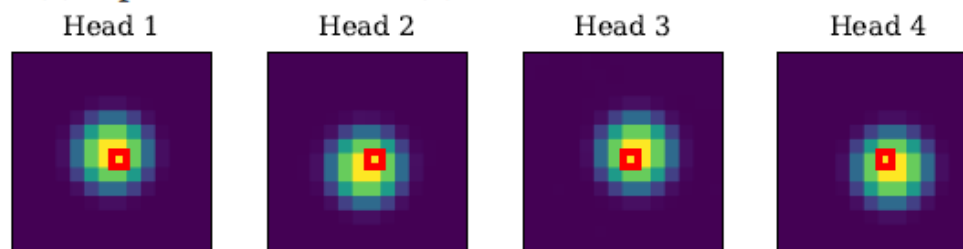
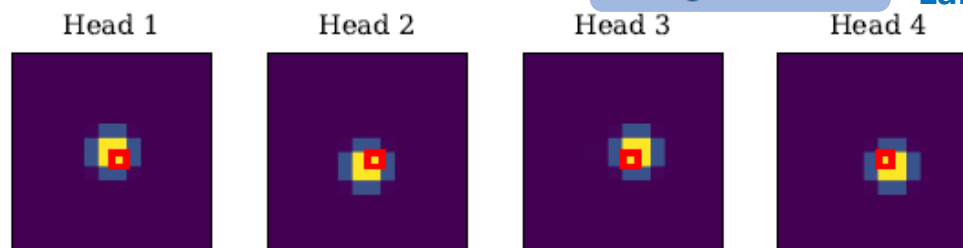
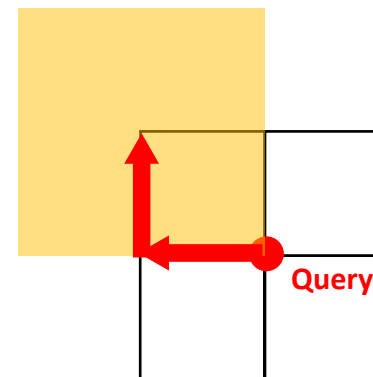
$$\begin{cases} v_{pos}^h := -\alpha^h (1, -2\Delta_1^h, -2\Delta_2^h, 0, \dots, 0) \\ r_{\delta} := (\|\delta\|^2, \delta_1, \delta_2, 0, \dots, 0) \\ W_{qry} = W_{key} := 0, \quad W_{val} := I \end{cases} \quad (5)$$

- Self-attention as a generalized convolution

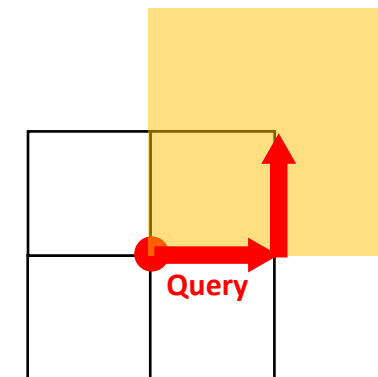


(a) Input

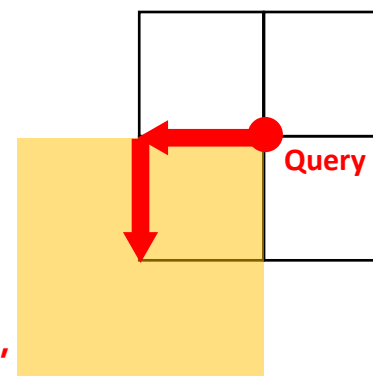
(b) Standard initialization

(c) Convolutional initialization, strength $\alpha = 0.5$ Less attended,
Larger window(d) Convolutional initialization, strength $\alpha = 2$ More attended,
Smaller window

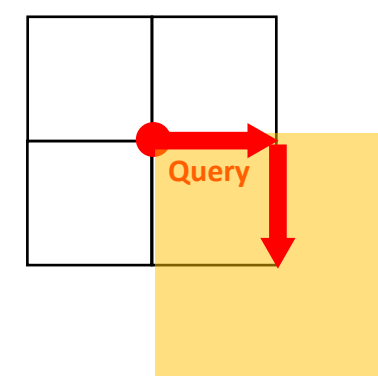
$$\Delta^1 = (-1, 1)$$



$$\Delta^3 = (1, 1)$$



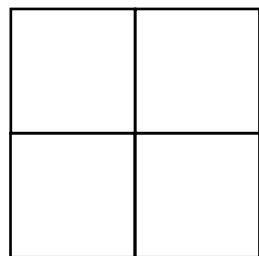
$$\Delta^2 = (-1, -1)$$



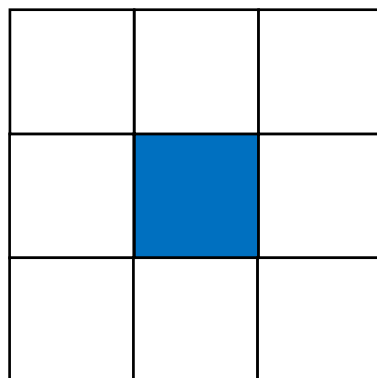
$$\Delta^4 = (1, -1)$$

- The center of attention : $\Delta^h \in \mathbb{R}^2$

- Convolutional Neural Network



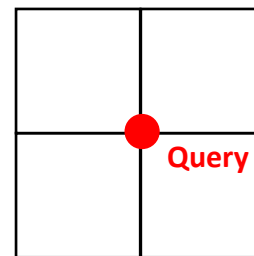
2 x 2



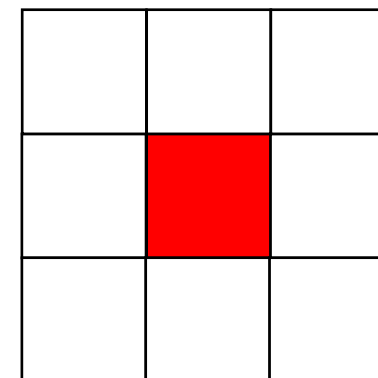
3 x 3

Impossible to centering

- Transformer



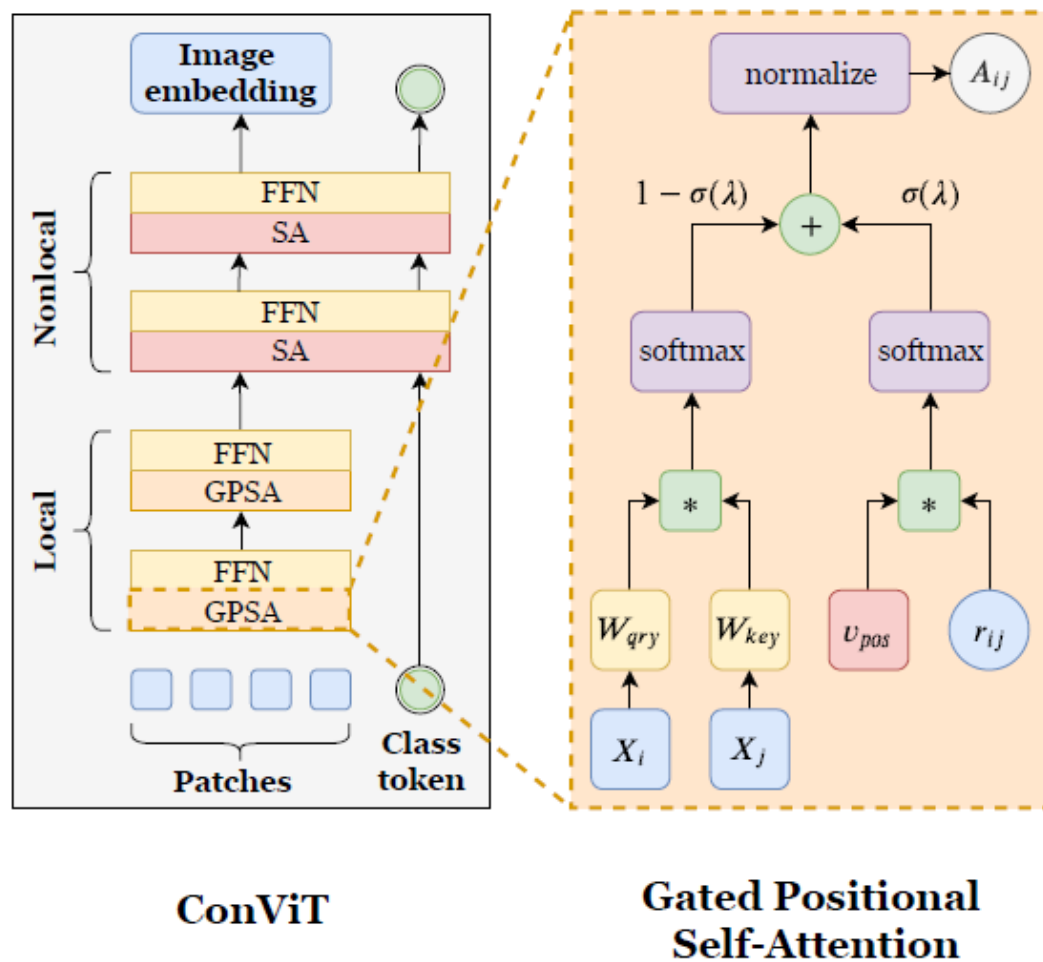
2 x 2



3 x 3

Possible to centering

- Architecture of the ConViT : Adaptive attention span + Positional Gating



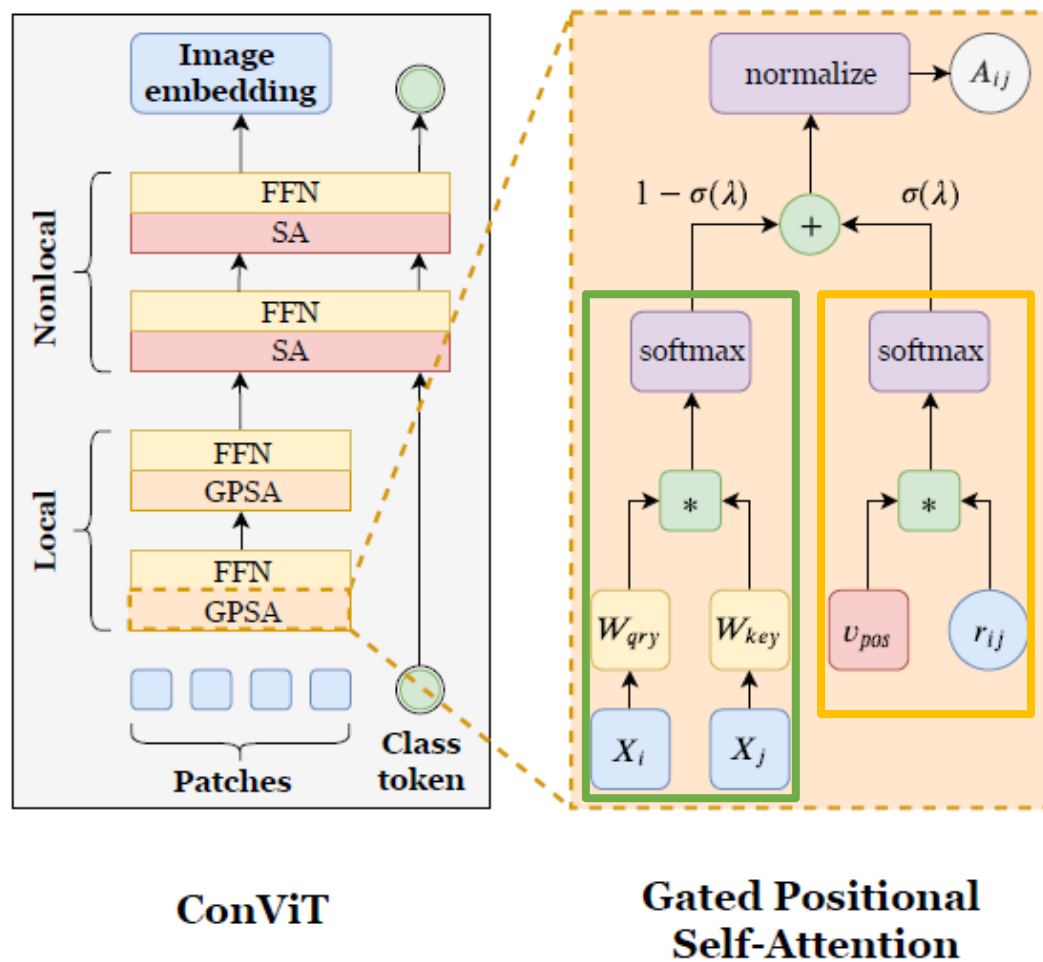
- Adaptive attention Span

$$A_{ij}^h := \text{softmax} (Q_i^h K_j^{h^T} + v_{pos}^{h^T} r_{ij}) \quad (4)$$

- ✓ the previous models: relative attention learn both $v_{pos}^{h^T}$ and r_{ij}
 - losing long-range information
 - ✓ To avoid the above problem,
 - Relative positional encodings r_{ij} : fixed
 - the embeddings $v_{pos}^{h^T}$: train
- (the embeddings $v_{pos}^{h^T}$ determine the center and span of the attention head)

$$\begin{cases} v_{pos}^h := -\alpha^h (1, -2\Delta_1^h, -2\Delta_2^h, 0, \dots, 0) \\ r_\delta := (\|\delta\|^2, \delta_1, \delta_2, 0, \dots, 0) \\ W_{qry} = W_{key} := 0, \quad W_{val} := I \end{cases} \quad (5)$$

- Architecture of the ConViT : Adaptive attention span + Positional Gating



- Positional Gating

$$\text{GPSA}_h(X) := \text{normalize} [A^h] X W_{val}^h \quad (6)$$

$$A_{ij}^h := (1 - \sigma(\lambda_h)) \text{softmax}(Q_i^h K_j^{h\top}) + \sigma(\lambda_h) \text{softmax}(v_{pos}^{h\top} r_{ij}), \quad (7)$$

More Global

$$\text{softmax}(Q_i^h K_j^{h\top})$$

>

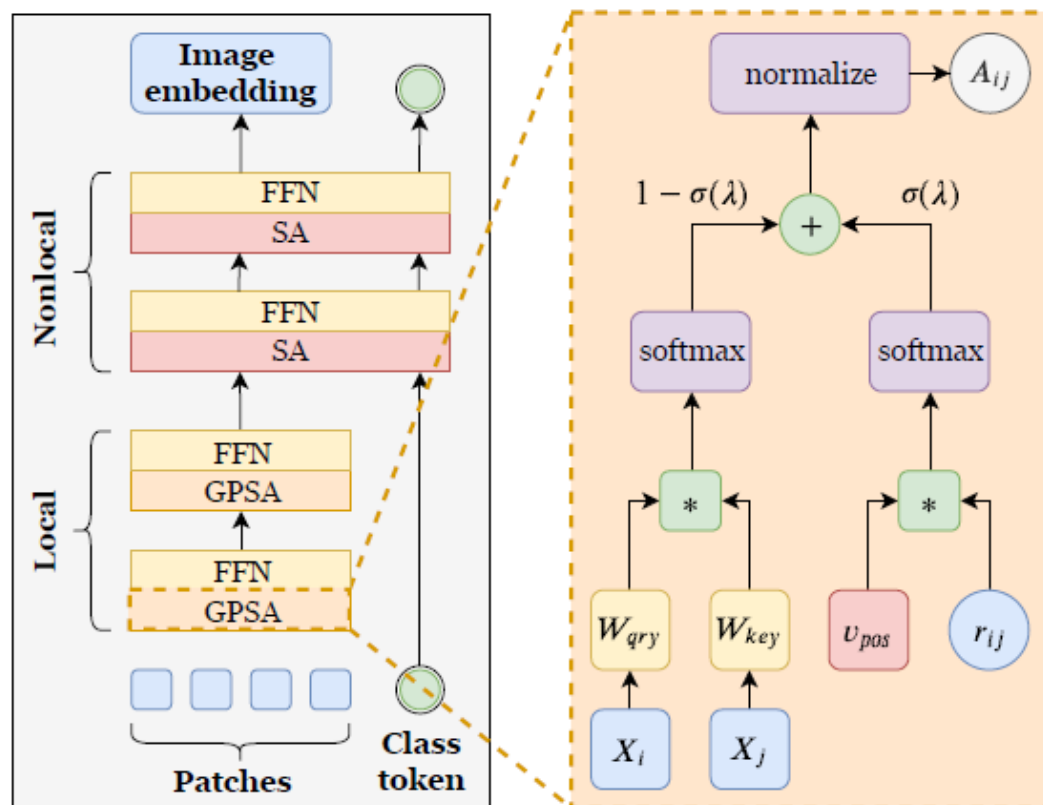
Relative position

$$\text{softmax}(v_{pos}^{h\top} r_{ij})$$

<

More local

- Architecture of the ConViT : Details

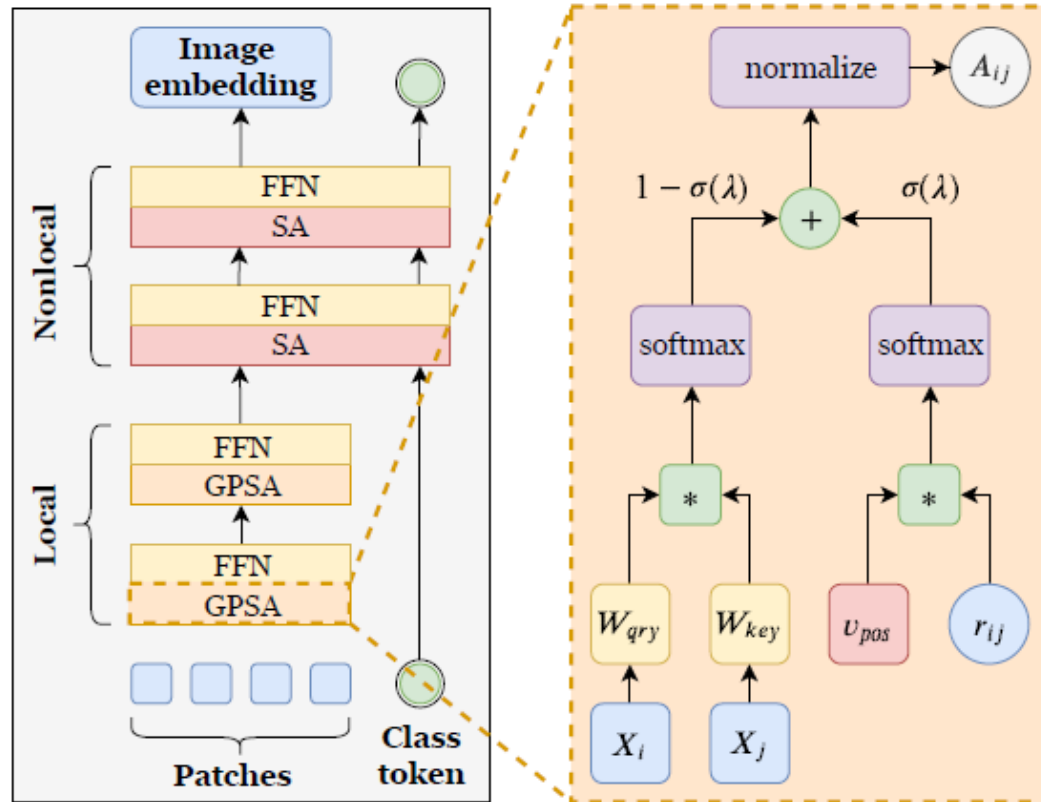


ConViT

Gated Positional
Self-Attention

- GPSA * 10 layer + SA * 2 layer

- Training



ConViT

Gated Positional
Self-Attention

- Hyperparameter-optimized version
- ConvViT models with 4, 9 and 16 attention heads
(to mimic 2 x 2, 3 x 3, and 4x4 Convolutional filters)

Background	Approach	Performance	Ablation Study
------------	----------	-------------	----------------

- Performance

Name	Model	N_h	D_{emb}	Size	Flops	Speed	Top-1	Top-5
Ti	DeiT	3	192	6M	1G	1442	72.2	-
	ConViT	4	192	6M	1G	734	73.1	91.7
Ti+	DeiT	4	256	10M	2G	1036	75.9	93.2
	ConViT	4	256	10M	2G	625	76.7	93.6
S	DeiT	6	384	22M	4.3G	587	79.8	-
	ConViT	9	432	27M	5.4G	305	81.3	95.7
S+	DeiT	9	576	48M	10G	480	79.0	94.4
	ConViT	9	576	48M	10G	382	82.2	95.9
B	DeiT	12	768	86M	17G	187	81.8	-
	ConViT	16	768	86M	17G	141	82.4	95.9
B+	DeiT	16	1024	152M	30G	114	77.5	93.5
	ConViT	16	1024	152M	30G	96	82.5	95.9

Table 1. Performance of the models considered, trained from scratch on ImageNet. Speed is the number of images processed per second on a Nvidia Quadro GP100 GPU at batch size 128. Top-1 accuracy is measured on ImageNet-1k test set without distillation (see SM. B for distillation). The results for DeiT-Ti, DeiT-S and DeiT-B are reported from (Touvron et al., 2020).

- Sample Efficiency

Train size	Top-1			Top-5		
	DeiT	ConViT	Gap	DeiT	ConViT	Gap
5%	34.8	47.8	37%	57.8	70.7	22%
10%	48.0	59.6	24%	71.5	80.3	12%
30%	66.1	73.7	12%	86.0	90.7	5%
50%	74.6	78.2	5%	91.8	93.8	2%
100%	79.9	81.4	2%	95.0	95.8	1%

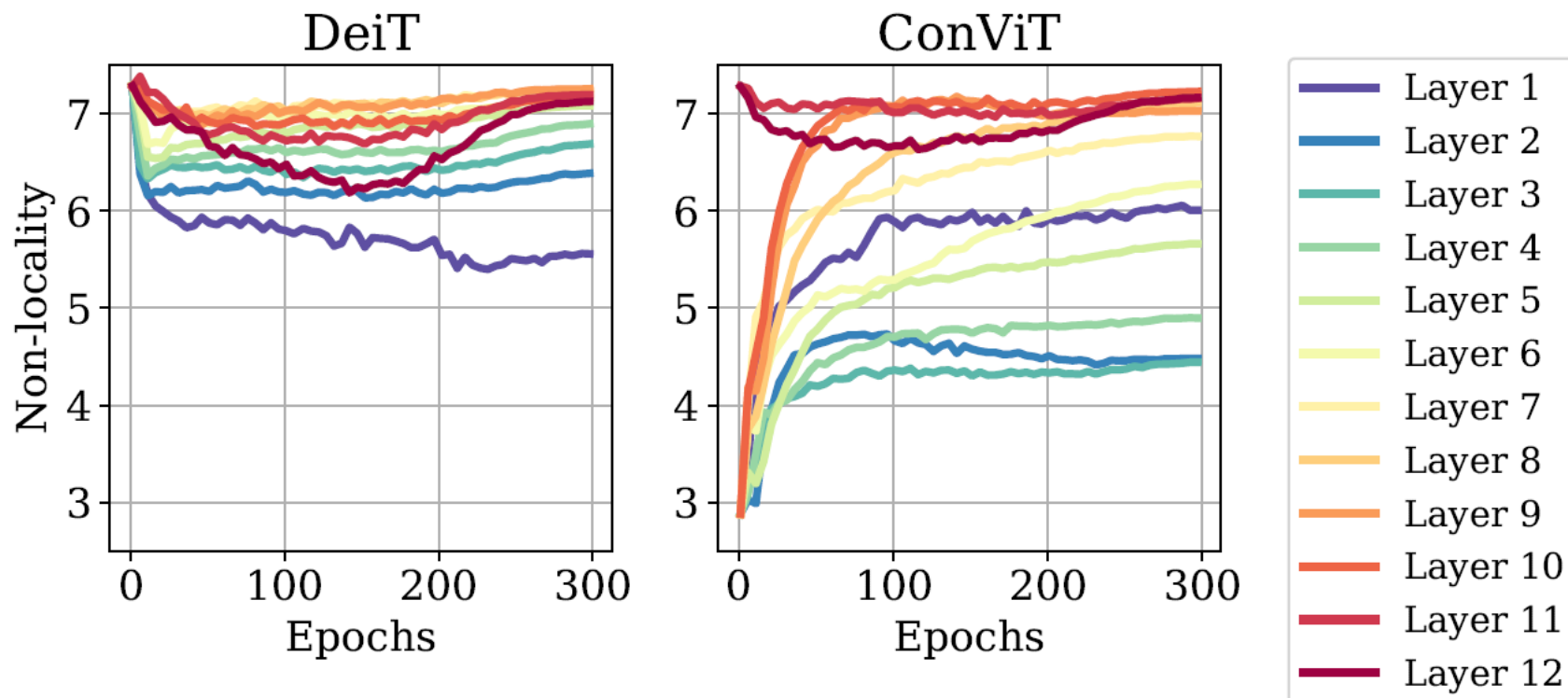
Table 2. **The convolutional inductive bias strongly improves sample efficiency.** We compare the top-1 and top-5 accuracy of our ConViT-S with that of the DeiT-S, both trained using the original hyperparameters of the DeiT (Touvron et al., 2020), as well as the relative improvement of the ConViT over the DeiT. Both models are trained on a subsampled version of ImageNet-1k, where we only keep a variable fraction (leftmost column) of the images of each class for training.

- **Role of locality: Non-locality Comparison between SA & GPSA**

- **Quantifying Non-locality**

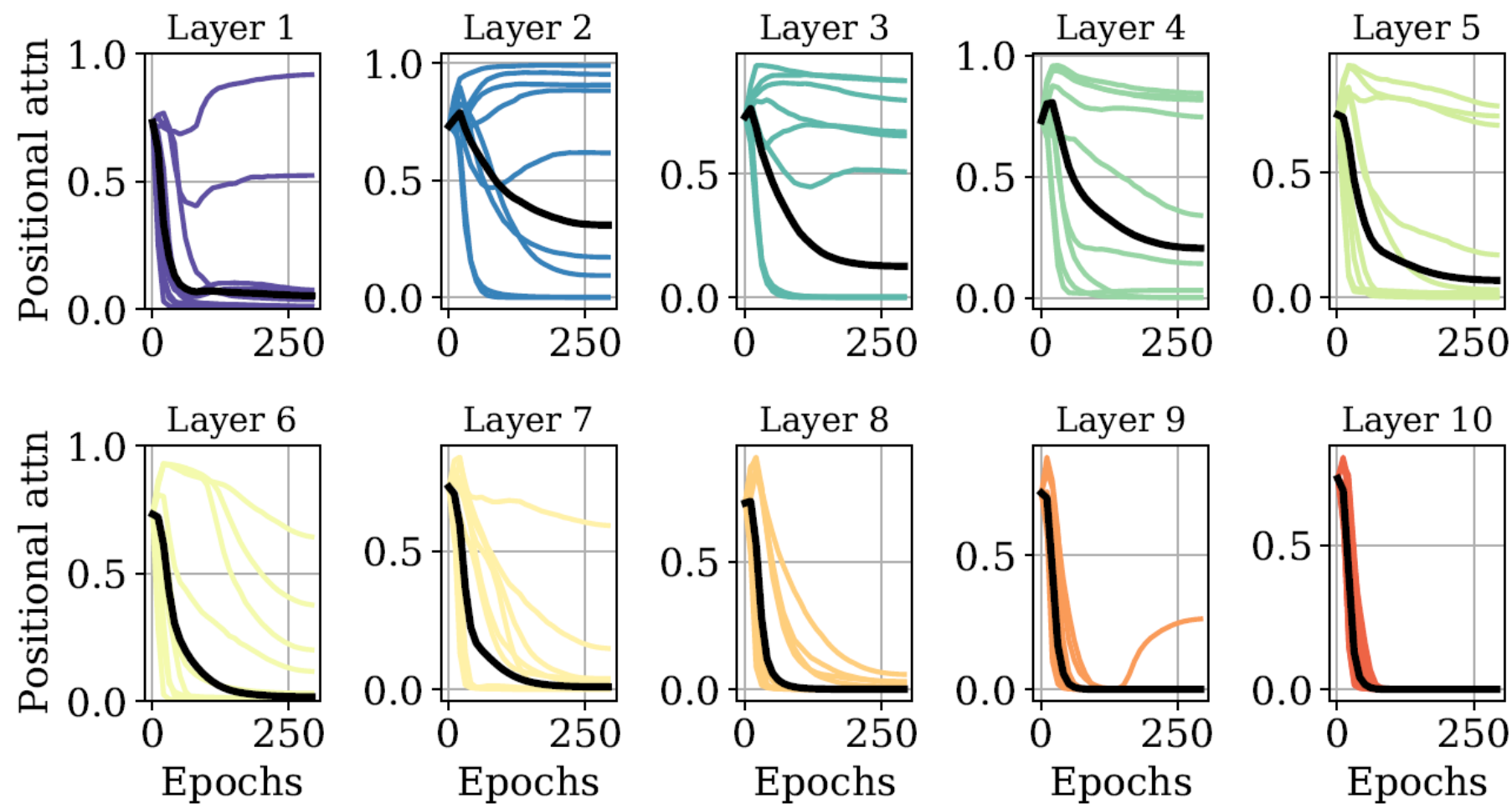
$$D_{loc}^{\ell,h} := \frac{1}{L} \sum_{ij} A_{ij}^{h,\ell} \|\delta_{ij}\|,$$

$$D_{loc}^{\ell} := \frac{1}{N_h} \sum_h D_{loc}^{\ell,h} \quad (8)$$

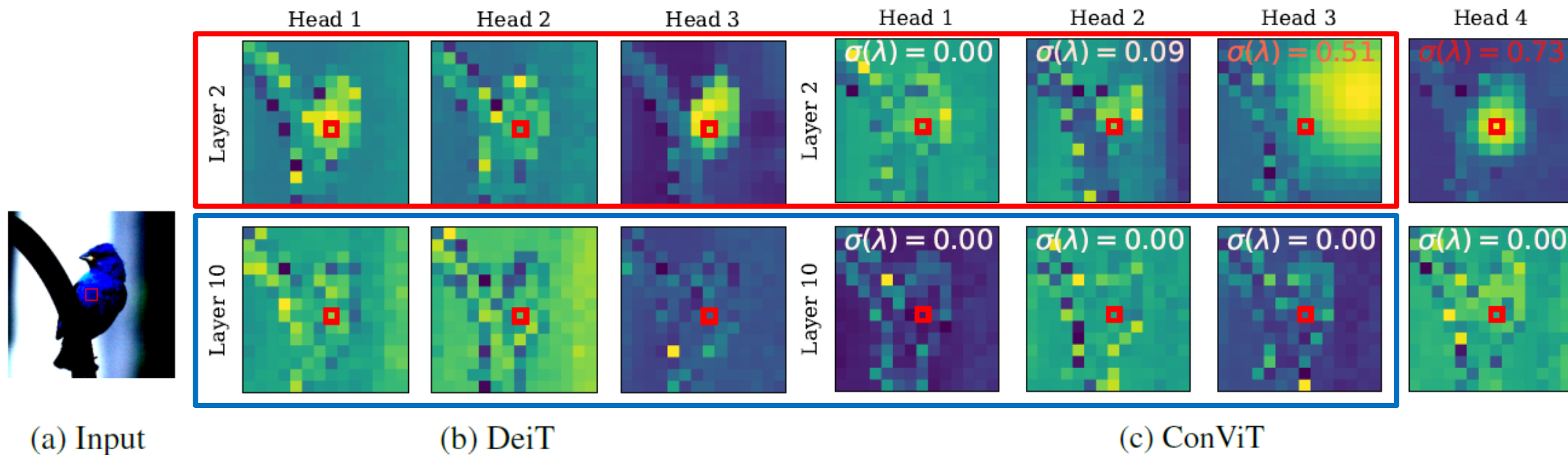


Background	Approach	Performance	Ablation Study
------------	----------	-------------	----------------

- Role of locality: Gating Parameter



- The ConViT learns more diverse attention maps

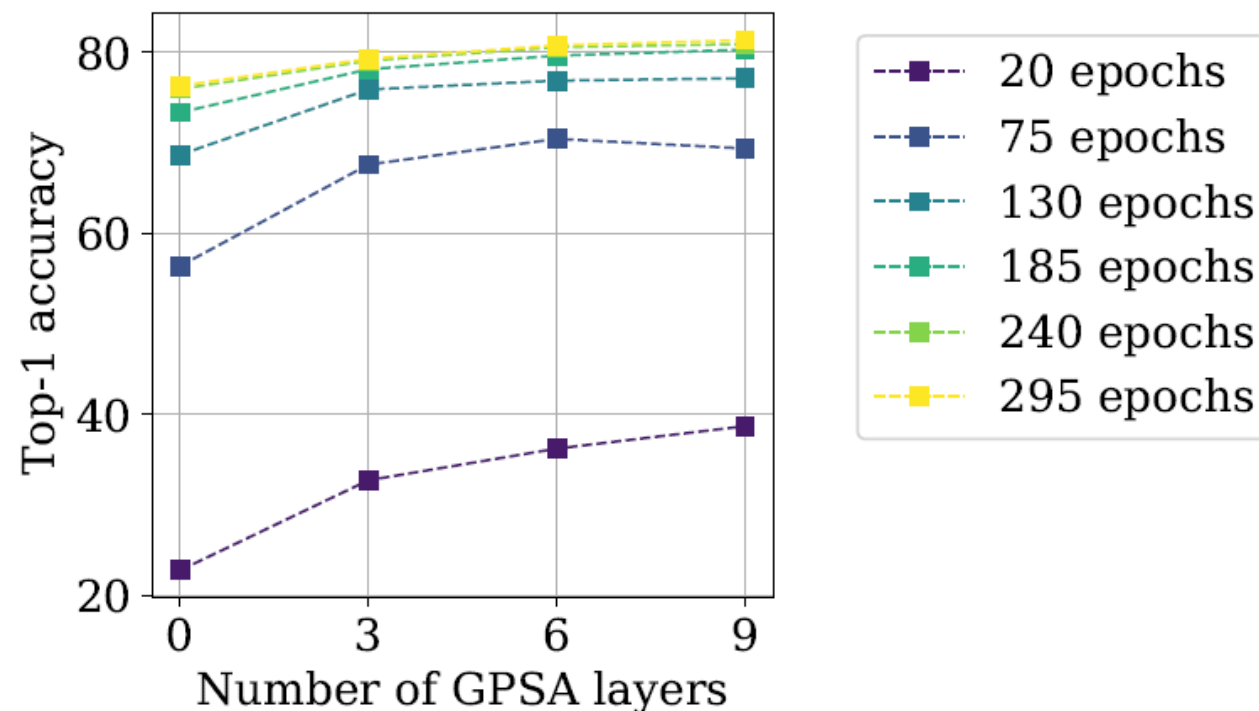
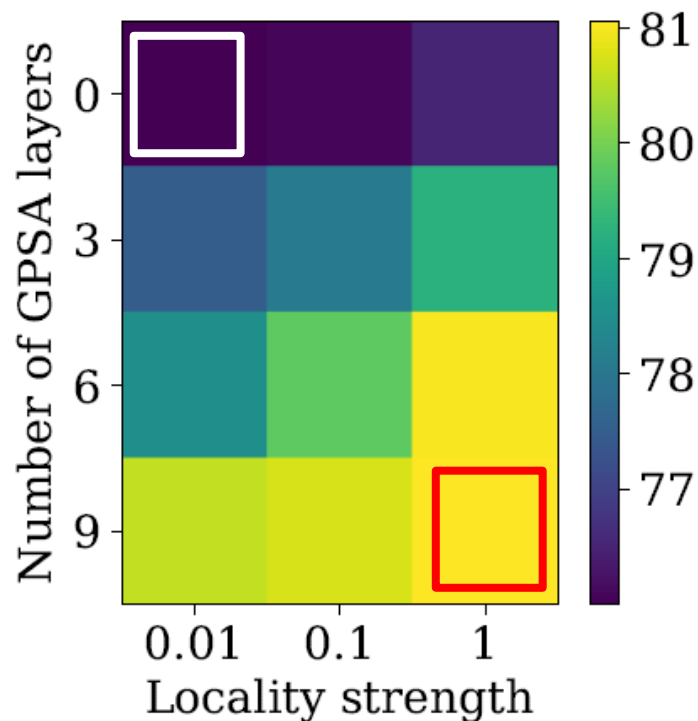


- Layer2 : position-based attention (More Varied attention)
- Layer10: contents-based attention (Similar Attention Map)

Localized

Globalized

- The beneficial effect of Locality
 - Strong Locality is desired (Locality Strength & α)



- Locality strength $\uparrow \propto$ Accuracy \uparrow
- # of GPSA layers $\uparrow \propto$ Accuracy \uparrow