ConViT: Improving Vision Transformers with Soft Convolutional Inductive Biases

Sabiq Muhtadi, Charlie Arleth and Cheng Che Tsai (Michael)

Background

Convolutional Neural Networks	
❑ Universal solution to CV tasks ❑ Characterized by hard-coded inductive biases: ▪ Locality ▪ Weight sharing ❑ Both sample efficient and parameter efficient ▪ High performance floor, low performance ceiling	

Background

Convolutional Neural Networks	Vision Transformers
❑ Universal solution to CV tasks ❑ Characterized by hard-coded inductive biases: ▪ Locality ▪ Weight sharing ❑ Both sample efficient and parameter efficient ▪ High performance floor, low performance ceiling	❑ Leverage self-attention (SA) to capture long range dependencies within the input. ❑ Performs SA across embeddings of patches of pixels. ❑ Matches or exceeds performance of CNN's ▪ Requires pre-training on vast amounts of data ▪ High performance ceiling, low performance floor

Background

Convolutional Neural Networks	Vision Transformers
❑ Universal solution to CV tasks ❑ Characterized by hard-coded inductive biases: ▪ Locality ▪ Weight sharing ❑ Both sample efficient and parameter efficient ▪ High performance floor, low performance ceiling	❑ Leverage self-attention (SA) to capture long range dependencies within the input. ❑ Performs SA across embeddings of patches of pixels. ❑ Matches or exceeds performance of CNN's ▪ Requires pre-training on vast amounts of data ▪ High performance ceiling, low performance floor

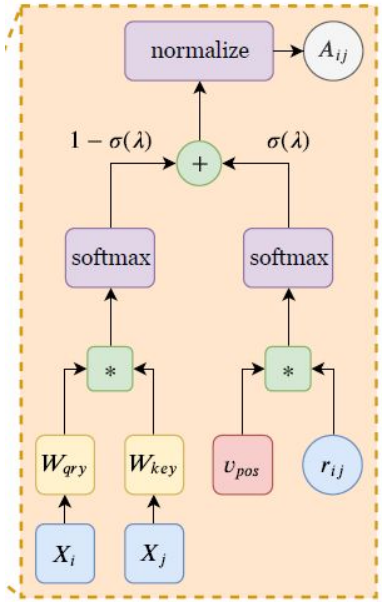
How can one get the best of both worlds?

Background

Solution: ‘softly’ introduce convolutional inductive bias into the ViT, by letting each SA layer **decide** whether to behave as a convolutional layer or not.

Background

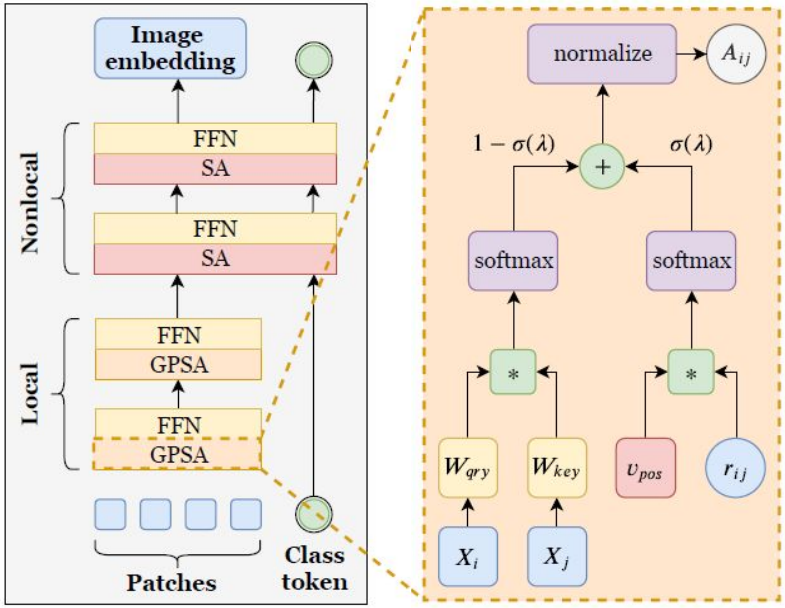
Solution: ‘softly’ introduce convolutional inductive bias into the ViT, by letting each SA layer **decide** whether to behave as a convolutional layer or not.



Gated Positional Self-Attention

Background

Solution: ‘softly’ introduce convolutional inductive bias into the ViT, by letting each SA layer **decide** whether to behave as a convolutional layer or not.



ConViT

Gated Positional Self-Attention

Background

SA:

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

Background

SA:

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

PSA:

$$\boldsymbol{A}_{ij}^h := \text{softmax}\left(\boldsymbol{Q}_i^h \boldsymbol{K}_j^{h\top} + \boldsymbol{v}_{pos}^{h\top} \boldsymbol{r}_{ij}\right)$$

Background

SA:

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

PSA:

$$\boldsymbol{A}_{ij}^h := \text{softmax}\left(\boldsymbol{Q}_i^h \boldsymbol{K}_j^{h\top} + \boldsymbol{v}_{pos}^{h\top} \boldsymbol{r}_{ij}\right)$$

$$\begin{cases} v_{pos}^h := -\alpha^h \left(1, -2\Delta_1^h, -2\Delta_2^h, 0, \ldots 0\right) \\ r_{\boldsymbol{\delta}} := \left(\|\boldsymbol{\delta}\|^2, \delta_1, \delta_2, 0, \ldots 0\right) \\ \boldsymbol{W}_{qry} = \boldsymbol{W}_{key} := \boldsymbol{0}, \quad \boldsymbol{W}_{val} := \boldsymbol{I} \end{cases}$$

For a PSA layer with N_h heads,

- ➔ Δ_h is the position which attention head $\boldsymbol{h}$ pays most attention to relative to the query patch.
- ➔ $\boldsymbol{\alpha}_h$ determines how focused the attention is around the query patch.

Background

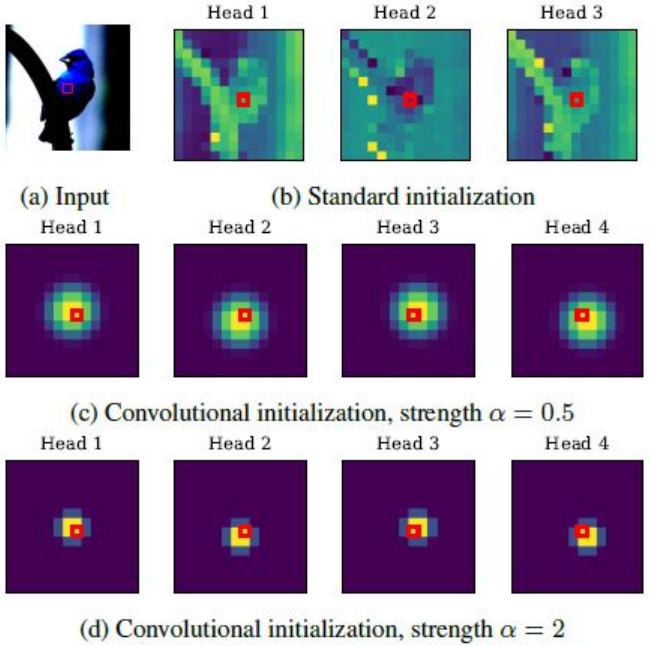


Figure 3. **Positional self-attention layers can be initialized as convolutional layers.** (a): Input image from ImageNet, where the query patch is highlighted by a red box. (b),(c),(d): attention maps of an untrained SA layer (b) and those of a PSA layer using the convolutional-like initialization scheme of Eq. 5 with two different values of the locality strength parameter, α (c, d). Note that the shapes of the image can easily be distinguished in (b), but not in (c) or (d), when the attention is purely positional.

Background

Modify PSA layer to Gated Positional Self Attention (GPSA):

- Restrict attention to subset of patches around query patch – adaptive attention span

Background

Modify PSA layer to Gated Positional Self Attention (GPSA):

- Restrict attention to subset of patches around query patch – adaptive attention span
- Sum content and positional terms after softmax, with their relative importance governed by a learnable gating parameter λ_h

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

Background

Modify PSA layer to Gated Positional Self Attention (GPSA):

- Restrict attention to subset of patches around query patch – adaptive attention span
- Sum content and positional terms after softmax, with their relative importance governed by a learnable gating parameter λ_h

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

- Normalize summation

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

Background

ConViT in comparison to ViT:

- Same as ViT, except in first 10 blocks SA layers are replaced by GPSA.
- Since GPSA layers involve positional attention, it is not suitable to insert the class token as in regular ViT. The class token is thus appended to the patches after the last GPSA layer.

Training details and Performance Comparison between ViT (DeiT) and ConViT

Training details

Distillation

ConViT is based on DeiT, which was chosen for its training efficiency

Several models were built with different number of attention heads to mimic the size of the convolutional filters

All hyperparameters were unchanged from DeiT to ensure performance boost was due to convolution not other factors

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**

ConViT consistently outperforms DeiT (model it was based on) on any amount of parameters and flops.

Distillation

ConViT is compatible with ResMLP distillation with no additional modifications. It can be distilled without being passed through a pre-trained Convolutional network.

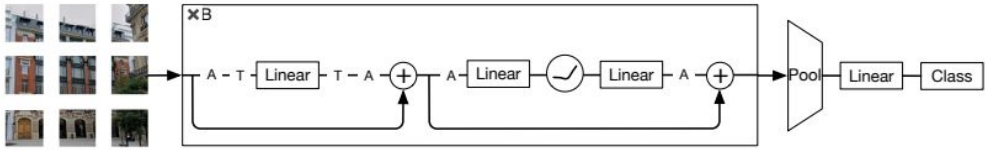


Figure 1: The ResMLP architecture: After linearly projecting the image patches, our network alternately processes them by (1) a communication layer between vectors implemented as a linear layer; (2) a two-layer residual perceptron. We denote by A the Affine element-wise transformation, and by T the transposition.

With distillation ConViT S+ can reach 82.9% top-1 accuracy

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%

ConViT doesn't suffer as much from decreased training size.

Investigating the role of locality

Quantitative definition of `non-locality`

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

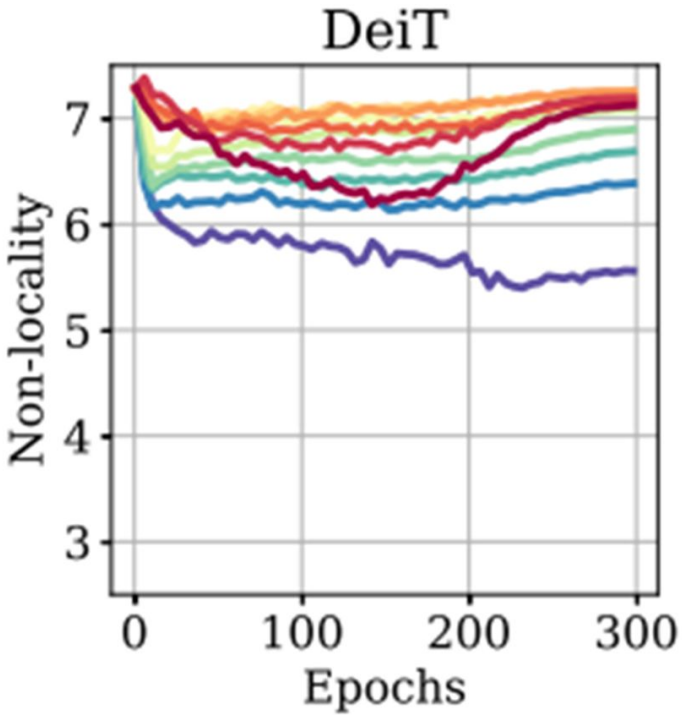
Attention matrix

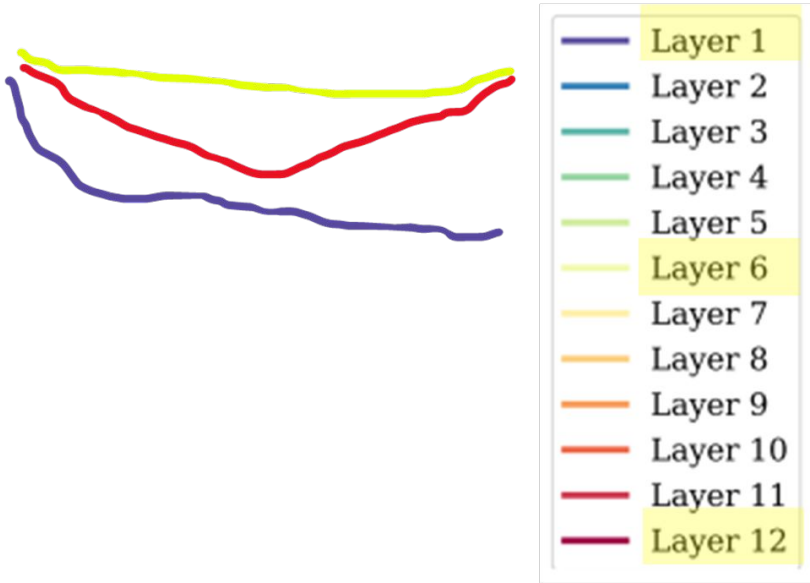
Distance between query and key

Average through multi-head attention

$$D_{loc}^{\ell} := \frac{1}{N_h} \sum_{h} D_{loc}^{\ell,h}$$

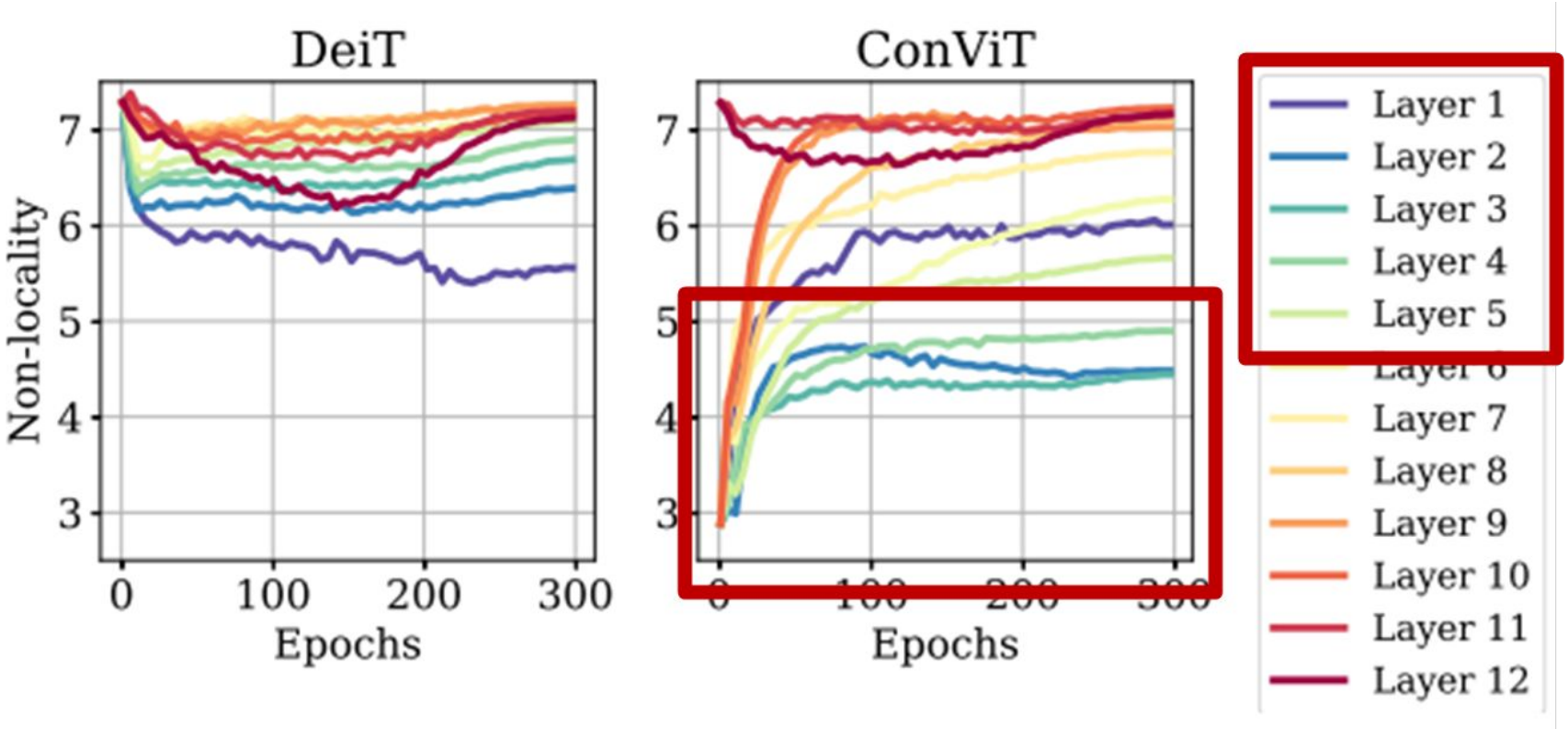
Investigating non-locality of ViT learned from CNN



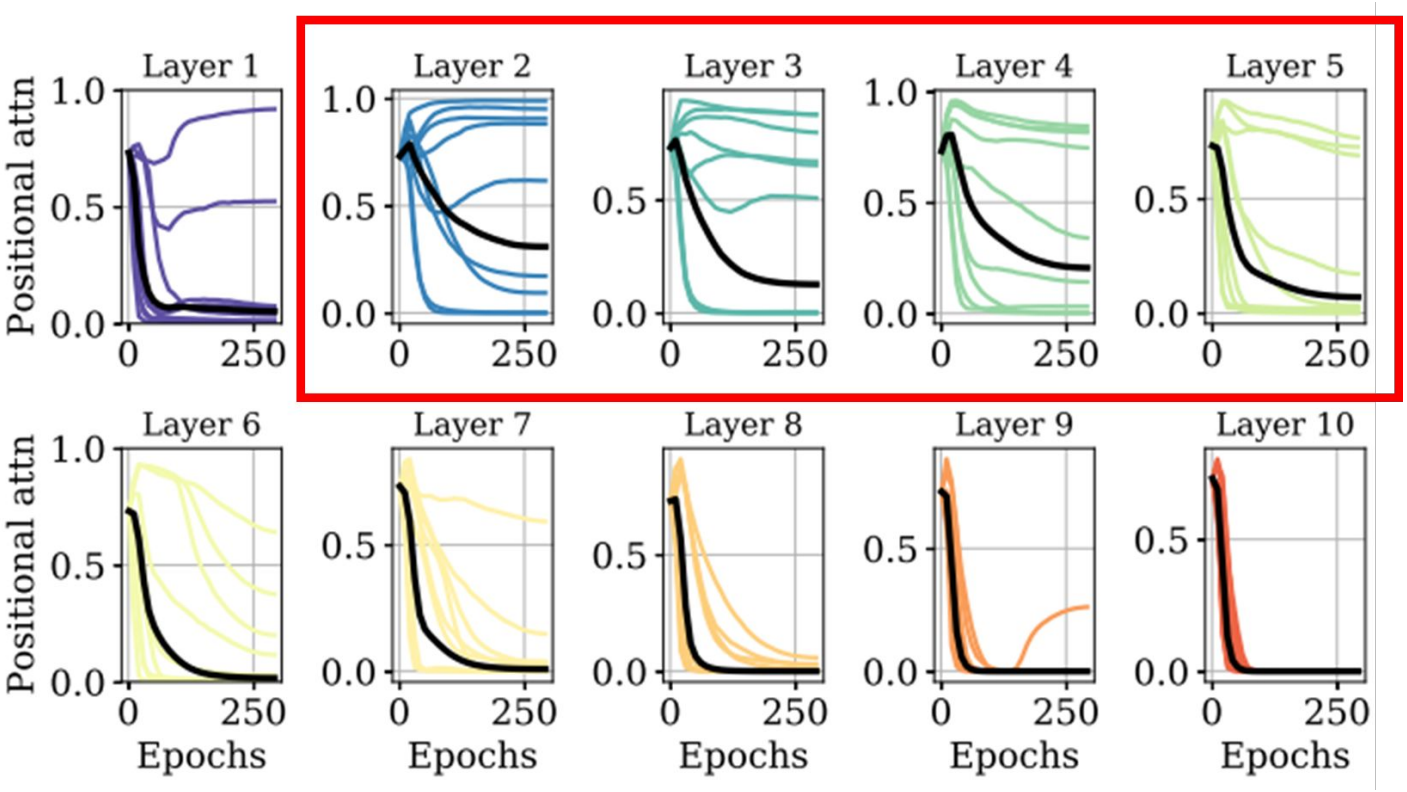


The ViT structure, built without convolution, is encouraged to learn locality in early layers.

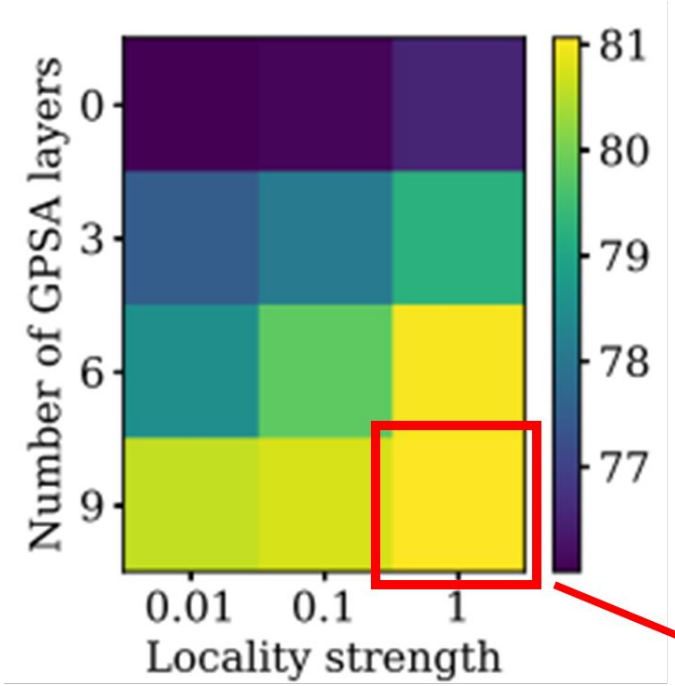
GPSA is still encouraged to learn locality without CNN bias



Investigate the gating parameter



Strong locality is desirable



multiple GPSA layers +
strongest locality

Brief summary

1. Locality is desirable for providing better accuracy.
2. Attention-based structure (ViT) can still learn locality.
3. ConViT learns locality flexibly and effectively, via GPSA.

Ablation

Ref.	Train gating	Conv init	Train GPSA	Use GPSA	Full data	10% data
a (ConViT)	✓	✓	✓	✓	**82.2**	**59.7**
b	✗	✓	✓	✓	82.0	57.4
c	✓	✗	✓	✓	81.4	56.9
d	✗	✗	✓	✓	81.6	54.6
e (DeiT)	✗	✗	✗	✗	79.1	47.8
f	✗	✓	✗	✓	78.6	54.3
g	✗	✗	✗	✓	73.7	44.8

Maybe Conv init. is even important than the gating?

Ref.	Train gating	Conv init	Train GPSA	Use GPSA	Full data	10% data
a (ConViT)	✓	✓	✓	✓	**82.2**	**59.7**
b	✗	✓	✓	✓	82.0	57.4
c	✓	✗	✓	✓	81.4	56.9
d	✗	✗	✓	✓	81.6	54.6
e (DeiT)	✗	✗	✗	✗	79.1	47.8
f	✗	✓	✗	✓	78.6	54.3
g	✗	✗	✗	✓	73.7	44.8