A. The importance of positional gating

In the main text, we discussed the importance of using GPSA layers instead of the standard PSA layers, where content and positional information are summed before the softmax and lead the attention heads to focus only on the positional information. We give evidence for this claim in Fig. 9, where we train a ConViT-B for 300 epochs on ImageNet, but replace the GPSA by standard PSA. The convolutional initialization of the PSA still gives the ConViT a large advantage over the DeiT baseline early in training. However, the ConViT stays in the convolutional configuration and ignores the content information, as can be seen by looking at the attention maps (not shown). Later in training, the DeiT catches up and surpasses the performance of the ConViT by utilizing content information.

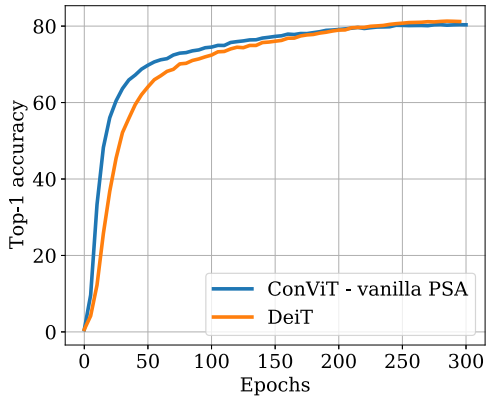


Figure 9. **Convolutional initialization without GPSA is helfpul during early training but deteriorates final performance.** We trained the ConViT-B along with its DeiT-B counterpart for 300 epochs on ImageNet, replacing the GPSA layers of the ConViT-B by vanilla PSA layers.

B. The effect of distillation

Nonlocality In Fig. 10, we compare the nonlocality curves of Fig. 5 of the main text with those obtained when the DeiT is trained via hard distillation from a RegNetY-16GF (84M parameters) (Radosavovic et al., 2020), as in Touvron et al. (2020). In the distillation setup, the nonlocality still drops in the early epochs of training, but increases less at late times compared to without distillation. Hence, the final internal states of the DeiT are more "local" due to the distillation. This suggests that knowledge distillation transfers the locality of the convolutional teacher to the student, in line with the results of (Abnar et al., 2020).

Performance The hard distillation introduced in Touvron et al. (2020) greatly improves the performance of the DeiT. We have verified the complementarity of their distillation methods with our ConViT. In the same way as in the DeiT paper, we used a RegNet-16GF teacher and experimented hard distillation during 300 epochs on ImageNet. The results we obtain are summarized in Tab. 4.

Method	DeiT-S (22M)	DeiT-B (86M)	ConViT-S+ (48M)
No distillation	79.8	81.8	**82.2**
Hard distillation	80.9	**83.0**	82.9

Table 4. Top-1 accuracies of the ConViT-S+ compared to the DeiT-S and DeiT-B, both trained for 300 epochs on ImageNet.

Just like the DeiT, the ConViT benefits from distillation, albeit somewhat less than the DeiT, as can be seen from the DeiT-B performing less well than the ConViT-S+ without distillation but better with distillation. This hints to the fact that the convolutional inductive bias transferred from the teacher is redundant with its own convolutional prior.

Nevertheless, the performance improvement obtained by the ConViT with hard distillation demonstrates that instantiating soft inductive biases directly in a model can yield benefits on top of those obtained by instantiating such biases indirectly, in

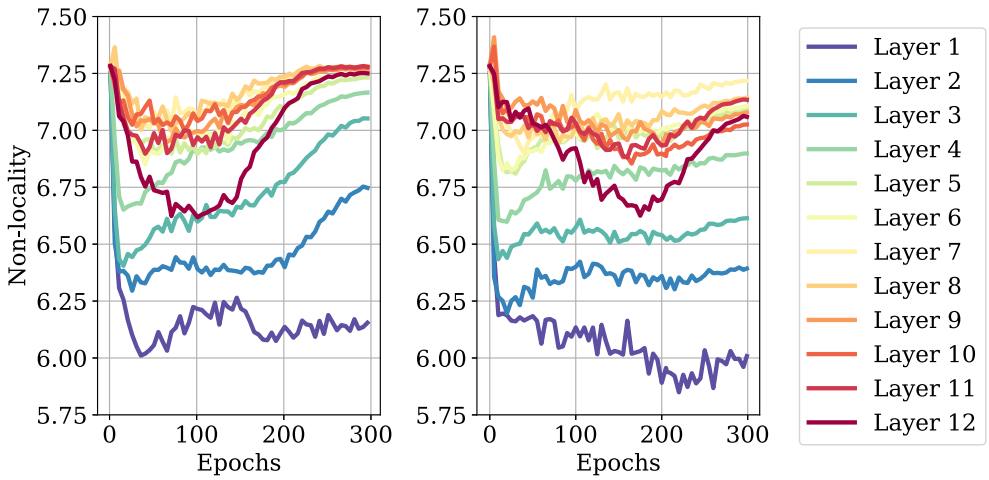


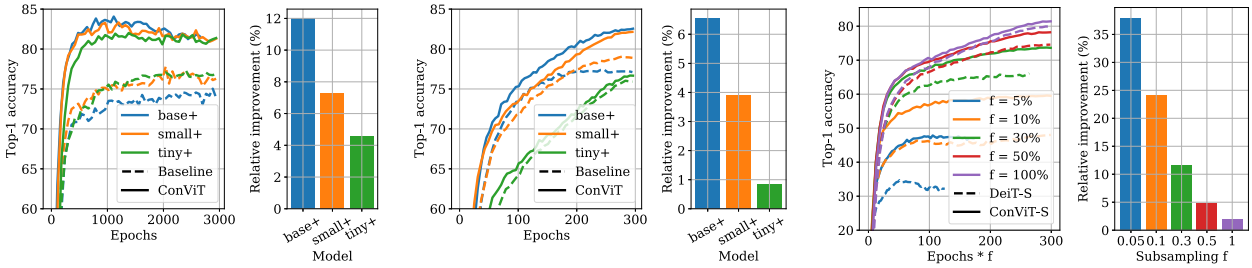
Figure 10. **Distillation pulls the DeiT towards a more local configuration.** We plotted the nonlocality metric defined in Eq. 8 throughout training, for the DeiT-S trained on ImageNet. *Left:* regular training. *Right:* training with hard distillation from a RegNet teacher, by means of the distillation introduced in (Touvron et al., 2020).

this case via distillation.

C. Further performance results

In Fig. 11, we display the time evolution of the top-1 accuracy of our ConViT+ models on CIFAR100, ImageNet and subsampled ImageNet, along with a comparison with the corresponding DeiT+ models.

For CIFAR100, we kept all hyperparameters unchanged, but rescaled the images to 224×224 and increased the number of epochs (adapting the learning rate schedule correspondingly) to mimic the ImageNet scenario. After 1000 epochs, the ConViTs shows clear signs of overfitting, but reach impressive performances (82.1% top-1 accuracy with 10M parameters, which is better than the EfficientNets reported in (Zhao et al., 2020)).



(a) CIFAR100 (b) ImageNet-1k (c) Subsampled ImageNet

Figure 11. **The convolutional inductive bias is particularly useful for large models applied to small datasets.** Each of the three panels displays the top-1 accuracy of the ConViT+ model and their corresponding DeiT+ throughout training, as well as the relative improvement between the best top-1 accuracy reached by the DeiT+ and that reached by the ConViT+. *Left:* tiny, small and base models trained for 3000 epochs on CIFAR100. *Middle:* tiny, small and base models trained for 300 epochs on ImageNet-1k. The relative improvement of the ConViT over the DeiT increases with model size. *Right:* small model trained on a subsampled version of ImageNet-1k, where we only keep a fraction $f \in \{0.05, 0.1, 0.3, 0.5, 1\}$ of the images of each class. The relative improvement of the ConViT over the DeiT increases as the dataset becomes smaller.

In Fig. 12, we study the impact of the various ingredients of the ConViT (presence and number of GPSA layers, gating parameters, convolutional initialization) on the dynamics of learning.

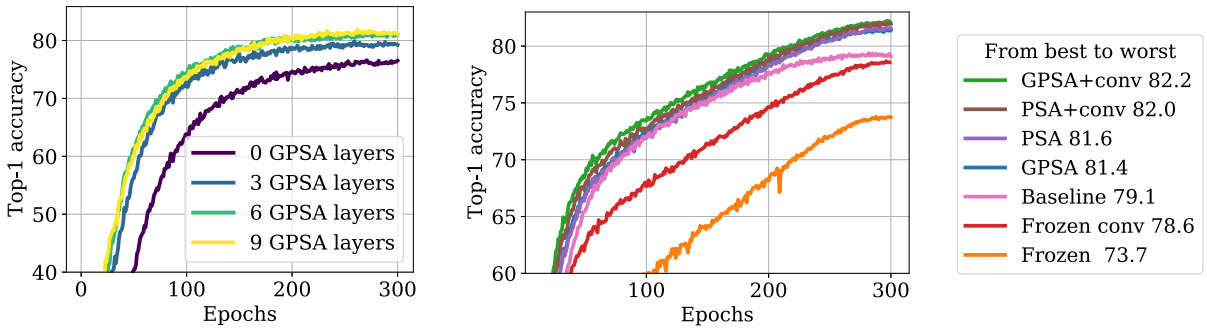
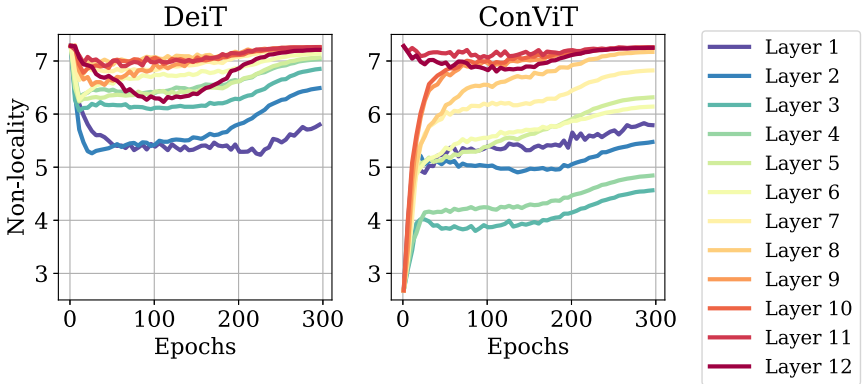


Figure 12. **Impact of various ingredients of the ConViT on the dynamics of learning.** In both cases, we train the ConViT-S+ for 300 epochs on first 100 classes of ImageNet. *Left:* ablation on number of GPSA layers, as in Fig. 8. *Right:* ablation on various ingredients of the ConViT, as in Tab. 3. The baseline is the DeiT-S+ (pink). We experimented (i) replacing the 10 first SA layers by GPSA layers ("GPSA") (ii) freezing the gating parameter of the GPSA layers ("frozen gate"); (iii) removing the convolutional initialization ("conv"); (iv) freezing all attention modules in the GPSA layers ("frozen"). The final top-1 accuracy of the various models trained is reported in the legend.

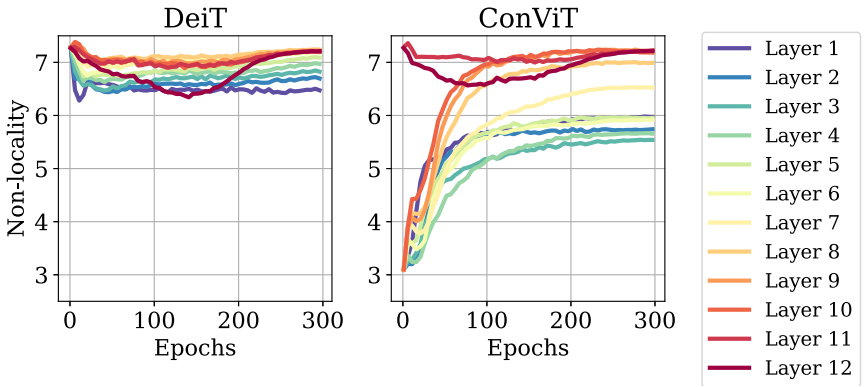
D. Effect of model size

In Fig. 13, we show the analog of Fig. 5 of the main text for the tiny and base models. Results are qualitatively similar to those observed for the small model. Interestingly, the first layers of DeiT-B and ConViT-B reach significantly higher nonlocality than those of the DeiT-Ti and ConViT-Ti.

In Fig. 14, we show the analog of Fig. 6 of the main text for the tiny and base models. Again, results are qualitatively similar: the average weight of the positional attention, $\mathbb{E}_h\sigma(\lambda_h)$, decreases over time, so that more attention goes to the content of the image. Note that in the ConViT-Ti, only the first 4 layers still pay attention to position at the end of training (average gating parameter smaller than one), whereas for ConViT-S, the 5 first layers still do, and for the ConViT-B, the 6 first layers still do. This suggests that the larger (i.e. the more underspecified) the model is, the more layers make use of the convolutional prior.

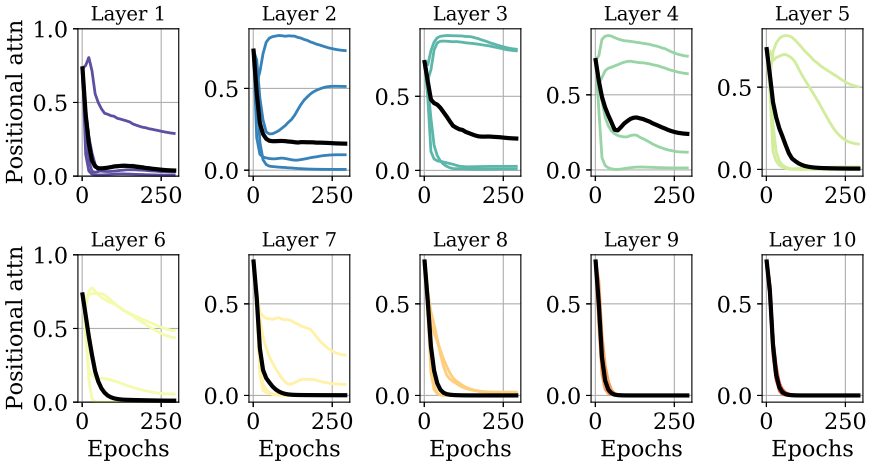


(a) DeiT-Ti and ConViT-Ti

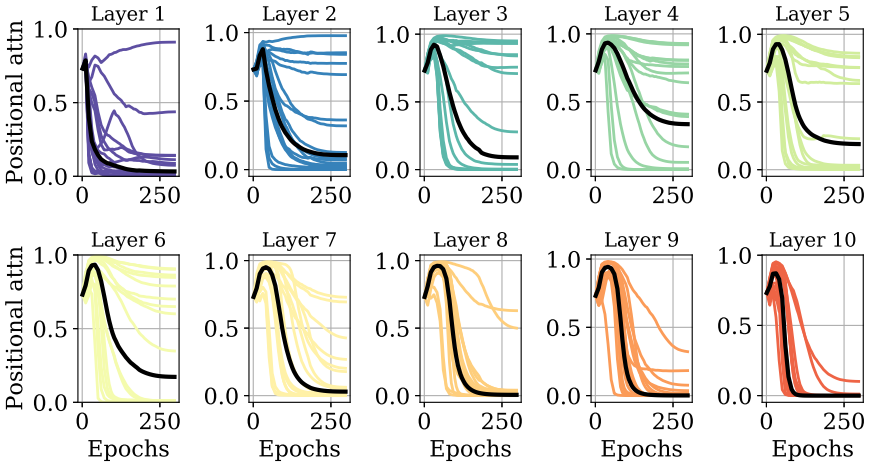


(b) DeiT-B and ConViT-B

Figure 13. **The bigger the model, the more non-local the attention.** We plotted the nonlocality metric defined in Eq. 8 of the main text (the higher, the further the attention heads look from the query pixel) throughout 300 epochs of training on ImageNet-1k.



(a) ConViT-Ti

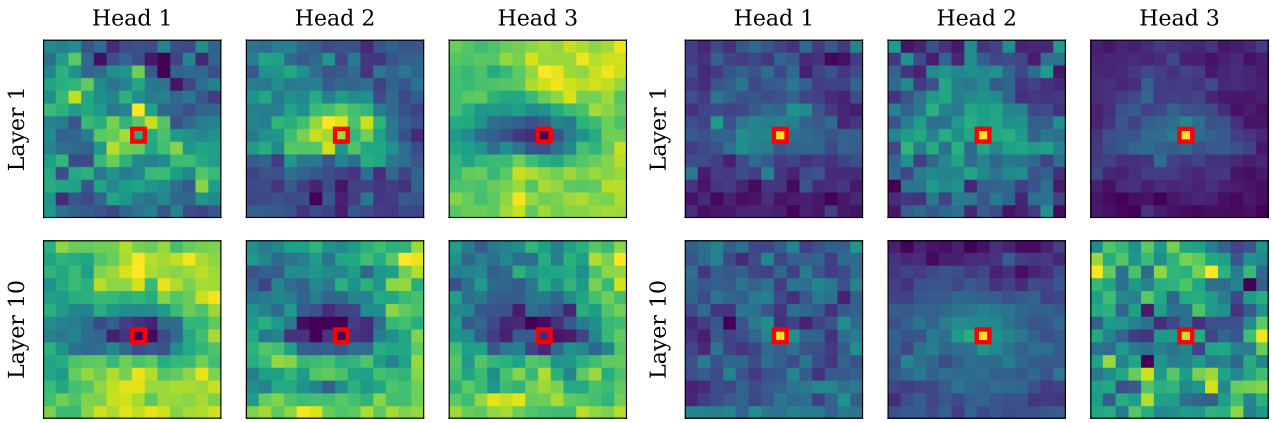


(b) ConViT-B

Figure 14. **The bigger the model, the more layers pay attention to position.** We plotted the gating parameters of various heads and various layers, as in Fig. 6 of the main text (the lower, the less attention is paid to positional information) throughout 300 epochs of training on ImageNet-1k. Note that the ConViT-Ti only has 4 attention heads whereas the ConViT-B has 16, hence the different number of curves.

E. Attention maps

Attention maps of the DeiT reveal locality In Fig. 15, we give some visual evidence for the fact that vanilla SA layers extract local information by averaging the attention map of the first and tenth layer of the DeiT over 100 images. Before training, the maps look essentially random. After training, however, most of the attention heads of the first layer focus on the query pixel and its immediate surroundings, whereas the attention heads of the tenth layer capture long-range dependencies.

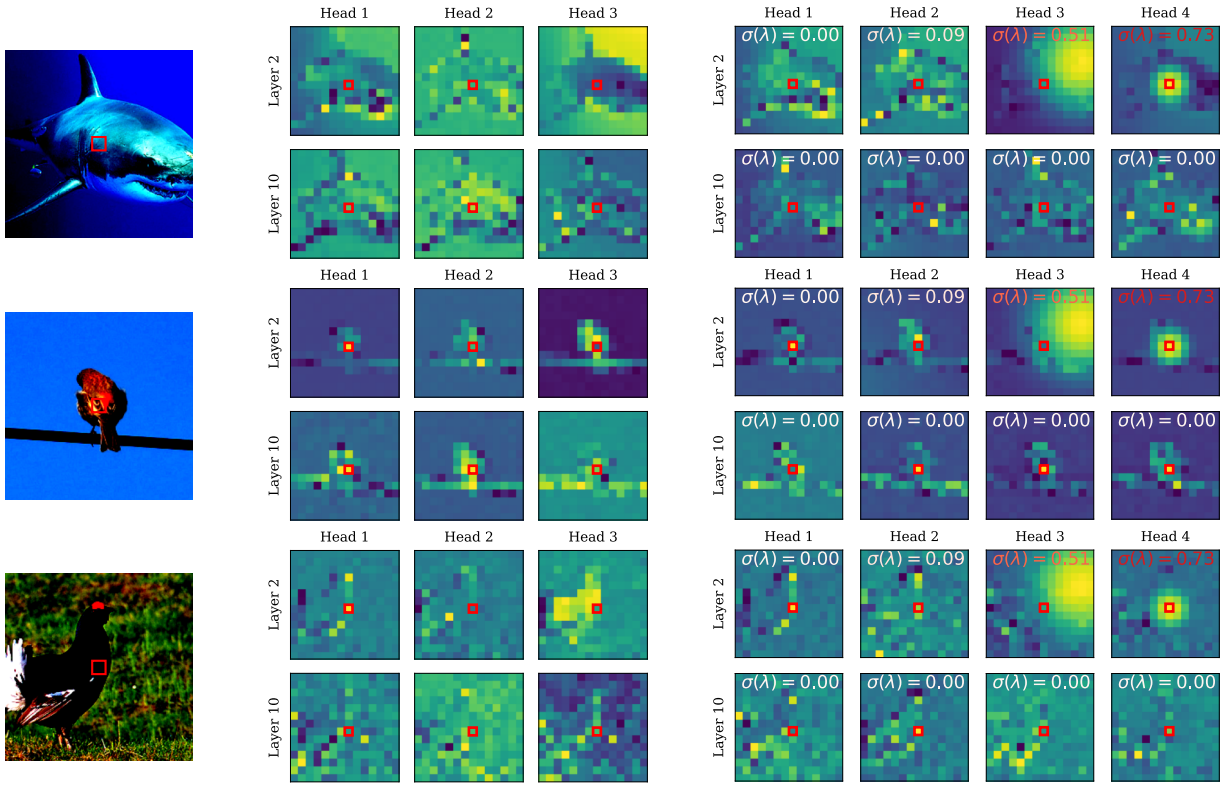


(a) Before training (b) After training

Figure 15. **The averaged attention maps of the DeiT reveal locality at the end of training.** To better visualise the center of attention, we averaged the attention maps over 100 images. *Top:* before training, the attention patterns exhibit a random structure. *Bottom:* after training, most of the attention is devoted to the query pixel, and the rest is focused on its immediate surroundings.

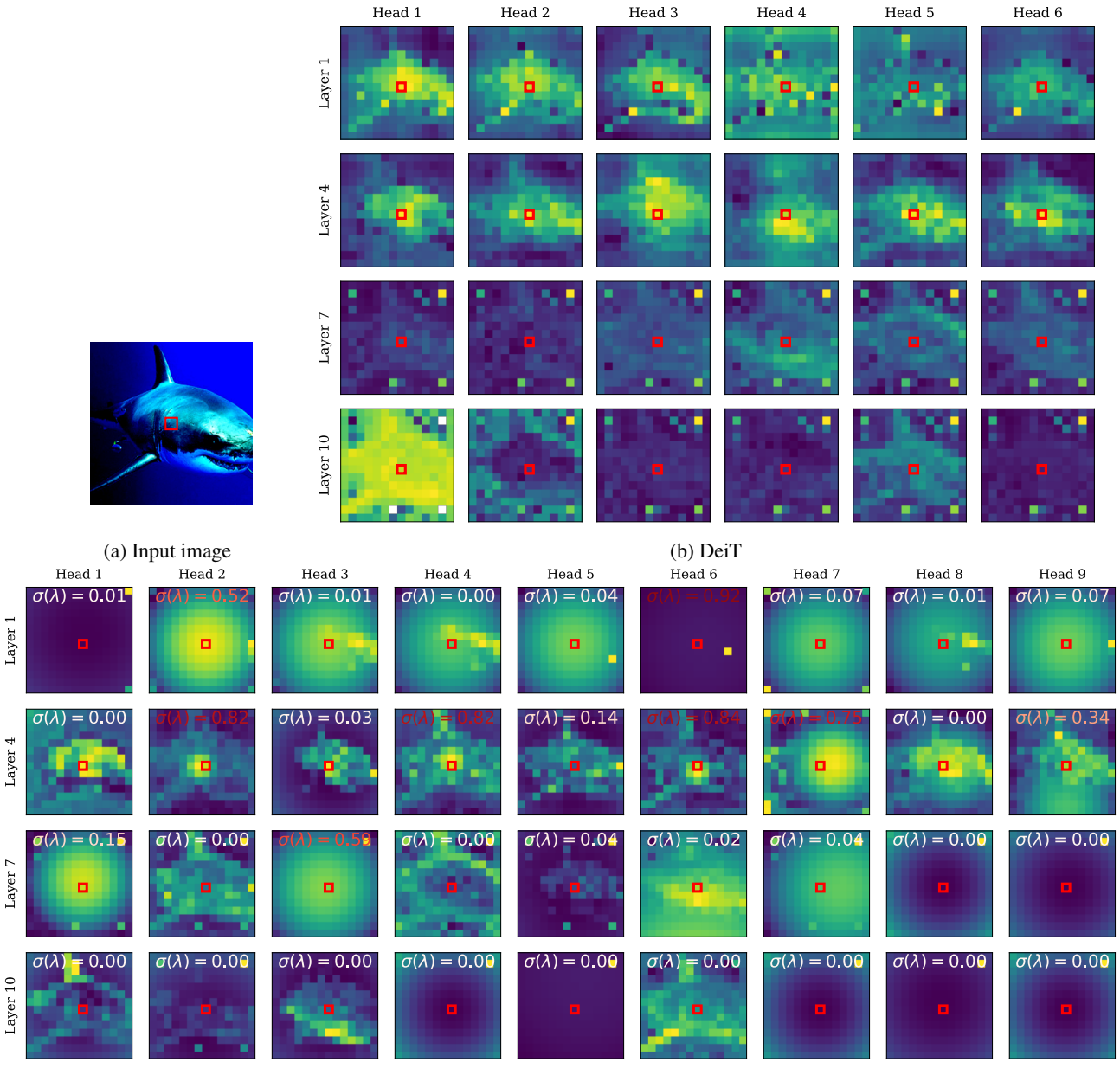
Attention maps of the ConViT reveal the diversity of the attention heads In Fig. 16, we show a comparison of the attention maps of Deit-Ti and ConViT-Ti for different images of the ImageNet validation set. In Fig. 17, we compare the attention maps of DeiT-S and ConViT-S.

In all cases, results are qualitatively similar: the DeiT attention maps look similar across different heads and different layers, whereas those of the ConViT perform very different operations. Notice that in the second layer, the third and forth head focus stay local whereas the first two heads focus on content. In the last layer, all the heads ignore positional information, focusing only on content.



(a) Input images (b) DeiT (c) ConViT

Figure 16. Left: input image which is embedded then fed into the models. The query patch is highlighted by a red box and the colormap is logarithmic to better reveal details. *Center:* attention maps obtained by a DeiT-Ti after 300 epochs of training on ImageNet. *Right:* Same for ConViT-Ti. In each map, we indicated the value of the gating parameter in a color varying from white (for heads paying attention to content) to red (for heads paying attention to position).



(a) Input image

(b) DeiT

(c) ConViT

Figure 17. Attention maps obtained by a DeiT-S and ConViT-S after 300 epochs of training on ImageNet. In each map, we indicated the value of the gating parameter in a color varying from white (for heads paying attention to content) to red (for heads paying attention to position).

F. Further ablations

In this section, we explore masking off various parts of the network to understand which are most crucial.

In Tab. 5, we explore the importance of the absolute positional embeddings injected to the input in both the DeiT and ConViT. We see that masking them off at test time a mild impact on accuracy for the ConViT, but a significant impact for the DeiT, which is expected as the ConViT already has relative positional information in each of the GPSA layers. This also shows that the absolute positional information contained in the embeddings is not very useful.

In Tab. 6, we explore the relative importance of the positional and content information by masking them off at test time. To do so, we manually set the gating parameter $\sigma(\lambda)$ to 1 (no content attention) or 0 (no positional attention). In the first GPSA layers, both procedures affect performance similarly, signalling that positional and content information are both useful. However in the last GPSA layers, masking the content information kills performance, whereas masking the positional information does not, confirming that content information is more crucial.

Model	Mask pos embed	No mask
DeiT-Ti	38.3	72.2
ConViT-Ti	67.1	73.1

Table 5. Performance on ImageNet with the positional embeddings masked off at test time.

# layers masked	Mask content	Mask position	No mask
3	62.3	63.5	73.1
5	35.0	53.1	73.1
10	1.3	46.8	73.1

Table 6. Performance of ConViT-Ti on ImageNet with positional or content attention masked off at test time.