Supplementary material for
LeViT: a Vision Transformer in ConvNet's Clothing for Faster Inference

In this supplementary material, we report details and results that did not fit in the main paper:

Appendix A details the timings of constituent block and provides more details about our ablation. We provide visualizations of the attention bias in Appendix B.

A. Detailed analysis

A.1. Block timings

In this section we compare the differences in design between DeiT and LeViT blocks from the perspective of a detailed runtime analysis. We measure the runtime of their constituent parts side-by-side in the supplementary Table 1. For DeiT-Tiny, we replace the GELU activation with Hardswish, as otherwise it dominates the runtime.

For DeiT, we consider a block from DeiT-tiny. For LeViT, we consider a block from the first stage of LeViT-256. Both operate at resolution 14×14 and have comparable run times, although LeViT is 33% wider ($C = 256$ vs $C = 192$). Note that stage 1 is the most expensive part of LeViT-256. In stages 2 and 3, the cost is lower due to the reduction in resolution (see Figure 4 of the main paper).

LeViT spends less time calculating the attention QK^T, but more time on the subsequent matrix product AV. Despite having the larger block width C, LeViT spends less time on the MLP component as the expansion factor is halved from four to two.

A.2. More details on our ablation

Here we give additional details of the ablation experiments in Section 5.6 and Table 4 of the main paper.

A1 – without pyramid shape. We test the effect of the LeViT pyramid structure, we replace the three stages with a single stage of depth 11 at resolution 14×14. To preserve the FLOP count, we take $D = 19$, $N = 3$ and $C = 2ND = 114$.

A6 – without wider blocks. Compared to DeiT, LeViT blocks are relatively wide given the number of FLOPs, with smaller keys and MLP expansion factors. To test this change we modify LeViT-128S to have more traditional blocks while preserving the number of FLOPs. We therefore take Q, K, V to all have dimension $D = 30$, and $C = ND = 120, 180, 240$ for the three stages. As in DeiT, the MLP expansion ratio is 4. In the subsampling layers we use $N = 4C/D = 16, 24$, respectively.

Table 1. Timings for the components of the LeViT architecture on an Intel Xeon E5-2698 CPU core with batch size 1.

Model	DeiT-tiny	LeViT-256
Dimensions	$C = 192$ $N = 3$ $D = 64$	$C = 256$ $N = 4$ $D = 32$
Component	Runtime (μs)	Runtime (μs)
LayerNorm	49	n/a
Keys Q, K	299	275
Values V	172	275
Product QK^T	228	159
Product Attention AV	161	206
Attention projection	175	310
MLP	1390	1140
Total	2474	2365

B. Visualizations: attention bias

The attention bias maps from Eqn. 1 in the main paper are just two-dimensional maps. Therefore we can vizualize them, see Figure 1. They can be read as the amount of attention between two pixels that are at a certain relative position. The lowest values of the bias are low enough (-20) to suppress the attention between the two pixels, since they are input to a softmax.

We can observe that some heads are quite uniform, while other heads specialize in nearby pixels (*e.g.* most heads of the shrinking attention). Some are clearly directional, *e.g.* heads 1 and 4 of Stage 2/block 1 handle the pixels adjacent vertically and horizontally (respectively). Head 1 of stage 2, block 4 has a specific period-2 pattern that may be due to the fact that its output is fed to a sub-sampling filter in the next shrinking attention block.

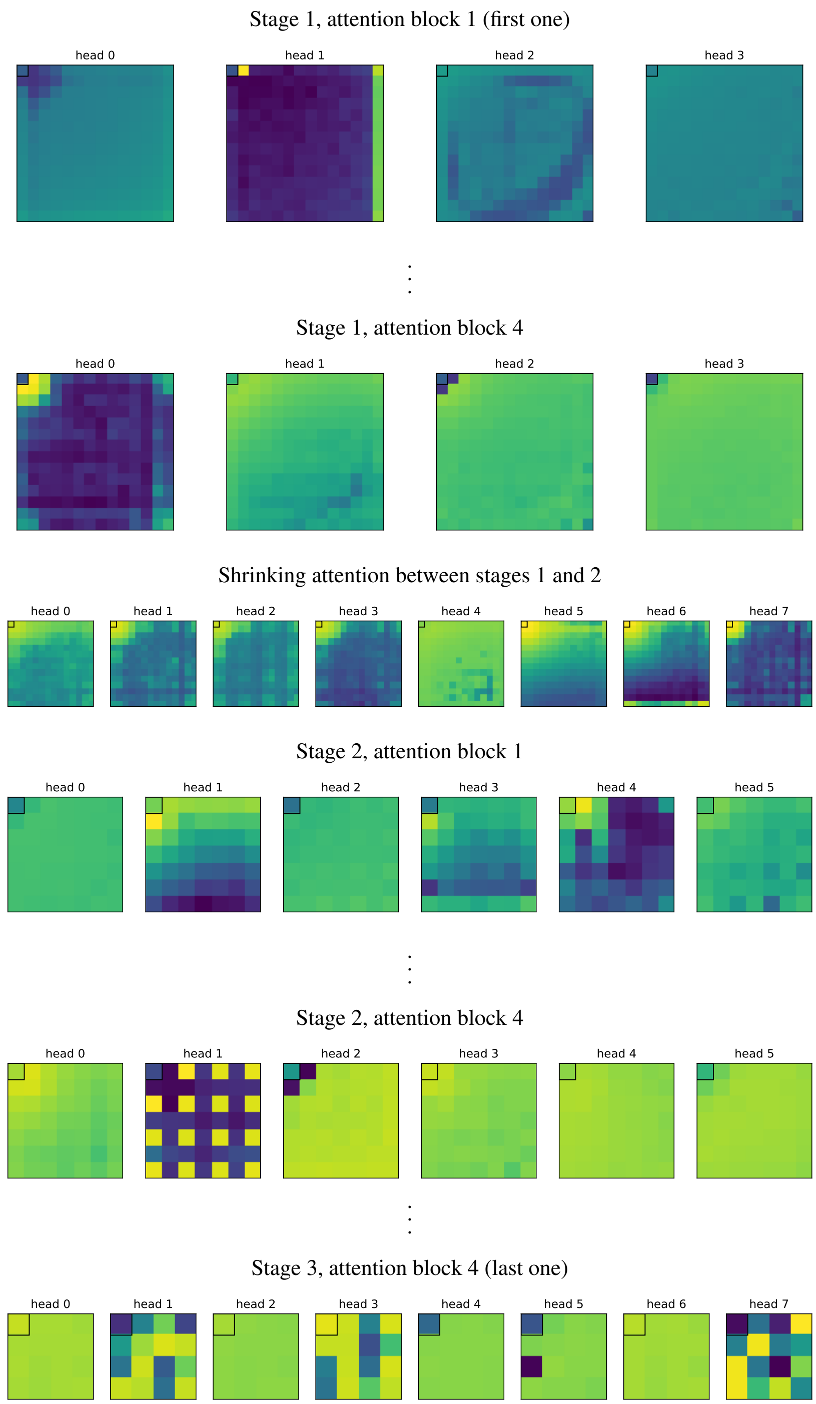

Figure 1. Visualization of the attention bias for several blocks of a trained LeViT-256 model. The center for which the attention is computed is the upper left pixel of the map (with a square). Higher bias values are in yellow, lower values in dark blue (values range from -20 to 7).