RapidNet: Multi-Level Dilated Convolution Based Mobile Backbone

Supplementary Material

A. Ablation Studies

The ablation studies are conducted on ImageNet-1K [9]. Table 4 reports the ablation study of RapidNet-Ti (RNet-Ti) on the effects of static graph convolution, pointwise convolution, and 3 $\times$ 3 convolution. Table 5 reports the effects of conditional positional encoding (CPE), the large kernel feedforward network (FFN), single-level dilated convolution (SLDC), and multi-level dilated convolution (MLDC).

A.1. Effect of Different Convolution Types and Knowledge Distillation

Starting with a RapidNet-Ti configuration with no CPE, no large kernel FFN, and using pointwise convolution instead of MLDC we achieve a top-1 accuracy of 75.2%. We can see this is a lower accuracy than the static graph convolution of MobileViG-Ti in Table 4, which achieves 75.7% with an increase of 0.3 GMACs (42.9% increase in GMACs). This shows that static graph convolution adds additional information beyond pointwise convolution. When we replace the PW convolution with a 3×3 kernel convolution we gain 0.5% in accuracy compared to the PW convolution, but we also gain 0.1 GMACs. This demonstrates that our RapidNet architecture can match MobileViG in accuracy with a lower amount of GMACs by simply using 3 $\times$ 3 kernel convolutions in our architecture.

Table 4. Ablation study for RapidNet-Ti on ImageNet-1K benchmark for how SVGA [46], pointwise (PW) convolution, 3×3 kernel convolution, and knowledge distillation (KD) affect performance. In each column yes means this form of convolution was used in the network in place of multi-level dilated convolution in the dilated convolution block. No means this form of convolution was not used to replace multi-level dilated convolution in the ablation study. For KD yes and no just mean whether KD was used. For the MLDC column yes means MLDC was used and no replacement method of convolution was used.

Model	GMACs	SVGA	MLDC	PW Conv	3×3 Conv	KD	Acc
MViG-Ti	0.7	Yes	No	Yes	No	Yes	75.7
RNet-Ti	0.4	No	No	Yes	No	Yes	75.2
RNet-Ti	0.5	No	No	No	Yes	Yes	75.7
MViG-Ti	0.7	Yes	No	No	No	No	74.5
RNet-Ti	0.6	No	Yes	No	No	No	75.1
RNet-Ti	**0.6**	**No**	**Yes**	**No**	**No**	**Yes**	**76.3**

When trying to determine the impact of knowledge distillation in Table 4 we can see RapidNet-Ti loses 1.2% in accuracy as does MobileViG-Ti. This shows that for both models knowledge distillation is beneficial to performance.

A.2. Effect of Dilated Convolutions and Positional Encoding

Replacing the 3×3 kernel convolution in Table 4 with a single-level dilated convolution in Table 5, we gain another 0.1% increase in accuracy with no increase in the number of GMACs. Adding CPE and the large kernel FFN increases the accuracy by 0.1% each with a near negligible gain in GMACs. Lastly adding MLDC to replace SLDC increases the accuracy to 76.3% providing another 0.3% increase with an increase of only 0.1 GMACs.

Table 5. Ablation study for RapidNet-Ti on ImageNet-1K benchmark for how CPE, large kernel feedforward network (LKFFN), single-level dilated convolution (SLDC), and multi-level dilated convolution (MLDC) affect performance. Results are averaged over two runs.

Model	GMACs	CPE	LKFFN	SLDC	MLDC	Acc (%)
RNet-Ti	0.5	No	No	Yes	No	75.8
RNet-Ti	0.5	Yes	No	Yes	No	75.9
RNet-Ti	0.5	Yes	Yes	Yes	No	76.0
RNet-Ti	**0.6**	**Yes**	**Yes**	**No**	**Yes**	**76.3**

A.3. Effect of Dilation Factors, Kernel Sizes, and Deformable Convolution

Replacing the 3×3 kernel convolution in MLDC with 5×5 kernel convolution in Table 6, we gain only 0.1% in accuracy, but we gain 2 million parameters. Due to this increased computational cost and minimal benefit in terms of accuracy we opt for 3×3 kernel convolution in our network. We also perform an ablation study using larger dilation factors in our RapidNet model by increasing the dilation factor in MLDC from 2 and 3 to 3 and 4. Increasing the dilation factor actually leads to a decrease in accuracy of 0.4% falling from 76.3% to 75.9%. Replacing MLDC with deformable convolution we gain 0.1 million parameters due to the learnable offsets, but we do not see an increase in accuracy. Instead we actually see a decrease in accuracy of 0.4%. For image classification tasks, dilated convolutions are more widely used as opposed to deformable convolutions which are more widely used for tasks like object detection and segmentation [4, 8].

Table 6. Ablation study for RapidNet-Ti on ImageNet-1K for how kernel size and dilation factors in MLDC affect performance.

Model	Params (M)	Kernel	Dilation	Deformable	Acc (%)
RNet-Ti	6.7	3x3	No	Yes	76.0
RNet-Ti	6.6	3x3	3,4	No	75.9
RNet-Ti	8.6	5x5	2,3	No	76.4
RNet-Ti	**6.6**	**3x3**	**2,3**	**No**	**76.3**

B. Further Latency Results

B.1. RapidNet ImageNet-1k Classification versus Latency

We visualize RapidNet's accuracy-latency tradeoff in Figure 4 to demonstrate we are not only optimal in terms of the accuracy-GMACs tradeoff as shown in Figure 1, but also in terms of accuracy versus latency.

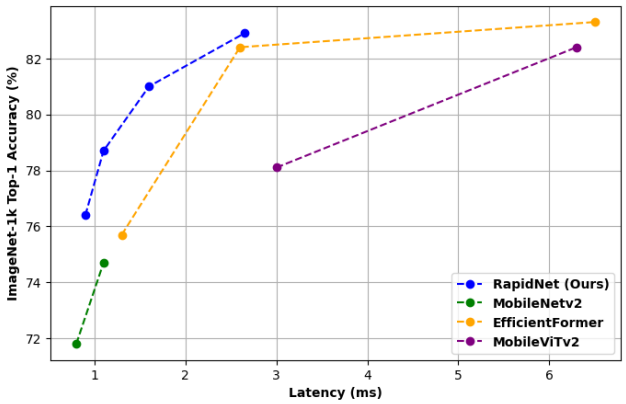

Figure 4. **Comparison of accuracy versus latency on ImageNet-1K**. RapidNet achieves the best accuracy-latency tradeoff on all model sizes compared.