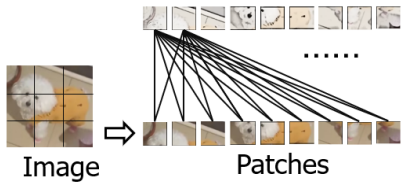
Window Shifted Big Kernel Depthwise Attention based Low-Complexity MobileViT

謝曉梦 池永研究室 修士課程修了

Background



MobileViT performs better than CNN and other ViT models, but it's still slow for 300-FPS tasks

Speeding up MobileViT is needed

Problems

Existing works only focus on convolution blocks in MobileViT

- Limited feature extraction ability of big kernel
- Ignore convolution blocks
- Accuracy drops after speeding up

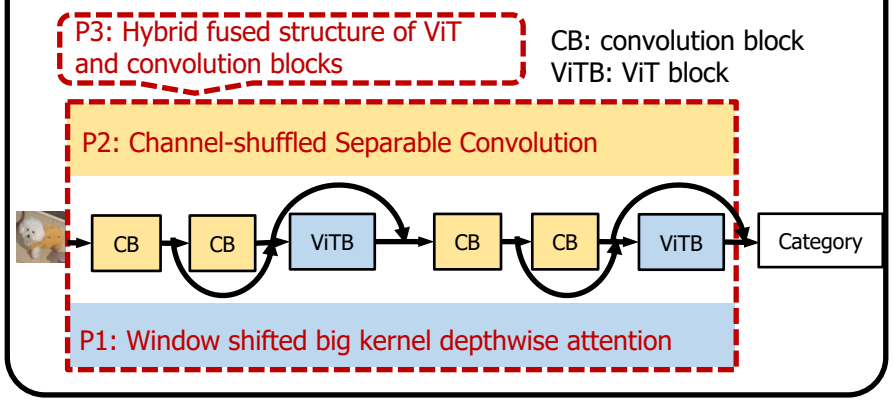
Solution

Both optimize convolution block and ViT block

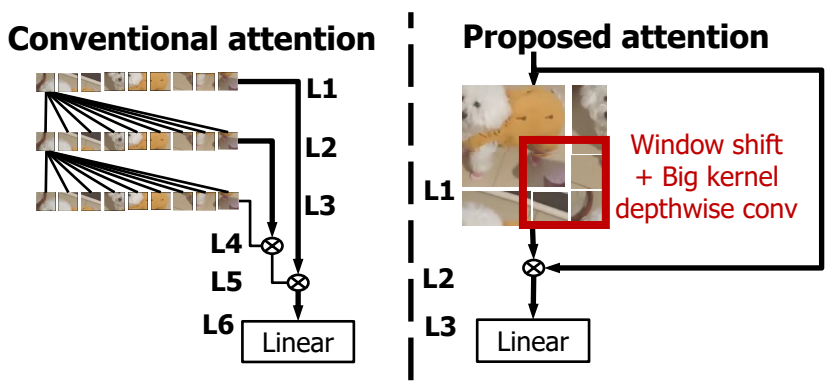
- **P1:** Window Shifted Big kernel Depthwise Attention
- **P2**: Channel-shuffled Separable Convolution
- **P3**: Hybrid fused structure of ViT and convolution blocks

Proposed Methods

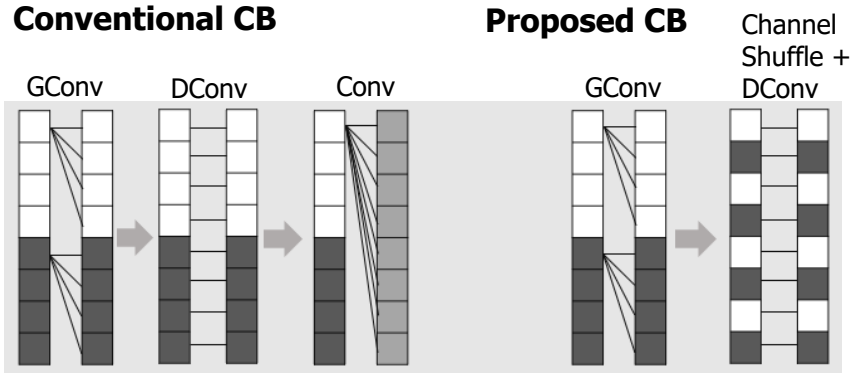
Overall framework



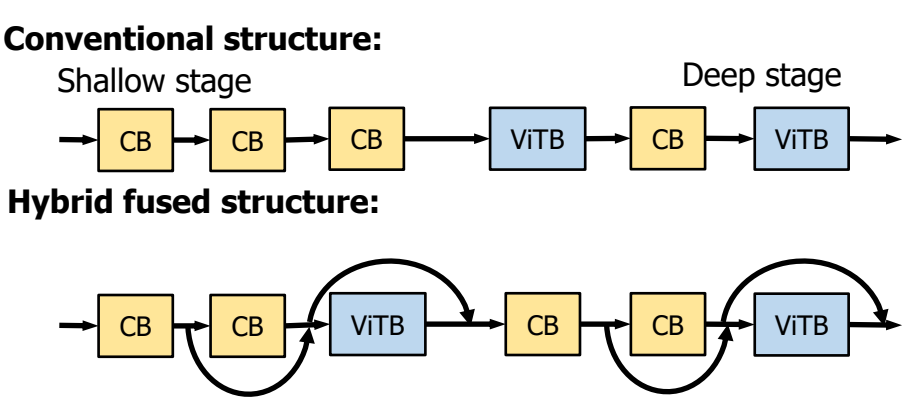
P1: Window Shifted Big kernel Depthwise Attention



P2: Channel-shuffled Separable Convolution



P3: Hybrid fused structure of ViT and convolution blocks



Experimental Results

FLOPs: floating point operations

Model	FLOPs (M)	Layers	Accuracy (%)	Inference Time on GPU 1080Ti (ms)	Inference Time on CPU I7 6700k (ms)
Original MobileViT	64	63	70.18	5.51	5.67
Light MobileViT	51	59	70.35	5.15	5.19
Conv2Former	45	55	70.16	4.73	4.76
+P1	40	45	69.13	3.98	4.28
+P1+P2	36	38	67.95	3.40	3.64
+P1+P2+P3	45	38	70.92	3.42	3.81

Conclusion

- **This work aims to speed up MobileViT while keeping accuracy**
- **Proposed MobileViT speeds up 38%, while accuracy is 0.74% higher than original MobileViT**

Graduate School of Information, Production and Systems
Waseda University