Drop an Octave: Reducing Spatial Redundancy in Convolutional Neural Networks with Octave Convolution

Yunpeng Chen et al.

Facebook AI, NUS, Qihoo360AI

https://arxiv.org/pdf/1904.05049.pdf

Presenter: Weilin Xu
`https://qdata.github.io/deep2Read`

Outline

1. Motivation
2. Method
3. Experiments
4. Conclusion

Motivation

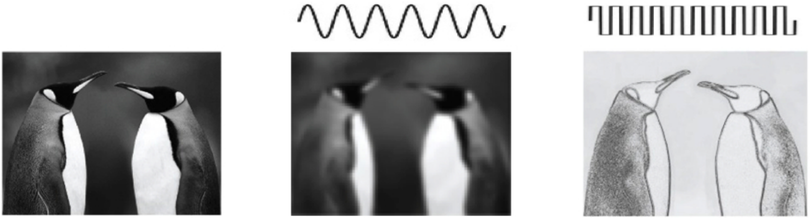
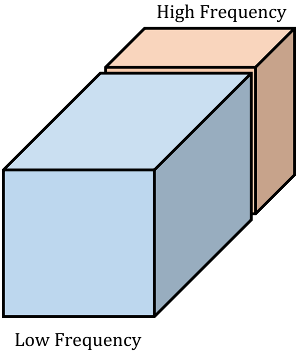
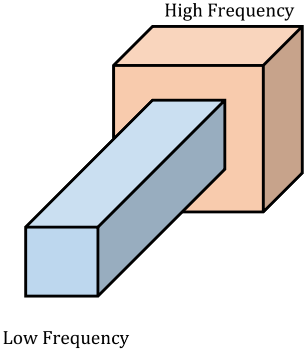


Figure: Decompose an natural image into a low and a high spatial frequency part.

Octave Feature Representation



(a) Group output maps by their spatial frequency.



(b) Multi-frequency feature representation, reducing space redundancy.

Scale-space Theory

- Principled way of creating scale-spaces of spatial resolutions.
- **Octave**: a division of the spatial dimensions by a power of 2 (only 2^1 in this work).
- **Low-frequency space**: reducing the spatial resolution of the feature maps by an octave.

Octave Convolution

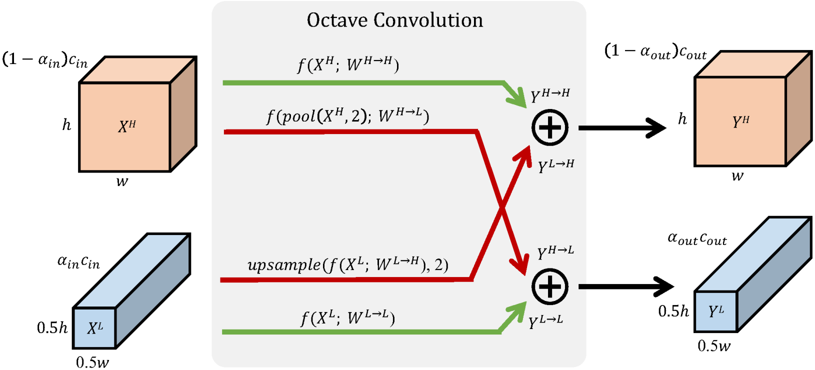


Figure: Detailed design of the Octave Convolution. Green arrows correspond to information updates while red arrows facilitate information exchange between the two frequencies.

Hyper-parameter: α

Table: Relative theoretical gains for the proposed multi-frequency feature representation over vanilla feature maps for varying choices of the ratio α of channels used by the low-frequency feature. When $\alpha = 0$, no low-frequency feature is used which is the case of vanilla convolution. Note the number of parameters in OctConv operator is constant regardless of the choice of ratio.

ratio (α)	.0	.125	.25	.50	.75	.875	1.0
#FLOPs Cost	100%	82%	67%	44%	30%	26%	25%
Memory Cost	100%	91%	81%	63%	44%	35%	25%

Integrate OctConv into Backbone Networks

- At the first OctConv layer: $\alpha_{in} = 0$, $\alpha_{out} = \alpha$
  Disable low-frequency input, only two paths.
- At the last OctConv layer: $\alpha_{out} = 0$
  Disable low-frequency output, single full resolution output.

Octave Convolution Kernel

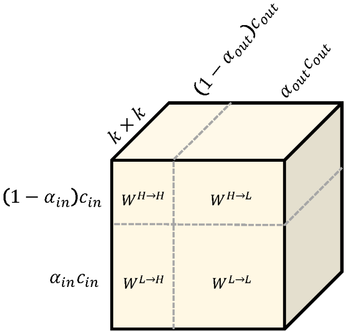


Figure: The Octave Convolution kernel. The $k \times k$ Octave Convolution kernel $W \in \mathbb{R}^{c_{in} \times c_{out} \times k \times k}$ is equivalent to the vanilla convolution kernel in the sense that the two have the exact same number of parameters.

Octave Convolution Kernel

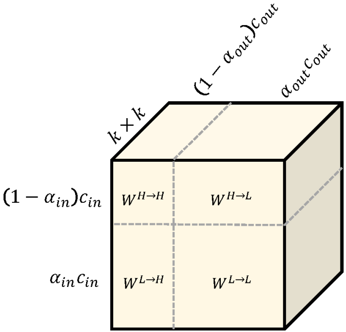


Figure: The Octave Convolution kernel. The $k \times k$ Octave Convolution kernel $W \in \mathbb{R}^{c_{in} \times c_{out} \times k \times k}$ is equivalent to the vanilla convolution kernel in the sense that the two have the exact same number of parameters.

Larger receptive field for low-frequency input due to the same kernel size.

Octave Convolution Implementation

$$
\begin{aligned}
Y^H &= f(X^H; W^{H\to H}) + \texttt{upsample}(f(X^L; W^{L\to H}), 2) \\
Y^L &= f(X^L; W^{L\to L}) + f(\texttt{pool}(X^H, 2); W^{H\to L})),
\end{aligned} \tag{1}
$$

- $f(X; W)$: convolution with parameters W
- $\texttt{pool}(X, k)$: average pooling operation with kernel size $k \times k$ and stride k.
- $\texttt{upsample}(X, k)$: up-sampling operation by a factor of k via nearest interpolation.

HF Output in Octave Convolution

Take low-frequency input with up-sampling.

$$
\begin{aligned}
Y_{p,q}^{H} =& Y_{p,q}^{H\to H} + Y_{p,q}^{L\to H} \\
=& \sum_{i,j\in\mathcal{N}_k} {W_{i+\frac{k-1}{2},j+\frac{k-1}{2}}^{H\to H}}^{\top} X_{p+i,q+j}^{H} \\
&+ \sum_{i,j\in\mathcal{N}_k} {W_{i+\frac{k-1}{2},j+\frac{k-1}{2}}^{L\to H}}^{\top} X_{(\lfloor\frac{p}{2}\rfloor+i),(\lfloor\frac{q}{2}\rfloor+j)}^{L},
\end{aligned} \tag{2}
$$

LF Output in Octave Convolution

Take high-frequency input with down-sampling (average pooling).

$$
\begin{aligned}
Y^{L}_{p,q} =& Y^{L\to L}_{p,q} + Y^{H\to L}_{p,q} \\
=& \sum_{i,j\in\mathcal{N}_k} {W^{L\to L}_{i+\frac{k-1}{2},j+\frac{k-1}{2}}}^{\top} X^{L}_{p+i,q+j} \\
&+ \sum_{i,j\in\mathcal{N}_k} {W^{H\to L}_{i+\frac{k-1}{2},j+\frac{k-1}{2}}}^{\top} X^{H}_{(2*p+0.5+i),(2*q+0.5+j)},
\end{aligned} \tag{3}
$$

Ablation study

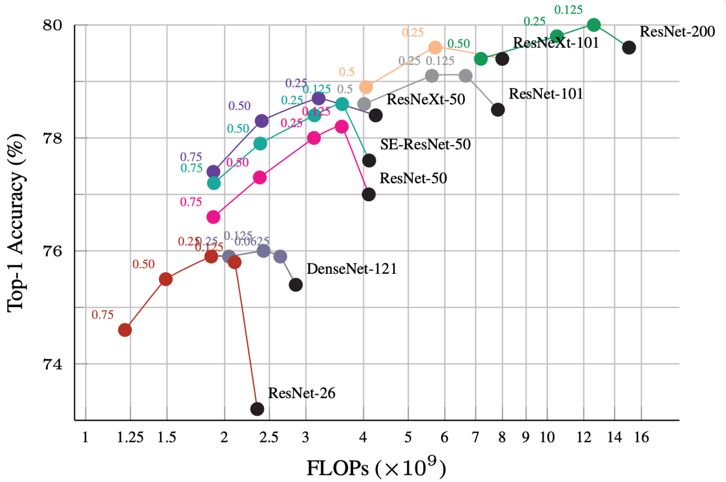


Figure: Ablation study results on ImageNet. OctConv-equipped models are more efficient and accurate than baseline models. Markers in black in each line denote the corresponding baseline models without OctConv. The colored numbers are the ratio α. Numbers in X axis denote FLOPs in logarithmic scale.

Conclusion

- Octave Convolution: Reduce spatial redundancy by separating low- and high-frequency features.
- Replace regular convolution in-place.
- Improve classification performance and reduce computational cost.

Discussions

- Simple and effective method.
- Adversarial implication.