Drop an Octave: Reducing Spatial Redundancy in Convolutional Neural Networks with Octave Convolution [1]

陈威宇

2023-12-10

摘要

在自然图像中，信息以不同的频率传达，其中较高的频率通常编码了细节，而较低的频率通常编码了全局结构。同样，卷积层的输出特征图也可以看作是在不同频率上混合的信息。在这项工作中，我们提出通过它们的频率对混合特征图进行因式分解，并设计了一种新颖的八度卷积（OctConv）操作，以存储和处理在较低空间分辨率下空间变化“较慢”的特征图，从而降低内存和计算成本。与现有的多尺度方法不同，OctConv 被制定为单一的、通用的、即插即用的卷积单元，可以直接替代（普通的）卷积，而无需调整网络架构。它还与建议更好的拓扑结构或减少通道冗余的方法（如组卷积或深度可分离卷积）正交且互补。我们通过实验证明，通过简单地用 OctConv 替换卷积，我们可以在图像和视频识别任务的准确性上保持一致提升，同时降低内存和计算成本。配备 OctConv 的 ResNet-152 在 ImageNet 上可以实现 82.9% 的 top-1 分类准确性，仅需 22.2 GFLOPs。

关键词： OctConv；ImageNet；减少冗余；不同频率.

1 引言

卷积神经网络（CNNs）的效率随着近期减少密集模型参数 [4, 13, 17] 和特征图通道维度中固有冗余的努力而不断提高 [2, 3, 6, 19]。然而，在 CNNs 生成的特征图的空间维度中也存在实质性的冗余，其中每个位置独立存储其自己的特征描述符，同时忽略了相邻位置之间可能共同存储和处理的公共信息。为了解决该难题，OctConv 主张接收包含两个频率张量的特征图，这两个特征频率相差一个八度。OctConv 直接从低频图中提取信息，无需将其解码回高频再运算。

如图1所示，在自然图像中，图像信息可以被分解为高频信息与低频信息，其中高频信息是图像中空间变化剧烈的内容，低频信息则是空间变化缓慢的内容。同样的，我们也可以将深度学习中得到的特征图分解为高频信息与低频信息，OctConv 便是一种探究分解整合不同频率信息的一种新形卷积。

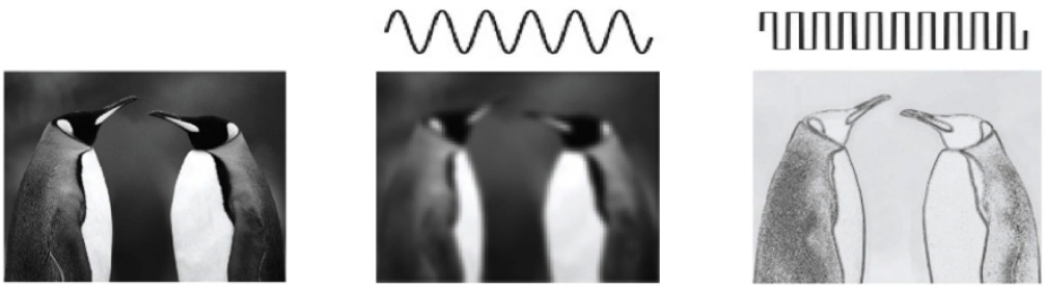


图 1. 图像的低频分量和高频分量

作为普通卷积的替代，OctConv 消耗的内存和计算资源大大减少。此外，OctConv 通过相应的（低频率的）卷积处理低频信息，有效地扩大了原始像素空间中的感知域，从而可以提高识别性能。

与利用多尺度信息的方法不同，OctConv 可以轻松部署为即插即用的单元，替换卷积，无需更改网络架构或进行超参数调整。由于 OctConv 主要专注于处理多个空间频率上的特征图并减少它们的空间冗余，因此它与现有方法相辅相成，这些方法专注于构建更好的 CNN 拓扑结构，减少卷积特征图中的通道冗余以及减少密集模型参数的冗余。此外，与密切相关的多网格卷积 [9] 相比，OctConv 基于频率模型提供了更多有关减少 CNN 中空间冗余的见解，并采用更有效的频率信息交换策略，性能更佳。

2 相关工作

2.1 传统 CNN

自 AlexNet [10] 和 VGG [16] 的开创性工作以来，研究人员已经付出了大量努力来提高 CNN 的效率。ResNet [5, 8] 和 DenseNet [7] 通过在早期层添加快捷连接来改进网络拓扑。ResNeXt [18] 和 ShuffleNet [20] 使用稀疏连接的组卷积来减少通道间的冗余。Xception [3] 和 MobileNet [6, 15] 采用深度卷积进一步减少连接密度。与此同时，NAS [21]、PNAS [12] 和 AmoebaNet [14] 提出了原子级地为给定任务找到最佳网络拓扑的方法。剪枝方法，如 DSD [4] 和 ThiNet [13]，专注于通过消除 CNN 中最不重要的权重或连接来减少模型参数中的冗余。然而，所有这些方法都忽视了特征图的空间维度上的冗余，而这正是提出的 OctConv 所解决的。

2.2 多尺度表征学习

在深度学习时代，多尺度表示也因其强大的鲁棒性和泛化能力而发挥着重要作用。但是，当将它们应用于 ResNet 之外的架构，如 MobileNetV1 [6]、DenseNet [7] 等时，仍然需要额外的专业知识和超参数调整。Multi-grid CNNs [9] 提出了一个多网格金字塔特征表示，并将 MG-Conv 运算符定义为卷积运算符的替代品，从概念上与我们的方法类似，但其动机是利用多尺度特征。与 MG-Conv 相比，OctConv 采用更高效的设计来交换不同频率之间的信息，并具有更高的性能。简而言之，OctConv 专注于减少 CNN 中的空间冗余，并被设计为替代普通卷积操作，而无需调整主干的 CNN 架构。

3 本文方法

3.1 方法概述

为了减少空间冗余，我们引入了八度特征表示法，将特征图张量明确分解为对应于低和高频率的组。尺度空间理论 [11] 为我们提供了一种按比例缩减空间分辨率的合理方式，并将一个八度定义为通过 2 的幂次对空间维度进行划分（在这项工作中，我们仅研究 2^1）。我们遵循这一模式，将低频特征图的空间分辨率降低了一个八度。

我们设计的目标是在相应的频率张量中有效处理低频和高频，并实现高效的频际通信。设 X、Y 分别为分解的输入和输出张量。那么输出 $Y = \{Y^H, Y^L\}$ 的高频和低频特征图将分别由 $Y^H = Y^{H\to H} + Y^{L\to H}$ 和 $Y^L = Y^{L\to L} + Y^{H\to L}$ 给出，其中 $Y^{A\to B}$ 表示从特征组 A 更新到组 B 的卷积更新。具体而言，$Y^{H\to H}$ $Y^{L\to L}$ 表示同频更新，而 $Y^{H\to L}$、$Y^{L\to H}$ 表示跨频通信。

3.2 OctConv 的特征表示

$X \in \mathbb{R}^{c\times h\times w}$ 表示卷积层的输入特征张量，其中 h 和 w 表示空间维度，c 表示特征图或通道的数量。我们明确地沿着通道维度将 X 因子化为 $X = \{X^H, X^L\}$，其中高频特征 $X^H \in \mathbb{R}^{(1-\alpha)c\times h\times w}$，捕捉细节，而低频特征图 $X^L \in \mathbb{R}^{\alpha\times c\times \frac{h}{2}\times \frac{w}{2}}$，在空间维度上变化较慢（相对于图像位置）。这里 $\alpha \in [0, 1]$ 表示分配给低频部分的通道比例，低频特征图的空间分辨率是高频特征图的一半，即在空间分辨率的一半处。

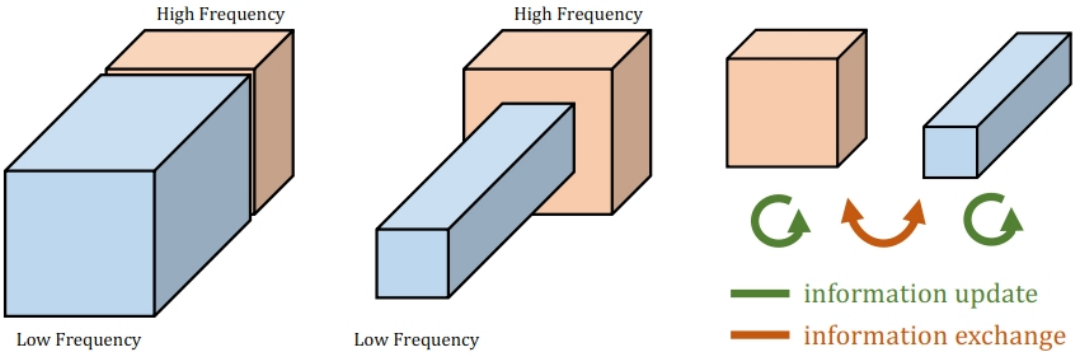


图 2. 方法示意图

如图2所示。如此，八度特征表示减少了空间冗余并比原始表示更紧凑。然而，由于输入特征在空间分辨率上存在差异，普通的卷积无法直接在这种表示上操作。一种绕过这个问题的朴素方法是将低频部分 X^L 上采样到原始空间分辨率，与 X^H 连接，然后进行卷积，但这样做会导致额外的计算和内存成本，并削弱来自压缩的所有节省。为了充分利用我们的紧凑多频特征表示，我们引入了八度卷积，它可以直接在分解的张量 $X = \{X^L, X^H\}$ 上进行操作，降低了计算开销。

3.3 实现细节

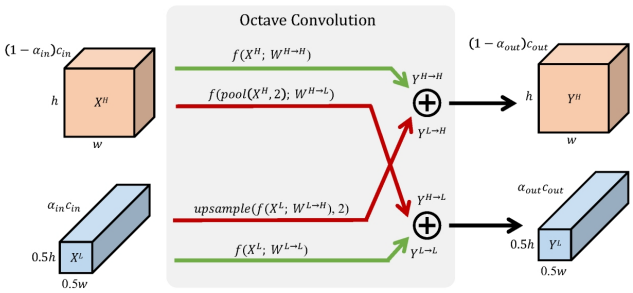


图 3. OctConv 实现细节

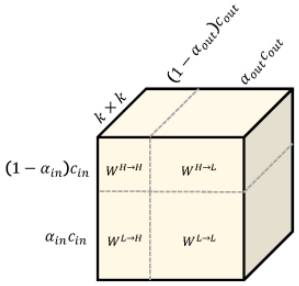


图 4. OctConv 的卷积核

OctConv 的实现细节如图3所示。它由四条计算路径组成，分别对应公式1和2中的四个项：两条绿色路径对应高频和低频特征图的信息更新，两条红色路径促进两个频率之间的信息交换。如下图所示，低频和高频的输入经过八度卷积操作得到了低频和高频的输出。

除此之外，如何将 OctConv 集成到主干网络中也是一个重要的问题。OctConv 与传统的卷积兼容，无需特殊调整即可插入常规的卷积网络。要将常规特征表示转换为八度特征表示，需要在第一个 OctConv 层，我们设置 $\alpha_{in} = 0$ 和 $\alpha_{out} = \alpha$。为了将多频特征表示转换回常规特征表示，即在最后一个 OctConv 层，我们设置 $\alpha_{out} = \alpha$ 和 $\alpha_{out} = 0$。而在中间层中，我们设置 $\alpha_{in} = \alpha_{out} = \alpha$。

如图4所示，为计算这些项，我们将卷积核 W 分成两个分量 $W = [W^H, W^L]$，分别用于与 X^H 和 X^L 进行卷积。每个分量可以进一步分为频内和频间两部分：$W^H = [W^{H\to H}, W^{L\to H}]$ 和 $W^L = [W^{L\to L}, W^{H\to L}]$，并使用平均池化表达 OctConv 的输出 $Y = \{Y^H, Y^L\}$，进行下采样，Y^H 的表达式如下所示：

$$Y^H = f(X^H; W^{H\to H}) + upsample(f(X^L; W^{L\to H}), 2) \tag{1}$$

Y^L 的表达式如下所示：

$$Y^L = f(X^L; W^{L\to L}) + f(pool(X^H, 2); W^{H\to L}) \tag{2}$$

其中，$f(X; W)$ 表示具有参数 W 的卷积，$pool(X, k)$ 是一个核大小为 $k \times k$ 且步幅为 k 的平均池化操作。$upsample(X, k)$ 是通过最近邻插值进行 k 倍上采样的操作。

4 复现细节

4.1 与已有开源代码对比

此工作已有开源代码 (https://github.com/facebookresearch/OctConv).

此工作已有的开源代码是使用的框架为 MXNet，但是研究组内的代码大都基于 Pytorch，因此本文将参考其余非官方的基于 Pytorch 的复现来重构一个 Pytorch 框架下的 OctConv 并进行后续的训练测试；此外，本文尝试了不同 α 的 OctConv，并对其进行评估；最后，囿于算力资源，本次复现无法与原论文中使用相同的数据集-ImageNet，为了评估 OctConv，本文

将数据集更改为更小体积的 CIFAR-10(仅含有 10 个类，每个类有 6000 张图片，每张图片的尺寸为 32×32).

4.2 实验环境

在本文的复现中所使用的实验环境如下：

- CPU: Intel(R) Xeon(R) CPU E5-2686 v4 @ 2.30GHz
- GPU: NVIDIA RTX A6000 48G
- OS: Linux 5.4.0-149-generic 166-Ubuntu
- python 3.8
- 框架: Pytorch
- 包管理: Anaconda3

4.3 创新点

综上，本文的贡献可以简单概括为以下几点：

- 将原文的开源代码由 MXNet 框架改写为 Pytorch 框架，重写的代码可读性更高。
- 使用不同的 α 对 OctConv 进行评估，扩充了实验的完整性。
- 囿于算力资源，将 ImageNet 数据集更换为了小体积的 CIFAR-10 数据集并成功验证。

5 实验结果分析

首先是原论文的消融实验，如图5所示，与基线模型相比，配备 OctConv 的模型更高效、更准确。

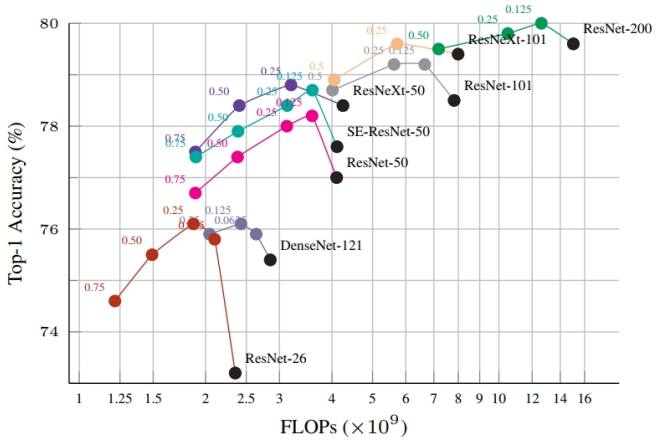


图 5. 消融实验

表 1. OctConv 在轻量级模型上的表现

Method	ratio(α)	Params (M)	FLOPs (M)	CPU (ms)	Top-1 (%)
CondenseNet (G = C = 8)	-	2.9	274	-	71.0
ShuffleNet (v1)	-	3.4	292	-	71.5
1.5 ShuffleNet (v2)	-	3.5	299	-	**72.6**
0.75 MobileNet (v1)	-	2.6	325	13.4	70.3*
0.75 Oct-MobileNet (v1) (ours)	.375	2.6	**213**	**11.9**	70.5
1.0 Oct-MobileNet (v1) (ours)	.5	4.2	321	18.4	**72.5**
1.0 MobileNet (v2)	-	3.5	300	24.5	72.0
1.0 Oct-MobileNet (v2) (ours)	.375	3.5	**256**	**17.1**	72.0
1.125 Oct-MobileNet (v2) (ours)	.5	4.2	295	26.3	**73.0**

在 OctConv 的评估中，我们采用最流行的轻量级网络作为基线，研究 OctConv 是否能在这些具有深度卷积功能的紧凑型网络上运行良好。特别是，我们使用了"0.75 MobileNet (v1)"和"1.0MobileNet (v2)" 。和"1.0MobileNet (v2) " 作为基线。作为基线，用我们提出的 OctConv 代替常规卷积。

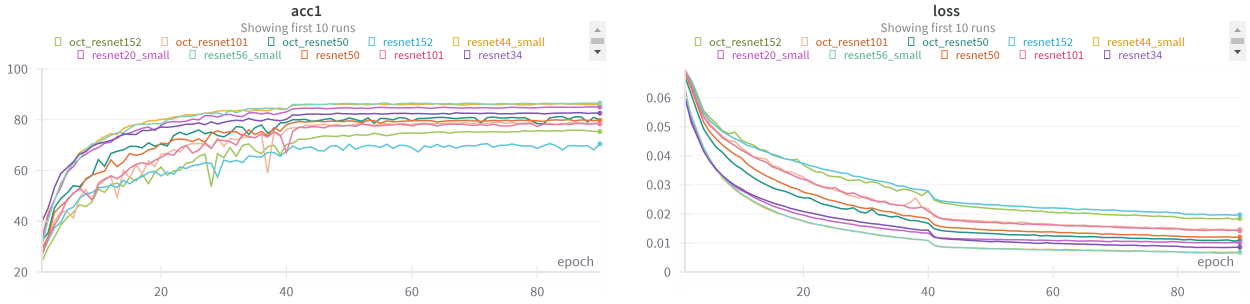
表1展示了 OctConv 在轻量级模型上的表现，可以从表中得出这样一个结论：OctConv 可以将 MobileNetV1 的 FLOPs 减少 34%，并在实际应用中提供更好的精度和更快的速度；它可以将 MobileNetV2 的 FLOPs 减少 15%，实现相同的精度和更快的速度。由于 OctConv 可以补偿额外的计算成本，因此可以采用更宽的模型来提高学习能力。特别是，配备了 OctConv 的网络在相同 FLOPs 条件下比 MobileNetV1 提高了 2%，比 MobileNetV2 提高了 1%。

受限于篇幅，本报告只展示部分原论文的评估工作。接下来是根据原论文，我们所做的复现工作评估。

5.1 实验评估

由于算力资源有限，在复现工作中和原论文一样使用 ImageNet 来训练评估是不现实的，故本文将数据集更换为了体积更小的 CIFAR-10 数据集，该数据集仅含 10 个类别，每个类别含有 6000 张图片，每张图片的尺寸为 35×35。不同于原文，本工作分别在 resnet18 [8], resnet34, resnet34, resnet50, resnet101, resnet152, resnet20-small, resnet44-small, resnet56-small 和配备了 OctConv 的 resnet50, resnet101, resnet152 上进行实验，实验的参数设置为：

- $\alpha = 0.125$
- Epoch=90
- Learning Rate=1e-3
- BatchSize=64



(a) Top-1(%) 变化曲线　　(b) loss 变化曲线

如图6a和6b所示，分别是训练过程的 Top-1(%) 准确率和 loss 的变化曲线。由此可以看出，配备了 OctConv 的网络与配备了传统卷积的并无太大差异，只是在运行时间和准确率上有所提升。

表 2. 我的复现结果

Model	Top-1(%)	Top-5(%)
resnet18	82.80	99.08
resnet34	82.58	98.9
resnet50	79.80	98.89
resnet101	78.35	98.53
resnet152	70.52	97.22
resnet20(small)	85.02	99.34
resnet44(small)	86.05	99.48
resnet56(small)	86.56	99.41
Oct-resnet50	**79.82**	**99.05**
Oct-resnet101	**78.86**	**98.85**
Oct-resnet152	**75.35**	**98.47**

如表2所示，当我们更换了不同于原论文的 ImageNet 数据集，而是使用 CIFAR-10 数据集时，配备了 OctConv 的 resnet50，resnet101，resnet152 的准确率仍比配备了传统卷积的要优。在 Top-1(%) 准确率上，分别提高了 0.02，0.51，4.83；在 Top-5(%) 准确率上，分别提升了 0.16，0.32，1.25。

可以注意到，不同于原文的结果，即模型参数越大时准确率越高，在我们复现的结果中，反而是参数较小的模型表现较好。我们推测，这可能是更换了 CiFAR-10 导致的，该数据集仅含 60000 张图片，当与原论文一样训练 90 个 Epoch 时，可能在某个 Epoch 中就已经过拟合。而找出模型过拟合的点，这正是我们在后续需要改进的地方。

除此之外，原论文中评估了每个模型的运行时间，限于现有条件，本次复现中没有开展此项实验，希望能够在后续的科研学习中继续完善。

5.2 参数调优

在原论文中只展示了 $\alpha = 0.125$ 时八度卷积的表现性能，而 α 为其他值时并没有详细评估。本着严谨的求真态度，在本次复现中将 α 分别更改为 0.25 和 0.5 继续在配备了 OctConv 的 resnet50，resnet101，resnet152 上进行实验，实验结果如表3所示。

表 3. 调优结果

α \ model	Oct-resnet50	Oct-resnet101	Oct-resnet152
0.125	**79.82**	**78.86**	**75.35**
0.25	78.76	76.45	72.92
0.5	76.18	73.63	67.26

可以看到，随着 α 的增大，模型的表现逐渐下降，相比之下还是原论文所使用的 0.125 性能最优。在自然图像中，高频信息是图像中空间变化剧烈的内容，低频信息则是空间变化缓慢的内容。如此，我们可以得知，图像的低频分量所含分量仅占原图像信息的 1/8，而那些急剧变化的像素包含了剩下 7/8 的信息。

6 总结与展望

6.1 总结

在我的复现工作中，成功验证了 OctConv 的有效性。接下来谈一谈我对八度卷积的看法：OctConv 操作类似于神经网络的中的双流网络设计，并且将双流网络的思想运用到了极致，取得了不错的效果。在论文解释上，文中采用了频率信息整合的解释，具有一定的启发意义。但是不足之处在于，作者没有结合数字图像处理，对 OctConv 得到的特征图进行频域分析，有些立意高而内容不充实。当然在神经网络作为一种无法解释的黑盒的今天，一种架构可以有不同解释也是可以理解。

在这次的复现工作中，相较于原论文，我做出了如下的改动：

- 原论文的开源代码是基于 MXNet 实现的，将其重构为 Pytroch 框架。
- 囿于算力资源，将 ImageNet 数据集更换为了小体积的 CIFAR-10 数据集并成功验证。
- 实验了 $\alpha = 0.25, 0.5$ 时的八度卷积的表现。

与此同时，我的工作也存在不足的地方：

- 算力资源紧张，所提及的结果没有进行多次实验取平均。
- 原论文中使用 ImageNet，没有完整地复现出原论文的工作。

6.2 展望

- 在 ImageNet 数据集下验证配备了 OctConv 的有效性。
- 在 CIFAR-10 数据集下，找出模型过拟合的点。
- 尝试更改网络结构，得到一个更优的八度卷积结构。
- 网格化搜索，找到适合不同任务的最佳超参 α。
- 完善更改过的开源代码，并将其上传至开源社区。
- 将报告详细化并上传开源社区，为后来者复现提供解决方案。

参考文献

[1] Yunpeng Chen, Haoqi Fan, Bing Xu, Zhicheng Yan, Yannis Kalantidis, Marcus Rohrbach, Shuicheng Yan, and Jiashi Feng. Drop an octave: Reducing spatial redundancy in convolutional neural networks with octave convolution. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3435–3444, 2019.

[2] Yunpeng Chen, Yannis Kalantidis, Jianshu Li, Shuicheng Yan, and Jiashi Feng. Multi-fiber networks for video recognition. In *Proceedings of the european conference on computer vision (ECCV)*, pages 352–367, 2018.

[3] C Fran et al. Deep learning with depth wise separable convolutions. In *IEEE conference on computer vision and pattern recognition (CVPR)*, 2017.

[4] Song Han, Jeff Pool, Sharan Narang, Huizi Mao, Enhao Gong, Shijian Tang, Erich Elsen, Peter Vajda, Manohar Paluri, John Tran, et al. Dsd: Dense-sparse-dense training for deep neural networks. *arXiv preprint arXiv:1607.04381*, 2016.

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 630–645. Springer, 2016.

[6] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

[7] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[8] S Jian, H Kaiming, R Shaoqing, and Z Xiangyu. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision & Pattern Recognition*, pages 770–778, 2016.

[9] Tsung-Wei Ke, Michael Maire, and Stella X Yu. Multigrid neural architectures. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6665–6673, 2017.

[10] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

[11] Tony Lindeberg. *Scale-space theory in computer vision*, volume 256. Springer Science & Business Media, 2013.

[12] Chenxi Liu, Barret Zoph, Maxim Neumann, Jonathon Shlens, Wei Hua, Li-Jia Li, Li Fei-Fei, Alan Yuille, Jonathan Huang, and Kevin Murphy. Progressive neural architecture search. In *Proceedings of the European conference on computer vision (ECCV)*, pages 19–34, 2018.

[13] Jian-Hao Luo, Hao Zhang, Hong-Yu Zhou, Chen-Wei Xie, Jianxin Wu, and Weiyao Lin. Thinet: pruning cnn filters for a thinner net. *IEEE transactions on pattern analysis and machine intelligence*, 41(10):2525–2538, 2018.

[14] Esteban Real, Alok Aggarwal, Yanping Huang, and Quoc V Le. Regularized evolution for image classifier architecture search. In *Proceedings of the aaai conference on artificial intelligence*, volume 33, pages 4780–4789, 2019.

[15] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018.

[16] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[17] Frederick Tung and Greg Mori. Clip-q: Deep network compression learning by in-parallel pruning-quantization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7873–7882, 2018.

[18] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017.

[19] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

[20] Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6848–6856, 2018.

[21] Barret Zoph, Vijay Vasudevan, Jonathon Shlens, and Quoc V Le. Learning transferable architectures for scalable image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8697–8710, 2018.