湖北工業大學

英文文献翻译

译文出处： Andrew Howard and et al.
Searching for MobileNetV3.
CoRR, 2019, abs/1905.02244:1-3

姓　　名　姓名

学　　号　学号

所在学院　所在学院

专业班级　专业班级

指导教师　指导教师

2020 年 03 月

译文要求

一、文献内容须与课题（或专业）相关，并在封面注明详细出处。

二、出处格式

图书：作者. 书名. 出版地：出版者，出版年：起止页.

期刊：作者. 文章名称. 期刊名称，年号，卷号（期号）：起止页.

三、译文不少于 3000 字。

四、文献翻译封面和译文要求部分采用 A4 号纸双面打印，其余部分均单面打印，顺序依次为封面、译文、英文文献。

对于 MobileNetv3 的探索

概述

我们推出了新一代的 MobileNet 网络，该网络基于一种新颖的补充搜索技术的组合。MobileNetV3 针对手机 CPU 进行优化，通过使用硬件感知网络结构搜索 (NAS) 同时利用 NetAdapt 算法进行补充，新颖的网络结构发展也对其产生影响。文章首先探索了如何将自动搜索算法与网络设计互补工作从而推进 SOTA 的高度。通过此研究，我们设计了两个新的 MobileNet 网络——MobileNetV3-Large 与 MobileNetV3-Small——这两种网络分别针对于高计算资源环境与低计算资源环境。这两种网络随后被适配与应用于物体检测与语义分割。在语义分割（或者任何密集像素预测）任务中，我们提出了一种新的高效分割解码器 LR-ASPP。移动端平台上，我们在分类、检测与分割应用上达到了 SOTA。相比于 MobileNetV2，MobileNetV3-Large 在 ImageNet 分类中提升了 3.2% 的正确率同时减小了 20% 延迟；而 MobileNetV3-Small 在相同延迟的情况下提升了 6.6% 的正确率。在 COCO 数据集上，MobileNetV3-Large 在与 MobileNetV2 正确率相同的情况下可以达到 25% 的速度提升。在城市景观分割中，MobileNetV3-Large LR-ASPP 在与 MobileNetV2 R-ASPP 相同正确率的情况下可以有 34% 的速度提升。

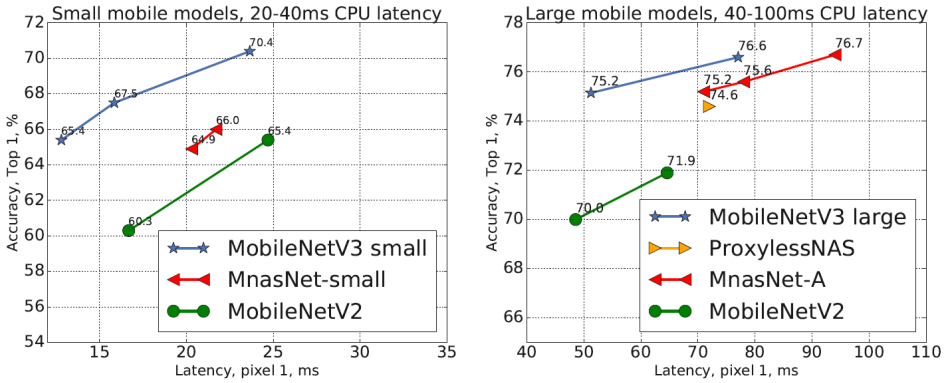

图 1. 在 Pixel1 手机上的延迟与 ImageNet top-1 正确率的权衡。所有的模型输入分辨率皆为 224.V3 large 和 V3 small 使用 0.75，1 和 1.25 的缩放倍数来展示最优边界。所有模型皆利用 TFLite 框架在单个大核心上运算。MobileNetV3-Small&Large 是我们推出的下一代移动端模型。

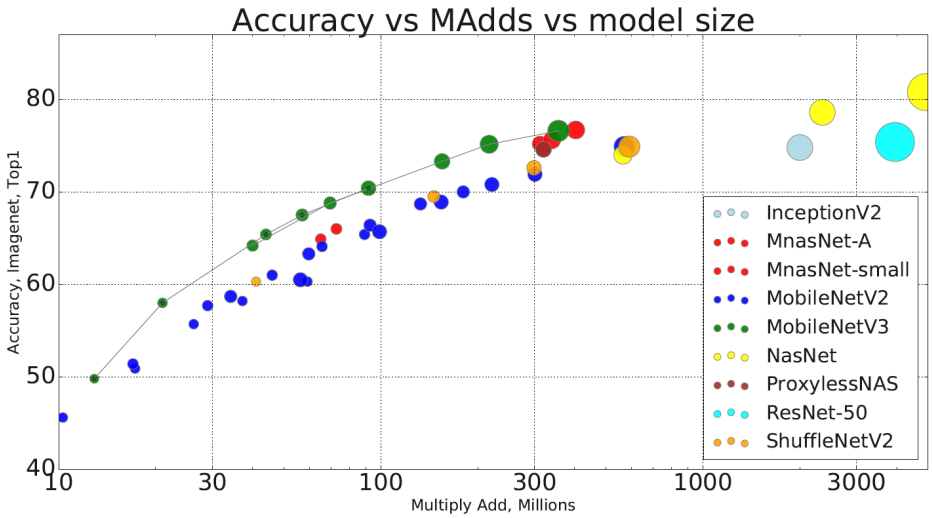

图 2. Madds 和 ImageNet top-1 正确率的权衡。这个比较允许模型运行在不同的软件框架与硬件平台上。所有的 MobileNetV3 输入分辨率为 224，使用 0.35，0.5，0.75，1 和 1.25 的缩放倍率。其他情况可以在第六章节看到。

1 介绍

在移动应用上开始普及的高效神经网络使得移动端的体验有了极大的改变。本地部署使得用户无需将自身隐私上传至服务器就可获得神经网络带来的益处。高效的神经网络不仅通过高正确率与低延迟带给用户体验的提升，同时减少了电能的损耗从而使移动设备获得更好的续航能力。

本文阐述了我们为了提升移动端机器视觉新能而推出的下一代高正确率、高效率神经网络模型，即 MobileNetV3-Large 与 MobileNetV3-Small 两种网络模型，的设计过程。这两种模型推动 SOTA 达到了新的高度，并且展示了如何将自动搜索算法与新的网络结构设计混合工作来设计有效的网络模型。

本文的目的是为了探索一种在移动设备上尽可能平衡正确率与延迟的移动端机器视觉网路模型。为了达到这个目标，我们引入了以下四种技术:

- 互补搜索技术
- 适用于移动设备的新型高效非线性模型
- 新的高效网络设计
- 新的高效分割解码器

我们通过实验，证明了以上几种技术在广泛的用途和手机上的价值和有效性。

论文结构如下。我们从第二部分对有关研究的讨论开始。第三部分回顾了

适用于移动设备的高效网络模型。第四部分回顾了结构搜索以及 MnasNet 和 NetAdapt 算法的互补性。第五部分阐述了新的网络结构设计通过联合搜索提升了模型搜索的效率。第六部分呈献了大量分类、检测与分割的实验数据，展示了模型的有效性与各个模块的功效。第七部分包括总结和今后的工作。

2 相关研究

近些年，设计一个在正确率与有效性之间平衡的 DNN 结构已经成为一个热点研究领域。对于推进此领域，手工设计结构与 NAS 算法都发挥了重要的作用。

SqueezeNet[22] 为了压缩参数广泛的使用了 1x1 卷积核与挤压-扩张模型。近期更多的模型将关注点从减少参数转变到减少计算量（MAdds）以及实际的延迟上。MobileNetV1[19] 引入的深度可分离卷积从实质上提升了计算效率。MobileNetV2[39] 在此之上引入了一个由翻转残差结构和线性瓶颈的高效资源块。ShuffleNet[49] 利用分组卷积与通道乱序进一步减少了 MAdds。CondenseNet[21] 在训练阶段学习分组卷积以致于维持层间有效连接从而使特征再利用。ShiftNet[46] 提出了将移位操作插入点卷积从而减少昂贵的空间卷积。

为了自动化设计网络结构，首先引入强化学习（RL）来寻找具有竞争性正确率的高效网络结构。一个完全可配置的搜索空间将会成指数增长且较难处理。所以结构搜索的前期工作着眼于单元层级的结构搜索。近期，[43] 探索了一种结构块层级分层搜索空间，该空间允许在一个网络中存在不同结构块，每个结构块中的分层结构也可不尽相同。为了减少搜索的计算量，在 [28,5,45] 中使用的可微结构搜索框架进行了梯度优化。而在通过调整现有网络以适应移动平台，[48,15,12] 提出了更有效的自动网络简化算法。

通过降低计算精度从而量化网络 [23,25,47,41,51,52,37] 也是增加网络效率的重要的补充研究方向。最后，知识蒸馏提供了一种从大型网络指导生成小型网络的附加补充方法。

3 高效的移动网络构建块

移动端模型建立在越来越多的更高效的基础之上。MobileNetV1 引入的深度可分离卷积是传统卷积层的有效替代。深度可分离卷积将空间滤波和特征生成机制分离，有效的分解了传统的卷积。深度可分离卷积有两层独立的层组成：较轻量的深度卷积负责空间滤波，计算量较大的 1x1 点卷积负责特征生成。

MobileNetV2[39] 引入了线性瓶颈和翻转的残差结构，目的是利用输入的低层性质使层结构更加有效。这个结构如图 3所示，由深度可分离卷积，一层 1x1 扩充卷积和一层 1x1 的投影卷积组成。瓶颈结构的输入和输出在当通道数相同时将由残差结构相连。该结构在输入和输出的表现形式紧凑的同时，在内部扩充到一个高维度的特征空间，增强了每个通道非线性转换的表达能力。

MnasNet[43] 在 MobileNetV2 的结构之上，在瓶颈块中引入了基于挤压和激励的轻量级注意模型。请注意，此模型中 SE 插入的位与 [20] 中基于 ResNet 中提出的不同。如图 4所示，该模型将 SE 置于深度可分离卷积层后的扩增空间中以便应用于最大的表示。

在 MobileNetV3 中，为了构建出最有效的模型，我们将上述模型组合成构建块。每层的非线性激活函数也升级为改进过的 swish。SE 结构和 swish 激活函数中都使用了 sigmoid 函数，但 sigmoid 函数很难在定点计算中保持精度，所系我们将其替换为 hard sigmoid，此部分我们将在 5.2 章节讨论。

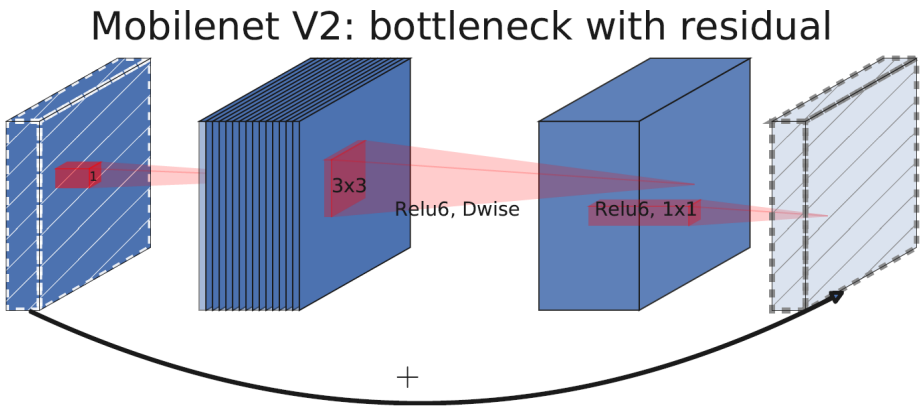

图 3. MobileNetV2[39] 逆残差和线性瓶颈结构块。每个结构块由没有非线性激活函数的较窄的输入层，将输入扩充到较高纬度的空间中扩充层，投影至输出的投影层以及同样没有非线性激活函数的较窄的输出层组成。残差结构连接输入和输出。

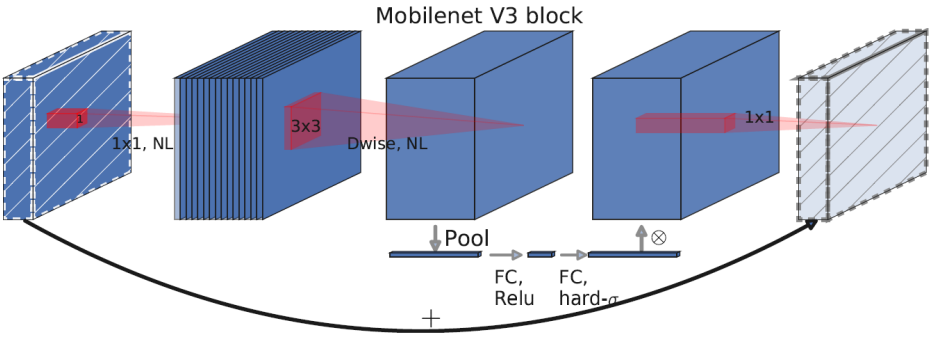

图 4. MobileNetV2 + SE 结构 [20]。对比文献 [20]，我们将 SE 结构插入残差层。我们在层间使用了不一样的非线性激活函数，具体细节在章节 5.2。

Search for MobileNetV3

Abstract

We present the next generation of MobileNets based on a combination of complementary search techniques as well as a novel architecture design. MobileNetV3 is tuned to mobile phone CPUs through a combination of hardwareaware network architecture search (NAS) complemented by the NetAdapt algorithm and then subsequently improved through novel architecture advances. This paper starts the exploration of how automated search algorithms and network design can work together to harness complementary approaches improving the overall state of the art. Through this process we create two new MobileNet models for release: MobileNetV3-Large and MobileNetV3-Small which are targeted for high and low resource use cases. These models are then adapted and applied to the tasks of object detection and semantic segmentation. For the task of semantic segmentation (or any dense pixel prediction), we propose a new efficient segmentation decoder Lite Reduced Atrous Spatial Pyramid Pooling (LR-ASPP). We achieve new state of the art results for mobile classification, detection and segmentation. MobileNetV3-Large is 3.2accurate on ImageNet classification while reducing latency by 20% compared to MobileNetV2. MobileNetV3-Small is 6.6% more accurate compared to a MobileNetV2 model with comparable latency. MobileNetV3-Large detection is over 25% faster at roughly the same accuracy as MobileNetV2 on COCO detection. MobileNetV3-Large LRASPP is 34% faster than MobileNetV2 R-ASPP at similar accuracy for Cityscapes segmentation.

1 Intrudction

Efficient neural networks are becoming ubiquitous in mobile applications enabling entirely new on-device experiences. They are also a key enabler of personal privacy allowing a user to gain the benefits of neural networks without needing to send their data to the server to be evaluated. Advances in neural network efficiency not only improve user experience via higher accuracy and lower latency, but also help preserve battery life through reduced power consumption.

This paper describes the approach we took to develop MobileNetV3 Large and

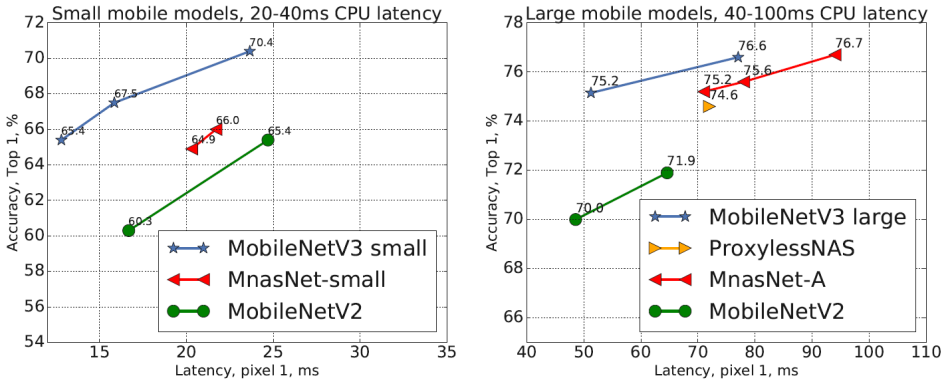

Figure 1. The trade-off between Pixel 1 latency and top-1 ImageNet accuracy. All models use the input resolution 224. V3 large and V3 small use multipliers 0.75, 1 and 1.25 to show optimal frontier. All latencies were measured on a single large core of the same device using TFLite[1]. MobileNetV3-Small and Large are our proposed next-generation mobile models.

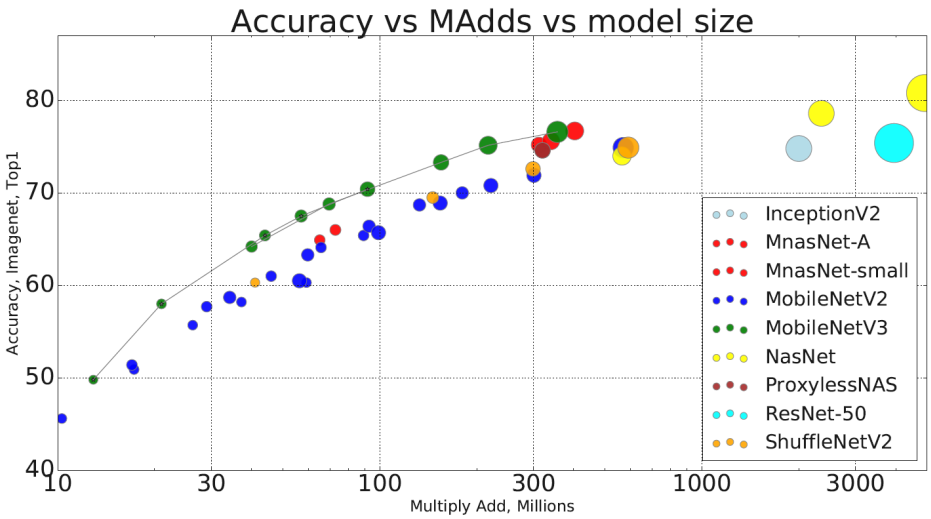

Figure 2. The trade-off between MAdds and top-1 accuracy. This allows to compare models that were targeted different hardware or software frameworks. All MobileNetV3 are for input resolution 224 and use multipliers 0.35, 0.5, 0.75, 1 and 1.25. See section 6 for other resolutions. Best viewed in color.

Small models in order to deliver the next generation of high accuracy efficient neural network models to power on-device computer vision. The new networks push the state of the art forward and demonstrate how to blend automated search with novel architecture advances to build effective models.

The goal of this paper is to develop the best possible mobile computer vision architectures optimizing the accuracylatency trade off on mobile devices. To accomplish this we introduce (1) complementary search techniques, (2) new efficient versions of nonlinearities practical for the mobile setting, (3) new efficient network design, (4) a new efficient segmentation decoder. We present thorough experiments demonstrating the efficacy and value of each technique evaluated on a wide range of use cases and mobile phones.

The paper is organized as follows. We start with a discussion of related work in Section 2. Section 3 reviews the efficient building blocks used for mobile models. Section 4 reviews architecture search and the complementary nature of MnasNet and NetAdapt algorithms. Section 5 describes novel architecture design improving on the efficiency of the models found through the joint search. Section 6 presents extensive experiments for classification, detection and segmentation in order do demonstrate efficacy and understand the contributions of different elements. Section 7 contains conclusions and future work.

2 Relate Work

Designing deep neural network architecture for the optimal trade-off between accuracy and efficiency has been an active research area in recent years. Both novel handcrafted structures and algorithmic neural architecture search have played important roles in advancing this field.

SqueezeNet[22] extensively uses 1x1 convolutions with squeeze and expand modules primarily focusing on reducing the number of parameters. More recent works shifts the focus from reducing parameters to reducing the number of operations (MAdds) and the actual measured latency. MobileNetV1[19] employs depthwise separable convolution to substantially improve computation efficiency. MobileNetV2[39] expands on this by introducing a resource-efficient block with inverted residuals and linear bottlenecks.

ShuffleNet[49] utilizes group convolution and channel shuffle operations to further reduce the MAdds. CondenseNet[21] learns group convolutions at the training stage to keep useful dense connections between layers for feature re-use. ShiftNet[46] proposes the shift operation interleaved with point-wise convolutions to replace expensive spatial convolutions.

To automate the architecture design process, reinforcement learning (RL) was first introduced to search efficient architectures with competitive accuracy [53, 54, 3, 27, 35]. A fully configurable search space can grow exponentially large and intractable. So early works of architecture search focus on the cell level structure search, and the same cell is reused in all layers. Recently, [43] explored a block-level hierarchical search space allowing different layer structures at different resolution blocks of a network. To reduce the computational cost of search, differentiable architecture search framework is used in [28, 5, 45] with gradient-based optimization. Focusing on adapting existing networks to constrained mobile platforms, [48, 15, 12] proposed more efficient automated network simplification algorithms.

Quantization [23, 25, 47, 41, 51, 52, 37] is another important complementary effort to improve the network efficiency through reduced precision arithmetic. Finally, knowledge distillation [4, 17] offers an additional complementary method to generate small accurate ”student” networks with the guidance of a large ”teacher” network.

3 Efficient Mobile Building Blocks

Mobile models have been built on increasingly more efficient building blocks. MobileNetV1 [19] introduced depthwise separable convolutions as an efficient replacement for traditional convolution layers. Depthwise separable convolutions effectively factorize traditional convolution by separating spatial filtering from the feature generation mechanism. Depthwise separable convolutions are defined by two separate layers: light weight depthwise convolution for spatial filtering and heavier 1x1 pointwise convolutions for feature generation.

MobileNetV2 [39] introduced the linear bottleneck and inverted residual structure in order to make even more efficient layer structures by leveraging the low rank nature of the problem. This structure is shown on Figure 3 and is defined by a 1x1 expansion

convolution followed by depthwise convolutions and a 1x1 projection layer. The input and output are connected with a residual connection if and only if they have the same number of channels. This structure maintains a compact representation at the input and the output while expanding to a higher-dimensional feature space internally to increase the expressiveness of nonlinear perchannel transformations.

MnasNet [43] built upon the MobileNetV2 structure by introducing lightweight attention modules based on squeeze and excitation into the bottleneck structure. Note that the squeeze and excitation module are integrated in a different location than ResNet based modules proposed in [20]. The module is placed after the depthwise filters in the expansion in order for attention to be applied on the largest representation as shown on Figure 4.

For MobileNetV3, we use a combination of these layers as building blocks in order to build the most effective models. Layers are also upgraded with modified swish nonlinearities [36, 13, 16]. Both squeeze and excitation as well as the swish nonlinearity use the sigmoid which can be inefficient to compute as well challenging to maintain accuracy in fixed point arithmetic so we replace this with the hard sigmoid [2, 11] as discussed in section 5.2.

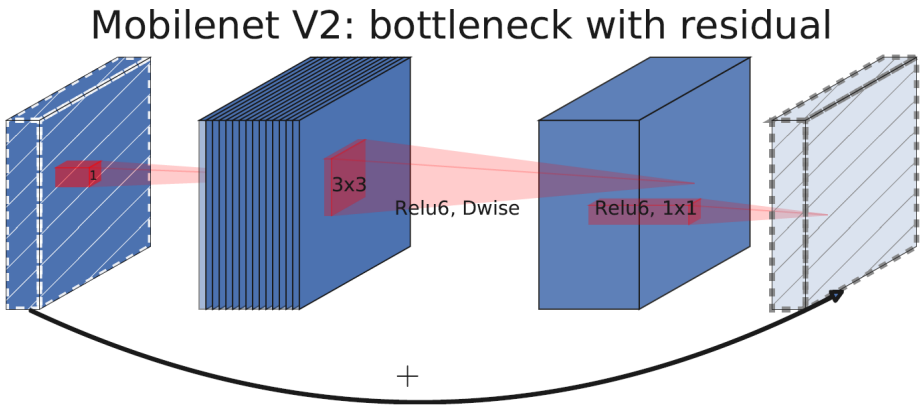

Figure 3. MobileNetV2 layer (Inverted Residual and LinearBottleneck). Each block consists of narrow input and output (bottleneck), which don't have nonlinearity, followed by expansion to a much higher-dimensional space and projection to the output. The residual connects bottleneck (rather than expansion).

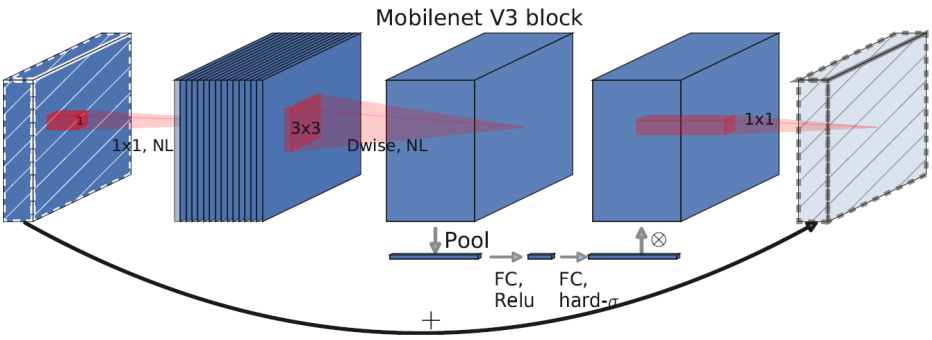

Figure 4. MobileNetV2 + Squeeze-and-Excite [20]. In contrast with [20] we apply the squeeze and excite in the residual layer. We use different nonlinearity depending on the layer, see section 5.2 for details.