**MobileNets：用于移动端视觉应用的高效卷积神经网络**

**摘要**

我们提供一类称为MobileNets的高效模型，用于移动和嵌入式视觉应用。 MobileNets是基于一个流线型的架构，它使用深度可分离的卷积来构建轻量级的深层神经网络。我们引入两个简单的全局超参数，在延迟度和准确度之间有效地进行平衡。这两个超参数允许模型构建者根据问题的约束条件，为其应用选择合适大小的模型。我们进行了资源和精度权衡的广泛实验，与ImageNet分类上的其他流行的网络模型相比，MobileNets表现出很强的性能。最后，我们展示了MobileNets在广泛的应用场景中的有效性，包括物体检测，细粒度分类，人脸属性和大规模地理定位。

1. **简介**

自从AlexNet 通过赢得ImageNet的挑战“ILSVRC 2012” 来推广深度卷积神经网络以来，卷积神经网络就已经在计算机视觉中变得无处不在。 总的趋势是建立更深，更复杂的网络以实现更高的准确性。 但是，这些提高准确性的进步并不一定会使网络在大小和速度方面更加高效。在许多现实世界的应用程序中，例如机器人技术，自动驾驶汽车和增强现实，识别任务需要在计算受限的平台上及时执行。

本文介绍了一种有效的网络体系结构和两个超参数集，以建立非常小的，低延迟的模型，这些模型可以轻松地与移动和嵌入式视觉应用的设计要求匹配。第2节回顾了构建小型模型的先前工作。 第3节介绍了MobileNet体系结构以及两个超参数宽度因子和分辨率因子，以定义更小，更高效的MobileNet。 第4节介绍了ImageNet上的实验以及各种不同的应用程序和用例。第5节以总结和结论结束。

1. **先前的工作**

在最近的文献中，人们对建立小型高效的神经网络越来越感兴趣。通常可以将许多不同的方法归类为压缩预训练网络或直接训练小型网络。本文提出了一类网络架构，允许模型开发人员专门为其应用选择与资源限制（延迟，大小）相匹配的小型网络。MobileNets主要专注于优化延迟，但也会产生小型网络。目前关于小型网络的许多论文只关注规模，却不考虑速度。

MobileNets主要是从深度方向可分卷积中构建的，该方法最初在Inception模型中引入，随后在Inception模型中使用，以减少前几层的计算量。扁平化网络由完全分解的卷积构建网络，并显示了高度分解的网络的潜力。独立于这篇论文之外，分解网络引入了类似的分解卷积以及拓扑连接的使用。随后，Xception网络演示了如何按比例扩展深度可分离滤波器以执行Inception V3网络。另一个小型网络是Squeezenet，它使用瓶颈方法设计了一个非常小的网络。其他简化的计算网络包括结构化转换网络和deep fried convnets。

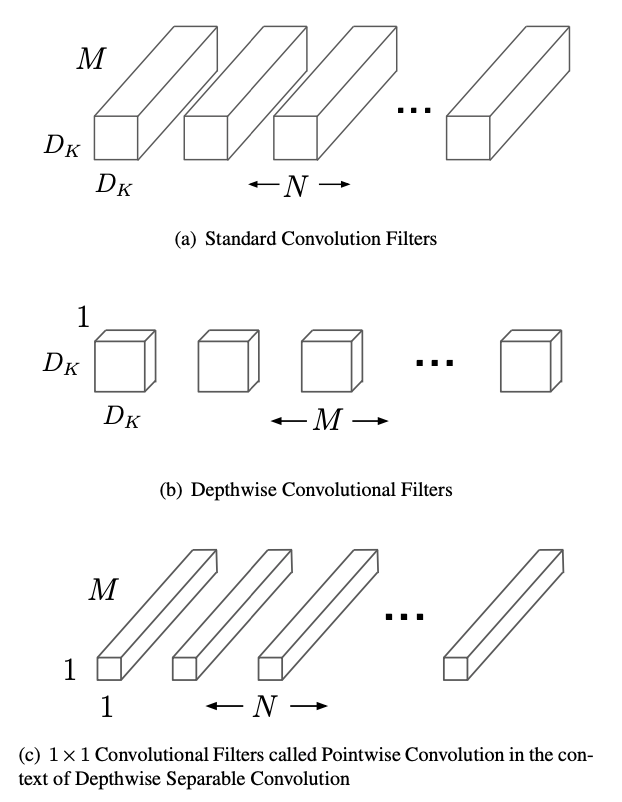
获得小型网络的另一种方法是收缩，分解或压缩预训练的网络。在文献中已经提出了基于乘积量化，散列以及修剪，矢量量化和霍夫曼编码的压缩。另外，已经提出了各种因式分解来加速预训练的网络。训练小型网络的另一种方法是蒸馏，该方法使用较大的网络来训练较小的网络。它是我们方法的补充，在第4节的某些用例中进行了介绍。另一种新兴方法是低位网络。

1. **MobileNet架构**

在本节中，我们首先描述MobileNet所基于的核心层，它们是深度可分离的过滤器。 然后，我们描述MobileNet网络结构，并以两个模型缩减超参数宽度因子和分辨率因子的描述作为结束。

**3.1深度可分离卷积**

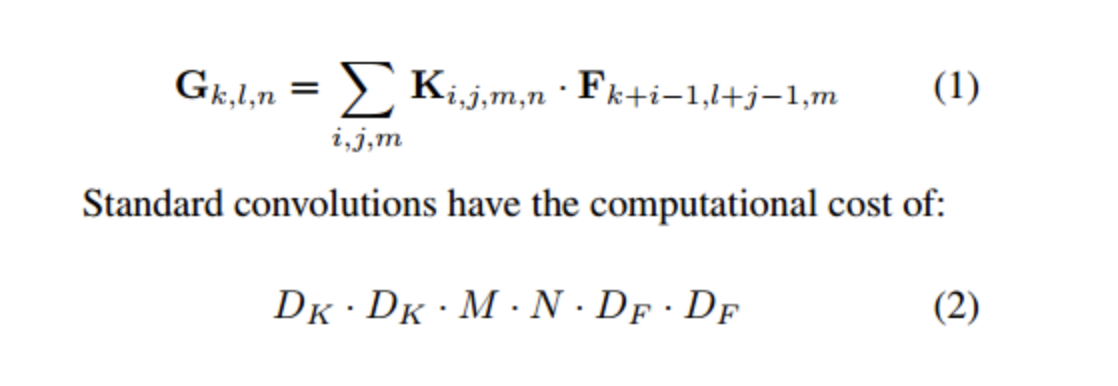
MobileNet模型基于深度可分离卷积，它是分解卷积的一种形式，它将标准卷积分解为深度卷积和称为点向卷积的1×1卷积。 对于MobileNet，深度卷积将单个滤波器应用于每个输入通道。 然后逐点卷积应用1×1卷积以组合输出深度卷积。 一个标准的卷积既可以过滤又可以将输入合并为一组新的输出。 深度可分离卷积将其分为两层，一个用于过滤的单独层和一个用于合并的单独层。 这种分解具有极大地减少计算和模型大小的效果。图2显示了如何将标准卷积2（a）分解为深度卷积2（b）和1×1点向卷积2（c）。



标准卷积层将DF×DF×M特征图F作为输入，并生成DF×DF×N特征图G，其中DF是正方形输入特征图的空间宽度和高度，M是输入通道的数量（输入深度），DG是方形输出特征图的空间宽度和高度，N是输出通道的数量（输出深度）

标准卷积层由大小为DK×DK×M×N的卷积核K进行参数化，其中DK是假定为正方形的核的空间尺寸，M是输入通道数，N是输出通道数，如先前定义 。

假设跨度为1和填充为标准卷积的输出特征图，计算公式为：

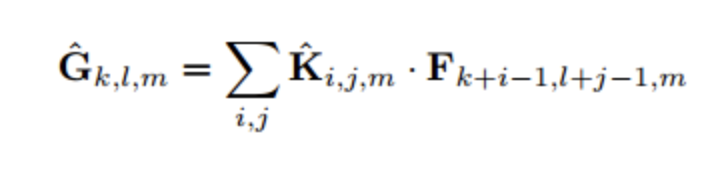


其中计算成本乘以输入通道数M，输出通道数 N，内核大小 Dk\*Dk 和特征图大小DF \* DF 。MobileNet模型处理了这些项中的每一个及其相互作用。首先，它使用深度可分离卷积来打破输出通道数与内核大小之间的相关作用。

标准卷积运算具有对基于卷积核的特征进行过滤并组合特征以产生新表示的效果。可以通过使用称为深度可分离卷积的因式分解将过滤和组合步骤分为两步，这样可以实质性地减少计算成本。

深度可分离卷积由两层组成：深度卷积和点卷积。我们使用深度卷积对每个输入通道（输入深度）应用单个滤波器。 然后使用逐点卷积（简单的1×1卷积）来创建深度层输出的线性组合。 MobileNets对这两个层都使用了batchnorm和ReLU非线性函数。

简单来说，每个输入通道（输入深度）带有一个滤波器的深度卷积可以写成下面：

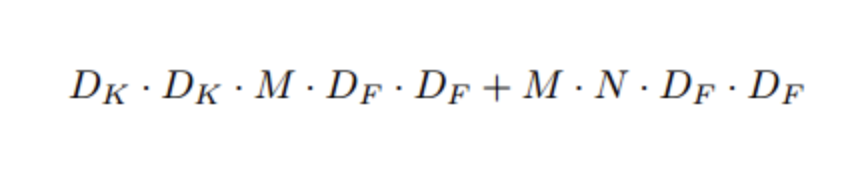


其中K表示深度卷积，卷积核为（Dk,  Dk,  M），其中第m个卷积核应用在F中第m个通道上，产生 G上第m个通道输出。

深度卷积计算量为：Dk\*Dk\*M\*Df\*Df

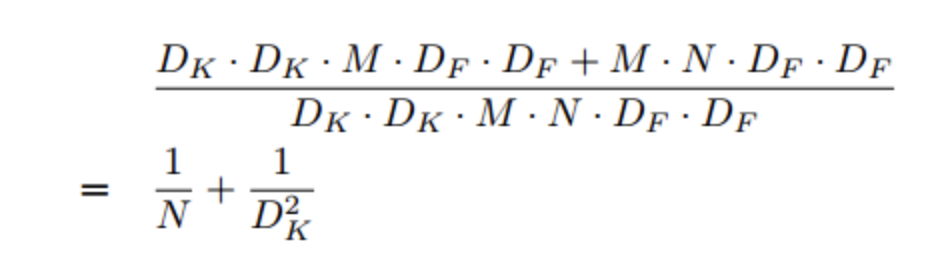
相对于标准于标准卷积，深度卷积非常有效。但是，它仅过滤输入通道，不会将他们组合以创建新功能。因此，需要额外的一层来计算通过 1\*1 卷积的深度卷积输出的线性组合，以便生出这些新特征。

深度卷积和 1\*1 （点向）卷积的组合被称为深度可分离卷积，最初在[26]中引入的。



上面是深度卷积和1\*1点卷积的总和。

通过将卷积表示为过滤和组合的两步过程，我们减少了以下计算：



MobileNet使用 3\*3 深度可分离卷积，使用的计算量比标准卷积小8到9倍，而准确率仅略有降低，如第四节所述。

在空间维度上进行额外的因式分解不会节省太多额外的计算，因为在深度卷积中只花费了很少的计算。

**3.2网络结构与训练**

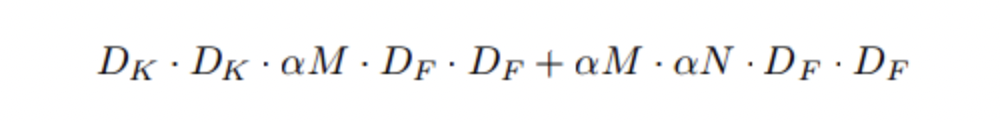
如前一节所述，MobileNet结构建立在深度可分离卷积之上，第一层是完整卷积。通过以如此简单的术语定义网络，我们能够轻松地探索网络拓扑以找到一个好的网络。表1中定义了MobileNet体系结构。所有层后面都带有一个batchnorm和ReLU非线性函数，最后一个完全连接的层除外，该层没有非线性函数，并馈入softmax层进行分类。图3将具有常规卷积，batchnorm和ReLU非线性的层与具有深度卷积，1×1点式卷积以及每个卷积层后的batchnorm和ReLU的分解层进行了对比。下采样是在深度卷积以及第一层中通过跨步卷积处理的。最终平均池在完全连接的层之前将空间分辨率降低为1。将深度和点积卷积计为单独的层，MobileNet有28层。

仅用少量的Mult-Adds定义网络是不够的。确保这些操作可以有效实施也很重要。例如，直到非常高的稀疏度，非结构化稀疏矩阵运算通常不比密集矩阵运算快。我们的模型结构将几乎所有计算都放入密集的1×1卷积中。这可以通过高度优化的通用矩阵乘法（GEMM）函数来实现。卷积通常由GEMM实现，但需要在内存中进行名为im2col的初始重新排序，才能将其映射到GEMM。例如，这种方法在流行的Caffe软件包中使用。 1×1卷积不需要在内存中进行重新排序，可以直接使用GEMM（这是最优化的数值线性代数算法之一）来实现。如表2所示，MobileNet将其95％的计算时间花费在1×1卷积中，其中也包含75％的参数。几乎所有其他参数都位于完全连接的层中。

使用RMSprop [33]在TensorFlow [1]中使用与Inception V3 [31]类似的异步梯度下降训练了MobileNet模型。但是，与训练大型模型相反，我们使用较少的正则化和数据增强技术，因为小型模型的过拟合麻烦较小。在训练MobileNets时，我们不使用侧边或标签平滑，并且通过限制大型Inception训练中使用的小作物的大小来减少失真的图像数量。另外，我们发现，对深度过滤器进行很少或几乎没有权重衰减是很重要的，因为它们中的参数太少了。对于下一部分中的ImageNet基准，无论模型的大小如何，所有模型都使用相同的训练参数进行训练。

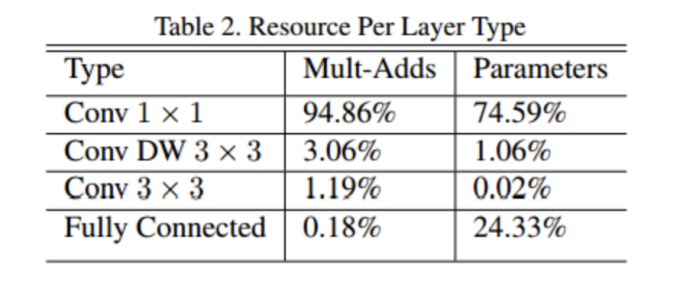
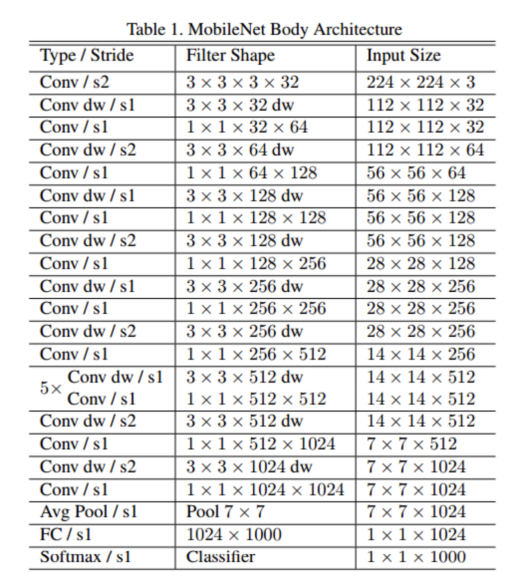
**3.3宽度参数：更小的模型**

尽管基本的MobileNet体系结构已经很小且延迟很低，但很多情况下特定用例或应用可能要求模型变得更小，更快。为了构建这些更小且计算成本更低的模型，我们引入了一个非常简单的参数 alpha，称为宽度乘数。宽度乘数 alpha 的作用是使每一层的网络均匀变薄。对于给定的层和宽度乘数 alpha，输入通道的数量变为 alpha\*M，而输出通道的数量N变为 alpha\*N。具有宽度乘数 alpha 的深度可分离卷积的计算成本为：



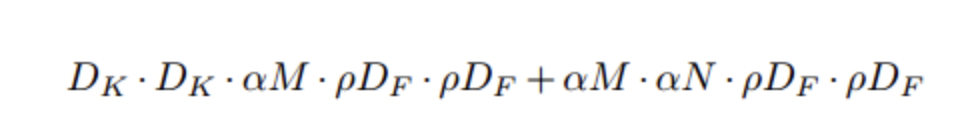
其中α∈（0，1]，典型设置为1、0.75、0.5和0.25。α= 1是基准MobileNet，α<1是减少的MobileNets。宽度乘数具有减少计算成本和二次方参数数量的作用宽度乘数可以应用于任何模型结构，以合理的精度，等待时间和大小折衷来定义新的较小模型，用于定义需要从头开始训练的新的简化结构。

下图为MobileNet的网络构成，它的95%的时间是在 1\*1 conv层上消耗的，另外 1\*1 的conv参数也占了所有可训练参数的 75%。



**3.4分辨率因子：减少表示**

减少神经网络计算成本的第二个超参数是分辨率乘数ρ。我们将其应用于输入图像，然后通过相同的乘数来减少每一层的内部表示。实际上，我们通过设置输入分辨率来隐式设置ρ。现在，我们可以将网络核心层的计算成本表示为具有宽度乘数α和分辨率乘数ρ的深度可分离卷积：



其中ρ∈（0，1]通常是隐式设置的，因此网络的输入分辨率为224、192、160或128。ρ= 1为基准值，而ρ<1是简化的计算MobileNets。将计算成本降低ρ的平方的效果。

作为示例，我们可以查看MobileNet中的典型层，并了解深度可分离卷积，宽度乘数和分辨率乘数如何减少成本和参数。表3显示了将体系结构收缩方法依次应用于该层时该层的参数的计算和数量。第一行显示了完整卷积层的Mult-Adds和参数，输入特征图的大小为14×14×512，内核K的大小为3×3×512×512。我们将在下一部分中详细介绍在资源和准确性之间进行权衡。

