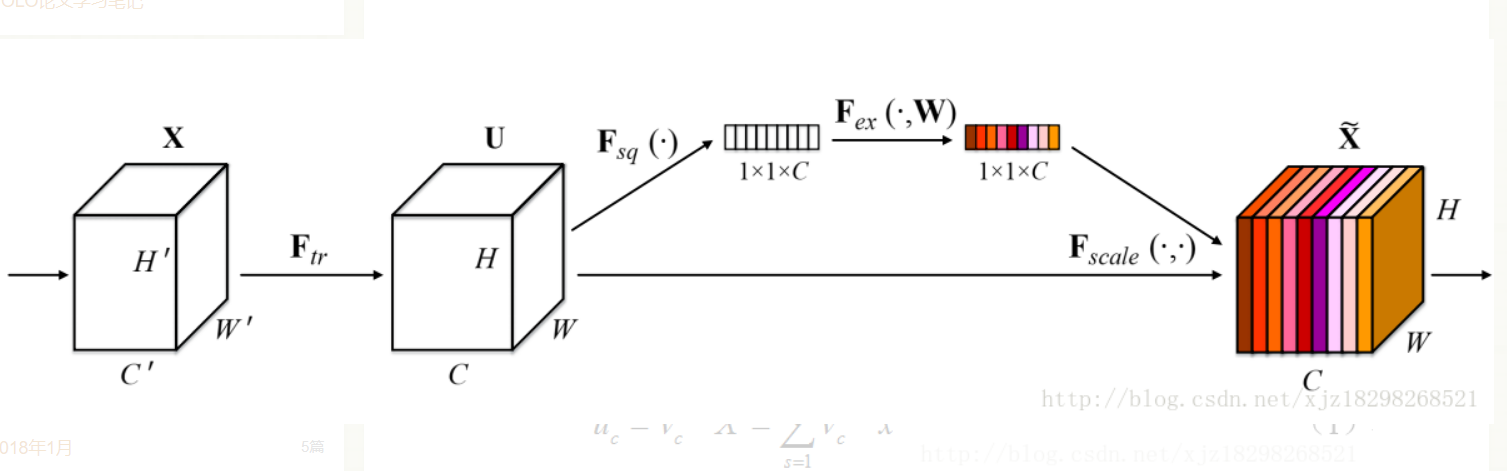
# 序列激活网络[[1]](#footnote-1)

赵淼译

(北京交通大学电子与信息工程学院，电子科学与技术，15650761201)

作者通过显式建模的方法利用全局信息获得了每个通道的重要性，然后进行增强或削弱，实现特征通道自适应校准。特征图U经过squeeze操作在空间维度H\*W进行融合形成一个通道描述子。这样可以达到对全局信息利用的目的。紧跟着是一个excitation操作，通过通道依赖的的自控机制获得每个通道的激活值。特征图进行重新加权送到后面的信息中。

这一模块可以直接替换以前的模块。这是一个通用的模块，但是在不同深度下适合网络的不同需要。在早期以类不可知的方式特性，支持共享的低级表示的质量。在以后的层中，SE块变得越来越专门化，并以高度特定于类的方式响应不同的输入。因此，通过整个网络可以累积由SE块进行的特性重新校准的好处。



## 相关工作

深度架构：1深度的优势（VGG,Inception）2归一化的操作3残差结构（使用跳跃性连接）4高速网络（使用门机制）。对网络层5、14层之间的连接进行了重新设计，以进一步提高深度网络的学习和代表性特性。

## 核心

论文的核心是Squeeze和Excitation模块。

卷积神经网络是建立在卷积运算的基础上的，它通过将空间和通道信息融合到局部的接受域中来提取信息性的特征。最近的一个研究重点是空间编码的增强。

最近一方面是对多尺度处理过程的嵌入（Inception Network）,另一方面是对空间依赖性和注意力机制的应用）。

前面是正常的卷积运算，V=[v1,v2,…,vc]是一系列卷积操作，然后vcs代表一个二维空间的卷积核运算，作用于X对应的通道。输出是各个通道的和，通道依赖关系隐含地嵌入到vc中，但是这些依赖关系与过滤器捕获的空间相关性纠缠在一起。这里显式的模拟了通道间的依赖关系，对滤波器响应进行了重新校准。

|  |  |
| --- | --- |
|  | 1 |

### Squeeze：全局信息嵌入

为了解决通道依赖性的问题，首先考虑每个通道的输出信号。在每个通道输出的特征图上的一个单元考虑的是局部感受野的信息，不能探索区域外的上下文信息，低层次的特征图感受野很小。

这里使用squeeze操作获得全局空间信息作为一个通道描述子。用全局平均池化获得通道统计信息。形式上，z属于C维向量。

|  |  |
| --- | --- |
|  | 1 |

转换输出U可以被解释为一个本地描述符的集合，这些描述符的统计信息是对整个图像的表达。利用这些信息在35、38、49中很普遍。我们选择最简单的全球平均池，并指出在这里也可以采用更复杂的聚合策略。

### Excitation：自适应校准

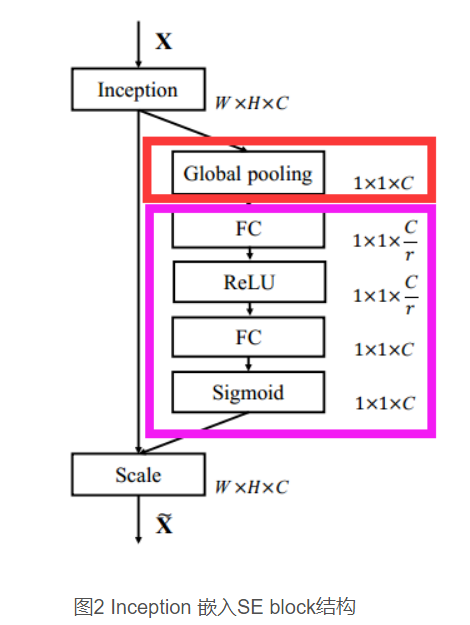
利用在squeeze中的信息聚合操作，我们接下来通过第二个操作来全面捕获通道依赖性。为了实现这个目标，这个功能必须符合两个标准：第一，它必须是灵活的（特别是它必须能够学习通道之间的非线性交互）；第二，它必须学习一个非互斥的关系，因为独热激活相反，这里允许强调多个通道。为了满足这些标准，我们选择采用一个简单的门机制，并使用sigmoid激活：

|  |  |
| --- | --- |
|  | 1 |

Z经过带有relu函数的全连接层进行降维，后又经过W2进行升维，最后对U进行变换。

|  |  |
| --- | --- |
|  | 1 |

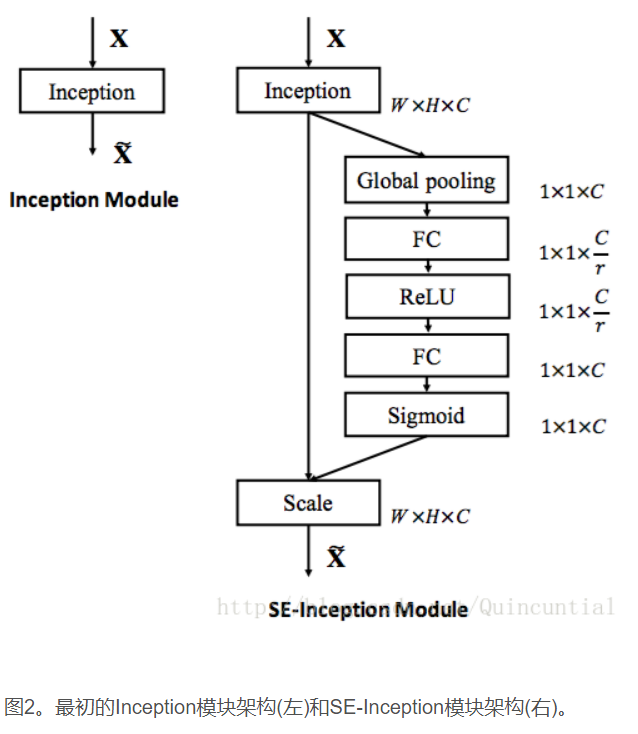
这里通过激活值得到各通道权值，获得全局性的信息决定通道信息利用。



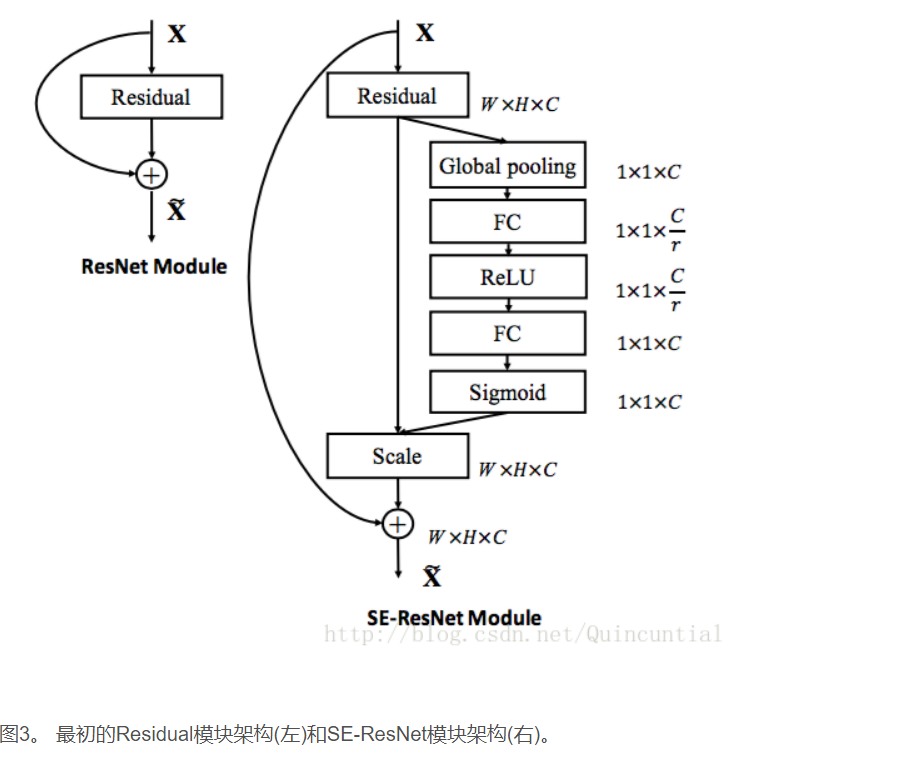
### SE Inception 和 SE ResNet

SE模块可以直接应用于AlexNet和VGGNet，这里作者通过一定的设计将SENet应用到了一些现代的架构中。

通过将变换Ftr看作一个整体的Inception模块（参见图2），为Inception网络构建SE块。通过对架构中的每个模块进行更改，我们构建了一个SE-Inception网络。



残差网络及其变种已经证明在学习深度表示方面非常有效。我们开发了一系列的SE块，分别与ResNet[9]，ResNeXt[43]和Inception-ResNet[38]集成。图3描述了SE-ResNet模块的架构。在这里，SE块变换FtrFtr被认为是残差模块的非恒等分支。压缩和激励都在恒等分支相加之前起作用。



接下来，我们考虑由提议的区块引入的附加参数。所有这些都被包含在两个FC层的控制机制中，这构成了整个网络容量的一小部分。更准确地说，引入的附加参数的数量是由：

|  |  |
| --- | --- |
|  | 1 |

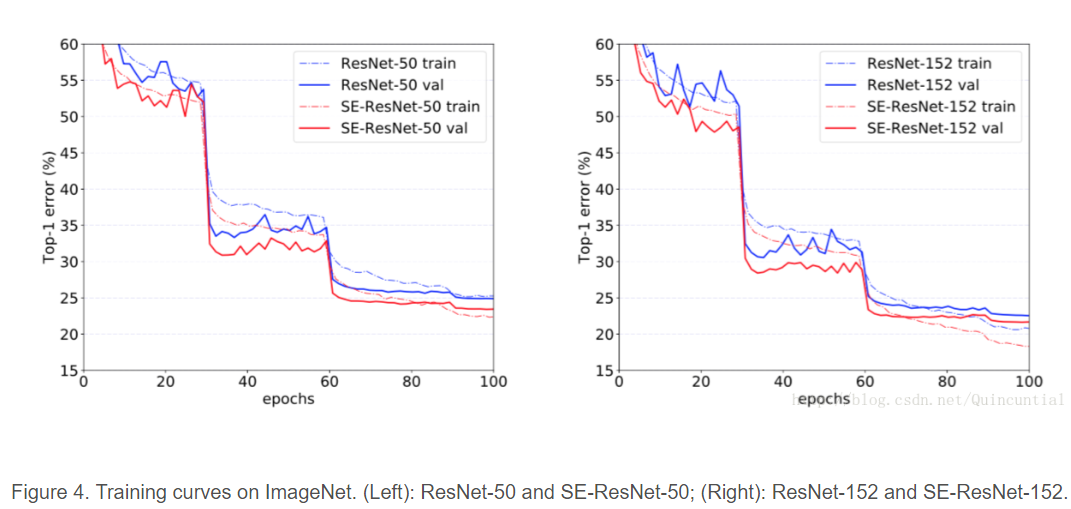
当r表示还原比时，S指的是阶段的数量（每个阶段指的是在一个公共空间维度的特征图上操作的块的集合），Cs表示输出通道的维数，而n表示阶段S的重复块数。这里计算代价最高的是最后阶段，这一阶段被移除在性能下降不明显的情况下获得计算复杂度的下降。

## 5 实施

在训练过程中，我们遵循标准的做法，使用随机大小裁剪[39]到224×224像素（299×299用于Inception-ResNet-v2[38]和SE-Inception-ResNet-v2）和随机的水平翻转进行数据增强。输入图像通过通道减去均值进行归一化。另外，我们采用[32]中描述的数据均衡策略进行小批量采样，以补偿类别的不均匀分布。网络在我们的分布式学习系统“ROCS”上进行训练，能够处理大型网络的高效并行训练。使用同步SGD进行优化，动量为0.9，小批量数据的大小为1024（在4个服务器的每个GPU上分成32张图像的子批次，每个服务器包含8个GPU）。初始学习率设为0.6，每30个迭代周期减少10倍。使用[8]中描述的权重初始化策略，所有模型都从零开始训练100个迭代周期。

## 6 实验

ImageNet 2012数据集包含来自1000个类别的128万张训练图像和5万张验证图像。我们在训练集上训练网络，并在验证集上使用中心裁剪图像评估来报告top-1和top-5错误率，其中每张图像短边首先归一化为256，然后从每张图像中裁剪出224×224个像素，（对于Inception-ResNet-v2和SE-Inception-ResNet-v2，每幅图像的短边首先归一化到352，然后裁剪出299×299个像素）。

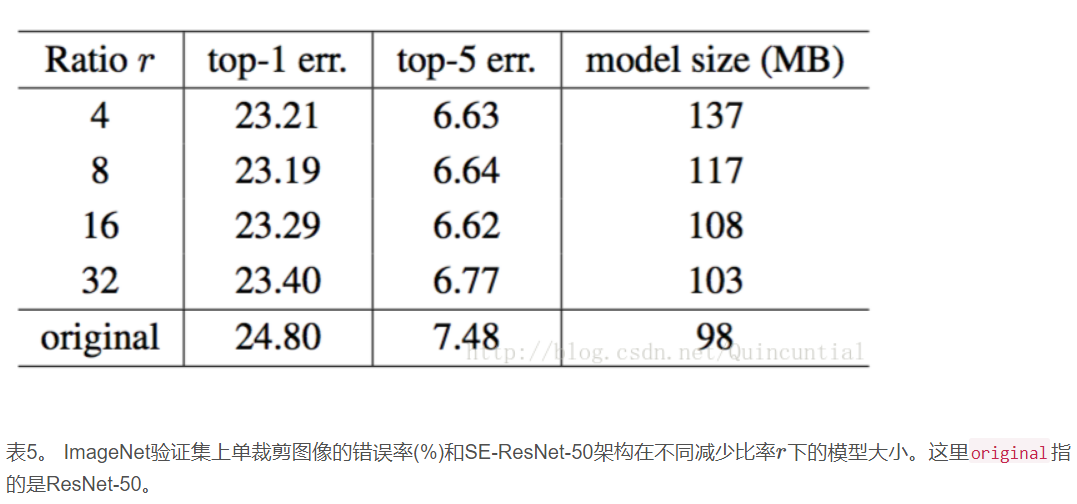


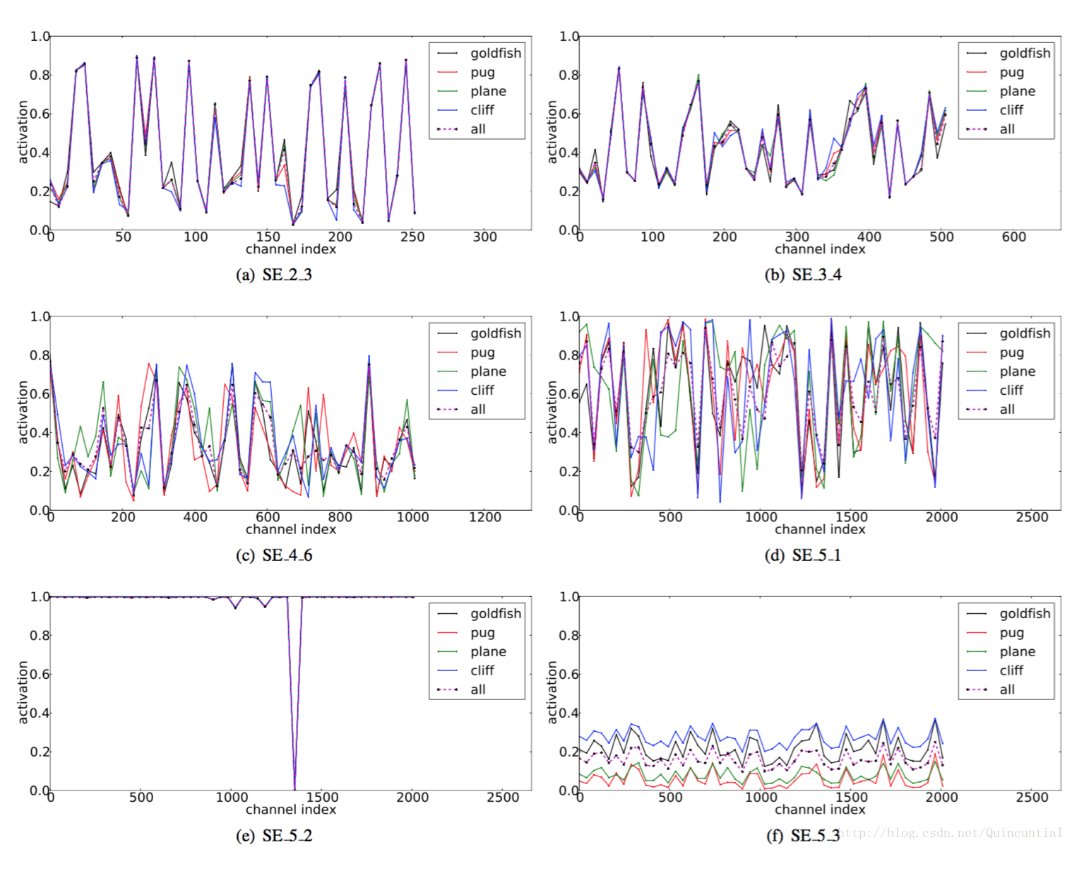
虽然应该注意SE块本身增加了深度，但是它们的计算效率极高，即使在扩展的基础架构的深度达到收益递减的点上也能产生良好的回报。而且，我们看到通过对各种不同深度的训练，性能改进是一致的，这表明SE块引起的改进可以与增加基础架构更多深度结合使用。

同时作者尝试了多种现代网络，证明了SE块可以应用到多种结构下。

### 分析与解释

减少比率。公式（5）中引入的减少比率r是一个重要的超参数，它允许我们改变模型中SE块的容量和计算成本。为了研究这种关系，我们基于SE-ResNet-50架构进行了一系列不同r值的实验。表5中的比较表明，性能并没有随着容量的增加而单调上升。这可能是使SE块能够过度拟合训练集通道依赖性的结果。尤其是我们发现设置r=16在精度和复杂度之间取得了很好的平衡，因此我们将这个值用于所有的实验





首先，不同类别的分布在较低层中几乎相同，例如，SE\_2\_3。这表明在网络的最初阶段特征通道的重要性很可能由不同的类别共享。然而有趣的是，第二个观察结果是在更大的深度，每个通道的值变得更具类别特定性，因为不同类别对特征的判别性值具有不同的偏好。SE\_4\_6和SE\_5\_1。这两个观察结果与以前的研究结果一致[21,46]，即低层特征通常更普遍（即分类中不可知的类别），而高层特征具有更高的特异性。因此，表示学习从SE块引起的重新校准中受益，其自适应地促进特征提取和专业化到所需要的程度。最后，我们在网络的最后阶段观察到一个有些不同的现象。SE\_5\_2呈现出朝向饱和状态的有趣趋势，其中大部分激活接近于1，其余激活接近于0。在所有激活值取1的点处，该块将成为标准残差块。在网络的末端SE\_5\_3中（在分类器之前紧接着是全局池化），类似的模式出现在不同的类别上，尺度上只有轻微的变化（可以通过分类器来调整）。这表明，SE\_5\_2和SE\_5\_3在为网络提供重新校准方面比前面的块更不重要。这一发现与第四节实证研究的结果是一致的，这表明，通过删除最后一个阶段的SE块，总体参数数量可以显著减少，性能只有一点损失（<0.1%的top-1错误率）。

## 结论

在本文中，我们提出了SE块，这是一种新颖的架构单元，旨在通过使网络能够执行动态通道特征重新校准来提高网络的表示能力。大量实验证明了SENets的有效性，其在多个数据集上取得了最先进的性能。此外，它们还提供了一些关于以前架构在建模通道特征依赖性上的局限性的洞察，我们希望可能证明SENets对其它需要强判别性特征的任务是有用的。最后，由SE块引起的特征重要性可能有助于相关领域，例如为了压缩的网络修剪。

## SENet154细节

SENet是通过将SE块集成到64×4d的ResNeXt-152的修改版本中构建的，通过遵循ResNet-152[9]的块堆叠来扩展原始ResNeXt-101[43]。更多设计和训练差异（除了SE块的使用之外）如下：（a）对于每个瓶颈构建块，首先1×1卷积通道的数量减半，以性能下降最小的方式降低网络的计算成本。（b）第一个7×7卷积层被三个连续的3×3卷积层所取代。（c）步长为2的1×1卷积的下采样投影被替换步长为2的3×3卷积以保留信息。（d）在分类器层之前插入一个丢弃层（丢弃比为0.2）以防止过拟合。（e）训练期间使用标签平滑正则化（如[40]中所介绍的）。（f）在最后几个训练迭代周期，所有BN层的参数都被冻结，以确保训练和测试之间的一致性。（g）使用8个服务器（64个GPU）并行执行培训，以实现大批量数据大小（2048），初始学习率为1.0。

1. 赵淼 电信研1706班 电子科学与技术 17120036 15650761201 [↑](#footnote-ref-1)