

MicroNet:2020

Towards Image Recognition with Extremely Low FLOPS

论文出发点或背景

当网络的计算成本变得极低的时候，网络的性能也会下降

思考一个问题：能否在6MFLOPS下，对输入为分辨率为 224×224 的图像进行超过1000种类别的分类

如果基于现在的网络解决这些问题就要显著减少其宽度与深度，但是换来的是网络性能的显著下降

论文创新思路

提出了微因子分解卷积，将驻点卷积和深度卷积分解为低秩矩阵，以便在通道数量和输入/输出连接之间进行较好的权衡

提出了一种新的激活函数Dynamic Shift-Max，通过最大化输入特征图与其循环通道偏移之间的多重动态融合，来增强非线性特征

设计了分组自适应卷积来分解点态卷积，通过一个平方根关系来调整group的数量到channel的数量

论文方法介绍

两个设计原则：

- 1.通过降低节点的联通来避免网络宽度的减少

当计算成本（与连接数量成正比）固定时，通道数量与连接性冲突。我们认为，它们之间的良好平衡可以有效地避免信道减少，提高一个层的表示能力。

我们假设卷积核 W 具有相同数量的输入和输出通道（ $C_{in} = C_{out} = C$ ），并忽略了偏差。核矩阵 W 被分解为两个组自适应卷积，其中组数 G 取决于通道 c 的数量，在数学上可以表示为： $W = P\Phi Q^T$

2.通过再每层中引入更为复杂的非线性变化进而弥补网络深度的减少

我们将一个 $k \times k$ 个深度卷积核分解为一个 $k \times 1$ 核和一个 $1 \times k$ 核。

其中 J 表示组数， i 表示通道数， K 表示融合后的输出数量。当 $J=K=2$ 时，可以在准确率和复杂度之间取得较好的折衷。

Micro-Block-A

Micro-Block-A如上图（a）所示使用微分解的pointwise和depthwise卷积的精简组合。通过微分解的depthwise卷积扩展了通道的数量，并通过组自适应卷积压缩它们。

Micro-Block-B

Micro-Block-B如上图（b）所示，用于连接Micro-Block-A和Micro-Block-C，它既压缩又扩展了通道的数量。

Micro-Block-C

Micro-Block-C如上图（c）所示，实现了微分解的pointwise卷积和depthwise卷积的常规组合。它最适合于更高的网络层，因为它比精简组合为通道融合分配更多的计算。

实际效果

个人理解

这篇论文主要的创新点：

- 1.通过分解卷积核，进一步压缩网络的连接性，使得网络更加稀疏化，提高计算性能。
- 2.设计了全新的动态激活函数，引入更多的非线性特征，增强模型表现力。

感觉文章数学性还是挺强的，在很多地方都是从线性代数的角度来讲。

