# RepVGG

***author: qiuchao***

# 

# 一，引入&介绍

目前大家想要得到相对SOTA的结果或者较低的计算量（FLOPs），都会设计比较复杂的模型结构来达到目标。但是这些复杂的结构很难去实际实现和修改，会增大内存访问的开销，并且对于不同设备的支持也不太友好。单独参考低FLOPs并不能说明模型推理的速度就更快。因此设计了一种更简洁的网络结构---RepVGG。

# 二，网络构建

这里所说的“VGG式”指的是：

1. 没有任何分支结构。即通常所说的plain或feed-forward架构。

2. 仅使用3x3卷积。

3. 仅使用ReLU作为激活函数。

下面用一句话介绍RepVGG模型的基本架构：将20多层3x3卷积堆起来，分成5个stage，每个stage的第一层是stride=2的降采样，每个卷积层用ReLU作为激活函数。

再用一句话介绍RepVGG模型的详细结构：RepVGG-A的5个stage分别有[1, 2, 4, 14, 1]层，RepVGG-B的5个stage分别有[1, 4, 6, 16, 1]层，宽度是[64, 128, 256, 512]的若干倍。这里的倍数是随意指定的诸如1.5,2.5这样的“工整”的数字，没有经过细调。

再用一句话介绍训练设定：ImageNet上120 epochs，不用trick，甚至直接用PyTorch官方示例的训练代码就能训出来！

# 三，VGG形式优点

1. 3x3卷积非常快。在GPU上，3x3卷积的计算密度（理论运算量除以所用时间）可达1x1和5x5卷积的四倍。

1. 单路架构非常快，因为并行度高。同样的计算量，“大而整”的运算效率远超“小而碎”的运算。

1. 单路架构省内存。例如，ResNet的shortcut虽然不占计算量，却增加了一倍的显存占用。

1. 单路架构灵活性更好，容易改变各层的宽度（如剪枝）。

1. RepVGG主体部分只有一种算子：3x3卷积接ReLU。在设计专用芯片时，给定芯片尺寸或造价，我们可以集成海量的3x3卷积-ReLU计算单元来达到很高的效率。别忘了，单路架构省内存的特性也可以帮我们少做存储单元。

# 四，训练和部署解耦

相比于各种多分支架构（如ResNet，Inception，DenseNet，各种NAS架构），近年来VGG式模型鲜有关注，主要自然是因为性能差。例如，有研究认为，ResNet性能好的一种解释是ResNet的分支结构（shortcut）产生了一个大量子模型的隐式ensemble（因为每遇到一次分支，总的路径就变成两倍），单路架构显然不具备这种特点。

既然多分支架构是对训练有益的，而想要部署的模型是单路架构，因此提出**解耦训练时和推理时架构**。通常使用模型的方式是：

1. 训练一个模型

2. 部署这个模型

但在这里，有一个新的做法：

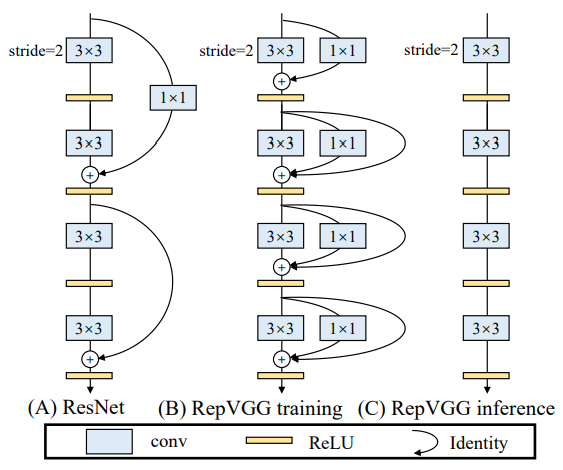
1. 训练一个多分支模型

2. 将多分支模型**等价转换**为单路模型

3. 部署单路模型

这样就可以同时利用**多分支模型训练时的优势**（**性能高**）和**单路模型推理时的好处**（**速度快、省内存**）。这里的关键在于这种多分支模型的构造形式和转换的方式。

具体实现方式是在训练时，为每一个3x3卷积层添加平行的1x1卷积分支和恒等映射分支，构成一个RepVGG Block。这种设计是借鉴ResNet的做法，区别在于ResNet是每隔两层或三层加一分支，RepVGG是每层都加。



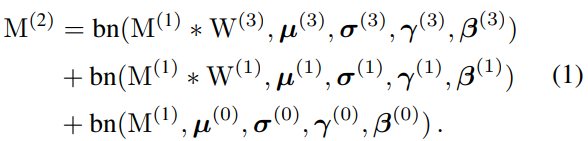
训练完成后，对模型做等价转换，得到部署模型。根据卷积的线性（具体来说是可加性），设三个3x3卷积核分别是W1，W2，W3，有 conv(x, W1) + conv(x, W2) + conv(x, W3) = conv(x, W1+W2+W3)）

因为RepVGG Block中的1x1卷积是相当于一个特殊（卷积核中有很多0）的3x3卷积，而恒等映射是一个特殊（以单位矩阵为卷积核）的1x1卷积，因此也是一个特殊的3x3卷积！只需要：

1. 把identity转换为1x1卷积，只要构造出一个以单位矩阵为卷积核的1x1卷积即可；

2. 把1x1卷积等价转换为3x3卷积，只要用0填充即可。

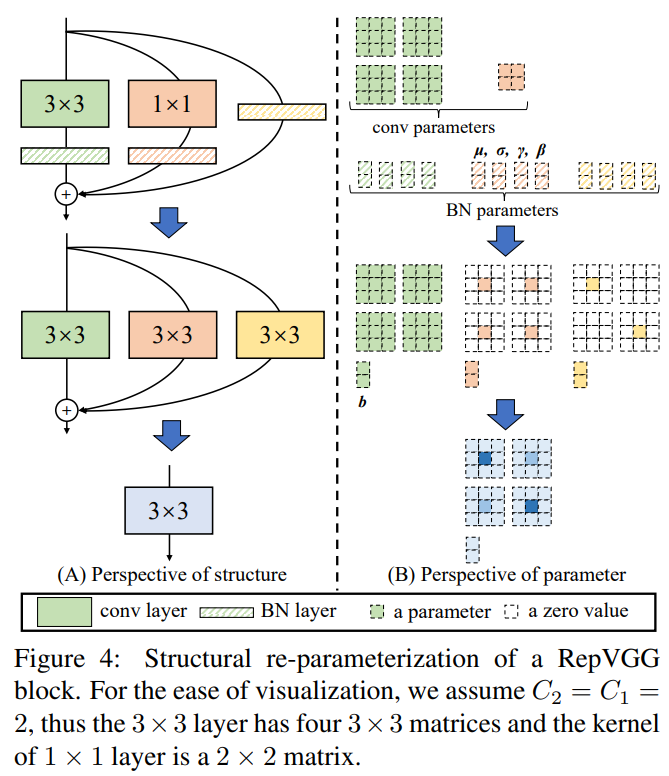
下图描述了这一转换过程。在这一示例中，输入和输出通道都是2，故3x3卷积的参数是4个3x3矩阵，1x1卷积的参数是一个2x2矩阵。注意三个分支都有BN（batch normalization）层，其参数包括累积得到的均值及标准差和学得的缩放因子及bias。这并不会妨碍转换的可行性，因为推理时的卷积层和其后的BN层可以等价转换为一个带bias的卷积层（也就是通常所谓的“吸BN”）。







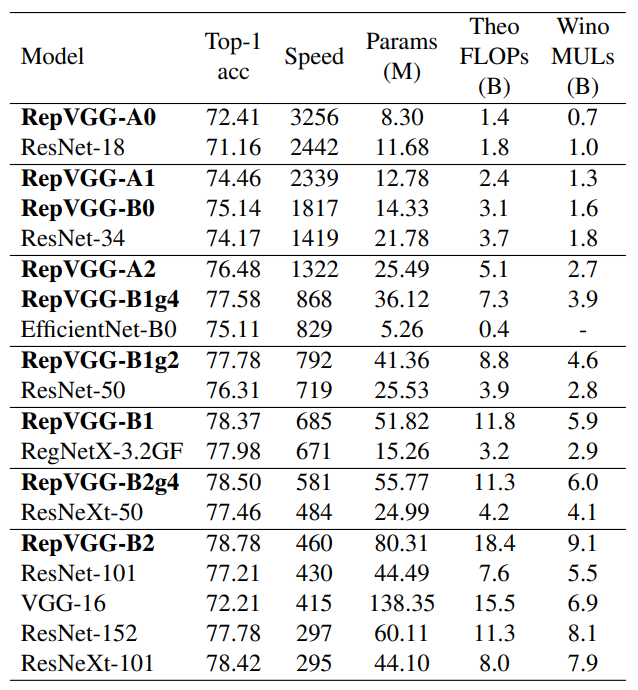
对三分支分别“吸BN”之后（注意恒等映射可以看成一个“卷积层”，其参数是一个2x2单位矩阵！），将得到的1x1卷积核用0给pad成3x3。最后，三分支得到的卷积核和bias分别相加即可。这样，每个RepVGG Block转换前后的输出完全相同，因而训练好的模型可以等价转换为只有3x3卷积的单路模型。



**结构重参数化”的实质：训练时的结构对应一组参数，推理时我们想要的结构对应另一组参数；只要能把前者的参数等价转换为后者，就可以将前者的结构等价转换为后者。**

# 五，实验结果

## 分类问题：



## 8155端侧测试：

