背景

深度网络比网络层数少的或者宽的网络更有效，出于这样的考虑，越来越多的研究倾向于精心设计的网络初始化策略、设计特殊的非线性激活函数、设计更好的优化器等方法用于训练更深的神经网络。当 ResNet、Inception 系列、DenseNet 等网络结构被提出，研究者发现使用多分支的思想设计很深的网络可以达到很好的性能。

但是多分支的网络可以给网络的训练带来很多的好处，在模型推理时有很多的缺点，因此提出了将多分支网络的训练和推理过程解耦，训练时使用多分支网络模型，推理时采用结构重参数化，将多分支网络转化成单路模型。

为什么推理过程要将多分支网络转化成单路模型？

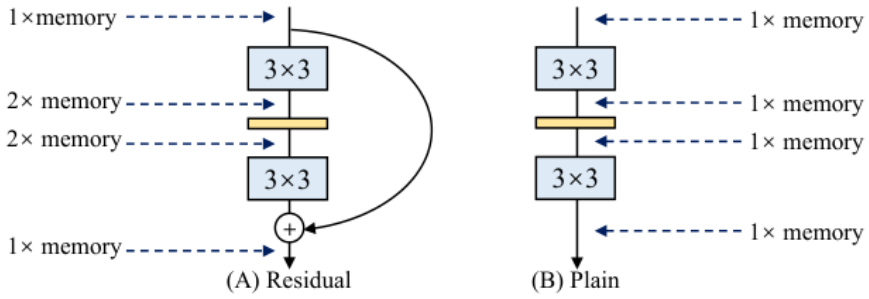
①从速度角度考虑

虽然多分支网络从 FLOPs 这个性能指标上看是优于单路网络模型的，但是实际上单路的运行速度比多分支的快。FLOPs 这个指标没有将内存访问成本(MAC,memory access cost)和并行程度考虑在内，而这两个因素对速度影响程度很大。对于多分支模型，硬件需要分别计算每个分支的结果，有的分支计算的快，有的分支计算的慢，而计算快的分支计算完后只能干等着，等其他分支都计算完后才能做进一步融合，这样会导致硬件算力不能充分利用，或者说并行度不够高。而且每个分支都需要去访问一次内存，计算完后还需要将计算结果存入内存（不断地访问和写入内存会在 IO 上浪费很多时间）

（此处参考博文 http://t.csdn.cn/07q41）

②从内存占用角度考虑

如下图的(A)结构所示，由于从输入引出了一个分支用于跳跃连接，所以在两路分支相加之前，都需要占用一个额外的内存，和(B)的单路结构相比，后者不需要保存中间的计算结果，所以更节省内存。



③优化灵活性

多分支网络在后期优化比单路结构受到的限制更多。

方法

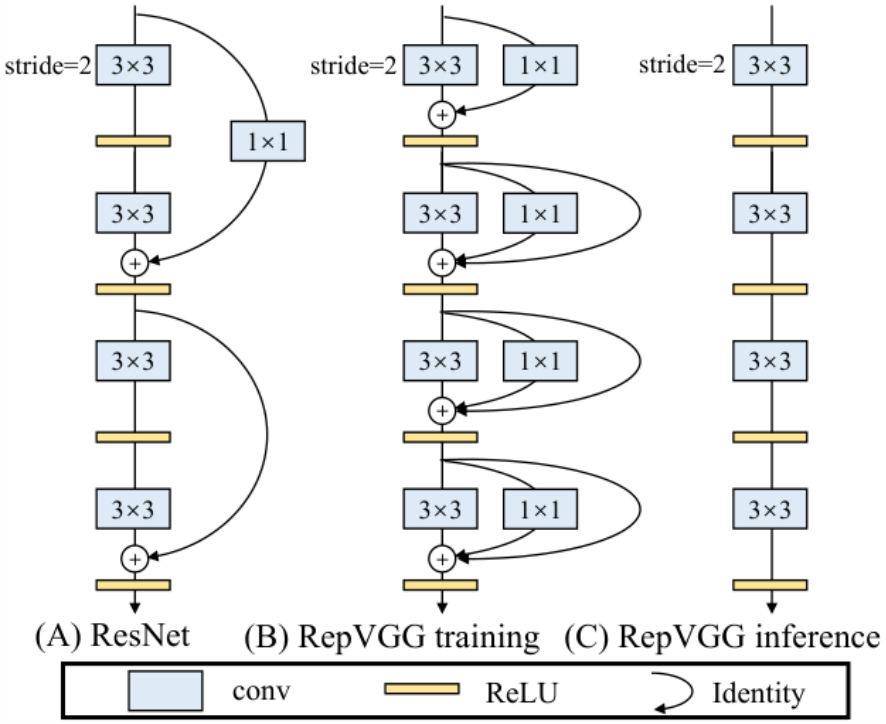
网络训练阶段

在网络训练时采用下图(B)的 RepVGG Block 结构，分为三路，一路是卷积核大小为 3×3 的卷积、一路是卷积核大小为 1×1 的卷积，一路是只有 BN 操

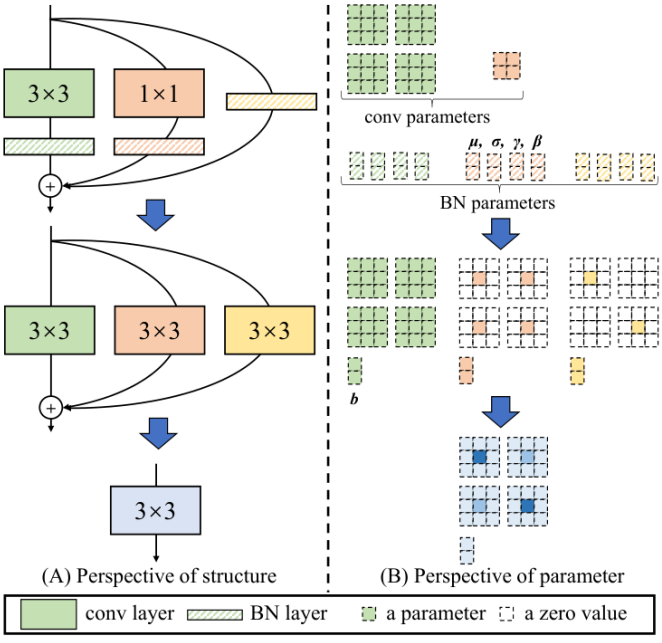
作。

✏网络推理阶段

模型推理采用下图(C)的单路结构。可以看出转化后的模型结构都采用的是3×3 的卷积，加速推理过程。



✏从 training👉inference 的结构变化



上图说明了结构重参数化的过程。

(A)图说明了结构重参数化的过程。首先将 bn 层和其前接卷积操作合并为一个带有偏置的卷积操作，以及将只有 bn 层的分支转换成卷积操作；其次将转后后的三个卷积操作融合成一个卷积。前述过程涉及到了 bn 和卷积的融合、将 1×1 卷积转换成 3×3 卷积、将 bn 转化成 3×3 卷积。

①bn 和卷积的融合

$$\mathrm{bn}(\mathrm{M},\boldsymbol{\mu},\boldsymbol{\sigma},\boldsymbol{\gamma},\boldsymbol{\beta})_{:,i,:,:}=(\mathrm{M}_{:,i,:,:}-\boldsymbol{\mu}_i)\frac{\boldsymbol{\gamma}_i}{\boldsymbol{\sigma}_i}+\boldsymbol{\beta}_i\,. \quad (2)$$

对于推理过程，特征图第 i 个通道的 bn 操作如(2)式所示，经转换后的权重和偏置如(3)所示：

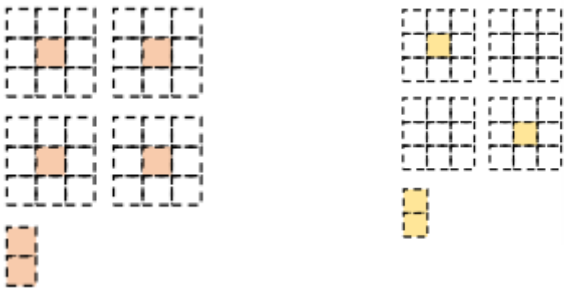
$$\mathrm{W}'_{i,:,:,:}=\frac{\boldsymbol{\gamma}_i}{\boldsymbol{\sigma}_i}\mathrm{W}_{i,:,:,:}\,,\quad \mathbf{b}'_i=-\frac{\boldsymbol{\mu}_i\boldsymbol{\gamma}_i}{\boldsymbol{\sigma}_i}+\boldsymbol{\beta}_i\,. \quad (3)$$

由于是推理过程，所以参数 γ，β，σ，μ 都是已知的，就相当于将第 i 个卷积核的权重乘以一个系数，第 i 个卷积核的偏置设为式(3)的 bi。

②将 1×1 卷积转化为 3×3 卷积

首先 1×1 卷积不改变输入特征图的大小，所以将其转换成 3×3 卷积时为了保证不改变特征图的大小，需要将 padding 设置为 1.

其次，1×1 卷积变为 3×3 卷积只需要将卷积核的权重参数扩充成 3×3 大小，将原来的权重参数放置在中心，其余位置的权重参数设置为 0，过程如下图的右图所示。再进行 bn 和卷积的融合操作。



③将单独的 bn 层转换为 3×3 卷积

由于单独的 bn 层实现的是恒等映射，所以在构建相应的卷积操作时，只需要将权重参数设置为 1，再利用 1×1 卷积变为 3×3 卷积时操作，将其转变为 3×3 卷积，过程如上图的左图所示。再进行 bn 和卷积的融合操作。

总结

在训练时利用多分支强大的特征表征能力，在推理时将多分支等价转换为单分支加速推理过程。

可以将结构重参数化的思想用到自己的模型推理中。