基于 RKNN 的深度神经网络模型设计建议

本文重点介绍了如何设计卷积神经网络，使其能在 RKNN 上实现最佳性能。

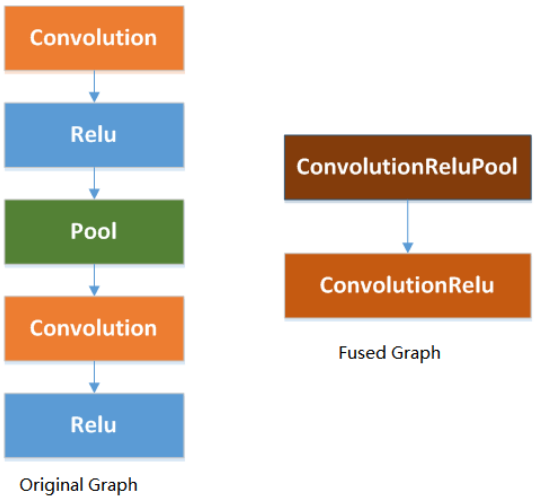
1. 卷积核设置

推荐在设计的时候尽量使用 3x3 的卷积核，这样可以实现最高的 MAC 利用率，使得 NPU 的性能最佳。

我们的 NPU 也可以支持大范围的卷积核。 支持的最小内核大小为[1]，最大值为[11 * stride - 1]。同时 NPU 也支持非平方内核，不过会增加一些额外的计算开销。

2. 融合结构设计

NPU 会对卷积后面的 ReLU 和 MAX Pooling 进行融合的优化操作，能在运行中减少计算和带宽开销。所以在搭建网络时，能针对这一特性，进行设计。



在设计网络时，卷积层后面的 ReLU 层将都会被融合。不过为了确保 MAX Pooling 层也能进行融合加速，需要尽量按照下面规则进

行设计：

(1) pool size 必须是 2x2 或者 3x3，而步长 stride=2

(2) 2x2 池化的输入图片尺寸必须是偶数，而且不能有填充

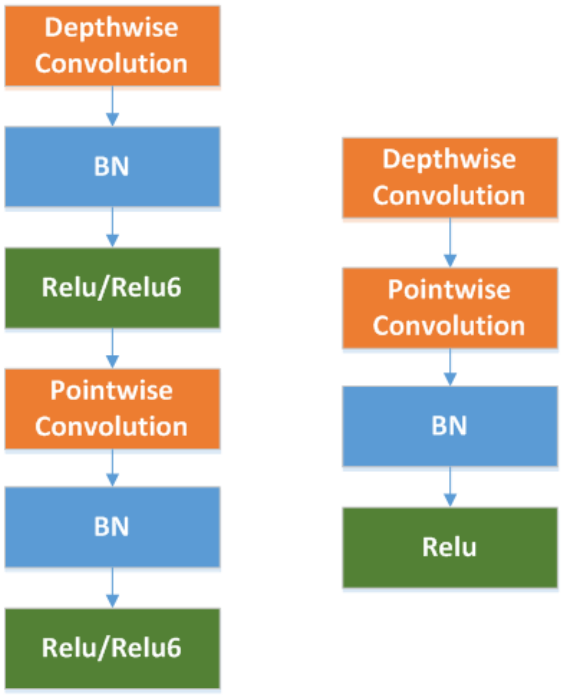
(3) 3x3 池化的输入图片尺寸必须是非 1 的奇数，而且不能有填充

(4) 如果是 3x3 的池化，则水平输入大小必须小于 64（8-bit 模型）或 32（16-bit 模型）

3. 关于 2D 卷积和 Depthwise 卷积的使用

NPU 支持常规 2D 卷积和 Depthwise 卷积加速。由于 Depthiwise 卷积特定的结构，使得它对于我们量化(int8)模型不太友好，而 2D 卷积的优化效果更好。我们设计网络时建议尽量使用 2D 卷积。

如果必须使用 Depthwise 卷积，我们建议按照下面的规则进行修改，能提高量化后模型的精度:



(1)如果网络中的激活函数使用的是 ReLU6，建议将其都改为 ReLU

(2)在 Depthwise 卷积层的 BN 层和激活层，建议去除。

(3)在训练时，针对 Depthwise 卷积层，对它的权重进行 L2 正则化

4. 网络稀疏化

当前的神经网络存在过度参数化现象，并且在其设计时会存在很多冗余。我们的 NPU 针对稀疏矩阵，有进行跳零计算和内存提取方面的优化。所以建议在设计网络的时候，可以针对性的进行网络稀疏化设计，以利用该技术进一步提高网络性能。