depthwise 加速实验（张少鹏）

1、理论依据

详细的计算量分析可参考 csdn 帖子 http://blog.csdn.net/u014380165/article/details/72938047?utm_source=itdadao&utm_medium=referral

论文：MobileNets Efficient Convolutional Neural Networks for Mobile Vision Applications

论文链接：https://arxiv.org/abs/1704.04861

2、实验过程

我们在 cifar10 数据集上进行了实验，初始模型为 3+2（3 个传统卷积层，两个全连接层），对初始模型在 cifar10 上进行训练，训练到测试精度约为 61%时保存模型，训练时学习率从 10-4 等比下降到 10-8。然后进行了如下几组实验：

2.1、将第一卷积层替换为 depthwise 的形式，随机初始化，第二三卷积层和全连接层用已训练并保存的初始模型的对应参数初始化

a）、对新模型进行训练，梯度下降的变量为 depthwise 形式的第一卷积层的参数，第二三卷积层和全连接层参数不变，训练了 13 万轮，准确率只有 37%左右，遂停止

b）、对新模型进行训练，梯度下降的变量为所有的参数，训练了 8 万轮，准确率达到 63%，对测试集的 10000 张图片，每次取一张图片预测，累计测试完 10000 张的时间，对比新模型(depthwise)和初始模型(gen)时间，重复进行了 10 次，时间(seconds)如下:

depthwise	16.207	15.623	15.645	15.693	15.680	15.735	15.767	15.681	15.678	15.765
gen	15.521	14.985	14.969	14.987	15.046	14.996	15.055	15.141	15.099	15.053

加速效果不好，可能的原因是第一卷积层的计算量占所有卷积层的计算量比例较小，depthwise 将一次卷积变为两次卷积可能增加的额外时间开销多余减小的计算开销

2.2、将第一、二卷积层替换为 depthwise 的形式，随机初始化，第三卷积层和全连接层用已训练并保存的初始模型的对应参数初始化

a）、对新模型进行训练，梯度下降的变量为 depthwise 形式的第一、二卷积层的参数，第三卷积层和全连接层参数不变，训练了 2 千轮，效果不太好，遂停止

b）、对新模型进行训练，梯度下降的变量为所有的参数，训练了 5 万轮，准确率只达到 47%，测试方法同 2.1（b），时间统计如下（时间加快不到半秒）；

depthwise	15.356	14.774	14.788	14.888	14.819
gen	15.727	15.189	15.146	15.161	

2.3、将第三卷积层替换为 depthwise 的形式，随机初始化，第一、二卷积层和全连接层用已训练并保存的初始模型的对应参数初始化

a）、同上效果不好

b）、对新模型进行训练，梯度下降的变量为所有的参数，训练了 7 万轮，准确率达到 61%，测试方法同 2.1（b），时间统计如下（时间加快了半秒左右）；

depthwise	14.608	14.087	14.089	14.136	14.125	14.115	14.121	14.162	14.136
gen	15.116	14.534	14.569	14.563	14.622	14.646	14.650	14.654	14.635

2.4、将第二、三卷积层替换为 depthwise 的形式，随机初始化，第一卷积层和全连接层用已训练并保存的初始模型的对应参数初始化

a）、对新模型进行训练，梯度下降的变量为所有的参数，训练了 5 万轮，准确率达到 52%（继续训练准确率还可以提升，预计能达到 60%左右），测试方法同 2.1（b），时间统计如下（时间加快了 1 秒左右）；

depthwise	14.665	14.122	14.162	14.150	14.133
gen	15.658	15.133	15.134	15.149	15.178

b）、对 1 中的方法对 a 中计算量分析，计算量至少减少为原先的 1/3,而时间减少不成比例，怀疑是文件读取、内外存访问用时开销较大，遂在 a 的基础上增加了一组测试，将测试集每次取一张图片进行预测改为每次取 100 张图片进行测试，对测试集进行了 6 轮比对，时间如下：

depthwise	5.494	4.974	4.976	4.975	4.977	4.988
gen	5.618	5.050	5.085	5.060	5.040	5.055

3、实验结论

3.1、depthwise 可以有效的减少计算量，从而减少计算时间，本例中计算时间的减少和计算量的减少出入较大，推测原因主要和数据集较小，计算时间在程序运行时间中占比较少有较大关系

3.2、将传统卷积替换为 depthwise 卷积时，尽量选择计算量大，且比较深的层，比较浅的层保留了较多的原始信息且计算量较少，不适宜做 depthwise 更改

3.3、模型某些层替换为 depthwise 形式之后，要做 fine-tune