

CSDN新首页上线啦，邀请你来立即体验！(http://blog.csdn.net/)

广告

立即体验

CSDN

博客 (//blog.csdn.net?ref=toolbar) 学院 (//edu.csdn.net?ref=toolbar)

下载 (//download.csdn.net?ref=toolbar)

GitChat (//gitbook.cn/?ref=csdn) 更多 ▾

1

🔍

📌

📁

登录 (https://passport.csdn.net/account/login?ref=toolbar) (//write.blog.csdn.net/postednews/toolbar)

注册 (https://passport.csdn.net/account/mobile/register?ref=toolbar&action=mobileRegister) (/activity?utm_source=csdnblog1)

论文笔记: DeepRebirth——从非权重层入手来进行模型压缩

原创 2017年07月25日 20:34:35

标签: 深度学习 (http://so.csdn.net/so/search/s.do?q=深度学习&t=blog) /

压缩 (http://so.csdn.net/so/search/s.do?q=压缩&t=blog) /

论文笔记 (http://so.csdn.net/so/search/s.do?q=论文笔记&t=blog)

XlyPb (http://blog.csdn.n...)

+ 关注

(http://blog.csdn.net/wspba)

原创

粉丝

喜欢

未开通

43

215

4

(https://gite

他的最新文章

更多文章 (http://blog.csdn.net/wspba)

从神经网络到深度学习（一） (http://blog.csdn.net/wspba/article/details/77621112)

论文笔记: ThiNet——一种filter级的模型裁剪算法 (http://blog.csdn.net/wspba/article/details/77427960)

深度学习模型压缩方法综述（三） (http://blog.csdn.net/wspba/article/details/76039135)

深度学习模型压缩方法综述（二） (http://blog.csdn.net/wspba/article/details/75675554)

深度学习模型压缩方法综述（一） (http://blog.csdn.net/wspba/article/details/75671573)

相关推荐

DeepRebirth——通过融合加速网络 (http://blog.csdn.net/shuzfan/article/details/53139224)

ImageNet中的LRN（Local Response Normalization） (http://blog.csdn.net/searobbers_duck/article/details/51645941)

LRN层的实现 (http://blog.csdn.net/u014696921/article/details/52873661)

内容举报

如何在Caffe中配置每一个层的结构 (http://blog.csdn.net/sherry_gp/article/details/50924481)

返回顶部

前言

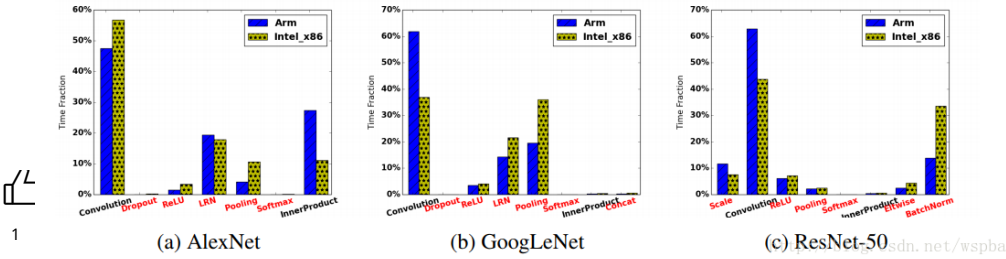
因为最近都在关注深度学习模型压缩相关的工作，所以今天给大家介绍的也是一篇关于模型压缩的方法。这是一篇非常有意思的工作，来自于三星研究院在ICLR2017上发表的论文：DeepRebirth: A General Approach for Accelerating Deep Neural Network Execution on Mobile Devices (https://openreview.net/pdf?id=SkwSJ99ex)。常用的模型压缩手段往往是从网络的参数，也就是weight入手（可以参考本人之前的几篇关于模型压缩方法的综述性博文），因为weight占用了大量的计算资源和存储空间，这往往是最直接有效的。而本文却另辟蹊径，从非权重层入手来进行模型压缩。

动机

作者认为，目前常用的模型，如ResNet、GoogLeNet等的卷积层都是由很小的卷积核组成，本身就非常紧致了，并且也去掉了非常占参数量的全连接层。而Non-tensor layer（也就是非权重层，如pooling、BN、LRN、ReLU等等）反而成为了模型在cpu以及其他嵌入式硬件上达到real-time的最大阻碍，见下图表：

Table 1: Percentage of Forwarding Time on Non-tensor Layers

Network	Intel x86	Arm	Titan X
AlexNet	32.08%	25.08%	22.37%
GoogLeNet	62.03%	37.81%	26.14%
ResNet-50	55.66%	36.61%	47.87%
ResNet-152	49.77%	N/A	44.49%
Average	49.89%	33.17%	35.22%

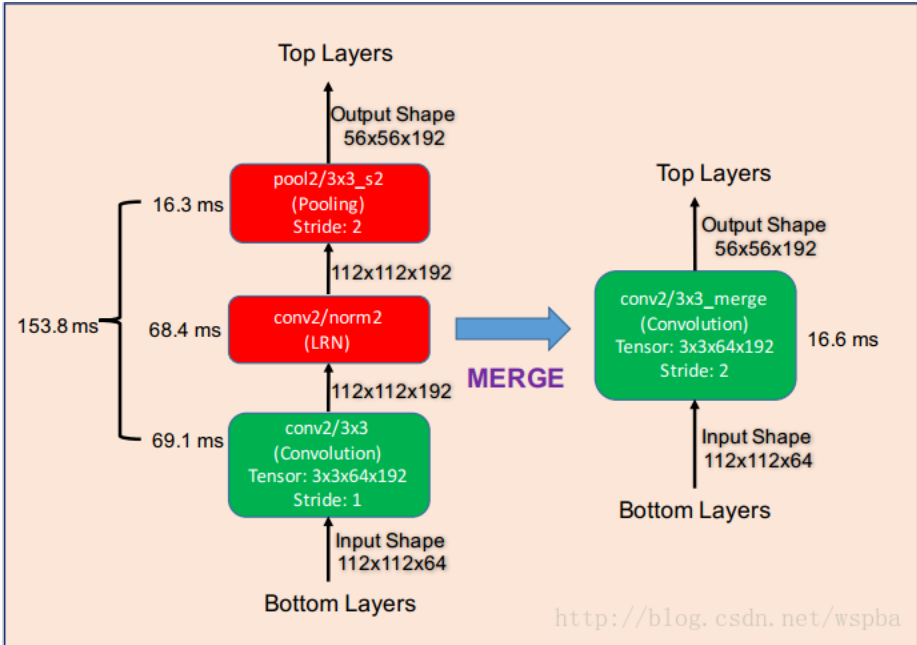


可见无论在哪一种硬件平台上，这些Non-Tensor层都占用了非常多的运行时间。因此作者认为，如果能够找到一种方法将这些Non-Tensor层剔除，就能够大幅度的提升模型的运算时间。

方法

作者提出了两种剔除Non-Tensor层的方法：StreamLine Merging和Branch Merging，如下图：

- StreamLine Merging



所谓的StreamLine，流水线，顾名思义，就是将这一连串(layer)合并起来，如上图，作者将这里的Non-Tensor层 (Pooling、LRN) 与它们相邻近的Tensor层 (Conv) 合并在一起，合并的方式也非常简单粗暴：对于Pooling层，将stride直接乘到Conv层中，然后将Pooling去掉；对于其他非Pooling层，如 BN、LRN、ReLU直接去掉。如上图中，这样一个合并的过程将运行时间153.8ms降到了16.6ms。

- Branch Merging



Unable to Conn

The Proxy was unable to connect to the remote site. responding to requests. If you feel you have reached please submit a ticket via the link provided below.

URL: <http://pos.baidu.com/s?hei=250&wid=300&di=u%2Fblog.csdn.net%2Fwspba%2Farticle%2Fdetails%2F76098493>

博主专栏

 The Path to Deep Learning
(<http://blog.csdn.net/wspba/article/details/14602>)
/column 32049

 论文笔记
(<http://blog.csdn.net/wspba/article/details/14646>)
/column 75012

他的热门文章

ResNet论文笔记 (<http://blog.csdn.net/wspba/article/details/56019373>)
19449

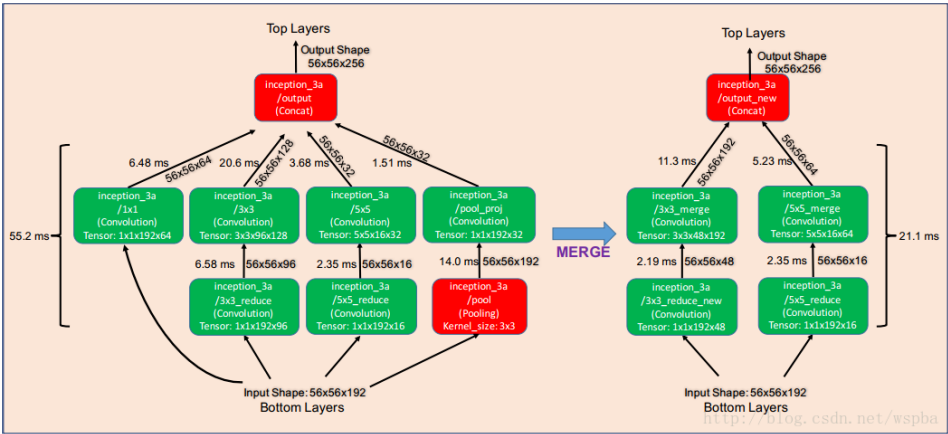
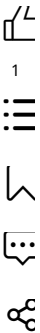
TensorFlow官方教程学习笔记（一）——起步 (<http://blog.csdn.net/wspba/article/details/54089132>)
9737

Generative Adversarial Nets (译) (<http://blog.csdn.net/wspba/article/details/54577236>)
5996

InfoGAN论文笔记+源码解析 (<http://blog.csdn.net/wspba/article/details/54808833>)
4240

DCGAN论文笔记+源码解析 (<http://blog.csdn.net/wspba/article/details/54730871>)
4196

广告



这种融合主要针对GoogleNet中的Inception结构。GoogleNet虽然参数比较少，但是由于细小层比较多，参数比较碎，因此运行起来速度也收到了一些阻碍。作者将比较细小的卷积层（1*1）以及Pooling层所在的分支，直接合并到和它并排的大卷积分支中，而合并后的分支的输出通道也进行了相应的增加。考虑到大卷积（3*3或5*5）的输出通道增加，会增大参数量和运算量，因此作者又将这些分支的输入通道进行了一定的缩减，以达到一个降低参数量和运算量的作用。

以上两种方法就是作者提出的将Non-Tensor层搞掉的方法，但是搞掉之后，模型的性能必然会降低（实验都是在预训练模型的基础上进行的），因此作者在这里使用一个retraining的方式来恢复性能，这个retraining的方式也其他模型压缩方法中所用到的retraining都有一些不同。对于一个预训练模型，作者逐层进行合并，合并得到的新层使用标准的初始化方式，其他层的参数保留原预训练模型的参数，然后将新层的学习率调高为其他层的10倍，进行finetuning，对于某些层，如GoogLeNet中的Inception 4b-4d可以一起进行合并并在finetuning。

实验结果

作者使用GoogLeNet在ImageNet上的结果如下表：

Step	Merged Layer(s)	Top-5 Accuracy
0	N/A	88.89%
1	conv1	88.73%
2	conv2	88.82%
3	inception_3a	88.50%
4	inception_3b	88.27%
5	inception_4a	88.60%
6	inception_4b-4d	88.61%
7	inception_4e	88.43%
8	inception_5a	88.41%
9	inception_5b	88.43%
Tucker Decomposition	N/A	86.54%

在经过一系列的合并之后，模型的准确率仅仅降低了0.4%。那么加速效果如何呢：

广告

Device	GoogLeNet	GoogLeNet -Tucker	GoogLeNet -Merge	GoogLeNet -Merge-Tucker
conv1	94.92 ms	87.85 ms	8.424 ms	6.038 ms
conv2	153.8 ms	179.4 ms	16.62 ms	9.259 ms
inception_3a	55.23 ms	85.62 ms	21.17 ms	9.459 ms
inception_3b	98.41 ms	66.51 ms	25.94 ms	11.74 ms
inception_4a	30.53 ms	36.91 ms	16.80 ms	8.966 ms
inception_4b	32.60 ms	41.82 ms	20.29 ms	11.65 ms
inception_4c	46.96 ms	30.46 ms	18.71 ms	9.102 ms
inception_4d	36.88 ms	21.05 ms	24.67 ms	10.05 ms
inception_4e	48.24 ms	32.19 ms	28.08 ms	14.08 ms
inception_5a	24.64 ms	14.43 ms	10.69 ms	5.36 ms
inception_5b	24.92 ms	15.87 ms	14.58 ms	6.65 ms
loss3	3.014 ms	2.81 ms	2.97 ms	2.902 ms
Total	651.4 ms	614.9 ms (1.06x)	210.6 ms (3.09x)	106.3 ms (6.13x)

上表是作者在Samsung Galaxy S5上测试得到的结果，发现经过一系列的合成，将运行所消耗的时间降低了3倍以上，并且这个合并的方法也可以和之前的一种叫做Tucker分解的方法一起使用，可以将时间消耗降低到6倍以上。

在不同硬件平台上的效果：

Device	GoogLeNet	GoogLeNet -Tucker	GoogLeNet -Merge	GoogLeNet -Merge-Tucker	SqueezeNet
Moto E	1168.8 ms	897.9 ms	406.7 ms	213.3 ms	291.4 ms
Samsung Galaxy S5	651.4 ms	614.9 ms	210.6 ms	106.3 ms	136.3 ms
Samsung Galaxy S6	424.7 ms	342.5 ms	107.7 ms	65.34 ms	75.34 ms
Macbook Pro (CPU)	91.77 ms	78.22 ms	23.69 ms	15.18 ms	17.63 ms
Titan X	10.17 ms	10.74 ms	6.57 ms	7.68 ms	3.29 ms

我们发现，由于GPU (Titan X) 强大的计算性能，这些Non-Tensor层所带来的时间消耗可能看起来没什么，但是在便携式的硬件平台上，如手机，这加速的效果就非常明显了。这也使得深度学习模型能在便携式硬件上达到real-time又更近了一步。

总结

本文确实是一篇非常有趣的论文，也是一项很有启发性的工作，我们在设计模型包括压缩和使用模型时，不能光考虑理论计算量，还应该考虑到硬件特性、带宽等因素，就比如在这里，Pooling层本来没有任何参数，不会带来任何的理论计算量，但是在CPU上却会带来额外的时间消耗。但是本文这种retraining的方式可能会给模型的性能大打折扣，因为新合并的层需要重新进行训练，虽然其它层的参数以及达到了一个非常好的状态，但是这个新的合并层能训到一个什么样的效果，在这里还是持怀疑态度。不过总的来说，这篇文章无论是从创新点还是实用性上，都是一项非常有意义的工作。

版权声明：本文为博主原创文章，未经博主允许不得转载。

本文已收录于以下专栏：论文笔记 (<http://blog.csdn.net/column/details/14646.html>)



内容举报



返回顶部



加入CSDN，享受更精准的内容推荐，与500万程序员共同成长！

登录

注册





广告

相关文章推荐



DeepRebirth——通过融合加速网络 (<http://blog.csdn.net/shuzfan/article/details/53139224>)

这里介绍2017ICLR OpenReview中的一篇有关网络加速的文章《DeepRebirth: A General Approach for Accelerating Deep Neural Ne...

 shuzfan (<http://blog.csdn.net/shuzfan>) 2016年11月12日 12:19  2133



ImageNet中的LRN（Local Response Normalization）(http://blog.csdn.net/searobbers_...)

LRN（Local Response Normalization）神经网络初学者，没有什么理论基础，偶然看到个ImageNet，就准备从其入手，先弄懂每层的含义，其中这个LRN层真是让人百思不得其...

 searobbers_duck (http://blog.csdn.net/searobbers_duck) 2016年06月12日 14:28  41622



LRN层的实现 (<http://blog.csdn.net/u014696921/article/details/52873661>)

版权声明：本文为卜居原创文章，未经博主允许不得转载。卜居博客地址：<http://blog.csdn.net/kkk584520> LRN全称为Local Response Normalizat...

 u014696921 (<http://blog.csdn.net/u014696921>) 2016年10月20日 16:08  5623



如何在Caffe中配置每一个层的结构 (http://blog.csdn.net/sherry_gp/article/details/50924...)

from: <http://demo.netfoucs.com/danieljianfeng/article/details/42929283> 最近刚在电脑上装好Caffe，由于神经网络中有不同的层结...

 sherry_gp (http://blog.csdn.net/sherry_gp) 2016年03月18日 16:12  895



Isolation Forest算法原理详解 (<http://blog.csdn.net/u013709270/article/details/73436588>)

本文只介绍原论文中的 Isolation Forest 孤立点检测算法的原理，实际的代码实现详解请参照我的另一篇博客：Isolation Forest算法实现详解。 或者读者可以到我的Gi...

 u013709270 (<http://blog.csdn.net/u013709270>) 2017年06月18日 18:39  5558



DeepRebirth——通过融合加速网络 (<http://blog.csdn.net/shuzfan/article/details/53139224>)

这里介绍2017ICLR OpenReview中的一篇有关网络加速的文章《DeepRebirth: A General Approach for Accelerating Deep Neural Ne...

 shuzfan (<http://blog.csdn.net/shuzfan>) 2016年11月12日 12:19  2133

keras学习笔记4——部分连接非共享权重层 (<http://blog.csdn.net/xyzhang16/article/deta...>)

最近一直在纠结一个问题，n久没有解决，幸运的是终于在昨天取得了突破，下面就听我细细道来。在工业应用中，模型效果的好坏主要取决于相关特征的提取，而在各类数据挖掘竞赛中，竞赛人员也将大部分比赛时间应用...

 xyzhang16 (<http://blog.csdn.net/xyzhang16>) 2016年06月28日 10:57  1203



内容举报



返回顶部



广告

信息检索导论学习笔记(6)-文档评分,词项权重计算及向量空间模型 (http://blog.csdn.net...

参数化索引及域索引迄今为止,我们都将文档看成一系列词项的序列.实际上,大多数文档都具有额外的结构信息.数字文档通常会把与之相关的元数据(metadata)以机读的方式一起编码.所谓元数据,指的是和文...

zins26914 (http://blog.csdn.net/zins26914) 2013年08月20日 15:47 2399



信息检索导论——六、文档评分、词项权重计算及向量空间模型 (http://blog.csdn.net/u...

参数化索引及域索引迄今为止,我们都将文档看成一系列词项的序列.实际上,大多数文档都具有额外的结构信息.数字文档通常会把与之相关的元数据(metadata)以机读的方式一起编码.所谓元数据,指的...

u013952285 (http://blog.csdn.net/u013952285) 2016年07月11日 17:00 717



tensorflow将训练好的模型freeze,即将权重固化到图里面,并使用该模型进行预测 (http://...

ML主要分为训练和预测两个阶段,此教程就是将训练好的模型freeze并保存下来.freeze的含义就是将该模型的图结构和该模型的权重固化到一起了.也即加载freeze的模型之后,立刻能够使用了。下面...

c2a2o2 (http://blog.csdn.net/c2a2o2) 2017年05月27日 09:47 1612

运动物体检测与跟踪——累积权重构建背景模型 (http://blog.csdn.net/dcrmg/article/det...

运动物体检测与跟踪中的帧差分法,除了相邻帧差分法和三帧差分法外,还有一种差分方法,可以通过建立不含前景的背景模型,用当前帧和背景模型做差,差值就可以体现运动物体大概的位置和大小信息。相比相邻帧差分法...

dcrmg (http://blog.csdn.net/dcrmg) 2016年08月21日 00:25 2199

tensorflow将训练好的模型freeze,即将权重固化到图里面,并使用该模型进行预测 (http://...

ML主要分为训练和预测两个阶段,此教程就是将训练好的模型freeze并保存下来.freeze的含义就是将该模型的图结构和该模型的权重固化到一起了.也即加载freeze的模型之后,立刻能够使用了。下...

lujiandong1 (http://blog.csdn.net/lujiandong1) 2016年11月28日 21:54 9746

tensorflow将训练好的模型freeze,即将权重固化到图里面,并使用该模型进行预测 (http://...

转载来自: http://blog.csdn.net/lujiandong1/article/details/53385092 ML主要分为训练和预测两个阶段,此教程就是将训练好的模型...

zhyl3038 (http://blog.csdn.net/zhyl3038) 2017年04月07日 17:22 392

数学笔记18——定积分的应用3（均值、权重、概率） (http://blog.csdn.net/sunbobosu...

均值均值与定积分的关系 在数学笔记14——微积分第一基本定理中曾介绍过定积分与均值关系,如果y = f(x),则当n → ∞时: 用定积分的几何意义解释这个等式,如下图...

sunbobosun56801 (http://blog.csdn.net/sunbobosun56801) 2017年11月09日 15:30 133

内容举报
返回顶部

Fielding的博士论文学习笔记（二）——概念和术语与现实网络模型的对应关系 (http://...

REST架构的主要元素分为三类,分别是Data Element、Connector、Component。以下是三种元素中主要术语所对应的网络现实模型中的对象。核心概念: Resour...

tf718339 (http://blog.csdn.net/tf718339) 2012年12月21日 09:27 485

加入CSDN,享受更精准的内容推荐,与500万程序员共同成长!

登录 注册 X

广告

- 

利用交叉权重质心对陈涛ECandARC进行改进 (http://download.csdn.n...

http://download.csdn.net/detail/tfg1025/1251111

2012年06月15日 14:58

944KB

下载 (0)
- 

论文阅读笔记——使用双向PCA进行行人检测 (http://blog.csdn.net/tfg1025/article/detail...

题目: Novel and efficient pedestrian detection using bidirectional PCA 作者: Thi-Hai-Binh Nguyen, Hakil K...

 tfg1025 (http://blog.csdn.net/tfg1025)

2014年03月02日 11:34

 813
- 

利用粗糙集确定权重利用粗糙集进行约减 (http://download.csdn.net/do...

http://download.csdn.net/detail/tfg1025/1251111

2009年06月18日 18:40

1.08MB

下载 (0)
- 

基于向量空间模型的文本分类特征权重算法研究_苏力华 (http://downl...

http://download.csdn.net/detail/tfg1025/1251111

2015年05月17日 17:11

254KB

下载 (0)
- 


"Gradient Domain Guided Image Filtering"论文中边缘权重函数matlab实现代码 (http://...


"Kou F, Chen W, Wen C, et al. Gradient Domain Guided Image Filtering[J]. Image Processing, IEEE Tran...

majinlei121 (http://blog.csdn.net/majinlei121)

2016年01月04日 20:31

 1408


内容举报


返回顶部