

(/apps/redirect?
utm_source=side-
banner-click)

TensorFlow Mobile模型压缩



Jcme \ Ls (/u/51d83682b7b9) + 关注

2017.06.29 16:22* 字数 1799 阅读 1616 评论 2 喜欢 5

(/u/51d83682b7b9)

前言

前文中我们把训练好的模型打包成GraphDef文件（PB文件）了，可是打包出来的文件还是有点大；移动设备的内存容量有限，而且我们需要下载模型到移动端去加载，所以一个大的模型要经过压缩之后才能放在Android或者IOS上跑的，不然会给RAM造成极大的负担，系统会自行kill掉一些进程，甚至你的APP自己crash了。所以模型压缩是十分必要的。

正文

在Google的优化方案中，主要有那么几种优化方式：删除未使用节点，删除training所需的ops，对权重进行四舍五入运算，对BN进行预处理，内存映射，以及终极伤敌1000自损800的方案——转换权重数据类型。本文将逐一介绍。

optimize_for_inference

调用optimize_for_inference脚本会删除输入和输出节点之间所有不需要的节点。同时该脚本还做了一些其他优化以提高运行速度。例如它把显式批处理标准化运算跟卷积权重进行了合并，从而降低了计算量。

使用方法：

```
bazel build tensorflow/python/tools:optimize_for_inference

bazel-bin/tensorflow/python/tools/optimize_for_inference \
-input=/tf_files/CTNModel.pb \
-output=/tf_files/optimized_graph.pb \
-input_names=inputs/X \
-output_names=output/predict
```

首先用bazel build出我们的optimize_for_inference脚本，再调用这个脚本，提供几个参数：输入的PB文件路径，输出的PB文件路径，输入节点名以及输出节点名。

经过这一次的优化，文件变小了一些，可是还不足以我们放到手机端去运行，所以我们要进一步的压缩模型，同时还要保证准确率。

我们可以使用之前的在Python端测试PB文件的代码测试优化后的PB文件是否可行。

quantize_graph

在IOS或者Android项目中，人们通常会把PB文件放在assets文件夹中加载，不管是一起打包进APP还是进入APP后再进行下载、解压、复制，在经过optimize_for_inference优化过后的模型依然是非常的大的。IOS在使用.ipa包发布APP的时候，所有内容都会经过zip压缩；Android从网络端下载文件也要经过压缩后解压的过程，所以有没有一种行之有效的方法在不过多的降低精确度的情况下压缩更大的空间呢？Google提供了这么一个脚本，经过这个脚本的PB文件原本的大小不会改变，但会有更多的可利用的重复性，所以压缩成zip包会缩小大约3~4倍的大小。使用方式很简单：



```
bazel build tensorflow/tools/quantization:quantize_graph

bazel-bin/tensorflow/tools/quantization/quantize_graph \
-input=/tf_files/optimized_graph.pb \
-output=/tf_files/rounded_graph.pb \
-output_node_names=output/predict \
-mode=weights_rounded
```

(/apps/redirect?utm_source=side-banner-click)

输入的参数依然是：输入的PB文件路径，输出的PB文件路径，输出节点名，这里还有个特别的参数mode，这个参数是告诉脚本我们选择哪种压缩方式，这里我们选择了对权重进行四舍五入。

graph_transforms

在2017年的Google I/O大会上，曾提到过的graph_transforms是一整套优化工具，基本上所有的优化方案都能在这里找到，包括之前的那两种优化方式。

Optimizing for Deployment

这个方式类似前面的optimize_for_inference，只是被整合到了transforms中，脚本进行了删除前向传播过程中未调用到的节点，通过预先乘卷积的权重来优化BN中的一些乘法运算。

示例代码：

```
bazel build tensorflow/tools/graph_transforms:transform_graph

bazel-bin/tensorflow/tools/graph_transforms/transform_graph \
--in_graph=CTNModel.pb \
--out_graph=optimized_graph.pb \
--inputs='inputs/X' \
--outputs='output/predict' \
--transforms='
strip_unused_nodes(type=float, shape="256*64")
remove_nodes(op=Identity, op=CheckNumerics)
fold_constants(ignore_errors=true)
fold_batch_norms
fold_old_batch_norms'
```

transforms的参数代表着想要做的操作，这里主要就是删除未调用的节点，以及优化BN算法，而且必须先运行fold_constants。

Fixing Missing Kernel Errors

由于TensorFlow在Mobile中使用的时候，默认情况下是只能推理预测，可是在build so文件、jar文件或者.a文件的时候，依赖文件是写入在tensorflow / contrib / makefile / tf_op_files.txt中的，里面还包含了一些training相关的ops，这些可能会导致我们在加载PB文件的时候报错—No OpKernel was registered to support Op，这时候我们可以通过这个脚本来修复

实际上，经过上面的脚本，已经删除了这部分节点，不需要重复使用。

示例代码：

```
bazel build tensorflow/tools/graph_transforms:transform_graph

bazel-bin/tensorflow/tools/graph_transforms/transform_graph \
--in_graph=CTNModel.pb \
--out_graph=fixed_kernel_graph.pb \
--inputs='inputs/X' \
--outputs='output/predict' \
--transforms='
  strip_unused_nodes(type=float, shape="256*64")
  fold_constants(ignore_errors=true)
  fold_batch_norms
  fold_old_batch_norms'
```

(/apps/redirect?
utm_source=side-
banner-click)

round_weights

上文中提及的quantize_graph，其实在graph_transforms也有，在graph_transforms中我们应该如何使用呢？

示例代码：

```
bazel build tensorflow/tools/graph_transforms:transform_graph

bazel-bin/tensorflow/tools/graph_transforms/transform_graph \
--in_graph=CTNModel.pb \
--out_graph=rounded_graph.pb \
--inputs='inputs/X' \
--outputs='output/predict' \
--transforms='
  strip_unused_nodes(type=float, shape="256*64")
  fold_constants(ignore_errors=true)
  fold_batch_norms
  fold_old_batch_norms
  round_weights(num_steps=256)'
```

Eight-bit

在graph_transforms中有一种更加残暴的文件大小压缩方式，就是把权重的数据类型由32位的float 32转成8位的int，在数据类型层面为模型减少3~4倍的占用大小。示例图如下：

image.png

原本80多M的文件经过这个转换缩减成20M左右，成效还是挺明显的，不过缺点就是相比之前的几种压缩方式，这种压缩方式损失精度较大，非必要情况，建议别使用。

使用方法：

```
bazel build tensorflow/tools/graph_transforms:transform_graph

bazel-bin/tensorflow/tools/graph_transforms/transform_graph \
--in_graph=CTNModel.pb \
--out_graph=eight_bit_graph.pb \
--inputs='inputs/X' \
--outputs='output/predict' \
--transforms='
strip_unused_nodes(type=float, shape="256*64")
fold_constants(ignore_errors=true)
fold_batch_norms
fold_old_batch_norms
quantize_weights'
```

(/apps/redirect?utm_source=side-banner-click)

再次提醒，非极端情况请勿使用！

由上面几种grapg转换功能来讲，graph_transforms确实是一个非常强大的优化脚本，它能够一步到位的优化我们的PB文件，不像之前需要使用多种脚本之后才能得到想要的结果。上文提供了几种常用的脚本代码示例，每次使用只需要选其一即可。

memmapped_format

在前面几种压缩之后，如果文件还是较大，那该怎么办呢，这时候可以采用内存映射的方式，这一种方式是在运行时控制内存占用的一种有效方式，只是使用起来与原本的PB文件调用方式有些不同（暂时找不到Android的资料，只有IOS的，代码详见example）

示例图如下：

image.png

从整个文件读到内存中变成内存的映射，这能大量的节省内存带宽以及占用量。

使用方法：

```
bazel build tensorflow/contrib/util:convert_graphdef_memmapped_format

bazel-bin/tensorflow/contrib/util/convert_graphdef_memmapped_format \
-in_graph=/tf_files/rounded_graph.pb \
-out_graph=/tf_files/mmapped_graph.pb
```

经过这个脚本输出的PB文件不能再Python中直接调用，所以我们暂时也无法检测其可行性，需要在IOS端调用。

后记

以上是PB文件常见的压缩方法，在TensorFlow移植Mobile的过程中还有其他的优化方案，比如优化TensorFlow依赖库的大小，对ops进行高级定制，以删除不需要的ops，此方法需要针对不同的model进行不同的定制，较为高端。还有就是对模型进行优化比如尝试减少层数、减少节点数、对卷积运行做优化（点卷积降维（inception）、卷积拆分（MobileNets））这些都是较为高级的Mobile端压缩方式，本文暂时不做介绍了。

欢迎各位指错讨论。

小礼物走一走，来简书关注我

赞赏支持

(/apps/redirect?utm_source=side-banner-click)

机器学习 (/nb/9018537) 举报文章 © 著作权归作者所有



Jcme \ Ls (/u/51d83682b7b9)
写了 16753 字，被 228 人关注，获得了 485 个喜欢
(/u/51d83682b7b9)

+ 关注

木讷、后知后觉geek一枚

喜欢 | 5



更多分享

(http://cwb.assets.jianshu.io/notes/images/1388637)



下载简书 App ▶
随时随地发现和创作内容

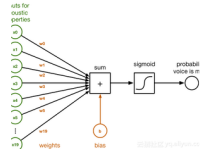


(/apps/redirect?utm_source=note-bottom-click)

被以下专题收入，发现更多相似内容

-  机器学习与数据挖掘 (/c/9ca077f0fae8?utm_source=desktop&utm_medium=notes-included-collection)
-  首页投稿 (暂停... (/c/bDHhpK?utm_source=desktop&utm_medium=notes-included-collection)
-  深度学习·计算... (/c/1249336e61cb?utm_source=desktop&utm_medium=notes-included-collection)
-  程序员 (/c/NEt52a?utm_source=desktop&utm_medium=notes-included-collection)
-  Tensorflow (/c/389b6fbe9805?utm_source=desktop&utm_medium=notes-included-collection)
-  我爱编程 (/c/7847442e0728?utm_source=desktop&utm_medium=notes-included-collection)

(/p/b370ac791613?



utm_campaign=maleskine&utm_content=note&utm_medium=seo_notes&utm_source=recommendation)
IOS平台TensorFlow实践 (/p/b370ac791613?utm_campaign=maleskine...

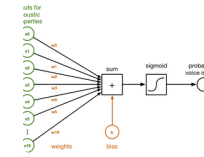
1 天前 作者简介： MATTHIJS HOLLEMANS 荷兰人，独立开发者，专注于底层编码，GPU优化和算法研究。目前研究方向为IOS上的深度学习及其在APP上的应用。推特地址：https://twitter.com/mhollemans ...



阿里云云栖社区 (/u/12532d36e4da?)

utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommendation)

(/p/abc4e084bdbc?)



(/apps/redirect?utm_source=side-banner-click)

utm_campaign=maleskine&utm_content=note&utm_medium=seo_notes&utm_source=recommendation)

深度学习指南：在iOS平台上使用TensorFlow (/p/abc4e084bdbc?utm_ca...

在利用深度学习网络进行预测性分析之前，我们首先需要对其加以训练。目前市面上存在着大量能够用于神经网络训练的工具，但TensorFlow无疑是其中极为重要的首选方案之一。这就是Tensor的全部含义。在卷...



BURIBURI_ZAEMON (/u/00d1ed2b53ae?)

utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommendation)

(/p/525a1c507d76?)



utm_campaign=maleskine&utm_content=note&utm_medium=seo_notes&utm_source=recommendation)

深度学习 - Tensorflow on iOS 入门 + MNIST (/p/525a1c507d76?utm_ca...

前言 本文主要参考了几篇文章，搭建了一个在iOS上跑Tensorflow MNIST模型的demo，本文会给出一个可用的Demo，写出当时我遇到的问题。想要把项目跑起来，需要详细的阅读我贴出来的几篇文章，某些具体步...



Thanatos_defy (/u/196c47bebfff?)

utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommendation)

掘金 Android 文章精选合集 (/p/5ad013eb5364?utm_campaign=maleski...

用两张图告诉你，为什么你的 App 会卡顿？ - Android - 掘金 Cover 有什么料？从这篇文章中你能获得这些料：知道setContentView()之后发生了什么？ ... Android 获取 View 宽高的常用正确方式，避免为零 - 掘金...



掘金官方 (/u/5fc9b6410f4f?)

utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommendation)

Android - 收藏集 (/p/dad51f6c9c4d?utm_campaign=maleskine&utm_c...

用两张图告诉你，为什么你的 App 会卡顿？ - Android - 掘金 Cover 有什么料？从这篇文章中你能获得这些料：知道setContentView()之后发生了什么？ ... Android 获取 View 宽高的常用正确方式，避免为零 - 掘金...



passiontim (/u/e946d18f163c?)

utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommendation)

觉察感受之午睡问题2017.4.20 (/p/59970ceef543?utm_campaign=malesk...

儿子没有睡午觉的习惯，周末我如果在卧室睡午觉，他一个人在客厅会害怕，不敢一个人在大厅，要跑到卧室的床上坐在我旁边，或者躺下玩，也有躺着躺着就睡着了。对比，我很反感，很愤怒。一他不睡的时...



李信兰 (/u/71459b214697?)

utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommendation)

暗恋和一切 (/p/823cd4510c17?utm_campaign=maleskine&utm_content...

从我的角度，今天也是有血有肉的那几个美好的人，我已经见过了一个世界以我为理由孕育，苦恼，爆炸，又往生了而我的故事刚好以失去主角为线用钱包，酒店，灰色的一排车辆以及浓重的黑色拳头把星...



彭先生10 (/u/e997ab8844f8?)


utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommendation)

(/p/b7b4cec87231?)

utm_campaign=maleskine&utm_content=note&utm_medium=seo_notes&utm_source=recommendation)

心理科故事：她和他（七） (/p/b7b4cec87231?utm_campaign=maleskin...

她60岁，他62岁。她去门诊看病时，他和他们的孩子一起陪着她去的，在门诊就诊时她就一会哭一会闹，像一个3岁的孩子一样肆无忌惮的倾诉她的故事。...

 吕毓萱 (/u/c5f34834ba72?)



(/apps/redirect?utm_source=side-banner-click)


utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommendation)

(/p/06b4c131b26b?



utm_campaign=maleskine&utm_content=note&utm_medium=seo_notes&utm_source=recommendation)
银饰：来自民间的冷门收藏 (/p/06b4c131b26b?utm_campaign=maleskin...

金银饰品被中国人使用的历史可以追溯到战国时期，明清至民国时期银饰更是成为中国妇女最常见的饰物之一，银饰更是成为中国妇女最常见的饰物之一，民间因此有了“无银不成饰”的说法。至今有些人的家中依然...

 芙莱尼珠宝 (/u/6a8749cafea8?)


utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommendation)

(/p/bdb027f08fd5?



utm_campaign=maleskine&utm_content=note&utm_medium=seo_notes&utm_source=recommendation)
热心解决实际困难，切实做好指导工作 (/p/bdb027f08fd5?utm_campaign...

为帮助都市丽茵小区业主解决实际困难，10月19日上午我中心邀请五里堡办事处二环西社区工作人员、都市丽茵小区业主代表召开专题协调会。会上，小区业主对都市丽茵业主大会的成立提出异议。对此，我中心...

 二七区住房保障服务中心 (/u/ee199a1adcd0?)

utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommendation)