



...

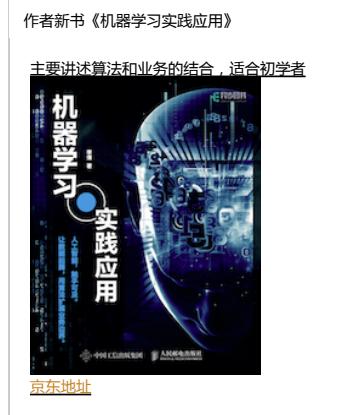
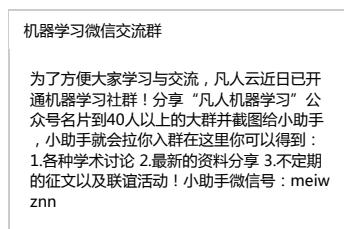
李博Garvin的专栏

阿里云机器学习PD

目录视图

摘要视图

RSS 订阅



十图详解TensorFlow数据读取机制(附代码)

2017-06-09 13:54

2969人阅读

评论(2)

收藏

举报

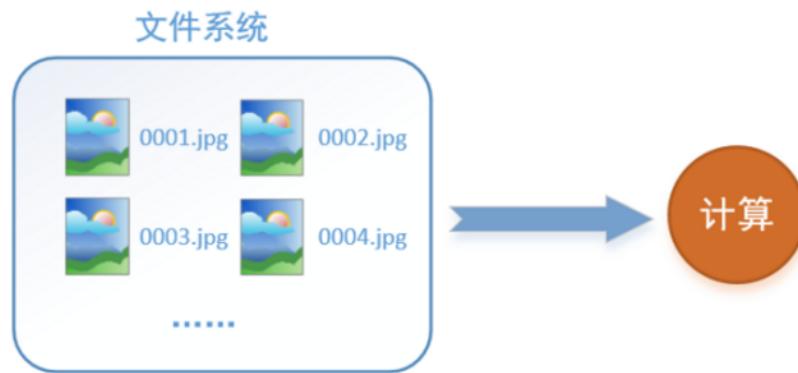
目录(?)

[+]

在学习TensorFlow的过程中，有很多小伙伴反映读取数据这一块很难理解。确实这一块官方的教程比较简略，网上也找不到什么合适的学习材料。今天这篇文章就以图片的形式，用最简单的语言，为大家详细解释一下TensorFlow的数据读取机制，文章的最后还会给出实战代码以供参考。

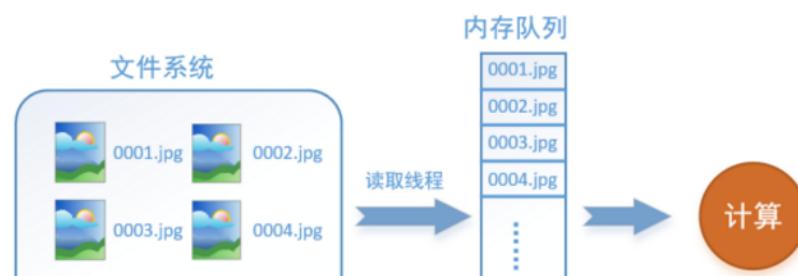
TensorFlow读取机制图解

首先需要思考的一个问题是，什么是数据读取？以图像数据为例，读取数据的过程可以用下图来表示：



假设我们的硬盘中有一个图片数据集0001.jpg, 0002.jpg, 0003.jpg.....我们只需要把它们读取到内存中，然后提供给GPU或是CPU进行计算就可以了。这听起来很容易，但事实远没有那么简单。事实上，我们必须要把数据先读入后才能进行计算，假设读入用时0.1s，计算用时0.9s，那么就意味着每过1s，GPU都会有0.1s无事可做，这就大大降低了运算的效率。

如何解决这个问题？方法就是将读入数据和计算分别放在两个线程中，将数据读入内存的一个队列，如下图所示：



排名：第1696名
原创：233篇 转载：40篇
译文：0篇 评论：460条

友情链接

czxttkl的专栏
wusuopu的专栏
buptpatriot的专栏

文章搜索



博客专栏



机器学习实践

文章：12篇
阅读：25870



LeetCode从零单排

文章：31篇
阅读：38845



git学习笔记

文章：5篇
阅读：7925



机器学习算法-python实现

文章：14篇
阅读：127475



android-tips

文章：20篇
阅读：99046



Cocos2d实例教程

文章：8篇
阅读：23253

文章分类

linux (11)
c语言 (2)
java (49)
c# (12)
百度地图api (5)
学习笔记 (63)
web互联网 (3)
android开发 (25)
DataMining (28)
Cocos2d实例教程 (8)
J2EE-ssh (2)
算法与数据结构 (47)
JDBC (3)
开源夏令营 (13)
python (16)
git (5)
面试 (5)

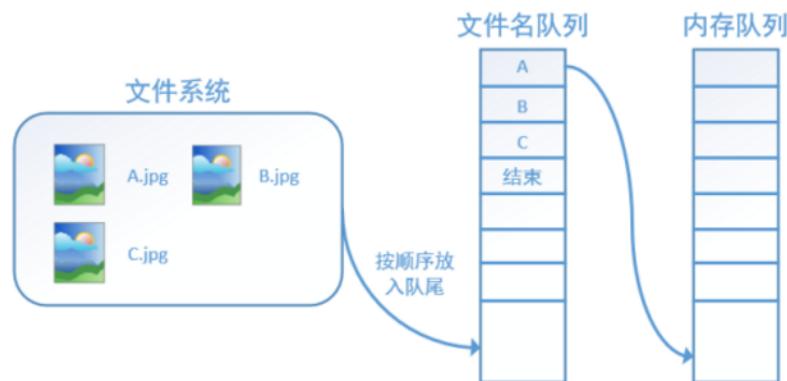


读取线程源源不断地将文件系统中的图片读入到一个内存的队列中，而负责计算的是另一个线程，计算需要数据时，直接从内存队列中取就可以了。这样就可以解决GPU因为IO而空闲的问题！

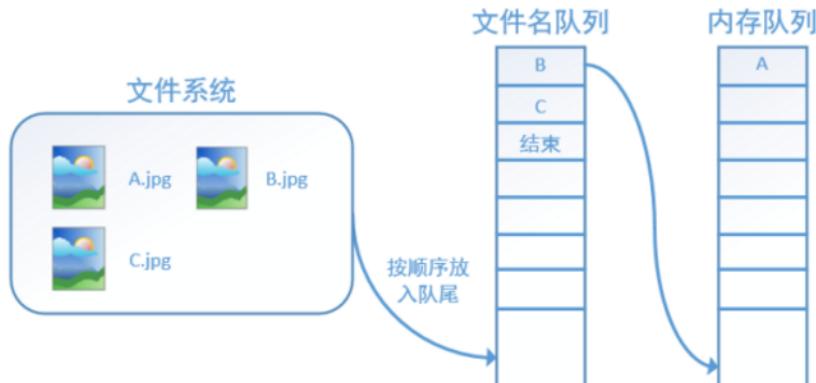
而在TensorFlow中，为了方便管理，在内存队列前又添加了一层所谓的“文件名队列”。

为什么要添加这一层文件名队列？我们首先得了解机器学习中的一个概念：epoch。对于一个数据集来讲，运行一个epoch就是将这个数据集中的图片全部计算一遍。如一个数据集中有三张图片A.jpg、B.jpg、C.jpg，那么跑一个epoch就是指对A、B、C三张图片都计算了一遍。两个epoch就是指先对A、B、C各计算一遍，然后再全部计算一遍，也就是说每张图片都计算了两遍。

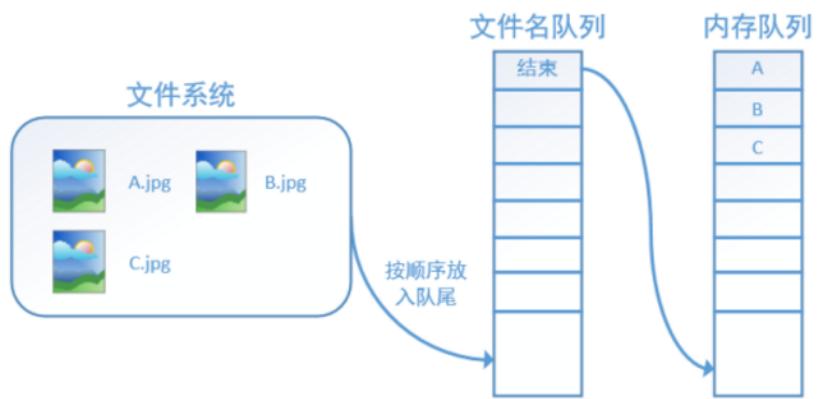
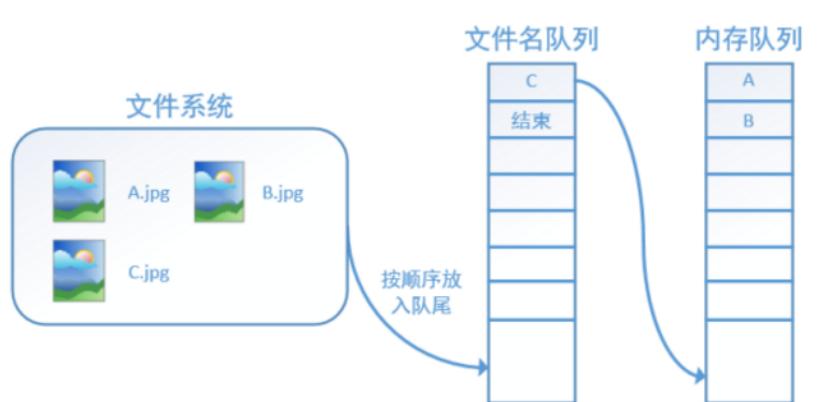
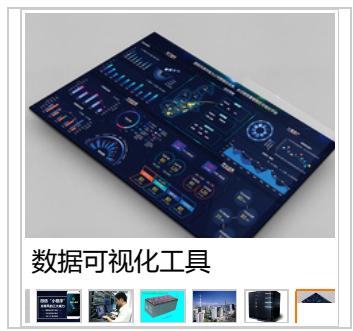
TensorFlow使用文件名队列+内存队列双队列的形式读入文件，可以很好地管理epoch。下面我们用图片的形式来说明这个机制的运行方式。如下图，还是以数据集A.jpg, B.jpg, C.jpg为例，假定我们要跑一个epoch，那么我们在文件名队列中把A、B、C各放入一次，并在之后标注队列结束。



程序运行后，内存队列首先读入A（此时A从文件名队列中出队）：



再依次读入B和C：



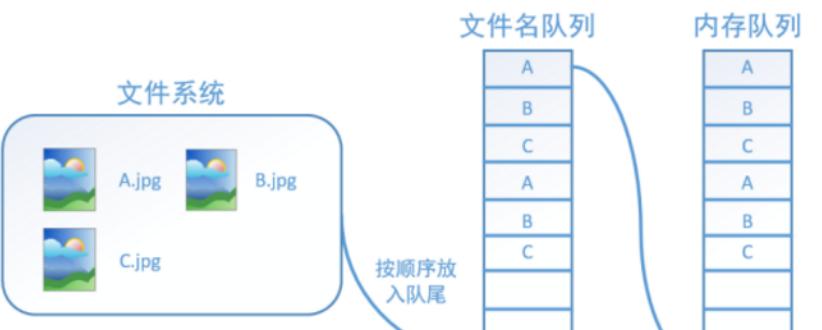
此时，如果再尝试读入，系统由于检测到了“结束”，就会自动抛出一个异常（`OutOfRange`）。外部捕捉到这个异常后就可以结束程序了。这就是TensorFlow中读取数据的基本机制。如果我们要跑2个epoch而不是1个epoch，那只要在文件名队列中将A、B、C依次放入两次再标记结束就可以了。

TensorFlow读取数据机制的对应函数

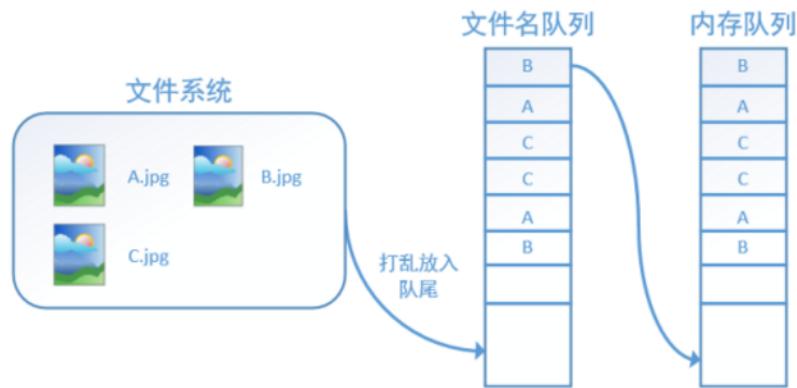
如何在TensorFlow中创建上述的两个队列呢？

对于文件名队列，我们使用`tf.train.string_input_producer`函数。这个函数需要传入一个文件名list，系统会自动将它转为一个文件名队列。

此外`tf.train.string_input_producer`还有两个重要的参数，一个是`num_epochs`，它就是我们上文中提到的epoch数。另外一个就是`shuffle`，`shuffle`是指在一个epoch内文件的顺序是否被打乱。若设置`shuffle=False`，如下图，每个epoch内，数据还是按照A、B、C的顺序进入文件名队列，这个顺序不会改变：



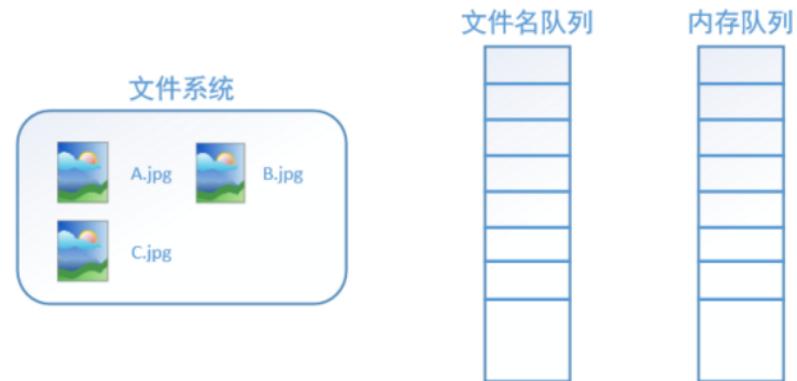
如果设置`shuffle=True`，那么在一个epoch内，数据的前后顺序就会被打乱，如下图所示：



在TensorFlow中，内存队列不需要我们自己建立，我们只需要使用reader对象从文件名队列中读取数据就可以了，具体实现可以参考下面的实战代码。

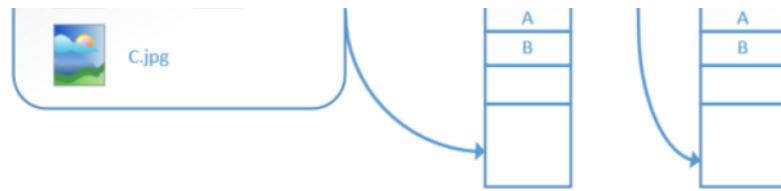
除了`tf.train.string_input_producer`外，我们还要额外介绍一个函数：`tf.train.start_queue_runners`。初学者会经常在代码中看到这个函数，但往往很难理解它的用处，在这里，有了上面的铺垫后，我们就可以解释这个函数的作用了。

在我们使用`tf.train.string_input_producer`创建文件名队列后，整个系统其实还是处于“停滞状态”的，也就是说，我们文件名并没有真正被加入到队列中（如下图所示）。此时如果我们开始计算，因为内存队列中什么也没有，计算单元就会一直等待，导致整个系统被阻塞。



而使用`tf.train.start_queue_runners`之后，才会启动填充队列的线程，这时系统就不再“停滞”。此后计算单元就可以拿到数据并进行计算，整个程序也就跑起来了，这就是函数`tf.train.start_queue_runners`的用处。





实战代码

我们用一个具体的例子感受TensorFlow中的数据读取。如图，假设我们在当前文件夹中已经有A.jpg、B.jpg、C.jpg三张图片，我们希望读取这三张图片5个epoch并且把读取的结果重新存到read文件夹中。



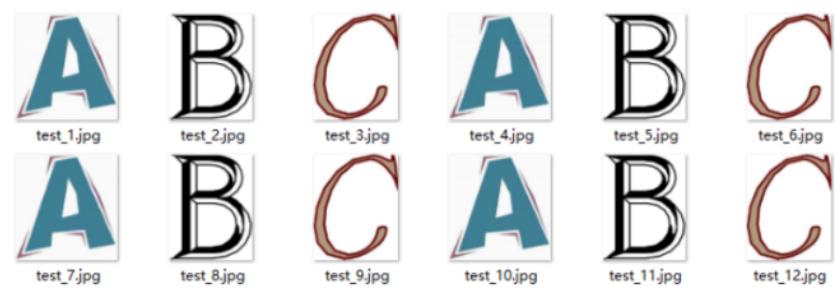
对应的代码如下：

```
# 导入TensorFlow
import tensorflow as tf

# 新建一个Session
with tf.Session() as sess:
    # 我们要读三幅图片A.jpg, B.jpg, C.jpg
    filename = ['A.jpg', 'B.jpg', 'C.jpg']
    # string_input_producer会产生一个文件名队列
    filename_queue = tf.train.string_input_producer(filename, shuffle=False, num_epochs=5)
    # reader从文件名队列中读数据。对应的方法是reader.read
    reader = tf.WholeFileReader()
    key, value = reader.read(filename_queue)
    # tf.train.string_input_producer定义了一个epoch变量，要对它进行初始化
    tf.local_variables_initializer().run()
    # 使用start_queue_runners之后，才会开始填充队列
    threads = tf.train.start_queue_runners(sess=sess)
    i = 0
    while True:
        i += 1
        # 获取图片数据并保存
        image_data = sess.run(value)
        with open('read/test_%d.jpg' % i, 'wb') as f:
            f.write(image_data)
```

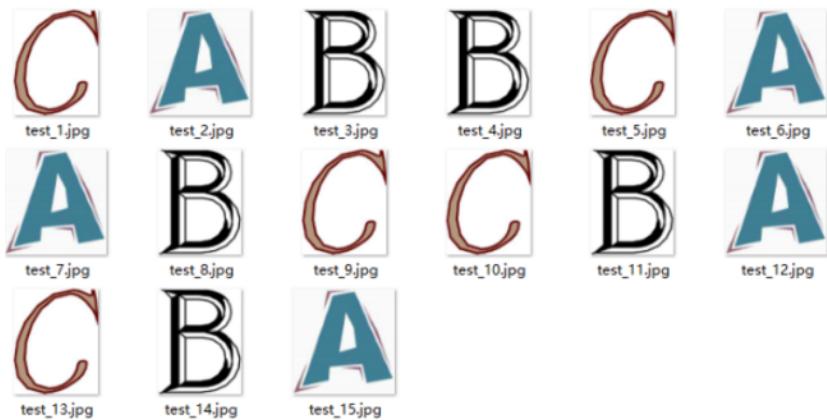
我们这里使用filename_queue = tf.train.string_input_producer(filename, shuffle=False, num_epochs=5)建立了一个会跑5个epoch的文件名队列。并使用reader读取，reader每次读取一张图片并保存。

运行代码后，我们得到就可以看到read文件夹中的图片，正好是按顺序的5个epoch：





如果我们设置`filename_queue = tf.train.string_input_producer(filename, shuffle=False, num_epochs=5)`中的`shuffle=True`，那么在每个epoch内图像就会被打乱，如图所示：



我们这里只是用三张图片举例，实际应用中一个数据集肯定不止3张图片，不过涉及到的原理都是共通的。

总结

这篇文章主要用图解的方式详细介绍了TensorFlow读取数据的机制，最后还给出了对应的实战代码，希望能够给大家学习TensorFlow带来一些实质性的帮助。

- 上一篇 五大机器学习微信公众号推荐
- 下一篇 【机器学习PAI实践八】用机器学习算法评估学生考试成绩

顶

6

踩

1



查看评论



韦德隆东强

2楼 2017-12-29 11:17发表

我试了一下，在循环`num_epochs`完了之后会报错，`FIFOQueue '_15_input_producer_2'` is closed and has insufficient elements, 循环队列`filename_queue`没有足够的空间。



小白和小白狼

写的很好，有帮助。

1楼 2017-12-07 12:30发表

TensorFlow高效读取数据的方法  u012759136 2016-08-17 19:20  40324

概述关于Tensorflow读取数据，官网给出了三种方法：供给数据(Feeding)：在TensorFlow程序运行的每一步，让Python代码来供给数据。从文件读取数据：在TensorFlow图的起始，让一个输入管线从文件中读取数据。预加载数据：在TensorFlow图中定义常量或...

Tensorflow数据读取方法  aitazhixin 2017-06-23 11:32  576

Tensorflow读取文件的三种方式：预读取，喂数据，读文件

你当初为什么想不开去当程序员？

你热爱编程吗？对一切新鲜的语言都热血沸腾吗？如果不，试试这个..

广告

Tensorflow从文件读取数据  zengxyuyu 2016-11-22 19:47  21552

作者：曾翔钰 && 石炜贤 @曾翔钰 @石炜贤 TensorFlow程序读取数据一共有3种方法：供给数据(Feeding)：在TensorFlow程序运行的每一步，让Python代码来供给数据。从文件读取数据：在TensorFlow图的起始，让一个输入管线从文件中读取...

13、Tensorflow : Tensorflow数据读取有三种方式 (next_batch)

一、Tensorflow数据读取有三种方式：Preloaded data: 预加载数据 Feeding: Python产生数据，再把数据喂给后端。Reading from file: 从文件中直接读取这三种有读取方式有什么区别呢？我们首先要要知道TensorFlow(TF)是怎...

 qq_38906523 2018-01-02 22:24  152

码农不会英语怎么行？英语文档都看不懂！



软件工程出身的英语老师，教你用数学公式读懂天下英文→

tensorflow--数据读取篇  u010540396 2017-10-26 20:31  214

最近，心血来潮搞一搞tensorflow,看着《tensorflow实战》码了几个简单的小网络，自以为蛮简单啊，当自己开始从头开始构建自己网络时候，就开始怀疑人生了。自己的数据读取都是一个大问题，今天解决了使用tensorflow读取csv文本数据，写到博客做个笔记。

tensorflow载入数据的三种方式  lujiandong1 2016-11-28 14:50  11622

Tensorflow数据读取有三种方式：Preloaded data: 预加载数据 Feeding: Python产生数据，再把数据喂给后端。Reading from file: 从文件中直接读取这三种有读取方式有什么区别呢？我们首先要要知道TensorFlow(TF)是怎么样工作的。TF...

tensorflow的数据输入  zzk1995 2017-01-09 20:57  7228

tensorflow有两种数据输入方法，比较简单的一种是使用feed_dict，这种方法在画graph的时候使用placeholder来站位，在真正run的时候通过feed字典把真实的输入传进去。比较简单不再介绍。比较恼火的是第二种方法，直接从文件中读取数据（其实第一种也可以我们自己从文件中读出来之...

TensorFlow——训练自己的数据（一）数据处理

参考：Tensorflow教程-猫狗大战数据集贴一
张自己画的思维导图数据集准备 kaggle猫狗大战数据集
(训练)，微软的不需要翻墙 12500张cat 12500张dog 生成图片路径和标签的List
step1：获取D:/Study/Python/Projects/Cats...

深度学习小白——Tensorflow(三) 读取数据

一个典型的文件读取管线会包含下面这些步骤 :  MargretWG 2017-04-16 15:43  3303

文件名列表可配置的 文件名乱序(shuffling)可配

置的最大训练迭代数(epoch limit)文件名队列针对输入文件格式的阅读器纪录解析器可配置的预处理器样本队列

tensorflow读取数据到队列当中

 s_sunnyy 2017-04-18 16:10  1690

原文地址 : <http://blog.csdn.net/lujiaodong1/article/details/53376134> TensorFlow是一种符号编程框架 (与theano类似) , 先构建数据流图再输入数据进行模型训练。Tensorflow支持很多种样例输入的方式。最容易的是使用place...

TensorFlow读取数据的方法

 tengxing007 2017-01-14 20:25  839

关于Tensorflow读取数据 , 官网给出了三种方法 : 供给数据(Feeding) : 在TensorFlow程序运行的每一步 , 让Python代码来供给数据。从文件读取数据 : 在TensorFlow图的起始 , 让一个输入管线从文件中读取数据。预加载数据 : 在TensorFlow图中定义常量或...

tensorflow读取数据之CSV格式

 sunquan_ok 2016-07-06 13:56  10705

tensorflow要想用起来 , 首先自己得搞定数据输入。官方文档中介绍了几种 , 1.一次性从内存中读取数据到矩阵中 , 直接输入 ; 2.从文件中边读边输入 , 而且已经给设计好了多线程读写模型 ; 3.把网络或者内存中的数据转化为tensorflow的专用格式tfRecord,存文件后再读取。其中 , 从文件中边...

Tensorflow分批量读取数据

 freedom098 2017-02-20 15:28  1988

Tensorflow分批量读取数据之前的博客里使用tf读取数据都是每次fetch一条记录 , 实际上大部分时候需要fetch到一个batch的小批量数据 , 在tf中这一操作的明显变化就是tensor的rank发生了变化 , 我目前使用的人脸数据集是灰度图像 , 因此大小是92*112的 , 所以最开始fetch拿到的...

TensorFlow 学习 (二) 制作自己的TFRecord数据集 , 读取 , 显示及代...

前言在跑通了官网的mnist和cifar10数据之后 , 笔者尝试着制作自己的数据集 , 并保存 , 读入 , 显示。TensorFlow可以支持cifar10的数据格式 , 也提供了标准的TFRecord 格式 , 而关于 tensorflow 读取数据 , 官网提供了3中方法 1 Feeding : 在tensor...

 miaomiaoyuan 2017-02-24 20:33  7034

Tensorflow图片数据读取

 woshilimengxi 2016-09-24 09:47  4976

Tensorflow数据经典处理方法

TensorFlow学习记录-- 7 .TensorFlow高效读取数据之tfrecord详细解读

— why tfrecord?对于数据量较小而言 , 可能一般选择直接将数据加载进内存 , 然后再分batch输入网络进行训练 (tip: 使用这种方法时 , 结合yield 使用更为简洁 , 大家自己尝试一下吧 , 我就不赘述了) 。但是 , 如果数据量较大 , 这样的方法就不适用了 , 因为太耗内存 , 所以这时最好使用tensorfl...

 qq_16949707 2016-12-06 10:06  9157

tensorflow学习笔记 (五) : TensorFlow变量共享和数据读取

1、变量共享 前面已经说过如何进行变量  woidapaopao 2017-06-13 16:28 5520
的生成和初始化内容，也用到了命名空间的概念，这里说一下什么是变量共享。当我们有一个非常庞大的模型的时候免不了需要进行大量的变量共享，而且有时候还希望能够在同一个地方初始化所有的变量，这就需要tf.variable_scope() 和 tf.get_varia..

TensorFlow 读取图片1：初探四种从文件读取的方式

本文记录一下TensorFlow的几种图片读取方法， Wayne2019 2017-09-07 17:27 3958
官方文档有较为全面的介绍。1.使用gfile读图片，decode输出是Tensor，eval后是ndarrayimport matplotlib.pyplot as plt import tensorflow as tf import numpy ...

TensorFlow数据读取方法

 u010329292 2017-03-30 11:02 2781
转自：<http://honggang.io/2016/08/19/tensorflow-data-reading/> 引言 Tensorflow的数据读取有三种方式：Preloaded data: 预加载数据Feeding: Python产生数据，再把数据喂给后端。...

【TensorFlow动手玩】数据导入2

TensorFlow的第二种数据导入机制：二进制  shenxiaolu1984 2016-11-05 11:20 6098
文件。

公司简介 | 招贤纳士 | 广告服务 | 联系方式 | 版权声明 | 法律顾问 | 问题报告 | 合作伙伴 | 论坛反馈

网站客服 | 杂志客服 | 微博客服 | webmaster@csdn.net | 400-660-0108 | 北京创新乐知信息技术有限公司 版权所有 | 江苏知之为计算机有限公司 |

江苏乐知网络技术有限公司

京 ICP 证 09002463 号 | Copyright © 1999-2017, CSDN.NET, All Rights Reserved 