

原

【论文笔记】SparkNET: 用Spark训练深度神经网络

2015年12月27日 23:00:47

阅 9183

7

SparkNet: Training Deep Network in Spark

写评论

目录

收藏

微信

原文是：《SparkNet: Training Deep Network in Spark》  
本博客是该论文的阅读笔记，不免有很多细节不对之处。  
还望各位看官能够见谅，欢迎批评指正。  
更多相关博客请猛戳：[http://blog.csdn.net/cyh\\_24](http://blog.csdn.net/cyh_24)  
如需转载，请附上本文链接：[http://blog.csdn.net/cyh\\_24/article/details/50414568](http://blog.csdn.net/cyh_24/article/details/50414568)

这篇论文是 **Berkeley** 大学 **Michael I. Jordan** 组的 ICLR2016(under review) 的最新论文，有兴趣可以看看原文和源码：[paper](#) [ib](#) .  
训练深度神经网络是一个非常耗时的过程，比如用卷积神经网络去训练一个目标识别任务需要好几天来训练。因此，充分利用算资源，加快训练速度成了一个非常重要的领域。不过，当前非常热门的批处理计算架构（例如：MapReduce 和 Spark）都不是设计异步计算和现有的一些通信密集型的深度学习系统。

SparkNet 是基于Spark的深度神经网络架构，

1. 它提供了便捷的接口能够去访问Spark RDDs；
2. 同时提供Scala接口去调用caffe；
3. 还拥有轻量级的tensor 库；
4. 使用了一个简单的并行机制来实现SGD的并行化，使得SparkNet能够很好的适应集群的大小并且能够容忍极高的通信延时；
5. 它易于部署，并且不需要对参数进行调整；
6. 它还能很好的兼容现有的caffe模型；

下面这张图是SparkNet的架构：

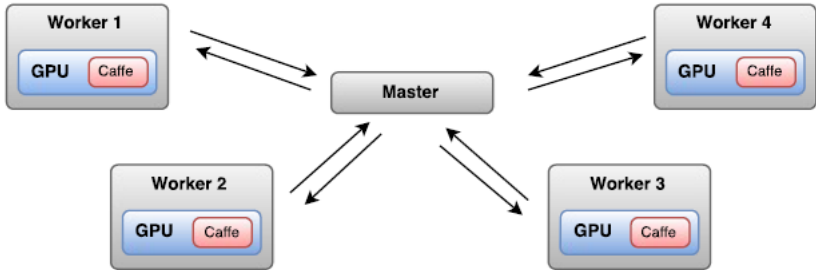


Figure 1: This figure depicts the SparkNet architecture.

从上图可以看出，Master 向每个worker 分发任务之后，各个worker都单独的使用Caffe（利用GPU）来进行训练。每个worker完成任务之后，把参数传回Master。论文用了5个节点的EC2集群，broadcast 和 collect 参数（每个worker几百M），耗时20秒，而一个minibatch的计算时间是2秒。

Implementation

SparkNet 是建立在Apache Spark和Caffe深度学习库的基础之上的。SparkNet 用Java来访问Caffe的数据，用Scala来访问Caffe的参数，用ScalaBuf来使得Caffe网络在运行时保持动态结构。SparkNet能够兼容Caffe的一些模型定义文件，并且支持Caffe模型参数的载入。

下面简单贴一下SparkNet的api和模型定义、模型训练代码。

```
class Net {
  def Net(netParams: NetParams): Net
  def setTrainingData(data: Iterator[(NDArray, Int)]): Net
  def setValidationData(data: Iterator[(NDArray, Int)]): Net
  def train(numSteps: Int): Net
  def test(numSteps: Int): Float
  def setWeights(weights: WeightCollection): Net
  def getWeights(): WeightCollection
}
```

Listing 1: SparkNet API

```
val netParams = NetParams(
  RDDLayer("data", shape=List(batchsize, 1, 28, 28)),
  RDDLayer("label", shape=List(batchsize, 1)),
  ConvLayer("conv1", List("data"), kernel=(5,5), numFilters=20),
  PoolLayer("pool1", List("conv1"), pool=Max, kernel=(2,2), stride=(2,2)),
  ConvLayer("conv2", List("pool1"), kernel=(5,5), numFilters=50),
  PoolLayer("pool2", List("conv2"), pool=Max, kernel=(2,2), stride=(2,2)),
  LinearLayer("ip1", List("pool2"), numOutputs=500),
  ActivationLayer("relu1", List("ip1"), activation=ReLU),
  LinearLayer("ip2", List("relu1"), numOutputs=10),
  SoftmaxWithLoss("loss", List("ip2", "label"))
)
```

Listing 2: Example network specification in SparkNet

```
var trainData = loadData(...)
var trainData = preprocess(trainData).cache()
var nets = trainData.foreachPartition(data => {
  var net = Net(netParams)
  net.setTrainingData(data)
  net
})
var weights = initialWeights(...)
for (i <- 1 to 1000) {
  var broadcastWeights = broadcast(weights)
  nets.map(net => net.setWeights(broadcastWeights.value))
  weights = nets.map(net => {
    net.train(50)
    net.getWeights()}).mean() // an average of WeightCollection objects
}
```

Listing 3: Distributed training example

## 并行化的SGD

为了让模型能够在带宽受限的环境下也能运行得很好，论文提出了一种SGD的并行化机制使得最大程度减小通信，这也是全文最大的亮点。这个方法也不是只针对SGD，实际上对Caffe的各种优化求解方法都有效。

在将SparkNet的并行化机制之前，先介绍一种Naive的并行机制。

### Naive SGD Parallelization

Spark拥有一个master节点和一些worker节点。数据分散在各个worker中的。

在每一次的迭代中，Spark master节点都会通过broadcast（广播）的方式，把模型参数传到各个worker节点中。

各个worker节点在自己分到的部分数据，在同一个模型上跑一个minibatch的SGD。

完成之后，各个worker把训练的模型参数再发送回master，master将这些参数进行一个平均操作，作为新的（下一次迭代）的模型参数。

这是很多人都会采用的方法，看上去很对，不过它有一些缺陷。

## Naive 并行化的缺陷

这个缺陷就是需要消耗太多的通信带宽，因为每一次minibatch训练都要broadcast 和 collect 一次，而这个过程特别消耗时间（20秒左右）。

令  $N_a(b)$  表示，在batch-size为  $b$  的情况下，到达准确率  $a$  所需要的迭代次数。  
令  $C(b)$  表示，在batch-size 为  $b$  的情况下，SGD训练一个batch的训练时间（约2秒）。  
显然，使用SGD达到准确率为 $a$ 所需要的时间消耗是：

$$N_a(b)C(b)$$

假设有 $K$ 个机器，通信（broadcast 和 collect）的时间为  $S$ ，那么Naive 并行 SGD 的时间消耗就是：

$$N_a(b)(C(b)/K + S)$$

## SparkNet 的并行化机制

基本上过程和Naive 并行化差不多。唯一的区别在于，各个worker节点每次不再只跑一个迭代，而是在自己分到的一个minibatch数据集上，迭代多次，迭代次数是一个固定值 $\tau$ 。

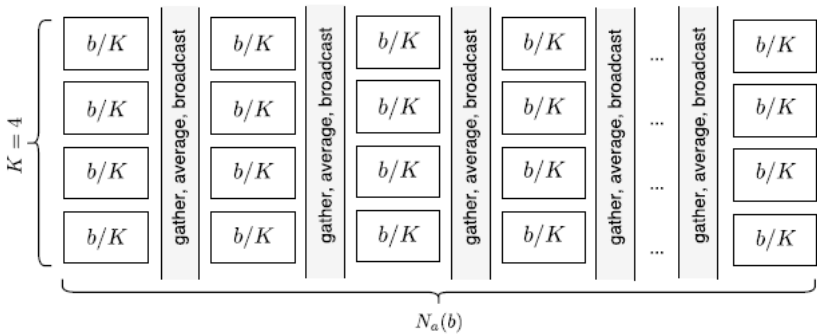
SparkNet的并行机制是分好几个rounds来跑的。在每一个round中，每个机器都在batch size为 $b$ 的数据集上跑  $\tau$  次迭代。没一个round结束，再把参数汇总到master进行平均等处理。

我们用 $M_a(b, K, \tau)$  表示达到准确率  $a$  所需要的 round 次数。  
因此，SparkNet需要的时间消耗就是：

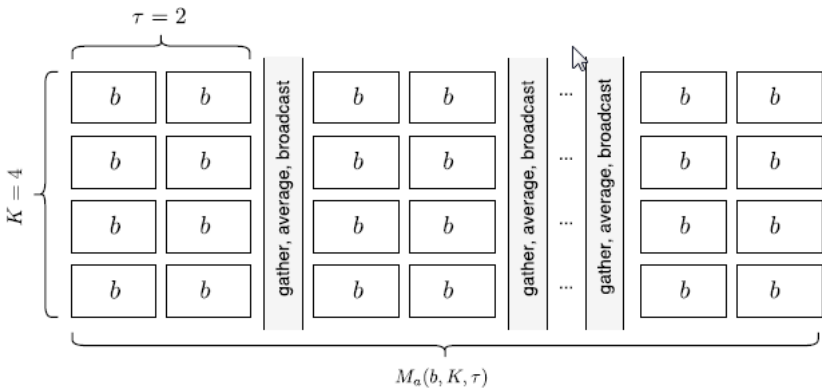
$$M_a(b, K, \tau) * (\tau C(b) + S)$$

下面这张图，很直观的对比了Naive 并行机制跟 SparkNet 并行机制的区别：

Naive 并行机制：



SparkNet 并行机制：



论文还做了各种对比实验，包括时间，准确率等。实验模型采用AlexNet，数据集是ImageNet的子集（100类，每类1000张）。

假设 $S = 0$ ，那么 $\tau M_a(b, K, \tau) / N_a(b)$  就是SparkNet的加速倍数。论文通过改变 $\tau$  和  $K$  得出了下面的表格（使准确率达到20%的耗时情况）：

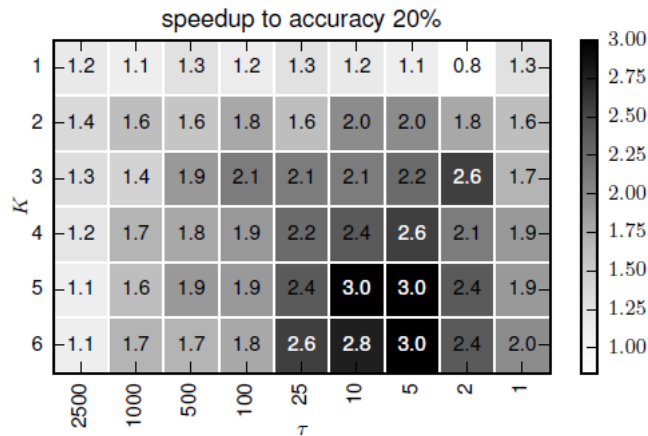


Figure 3: This figure shows the speedup  $\tau M_a(b, \tau, K) / N_a(b)$  given by SparkNet’s parallelization

- 上面的表格还是体现了一些趋势的:
- (1). 看第一行，当 $K = 1$ ，因为只有一个worker节点，所以异步计算的 $\tau$ 这时并没有起到什么作用，可以看到第一行基本的值基本都是接近1.
  - (2). 看最右边这列，当 $\tau = 1$ ，这其实就相当于Naive并行机制，只不过，Naive的batch是 $b/K$ ，这里是 $b$ . 这一列基本上是跟 $K$ 成正比.
  - (3). 注意到每一行的值并不是从左到右一直递增的.

当 $S! = 0$ 的时候，naive跟SparkNet的耗时情况又是怎样的呢？作者又做了一些实验。

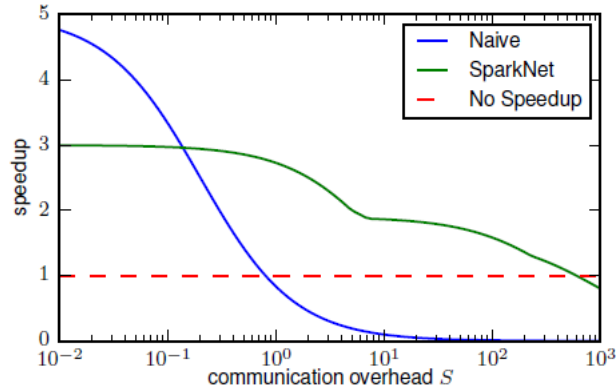


Figure 4: This figure shows the speedups obtained by the naive parallelization scheme and by SparkNet as a function of the cluster’s communication overhead (normalized so that  $C(b) = 1$ ).

可以看到，当 $S$ 接近与0的时候（带宽高），Naive会比SparkNet速度更快，但是，当 $S$ 变大（带宽受限），SparkNet的性能将超过Naive，并且可以看出，Naive受 $S$ 变化剧烈，而SparkNet相对平稳。

而作者实验用EC2环境， $S$ 大概是20秒，所以，显然，SparkNet会比Naive好很多。

论文还做了一些事情，比如：

- 令 $\tau = 50$ ，分别测试 $K = 1、3、5、10$ 时，准确率与时间的关系；
- 当 $K = 5$ ，分别测试 $\tau = 20、50、100、150$ 时，准确率与时间的关系。

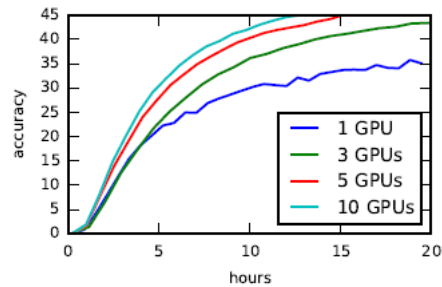


Figure 5: This shows the scaling of SparkNet with 3, 5, and 10 GPUs and  $\tau = 50$ . The 1 GPU plot was obtained by running Caffe with no communication, whereas the other experiments communicate parameters between machines incurring an overhead of about 20 seconds per synchronization.

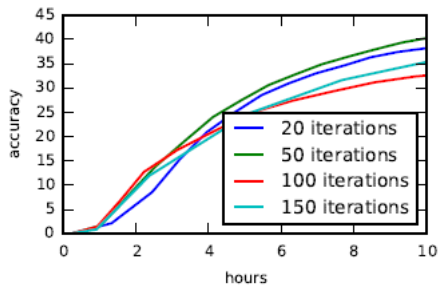


Figure 6: This figure shows the dependence of the parallelization scheme described in Section 2.1 on  $\tau$ . Each experiment was run with  $K = 5$  workers. This figure shows that there is no need to collect and broadcast the model more frequently than every 50 iterations in our bandwidth-limited cluster.

总结一下，这篇论文其实没有太多复杂的创新（除了SGD并行化时候的一点小改进），不过我很期待后续的工作，同时也希望这个SparkNet能够维护的越来越好。有时间的话，还是很想试试这个SparkNet的。

版权声明：如需转载，请附上本文链接。作者主页：[http://blog.csdn.net/cyh\\_24](http://blog.csdn.net/cyh_24) <https://blog.csdn.net/cyh24/article/details/50414568>

所属专栏：[游戏编程模式](#) [PySpark-机器学习](#)

CSDN学院

人工智能工程师

火热报名中

四个月搞定机器学习，快速进入人工智能领域

点击了解

想对作者说点什么？

我来说两句

大愚若智\_ 2018-01-23 09:23:31 #1楼

赞

上一页

1

下一页

### Spark2.0机器学习系列之6：MLPC（多层神经网络）

Spark2.0 MLPC（多层神经网络分类器）算法概述 MultilayerPerceptronClassifier（MLPC）这是一个基于前馈神经网络的分类器，它是一种在输入层与输出层之间含有...

qq\_34531825 2016-08-31 09:11:16 阅读数：4225

### Spark MLlib Deep Learning Neural Net(深度学习-神经网络)1.1

Spark MLlib Deep Learning Neural Net(深度学习-神经网络)1.1 <http://blog.csdn.net/sunbow0/> Spark MLlib Deep L...

sunbow0 2015-05-28 18:01:09 阅读数：7581

### SparkNET: 用Spark训练深度神经网络 - CSDN博客

训练深度神经网络是一个非常耗时的过程,比如用卷积神经网络去训练一个目标识别任务...作者又做了一些实验。可以看到,当接近于0的时候(带宽高),Naive会比SparkNet...

2018-6-6

## 基于Apache Spark的机器学习及神经网络算法和应用 - CSDN博客

《基于Apache Spark的机器学习及神经网络算法和应用》的课程,介绍了大规模分布式机器学习在欺诈检测、用户行为预测(稀疏逻辑回归)中的实际应用,以及英特尔在LDA、Word2...

2018-5-25

### 一个简单的生长头发方法，脱发掉发的人欢呼了！

美兰 · 顶新

## 基于Apache Spark的机器学习及神经网络算法和应用

使用高级分析算法（如大规模机器学习、图形分析和统计建模等）来发现和探索数据是当前流行的思路，在IDF16技术课堂上，英特尔公司软件开发工程师王以恒分享了《基于Apache Spark的机器学习及神经...

 happytofly 2016-04-14 00:53:44 阅读数：151

## 用spark训练深度神经网络 - CSDN博客

训练深度神经网络是一个非常耗时的过程,比如用卷积神经网络去训练一个目标识别任务...作者又做了一些实验。 可以看到,当接近与0的时候(带宽高),Naive会比SparkNet...

2018-6-19

## 用spark训练深度神经网络 - CSDN博客

训练深度神经网络是一个非常耗时的过程,比如用卷积神经网络去训练一个目标识别任务...作者又做了一些实验。 可以看到,当接近与0的时候(带宽高),Naive会比SparkNet...

2018-6-19


## spark-BigDL:深度学习之神经网络编写

BigDL主要实现了各种深度学习神经网络算法，当然也可以构建简单的神经网络。 一、下载依赖包和初始化系统 1.第一步是下载适合本地spark版本的 [https://github.com/intel...](https://github.com/intel)

 qq\_30232405 2017-06-21 14:32:40 阅读数：643

## 深度学习-基于spark的多层神经网络

最后我们再写3篇基于spark的深度学习，这篇是手写识别的，用的是spark的local模式，如果想用集群模式在submit的时候设置-useSparkLocal false，或者在程序中设置useS...

 chencheng12077 2017-01-20 17:14:39 阅读数：1788

## 深度学习-基于spark的多层神经网络 - CSDN博客

深度学习-基于spark的多层神经网络2017年01月20日 17:14:39 阅读数:1627 最后我们再写3篇基于spark的深度学习,这篇是手写识别的,用的是spark的local模式,如果想...

2018-6-7


## pyspark 多层神经网络 - CSDN博客

pyspark 多层神经网络2018年02月24日 17:45:43 阅读数:220 from pyspark import SparkContext from pyspark.sql import SQLContext from pyspark.sql import Spark...

2018-6-3

## 分布式TensorFlow：在Spark上将谷歌的深度学习库进行尺度变换

本文为数盟原创译文，转载请注明出处为“数盟社区”。 介绍 Arimo的日益增长的数据科学团队包括研究和开发机器学习和深度学习新的方法和应用。 我们正在调查的一个主题是分布式的深度学习...

 u013886628 2016-07-04 11:26:02 阅读数：8637

### 老中医说：男人用这个方法，时间提高20分钟

番当 · 顶新



## 【论文笔记】SparkNET: 用Spark训练深度神经网络 - CSDN博客

【论文笔记】SparkNET: 用Spark训练深度神经网络 标签: 深度学习 机器学习 spark神经网络 2015-12-27 23:00 1368人阅读 评论(0) 收藏 举报 分类:...

2018-6-6

## 基于Apache Spark的机器学习及神经网络算法和应用 - CSDN博客

《基于Apache Spark的机器学习及神经网络算法和应用》的课程,介绍了大规模分布式机器学习在欺诈检测、用户行为预测(稀疏逻辑回归)中的实际应用,以及英特尔在LDA、Word2...

2018-4-14


## Spark MLlib Deep Learning Neural Net(深度学习-神经网络)1.2

[原]Spark MLlib Deep Learning Neural Net(深度学习-神经网络)1.2 2015-5-28阅读62 评论0 Spark MLlib Deep Lea...

 javastart 2015-05-30 10:36:02 阅读数 : 4358

## spark深度学习算法(CNN卷积神经网络)的测试与分析

卷积神经网络 (Convolutional Neural Network, CNN) 是一种前馈神经网络, 它的人工神经元可以响应一部分覆盖范围内的周围单元, 对于大型图像处理有出色表现。 关于CNN...

 sparkexpert 2015-11-03 10:39:34 阅读数 : 5234

## 深度学习-基于spark的多层神经网络 - CSDN博客

深度学习-基于spark的多层神经网络2017年01月20日 17:14:39 阅读数:1718 最后我们再写3篇基于spark的深度学习,这篇是手写识别的,用的是spark的local模式,如果想...

2018-7-3

## Spark MLlib Deep Learning Convolution Neural Network (深度学习-卷积神经网络)3.3

3、Spark MLlib Deep Learning Convolution Neural Network(深度学习-卷积神经网络)3.3 <http://blog.csdn.net/sunbow0> ...

 sunbow0 2015-07-22 20:33:42 阅读数 : 4766

## 跟着吴恩达学深度学习：用Scala实现神经网络-第二课：用Scala实现多层神经网络

上一章我们讲了如何使用Scala实现LogisticRegression，这一章跟随着吴恩达的脚步我们用Scala实现基础的深度神经网络。顺便再提一下，吴恩达对于深度神经网络的解释是我如今听过的最清楚...

 pan5431333 2017-08-29 14:00:23 阅读数 : 1156

## Spark与深度学习框架——H2O、deeplearning4j、SparkNet

阅读原文请点击 摘要： 引言：你可能对使用Spark服务比较感兴趣。Spark已经提供了很多功能，也有一个好用的界面，而且背后有强大的社区，开发者十分活跃，这也是人们对Spark寄予厚望的原...

 qq\_35267530 2017-06-22 14:24:33 阅读数 : 2013

## Spark(一): 基本架构及原理

Apache Spark是一个围绕速度、易用性和复杂分析构建的大数据处理框架，最初在2009年由加州大学伯克利分校的AMPLab开发，并于2010年成为Apache的开源项目之一，与Hadoop和St...

 swing2008 2017-03-08 11:26:45 阅读数 : 65534

## 简单神经网络实现 02

误差选择均方误差 梯度下降步骤：数据集为研究生院录取数据，来源。数据格式：admit这一栏为标签，其余的栏目是特征。网络没有设置隐层。代码：import panda...

 y12345678904 2017-10-12 13:04:03 阅读数 : 167



## 50万码农评论：英语对于程序员有多重要！

不背单词和语法，老司机教你一个数学公式秒懂天下英语

## 神经网络加速器的兴起

自从投身智能硬件以来，又开始重新关注嵌入式领域的相关技术。这是“2018嵌入式处理器报告: 神经网络加速器的兴起”(http://www.embedded-computing.com/processi...

wireless\_com 2018-02-07 00:00:00 阅读数：654

小白都理解的人工智能系列（13）——如何加速神经网络训练过程

问题1：如何加速训练？SGD将大的数据拆分成一块块小的数据进行训练。Momentum通过不断调整w值的形式，朝着梯度往下走，以期尽快达到终点。AdaGrad...

taczeng 2018-01-27 20:37:53 阅读数：163

Spark在Windows下的环境搭建

由于Spark是用Scala来写的，所以Spark对Scala肯定是原生态支持的，因此这里以Scala为主来介绍Spark环境的搭建，主要包括四个步骤，分别是：JDK的安装，Scala的安装，Spar...

u011513853 2016-10-19 23:40:21 阅读数：66540

Spark在Win10下的环境搭建

前言本章将介绍如何在WIN10下实现spark环境搭建。本章概要1、版本说明2、环境准备：jdk配置；scala安装与配置；spark安装与配置；hadoop安装与配置；版本说明jdk：1.8scal...

songhaifengshuaige 2018-03-08 10:15:06 阅读数：1626

Ubuntu Spark 环境搭建

在安装Spark之前，我们需要在自己的系统当中先安装上jdk和scala 可以去相应的官网下载：JDK：http://www.oracle.com/technetwork/java/javas...

u010171031 2016-07-07 11:04:32 阅读数：17034



python的正确学习路线,你一定不知道

python学习路线

spark开发环境搭建（基于idea 和maven）

使用idea构建maven 管理的spark项目，默认已经装好了idea 和Scala,mac安装Scala 那么使用idea 新建maven 管理的spark 项目有以下几步: scala插件...

u012373815 2016-11-22 00:26:56 阅读数：23007

spark学习笔记（4）IntelliJ IDEA搭建Spark开发环境

基于IntelliJ IDEA开发Spark的Maven项目——Scala语言 1、Maven管理项目在JavaEE普遍使用，开发Spark项目也不例外，而Scala语言开发Spark项目的首...

oh\_Mourinho 2016-09-29 14:36:47 阅读数：10131

Spark入门三部曲之第二步Spark开发环境搭建

使用Scala+IntelliJ IDEA+Sbt搭建开发环境提示搭建开发环境常遇到的问题：1.网络问题，导致sbt插件下载失败，解决方法，找到一个好的网络环境，或者预先从我提供的网盘中下载jar(链...

maixia24 2015-08-04 13:13:01 阅读数：3070

spark环境构建及示例

spark环境构建及示例

flykinghg 2016-11-04 15:59:52 阅读数：3968

spark-2.2.0安装和部署——Spark集群学习日记

Spark-2.2.0安装和部署教程

weixin\_36394852 2017-07-24 17:03:03 阅读数：29090




50万码农评论：英语对于程序员有多重要？

不背单词和语法，一个公式学好英语



### Scala 基础入门教程

1.前言最近在参加Hadoop和Spark培训，需要使用Scala，自学了一下作为入门，这里作一个记录。2.下载1) 在scala官网下载，地址: <http://www.scala-lang.org/>...

afandaafandaafanda

2016-03-16 16:14:01

阅读数：13575

### spark安装及环境搭建

安装 版本配套 Spark: 1.6.2 Scala: 2.12.1 软件安装 1、安装JDK 手工配置JAVA\_HOME环境变量，并将JDK的bin目录加入Path环境变量中。 ...

yunini2

2017-02-17 15:31:38

阅读数：591

### spark完全分布式集群搭建

最近学习Spark，因此想把相关内容记录下来，方便他人参考，也方便自己回忆吧 spark开发环境的介绍资料很多，大同小异，很多不能一次配置成功，我以自己的实际操作过程为准，详细记录下来。 1、基本...

hit0803107

2016-10-12 10:56:10

阅读数：17895

### SparkML (一) Spark的环境搭建与运行

做Spark也有段时间了，主要是平台方面的东西源码也改过些。不过总觉得还是应用才是王道，加上现在AI日趋火爆，抽点时间学习下SparkML吧。—前言我博客里SparkML系列的文章是基于Spark机器...

a071800

2017-09-11 22:26:38

阅读数：451

### spark本地java开发环境的搭建

基于Java开发Spark HelloWorld 绪论 对于学习任何一门新的开发语言或者新的技术，常常都是从HelloWorld开发写起，文章主要介绍在本地环境下如何构建Spar...

liujianhuiouc

2015-12-11 15:18:06

阅读数：1662



客户管理系统  
免费开源CRM客户关系管理软件系统

### Intellij搭建spark开发环境

spark怎么学习呢？在一无所知的前提下，首先去官网快速了解一下spark是干什么的，官网在此。然后，安装开发环境，从wordcount开始学习。第三，上手以后可以学习其他算法了。最后，不要放弃，继续...

pirage

2015-12-08 11:18:42

阅读数：18257

### Intellij Idea搭建Spark开发环境

在Spark快速入门指南 – Spark安装与基础使用中介绍了Spark的安装与配置，在那里还介绍了使用spark-submit提交应用，不过不能使用vim来开发Spark应用，放着IDE的方便不用。 ...

u012877472

2016-03-30 14:24:27

阅读数：16613

#### 个人资料



仙道菜

关注

原创	粉丝	喜欢	评论
52	855	309	217

等级：  博客 5      访问：44万+

积分：3691      排名：1万+

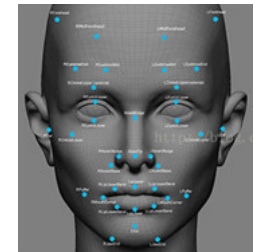
勋章：



杭州单身公寓



关于我



北京航空航天大学 - 研三

关注：计算机视觉、机器学习等

微博：Libra\_Leo\_

知乎专栏：AutoVision

邮箱：cyh@buaa.edu.cn

博主专栏



游戏编程模式

阅读量：197838    18 篇



PySpark-机器学习

阅读量：181067    10 篇

个人分类

- 【Algorithm】6篇
  - 【计算机视觉】6篇
  - 【技术大乱炖】23篇
  - 【编程设计模式】6篇
  - 【pySpark 教程】2篇
- 展开

归档

- 2018年1月 1篇
  - 2016年6月 3篇
  - 2016年5月 2篇
  - 2016年2月 2篇
  - 2016年1月 2篇
- 展开

心理抑郁测试题



联系我们



请扫描二维码联系客服  
✉ webmaster@csdn.net  
☎ 400-660-0108  
💬 QQ客服    💬 客服论坛

关于    招聘    广告服务    网站地图  
©2018 CSDN版权所有 京ICP证09002463号  
🔗 百度提供支持

经营性网站备案信息  
网络110报警服务  
中国互联网举报中心  
北京互联网违法和不良信息举报中心