Dl4j使用Spark分布式训练指定CPU后端训练

问题描述

1、打包dl4j应用程序，使用`spark-submit`在spark集群上分布式运行；
示例提交命令：

```
spark-submit --class cn.nd4jonSpark.Nd4jTest
    --master spark://storm6:7077
    --deploy-mode client
    --driver-memory 4g
    --executor-memory 2g hdfs://ns1/spark_lib/Nd4jTestOnMllib-0.0.1-SNAPSHOT.jar > ../logs/err.log
```

2、Spark分布式集群上均无英伟达显卡，并且没有安装cuda；
3、dl4j应用打包中的pom文件，已经指定backend为`nd4j-native-platform`

但是在运行Spark分布式运行的时候，回去默认寻找cuda后端，尝试使用显卡进行分布式训练，导致报错：

```
WARN scheduler.TaskSetManager: Lost task 1.0 in stage 0.0 (TID 1, 10.100.2.10 , executor 8): java.lang.NoClassDefFoundError: Could not initialize class org.nd4j.linalg.factory.Nd4j
        at org.nd4j.Nd4jRegistrator.registerClasses(Nd4jRegistrator.java:20)
        at org.apache.spark.serializer.KryoSerializer$$anonfun$newKryo$6.apply(KryoSerializer.scala:135)

....

Caused by: java.lang.RuntimeException: No CUDA devices were found in system
```

解决方案

在官方文档 https://deeplearning4j.org/cn/gpu 中提到两个属性

设置环境变量BACKEND_PRIORITY_CPU和BACKEND_PRIORITY_GPU

环境变量BACKEND_PRIORITY_CPU和BACKEND_PRIORITY_GPU的设置可以决定采用的是GPU还是CPU后端。具体用法是将BACKEND_PRIORITY_CPU和BACKEND_PRIORITY_GPU设置为整数。最高的值对应的后端将被采用。

这两个属性会决定dl4j应用程序使用什么样的后端进行模型的训练。

编辑 `/spark/conf/spark_env.sh` 文件，修改其中的环境变量(使用ssh将该文件传输到每一个worker机器上，更改其文件配置)：

```
export BACKEND_PRIORITY_CPU=110
export BACKEND_PRIORITY_GPU=0
```

让`BACKEND_PRIORITY_CPU`的值大于`BACKEND_PRIORITY_GPU`，就可以指定分布式集群中的每一个worker节点都是用CPU后端进行训练。

由群友 @赵彦辉-大连 发现并提供解决方案

更多文档可以查看 https://github.com/sjsdfg/deeplearning4j-issues。
欢迎star