WOTD

51CTO

全球 软件开发技术峰会

World Of Tech 2017

2017年12月1日-2日 • 深圳中洲万豪酒店

DEVELOPMENT

WOTD

51CTO

深度学习Deep CTR 套件

胡南炜

微博 架构师

目录

01 概述

- 2016-06 Google, Play 应用推荐引入 Wide & Deep，3.9%
- 2016-09 Google, YouTube 视频推荐引入 DNN 模型
- 2017-05 Twitter, 时间序引入深度学习，互动率和时长显著 增长
- 2017-07 美团，搜索推荐引入深度学习，3%AUC

背景

- 标准化 更省心
- 工具化 更简单
- 服务化 更快速

功能

- 输入：csv格式，数据清洗等
- 特征：连续、离散、文本、标签...
- 模型：DNN、Wide&Deep
- 预测：开箱即用
- 分布式计算

02 训练与分析

- 配置
- 训练
- 分析

02 训练与分析

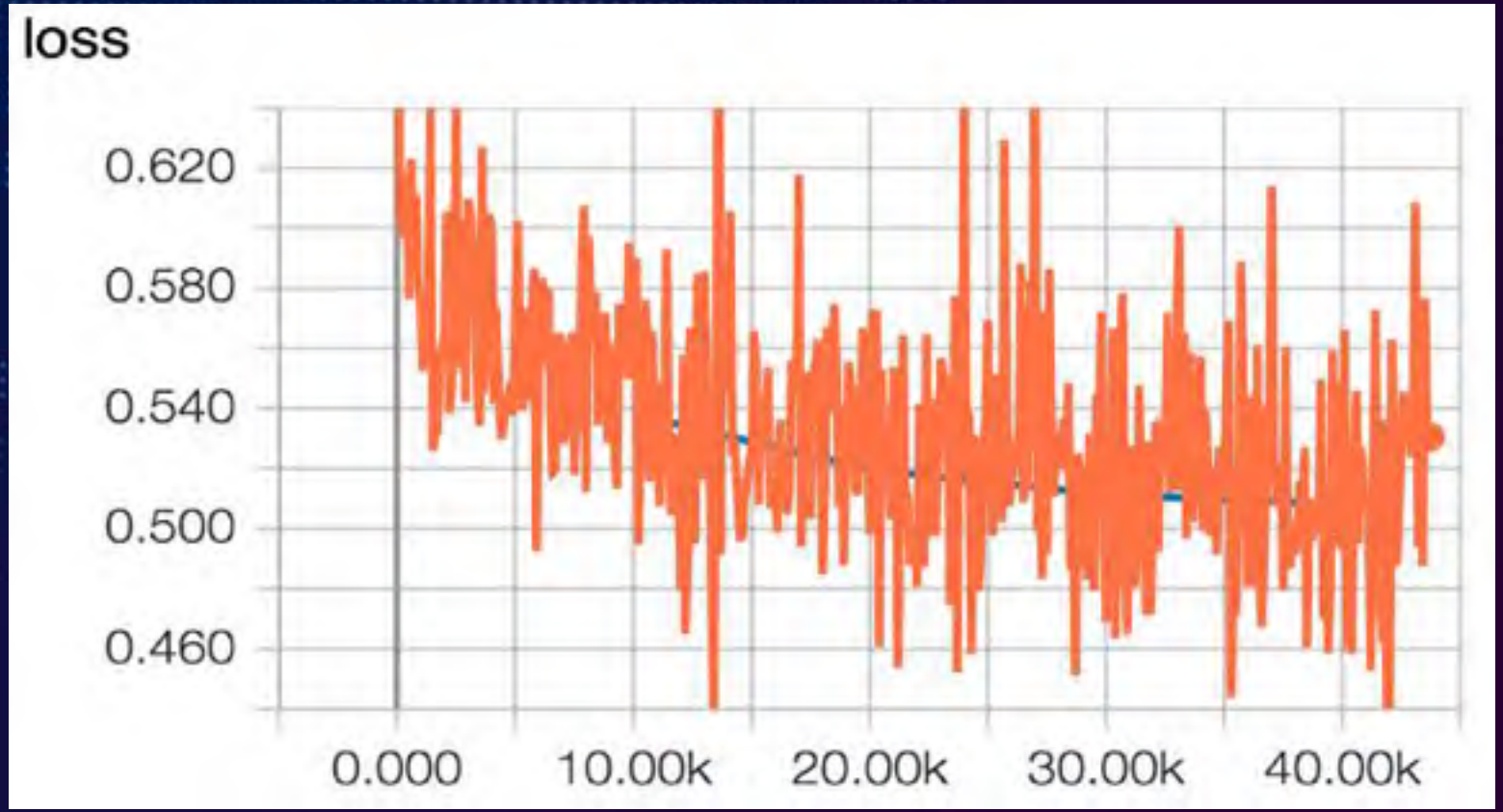
- 配置
- 训练
- 分析

配置

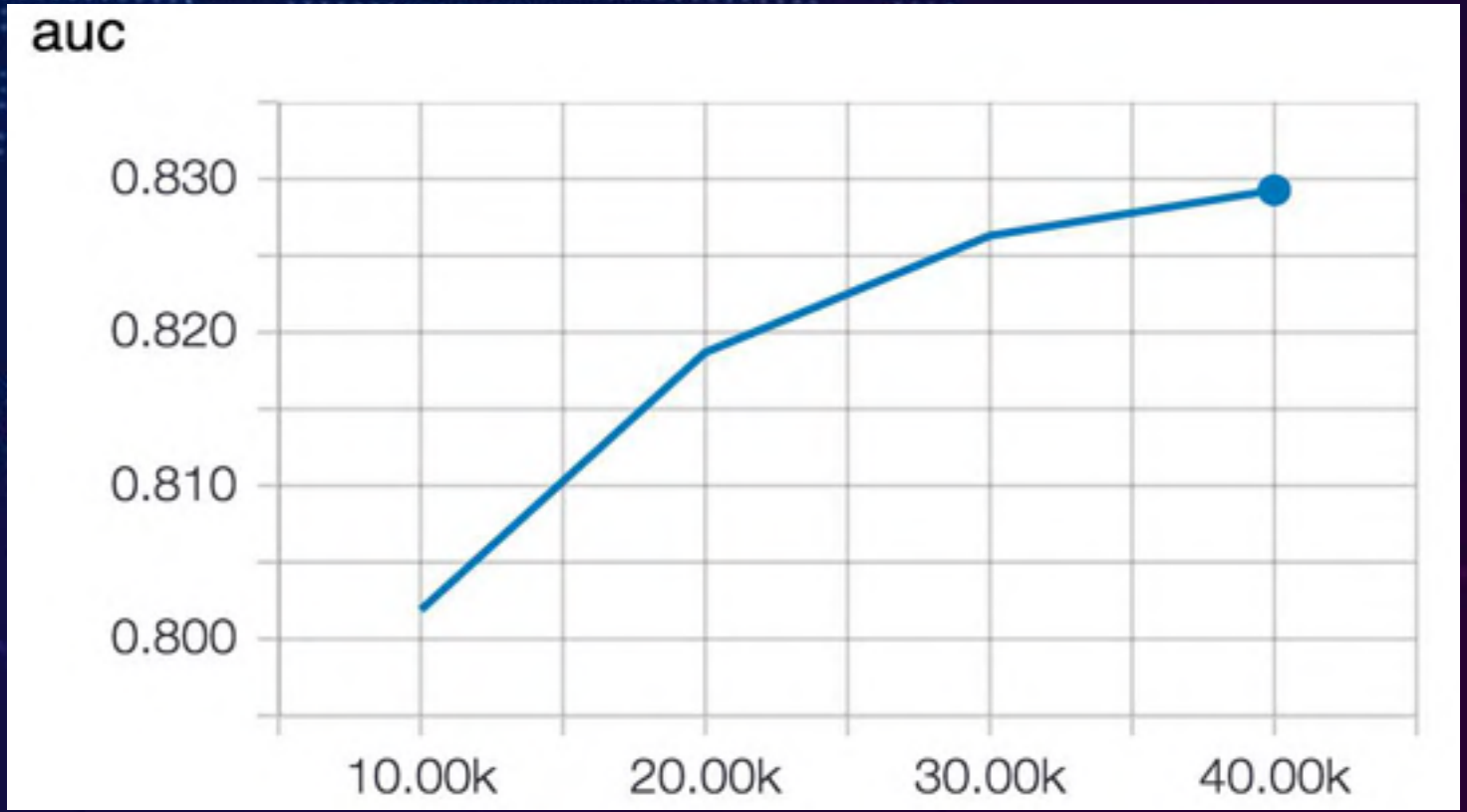
```
<config version="0.0.3" model="prometheus">
    <input>
        <train_data>./data/train_ori.csv</train_data>
        <test_data>./data/test_ori.csv</test_data>
        <batch_size>256</batch_size>
        <header>True</header>
        <label_index>3</label_index>
    </input>
    <columns>
        <column name="index" id="0" cls="continuous" dtype="float" unused="True"></column>
        <column name="uid" id="1" cls="category" dtype="int">
            <num_buckets>6041</num_buckets>
            <dimension>32</dimension>
        </column>
        <column name="mid" id="2" cls="category" dtype="int">
            <num_buckets>3953</num_buckets>
            <dimension>32</dimension>
        </column>
        <column name="score" id="3" cls="continuous" dtype="int"></column>
        <column name="time" id="4" cls="continuous" dtype="int" unused="True"></column>
        <column name="gender" id="5" cls="category" dtype="str">
            <vocab_list>M, F</vocab_list>
            <dimension>16</dimension>
        </column>
        <column name="age" id="6" cls="category" dtype="int">
            <num_buckets>57</num_buckets>
            <dimension>16</dimension>
        </column>
        <column name="occupation" id="7" cls="category" dtype="int">
            <num_buckets>22</num_buckets>
            <dimension>16</dimension>
        </column>
```

```
NORMAL  TST/main_feed>  config/config.xml                    xml  utf-8[unix]   1%  :  1/76  :  1
[prome] 0:facai077-112-153-hadoop-test:~/first_features_96-  1:vim  2:vim*        "bj-m-203050a.local" 18:46 22-Sep-17
```

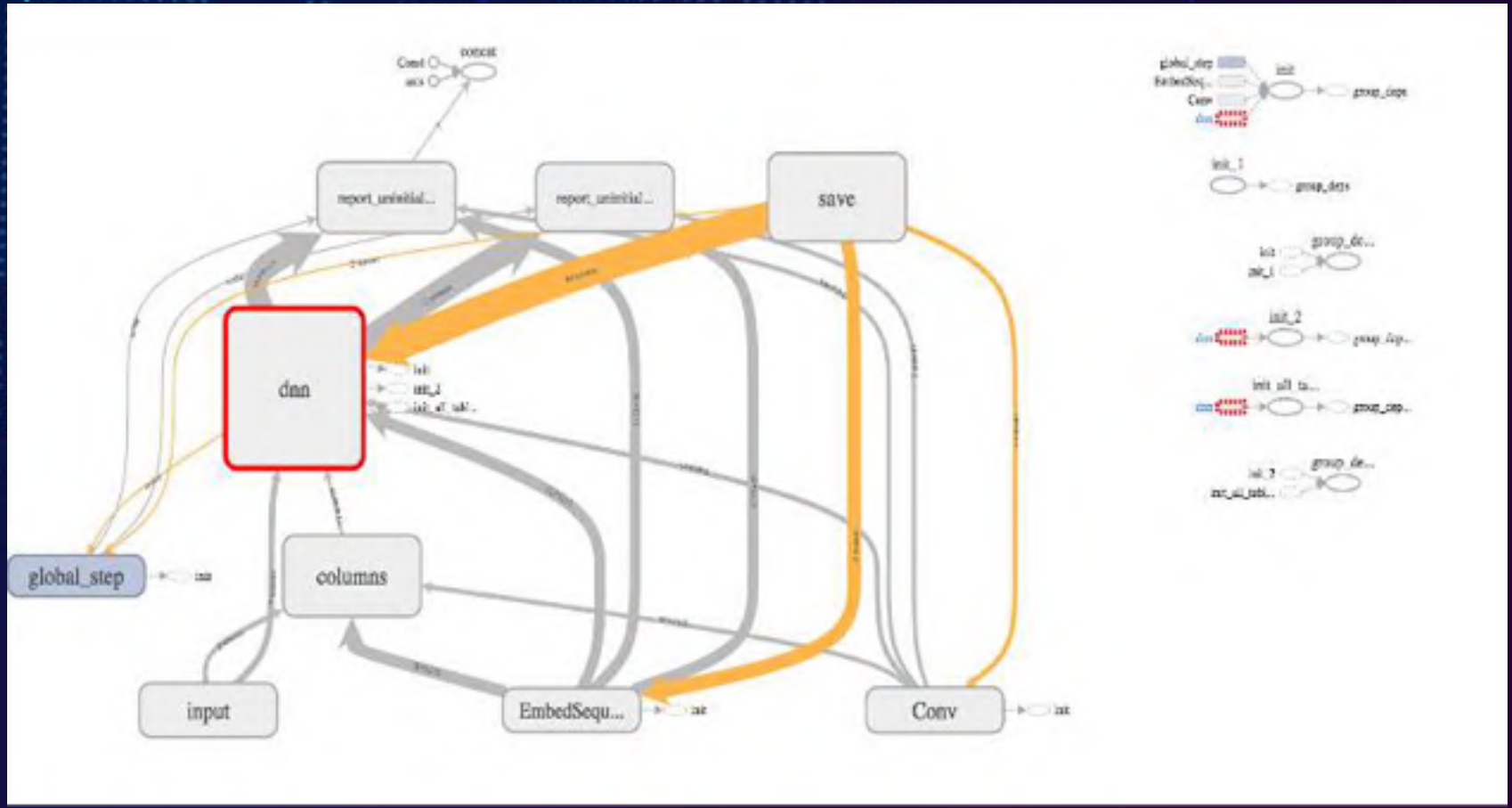
训练



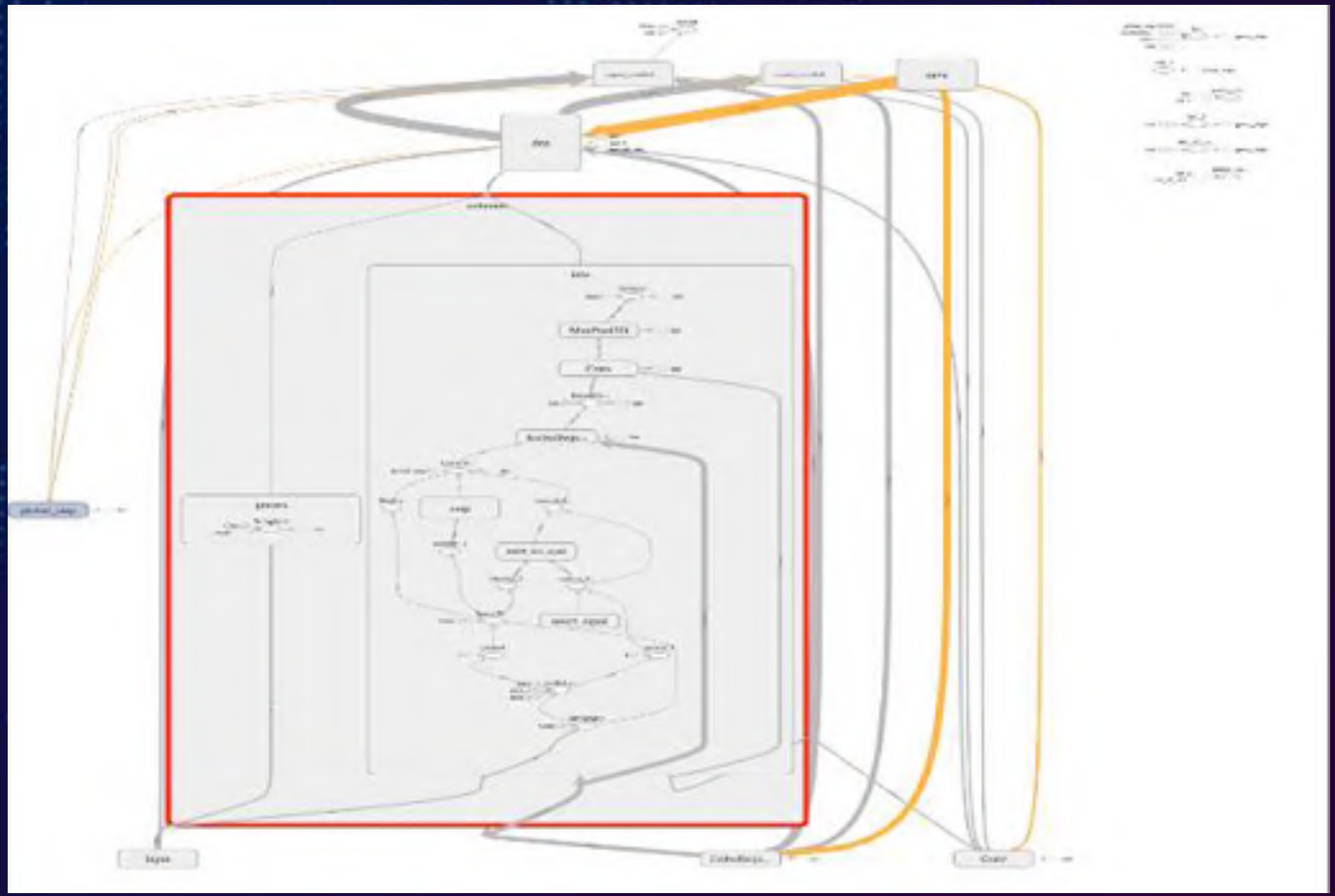
训练



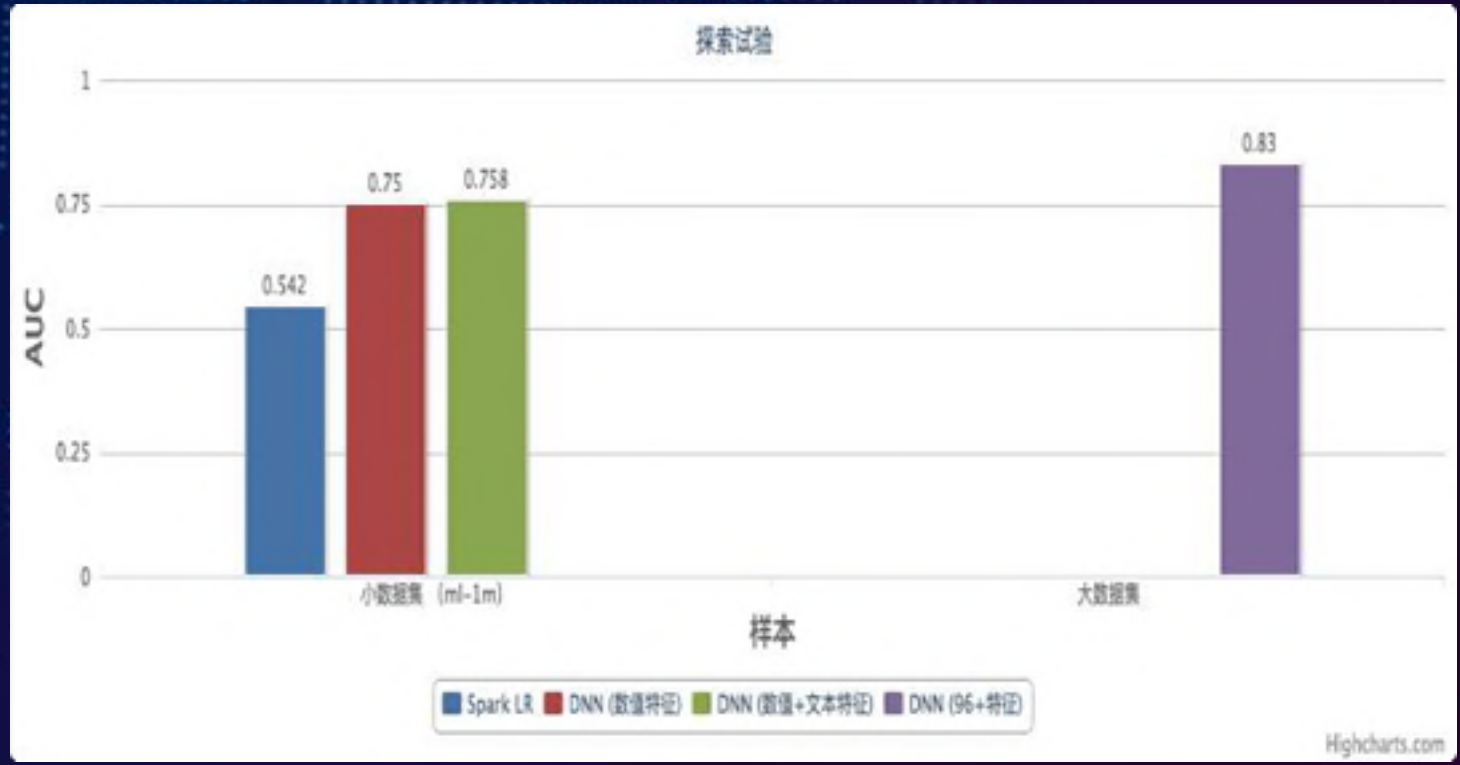
分析


concat
report_uninitial...
report_uninitial...
save
dnn
global_step
columns
input
EmbedSequ...
Conv

分析



分析

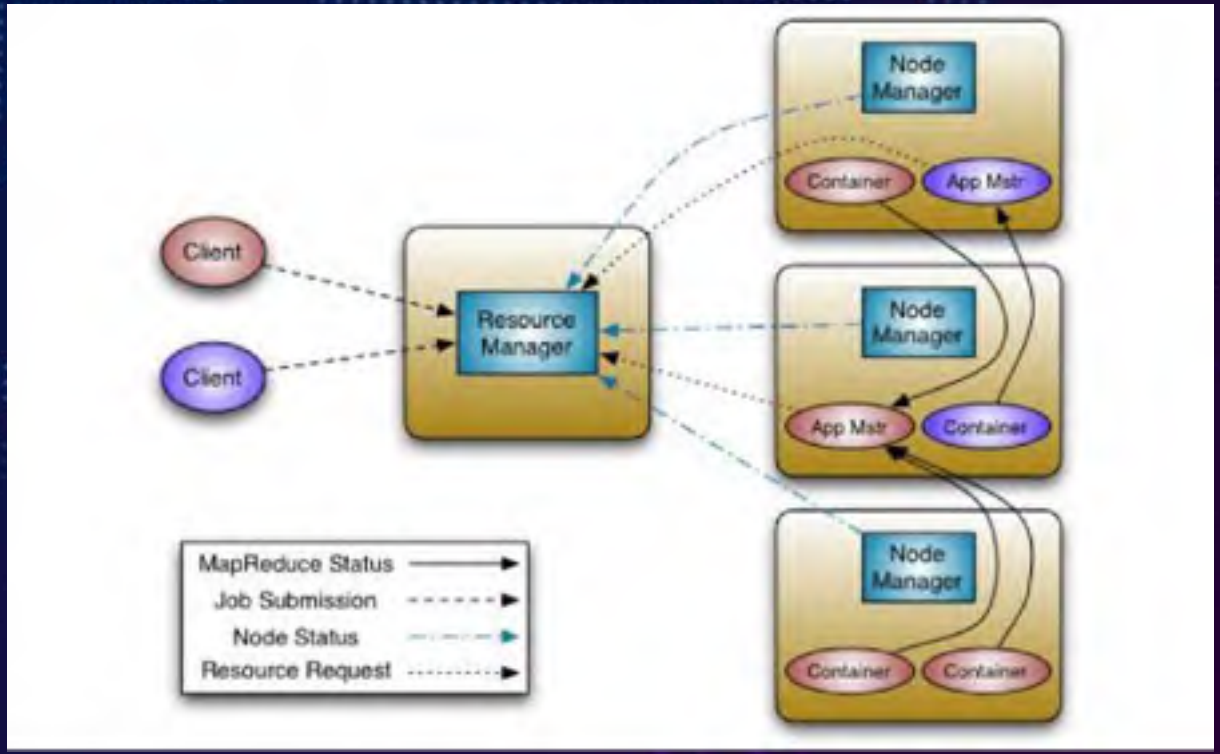

探索试验
1
0.75
0.5
0.25
0
AUC
0.542
0.75
0.758
0.83
小数据集（ml-1m）
大数据集
样本
Spark LR
DNN (数值特征)
DNN (数值+文本特征)
DNN (96+特征)
Highcharts.com

03 导出与预测

```
<config version="0.0.3" model="prometheus">
    <output>
        <export_dir_base>file:///tmp/facai/prome/export</export_dir_base>
    </output>
</config>
```

```
client = dnn_tfserving_client.DnnClient(FLAGS.server)
inputs_dict = {
    "uid": _int64_feature(value=1),
    "mid": _int64_feature(value=1193),
    "time": _float_feature(value=0),
    "gender": _bytes_feature(value='M'.encode()),
    "age": _int64_feature(value=25),
    "job": _int64_feature(value=0),
    "zip_code": _float_feature(value=0),
    "title": _bytes_feature(value='Shine'.encode()),
    "genres": _bytes_feature(value='Drama'.encode())}
result = client.do_predict(FLAGS.model, inputs_dict, FLAGS.timeout)
```

04 分布式计算-TensorFlow on yarn



04 为什么TensorFlow on yarn

- 集群资源的统一管理和分配
- 作业统一管理，运行状态实时跟踪，在线的log查看
- 作业进程的资源隔离
- 利用微博现有的Hadoop集群

04 TensorFlow on yarn-资源分配

- AppMaster自动分配PS hosts和Worker Hosts，并自动分配端口号
- 自动分配TensorBoard可视化所需的host和端口号
- 对训练数据自动进行数据分片
- 启动PS Task和Worker Task，启动TensorBoard

04 TensorFlow on yarn-模型训练

- TensorFlow根据PS hosts和Worker hosts组装成ClusterSpec
- Worker task从HDFS读取其分片的训练数据，进行训练
- 训练过程中定期将checkpoint保存到HDFS中
- 通过TensorBoard实时查看计算情况

04 TensorFlow on yarn-资源回收

- 训练结束后模型保存到HDFS
- AppMaster强制结束PS进程
- AppMaster强制结束TensorBoard进程

04 TensorFlow on yarn-示例启动脚

```
# ----------------------------------------------------------------------
# Start Script for launching the Tensorflow job in cluster.
# ----------------------------------------------------------------------
hadoop jar agent.jar com.weibo.flyhorse.yarn.Client hdfs:///user/weibo/agent.jar TF " \
--command_file train.py \
--command_path hdfs:///user/weibo/train.py \
--input hdfs://77-16-121-hadoop:9000/user/weibo/feed_data_v0/train/20170703 \
--output hdfs://77-16-121-hadoop:9000/user/weibo/tensorflow/train/context/output \
--num_of_worker 10 \
--num_of_ps 5 \
--memory_of_worker 4096 \
--memory_of_ps 1024 \
--num_epochs 1 \
--board_enable true"
```

总结

- 标准化 ：支持全部微博特征
- 工具化 ：简化了深度学习引入排序模型
- 服务化 ： 支持Tf-Serving预测服务

Thank you!