

▲

-1

▼

华为在深度学习平台上的优化实践

- 深度学习 (<http://www.csdn.net/tag/深度学习/news>)
- Kubernetes (<http://www.csdn.net/tag/Kubernetes/news>)
- 华为 (<http://www.csdn.net/tag/华为/news>)

阅读 591

⚠

“Kubernetes Meetup 中国 2017”——北京站3.18落幕啦！本次分享嘉宾彭靖田来自华为，他的分享题目是《华为在深度学习平台上的优化实践》。实录将从深度学习平台的架构、优化等几个方面，介绍华为在深度学习平台上的实践。本文由才云科技供稿。

彭靖田：华为中软大数据工程师。2016年毕业于浙大竺可桢学院求是科学班，加州大学圣迭戈分校访问学者。毕业加入华为后，主要从事深度学习平台的设计和研发工作，专注开源社区。先后在 TensorFlow 社区独立贡献了 Mnist 分布式模型、VariableAsGradients、InitWithoutInitializer 等特性。同时，从 Kubernetes v1.3开始，参与维护 Kubernetes 社区 CentOS 平台相关脚本。

今天第一部分主要讲华为在深度学习方面的应用需求以及华为在深度学习平台遇到的一些挑战。第二部分是讲华为深度学习平台的架构、优化以及经验。

提纲

- 华为深度学习的应用需求
- 华为深度学习平台的挑战
- 华为深度学习平台的架构
- 华为深度学习平台的优化
- 华为深度学习平台的应用
- 总结



这两年，深度学习取得了突破性发展。尤其是在语音识别和图像识别这两方面。

在 ImageNet 图像分类任务上，AI 现在的错误率2.9%已经超越人类5%了。去年的 AlphaGo 又一次在围棋领域打败了人类顶尖高手李世石。今年年初的时候，AlphaGo 2.0 Master 大败中日韩三国高手，围棋领域也被 AI 突破。最近深度学习还被应用在在图像风格迁移 Prisma 和皮肤癌诊断。

背景：深度学习快速发展

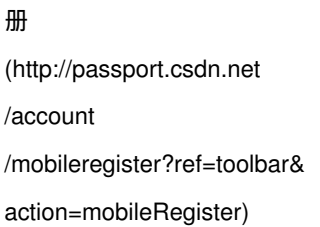
- 深度学习近年成为最热门的机器学习分支
 - 语音识别、图像识别取得突破性进展；
 - ImageNet：分类错误率**2.9%**，优于人类**5%**；
 - 围棋：AlphaGo战胜李世石，Master连胜60局
 - 各种深度学习框架TensorFlow、MXNet等；
- 深度学习的发展机遇
 - 持续不断的海量数据
 - 高速增长的计算能力
 - 日渐降低的带宽成本



登录

(<https://passport.csdn.net>
华为路由器，但
/account
的人口，业务开
login?ref=toolbar)注
册
(<http://passport.csdn.net>
/account
/mobileregister?ref=toolbar&
action=mobileRegister)

(https://passport.csdn.net
华为路由器，但
/account
的人口。业务开
login?ref=toolbar)/注
册
(http://passport.csdn.net
/account
/mobileregister?ref=toolba
action=mobileRegister)



```
action=mobileRegister)
```

- ```
action=mobileRegister)
```



```
action=mobileRegister)
```





(http://www.csdn.net/ref=toolbar)

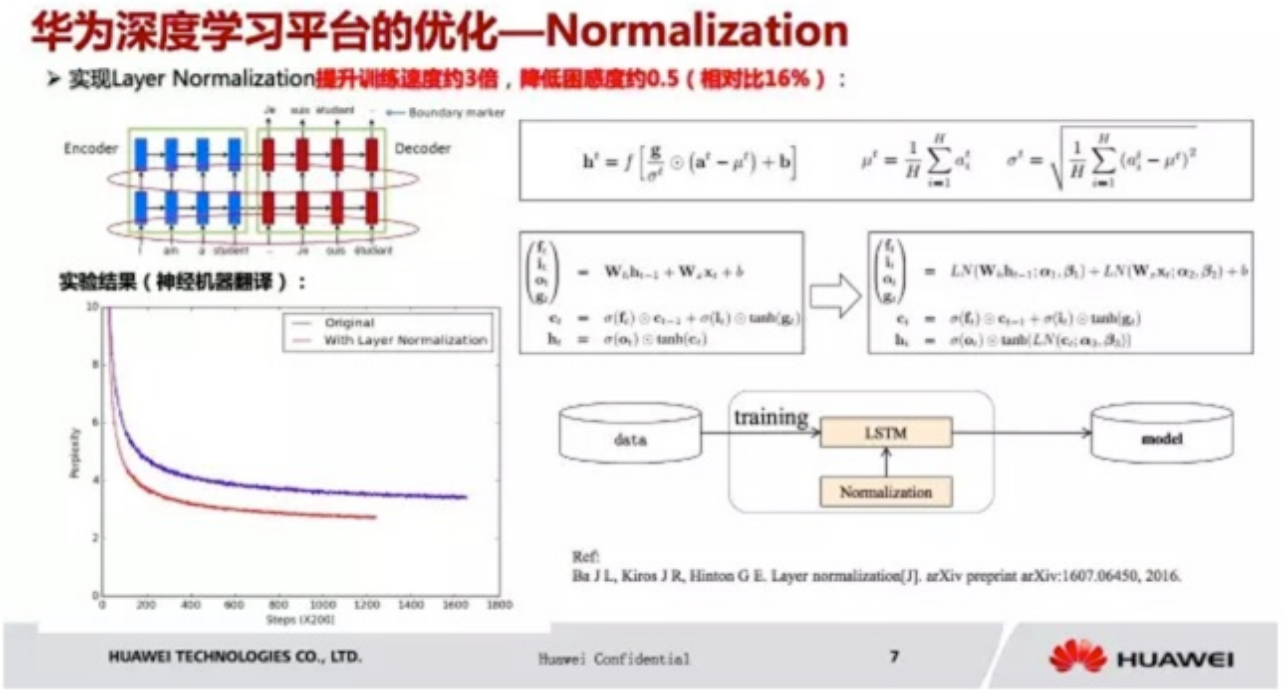


登录

(https://passport.csdn.net/account/login?ref=toolbar)| 注册

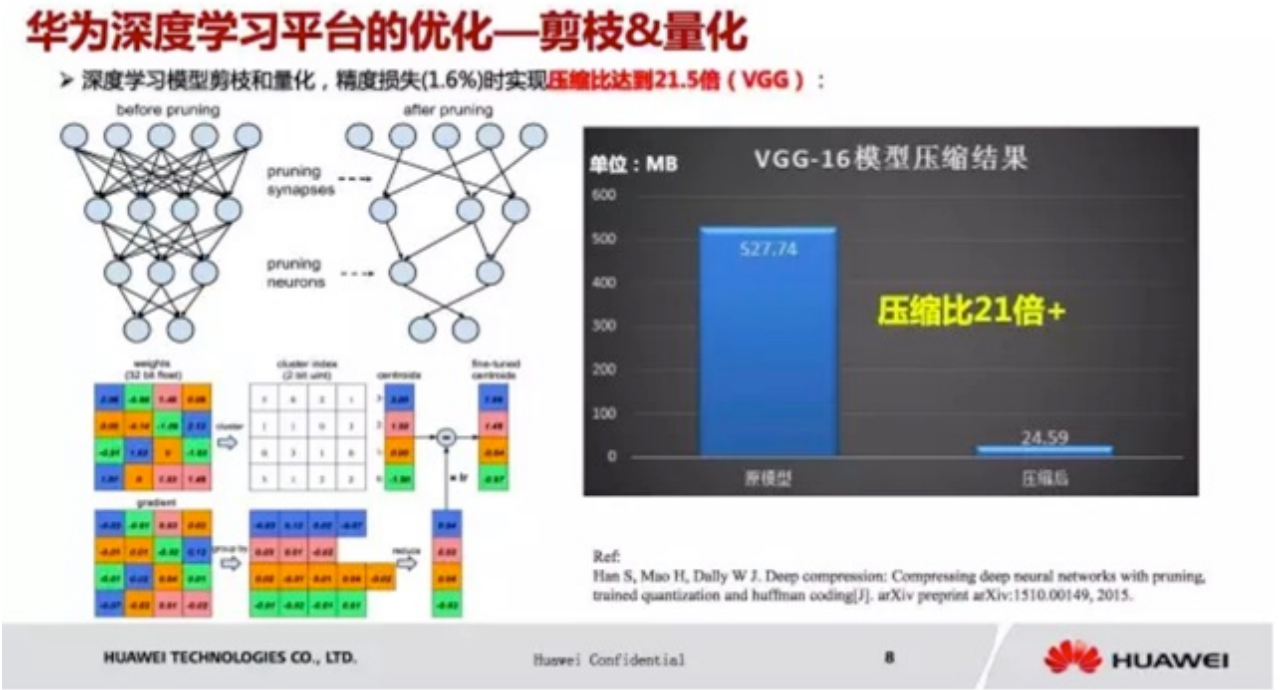
(http://passport.csdn.net/account/mobileregister?ref=toolbar&action=mobileRegister)

Normalization 是深度学习领域常用的优化技术，Batch Normalization 已经在 CNN 领域取得了广泛应用。ICLR 2017 Hinton 提出的 Layer Normalization 则有效解决了 RNN 的计算加速问题，两者本质都是在解决深层网络梯度弥散/爆炸的问题。



接下来讲讲基于剪枝和量化的模型压缩。

通常，在训练结束后，全连接层和卷积层有大量参数值是接近于0的，对于 inference 的贡献很少。我们可以设置一个接近0的 threshold，将小于它的值设为0。这样可以将稠密参数矩阵中大量0值点剪掉，转换为稀疏的参数矩阵，这一步剪枝可以实现9倍左右的压缩。但是压缩完之后，我们还是觉得不够，我们还可以做量化，虽然我们用32bit 的单精度去存储每一个 weight，但是这个 weight 真正的值的分布类型其实比较少。大部分的 weights 是比较靠近的，这个时候我们想做量化。我们把 32bit 的变成一个 2bit 的 Unit，整个参数的值，比如这个是I类，这个类型附近的值都用它来代替，其它也是一样的，那这样的话，我们就把 32bit 的 weights 变成了 2bit 的 weights。这样的压缩效果是16倍，但是后来发现，这样虽然压缩效果很好，但是对于准确率的影响是比较大的。所以我们就采用了8bit一个量化，压缩效果是4倍。

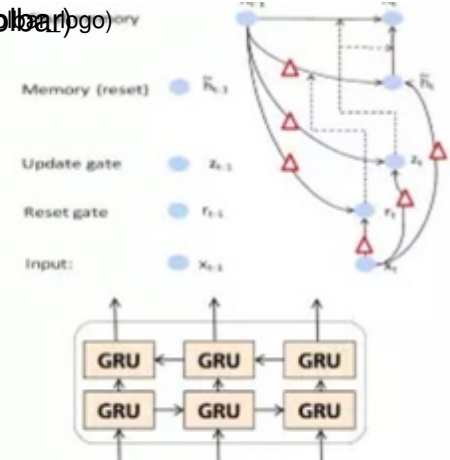


接下来讲的是 SVD 分解。

当参数很大的时候要怎么办呢，这个时候可以使用 SVD 分解，比如说我有一个大矩阵，换成三个小矩阵，进行瘦身，这样的好处是，被压缩了，因为存储变小了；第二个是量变小了，小了之后可以进行加速，SVD 分解，所以模型压缩也是一个很直观的方法。大家可以看到，经过5倍的压缩之后，损失几乎没有。



(http://www.csdn.net?ref=toolbar)



|                 | (Fe1,Fe2,Fe3) | 原始模型   | (conv=6,4096,4096,1000) | (conv=6,512,256,1000) |
|-----------------|---------------|--------|-------------------------|-----------------------|
| Accuracy        | Top-1         | 70.64% | 68.95%                  | 68.67%                |
|                 | Top-5         | 89.66% | 88.67%                  | 88.56%                |
| Model size      |               | 540MB  | 498MB                   | 98MB                  |
| Compress Ratio  |               | 1×     | 1.08×                   | 5.51×                 |
| Actual speed up |               | 1×     | 2×                      | 2×                    |

基于CNN的实验结果：VGG模型压缩5倍，运行时加速2倍。

Ref: Zhang X, Zou J, He K, et al. Accelerating very deep convolutional networks for classification and detection[J]. IEEE transactions on pattern analysis and machine intelligence, 2016, 38(10): 1943-1955.

HUAWEI TECHNOLOGIES CO., LTD.

Huawei Confidential

9

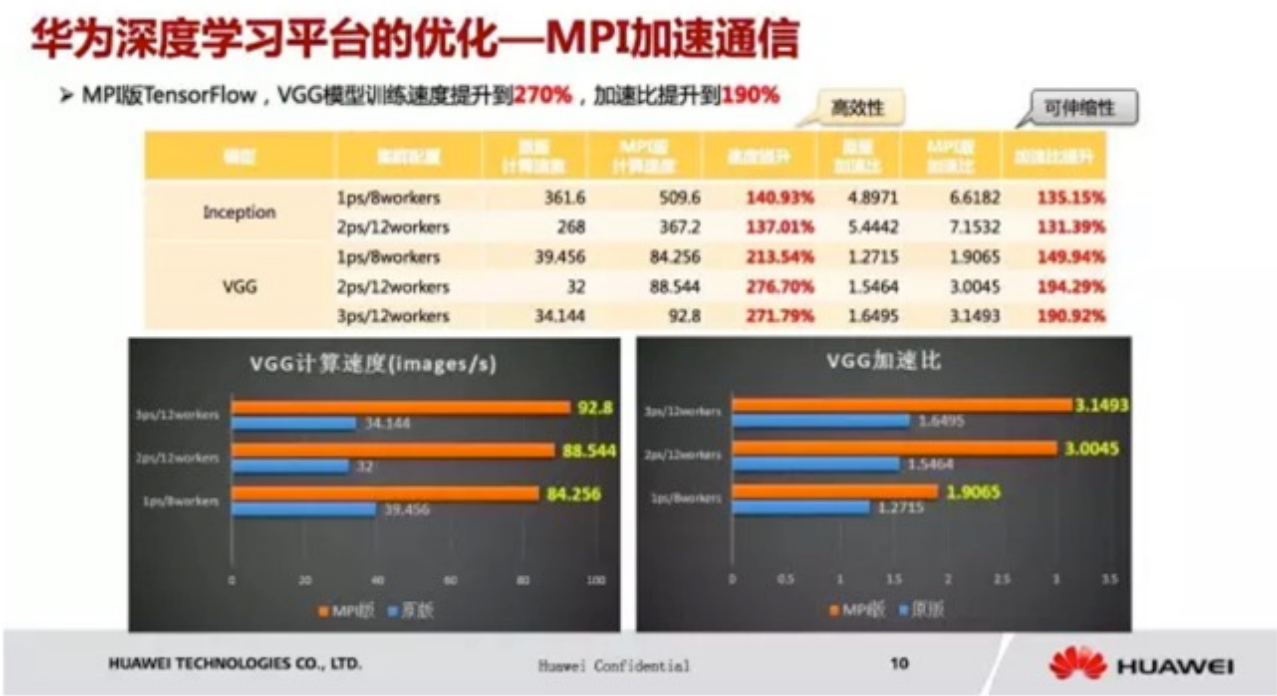
HUAWEI

登录

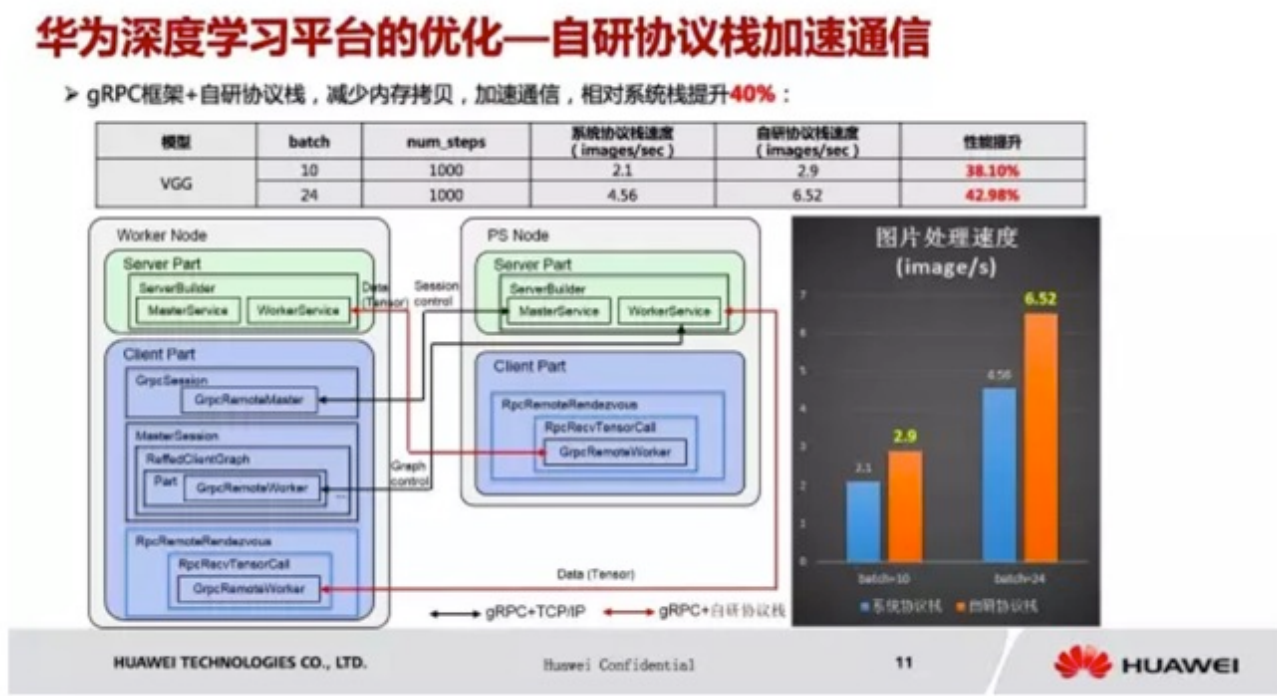
(https://passport.csdn.net/account/login?ref=toolbar)| 注册

(http://passport.csdn.net/account/mobileregister?ref=toolbar&action=mobileRegister)

通信加速我们选择在高性能计算领域广泛使用的 MPI，我们开发了 MPI 版 TensorFlow,使用 MPI 传输数据，控制面仍然走 gRPC。我们用 MPI 的方式进行通信，在 VGG 上面，我们看到它提速了2.7倍，计算速度明显提升；加速比增长幅度也很大。



这个是我们自主研发的协议栈，大家看到，TensorFlow 在分布式的时候，如果要发送到另一个节点，减少了内核态向用户态的数据拷贝，仍然使 TensorFlow 的 gRPC，大概有40%的性能提升。



简单讲下我们在深度学习平台上的应用，神经机器翻译。大家知道，现在官方的神经机器翻译只有一个单机单卡版本的，训练速度比较慢。

我们以英语到法语的翻译为例，它其实先建立了英语和法语的语言模型，每一个词在它的语言模型里面都有一个对应的 embedded vector，比如我输入的是一个单词，再到 LSTM 里面，它会把所有的句子之类都存下来，这个信息的多少跟参数是有关系的，然后放到 decoder 那里，最后再放到 RNN 的网络，然后这边对应的也是一个一个的单词，这些单词会被映射到法语里面。



(http://www.csdn.net/ref=toolbar)

### 华为深度学习平台的应用—分布式神经机器翻译

➢ Seq2seq神经机器翻译模型，开源单机单卡版，训练速度慢：

Target Sentence

Softmax

Word Probability

LSTM/GRU

Continuous-Space Word Representation

RNN LM

One-hot Coding

Source Sentence

Ref: Sutskever I, Vinyals O, Le Q V. Sequence to sequence learning with neural networks[C]//Advances in neural information processing systems. 2014: 3104-3112.

HUAWEI TECHNOLOGIES CO., LTD. Huawei Confidential 12 HUAWEI

登录

(https://passport.csdn.net/account/login?ref=toolbar)| 注册

(http://passport.csdn.net/account/mobileregister?ref=toolbar&action=mobileRegister)

这一块的话，我们实现分布式神经机器翻译，2个节点，8个 GPU，处理速度提升4倍多，达到23316（ words/s ）。

### 华为深度学习平台的应用—分布式神经机器翻译

➢ 实现分布式神经机器翻译，2节点8GPU，训练速度提升**4.42倍**，处理速度达**23316(words/s)**：

| 模型参数               | layers = 3, neural size=128, batch_size=32, RNN_cell=GRU |         |         |         |         |         |         |         | 合计    | 加速比  |
|--------------------|----------------------------------------------------------|---------|---------|---------|---------|---------|---------|---------|-------|------|
| 单机单卡               | worker0                                                  | worker1 | worker2 | worker3 | worker4 | worker5 | worker6 | worker7 | 5277  | 1.00 |
| 1ps/4workers       | 1436                                                     | 2077    | 2159    | 1917    |         |         |         |         | 7589  | 1.44 |
| 1ps/4workers(ps同机) | 3956                                                     | 3613    | 4144    | 3761    |         |         |         |         | 15473 | 2.93 |
| 1ps/8workers       | 3731                                                     | 3550    | 3661    | 3423    | 2073    | 1675    | 1214    | 1708    | 21035 | 3.99 |
| 2ps/8workers       | 3935                                                     | 4234    | 4235    | 3554    | 1574    | 1892    | 1872    | 2020    | 23316 | 4.42 |

计算速度(words/s)

23316 words/s

加速比

加速比4.42倍

HUAWEI TECHNOLOGIES CO., LTD. Huawei Confidential 13 HUAWEI

这个是在我们华为应用商城上线的伏羲推荐系统。现在大概有百万级的 App 在各大应用市场，同质化很严重。比如我想选用一个好用的 App，同类型的APP有好几十种，用户选择成本非常高。

现在的话，我们就去学习用户搜索和浏览习惯，做到真正去了解用户想要什么。第二个方法是抽取百亿规模的用户特征，画一个用户画像，实现华为有亿级的注册用户，千万级日活用户，所以点击转化量是非常高的。

### 华为深度学习平台的应用—伏羲推荐系统

➢ 伏羲手机应用推荐系统，应用在国内**上亿部**华为手机：

- 行业现状
  - 百万APP -> 用户选择成本太高；
  - 各大应用市场趋于同质化；
- 解决方案
  - 学习用户搜索和浏览习惯 -> “懂” 用户
  - 抽取**百亿**规模用户特征 -> 用户画像
- 华为手机应用市场
  - 亿级**注册用户
  - 千万级**日活用户
  - 亿级**日均分发量

14

HUAWEI TECHNOLOGIES CO., LTD. Huawei Confidential HUAWEI

总结一下，华为有着广泛的深度学习应用需求和优势，我们在为全球170多个国家提供服务，拥有万亿美元的网络存量和上亿终端用户数据，我们希望通过深度学习等技术将这些高价值数据利用起来，加速华为业务智能化转型。

(http://www.csdn.net/2018/03/22/5301290.html) 第三，我们兼容原生的 TensorFlow 和 MXNet 接口。

第三，我们拥有自主研发的 MPI 版 TensorFlow 和自研协议栈。

### 总结：华为深度学习平台

- 华为有广泛的深度学习应用需求和优势：
  - 数据：服务全球**170个**国家，**万亿**美元的网络存量，**上亿部**华为智能手机；
  - 开放：**支持原生**TensorFlow和MXNet接口；
  - 高性能：自研MPI版TensorFlow和协议栈；
- 华为深度学习平台的挑战：
  - 模型复杂，计算量大，训练数据多，通信密集；
  - 大模型、多框架、大规模GPU集群；
- 华为深度学习平台已成功应用到华为手机**应用市场个性推荐**，覆盖**亿级**用户，推荐点击率明显提升，机器翻译也已取得初步应用



(http://geek.csdn.net/user/publishlist/karamos)  
CSDN魏伟 (http://geek.csdn.net/user/publishlist/karamos)  
发布于 Container (http://geek.csdn.net/forum/53) 2017-03-22 09:58

分享到：

已有0条评论

最新 ▾

评论

还没有评论，赶快来抢沙发吧。

登录

(https://passport.csdn.net/account/login?ref=toolbar)| 注册

(http://passport.csdn.net/account/mobileregister?ref=toolbar&action=mobileRegister)