
欺诈检测中的自动特征选择

设备数据

Mac地址,ip,基站

转账记录

转账时间,身份,金额

面对海量数据不知如何是好
传统的基于规则的金融模型

活动动轨迹

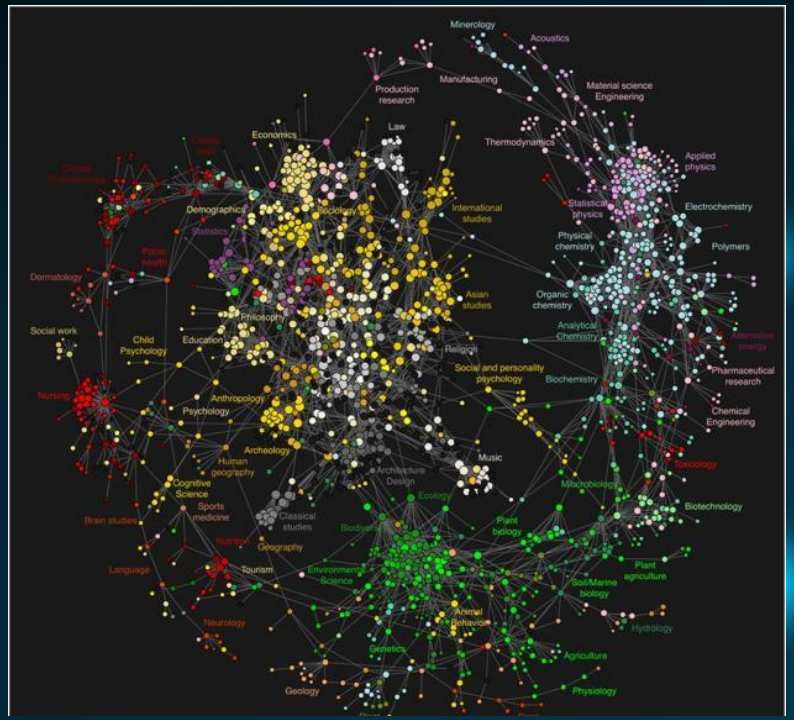
生物特征(输入密码节奏),

电商数据

点击,下单,加入购物车,收藏

关系数据

同人,好友,接近程度...



量大

50亿个点,8000亿条边

异构

每个点有不同的属性

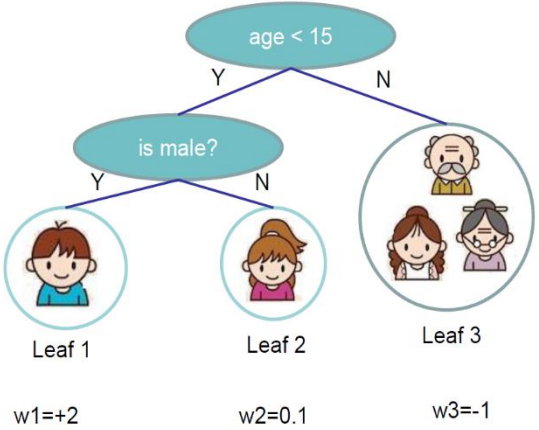
样本少

负样本占比一般在1%以下

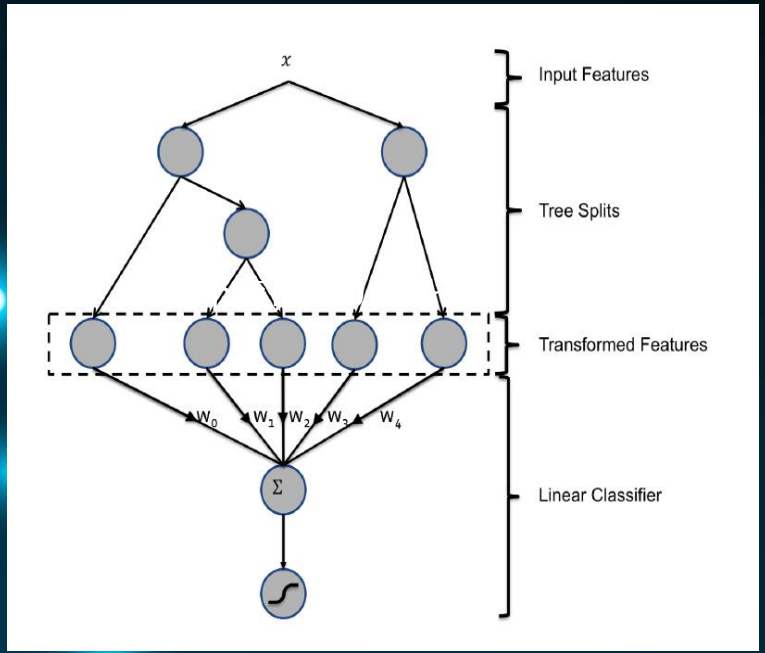
- Define complexity as (this is not the only possible definition)

$$\Omega(f_t) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2$$

Number of leaves　　L2 norm of leaf scores



$$\Omega = \gamma 3 + \frac{1}{2}\lambda(4 + 0.01 + 1)$$



第一阶段:借鉴广告,gbdt数结构作为维度

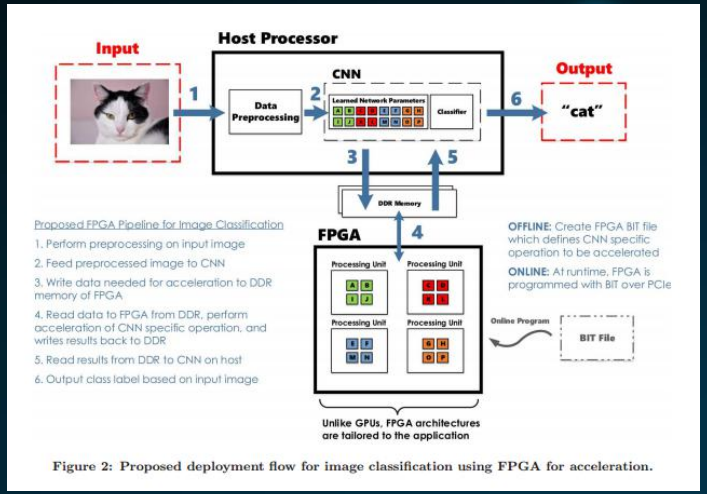
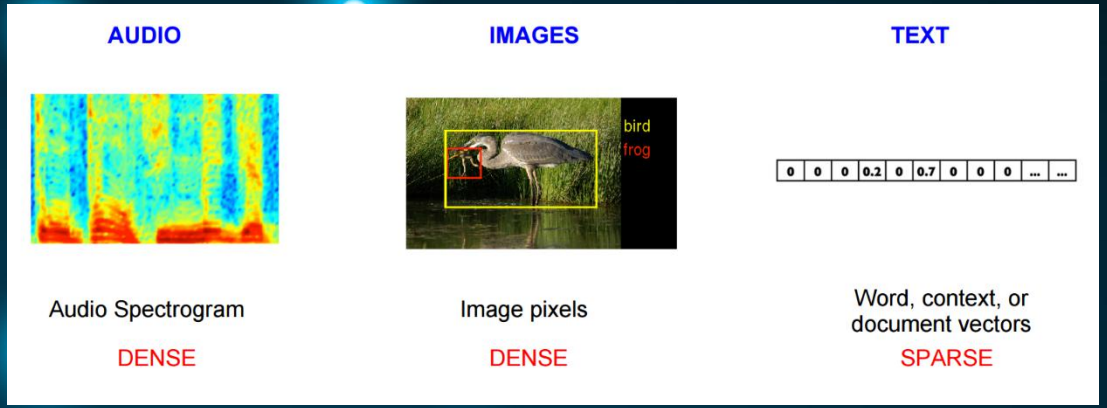
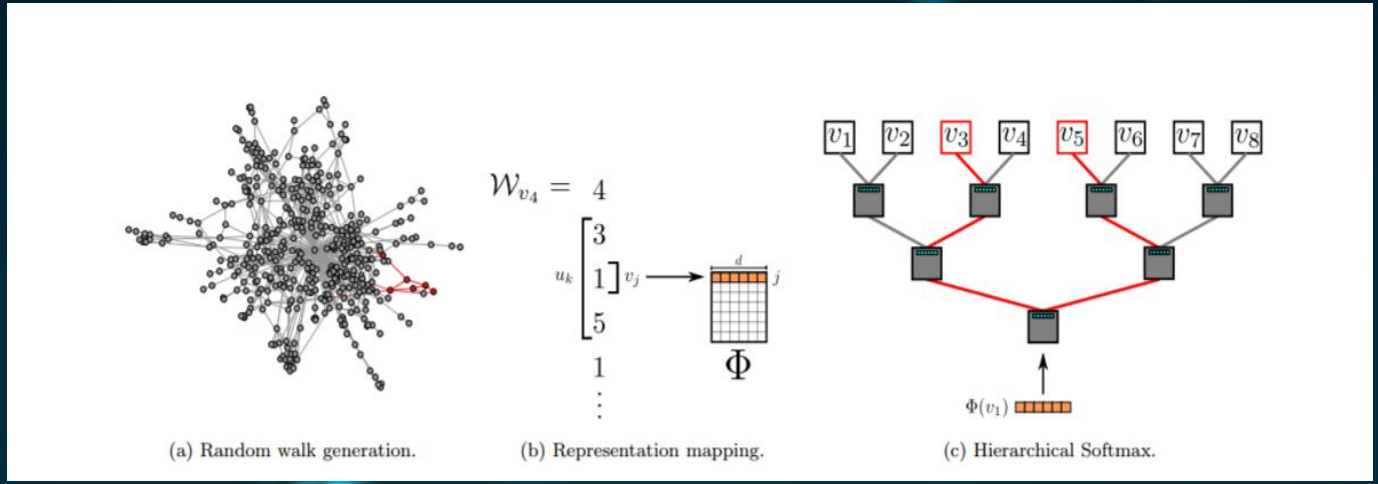


Figure 2: Proposed deployment flow for image classification using FPGA for acceleration.

DL应用在表征学习

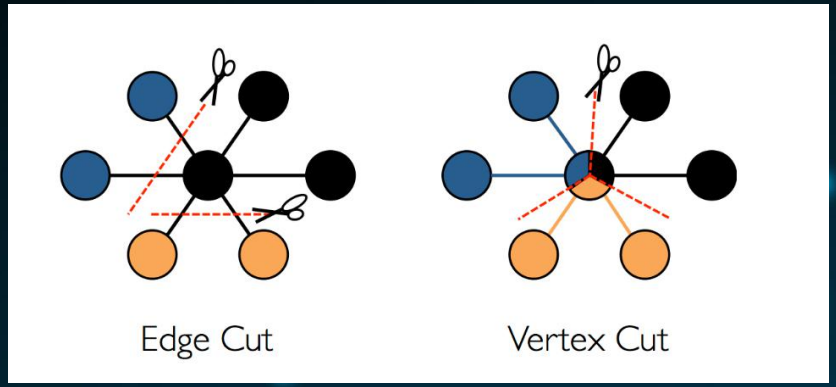


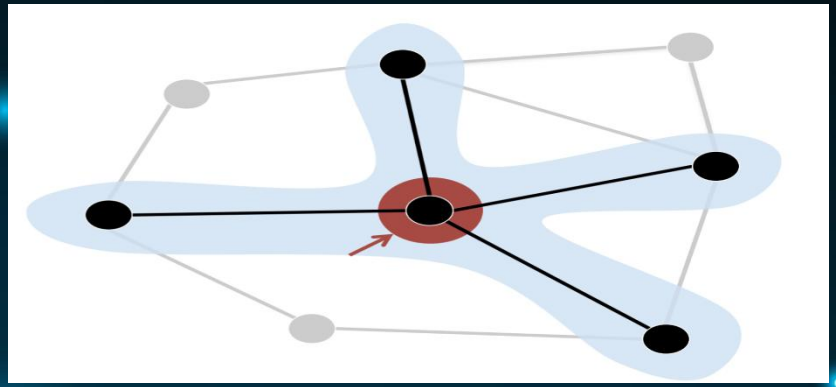
w2v



(a) Random walk generation. (b) Representation mapping. (c) Hierarchical Softmax.

G2V





使用工具:提取维度**graphx,titan**

Machine 1

Machine 2

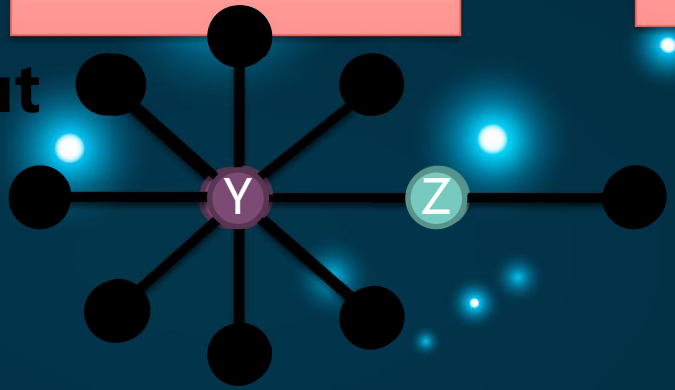
Machine 3

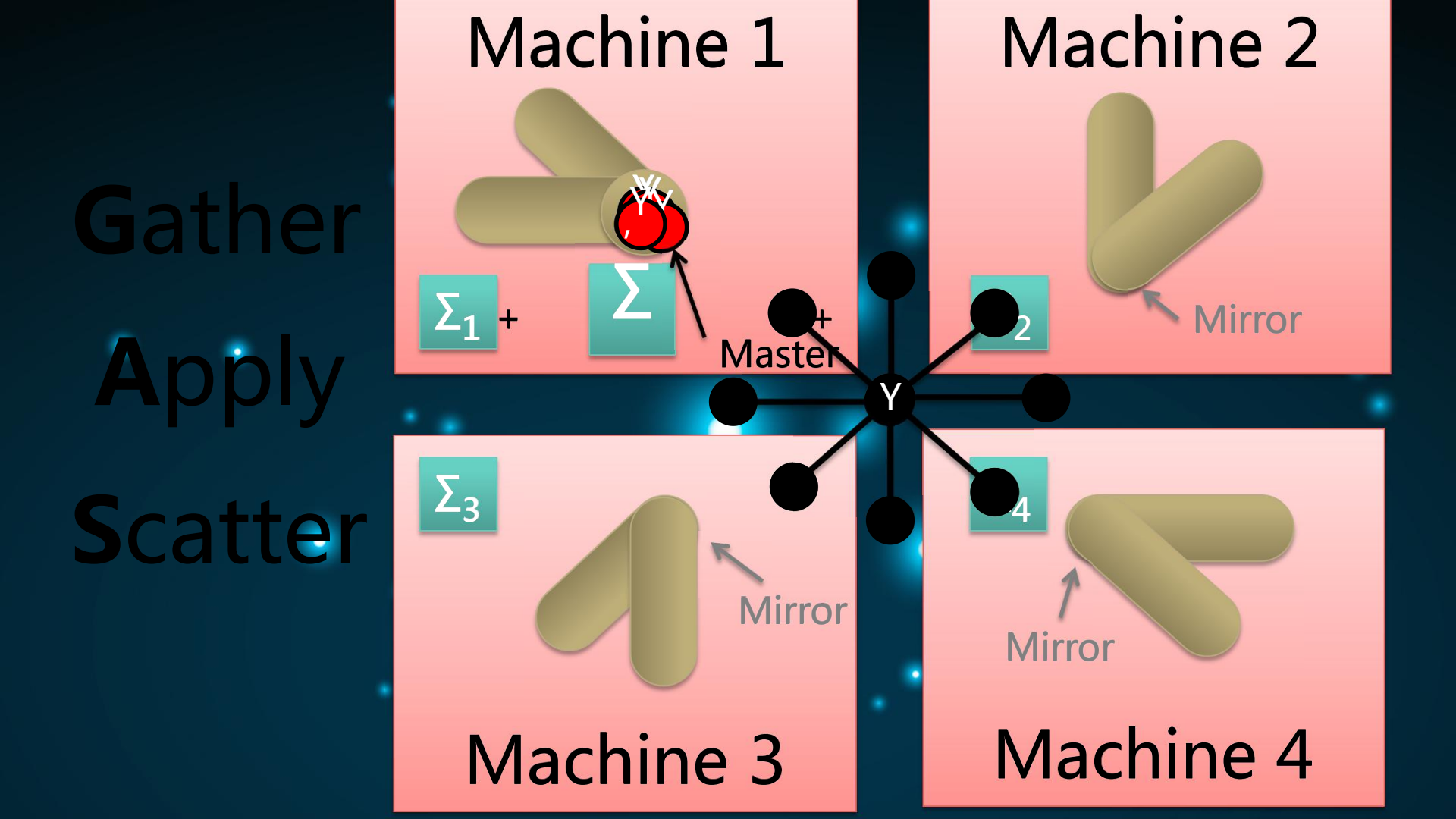
Balanced Vertex-Cut

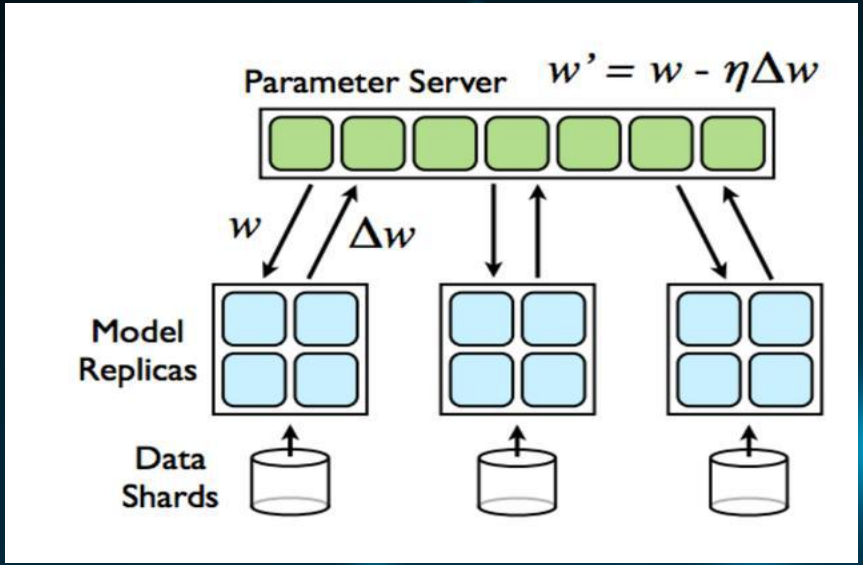
Y Spans 3 Machines

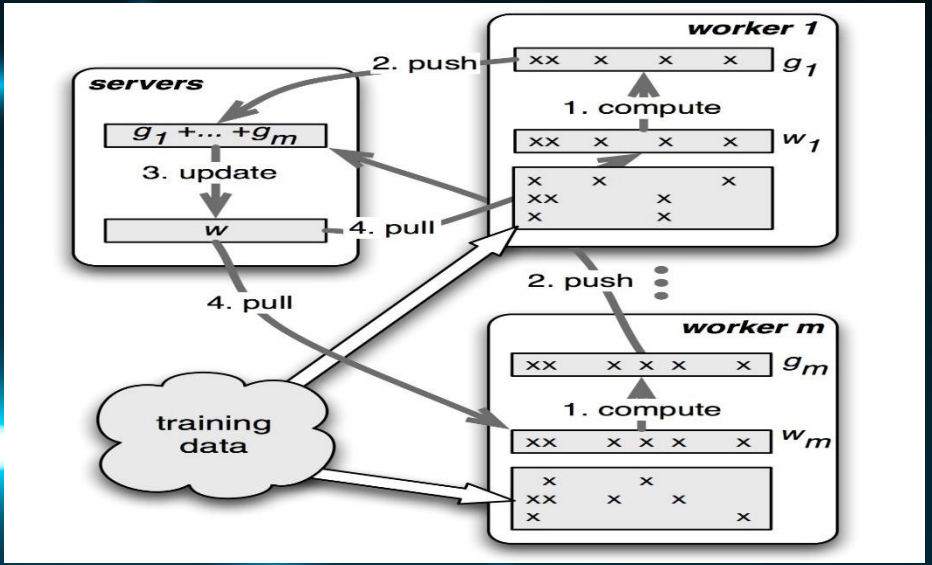
Z Spans 2 Machines

Not cut!

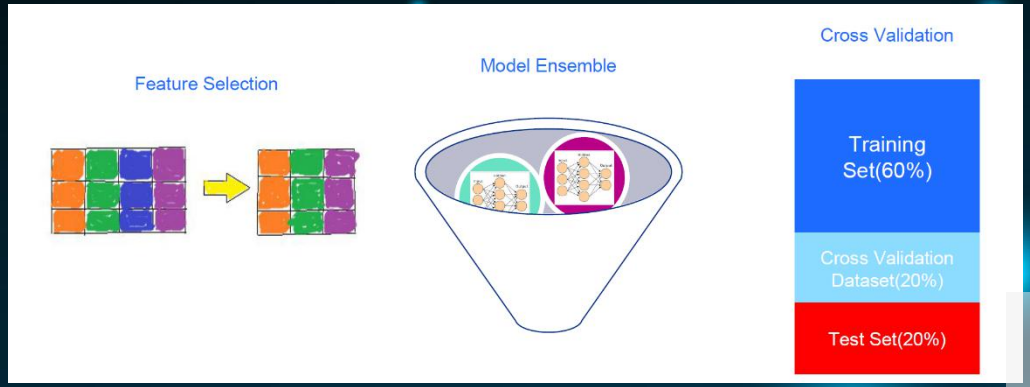



Gather
Apply
Scatter
Machine 1
Machine 2
Machine 3
Machine 4
Master
Mirror
Mirror
Mirror
Y
Y
Y
Y
Σ
Σ1 +
+
2
Σ3
4

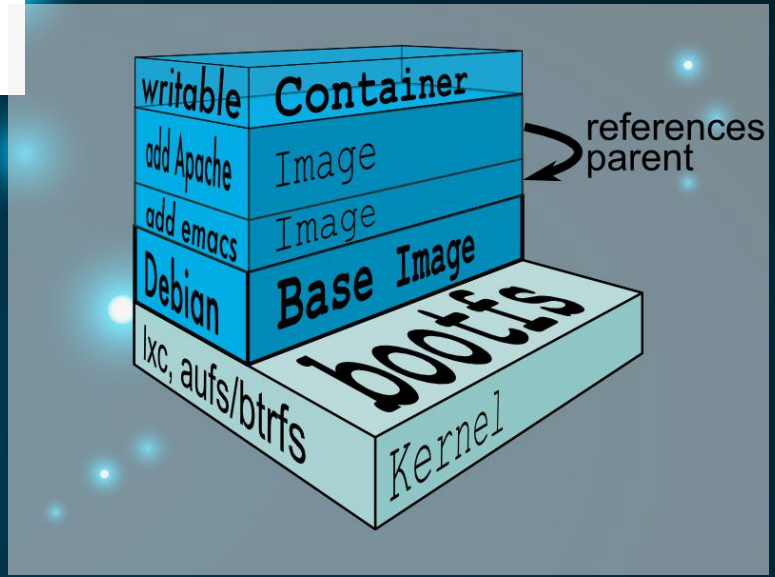




使用工具:模型训练**parameter server**



数据流:高可用,docker化



效果分析:
1:批量加入与去除
2:训练速度
3:模型维护


Thanks
微博账号：吴炜_机器学习数据挖掘