机器视觉

（嵌入式计算）

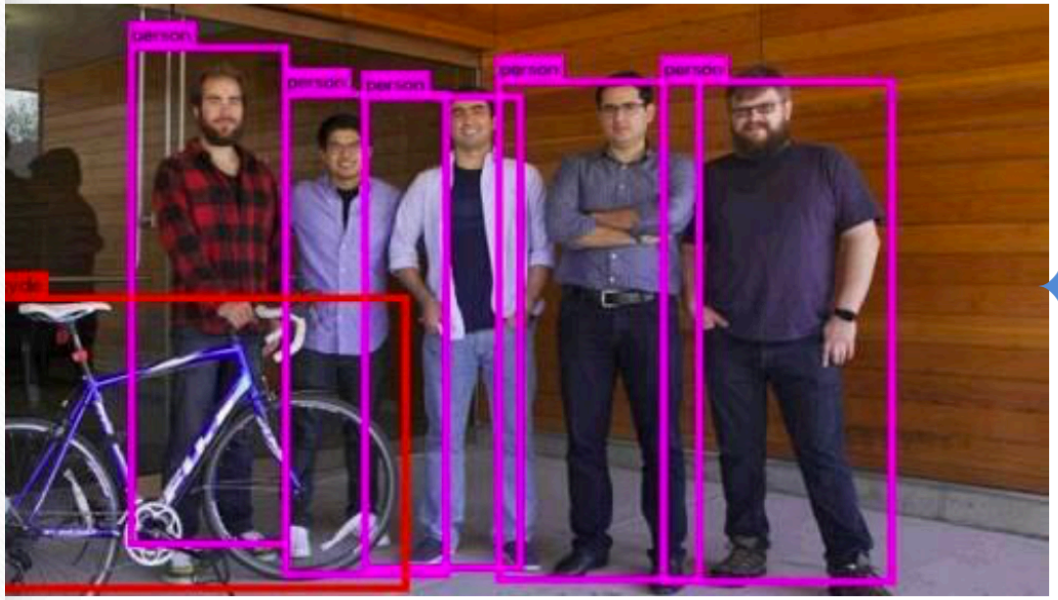
学生：屠晓涵

内容

- 做项目
  - 目标检测和定位、测移动物体的速度
- 读论文
  - 重点关注嵌入式平台模型的计算

研究团队了解

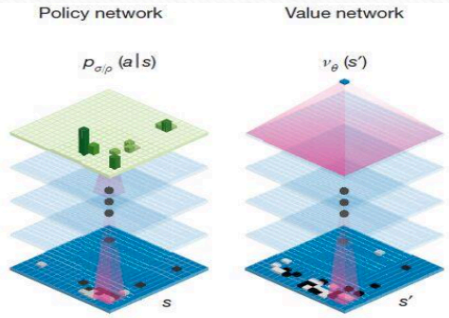
2017年2月2日，位于美国西雅图的 AI 创业公司 XNOR.AI 宣布获得来自Madrona Venture Group和艾伦人工智能研究所（Allen Institute for Artificial Intelligence）等机构的260万美元的种子融资。XNOR.AI 利用二值化神经网络等技术对深度学习网络进行压缩，致力于开发有效地在移动端或嵌入式设备上运行的深度学习算法。



XNOR.AI团队CEO Ali Farhadi是华盛顿大学计算机系教授，同时也是艾伦人工智能研究所的计算机视觉方向的负责人。是非常惊艳的实时物体检测框架YOLO的主要贡献者

XNOR.AI的CTO Mohammad Rastegari是艾伦人工智能研究所研究科学家，也在计算机视觉领域有接近十年的研究经历。2016 年 3 月，Mohammad Rastegari 等人在ECCV论文（XNOR-Net: ImageNet Classification Using Binary Convolutional Neural Networks）中首次提出了 XNOR-Net 的概念

深度学习>传统方法



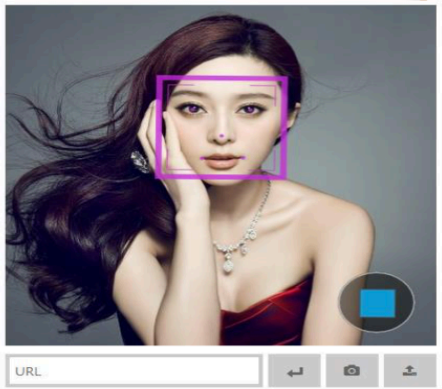
?

AlphaGo

深度学习的普遍应用

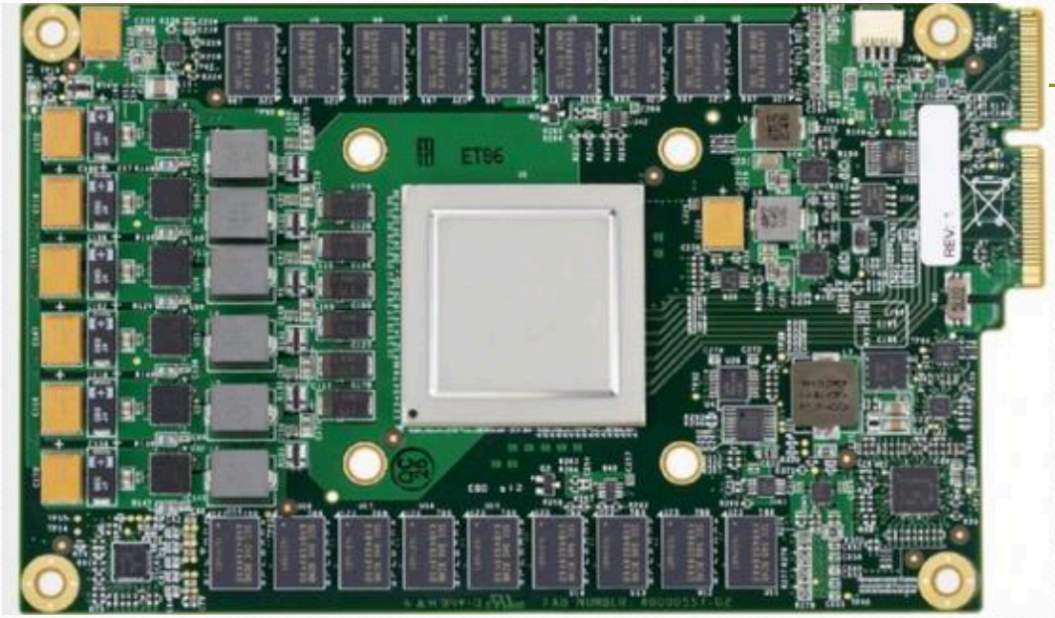
语音识别

手机上的人脸检测



军队机器人

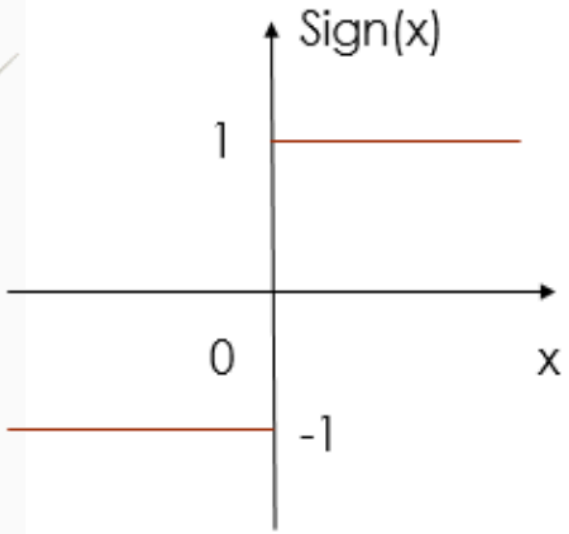
硬件了解TPU（Tensor processing uni）

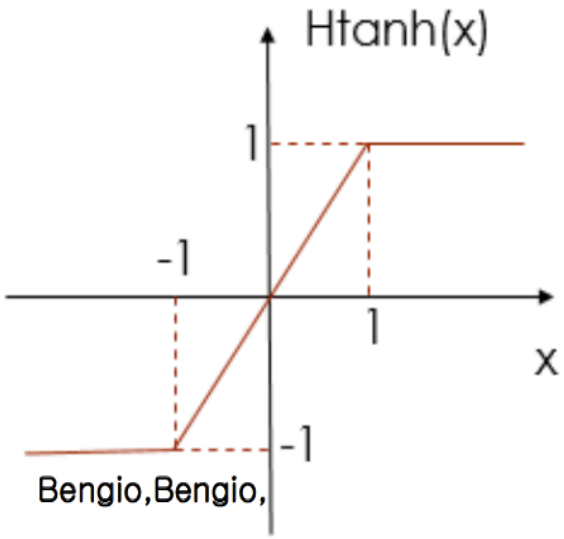


- 谷歌的专用机器学习芯片TPU 处理速度要比GPU和CPU快15 -30倍(和TPU对比的是英特尔Haswell CPU以及Nvidia Tesla K80 GPU)，而在能效上，TPU更是提升了30到80倍
- 这块芯片不大，主要用在前向传播的时候。

模型二值化-1

$$q = \mathrm{Sign}(r)$$





$$g_r = g_q 1_{|r|\leq 1}.$$

Htanh

- 方法：对weights和activations进行二值化。如图左所示，Binarization function 很简单，就是一个符号函数。但符号函数不好进行梯度的反向传播，因此就把它近似成了右边的Htanh(x)的函数，这样在[-1,1]区间内导数就等于1。

Paper: Binarized Neural Networks: Training Neural Networks with Weights and Activations Constrained to +1 or 1

模型二值化-2

Ensure: updated weights W^{t+1}, updated BatchNorm parameters θ^{t+1} and updated learning rate η^{t+1}.

{1. Computing the parameters' gradient:}

{1.1. Forward propagation:}

for $k = 1$ to L **do**

  $W_k^b \leftarrow \text{Binarize}(W_k)$

  $s_k \leftarrow a_{k-1}^b W_k^b$

  $a_k \leftarrow \text{BatchNorm}(s_k, \theta_k)$

  if $k < L$ **then**

    $a_k^b \leftarrow \text{Binarize}(a_k)$

  end if

end for

- 方法：权重Wk 经过二值化，然后与上层二值化后的激活值abk−1相乘，再进行BatchNormalization得到这一层的激活值ak，由于 BatchNorm的参数θk不是二值的，因此ak也不是二值的，论文再对它做二值化得到二值化后的激活值abk。

Paper: Binarized Neural Networks: Training Neural Networks with Weights and Activations Constrained to +1 or 1

模型二值化-3

{1.2. Backward propagation:}
{Please note that the gradients are not binary.}
Compute $g_{a_L} = \frac{\partial C}{\partial a_L}$ knowing a_L and a^*
for $k = L$ to 1 **do**
  if $k < L$ **then**
    $g_{a_k} \leftarrow g_{a_k^b} \circ 1_{|a_k|\leq 1}$
  end if
  $(g_{s_k}, g_{\theta_k}) \leftarrow \text{BackBatchNorm}(g_{a_k}, s_k, \theta_k)$
  $g_{a_{k-1}^b} \leftarrow g_{s_k} W_k^b$
  $g_{W_k^b} \leftarrow g_{s_k}^\top a_{k-1}^b$
end for

Algorithm 4 Running a BNN. L is the number of layers.

Require: a vector of 8-bit inputs a_0, the binary weights W^b and the BatchNorm parameters θ.
Ensure: the MLP output a_L.
{1. First layer:}
$a_1 \leftarrow 0$
for $n = 1$ to 8 **do**
  $a_1 \leftarrow a_1 + 2^{n-1} \times \text{XnorDotProduct}(a_0^n, W_1^b)$
end for
$a_1^b \leftarrow \text{Sign}(\text{BatchNorm}(a_1, \theta_1))$
{2. Remaining hidden layers:}
for $k = 2$ to $L-1$ **do**
  $a_k \leftarrow \text{XnorDotProduct}(a_{k-1}^b, W_k^b)$
  $a_k^b \leftarrow \text{Sign}(\text{BatchNorm}(a_k, \theta_k))$
end for
{3. Output layer:}
$a_L \leftarrow \text{XnorDotProduct}(a_{L-1}^b, W_L^b)$
$a_L \leftarrow \text{BatchNorm}(a_L, \theta_L)$

- 反向传播过程如左图，权重和激活值的更新并不是二值的，因为这样误差会很大。
- 输入层的特征是没有进行二值化的，输入的图像像素值分布在[0,255] 之间，用8比特来表示，这样就能将输入的实值像素值变成二值化的编码。整体BNN的流程如左图，将乘法运算都变成XNOR运算
- 缺点:简单粗暴，精度较低，可以 进行模型加速

Paper: Binarized Neural Networks: Training Neural Networks with Weights and Activations Constrained to +1 or 1

论文总结

剪枝：剪掉一些权值低的连接，可以减少10倍权值数目

量化：对权值进行编码，相当于对权值进行聚类，用聚类的中心代替这一个类别的权值

哈弗曼编码：聚类得到的中心长度不一致，进行哈弗曼编码可有效减少存储空间

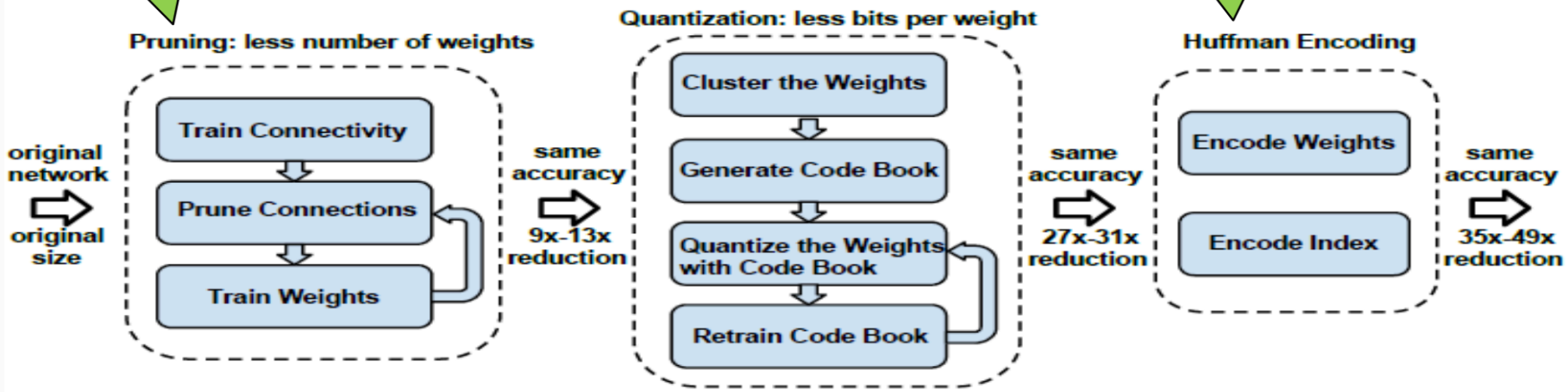


Figure 1: The three stage compression pipeline: pruning, quantization and Huffman coding. Pruning reduces the number of weights by 10×, while quantization further improves the compression rate: between 27× and 31×. Huffman coding gives more compression: between 35× and 49×. The compression rate already included the meta-data for sparse representation. The compression scheme doesn't incur any accuracy loss.

Paper：Deep Compression: Compressing Deep Neural Networks with Pruning, Trained Quantization and Huffman Coding. ICLR

论文总结

三个方面

剪枝：
权值小的，如接近0的。0.1，-0.23等全部剪掉，需要遍历整个权值空间
O(n)

量化：
权值1.8、2.1、2、2.2全部用2代替。这个处理起来没有一定的定论。看你自己怎么操作了。1.6是用1.5还是用2呢？

哈弗曼编码：
这个就是数据结构里的。

- **缺点：**
- 1、编码实现较复杂
- 2、是对训练好的模型进行的处理，而不是在训练过程中的处理，因而有精度损失。
- 3、还是陷入权值具体的值中，压缩比有限
- 4、只有压缩，没有加速。
- Paper：Deep Compression: Compressing Deep Neural Networks with Pruning, Trained Quantization and Huffman Coding. ICLR

疑问

- 为什么适合嵌入式平台的二值化权值网络准确率没什么变化，而一起二值化图像像素卷积特征精度降低10个百分点?
- 把一些权重(矩阵)、数据(矩阵)简单化表示，1.4也是1，2.3是2，这样简单化表示能丰富表达模型结构和样本空间吗?

参考文献

- Deep Compression: Compressing Deep Neural Networks with Pruning, Trained Quantization and Huffman Coding
- **XNOR-Net: ImageNet Classification Using Binary Convolutional Neural Networks**
- **Binarized Neural Networks: Training Neural Networks withWeights and Activations Constrained to +1 or 1**

谢谢观看指导！