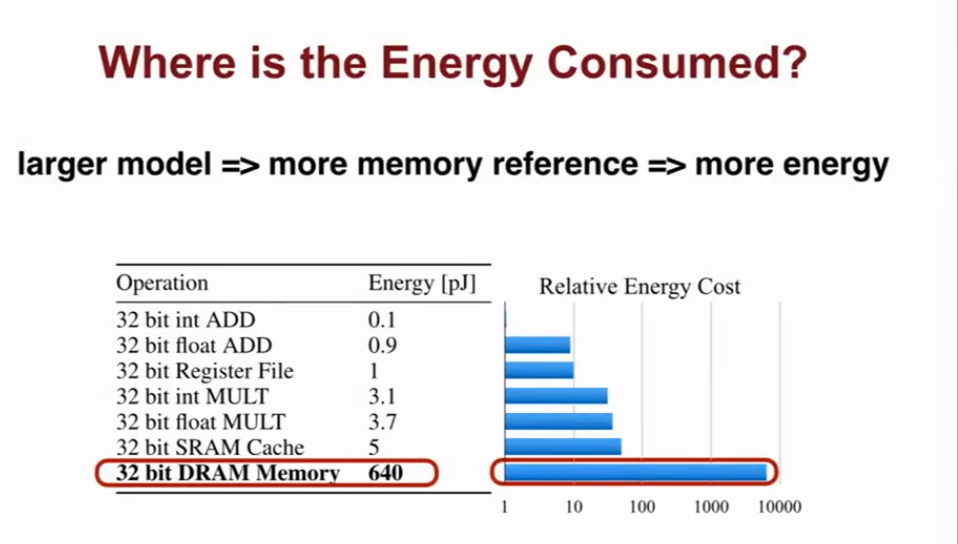
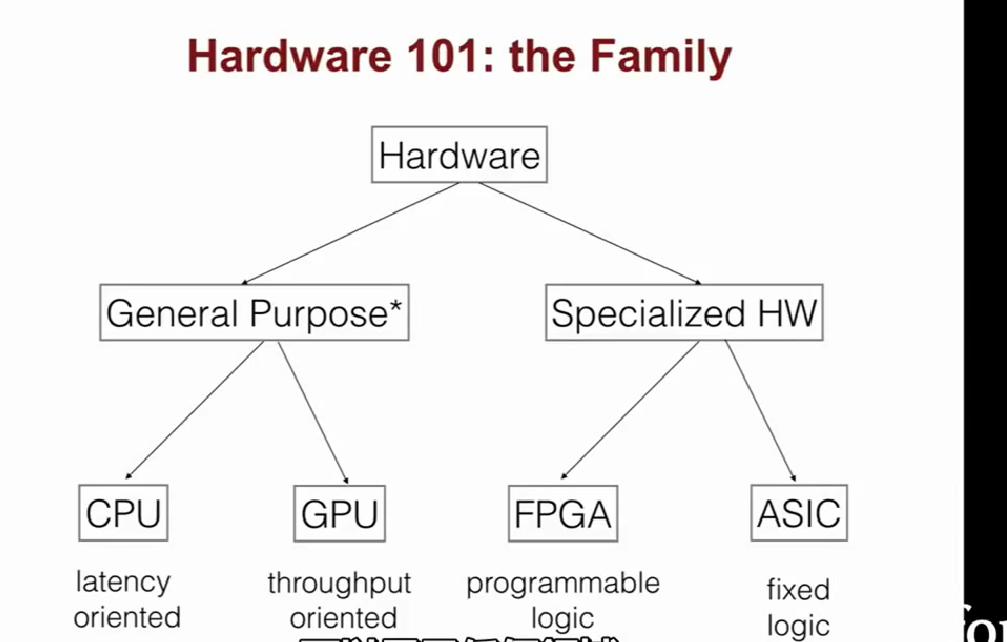
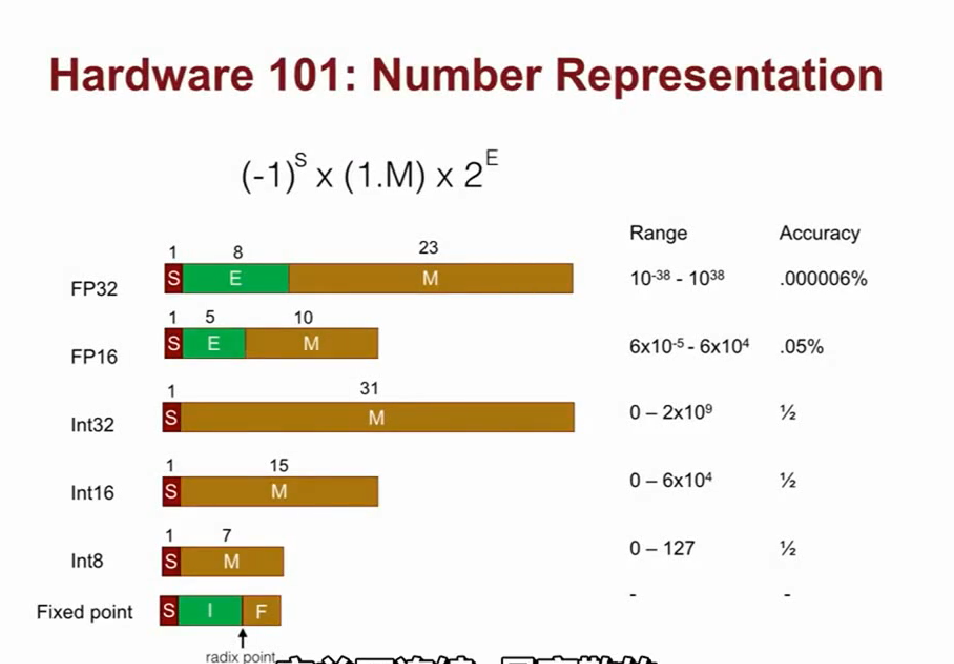
larger model => more memory reference => more energy



Improve the Efficiency of Deep Learning by Algorithm-Hardware Co-Design





32位的计算开销要比16位的计算开销大大约4倍。

Pruning Neural Networks

对神经网络进行剪枝，删除不需要的一些权重或者参数，删除一些连接。

通过反复训练，在减少90%的参数后仍然可以回到最初的准确度。

设置一个阈值，小于这个阈值的就可以删减掉。

Weights Sharing

对权重进行聚类，不要保留一切精度。

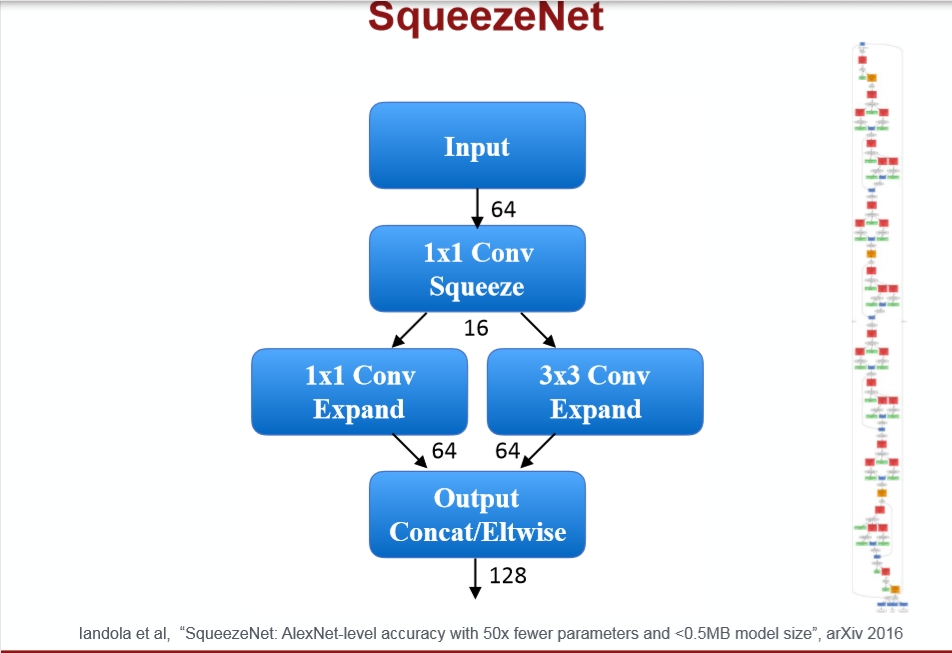
计算梯度也可也进行聚类，分为几类，通过改为加法计算结果。

Huffman Code

使用Huffman编码用更短的编码来表示常见的权重。

精简模型举例:

1.Squeeze Net



Quantization（量化)

Train with float

Quantizing the weight and activation:

      Gather the statistics for weights and activation

      Choose proper radix point position

Fine-tune in float format

Convert to fixed-point format

用定点数前向传播，用浮点数反向传播

Low Rank Approximate (低秩近似)

对全连接使用SVD分解

Binary / Ternary Net（二元 / 三元网络）

Winograd convolution

