机器视觉

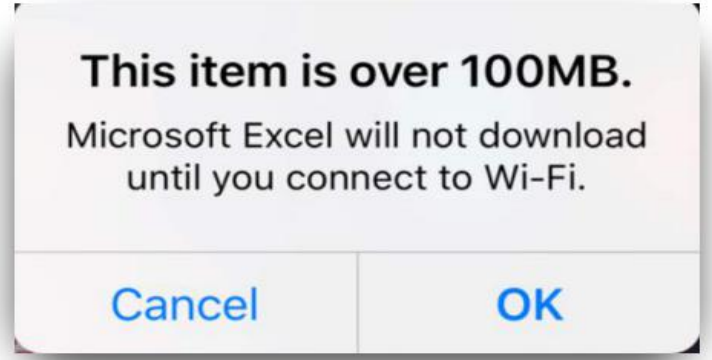
学生：屠晓涵

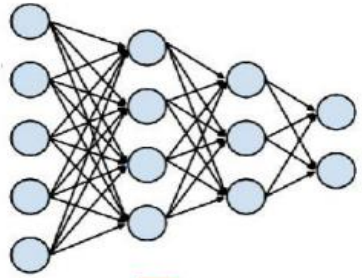
研究背景和研究问题

- 研究背景
  - 目标检测，避障。
  - 高铁接触网几何参数检测。
  - 深度学习 > 传统方法
  - 模型直接运行在嵌入式设备上，达到精度高、反应快、保护隐私等独特优势
  - 嵌入式平台挑战：模型大小、速度、能耗
- 研究问题：模型压缩

嵌入式平台挑战：模型大

- 如何在嵌入式设备上布置大模型





嵌入式平台挑战：速度

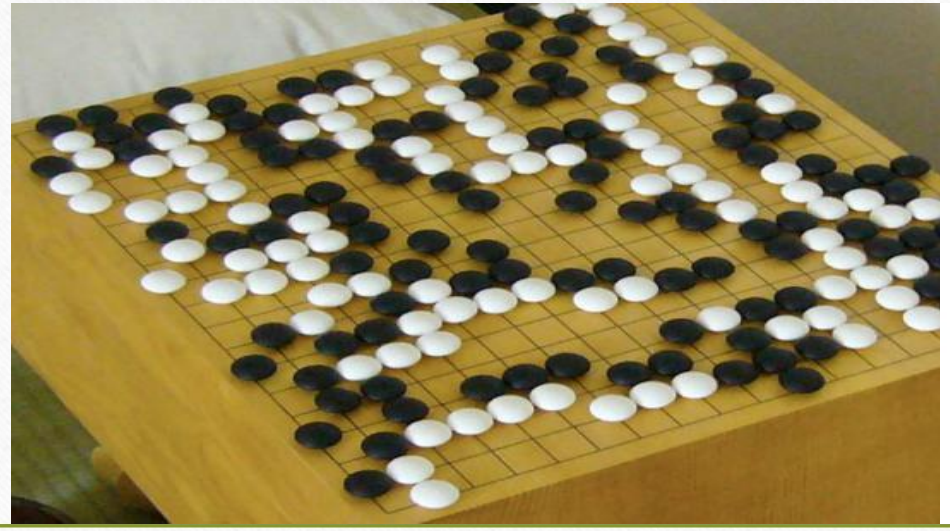
模型层数	误差率	训练时间
• ResNet18:	10.76%	2.5 天
• ResNet50:	7.02%	5 天
• ResNet101:	6.21%	1 周
• ResNet152:	6.16%	1.5 周

嵌入式平台现状：能耗

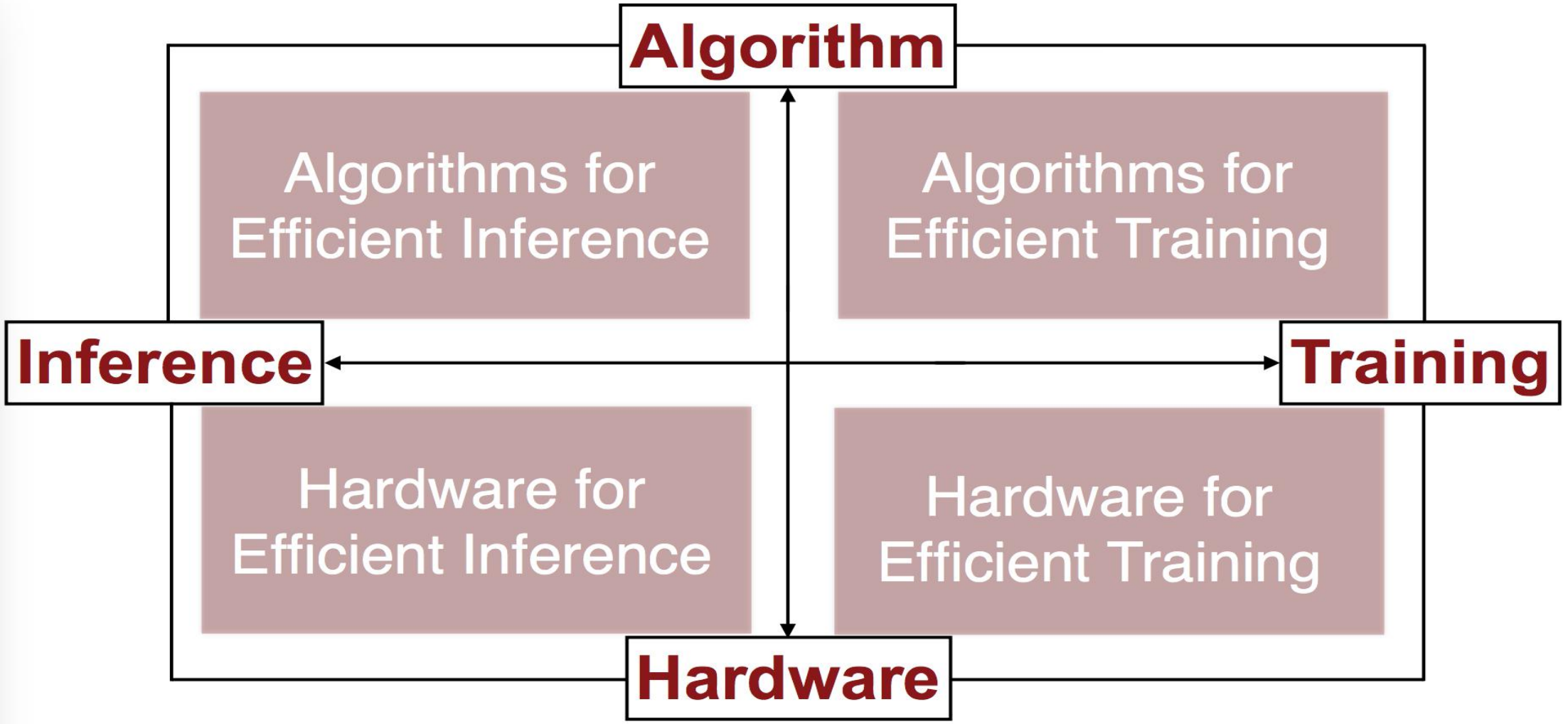
- AlphaGo：1920块 CPU and 280块 GPU，$3000 电费/每次游戏

- 嵌入式设备：电量耗尽　　数据中心：增加所有成本

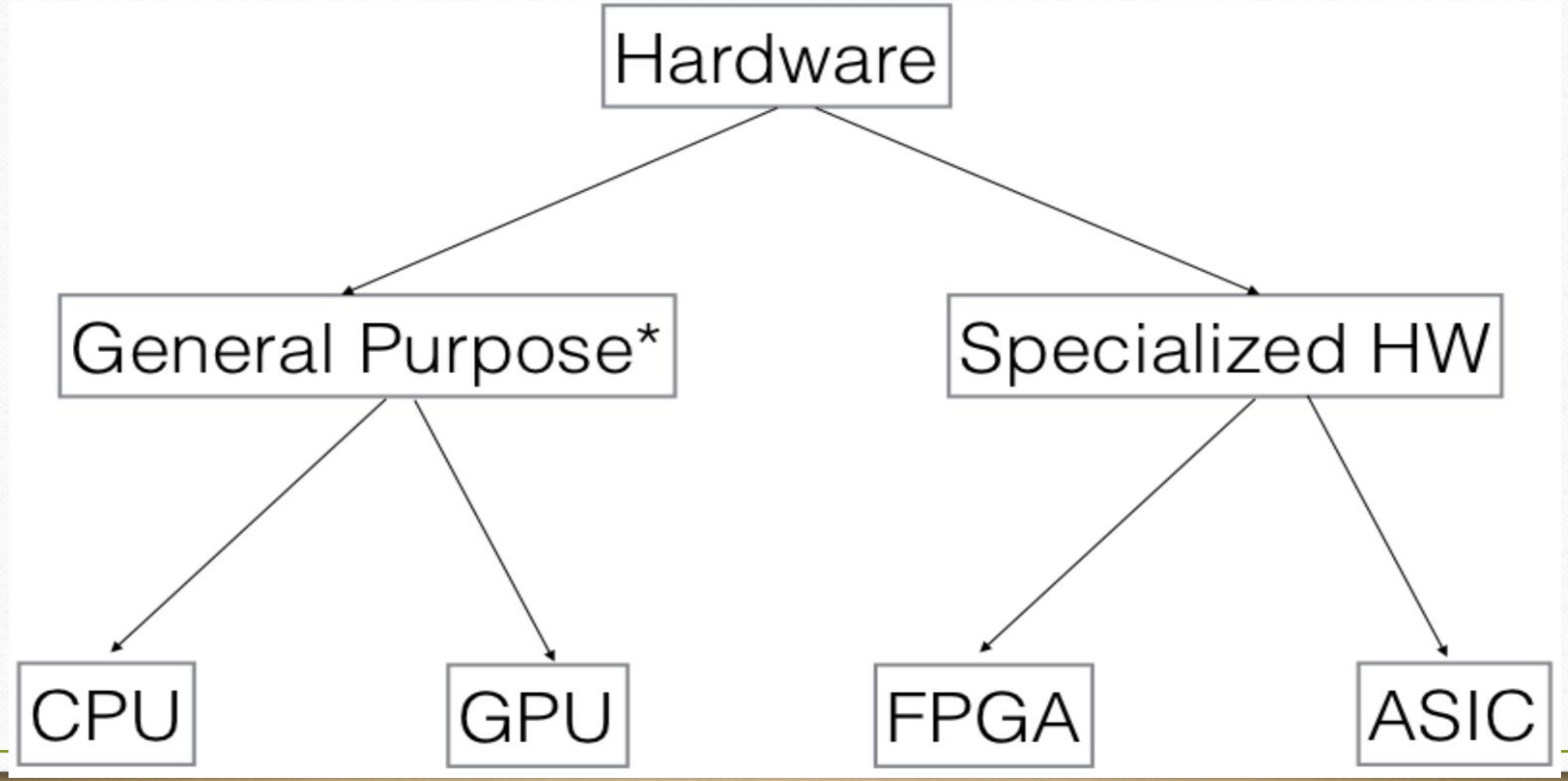
- 更大的模型——更多的内存占用——更加耗能



解决方案


Algorithm
Algorithms for Efficient Inference
Algorithms for Efficient Training
Inference
Training
Hardware for Efficient Inference
Hardware for Efficient Training
Hardware

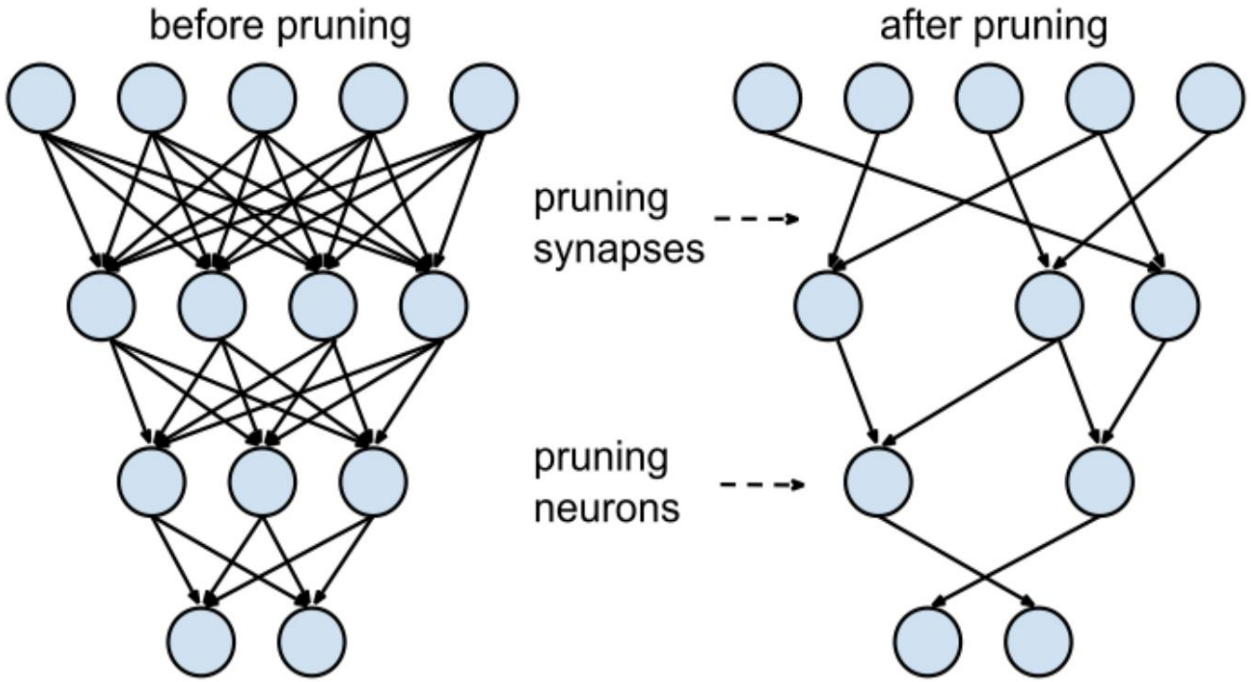
解决方案：硬件加速

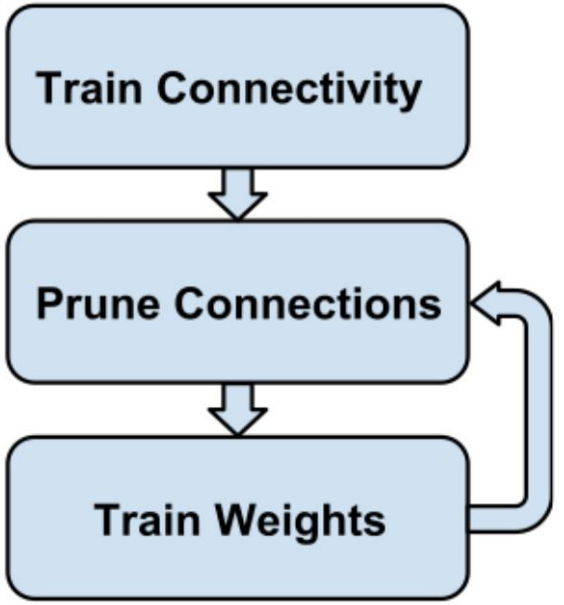


解决方案：算法改进

- 剪枝
- 权重共享
- 量化
- 低跌近似
- 二值化

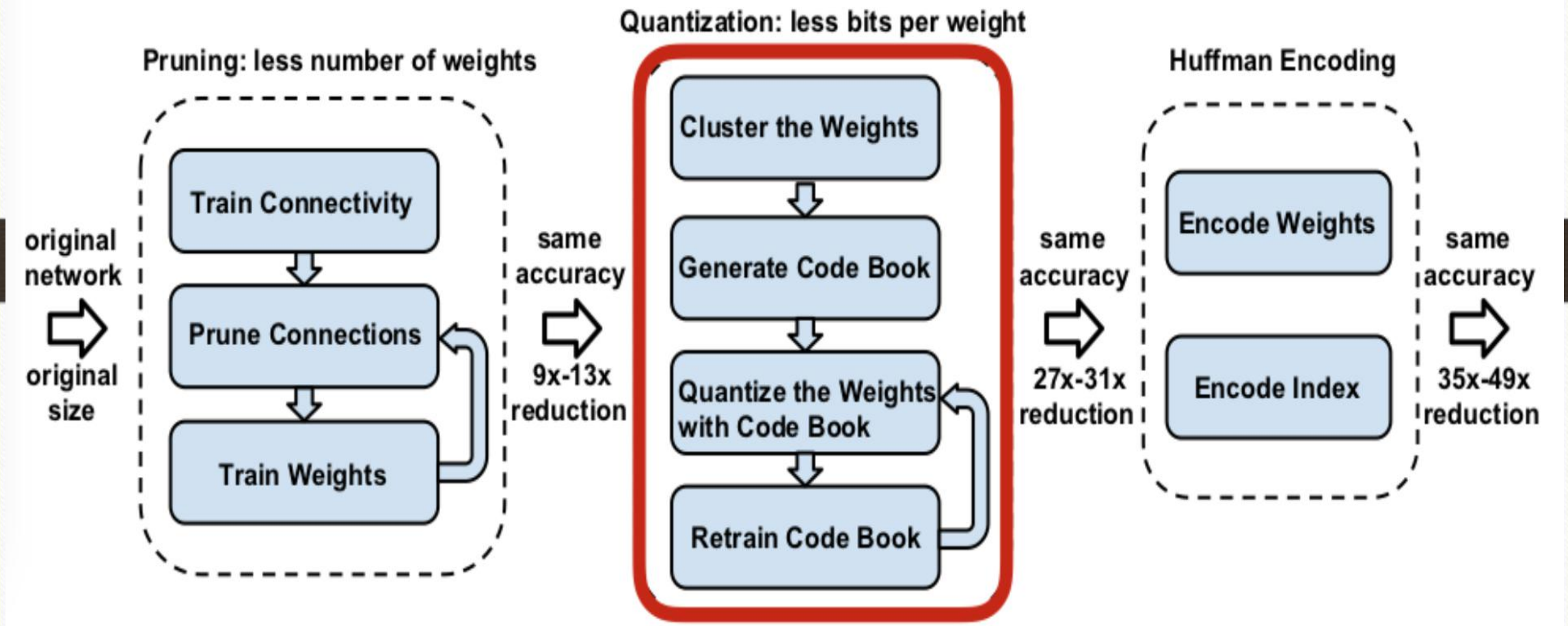
算法改进：剪枝





[1] LeCun et al. Optimal Brain Damage NIPS'90
[2] Hassibi, et al. Second order derivatives for network pruning: Optimal brain surgeon. NIPS'93
[3] Han et al. Learning both Weights and Connections for Efficient Neural Networks, NIPS'15

算法改进：权重共享


Pruning: less number of weights
Train Connectivity
Prune Connections
Train Weights
original network
original size
same accuracy
9x-13x reduction
Quantization: less bits per weight
Cluster the Weights
Generate Code Book
Quantize the Weights with Code Book
Retrain Code Book
same accuracy
27x-31x reduction
Huffman Encoding
Encode Weights
Encode Index
same accuracy
35x-49x reduction

深度压缩效果

Network	Original Size	Compressed Size	Compression Ratio	Original Accuracy	Compressed Accuracy
LeNet-300	1070KB	27KB	**40x**	98.36%	98.42%
LeNet-5	1720KB	44KB	**39x**	99.20%	99.26%
AlexNet	240MB	6.9MB	**35x**	80.27%	80.30%
VGGNet	550MB	11.3MB	**49x**	88.68%	89.09%
GoogleNet	28MB	2.8MB	**10x**	88.90%	88.92%
ResNet-18	44.6MB	4.0MB	**11x**	89.24%	89.28%

现阶段个人想法

- 推理时间
  - 单图像运算推理时间
  - 推理时间随batch size大小的变化情况。考察多图像并行处理的时间消耗。
- 内存消耗
- 模型性能与吞吐量
- 运算量

本学期目标

- 研究适合于嵌入式平台的模型
- 嵌入式系统机器视觉目标检测综述
- 熟练掌握实验平台—tensorflow
- 复现经典论文实验结果
- 实验验证自己的想法，应用于机器人，替换原有的传统方法。

参考文献

- Han S, Pool J, Tran J, et al. Learning both weights and connections for efficient neural network[C]//Advances in Neural Information Processing Systems. 2015: 1135-1143.
- Han S, Liu X, Mao H, et al. EIE: efficient inference engine on compressed deep neural network[C]//Proceedings of the 43rd International Symposium on Computer Architecture. IEEE Press, 2016: 243-254.
- Han S, Mao H, Dally W J. Deep Compression: Compressing Deep Neural Networks with Pruning, Trained Quantization and Huffman Coding. ICLR, 2016, 56(4):3--7.
- Han S, Pool J, Narang S, et al. Dsd: Dense-sparse-dense training for deep neural networks. ICLR, 2017.

谢谢观看指导！