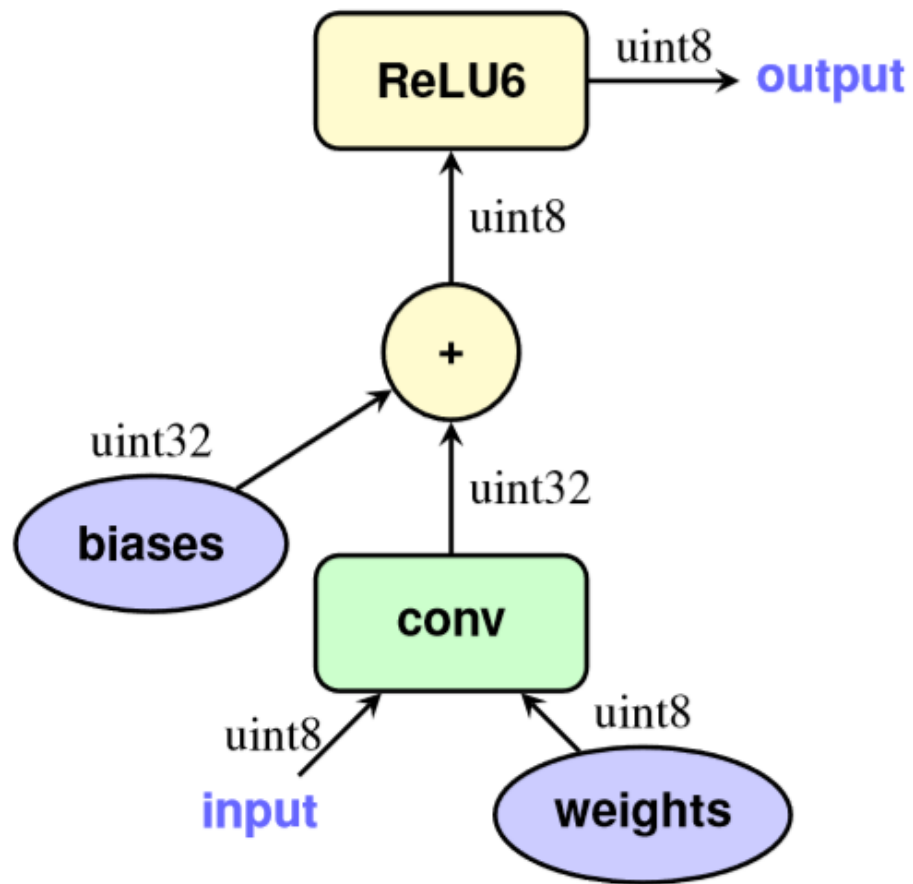


量化神经网络

赵恒锐

2018.12.18

Introduction



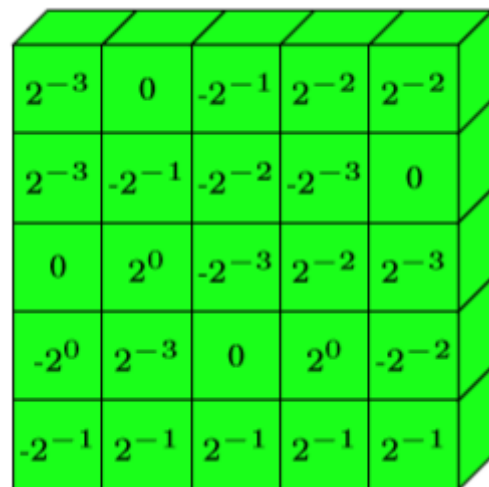
只使用整数运算进行推理，节省时间和空间

Reference

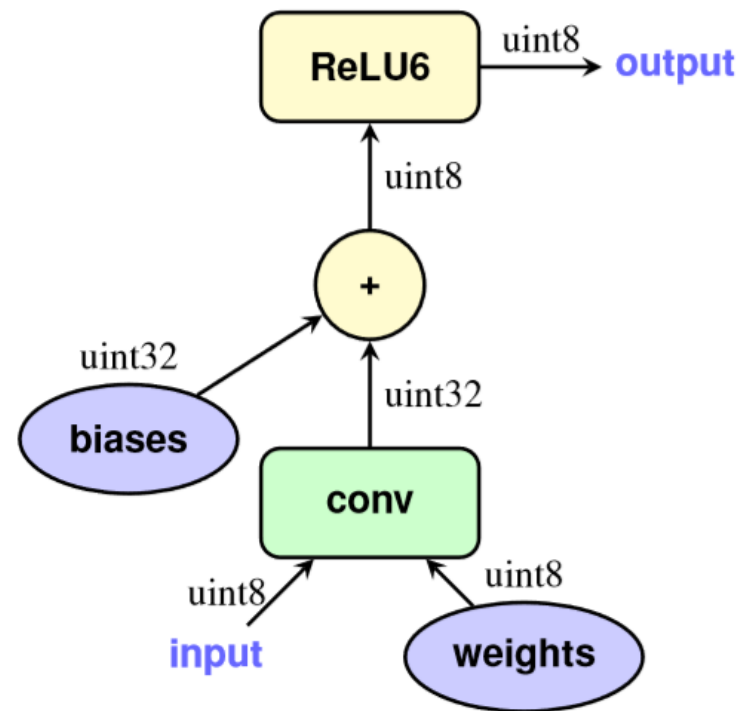
- Quantization and training of neural networks for efficient integer-arithmetic-only inference, Google, CVPR2018
- Discovering Low-Precision Networks Close to Full-Precision Networks for Efficient Embedded Inference, IBM, arxiv
- Incremental Network Quantization Towards Lossless CNNs with Low-precision Weights, Intel, ICLR2017

量化目标

- 兼顾硬件与应用
- 主要优化推理阶段

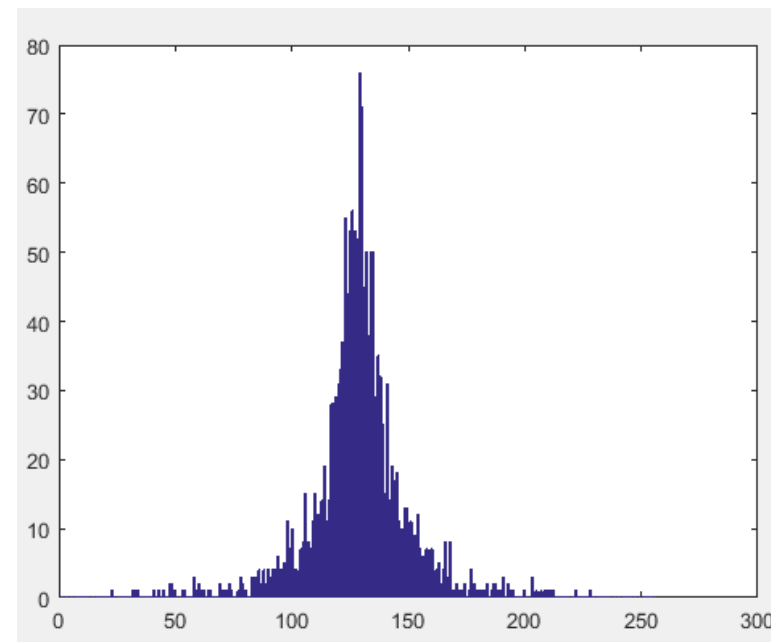
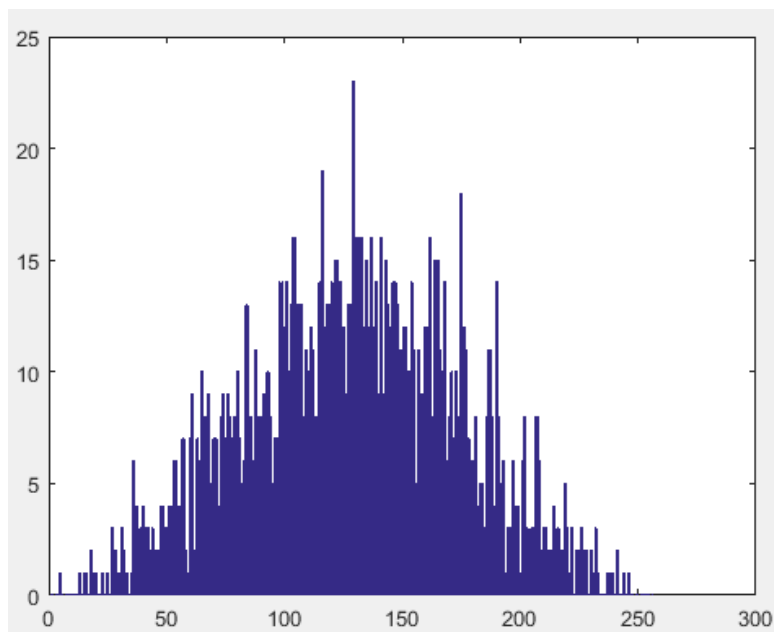


2^{-3}	0	-2^{-1}	2^{-2}	2^{-2}
2^{-3}	-2^{-1}	-2^{-2}	-2^{-3}	0
0	2^0	-2^{-3}	2^{-2}	2^{-3}
-2^0	2^{-3}	0	2^0	-2^{-2}
-2^{-1}	2^{-1}	2^{-1}	2^{-1}	2^{-1}



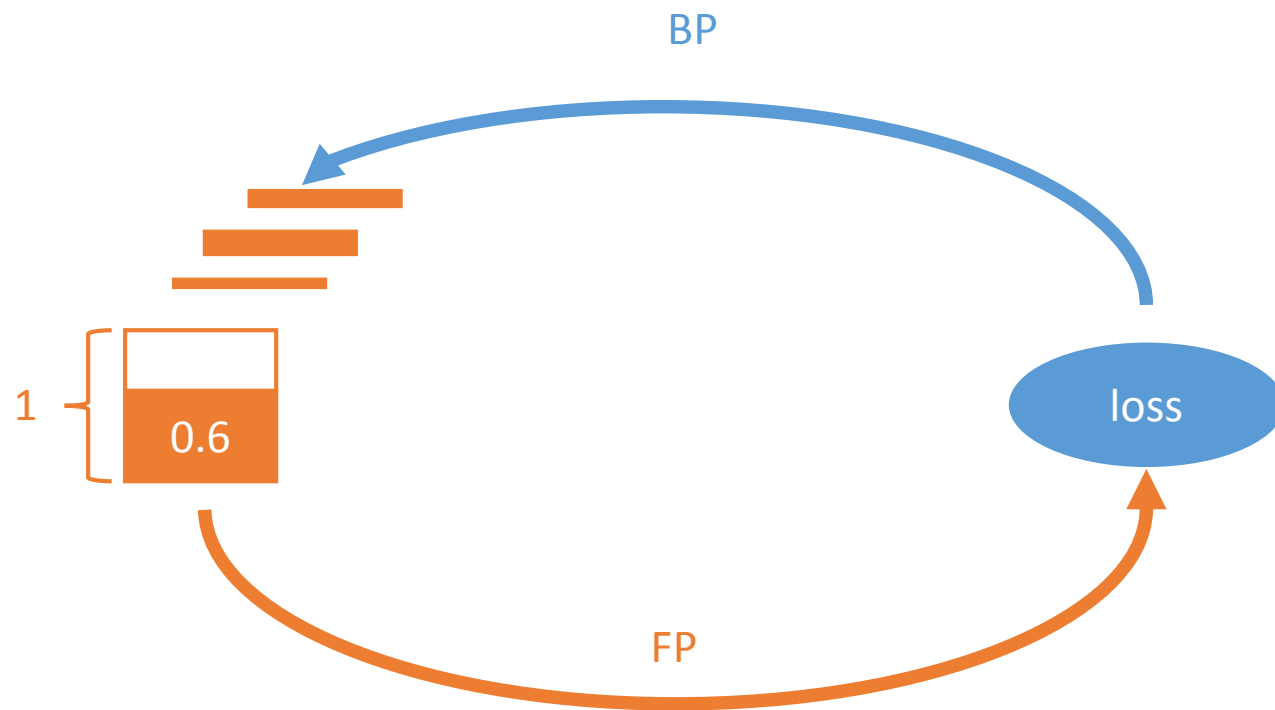
模型量化

- 线性 or 非线性
- 区间的选择



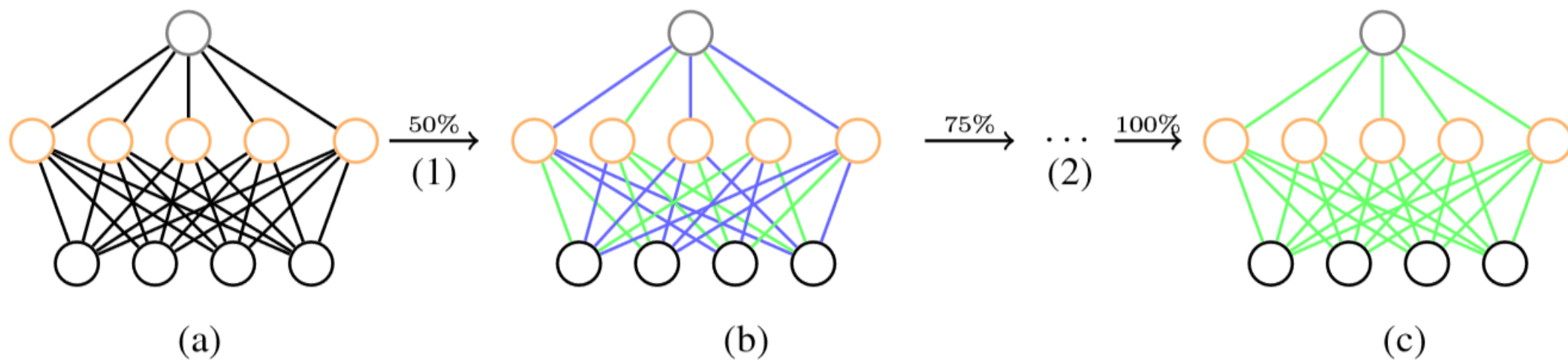
Fine-tune

- 累积更新



Fine-tune

- 全局 or 局部



INQ

Fine-tune数据

- $k \leq \frac{(\sigma^2 + L * \|x_0 - x^*\|_2^2)^2}{\epsilon^2}$
- 对于全局Fine-tune只需要1 epoch
- 对于INQ需要8 epoch左右

性能

大多数论文宣称：
同等模型量化后性能更优

Network	Method	Precision (w,a)	Accuracy (% top-1)	Accuracy (% top-5)
ResNet-18	baseline	32,32	69.76	89.08
ResNet-18	Apprentice	4,8	70.40	-
ResNet-18	FAQ (This paper)	8,8	70.02	89.32
ResNet-18	FAQ (This paper)	4,4	69.82	89.10
ResNet-18	Joint Training	4,4	69.3	-
ResNet-18	UNIQ	4,8	67.02	-
ResNet-18	Distillation	4,32	64.20	-
ResNet-34	baseline	32,32	73.30	91.42
ResNet-34	FAQ (This paper)	8,8	73.71	91.63
ResNet-34	FAQ (This paper)	4,4	73.31	91.32
ResNet-34	UNIQ	4,32	73.1	-
ResNet-34	Apprentice	4,8	73.1	-
ResNet-34	UNIQ	4,8	71.09	-
ResNet-50	baseline	32,32	76.15	92.87
ResNet-50	FAQ (This paper)	8,8	76.52	93.09
ResNet-50	FAQ (This paper)	4,4	76.27	92.89
ResNet-50	IOA	8,8	74.9	-
ResNet-50	Apprentice	4,8	74.7	-
ResNet-50	UNIQ	4,8	73.37	-
ResNet-152	baseline	32,32	78.31	94.06
ResNet-152	FAQ (This paper)	8,8	78.54	94.07
Inception-v3	baseline	32,32	77.45	93.56
Inception-v3	FAQ (This paper)	8,8	77.60	93.59
Inception-v3	IOA	8,8	74.2	92.2
Densenet-161	baseline	32,32	77.65	93.80
Densenet-161	FAQ (This paper)	8,8	77.84	93.91
VGG-16bn	baseline	32,32	73.36	91.50
VGG-16bn	FAQ (This paper)	8,8	73.66	91.56

实际速度

5*5卷积, cudnn7.0

size	<i>Float32</i>	<i>int8_{dynamic}</i>	<i>int8_{static}</i>
416 * 240	11.761	18.506	7.788
832 * 480	22.683	34.154	11.867
1280 * 720	42.634	55.122	22.329
1920 * 1080	86.831	107.609	42.388
2560 * 1600	155.432	182.347	79.333

谢谢！

- Q&A