Deep Learning with Low Precision by Half-wave Gaussian Quantization

2018-10-20

Overview

- 作者：Zhaowei Cai UC San Diego，Xiaodong He Microsoft Research Redmond，Jian Sun Megvii Inc，Nuno Vasconcelos UC San Diego
- CVPR 2017 引用35
- 加速压缩 32~
- 做了1 bit weight和2-bit的activation
- https://github.com/zhaoweicai/hwgq
- caffe
- 主要是改变了activation的二值化，采用的是ReLU高斯量化，weight的二值化和xnor-net一样， 提出的HGWQ以及对应的backprop近似解决gradient mismatch的问题，主要是在alexnet, googleNet, resnet, vgg的imagenet实验，还有cifar10的实验，针对xnor还是有一定的提升，但是其实差距还是很大，主要差别是在大数据集上做了这些个实验，xnor只做了alexnet。

Weight Quantization

- Binary Networks $$z = g(\mathbf{w}^T \mathbf{x}), \quad (1)$$
- Weight Binarization处理类似于XNOR

$$\mathbf{I} * \mathbf{W} \approx \alpha(\mathbf{I} \oplus \mathbf{B}), \quad (2)$$

$$\mathbf{B} \in \{+1, -1\}^{c \times w \times h} \text{ and a scaling factor } \alpha \in \mathbb{R}^+$$

$$\mathbf{W} \approx \alpha \mathbf{B}.$$

$$\mathbf{B}^* = sign(\mathbf{W}) \text{ and } \alpha^* = \frac{1}{cwh} \|\mathbf{W}\|_1.$$

Activation Binarization

- 原来都是直接取sign操作

$$z = sign(x) = \begin{cases} +1, & \text{if } x \geq 0, \\ -1, & \text{otherwise} \end{cases} \quad (3)$$

- 所以反向传播用的是hard tanh近似，但是本文使用hwgq量化，现在用ReLU近似

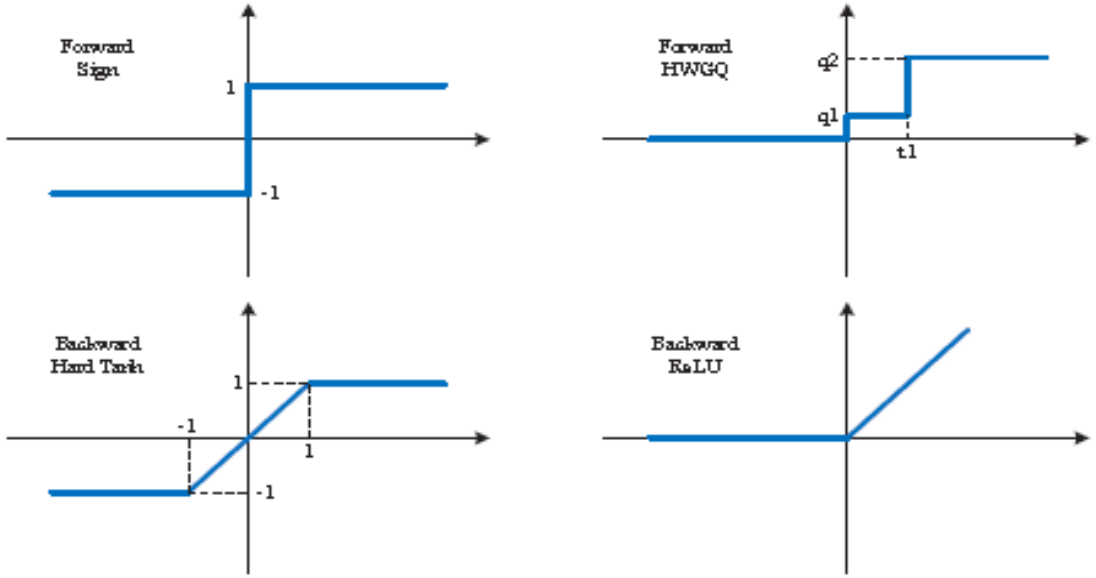


Figure 1. Forward and backward functions for binary $sign$ (left) and half-wave Gaussian quantization (right) activations.

Half-wave Gaussian Quantization

- 前向过程中是类似ReLU的一个分段函数：

$$Q(x) = q_i, \quad if \quad x \in (t_i, t_{i+1}], \tag{7}$$

that maps all values of x within quantization interval $(t_i, t_{i+1}]$ into a quantization level $q_i \in \mathbb{R}$, for $i = 1, \cdots, m$.

$$q_{i+1} - q_i = \Delta, \quad \forall i, \tag{8}$$

Δ is a constant quantization step.

q_i act as the reconstruction values for x,

- 求解目标：

$$Q^*(x) = \arg\min_Q E_w[(Q(x) - x)^2] \tag{9}$$

$$= \arg\min_Q \int p(x)(Q(x) - x)^2 dx$$

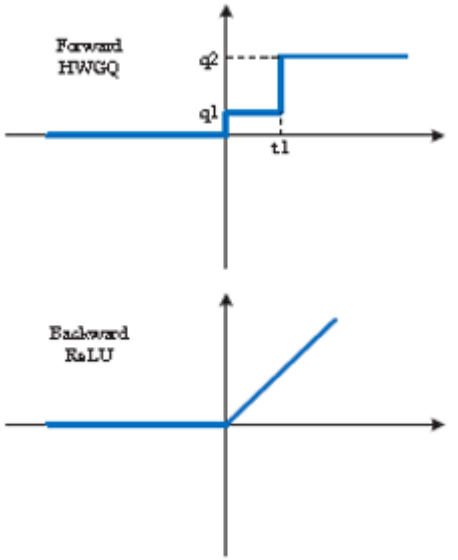
$p(x)$ is the probability density function of x.

- 近似于高斯函数处理，超参数用Lloyd算法求解

Half-wave Gaussian Quantization

- 反向传播用ReLU或者其变型近似：出发点在于希望一些很大的点的梯度尽量小的影响到网络
- Vanilla ReLU

$$\widetilde{Q}'(x) = \begin{cases} 1, & \text{if } x > 0, \\ 0, & \text{otherwise} \end{cases} \tag{11}$$



- Clipped ReLU

$$\widetilde{Q}_c(x) = \begin{cases} q_m, & x > q_m, \\ x, & x \in (0, q_m], \\ 0, & \text{otherwise.} \end{cases} \tag{12}$$

- Log-tailed ReLU

$$\widetilde{Q}_l(x) = \begin{cases} q_m + \log(x - \tau), & x > q_m, \\ x, & x \in (0, q_m], \\ 0, & x \leq 0, \end{cases} \tag{13}$$

$$\widetilde{Q}_l'(x) = \begin{cases} 1/(x - \tau), & x > q_m, \\ 1, & x \in (0, q_m], \\ 0, & x \leq 0. \end{cases} \tag{14}$$

Results

- **Full-precision Activation Comparison**

Table 1. Full-precision Activation Comparison for AlexNet.

	Full	FW+$sign$	FW+$\tilde{Q}$	BW+$sign$	BW+$\tilde{Q}$
Top-1	55.7	46.7	55.7	43.9	53.9
Top-5	79.3	71.0	79.3	68.3	77.3

- 本文的Q比sign来说是个更好的选择

Results

- Low-bit Activation Quantization Results

Table 2. Low-bit Activation Comparison.

Model		Full	BW	FW+Q	BW+$sign$	BW+Q
AlexNet	Top-1	55.7	52.4	49.5	39.5	46.8
	Top-5	79.3	75.9	73.7	63.6	71.0
ResNet-18	Top-1	66.3	61.3	37.5	42.1	33.0
	Top-5	87.5	83.6	61.9	67.1	56.9
VGG-Variant	Top-1	68.6	65.5	48.3	50.1	44.1
	Top-5	88.9	86.5	72.3	74.3	68.7

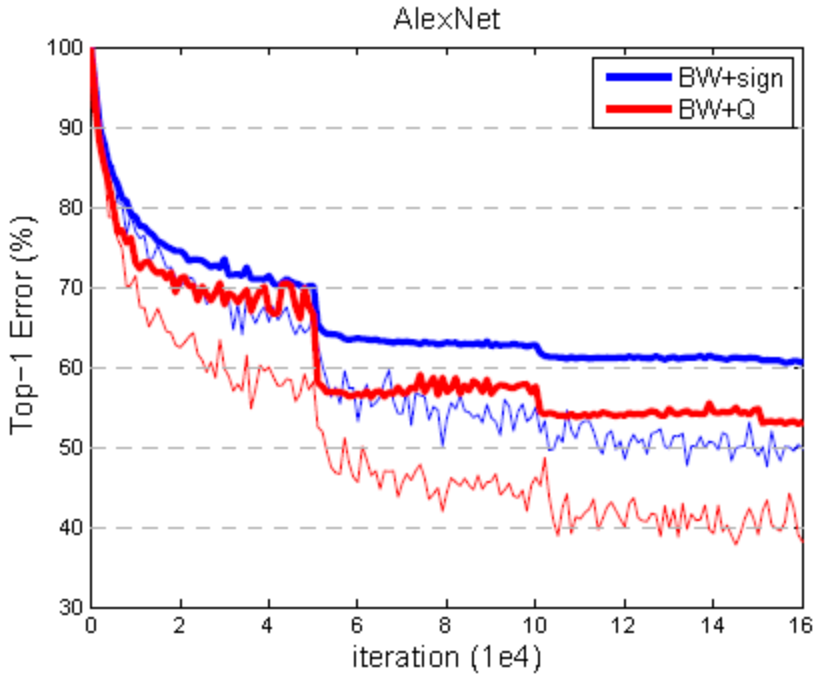


Figure 4. The error curves of training (thin) and test (thick) for $sign(x)$ and $Q(x)$ (HWGQ) activation functions.

Results

- Backward Approximations Comparison

Table 3. Backward Approximations Comparison.

Model		BW	no-opt	vanilla	clipped	log-tailed
AlexNet	Top-1	52.4	30.0	46.8	48.6	49.0
	Top-5	75.9	53.6	71.0	72.8	73.1
ResNet-18	Top-1	61.3	34.2	33.0	54.5	53.5
	Top-5	83.6	59.6	56.9	78.5	77.7
VGG-Variant	Top-1	65.5	42.8	44.1	60.9	60.6
	Top-5	86.5	68.3	68.7	83.2	82.9

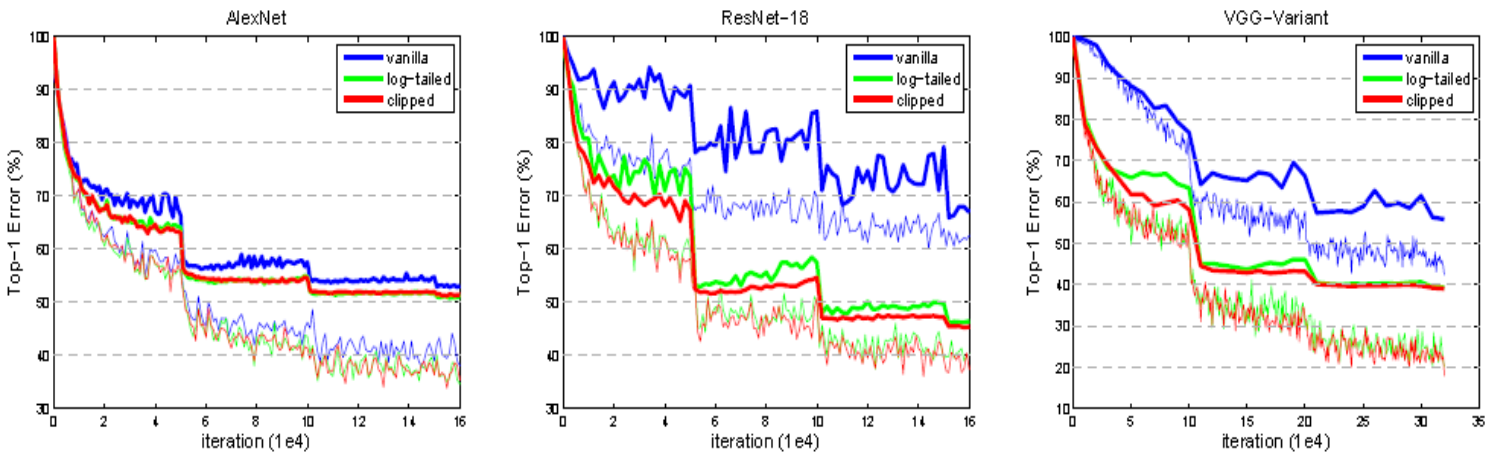


Figure 5. The error curves of training (thin) and test (thick) for alternative backward approximations.

Results

- **Bit-width Impact**

Table 4. Bit-width Comparison of Activation Quantization.

quantization type		non-uniform				uniform		none
# levels		2	3	7	15	3*	7*	BW
AlexNet	Top-1	48.6	50.6	52.4	52.6	50.5	51.9	52.4
	Top-5	72.8	74.3	75.8	76.2	74.6	75.7	75.9
ResNet-18	Top-1	54.5	57.6	60.3	60.8	56.1	59.6	61.3
	Top-5	78.5	81.0	82.8	83.4	79.7	82.4	83.6

Results

- **Comparison with the state-of-the-art**

Table 5. The results of various popular networks.

Model		Reference	Full	HWGQ
AlexNet	Top-1	57.1	58.5	52.7
	Top-5	80.2	81.5	76.3
ResNet-18	Top-1	69.6	67.3	59.6
	Top-5	89.2	87.9	82.2
ResNet-34	Top-1	73.3	69.4	64.3
	Top-5	91.3	89.1	85.7
ResNet-50	Top-1	76.0	71.5	64.6
	Top-5	93.0	90.5	85.9
VGG-Variant	Top-1	-	69.8	64.1
	Top-5	-	89.3	85.6
GoogLeNet	Top-1	68.7	71.4	63.0
	Top-5	88.9	90.5	84.9

Table 6. Comparison with the state-of-the-art low-precision methods. Top-1 gap to the corresponding full-precision networks is also reported.

Model	AlexNet			ResNet-18	
	XNOR	DOREFA	HWGQ	XNOR	HWGQ
Top-1	44.2	47.7	52.7	51.2	59.6
Top-5	69.2	-	76.3	73.2	82.2
Top-1 gap	-12.4	-8.2	-5.8	-18.1	-7.7

Table 7. The results on CIFAR-10. The bit width before and after "+" is for weights and activations respectively.

precision	Method	error (%)
Full + Full	Maxout [9]	9.38
	NIN [27]	8.81
	DSN [23]	8.22
	FitNet [34]	8.39
	ResNet-110 [13]	6.43
	VGG-Small	6.82
1-bit + Full	BinaryConnect [3]	8.27
2-bit + Full	Ternary Weight Network [24]	7.44
1-bit + 1-bit	BNN [4]	10.15
1-bit + 2-bit	VGG-Small-HWGQ	7.49