基于聚类正则量化的深度神经网络压缩方法

Cluster Regularized Quantization for Deep Networks Compression

○胡 一鸣, 李 建权, 龙 宪磊, 胡 申华, 朱 佳刚, 王 欣刚, 顾 庆毅 (中国科学院自动化研究所)

○ Yiming Hu, Jianquan Li, Xianlei Long, Shenhua Hu, Jiagang Zhu, Xingang Wang, Qingyi Gu (CASIA)

Introduction

Quantization

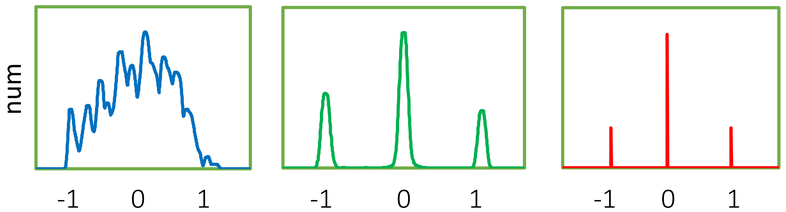
Learn an optimal quantizer that can minimize the quantization error for the input data distribution:

$$Q^*(x) = \underset{Q}{argmin} \int p(x)(Q(x) - x)^2 dx$$

Existing Quantization Methods

	Can achieve integer quantization
	Can achieve extreme low-bit quantization
☒	High quantization error and low accuracy

Stage Ⅰ Original Weight | **Stage Ⅱ** Cluster Regularized training | **Stage Ⅲ** Quantization



The proposed CRQ

Cluster Regularization

Formulation

To reduce the quantization error, we introduce a cluster regularization to encourage weight cluster around target values:

$$J(Z, \alpha) = ||W - \alpha M^T Z||^2$$

$$\alpha^*, Z^* = \underset{\alpha, Z}{argmin}\, J(Z, \alpha)$$

$$s.t. Z \in \{0,1\}^{3\times N}, \sum_i^3 Z_j^i = 1, \forall j \in [1, N]$$

Optimization

Solve Z with α fixed

$$z_i = \begin{cases} 1 & i = H(M, \frac{w}{\alpha}) \\ 0 & otherwise \end{cases}$$

Solve α with Z fixed

$$\alpha = \frac{WZ^TM}{M^TZZ^TM}$$

In this way, α and Z are updated iteratively until convergence.

Experimental Results

Results on ImageNet

Method	Model	Top1	Top1↓	Top5	Top5↓
Ref	AlexNet	42.80%	-	19.70%	-
	ResNet18	30.40%	-4.30%	13.80%	-3.00%
TWN	AlexNet	45.50%	-2.70%	23.20%	-3.50%
	ResNet18	34.70%	-4.30%	13.80%	-3.00%
TTQ	AlexNet	42.50%	0.30%	20.30%	-0.60%
	ResNet18	33.40%	-3.00%	12.80%	-2.00%
Our Ref	AlexNet	42.77%	-	19.79%	-
	ResNet18	30.89%	-	11.18%	-
Our CRQ	AlexNet	42.02%	**0.75%**	19.18%	**0.61%**
	ResNet18	32.88%	-1.99%	12.56%	-1.38%

Results on CIFAR

Model	Method	Ref	Error	Error↓
ResNet-32	TTQ	7.67%	7.63%	0.04%
ResNet-32	CRQ	7.74%	7.61%	**0.13%**
ResNet-44	TTQ	7.18%	7.02%	0.16%
ResNet-44	CRQ	7.21%	6.95%	**0.26%**