模型量化

问题

不同的网络结构，如何保精度地量化到 8 比特甚至更低的比特位数？

量化研究分类

按照量化方式可以分为

- 线性量化：即量化分立值是均匀的，绝大多数文章研究线性量化
- 非线性量化：量化分立值非均匀

按照是否从 pre-trained 模型出发，可以分为

- 网络模型量化：即对一个 pre-trained full-precision 模型量化到 fixed-point precision 网络
- 量化网络：从头训练一个量化的网络

对网络模型量化，按照训练方式分为

- Post-training，这种不需要训练，基本只是做 calibaration
- Training-aware，这种需要量化模拟训练，
    - 按照使用的 loss 分类，可以分成
        - 最小化量化误差 (QE Minimization)，通过最小化量化误差来决定量化超参
        - 数据统计形式(Data Statics)，通过数据统计信息决定量化超参
        - 量化超参 BP(BackProp)，即通过网络总的 loss 对超参进行回传，按照研究 BP 的类型，可以分为
            - 使用 STE 近似：即在 STE 近似基础下是 BP 更加有效
            - 对 STE 近似改进：对 STE 本身进行改进使得 BP 更有效

量化策略

量化理论

论文列表

Quantization Strategies

论文题目	tags	评价	相关资料
Effective Training of Convolutional Neural Networks with Low-bitwidth Weights and Activations		低比特量化训练时难点在于量化函数不可导，训练时梯度不能很有效地回传，在网络非常深的时候很难收敛足够高的精度。作者提出三种策略来量化：（1）渐进量化，（2）随机量化，(3)双向知识蒸馏。作者对这些策略分别进行了测试并组合，发现在低比特量化时相比基准线都有提升，组合这些策略可以进一步提升量化精度。	https://arxiv.org/abs/1908.04680

WRPN: WIDE REDUCED-PRECISION NETWORKS		受 Wide ResNet 启发，作者尝试了将网络的 channel 数增加再进行量化，发现在极低比特量化后网络仍能保持和 full-precision 一样的精度，部分实验甚至发现量化后网络速度比 full-precision 快的时候精度甚至提高了	https://arxiv.org/abs/1709.01134
Incremental Network Quantization: Towards Lossless CNNs with Low-Precision Weights			http://arxiv.org/abs/1612.01064

Post-training

论文题目	tags	评价	相关资料
【Presentation】8-bit Inference with TensorRT		TensorRT Calibration from Nvidia	http://on-demand.gputechconf.com/gtc/2017/presentatio tensorrt.pdf
【Presentation】Low Precision Inference on GPU		Int8 Quantization General Introduction from Nvidia	https://developer.download.nvidia.com/video/gputechco inference-at-reduced-precision-on-gpus.pdf
Fighting Quantization Bias With Bias		不用训练的缓解 mobileNet 内的 MAS 现象	
Post training 4-bit quantization of convolutional networks for rapid-deployment		无训练 4bit，很多计算近似，效果比 TensorRT 好一点	
Bit Efficient Quantization for Deep Neural Networks			https://arxiv.org/abs/1910.04877

Training-aware

流派	论文题目	tags	评价	相关资料
Fixed	A Quantization-Friendly Separable Convolution for MobileNets		解决了 mbv2 的 dw-conv 的掉点问题	

	Convolutional networks for fast, energy-efficient neuromorphic computing			
	Discovering low-precision networks close to full-precision networks for efficient embedded inference			
	Quantizing deep convolutional networks for efficient inference: A whitepaper		google 的经典白皮书，量化入门必读	
QE Minimization	Lq-nets: Learned quantization for highly accurate and compact deep neural networks		本文提出了一个有效训练 quantizer 的方法，使得在 very low-bit 量化时，相比这篇文章之前的 SOTA 方法都有提升	https://arxiv.org/abs/1807.10029

	Weight Normalization based Quantization for Deep Neural Network Compression			
BackProp	Joint training of low-precision neural network with quantization interval parameters	QIL	和 QIL 同一篇，包含更多细节	https://deeplearn.org/arxiv/44389/joint-t precision-neural-network-with-quantizatio
	Nice: Noise injection and clamping estimation for neural network quantization	NICE	作者在量化训练的同时逐步在 weights 中注入噪声，使得量化具有 dropout 的效果，通过实验发现在量化比特数 b≥4 时相比其他方法如 PACT/LQ-NET 等有所提升. 这个 trick 可以用在一般量化感知训练中	https://arxiv.org/abs/1810.00162
	Learned Step Size Quantization	LSQ	对 scale 阈值进行 online 训练，相比 PACT 完善了 round 函数的梯度反向传递，效果更好	https://arxiv.org/abs/1902.08153

	Trained Uniform Quantization for Accurate and Efficient Neural Network Inference on Fixed-Point Hardware	ALT	作者在 log-domain 对 clip threshold 进行训练，并克服了 LSQ 具体训练过程的困难，在 MobileNets v1/v2 中测试只需 5 个 epoch FLQ 训练，不掉点	https://arxiv.org/abs/1903.08066
	Learning to Quantize Deep Networks by Optimizing Quantization Intervals with Task Loss	QIL	作者提出了一种同时包含 pruning 和 clipping 的 non-linear quantizer，利用 task loss 学习 quantizer 参数可以使得在 4-bit 下也能保持 full-precision 的精度。据说是三星 npu 芯片的 4-bit 核心算法	https://arxiv.org/abs/1808.05779

	Two-Step Quantization for Low-bit Neural Networks	TSQ	作者将 low-bit 训练分成两步：稀疏量化学习和在 low-bit constraints 下的非线性回归。在 AlexNet 上做 2-bit 量化，该方法相对于 full-precision 只掉了 0.5 个点，相较于其他方法掉 5 个点提升显著	http://openaccess.thecvf.com/content_cvpr_ Step_Quantization_for_CVPR_2018_paper.pdf
	Training Quantized Network with Auxiliary Gradient Module		低比特量化新的 SOTA，2bit 量化相比较之前的方法不少提升，效果比 KD 有优势。 作者提出了一种 Auxilary gradient module 代替 KD 使得量化训练梯度回传变得容易，并利用了最新的量化策略，在目前最好的低比特量化方法基础上都有提升。	https://arxiv.org/abs/1903.11236

	Learning Low-precision Neural Networks without Straight-Through Estimator (STE)	AB		https://arxiv.org/abs/1903.01061
	Relaxed Quantization For Discretized Neural Networks	RQ		https://arxiv.org/abs/1810.01875
	ProxQuant: Quantized Neural Networks via Proximal Operators			http://arxiv.org/abs/1810.00861
	Mirror Descent View for Neural Network Quantization			https://arxiv.org/abs/1910.08237
BackProp+KD				
	Apprentice: Using knowledge distillation techniques to improve low-precision network accuracy			https://openreview.net/forum?id=B1ae1lZRb

	Model compression via distillation and quantization		蒸馏量化的一篇经典文章。作者提出了两种蒸馏量化方式：(1) quantized distillation: 在训练量化分立的 student 网络的过程中使用蒸馏 loss，（2) differentiable quantization: 通过 SGD 优化量化的分立值位置，使得模型更加 fit teacher 网络。实验发现在蒸馏的 4bit 量化 2xResNet18 网络上达到 73.31% 测试精度，甚至比未量化的 ResNet18 模型 (69.75%) 精度高。	https://openreview.net/forum?id=S1XolQbRW https://openreview.net/forum?id=S1XolQbRW
Others	EIE: Efficient Inference Engine on Compressed Deep Neural Network		介绍量化底层实现	https://arxiv.org/pdf/1602.01528
	GDRQ: Group-based Distribution Reshaping for Quantization Haibao			

Quantized Network

论文题目	tags	评价	相关资料
BinaryConnect: Training Deep Neural Networks with binary weights during propagations	BNN	作者提出了一种 1-bit 的 Quantized Neural Networks (QNN) ，在 MNIST 上测试速度快了 7 倍同时并没有带来精度上的降低	
Binarized Neural Networks: Training Neural Networks with Weights and Activations Constrained to +1 or −1			https://arxiv.org/abs/1609.07061
Towards Accurate Binary Convolutional Neural Network			
Quantized Neural Networks: Training Neural Networks with Low Precision Weights and Activations			http://arxiv.org/abs/1609.07061
Ternary weight networks			
Trained ternary quantization			

DoReFa-Net: Training Low Bitwidth Convolutional Neural Networks with Low Bitwidth Gradients			
Xnor-net: Imagenet classification using binary convo- lutional neural networks			
Bridging the accuracy gap for 2-bit quantized neural networks			https://arxiv.org/abs/1807.06964
Deep learning with low precision by half-wave gaussian quantization			https://arxiv.org/abs/1702.00953
Learning Discrete Weights Using the Local Reparameterization Trick		作者提出了一种新的训练 Binary/Ternary 网络的方法，通过一种 local reparameterization trick 可以成功地训练 discrete weights，效果比之前的 stochastic 或者 STE 的方法好	https://openreview.net/forum?id=BySRH6CpW https://discourse.brainpp.cn/t/topic/21524
Structured Binary Neural Networks for Accurate Image Classification and Semantic Segmentation			https://arxiv.org/abs/1811.10413

BNN+: Improved Binary Network Training			https://arxiv.org/abs/1812.11800

Nonlinear Quantization

论文题目	tags	**评价**	**相关资料**
Weighted-entropy-based quantization for deep neural networks		作者对 logQuant 非线性量化器做了改进，将 weights 分块并按照重要性度量搜索找到其应该量化成的值，实验发现效果比 XNOR-Net 和 DoReFa-Net 要好	https://ieeexplore.ieee.org/document/8100244

Additive Powers-of-Two Quantization: A Non-uniform Discretization for Neural Networks		作者通过将几个power-of-two非线性量化器叠加增加了power-of-two量化的精度，同时可以利用其计算速度优势。实验结果表明，和最近新的量化方法如 QIL, DSQ, LQ-Net, PACT, DoReFa相比精度要高	https://openreview.net/forum?id=BkgXT24tDS
Joint training of low-precision neural network with quantization interval parameters	QIL		https://deeplearn.org/arxiv/44389/joint-training-of-low-precision-neural-network-with-quantization-interval-parameters

Lq-nets: Learned quantization for highly accurate and compact deep neural networks	LQ-Net		

Quantization Theory

论文题目	tags	评价	相关资料
The High-Dimensional Geometry of Binary Neural Networks			https://arxiv.org/abs/1705.07199
Training Quantized Nets: A Deeper Understanding			https://arxiv.org/abs/1706.02379

Review Articles

论文题目	tags	评价	相关资料

A Survey on Methods and Theories of Quantized Neural Networks		主要讲了量化网络方面的进展，可惜对网络量化部分涉及不足	https://arxiv.org/abs/1808.04752