DESIGNING STRONG BASELINES FOR TERNARY NEURAL NETWORK QUANTIZATION THROUGH SUPPORT AND MASS EQUALIZATION

Edouard Yvinec[1,2] , Arnaud Dapogny[2] , Kevin Bailly[1,2]

Sorbonne Université[1], CNRS, ISIR, f-75005, 4 Place Jussieu 75005 Paris, France
Datakalab[2], 114 boulevard Malesherbes, 75017 Paris, France

The chat GPT slide + some green washing ¯_(ツ)_/¯

- training GPT-3 requires **1,300MWh** (320 homes/year), or **552 tons of CO2**
- one instance of GPT-3 generates **8.4 tons of CO2 per year** (5 homes/year)
- initially the price to run GPT-3 with **100,000$** per month
- It is reported that OpenAI currently uses **700,000$** worth of resources per day

no sources as I do not want to give too much credit to these whispers.

How do we compress models for deployment?

Pruning

- remove operations or computations
- granularity level: structure[4] of the removal
- usage of data: pruning at initialization[1], pruning post-training[3], iterative pruning[2]
- paradigm: importance based[4] or similarity based[3]

Quantization

- simplify the individual computations (e.g. from fp32 to int4)
- granularity level[5]: per-tensor, per-channel, per-group
- usage of data[6]: data-free, gptq, or qat
- mixed-precision[7]: use the adequate precision (RL, heuristics)
- quantization space[8]: uniform, log, power,...

NAS and Distillation

- Use the adequate architecture to begin with
- Search for efficient architectures[9]
- Use adapters and distillation to compress models (very trending on LLMs)[10]

[1] Tanaka, Hidenori, et al. "Pruning neural networks without any data by iteratively conserving synaptic flow." NIPS 2020

[2] Frankle, Jonathan, and Michael Carbin. "The Lottery Ticket Hypothesis: Finding Sparse, Trainable Neural Networks." ICLR 2018

[3] Yvinec, Edouard, et al. "Red: Looking for redundancies for data-free structured compression of deep neural networks." NIPS 2021

[4] Yvinec, Edouard, et al. "SInGE: Sparsity via Integrated Gradients Estimation of Neuron Relevance." NIPS 2022

[5] Yvinec, Edouard, et al. "SPIQ: Data-Free Per-Channel Static Input Quantization." WACV 2023

[6] Nagel, Markus, et al. "Data-free quantization through weight equalization and bias correction." ICCV 2019.

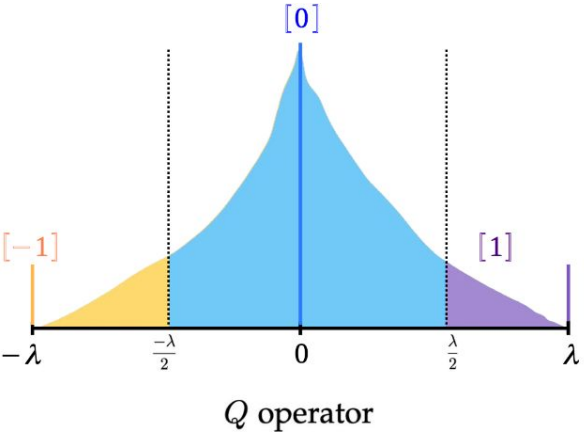
[7] Wang, Kuan, et al. "Haq: Hardware-aware automated quantization with mixed precision." CVPR 2019.

[8] Yvinec, Edouard, et al. "Powerquant: Automorphism search for non-uniform quantization." ICLR 2023

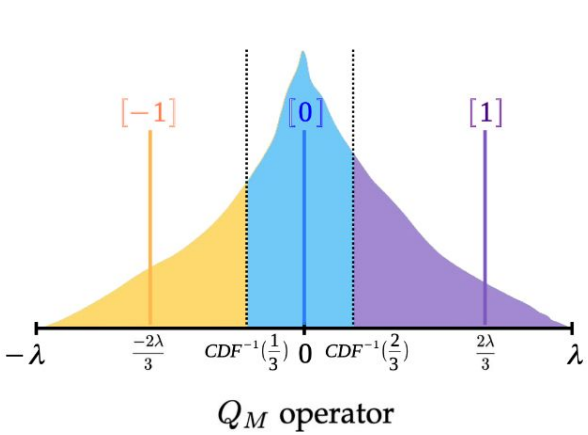
[9] Tan, Mingxing, and Quoc Le. "Efficientnet: Rethinking model scaling for convolutional neural networks." ICML 2019.

[10] Hu, Edward J., et al. "LoRA: Low-Rank Adaptation of Large Language Models." ICLR 2023

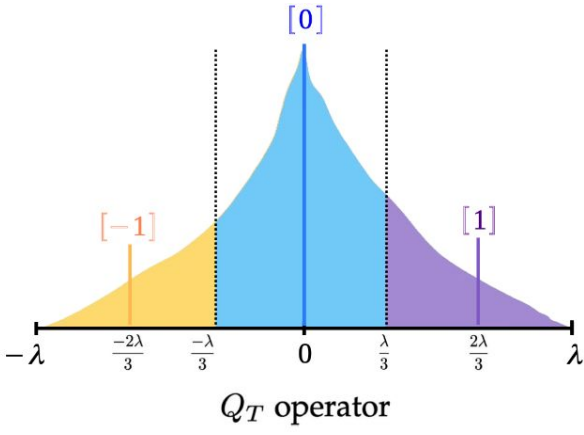
Ternay Quantization



Q operator



Q_M operator

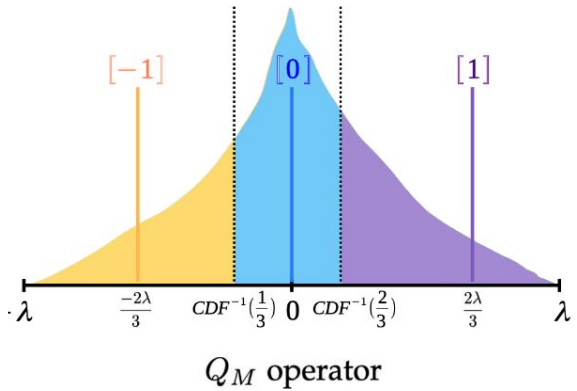


Q_T operator

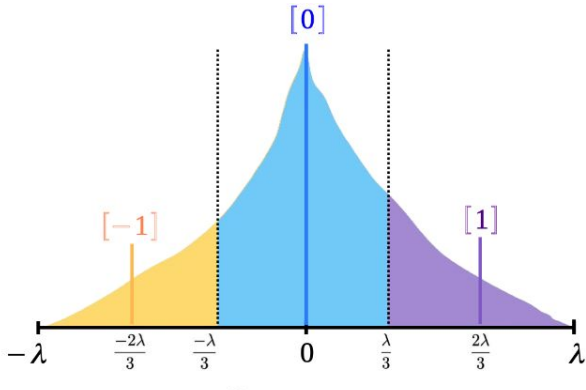
Ternary quantization has a specificity: the zero value is assigned many more values as compared to 1 and -1. To fight this phenomenon, we propose two novel ternary quantization operators.

A little bit of Maths

- When we equalize the mass (top right), we minimize the expected error introduced by quantization
- When we equalize the support (bottom right), we minimize the maximum error introduced by quantization.

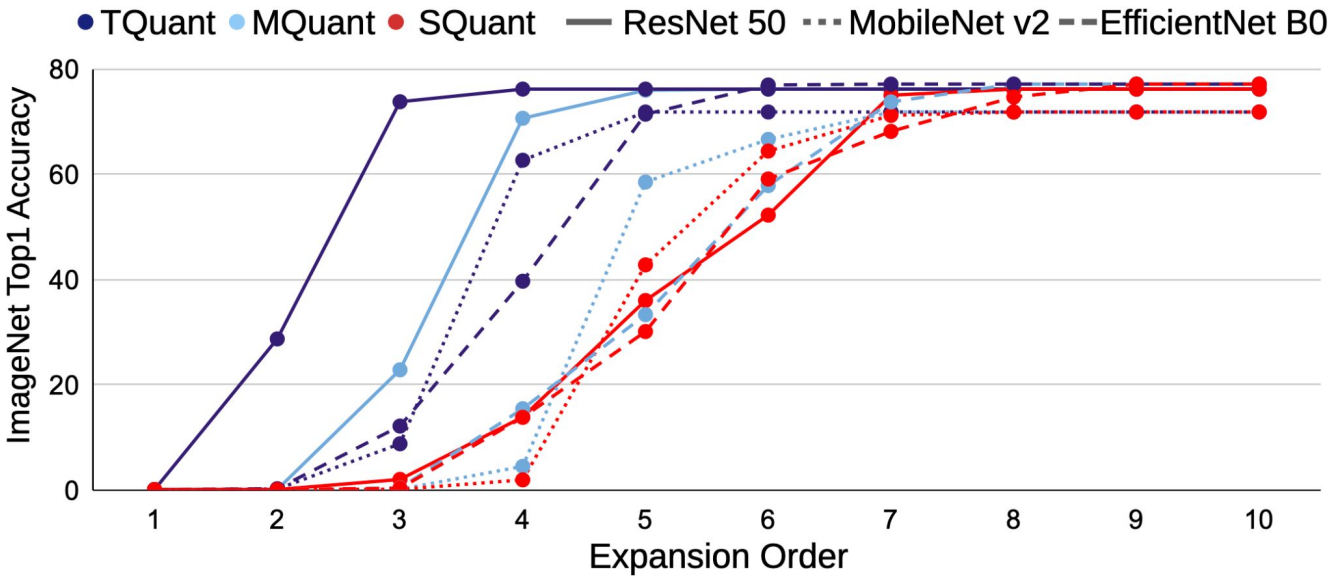


Q_M operator



Q_T operator

Evaluation with data-free quantization



Evaluation with gradient-based post-training quantization

PTQ method	operator	accuracy	Processing Time
-	-	89.100	-
AdaRound	native	11.790 ± 3.210	5m01
	MQuant	**42.910** ± 0.620	5m18
	TQuant	40.490 ± 0.250	5m18
BrecQ	native	25.780 ± 2.440	3m45
	MQuant	**63.540** ± 0.850	3m50
	TQuant	58.000 ± 1.120	3m50

Evaluation with quantization aware training

	Baseline	MQuant	TQuant
accuracy	42.910 ± 14.61	68.250 ± 6.26	$\mathbf{82.620} \pm 2.43$