Low-bit Quantization of Neural Networks for Efficient Inference

Yoni Choukroun, Eli Kravchik, *Fan Yang*, Pavel Kisilev

Huawei Technologies Co., Ltd

Introduction

Background

- Recent breakthroughs in machine learning make use of large deep Neural Networks (NN) deployed on powerful GPUs.
- For applications running on limited hardware, real-time processing with high accuracy is still a challenge.
- Straightforward implementation of these models puts a heavy strain on resources, increasing memory loading, computation latency and power consumption.
- Currently, the most popular method to tackle these constraints is **to reduce the precision of weights and activations**, such that their memory loading and computations on dedicated hardware will be handled more efficiently.

Motivation

Quantization scenario:

- **Hardware compliant low bit (4 bits) linear post-training quantization of NN without retraining**
- Dataset access limitation：
  - **No retraining** (small calibration set is available)
- Hardware limitation:
  - **Linear quantization** (no LUT, no tensor partitioning)
  - **No mixed precision** (INT4 only)

How to solve the problem:

- No retraining => **Minimum Mean Squared Error for quantization error optimization + unlabeled calibration**
- Linear quantization => **Kernel wise quantization**
- No mixed precision => **Dual tensors approximation for badly approximated layers**

Linear quantization

Linear quantization

- Hardware friendly: transfer the high computational cost FP convolution to low-bit integer convolution + matrix-vector multiplication

$$T_3 = T_1 T_2 \approx (\alpha_1 \tilde{T}_1)(\alpha_2 \tilde{T}_2) = (\alpha_1 \alpha_2)(\tilde{T}_1 \tilde{T}_2) = \alpha_3 \tilde{T}_3$$

- Uniform quantization scheme:

$$\hat{T} = \alpha \left[\frac{T - \delta}{\alpha} \right]_{\mathbb{Z}_p} + \delta$$

$$\delta = \min_i (T_i)$$

$$\alpha = \frac{\max_i (T_i) - \delta}{2^p - 1}$$

- Quantization offset δ can be of major importance for non symmetric distributions, but increases the computational complexity of tensor multiplications
  - Without quantization offset : $\delta = 0, \alpha = \max_i(|T_i|)/(2^{p-1} - 1)$
  - With quantization offset : $\mathbb{Z}_p \in \{-2^{p-1}, \ldots, 2^{p-1} - 1\}$

Mean Squared Error Analysis

Mean Squared Error Analysis

- Approximation of single layer quantization:

$$\begin{aligned}\hat{Y} &= \hat{W}\hat{X} = (W + n_W)(X + n_X) \\ &= WX + Wn_x + n_W X + n_W n_X \\ &\approx WX + Wn_X + n_W X = Y + Wn_X + n_W X\end{aligned}$$

<= n_X, n_W denote the quantization noise of the weights

- Consider the case where the NN is composed of linear layers only: $Y = W_L(W_{L-1}...(W_1 X_1)) = W_L X_L$

- Defining e_L^2, the MSE between the original model output and the quantized model output, we will have:

$$\begin{aligned}\mathbb{E}(e_L^2) &= \mathbb{E}\,\|Y - \hat{Y}\|_F^2 = \mathbb{E}\,\|W_L X_L - \hat{W}_L \hat{X}_L\|_F^2 \\ &= \mathbb{E}\,\|W_L n_{X_L} + n_{W_L} X_L\|_F^2 \\ &= \mathbb{E}\,\|W_L n_{X_L}\|_F^2 + 2\,\mathbb{E}\,trace(n_{X_L}^T W_L^T X_L n_{W_L}) \\ &+ \mathbb{E}\,\|n_{W_L} X_L\|_F^2 \\ &\approx \mathbb{E}\,\|W_L n_{X_L}\|_F^2 + \mathbb{E}\,\|n_{W_L} X_L\|_F^2 \\ &\leq \mathbb{E}\,\|n_{W_L}\|_F^2\,\mathbb{E}\,\|X_L\|_F^2 + \mathbb{E}\,\|W_L\|_F^2\,\mathbb{E}\,\|n_{X_L}\|_F^2 \\ &= \mathbb{E}\,\|n_{W_L}\|_F^2\,\mathbb{E}\,\|X_L\|_F^2 + \mathbb{E}\,\|W_L\|_F^2\,\mathbb{E}(e_{L-1}^2),\end{aligned}$$

Insights:

1. MMSE should be minimal
2. First layers have bigger impact
3. Weights have bigger impact (recursion factor)

PS : the approximation is obtained by assuming zero mean noise and where $\|\cdot\|_F$ denotes the Frobenius norm. The inequality is obtained using Cauchy Schwarz inequality and by assuming the weights/activations and the noises are statistically independent.

Quantization of Weights (1/4) - Kernel wise quantization

Kernel wise quantization

- Hardware friendly, no supplementary computation cost
- Scaling factor α :
  - Fineness: shared per layer => shared per output channel
  - Shape : single value => vector with length of output channel
- Quantizing each kernel separately with a different scaling factor $W_l \cong \{\alpha_{lk}\widetilde{W}_{lk}\}_{k=1}^{K_l}$,maintains the linearity of the dot product is: $A_l = \{A_{lk}\}_{k=1}^{K_l} = \{W_{lk} \cdot X_l\}_{k=1}^{K_l} \cong \{(\alpha_{lk}\widetilde{W}_{lk}) \cdot (\beta_l\tilde{X}_l)\}_{k=1}^{K_l} = \{(\alpha_{lk} \cdot \beta_l) \cdot (\widetilde{W}_{lk} \cdot \widetilde{X}_l)\}_{k=1}^{K_l}$, where W is weight, X is activations, A is convolution result, l is the layer index and k is the kernel index.
- **Improvement :**

Architecture	Original	Global	Kernel-wise
Alexnet [22]	56.624%	0.694%	**46.796%**
VGG16bn [32]	73.476%	3.034%	**65.23%**
Inception v3 [35]	76.226%	0.106%	**12.564%**
Resnet18 [13]	69.644%	1.83%	**44.082%**
Resnet50 [13]	76.012%	0.172%	**62.242%**
Resnet101 [13]	77.314%	0.148%	**64.434%**
SqueezeNet [19]	58.0%	1.528%	**29.908%**
DenseNet [17]	74.472%	0.58%	**57.072%**

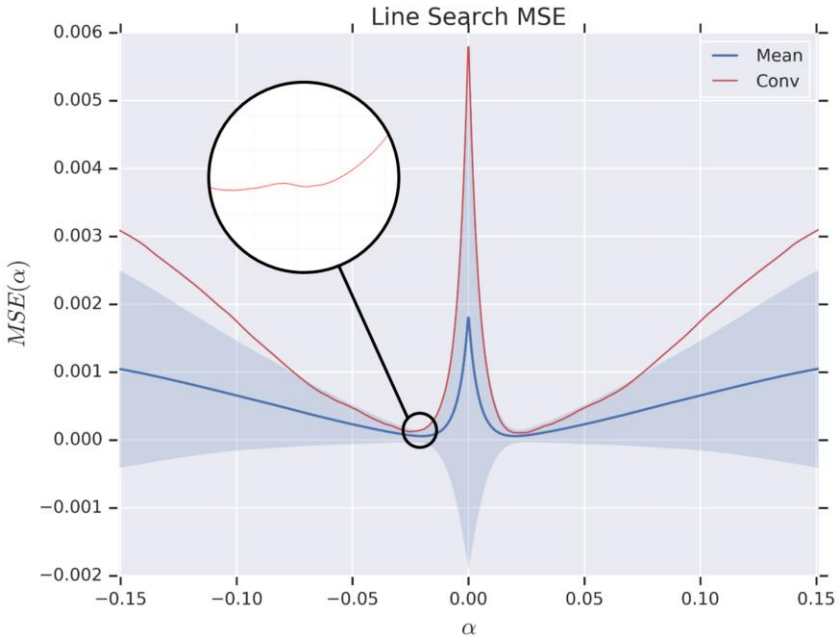
Quantization of Weights (2/4) - Minimum MSE Quantization

Minimum MSE Quantization

- Key point: a defines both the quantization precision and the saturation value in case of linear quantization without offset

$$\alpha_{lk} = argmin_a \left\{ \left\| W_{lk} - a \left[\frac{W_{lk}}{a} \right]_{Z_4} \right\|^2 \right\}$$

- Typical MSE as a function of α is presented :



- Several methods for finding optimal α can be considered, we propose to use one dimensional exact line-search for optimal results on non convex problem.
- **Results :**

Architecture	Uniform	Altern.	Golden	OMSE
Alexnet	46.796%	40.962%	46.070%	**46.892%**
VGG16bn	65.23%	55.936%	62.250%	**65.414%**
Inception v3	12.564%	4.368%	7.408%	**22.028%**
Resnet18	44.082%	52.646%	52.398%	**56.688%**
Resnet50	62.242%	60.186%	63.178%	**67.356%**
Resnet101	64.434%	56.282%	63.052%	**65.066%**
SqueezeNet	29.908%	24.040%	**34.036%**	32.262%
DenseNet	57.072%	45.668%	50.058%	**59.012%**

Quantization of Weights (3/4) - Multiple Tensor Quantization

Multiple Tensor Quantization

- **Problem :**
  - Some layers can be harder to approximate than others, and have larger impact on the model accuracy loss
  - No support of mixed precision by hardware
- **Solution：** multiple low precision quantized tensors

$$TX \approx \sum_{i=1}^{n} \alpha_i \tilde{T}^i X \approx \sum_{i=1}^{n} \alpha_i \beta (\tilde{T}^i \tilde{X})$$

Algorithm 1 Alternating Optimization for Multiple Quantization of high MSE layer

Input: Tensor T, desired precision $\{p_i\}_{i=1}^n$ and quantization mapping ϕ.

Output: $\{\alpha_i\}_{i=1}^n$ and $\{\tilde{T}_i\}_{i=1}^n$

1: **while** convergence rate $> \epsilon$ **do**
2: **for** $j \in [1, ..., n]$ **do**
3: $(\alpha_j, \tilde{T}^j) = \phi(T - \sum_{i=1, i \neq j}^{n} \alpha_i \tilde{T}^i, p_j)$
4: **end for**
5: **end while**

- Due to supplementary computation cost, the **Line Search** dual tensor quantization is only recommended to use in difficult layer.
- **Results：**

Architecture	Original	OMSE	Dual	CR
Alexnet	56.624%	46.892%	**54.408%**	0.1256
VGG16bn	73.476%	65.414%	**66.932%**	0.125
Inception v3	76.226%	22.028%	**51.642%**	0.1263
Resnet18	69.644%	56.688%	**64.034%**	0.1318
Resnet50	76.012%	67.356%	**70.060%**	0.1261
Resnet101	77.314%	65.066%	**71.492%**	0.1261
SqueezeNet	58.0%	32.262%	**53.466%**	0.1493
DenseNet	74.472%	59.012%	**64.400%**	0.1432

Quantization of Weights (4/4) - Scaling Factors Refinement

Scaling Factors Refinement

- **Problem** : Presented quantization methods based on several approximation algorithms, and quantization errors accumulate
- **Solution**： additional (re)scaling factors γ to better approximate the full precision model, trained from unlabeled calibration data

$$\hat{T}(\gamma) = \gamma\left(\alpha\left[\frac{T}{\alpha}\right]_{\mathbb{Z}_p}\right)$$

- **Optimization objective**:

$$\min_{\gamma_l=\{\gamma_{lk}\}_{lk}} \sum_{i}^{M} \|f(X_i, \{W_l\}_l) - f(X_i, \{\hat{W}_l(\gamma_l)\}_l)\|_F^2$$

f(X, $\{w_l\}_l$): NN mapping function, M: size of the calibration set

Advantages:

- Small number of values to optimize
- Fast convergence with small calibration set
- No label requirement

- **Results**：

Architecture	Dual	OMSE+Opt.	Dual+Opt.
Alexnet	54.408%	53.306%	**55.226** %
VGG16bn	66.932%	72.294%	**72.576**%
Inception v3	51.642%	73.656%	**74.790**%
Resnet18	64.034%	67.120%	**68.806**%
Resnet50	70.060%	74.672%	**74.976**%
Resnet101	71.492%	76.226%	**76.402**%
SqueezeNet	53.466%	54.514%	**56.248**%
DenseNet	64.400%	71.730%	**73.600**%

Quantization of Activations

Problem:

- Unlike weights, activations can't be quantized offline.

Solution:

- Calculate optimal scaling factors offline on a small unlabeled calibration set by approximating the activations distribution
- Use line search optimization, for each layer separately, a MSE minimization problem over a small calibration set :

$$\beta_l = argmin_b \left\{ \sum_{m \in M} \left\| X_l^m - b \left[\frac{X_l^m}{b} \right]_{Z_4} \right\|^2 \right\}$$

- Key layers with high quantization noise are approximated by dual Int4 tensor representation.
- Dual Suffer from overfitting, so we opt for learning residual parameters such that

$$X \approx \hat{X} = \beta_1 \tilde{X}_1 + \beta_2 \tilde{X}_2$$
$$= \beta_1 \left[\frac{X}{\beta_1} \right]_{\mathbb{Z}_p} + \beta_2 \left[\frac{X - \beta_1 \tilde{X}_1}{\beta_2} \right]_{\mathbb{Z}_p}$$

Results:

Architecture	Original	Baseline	KLD	OMSE	Dual (CR)	ACIQ*	OMSE*	Dual*
Alexnet	56.624%	38.616%	44.23%	**48.908%**	**54.552%** (0.195)	-	**49.122%**	**54.994%**
Alexnet (offset)	56.624%	51.774%	52.948%	**53.286%**	**55.522%** (0.195)	52.304%	**53.998%**	**55.508%**
VGG16bn	73.476%	33.276%	53.686%	**62.168%**	**68.120%** (0.127)	-	**67.726%**	**68.334%**
VGG16bn (offset)	73.476%	58.832%	64.336%	**67.198%**	**71.478%** (0.127)	67.716%	**71.260%**	**71.260%**
Inception v3	76.226%	4.042 %	23.658 %	**40.916%**	**66.176%** (0.154)	-	**43.184%**	**68.608%**
Inception v3 (offset)	76.226%	57.516 %	64.504%	**67.964%**	**73.060%** (0.151)	59.826%	**69.528%**	**74.486%**
Resnet18	69.644%	48.37%	56.728%	**61.268%**	**66.522%** (0.150)	-	**63.744%**	**66.628%**
Resnet18 (offset)	69.644%	63.106%	64.486%	**64.992%**	**68.380%** (0.148)	65.694%	**67.508%**	**68.316%**
Resnet50	76.012%	40.786%	57.57%	**64.878%**	**70.368%** (0.129)	-	**66.562%**	**70.202%**
Resnet50 (offset)	76.012%	65.338%	68.328%	**71.274%**	**73.252%** (0.126)	71.362%	**73.392%**	**73.392%**
Resnet101	77.314%	36.494%	58.744%	**65.316%**	**70.770%** (0.129)	-	**65.350%**	**70.806%**
Resnet101 (offset)	77.314%	65.552%	71.412%	**72.750%**	**74.266%** (0.126)	69.544%	**74.332%**	**74.332%**
SqueezeNet	58.0%	3.282%	9.582%	**18.630%**	**52.382%** (0.216)	-	**20.448%**	**52.430%**
SqueezeNet (offset)	58.0%	24.97%	35.806%	**39.820%**	**56.150%** (0.203)	-	**42.026%**	**56.284%**
DenseNet	74.472%	47.808%	64.062%	**65.032%**	**67.952%** (0.132)	-	**67.558%**	**68.048%**
DenseNet (offset)	74.472%	67.676%	69.578%	**70.118%**	**72.282%** (0.132)	-	**72.304%**	**72.310%**

INT4 quantization for activations and INT8 for weights. Test result with calibration set of 250 images, *means no quantization of the first layer of NN. Offset means activation quantization with offset.

Review of our Quantization Pipeline

Quantization of weights

- Kernel-wise quantization allows better reconstruction while remaining HW compliant
- Optimal MMSE quantization:

$$\alpha_{lk} = argmin_a \left\{ \left\| W_{lk} - a \left[\frac{W_{lk}}{a} \right]_{Z_4} \right\|^2 \right\}$$

- Dual tensors approximation for badly approximated layers: $W_{lk} \cong \alpha_{lk}^1 \widetilde{W}_{lk}^1 + \alpha_{lk}^2 \widetilde{W}_{lk}^2$
- Refinement of scaling factors α

Quantization of activations

- Optimal MMSE quantization (saturation) of activations collected from small calibration set
- Dual tensors approximation using residuals in order to avoid overfitting

Experiments and Results (1/2)

Performance of the proposed framework on INT4 quantization with corresponding CR

Model	Original		KLD [1]		Our		Our+offset		CR(W,A)	% Dual
Alexnet	56.624%	79.056%	35.590%	59.326%	53.632%	77.244%	**54.476%**	**77.846%**	(0.125,0.195)	25%
VGG16bn	73.476%	91.536%	41.530%	66.008%	67.492%	88.016%	**70.658%**	**90.136%**	(0.125,0.127)	6.25%
Resnet18	69.644%	88.982%	31.934%	55.510%	65.456%	86.630%	**67.416%**	**87.716%**	(0.126,0.148)	23.80%
Resnet50	76.012%	92.934%	46.190%	70.162%	69.370%	89.204%	**72.602%**	**90.852%**	(0.126,0.129)	3.70%
Resnet101	77.314%	93.556%	49.948%	73.034%	69.700%	89.686%	**73.602%**	**91.526%**	(0.126,0.128)	1.90%
Inception v3	76.226%	92.984%	1.84%	4.848%	64.572%	85.852%	**71.606%**	**90.470%**	(0.126,0.154)	11.57%
SqueezeNet	58.184%	80.514%	8.224%	19.348%	50.722%	74.634%	**55.358%**	**78.482%**	(0.149,0.216)	68%
DenseNet	74.472%	91.974%	44.05%	68.52%	66.832%	87.518%	**71.558%**	**90.532%**	(0.143,0.132)	2.47%
SSD300	77.43%		47.38%		73.94%		**75.77%**		(0.125,0.126)	2.85%

Performance of the proposed framework for INT4 weights and activations obtained using $\tau = 8 \cdot 10{-}5$. Compression ratios of both the weights and the activations and the percentage of dual layers are provided in the last two columns, respectively. The mean top-1 decay is of 6.7% and 3% for the regular and offset (unsigned) versions, respectively. Similarly, mean top-5 decay is of 4% and 1.7% respectively.

Experiments and Results (2/2)

Trade-off with quantization threshold

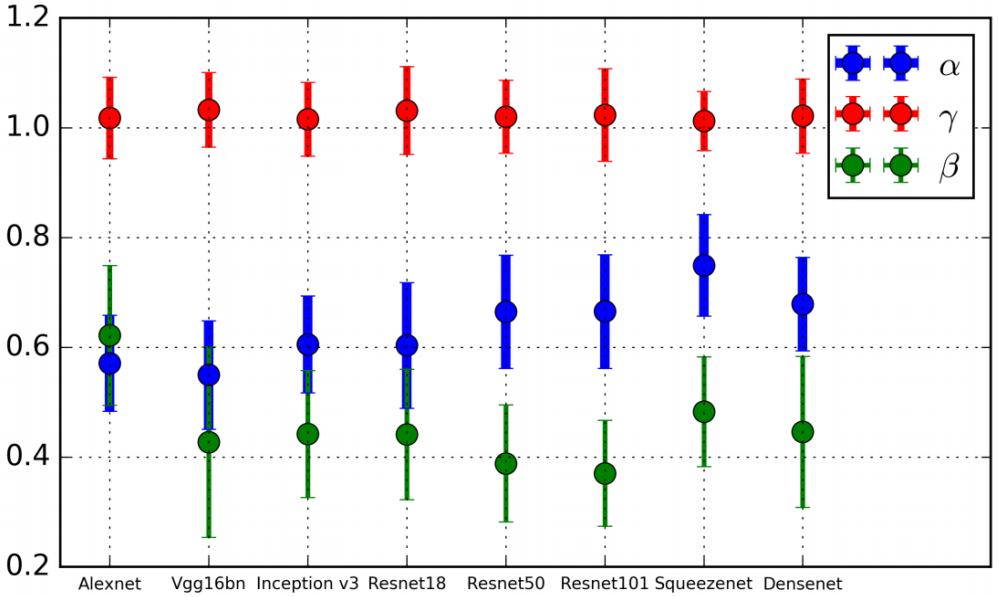
- Modification of the MSE threshold τ influences the accuracy-compression ratio trade-off

Architecture	Original	Our	CR(W,A)	Our +offset
Alexnet	56.624%	49.570%	(0.125,0.178)	52.274%
		54.186%	(0.127,0.206)	54.936%
		54.720%	(0.130,0.245)	56.068%
Resnet50	76.012%	69.472%	(0.126,0.129)	72.530%
		73.078%	(0.137,0.156)	74.826%
		74.336%	(0.154,0.203)	75.198%
Inception v3	76.226%	61.856%	(0.126,0.146)	71.662%
		74.036%	(0.150,0.215)	75.354%
		74.766%	(0.186,0.238)	75.870%
DenseNet	74.472%	65.526%	(0.126,0.132)	70.730%
		70.716%	(0.155,0.154)	73.114%
		73.382%	(0.226,0.241)	74.116%

Trade-off analysis of the proposed method without and with offset for three thresholding values τ = {10, 2, 0.9} · 10−5 . The compression ratio refers to the method without offset.

Scaling factors distribution

- Scaling factor for weights and activations varies with different NN models, no strong link with the compatibility of models
- Refined factors remains close to 1 with small vibration



Mean and standard deviation of the set of scaling factors of the weights (α), the activations (β) on the calibration set (not dual) and the refined factors (γ), all normalized by the maximum value of the tensor they approximate

Conclusion

- Presented algorithm for quantization of NN (both weights and activations) achieves **state of the art results for INT4 quantization**.
- Quantized NN can be easily deployed on **deep learning dedicated hardware.**
- Accuracy of quantized NN can be improved according to a trade-off with NN compression.
- Only a **small unlabeled calibration set** is used.
- **No re-training** process is performed.
- Method works well even for **non over-parameterized architectures** (e.g squeezenet or densenet) which are considered hard to quantize properly.