Paper ID : 511
Poster ID : 61

McQueen : Mixed Precision Quantization of Early Exit Networks

Utkarsh Saxena and Kaushik Roy (Purdue University)

Elmore Family School of Electrical and Computer Engineering

MOTIVATION

Goal

Improve efficiency of deep neural networks (DNNs) by leveraging mixed precision quantization and dynamic execution via early exit

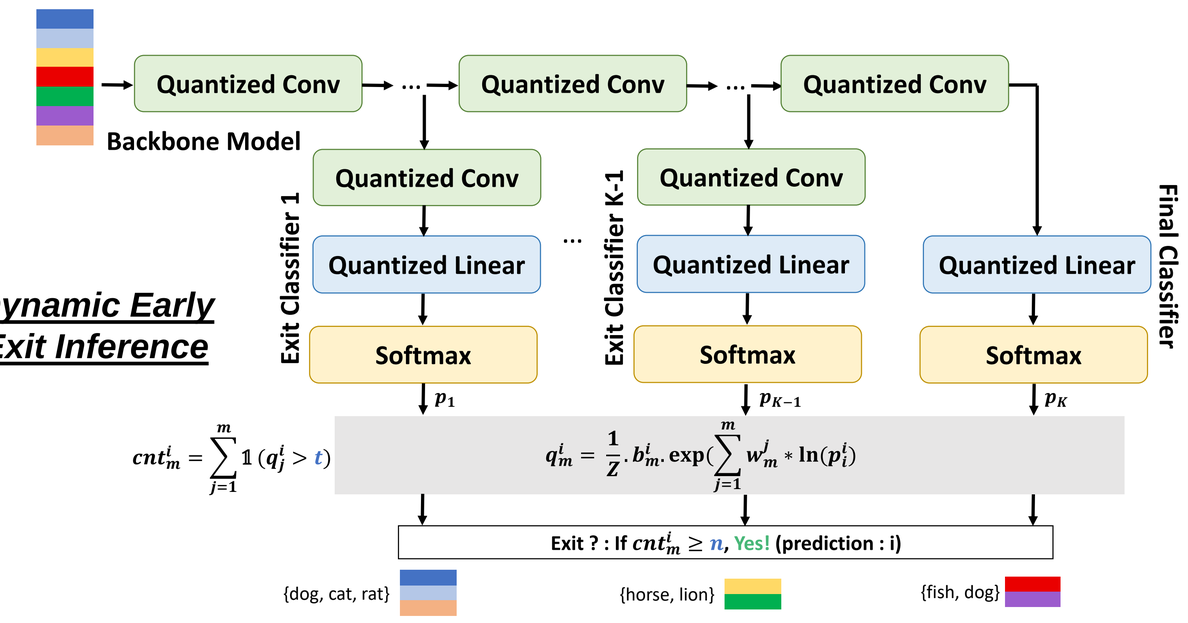
Challenges

Precision Selection : Assigning different precision to layer-wise weights and activations of multi-exit model presents a huge design space.

Training multi-exit model : Naïve training of quantized multi-exit model leads to considerable drop in workload accuracy.

BACKGROUND: EARLY EXIT NETWORKS

- Backbone model is augmented with shallow exit classifiers to terminate classification of "easy" samples early saving computational effort.



MCQUEEN FRAMEWORK

- Training multi-exit model leads to a considerable drop in final classifier accuracy.

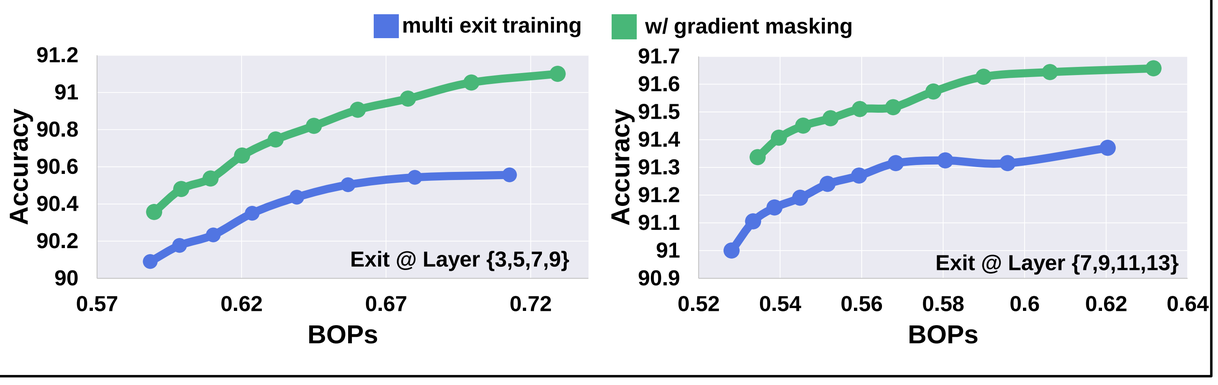
Model/Dataset	Precision	w/o EE	w/ EE
ResNet-20/CIFAR-10	2/2	89.3	88.9
ResNet-18/ImageNet	2/2	67.6	66.4
ResNet-18/ImageNet	$3_{MP}/3_{MP}$	69.8	69.1
ResNet-18/ImageNet	8/8	69.7	69.5

- Accuracy drop attributed to high gradient interference (low similarity) between gradients from exit classifiers (g_{exit}) and gradients from final classifier (g_{final}). **Gradient masking** is proposed to mitigate this:

$$g_{layer} = g_{final} + mask \odot g_{exit},$$

$$mask = \begin{cases} 1, & sign(g_{exit}) = sign(g_{final}) \\ 0, & otherwise \end{cases}$$

- Gradient Masking improves on inference performance when compared with naïve multi-exit training.



- **Parametric Differentiable Quantizer** learns weight/activation quantizer precision during training.

$$x_q = \alpha \cdot clip\left(round\left(\frac{x}{\beta}\right), Q_n, Q_p\right), \qquad Q_n = -2^{n-1}, Q_p = 2^{n-1} - 1$$

- We introduce following gradient to learn quantizer precision :

$$\frac{\partial x_q}{\partial n} = 2^{n-1} \ln(2) \cdot \left\{\frac{\partial x_q}{\partial Q_p} - \frac{\partial x_q}{\partial Q_n}\right\}$$

$$\frac{\partial x_q}{\partial Q_p} = \begin{cases} \alpha, & \frac{x}{\beta} \geq Q_p \\ 0, & otherwise \end{cases} \qquad \frac{\partial x_q}{\partial Q_n} = \begin{cases} \alpha, & \frac{x}{\beta} \leq Q_n \\ 0, & otherwise \end{cases}$$

- To constrain precision to low values, we add a regularization penalty to the loss function based on bit-wise operations (BOPs) of the multi-exit model. Total loss is :

$$Loss = \sum_{k=1}^{K} l_{CE}^k + \gamma \cdot \left| \sum_{l=1}^{L} bop_l - bop_{target} \right|$$

RESULTS ON IMAGENET

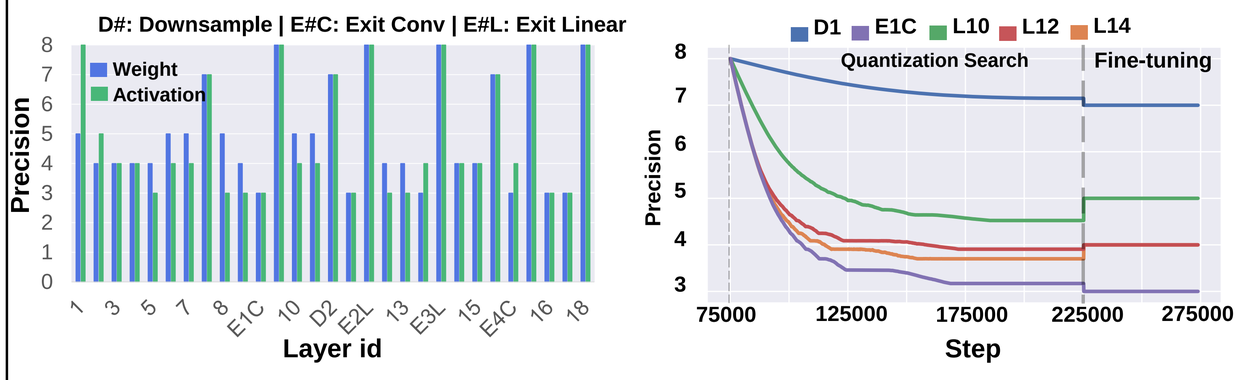
Comparison with homogenous quantization.

Method	Precision	top-1	Delta	BOPs	FP top-1
DoReFa [38]	2/2	64.7	-5.0	14.36	69.7
PACT [4]	2/2	64.4	-5.8	14.36	70.2
LSQ [7]	2/2	67.6	-2.9	14.36	70.5
N2UQ [22]	2/2	69.4	-2.4	14.36	71.8
McQueen	**2/2**	**67.4**	**-2.3**	**9.38**	**69.7**
DoReFa [38]	3/3	67.5	-2.2	22.84	69.7
PACT [4]	3/3	69.2	-1.0	22.84	70.2
LSQ [7]	3/3	70.2	-0.3	22.84	70.5
N2UQ [22]	3/3	70.0	0.1	22.84	71.8
McQueen	**3/3**	**70.0**	**0.3**	**17.0**	**69.7**

Comparison with mixed precision quantization.

Method	Precision	top-1	Delta	BOPs	FP top-1
SPOS [9]	$3_{MP}/3_{MP}$	69.4	-1.5	21.92	70.9
FracBits [36]	$3_{MP}/3_{MP}$	69.4	-0.8	22.93	70.2
LLI [28]	$3_{MP}/3_{MP}$	69.0	-0.6	23.02	69.6
DQ-Net [23]	$3_{MP}/3_{MP}$	69.8	0.0	27.18	69.8
McQueen	$\mathbf{3_{MP}/3_{MP}}$	**70.0**	**0.3**	**23.15**	**69.7**
SPOS [9]	$4_{MP}/4_{MP}$	70.5	-0.4	31.81	70.9
FracBits [36]	$4_{MP}/4_{MP}$	70.6	0.4	34.7	70.2
LLI [28]	$4_{MP}/4_{MP}$	70.1	0.5	33.05	69.6
DQ-Net [23]	$4_{MP}/4_{MP}$	70.4	0.6	42.49	69.8
McQueen	$\mathbf{4_{MP}/4_{MP}}$	**70.8**	**1.0**	**32.3**	**69.7**

Learned Precisions



Key Insight : Combining dynamic execution with quantization can enhance gains in efficiency.

ACKNOWLEDGEMENT

This work is supported by C-BRIC, Semiconductor Research Corporation (SRC),DARPA, DoE, DARPA AIE, NSF, and DoD