DBQ: A Differentiable Branch Quantizer for Lightweight Deep Neural Networks

***Hassan Dbouk**[1,2], Hetul Sanghvi[2], Mahesh Mehendale[2], and Naresh Shanbhag[1]*

[1]Dept. of Electrical and Computer Engineering, University of Illinois at Urbana Champaign

[2]Kilby Labs, Texas Instruments Inc.

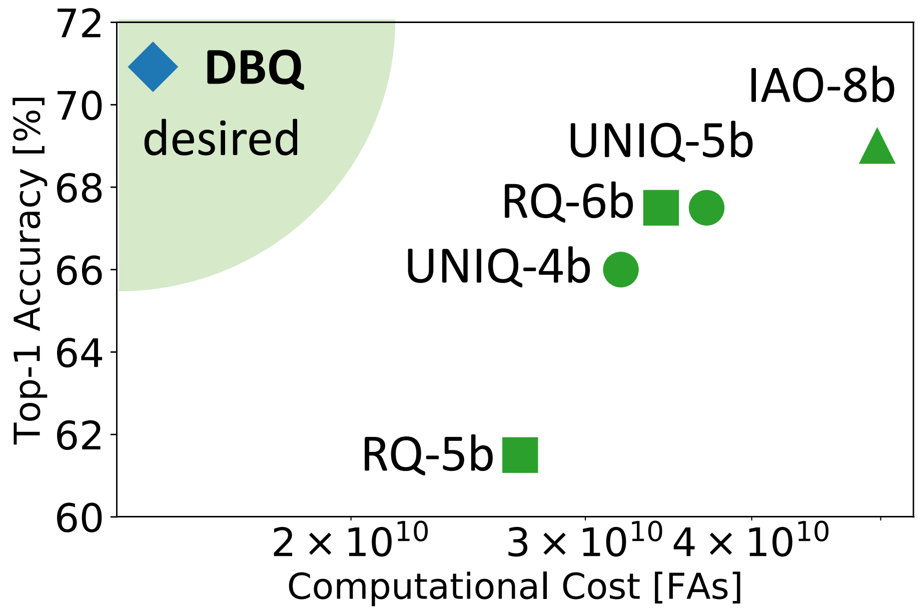
TEXAS INSTRUMENTS

ECCV'20 ONLINE 23-28 AUGUST 2020

ILLINOIS
Electrical & Computer Engineering
COLLEGE OF ENGINEERING

Motivation

- The complexity of DNNs inhibits their deployment on resource-constrained devices
- Current quantization methods offer *conservative* complexity reduction for *lightweight* networks:
- A ternarized MobileNetV1 incurs a massive (6%) accuracy drop

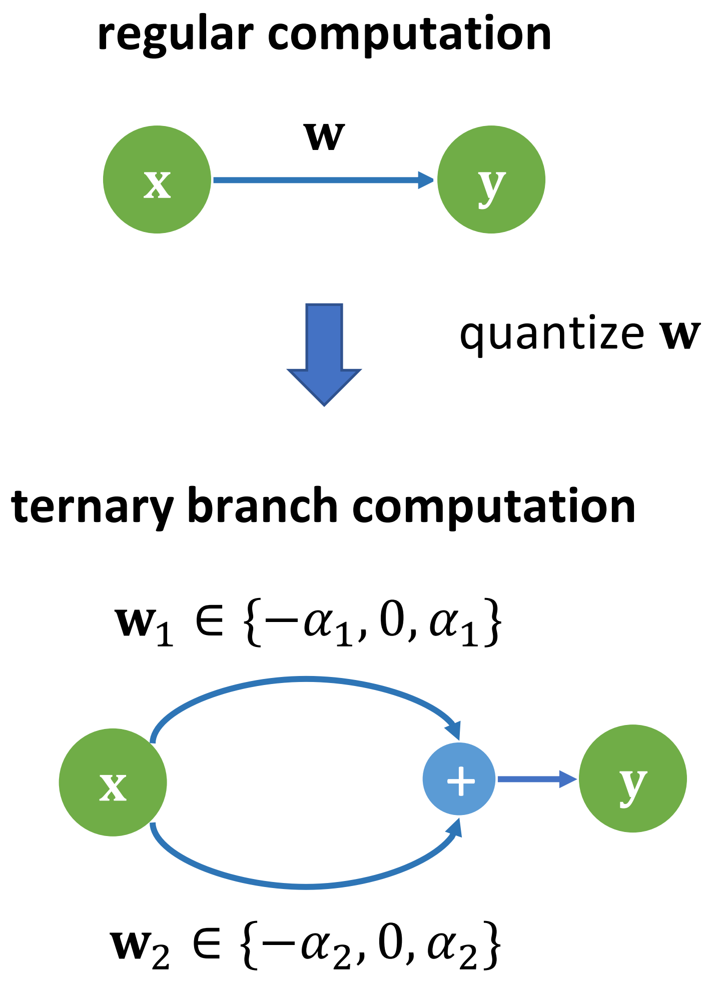


Goal: aggressively quantize lightweight networks while maintaining accuracy

Ternary Branch Quantization

- Quantizing parameters to two ternary branches:
  - utilizes efficient ternary arithmetic
  - offers a 9-level non-uniform quantizer

- How to train networks with such a structure **efficiently**?



Differentiable Branch Quantizer (DBQ)

- Formulate a B-branch ternary quantizer as a non-uniform quantizer with $N = 3^B$ levels:

$$Q(\mathbf{w}) = \gamma_2 \left[\sum_{i=1}^{N-1} \left[f(\gamma_1 \mathbf{w} - t_i) \sum_{j=1}^{B} b_{i,j} \alpha_j \right] - \sum_{j=1}^{B} \alpha_j \right]$$

Where:

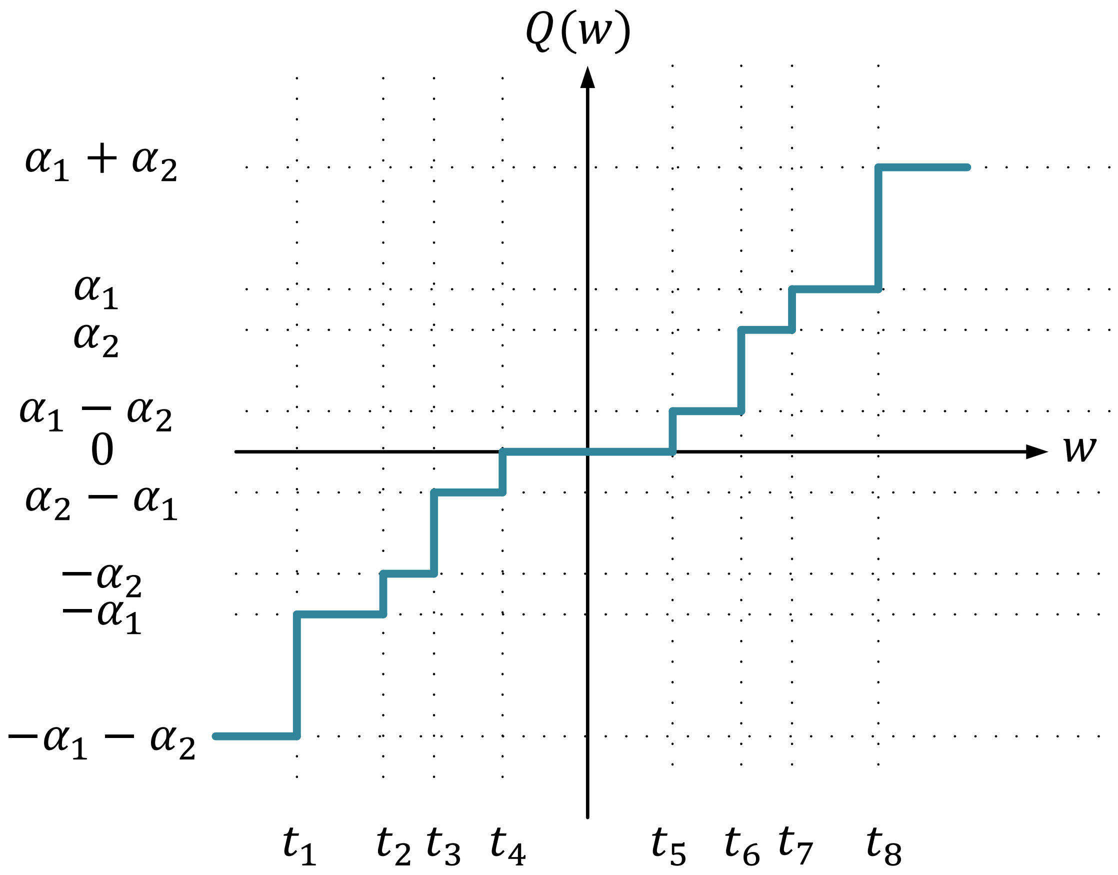
- f is an ideal step function
- $\{\alpha_j\}_{j=1}^{B}$ are the branch scales
- $\{t_i\}_{i=1}^{N-1}$ are the quantizer thresholds
- γ_1& γ_2 are pre/post quantization scales

- The ternary structure is enforced by the choice of $b_{i,j}$'s

Example: Two Ternary Branches

- A two-ternary branch quantizer can be written as:

$$
\begin{aligned}
Q(\mathbf{w}) = \gamma_2 \Big[& \alpha_2 f(\gamma_1 \mathbf{w} - t_1) + (\alpha_1 - \alpha_2) f(\gamma_1 \mathbf{w} - t_2) \\
& + (2\alpha_2 - \alpha_1) f(\gamma_1 \mathbf{w} - t_3) + (\alpha_1 - \alpha_2) f(\gamma_1 \mathbf{w} - t_4) \\
& + (\alpha_1 - \alpha_2) f(\gamma_1 \mathbf{w} - t_5) + (2\alpha_2 - \alpha_1) f(\gamma_1 \mathbf{w} - t_6) \\
& + (\alpha_1 - \alpha_2) f(\gamma_1 \mathbf{w} - t_7) + \alpha_2 f(\gamma_1 \mathbf{w} - t_8) \\
& - (\alpha_1 + \alpha_2) \Big]
\end{aligned}
$$



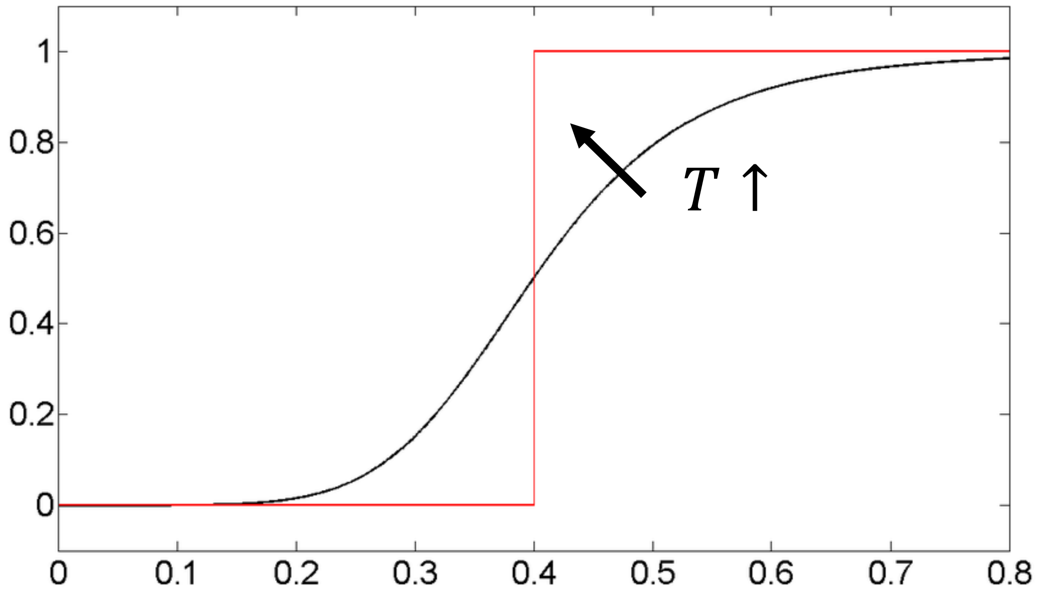
Differentiability in DBQ

- The non-differentiability of the quantizer comes from the step function $f()$
- **Solution**: use an approximate smooth function $\hat{f}_T()$ (e.g. Sigmoid) with a temperature parameter T that controls the approximation error:

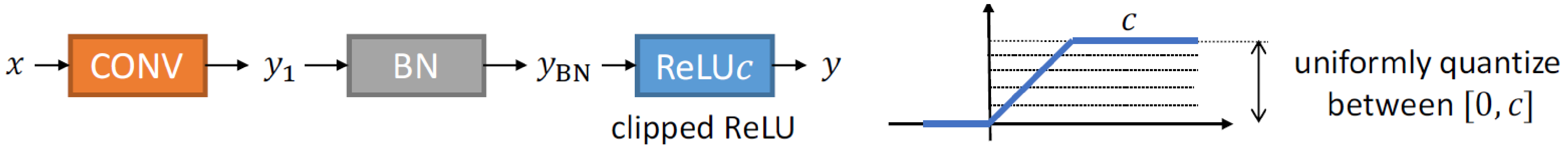
$$e_T(u) = \hat{f}_T(u) - f(u) \xrightarrow[T \to \infty]{} 0$$

$f()$: step function

$\hat{f}_T()$: temperature controlled Sigmoid



Activation Quantization



- Quantizing activations requires an appropriate clipping value c
- Leverage activation statistics offered by BatchNorm (BN) layers to choose c:

$$c = \max_{i \in [C]} \beta_i + k\gamma_i$$

β_i and γ_i: the per-channel BN shift and scale parameters; k controls clipping probability

Complexity Metrics

- **Computational Cost ($\mathcal{C}_C$):** captures the number of 1-b full adders (FA) needed to implement the dot-products required for a single inference
- **Sparsity-Aware Computational Cost ($\mathcal{C}_S$):** analogous to $\mathcal{C}_C$, defined in order to leverage weight-sparsity in different models that can be reflected on the model complexity
- **Representational Cost ($\mathcal{C}_R$):** measures the number of bits needed to represent the entire network (both weights and activations) for a single inference
- **Model Storage Cost ($\mathcal{C}_M$):** analogous to $\mathcal{C}_R$, but only accounts for the weight storage as it is useful for studying model compression

CIFAR10 Results: ResNet-20

Compared to a binary branch quantizer for ResNet-20:

- DBQ achieves higher accuracy with lower complexity at iso-number of branches (by exploiting weight sparsity)
- DBQ-2T achieves a 56% reduction in $\mathcal{C}_S$, at iso-accuracy

Method	Acc. (Δ) [%]	$\mathcal{C}_C$ ($\mathcal{C}_S$) [10^9FA]	$\mathcal{C}_R$ ($\mathcal{C}_M$) [10^6b]
FP [30]	92.10 (/)	23.73 (23.73)	14.63 (8.63)
LQNet-1B [30]	90.10 (−2.171)	1.60 (1.60)	6.34 (0.35)
LQNet-2B [30]	91.80 (−0.325)	2.83 (2.83)	6.61 (0.61)
LQNet-3B [30]	92.00 (−0.108)	4.07 (4.07)	6.88 (0.88)
FP (Ours)	92.00 (/)	23.73 (23.73)	14.63 (8.63)
DBQ-1T (Ours)	**91.06 (−1.021)**	**1.60 (0.92)**	6.61 (0.61)
DBQ-2T (Ours)	**91.93 (−0.076)**	**2.83 (1.79)**	7.15 (1.15)

Ablation Study: MobileNetV1 on ImageNet

- DBQ-1T (one ternary branch) achieves a massive reduction in $\mathcal{C}_C$ compared to FP but at a catastrophic loss of 5.67% in accuracy
- DBQ-2T-1 (two ternary branches) recovers the accuracy to within 1.03% of FP while also achieving massive savings in $\mathcal{C}_C$ of 84%

Model Name	Activations	FL	DW	PW	FC	Top-1/5 Acc. [%]	$\mathcal{C}_C$ ($\mathcal{C}_S$) [10^{10}FA]	$\mathcal{C}_R$ ($\mathcal{C}_M$) [10^7b]
FP	ReLU - 32b	32b	32b	32b	32b	**72.12/90.43**	33.37 (33.37)	30.00 (13.54)
FX8-1	ReLU6 - 8b	32b	8b	8b	32b	71.65/90.17	5.78 (5.39)	10.38 (5.90)
FX8-2	ReLU6 - 8b	8b	8b	8b	8b	71.60/90.19	5.24 (4.85)	7.56 (3.44)
FX8-3	ReLUx - 8b	8b	8b	8b	8b	**71.86/90.26**	5.24 (4.85)	7.56 (3.44)
DBQ-1T	ReLU - 32b	32b	32b	1T	32b	66.45/86.72	3.60 (2.61)	20.58 (4.12)
DBQ-2T-1	ReLU - 32b	32b	32b	2T	32b	71.09/89.71	5.23 (3.77)	21.21 (4.75)
DBQ-2T-2	ReLU6 - 8b	32b	8b	2T	32b	70.25/89.42	2.73 (1.97)	9.12 (4.64)
DBQ-2T-3	ReLUx - 8b	32b	8b	2T	32b	70.80/89.75	2.73 (1.97)	9.12 (4.64)
DBQ-2T-4	ReLUx - 8b	8b	8b	2T	8b	**70.92/89.61**	**2.18** (**1.42**)	**6.30** (**2.18**)

Ablation Study: MobileNetV1 on ImageNet

- The Top-1 accuracy of **FX8-3** (BN-based clipping) is better than **FX8-2** (ReLU6-based clipping) without any overhead in training or inference
- Similarly for **DBQ-2T-3** and **DBQ-2T-2**

Model Name	Activations	FL	DW	PW	FC	Top-1/5 Acc. [%]	$\mathcal{C}_C$ ($\mathcal{C}_S$) [10^{10}FA]	$\mathcal{C}_R$ ($\mathcal{C}_M$) [10^7b]
FP	ReLU - 32b	32b	32b	32b	32b	**72.12/90.43**	33.37 (33.37)	30.00 (13.54)
FX8-1	ReLU6 - 8b	32b	8b	8b	32b	71.65/90.17	5.78 (5.39)	10.38 (5.90)
FX8-2	ReLU6 - 8b	8b	8b	8b	8b	71.60/90.19	5.24 (4.85)	7.56 (3.44)
FX8-3	ReLUx - 8b	8b	8b	8b	8b	**71.86/90.26**	5.24 (4.85)	7.56 (3.44)
DBQ-1T	ReLU - 32b	32b	32b	1T	32b	66.45/86.72	3.60 (2.61)	20.58 (4.12)
DBQ-2T-1	ReLU - 32b	32b	32b	2T	32b	71.09/89.71	5.23 (3.77)	21.21 (4.75)
DBQ-2T-2	ReLU6 - 8b	32b	8b	2T	32b	70.25/89.42	2.73 (1.97)	9.12 (4.64)
DBQ-2T-3	ReLUx - 8b	32b	8b	2T	32b	70.80/89.75	2.73 (1.97)	9.12 (4.64)
DBQ-2T-4	ReLUx - 8b	8b	8b	2T	8b	**70.92/89.61**	**2.18** (**1.42**)	**6.30** (**2.18**)

Ablation Study: MobileNetV1 on ImageNet

- DBQ-2T-4, which is DBQ-2T-1 with the remaining layers quantized to 8b, incurs a minimal loss in accuracy (1.2%) compared to FP while also achieving even greater reduction in both $\mathcal{C}_C$ (93%) and $\mathcal{C}_R$(70%). The reduction in $\mathcal{C}_S$ increases to 96% when branch sparsity is exploited to skip computations.

Model Name	Activations	FL	DW	PW	FC	Top-1/5 Acc. [%]	$\mathcal{C}_C$ ($\mathcal{C}_S$) [10^{10}FA]	$\mathcal{C}_R$ ($\mathcal{C}_M$) [10^{7}b]
FP	ReLU - 32b	32b	32b	32b	32b	**72.12/90.43**	33.37 (33.37)	30.00 (13.54)
FX8-1	ReLU6 - 8b	32b	8b	8b	32b	71.65/90.17	5.78 (5.39)	10.38 (5.90)
FX8-2	ReLU6 - 8b	8b	8b	8b	8b	71.60/90.19	5.24 (4.85)	7.56 (3.44)
FX8-3	ReLUx - 8b	8b	8b	8b	8b	**71.86/90.26**	5.24 (4.85)	7.56 (3.44)
DBQ-1T	ReLU - 32b	32b	32b	1T	32b	66.45/86.72	3.60 (2.61)	20.58 (4.12)
DBQ-2T-1	ReLU - 32b	32b	32b	2T	32b	71.09/89.71	5.23 (3.77)	21.21 (4.75)
DBQ-2T-2	ReLU6 - 8b	32b	8b	2T	32b	70.25/89.42	2.73 (1.97)	9.12 (4.64)
DBQ-2T-3	ReLUx - 8b	32b	8b	2T	32b	70.80/89.75	2.73 (1.97)	9.12 (4.64)
DBQ-2T-4	ReLUx - 8b	8b	8b	2T	8b	**70.92/89.61**	**2.18** (**1.42**)	**6.30** (**2.18**)

ImageNet Results: MobileNetV1

- DBQ-2T achieves the lowest computational cost compared to previously published works, while achieving the highest Top-1 accuracy 70.92%

Method	Act.	FL	DW	PW	FC	Top-1 Acc. [%]	$\mathcal{C}_C$ ($\mathcal{C}_S$) [10^{10}FA]	$\mathcal{C}_R$ ($\mathcal{C}_M$) [10^7b]
IAO* [12]	8b	8b	8b	8b	8b	**69.00***	4.97 (/)	7.49 (3.37)
UNIQ [1]	8b	5b	5b	5b	5b	67.50	3.70 (/)	6.29 (2.18)
UNIQ [1]	8b	4b	4b	4b	4b	66.00	3.19 (/)	5.87 (**1.76**)
UNIQ [1]	8b	8b	8b	8b	8b	68.25	5.24 (/)	7.56 (3.44)
QSM* [27]	8b	8b	8b	8b	8b	68.03	4.97 (/)	7.49 (3.37)
RQ [19]	5b	5b	5b	5b	5b	61.50	**2.68** (/)	**4.75** (2.18)
RQ [19]	6b	6b	6b	6b	6b	67.50	3.42 (/)	5.69 (2.60)
HAQ cloud [28]	mixed	8b	mixed	mixed	8b	65.33 – 71.20†	2.73 (/)	5.09 (3.12)
HAQ edge [28]	mixed	8b	mixed	mixed	8b	67.40 – 71.20†	4.06 (/)	5.87 (2.49)
FP (Ours)	32b	32b	32b	32b	32b	**72.12**	33.37 (33.37)	30.00 (13.54)
FX8 (Ours)	8b	8b	8b	8b	8b	**71.86**	5.24 (4.85)	7.56 (3.44)
DBQ-2T (Ours)	8b	8b	8b	2T	8b	**70.92**	**2.18** (**1.42**)	6.30 (2.18)

*models with BN folding *results extracted from a plot †exact accuracy not reported

ImageNet Results: MobileNetV2 & ShuffleNetV2

- Inline with our experiments on MobileNetV1, we quantize all PW layers using 2T, with the remaining layers and activations quantized to 8b fixed-point.
- Observe a minimal 1.3% (MobileNetV2) and 2.6% (ShuffleNetV2) drop in accuracy compared to FP, while achieving massive (77% - 95%) reductions in all the complexity metrics.

Model	**Act.**	**FL**	**DW**	**PW**	**FC**	**Top-1 Acc.** [%]	$\mathcal{C}_C$ ($\mathcal{C}_S$) [10^{10}FA]	$\mathcal{C}_R$ ($\mathcal{C}_M$) [10^7b]
MobileNetV2-FP	32b	32b	32b	32b	32b	71.88	17.83 (17.83)	32.87 (11.22)
MobileNetV2-2T	8b	8b	8b	2T	8b	**70.54**	**1.42 (1.11)**	**7.45 (2.04)**
ShuffleNetV2-FP	32b	32b	32b	32b	32b	69.36	8.52 (8.52)	13.81 (7.29)
ShuffleNetV2-2T	8b	8b	8b	2T	8b	**66.74**	**0.64 (0.46)**	**3.21 (1.38)**

Accuracy-Precision-Complexity Trade-off: Dataset

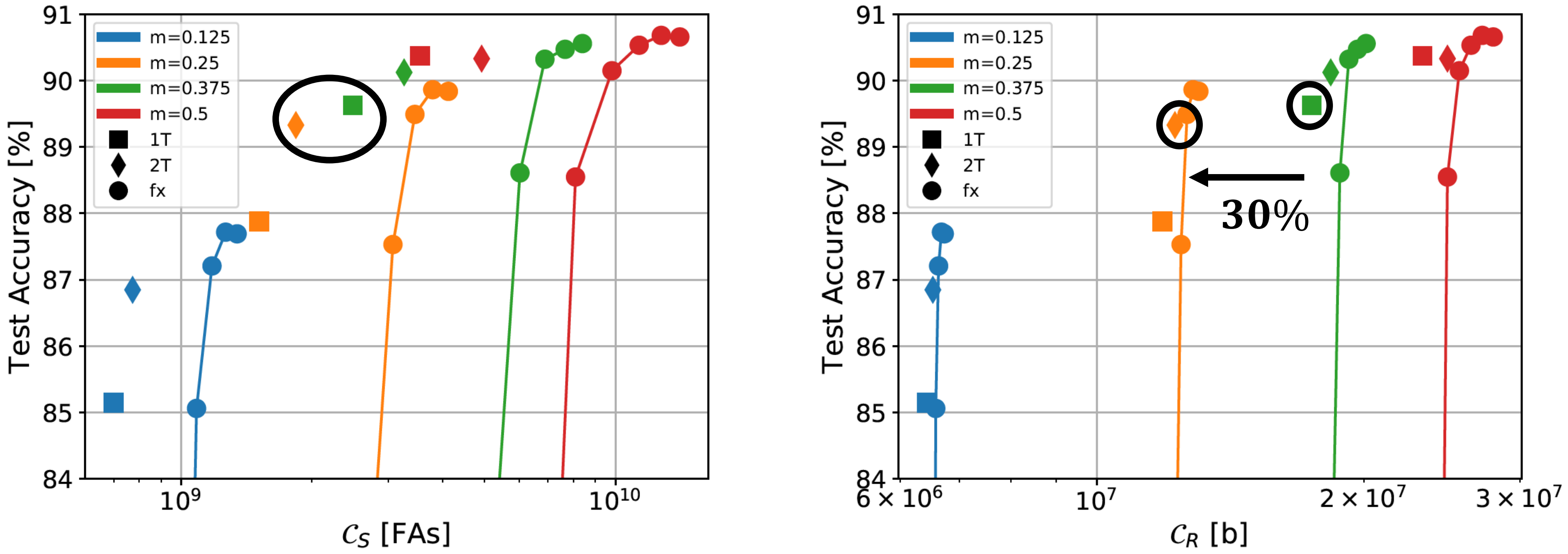
[Chowdhery, arXiv'19]

Google's Visual Wake Words (VWW) Dataset:

- Binary classification problem (person, no-person)
- Images taken from COCO'14 dataset
- Contains 115k training images and 8k validation images
- Reflects a real-life detection scenario for always-on resource-constrained Edge devices

Accuracy-Precision-Complexity Trade-off: Results

- MobileNetV1 complexity is varied via the width multiplier m which controls the number of channels
- DBQ models form a pareto-optimal curve
- For lightweight models: going from 1T to 2T is better than increasing m



Thank You!

paper link:
https://arxiv.org/abs/2007.09818