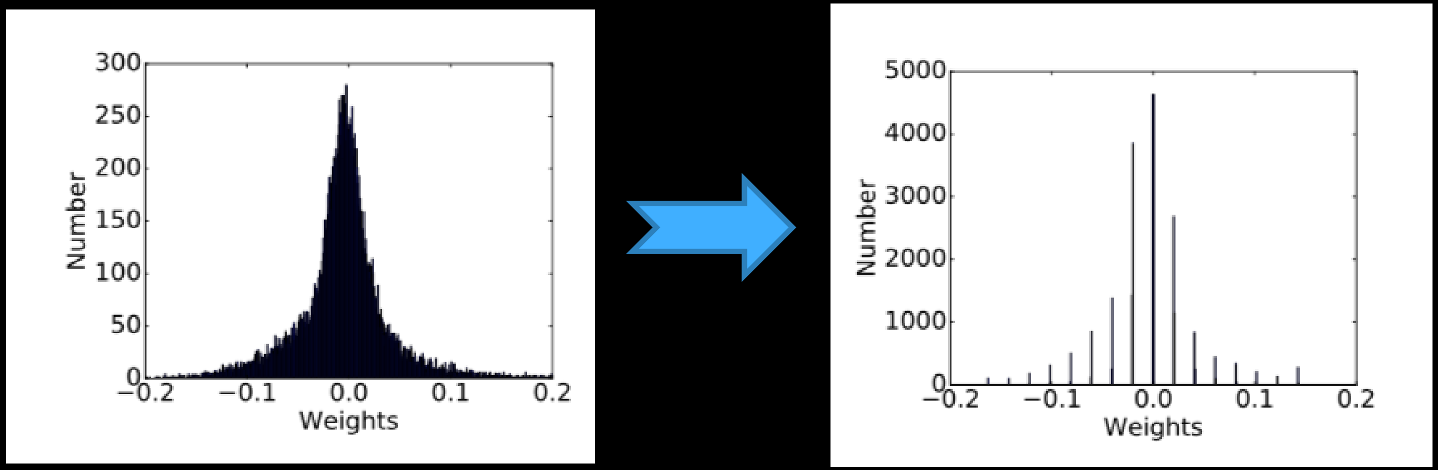
# Discovering Low-Precision Networks Close to Full-Precision Networks for Efficient Inference

Jeffrey L. McKinstry*, Steven K. Esser, Rathinakumar Appuswamy, Deepika Bablani, John V. Arthur, Izzet B. Yildiz & Dharmendra S. Modha
IBM Almaden Research Center, San Jose

# Network Quantization Problem

- Pretrained, high-precision networks must be prepared to run on low precision hardware for low cost/energy efficient inference (8 or 4-bits for weights and activations)
  - Quantize weights (uniform)
  - Quantize activations (uniform)
  - Finetune to improve score




- 8-bit networks: accuracy is typically lower then full precision scores when just quantizing (Migacz, 2017), or when training from scratch (Jacob et al., 2017)

- 4-bit networks: only 1 method had been shown to match full precision accuracy by combining several finetuning techniques (ResNet-50 net on imagenet) (Zhuang et al., 2018)

- Are complex training techniques required? Do 4-bits suffice for classification for other networks?

# Proposed solution

- Train model with low precision quantization in forward pass (Courbariaux et al, 2015)

- Hypothesis: noise due to quantization (Polino et al. 2018) hinders low precision training

- SGD requires

$$k \leq (\sigma^2 + L * \|x_0 - x^*\|_2^2)^2 / \epsilon^2$$

  iterations to find a $2\epsilon$-approximate optimal value, where $\sigma$ is the gradient noise level, $L$ relates to curvature, $x_0$ and $x^*$ are initial and optimal network parameters, $\epsilon$ is error tolerance

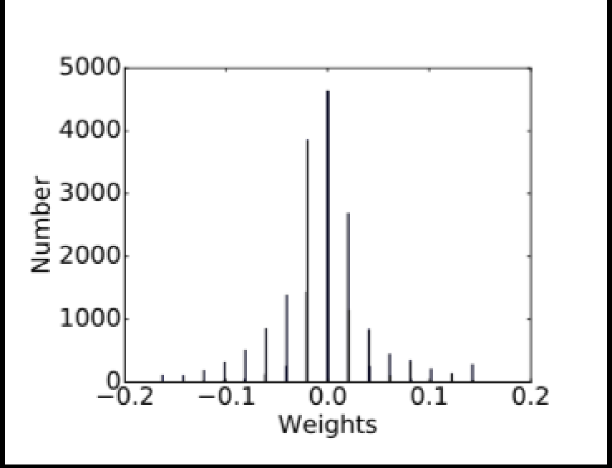
- Suggests that to overcome noise due to quantization:
  - Finetune to start closer to solution (Zhou et al., 2017)
  - Learning rate annealing to lower learning rates ($10^{-6}$) to average over more batches (Smith et al., 2017)
  - Finetune longer: 110 epochs to achieve better accuracy

- In addition, use empirically optimal quantization step size for both weights and activations that is a function of the precision.

- Finetuning after Quantization: FAQ

# FAQ: Methods

- Uniform quantization of weights and activations with quantization bin width, **Δ**, a power of 2 (fixed point representation)




- Weight quantization for 4-bits: $\Delta = \left[\frac{4.12\sigma^l}{8}\right[$

  where $[x[ = 2^{[log_2(x)]}$, and $\sigma^l$ is the standard deviation of the weights for layer *l*. Weights outside range are clipped during training.

- Activation quantization for 4-bits: $\Delta = [y_{max}/16[$

  where $[x[ = 2^{[log_2(x)]}$, and $y_{max}$ is the maximum 99.9$^{th}$ percentile of activations for layer *l* among 5 calibration batches from training set. Activations outside range are clipped during training.

- Training
  - Imagenet dataset
  - SGD with momentum (PyTorch)
  - 4 bit networks use 110 epochs, learning rate 0.0015 with an exponential decay to a final value of $10^{-6}$, 1$^{st}$ and last layers are 8-bits
  - Batch size: 256

# FAQ Results on Imagenet classification benchmark

- 8-bit nets exceed full-precision accuracy in 1 epoch on Imagenet
- 4-bit nets match original full precision net accuracy for a wide range of networks in PyTorch model zoo
- 4-bit solutions are close to the full-precision solution
  - Mean cosine similarity 0.994

| Network | Method | Precision (w,a) | Accuracy (% top-1) | Accuracy (% top-5) |
|---|---|---|---|---|
| ResNet-18 | baseline | 32,32 | 69.76 | 89.08 |
| **ResNet-18** | **Apprentice** | **4,8** | **70.40** | **-** |
| **ResNet-18** | **FAQ (This paper)** | **8,8** | **70.02** | **89.32** |
| **ResNet-18** | **FAQ (This paper)** | **4,4** | **69.78±0.04** | **89.11±0.03** |
| **ResNet-18** | Joint Training | 4,4 | 69.3 | - |
| ResNet-18 | UNIQ | 4,8 | 67.02 | - |
| ResNet-18 | Distillation | 4,32 | 64.20 | - |
| ResNet-34 | baseline | 32,32 | 73.30 | 91.42 |
| **ResNet-34** | **FAQ (This paper)** | **8,8** | **73.71** | **91.63** |
| **ResNet-34** | **FAQ (This paper)** | **4,4** | **73.31** | **91.32** |
| ResNet-34 | UNIQ | 4,32 | 73.1 | - |
| ResNet-34 | Apprentice | 4,8 | 73.1 | - |
| ResNet-34 | UNIQ | 4,8 | 71.09 | - |
| ResNet-50 | baseline | 32,32 | 76.15 | 92.87 |
| **ResNet-50** | **FAQ (This paper)** | **8,8** | **76.52** | **93.09** |
| **ResNet-50** | **FAQ (This paper)** | **4,4** | **76.27** | **92.89** |
| **ResNet-50** | **EL-Net** | **4,4** | **75.9** | **92.4** |
| ResNet-50 | IOA | 8,8 | 74.9 | - |
| ResNet-50 | Apprentice | 4,8 | 74.7 | - |
| ResNet-50 | UNIQ | 4,8 | 73.37 | - |
| ResNet-152 | baseline | 32,32 | 78.31 | 94.06 |
| **ResNet-152** | **FAQ (This paper)** | **4,4** | **78.64** | **94.12** |
| **ResNet-152** | **FAQ (This paper)** | **8,8** | **78.54** | **94.07** |
| Inception-v3 | baseline | 32,32 | 77.45 | 93.56 |
| **Inception-v3** | **FAQ (This paper)** | **8,8** | **77.60** | **93.59** |
| Inception-v3 | FAQ (This paper) | 4,4 | 77.33 | 93.59 |
| Inception-v3 | IOA | 8,8 | 74.2 | 92.2 |
| Densenet-161 | baseline | 32,32 | 77.65 | 93.80 |
| **Densenet-161** | **FAQ (This paper)** | **4,4** | **77.90** | **93.83** |
| **Densenet-161** | **FAQ (This paper)** | **8,8** | **77.84** | **93.91** |
| VGG-16bn | baseline | 32,32 | 73.36 | 91.50 |
| **VGG-16bn** | **FAQ (This paper)** | **4,4** | **73.87** | **91.67** |
| **VGG-16bn** | **FAQ (This paper)** | **8,8** | **73.66** | **91.56** |

# Conclusion

- Ablation study on ResNet-18 indicates that longer training, finetuning, and proper activation stepsize calibration were essential

| Epochs | Pre-trained | Batch size | Learning rate schedule | Weight decay | Activation calibration | Accuracy (% top-1) | Change |
|---|---|---|---|---|---|---|---|
| 110 | Yes | 256 | exp. | 0.00005 | Yes | 69.82 | - |
| **60** | Yes | **400** | exp. | 0.00005 | Yes | 69.40 | -0.22 * |
| 110 | **No** | 256 | exp. | 0.00005 | Yes | 69.24 | -0.58 * |
| 165* | Yes | **256-2048** | exp. | 0.00005 | Yes | 69.96 | +0.14 |
| 110 | Yes | 256 | **step** | 0.00005 | Yes | 69.90 | +0.08 |
| 110 | Yes | 256 | exp. | **0.0001** | Yes | 69.59 | -0.23 |
| 110 | Yes | 256 | exp. | 0.00005 | **No** | 69.19 | -0.63 * |

- Results provide empirical evidence that 4-bits suffice for classification – simply Finetune After Quantization (FAQ)

- We are hiring