Deep $k$-Means: Re-Training and Parameter Sharing with Harder Cluster Assignments for Compressing Deep Convolutions

Junru Wu$^1$ Yue Wang$^2$ Zhenyu Wu$^1$ Zhangyang Wang$^1$
Ashok Veeraraghavan$^2$ Yingyan Lin$^2$
Motivation

• Deploying CNNs on resource-constrained platforms
• Two important concerns: Model Size + Energy Efficiency
• They are often not aligned*, so need to consider both in implementation


Wu et al. Deep k-Means: Re-Training and Parameter Sharing with Harder Cluster Assignments for Compressing Deep Convolutions
1. Reshape into $W \in \mathbb{R}^{s \times N}$
   ($N = s \times c \times m$)

2. Apply $k$-Means on $W$, cluster $N$ samples into $K$ clusters

3. Reshape back into $W \in m \times \mathbb{R}^{s \times c \times s}$

Parameter Sharing via Row-wise $k$-Means

$W \in m \times \mathbb{R}^{s \times c \times s}$

$W \in m \times \mathbb{R}^{s \times c \times s}$

$W \in \mathbb{R}^{s \times N}$

$K$ clusters

$W \in \mathbb{R}^{s \times N}$

$W \in m \times \mathbb{R}^{s \times c \times s}$

$W \in m \times \mathbb{R}^{s \times c \times s}$
Parameter Sharing via Row-wise k-Means

• For a conv layer with $m$ filters each of size $s \times s \times c$

• Original Memory Consumption can be represented as:

  \[ \text{MEM}_{\text{org}} = s \times s \times c \times m + m \]

  Weights Bias

• Applying K-Means* to assign weights with K clusters, the memory consumption is reduced to:

  \[ \text{MEM}_{\text{comp}} = K \times s + \left( -\sum_{i=1}^{N} p_i \log_2 p_i \right) + m \]

  Weights Weight Assignment Indexes Bias

  \[ p_i : \text{occurrence probability of samples in the } i^{th} \text{ cluster.} \]

* Compressing deep convolutional networks using vector quantization, ICLR 2015

Wu et al. Deep k-Means: Re-Training and Parameter Sharing with Harder Cluster Assignments for Compressing Deep Convolutions
Filter Visualization on Wide ResNet

Pre-Trained Model

Compressed Model w/o Re-Training

Wu et al. Deep k-Means: Re-Training and Parameter Sharing with Harder Cluster Assignments for Compressing Deep Convolutions
Deep k-Means w/o Re-Training

<table>
<thead>
<tr>
<th>Model</th>
<th>Δ (%)</th>
<th>CR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Soft Weight-Sharing</td>
<td>-2.02</td>
<td>45</td>
</tr>
<tr>
<td>Deep k-Means WR</td>
<td>-16.02</td>
<td>45</td>
</tr>
<tr>
<td>Deep k-Means WR</td>
<td>-25.45</td>
<td>47</td>
</tr>
<tr>
<td>Deep k-Means WR</td>
<td>-45.08</td>
<td>50</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model</th>
<th>Δ⁺ %</th>
<th>Δ⁻ %</th>
<th>CR</th>
</tr>
</thead>
<tbody>
<tr>
<td>One-shot (Kim et al., 2015)</td>
<td>N/A</td>
<td>-0.24</td>
<td>1.28</td>
</tr>
<tr>
<td>Low-rank (Tai et al., 2015)</td>
<td>N/A</td>
<td>-0.42</td>
<td>2.84</td>
</tr>
<tr>
<td>Deep k-Means WR</td>
<td>-1.22</td>
<td>-0.65</td>
<td>1.5</td>
</tr>
<tr>
<td>Deep k-Means WR</td>
<td>-3.7</td>
<td>-2.46</td>
<td>2</td>
</tr>
<tr>
<td>Deep k-Means WR</td>
<td>-13.72</td>
<td>-10.05</td>
<td>3</td>
</tr>
<tr>
<td>Deep k-Means WR</td>
<td>-48.95</td>
<td>-48.82</td>
<td>4</td>
</tr>
</tbody>
</table>

- CR: Compression Ratio, same as defined in (Han et. al., 2015)
- Considerable Performance Drop!
- Design a re-training process that is more “suitable” for k-means?

Wu et al. Deep k-Means: Re-Training and Parameter Sharing with Harder Cluster Assignments for Compressing Deep Convolutions
k-Means Regularized Re-Training

• Spectrally Relaxation* of k-means ($W \in \mathbb{R}^{s \times N}$ denotes the sample matrix):

  1. Rewrite k-means objective: 
     $$\min_{W,F \in \mathcal{F}} \text{Tr}(W^T W) - \text{Tr}(F^T W^T W F),$$
     ($F \in \mathbb{R}^{N \times k}$: cluster index matrix with special structure)

  2. Since $W$ as given:
     $$\max_{F \in \mathcal{F}} \text{Tr}(F^T W^T W F)$$

  3. Relax the structure of $F$:
     $$\max_F \text{Tr}(F^T W^T W F), \text{ s.t. } F^T F = I$$

---

* H Zha, X He, C Ding, M Gu, HD Simon “Spectral relaxation for k-means clustering”, NIPS 2001

---

Wu et al. Deep k-Means: Re-Training and Parameter Sharing with Harder Cluster Assignments for Compressing Deep Convolutions
k-Means Regularized Re-Training

• Spectrally Relaxation of k-means ($W \in \mathbb{R}^{S \times N}$ denotes the sample matrix):

  • 1. Rewrite k-means objective:
    $$\min_{W;F \in F} Tr(W^T W) - Tr(F^T W^T W F),$$
    ($F \in \mathbb{R}^{N \times k}$: cluster index matrix with special structure)

  • 2. Since $W$ as given:
    $$\max_{F \in F} Tr(F^T W^T W F)$$
    - No longer true for $W$ as a variable during re-training!

  • 3. Relax the structure of $F$:
    $$\max_F Tr(F^T W^T W F), \text{ s.t. } F^T F = I$$
k-Means Regularized Re-Training

• Use k-means spectrally relaxation to design a new regularizer, that keeps weights $W$ “suitable” for k-means clustering

• Assume the original training objective: $E(W)$

• The new regularized re-training objective:

$$\min_{W,F} E(W) + \frac{\lambda}{2} [Tr(W^TW) - Tr(F^TF^TWF)],$$

$$s.t. F^TF = I$$
Filter Visualization on Wide ResNet

MMSE: 1.5e-08

Accuracy: 94.69%
Pre-Trained Model

Accuracy: 92.89%
Pre-Trained Model w/ Re-Training

Accuracy: 93.06%
Compressed Model w/ Re-Training

Wu et al. Deep k-Means: Re-Training and Parameter Sharing with Harder Cluster Assignments for Compressing Deep Convolutions
Deep k-Means w/ Re-Training

<table>
<thead>
<tr>
<th>Model</th>
<th>$\Delta$ (%)</th>
<th>CR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Soft Weight-Sharing</td>
<td>-2.02</td>
<td>45</td>
</tr>
<tr>
<td>Deep $k$-Means WR</td>
<td>-16.02</td>
<td>45</td>
</tr>
<tr>
<td>Deep $k$-Means WR</td>
<td>-25.45</td>
<td>47</td>
</tr>
<tr>
<td>Deep $k$-Means WR</td>
<td>-45.08</td>
<td>50</td>
</tr>
<tr>
<td>Deep $k$-Means WR</td>
<td>-1.63</td>
<td>45</td>
</tr>
<tr>
<td>Deep $k$-Means WR</td>
<td>-2.23</td>
<td>47</td>
</tr>
<tr>
<td>Deep $k$-Means WR</td>
<td>-4.49</td>
<td>50</td>
</tr>
</tbody>
</table>

Table 3. Compressing Wide ResNet in comparison to soft weight-sharing (Ullrich et al., 2017).

Wide ResNet

GoogLeNet

<table>
<thead>
<tr>
<th>Model</th>
<th>$\Delta^\dag$ %</th>
<th>$\Delta^\ddagger$ %</th>
<th>CR</th>
</tr>
</thead>
<tbody>
<tr>
<td>One-shot (Kim et al., 2015)</td>
<td>N/A</td>
<td>-0.24</td>
<td>1.28</td>
</tr>
<tr>
<td>Low-rank (Tai et al., 2015)</td>
<td>N/A</td>
<td>-0.42</td>
<td>2.84</td>
</tr>
<tr>
<td>Deep $k$-Means WR</td>
<td>-1.22</td>
<td>-0.65</td>
<td>1.5</td>
</tr>
<tr>
<td>Deep $k$-Means WR</td>
<td>-3.7</td>
<td>-2.46</td>
<td>2</td>
</tr>
<tr>
<td>Deep $k$-Means WR</td>
<td>-13.72</td>
<td>-10.05</td>
<td>3</td>
</tr>
<tr>
<td>Deep $k$-Means WR</td>
<td>-48.95</td>
<td>-48.82</td>
<td>4</td>
</tr>
<tr>
<td>Deep $k$-Means WR</td>
<td>-0.26</td>
<td>0.00</td>
<td>1.5</td>
</tr>
<tr>
<td>Deep $k$-Means WR</td>
<td>-0.17</td>
<td>+0.06</td>
<td>2</td>
</tr>
<tr>
<td>Deep $k$-Means WR</td>
<td>-0.36</td>
<td>+0.03</td>
<td>3</td>
</tr>
<tr>
<td>Deep $k$-Means WR</td>
<td>-1.95</td>
<td>-1.14</td>
<td>4</td>
</tr>
</tbody>
</table>

Table 4. Compressing GoogLeNet on ILSVRC12 ($^\dag$ and $^\ddagger$ are top-1 and top-5 accuracies respectively).

- **Minimum Performance Drop!**

Wu et al. Deep k-Means: Re-Training and Parameter Sharing with Harder Cluster Assignments for Compressing Deep Convolutions
More Experiments on CR

<table>
<thead>
<tr>
<th>Model</th>
<th>Δ (%)</th>
<th>CR</th>
</tr>
</thead>
<tbody>
<tr>
<td>TT-conv (naive)</td>
<td>-2.4</td>
<td>2.02</td>
</tr>
<tr>
<td>TT-conv (naive)</td>
<td>-3.1</td>
<td>2.90</td>
</tr>
<tr>
<td>TT-conv</td>
<td>-0.8</td>
<td>2.02</td>
</tr>
<tr>
<td>TT-conv</td>
<td>-1.5</td>
<td>2.53</td>
</tr>
<tr>
<td>TT-conv</td>
<td>-1.4</td>
<td>3.23</td>
</tr>
<tr>
<td>TT-conv</td>
<td>-2.0</td>
<td>4.02</td>
</tr>
<tr>
<td>Deep k-Means</td>
<td>+0.05</td>
<td>2</td>
</tr>
<tr>
<td>Deep k-Means</td>
<td>-0.04</td>
<td>4</td>
</tr>
</tbody>
</table>

Table 1. Compressing TT-conv-CNN in (Garipov et al., 2016).

<table>
<thead>
<tr>
<th>Model</th>
<th>Δ (%)</th>
<th>CR</th>
</tr>
</thead>
<tbody>
<tr>
<td>LRD</td>
<td>-8.32</td>
<td>16</td>
</tr>
<tr>
<td>HashedNet</td>
<td>-9.79</td>
<td>16</td>
</tr>
<tr>
<td>FreshNet</td>
<td>-6.51</td>
<td>16</td>
</tr>
<tr>
<td>Deep k-Means WR</td>
<td>-5.95</td>
<td>16</td>
</tr>
<tr>
<td>Deep k-Means</td>
<td>-1.30</td>
<td>16</td>
</tr>
</tbody>
</table>

Table 2. Compressing FreshNet-CNN in (Chen et al., 2016a).
Computational cost *

- Measure the computational resources needed to generate a single decision (1 bit full adders)

\[ DB_w B_x + (D - 1)(B_x + B_w + [\log_2 D] - 1) \]

- \( B_w \): weight precision
- \( B_x \): activation precision
- \( D \) is the dimensional of dot product.

*Charbel Sakr, Yongjune Kim, Naresh R. Shanbhag, "Analytical Guarantees on Numerical Precision of Deep Neural Networks" ICML, 2017

Wu et al. Deep k-Means: Re-Training and Parameter Sharing with Harder Cluster Assignments for Compressing Deep Convolutions
Weight/ Activation Representational Cost *

• Measure the storage complexity and communication costs associated with data movement

\[ N_w |W| B_w + N_x |\chi| B_x \]

• \( N_w, N_x \): total number of times weight/ activation is used for convolution
• \( |W| \): index sets of weights \( |\chi| \): index sets of activation
• \( B_w \): weight precision \( B_x \): activation precision

*Charbel Sakr, Yongjune Kim, Naresh R. Shanbhag, "Analytical Guarantees on Numerical Precision of Deep Neural Networks" ICML, 2017
Verification of Energy-Aware Metrics

• We verify our Energy-Aware Metrics with MIT energy estimation* tool whose results are extrapolated from actual hardware measurements.

\[ R^2 \text{ Coefficient:} \]
• AlexNet: 0.9931
• GoogLeNet_v1: 0.9675

Highly aligned!


Wu et al. Deep k-Means: Re-Training and Parameter Sharing with Harder Cluster Assignments for Compressing Deep Convolutions
Computational Resources Used in Project

High Performance Research Computing
A Resource for Research and Discovery

• Hardware Stack
  • Texas A&M HRPC Terra GPU Cluster
    • Intel Xeon E5-2680 v4 2.40GHz 14-core
    • NVIDIA Tesla K80 Accelerator

• Software Stack:
  • CUDA 8.0
  • PyTorch 0.3.1

Wu et al. Deep k-Means: Re-Training and Parameter Sharing with Harder Cluster Assignments for Compressing Deep Convolutions