Long-tailed Recognition by Routing Diverse Distribution-Aware Experts

Xudong Wang¹, Long Lian¹, Zhongqi Miao¹, Ziwei Liu² and Stella Yu¹

¹ UC Berkeley / ICSI
² Nanyang Technological University
Natural Data Are Often Long-tailed Distributed Over Semantic Classes


Long-tailed Recognition: Imbalance + Few-shot Learning

- Training set: long-tailed distribution
  - Many-shot: #samples > 100
  - Medium-shot: #samples < 100 & > 20
  - Few-shot: #samples < 20
- Testing set: balanced distribution
- Evaluation:
  - Overall testing set
  - Three splits based on class size
Previous Methods

Methods Overview

1. Instance-wise Balancing (current SOTA)
   - Up/Down sampling tail/head classes. (e.g., Decouple [ICLR 2020], BBN [CVPR 2020])

2. Weighted Loss
   - Assign larger/smaller weights to tail/head classes. (e.g., LDAM [NeurIPS 2019], CB-Loss [CVPR 2019])

3. Feature Enhancement
   - Use the memory enhanced feature learned from both head and tail classes. (e.g., OLTR [CVPR 2018])
Previous Methods

Methods Overview

1. Instance-wise Balancing (current SOTA)
   - Up/Down sampling tail/head classes. (e.g., Decouple [ICLR 2020], BBN [CVPR 2020])

2. Weighted Loss
   - Assign larger/smaller weights to tail/head classes. (e.g., LDAM [NeurIPS 2019], CB-Loss [CVPR 2019])

3. Feature Enhancement
   - Use the memory enhanced feature learned from both head and tail classes. (e.g., OLTR [CVPR 2018])

Caveats

- All these methods generally **gain accuracy on tail classes** at the cost of **performance loss on head classes**.
Previous Methods

Methods Overview

1. Instance-wise Balancing (current SOTA)
   - Up/Down sampling tail/head classes. (e.g., Decouple [ICLR 2020], BBN [CVPR 2020])

2. Weighted Loss
   - Assign larger/smaller weights to tail/head classes. (e.g., LDAM [NeurIPS 2019], CB-Loss [CVPR 2019])

3. Feature Enhancement
   - Use the memory enhanced feature learned from both head and tail classes. (e.g., OLTR [CVPR 2018])

Caveats

- All these methods generally gain accuracy on tail classes at the cost of performance loss on head classes.

In order to understand the cause of caveats, we decoupled the model error with bias-variance decomposition.
Bias-variance Decomposition with Respect to the Variation in Dataset D

$$\text{Error}(x; h) = \text{Bias}(h)^2 + \text{Variance}(h) + \text{irreducible error}. $$

How to Obtain Bias and Variance of Each Method?

Stage 1: Training D models on D data subsets

How to Obtain Bias and Variance of Each Method?

Stage 1: Training D models on D data subsets

Stage 2: Collect predictions

How to Obtain Bias and Variance of Each Method?

Stage 1: Training D models on D data subsets

Stage 2: Collect predictions

Stage 3: Calculate Bias/Variance

How to Obtain Bias and Variance of Each Method?

Few-shot Accuracy Gain at The Cost of Many-shot Drop
Bias Reduction Tends to Be Greater for Tail Classes
Variance Is Increased Throughout The Class Spectrum

Accuracy ↑

Bias ↓

Variance ↓
### Our Key Insights

<table>
<thead>
<tr>
<th></th>
<th>Head Classes</th>
<th></th>
<th>Tail Classes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Acc</td>
<td>Bias</td>
<td>Variance</td>
</tr>
<tr>
<td>Current SOTAs</td>
<td>Worse</td>
<td>Comparable</td>
<td>Worse</td>
</tr>
</tbody>
</table>

Why previous methods get worse accuracy on many-shot classes?  
*The increased variance leads to a worse bias-variance trade-off.*

How to further improve the performance on few-shot classes?  
*Obtaining the optimal bias-variance trade-off by further reducing variance *and* bias.*
Reducing Model Variance with Multi-expert Framework

Stage One: Jointly Optimize Diverse Distribution-aware Experts

- $f_{\theta}$
- $\psi_{\theta_1}$
- $\psi_{\theta_2}$
- $\psi_{\theta_n}$

Distribution-aware diversity loss $\mathcal{L}_D$ + Classification loss $\mathcal{L}_{\text{Classify}}$

- Trainable
- Frozen
Reducing Model Bias with Individual Loss

Using Individual Loss Instead of Collaborative Loss

Collaborative loss leads to correlated experts

Individual loss decorrelates experts
The distribution-aware diversity loss is proposed to penalize the inter-expert correlation, formulated as:

\[ \mathcal{L}_{i-D\text{-Diversify}} = -\frac{\lambda}{k-1} \sum_{j \neq i}^n \mathcal{D}_{KL}(\phi^i(x, \vec{T}), \phi^j(\bar{x}, \vec{T})) \]

KL divergence  Softmax with temperature
Total Loss for Stage One

\[ L^i_{\text{Total}} = L^i_{\text{Classify}}(\phi^i(\bar{x}), y) - \frac{\lambda}{n-1} \sum_{j \neq i}^n D_{KL}(\phi^j(\bar{x}, \bar{T}), \phi^j(\bar{x}, \bar{T})) \]

where \( i \) is the expert index, \( L^i_{\text{Classify}}(\ldots) \) can be LDAM loss, focal loss, etc., depending on the training mechanisms we choose.
Reducing the Computational Complexity with Routing Module

Stage Two: Routing Diverse Experts
Routing Loss of Expert Assignment

The expert assignment is optimized with the routing loss, a weighted variant of binary cross entropy loss:

\[
\mathcal{L}_{\text{Routing}} = -\omega_p y \log \left( \frac{1}{1 + e^{-y_{ea}}} \right) - \omega_n (1 - y) \log \left( 1 - \frac{1}{1 + e^{-y_{ea}}} \right)
\]
Method Overview

Stage One: Jointly Optimize Diverse Distribution-aware Experts

Stage Two: Routing Diverse Experts

Distribution-aware diversity loss $L_{D\text{-Diversify}}$ + Classification loss $L_{\text{Classify}}$

Accuracy ↑

LDAM or RIDE vs. CE

- Few
- Medium
- Many

RIDE vs. LDAM
RIDE Decreases Bias More Than Other Methods on Few-shot Classes

Accuracy ↑

Bias ↓
RIDE Reduces Variances Throughout the Class Spectrum

Accuracy ↑

Bias ↓

Variance ↓
## RIDE vs Current SOTAs

<table>
<thead>
<tr>
<th></th>
<th>Head Classes</th>
<th></th>
<th>Tail Classes</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Acc</td>
<td>Bias</td>
<td>Variance</td>
<td>Acc</td>
</tr>
<tr>
<td>Current SOTAs</td>
<td>Worse</td>
<td>Comparable</td>
<td>Worse</td>
<td>Better</td>
</tr>
<tr>
<td>RIDE</td>
<td>Better</td>
<td>Better</td>
<td>Better</td>
<td>Better</td>
</tr>
</tbody>
</table>

Better accuracy for all splits.
Better bias-variance trade-off for all splits.
CIFAR100-LT (100 Classes)

SOTA performance on few-shot classes with 5.8% improvements.

<table>
<thead>
<tr>
<th>Methods</th>
<th>MFlops</th>
<th>Acc. (%)</th>
<th>Many</th>
<th>Med</th>
<th>Few</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cross Entropy (CE) ‡</td>
<td>69.5 (1.0x)</td>
<td>38.3</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Cross Entropy (CE) †</td>
<td>69.5 (1.0x)</td>
<td>39.1</td>
<td>66.1</td>
<td>37.3</td>
<td>10.6</td>
</tr>
<tr>
<td>Focal Loss ‡ (Lin et al., 2017)</td>
<td>69.5 (1.0x)</td>
<td>38.4</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>OLTR † (Liu et al., 2019)</td>
<td>-</td>
<td>41.2</td>
<td>61.8</td>
<td>41.4</td>
<td>17.6</td>
</tr>
<tr>
<td>LDAM + DRW (Cao et al., 2019)</td>
<td>69.5 (1.0x)</td>
<td>42.0</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>LDAM + DRW † (Cao et al., 2019)</td>
<td>69.5 (1.0x)</td>
<td>42.0</td>
<td>61.5</td>
<td>41.7</td>
<td>20.2</td>
</tr>
<tr>
<td>BBN (Zhou et al., 2020)</td>
<td>74.3 (1.1x)</td>
<td>42.6</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>$\tau$-norm † (Kang et al., 2020)</td>
<td>69.5 (1.0x)</td>
<td>43.2</td>
<td>65.7</td>
<td>43.6</td>
<td>17.3</td>
</tr>
<tr>
<td>cRT † (Kang et al., 2020)</td>
<td>69.5 (1.0x)</td>
<td>43.3</td>
<td>64.0</td>
<td>44.8</td>
<td>18.1</td>
</tr>
<tr>
<td>M2m (Kim et al., 2020)</td>
<td>-</td>
<td>43.5</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>LFME (Xiang et al., 2020)</td>
<td>-</td>
<td>43.8</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>RIDE (2 experts)</td>
<td>64.8 (0.9x)</td>
<td>47.0 (+3.2)</td>
<td>67.9</td>
<td>48.4</td>
<td>21.8</td>
</tr>
<tr>
<td>RIDE (3 experts)</td>
<td>77.8 (1.1x)</td>
<td>48.0 (+4.2)</td>
<td>68.1</td>
<td>49.2</td>
<td>23.9</td>
</tr>
<tr>
<td>RIDE (4 experts)</td>
<td>91.9 (1.3x)</td>
<td>49.1 (+5.3)</td>
<td>69.3</td>
<td>49.3</td>
<td>26.0</td>
</tr>
</tbody>
</table>
## ImageNet-LT (1000 Classes)

Consistent improvements to various backbones by 6.9~7.7%

<table>
<thead>
<tr>
<th>Methods</th>
<th>ResNet-50 GFlops</th>
<th>ResNetXt-50 GFlops</th>
<th>ResNet-50 Acc. (%)</th>
<th>ResNetXt-50 Acc. (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cross Entropy (CE) †</td>
<td>4.11 (1.0x)</td>
<td>4.26 (1.0x)</td>
<td>41.6</td>
<td>44.4</td>
</tr>
<tr>
<td>OLTR † (Liu et al., 2019)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>46.3</td>
</tr>
<tr>
<td>NCM (Kang et al., 2020)</td>
<td>4.11 (1.0x)</td>
<td>4.26 (1.0x)</td>
<td>44.3</td>
<td>47.3</td>
</tr>
<tr>
<td>τ-norm (Kang et al., 2020)</td>
<td>4.11 (1.0x)</td>
<td>4.26 (1.0x)</td>
<td>46.7</td>
<td>49.4</td>
</tr>
<tr>
<td>cRT (Kang et al., 2020)</td>
<td>4.11 (1.0x)</td>
<td>4.26 (1.0x)</td>
<td>47.3</td>
<td>49.6</td>
</tr>
<tr>
<td>LWS (Kang et al., 2020)</td>
<td>4.11 (1.0x)</td>
<td>4.26 (1.0x)</td>
<td>47.7</td>
<td>49.9</td>
</tr>
<tr>
<td>RIDE (2 experts)</td>
<td>3.71 (0.9x)</td>
<td>3.92 (0.9x)</td>
<td>54.4 (+6.7)</td>
<td>55.9 (+6.0)</td>
</tr>
<tr>
<td>RIDE (3 experts)</td>
<td>4.36 (1.1x)</td>
<td>4.69 (1.1x)</td>
<td>54.9 (+7.2)</td>
<td>56.4 (+6.5)</td>
</tr>
<tr>
<td>RIDE (4 experts)</td>
<td>5.15 (1.3x)</td>
<td>5.19 (1.2x)</td>
<td><strong>55.4 (+7.7)</strong></td>
<td><strong>56.8 (+6.9)</strong></td>
</tr>
</tbody>
</table>
iNaturalist (8000 Classes)

Significantly better performance on many-shot than current SOTA BBN.

<table>
<thead>
<tr>
<th>Methods</th>
<th>GFlops</th>
<th>All</th>
<th>Many</th>
<th>Medium</th>
<th>Few</th>
</tr>
</thead>
<tbody>
<tr>
<td>CE †</td>
<td>4.14 (1.0x)</td>
<td>61.7</td>
<td>72.2</td>
<td>63.0</td>
<td>57.2</td>
</tr>
<tr>
<td>CB-Focal †</td>
<td>4.14 (1.0x)</td>
<td>61.1</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>OLTR</td>
<td>4.14 (1.0x)</td>
<td>63.9</td>
<td>59.0</td>
<td>64.1</td>
<td>64.9</td>
</tr>
<tr>
<td>LDAM + DRW †</td>
<td>4.14 (1.0x)</td>
<td>64.6</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>cRT</td>
<td>4.14 (1.0x)</td>
<td>65.2</td>
<td>69.0</td>
<td>66.0</td>
<td>63.2</td>
</tr>
<tr>
<td>τ-norm</td>
<td>4.14 (1.0x)</td>
<td>65.6</td>
<td>65.6</td>
<td>65.3</td>
<td>65.9</td>
</tr>
<tr>
<td>LWS</td>
<td>4.14 (1.0x)</td>
<td>65.9</td>
<td>65.0</td>
<td>66.3</td>
<td>65.5</td>
</tr>
<tr>
<td>BBN</td>
<td>4.36 (1.1x)</td>
<td>66.3</td>
<td>49.4</td>
<td>70.8</td>
<td>65.3</td>
</tr>
<tr>
<td>RIDE (2 experts)</td>
<td>3.67 (0.9x)</td>
<td>71.4 (+5.1)</td>
<td>70.2 (+1.2)</td>
<td>71.3 (+0.5)</td>
<td>71.7 (+5.8)</td>
</tr>
<tr>
<td>RIDE (3 experts)</td>
<td>4.17 (1.0x)</td>
<td>72.2 (+5.9)</td>
<td>70.2 (+1.2)</td>
<td>72.2 (+1.4)</td>
<td>72.7 (+6.8)</td>
</tr>
<tr>
<td>RIDE (4 experts)</td>
<td>4.51 (1.1x)</td>
<td>72.6 (+6.3)</td>
<td><strong>70.9 (+1.9)</strong></td>
<td>72.4 (+1.6)</td>
<td><strong>73.1 (+7.2)</strong></td>
</tr>
</tbody>
</table>
iNaturalist (8000 Classes)

SOTA performance on iNaturalist with the largest improvements from few-shot classes.

<table>
<thead>
<tr>
<th>Methods</th>
<th>GFlops</th>
<th>All</th>
<th>Many</th>
<th>Medium</th>
<th>Few</th>
</tr>
</thead>
<tbody>
<tr>
<td>CE †</td>
<td>4.14 (1.0x)</td>
<td>61.7</td>
<td>72.2</td>
<td>63.0</td>
<td>57.2</td>
</tr>
<tr>
<td>CB-Focal †</td>
<td>4.14 (1.0x)</td>
<td>61.1</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>OLTR</td>
<td>4.14 (1.0x)</td>
<td>63.9</td>
<td>59.0</td>
<td>64.1</td>
<td>64.9</td>
</tr>
<tr>
<td>LDAM + DRW †</td>
<td>4.14 (1.0x)</td>
<td>64.6</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>cRT</td>
<td>4.14 (1.0x)</td>
<td>65.2</td>
<td>69.0</td>
<td>66.0</td>
<td>63.2</td>
</tr>
<tr>
<td>τ-norm</td>
<td>4.14 (1.0x)</td>
<td>65.6</td>
<td>65.6</td>
<td>65.3</td>
<td>65.9</td>
</tr>
<tr>
<td>LWS</td>
<td>4.14 (1.0x)</td>
<td>65.9</td>
<td>65.0</td>
<td>66.3</td>
<td>65.5</td>
</tr>
<tr>
<td>BBN</td>
<td>4.36 (1.1x)</td>
<td>66.3</td>
<td>49.4</td>
<td>70.8</td>
<td>65.3</td>
</tr>
<tr>
<td>RIDE (2 experts)</td>
<td><strong>3.67 (0.9x)</strong></td>
<td><strong>71.4 (+5.1)</strong></td>
<td><strong>70.2 (+1.2)</strong></td>
<td><strong>71.3 (+0.5)</strong></td>
<td><strong>71.7 (+5.8)</strong></td>
</tr>
<tr>
<td>RIDE (3 experts)</td>
<td>4.17 (1.0x)</td>
<td>72.2 (+5.9)</td>
<td>70.2 (+1.2)</td>
<td>72.2 (+1.4)</td>
<td>72.7 (+6.8)</td>
</tr>
<tr>
<td>RIDE (4 experts)</td>
<td>4.51 (1.1x)</td>
<td><strong>72.6 (+6.3)</strong></td>
<td><strong>70.9 (+1.9)</strong></td>
<td><strong>72.4 (+1.6)</strong></td>
<td><strong>73.1 (+7.2)</strong></td>
</tr>
</tbody>
</table>
RIDE is a Universal Framework

Consistent improvements to various methods can be obtained
Expert Assignment: Tail Classes Require More Experts

More than half samples in few-shot require more than one expert
More than half samples in many-shot only require one expert
RIDE is the first paper to theoretically analyze the long tail problem from the perspective of bias-variance decomposition.

RIDE is the first paper that increases the performances on all three splits (many-/med-/few-shot).

RIDE significantly outperforms current state-of-the-arts on all experimented benchmarks by 5%~8%, including CIFAR100-LT, ImageNet-LT and iNaturalist.

RIDE is a universal framework that can be integrated with various existing methods, which provides a strong framework for future research in long-tailed recognition.
LONG-TAILED RECOGNITION BY ROUTING DIVERSE DISTRIBUTION-AWARE EXPERTS

Xudong Wang¹, Long Lian¹, Zhongqi Miao¹, Ziwei Liu², Stella X. Yu¹
¹UC Berkeley / ICSI, ²Nanyang Technological University
{xdwang, longlian, zhongqi.miao, stellayu}@berkeley.edu
ziwei.liu@ntu.edu.sg