Iterative Self Knowledge Distillation – From Pothole Classification to Fine-Grained and COVID Recognition
Kuan-Chuan Peng
Mitsubishi Electric Research Laboratories (MERL), Cambridge, MA, USA
kpeng@merl.com

Motivation, Constraints & Task

**Motivation:** ~$3 billion annual vehicle repair bills related to pothole damage in the U.S.

**Constraints:** limited computational power (e.g., no GPU) on edge devices installed on the road inspection vehicles; need fast inference speed.

**Task:** Given a fixed number of training epochs and a lightweight model to be trained, what can practitioners do to improve the pothole classification accuracy?

Contributions

- We propose Iterative Self Knowledge Distillation (ISKD), which outperforms the SOTA self KD methods from pothole classification (RDD, simplex, complex) to generic (CIFAR-10, CIFAR-100), fine-grained (Oxford 102 Flower, Oxford IIIT Pet, Caltech-UCSD Birds 200), and medical imaging classification (COVID-19 Radiography).
- We provide more evidence showing that a teacher model with accuracy lower than the baseline can still result in a student model outperforming the baseline.
- ISKD is flexible with respect to parameter selection.

Dataset Statistics & Quantitative Results

<table>
<thead>
<tr>
<th>dataset</th>
<th>RDD (13)</th>
<th>RDD (13)</th>
<th>CIFAR-10</th>
<th>CIFAR-100</th>
<th>Oxford-102</th>
<th>Oxford-102</th>
<th>CUB-200</th>
<th>COVID</th>
</tr>
</thead>
<tbody>
<tr>
<td># training images</td>
<td>8511</td>
<td>4736</td>
<td>7489</td>
<td>5000</td>
<td>5000</td>
<td>6552</td>
<td>3680</td>
<td>3000</td>
</tr>
<tr>
<td># testing images</td>
<td>2100</td>
<td>650</td>
<td>604</td>
<td>1000</td>
<td>1000</td>
<td>102</td>
<td>3697</td>
<td>200</td>
</tr>
</tbody>
</table>

**Motivation:**

∼$3 billion annual vehicle repair bills related to pothole damage in the U.S.

**Constraints:** limited computational power (e.g., no GPU) on edge devices installed on the road inspection vehicles; need fast inference speed.

**Task:** Given a fixed number of training epochs and a lightweight model to be trained, what can practitioners do to improve the pothole classification accuracy?

Contributions

- We propose Iterative Self Knowledge Distillation (ISKD), which outperforms the SOTA self KD methods from pothole classification (RDD, simplex, complex) to generic (CIFAR-10, CIFAR-100), fine-grained (Oxford 102 Flower, Oxford IIIT Pet, Caltech-UCSD Birds 200), and medical imaging classification (COVID-19 Radiography).
- We provide more evidence showing that a teacher model with accuracy lower than the baseline can still result in a student model outperforming the baseline.
- ISKD is flexible with respect to parameter selection.

Iterative Self Knowledge Distillation

**Iterative Self Knowledge Distillation:**

\[ L_{KD} = (1 - \alpha) L_{ce} + \alpha KL D(z, z') \]

**KL D:** KL divergence; \( z, z' \): output probability distribution of \( T_k \) and \( S_k \); \( \alpha \): weight of KL.

**Accuracy under Different \( \alpha \) Values**

![Graph showing accuracy under different values of \( \alpha \)]

**Teacher-Student Accuracy Relation**

![Diagram showing teacher-student accuracy relation]

**Parameter Efficiency**

<table>
<thead>
<tr>
<th>dataset/method</th>
<th>ISKD</th>
<th>prior work (backbone)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CIFAR-100 [26]</td>
<td>82.67</td>
<td>81.60 [27] (Wide-ResNet-28-10)</td>
</tr>
<tr>
<td>Oxford-37 [18]</td>
<td>91.80</td>
<td>91.60 [28] (ResNet50-SAM)</td>
</tr>
</tbody>
</table>

Table 2: The comparison of classification accuracy (%) between ISKD (backbone: ResNet-18 [23]) and the prior works using backbones with more parameters.

<table>
<thead>
<tr>
<th>dataset</th>
<th>ISKD</th>
<th>prior work (backbone)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CIFAR-100 [26]</td>
<td>82.67</td>
<td>81.60 [27] (Wide-ResNet-28-10)</td>
</tr>
<tr>
<td>Oxford-37 [18]</td>
<td>91.80</td>
<td>91.60 [28] (ResNet50-SAM)</td>
</tr>
</tbody>
</table>

Table 1: Comparing the classification accuracy of the iterative self knowledge distillation (KD) method versus the baselines. The numbers are in the format of [accuracy]_{\alpha}, where \( \alpha \) is the number of epochs when the student model is trained for.

Example Images of the Datasets

![Examples of datasets images]