**Progressive Knowledge Distillation for Generative Modeling**

Adrien Barde$^1,2$  Yu-Xiong Wang$^1$  Ruslan Salakhutdinov$^1$  Martial Hebert$^1$

$^1$Carnegie Mellon University  $^2$Ecole Normale Superieure Paris-Saclay

---

**Motivation**

- **Problem statement:** few-shot learning
- **Key insight:** leverage the knowledge of a high capacity model
- **Our approach:** knowledge distillation
  - Guide the learning of a generative model that "hallucinates" novel examples
  - Maintain decision boundary

---

**Knowledge Distillation for Few-Shot Learning**

- **Few-shot learning setting:** learn from a meta-set—a collection of few-shot classification tasks
- **Meta-training:** train a classification algorithm $\theta$ on small $S_{train}$ that achieves high accuracy on $S_{task}$ for different tasks on $\mathcal{D}_{task-train}$
- **Meta-testing:** use $\theta$ to solve novel classification tasks on $\mathcal{D}_{task-test}$

---

**Knowledge Distillation**

- **Transfer knowledge:** minimize discrepancy between a student model and its teacher
- **Knowledge distillation loss:** Hinton et al. (2015)
  \[L_{CE}(s,t,y) = L_{CE}(s(t),s(t)) + \gamma T L_{CE}(s(t/T),s(t/T))\]
- **Teacher knowledge:** build the teacher model with a large set $S_{teacher}$

---

**Visualization of Knowledge Distillation**

- **Evolution of the decision boundary**
  - real examples (dots) | synthesized examples (triangle)
  - real boundary (red line) | student boundary (black line)
- **Nearest neighbours of synthesized examples**
  - seed (black frame) | neighbor sample (colored frame)
- **Classification results:** generator with distillation (D) against simple generator (G)

---

**Progressive and Ensemble Distillation**

- **Progressive distillation:** progressively remove real examples and synthesize useful examples
  - **Goal:** maintain the optimal decision boundary over the training
  - **Benefit:** make the optimization process smoother than non-progressive approach
- **Ensemble distillation:** learn from multiple teacher models
  - **Diversity** of the generated samples
  - **Robustness** to variations and better generalization

---

**Experimental Results**

- **Benchmarks:** large and small scale classification tasks
  - ImageNet1k: 389 base classes, 196 validation classes, 311 test classes, ResNet-10 pretrained features
  - Mini-ImageNet: 64 base classes, 16 validation classes, 20 test classes, ResNet-18 pretrained features
- **Meta-learner:** prototypical networks with cosine distance
- **Hallucinator:** G: three layer MLP with ReLU
- **Impact of knowledge distillation**
  - Significant improvements on novel classes
  - Progressive distillation outperforms plain distillation
  - Ensemble distillation further improves the performance

---

**Conclusion and Future Work**

- A general framework of meta-learning with generative modeling
- Knowledge distillation to guide the learning of generative models
- State-of-the-art few-shot recognition results on MiniImageNet and ImageNet1K
- Future work: domain adaptation & pixel space generation

---

**Comparison to prior work**

- Cosine classifier: prototypical network with cosine distance
- Cosine classifier with generator

---

**Consistent improvements over baselines on MiniImageNet and ImageNet1k**

<table>
<thead>
<tr>
<th></th>
<th>k=1</th>
<th>5</th>
<th>10</th>
<th>20</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Mini-ImageNet</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cosine Classifier (baseline)</td>
<td>51.99</td>
<td>74.32</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cosine Classifier Gen (baseline)</td>
<td>57.88</td>
<td>75.95</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cosine Classifier Dist</td>
<td>59.20</td>
<td>76.36</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cosine Classifier Dist</td>
<td>59.56</td>
<td>76.57</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cosine Classifier Dist</td>
<td>60.21</td>
<td>77.52</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cosine Classifier Dist</td>
<td>Ensemble</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>ImageNet1k</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cosine Classifier (baseline)</td>
<td>37.6</td>
<td>65.5</td>
<td>72.5</td>
<td>76.6</td>
</tr>
<tr>
<td>Cosine Classifier Gen (baseline)</td>
<td>42.6</td>
<td>66.4</td>
<td>72.6</td>
<td>76.3</td>
</tr>
<tr>
<td>Cosine Classifier Dist</td>
<td>44.5</td>
<td>68.8</td>
<td>74.2</td>
<td>77.3</td>
</tr>
<tr>
<td>Cosine Classifier Dist</td>
<td>45.1</td>
<td>68.8</td>
<td>74.8</td>
<td>79.3</td>
</tr>
<tr>
<td>Cosine Classifier Dist</td>
<td>Ensemble</td>
<td>46.2</td>
<td>70.0</td>
<td>75.6</td>
</tr>
<tr>
<td>Cosine Classifier Dist</td>
<td>Ensemble Mixed</td>
<td>46.9</td>
<td>70.4</td>
<td>75.8</td>
</tr>
</tbody>
</table>