How Well Does Self-Supervised Pre-Training Perform with Streaming ImageNet?

1. Sequential Training vs. Joint Training

- In the upstream pre-training, we propose four types of streaming data with different degrees of distribution shift to mimic real-world data collection scenarios. MoCo-v2 is adopted as the pre-training method.
- In the downstream evaluation, we adopt three common downstream tasks, including many-shot evaluation and few-shot evaluation on 12 image classification tasks and object detection on PASCAL VOC.

2. Dissection of Sequential Self-Supervised Pre-Training

- For streaming data, severe distribution shift in streaming data challenges the sequential SSL.
- For downstream tasks, SSL representations are poor at few-shot transfer. Good upstream performances do not ensure good downstream transfer.
- For SSL methods, BYOL performs similarly to MoCo-v2 on the distant class incremental sequence.
- For continual learning methods, both replay and regularization are effective and can work together to improve the performance.

3. SSL Models Forget Less than SL Models

- Backward transfer:
  - In the sequential training, we use fine-tuning to perform with streaming ImageNet.
  - How well does self-incremental sequence random class incremental instance domain incremental sequence distant class random class incremental instance domain incremental sequence.
  - SSL Models Forget Less than SL Models.

- Sequential Training vs. Joint Training

- For streaming data, severe distribution shift in streaming data challenges the sequential SSL.
- For downstream tasks, SSL representations are poor at few-shot transfer. Good upstream performances do not ensure good downstream transfer.
- For SSL methods, BYOL performs similarly to MoCo-v2 on the distant class incremental sequence.
- For continual learning methods, both replay and regularization are effective and can work together to improve the performance.

- SSL Models Forget Less than SL Models.

- Smaller negative transfer even positive transfer of SSL models.
- SSL representations retain more information in the sequential training process.

4. Our Hypothesis: SSL Models Have Flatter Minima

- The sharpness metric based on representations:

\[ C = \{ x \in \mathbb{R}^n : -\epsilon < \| \theta \| \leq \epsilon \} \]

\[ h_{\theta}(x) = \max_{\theta \in C} \psi(\theta) = \max_{x} f(\theta) - \text{min}_{x} f(\theta) \]

- The sharpness results via perturbation:

<table>
<thead>
<tr>
<th>Instance</th>
<th>Random Class</th>
<th>Distant Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>e = 0.1</td>
<td>e = 0.3</td>
<td>e = 0.1</td>
</tr>
<tr>
<td>SL</td>
<td>0.67</td>
<td>0.21</td>
</tr>
<tr>
<td>SSL</td>
<td>0.14</td>
<td>0.68</td>
</tr>
<tr>
<td></td>
<td>0.08</td>
<td>0.66</td>
</tr>
<tr>
<td></td>
<td>0.71</td>
<td></td>
</tr>
</tbody>
</table>

- For streaming data with mild distribution shift, sequential SSL is a good choice. For streaming data with severe distribution shifts or longer sequences, sequential SSL with unsupervised parameter regularization and simple data replay is efficient.

- Compared with supervised learning (SL) models, SSL models consistently show smaller performance gaps between ST and JT. Our comprehensive investigation of learned representations demonstrates that sequential SSL models are less prone to catastrophic forgetting than SL models.

- Through the empirical analysis on the sharpness minima in the loss landscape, we find that SSL models have wider minima than SL models, which we argue is the probable reason for less forgetting of SSL models.

Takeaways:

- For streaming data with mild distribution shift, sequential SSL is a good choice. For streaming data with severe distribution shifts or longer sequences, sequential SSL with unsupervised parameter regularization and simple data replay is efficient.

- Compared with supervised learning (SL) models, SSL models consistently show smaller performance gaps between ST and JT. Our comprehensive investigation of learned representations demonstrates that sequential SSL models are less prone to catastrophic forgetting than SL models.

- Through the empirical analysis on the sharpness minima in the loss landscape, we find that SSL models have wider minima than SL models, which we argue is the probable reason for less forgetting of SSL models.