We propose a general feature-propagation based contrastive continual learning method in an online fashion:

- feature propagation to align the current and previous representation spaces;
- contrastive representation learning to bridge the domain shifts among distinct tasks;
- supervised contrastive learning to further mitigate the class-wise shifts in the feature space.

The extensive experiments are implemented in multiple image classification tasks.

**Problem Setting**

- Consider a data stream of unknown distributions \(D_t, \ldots, D_T\).
- In the continual learning (CL), the model \(g\) is expected to learn the tasks sequentially - it can only learn one task at a time without forgetting what has learned in the past.
- In the online CL, the model can only experience the current data \(D_t\) on line.
- For rehearsal-based CL, a small memory set \(M\) storing a few past examples is reserved and can be revisited multiple times along the learning of new tasks.

Thus, the objective of the continual learner \(g\) at task \(t\) is to minimize the following loss:

\[
\ell_t = \mathbb{E}_{(x, y) \sim D_t} [\ell_t(x, y)]
\]

which is also known as experience replay.

**Methodology**

**Experience Replay with Feature Propagation**

- Suppose the model consists of a feature extractor \(\psi\) and classifier \(f\).
- Denote \(\psi_i\) is the state model before learning current task.
- Apply feature propagation procedure to training data \(D_t \cup M\) in the feature space.

- The example embeddings derived from \(\psi_i\) are amended by fusing a weighted sum of all example embeddings come from \(\psi_i\):

\[
\bar{\psi}(x) = (1 - w)\psi_i(x) + w \sum_{t \neq i} A_t \psi_t(x)
\]

- The propagation weight \(A_t\) for examples \(x_i\) and \(x_j\) is set as:

\[
A_t = \exp(-\|\phi_t(\psi(x_i)) - \phi_t(\psi(x_j))\|_2^2) / \sum_{s \neq t} \exp(-\|\phi_t(\psi(x_i)) - \phi_t(\psi(x_j))\|_2^2)
\]

where \(\phi_t\) is the Euclidean distance.

- The fusion of the two tasks is motivated by the exponential moving averaging which can move forward without forgetting the past.

- In consequence, the amended experience replay loss is

\[
\ell_t = \mathbb{E}_{(x, y) \sim D_t} [\ell_t(\phi(\psi_i(x)), y)]
\]

**Contrastive Representation Rehearsal**

- Enforce the example embeddings to stay near the previous corresponding ones;

- Protect the representation space from dramatically changing after being retrained on new tasks.

The proposed contrastive loss is defined as follows:

\[
\ell_{\text{con}} = -\sum_{t \neq i} \log \exp(-\|\phi_t(\psi(x_i)) - \phi_t(\psi(x_j))\|_2^2) / \sum_{s \neq t} \exp(-\|\phi_t(\psi(x_i)) - \phi_t(\psi(x_j))\|_2^2)
\]

\[\uparrow \text{To this point, the contrastive continual learning with feature propagation (CCL-PP)} \]

**Supervised Contrastive Replay**

- To further improve the feature space by a supervised contrastive loss.

- Try to eliminate the domain and class shifts between tasks.

The supervised contrastive loss is defined as follows,

\[
\ell_{\text{sup}} = -\sum_{t \neq i} \log \exp(-\|\phi_t(\psi(x_i)) - \phi_t(\psi(x_j))\|_2^2) / \sum_{s \neq t} \exp(-\|\phi_t(\psi(x_i)) - \phi_t(\psi(x_j))\|_2^2)
\]

- By integrating the supervised contrastive loss into the overall objective, we obtain our integrated model named by CCL-PP:

\[
\ell_{\text{CCL-PP}} = \ell_{\text{con}} + \alpha \ell_{\text{sup}} + \beta \ell_{\text{sup}}
\]

**An overview of the proposed method:**

![Diagram of the proposed method]

---

**Experiments**

**Datasets**

We compare our models with several state-of-the-art continual learning methods on Split MNIST, Permuted MNIST, Rotated MNIST, Split CIFAR-10, Split CIFAR-100 and Split Tiny ImageNet.

**Evaluation metric**

We use average accuracy \(\bar{A}\) to evaluate the overall performance of models on test data of all seen tasks, defined as follows,

\[
\bar{A} = \frac{1}{K} \sum_{k=1}^{K} \frac{1}{N_k} \sum_{n=1}^{N_k} \delta(y_{nk} = \hat{y}_{nk})
\]

which indicates the average accuracy on test data of task 1 to \(K\) after the model has learned continually up till task \(K\).

**Experimental Results**

![Graphs showing experimental results]

---

**Table 1**: The average accuracy ± standard deviation (%) by the end of training for baselines and our models across 5 runs with different random seeds. The results for joint training, i.e. the upper bound, and the best accuracies for CL models on each benchmark are marked in bold.

```plaintext
<table>
<thead>
<tr>
<th>Model</th>
<th>S-MNIST</th>
<th>S-CIFAR-100</th>
<th>S-CIFAR-10</th>
<th>S-CIFAR-100</th>
<th>Domain-IL</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Class-IL</td>
<td>Task-IL</td>
<td>Class-IL</td>
<td>Task-IL</td>
<td>Domain-IL</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ER-Reservoir</td>
<td>30.70± 0.14</td>
<td>58.25± 0.09</td>
<td>30.70± 0.14</td>
<td>58.25± 0.09</td>
<td>30.70± 0.14</td>
</tr>
<tr>
<td>HAL</td>
<td>34.00± 0.00</td>
<td>59.35± 0.00</td>
<td>34.00± 0.00</td>
<td>59.35± 0.00</td>
<td>34.00± 0.00</td>
</tr>
<tr>
<td>A-GEM</td>
<td>37.00± 0.00</td>
<td>62.00± 0.00</td>
<td>37.00± 0.00</td>
<td>62.00± 0.00</td>
<td>37.00± 0.00</td>
</tr>
<tr>
<td>GEM</td>
<td>37.00± 0.00</td>
<td>62.00± 0.00</td>
<td>37.00± 0.00</td>
<td>62.00± 0.00</td>
<td>37.00± 0.00</td>
</tr>
<tr>
<td>SCR</td>
<td>37.00± 0.00</td>
<td>62.00± 0.00</td>
<td>37.00± 0.00</td>
<td>62.00± 0.00</td>
<td>37.00± 0.00</td>
</tr>
</tbody>
</table>
```

---

**Figure Left**: The evolution curves of average accuracy on test data of all seen tasks as new tasks are learned, on all datasets in task-1, class-4 and domain-4 settings.

1. Table above: The ablation study on S-MNIST, S-CIFAR-10 and R-MNIST datasets. All results are the average accuracy across 5 runs with different random seeds.

2. Table above: The accuracy ± standard deviation (%) by the end of training for baselines and our models across 5 runs with different random seeds.