Introduction & Related Work

Good representations should at least be Parsimonious, Interpretable and Generalizable. Many existing methods try to achieve some of these goals. Methods like weight decay put regularizer on weights while Dropout, Denoise AutoEncoder, Contractive AutoEncoder, DeCov directly regularize the hidden representations. Here we exploit clustering loss $R$ as a regularizer which leads to the overall objective $\mathcal{L} = \mathcal{L}_c + \lambda \mathcal{L}_r$, where $\mathcal{L}_c$ is the conventional loss function, like cross-entropy.

Clustering Regularization

\[
\mathcal{L}_c = \sum_{i=1}^{N} \sum_{j=1}^{C} \left[ (\mathbf{y}_i^T \mathbf{W}^j \mathbf{x}_i - \mu_j)^2 \right]
\]

Spatial-Clustering

\[
\mathcal{L}_r = \sum_{i=1}^{N} \sum_{j=1}^{C} \left[ \left( \mathbf{x}_i \mathbf{W}^j \mathbf{x}_i - \mu_j \right)^2 \right]
\]

Channel Co-Clustering

\[
\mathcal{L}_r = \sum_{i=1}^{N} \sum_{j=1}^{C} \left[ \left( \mathbf{y}_i \mathbf{W}^j \mathbf{y}_i - \mu_j \right)^2 \right]
\]

Learning Parsimonious Representations

1. Initialization: Maximum training iteration $R$, batch size $B$, smooth weight $\alpha$, set of clustering layers $S$ and set of cluster centers $\{\nu_k|k \in [K]\}$, update period $M$.
2. For iteration $i = 1, 2, ..., R$.
3. For layer $t = 1, 2, ..., L$.
4. Compute the output representation of layer $i$ as $x$.
5. If $t \in S$: Assigning cluster $z_{nt} = \arg\max_{k} \{X_n \mathbf{W}^j \mathbf{x}_i - \nu_{jk}\}$, $\nu_k \in [K]$.
6. Compute cluster center $\nu_k = \frac{1}{|N_k|} \sum_{n \in N_k} X_n$, where $N_k = \{n|z_{nt} = k\}$.
7. Smooth cluster center $\nu_k' = \alpha \nu_k + (1 - \alpha)\nu_k^{-1}$.
8. End
9. End
10. Compute the gradients with cluster centers $\nu_k'$ fixed.
11. Update weights.
12. Update drifted cluster centers using Kmeans++ every $M$ iterations.

Experiments

Auto-Encoder

<table>
<thead>
<tr>
<th>Measurement</th>
<th>Train Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>CIFAR10</td>
<td>69.32</td>
</tr>
<tr>
<td>CIFAR100</td>
<td>68.01</td>
</tr>
</tbody>
</table>

Table 1: Autoencoder Experiments on MNIST. We report the average of mean reconstruction error over 4 trials and the corresponding standard deviation.

Visualization on CIFAR10

Figure 2: Visualization of clusterings on CIFAR10 dataset. Rows 1-4 and 5-8 show sample and spatial clusters respectively. Receptive fields are truncated to fit images.

Zero-Shot Learning

Based on the learned representations, we perform zero-shot learning via solving the following unregularized structure SVM,

\[
\min_{W} \frac{1}{N} \sum_{n=1}^{N} \max_{y \in [K]} \left\{ b_i + x_i^T W y - \frac{1}{2} \| W \|_F^2 \right\}
\]

The results are listed as below.

Table 5: Zero-shot learning on CIFAR-20011.

Future Work

- Back-propagate the gradient through unrolled steps of K-means.
- Exploit soft cluster assignments.
- Apply to semi-supervised tasks.

Code is available at https://github.com/lrjconan/deep_parsimonious.