Introduction

Continual Learning Problem: continuously training neural networks for a new task without information trained on the previous tasks. The goal is to make the network perform well for both tasks.

Catastrophic Forgetting: neural networks lose the performance for the previous tasks after training the new task.

Continual Learning Problem: new task without information trained on the previous tasks. The goal is to make neural networks lose the performance for the previous tasks after training the new task.

Propose two types of moment matching (IMM): incrementally matching the posterior distribution of the neural network which is trained on previous tasks after training the new task.

Incremental moment matching (IMM): incrementally matching the posterior distribution of the neural network which is trained on previous tasks after training the new task.

Contribution

1. Propose two types of incremental matching (IMM) methods for overcoming catastrophic forgetting - Mean-Incremental Moment Matching (mean-IMM) - Mode-Incremental Moment Matching (mode-IMM)
2. Interpret the IMM as the Bayesian perspectives
3. Propose drop-transfer as both a knowledge transfer method for IMM and a continual learning method
4. Apply various transfer techniques in the IMM procedure to make our assumption of Gaussian distribution reasonable

Incremental Moment Matching

Geometric illustration of incremental moment matching (IMM). Mean-IMM simply averages the parameters of two neural networks, whereas mode-IMM tries to find a maximum of the mixture of Gaussian posteriors.

Transfer Techniques for IMM

Mean-IMM: minimize local KL-divergence

\[ \hat{\mu}_k \rightarrow \mu_k \]

Weight-transfer

\[ \hat{\mu}_k = \frac{1}{1-p} \frac{1}{p} \mu_{k-1} + p \mu_k \]

1-step-transfer

\[ \hat{\mu}_k = \frac{1}{1-p} \frac{1}{p} \mu_{k-1} \]

Drop-transfer

\[ \hat{\mu}_k = \frac{1}{1-p} \frac{1}{p} \mu_{k-1} \]

If the th node is turned off

\[ \hat{\mu}_k = \mu_{k-1} \]

To make IMM be reasonable, the research space of the low function between two posterior means and should be reasonably smooth and convex-like. To find a \( \hat{\mu}_k \) which satisfies this condition of a smooth and convex-like path from \( \mu_k \), we apply proposing various transfer techniques for the IMM procedure.

Merging by Approximating Mixture of Gaussian Posteriors

Transfer Techniques for IMM

Weight-transfer makes the search space convex-like (CIFAR-10)

Smooth Search Space

Use inverse Fisher matrix as covariance matrix

\[ \Sigma_k = \frac{1}{1-p} \frac{1}{p} \Sigma_{k-1} + \frac{1}{1-p} \frac{1}{p} \Sigma_k \]

Assume local posterior and approximated global posterior is Gaussian

\[ \Sigma_k \approx \Sigma_k \]

Various transfer techniques for IMM makes the search space in the line/curve smooth (Disjoint MNIST)

Comparison on Disjoint MNIST and Shuffled MNIST Datasets

Comparison on Lifelog Dataset

Experimental Results

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Transfer</th>
<th>Hyperparameter</th>
<th>Hyperparameter</th>
<th>Accuracy</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lifelog</td>
<td>Drop</td>
<td>0.48</td>
<td>0.48</td>
<td>0.60</td>
<td>0.60</td>
</tr>
<tr>
<td>Lifelog</td>
<td>Drop</td>
<td>0.35</td>
<td>0.35</td>
<td>0.70</td>
<td>0.70</td>
</tr>
<tr>
<td>Lifelog</td>
<td>Drop</td>
<td>0.25</td>
<td>0.25</td>
<td>0.75</td>
<td>0.75</td>
</tr>
<tr>
<td>Lifelog</td>
<td>Drop</td>
<td>0.15</td>
<td>0.15</td>
<td>0.80</td>
<td>0.80</td>
</tr>
<tr>
<td>Lifelog</td>
<td>Drop</td>
<td>0.05</td>
<td>0.05</td>
<td>0.85</td>
<td>0.85</td>
</tr>
<tr>
<td>Lifelog</td>
<td>Drop</td>
<td>0.05</td>
<td>0.05</td>
<td>0.90</td>
<td>0.90</td>
</tr>
</tbody>
</table>

ACKNOWLEDGMENT

This work was supported by Naver Corp. and partly by the Korea government (IITP-2017-0-01772-VTT, IITP-R0126-16-1072-SW.StarLab, KEIT-10044009-HRI.MESSI, KEIT-10060086-RISF). JK is supported by 2017 Google PhD Fellowship.

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.