EWC-detection instantiation

Peng YUN

20200921

We want to find the most probable parameters given the data of old and new tasks $\mathcal{D} = \mathcal{D}_A \cup \mathcal{D}_B$, where $\mathcal{D}_A$ denotes the data of the old task-A, and $\mathcal{D}_B$ denotes the data of the new task-B. The goal can be achieved by minimizing the negative logarithm of the posterior distribution $-\log p(\boldsymbol{\theta}|\mathcal{D}_A \cup \mathcal{D}_B)$. According to the bayesian rule, we can write the posterior term into:

$$-\log p(\boldsymbol{\theta}|\mathcal{D}_A \cup \mathcal{D}_B) = -\log p(\mathcal{D}_B|\boldsymbol{\theta}) - \log p(\boldsymbol{\theta}|\mathcal{D}_A) + \log p(\mathcal{D}_B) \tag{1}$$

To get the optimal parameters $\boldsymbol{\theta}$, we solve the following optimization problem:

$$\boldsymbol{\theta}^*_{AB} = \arg\min_{\boldsymbol{\theta}} -\log p(\mathcal{D}_B|\boldsymbol{\theta}) - \log p(\boldsymbol{\theta}|\mathcal{D}_A) \tag{2}$$

It is noted that we get rid of the term $\log p(\mathcal{D}_B)$ in this optimization problem since the data distribution prior can be considered as a constant.

For the objective function, we can easily compute the likelihood term $-\log p(\mathcal{D}_B|\boldsymbol{\theta})$ with the task-B dataset at hand, and it is exactly the linear combination of classification and regression loss for object detection, which we denote it as $\mathcal{L}_{det}(\boldsymbol{\theta}, \mathcal{D}_B)$.

The term $-\log p(\boldsymbol{\theta}|\mathcal{D}_A)$ is intractable, since we do not have $\mathcal{D}_A$ at hand at the time of incrementally training task-B. The mechanism behind the prior-based method EWC is to restore the prior information of old task and then use the priors instead of old-task data to preserve learned knowledge. What we have at hand is $\boldsymbol{\theta}^*_A$ which minimizes the posterior term $-\log p(\boldsymbol{\theta}|\mathcal{D}_A)$, and therefore $\frac{\partial}{\partial \boldsymbol{\theta}} \log p(\boldsymbol{\theta}|\mathcal{D}_A)|_{\boldsymbol{\theta}^*_A} = 0$. We unfold the $\log p(\boldsymbol{\theta}|\mathcal{D}_A)$ at $\boldsymbol{\theta}^*_A$ with Taylor expansion:

$$\log p(\boldsymbol{\theta}|\mathcal{D}_A) \approx \log p(\boldsymbol{\theta}^*_A|\mathcal{D}_A) + \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)^T \frac{\partial^2}{\partial^2 \boldsymbol{\theta}} \log p(\boldsymbol{\theta}|\mathcal{D}_A)|_{\boldsymbol{\theta}^*_A}(\boldsymbol{\theta} - \boldsymbol{\theta}^*) \tag{3}$$

We denote the $\frac{\partial^2}{\partial^2 \boldsymbol{\theta}} \log p(\boldsymbol{\theta}|\mathcal{D}_A)|_{\boldsymbol{\theta}^*_A}$ with $\boldsymbol{H}(\mathcal{D}_A, \boldsymbol{\theta}^*_A)$. It demonstrates the logarithm posterior distribution $\log p(\boldsymbol{\theta}|\mathcal{D}_A) \sim \mathcal{N}(\log p(\boldsymbol{\theta}^*_A|\mathcal{D}_A), -\boldsymbol{H}(\mathcal{D}_A, \boldsymbol{\theta}^*_A)^{-1})$ according to Laplacian approximation.

We can compute $\boldsymbol{H}(\mathcal{D}_A, \boldsymbol{\theta}^*_A)$ with empirical fisher information matrix (FIM):

$$\begin{gathered} \boldsymbol{H}(\mathcal{D}_A, \boldsymbol{\theta}^*_A) = -\mathbb{F}(\mathcal{D}_A, \boldsymbol{\theta}^*_A) \\ \mathbb{F}(\mathcal{D}_A, \boldsymbol{\theta}^*_A) = \frac{1}{|S|} \sum_{\tilde{\mathcal{D}}_A \sim \mathcal{D}_A} [(\frac{\partial}{\partial \boldsymbol{\theta}} \log p(\boldsymbol{\theta}|\tilde{\mathcal{D}}_A)|_{\boldsymbol{\theta}^*_A})^T (\frac{\partial}{\partial \boldsymbol{\theta}} \log p(\boldsymbol{\theta}|\tilde{\mathcal{D}}_A)|_{\boldsymbol{\theta}^*_A})] \end{gathered} \tag{4}$$

where $|S|$ denotes the number of times sampling $\tilde{\mathcal{D}}_A$ from $\mathcal{D}_A$. Consider the task-A is the first task in the task sequence, thus there is no prior information related to $\boldsymbol{\theta}$:

$$\begin{aligned}\log p(\boldsymbol{\theta}|\mathcal{D}_A) &= \log p(\mathcal{D}_A|\boldsymbol{\theta}) + \log p(\boldsymbol{\theta}) - \log p(\mathcal{D}_A) \\ &= \log p(\mathcal{D}_A|\boldsymbol{\theta}) + \text{const.}\end{aligned} \tag{5}$$

Therefore, we can compute the empirical FIM right after training task-A with the equation:

$$\mathbb{F}(\mathcal{D}_A, \boldsymbol{\theta}^*_A) = \frac{1}{|S|} \sum_{\tilde{\mathcal{D}_A} \sim \mathcal{D}_A} [(\frac{\partial}{\partial \boldsymbol{\theta}} \mathcal{L}_{det}(\tilde{\mathcal{D}_A}, \boldsymbol{\theta})|_{\boldsymbol{\theta^*_A}})^T (\frac{\partial}{\partial \boldsymbol{\theta}} \mathcal{L}_{det}(\tilde{\mathcal{D}_A}, \boldsymbol{\theta})|_{\boldsymbol{\theta^*_A}})] \tag{6}$$

As a result, the objective function of optimizing θ for incrementally learning task-B is:

$$\boldsymbol{\theta}^*_{AB} = \arg\min_{\boldsymbol{\theta}} \mathcal{L}_{det}(\boldsymbol{\theta}, \mathcal{D}_B) + \lambda(\boldsymbol{\theta} - \boldsymbol{\theta}^*_A)^T \mathbb{F}(\mathcal{D}_A, \boldsymbol{\theta}^*_A)(\boldsymbol{\theta} - \boldsymbol{\theta}^*_A) \tag{7}$$

where the hyperparameter λ balances the weights between the detection loss and the L2-norm prior constraints.

Then we extend to the third task task-C, and derive the empirical FIM propagation. For three task cases, we want to find the optimal parameter $\boldsymbol{\theta}$ by minimizing the $-\log p(\boldsymbol{\theta}|\mathcal{D}_A \cup \mathcal{D}_B \cup \mathcal{D}_C)$. The optimization problem is seen as:

$$\boldsymbol{\theta}^*_{ABC} = \arg\min_{\boldsymbol{\theta}} - \log p(\mathcal{D}_C|\boldsymbol{\theta}) - \log p(\boldsymbol{\theta}|\mathcal{D}_A \cup \mathcal{D}_B) \tag{8}$$

As before, the likelihood term $-\log p(\mathcal{D}_C|\boldsymbol{\theta}) = \mathcal{L}_{det}(\boldsymbol{\theta}, \mathcal{D}_C)$ and can be computed easily with the task-C dataset at hand. What we also have at hand is $\boldsymbol{\theta}^*_{AB}$ which minimizes the posterior term of the two-task case $-\log p(\boldsymbol{\theta}|\mathcal{D}_A \cup \mathcal{D}_B)$, thus $\frac{\partial}{\partial \boldsymbol{\theta}} \log p(\boldsymbol{\theta}|\mathcal{D}_A \cup \mathcal{D}_B) = 0$. Please recap the equation 1. We can unfold its RHS at $\boldsymbol{\theta}^*_{AB}$ and get:

$$\begin{aligned}&\log p(\boldsymbol{\theta}|\mathcal{D}_A \cup \mathcal{D}_B) \\ &= \log p(\mathcal{D}_B|\boldsymbol{\theta}^*_{AB}) + \log p(\boldsymbol{\theta}^*_{AB}|\mathcal{D}_A) \\ &= \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}^*_{AB})^T \frac{\partial^2}{\partial^2 \boldsymbol{\theta}}[\log p(\mathcal{D}_B|\boldsymbol{\theta}) + \log p(\boldsymbol{\theta}|\mathcal{D}_A)]_{\boldsymbol{\theta}^*_{AB}}(\boldsymbol{\theta} - \boldsymbol{\theta}^*_{AB}) + \text{const.} \\ &= \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}^*_{AB})^T [\frac{\partial^2}{\partial^2 \boldsymbol{\theta}} \log p(\mathcal{D}_B|\boldsymbol{\theta})|_{\boldsymbol{\theta}^*_{AB}} + \frac{\partial^2}{\partial^2 \boldsymbol{\theta}} \log p(\boldsymbol{\theta}|\mathcal{D}_A)|_{\boldsymbol{\theta}^*_{AB}}](\boldsymbol{\theta} - \boldsymbol{\theta}^*_{AB}) + \text{const.}\end{aligned} \tag{9}$$

For $\frac{\partial^2}{\partial^2 \boldsymbol{\theta}} \log p(\mathcal{D}_B|\boldsymbol{\theta})|_{\boldsymbol{\theta}^*_{AB}}$, we can approximate it with the negative empirical FIM as before:

$$\mathbb{F}(\mathcal{D}_B, \boldsymbol{\theta}^*_{AB}) = \frac{1}{|S|} \sum_{\tilde{\mathcal{D}}_B \sim \mathcal{D}_B} [(\frac{\partial}{\partial \boldsymbol{\theta}} \mathcal{L}_{det}(\tilde{\mathcal{D}}_B, \boldsymbol{\theta})|_{\boldsymbol{\theta^*_B}})^T (\frac{\partial}{\partial \boldsymbol{\theta}} \mathcal{L}_{det}(\tilde{\mathcal{D}}_B), \boldsymbol{\theta})|_{\boldsymbol{\theta^*_B}})] \tag{10}$$

For $\frac{\partial^2}{\partial^2 \boldsymbol{\theta}} \log p(\boldsymbol{\theta}|\mathcal{D}_A)|_{\boldsymbol{\theta}^*_{AB}}$, we have already derived the following equation (equation 3, 6):

$$\log p(\boldsymbol{\theta}|\mathcal{D}_A) \approx \log p(\boldsymbol{\theta}^*_A|\mathcal{D}_A) + \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}^*_A)^T \mathbb{F}(\mathcal{D}_A, \boldsymbol{\theta}^*_A)(\boldsymbol{\theta} - \boldsymbol{\theta}^*_A) \tag{11}$$

Thus, we can get:

$$\frac{\partial^2}{\partial^2 \boldsymbol{\theta}} \log p(\boldsymbol{\theta}|\mathcal{D}_A)|_{\boldsymbol{\theta}^*_{AB}} = \mathbb{F}(\mathcal{D}_A, \boldsymbol{\theta}^*_A) \tag{12}$$

As a result, the objective function for incrementally learning task-C is:

$$\boldsymbol{\theta}^*_{ABC} = \arg\min_{\boldsymbol{\theta}} \mathcal{L}_{det}(\boldsymbol{\theta}, \mathcal{D}_C) + \lambda(\boldsymbol{\theta} - \boldsymbol{\theta}^*_{AB})^T [\mathbb{F}(\mathcal{D}_A, \boldsymbol{\theta}^*_A) + \mathbb{F}(\mathcal{D}_B, \boldsymbol{\theta}^*_{AB})](\boldsymbol{\theta} - \boldsymbol{\theta}^*_{AB}) \tag{13}$$

By recursively applying the above equation, the general objective function for incrementally learning task-$\mathcal{T}$ is:

$$\boldsymbol{\theta}^*_{A..\mathcal{T}} = \arg\min_{\boldsymbol{\theta}} \mathcal{L}_{det}(\boldsymbol{\theta}, \mathcal{D}_{\mathcal{T}}) + \lambda(\boldsymbol{\theta} - \boldsymbol{\theta}^*_{AB})^T [\sum_t \mathbb{F}(\mathcal{D}_t, \boldsymbol{\theta}^*_{A..t})](\boldsymbol{\theta} - \boldsymbol{\theta}^*_{AB}) \tag{14}$$