Student ID : 20194293
Name : Go, Kyeong Ryeol

[AI 602] Variational Continual Learning

1. Paper Summary

When data arrives in a non-i.i.d. fashion, task changes over time, or entirely new task emerges, the previously trained model should be modified and minimum changes is preferable. Therefore, balancing the tendency between the adapting to recent data and sticking to old data is critical. As a prior work, individual models were trained for each task and combined in the later step or parameters were separated for whether it allows to be changed restrictively or freely.

The author proposed to use the Bayesian inference which inherently proceed with this regime where the previous posterior and the likelihood on the new data are utilized to model the new posterior. More specifically, the online variational inference with Monte Carlo VI for neural networks are leveraged so that Variational Continual Learning (VCL) is introduced which has shown excellent performance on both deep discriminative and generative model when used along with coreset data summarization method.

As mentioned above, the posterior can be computed through recursion; $\mathrm{p}(\theta|\mathrm{D}_{1:\mathrm{t}}) \approx q_t(\theta) = proj\big(q_{t-1}(\theta)p(D_t|\theta)\big)$. Here, the projection operation implies the approximate inference and among the well-known methods, the author chose the variational KL minimization as it typically outperformed the others for complex models in the static setting. As a result, it can be summarized as follow.

$$\mathrm{q_t}(\theta) = argmin_q KL\left(q(\theta)\left\|\frac{1}{Z_t}q_{t-1}(\theta)p(D_t|\theta)\right.\right) \;\; for \;\; t = 1,2,\dots,T$$

To mitigate the bias occurred by the repeated approximation by recursion and the minimization at each step, VCL is extended to include the coreset as an episodic memory to transfer the key information from the previous tasks. Then, a variational recursion is developed as follow where the likelihood is only computed from the coreset of the corresponding task. Therefore, the propagation is revised as $\tilde{q}_t(\theta) = proj\big(\tilde{q}_{t-1}(\theta)p(D_t \cup C_{t-1} \backslash C_t|\theta)\big)$ and further projection step is required for performing prediction $\mathrm{q_t}(\theta) = proj\big(\tilde{q}_t(\theta)p(C_t|\theta)\big)$. When applied to deep discriminative and generative models, the variational learning objective can be summarized as follow.

($* \mathrm{q}_\phi\left(\mathrm{z}_\mathrm{t}^{(n)}\middle|\mathrm{x}_\mathrm{t}^{(n)}\right)$ is an encoder network and ϕ is the task-specific parameter.)

$$\text{(DDM)}: L_{VCL}^t\big(q_t(\theta)\big) = \sum_{n=1}^{N_t} E_{\theta\sim q_t(\theta)}\big[\log p\big(y_t^{(n)}|\theta, x_t^{(n)}\big)\big] - KL(q_t(\theta)\|q_{t-1}(\theta))$$

$$\text{(DGM)}: L_{VCL}^t(q_t(\theta),\phi) = E_{\theta\sim q_t(\theta)}\left\{\sum_{n=1}^{N_t} E_{q_\phi(z_t^{(n)}|x_t^{(n)})}\left[\log\frac{p\big(x_t^{(n)}|z_t^{(n)},\theta\big)p\big(z_t^{(n)}\big)}{q_\phi\big(z_t^{(n)}|x_t^{(n)}\big)}\right]\right\} - KL(q_t(\theta)\|q_{t-1}(\theta))$$

2. In-depth discussions

I. Rather than considering the task relation only through the shared parameters, would it be beneficial if we introduce the hierarchy among the tasks? In the extreme, what about keep extending the hierarchy so that every node in the tree structure only contain one task?