Student ID : 20194293
Name : Go, Kyeong Ryeol

[AI 602] Meta-Learning Probabilistic Inference for Prediction

1. Paper Summary

The author pointed out that there lacks a general method for flexible and data-efficient learning method for meta-learning. Here, a new unifying framework, Meta-Learning approximate Probabilistic Inference for Prediction (ML-PIP) is proposed, which utilized a hierarchical probabilistic model to share the statistical structure between the tasks. Most importantly, it amortized the posterior predictive distributions $q_\phi(\tilde{y}|D) = \int p_\theta(\tilde{y}|\psi)q_\phi(\psi|D)d\psi$ which enables to flexibly and versatilely handle a wide range of tasks. (θ: shared parameter, ψ: task specific parameter, ϕ: amortization parameter) Here, the factorized gaussian distribution with means and variances are set by the amortization network where reparameterization trick can be applied.

Specifically, the goal of learning is to minimize the expected value of the KL averaged over tasks $E_{p(D)}KL\left[p(\tilde{y}|D)\middle\|q_\phi(\tilde{y}|D)\right]$. Therefore, the end-to-end stochastic training objective for θ and ϕ can be computed as below where Monte Carlo sampling is applied to approximate the integral of task specific parameter ψ. This has its strength on the fact that no explicit prior specification is required where the typical choice in the stochastic neural network as the standard normal may incur the information loss.

$$\hat{L}(\theta,\phi) = -\frac{1}{MT}\sum_{M,T}\log\frac{1}{L}\sum_{i=1}^{L} p_\theta(\tilde{y}_m|\tilde{x}_m^{(t)},\psi_l^{(t)}) \;\; where \;\; \psi_l^{(t)} \sim q_\phi(\psi|D^{(t)},\theta)$$

Furthermore, the author specified the amortized inference framework VERSA which substitute the optimization procedure at test time with simple feed forward passes. Here, instance pooling is applied for permutation invariant set encoding property and context independent approximation is considered to address the limitations of a naïve amortization.

2. In-depth discussions

I. What may be the pros and cons of minimizing the KL divergence of predictive distributions comparing to minimizing the KL divergence of posterior distributions of task-specific parameter which is widely leveraged in variational inference technique?

II. What may be the pros and cons for substituting the optimization step in the test phase by the feed forward process of the amortization network? Will it be still valid in complex dataset?