Regularization-based incremental learning derivation

Peng YUN

20200818

1 prior-based incremental learning [1]

Incremental learning assumes that an incremental dataset ϵ is provided in addition to $\mathcal{D}$. If these two are disjoint, we can write

$$L(\mathbf{w}; \mathcal{D} \cup \epsilon) = L(\mathbf{w}; \mathcal{D}) + L(\mathbf{w}; \epsilon) \tag{1}$$

If L is differentiable with respect to $\mathbf{w}$ and we train until convergence ($\nabla_{\mathbf{w}} L(\mathbf{w}_0, \mathcal{D}) = 0$), we can expand L to second-order around the previous parameters $\mathbf{w}_0$ to obtain

$$\begin{aligned} L(\mathbf{w}; \mathcal{D} \cup \epsilon) &= L(\mathbf{w}; \mathcal{D}) + L(\mathbf{w}; \epsilon) \\ &\simeq L(\mathbf{w}_0 + \delta\mathbf{w}; \epsilon) + L(\mathbf{w}_0; \mathcal{D}) \\ &+ \delta\mathbf{w}^T H(\mathbf{w}_0; \mathcal{D}) \delta\mathbf{w} \end{aligned} \tag{2}$$

where $\mathbf{w} = \mathbf{w}_0 + \delta\mathbf{w}$ and $H(\mathbf{w}_0; \mathcal{D})$ is the Hessian of the Loss $L(\mathbf{w}; \mathcal{D})$ computed at $\mathbf{w}_0$. Ignoring the constant term $L(\mathbf{w}_0; \mathcal{D})$ yields the derived loss

$$L(\mathbf{w}; \epsilon) = L(\mathbf{w}_0 + \delta\mathbf{w}; \epsilon) + \delta\mathbf{w}^T H(\mathbf{w}_0; \mathcal{D}) \delta\mathbf{w} \tag{3}$$

minimizing which corresponds to fine-tuning the based model for the new task while ensuring that the parameters change little.

2 distillation-based incremental learning [1]

Distillation is based on approximating the loss not by perturbing the weights, $\mathbf{w}_0 \to \mathbf{w}_0 + \delta\mathbf{w}$, but by perturbing the discriminant function, $p_{\mathbf{w}_0} \to p_{\mathbf{w}_0 + \delta\mathbf{w}}$, which can be done by minimizing

$$L(\mathbf{w}) = L(\mathbf{w}; \epsilon) + \lambda \mathbb{E}_{x \sim \mathcal{D}} KL(p_{\mathbf{w}_0}(y|x) || p_{\mathbf{w}}(y|x)), \tag{4}$$

where the KL divergence measures the perturbation of the new discriminant p_w with respect to the old one $p_{\mathbf{w}_0}$ in units λ.

3 Connection between the regularization-based and distillation-based methods [1]

The losses in equation 3 and 4 are equivalent up to first-order, meaning that a local first-order optimization would yield the same initial step when minimizing them.

References

[1] Qing Liu, Orchid Majumder, Alessandro Achille, Avinash Ravichandran, Rahul Bhotika, and Stefano Soatto. Incremental Meta-Learning via Indirect Discriminant Alignment. 2020.