《EFFICIENT SHARPNESS-AWARE MINIMIZATION FOR IMPROVED TRAINING OF NEURAL NETWORKS》- ICLR2022

[ICLR2022]https://arxiv.org/pdf/2110.03141.pdf

Use ESAM(Stochastic Weight Perturbation and Sharpness-Sensitive Data Selection) to boosts SAM's efficiency at no cost to its generalization performance.

1. 背景介绍

考虑到SAM的计算成本，作者提出了ESAM，用于提升计算效率。

2. METHODOLOGY

ESAM包括Stochastic Weight Perturbation和Sharpness-Sensitive Data Selection。

Stochastic Weight Perturbation（用于第一步分的梯度计算）：在所有的参数里面随机选出一部分（随机按照伯努利0-1分布只更新部分权重），m为mask（掩码）(与《Make Sharpness-Aware Minimization Stronger: A Sparsified Perturbation Approach》思想类似)

$$\boldsymbol{a}(\theta, \mathbb{B}) = \frac{\mathbf{m}^\top \hat{\epsilon}(\theta, \mathbb{B})}{\beta}$$

β为伯努利分布参数，除β是为了保持强度。

Sharpness-Sensitive Data Selection（用于第二部分loss的梯度计算）：对数据进行选择，

$$\mathbb{B}^+ := \{(x_i, y_i) \in \mathbb{B} : \ell(f_{\theta+\hat{\epsilon}}, x_i, y_i) - \ell(f_\theta, x_i, y_i) > \alpha\},$$
$$\mathbb{B}^- := \{(x_i, y_i) \in \mathbb{B} : \ell(f_{\theta+\hat{\epsilon}}, x_i, y_i) - \ell(f_\theta, x_i, y_i) < \alpha\},$$

用阈值控制，只选出最敏感的样本，α值较难选择，文中作者用$\gamma = |\mathrm{B}+|/|\mathrm{B}|$来决定。

SWP的泛化（为什么可以得到和SAM一样的泛化能力）：

$$\bar{\boldsymbol{a}}(\theta, \mathbb{B})_{[i]} = \mathbb{E}[\boldsymbol{a}(\theta, \mathbb{B})_{[i]}] = \frac{1}{\beta} \cdot \beta\hat{\epsilon}(\theta, \mathbb{B})_{[i]} = \hat{\epsilon}(\theta, \mathbb{B})_{[i]},$$