MAPLE

key idea：找到一个 reweighting method. 在加权训练数据上对大model进行ERM可以得到好的OOD泛化性能.

在联知的训练数据上衡量其性能.

经验风险.

Regularization based → DRO / IRM

小模型 perform well. 对于 overparameterized models 很脆弱 vulnerable.

本文改进：weighted-ERM；把模型参数的搜索空间转换成样本权重的搜索空间.

Reweighting based →

① 重新加权数据的分布. 启发式方法 ⟶ 需要先验知识，难.

本文改进：自动学习样本权重

② 在加权后的分布上进行ERM.

本文：用weighted-ERM解决正则化方法过参数化的问题（过拟合）.

自动学习样本权重、解决重新加权方法的限制.

model agnostic 的两层优化:

内循环: 在加权训练样本上训练DNN.(ERM训练), 得到模型θ.

外循环: 在验证集上评估的OOD标准作为外部目标来学习样本权重

自动权重学习

不容易过拟合的原因: 搜索权重、而不是模型参数、

Question: 权重空间为什么以样本数作为参考?

有可能是因为样本的权重、对样本操作

举例: CIFAR-10 : 50K 训练样本

ResNet-18 : 11.4M 参数

• 在 waterbirds 数据集中比 G-DRO 得分更高.

数据集 $D := \{(x_i, y_i)\}_{i=1}^{n}$, $(x_i, y_i) \in X \times Y$

加权经验损失 $\mathcal{L}(D, \theta; w) := \frac{1}{n} \sum_{i=1}^{n} \underline{w_i} \, \underline{l}(\underline{f}(x_i; \theta), y_i)$

≥ 0　损失函数 CE, LS.　网络

$\mathcal{L}(D, \theta)$: 无偏损失 $= \mathcal{L}(D, \theta; 1)$

Z_c : bias-conflict (core)　　Z_s : bias-aligned (spurious)

Validation ⟶ 相比于正则化方法：使用验证集缓解训练集的过拟合

$$\min_{w \in C_1} \underline{R}(\underline{D_v}, \theta^*(w))$$

IRM/DRO

training

$$\text{s.t.} \quad \theta^*(w) \in \arg\min_{\theta} \underline{\mathcal{L}}(\underline{D_{tr}}, \theta; w), \quad C = \{w : w \geq 0, \|w_0\| \leq k\}$$

$w \leftarrow \text{proj}_C\left(w - \eta \nabla_\theta R|_{\theta_T} \frac{\partial^2 \mathcal{L}}{\partial\theta \partial w^T}\Big|_{\theta_{T-1}}\right)$

ERM loss　$\hat{\theta}_{ERM} := \arg\min_{\theta \in \Theta} \mathbb{E}_{(x,y) \in \hat{p}} [l(\theta; (x, y))]$

训练集上的经验分布.

Sparse. 减小 cost

$$\min_{(w,s) \in C'} \Phi(w, s) = \mathbb{E}_{p(m|s)} R(D_v, \theta^*(w, m))$$

$C' = \{(w, s) : w \geq 0, 0 \leq s \leq 1, \|s\|_1 \leq k\}$

$$\text{s.t.} \quad \theta^*(w, m) \in \arg\min_{\theta} \mathcal{L}(D_{tr}, \theta; w \circ m)$$

m_i : Bernoulli Random Variable.

$p(m_i = 1) = s_i$,

$p(m_i = 0) = 1 - s_i$

Gumbel(0,1) 中随机选的两个随机变量 iid

$p(m|s) = \prod_{i=1}^{n} (s_i)^{m_i} (1 - s_i)^{(1 - m_i)}$

$$\nabla_{w,s} \Phi \approx \nabla_{w,s} R\left(\theta^*\left(w, \mathbb{1}\left(\log\left(\frac{s}{1-s}\right) + \underline{g_1 - g_0} \geq 0\right)\right)\right)$$

$(w, s) \leftarrow \text{proj}_{C'}(w - \eta \nabla_w \Phi, s - \eta \nabla_s \Phi)$　projected gradient descent.

Datasets: Colored MNIST, CIFARMNIST, Waterbirds, CelebA

validate MAPLE on DRO

Baseline: ERM, IRMv1, REx, MRM, Sparse IRM, BayesianIRM → IRM系列

ERM, CVaRDRO, LfF, JTT, Up Weighting, Group DRO → DRO系列.

Algorithm 1 Model Agnostic Sample Reweighting (MAPLE)

Input: a network θ, remaining training sample size K, training set $\mathcal{D}_{tr}$ and validation set $\mathcal{D}_v$.

Initialize sample weights $\boldsymbol{w} = \mathbf{1}$ and probabilities $\boldsymbol{s} = \frac{K}{|\mathcal{D}_{tr}|}\mathbf{1}$.

for training iteration $i = 1, 2 \ldots I$ **do**

  Sample mask $\boldsymbol{m}$ according to the probability distribution $p(\boldsymbol{m}|\boldsymbol{s}) = \Pi_{i=1}^{n}(s_i)^{m_i}(1 - s_i)^{(1-m_i)}$. After Sparse.

  Train the inner loop to converge: $\boldsymbol{\theta}^*(\boldsymbol{w}, \boldsymbol{m}) \leftarrow \arg\min_{\boldsymbol{\theta}} \mathcal{L}(\mathcal{D}_{tr}, \boldsymbol{\theta}; \boldsymbol{w}, \boldsymbol{m})$ started from randomly initialized $\boldsymbol{\theta}$.

  Estimate $\nabla_{\boldsymbol{s}}\Phi(\boldsymbol{w}, \boldsymbol{s})$ and $\nabla_{\boldsymbol{w}}\Phi(\boldsymbol{w}, \boldsymbol{s})$ by Straight-through Gumbel-softmax and 1-step truncated backpropagation.

  Perform projected gradient descent: $(\boldsymbol{w}, \boldsymbol{s}) \leftarrow \text{proj}_{\mathcal{C}'}(\boldsymbol{w} - \eta\nabla_{\boldsymbol{w}}\Phi(\boldsymbol{w}, \boldsymbol{s}), \boldsymbol{s} - \eta\nabla_{\boldsymbol{s}}\Phi(\boldsymbol{w}, \boldsymbol{s}))$

end for

output The weighted set $\{(\mathbf{x}_i, \mathbf{y}_i, w_i) : m_i \neq 0, (\mathbf{x}_i, \mathbf{y}_i) \in \mathcal{D}_{tr}\}$ with $\boldsymbol{m}$ sampled from $p(\boldsymbol{m}|\boldsymbol{s})$