【周报9.13~9.19】 王启迅

一、实验

这周大半的时间花在了配服务器环境上。。
受Sepctral Decoupling方法：

$$\mathcal{L}(\theta)=1\cdot(log[1+exp(-Y\hat{y})]+\frac{\lambda}{2}||\hat{y}||^2)$$

对$||\hat{y}||^2$做正则的想法启发，想到可以对同样是能反映模型生成能力的，由模型生成的$p_\theta(x)$做正则。
大致思路是：根据Energy Based Model中提出的数据x概率密度：

$$p_\theta(x)=\frac{\sum_y exp(f_\theta(x)[y])}{Z(\theta)}$$

忽略掉分母的配分函数，把分子的$log-sum-exp$项拿出来作正则，来增强OoD robustness。

具体实验：在cifar10数据集上，用resnet18跑了135个epoch，测试集正确率能到92.070%。
与Spectral Decoupling在colored mnist上对比，2000次迭代下，运行在测试集正确率为67.77%，我们的方法只有29.02%，接近ERM（29.48%）。分析后发现colored mnist是二分类任务，logsumexp函数对输出logits的标签维度求和没意义，因为dim=1那维只有一个数，求和了也没用，所以相当于只是在原本的loss上加了一个一阶的logits $|y|$。

二、论文阅读

1、Gradient Matching for Domain Generalization
在上次讲的SAND-mask基础上，又看了一篇用梯度方法解决domain generalization的文章:Gradient Matching for Domain Generalization.文章提出了$inner-domain\ gradient\ matching(IDGM)$方法，即通过最大化不同environment的loss梯度的内积，来寻找domain invariant features。损失函数于是可以写为：

$$\mathcal{L}_{idgm}=\mathcal{L}_{erm}(\mathcal{D}_{tr};\theta)-\gamma\frac{2}{S(S-1)}\sum_{i,j\in S}^{i\neq j}G_iG_j$$

其中，$G_i=E_{\mathcal{D}_i}\frac{\partial l(x,y;\theta)}{\partial\theta}$.

第一项为ERM。记第二项为$GIP(gradient\ inner\ product)$。由于直接优化上述损失函数涉及二阶导数的计算，为了避免这一点，文章提出了一阶近似算法Fish：用 $\tilde{\theta}-\theta$每次迭代作为更新量更新参数，其中$\tilde{\theta}$由如下迭代得到：

for $\mathcal{D}_i\in pemute(\mathcal{D}_1,\mathcal{D}_2,...,\mathcal{D}_S)$ **do**
sample batch $d_i\sim\mathcal{D}_i$

$\tilde{g}_i = E_{d_i}(\frac{\partial l(x,y;\tilde{\theta})}{\partial \tilde{\theta}})$
Update $\tilde{\theta} \leftarrow \tilde{\theta} - \alpha \tilde{g}_i$

关于Fish为什么能作为GIP的近似，作者通过下面的定理进行了说明：

Theorem 3.1 定义两个量：

$$G_f = E[(\theta - \tilde{\theta}) - \alpha S \cdot \overline{G}]$$

$$G_g = -\frac{\partial \hat{G}}{\partial \theta}$$

其中：$\overline{G} = \frac{1}{S}\sum_{s=1}^{S} G_s$，为ERM的全梯度.$\hat{G}$为GIP项，即$\frac{2}{S(S-1)}\sum_{i,j\in S}^{i\neq j} G_i G_j$
我们有：

$$\lim_{\alpha\to 0}\frac{G_f \cdot G_g}{||G_f|| \cdot ||G_g||} = 1$$

上述定理说明了G_f和G_g在$\alpha \to 0$时方向相同，于是Fish的更新项$\tilde{\theta} - \theta$使得对损失函数的优化同时朝着ERM和GIP的方向进行，实现了对IDGM的一阶近似。

作者在多个数据集上进行了实验。
①CDSPRITES-N
作者提出的一类二分类任务OoD数据集，训练时颜色和标签之间存在spurious correlation，测试时这一关联被破坏，而我们希望分类器学习到形状这一variant feature。实验结果表明IDGM性能优于ERM，且Fish相比于直接优化IDGM，在性能上并无差别。
①WILDS
选了WILDS最难的6个数据集，在除了AMAZON之外的数据集上均优于baseline。在AMAZON上，ERM还是最优算法。其他domain generalization算法的失败可能是由于其过大的环境数(7,676个环境)。
③domain bed
在七个测试集上，之比carol低了0.1%，位于domain bed上已有算法的第二。

一些其他的分析：
(1)Fish方法的优势来自所maximize的梯度乘积必须是来自不同环境之间的。作者做了实验，如果把不同环境的样本放在一起进行random grouping，那么在CDSPRITES上的性能会从100%降到50%（等同于ERM）。

2、Fishr: Invariant Gradient Variances for Out-of-distribution Generalization
这篇文章指出了之前一些基于match不同domain梯度的工作的不足：把各个domain内部的梯度进行了batch average，导致损失了更多的granular statistics。基于此，作者提出了Fishr loss：

$$\mathcal{L}_{Fishr}(\theta) = \frac{1}{|\mathcal{E}|}\sum_{e\in\mathcal{E}}||C_e - C||_F^2$$

其中$C_e = \frac{1}{n_e-1}\sum_i^{n_e} = 1(\nabla_\theta l(f_\theta(x_e^i), y_e^i) - g_e)(\nabla_\theta l(f_\theta(x_e^i), y_e^i) - g_e)^\top$为domain e所有样本梯度G_e的协方差，g_e是G_e在domain e的batch上的的均值，C是C_e在所有domain的均值。

2.1、对covariance进行match的主要动机：(这部分引用较多我没来得及看，抽时间补一下)

①设经验费舍信息矩阵(empirical Fisher Information Matrix)$\tilde{F} = G_e G^\top = \sum_{i=1}^n \nabla_\theta log\ p_\theta(y^i|x^i)\nabla_\theta log\ p_\theta(y^i|x^i)^\top$。当损失函数l是负对数似然时，$\tilde{F}$就是$C$.所以它们在驻点是highly related且equivalent的（作者在table6的实验中也验证了这一点）。

②另外，原始版本的费舍信息矩阵$\tilde{F} = \sum_{i=1}^n E_{\hat{y}\sim P_\theta(\cdot|x^i)}[\nabla_\theta log\ p_\theta(\hat{y}|x^i)\nabla_\theta log\ p_\theta(\hat{y}|x^i)^\top]$在一个较弱的假设下，可以在有限的误差内近似海森矩阵$H = \sum_{i=1}^n \nabla_\theta^2 l(f_\theta(x^i), y^i)$.(Theorem 1. Kunster et al. 2019)

③在回归和分类任务上，$\tilde{F} \propto F \approx H$(Thomas et al. 2020)(Singh et al.2020;Li et al.2020)

而match$H$又可以增强泛化能力（下一节中有阐述）。于是，便采用C_e近似H_e。

2.2、 对海森矩阵进行match的动机

这里作者通过推导证明了AND-mask那篇文章的inconsistent score:

$$I^\epsilon(\theta^*) = max\{\max_{|R_A(\theta)-R_A(\theta^*)|\le\epsilon}|R_B(\theta) - R_A(\theta^*)|, max\{\max_{|R_B(\theta)-R_B(\theta^*)|\le\epsilon}|R_A(\theta) - R_B(\theta^*)|\}$$

与domain A、B的海森矩阵H_A、H_B的特征值λ_A、λ_B的关系为:

$$I^\epsilon(\theta^*) \lessapprox \epsilon \cdot max\ max\{\lambda_i^B/\lambda_i^A, \lambda_i^A/\lambda_i^B\}$$

从而证明了当不同domain的loss的海森矩阵具有相同的特征值时，inconsistent score最小，不同domain达成了agreement。

三、plans

接下来一周计划:

1、尽快把domainbed上的ood算法都了解完，这周可能再看1~2个算法

2、再看一些基于梯度的工作，总结一下相关的工作和思路，写一个notes

（Input similarity from the neural network perspective. In NeurIPS, 2019：梯度的激活由对预测的重要性决定）