“基于因果关系和特征对齐的图像分类域泛化模型”的补充材料

明水根，张洪

附录 A. 自适应权重的推导

在深度学习中，模型通常根据标签分类器的输出预测输入的标签。本文中的标签分类器可用条件分布表示为 $p(y|f_y(\boldsymbol{x}))$，其中 $\boldsymbol{z}_y = f_y(\boldsymbol{x})$ 为 ADIVA 中输入标签分类器的特征，故这里关注不同域的标签后验概率 $p(y|f_y(\boldsymbol{x}))$ 在特征对齐的过程中的变化。注意到文中的标签分类器在预测样本标签时不会考虑样本的域属性，换句话说，将所有源域视为同一个域。因此，本文重新定义标签分类器的概率分布表示如下：

$$\bar{p}(y|f_y(\boldsymbol{x})) = \frac{\bar{p}(f_y(\boldsymbol{x})|y)\bar{p}(y)}{\bar{p}(f_y(\boldsymbol{x}))}, \tag{1}$$

其中 $\bar{p}(y|f_y(\boldsymbol{x}))$ 表示由标签分类器输出的标签后验分布，$\bar{p}(f_y(\boldsymbol{x})|y)$、$\bar{p}(f_y(\boldsymbol{x}))$ 和 $\bar{p}(y)$ 分别表示整个训练集的条件特征分布、边际特征分布和标签分布。对于某个源域 $d_i \in D_k$ 中的数据，也有类似的公式表示：

$$p^i(y|f_y(\boldsymbol{x})) = \frac{p^i(f_y(\boldsymbol{x})|y)p^i(y)}{p^i(f_y(\boldsymbol{x}))}. \tag{2}$$

假设在模型训练过程中，通过特征对齐已经将分布 $p(f_y(\boldsymbol{x})|y)$ 对齐了，那么 $p(f_y(\boldsymbol{x}))$ 也会对齐[1]。于是就有如下的方程：

$$\frac{p^i(f_y(\boldsymbol{x})|y)}{p^i(f_y(\boldsymbol{x}))} = \frac{\bar{p}(f_y(\boldsymbol{x})|y)}{\bar{p}(f_y(\boldsymbol{x}))}. \tag{3}$$

结合式(1)-(3)，就得到：

$$p^i(y|f_y(\boldsymbol{x})) = \frac{p^i(y)}{\bar{p}(y)}\bar{p}(y|f_y(\boldsymbol{x})), \tag{4}$$

其中 $p^i(y|f_y(\boldsymbol{x}))$ 是域 d_i 数据的标签后验分布。理论上，本应该由 $p^i(y|f_y(\boldsymbol{x}))$ 与真实标签分布 $p^i(y)$ 一起计算标准分类损失，而 $\bar{p}(y|f_y(\boldsymbol{x}))$ 才是 ADIVA 模型中标签分类器的输出，所以若要用标签分类器的输出 $\bar{p}(y|f_y(\boldsymbol{x}))$ 来计算标签分类损失就得到了如下的自适应权重：

$$\omega^i(y) = \frac{p^i(y)}{\bar{p}(y)}. \tag{5}$$

附录 B. 可识别性定理的证明

在证明 ADIVA 的可识别性定理之前，本文在这里介绍一些预备知识和相关记号。设 $\boldsymbol{x} \in \mathcal{X} \subset \mathbb{R}^d$ 为代表数据的随机变量，$\boldsymbol{u} \in \mathbb{R}^m$ 代表数据中附加的观察变量（如标签、域），$\boldsymbol{z} \in \mathcal{Z} \subset \mathbb{R}^n$ 代表隐变量（潜在因子）。Khemakhem 等[2] 为深度隐变量模型的可识别性提出了一些设定，而最主要的设定是要求隐变量的先验是条件因子分布。此时，VAE 可以由以下公式来描述：

$$p_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{z}|\boldsymbol{u}) = p_{\boldsymbol{g}}(\boldsymbol{x}|\boldsymbol{z}) p_{\boldsymbol{T},\boldsymbol{\lambda}}(\boldsymbol{z}|\boldsymbol{u}), \tag{6}$$

其中 $\boldsymbol{\theta} = (\boldsymbol{g}, \boldsymbol{T}, \boldsymbol{\lambda})$ 均为模型参数，$p_{\boldsymbol{T},\boldsymbol{\lambda}}(\boldsymbol{z}|\boldsymbol{u})$ 为条件因子先验。在 VAE 中，数据 $\boldsymbol{x}$ 可以用隐变量的函数表示为：

$$\boldsymbol{x} = \boldsymbol{g}(\boldsymbol{z}) + \boldsymbol{\epsilon}, \tag{7}$$

其中 $\boldsymbol{\epsilon}$ 表示独立的噪声，相应的概率分布记为 $p_{\boldsymbol{\epsilon}}(\boldsymbol{\epsilon})$。此时，$p_{\boldsymbol{g}}(\boldsymbol{x}|\boldsymbol{z})$ 也可以表示为：

$$p_{\boldsymbol{g}}(\boldsymbol{x}|\boldsymbol{z}) = p_{\boldsymbol{\epsilon}}(\boldsymbol{x} - \boldsymbol{g}(\boldsymbol{z})). \tag{8}$$

VAE 中的条件因子先验 $p_{\boldsymbol{T},\boldsymbol{\lambda}}(\boldsymbol{z}|\boldsymbol{u})$ 一般设定为高斯分布，用公式可表示为：

$$p_{\boldsymbol{T},\boldsymbol{\lambda}}(\boldsymbol{z}|\boldsymbol{u}) = \prod_i \frac{Q_i(z_i)}{Z_i(\boldsymbol{u})} \exp\left[\sum_{j=1}^{k} T_{i,j}(z_i)\, \lambda_{i,j}(\boldsymbol{u})\right], \tag{9}$$

其中，Q_i 为定义在 z_i 所在空间的基础测度，$Z_i(\boldsymbol{u})$ 为相应的归一化常数，$\boldsymbol{T}_i = (T_{i,1}, \ldots, T_{i,k})$ 为与 $\boldsymbol{z}_i$ 相对应的充分统计量。$\boldsymbol{\lambda}_i(\boldsymbol{u}) = (\lambda_{i,1}(\boldsymbol{u}), \ldots, \lambda_{i,k}(\boldsymbol{u}))$ 为充分统计量 $\boldsymbol{T}_i$ 对应的参数。

模型的可识别性可简单理解为参数学习的唯一性，其严谨数学定义如下：

定义 1 (可识别性)**.** *对于模型参数空间中的任意两个参数 $\boldsymbol{\theta}$ 和 $\boldsymbol{\theta}'$，若有 $p_{\boldsymbol{\theta}}(x) = p_{\boldsymbol{\theta}'}(x)$，则蕴含着 $\boldsymbol{\theta} = \boldsymbol{\theta}'$。*

从定义可以看出，如果一个模型是可识别的，那么说明从给定的数据中学习到的参数是唯一的，在这种情况下，模型是有可能学习到真实参数的。

但是证明深度隐变量模型的可识别性是有很大难度的。于是 Khemakhem 等[2] 提出了弱可识别性的概念，可以简单理解为参数能在模去一个等价关系后保证唯一。在这里取这个等价关系为线性变换，则 VAE 中的弱可识别性的严谨数学定义如下：

定义 2 (弱可识别性)**.** *对于 VAE 参数空间中的任意两个参数* $\boldsymbol{\theta} = (\boldsymbol{g}, \boldsymbol{T}, \boldsymbol{\lambda})$ *和* $\tilde{\boldsymbol{\theta}} = (\tilde{\boldsymbol{g}}, \tilde{\boldsymbol{T}}, \tilde{\boldsymbol{\lambda}})$，*若有* $p_{\boldsymbol{\theta}}(x) = p_{\boldsymbol{\theta}'}(x)$，*则存在* A, c，*使得对任意的* $\boldsymbol{x} \in \mathcal{X}$ *有：*

$$\boldsymbol{T}(\boldsymbol{g}^{-1}(\boldsymbol{x})) = A\tilde{\boldsymbol{T}}(\tilde{\boldsymbol{g}}^{-1}(\boldsymbol{x})) + \boldsymbol{c}, \tag{10}$$

其中 A *为一个* $nk \times nk$ *的可逆矩阵，而* $\boldsymbol{c}$ *为一个* nk *维的向量。*

基于以上预备知识，Khemakhem 等[2] 证明了如下的弱可识别性引理：

引理 1 (弱可识别性引理)**.** *假设给定的观测数据集是采样于一个根据*(6)—(9)*式定义的 VAE 模型，其模型参数为* $(\boldsymbol{g}, \boldsymbol{T}, \boldsymbol{\lambda})$。*再假设如下条件成立：*

(i) *集合* $\{\boldsymbol{x} \in \mathcal{X} | \phi_{\boldsymbol{\epsilon}}(\boldsymbol{x}) = 0\}$ *为零测度集，其中* $\phi_{\boldsymbol{\epsilon}}$ *为*(8)*式中的分布* $p_{\boldsymbol{\epsilon}}$ *对应的特征函数。*

(ii) (8)*式中的函数* $\boldsymbol{g}$ *为单射。*

(iii) (9)*式中的任意一个充分统计量* $T_{i,j}$ *几乎处处可微，且在任意一个测度不为 0 的* $\mathcal{X}$ *的子集上都有：*$(T_{i,j})_{1 \le j \le k}$ *是线性独立的，对* $\forall i, 1 \le i \le n$ *成立。*

(iiii) *存在* $nk + 1$ *个不同的点* $\boldsymbol{u}_0, ..., \boldsymbol{u}_{nk}$，*使得这个* $nk \times nk$ *的矩阵*

$$L = (\boldsymbol{\lambda}(\boldsymbol{u}_1) - \boldsymbol{\lambda}(\boldsymbol{u}_0), ..., \boldsymbol{\lambda}(\boldsymbol{u}_{nk}) - \boldsymbol{\lambda}(\boldsymbol{u}_0))$$

是可逆的。

则此模型是弱可识别的。

现在回到图像分类的域泛化研究中。注意到在域泛化任务中，除了图像数据（$\boldsymbol{x}$），还可以观察到对应的标签（y）和域（d）变量，若在隐变量模型中将这两者看做额外观测到的变量（$\boldsymbol{u}$），则可以为隐变量（$\boldsymbol{z}$）设定一个条件因子先验分布。而本文针对域泛化任务提出的 ADIVA 模型，是基于 VAE 框架的。VAE 作为一种经典的深度隐变量模型，显然也适用于以上对模型可识别性进行的讨论与研究。如果 ADIVA 模型满足引理 1中的四条假定，则可以证明 ADIVA 模型具有弱可识别性，故本文提出了如下定理：

定理 2 (可识别性定理)**.** 假设 *ADIVA* 模型中，隐变量 $(\boldsymbol{z}_d, \boldsymbol{z}_y, \boldsymbol{z}_{\boldsymbol{x}})$ 的先验分布属于高斯分布族，且具有如下形式：

$$p(\boldsymbol{z}_d, \boldsymbol{z}_y, \boldsymbol{z}_{\boldsymbol{x}}|y, d) = p(\boldsymbol{z}_d|d)p(\boldsymbol{z}_y|y)p(\boldsymbol{z}_{\boldsymbol{x}}). \tag{11}$$

则 *ADIVA* 是可识别的，至多相差一个线性变换。

证明. 只需要验证 ADIVA 模型满足引理 1中的四条假定即可，验证过程如下：

首先，ADIVA 模型中的解码器（$p_{\boldsymbol{g}(\boldsymbol{x}|\boldsymbol{z})}$）部分和变分自编码器的解码器部分一致，只接受从隐变量分布（$q_\phi(\boldsymbol{z}|\boldsymbol{x})$）中采样出的值为输入，没有添加噪音的过程，此时(8)式中 $p_{\boldsymbol{\epsilon}}$ 可看成一个具有无穷小方差的高斯分布，因此满足引理 1中的第一条假定。

其次，ADIVA 模型中解码器（$\boldsymbol{f}$）由基于 ReLU 激活函数的深度神经网络构成，具有单射性[3]，因此满足引理 1中的第二条假定。

再其次，ADIVA 中对隐变量 $\boldsymbol{z}$ 设定的先验分布属于高斯位置尺度族，因此隐变量每个维度的分量变量都服从单变量高斯分布，则其充分统计量为：

$$\boldsymbol{T}_i(z_i) = (z_i, z_i^2) \quad \forall i, 1 \leq i \leq n.$$

显然满足引理 1中的第三条假定。

最后，在 ADIVA 模型中，$\boldsymbol{\lambda}(\boldsymbol{u})$ 设定为一个浅层神经网络，由 Khemakhem 等[2] 的研究，这样的 $\boldsymbol{\lambda}$ 满足引理（1）中的第四条假定。

由引理 1知，ADIVA 模型是弱可识别的, 即 ADIVA 是可识别的，至多相差一个线性变换。 □

注意到，(11)式中的因子分解形式是有逻辑的。因为在 ADIVA 中，模型学习好的隐变量 $\boldsymbol{z}_d, \boldsymbol{z}_y$ 和 $\boldsymbol{z}_{\boldsymbol{x}}$ 应该是相互独立的，在给定 (y, d) 的条件下。更进一步的讲，$\boldsymbol{z}_d$ 只包含域信息，则有 $\boldsymbol{z}_d \perp y$。类似地有：$\boldsymbol{z}_y \perp d$, $\boldsymbol{z}_{\boldsymbol{x}} \perp (y, d)$。因此，(11)式的形式假定是合理的。

附录 C. PACS 数据集的介绍

PACS 数据集[4] 由来自七个共享类别（狗、大象、长颈鹿、吉他、马、房子和人）的 9991 张 RGB 图像组成，包含四个域（照片、艺术画、卡通和素描）。每个域中每个标签的样本数量如表 1所示。显然，PACS 数据集中存在严重的标签漂移问题。

表 1: PACS 数据集中每一个（域，标签）对的样本数量。

	吉他	房子	长颈鹿	人	马	狗	大象
艺术画	184	295	285	449	201	379	255
卡通	135	288	346	405	324	389	457
素描	608	80	753	160	816	772	740
照片	186	280	182	432	199	189	202

参考文献

[1] K. Zhang, B. Schölkopf, K. Muandet, Z. Wang, Domain adaptation under target and conditional shift, in: 2013 International Conference on Machine Learning, PMLR, 2013, pp. 819–827.

[2] I. Khemakhem, D. Kingma, R. Monti, A. Hyvarinen, Variational autoencoders and nonlinear ICA: A unifying framework, in: 2020 International Conference on Artificial Intelligence and Statistics, PMLR, 2020, pp. 2207–2217.

[3] M. Puthawala, K. Kothari, M. Lassas, I. Dokmanić, M. de Hoop, Globally injective ReLU networks, Journal of Machine Learning Research 23 (105) (2022) 1–55.

[4] D. Li, Y. Yang, Y. Song, T. M. Hospedales, Deeper, broader and artier domain generalization, in: 2017 IEEE International Conference on Computer Vision (ICCV), 2017, pp. 5543–5551.