Supplementary Materials for "Domain Generalization in Deep Learning-Based Image Classification" by

Shuigen Ming and Hong Zhang

Appendix A. Derivation of the weights in the weighted label classification loss

Generally, in deep learning, models predict the label of input based on the output of label classifier, i.e., the conditional distribution $p(y|f_y(\mathbf{x}))$ in this paper. Now we focus on the label posterior $p(y|f_y(\mathbf{x}))$ during the feature alignment. Noticed that the label classifier predicts the labels of samples regardless of their domains, i.e., it treats all the source domains as a single domain. Then, we can reformulate the label posterior as follows:

$$\bar{p}(y|f_y(\mathbf{x})) = \frac{\bar{p}(f_y(\mathbf{x})|y)\bar{p}(y)}{\bar{p}(f_y(\mathbf{x}))}, \tag{S1}$$

where $\bar{p}(y|f_y(\mathbf{x}))$ denotes the output of label classifier, $\bar{p}(f_y(\mathbf{x})|y)$, $\bar{p}(f_y(\mathbf{x}))$, and $\bar{p}(y)$ are the conditional latent distribution, marginal latent distribution, and label distribution of the whole training data, respectively. For data from domain $d_i \in D_k$, we have a similar equation:

$$p^i(y|f_y(\mathbf{x})) = \frac{p^i(f_y(\mathbf{x})|y)p^i(y)}{p^i(f_y(\mathbf{x}))}. \tag{S2}$$

Assume that we have aligned the estimated conditional latent distribution $p(f_y(\mathbf{x})|y)$ in ADIVA, $p(f_y(\mathbf{x}))$ is also aligned[1]. Then, we have:

$$\frac{p^i(f_y(\mathbf{x})|y)}{p^i(f_y(\mathbf{x}))} = \frac{\bar{p}(f_y(\mathbf{x})|y)}{\bar{p}(f_y(\mathbf{x}))}. \tag{S3}$$

Combining Equation (S1)-(S3), we have:

$$p^i(y|f_y(\mathbf{x})) = \frac{p^i(y)}{\bar{p}(y)}\bar{p}(y|f_y(\mathbf{x})), \tag{S4}$$

where $\bar{p}(y|f_y(\mathbf{x}))$ is the label posterior outputted by the label classifier, $p^i(y|f_y(\mathbf{x}))$ is the estimated label posterior of samples from domain d_i. Finally, we have the weights for the adaptive weighting as follows:

$$\omega^i(y) = \frac{p^i(y)}{\bar{p}(y)}. \tag{S5}$$

Appendix B. Proof of the identifiability theorem in the main text

Before proofing the identifiability theorem of ADIVA, we introduce some pre-knowledge and notations here. Let $\mathbf{x} \in \mathcal{X} \subset \mathbb{R}^d$ be the random variable of data, and $\mathbf{u} \in \mathbb{R}^m$ be additionally observed variable (such as

label, domain), and $\mathbf{z} \in \mathcal{Z} \subset \mathbb{R}^n$ be a latent variable. The VAE can be described by the following formula:

$$p_\theta(\mathbf{x}, \mathbf{z}|\mathbf{u}) = p_{\mathbf{g}}(\mathbf{x}|\mathbf{z}) p_{\mathbf{T},\lambda}(\mathbf{z}|\mathbf{u}), \tag{S6}$$

where $\theta = (\mathbf{g}, \mathbf{T}, \lambda)$ are model parameters. Data $\mathbf{x}$ can be represented by $\mathbf{x} = \mathbf{g}(\mathbf{z}) + \epsilon$, where ϵ is an independent noise with the probability distribution $p_\epsilon(\epsilon)$. $p_{\mathbf{g}}(\mathbf{x}|\mathbf{z})$ can also be described as follows:

$$p_{\mathbf{g}}(\mathbf{x}|\mathbf{z}) = p_\epsilon(\mathbf{x} - \mathbf{g}(\mathbf{z})). \tag{S7}$$

In VAE, the prior distributions of latent variables are Gaussian distributions. As stated above, VAE with conditional factorial priors can be identifiable up to a linear transformation[2]. Then, the conditional factorial prior distribution can be described as

$$p_{\mathbf{T},\boldsymbol{\lambda}}(\mathbf{z}|\mathbf{u}) = \prod_i \frac{Q_i(z_i)}{Z_i(\mathbf{u})} \exp\left[\sum_{j=1}^k T_{i,j}(z_i)\, \lambda_{i,j}(\mathbf{u})\right], \tag{S8}$$

where Q_i is the base measure defined on the space where z_i is located, $Z_i(\mathbf{u})$ is a normalization constant, $\mathbf{T}_i = (T_{i,1}, \ldots, T_{i,k})$ are the sufficient statistics corresponding to $\mathbf{z}_i$, and $\lambda_i(\mathbf{u}) = (\lambda_{i,1}(\mathbf{u}), \ldots, \lambda_{i,k}(\mathbf{u}))$ are parameters correspondig to each sufficient statistic in $\mathbf{T}_i$.

Based on the above pre-knowledge, Khemakhem et al.[2] proved the following lemma:

Lemma 0.1. *Given data sampled from a VAE model defined by Equation* (S6)-(S8)*. Assume the following conditions hold:*

1. *The set* $\{\mathbf{x} \in \mathcal{X} | \phi_\epsilon(\mathbf{x}) = 0\}$ *has zero measure, where* ϕ_ϵ *is the characteristic function of the distribution* p_ϵ *defined in Equation* (S7)*.*

2. *The function* $\mathbf{g}$ *in Equation* (S7) *is injective.*

3. *The sufficient statistics* $T_{i,j}$ *in Equation* (S8) *are differentiable almost everywhere, and* $\mathbf{T}_i$ *are linearly independent on any nonzero measure subset of* $\mathcal{X}$*.*

4. *There exists* $nk + 1$ *distinct points* $\mathbf{u}_0, \ldots, \mathbf{u}_{nk}$ *such that the* $nk \times nk$ *matrix*

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

is invertible.

Then, the parameters $(\mathbf{g}, \mathbf{T}, \lambda)$ *are identifiable up to a linear transformation.*

Noticed that in the domain generalization task, the label (y) and domain (d) of a sample can be observed except for the image ($\mathbf{x}$). Here the label and domain are the additional observed variables, i.e., $\mathbf{u} = (y, d)$. We can easily assume a conditional factorial prior distribution for the latent variable ($\mathbf{z}$). ADIVA is based

on the VAE framework with a conditional factorial prior distribution, so it obviously applies to the above discussion. If ADIVA satisfies the four conditions in Lemma 0.1, it is identifiable. Therefore, the following theorem is proposed:

Theorem 0.2. *Assume that the prior of the latent vector* $(\mathbf{z}_d, \mathbf{z}_y, \mathbf{z}_\mathbf{x})$ *in ADIVA belongs to the Gaussian distribution family and has the following form:*

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

Then ADIVA is identifiable up to a linear transformation.

Proof. We just verify that ADIVA satisfies the four conditions in Lemma 0.1, the main process of proof is as follows:

- The decoder ($p_\mathbf{g}(\mathbf{x}|\mathbf{z})$) of ADIVA only takes the latent vector ($\mathbf{z}$) as input without adding noise. Then, the distribution p_ϵ in Equation (S7) can be regarded as a Gaussian distribution with an infinitesimal variance, which satisfies the first condition in Lemma 0.1.

- The decoder ($\mathbf{g}$) of ADIVA is a neural network with ReLu activation function, which means $\mathbf{g}$ is injective according to Puthawala et al.[3]. The second condition in Lemma 0.1 is confirmed.

- The prior distributions of latent variables ($\mathbf{z}$) in ADIVA belong to Gaussian location-scale family, and they can be factorized into the product of marginal distributions. Then, the sufficient statistics can be easily obtained as follows:

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

  It satisfies the third condition in Lemma 0.1.

- In ADIVA, $\lambda(\mathbf{u})$ is a shallow feedforward neural network. According to the study in Khemakhem et al.[2], such $\lambda(\mathbf{u})$ satisfies the fourth condition in Lemma 0.1.

Above all, ADIVA is identifiable up to a linear transformation according to Lemma 0.1.

□

Appendix C. A summary of the PACS dataset

The PACS dataset consists of 9991 RGB images from seven shared classes (dog, elephant, giraffe, guitar, horse, house, and person) with four domains (Photo, Art Painting, Cartoon, and Sketch). The detailed samples in each domain and each label are shown in Table S1. Evidently, the PACS dataset suffered a lot from the target shift.

Table S1: A summary of the PACS dataset.

	dog	elephant	giraffe	guitar	horse	house	person
Art painting	379	255	285	184	201	295	449
Cartoon	389	457	346	135	324	288	405
Sketch	772	740	753	608	816	80	160
Photo	189	202	182	186	199	280	432

References

[1] Kun Zhang, Bernhard Schölkopf, Krikamol Muandet, and Zhikun Wang. Domain adaptation under target and conditional shift. In: *2013 International Conference on Machine Learning*, 819–827. PMLR, **2013**.

[2] Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvarinen. Variational autoencoders and nonlinear ICA: A unifying framework. In: *2020 International Conference on Artificial Intelligence and Statistics*, 2207–2217. PMLR, **2020**.

[3] Michael Puthawala, Konik Kothari, Matti Lassas, Ivan Dokmanić, and Maarten de Hoop. Globally injective ReLU networks. **2022**, *Journal of Machine Learning Research*, *23* (105): 1–55.