Supplementary Materials: Gradient Distribution Alignment Certificates Better Adversarial Domain Adaptation

Zhiqiang Gao[1], Shufei Zhang[1], Kaizhu Huang*[1], Qiufeng Wang[1], and Chaoliang Zhong[2]

[1]School of Advanced Technology, Xi'an Jiaotong-liverpool University, Suzhou, China
{zhiqiang.gao, shufei.zhang, kaizhu.huang, qiufeng.wang}@xjtlu.edu.cn
[2]Fujitsu RD Center, Beijing, China, clzhong@fujitsu.com

1. Proof of Theorems 1 and 2

Theorem 1 Let G be a fixed representation function from $\mathcal{X}$ to $\mathcal{F}$, and $\mathcal{H}$ be a hypothesis space of VC-dimension d. If a random labeled sample of size m is generated by applying G to a $\mathcal{D}_s$ - i.i.d. The feature f is drawn from $\tilde{\mathcal{D}_S}$ or $\tilde{\mathcal{D}_T}$ with corresponding label y. Denote that $\tilde{\mathcal{U}_S}, \tilde{\mathcal{U}_T}$ are the set of unlabeled samples of size m' each, drawn from $\tilde{\mathcal{D}_S}$ and $\tilde{\mathcal{D}_T}$ respectively. Then with the probability at least $1-\delta$ (over the choice of the samples), for every $C \in \mathcal{H}$:

$$\begin{aligned} \epsilon_T(C) \leq \quad & \hat{\epsilon}_S(C) + \lambda + d_{\nabla}\left(\tilde{\mathcal{U}_S}, \tilde{\mathcal{U}_T}\right) \\ & + \frac{4}{m}\sqrt{\left(d \log \frac{2em}{d} + \log \frac{4}{\delta}\right)} \\ & + 4\sqrt{\frac{d \log\left(2m'\right) + \log\left(\frac{4}{\delta}\right)}{m'}} \\ & = const + d_{\nabla}\left(\tilde{\mathcal{U}_S}, \tilde{\mathcal{U}_T}\right) \end{aligned} \tag{1}$$

where $\hat{\epsilon}_S(C)$ is empirical error of source samples, λ is a very small constant, e represents the base of the natural logarithm, $d_{\nabla}\left(\tilde{\mathcal{U}_S}, \tilde{\mathcal{U}_T}\right) = a \sup_{D_g \mathcal{H}_D} \left| E_{\boldsymbol{f} \in \tilde{\mathcal{U}_S}} D_g(\nabla_{\boldsymbol{f}} \mathcal{L}) - E_{\boldsymbol{f} \in \tilde{\mathcal{U}_T}} D_g(\nabla_{\boldsymbol{f}} \mathcal{L}) \right|$ is the introduced ∇-distance, D_g is the discriminator and $a = \frac{1}{\min_{C(\boldsymbol{f}) \in [0,1]} \nabla_C \mathcal{L}(C(\boldsymbol{f}), y)}$. Here $\mathcal{L}(\cdot)$ denotes the loss function.

Theorem 2 When $a \leq 1$, our method can obtain a tighter upper bound than traditional domain adaptation methods:

$$const + d_{\nabla}\left(\tilde{\mathcal{U}_S}, \tilde{\mathcal{U}_T}\right) \leq const + d_{\mathcal{H}}\left(\tilde{\mathcal{U}_S}, \tilde{\mathcal{U}_T}\right), \text{where}$$

$$d_{\mathcal{H}}\left(\tilde{\mathcal{U}_S}, \tilde{\mathcal{U}_T}\right) = \sup_{D_g \in \mathcal{H}_D} \left| E_{\boldsymbol{f} \in \tilde{\mathcal{U}_S}} D_g(\boldsymbol{f}) - E_{\boldsymbol{f} \in \tilde{\mathcal{U}_T}} D_g(\boldsymbol{f}) \right|.$$

Proof According to [3, 1], the domian discrepancy is defined as $\left|\epsilon_T\left(C, C^*\right) - \epsilon_S\left(C, C^*\right)\right|$ where C^* is the ideal joint hypothesis defined as $C^* = \operatorname{argmin}_C \epsilon_S(C) + \epsilon_T(C)$. We assume that the most misclassified examples are classified as wrong classes with low confidence (close to the decision boundary). Thus, there exists a small perturbation vector $\boldsymbol{\gamma} \in \mathcal{F}$ for each feature $\boldsymbol{f}$ such that $C(\boldsymbol{f} + \boldsymbol{\gamma}) = y$, where y is the true label. The domain discrepancy can be reformulated as

$$\begin{aligned} & \left|\epsilon_T\left(C, C^*\right) - \epsilon_S\left(C, C^*\right)\right| \\ & = \left|\mathrm{E}_{\boldsymbol{f} \sim \tilde{\mathcal{D}_T}}\left[C(\boldsymbol{f}) \neq C^*(\boldsymbol{f})\right] - \mathrm{E}_{\boldsymbol{f} \sim \tilde{\mathcal{D}_S}}\left[C(\boldsymbol{f}) \neq C^*(\boldsymbol{f})\right]\right| \\ & = \left|\mathrm{E}_{\boldsymbol{f} \sim \tilde{\mathcal{D}_T}}\left[C(\boldsymbol{f}) - C^*(\boldsymbol{f})\right] - \mathrm{E}_{\boldsymbol{f} \sim \tilde{\mathcal{D}_S}}\left[C(\boldsymbol{f}) - C^*(\boldsymbol{f})\right]\right| \\ & = \left|\mathrm{E}_{\boldsymbol{f} \sim \tilde{\mathcal{D}_T}}\left|y - C^*(\boldsymbol{f}) + C(\boldsymbol{f}) - C(\boldsymbol{f} + \boldsymbol{\gamma})\right|\right. \\ & \quad \left. - \mathrm{E}_{\boldsymbol{f} \sim \tilde{\mathcal{D}_S}}\left|y - C^*(\boldsymbol{f}) + C(\boldsymbol{f}) - C(\boldsymbol{f} + \boldsymbol{\gamma})\right|\right| \\ & \leq \left|\mathrm{E}_{\boldsymbol{f} \sim \tilde{\mathcal{D}_T}}\left|y - C^*(\boldsymbol{f})\right| - \mathrm{E}_{\boldsymbol{f} \sim \tilde{\mathcal{D}_S}}\left|y - C^*(\boldsymbol{f})\right|\right. \\ & \quad + \mathrm{E}_{\boldsymbol{f} \sim \tilde{\mathcal{D}_T}}\left|C(\boldsymbol{f}) - C(\boldsymbol{f} + \boldsymbol{\gamma})\right| \\ & \quad \left. - \mathrm{E}_{\boldsymbol{f} \sim \tilde{\mathcal{D}_S}}\left|C(\boldsymbol{f}) - C(\boldsymbol{f} + \boldsymbol{\gamma})\right|\right| \\ & \leq \left|\mathrm{E}_{\boldsymbol{f} \sim \tilde{\mathcal{D}_T}}\left|y - C^*(\boldsymbol{f})\right| - \mathrm{E}_{\boldsymbol{f} \sim \tilde{\mathcal{D}_S}}\left|y - C^*(\boldsymbol{f})\right|\right| \\ & \quad + \left|\mathrm{E}_{\boldsymbol{f} \sim \tilde{\mathcal{D}_T}}\left|C(\boldsymbol{f}) - C(\boldsymbol{f} + \boldsymbol{\gamma})\right|\right. \\ & \quad \left. - \mathrm{E}_{\boldsymbol{f} \sim \tilde{\mathcal{D}_S}}\left|C(\boldsymbol{f}) - C(\boldsymbol{f} + \boldsymbol{\gamma})\right|\right|. \end{aligned} \tag{2}$$

Since the ideal joint hypothesis C^* is expected to minimize the joint error on two domains, $\left|\mathrm{E}_{\boldsymbol{f} \sim \tilde{\mathcal{D}_T}}\left|y - C^*(\boldsymbol{f})\right| - \mathrm{E}_{\boldsymbol{f} \sim \tilde{\mathcal{D}_S}}\left|y - C^*(\boldsymbol{f})\right|\right|$ should be a very small constant and we treat it as zero. Then, we

*Corresponding author: Kaizhu Huang.

reformulate the bound of target error in [3, 1] as

$$\begin{aligned} \epsilon_T(C) &\leq \epsilon_S + \lambda + \left|\epsilon_T\left(C, C^*\right) - \epsilon_S\left(C, C^*\right)\right| \\ &\leq \epsilon_S + \lambda \\ &+ |\mathrm{E}_{\boldsymbol{f}\sim\tilde{\mathcal{D}_T}}\,|C(\boldsymbol{f}+\boldsymbol{\gamma}) - C(\boldsymbol{f})| \\ &- \mathrm{E}_{\boldsymbol{f}\sim\tilde{\mathcal{D}_S}}\,|C(\boldsymbol{f}+\boldsymbol{\gamma}) - C(\boldsymbol{f})|\,|. \end{aligned} \tag{3}$$

we can find $b = \min\limits_{C(\boldsymbol{f})\in[0,1]} \nabla_C \mathcal{L}(C(\boldsymbol{f}), y)$ and ignore the second and higher order terms in Taylor's expansion of $\mathcal{L}(C(\boldsymbol{f}+\boldsymbol{\gamma}), y)$ ($\boldsymbol{\gamma}$ are small and we can regularize the function to be with the small second order term). Then, we have

$$\begin{aligned} b|C(\boldsymbol{f}+\boldsymbol{\gamma}) - C(\boldsymbol{f})| &\leq |\mathcal{L}(C(\boldsymbol{f}+\boldsymbol{\gamma}, y)) - \mathcal{L}(C(\boldsymbol{f}, y))| \\ &= |\boldsymbol{\gamma}\nabla_{\boldsymbol{f}}\mathcal{L}|\,. \end{aligned} \tag{4}$$

By defining the $a = \frac{1}{b}$, we obtain

$$|C(\boldsymbol{f}+\boldsymbol{\gamma}) - C(\boldsymbol{f})| \leq a|\mathcal{L}(C(\boldsymbol{f}+\boldsymbol{\gamma}, y)) - \mathcal{L}(C(\boldsymbol{f}, y))|, \tag{5}$$

such that

$$\begin{aligned} &|\mathrm{E}_{\boldsymbol{f}\sim\tilde{\mathcal{D}_T}}\,|C(\boldsymbol{f}+\boldsymbol{\gamma}) - C(\boldsymbol{f})| - \mathrm{E}_{\boldsymbol{f}\sim\tilde{\mathcal{D}_S}}\,|C(\boldsymbol{f}+\boldsymbol{\gamma}) - C(\boldsymbol{f})|\,| \\ &= a|\mathrm{E}_{\boldsymbol{f}\sim\tilde{\mathcal{D}_S}}\,|\boldsymbol{\gamma}\nabla_{\boldsymbol{f}}\mathcal{L}| - \mathrm{E}_{\boldsymbol{f}\sim\tilde{\mathcal{D}_T}}\,|\boldsymbol{\gamma}\nabla_{\boldsymbol{f}}\mathcal{L}|\ \ .| \end{aligned} \tag{6}$$

Accordingly, the probabilistic bound of target error is changed to

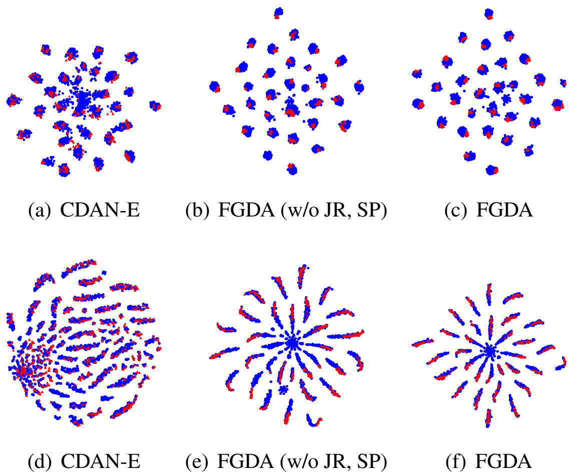
$$\begin{aligned} \epsilon_T(C) &\leq \epsilon_S + \lambda + \left|\epsilon_T\left(C, C^*\right) - \epsilon_S\left(C, C^*\right)\right| \\ &\leq \epsilon_S + \lambda \\ &\quad + a|\mathrm{E}_{\boldsymbol{f}\sim\tilde{\mathcal{D}_S}}\,|\boldsymbol{\gamma}\nabla_{\boldsymbol{f}}\mathcal{L}| - \mathrm{E}_{\boldsymbol{f}\sim\tilde{\mathcal{D}_T}}\,|\boldsymbol{\gamma}\nabla_{\boldsymbol{f}}\mathcal{L}|\,| \\ &\leq \epsilon_S + \lambda + d_{\nabla}(\tilde{\mathcal{D}}_S, \tilde{\mathcal{D}}_T), \end{aligned} \tag{7}$$

where the $d_{\nabla}(\tilde{\mathcal{D}}_S, \tilde{\mathcal{D}}_T)$ is a proposed ∇-distance to upper bound the target error. Since $|C(\boldsymbol{f}+\boldsymbol{\gamma}) - C(\boldsymbol{f})| \in [0, 1]$, we define a distance

$$\begin{aligned} &d_{\nabla}\left(\tilde{\mathcal{D}}_S, \tilde{\mathcal{D}}_T\right) = \\ &a \sup_{D_g\in\mathcal{H}_D} \left|E_{\boldsymbol{f}\in\mathcal{D}_S} D_g(\nabla_{\boldsymbol{f}}\mathcal{L}) - E_{\boldsymbol{f}\in\mathcal{D}_T} D_g(\nabla_{\boldsymbol{f}}\mathcal{L})\right|, \end{aligned} \tag{8}$$

where D_g is a binary classifier of a hypothesis space $\mathcal{H}_D$ over $\nabla_{\boldsymbol{f}}\mathcal{L}$.

Next, we prove that the proposed ∇-distance enables a tighter bound of target error compared with original theory [1]. Considering that $\nabla_{\boldsymbol{f}}\mathcal{L}$ can be obtained by a function $\nabla_{\boldsymbol{f}}\mathcal{L} = D_1(\boldsymbol{f})$ with respect to $\boldsymbol{f}$ ($D_1 \in \mathcal{H}_1 : \mathbb{R}^d \to \mathbb{R}^d$),


(a) CDAN-E (b) FGDA (w/o JR, SP) (c) FGDA

(d) CDAN-E (e) FGDA (w/o JR, SP) (f) FGDA

Figure 1. t-SNE visualizations of feature distribution and gradient distribution for CDAN-E, proposed FGDA (w/o JR, SP) and FGDA on the A→W task of the Office-31 dataset. Blue and red points denote the source and target domain samples respectively.

when $a \leq 1$, we can re-write Equ. 7 as

$$\begin{aligned} \epsilon_T(C) &\leq \epsilon_S + \lambda + \left|\epsilon_T\left(C, C^*\right) - \epsilon_S\left(C, C^*\right)\right| \\ &\leq \epsilon_S + \lambda \\ &+ a \sup_{D_g\in\mathcal{H}_D} |E_{\boldsymbol{f}\in\mathcal{D}_S} D_g(D_1(\boldsymbol{f})) \\ &- E_{\boldsymbol{f}\in\mathcal{D}_T} D_g(D_1(\boldsymbol{f}))| \\ &\leq \epsilon_S + \lambda \\ &+ a \sup_{D_g\in\mathcal{H}_D, D_f\in\mathcal{H}_1} |E_{\boldsymbol{f}\in\mathcal{D}_S} D_g(D_f(\boldsymbol{f})) \\ &- E_{\boldsymbol{f}\in\mathcal{D}_T} D_g(D_f(\boldsymbol{f}))| \\ &\leq \epsilon_S + \lambda + d_{\mathcal{H}}(\tilde{\mathcal{D}}_S, \tilde{\mathcal{D}}_T). \end{aligned} \tag{9}$$

where $D_f \in \mathcal{H}_1 : \mathbb{R}^d \to \mathbb{R}^d$ and $d_{\mathcal{H}}\left(\tilde{\mathcal{U}}_S, \tilde{\mathcal{U}}_T\right) = a \sup_{D_g\in\mathcal{H}_D} |E_{\boldsymbol{f}\in\tilde{\mathcal{U}_S}} D_g(\boldsymbol{f}) - E_{\boldsymbol{f}\in\tilde{\mathcal{U}_T}} D_g(\boldsymbol{f})|$.

2. Visualization Results

We provide additional visualization results to demonstrate the effectiveness of the proposed method, as shown in Fig. 1. Although the CDAN-E [3] conditions on a discriminator on the joint variable of feature and classifier's output, its feature and gradient distribution are still not well aligned. Utilizing merely gradient discriminator in FGDA (w/o JR, SP), the gradient distribution discrepancy is reduced such that the domain shift is reduced. In FGDA, target discriminative feature is obtained due to the gradient regularization and soft supervised learning of gradient alignment, comparing with CDAN-E.

r (in λ_1)	A-W	D-W	A-D	D-A	W-A
0.25	95.2	98.6	94.8	75.8	76.5
0.50	95.3	98.9	95.4	77.7	76.1
0.75	**95.7**	**98.9**	**95.6**	77.6	76.0
1.00	95.1	98.7	95.4	**78.1**	**76.5**
1.25	94.5	98.7	95.2	76.4	76.4

Table 1. Accuracy (%) of FGDA+MDD on Office-31 for UDA (ResNet-50)

3. Sensitivity Analysis for Adversarial Loss

We provide the sensitivity analysis for balancing parameters of adversarial learning loss, such as λ_1 and λ_3. Similar to the original adversarial domain adaptation [2], the above balancing parameters are changed with the following formula

$$\lambda = \frac{2}{1 + \exp(-\gamma \cdot \frac{r \cdot t}{10000})} - 1, \tag{10}$$

where $\gamma = 10$ as empirically used in the experiments, t is the number of iteration in training, and r is a defined parameter for controlling the combination ratio between our adversarial loss $\tilde{\mathcal{L}}_{adv}$ and feature-based adversarial loss $\mathcal{L}_{fada}$. For convenience, in part of implementation, $\lambda_3 = 0.5$ means $r = 0.5$ in λ_3.

Here, we couple our FGDA with MDD [4] and conduct experiments on the Office-31 dataset to analyze the sensitivity of λ_1 and λ_3. Specifically, we select r of λ_1 from $[0.25, 0.50, 0.75, 1.00, 1.25]$, and keep $r = 0.5$ of λ_3. As shown in Table 1, there is a positive correlation between accuracy and value of r in λ_1, and best results are achieved when $r \leq 1$.

References

[1] S. Ben-David, J. Blitzer, K. Crammer, and F. Pereira. Analysis of representations for domain adaptation. In *Advances in neural information processing systems*, pages 137–144, 2007.

[2] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030, 2016.

[3] M. Long, Z. Cao, J. Wang, and M. I. Jordan. Conditional adversarial domain adaptation. In *Advances in neural information processing systems*, pages 1640–1650, 2018.

[4] Y. Zhang, T. Liu, M. Long, and M. Jordan. Bridging theory and algorithm for domain adaptation. In *International Conference on Machine Learning*, pages 7404–7413, 2019.