Supplemental Material: Deep Least Squares Alignment for Unsupervised Domain Adaptation

Youshan Zhang
https://sites.google.com/view/youshanzhang
Brian D. Davison
http://www.cse.lehigh.edu/~brian/

Computer Science and Engineering
Lehigh University
Bethlehem, PA, USA

1 Model algorithm

The details about the pipeline of our method can be found in the algorithm below.

Algorithm 1 Deep Least Squares Alignment (DLSA) for Unsupervised Domain Adaptation. I is the number of iterations.

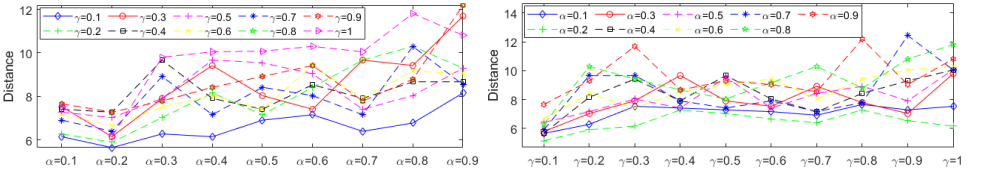
Input: $\mathcal{N}_\mathcal{S}$ labeled source samples $\mathcal{D}_\mathcal{S} = \{\mathcal{X}_\mathcal{S}^i, \mathcal{Y}_\mathcal{S}^i\}_{i=1}^{\mathcal{N}_\mathcal{S}}$ and $\mathcal{N}_\mathcal{T}$ unlabeled target samples $\mathcal{D}_\mathcal{T} = \{\mathcal{X}_\mathcal{T}^j\}_{j=1}^{\mathcal{N}_\mathcal{T}}$
Output: optimized parameters of classifier $\mathcal{F}$
repeat
  for $t = 1$ **to** T **do**
    Derive source and target batch from $\mathcal{D}_\mathcal{S}$ and $\mathcal{D}_\mathcal{T}$
    Initialize $\mathcal{F}$ using Eq. 1 and Eq. 8
    Generate pseudo-labels ($\mathcal{Y}_{\mathcal{T}_\mathcal{P}}$) for the target domain with the trained classifier $\mathcal{F}$
    Minimize the conditional adaptation loss using Eq. 12
  end for
until converged

2 Theoretical analysis

In this section, we theoretically show the error bound of the target domain for the proposed DLSA with domain adaptation theory [1] by three elements: (1) $R_\mathcal{S}(h)$: source domain risk; (2) $d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_\mathcal{S}, \mathcal{D}_\mathcal{T})$: the divergence between two domains; (3): β: adaptability.

In our DLSA, $R_\mathcal{S}(h)$ can be small by training the labeled source domain in Eq. 1 of main file. During the training, the domain discrepancy distance $d_{\mathcal{H}\Delta\mathcal{H}}$ can be minimized by reducing the divergence between the distribution of latent feature space of two domains. Specifically, $d_{\mathcal{H}\Delta\mathcal{H}} = \mathcal{L}_\mathcal{M} + \mathcal{L}_\mathcal{C} = \min(\theta_\mathcal{M} + \mathcal{B}_\mathcal{M} + \frac{1}{C}\sum_{c=1}^{C}(\theta_\mathcal{C}^c + \mathcal{B}_\mathcal{C}^c))$. Ideally, the domain

discrepancy distance will be perfectly removed if all parameters $(\theta_{\mathcal{M}}, \mathcal{B}_{\mathcal{M}}, \theta^c_{\mathcal{C}}, \mathcal{B}^c_{\mathcal{C}})$ are close to 0. However, it can be achieved if and only if $\mathcal{D}_{\mathcal{S}} = \mathcal{D}_{\mathcal{T}}$. Therefore, minimizing $d_{\mathcal{H}\Delta\mathcal{H}}$ is equivalent to minimizing $(\theta_{\mathcal{M}}, \mathcal{B}_{\mathcal{M}}, \theta^c_{\mathcal{C}}, \mathcal{B}^c_{\mathcal{C}})$, which represents that the marginal and conditional distributions alignment can be parameterized by these four key parameters. With the minimized four components, an ideal hypothesis exists with a small β and the error bound of $R_{\mathcal{T}}(h)$ can be minimized.



(a) Effects of different α (b) Effects of different γ

Figure 1: Effects of parameter α and γ on domain discrepancy distance of task A→D in Office-31. Best viewed in color.

Table 1: Least squares estimated parameters of task C→W on Office + Caltech-10 dataset ($\mathcal{M}$: marginal distribution, $\mathcal{C}^c$: conditional distribution of each class c, where $c = \{1, 2, 3, \cdots, 10\}$).

Distribution	$\mathcal{M}$	$\mathcal{C}^1$	$\mathcal{C}^2$	$\mathcal{C}^3$	$\mathcal{C}^4$	$\mathcal{C}^5$	$\mathcal{C}^6$	$\mathcal{C}^7$	$\mathcal{C}^8$	$\mathcal{C}^9$	$\mathcal{C}^{10}$	**Ave.**
θ	0.074	29.786	35.871	24.615	16.451	7.202	11.924	37.778	0.355	33.137	16.594	21.371
$\mathcal{B}$	1.415	33.137	9.875	17.656	3.622	3.626	0.250	15.820	3.097	3.358	16.237	10.668

3 Parameter analysis

The penalty parameters α and γ are two hyperparameters in our model. We search the optimal hyperparameter values by randomly selecting the task A→D in Office-31 datasets in Fig. 1. Considering no labels in the target domain, we then investigate how different parameters affect the domain discrepancy distance. As mentioned in Sec. 2, the domain distance = $(\theta_{\mathcal{M}} + \mathcal{B}_{\mathcal{M}} + \frac{1}{C}\sum_{c=1}^{C}(\theta^c_{\mathcal{C}} + \mathcal{B}^c_{\mathcal{C}}))$. Therefore, we can select the optimal parameters to minimize the distance. α and γ are tested for each value in the groups $\{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$, and $\{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1\}$. From the results in Fig. 1, we can clearly observe that the domain discrepancy distance varied with different choice of α and γ. It indicates the necessity to consider the different effects between angle and intercept differences and the balance between marginal and conditional distributions. We find that when $\alpha = 0.2$ (Fig. 1(a)) and $\gamma = 0.1$ (Fig. 1(b)), the distance achieves the minimum value. Therefore, setting $(\alpha, \gamma) = (0.2, 0.1)$ provides optimal parameters in DLSA.

4 Dimensionality validation

We assume that the first dimension has a linear relationship with the remaining $d-1$. However, other dimensions can be selected (e.g., the second dimension has a relationship with the remaining dimensions). We now investigate whether choosing a different dimension will affect the domain discrepancy distance. As shown in Fig. 2, our model is robust to the selection of dimensions, as all curves are generally flat and stable in all tasks. Therefore, we can select the first dimension as the independent variable and the remaining $d-1$ dimensions as dependent variables.

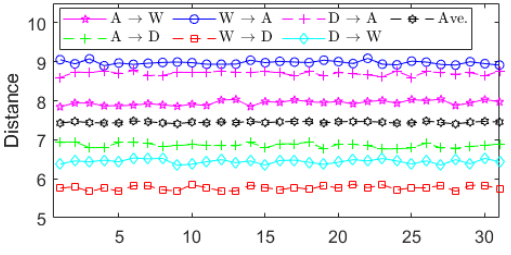


Figure 2: The sensitivity of dimensions on domain discrepancy distance using Office-31 dataset (x-axis: different dimensions for the least squares). Best viewed in color.

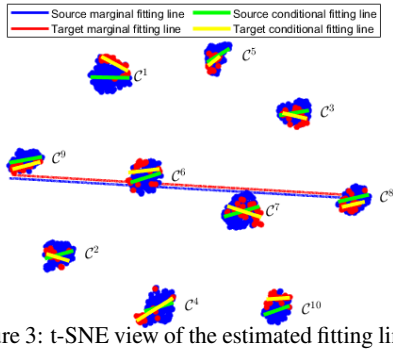


Figure 3: t-SNE view of the estimated fitting lines of marginal and conditional distributions of task C→W in Office + Caltech-10 dataset, which corresponds to Tab. 1. (Blue dots represent the source and red dots, the target domain).

5 Distribution alignment estimated parameters

To statistically explore how our proposed least squares method affects the marginal and conditional distribution alignment process, we show the statistics of the task C→W in the Office + Caltech-10 dataset. Fig. 3 shows the estimated fitting lines, and Tab. 1 lists values of estimated parameters of marginal and conditional distributions. Here, we return the value of angle $\theta_{\mathcal{M}}$ and $\theta_{\mathcal{C}}^{c}$ in degrees and show acute angle between the source and target fitting lines, which is in the range of $[0,90]$. We can find that the source marginal fitting line is almost overlapping with the target marginal fitting line, and the overall classes are discriminated against each other. Also, the estimated $\theta_{\mathcal{M}}$ and $\mathcal{B}_{\mathcal{M}}$ are small in Tab. 1, which represents good alignment of the marginal distributions. However, we notice that the estimated conditional $\theta_{\mathcal{C}}$ and $\mathcal{B}_{\mathcal{C}}$ are relatively large in Tab. 1. Although samples of ten classes in domain W are overlapping with domain C as shown in Fig. 3, the source and target fitting lines of each class (green and yellow line) are not close to each other. The underlying reason is that each class of the target domain (W) is only a part of the source domain (C). Hence, the estimated fitting line of each class in the target domain can be different from that in the source domain.

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.