Moment Matching for Multi-Source Domain Adaptation

Shubhangi Jain, Shoumik Majumdar
Boston University

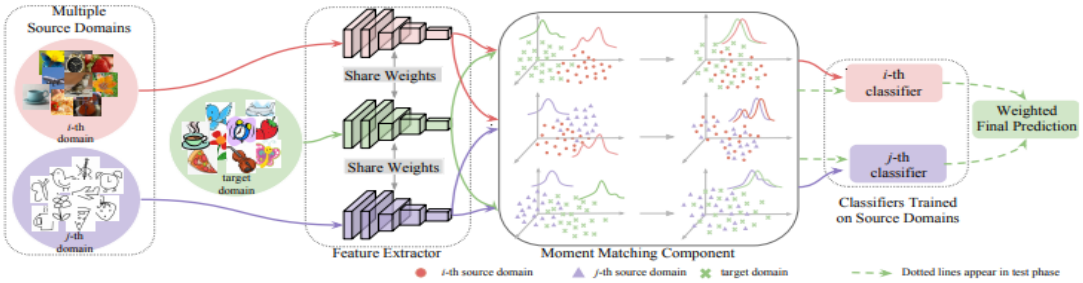
Abstract

Conventional unsupervised domain adaptation (UDA) assumes that training data are sampled from a single domain. This neglects the more practical scenario where training data are collected from multiple sources, requiring multi-source domain adaptation. To address this problem, a paper was published "Moment Matching for Multi-Source Domain Adaptation"[1] in which a dataset called Domain Net, which contain six domains and 0.6 million images distributed among 345 categories, was created. In the same paper they proposed a new deep learning approach, Moment Matching for Multi-Source Domain Adaptation (M3SDA). As part of our project, we have extended work of this paper. We have written code for image's dataset and changed the distance function used in original paper to calculate loss. We have used two different distance functions to train our model : **Dynamic Partial Distance Function** and **Mahalanobis distance**.

Introduction

Generalizing models learned on one visual domain to novel domains has been a major obstacle in the quest for universal object recognition. The performance of the learned models degrades significantly when testing on novel domains due to the presence of domain shift. In the past, transfer learning and domain adaptation methods have been proposed to mitigate the domain gap. For example, several UDA methods incorporate Maximum Mean Discrepancy loss into a neural network to diminish the domain discrepancy; other models introduce different learning schema to align the source and target domains, including aligning second order correlation,moment matching, adversarial domain confusion and GAN-based alignment . However, before proposal of "Moment Matching for Multi-Source Domain Adaptation" most of UDA methods assumed that source samples are collected from a single domain. This assumption neglected the more practical scenarios where labeled images are typically collected from multiple domains. For example, the training images can be taken under different weather or lighting conditions, share different visual cues, and even have different modalities .

We have taken a more difficult but practical problem of knowledge transfer from multiple distinct domains to one unlabeled target domain. The main challenges we faced while implementing this: (1) the source data has multiple domains, which hampers the effectiveness of mainstream single UDA method; (2) source domains also possess domain shift with each other; (3) the lack of large-scale multi domain dataset hinders the development of MSDA models. In the context of MSDA, some theoretical analysis has been proposed for multi-source domain adaptation (MSDA). Ben-David et al pioneer this direction by introducing an HΔH-divergence between the weighted combination of source domains and target domain. More applied works use an adversarial discriminator to align the multi-source domains with the target domain. However, these works focus only on aligning the source domains with the target, neglecting the domain shift between the source domains. Moreover, $H\Delta H$ divergence based analysis does not directly correspond to moment matching approaches. In terms of data, research has been hampered due to the lack of large-scale domain adaptation datasets, as state-of the-art datasets contain only a few images or have a limited number of classes. Many domain adaptation models exhibit saturation when evaluated on these datasets. For example, many methods achieve $\approx 90\%$ accuracy on the popular Office dataset; Self-Ensembling reports $\approx 99\%$ accuracy on the "Digit-Five" dataset and $\approx 92\%$ accuracy on Syn2Real dataset.



In our code we have taken component, introduced in Moment Matching paper[1], called M3SDA to tackle MSDA task by aligning the source domains with the target domain, and aligning the source domains with each other simultaneously. In the procedure multiple complex adversarial training procedures were disposed, and were directly aligned the moments of their deep feature distributions.

Related Work

Single-source UDA Over the past decades, various single source UDA methods have been proposed. These methods can be taxonomically divided into three categories. The first category is the discrepancy-based DA approach, which utilizes different metric learning schemas to diminish the domain shift between source and target domains. Inspired by the kernel two-sample test, Maximum Mean Discrepancy (MMD) is applied to reduce distribution shift in various methods. Other commonly used methods include correlation alignment , Kullback-Leibler (KL) divergence, and H divergence. The second category is the adversarial-based approach. A domain discriminator is leveraged to encourage the domain confusion by an adversarial objective. Among these approaches, generative adversarial networks are widely used to learn domain-invariant features as well to generate fake source or target data. Other frameworks utilize only adversarial loss to bridge two domains. The third category is reconstruction-based, which assumes the data reconstruction helps the DA models to learn domain-invariant features. The reconstruction is obtained via an encoder-decoder or a GAN discriminator, such as dualGAN , cycle-GAN, disco-GAN, and CyCADA . Though these methods make progress on UDA, few of them consider the practical scenario where training data are collected from multiple sources. Our paper proposes a model to tackle multi-source domain adaptation, which is a more general and challenging scenario.

Multi-Source Domain Adaptation Compared with single source UDA, multi-source domain adaptation assumes that training data from multiple sources are available. Originated from the early theoretical analysis , MSDA has many practical applications. Ben-David et al introduce an HΔH-divergence between the weighted combination of source domains and target domain. Crammer et al establish a general bound on the expected loss of the model by minimizing the empirical loss on the nearest k sources. Mansour et al claim that the target hypothesis can be represented by a weighted combination of source hypotheses. In the more applied works, Deep Cocktail Network (DCTN) proposes a k-way domain discriminator and category classifier for digit classification and real-world object recognition. Hoffman et al propose normalized solutions with theoretical guarantees for cross-entropy loss, aiming to provide a solution for the MSDA problem with very practical benefits. Duan et al propose Domain Selection Machine for event recognition in consumer videos by leveraging a large number of loosely labeled web images from different sources. Different from these methods, we directly matches all the distributions by matching the moments.

Moment Matching The moments of distributions have been studied by the machine learning community for a long time. In order to diminish the domain discrepancy between two domains, different moment matching schemes have been proposed. For example, MMD matches the first moments of two distributions. Sun et al propose an approach that matches the second moments. Zhang et al propose to align infinite-dimensional covariance matrices in RKHS. Zellinger et al introduce a moment matching regularizer to match high moments. As the generative adversarial network (GAN) becomes popular, many GAN based moment matching approaches have been proposed. McGAN utilizes a GAN to match the mean and covariance of feature distributions. GMMN and MMD GAN are proposed for aligning distribution moments with generative neural networks. Compared to these methods, we focuses on matching distribution moments for multiple domains as this is crucial for multi-source domain adaptation (demonstrated in Moment Matching paper[1])

Method Used : Moment Matching for Multi-Source DA

Given $D_S = D_1, D_2, ..., D_N$ the collection of labeled source domains and D_T the unlabeled target domain, where all domains are defined by bounded rational measures on input space X , the multi-source domain adaptation problem aims to find a hypothesis in the given hypothesis space H, which minimizes the testing target error on D_T .
Definition 1. Assume $X_1, X_2, ..., X_N, X_T$ are collections of i.i.d. samples from $D_1, D_2, ..., D_N, D_T$ respectively, then the Moment Distance between D_S and D_T is defined as

$$MD^2(D_S, D_T) = \Sigma_{k=1}^2 (1/N \Sigma_{i=1}^N ||E(X_i^k) - E(X_T^k)||_r)$$

Theoretical Insight

Following [1], we introduce a rigorous model of multi source domain adaptation for binary classification. A domain D = (μ, f) is defined by a probability measure (distribution) μ on the input space X and a labeling function f : $X \to \{0, 1\}$. A hypothesis is a function $h : X \to \{0, 1\}$. The probability that h disagrees with the domain labeling function f under the domain distribution μ is defined as:

$$\epsilon_D(h) = \epsilon_D(h, f) = E_\mu[|h(x)f(x)|]$$

For a source domain D_S and a target domain D_T , we refer to the source error and the target error of a hypothesis h as $\epsilon_S(h) = \epsilon_{D_S}(h)$ and $\epsilon_T(h) = \epsilon_{D_T}$ (h) respectively.

When the expectation in Equation 5 is computed with respect to an empirical distribution, we denote the corresponding empirical error by $\hat{\epsilon}_D(h)$, such as $\hat{\epsilon}_S(h)$ and $\hat{\epsilon_T}(h)$. In particular, we examine algorithms that minimize convex combinations of source errors, i.e., given a weight vector $\alpha = (\alpha_1, ..., \alpha_N)$ with $\Sigma_{j=1}^N \alpha_j = 1$, we define the $\alpha-$weighted source error of a hypothesis h as $\epsilon_\alpha(h) = \Sigma_{j=1}^N \alpha_j \epsilon_j(h)$, where $\epsilon_j(h)$ is the shorthand of $\epsilon_{D_j}(h)$. The empirical $\alpha-$weighted source error can be defined analogously and denoted by $\hat{\epsilon_\alpha}(h)$.

Previous theoretical bounds on the target error are based on the HΔH-divergence between the source and target domains. While providing theoretical insights for general multi-source domain adaptation, these HΔH divergence based bounds do not directly motivate moment based approaches. In order to provide a specific insight for moment-based approaches, we introduce the k-th order cross-moment divergence between domains, denoted by $d_{CM^k}(\Delta, \Delta)$, and extend the analysis in to derive the following moment-based bound for multi-source domain adaptation.

Theorem 1. Let H be a hypothesis space of V C dimension d. Let m be the size of labeled samples from all sources $\{D_1, D_2, ..., D_N\}$, Sj be the labeled sample set of size β_jm ($\Sigma_j \beta_j = 1$) drawn from μ_j and labeled by the ground truth labeling function $_j$. If $\hat{h} \in H$ is the empirical minimizer of $\hat{\epsilon_\alpha}(h)$ for a fixed weight vector α and $h_T = min_{h \in H} \epsilon_T(h)$ is the target error minimizer, then for any $\delta \in (0, 1)$ and any $\epsilon > 0$, there exist N integers $\{n_\epsilon^j\}_{j=1}^N$ and N constants $\{a_{n_\epsilon^j}\}_{j=1}^N$, such that with probability at least 1 δ,

$$\epsilon_T(\hat{h}) \le \epsilon_T(h_T^*) + \eta_{\boldsymbol{\alpha},\boldsymbol{\beta},m,\delta} + \epsilon + \sum_{j=1}^{N} \alpha_j \Big(2\lambda_j + a_{n_\epsilon^j} \sum_{k=1}^{n_\epsilon^j} d_{CM^k}(\mathcal{D}_j, \mathcal{D}_T)\Big), \quad (6)$$

where $\eta_{\boldsymbol{\alpha},\boldsymbol{\beta},m,\delta} = 4\sqrt{(\sum_{j=1}^N \frac{\alpha_j^2}{\beta_j})(\frac{2d(\log(\frac{2m}{d})+1)+2\log(\frac{4}{\delta})}{m})}$
and $\lambda_j = \min_{h \in \mathcal{H}}\{\epsilon_T(h) + \epsilon_j(h)\}$.

Theorem 1 shows that the upper bound on the target error of the learned hypothesis depends on the pairwise moment divergence $d_{CM^k}(D_S, D_T)$ between the target domain and each source domain. This provides a direct motivation for moment matching approaches beyond ours. In particular, it motivates our multi-source domain adaptation approach to align the moments between each target-source pair. Moreover, it is obvious that the last term of the bound, $\Sigma_k d_{CM^k}(D_j, D_T)$, is lower bounded by the pairwise divergences between source domains. To see this, consider the toy example consisting of two sources D_1, D_2, and a target D_T , since $d_{CM^k}(\Delta, \Delta)$ is a metric, triangle inequality implies the following lower bound:

$$d_{CM^k}(D_1, D_T) + d_{CM^k}(D_2, D_T) d_{CM^k}(D_1, D_2)$$

This motivates our algorithm to also align the moments between each pair of source domains. Intuitively, it is not possible to perfectly align the target domain with every source domain, if the source domains are not aligned themselves.

Experiments on Domain Net

For our first initial experiments, we ran the image M3SDA code and below are the results for same (**Euclidean Distance**) :

Source	Target	Loss	Discrepancy
Real | Paint	Sketch	1.81	0.0068
Real | Sketch	Paint	1.79	0.0086
Sketch | Painting	Real	1.815	0.0061

After our initial run we changed the distance functions. Below are the results obtained with **dynamic distance function**:

Source	Target	Loss	Discrepancy
Real | Paint	Sketch	1.797	0.0068
Real | Sketch	Paint	1.8761	0.0091
Sketch | Painting	Real	1.788	0.0058

Below are the results obtained after with **Mahalanobis distance function**:

Source	Target	Loss	Discrepancy
Real | Paint	Sketch	1.784	0.0061
Real | Sketch	Paint	1.831	0.0083
Sketch | Painting	Real	1.792	0.0052

Discussion

Initially we run our model for image dataset taking euclidean distance. As part of enhancement we changed distance function to improve the performance of model. We examined two of the most popular distance functions used for measuring image similarity: **Dynamic Partial Distance Function**.

The Dynamic Partial Distance Function $d(X, Y)$ is defined as

$$d(m, r) = \Big(\sum_{\delta_i \in \Delta_m} \delta_i^r \Big)^{\frac{1}{r}},$$

DPF has two adjustable parameters: m and r. Parameter m can range from 1 to p. When m = p, it degenerates to the Minkowski metric. When $m < p$, it counts only the smallest m feature distances between two objects, and the influence of the $(p - m)$ largest feature distances is eliminated. It helped us in improving our result as rather than comparing all features of all images, with Dynamic Partial Distance Function we are just comparing most similar features of images. When comparing most similar features our model converges fast resulting in less loss value.

Another distance function we used is **Mahalanobis distance** :

The Mahalanobis distance of an observation $x = (x_1, x_2, x_3, ..., x_n)^T$ from a set of observations with mean $\mu = (\mu_1, \mu_2, ..., \mu_n)^T$ and covariance matrix S is defined as:

$$D_m(x) = \sqrt{(x - \mu)^T S^{-1} (x - \mu)}$$

Conclusion

In the code, first we have extended M3SDA code on images. This was particularly challenging as at first it seems we are using M3SDA code as our baseline, but the code for M3SDA was run on digit dataset whereas we have done our complete implementation on image dataset. We hope it will be beneficial to evaluate future single- and multi-source UDA methods on image datasets.

In addition, we have also proposed use of different distance function for image dataset to calculate loss.

Future Work

Due to the computation constraint we could not test on complete 345 classes for all 6 categories in one go. In future we aim to test our model on complete dataset in one go.

Also, as part of enhancement we aim to do pseudo labeling. As part of pseudo labelling we run our moment matching model on unlabelled target domain. We assume that label predicted by model are accurate with some probability. After each iteration we add these now labelled target in source domain. Purpose of this approach is it train model with more number of categories which in turn result in labelling more accurate unlabelled target images from diverse categories.

Acknowledgments

We thank you Professor Sarah Adel Bargal, Professor Margrit Betke for her useful discussions and suggestions.

References

[1] Moment Matching for Multi-Source Domain Adaptation Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, Bo Wang

[2] Discovery of A Perceptual Distance Function for Measuring Image Similarity Beitao Li, Edward Chang, and Yi Wu

[3] Similarity Measurement Between Images Chaur-Chin Chen and Hsueh-Ting Chu

[4] Pseudo-labeling and confirmation bias in deep semi-supervised learning Eric Arazo, Diego Ortego, Paul Albert, Noel E. O'Connor Kevin McGuinness

[5] Towards Highly Accurate Coral Texture Images Classification Using Deep Convolutional Neural Networks and Data Augmentation. Anabel G omez-R ıosa, Siham Tabik , Juli an Luengo, ASM Shihavuddin ,Bartosz Krawczyk , Francisco Herrera