Attending to Discriminative Certainty for Domain Adaptation

Vinod Kumar Kurmi,*  Shanu Kumar,*  Vinay P. Namboodiri
Indian Institute of Technology Kanpur
India
{vinodkk,sshanu,vinaypn}@iitk.ac.in

Abstract

In this paper, we aim to solve for unsupervised domain adaptation of classifiers where we have access to label information for the source domain while these are not available for a target domain. While various methods have been proposed for solving these including adversarial discriminator based methods, most approaches have focused on the entire image based domain adaptation. In an image, there would be regions that can be adapted better, for instance, the foreground object may be similar in nature. To obtain such regions, we propose methods that consider the probabilistic certainty estimate of various regions and specify focus on these during classification for adaptation. We observe that just by incorporating the probabilistic certainty of the discriminator while training the classifier, we are able to obtain state of the art results on various datasets as compared against all the recent methods. We provide a thorough empirical analysis of the method by providing ablation analysis, statistical significance test, and visualization of the attention maps and t-SNE embeddings. These evaluations convincingly demonstrate the effectiveness of the proposed approach.

1. Introduction

With the advent of deep learning, there has been substantial progress for solving image classification tasks with state of the art methods obtaining lesser than 3% error (on top five results) on the imagenet dataset. However, it was observed that these results do not transfer to other datasets [25] due to the effect of dataset bias [44]. The classifiers trained on one dataset (termed source dataset) show a significant drop in accuracy when tested on another dataset (termed target dataset). To address this issue, some methods have been proposed for adapting domains [7, 50]. One of the more successful approaches towards addressing this domain shift has been based on the adversarial adaptation of features using an adversarial discriminator [12] that aims to distinguish between samples drawn from the source and target datasets. Due to the adversarial training, the feature representations are brought close such that the discriminator is not able to distinguish between samples from source and target dataset. However, all approaches that are based on this idea consider the whole image as being adapted. This usually is not the case as there are predominant regions in an image that may be better adapted and useful for improving classification on target dataset. We address this issue and propose a simple approach for solving this problem.

To specify regions that can be adapted we propose the use of certainty of a probabilistic discriminator. During training, we identify regions where the discriminator is certain, i.e., the probabilistic uncertainty for these regions is

*Equal contributions from both authors.
low. These regions can be adapted because there exists a clear distinction between the source and target regions. Figure 1 shows that using measures such as data uncertainty (known as aleatoric uncertainty) [19] and predictive uncertainty [24], we can obtain regions that can be adapted better. We also observe from the Figure 1 and 5 that for most of the duration during training, discriminator is certain on the foreground regions, as the foreground is hard to adapt. Hence, when the classifier is trained with the emphasis being placed on these regions, then we observe that the classifier focuses on these regions during prediction and therefore generalizes better on target dataset. Quantitatively we observe results that are up to 7.4% better than the current state of the art methods on Office-Home dataset [49].

To summarize, through this paper we make the following contributions:

- We propose a method to identify adaptable regions using the certainty estimate of the discriminator, and this is evaluated using various certainty estimates.
- We use these certainty estimate weights for improving the classifier performance on target dataset by focusing the training of the classifier on the adaptable regions of the source dataset.
- We provide a thorough evaluation of the method by considering detailed comparison on standard benchmark datasets against the state of the art methods and also provide an empirical analysis using statistical significance analysis, visualization and ablation analysis of the proposed method.
- An additional observation is that by using Bayesian classifiers we also improve the robustness of the classifier in addition to obtaining certainties of classification accuracy. This aids in better understanding of the proposed method.

2. Literature Survey

Some studies have examined different adaptation methods. One study by [47] examined domain adaptation by minimizing the maximum mean discrepancy distance. The maximum mean discrepancy based approaches were further extended to multi-kernel MMD in [25]. In adversarial learning framework [12] has proposed a method to minimize the source target discrepancy using a gradient reversal layer at discriminator. Recently many adversarial methods have been applied in the domain adaptation task to bring the source and target distribution closer. Adversarial discriminative domain adaptation [46] considers an inverted label GAN loss. Wasserstein distance based discriminator was used in [39] to bring the two distributions closer. Domain confusion network [45] was also used to solve the adaptation problem in two domains by minimizing the discriminate distance between two domains. The discriminative feedback of the discriminator also applied in the paraphrase generation problem [33]. Another adversarial discriminator based model is [34], where multiple discriminators (MADA) have been used to solve the mode collapse problem in the domain adaptation. Some works closely related to MADA have been proposed in [4, 3]. The labeled discriminator [23] used to tackle the mode collapse problem in domain adaptation. The adversarial domain adaptation also explored in scene graph [22]. Other source and target discrepancy minimization based methods such as [37, 54] also address the domain adaptation problem. [32, 31] have proposed an exemplar based discrepancy minimization method. Recently [2, 5] have applied generative adversarial network [14] for the domain adaptation problems. Image generation methods are used to adapt the source and target domain [13, 38, 30]. Other work in [16] and [42] have used Cycle consistency [55] loss and deep coral loss [41, 40] for ensuring the closeness of source and target domain respectively. Deep Bayesian models have been used to play an important role in the estimation of the deep model uncertainty. The Bayesian formulation in domain adaptation natural language processing has been proposed in [9]. In [11], it has been justified that dropout can also work as an approximation of the deep Bayesian network. Works on the uncertainty estimation have been reported in [10, 19]. In [29] predictive uncertainty has been calculated over the prior networks. Uncertainty in the ensemble model along with adversarial training has been discussed in [24]. Another work on the Bayesian uncertainty estimation has been reported in the [43].

Attention-based networks have been widely applied in many computer vision applications such as image captioning [51, 53], visual question answer [32, 31, 18] and speech recognition [6]. The advantage of the attention model is that it helps to learn some set of weights over a set of representation input which has relatively more importance than others. Recently [48] showed that the attention mechanism can also be achieved by dispensing with recurrence and convolutions. A recent work [17] addresses the domain adaptation problem by obtaining the synthetic source and target images from CycleGAN [55], and then aligned the attention map of all the pairs.

3. Methodology

In the unsupervised domain adaptation problem, the source dataset \( \mathcal{D}_s = (x_i^s, y_i^s) \) consists of data sample \( x_i^s \) with corresponding label \( y_i^s \) where \( D_s \subseteq P_s \) and the target dataset \( \mathcal{D}_t(x_i^t) \) consists of unlabeled data samples \( x_i^t \) where \( D_t \sim P_t \cdot P_s \) and \( P_s \) and \( P_t \) are the source and target distributions. We further assume that both the domains are complex and unknown. For solving this task, we are following the adversarial domain adaptation framework, where a discriminator is trained to learn domain invariant features domain invariant while a classifier is trained to learn class
Bayesian framework similar to the Bayesian framework is one of the efficient ways to predict uncertainty. Gal et.al [10] has shown that by applying dropout after every fully connected (fc) layer, we can perform probabilistic inference for deep neural networks. Hence we have followed a similar approach for defining the classifier. For estimating uncertainty, similar to [19], we trained the classifier to output class probabilities along with aleatoric uncertainty (data uncertainty). The predictive uncertainty includes both model uncertainty and data uncertainty, where model uncertainty results from uncertainty in model parameters. Estimation of aleatoric uncertainty for the classifier makes the features more robust for prediction, and estimation of predictive uncertainty provides a tool for visualizing model’s predictions.

For the input sample \( x_i \), the feature extractor \( G_f \) outputs features \( f_i \), represented by \( f_i = G_f(x_i) \). The predicted class logits \( y^c_i \) and aleatoric uncertainty \( v^c_i \) are obtained as:

\[
y^c_i = G_{cy}(G_c(f_i)), \quad v^c_i = G_{cu}(G_c(f_i))
\]

where \( G_{cy} \) and \( G_{cu} \) are the logits and aleatoric uncertainty prediction modules of the classifier \( G_c \) respectively. The classification loss for predicted logits is defined as:

\[
\mathcal{L}_{cy} = \frac{1}{n_s} \sum_{x_i \in D_s} \mathcal{L}(y^c_i, y_i)
\]

where \( \mathcal{L} \) is the cross entropy loss function and \( y_i \) is the true class label for the input \( x_i \). The total number of data samples in the source domain is denoted as \( n_s \). The classifier aleatoric loss \( \mathcal{L}_{cu} \) for predicted uncertainty \( v^c_i \) is defined as:

\[
\bar{y}^c_{i,t} = y^c_i + \sigma^c_i * \epsilon_t, \quad \epsilon_t \sim \mathcal{N}(0, I)
\]

\[
\mathcal{L}_{cu} = -\frac{1}{n_s} \sum_{x_i \in D_s} \log \frac{1}{T} \sum_{t} \mathcal{L}(\bar{y}^c_{i,t}, y_i)
\]

where \( \sigma^c_i \) is the standard deviation, \( v^c_i = (\sigma^c_i)^2 \). The classifier is trained by jointly minimizing both the classification loss \( \mathcal{L}_{cy} \) and aleatoric loss \( \mathcal{L}_{cu} \).

### 3.2. Bayesian Discriminator

In the proposed method, the discriminator is also modeled in the Bayesian framework similar to the Bayesian classifier. The uncertainty in the discriminator network implies the region where it is uncertain about its prediction about the domain. The uncertainty estimation of the discriminator can guide the feature extractor more efficiently for domain adaptation. All real-world images contain some type of aleatoric uncertainty or noise. These regions which contain aleatoric uncertainty, are not adaptable. By aligning these uncertain regions, we are corrupting the feature representation, thus confusing the classifier during predictions for the target domain. So, by estimating aleatoric uncertainty, the discriminator is avoiding the learning of feature representations for these regions, which also reduces negative transfer [34]. The negative transfer introduces false alignment of the mode of two distributions across domains, which needs to be prevented during adaptation. Similarly, the predictive uncertainty tells us about the model’s incapability to classify the domains, as the discriminator is not sure about the domain. Predictive uncertainty occurs in the region where either it is already adapted, or there is noise which corresponds to aleatoric uncertainty. We obtain the discriminator predicted logits and variance using the following equations:

\[
y^d_i = G_{dy}(G_d(f_i)), \quad v^d_i = G_{dv}(G_d(f_i))
\]

where \( G_{dy} \) and \( G_{dv} \) predict domain class logits \( y^d_i \) and domain aleatoric uncertainty \( v^d_i \) respectively using features from \( G_d \). The domain classification loss \( \mathcal{L}_{dy} \) is defined as:

\[
\mathcal{L}_{dy} = \frac{1}{n_s + n_t} \sum_{x_i \in D_s \cup D_t} \mathcal{L}(y^d_i, d_i)
\]

where \( \mathcal{L} \) is the cross entropy loss function, \( d_i \) is the true domain of the image, and \( n_s \) and \( n_t \) are the number of source and target samples. The domain label \( d_i \) is defined to be 0 if \( x_i \in D_s \) and 1 if \( x_i \in D_t \). The discriminator aleatoric loss \( \mathcal{L}_{dv} \) is defined as:

\[
\mathcal{L}_{dv} = -\frac{1}{n_s + n_t} \sum_{x_i \in D_s \cup D_t} \log \frac{1}{T} \sum_{t} \mathcal{L}(\hat{y}^d_{i,t}, d_i)
\]

where \( v^d_i = (\sigma^d_i)^2 \). Discriminator is trained by jointly minimizing both the domain classification loss \( \mathcal{L}_{dy} \) and discriminator aleatoric loss \( \mathcal{L}_{dv} \).

### 3.3. Certainty Based Attention

Uncertainty estimation of the discriminator can help in identifying those regions which can be adapted, cannot be adapted, or already adapted. The regions which are already aligned will confuse the discriminator for predicting the domain. Hence discriminator will be highly uncertain on these regions. The discriminator will also be highly uncertain on
Figure 2: The architecture of Certainty based Attention for Domain Adaptation (CADA), consists of a shared feature extractor, Bayesian classifier and Bayesian discriminator where both the classifier and discriminator predict the variance value along with the prediction score. Discriminator’s predictive or aleatoric uncertainty is used to highlight the regions where the discriminator is certain about its predictions.

The Certainty Attention based Domain Adaption (CADA) model is shown in Figure 2. In the proposed work, we propose the two variants of CADA: aleatoric certainty based attention (CADA-A) and predictive certainty based attention (CADA-P).

3.3.1 Aleatoric Certainty based Attention

The aleatoric uncertainty \( (v_d^t) \) of the domain discriminator occurs due to corruption or noise in the regions. These regions can’t be adapted. Uncertainty estimation can also help to identify regions where the discriminator is certain or the regions which can be further aligned.

In most of the datasets, the discriminator can easily discriminate between the source and target images by only attending on the background during the initial phase of the training. Hence, the discriminator will be more certain in these regions, which results in easier adaptation of background regions after some adversarial training. But the foreground regions are difficult to adapt, as foreground varies a lot across all the classes and images. Therefore for most of the span during training, the discriminator will be certain on the transferable regions of the foreground. Thus, if the classifier attends to the certain regions of the discriminator, it will focus more on the transferable regions of the foreground during training.

The Certainty Attention based Domain Adaption (CADA) model is shown in Figure 2. In the proposed work, we propose the two variants of CADA: aleatoric certainty based attention (CADA-A) and predictive certainty based attention (CADA-P).

3.3.2 Predictive Certainty based Attention

The predictive uncertainty measures the model’s capability for prediction. It occurs in the regions which are either
already domain invariant or which contain noise. The regions corresponding to discriminator’s predictive certainty are domain transferable regions and should be attended during classification. We follow the approach proposed in [19] for computing the predictive uncertainty of discriminator. It is obtained by the entropy of the average class probabilities of the Monte Carlo (MC) samples. We sample weights from $G_d$, and perform MC simulations over domain probabilities $p(y^d_{i,t})$ and aleatoric uncertainty $v^d_{i,t}$ for estimating predictive uncertainty $\text{Var}^d(f_i)$.

$$g_{i,t} = G_d^d(f_i), \quad G_d^d \sim G_d$$

Using the sampled model, we calculate domain probabilities and aleatoric uncertainty.

$$p(y^d_{i,t}) = \text{Softmax}(G_d(y^d_{i,t})), \quad v^d_{i,t} = G_d(c, (g_{i,t}))$$

$$H(y^d_{i,t}) = -\sum_{c=1}^{2} p(y^d_{i,t} = c) \log(p(y^d_{i,t} = c))$$

$$\text{Var}^d(f_i) \approx \frac{1}{T} \sum_{t=1}^{T} (v^d_{i,t} + H(y^d_{i,t}))$$

where $H(y^d_{i,t})$ is the Entropy of the $p(y^d_{i,t})$. For identifying the predictive uncertain regions, we compute gradients of the predictive uncertainty $\text{Var}^d(f_i)$ of the discriminator with respect to the features $f_i$, i.e. $\frac{\partial \text{Var}^d(f_i)}{\partial f_i}$, and negative of these gradients which returns through the gradient reversal layer will correspond to the gradients of the predictive certainty, i.e., $-\frac{\partial \text{Var}^d(f_i)}{\partial f_i}$.

$$p_i = f_i * -\frac{\partial \text{Var}^d(f_i)}{\partial f_i}$$

$$a_i = \text{ReLU}(p_i) + c * \text{ReLU}(-p_i)$$

$$w_i = (1 - \text{Var}^d(f_i)) * \text{Softmax}(a_i)$$

Similar to aleatoric certainty based attention, for obtaining the predictive certain regions we apply ReLU function to $p_i$ and ignoring its negative values (corresponds to uncertain regions), a ReLU function is applied to negative of $p_i$, and multiply it with a large number $c$ using Eq. 16. After applying the Softmax, the features that are activated by the predictive uncertainty will have zero weight and for features that are highly activated by the predictive certainty will get more weight. The residual weighted features are obtain by following equations

$$h_i = f_i * (1 + w_i)$$

Therefore the images with high predictive uncertainty have lower value of $w_i$, have less attention, while images have high predictive certainty have high value of $w_i$, produce high attentive features. This will ensure that already adapted regions or non-adaptive region (both cases have the lower attention value).

3.4. Training Algorithm

We employ Certainty Attention based Domain Adaption (CADA) models for solving the task of unsupervised domain adaptation. Both the CADA-P and CADA-A models jointly learn domain invariant and class invariant features, by focusing the classifier’s attention on the discriminator’s certain regions. So, the class probabilities $y^c_i$ and aleatoric uncertainty $v^c_i$ for the classifier will be estimated using weighted feature $h_i$.

$$y^c_i = G_c(G_c(h_i)), \quad v^c_i = G_{cv}(G_c(h_i))$$

The final objective function $J$ for optimizing both the models is defined as:

$$J = \mathcal{L}_{cy} + \mathcal{L}_{cv} - \lambda * (\mathcal{L}_{dy} + \mathcal{L}_{dc})$$

where $\lambda$ is a trade-off parameter between classifier and discriminator. The optimization problem is to find the parameters $\theta_f, \theta_c, \theta_{cy}, \theta_{cv}, \theta_d, \theta_{dy}, \theta_{dv}$ that jointly satisfy:

$$(\hat{\theta}_f, \hat{\theta}_c, \hat{\theta}_{cy}, \hat{\theta}_{cv}) = \arg \min_{\theta_f, \theta_c, \theta_{cy}, \theta_{cv}, \theta_d, \theta_{dy}, \theta_{dv}} J(\theta_f, \theta_c, \theta_{cy}, \theta_{cv}, \theta_d, \theta_{dy}, \theta_{dv})$$

$$(\hat{\theta}_d, \hat{\theta}_{dy}, \hat{\theta}_{dv}) = \arg \max_{\theta_d, \theta_{dy}, \theta_{dv}} J(\theta_f, \theta_c, \theta_{cy}, \theta_{cv}, \theta_d, \theta_{dy}, \theta_{dv})$$

The implementation details are provided in supplementary material and other details are provided in the project page.

4. Experiments and Results

4.1. Datasets

Office-31 Dataset The Office-31 [36] consists of three domains: Amazon, Webcam and DSLR with 31 classes in each domain. There are 2817 images in Amazon (A) domain, 795 images in Webcam (W) and 498 images are in DSLR (D) domain makes total 4,110 images. To enable unbiased evaluation, we evaluate all methods on 6 transfer tasks A→W, D→A, W→A, A→D, D→W and W→D.

Office-Home Dataset We also evaluated our model on the Office-Home dataset [49] for unsupervised domain adaptation. This dataset consists of four domains: Art (Ar), Clip-art (Cl), Product (Pr) and Real-World (Rw). Each domain has common 65 categories and total 15,500 images. We evaluated our model by considering all the 12 adaptation tasks. The performance is reported in the Table 3.

ImageCLEF Dataset ImageCLEF-2014 dataset consists of three datasets: Caltech-256 (C), ILSVRC 2012 (I), and Pascal VOC 2012 (P). There are 12 common classes and total 600 images in each domain. We evaluate our model on all the 6 transfer tasks and results are reported in Table 4.

1https://delta-lab-ntt.github.io/CADA/
4.2. Results

Following the common setting in unsupervised domain adaptation, we used the pre-trained Alexnet [21] and pre-trained ResNet [15] architecture as our base model. For Office-31 dataset, results are reported in the Table 1 and Table 2. From the table, it is clear that the proposed CADA outperforms the other methods on most transfer tasks, where CADA-P is the top-performing variant for both Alexnet and Resnet model. On average, we obtain improvements of 3.3% and 1.8% over the state of the art methods.
Table 4: Classification accuracy (%) on ImageCLEF dataset for unsupervised domain adaptation (ResNet-50 [15])

<table>
<thead>
<tr>
<th>Method</th>
<th>I → P</th>
<th>P → I</th>
<th>I → C</th>
<th>C → I</th>
<th>C → P</th>
<th>P → C</th>
<th>Avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>ResNet [15]</td>
<td>74.8</td>
<td>83.9</td>
<td>91.5</td>
<td>78.0</td>
<td>65.5</td>
<td>91.2</td>
<td>80.7</td>
</tr>
<tr>
<td>DAN [25]</td>
<td>75.0</td>
<td>86.2</td>
<td>93.3</td>
<td>84.1</td>
<td>69.8</td>
<td>91.3</td>
<td>83.3</td>
</tr>
<tr>
<td>RTN [27]</td>
<td>75.6</td>
<td>86.8</td>
<td>95.3</td>
<td>86.9</td>
<td>72.7</td>
<td>92.2</td>
<td>84.9</td>
</tr>
<tr>
<td>GRL [12]</td>
<td>75.0</td>
<td>86.0</td>
<td>96.2</td>
<td>87.0</td>
<td>74.3</td>
<td>91.5</td>
<td>85.0</td>
</tr>
<tr>
<td>JAN [28]</td>
<td>76.8</td>
<td>88.0</td>
<td>94.7</td>
<td>89.5</td>
<td>74.2</td>
<td>91.7</td>
<td>85.8</td>
</tr>
<tr>
<td>MADA [34]</td>
<td>75.0</td>
<td>87.9</td>
<td>96.0</td>
<td>88.8</td>
<td>75.2</td>
<td>92.2</td>
<td>85.8</td>
</tr>
<tr>
<td>CDAN [26]</td>
<td>77.2</td>
<td>88.3</td>
<td>98.3</td>
<td>90.7</td>
<td>76.7</td>
<td>94.0</td>
<td>87.5</td>
</tr>
<tr>
<td>CADA-A</td>
<td><strong>78.0</strong></td>
<td><strong>91.5</strong></td>
<td>96.3</td>
<td>91.0</td>
<td>77.1</td>
<td>95.3</td>
<td>88.2</td>
</tr>
<tr>
<td>CADA-P</td>
<td><strong>78.0</strong></td>
<td>90.5</td>
<td>96.7</td>
<td><strong>92.0</strong></td>
<td><strong>77.2</strong></td>
<td><strong>95.5</strong></td>
<td><strong>88.3</strong></td>
</tr>
</tbody>
</table>

Figure 3: The t-SNE visualization of representations learned by (a) ResNet, (b) DANN, (c) CADA-A, and (d) CADA-P (blue: A; red: W), (e) shows Proxy A-distance for A → W task for method Reset [15], GRL [12] and the proposed model CADA-P.

Figure 4: Analysis of statistically significant difference for A→W and A→D in ResNet [15], GRL [12], CADA-A, and CADA-P methods, with a significance level of 0.05.

5. Ablation Study

We investigate the Bayesian model with and without attention for both Alexnet and ResNet model on the office-31 dataset. From Table 1 and 2, it is clear that the Bayesian model without attention (CADA-W) performs significantly better than the most of the other previous models, as predicting uncertainty for discriminator reduces negative transfer, by neglecting the regions which contain data uncertainty.

6. Empirical Analysis

We further provide empirical analysis in terms of qualitative analysis of attention maps, feature visualization, discrepancy distance and statistical significance for additional insights about the performance of our method.

6.1. Qualitative Analysis of Attention Maps

To provide the effectiveness of proposed certainty based adaptation, we provide the certainty map of the discriminator at different training stages (chosen randomly) in the Figure 5. In the figure, we see that at the initial phase of the training (after 4 epochs), the discriminator discriminates the source and target domains just by some random location. As the training progress, the discriminator learns...
the domain by attending the background of the images (In A→W, domains are mostly dissimilar in the background). After some more training, the background is adapted (after 125 epochs), and now the discriminator attends to foreground part of the image to differentiate the domains (after 535 epochs). However, the foreground varies a lot across all the images. Hence discriminators are highly certain in the class object regions. Now with further training of the model, these class object regions are also adapted (after 1300 epochs). The remaining regions of the image cannot be further adapted because there is data uncertainty. At the end of the training, the discriminator will be highly certain regarding the domain, and the attention weight on the regions which are not adaptable will have low weights as we are using the certainty of the discriminator as the measure for the weights. Note that at the time of inference, we do not use the attention weights obtained by certainty for aiding classifier. These are used only at the time of training. The results show that these attention weights based training aids the classifier to better generalize to the target domain. We have provided more visualization examples in the supplementary material for further justification.

### 6.2. Feature Visualization

Adaptability of target to source features can be visualized using the t-SNE embeddings of images feature. We follow similar setting in [47, 12, 34] and plot t-SNE embeddings of the dataset in the Figure 3. We can observe that the proposed model correctly aligns the source and target domain images with exactly 31 clusters which is equal to the number of class labels with clear boundaries.

### 6.3. Discrepancy Distance

$A$-distance is a measure of cross domain discrepancy [1], which, together with the source risk, will bound the target risk. The proxy $A$-distance is defined as $d_A = 2(1 - 2e)$, where $e$ is the generalization error of a classifier (e.g. kernel SVM) trained on the binary task of discriminating source and target. Figure 3 (e) shows $d_A$ on tasks $A \rightarrow W$ with features of ResNet[15], GRL[12], and our model. We observe that $d_A$ using our model features is much smaller than $d_A$ using ResNet and GRL features, which suggests that our features can reduce the cross-domain gap more effectively.

### 6.4. Statistical Significance Test

We analyzed statistical significance [8] for variants of proposed method against GRL [12]. The Critical Difference (CD) for Nemenyi test depends upon the given confidence level (which is 0.05 in our case) for average ranks and number of test datasets. If the difference in the rank of the two methods lies within CD, then they are not significantly different. Figure 4 visualizes the post hoc analysis using the CD diagram for A→W dataset. From the figures, it is clear that our models are significantly different from GRL[12].

---

7. Conclusion

In this paper, we propose the use of certainty estimates of the discriminator to aid the generalization of the classifier by increasing the attention of the classifier on these regions. As it can be observed through our results, the attention maps obtained through certainty agree well with the classifier certainty for true labels and this aids in the generalization of the classifier for the target domain as well. The proposed method is thoroughly evaluated by comparison with state of the art methods and shows improved performance over all the other methods. Further, the analysis is provided in terms of statistical significance tests, discrepancy distance, and visualizations for better insight about the proposed method. The proposed method shows a new direction of using probabilistic measures for domain adaptation, and in the future, we aim to further explore this approach.

Acknowledgment: We acknowledge travel support from Microsoft Research India and Google Research India. We also acknowledge resource support from Delta Lab, IIT Kanpur. Vinod Kurmi acknowledges support from TCS Research Scholarship Program.
References


[23] Vinod Kumar Kurmi and Vinay P. Namboodiri. Looking back at labels: A class based domain adaptation technique, 2019. 2


[35] Pedro O Pinheiro and AI Element. Unsupervised domain adaptation with similarity learning. 6


[40] Baochen Sun, Jiashi Feng, and Kate Saenko. Return of frustratingly easy domain adaptation. In AAAI, volume 6, page 8, 2016. 2

[41] Baochen Sun, Jiashi Feng, and Kate Saenko. Correlation alignment for unsupervised domain adaptation. Domain Adaptation in Computer Vision Applications, page 153, 2017. 2


