Supplementary: 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 supplementary material we aim to provide implementaion details and visualization based analysis. Main results in the supplementary material include ablation analysis, visualization of results as the training progresses and details regarding the network architecture.

1. Setup

1.1. Architecture

For both Alexnet [4] and ResNet-50 [3] architecture, we used the pre-trained model (trained on Imagenet) for feature extractor, then a bottleneck layer of 256 dimension followed by a two-layer classification network. This classification network also predicts the variance. Similarly, the discriminator is a two-layer network that predicts the domain label as well as the variance score.

1.2. Training details

We use 0.0002 learning rate and anneal the learning rate over training using the exponential decay. The updated learning rate factor is $\mu_p = \frac{\mu_0}{(1+\alpha \dot{p})^{\beta}}$ where p is the training progress linearly changing from 0 to 1, μ_0 = 0.01, α = 10 and β = 0.75 similar to [1]. Classifier and discriminator layers are trained from scratch, and their learning rate is set to be 10 times that of the other layers. We follow the Xavier initialization [2] method to initialize the classifier and discriminator network. Stochastic gradient descent (SGD) is used to optimize the model parameters. Alexnet Model is trained with a batch size of 128 (64 source and 64 target samples) and ResNet Model is trained with a batch size of 40 (20 source and 20 target samples). The adaptation parameter λ is set to 1 for all the experiments. We apply label smoothing in classifier similar to [7, 6, 5]. The code is implemented in Torch-Lua. Other details and codes are provided in the project page [1]

*Equal contributions from both authors.
[1]https://delta-lab-iitk.github.io/CADA/

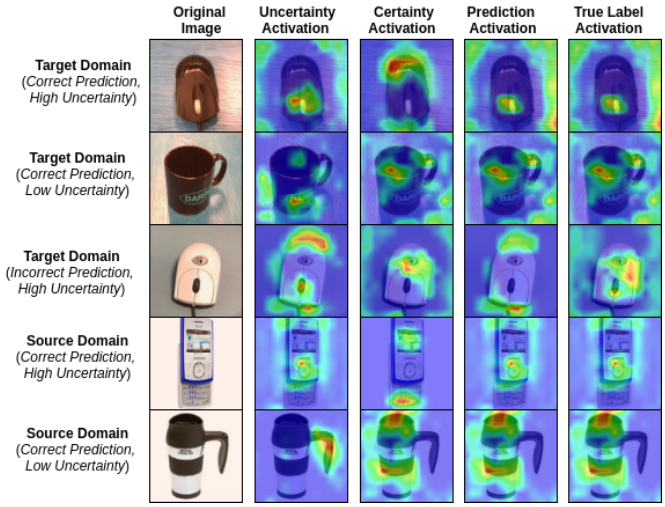
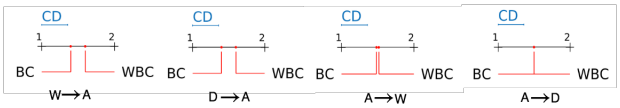


Figure 1: Qualitative Analysis of Certainty Map vs CAM

1.3. Predictive uncertainty over Class activation map(CAM)

In **Fig.**1 we compare the CAM [8] based activation and certainty based activation in different scenarios of model predictions after training the model. The uncertainty activation and certainty activation are generated through the positive and negative activation of uncertainty of classifier respectively. The prediction activations are generated through the predicted label while true label activations are generated through the ground truth label. We observe that, when the prediction is correct, with low uncertainty (2nd and 5th row), the certainty activation, prediction activation and true label activation all are similar. But if prediction is correct with high uncertainty (1st and 4th row), the certainty ac-



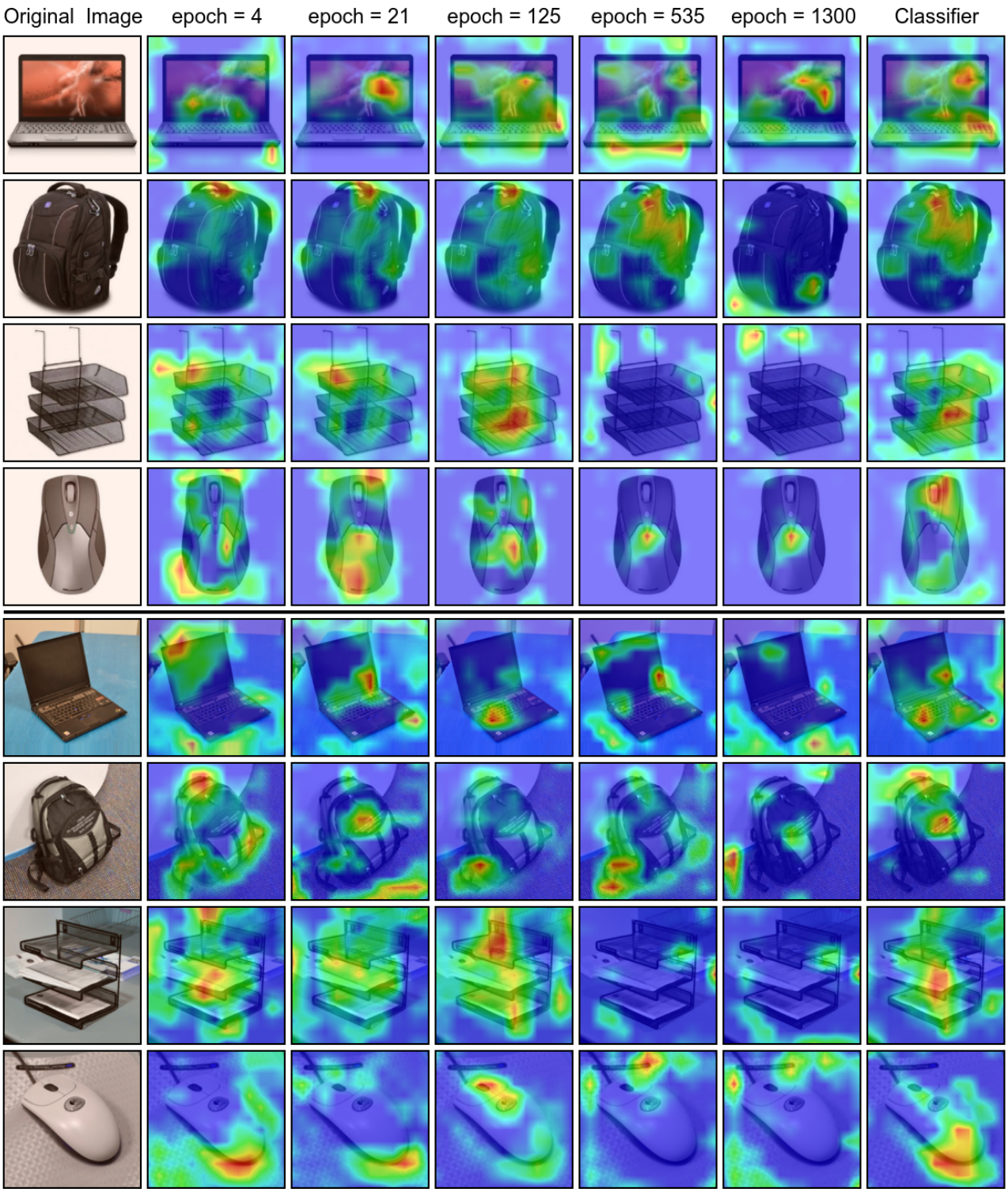


Figure 2: Attention visualization of the last convolutions layer of AlexNet [4]. Images from the source domain (**A**) are shown in the top four rows, and images from the target domain (**W**) are shown in the bottom four rows. In each row, the leftmost image represents the original image and the rightmost image represents the classifier's attention map for ground truth class label. From left to right, the attention map of discriminator's predictive certainty is illustrated at different training stages. We can see as the training progress, the discriminator's attention map changes from the background to the foreground, and then to the regions which can not be adapted further.

tivation is different than the prediction and true label activation, whereas uncertainty map is similar to prediction and true label activation. In this case CAM (true label or predicated activation maps) is activated from the uncertain region, thus CAM is not reliable as it does not incorporates uncertainty. While certainty maps not affected by the uncertain regions and can provide better explanation. For wrong predictions, with high uncertainty (3r row) the certainty activation matches to the true label activation, while the prediction activation is more similar to uncertainty activation, thus prediction is made using uncertain regions. Hence, certainty activation maps are more reliable.

1.3.1 Ablation Study: Without Bayesian Classifier

We use the Bayesian classifier to estimate the prediction uncertainty for the source and target for better visualization of domain adaption task using Certainty Maps rather than CAM. For fair comparison, we have also evaluated our architecture without Bayesian classifier, and reported results in the table for Office-31 dataset on Alexnet model. The performance does not change much without Bayesian classifier than the reported model (CADA-P) in the main paper. We also performed statistical significant analysis between the Bayesian and non-Bayesian classifier, shown in figure and it can be observed that both models are not statistically different.

Method	A→W	A→D	D→A	W→A	Avg
w/o Bayesian	83.30	80.20	61.27	57.05	70.46
with Bayesian	83.40	80.10	59.80	59.50	70.70

References

[1] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International Conference on Machine Learning*, pages 1180–1189, 2015. 1

[2] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256, 2010. 1

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1

[4] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, 2012. 1, 2

[5] Andrey Malinin and Mark Gales. Predictive uncertainty estimation via prior networks. In *Advances in Neural Information Processing Systems*, pages 7047–7058, 2018. 1

[6] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016. 1

[7] Christian Thiel. Classification on soft labels is robust against label noise. In *International Conference on Knowledge-Based and Intelligent Information and Engineering Systems*, pages 65–73. Springer, 2008. 1

[8] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016. 1