CSE 291 A00

Nitesh Gundavarapu (A53317353) January 24, 2020

1 Title:

Deep Transfer Learning with Joint Adaptation Networks

2 Summary:

This paper proposes to align joint distribution of activations across task-specific layers in DNNs to achieve domain adaptation. Joint Maximum Mean Discrepancy is formulated in tensor product RKHS and an empirical O(n) strategy over mini-batches is proposed to minimize it. Further, optimal MMD is learnt by posing it as minimax problem where the network learns weights to minimize worst case discrepancy. Experimental comparisons are presented on Office-31 and ImageCLEF-DA datasets against previous methods.

3 Strengths:

i) Joint distribution across layers is aligned instead of respective marginals. It is claimed that this is equivalent to aligning the joint distribution of labels and inputs, and thus doesn't require the assumption of same conditional distributions.
ii) Strong experimental results over the multi-layer counterparts, especially DAN, which aligns each layer's marginals separately.
iii) $O(n)$ approximation allowing for mini-batch updates.
iv) Optimal MMD is figured out adversarially, which circumvents the problem of choosing *rich* function class for kernel methods.

4 Weaknesses:

i) The claim that aligning joint distribution of multiple task-specific layers is same as aligning joint distribution of labels and inputs is not theoretically or empirically justified. ii) Adversarial variant is computationally expensive as shown in Figure 3d. with not much improvement over non-adversarial variant. iii) It's unclear which layers to choose for adaptation and on what basis iv) Not scalable for higher dimensional layers such as convolution blocks

5 Analysis of Experiments:

i) Strong improvement on Amazon->Webcam task (JAN 74.9%, AlexNet 61.6% and DAN 68.5%) and Amazon->DSLR task (JAN 71.8%, AlexNet 63.8% and DAN 67%) on Office-31 dataset using AlexNet base network. Similar strong results are shown in ImageCLEF-DA on most of the tasks proving the effectiveness of matching joint distributions.
ii) Results using ResNet (A->W 85.4%) base network are better than AlexNet (A->W 71.8%) and so are the improvement over ResNet baselines. This shows method scales well for deeper networks with more transferable layers.
iii) The improvement between the adversarial variant and non-adversarial variant in almost all settings isn't too great (<1%).

6 Possible Extensions:

Weaknesses ii,iii and iv can be addressed by following process: First, experimentally figure out which layers are transferable and which layers are not, say using Kernel MMD on pre-trained network, and then apply JMMD. Further, to reduce the computational cost of both adversarial and non-adversarial variants, even for convolutional blocks, the approximation of using mean in high dimensional space as presented in *Deep Domain Confusion* paper is very helpful.