1 Summary

The authors address the issue of domain adaptation by using **joint** distribution of the activations in domain-specific layers as a substitute for original joint distribution and use joint MMD to align the source and target domain features. JMMD loss is incorporated in domain-specific layers of AlexNet and ResNet and experiments are run on Office-31 and ImageCLEF-DA datasets and it sets new state of the art record.

2 Strengths

- The addition of a neural network to approximate a rich enough kernel for differentiating the joint distribution of source and target domains and using adversarial training strategy to maximize JMMD is clever.
- Directly aligns the joint distributions of source and target domains without any underlying assumption of changes in only marginal distribution shift or only in conditional distribution shift or both.
- Using the kernel trick, exact computation of JMMD takes $O(n^2)$ time-complexity which is rather undesirable. The paper uses unbiased estimate of JMMD which can be computed in $O(n)$.

3 Weaknesses

- The authors take joint distribution of last 3 layers as surrogate of original $P(X_s, Y_s)$ and $P(X_t, Y_t)$. Its not discussed how good of an approximation this is.
- Use of unbiased estimate of JMMD in $O(n)$ does lead to faster computation but it also has a higher variance than the $O(n^2)$ exact computation. The paper does not address this issue nor does it say that it wont lead to decrease in performance.

4 Critique of Experiments

- The t-SNE plots for activations in ResNet based DAN and JAN models are analyzed. It shows that target categories are discriminated much more clearly by the JAN source classifier.
- To measure and compare cross-domain discrepancy, the authors compute JMMD loss for CNN, DAN and JAN models in the last two layers of AlexNet. I am not sure how much insightful this observation is since the JMMD loss will be smallest for JAN model (which is what they found in experiments) since it explicitly reduces that metric by incorporating it in loss function.
- Test error vs epochs plot for JAN and JAN-A reveal that JAN converges much faster than JAN-A. This is because JAN uses non-parametric JMMD loss while JAN-A uses JMMD parameterized by θ.
- Plot of λ, relative JMMD weight in total loss, on test accuracy reveals that there exists a λ for which accuracy is maximum. This shows that there is a proper trade-off between joint distribution adaptation and source risk loss that enhances transferability.

5 Follow Ups/Extensions

- As an extension, we can vary the strength of the adversarial network parametrized by θ and see the effect on test accuracy. Intuitively, with a very powerful adversarial network, the network should have a hard time aligning the source and target domain joint distributions.