Cluster Triplet Loss for Unsupervised Domain Adaptation on Histology Images

Supplementary Material

6. Additional Ablation Studies

Clustering methods To find the representative cluster centres from our source dataset we also tried a consensus clustering approach with hierarchical clustering. Despite the theory behind this method we found it not to be robust, with unrealistic cluster distributions. Furthermore, due to the large sample size of the node-level data, we had to drastically reduce the size of the dataset before applying the consensus clustering, which could lead to loss of information. The number of features was small enough to ensure that dimensionality reduction on the number of features was not required. We tried a self-organizing map (SOM) to reduce the dimension of the data, with different map sizes, and extracted the resulting cluster labels from the consensus clustering on the reduced SOM sample set. To calculate the cluster centres, for each cluster label we then collected all the SOM samples with that label, and calculated the mean of their feature set. However, we found the traditional KMeans approach more robust, explainable and efficient for finding the cluster centres of the source data. The only downside of KMeans compared to hierarchical consensus clustering is that the user must choose the number of clusters before applying the model, but this can be selected in a methodical way using clustering metrics.

7. Implementation of SOTA Methods

In all of our SOTA implementations we had to adapt the method to our prediction problem. Firstly, these methods were implemented for multi-class classification or segmentation, so we had to adapt the code for a binary prediction problem, which in some cases meant there was less information from class pseudo-labels (since there are maximum two classes here). Secondly, most of these methods use either the original image or a pixel-level representation as input, so we had to adapt the methods to work on features of segmented tissue regions within each image.

Though the SRDC method is unsupervised [30], it requires the full source dataset, including source labels, during training, and the target labels are used to validate the model during training. Hence for this method we use the WSI label as the label for each individual segmented tissue region within the WSI.

In our implementation of Distill-SODA [32], we were unable to use their proposed adversarial data augmentation method, AdvStyle, since our source model is frozen for this research, and we aim to adapt it as is. It's possible this method could work better on our data if we were to use this pre-training method on our source data.

In the implementation of TCL [10], this method is originally introduced in a supervised setting, using the target labels to identify which class centre it should be using as the positive sample in the triplet loss function. We implement an unsupervised variant of TCL, using our fixed source cluster centres instead of learnable class centres, but keeping the idea that the negative example in the triplet loss function is the nearest negative cluster centre i.e. the second closest cluster centre, where the closest is our positive example.