Domain Adaptation using Optimal Transport for Invariant Learning in Histopathology Datasets

Kianoush Falahkheirkhah [1], Alex Lu [2], David Alvarez-Melis [2], Grace Huynh [2]

[1]University of Illinois Urbana-Champaign [2]Microsoft Corporation

Objective

We aim to develop a robust machine learning model for histopathology data analysis that can overcome batch effects and ensure consistent performance across different healthcare institutions.

Background

Batch effects induce systematic differences in histopathology images, arising from varying slide preparation methods, staining protocols, and data processing techniques. These can be institution-specific or even be found within the same institution. While human pathologists are capable of disregarding these differences, they significantly impact machine learning models. This leads to a degradation of model performance when deployed in different hospitals. Our approach addresses these issues by integrating an optimal transport (OT) based loss function during model training[1].

Optimal Transport

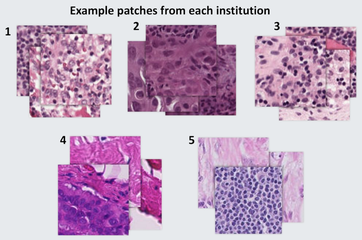
OT is a mathematical framework for comparing, aligning, and transforming probability distributions. It defines a distance between distributions that captures the 'cost' of transforming one distribution into another. Given two discrete distributions $P = \{(x_i, p_i)\}$ and $Q = \{(y_j, q_j)\}$, the OT problem can be defined as:

$$W(P, Q) = \min_{\Gamma \in \Pi(p,q)} \sum_{i,j} \Gamma_{ij} c(x_i, y_j) \quad (1)$$

where $c(x_i, y_j)$ is a cost function that quantifies the 'distance' between points x_i and y_j, and $\Pi(p, q)$ is the set of all joint distributions Γ whose marginals are respectively P and Q.

Dataset

We utilized the Camelyon17-WILDS benchmark dataset[2], specifically designed to tackle distribution shift in machine learning applications. This dataset, derived from Camelyon-17, contains Hematoxylin and Eosin (H&E) stained images of breast lymph node resections from five institutions. The dataset comprises 96x96 pixel patches from a whole-slide image of a lymph node, each annotated by experts as tumor or non-tumor.



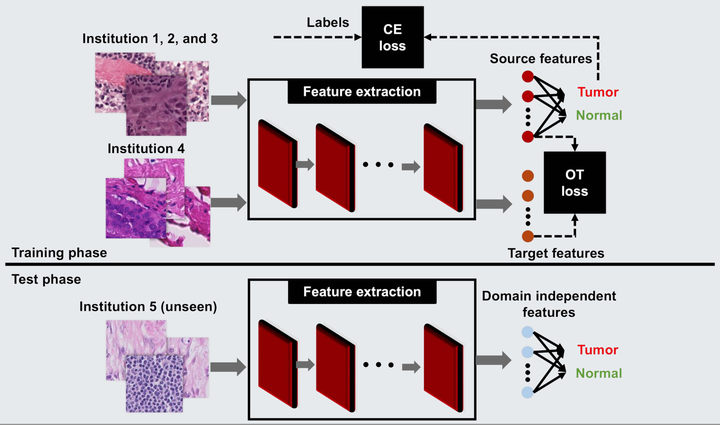
Dataset	# Patches	# WSI
Train	302,436	30
Validation	34,904	10
Test	85,054	10

OT-regularized Approach

We propose an "OT-regularized" approach where we integrate OT loss during model training to discover more robust, domain invariant features. The total objective function is defined as:

$$L_{\text{total}} = L_{CE} + \alpha * L_{OT}, \quad (2)$$

where α is a hyperparameter controlling the strength of cross-domain regularization. The cross-entropy loss L_{CE} is defined inline as $L_{CE} = -\sum(\log(C(\phi(X_S; \theta_\phi); \theta_C))$, where X_S is a batch of samples from the source domain, and θ_ϕ and θ_C are the weights of the featurizer ϕ and classifier C, respectively. The domain generalization L_{OT} is defined inline as $L_{OT} = \sum(\text{OT}(\phi(X_S, \theta_S), \phi(X_T, \theta_S)))$, where X_T is a batch of samples from the target domain (validation). Training on two sets of institutions – the labeled institutions 1-3, and the unlabeled institution 4 – enables this robust feature identification. The model is then evaluated at test time using these domain-independent features.



Choosing optimum value of α

We trained a ResNet50[3] model for this image classification task. We calculated the Optimal Transport (OT) distance between the final-layer features of training and validation batches during training. The total loss was a sum of cross-entropy and OT loss, regulated by tuning parameter α. The optimal $\alpha = 0.1$ was selected for highest validation dataset accuracy.

α	**0.00001**	**0.0001**	**0.001**	**0.01**	**0.1**	**1**
Accuracy of validation	0.882 (0.011)	0.882 (0.002)	0.871 (0.004)	0.882 (0.007)	**0.891 (0.005)**	0.712 (0.165)
Accuracy of test	0.799 (0.029)	**0.857 (0.011)**	0.811 (0.019)	0.826 (0.012)	0.850 (0.019)	0.733 (0.150)

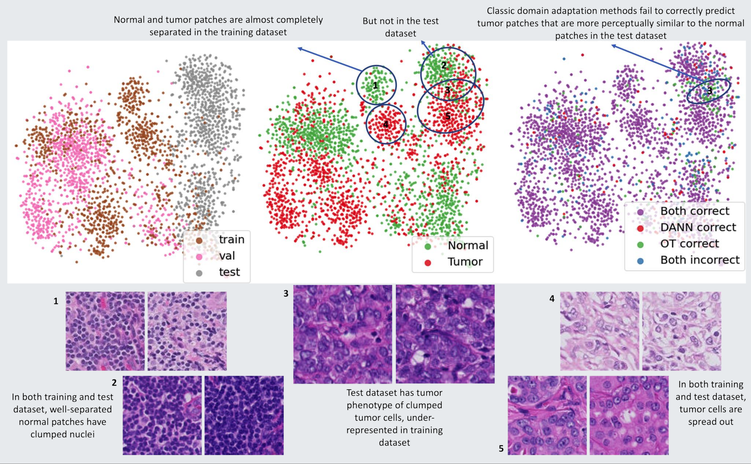
Comparing with DANN Method

Our approach was compared with the established Domain Adaptive Neural Network (DANN) method[4], a popular approach for domain adaptation. Our goal was to demonstrate that replacing DANN's adversarial objective with an OT loss could lead to a more nuanced correction. For fairness, we trained a DANN model using the same architecture and data splits as our method. The results showed our method surpassed DANN in performance on both validation and test datasets, notably with a wider margin on the latter.

	OT-regularized	**DANN**
Accuracy of validation	0.891 (0.005)	0.873 (0.014)
Accuracy of test	0.850 (0.019)	0.796 (0.052)

Exploring Feature Space

Our exploration of the feature space, conducted through a qualitative evaluation using features extracted by a pre-trained ResNet-50 and t-SNE projection, revealed distinct clustering patterns. In the training dataset, tumor and normal tiles formed distinct clusters, while in the test dataset, they formed a continuous single cluster. We noted that a significant portion of the test tiles contained image features that were underrepresented in the training/validation set. Further visualization showcased notable differences in the pathology tiles across all datasets, including variations in color, staining, and biological features. Despite these variations, our OT method consistently outperformed DANN, even in the poorly represented feature spaces during model training, indicating its potential for capturing the full distribution of image variability on the feature representation level.



Synergistic Opportunities

performance on the WILDS-Camelyon17 dataset. These include the Empirical Risk Minimization (ERM), CORAL, IRM for domain generalization, and Group DRO for subpopulation shifts. While these methods exhibit variable efficacy, none surpassed the ERM baseline in test performance, highlighting the challenge of domain adaptation on histopathology images. Contrarily, our method significantly outperformed the ERM baseline. Advanced approaches for batch effect correction, such as synthetic images (MBDG) and vision transformers (SGD Freeze Embed) occasionally outperform our OT method. However, we recognize potential synergies between these methods.

Methods	**ERM**	**CAROL**	**IRM**	**Group DRO**	**MBDG**	**SGD (Freeze-Embed)**	**OT-regularized**
Accuracy of validation	0.849 (0.031)	0.862 (0.014)	0.862 (0.014)	0.855 (0.022)	0.881 (0.018)	0.952 (0.003)	0.891 (0.005)
Accuracy of test	0.703 (0.064)	0.595 (0.077)	0.642 (0.081)	0.684 (0.073)	0.933 (0.010)	0.965 (0.004)	0.850 (0.019)

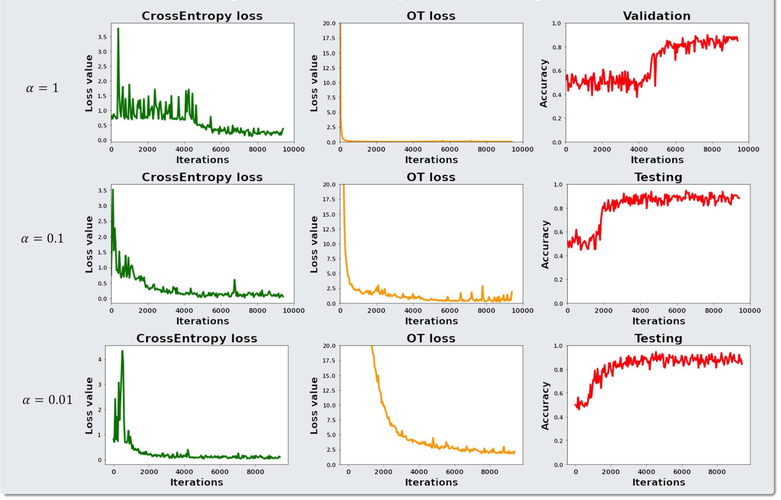
Effects of Stain Augmentation

We evaluated the influence of stain augmentation on model generalization in this dataset. Our analysis involved a comparison between the OT-regulated method, ERM with data augmentation (by WILDS), and the H&E-tailored RandAugment approach. The H&E-tailored RandAugment method showed superior performance over the OT-regulated approach on the Camelyon17-WILDS dataset. It is noteworthy that while stain augmentation was effective in this dataset, it might not be as successful in others due to inherent morphological differences that could arise from batch effects.

Methods	**ERM w/ data aug**	**H&E-tailored RandAugment**	**OT-regularized**	**H&E-tailored RandAugment w/ OT-regularized**
Accuracy of validation	0.906 (0.012)	0.914 (0.006)	0.891 (0.005)	0.912 (0.005)
Accuracy of test	0.820 (0.074)	0.922 (0.022)	0.850 (0.019)	0.924 (0.006)

Analysis of Convergence

This section evaluates the training process with varied values of α. The leftmost plots display the Cross-Entropy (CE) loss, the middle plots show the Optimal Transport (OT) loss, and the rightmost plots indicate model performance on the validation set throughout training. Each row represents a different value of α, showcasing the impacts of tuning on model convergence.



Conclusion

- Our study demonstrates the effective use of optimal transport (OT) in mitigating domain differences during model training, enhancing overall adaptability.
- The OT approach can accurately classify test image tiles even in regions with scarce training examples.
- Our OT loss demonstrates robustness against shifts in image feature space between training and test sets.
- While the OT loss doesn't outperform all current batch correction methods, it opens avenues for synergy, potentially enhancing methods like stain augmentation and advanced architecture designs.
- Future work should explore the performance of our method in wider fields-of-view, more challenging tasks, and in additional unseen domains.

References

[1] F. M. Howard, J. Dolezal, S. Kochanny, J. Schulte, H. Chen, L. Heij, D. Huo, R. Nanda, O. I. Olopade, J. N. Kather, et al. "The impact of site-specific digital histology signatures on deep learning model accuracy and bias." Nature communications, 12(1):1–13, 2021.

[2] P. W. Koh, et al. "Wilds: A benchmark of in-the-wild distribution shifts." International Conference on Machine Learning. PMLR, 2021.

[3] K. He, X. Zhang, S. Ren, and J. Sun. "Deep residual learning for image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition, 2016.

[4] H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, and M. Marchand. "Domain-adversarial neural networks." arXiv preprint arXiv:1412.4446, 2014.