STRENGTH IN DIVERSITY: UNDERSTANDING THE IMPACTS OF DIVERSE TRAINING SETS FOR SELF-SUPERVISED PRE-TRAINING FOR HISTOLOGY IMAGES

Kristina L. Kupferschmidt[1,2], Eu-Wern Teh[1,2], Graham W. Taylor[1,2]

BACKGROUND

DIGITAL HISTOPATHOLOGY

- Whole slide images (WSIs) are very large (100k x 100k)
- Data annotation is expensive and time consuming
- Limited labelled data exists despite the fact that large digital archives of WSIs exist [1,2]

TRANSFER LEARNING

- Learns low-level features by pre-training on data often from **different domains** (e.g. natural images)
- Improves performance in histopathology and other medical imaging datasets [3]
- Self-supervised learning (SSL) is a subcategory

SELF-SUPERVISED LEARNING

- Pre-trained on **alternative** task using unlabeled data from the **same domain**
- Labels generated algorithmically with no human intervention
- Provides superior initialization to random weights [4]
- ImageNet pre-training often outperforms SSL

HOW IT WORKS:

STAGE 1: SELF-SUPERVISED PRE-TRAINING
Train a CNN to complete one of the self-supervised tasks with unlabeled data

STAGE 2: SUPERVISED TRAINING
Initialize Stage 2 model with weights from Stage 1 and fine-tune model with labelled data

STAGE 3: TESTING CLASSIFICATION ACCURACY
Load model from Stage 2 and test in unseen data

MOTIVATION

CROSS-DOMAIN PRE-TRAINING IS POORLY UNDERSTOOD:

- Domain specific data for pre-training may improve performance [3]
- Using embeddings from other non-medical datasets may improve performance [3]
- Few evaluations in histopathology

COMPOSITION OF SOURCE DATASETS:

- Diversity of examples included?
- Number of examples?
- Similarity to target dataset?

OBJECTIVE

To explore if using source data from different domains for simple SSL pre-training tasks can provide a superior initialization for digital histopathology images

METHODS

EXPERIMENTAL SETUP

1. **SSL pre-training**
   - Subsample 4000 images from one source dataset
   - Apply one SSL task (rotation or jigsaw)
2. **Target task**
   - Initialize with (i) SSL pre-trained, (ii) random, or (iii) ImageNet
   - Fine-tune using subset of target data (N_c=100, N_c=1000)

DATASETS

TARGET DATASET

PatchCamelyon (PCam)

- Lymphatic tissue histology images
- Binary classification (1) Benign (2) Metastatic
- 327,680 patches (96 x 96 px)

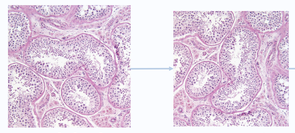
SOURCE DATASETS

1) Patch Camelyon (PCam)
2) TinyImageNet
3) Amsterdam Library of Textures (ALOT)
4) Colorectal Cancer Dataset (CRC)

SELF-SUPERVISED TASKS

1 ROTATION [4]

IMAGE PRE-PROCESSING



original IMAGE — rotated IMAGE

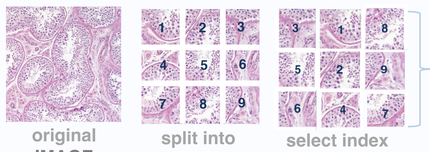
(0, 90, 180, or 270°)

PROXY TASK: PREDICT ROTATION CLASS → CNN Model

- CLASS 1 → 0°
- CLASS 2 → 90°
- CLASS 3 → 180°
- CLASS 4 → 270°

2 JIGSAW PUZZLE [5]

IMAGE PRE-PROCESSING



original IMAGE — split into TILES — select index JUMBLE TILES

Generate 100 puzzle combinations

INDEX	TILE ORDER
1	3,1,8,5,2,9,6,4,7
…	
100	2,4,6,8,1,3,5,7,9

PROXY TASK: PREDICT PERMUTATION INDEX → CNN Model

- CLASS 1 → idx 1
- CLASS 2 → idx 2
- CLASS N → idx N
- CLASS 100 → idx 100

METHODS

DIVERSITY EVALUATION

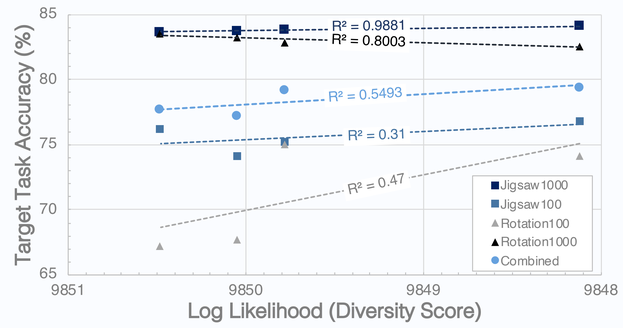
- Source datasets embedded into a pre-trained feature embedding (ResNeXt-10 with Instagram 1B)
- Fit a single Gaussian distribution to each dataset and compute average likelihood [6]

RESULTS

Table 1. Classification performance on PCam dataset across different initializations

Initialization	Likelihood (Div. Rank)	SSL Initialized Jigsaw Accuracy (%) (Rank)		SSL Initialized Rotation Accuracy (%) (Rank)	
		N_c = 100	N_c = 1,000	N_c = 100	N_c = 1,000
ImageNet	-	75.5 ± 2.1	83.3 ± 0.9	75.5 ±2.1	83.3 ± 0.9
Random	-	74.1 ± 1.0	77.9 ± 0.7	74.1 ±1.0	77.9 ± 0.7
SSL PCam	9850.05 (3)	74.2 ± 4.9 (4)	83.8 ± 0.3 (3)	67.8 ± 2.1 (3)	83.3 ± 0.9 (2)
SSL CRC	9848.12 (1)	76.8 ± 3.9 (1)	84.1 ± 0.7 (1)	74.2 ± 3.6 (2)	82.5 ± 0.7 (4)
SSL TinyImgNet	9849.78 (2)	75.2 ± 6.3 (3)	83.8 ± 1.7 (2)	75.0 ± 2.2 (1)	82.8 ± 1.0 (3)
SSL ALOT	9850.48 (4)	76.2 ± 6.3 (2)	83.7 ± 1.0 (4)	67.3 ± 5.0 (4)	83.6 ±0.7 (1)

Relationship of source data diversity and target task performance



CONCLUSION

- Using source data from different domains provides comparable improvement to using the same as the target dataset
- Diversity and model performance were correlated
- Selecting more diverse source data may improve target performance in low data conditions
- Diverse histopathology datasets for SSL pre-training may show the greatest promise

FUTURE WORK

- Implement additional self-supervised techniques
- Investigate source dataset combinations as additional sources of diversity
- Evaluate performance in additional histology target datasets

ACKNOWLEDGMENTS

NSERC CRSNG | VECTOR INSTITUTE | UNIVERSITY of GUELPH

REFERENCES

[1] Hegde, N., Hipp, J. D., Liu, Y., Emmert-Buck, M., Reif, E., Smilkov, D., ... & Stumpe, M. C. (2019). Similar image search for histopathology: SMILY. NPJ digital medicine, 2(1), 1-9
[2] Madabhushi, A., & Lee, G. (2016). Image analysis and machine learning in digital pathology: Challenges and opportunities. Medical Image Analysis, 33, 170–175.
[3] Cheplygina, V., de Bruijne, M., & Pluim, J. P. W. (2019). Not-so-supervised: A survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. Medical Image Analysis, 54, 280–296.
[4] Kolesnikov, Alexander, Xiaohua Zhai, and Lucas Beyer. 2019. "Revisiting Self-Supervised Visual Representation Learning." *arXiv [cs.CV]*. arXiv. http://arxiv.org/abs/1901.09005.
[5] Gidaris, Spyros, Praveer Singh, and Nikos Komodakis. 2018. "Unsupervised Representation Learning by Predicting Image Rotations." *arXiv [cs.CV]*. arXiv. http://arxiv.org/abs/1803.07728.
[6] Noroozi, Mehdi, and Paolo Favaro. 2016. "Unsupervised Learning of Visual Representations by Solving Jigsaw Puzzles." In *Computer Vision – ECCV 2016*, 69–84. Springer International Publishing.
[7] De Vries, T., Drozdzal, M., & Taylor, G. W. (2020) Instance Selection for GANs. https://github.com/uoguelph-mlrg/instance_selection_for_gans

AFFILIATIONS

[1] School of Engineering, University of Guelph – Guelph, ON
[2] The Vector Institute for Artificial Intelligence - Toronto, ON