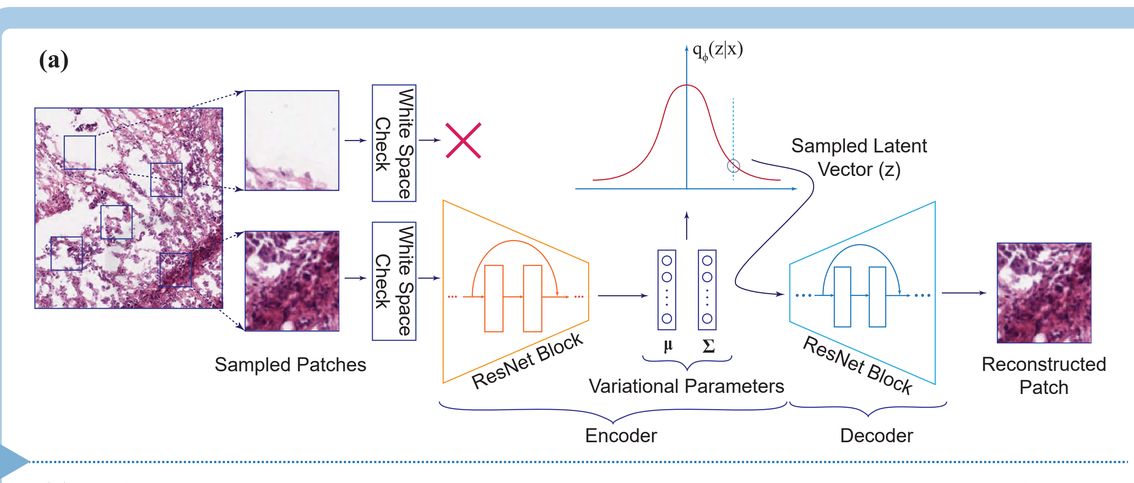
Clinically Relevant Latent Space Embedding of Cancer Histopathology Slides Through Variational Autoencoder Based Image Compression

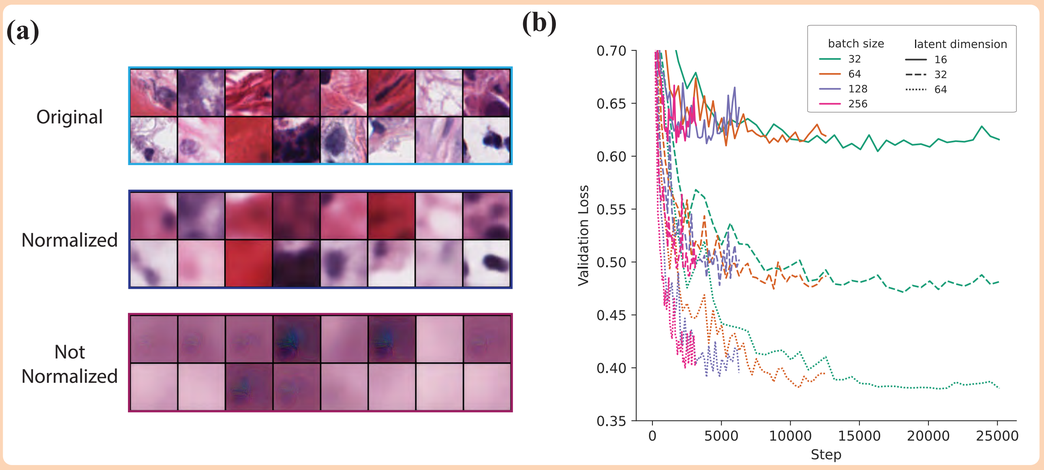
Mohammad Sadegh Nasr[*,1,2], Amir Hajighasemi[*,1,2], Paul Koomey[1,2], Parisa Boodaghi Malidarreh[1,2], Michael Robben[2,4], Jillur Rahman Saurav[1,2], Helen H. Shang[1,3], Manfred Huber[1], Jacob M. Luber[†,1,2,4]

* These authors contributed equally to this work. † Responsible author. Email: jacob.luber@uta.edu **1** Department of Computer Science and Engineering, University of Texas at Arlington **2** Multi-Interprofessional Center for Health Informatics, University of Texas at Arlington **3** Division of Internal Medicine, Ronald Reagan University of California Los Angeles Medical Center **4** Department of Bioengineering, University of Texas at Arlington

Introduction

This study aims to develop an efficient storage and indexing method for histopathological images using Variational Auto Encoders (VAEs). With databases like the NIH Genomic Data Commons (GDC) containing tens of thousands of Whole Slide Images (WSIs) for cancer research, effective compression and indexing are crucial. While traditional methods like JPEG2000 have limitations, neural networks show greater efficiency and fidelity. VAEs, in particular, maintain high image quality and low noise ratios at extreme compression levels. The proposed VAE-based approach compresses and indexes images in latent space up to a state-of-the-art 1:512 compression ratio, facilitating rapid and complex searches of whole slide H&E cancer images.





Importance of Normalization and Hyperparameter Tuning

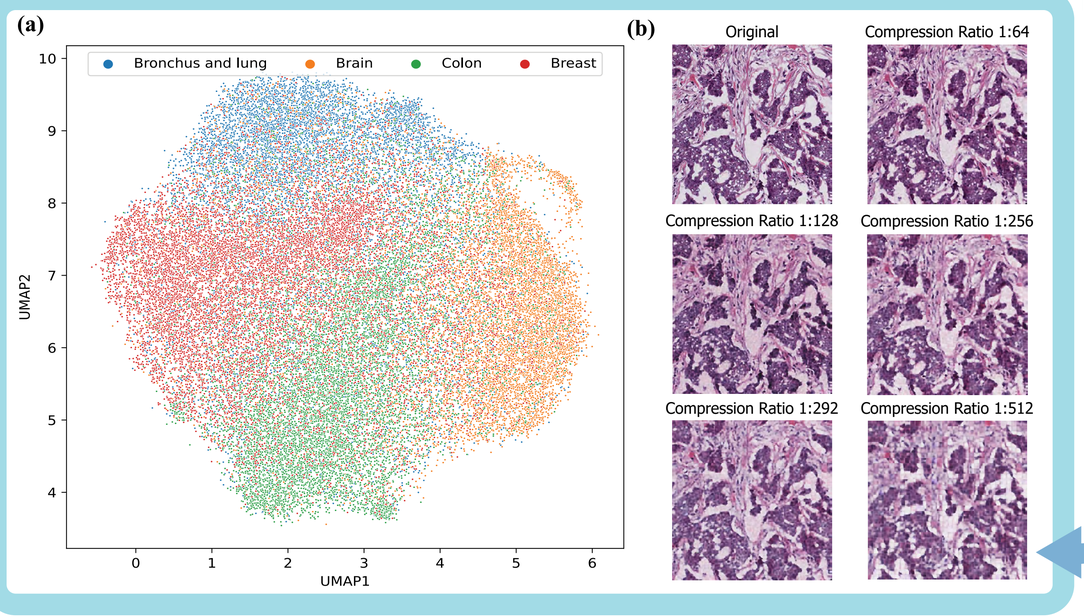
(a) Effect of Normalization. Normalization had a considerable effect on the pipeline's outcome, greatly improving the quality of the results. As confirmed through visual assessment, normalization is crucial for achieving acceptable outcomes.

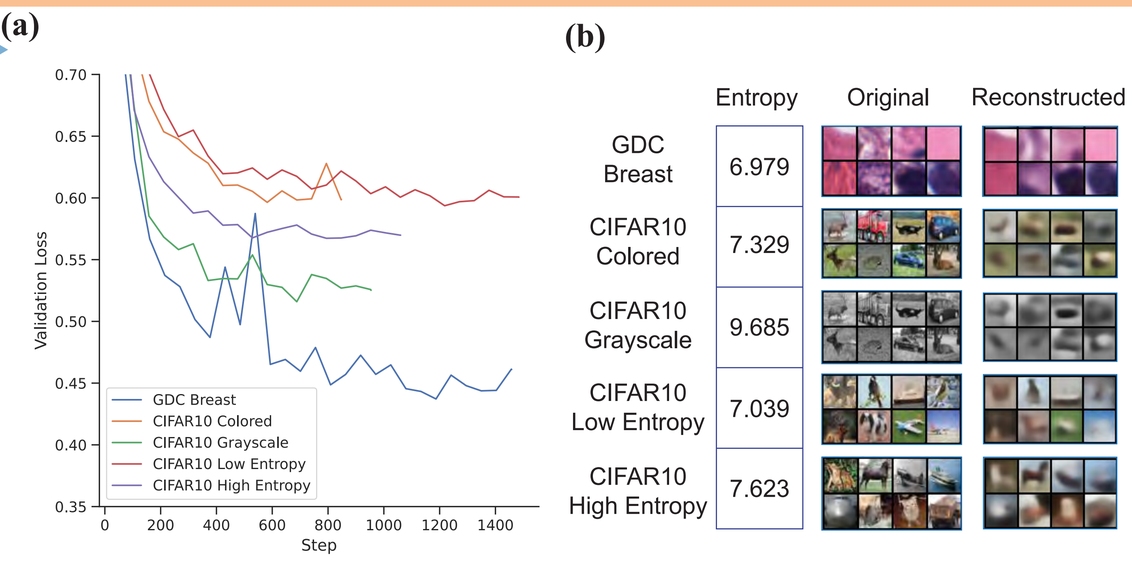
(b) Hyperparameter Tuning. The validation loss served as a metric to evaluate the performance of various models across different datasets. While larger batch sizes expedited objective minimization, smaller batch sizes yielded superior validation loss owing to heightened regularization effects. To strike a balance, experiments employed a batch size of 128. As anticipated, increased latent dimensions contributed to enhanced performance.

Training and Inference Pipelines

(a) Training Pipeline. During the training phase, non-blank and non-overlapping patches are randomly sampled from Whole Slide Images (WSIs) in the training and validation sets. The patches are normalized using the standard score method, and inverse transformations are stored for recreating slides. The model assumes a Gaussian prior and approximate posterior, with both the encoder and decoder utilizing ResNet18 architectures to minimize the number of model parameters for future clinical deployment.

(b) Inference Pipeline. During the inference phase, test images are fully tiled and each patch is processed through the trained networks before being reconstructed. For the UMAP experiment, the same patch sampling algorithm is used, but only sample latent variables are required, rather than the whole image.





Superior Performance on H&E Compared to Everyday Objects

Our compression model demonstrated superior performance (1:512 compression ratio) on histopathology slides in comparison to everyday object datasets like CIFAR10. Initially, we attributed the performance discrepancy to differences in entropy between the average images in each dataset. However, experiments conducted on high and low entropy folds of CIFAR10 contradicted this assumption. Despite having lower entropy, breast cancer H&E slides performed better, indicating that entropy isn't a reliable factor for explaining our VAE compression pipeline's performance on cancer imaging data.

We then explored color distribution as a possible contributor, given that H&E slides have a more limited color palette than CIFAR10 images. Experiments conducted on both colored and grayscale images revealed that the grayscale dataset experienced a lower validation loss, suggesting that a decrease in color content enhances compressibility.

Model Performance

(a) UMAP Plot. We used UMAP to visualize the latent space vectors learned by our pipeline, which captures both intra-tumor and across-tumor relationships, separating all four tissue types into distinct clusters. Furthermore, the UMAP identifies a unique sub-cluster of brain tumor samples that does not overlap with any other cancer types.

(b) Effect of Compression Ratios. Reconstructed images more closely resemble the original input images at lower compression ratios, with noticeable improvements in critical histologic features such as well-defined cell-to-cell borders and clearer distinctions between cytoplasmic and nuclear compartments.

Clinical Impact

Our pipeline can benefit clinicians and researchers by aiding in the accurate sub-typing and diagnosis of poorly understood cancers, such as brain cancer subtypes.

The UMAP visualization of the latent space reveals distinct cancer clusters and unique relationships between various cancers, preserving important histological features.

In the future, this embedding approach could be integrated with hospital systems and electronic health records to help clinicians diagnose patients with rare disorders by comparing UMAP embeddings of patient images to similar cases.

4 Minute Presentation

Paper Preprint

arXiv.org

Github Code

Positions @ LuberLab