Attention Pooling and Latent Space Analysis of WSI BYOL Embeddings

John Miller **Gabe Marx**

1. Introduction

Pathology is one of the last medical fields to join the digital revolution of medicine. The recent digitization of whole slide images (WSI) has opened opportunities for computational approaches to pathology problems. However, due to the recency of digital pathology there is a severe lack of annotated datasets. Acquisition of detailed pathology annotations presents a unique challenge. Each WSI, often scanned at 20x magnification (0.5 microns per pixel), contains several gigapixels of data. Detailed annotations of a single WSI could take several hours for a trained pathologist to perform. Additionally, their size provides a significant obstacle to computational modeling, as single images can take up prohibitively large amounts of memory. One approach to this problem is using a weakly supervised approach to training in which a single label is used for an entire WSI's worth of data. In this paradigm, a WSI is broken down into small (256x256 pixel) tiles. At prediction time, tiles are pooled into a single instance which represents the WSI along with its representative label.

In these experiments, we utilize Bootstrap Your Own Latent (BYOL), a state-of-the-art SSL technique[1]. BYOL was selected for numerous reasons: it provides superior performance, is relatively lightweight, and is flexible in its encoder network. We have previously trained a BYOL model on tiles drawn from WSIs, and will focus our study on an attention pooling mechanism and gaining a deeper understanding of the embeddings generated by BYOL.

2. Related Work

Much of the previous work in deep learning pathology has been in the field of cancer detection and classification. This work faces a similar difficult task in which there is a single label for a WSI which represents a large amount of data with a great deal of variance.

Campanella et al (2019) proposed a solution to the task of classifying WSI's as containing cancerous vs non-cancerous tissue. They approached the problem of representing a WSI into a single feature-vector from a bag of many patches tiles by implementing a

max-pooling. In this case, a slide was represented by the tile which had the highest probability of belonging to the positive class (cancerous). They were successfully able to detect cancer in WSI across a large array of different cancer types and tissues with an area under the curve of 0.98.

While Campenella et al was successful at recognizing cancerous tissue, their model and max-pooling approach was unable to satisfactorily accomplish more complex classification tasks such as cancer subtyping or assessing survivability. Lu et al (2021) provided a more complex architecture of attention based pooling. In this schema, features vectors are passed through a permutation invariant attention layer which yields an attention value for each tile in the slide. The slide is represented by taking a weighted average of all their tiles multiplied by their attention value, thus the tiles with the highest attention values represent the slide level feature vector the most while the tiles with the lowest attention values represent the slide the least. Their model was able to achieve an AUC of 0.991 on a cancer subtyping task.

3. Attention Pooling

In the past, the most common method for pooling multiple views of the same image has been average pooling. However, recent progress in the field of attention has offered many new insights into successful pooling techniques. Attention allows for the weighting of members of a latent space to be weighted in their importance to the final results. In our case, it weights important tiles in a WSI for the prediction of Braak.

Implementation

We present a relatively simple implementation of attention. First, a linear layer downscales the dimensionality of the feature vectors output by BYOL. Then, two identical attention heads are computed and separate activation functions are used (tanh and sigmoid, respectively). These attention heads are multiplied together and softmaxed to receive a proportional weight to apply to each tile. The tiles are then weighted and summed into a single feature vector representative of the entire WSI which then enters a linear ordinal regression layer which outputs probabilities for each Braak score.

More formally, for a given slide, S, has associated Braak score Y and is represented by N tile feature vectors $\{z_{1 \ldots} z_{N}\}$ where $z \in \mathbb{R}^{2048\ X\ 1}$. Each feature vector passes through a an initial linear layer $W_{1} \in \mathbb{R}^{512\ X\ 2048}$ such that $h_{k} = W_{1} z_{k}$ where $z \in \mathbb{R}^{512\ X\ 1}$.

The feature vector then gets passed through the two parallel attention backbone layers, $U_a \in \mathbb{R}^{384\ X\ 512}$ and $V_a \in \mathbb{R}^{384\ X\ 512}$ which then enters a nonlinear tanh function and sigmoid function respectively. The two outputs are multiplied together and put through a final attention layer, $W_a \in \mathbb{R}^{1X384}$. Then a softmax is performed across each tile to yield a scalar value a_j for each slide. The resulting attention value for each slide is then used to weight each tile level feature vector h_j and summed to create a composite slide level feature vector representation $s \in \mathbb{R}^{512X1}$.

$$a_k = \frac{\exp\{\mathbf{W_a}(\tanh(\mathbf{V_a}h_k) \odot \mathrm{sigm}(\mathbf{U_a}h_k))\}}{\sum_{j=1}^{N} \exp\{\mathbf{W_a}(\tanh(\mathbf{V_a}h_j) \odot \mathrm{sigm}(\mathbf{U_a}h_j))\}}$$

$$s = \sum_{k=1}^{N} a_k h_k$$

Each slide level vector, *s*, is then passed through a an ordinal regression layer, $W_c \in \mathbb{R}^{5\ X\ 512}$, developed by Cao et al 2020, which provides probabilities for the slide being greater than each Braak score.

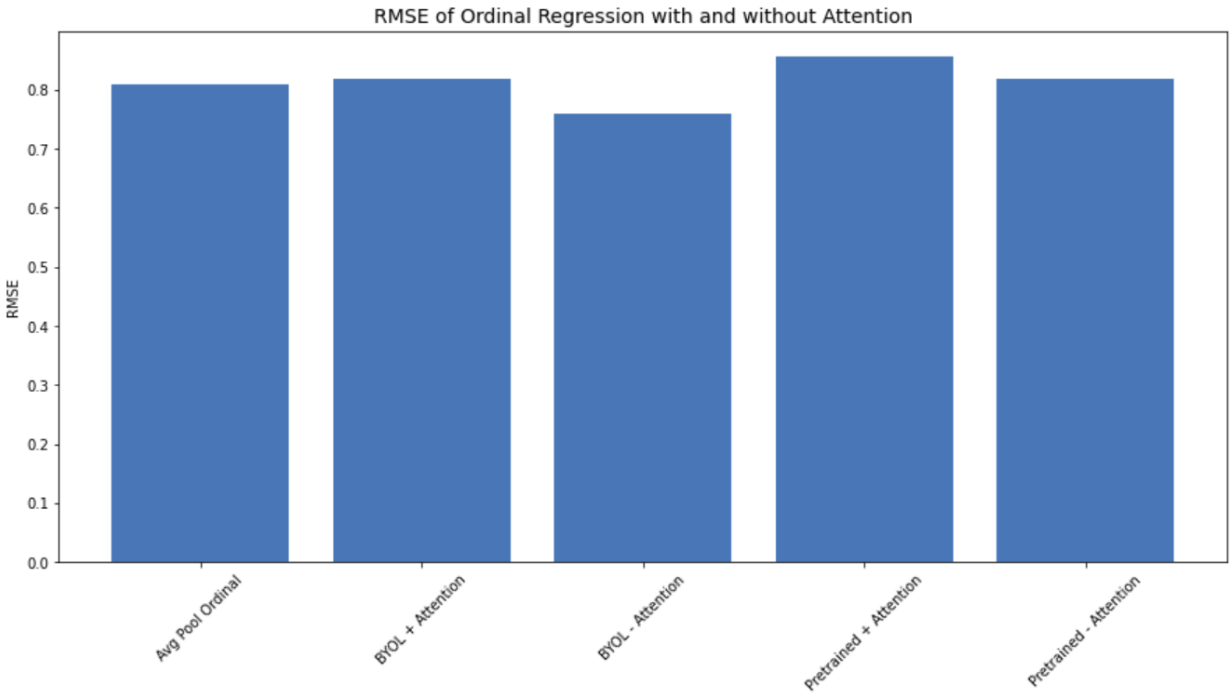


Figure X: RMSE of ordinal regression on both BYOL-trained and Imagenet-pretrained embeddings. (+ attention) are the groups which were pooled by attention, (- attention) denotes average pooling. **This will be updated with newest results.**

Attention Analysis

In addition to evaluating our model through RMSE, we also investigated the points of high and low attention through the WSIs. To do so, we generated heatmaps of the WSIs, where each tile was replaced by its attention value. We found that the attention primarily focused on physiologically relevant areas to tau burden such as the Dentate Gyrus and CA1. Interestingly, areas with dense white matter, such as the brain stem and neurofibril tracks, are given minimal attention. This is supportive of the underlying physiology of tau buildup and disease.

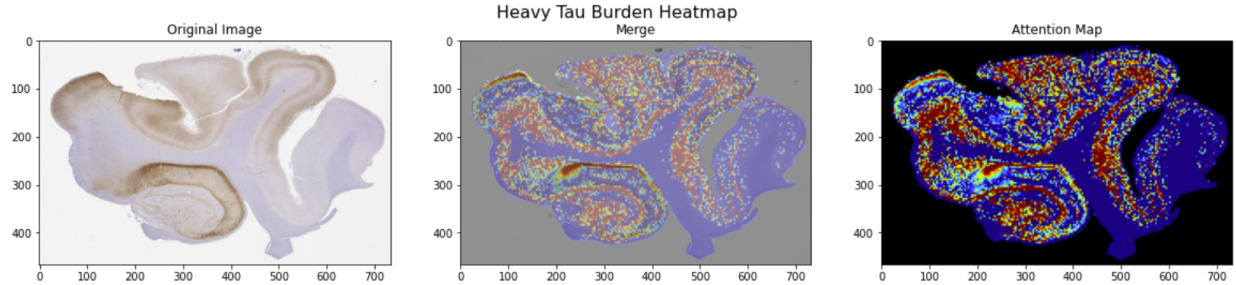
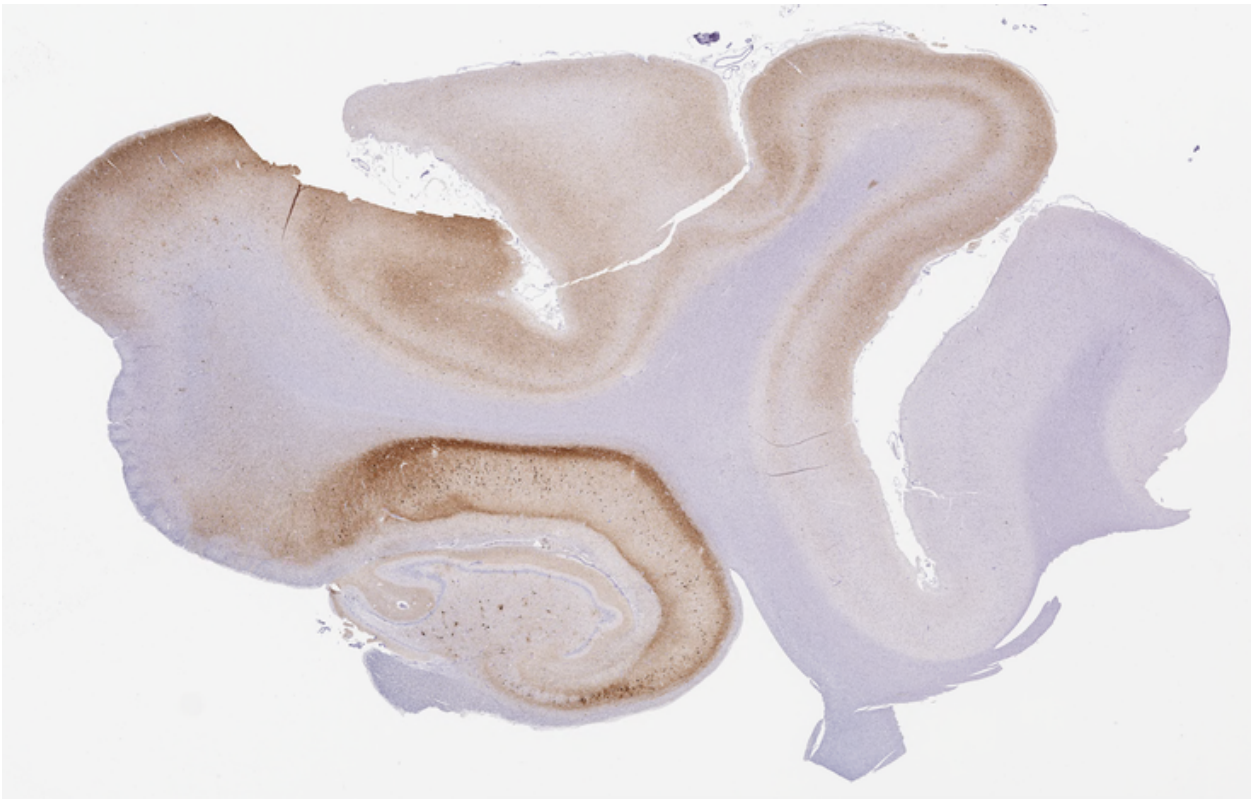


Figure X: A representative attention heatmap over a Braak 3 image.

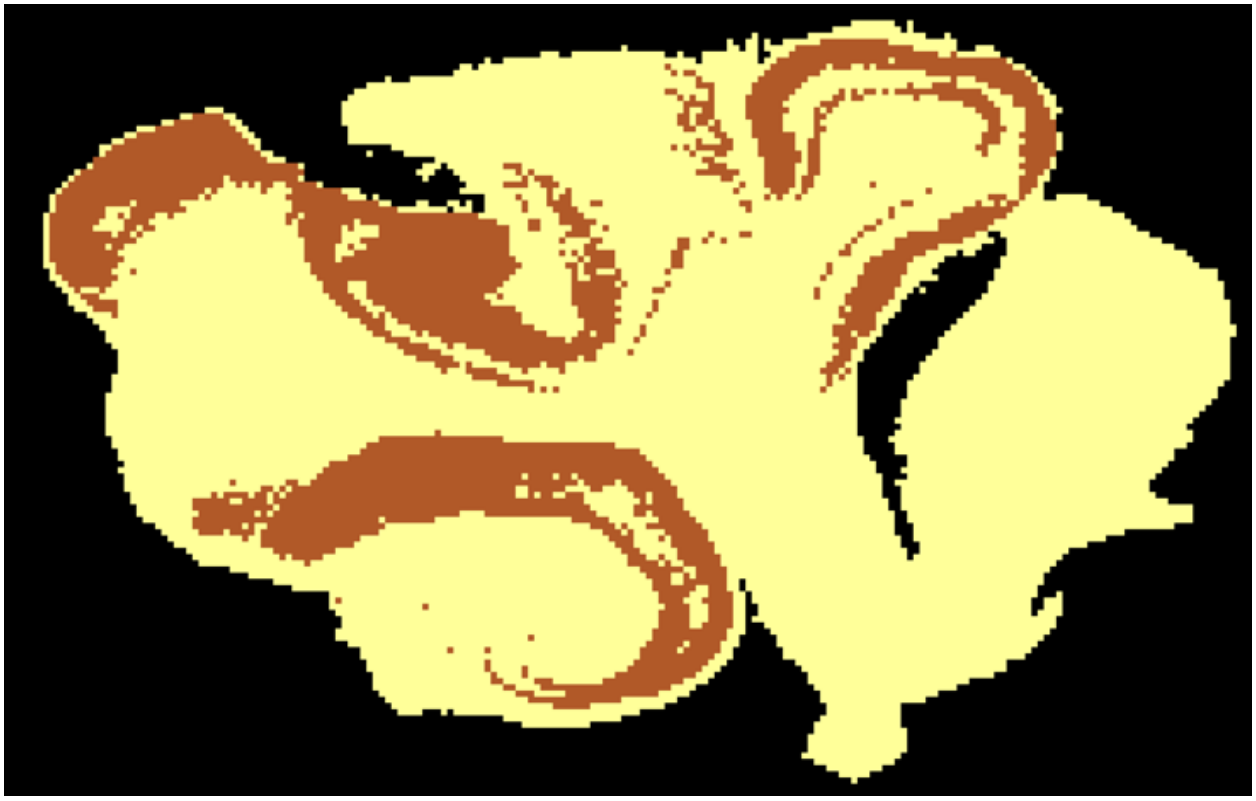
4. Latent Space Analysis

To investigate the quality of our feature space we employed k-means clustering and qualitatively assessed how well the clusters mapped onto the slide's anatomy and pathology. We used this approach to directly compare the quality of feature vectors created by our BYOL model versus feature vectors created by the Resnet50 pretrained with Imagenet.

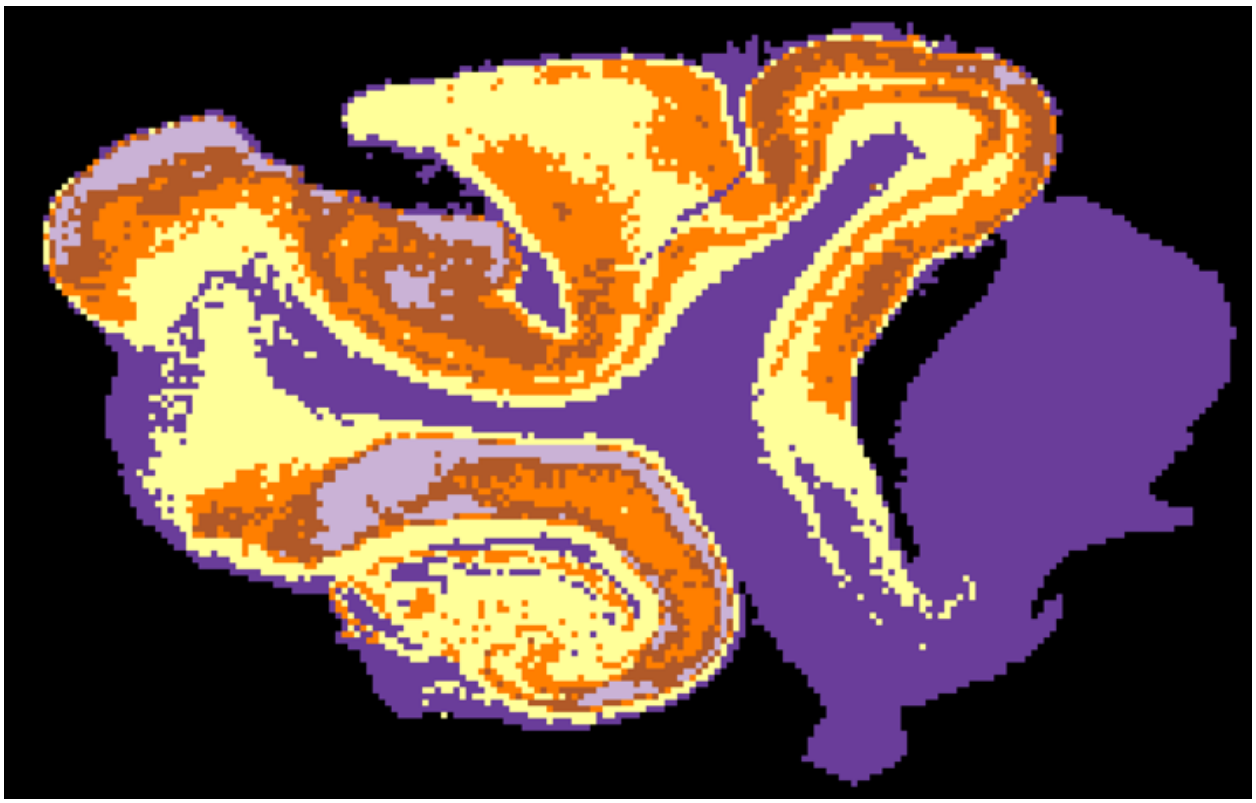
Here is the original image:



Here is the BYOL derived clusters at k = 2:

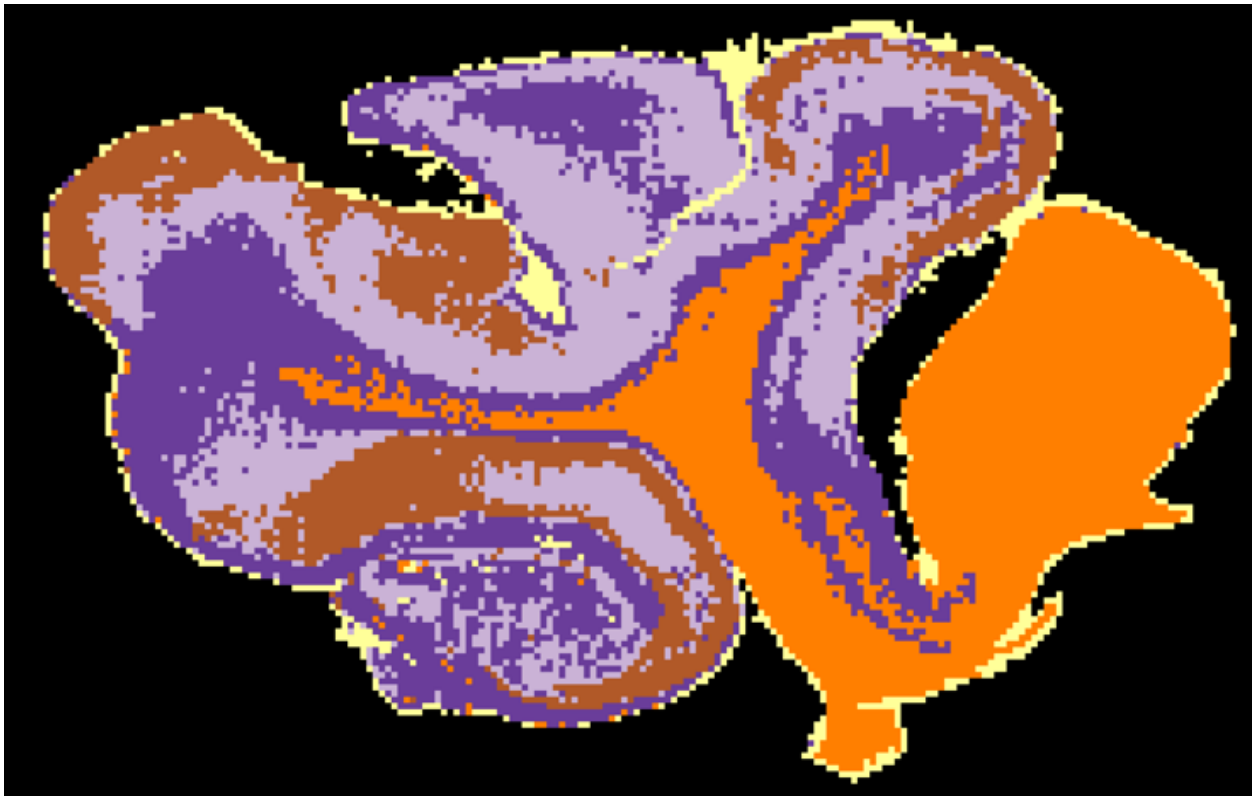


Here is the BYOL derived clusters at k = 5:



Here is the pretrained Resnet50 derived clusters at k = 2:

Here is the pretrained Resnet50 derived clusters at k = 5:



5. Discussion

References:

SSL: BYOL

MIL: CLAM, Dual Stream MIL, Campanella

https://arxiv.org/pdf/2004.09666.pdf

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7418463/pdf/nihms-1609511.pdf

https://arxiv.org/pdf/1901.07884.pdf