Enabling Interpretable Histopathology Representation Learning via Multimodal Language Guided Self-Supervision

Stanford CS224N Custom Project

Ekin Tiu
Department of Computer Science
Stanford University
ekintiu@stanford.edu

Anh (Tom) Nguyen
Department of Computer Science
Stanford University
anhn@stanford.edu

Abstract

Recent advancements in deep learning have led to notable improvements in the ability to learn meaningful embeddings from whole slide images of tumor (WSIs), applicable to clinically relevant tasks such as patient survival analysis or disease classification. These works are only guided by image based semi-supervised pre-training, however several works suggest that robust, human-interpretable image representations can be learned through supervision via unstructured language. To address this gap, we are the first to leverage unstructured pathology reports to guide representation learning of WSIs through neural natural language methods. We propose a contrastive pre-training pipeline that peforms co-attention on reports and WSI patches, enabling slide-level attention heatmap generation and zero-shot classification on downstream tasks. By comparing multi-class AUC to existing baselines on disease classification, we find that domain-specific pre-training on pathology data improves representation quality of image embeddings.

1 Key Information to include

- Mentor (custom project only): Elaine Sui

2 Introduction

The field of computational pathology involves extracting and interpreting meaningful features from whole slide images (WSI) of tumor. In recent years, the use of deep learning in computational pathology has led to notable advancements in the ability to learn and visualize meaningful representations applicable to downstream tasks such as patient survival analysis or disease classification Chen et al. (2022). In particular, since WSIs are gigapixels in size and are therefore traditionally complex to model, prior works have focused computational efforts towards learning regions of interest that are most relevant for clinical prognoses (Chen et al., 2021). These works are primarily guided by image based semi-supervised pre-training and do not leverage any other modalities. However, several works in the field of multimodal natural language processing demonstrate that better representations can be learned through supervision via unstructured language.

For instance, Radford et al. (2021) propose a joint image-language model that can map images and text into the same joint embedding space via self-supervised contrastive learning, which they name contrastive language image pre-training (CLIP). In addition to improving image embedding quality, CLIP enables multi-class classification to be performed in a zero-shot fashion, in which no fine-tuning on labeled data is required to classify images at high performance.

In our work, we are therefore motivated to extend prior joint image-language methods to guide interpretable representation learning of histopathology slides. To do so, we leverage raw pathology

reports paired with whole slide images (WSI) from The Cancer Genome Atlas (TCGA) to pre-train a joint image-language model. Unlike prior contrastive methods however, we propose the addition of a co-attention mechanism between text and image embeddings which enables an attention map to be created per slide for a text query. This contribution is motivated by the need to better understand and interpret clinically important regions on a large WSI.

Our contribution is therefore two-fold: 1) we are the first to apply joint image-text contrastive pre-training in the domain of histopathology slides and reports to improve embedding quality and enable zero-shot classification, and 2) introduce co-attention mappings between reports and slides to increase model interpretability with respect to text queries. To evaluate the performance of the pretrained model, we perform disease type classification on a held out test dataset. We compare our method to a ViT baseline, a CLIP pre-trained model, and compare results with and without language guided co-attention in both the linear probe and zero-shot settings.

3 Related Work

While language has guided the improvement of downstream tasks in other biomedical fields such as radiology, computational pathology has been slow in this adoption. Because WSIs are a complex and information rich data source, recent works have prioritized extracting signal via traditional neural image processing pipelines rather than drawing signal from additional modalities. Chen et al. (2022) apply DINO, a self-supervised pre-training method, to learn scale-invariant image embeddings. (Lee et al., 2022) apply graph neural networks on WSIs to derive histopathological features that have prognostic context. However, recent works have demonstrated that language can be used to improve image embeddings, even for complex medical image interpretation tasks, which motivates us to explore the untapped intersection of language and images in pathology.

Tiu et al. (2022) build upon the self-supervised joint image-language model CLIP proposed by (Radford et al., 2021) to demonstrate that a similar language-guided pre-training method can be applied for complex medical image interpretation tasks. In particular, they show that chest x-ray image embeddings can be used for downstream disease classification tasks and can match radiologist performance. They demonstrate this in the zero-shot setting, claiming that these models can perform well without training on any expert annotated datasets. Due to recent success of language models improving model performance in complex medical tasks, we are motivated to extend these methods to pathology to both improve embedding quality of histopathology slides and enable zero-shot classification via text queries.

Although prior work has not leveraged natural language, more recent methods have aimed to use multiple modalities to guide learning of histopathology slides. (Chen et al., 2021) apply co-attention on WSIs and genomic features to generate genomic guided attention maps, and obtain genome guided slide-based representations. We extend this method to natural language by using pathology reports instead of genomic features, to enable model interpretability as a tool to augment pathology analyses. However, (Chen et al., 2021) perform survival analysis in a fully supervised manner by training directly on labeled data. Our method aims to bypass the need for fully supervised training by using a contrastive loss proposed by (Radford et al., 2021) over the image and text embeddings from coattention. Thus, our work draws upon the joint contrastive image-pretraining methods of (Radford et al., 2021) and (Tiu et al., 2022) to enable zero-shot classification, while simultaneously introducing additional interpretability through co-attention methods proposed by (Chen et al., 2021).

4 Approach

4.1 Method

In this section, we present the overall framework, PathZero, which consists of contrastive pretraining on slide level embeddings learned from language-guided co-attention and corresponding pathology report text embeddings. The image enoder pipeline is referred to as Path-Coattn-ViT in the remaining sections of this paper.

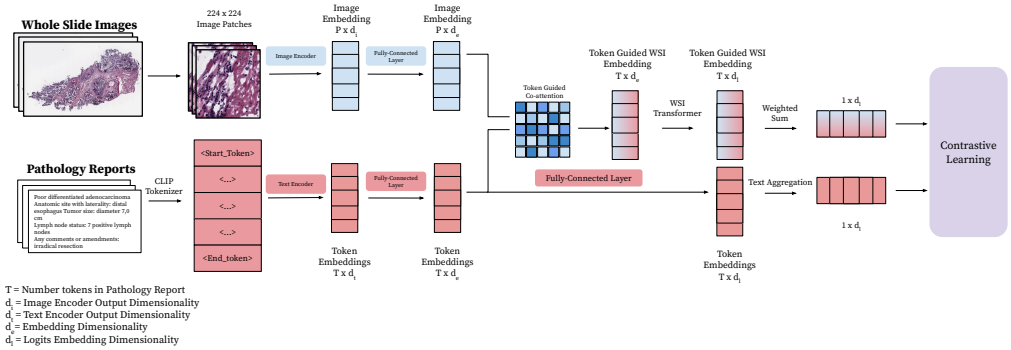

Figure 1: The contrastive training pipeline with co-attention. The image model learns features from raw pathology reports (n=1401) which act as a natural source of supervision.

4.1.1 Problem Formulation and Notations

Contrastive pre-training is a self-supervised task and framework that maximizes similarity of positive pair embeddings while minimizing the similarity of negative pair embeddings. In our setup, we propose a joint image-language contrastive pre-training setup where positive pairings consist of a whole slide image and its corresponding pathology report. In the framework of Multiple Instance Learning (MIL), whole slide images can be represented as a bag of patches where P is the number of patches and each i_t is a $224 \times 224 \times 3$ image. We represent a single image and a single report as

$$Im_i = \{i_1, \ldots, i_P\} \in \mathbb{R}^{P \times d_{im}}, Re_i \in \mathbb{R}^{T \times d_{txt}}$$

where P is the number of patches and T is the number of tokens. Batches of images and reports as $I_t \in \mathbb{R}^{N \times P \times d_{im}}$ and $T_t \in \mathbb{R}^{N \times T \times d_{txt}}$ respectively. We represent the full image encoder and text encoder as ϕ and γ, where for a single Im_i and Txt_i. $\phi : \mathbb{R}^{P \times d_{im}} \rightarrow \mathbb{R}^{d_l}$ and $\gamma : \mathbb{R}^{T \times d_{txt}} \rightarrow \mathbb{R}^{d_l}$. ϕ and γ are applied to each image and text within a batch to get $I_l \in \mathbb{R}^{N \times d_l}$ and $T_l \in \mathbb{R}^{N \times d_l}$. The objective is to minimize cross entropy loss over similarity scores computed pairwise between I_l and T_l.

4.1.2 Input preprocessing

Images Whole slide images are patched into 224x224 pixel images at 10x resolution after Otsu thresholding to retain relevant tissue regions and remove whitespace. We perform stain normalization on all patches to minimize the effect of domain shift across different sites.

Text The CLIP Text Encoder has a max token size of 76. To abide by this token constraint, we extracted only the "Diagnosis" section of the pathology reports. The TCGA pathology reports are not uniformly formatted, so we extracted with the heuristic of matching the starting token "Diagnosis" and extracting the following 200 tokens.

4.1.3 Training

As depicted in 1, our co-attention model can be broken up into three primary components: an image pipeline, a text pipeline, and the contrastive learning objective which maximizes the similarity between image and text embedding pairs.

Image Encoder Pipeline (ϕ) Due to computational constraints on storage, embeddings for all patches are passed through an ImageNet pre-trained ResNet-50 model, creating image embeddings with size d_i. This ResNet-50 embedding is then passed through a linear layer to create an image embedding that is of dimension $\in \mathbb{R}^{d_e}$ which is the same corresponding pathology report embedding before co-attention. We then co-attend report tokens across a slide to develop Token Guided WSI embeddings that are of shape T x d_l, where T is the number of tokens for a pathology report. Self-attention is then performed across this embedding resulting in a embedding $\in \mathbb{R}^{d_l}$ before being used in our contrastive learning framework. When applied to all images in a batch, we obtain $I_l \in \mathbb{R}^{N \times d_l}$

Text Encoder Pipeline (γ) Pathology report sections are first tokenized using the Byte Pair Encoding (BPE) scheme used by (Radford et al., 2021) to obtain a T x d_t sized embedding. Following this is a fully connected layer to create token embeddings $\in \mathbb{R}^{T \times d_e}$ used for co-attention. These token embeddings are passed through a fully connected layer to reach a dimensionality (T x d_l) which matches corresponding Token-Guided WSI embeddings post self-attention. This $T_l \in \mathbb{R}^{T \times d_l}$ is then used in the contrastive learning framework.

Contrastive learning Given a batch of text and co-attended image features in shape (N, d_l) and (N, d_l); these are I_l and T_l respectively. The $N \times N$ matrix of cosine similarities, also referred to as the logits by the authors, is then created by computing the dot product of I_l and $T_l^\top$. A cross-entropy loss is then computed between logits $\hat{y}_i$ and a pre-defined labels vector y_i of length N which contains values $0 \ldots N-1$.

$$\mathcal{L}_{CE} = -\sum_{i=1}^{N} y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)$$

4.1.4 Co-Attention

Our co-attention mechanism is a single-head attention layer which attempts to relate pathology token word embeddings to pathology slide pairs using the method described by Chen and He. We use $X \in \mathbb{R}^{T \times d_e}$ to guide features in the image embeddings $Y \in \mathbb{R}^{P \times d_e}$ using the following mapping:

$$\text{CoAttn}_{X \to Y}(X, Y) = \text{softmax}(\frac{W_q X Y^\top W_k^\top}{\sqrt{d_k}} W_v Y) \to A_{\text{coattn}} W_v X -> \hat{X}$$

Here the tokens embeddings X serve as our queries and our patches Y serve as our key-value pairs. $A_{coattn} \in \mathbb{R}^{T \times P}$ is the co-attention matrix used to computed weighted averages across our patches Y. Looking at this equation, we see that a single token $x_t \in X$ scores the pairwise similarity for how much each patch y_p attends to x_t. W_q, W_k, and $W_v \in R^{d_e \times d_e}$ are trainable weight matrices. A visual example can be seen in 5.

4.1.5 Zero-Shot Classification

For our zero-shot evaluation pipeline outlined in 2, we performed multi-class classification across a held out test set of whole slide images. Given an encoded input slide $Im \in \mathbb{R}^{d_{im}}$, and M class label queries $Q \in \mathbb{R}^{M \times T \times d_{txt}}$, we compute $\phi(Im)$ and all $\{\gamma(Re_i) \mid \forall Re_i \in Q\}$ to obtain an image embedding and M query embeddings. Cosine similarity is computed between the image embedding and query embeddings to obtain probability scores for each slide, $P \in \mathbb{R}^M$. We compute `Softmax(P)` to obtain a probability distribution over classes for a slide.

4.2 Baseline

The goal of this project is to see if we can improve image embeddings of pathology slides to be utilized in clinically relevant downstream tasks. Therefore, as a baseline, we performed our linear probe evaluation on a ViT-B@32 with no pre-trained weights. The ViT we used for the baseline has the same vision transformer model architecture as the pre-trained models. We expect meaningful performance improvements over these baseline-embeddings because they have not been pre-trained on natural images and language, nor on pairs of pathology slides and reports.

5 Experiments

5.1 Data

To evaluate the quality of embeddings, we perform 5-way disease type classification on ($n = 561$) images taken from cancer patients across 5 different disease types: Adenomas and Adenocarcinomas, Cystic, Mucinous and Serous Neoplasms, Ductal and Lobular Neoplasms, Gliomas, and Squamous Cell Neoplasm. The dataset we have chosen to use is The Cancer Genome Atlas (TCGA) dataset. The dataset includes roughly 11,000 pathology reports with 1 or more corresponding slide for each report. Our pretraining dataset consists of $n = 1401$ slide and pathology report pairs, randomly subselected from the full TCGA dataset due to storage constraints. For linear probe, we use a train

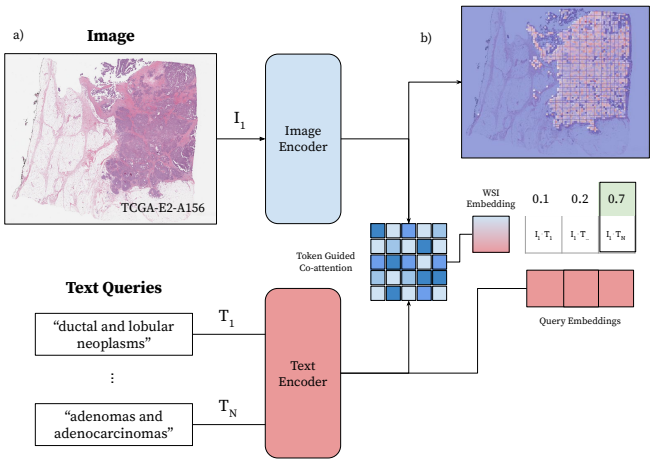

Figure 2: a) Diagram of zero-shot classification pipeline. b) Generation of attention heatmaps on slide based on textual query. Method allows model to map language to visual concepts with semantic meaning, enabling language-guided interpretations of predictions.

dataset with $n = 562$ and we evaluate all methods on a held-out test dataset with $n = 561$ randomly subsampled from the original TCGA cohort of 11,000.

5.2 Evaluation method and Metrics

We evaluate our model both using a linear probe as well as in a zero-shot fashion. We compare our method to two baselines, a ViT with random initialized weights and a model with CLIP pretrained weights from Radford et al. (2021). Additionally, we perform an ablation of our model embeddings with and without a co-attention pipeline appended to obtain slide level embeddings. To adapt this binary classification metric to the multi-class setting, we computed one-vs-rest (OvR) AUROC on each class, and computed the macro-average AUROC across all classes. For each class, the correct class is treated as a positive value whereas all other classes are considered to be negative. Due to class imbalance, both precision and recall are computed on the outputs of each of the classes. We report macro averaged precision and recall for each method.

5.3 Experimental Details

We performed the disease type classification task with 4 different models, all of which use the hyperparameters proposed by Tiu et al. (2022), and trained with an SGD optimizer at a learning rate of 1e-4 and momentum 0.9. All models are trained on a single Tesla T4 GPU with NVIDIA CUDA with 14GB of RAM. All slides, patches, and reports were stored on at 7TB Volume on AWS.

ViT Baseline A ViT-B@32 with an embedding dimension of 512. The baseline model uses no pre-trained weights. Additional model hyperparameters are selected based on the suggestions of (Radford et al., 2021). This model takes as input slide patches at a resolution of 224 x 224.

CLIP-ViT A ViT-B@32 that uses image and text encoder weights from CLIP pre-training on natural images and language. No further fine-tuning was performed on CLIP-ViT. We compare Path-ViT to CLIP-ViT to see the impact of in-domain self-supervised pre-training on embedding quality for pathology based tasks. To abide by memory constraints, CLIP-ViT is trained with a batch size of 64 patches.

Path-Coattn-ViT Image pipeline depicted in 1. Slide patches are passed through a frozen ResNet-50 pretrained on ImageNet to generate image embeddings. These embeddings are passed through token guided co-attention to get language-gudided WSI level representations. A smaller batch size of 8 is used for pre-training since unlike other self-supervised methods, we perform the contrastive

learning objective on the slide level rather than the patch level, so for each slide the data for all patches must be loaded into memory at once.

Path-ViT A ViT-B@32 that is pretrained on pathology images and pathology reports. The same ViT model architecture as CLIP-ViT is used. Both the image encoder and the text encoder use pre-trained weights of the CLIP model from Radford et al. (2021).

5.4 Results

Linear Probe Results Path-ViT performs better than CLIP-ViT after pre-training on pathology slide and report pairs, indicating that pathology reports act as a reasonable supervisory signal to learn representations of pathology slides. We observe that our Path-Coattn-ViT performs better than our ViT Baseline, but performs significantly worse than both our CLIP and Path ViT. We expected the supervisory signal provided by the pathology reports to create meaningful class clusterings of the image embeddings. However, our results and t-SNE plots of the co-attention image embeddings that minimal clustering occured 4

	AUROC	Accuracy	Precision	Recall
Linear Probe				
ViT Baseline	0.483	0.189	0.286	0.203
CLIP-Vit	0.792	0.434	0.657	0.665
Path-Coattn-ViT	0.515	0.200	0.329	0.574
Path-ViT	0.817	0.550	0.686	0.697
Zero Shot				
ViT Baseline	n/a	n/a	n/a	n/a
CLIP-Vit	0.498	0.189	0.194	0.057
Path-Coattn-ViT	0.557	0.228	0.355	0.078
Path-ViT	0.546	0.229	0.433	0.155

Zero-Shot Results We expected Path-Coattn-Vit to close the modality gap between queries and slide images, but this did not occur with Path-Coattn-ViT performing worst than Path-Vit across 3 different metrics. However, all self-supervised methods underperform in the zero-shot setup.

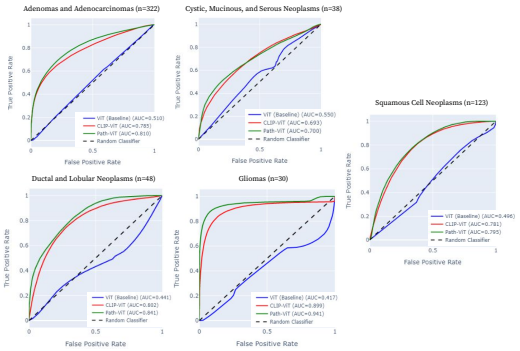

Figure 3: AUC curves for each class from linear probe evaluation. Path-ViT outperforms other self-supervised baselines across all classes.

6 Analysis

In this section, we perform error analysis on the zero-shot classification of the semi-supervised models, explore the performance of co-attention, and suggest potential avenues to address issues.

6.1 Correct and Incorrect Classifications

We will begin our analysis by examining correctly classified and incorrectly classified examples by Path-ViT on linear probing. After looking at 3 we chose these 2 classes as an example because Gliomas had the highest AUC value and Cystic, Mucinous, and Serous Neoplasms had the lowest. Our most notable observation was the low quality of the pathology reports which often included extraneous information and misspellings. 6 in the Appendix for image and report samples.

6.2 Zero-Shot Inference

Our original hypothesis was that language pre-training should enable zero-shot learning, however we find that all self-supervised models perform poorly in the zero-shot setting. Our initial set of hypotheses for why we observed poor performance was that a large proportion of the reports that the model was pre-trained on may not have contained the corresponding class label. Therefore, the model would not have learned to relate those class labels to the corresponding pathology slides. In a similar vein, since the model was not explicitly trained for this task, more careful prompt engineering may be required to make text queries closer in similarity to the report embeddings the model was pre-trained on. In this section, we conduct a series of analyses to shed light on the validity of our

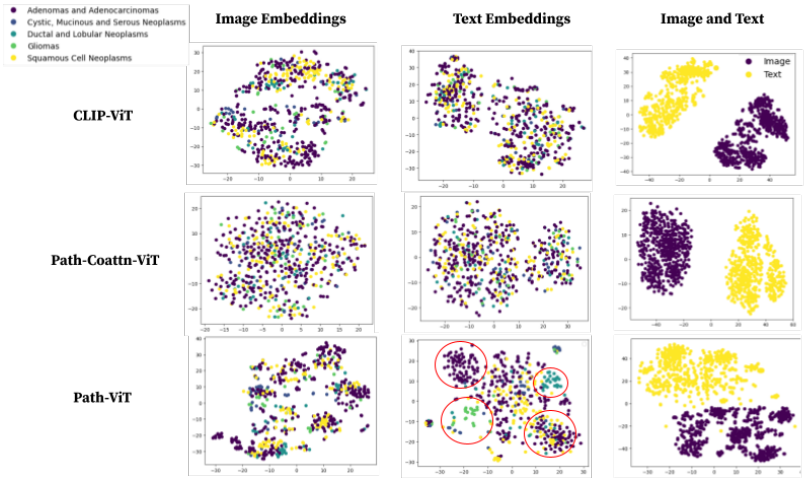

Figure 4: t-SNE plots of image and text embeddings for each self-supervised method, labeled by class. Image and text embeddings are additionally plotted in the same embedding space to highlight the effect of in-domain pre-training.

original hypotheses. We also discuss the lack in quality and non-uniformity of the reports themselves, and how this may have contributed to noisy pre-training.

First, we check whether class labels generally appear in the corresponding pathology reports. For instance, for all reports categorized as "Adenomas and Adenocarcinomas", we check how many contain the words "adenomas" or "adenocarcinomas". We find that out of all reports in the TCGA dataset, 8.440% of reports contain a subset of the class label. We include a class-specific breakdown in the Appendix A. Based on the low proportion for most classes, it suggests that using the class labels as zero-shot queries may not reflect the text embeddings that we would expect to obtain from corresponding reports. Thus, further prompt engineering strategies, such as using a description of a particular disease rather than the disease name itself, could improve model performance.

However, while prompts are a potential source of error, it is still unclear whether or not the text representations learned by the model have any semantic meaning to begin with. To gain clarity on this matter, we perform t-SNE on the text embeddings outputted by each self-supervised method to determine whether the model learns text representations with semantic meaning. In Path-ViT, the most performant model on all classes, we observe distinct learned clusters. For instance, we observe "gliomas" and "ductal and lobular neoplasms" clustered in 4. This supports the hypothesis that text representations have semantic meaning based on disease type, thereby highlighting the promise of this method for zero-shot tasks given more careful prompt engineering. We note that the image embeddings show some semblance of clustering, but not to the degree of separation of the text embeddings. Since text embeddings show more distinct clusters, this could reflect that longer training on more data could allow image embeddings to cluster in a similar manner to the text embeddings.

Additionally, as a sanity check, we observe that Path-ViT image and text embeddings are the closest amongst the methods, suggesting that self-supervised pretraining in the pathology domain is contributing to the model's ability to learn a joint embedding space. This supports the claim that. Based on these analyses, a future experiment could involve more careful prompting in the zero-shot setup, and more model pre-training on additional data for more epochs. Other ideas include performing binary classification on each class rather than multi-class classification, using a biomedically pre-trained text encoder to speed up learning, and using descriptions of the disease as a query rather than the disease type itself.

6.3 Co-Attention

In our results, we observed that the embeddings obtained from co-attention did not improve performance in both linear probe and zero-shot settings. To analyze these results further, we generate language guided attention heatmaps to determine if any language-relevant patches were displayed.

Furthermore, we conduct a case-study analysis where we visualize co-attention with different prompts on a single slide to better understand if the co-attention maps reflect any noticeable semantic visual patterns.

In 5, we present co-attention heatmaps generated from different queries on a single slide selected from the "Ductal and lobular neoplasms" class. We first examine the effect of different queries including the class label, the full report, and another class label not for the slide, to determine if semantic regions of attention varied depending on the query. We observe that regardless of the prompt, there do not seem to be any visual clusters of attention on the slide, which we would expect to see since such neoplasms are generally clustered in distinct regions (Chaudhary et al., 2013). We test the full corresponding report as a query as well to determine if the issue was a result of distribution shift between queries and the pathology reports, however the full report co-attention map also does not highlight any noticeable clusters. We perform a patch level analysis in the Appendix A.1.

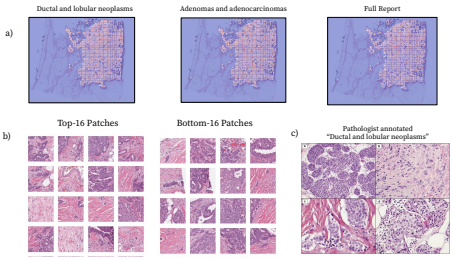

Figure 5: a) Attention maps for different queries on a patient with "Ductal and lobular neoplasms". Each title represents the text query used to generate the respective heatmap. b) Top-16 and bottom-16 patches taken from the attention map generated by "Ductal and lobular neoplasms" query. Selected to highlight any evident qualitative differences between patches with high attention vs. low attention. c) Sample images of "ductal and lobular neoplasms" expressed on a whole slide image annotated by pathologist. Used as a baseline for what we expect the model to attend to in a slide. Figure taken from (Chaudhary et al., 2013)

Based on the attention heatmaps, co-attention outputs have likely not learned semantically meaningful regions of interest. To help explain this finding, we analyze the t-SNE plots from 4. We hypothesize a potential reason that the model did not learn is that the ImageNet pre-trained image encoder was frozen and not in-domain, so initial image representations were poor. However, the text encoder was also not in-domain, and is trying to learn from image embeddings that were not representative of the original images. Since the image encoder was frozen, it could not update its weights to reflect improved signal from the text encoder, making it difficult for the model to learn.

To address these issues, we can 1) unfreeze the image encoder that was frozen due to memory constraints, and use Path-ViT instead of a ResNet-50, since Path-ViT image embeddings show clustering on the t-SNE, 2) use the text encoder that was pretrained from Path-ViT so that it is in domain, and 3) increase the amount of data the model was trained on and subselect higher quality reports to eliminate noise.

7 Conclusion

We build a contrastive pre-trained image-language model that leverages unstructured pathology reports to improve embedding quality of histopathology images. Our proposed method has the potential to enable language-guided interpretability of model outputs via co-attention as well as zero-shot classification on clinically relevant downstream tasks. We find that two modes of potential improvement include zero-shot evaluation and co-attention pre-training. One limitation was the lack of richness within our pathology reports; because class labels rarely appeared in a report, this inhibited the model from learning to relate class labels to the corresponding pathology slides. Second, computational contraints prevented the model from directly learning features from images. In future works, we envision expanding the pre-training dataset to more slide-report pairings from TCGA, or even pathology textbooks to provide a more salient substitute over pathology reports. Lastly, we hope to apply our method to clinically relevant tasks such as survival analysis to influence patient treatment decisions and personalize patient care.

References

Shweta Chaudhary, Loretta Lawrence, Geraldine McGinty, Karen Kostroff, and Tawfiqul Bhuiya. 2013. Classic lobular neoplasia on core biopsy: a clinical and radio-pathologic correlation study with follow-up excision biopsy. *Modern Pathology*, 26(6):762–771.

Richard J. Chen, Chengkuan Chen, Yicong Li, Tiffany Y. Chen, Andrew D. Trister, Rahul G. Krishnan, and Faisal Mahmood. 2022. Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16144–16155.

Richard J Chen, Ming Y Lu, Wei-Hung Weng, Tiffany Y Chen, Drew FK Williamson, Trevor Manz, Maha Shady, and Faisal Mahmood. 2021. Multimodal co-attention transformer for survival prediction in gigapixel whole slide images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4025.

Xinlei Chen and Kaiming He.

Yongju Lee, Jeong Hwan Park, Sohee Oh, Kyoungseob Shin, Jiyu Sun, Minsun Jung, Cheol Lee, Hyojin Kim, Jin-Haeng Chung, Kyung Chul Moon, et al. 2022. Derivation of prognostic contextual histopathological features from whole-slide images of tumours via graph deep learning. *Nature Biomedical Engineering*, pages 1–15.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Ekin Tiu, Ellie Talius, Pujan Patel, Curtis P Langlotz, Andrew Y Ng, and Pranav Rajpurkar. 2022. Expert-level detection of pathologies from unannotated chest x-ray images via self-supervised learning. *Nature Biomedical Engineering*, pages 1–8.

A Appendix

Class-wise percentages for class labels contained in reports. Adenomas and adenocarcinomas: 589 / 4373 = 13.466%
Cystic, mucinous, and serous neoplasms: 85 / 893 = 9.518%
Ductal and lobular neoplasms: 47 / 1204 = 3.904%
Gliomas: 0 / 1101 = 0%
Squamous cell neoplasms: 589 / 1339 = 43.988%

A.1 Co-attention Patch Analyses

For further analysis, we additionally sampled top-100 patches with the highest attention and the bottom-100 patches with the lowest attention to determine if there were any noticeable qualitative distribution differences. We observed that patches sampled from either high attention or low attention appeared to have been sampled from a similar distribution, meaning that attention maps still have not learned to differentiate patches with high and low semantic importance conditioned on the query. We also compare to pathologist annotated images of "ductal and lobular neoplasms" as a baseline for what we would expect relevant patches to look like. Since both high and low attention patches contain images that appear similar to this ground truth, we can not conclude that high attention patches reflect the actual regions of interest.

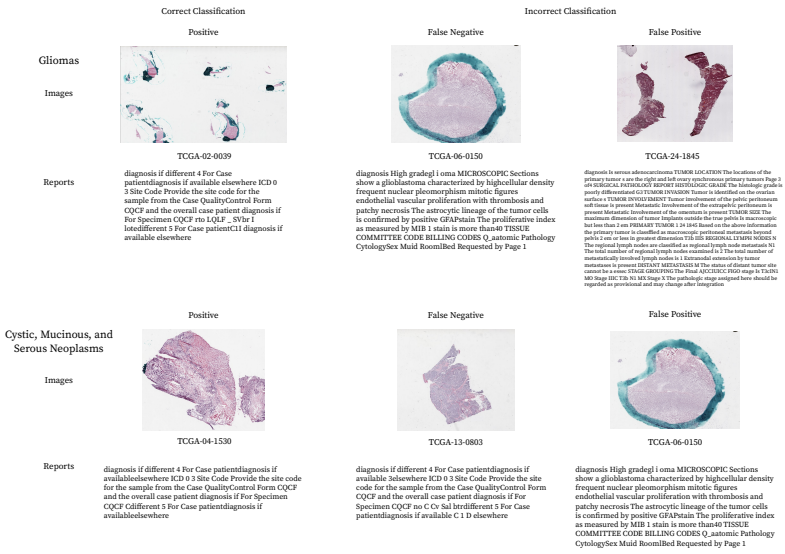

Figure 6: Examples of a positive, false negative, and false positive, for both the Gliomas class and Cystic, Mucinous, and Serous Neoplasms class