COS 529: Matching Pathology Notes with Tissue Images

Dillon Lue
Princeton University
dl4257@princeton.edu

Anirudh Ajith
Princeton University
aa8052@princeton.edu

Abstract

In the modern medical industry, highly-skilled experts such as doctors or diagnosticians are required to perform diagnoses/analyses using images of tissue samples. We attempt to automate this process by using the CLIP framework to create an image-captioning tool that can be used to retrieve useful textual pathology notes given histology images. We implement a modular CLIP framework and experiment with candidate choices of encoders. We conduct qualitative and quantitative analyses of the efficacy of our trained models. We demonstrated the ability of our model to identify the pathology of a given tissue image from unstructured data.

1. Introduction

In today's medical industry, highly-skilled experts such as doctors or diagnosticians with multiple years of training are required to perform diagnoses/analyses given images of tissue samples. Labeling tissue samples and writing histology reports at scale remains an expensive endeavor. In this project, we aim to use the CLIP framework [5] to create an image-captioning tool that can retrieve useful textual pathology notes given previously unseen histology images. This tool could help reduce the reliance on highly trained human experts and make diagnoses cheaper and more accessible to patients in need of medical attention.

Additionally, metadata for biomedical datasets are inherently messy, often containing unstructured text. The CLIP framework [5] is suitable for use in this setting since it produces a score measuring the compatibility of a `(image, caption)` pair. In this project, we use CLIP to create a scoring function that can be used to rank the best pathology caption for a tissue image from a set of candidates

2. Related Work

In the recent past, many contrastive-learning-based approaches in the supervised learning setting have shown promise in creating robust representations that are amenable to various downstream tasks. Loss functions such as triplet loss [6] have successfully been used to create robust embeddings of human faces which can be used downstream for facial recognition. Later work [8, 7, 3] focused on using other training examples within the same batch as the 'negatives' in the contrastive loss formulation to increase the strength of the training signal per batch.

The CLIP framework [5] introduced by OpenAI in 2021 offers one such formulation of a contrastive loss which can be used to train image and text encoders in conjunction such that the embeddings of the `image` and the `caption` from the legitimate `(image, caption)` pairs are maximized and the similarities between embeddings from illegitimate `(image, caption)` pairs are minimized.

CLIP has since been used for many downstream applications across many different domains. Notably, the framework has also been used in the biomedical domain to match chest X-ray images against diagnoses of pathological conditions at accuracy rates which are comparable to those of a trained radiologist. [9]

3. Dataset preparation

3.1. Original dataset

The data we used to train our CLIP implementation came from the dataset of high-quality histology images maintained by the Genotype Tissue Expression Consortium (GTEx) [1]. The dataset includes ∼25,000 pairs of histology images and associated metadata annotating these tissues. Specifically, each image is associated with 6 fields: Tissue type, Sex, Age bracket, Hardy Scale, Pathology Category and Pathology Notes. Each image, however, contains multiple (between 1 and 6) slices of the tissue sample separated from each other by whitespace.

3.2. Collection and post-processing

A large proportion of area of the images from the GTex dataset was usually occupied solely by whitespace. To increase the strength of the training signal and reduce the effect of confounding factors and spurious correlations such as the number of slices per slide, we applied the following three steps of postprocessing using the OpenCV Python

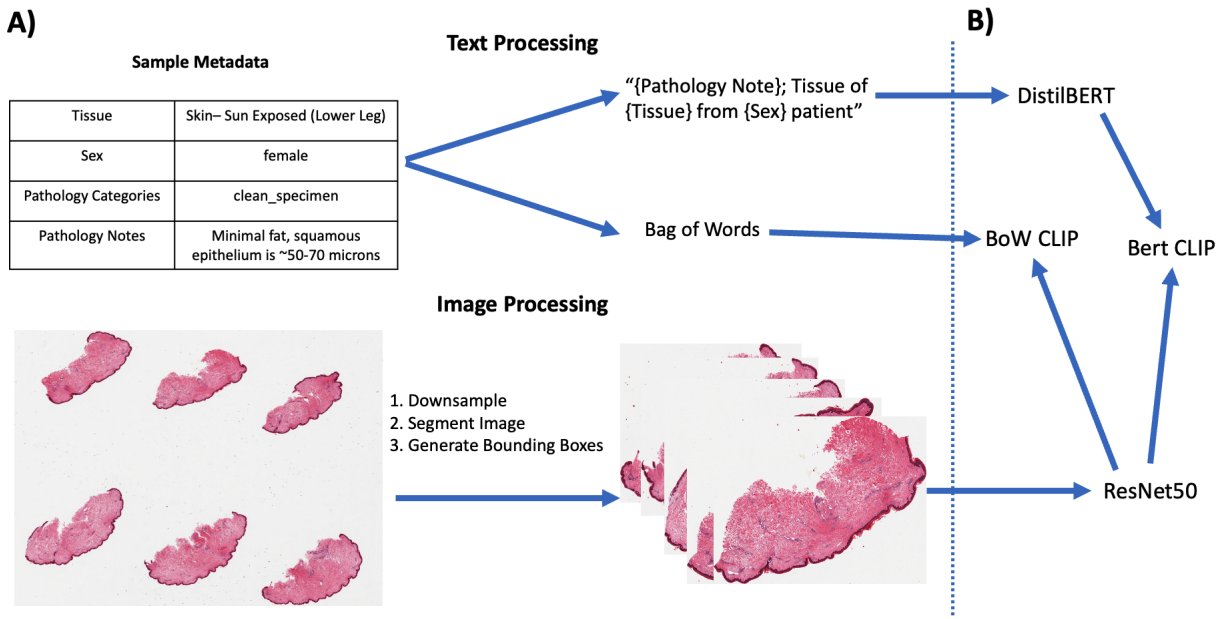

Figure 1. Data Preparation and Models

package [2]

1. Downsampling
2. Segmentation of each slice
3. Bounding box generation
4. Cropping

to split each original image into multiple tightly framed images containing the individual tissue samples. We have uploaded our post-processed images to Google Drive.

We created our final dataset by pairing up these post-processed images with captions describing the samples. The captions we created followed the template {`pathology note`}; *Tissue of* {`tissue origin`} *from* {`sex`} *patient*.

3.3. Statistics

We ended up with a dataset of ~80,000 (`image, caption`) pairs spread almost uniformly over 40 different tissue types. We split the dataset into training and validation datasets using a 80/20% split.

4. Model framework

4.1. Motivation

We used the CLIP [5] framework to learn a function that scores (`image, caption`) pairs from this dataset. Briefly, CLIP trains a text encoder and an image encoder to map paired images and text to similar low-dimensional representations. CLIP aims to maximize the similarity between valid (`image, caption`) pairs and minimize similarity between invalid ones. The core of our implementation is based on https://github.com/moein-shariatnia/OpenAI-CLIP.

4.2. Implementation details

Due to computational constraints, we chose a ResNet-50 architecture. For the text encoder, we experimented with using a Bag of Words (BoW) encoding and the DistilBERT model. For the BoW, we only used words that were found in at least instances. This resulted in a vocabulary size of 1277. For the BoW CLIP model, we term the text encodings the BoW encodings. For the DistilBERT CLIP model, we downloaded a pretrained DistilBERT model and considered the [CLS] token as the text encoding. From the text encoding we used two fully connected layers to project to a shared latent representation. We term this the text embedding. Similarly, we took the last layer of the ResNet-50 output and used two fully connected layers to project into a shared latent representation. We term this the image embedding. The embedding dimension was 256. We applied the CLIP contrastive loss to these embeddings across a batch size of 32. We trained on a NVIDIA V100 GPU using the Adam optimizer. We stopped training after 30 epochs.

Because our metadata is partially categorical (tissue type, sex, and age group) and partially unstructured (pathology notes), we also tested using a modified CLIP model that had a loss function of:

$$L = loss_{CLIP} + loss_{tissue} + 0.5 \cdot loss_{sex} + 0.5 \cdot loss_{age_group}$$

The categorical losses were calculated by doing a cross

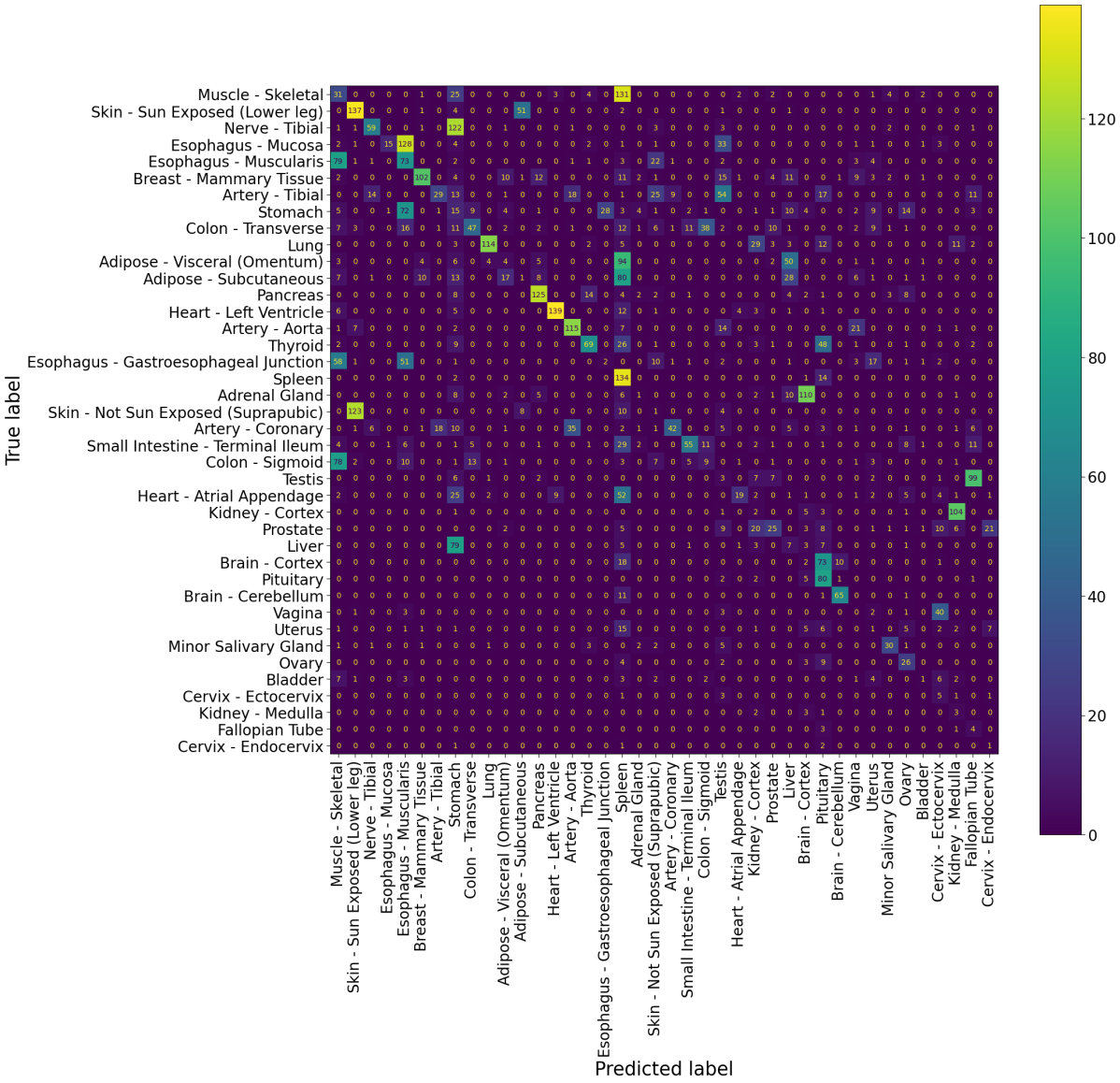

Figure 2. DistilBERT CLIP Model Tissue Prediction Confusion Matrix

entropy loss against the target labels and a linear layer straight from the embedded dimension. We created this combined loss function to try to take advantage of the structure of the data.

As a baseline for some of our experiments, we use a ResNet-50 to directly predict the categorical variables.

5. Results

5.1. Quantitative measures

5.1.1 Direct Categorical Prediction

To have a baseline for what features might be important or not, we first trained ResNet-50 to predict the categorical variables in our dataset. This included trying to predict `(Tissue, sex, age)` fields. Here, we simply used a cross entropy loss. We achieved an accuracy of 87% accuracy on Tissue prediction and 35% on Age prediction. For sex predictions, we found the model would always predict male even though there was a 60/40% split of male to female.

5.1.2 CLIP Model Categorical Prediction

OpenAI CLIP model was shown to have comparable accuracy to direct classifiers in the zero-shot setting [5]. To see if we could do the same on our dataset, we calculated the Tissue prediction accuracy. For the BoW model, we

encoded each of the 40 tissue into a BoW representation. For each image, we calculated the image embedding and matched the image embedding with the tissue embedding that had highest cosine similarity. We performed the same test with the DistilBERT model but using the caption "Tissue of {Tissue}". The BoW CLIP model had 12% accuracy while the DistilBERT CLIP model had 32% on tissue prediction. Figure 4.1 shows the tissue accuracy for DistilBERT CLIP model; there were some cases where tissues were misclassified into similar tissue types for example 104 predictions that were predicted to be the "Kidney - Medulla" but were actually "Kidney - Cortex". We applied the same strategy for sex prediction and found 40% accuracy for both models which is worse than a random classifier. Considering the fact that none of our predictors could guess the sex, we performed no further sex analysis. Overall, we found that the DistilBERT CLIP model more accurately predicted Tissue than the BoW CLIP; however, this is much lower than our baseline.

Next, we tested our model with the combined contrastive and categorical loss and found that it had 84% accuracy on Tissue prediction and 35% accuracy on age prediction. This was unsurprising given that the model is nearly like a direct prediction.

5.1.3 Cosine Similarity Between Paired Image and Text Compared to Random

The CLIP framework aims to minimize the cosine similarity between pairs of images and text. Next, we tested if paired image and captions had higher cosine similarity than randomly chosen image and captions. To perform this test, first we zero centered the image and text embeddings. Next, we calculated the cosine similarity between all pairs of images and captions. For BoW CLIP model, the average similarity between unpaired images and captions was 0 and the average similarity between paired images and captions was 0.4839. For BoW CLIP model, the average similarity between unpaired images and captions was 0 and the average similarity between paired images and captions was 0.3878. This suggests that indeed DistilBERT CLIP model was able to align image and text embeddings better than the BoW model.

Next, we performed the same test on the combined contrastive and categorical loss model and found that the average similarity between paired images and captions was 0.4105.

In summary, the combined contrastive and categorical loss model had higher tissue prediction accuracy than the pure contrastive models; however, the contrastive and categorical loss model had lower text and image latent similarity.

In this paper, we are primarily interested in trying to learn from unstructured data to generate meaningful pathology notes given an image; thus, we ended up choosing our DistilBERT CLIP model for further analysis.

5.2. Qualitative measures

5.2.1 Clustering of Tissues

To see if CLIP's latent space is learning any useful pairing of image and text data, we performed t-SNE in the latent space. While CLIP does not specifically optimize for the Euclidean distance between image and text embeddings, we would expect to see the image and text embeddings to be near each other in the t-SNE plot. Overall we did observe this pattern. For example on the bottom left, we can see that the lung image and text embeddings are close to each other. Furthermore, we can see that within a single tissue and modality combination, the points cluster closely together.

If the single tissue and modality combination clusters were so close to each other, this raises the question of why the CLIP categorical prediction was so much lower than the direct predictor. Perhaps the current CLIP model works better at aligning images to their full caption. To mitigate this problem, we could have occasionally used only the tissue context (and skipped the pathology notes) as a form of data augmentation.

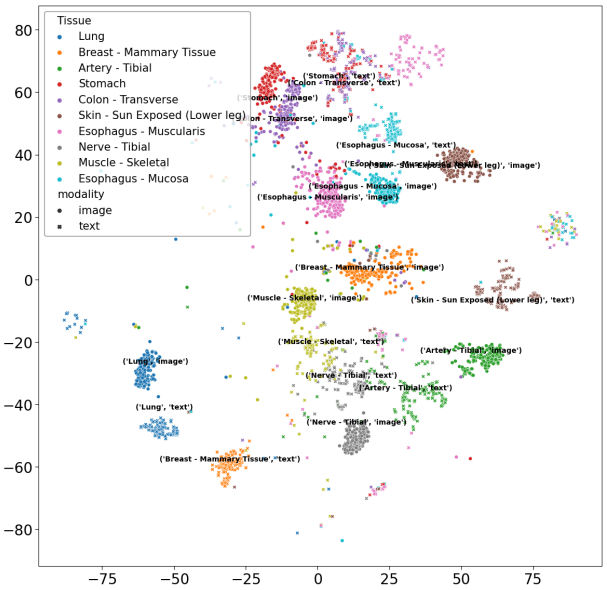

Figure 3. t-SNE of Image and Text Embedding Latent Space for Validation Images (only top 10 most frequent tissues are shown)

5.2.2 Matching validation images to captions

While the previous measures are nice to ensure that CLIP is in fact pairing images and text together, we aimed to show

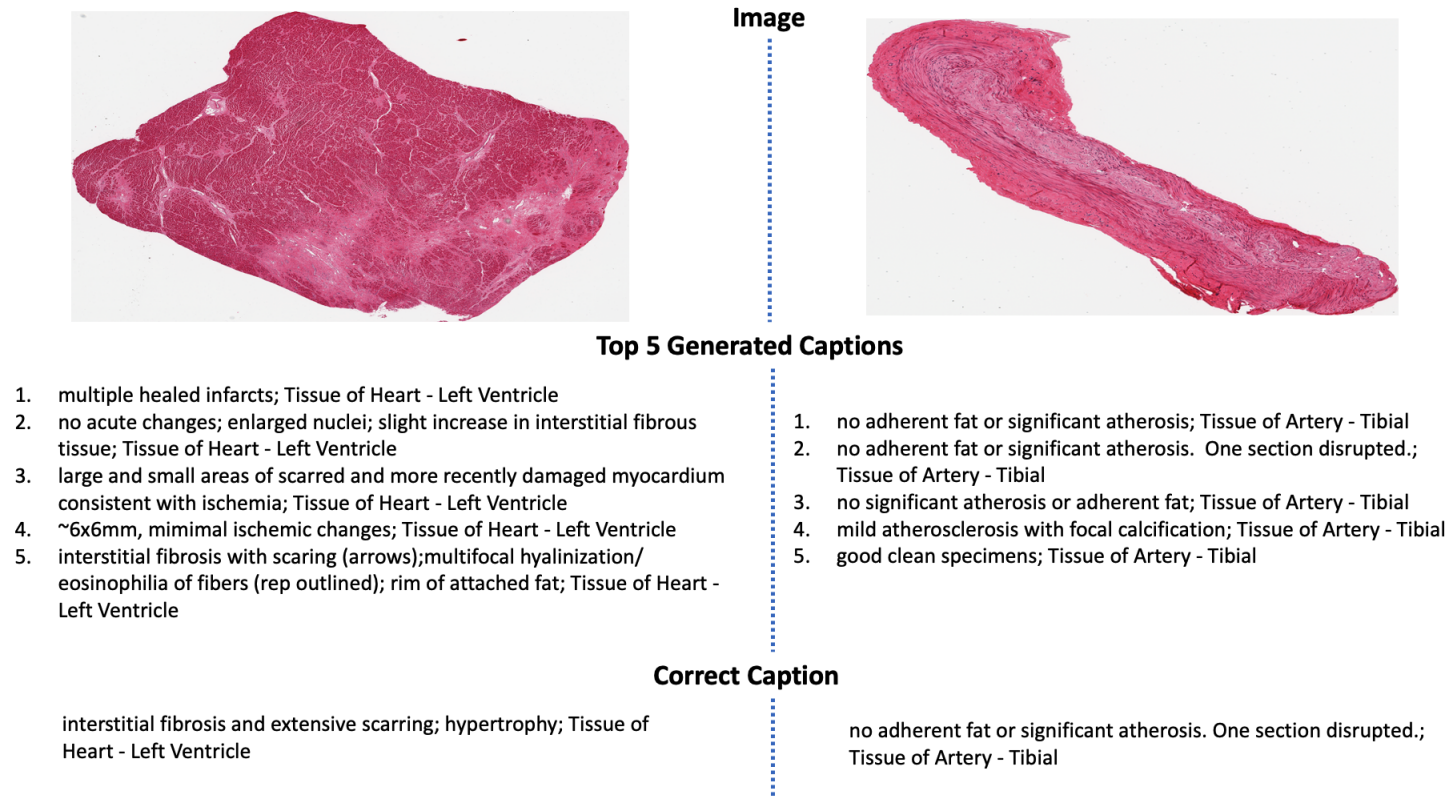

Figure 4. DistilBERT CLIP Generating Caption from Images

some examples of how such a model could be used in practice. First, we might wish to generate a caption given an image. Given an image of a tissue, what kind of pathology did the person likely have? For this test, we take an image from our validation set and identify the 10 text embeddings with the highest cosine similarity to the query image embedding. The results for 2 examples are shown in Figure 5.2.1. On the right, the DistilBERT CLIP model was able to correctly identify that the left ventricle heart image was indeed from the Left Ventricle and had some examples of scarring. On the left, the DistilBERT CLIP model was able to correctly identify an image of a healthy tibial artery. This suggests that the our model may have the ability to diagnosis/post analyze histology images in an automated manner.

5.2.3 Matching validation captions to images

The CLIP model can be also used to find tissue images that match a caption. Such a model might be useful when someone wants image examples of a particular pathology. This is like a google search for images with a particular disease. The results for 2 such examples are shown in Figure 5.2.1. Problematically, we are not tissue experts; however, we can evaluate our model based on how well the captions for the nearest image fit the text prompt. In fact we see that the model finds images from the correct tissue and with similar descriptions to the query.

6. Discussion

In this work, we preprocessed a large collection of histology images and used the CLIP framework to create an image-captioning tool that can retrieve useful textual pathology notes given a histology images. We experimented with different text encoders to improve the performance of our model and found a pretrained DistilBERT outperformed our other models in the joint embedding space. Our approach used the CLIP framework which acted as a scoring function that could be used to rank the best pathology caption for a tissue image from the 20,000 captions in our training data. This allowed us to effectively retrieve relevant pathology notes for a given tissue image, even when the available options are not necessarily perfect matches.

An advantage to our CLIP model is that it can deal with messy unstructured metadata text which traditional supervised learning techniques cannot handle.

Our image-captioning tool demonstrates a first step for reducing the reliance on highly trained human experts and make diagnoses cheaper and more accessible to patients in need of medical attention. In the research setting, it could

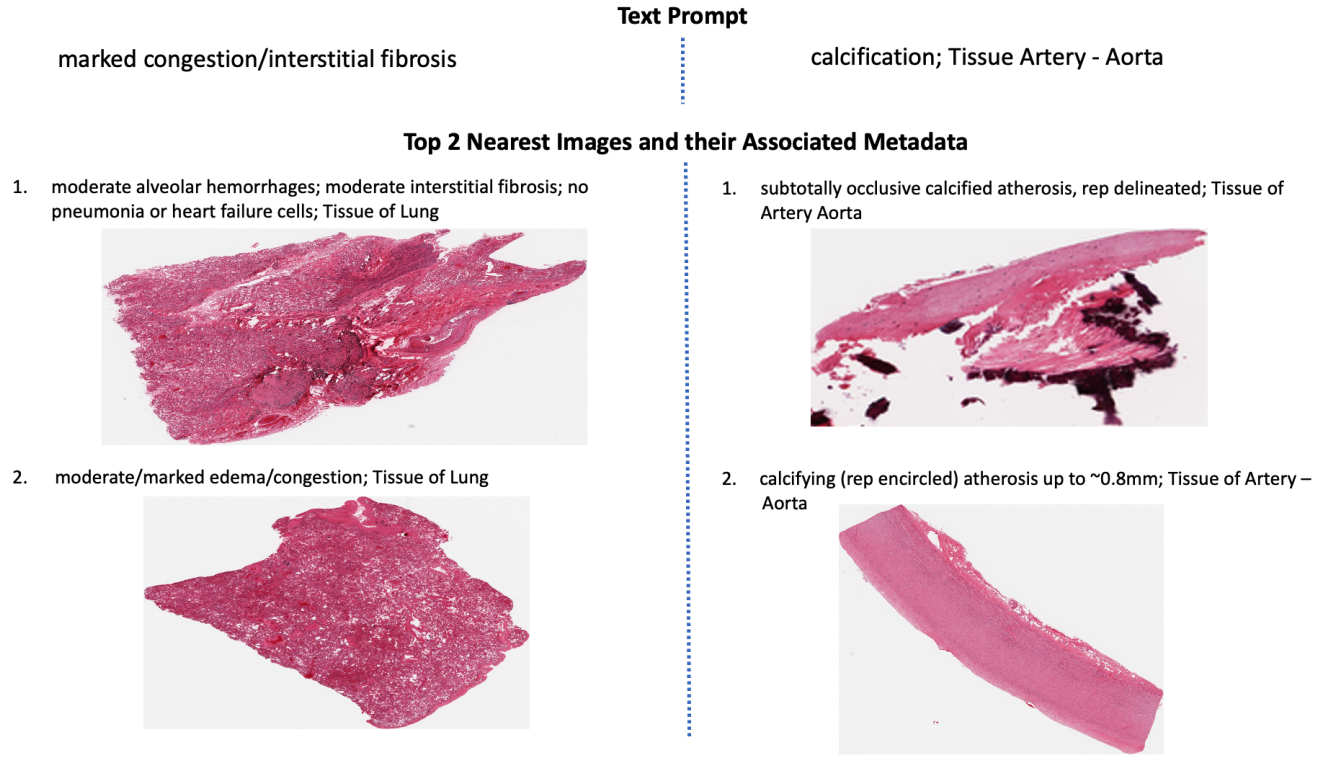

Figure 5. DistilBERT CLIP Finding Relevant Images for a Caption (note that the associated metadata are never used to find the nearest images)

be used to label tissue samples and write histology reports or find outliers and mislabeled images.

6.1. Limitations and Further Work

One limitation of our approach is that it still requires a set of candidate pathology notes to choose from, which may not always be available in practice. Most importantly, we found our model hard to validate since we are not tissue experts. To truly validate the model, we would have liked to talk to experts to see if the generated images or captions actually make biological sense.

On the data side there were various improvements that could have been made. The original images are very high resolution at 40,000 x 40,000 (0.25 GB). To make them workable we scaled them down 150 fold to 224 x 224. Perhaps the high resolution keeps some of the fine grained information about cells that would make disease prediction more reliable. Additionally, we found that some images were necessarily longer or skinnier than other tissues. Our current pipeline scales all the images to squares unnecessarily elongating some axis of the image. This may have contributed to learning particular features unrelated to tissue type. Furthermore, we would have liked to do data augmentation with images.

There are various model improvements we would have liked to make. Due to CLIP being a large model, we did not perform very much hyperparameter tuning. Furthermore, we stopped training after just 30 epochs when the validation loss was still likely decreasing. Additionally, we originally proposed testing vision transformers which is what the original CLIP paper [5] found to be most effective. Additionally, we would also consider using a language model that was trained on biomedical text [4].

7. Conclusion

In this paper, we prepossessed a large tissue database consortium to make the images more amenable to machine learning pipelines. We trained mutiple CLIP like model that embedded histology images and captions in a shared representation space. Lastly, we found that our tissue CLIP model could plausibly make non-obvious diagnosis/analysis about tissue images using a nearest neighbors approach in the representation space.

References

[1] François Aguet and Shankara Anand et. al. The gtex consortium atlas of genetic regulatory effects across human tissues. *Science*, 369(6509):1318–1330, 2020. 1

[2] G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000. 2

[3] Nicholas Frosst, Nicolas Papernot, and Geoffrey Hinton. Analyzing and improving representations with the soft nearest neighbor loss, 2019. 1

[4] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, sep 2019. 6

[5] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *CoRR*, abs/2103.00020, 2021. 1, 2, 3, 6

[6] Florian Schroff, Dmitry Kalenichenko, and James Philbin. FaceNet: A unified embedding for face recognition and clustering. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jun 2015. 1

[7] Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. 1

[8] Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese. Deep metric learning via lifted structured feature embedding. *CoRR*, abs/1511.06452, 2015. 1

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