Interpretable Lung X-Ray Classification using CRATE: White-Box Transformers

Saaketh Medepalli*, Hemit Shah*, Sai Koushik Guntakanti*

September 17, 2023

1 Motivation

The use of deep learning methods for medical image analysis is becoming increasingly popular [YZL+21], and transformer-based encoders have enabled state-of-the-art performance on a variety of downstream tasks. One such downstream task is the classification of radiological images (X-Rays, CT Scans, etc.) into disease diagnosis categories. On these types of tasks, recent work in medical image analysis achieves near perfect results with end-to-end models that receive images and predict the correct diagnosis or disease. Unlike radiologists and doctors, however, most deep learning methods are unable to provide any insights into the lines of reasoning used to reach a specific conclusion regarding an image [SWS17].

If we seek to empower medical professionals with machine learning methods, this should not be the case. Not only would having greater interpretability be incredibly useful, but it would also improve trust in deep learning methods if doctors could see a representation of which areas in medical images led a model to reach its conclusion regarding a diagnosis.

In this report, our goal is to present a proof-of-concept of how traditionally black-box models may be deployed for safer usage in a high-risk application such as medical image analysis. In our approach, we apply a novel white-box model capable of *implicitly* producing segmentation masks *only* by training on a supervised learning objective. By applying pre-trained models of this variety to medical datasets with some key modifications, we show how the decisions of such models can be better interpreted.

- White-box transformers are a novel variety of transformers which perform very similarly to regular ViTs (Dostovitsky...) while being fully mathematically interpretable due to recovering a sparse set of low-dimensional vectors

2 Methods

In this project, we attempt to apply a recent a line of work seeking to implement more interpretable transformer architectures while not compromising on accuracy. A group led by Yu et al. demonstrates a 'White-Box' Transformer known as a Coding RAte reduction TransformEr, or CRATE for short [YBP+23]. This model defines more interpretable objective functions in order to understand what representations the model is learning to perform the task. In their approach, they do this by transforming the input into *piecewise linear and sparse, compact* representations, which can then be used to classify the output. Mathematically, this takes the form of what they term the sparse rate reduction objective:

In the objective above, $\mathbf{Z}$ represents the features in the current layer and $\mathbf{U}_{[K]}$ represents the low-dimensional subspace which is used to find the attention maps. To solve this optimization problem, a 2-stage alternating approach is taken between compression and sparsification, for which

$$\max_{f \in \mathcal{F}} \mathbb{E}_{\boldsymbol{Z}} \left[R(\boldsymbol{Z}) - \lambda \|\boldsymbol{Z}\|_0 - R^c(\boldsymbol{Z}; \boldsymbol{U}_{[K]}) \right]$$

more details can be found in [YBP+23].

A surprising consequence in addition to the principled approach to designing the network is the ability for the model to produce semantically-interpretable segmentation masks (via its self-attention layers) despite only optimizing for classification accuracy [YCT+23]. The results in the paper are tested only on the ImageNet and CIFAR datasets, which we hope to extend to real-world and high-impact data. In particular, we would like to see whether medical images can be classified with high accuracy while also producing interpretable representations of image encodings. These encodings, as they are based on attention maps over the input images, can inform doctors of the "reasoning" behind a model's conclusion as the map would highlight key areas that have a large influence on the final classification produced.

2.1 Dataset

To see whether White-Box transformers are a reasonable approach to solve the problems described in the Motivation section, we chose to apply the CRATE model to the Kaggle Brain Tumor MRI Dataset as well as the Chest X-Ray 14 Dataset [mas, WPL+17] for a classification task.

The former dataset contains 7023 MRI images from the human brain in .JPG format, gathered across four different classes. See Figure 1 and Figure 2 for more details.

The latter dataset contains 112,120 frontal-view chest X-ray PNG images in 1024×1024 resolution. Due to class imbalances we decided to filter all images with multiple labels, and several classes of images that were over or under represented. 5 classes of diagnoses were left after preprocessing. See Figure 3 and Figure 4 for more details.

The reason we selected the latter dataset is that it not only included labels for diagnoses verified by doctors, but also included manually drawn bounding boxes for 158 of the images with at least around 20 for each of the 5 included classes of diagnoses. This would allow us to compare the attention maps generated from the test images and the bounding boxes drawn by domain experts to see if the model is paying attention to the correct regions to make its decision based on existing medical knowledge.

3 Results

While keeping in mind the accuracy of the model on any given classification task is important, the goal of our approach is to offer medical professionals greater transparency and insight into the decisions made by machine learning models.

If the model bases its decisions on the same features in the radiological images as those highlighted by the bounding boxes drawn by doctors, then we can assure ourselves that the model is in line with existing knowledge. On the other hand, if areas with high attention in the model are not at all overlapping or close to such bounding boxes, then offering these maps as information to medical professionals using the model is doubly important. They can choose to ignore the findings of the model if this is the case, or investigate whether the model is finding patterns which current medical knowledge is unaware of (given that the classification accuracy of the model is high for the latter case).

To start, we began by running the MRI images through the pre-trained ImageNet-21k model to observe the areas to which the model was attending to from its learned representations. We then fine tuned their pretrained model on our specific classification task, before doing the same to observe differences in the areas attended to by both the pre-trained and fine-tuned models.

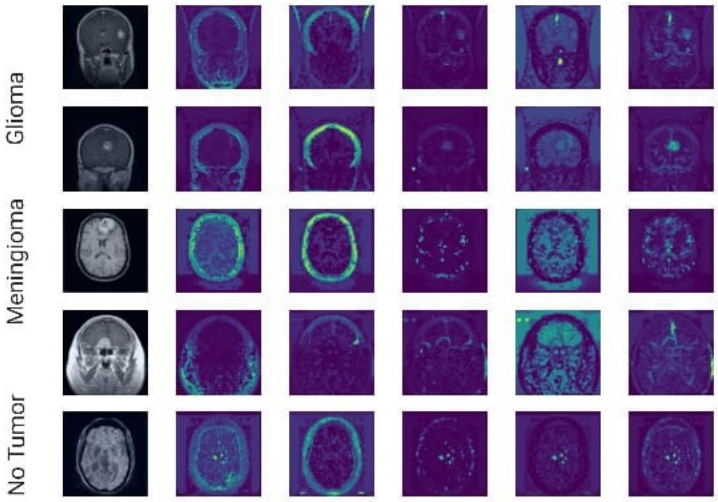

Figure 1: Last-layer Attention Masks from pretrained ImageNet-21k model on Brain Tumor MRI dataset. 5 images were chosen from the test dataset, including 2 Glioma, 2 Meningioma and 1 No Tumor image. Out of 6 attention heads, five were chosen for clarity. Dark purple to brighter green indicates increasing activation.

It is readily apparent above that many of the extracted self-attention maps indicate that the heads of the transformer are not looking at brain tissue at all to perform the classification task. This is a sign that the model may not be using the correct information to make its prediction.

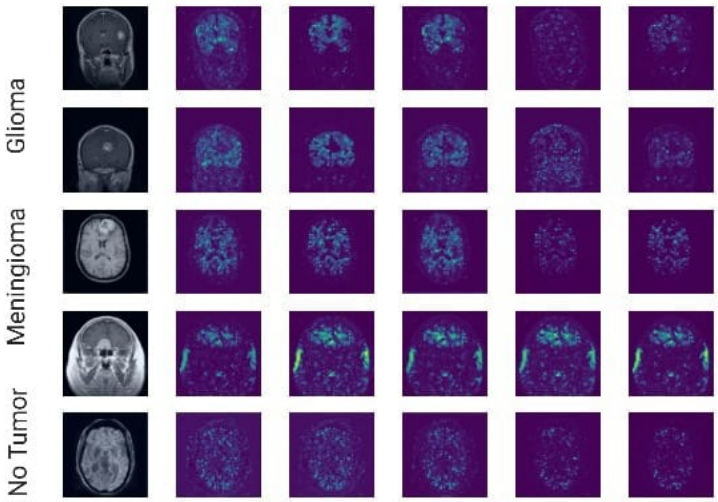

Figure 2: Last-layer attention masks from fine-tuned ImageNet-21k model on Brain Tumor MRI dataset. The model achieved 80% accuracy on the test dataset. The same images and columns were chosen for purposes of comparison.

Compared to the pre-trained model, our fine-tune model not only performs better on the brain MRI dataset in terms of accuracy, but the self-attention maps also indicate that the transformer is generally focusing much more on the brain tissue in the MRI images. Not only does this show

that the model is more transparent, but would also improve trust as medical professional's could be assured that the model is basing its decisions on truly important features in the image.

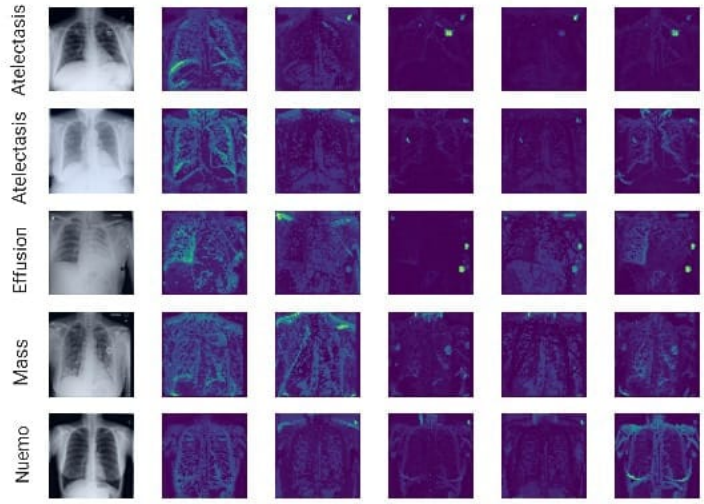

Figure 3: Last-layer attention masks from pre-trained ImageNet-21k model on chest X-ray images. 5 images were again chosen from the test dataset, each of which comes from a different class describing some condition afflicting the lungs.

In the above case, the pre-trained model produces self-attention maps that are not very consistent with areas that medical professionals would focus on when analyzing chest X-rays for specific conditions. In fact, lung tissue seem to be the least important feature for many of the transformer heads.

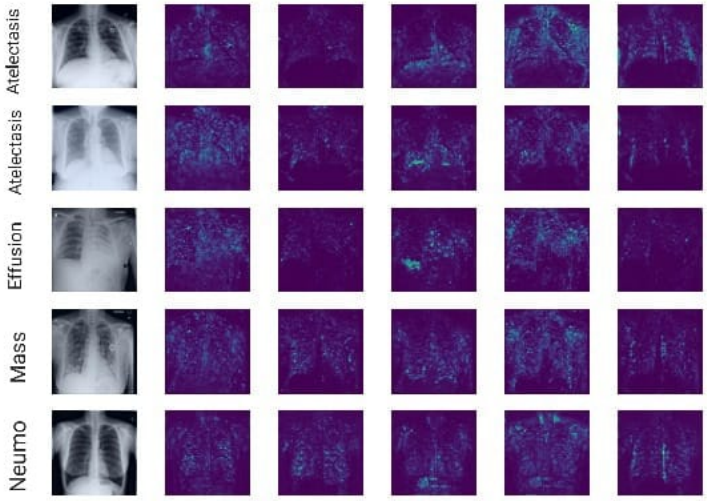

Figure 4: Similar to 3, last-layer attention masks were captured on the fine-tuned ImageNet-21k model. The model achieved 35% accuracy on the test dataset. The same images and attention heads were chosen as before for comparison purposes.

Although our fine-tuned model did not perform as well on the chest X-ray dataset in terms of

classification accuracy, we believe that given more training time it would be possible to achieve higher performance. It is imnportant to note that compared to the pre-trained model, the transformer heads in this case seem to be giving greater importance to the lung tissue itself when considering the self-attention maps.

4 Future Work

Future works in computer-aided medical diagnoses can take advantage of similarly interpretable transformer architectures to produce both accurate down-stream results while also providing medical professionals insights into the model's line of "reasoning".

If white-box Transformers can be applied to medical image datasets in other domains of diagnoses (such as scans of the fundus in the eye), we can further our understanding of which features are deemed important by models relying on transformers. Most suitable for this task would be datasets with both classification labels, and either manually drawn bounding boxes or segmentation maps for features impacting diagnoses made by domain experts. This will allow the classification accuracy of models to remain a key measure of performance, while also ensuring that transformer architectures aren't using knowledge extraneous to the current domain of medical knowledge to make such decisions on images.

We hope this work also motivates other groups to pursue work in a similar vein to other high-impact industries, including autonomous vehicles, hiring for jobs, justice system, etc.

References

[mas] Brain tumor mri dataset.

[SWS17] Dinggang Shen, Guorong Wu, and Heung-Il Suk. Deep learning in medical image analysis. *Annual Review of Biomedical Engineering*, 19(1):221–248, 2017. PMID: 28301734.

[WPL+17] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2097–2106, 2017.

[YBP+23] Yaodong Yu, Sam Buchanan, Druv Pai, Tianzhe Chu, Ziyang Wu, Shengbang Tong, Benjamin D Haeffele, and Yi Ma. White-box transformers via sparse rate reduction. *arXiv preprint arXiv:2306.01129*, 2023.

[YCT+23] Yaodong Yu, Tianzhe Chu, Shengbang Tong, Ziyang Wu, Druv Pai, Sam Buchanan, and Yi Ma. Emergence of segmentation with minimalistic white-box transformers. *arXiv preprint arXiv:2308.16271*, 2023.

[YZL+21] Sijie Yang, Fei Zhu, Xinghong Ling, Quan Liu, and Peiyao Zhao. Intelligent health care: Applications of deep learning in computational medicine. *Frontiers in Genetics*, 12, 2021.