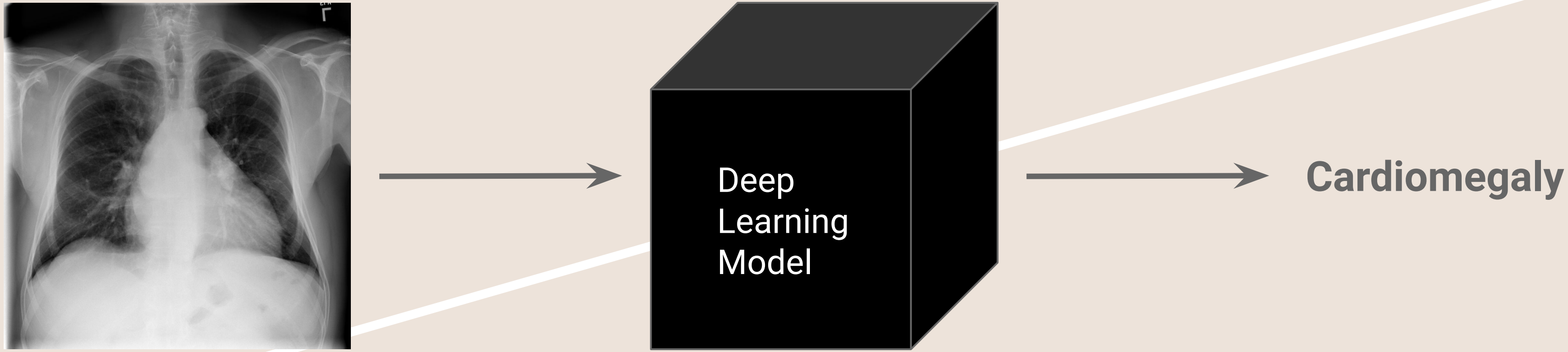
Interpretable Medical Image Classification using CRATE: White-Box Transformers

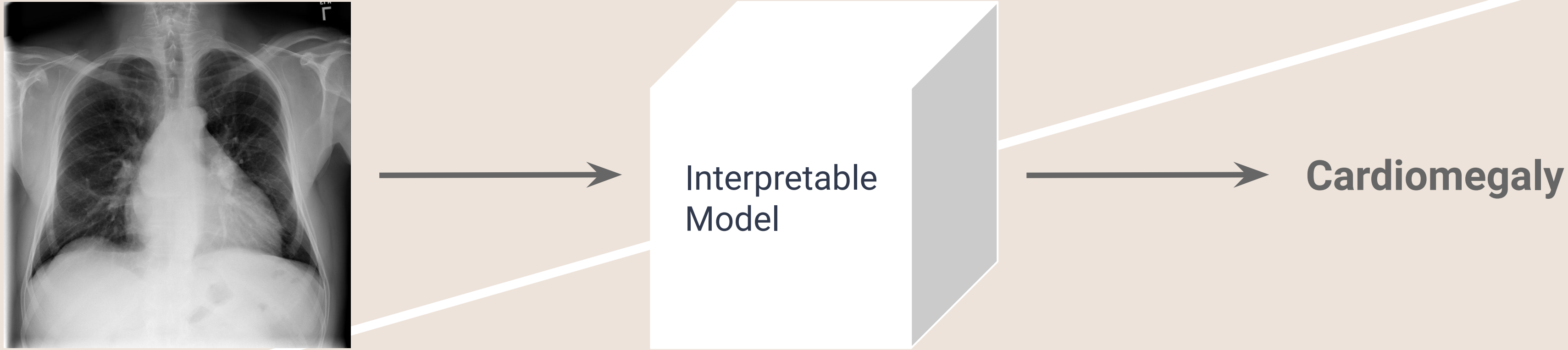
Challenge Topic 1: AI Safety

Saaketh Medepalli, Sai Koushik Guntakanti, Hemit Shah

Medical Image Classification (Currently)



Medical Image Classification (Goal)



White-Box Transformers (Yu et al., 2023)

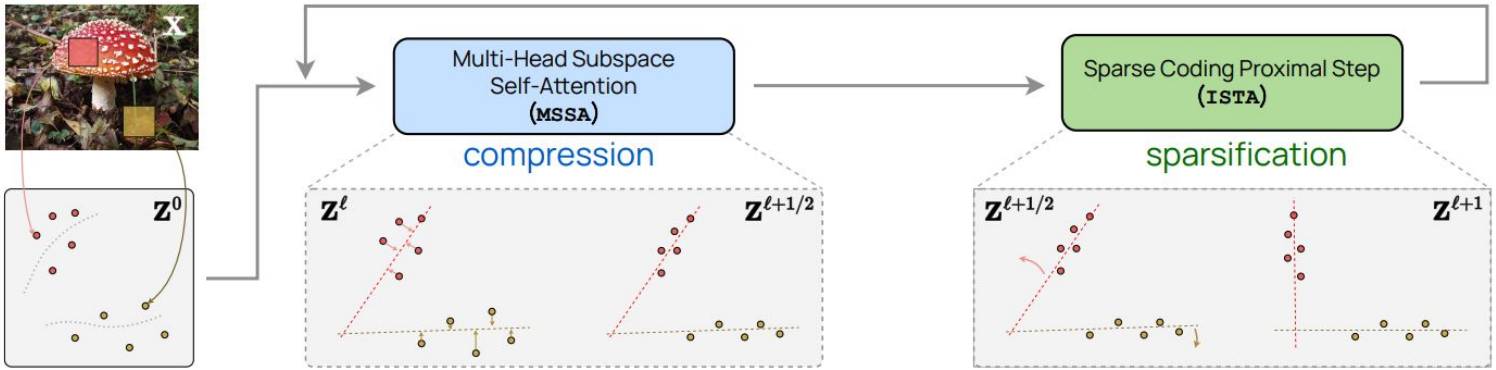


Figure 1: The 'main loop' of the CRATE white-box deep network design. After encoding input data $\boldsymbol{X}$ as a sequence of tokens $\boldsymbol{Z}^0$, CRATE constructs a deep network that transforms the data to a canonical configuration of low-dimensional subspaces by successive ***compression*** against a local model for the distribution, generating $\boldsymbol{Z}^{\ell+1/2}$, and ***sparsification*** against a global dictionary, generating $\boldsymbol{Z}^{\ell+1}$. Repeatedly stacking these blocks and training the model parameters via backpropagation yields a powerful and interpretable representation of the data.

CRATE:

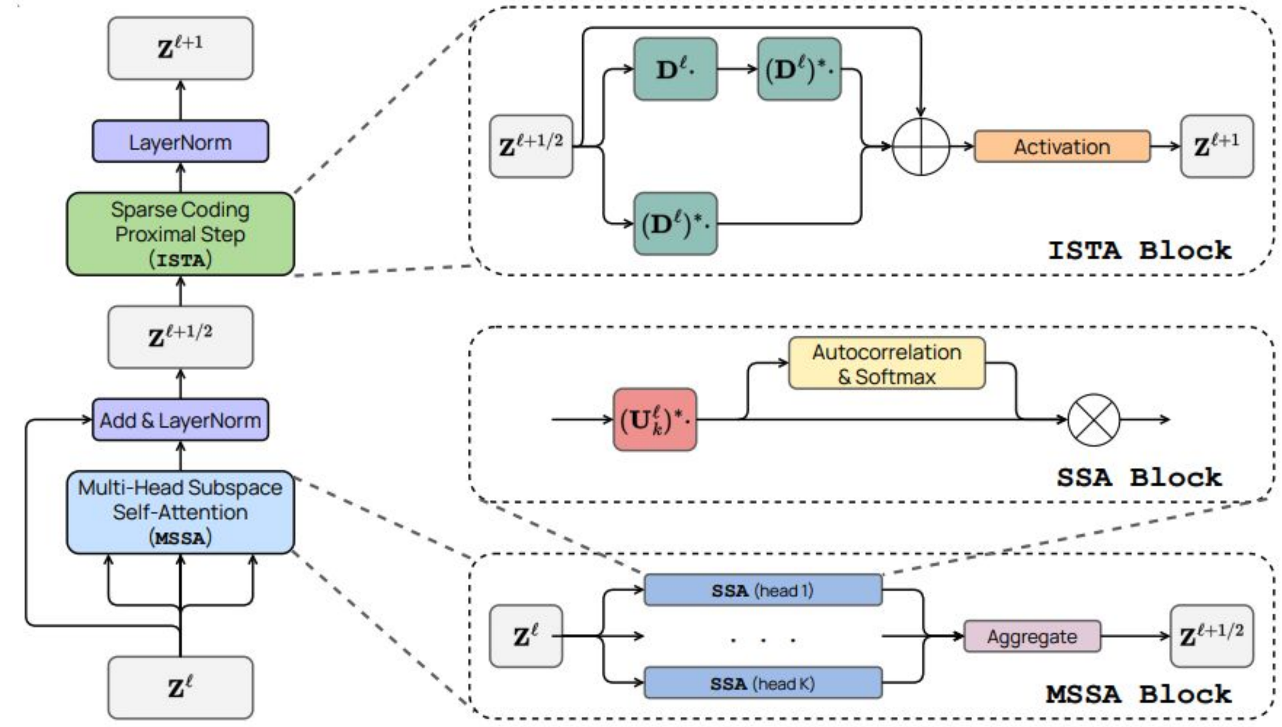


Figure 2: One layer of the CRATE architecture. The full architecture is simply a concatenation of such layers, with some initial tokenizer and final task-specific architecture (i.e., a classification head).

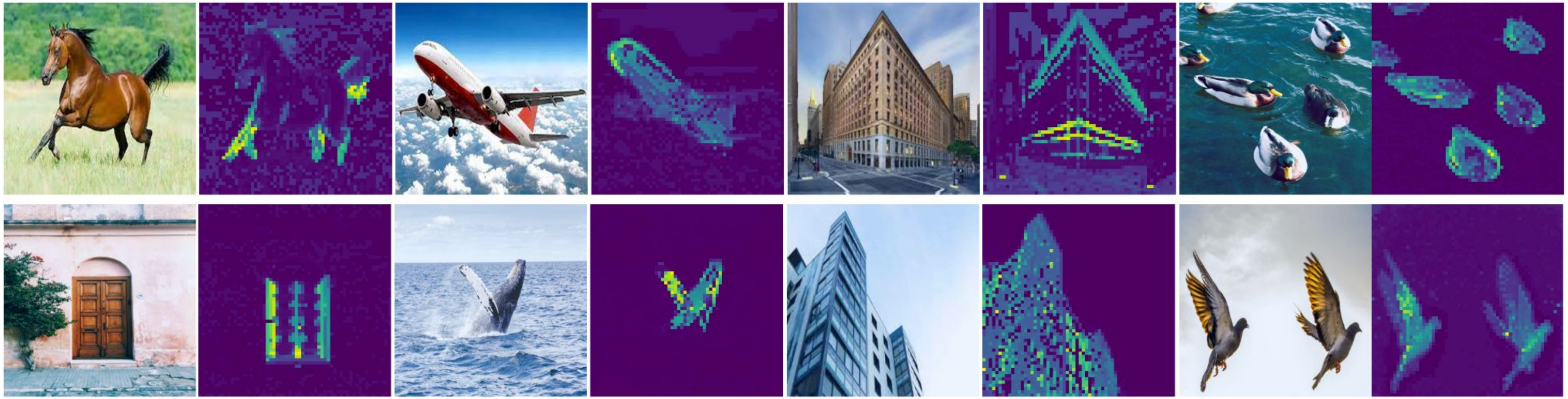


Figure 4: Self-attention maps from a supervised CRATE with 8×8 **patches** trained using classification. The CRATE architecture automatically learns to perform object segmentation without a complex self-supervised training recipe or any fine-tuning with segmentation-related annotations. For each image pair, we visualize the original image on the left and the self-attention map of the image on the right.

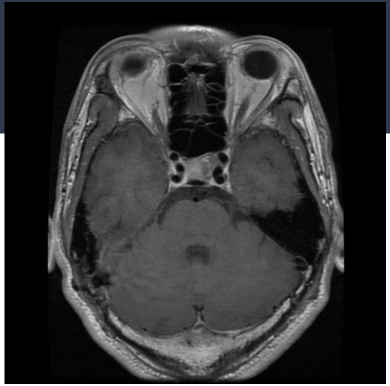
Emergence of Segmentation with Minimalistic White-box Transformers (Yu et al., 2023)

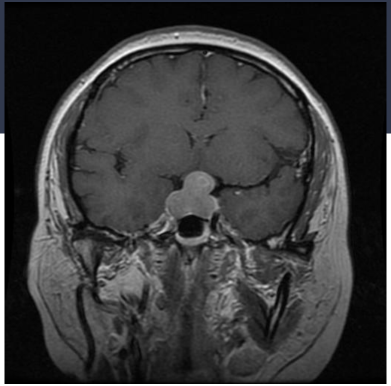
Key takeaways

- The self-attention maps extracted from the heads of the last layer show automatic learning of object segmentation *without explicit training!*
- The architecture learns which features of the image are most important for the final class prediction in an *interpretable* manner

Can we classify medical images accurately while also providing interpretable explanations for model decisions to medical professionals?

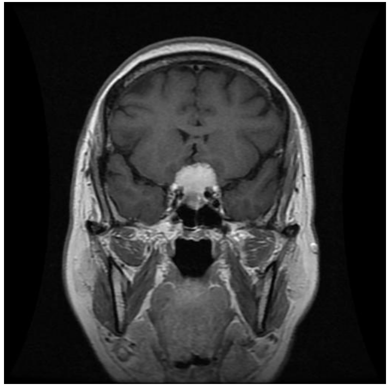
Test Dataset 1

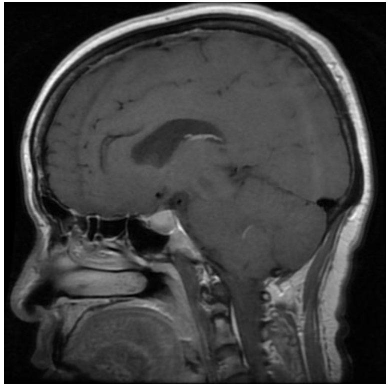




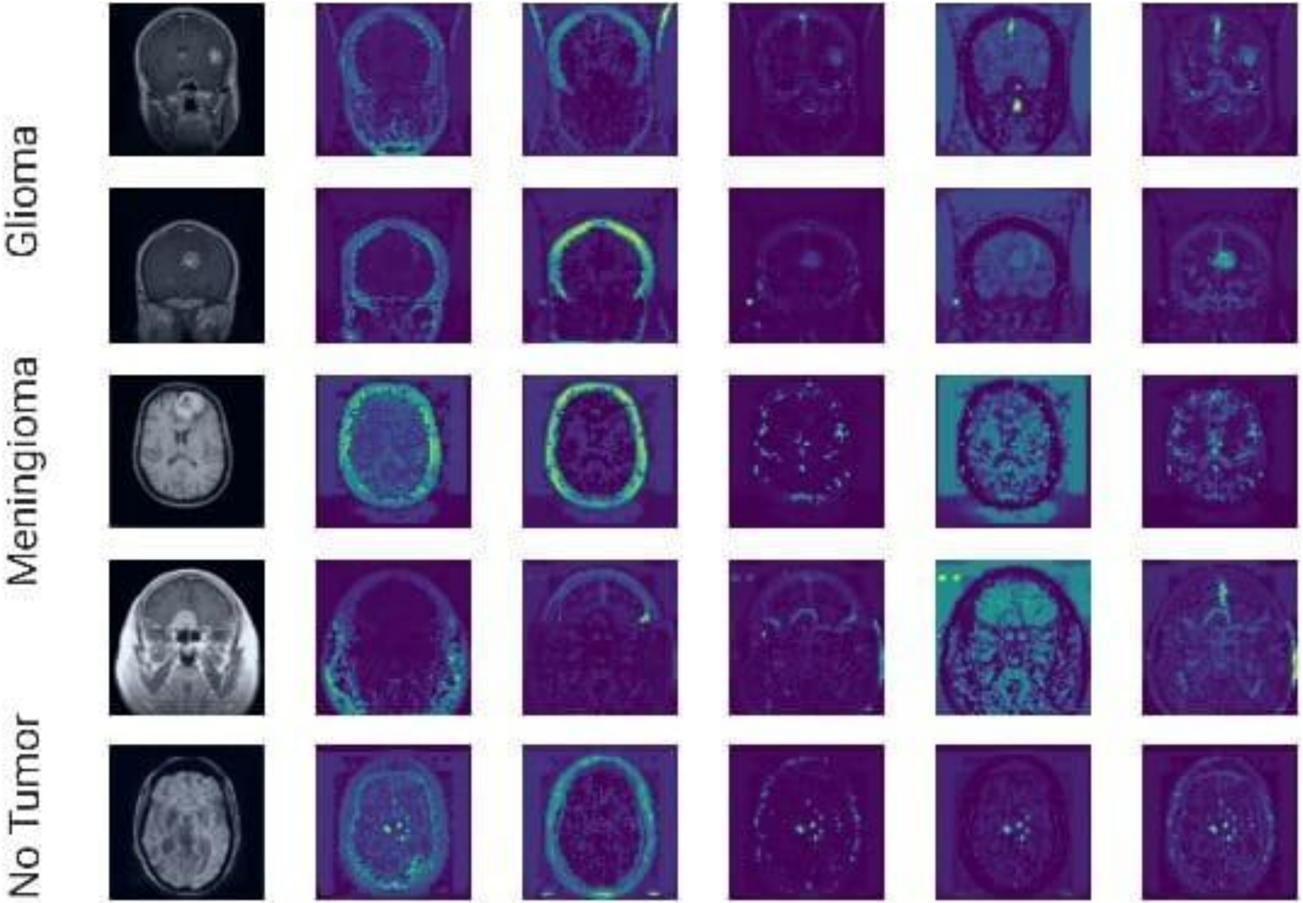
Details:

- Brain tumor MRI dataset
- 7032 images of human brain MRI images classified into 4 classes

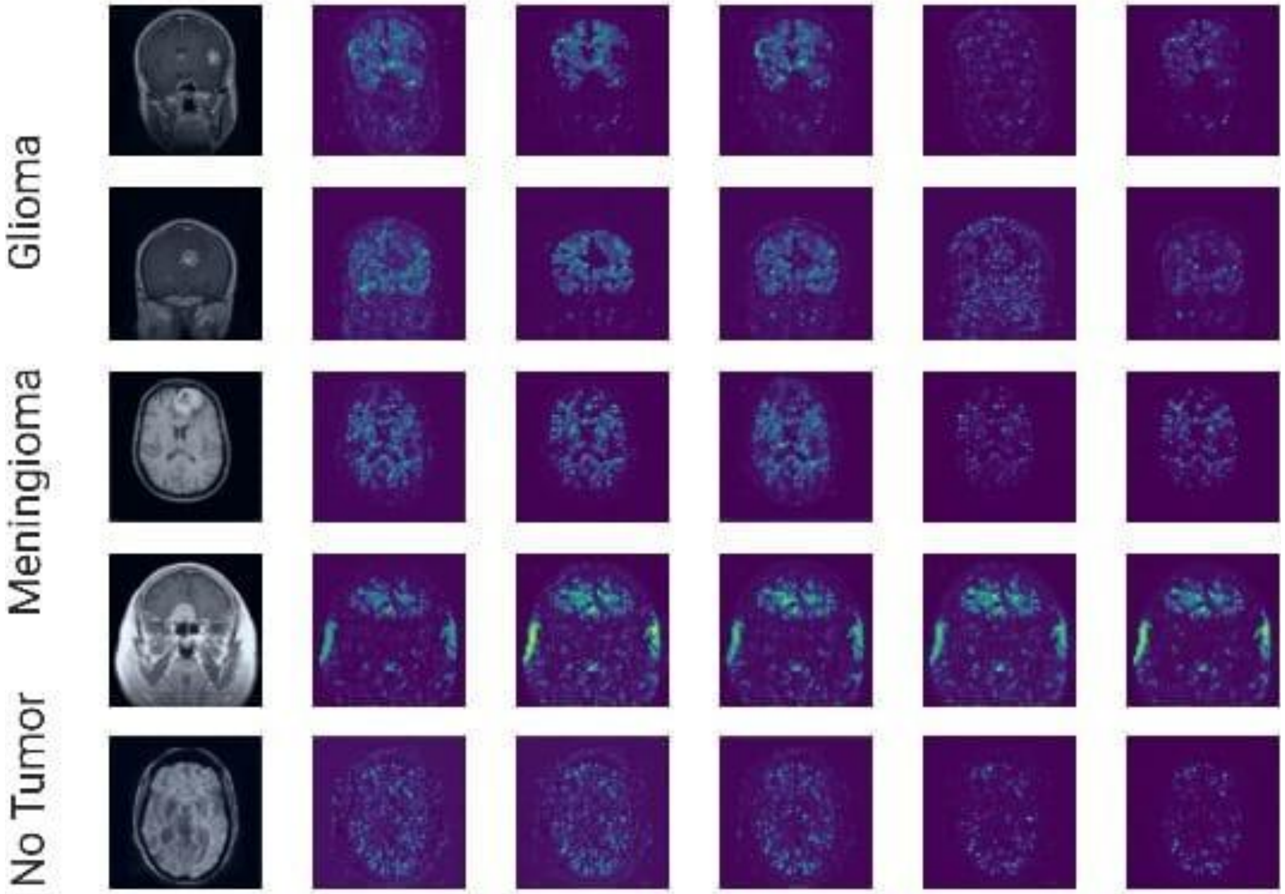




Where is model looking when decisions?



Self-attention maps extracted from their pre-trained model



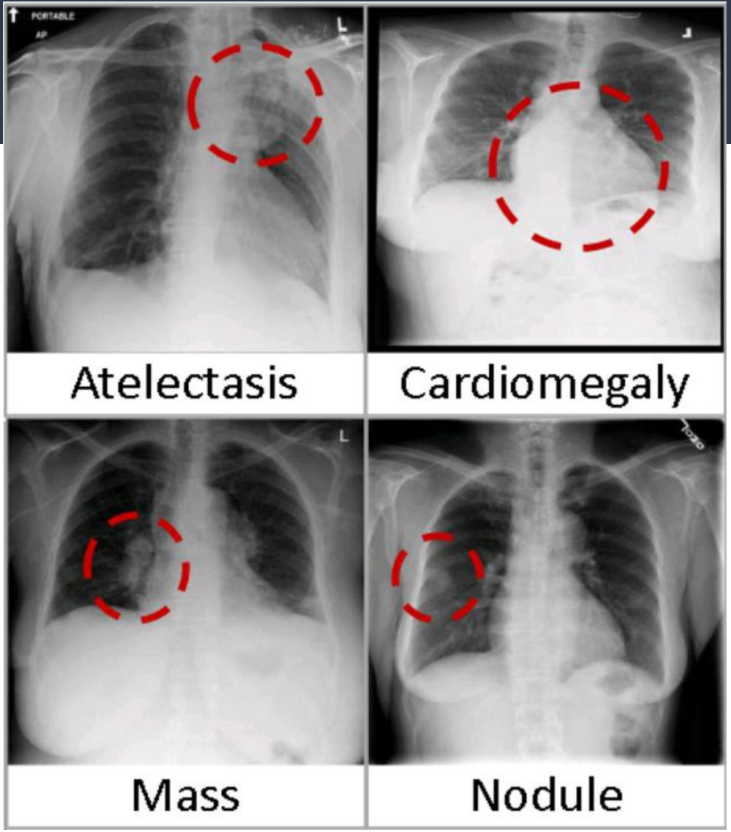
Self-attention maps extracted from our fine-tuned model

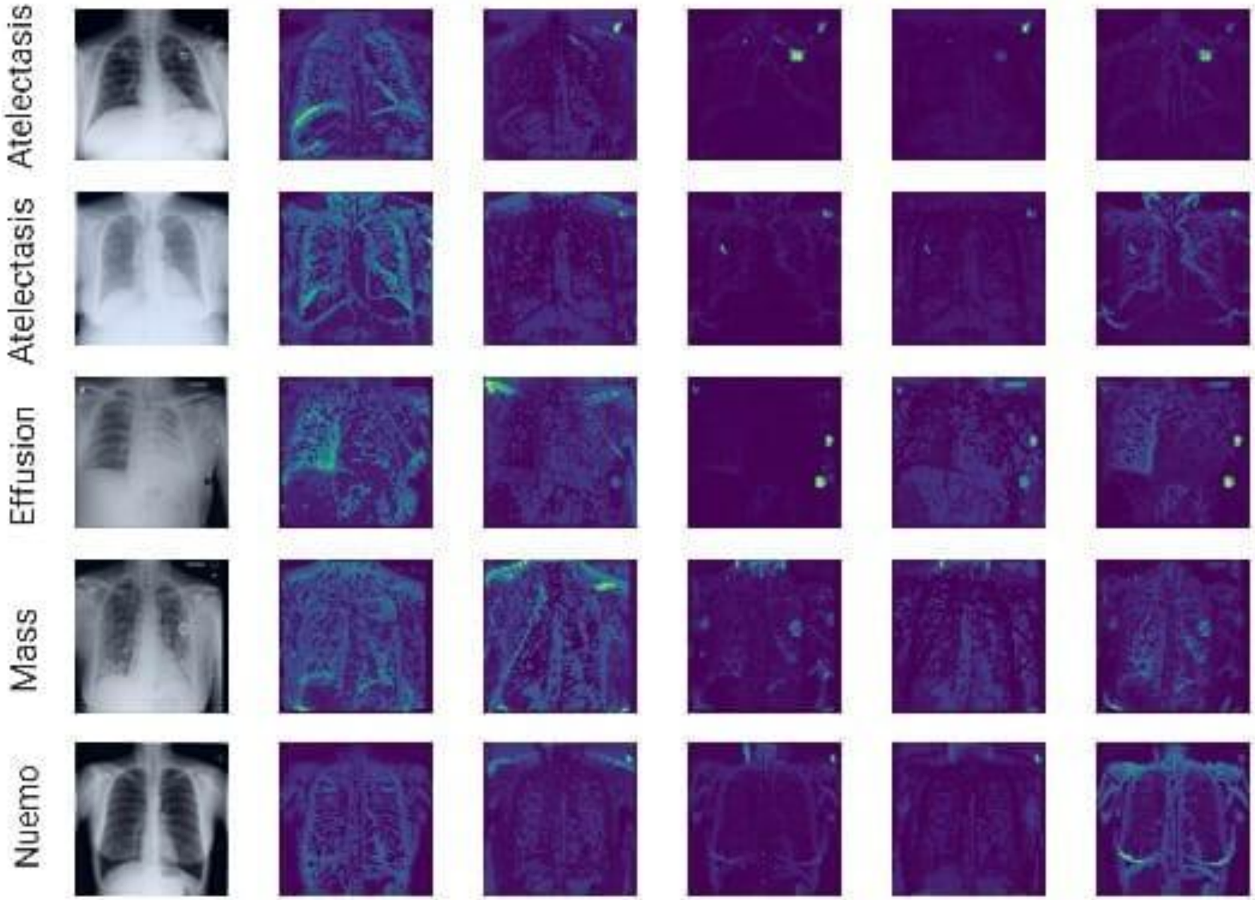
Test Dataset 2

<u>Details:</u>

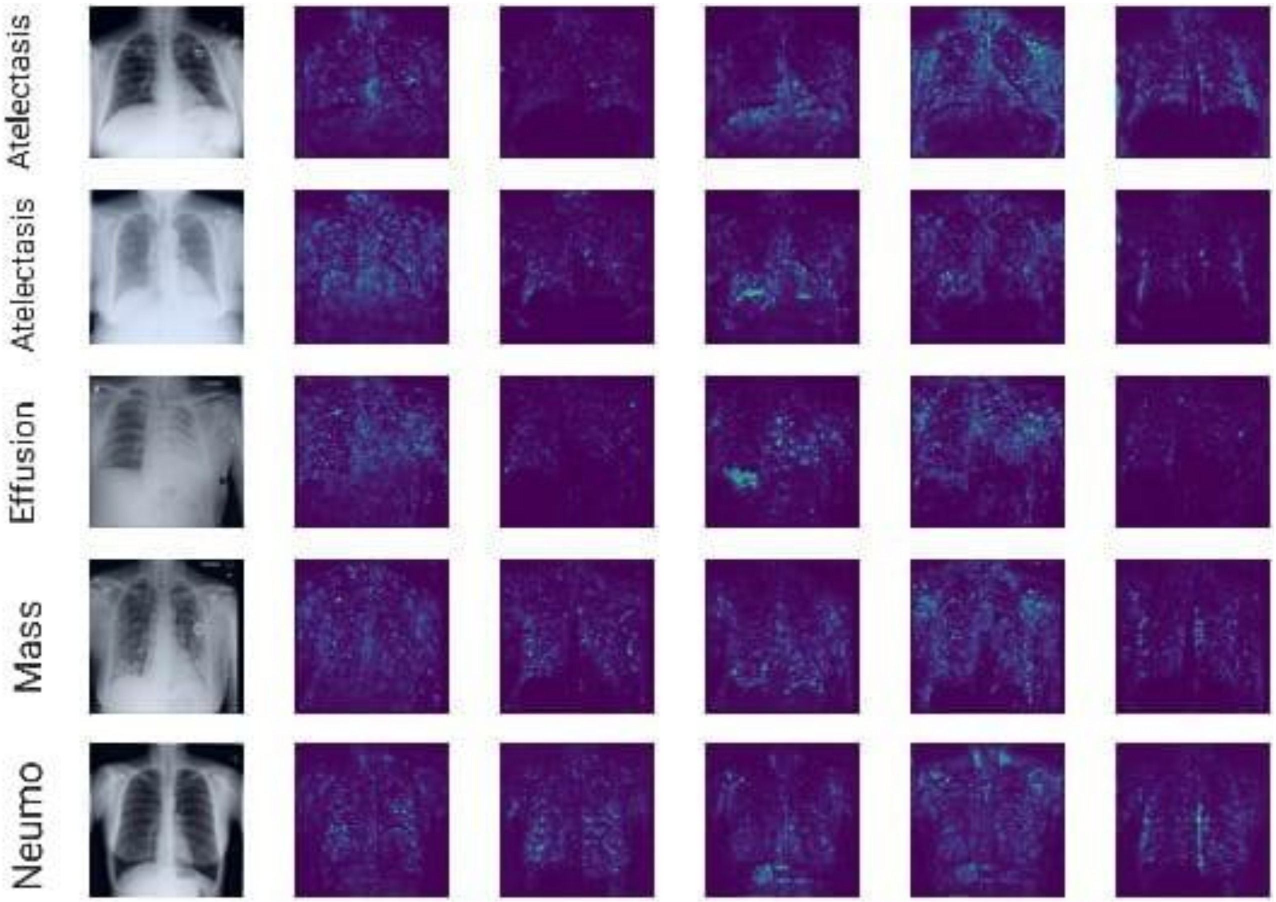
- Chest X-Ray 14 (CXR14)
- Includes 112,120 chest X-ray images (1024x1024) with one of 14 classes of diagnoses for each
- Filtered down to 5 balanced classes with ~1500 examples/class

Where is model looking when decisions?





Self-attention maps extracted from their pre-trained model



Self-attention maps extracted from our fine-tuned model

Future Directions

Computational

- Train for longer on both datasets to see if SOTA accuracy can be achieved with White-Box transformers
- Compare the extracted self-attention maps to the bounding boxes
- Apply the technique to other domain datasets

Domain

- Ask domain experts to segment areas of interest in image datasets with class labels/bounding boxes
- Ask domain experts to label images with the presence of certain diseases