

Unsupervised Pretraining of Foundation Models for Medical Imaging

Scott Chase Waggener

October 24, 2023

Overview

- Motivation
 - Convolutional networks and their limitations
 - Transformers, strengths and weaknesses
 - Desired properties of a medical imaging foundation model
- Unsupervised Pretraining Methods
 - Masked Autoencoder
 - Contrastive Embedding
 - Query Box Localization
- Results
 - APTOS2019 - Fundus photography
 - RSNA2023 - Abdominal CT scans
- Conclusions
- Future Work

Motivation

- Deep learning models can analyze medical images (such as mammograms) to assist in diagnosis [8]
- Many such models use convolutional architectures [3], which incorporate a locality prior
- Such a prior can accelerate learning, but can also be restrictive

Motivation

- Mammographic screenings capture four standard views of the breasts
 - Medio-lateral oblique (MLO)
 - Cranio-caudal (CC)
 - MLO and CC views are approximately orthogonal



Figure: MLO View

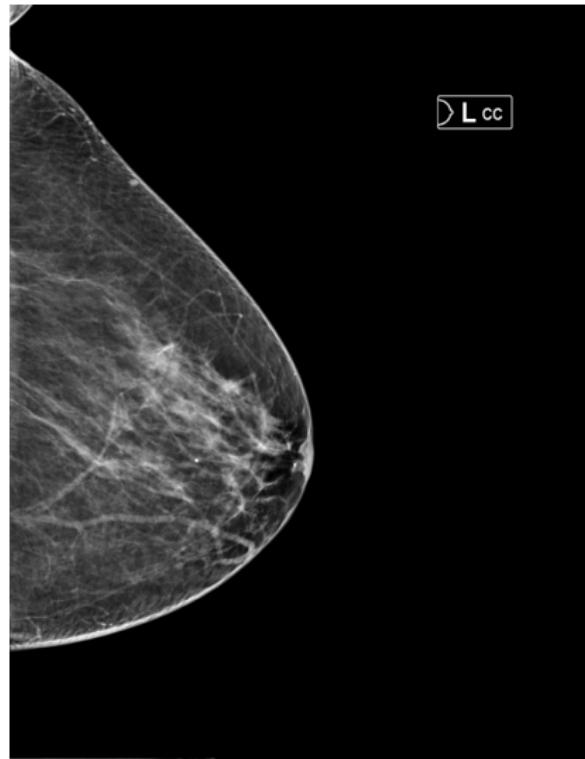


Figure: CC View

Motivation

- The standard views are typically examined together to leverage bilateral symmetry
- Reference images are used (when available) to compare to the standard views
- Sometimes lesions will appear in multiple views

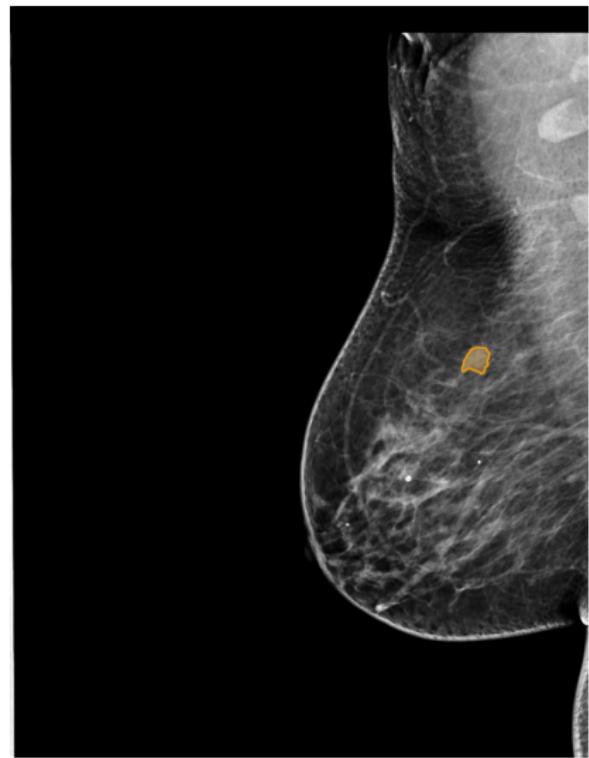


Figure: Lesion (MLO View)

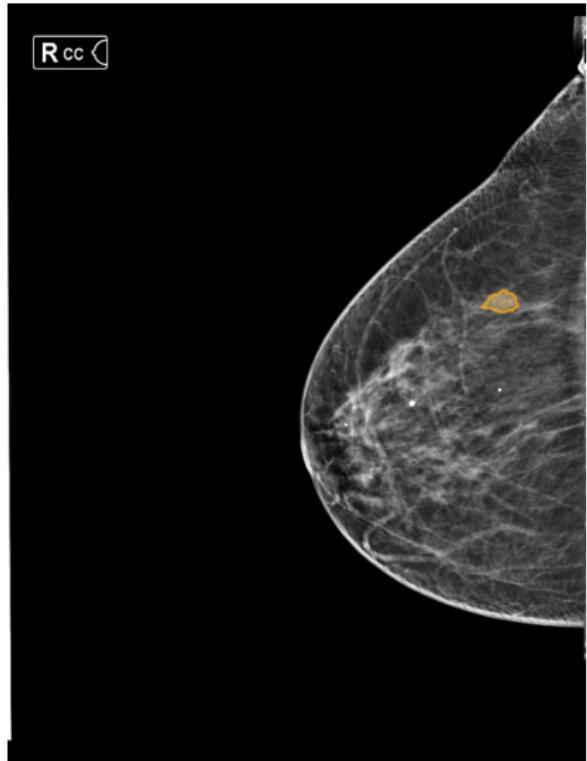


Figure: Lesion (CC View)

Motivation

A strong mammography model should incorporate all available information:

- Orthogonal nature of MLO and CC views is incompatible with a locality prior
- Additional imaging may be available with similar incompatible relationships
- Textual information may also be available (medical reports)

Similar considerations apply to other medical imaging modalities

Motivation

Objectives to improve mammographic performance:

- Relax the locality prior
- Support a variable number of additional context images
- Support textual information

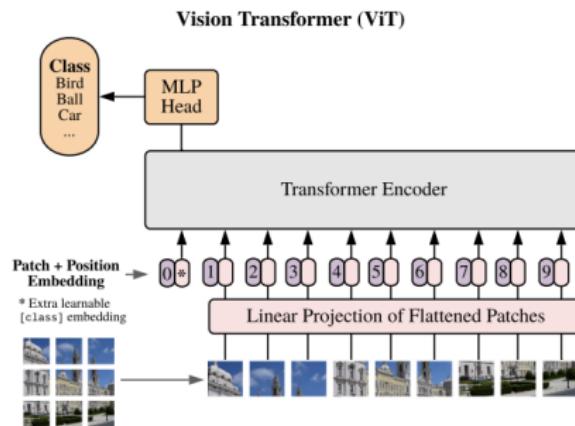
These objectives will also improve performance on other medical imaging modalities like:

- X-ray
- CT
- MRI
- Ultrasound

Motivation

Vision Transformers (ViTs) can address these objectives:

- Vision Transformers (ViTs) can achieve state-of-the-art results on image classification tasks [6]
- Attention is not restricted by a locality prior
- Transformers are cardinality invariant
- Transformers can support multiple modalities (Med-PaLM2) [9]



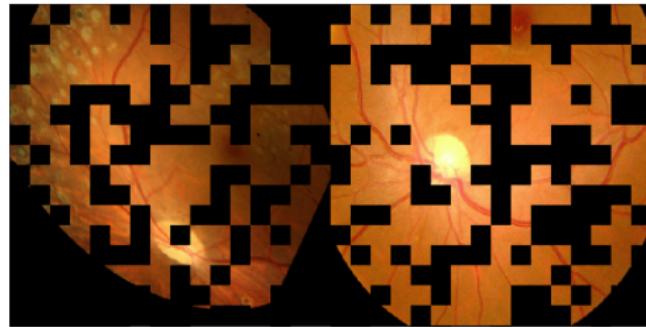
Motivation

ViTs are not without drawbacks:

- They require a large amount of labeled data (JFT-300M) [6]
 - Medical imaging datasets are often small and expensive to label
 - Thousands of images instead of millions or billions
- Or they require clever training methods (DeiT) [10]
- Self attention is expensive to compute (quadratic)
- Relatively difficult to train from scratch
 - Numerical instability
 - Sensitivity to batch size
 - Resource intensive

Masked Autoencoder

- Follows from masked language modeling (BERT) [5]
- Mask a subset of the input patches, regress to the original input



Contrastive Embedding

- Follows from popular contrastive learning methods (DINO) [2]
- Model creates an embedding vector for each image
- Embeddings should be similar for augmented versions of the same image
- Embeddings should be dissimilar for different images

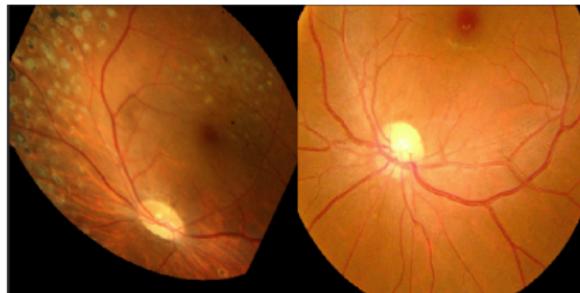


Figure: Global augmentation



Figure: Local augmentation

Query Box Localization

- Inspired by UP-DETR, a pretraining method for detection models [4]
- Regions of interest are randomly selected and augmented
- Given the original image and the augmented image, the model should predict the bounding box of the region of interest



Figure: Original image

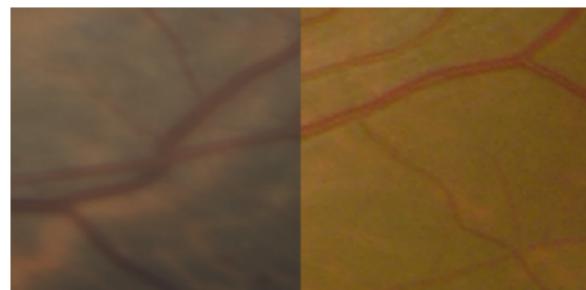


Figure: Augmented ROIs

Methods

Architecture:

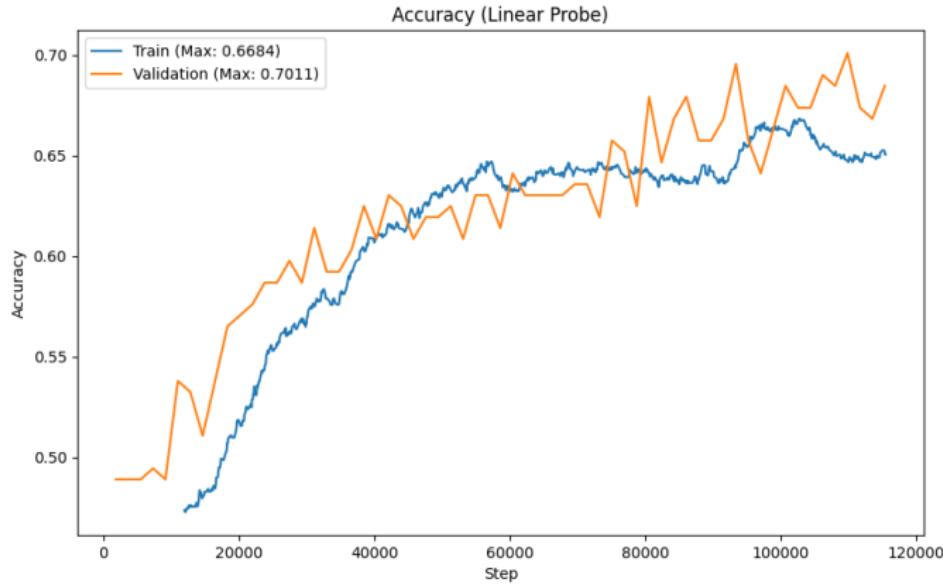
- Inspired by ViTDet [7]
- Standard patch embedding with log-spaced sinusoidal position embeddings
- Window attention without shifting
- Global attention at periodic intervals

Training:

- One or more of the pretraining methods are incorporated into the training process
- Tasks are cyclically sampled at each minibatch
- MAE and CE

Results (APTOs2019)

- Maximum validation accuracy of 0.7011 achieved under linear probing.
- Despite small dataset size, no overfitting observed



Results (APTOs2019)



Figure: Masked Input

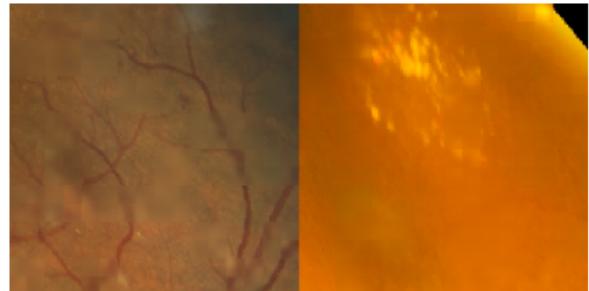


Figure: Model Prediction

Results (APTOs2019)

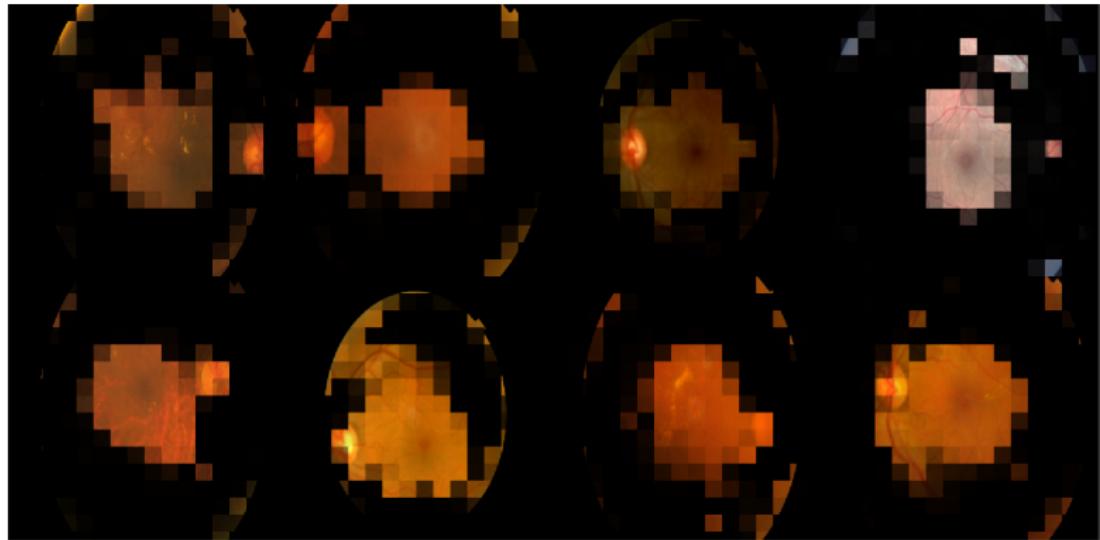


Figure: Contrastive embedding attention visualization

Results (APTOs2019)

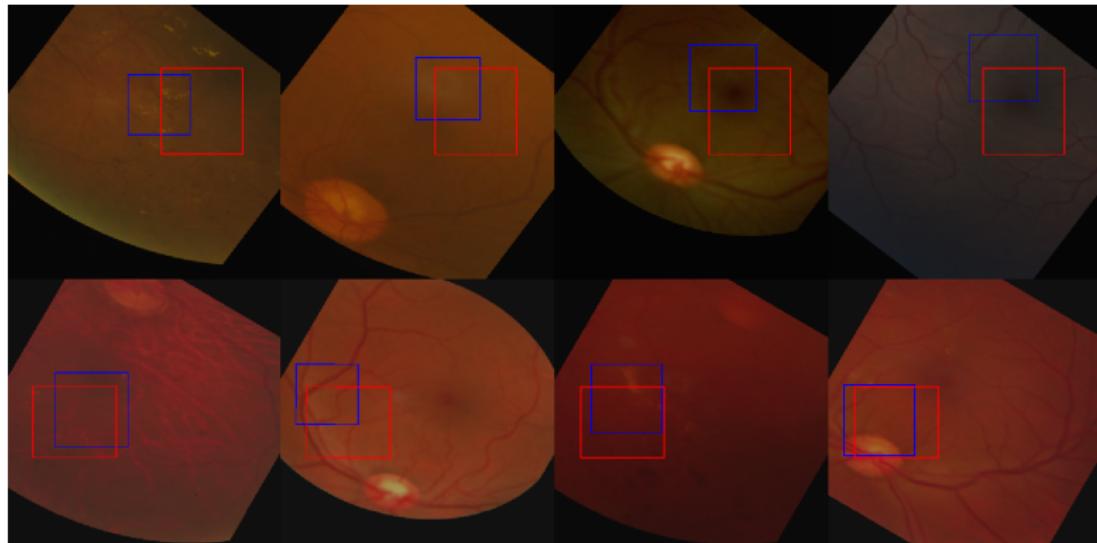
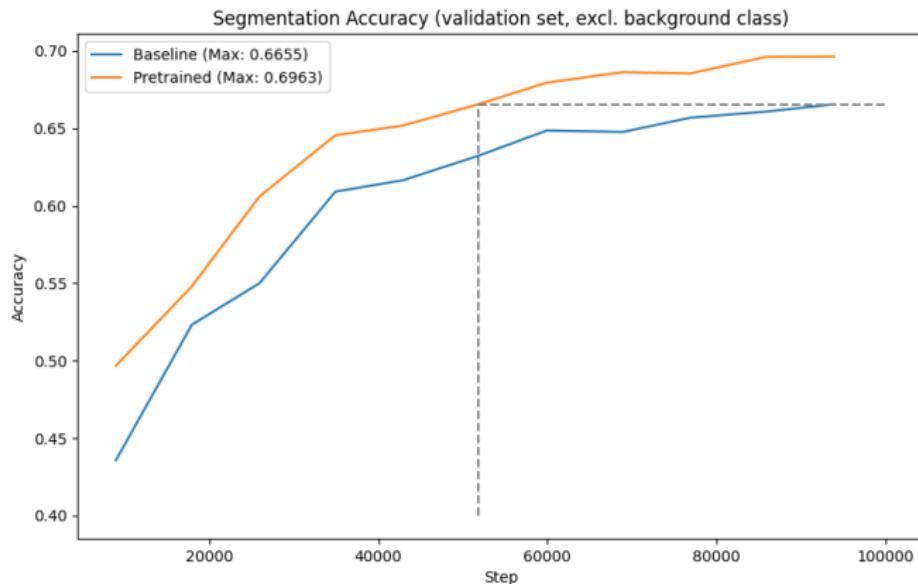


Figure: Query Box Localization. Predictions in blue, ground truth in red.

Results (RSNA2023)

- Pretraining improved maximum accuracy by 4.62%
- Pretrained model achieved the maximum accuracy of the non-pretrained model in $\approx 1/2$ as many training steps.



Conclusions

- Transformers can be pretrained on small datasets with unsupervised methods
- Pretraining can improve performance on medical imaging tasks
- Pretraining can reduce training time required to reach a given performance
- Multiple tasks can be trained simultaneously
- Toolkit for applying transformers to limited medical image datasets

Future Work

- Implement additional pretraining tasks (I-JEPA [1])
- Expand existing pretraining tasks to 3D
- Explore alternative architectures
- Optimize data annotation process / pipeline

References

- [1] Mahmoud Assran et al. *Self-Supervised Learning from Images with a Joint-Embedding Predictive Architecture*. 2023. arXiv: 2301.08243 [cs.CV].
- [2] Mathilde Caron et al. *Emerging Properties in Self-Supervised Vision Transformers*. 2021. arXiv: 2104.14294 [cs.CV].
- [3] Timothy Cogan, Maribeth Cogan, and Lakshman Tamil. "RAMS: Remote and automatic mammogram screening". In: *Computers in Biology and Medicine* 107 (2019), pp. 18–29. ISSN: 0010-4825. DOI: <https://doi.org/10.1016/j.combiomed.2019.01.024>. URL: <https://www.sciencedirect.com/science/article/pii/S0010482519300307>.
- [4] Zhigang Dai et al. "Unsupervised Pre-Training for Detection Transformers". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2022), pp. 1–11. DOI: 10.1109/tpami.2022.3216514. URL: <https://doi.org/10.1109%2Ftpami.2022.3216514>.
- [5] Jacob Devlin et al. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. 2019. arXiv: 1810.04805 [cs.CL].
- [6] Alexey Dosovitskiy et al. *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*. 2021. arXiv: 2010.11929 [cs.CV].
- [7] Yanghao Li et al. *Exploring Plain Vision Transformer Backbones for Object Detection*. 2022. arXiv: 2203.16527 [cs.CV].
- [8] Scott Mayer McKinney et al. "International evaluation of an AI system for breast cancer screening". In: *Nature* 577.7788 (2020), pp. 89–94. ISSN: 1476-4687. DOI: 10.1038/s41586-019-1799-6. URL: <https://doi.org/10.1038/s41586-019-1799-6>.
- [9] Karan Singhal et al. *Towards Expert-Level Medical Question Answering with Large Language Models*. 2023. arXiv: 2305.09617 [cs.CL].
- [10] Hugo Touvron et al. "Training data-efficient image transformers & distillation through attention". In: *International Conference on Machine Learning*. Vol. 139. July 2021, pp. 10347–10357.