Rensselaer

MICCAI 2024 Marrakesh MOROCCO

Foundation Model for Long-tailed, Multilabel Classification on Chest X-rays

Zefan Yang, *PhD Student*

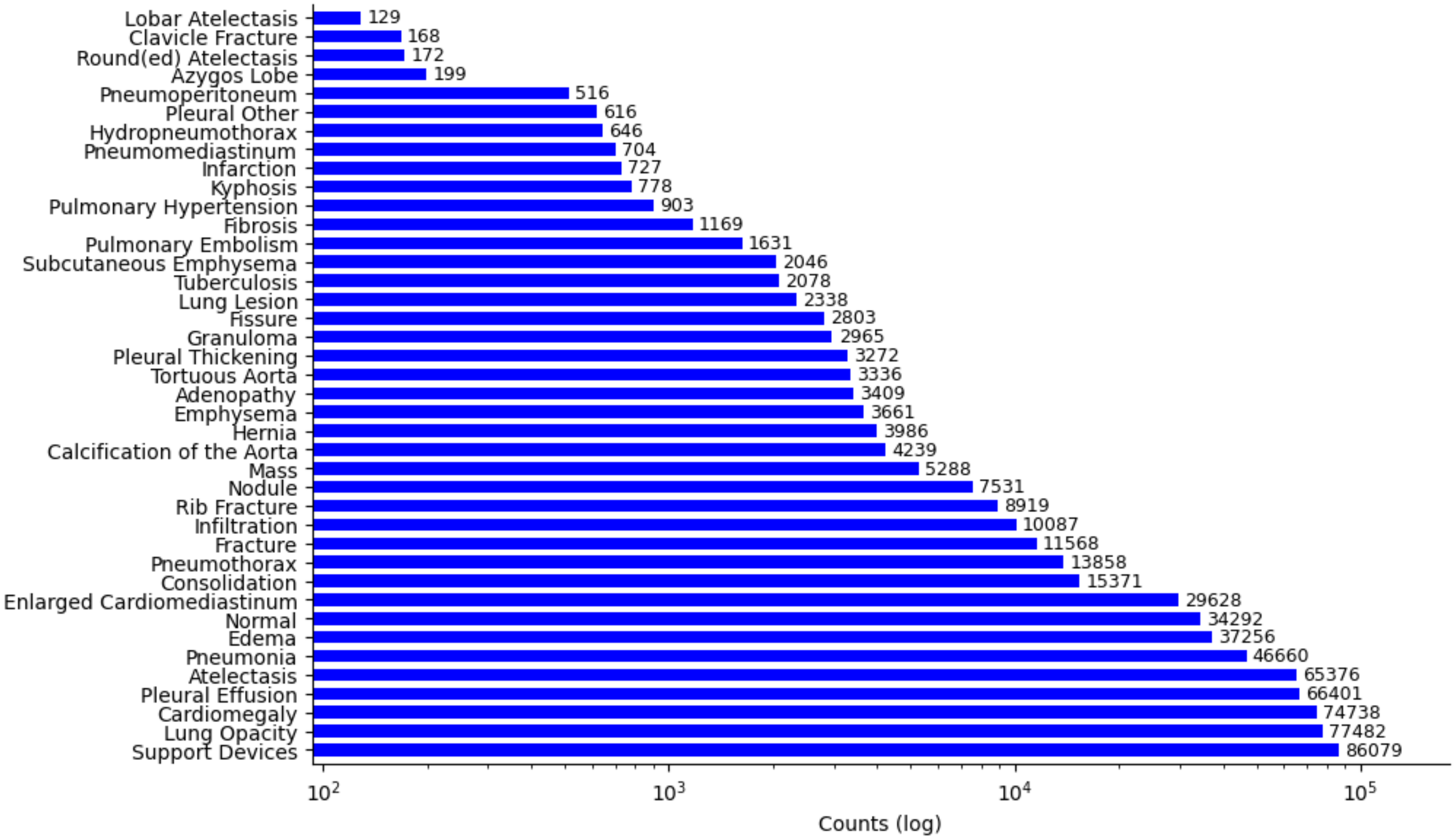
Department of Biomedical Engineering and the Center for Biotechnology and Interdisciplinary Studies, Rensselaer Polytechnic Institute, Troy, NY, USA

MICCAI 2024 CXR-LT Challenge Event

October 10, 2024

Background: MICCAI 2024 CXR-LT

- Classifying 40 disease findings with different levels of prevalence
- Challenges: Long-tailed distributions, multilabel classification problem



Data from: https://bionlplab.github.io/2024_MICCAI_CXRLT/

Background: General-purpose representations

Self-supervised learning methods, such as DINO, iBOT, and DINOv2, can obtain general-purpose representations useful for classification of 1K object classes (ImageNet-1k).

Caron, Mathilde, et al. "Emerging properties in self-supervised vision transformers." *Proceedings of the IEEE/CVF international conference on computer vision.*

Oquab, Maxime, et al. "Dinov2: Learning robust visual features without supervision." *arXiv preprint arXiv:2304.07193* (2023).

Such robust representations motivate us to develop a chest X-ray foundation model for the classification of the 40 disease findings from CXR-LT

Zhou, Jinghao, et al. "Image BERT Pre-training with Online Tokenizer." *International Conference on Learning Representations.*

Overview of our method

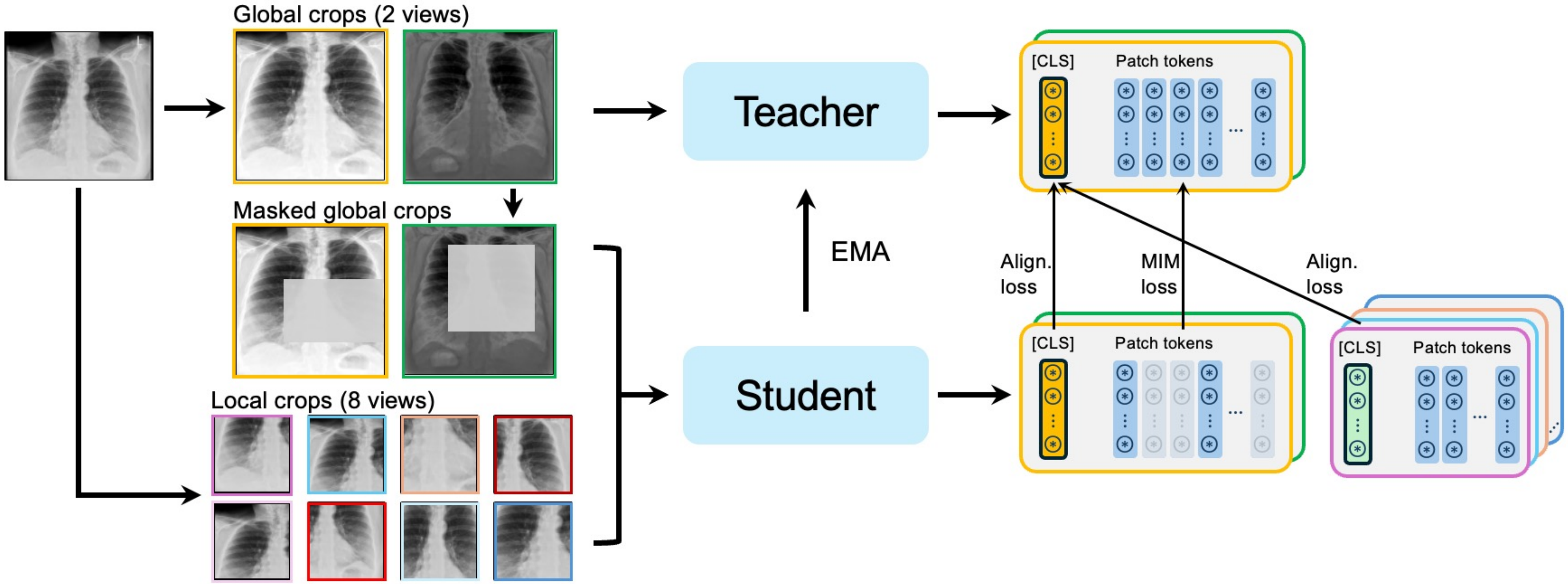
Phase 1: Self-supervised pretraining of the foundation model

Phase 2: Prediction head training for disease classification

Method: Foundation model pretraining

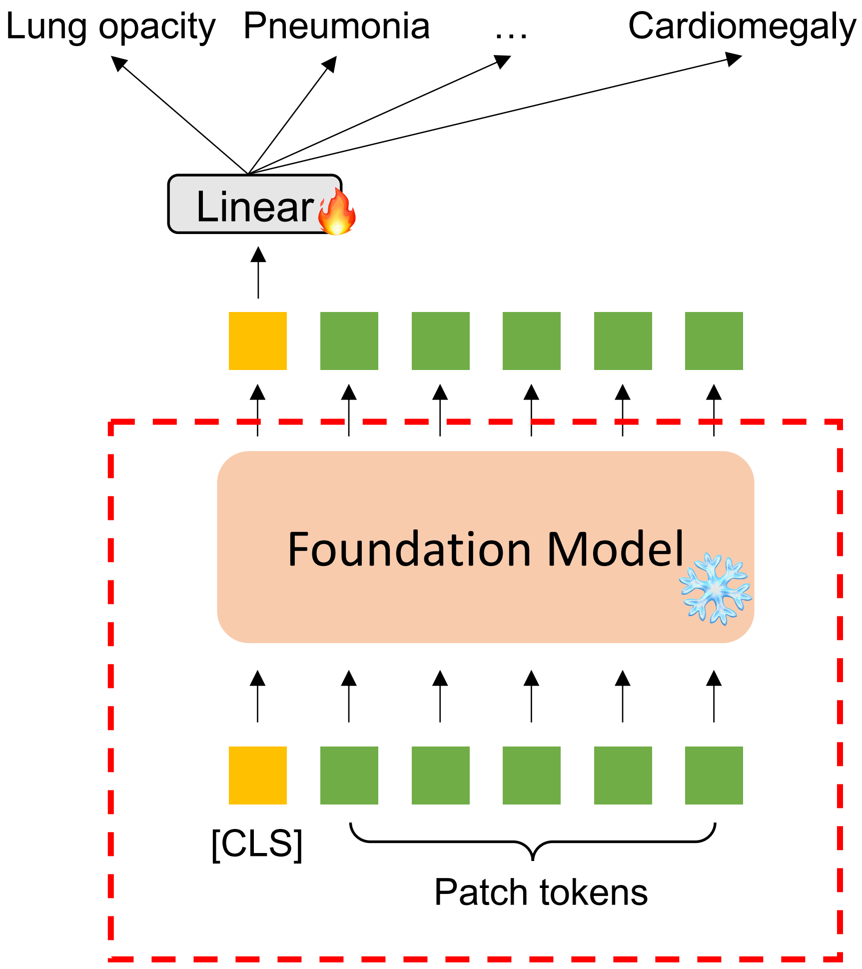
DINOv2 self-supervised pretraining

- More than 700K chest X-rays w/o labels
- Vision Transformers
- Masked image modeling
- [CLS] token alignment



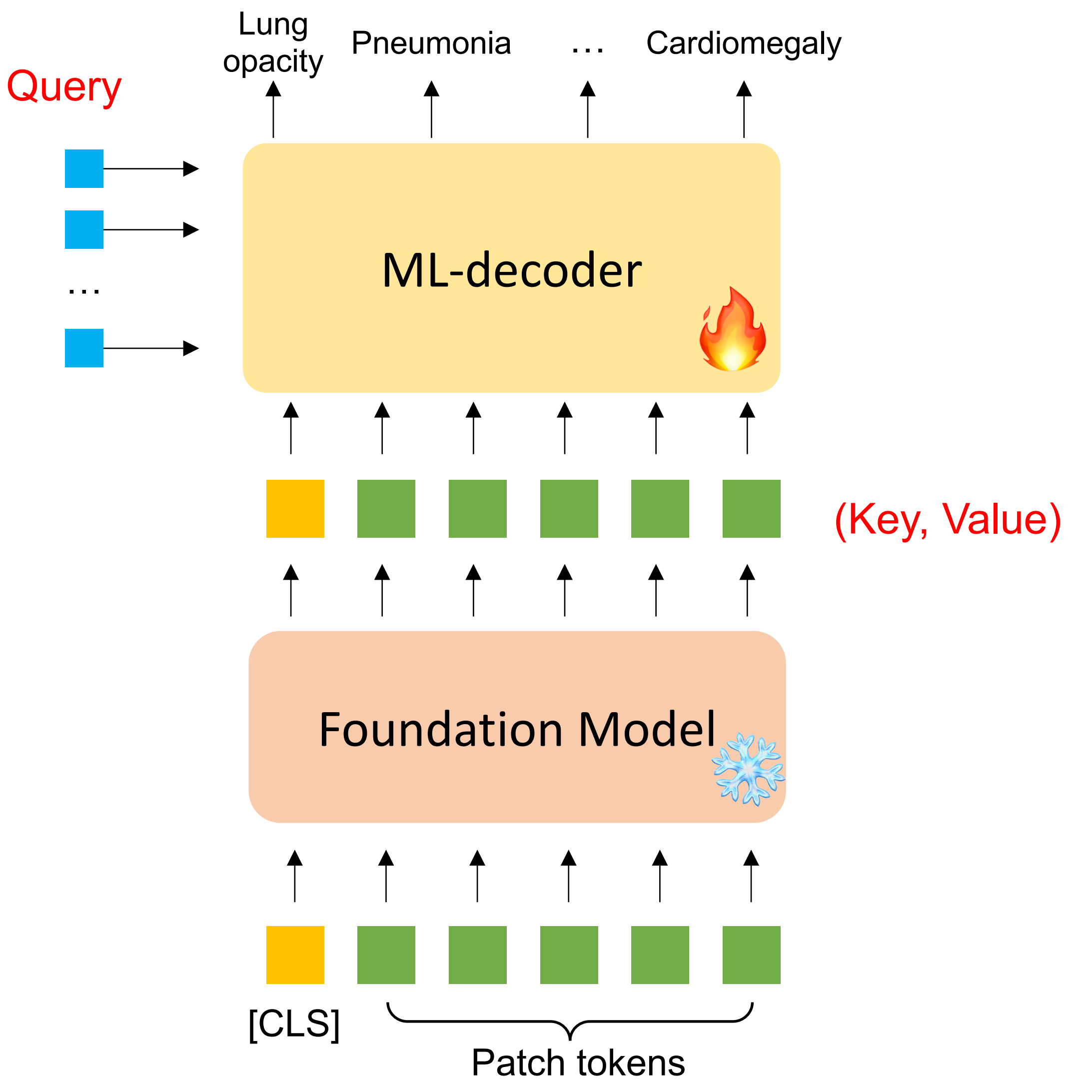
Oquab, Maxime, et al. "Dinov2: Learning robust visual features without supervision." *arXiv preprint arXiv:2304.07193* (2023).

Method: Prediction head training



The *[CLS] token* summarizes global information, lacking local features for classification of diverse disease findings

Method: Multilabel decoder

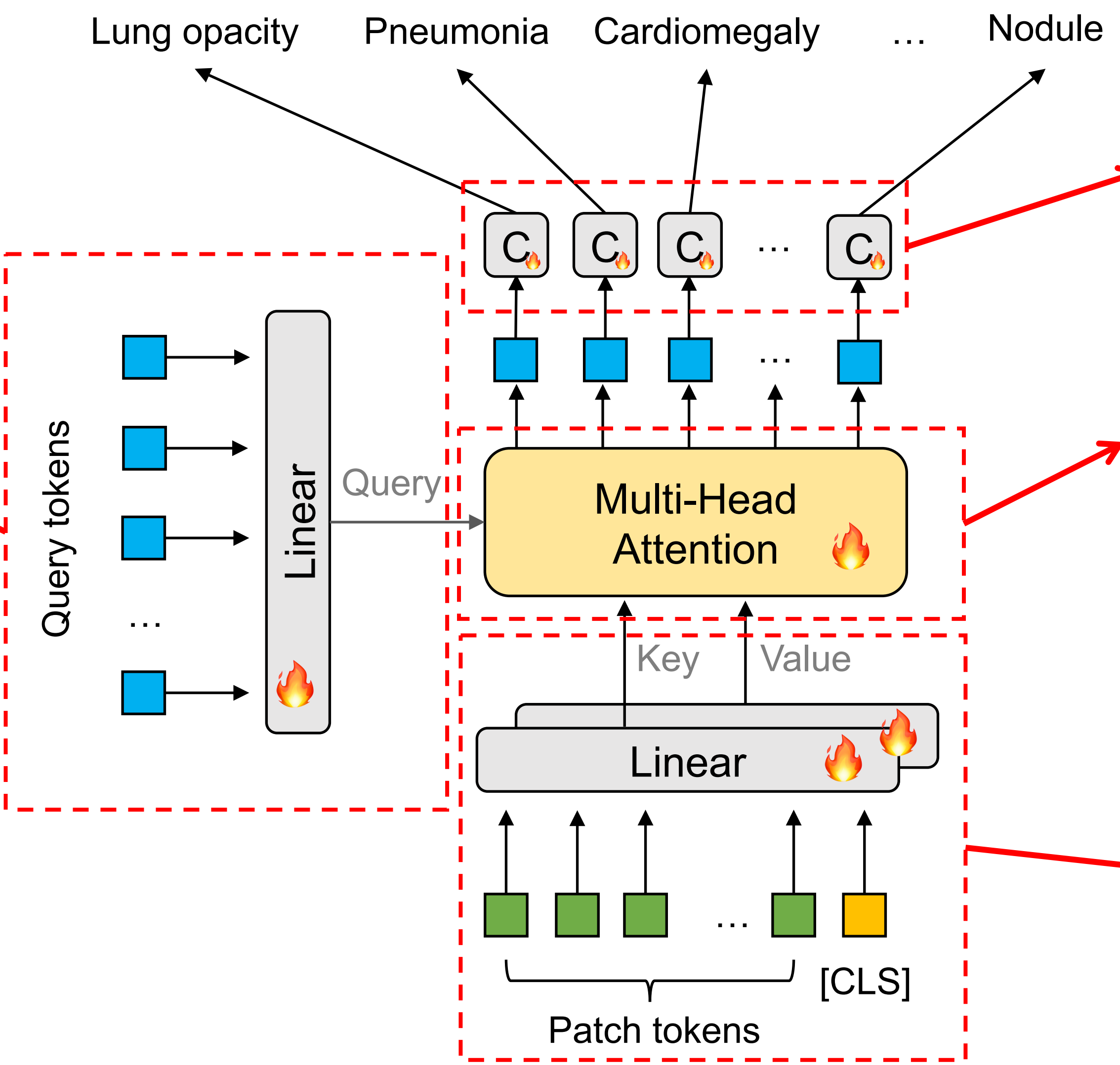


Multilabel decoder (ML-Decoder) uses query tokens to attend to class-specific patch tokens for classification of disease findings.

Ridnik, Tal, et al. "Ml-decoder: Scalable and versatile classification head." *Proceedings of the IEEE/CVF winter conference on applications of computer vision*. 2023.

Method: Details of ML-decoder

Query tokens each for a disease finding



Independent linear layers for each query token

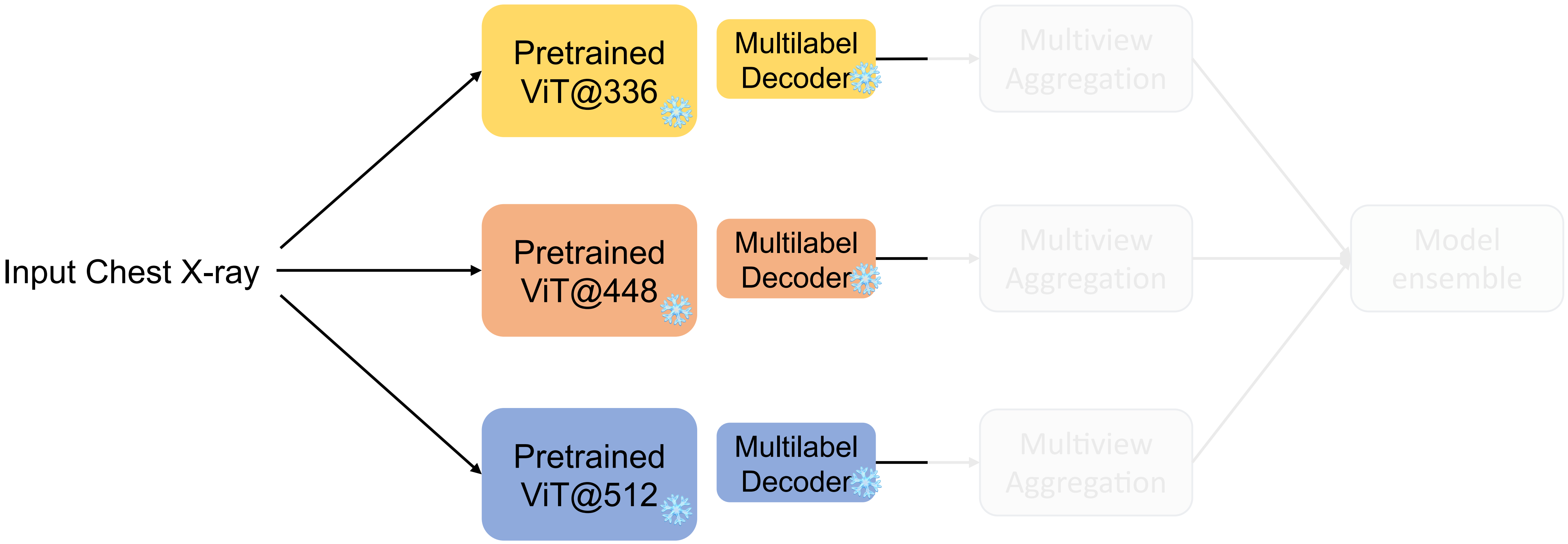
Compute attention weights to patch tokens

[CLS] and patch tokens as keys and values

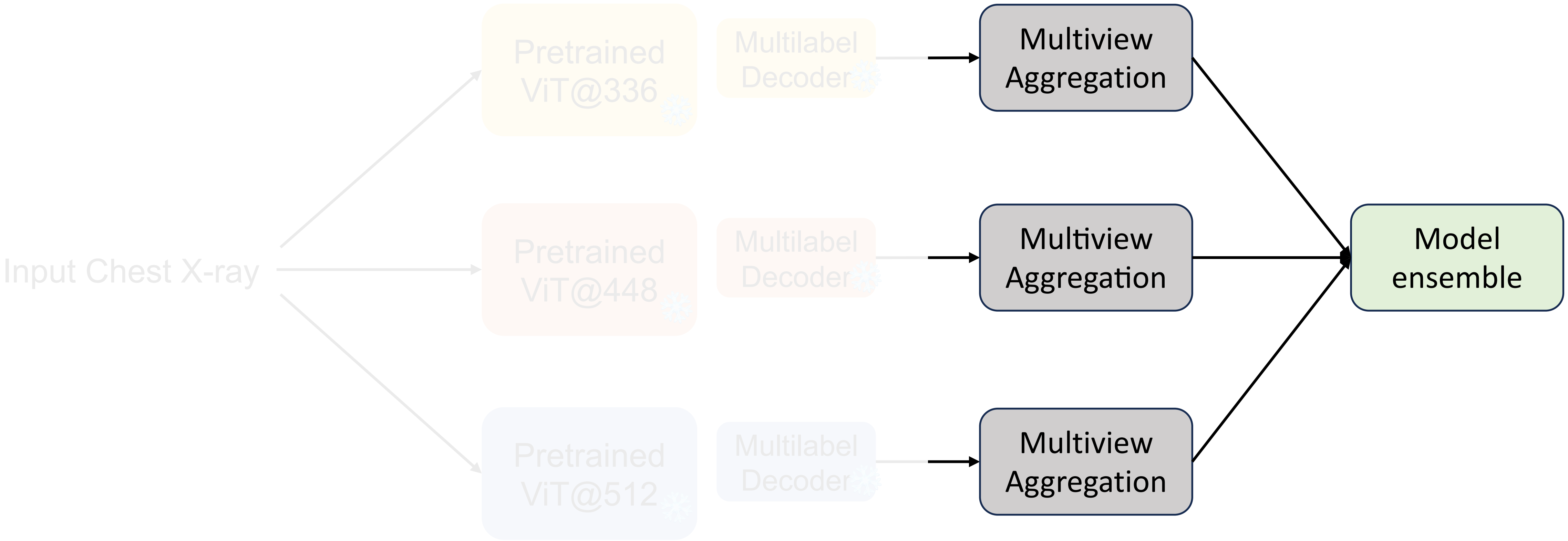
Ridnik, Tal, et al. "Ml-decoder: Scalable and versatile classification head." *Proceedings of the IEEE/CVF winter conference on applications of computer vision*. 2023.

Method: Foundation models at different input resolutions

Perform inference-time ensemble with foundation models pretrained on chest X-rays of 336^2, 448^2, and 512^2 input size

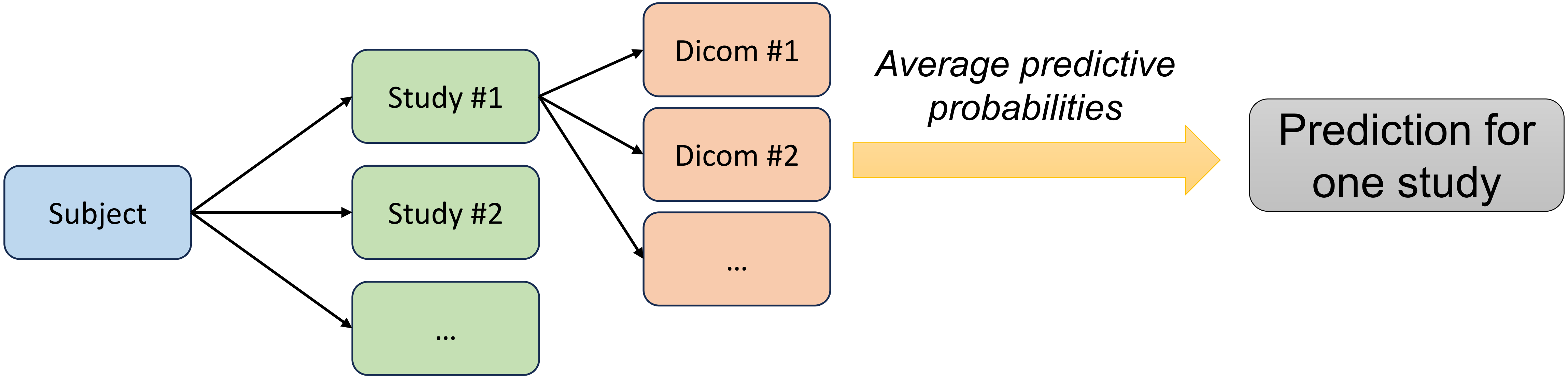


Method: Aggregation within one study



Kim, Dongkyun. "Chexfusion: Effective fusion of multi-view features using transformers for long-tailed chest x-ray classification." *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023.

Method: Multiview aggregation within one study



Put a constraint of prediction consistency within one chest X-ray study

Kim, Dongkyun. "Chexfusion: Effective fusion of multi-view features using transformers for long-tailed chest x-ray classification." *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023.

Experiments: Official training set

Development phase

- Chest X-rays w/ labels: 258,871

Training:
207,096 CXRs
(80%)

Validation:
51,775
(20%)

Experiments: Dataset for self-supervised pretraining

Dataset	# images
CXR-LT	207,096
CheXpert [1]	223,648
PadChest [2]	160,861
NIH Chest X-ray14 [3]	86,524
BRAX [4]	32,748
Total	710,877

Training:
207,096 CXRs
(80%)

Only the held-out subset of the official training set were used to prevent any data leakage

[1] Irvin, Jeremy, et al. "Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison." Proceedings of the AAAI conference on artificial intelligence. Vol. 33. No. 01. 2019.
[2] Bustos, Aurelia, et al. "Padchest: A large chest x-ray image dataset with multi-label annotated reports." Medical image analysis 66 (2020): 101797.
[3] Wang, Xiaosong, et al. "Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.
[4] Reis, Eduardo P., et al. "BRAX, Brazilian labeled chest x-ray dataset." Scientific Data 9.1 (2022): 487.

Experiments: Self-supervised pretraining settings

Hyperparameters

- Network architecture: ViT-L
- Patch size: 16
- iBOT loss weight: 3
- DINO loss weight: 1
- Epoch length: 25,000
- Epochs: 100
- (Local, global) crop size:
  (96, 224), (128, 336), (144, 448), (144, 512)

Pretraining input resolution (in pixel)

- $224^2 \rightarrow 336^2, 448^2, 512^2$

Experiments: Prediction head training

Official training set

- Train: 207,096
- Val: 51,775

Hyperparameter search

- Learning rates: [1e-5, 2e-5, 5e-5, 1e-4, 2e-4, 5e-4, 1e-3, 2e-3, 5e-3]
- Train for 10 epochs
- Binary cross-entropy loss

Learning rate that gives best performance at val

- Combine train and val sets
- Learn the prediction head for 10 epochs at the optimal learning rate
- Early stopping at the 2nd epoch for test-time inference

Results: Task 1 development set

Ablation studies:

- Network architecture
- Pretraining resolution
- Prediction head
- Multiview aggregation

CXR-LT results on Task-1 development set

Configuration	Multiview Aggregation	Head	mAP
ViT-Base@224	-	Linear	0.1795
ViT-Base@336	-	Linear	0.1852
ViT-Base@336	-	Linear Epoch 50	0.1842

Configuration	Multiview Aggregation	Head	mAP
ViT-Large@224	-	Linear	0.1798
ViT-Large@336	-	Linear	0.1922
ViT-Large@336	Average aggregation	Linear	0.2031
ViT-Large@336	[CLS] token concat.	Linear	0.1943

Configuration	Multiview Aggregation	Head	mAP
ViT-Large@336	-	Multilabel decoder	0.2311
ViT-Large@336	Average aggregation	Multilabel decoder	**0.2449**

Results: Task 2 development set

CXR-LT results on Task-2 development set

Configuration	Multiview Aggregation	Head	mAP
ViT-Base@224	-	Linear	0.2585
ViT-Base@336	-	Linear	0.2649
ViT-Base@336	-	Linear Epoch 50	0.2642

Configuration	Multiview Aggregation	Head	mAP
ViT-Large@224	-	Linear	0.2605
ViT-Large@336	-	Linear	0.2766
ViT-Large@336	Average aggregation	Linear	0.2896
ViT-Large@336	[CLS] token concat.	Linear	0.2801

Configuration	Multiview Aggregation	Head	mAP
ViT-Large@336	-	Multilabel decoder	0.3194
ViT-Large@336	Average aggregation	Multilabel decoder	**0.3330**

Results: Role of multi-resolution ensemble

CXR-LT results on Task-1 development set

Configuration	Label Aug.	Multiview Aggregation	Head	mAP
ViT-Large@336	True	True	Multilabel decoder	0.235
ViT-Large@448	True	True	Multilabel decoder	0.236
ViT-Large@512	Ture	True	Multilabel decoder	0.259
Ensemble	-	-	-	**0.271**

CXR-LT results on Task-2 development set

Configuration	Label Aug.	Multiview Aggregation	Head	mAP
ViT-Large@336	True	True	Multilabel decoder	0.323
ViT-Large@448	True	True	Multilabel decoder	0.323
ViT-Large@512	Ture	True	Multilabel decoder	0.342
Ensemble	-	-	-	**0.354**

Results: Task 2 test set

Results							
#	User	Entries	Date of Last Entry	mAP ▲	mAUC ▲	mF1 ▲	mECE ▲
1	XYPB	3	08/28/24	0.526 (1)	0.833 (3)	0.499 (1)	0.464 (6)
2	zguo	2	08/31/24	0.519 (2)	0.834 (2)	0.471 (4)	0.457 (3)
3	yangz16	3	08/29/24	0.511 (3)	0.836 (1)	0.265 (9)	0.744 (10)
4	YYama	3	08/29/24	0.509 (4)	0.829 (5)	0.474 (3)	0.462 (5)
5	lynnj	3	08/29/24	0.506 (5)	0.817 (7)	0.283 (8)	0.766 (13)
6	tianjie_dai	4	08/29/24	0.505 (6)	0.822 (6)	0.461 (5)	0.476 (7)
7	dongkyunk	1	08/30/24	0.505 (7)	0.832 (4)	0.486 (2)	0.460 (4)
8	pamessina	2	09/03/24	0.484 (8)	0.806 (9)	0.440 (6)	0.612 (8)
9	yyge	3	08/29/24	0.466 (9)	0.809 (8)	0.432 (7)	0.451 (2)
10	damanti	2	09/05/24	0.456 (10)	0.791 (10)	0.236 (11)	0.758 (12)
11	haoliu	2	08/27/24	0.389 (11)	0.758 (12)	0.130 (13)	0.673 (9)
12	phphuc612	1	09/05/24	0.371 (12)	0.759 (11)	0.261 (10)	0.312 (1)
13	StefanDenner	1	09/03/24	0.370 (13)	0.740 (13)	0.183 (12)	0.754 (11)

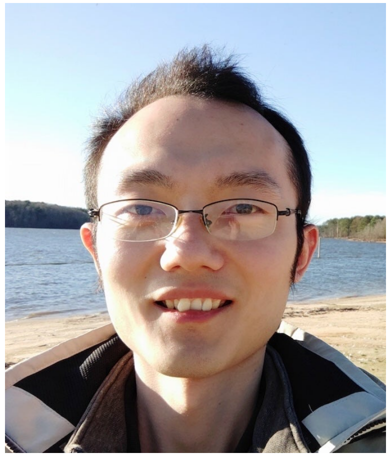
Summary

- We developed a chest X-ray foundation model that learns robust representations for the classification of disease findings on CXR-LT

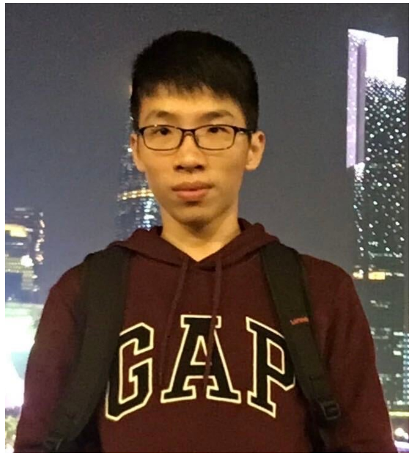
- We utilized a multilabel decoder to learn class-specific local features, attaining better results than a linear prediction head on the [CLS] token.

- Limitations: The proposed method has limited scalability. More work is to make the model capable of learning increasing disease findings while improving or maintaining its performance.

Acknowledgements



Xuanang Xu, PhD
Postdoctoral Research Associate



Zefan Yang
PhD Student

Pingkun Yan, PhD
Associate Professor

*We thank **Dr. Xuanang Xu** and **Prof. Pingkun Yan** at Rensselaer Polytechnic Institute for providing constructive comments on the method development.*

Rensselaer

110 8th St, Troy, New York 12180

Thank you for listening!

E-mail: yangz16@rpi.edu