Comparing Foundation Models and Fine-Tuning Methods for Carotid Atherosclerosis Prediction Using Retinal Images with Quantified Explainability via Vessel Segmentation Model

2025년도 1학기

[의료 데이터 기반 인공지능과 기계학습의 기초]

기말고사 프로젝트

서울대학교 의과학과

2023-24885 이설하

2023-24526 이혁종

Comparing Foundation Models and Fine-Tuning Methods for Carotid Atherosclerosis Prediction Using Retinal Images with Quantified Explainability via Vessel Segmentation Model

Foundation model 과 Fine tuning method 를 활용하여,

안저 이미지로 부터 동맥 경화를 예측하는 모델을 만들고,

단순 성능 평가 뿐 아니라 정량적으로 설명 가능성도 평가

Outline

1. Introduction
2. Method
3. Results and Discussion
4. Conclusion

1. Introduction

1.1 Opportunistic Learning

1.2 Retinal Images and Deep Learning

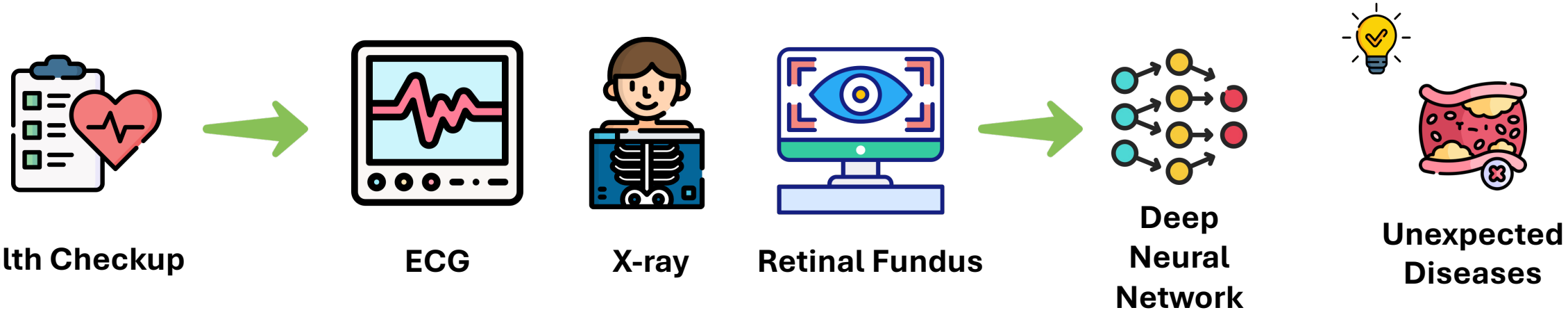
1.3 Foundation Models and Fine-tuning Methods

1.4 Explainability

1.5 Purpose of this Research

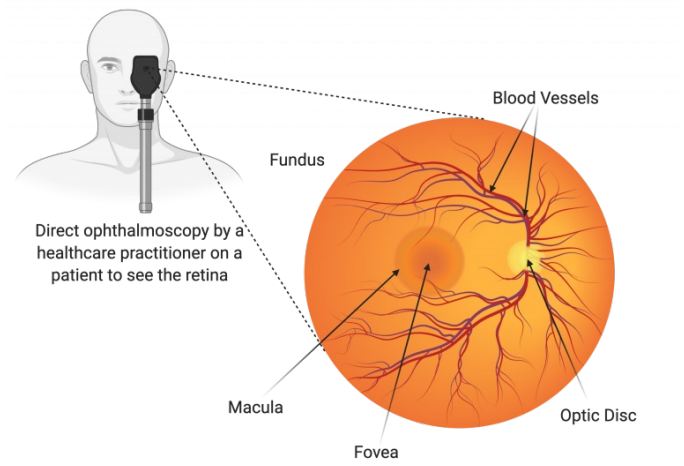
1.1 Opportunistic Learning

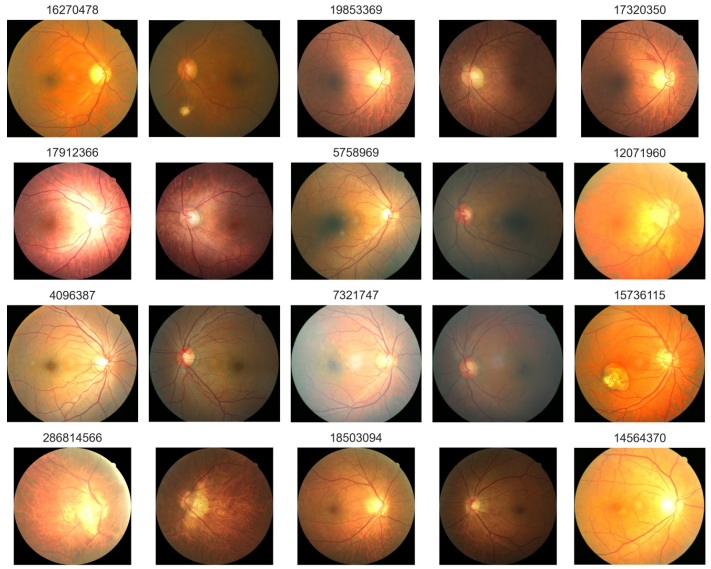
Identifying diseases incidentally during tests conducted for other reasons.

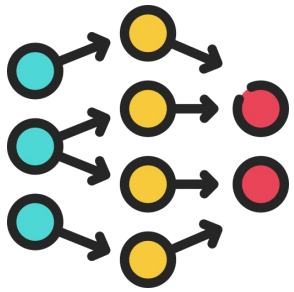


- ✓ Uses existing data
- ✓ No need for additional procedures (low cost and efficient)
- ✓ Enables early detection and efficient diagnosis

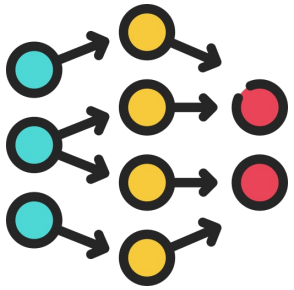
1.2 Retinal Images and Deep Learning







- Diabetic Retinopathy
- Age-related Macular Degeneration (AMD)
- Glaucoma
- Retinal Detachment



- Cardiovascular Disease
- Hypertension
- Atherosclerosis
- Alzheimer’s Disease
- Parkinson’s Disease

1.2 Retinal Images and Deep Learning

Association of Cardiovascular Mortality and Deep Learning-Funduscopic Atherosclerosis Score derived from Retinal Fundus Images

JOOYOUNG CHANG, AHRYOUNG KO, SANG MIN PARK, SEULGGIE CHOI, KYUWOONG KIM, SUNG MIN KIM, JAE MOON YUN, UK KANG, IL HYUNG SHIN, JOO YOUNG SHIN, TAEHOON KO, JINHO LEE, BAEK-LOK OH, AND KI HO PARK

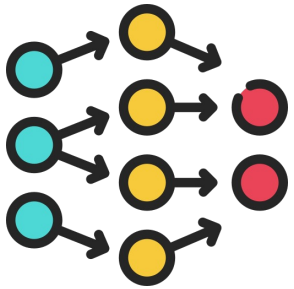
Deep learning predicts prevalent and incident Parkinson's disease from UK Biobank fundus imaging

Charlie Tran[1], Kai Shen[1], Kang Liu[2], Akshay Ashok[3], Adolfo Ramirez-Zamora[4], Jinghua Chen[5], Yulin Li[6] & Ruogu Fang[1,7,8]✉

A novel model for retinal imaging in the diagnosis of Alzheimer's disease

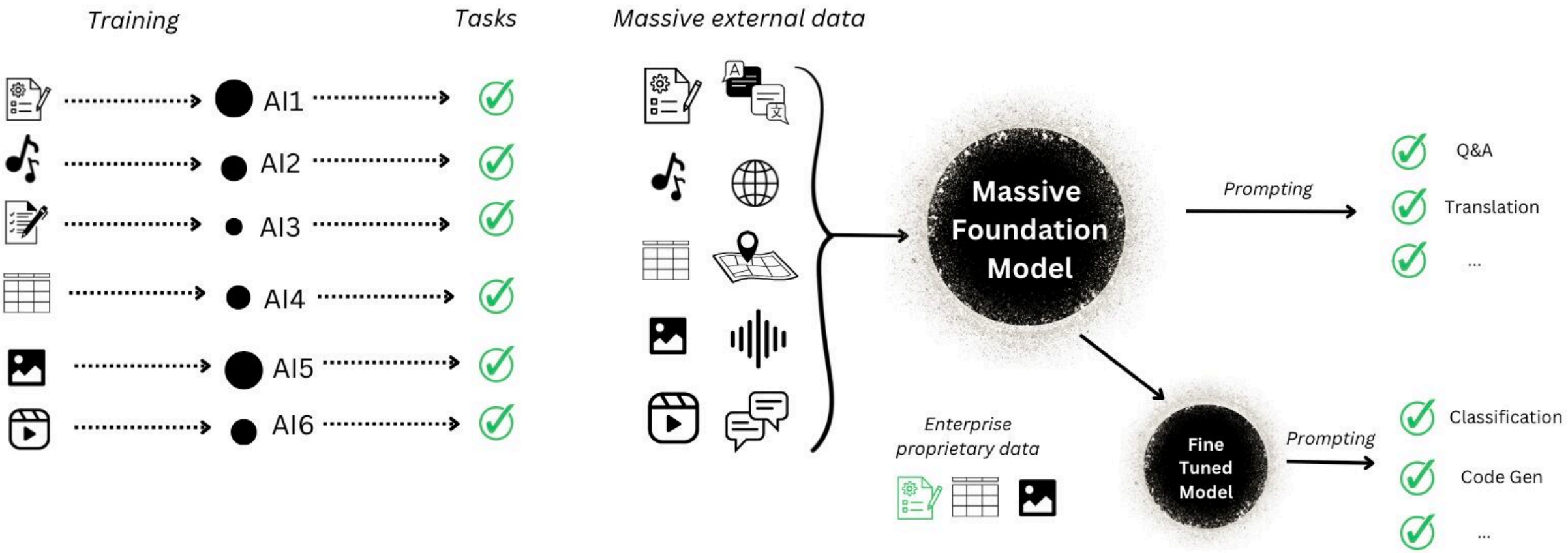
Kimia Heydari [1], Elizabeth J Enichen [2], Serena Wang [2], Grace C Nickel [2], Joseph C Kvedar [2]

Opportunistic Learning with Retinal Images



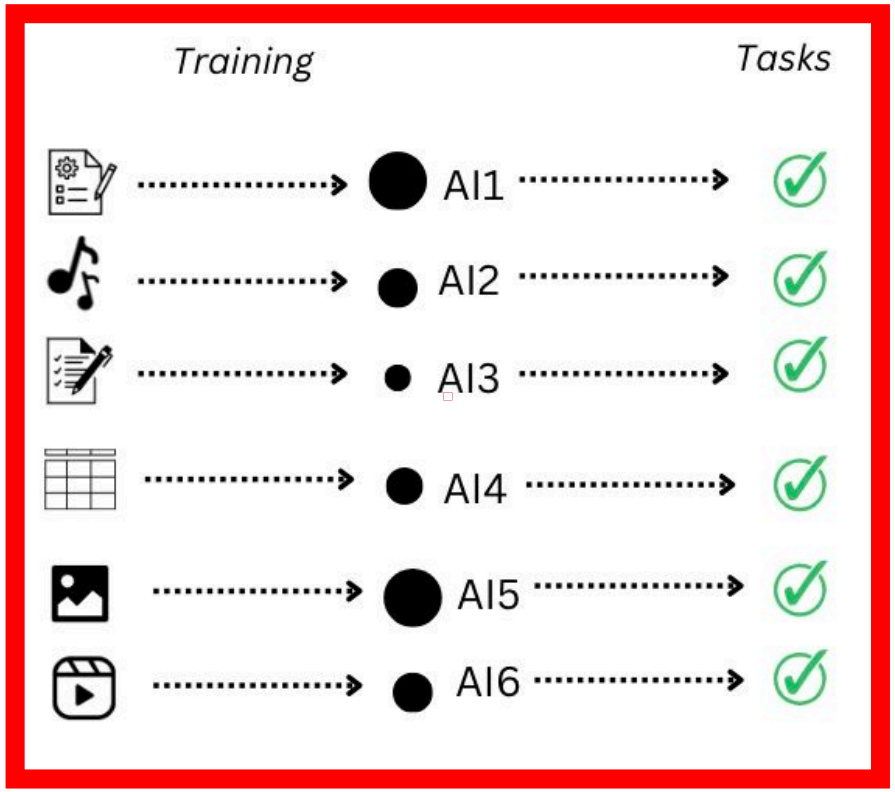
- Cardiovascular Disease
- Hypertension
- Atherosclerosis
- Alzheimer's Disease
- Parkinson's Disease

1.3 Foundation Models and Fine-tuning Methods



https://humanloop.com/blog/foundation-models

1.3 Foundation Models and Fine-tuning Methods

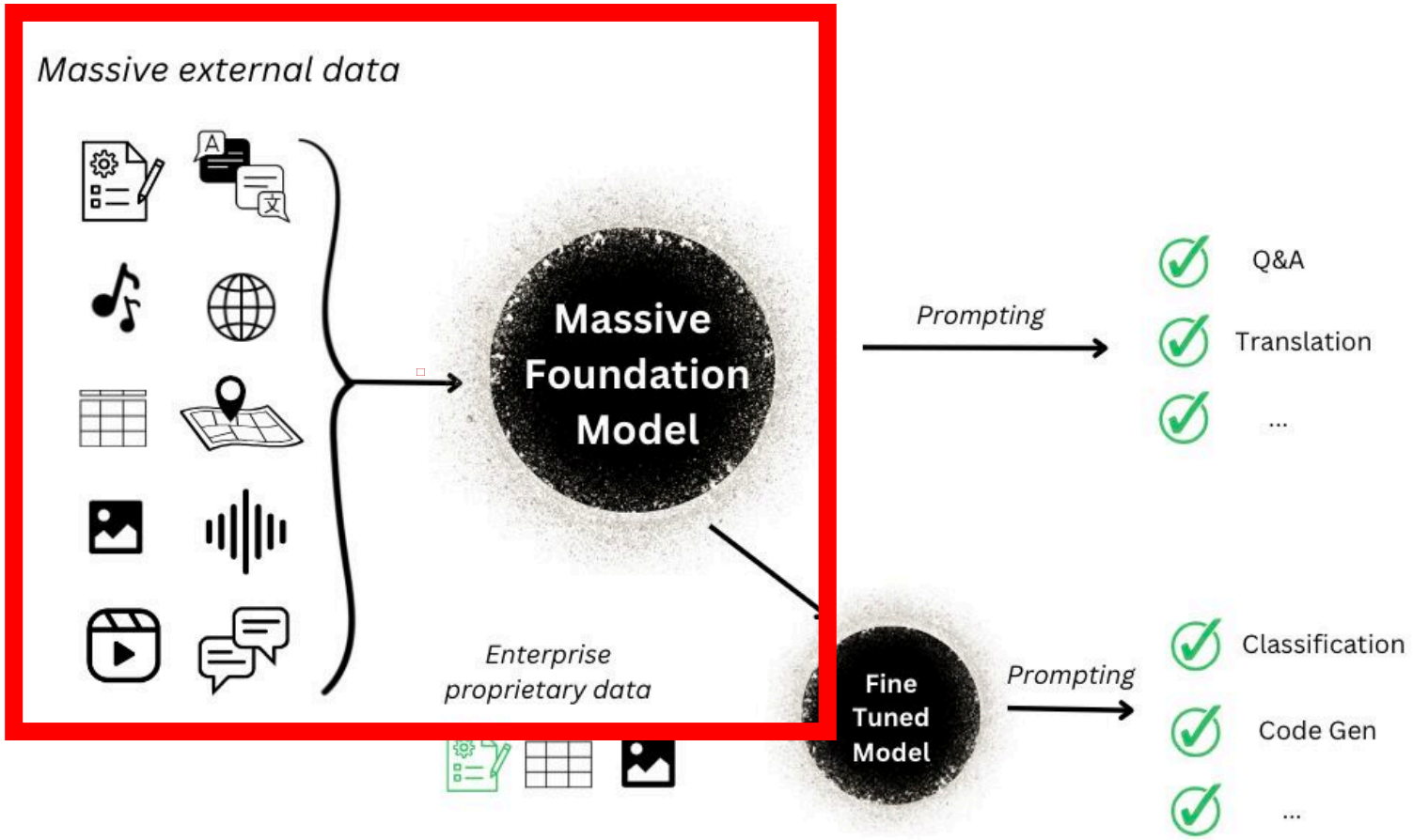


Traditional ML / DL models are [Task Specific] models that works well only specific task or dataset

https://humanloop.com/blog/foundation-models

1.3 Foundation Models and Fine-tuning Methods

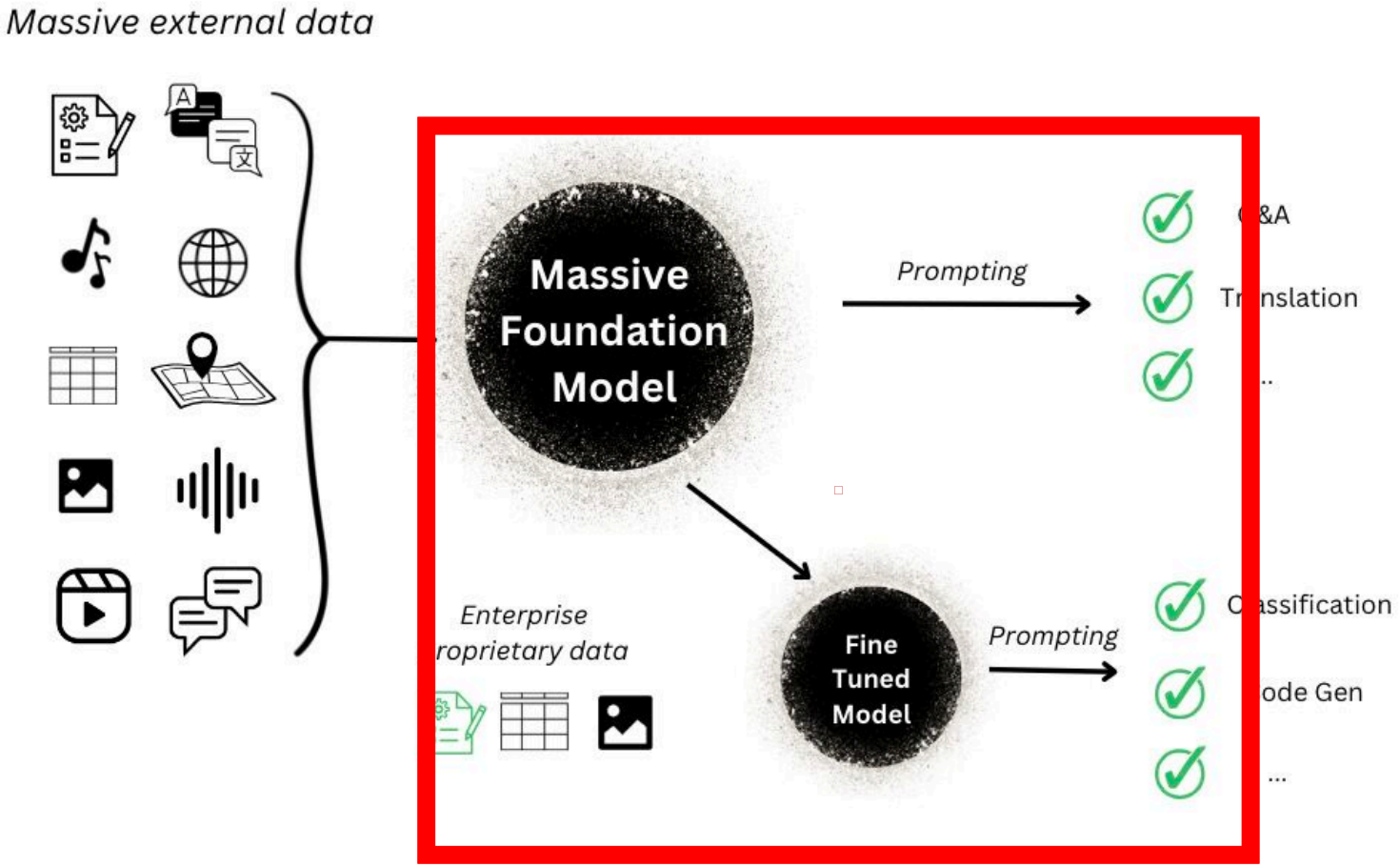
Foundation models are large-scale models trained on massive datasets and are capable of performing well across a wide range of general tasks.



https://humanloop.com/blog/foundation-models

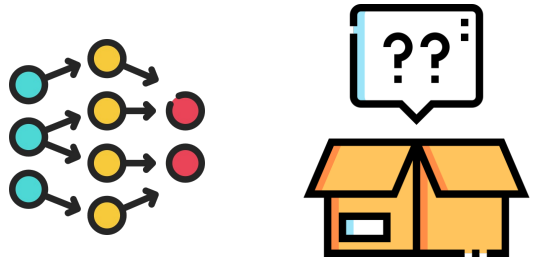
1.3 Foundation Models and Fine-tuning Methods

Since foundation models are large, they require [efficient fine-tuning methods] for downstream tasks.



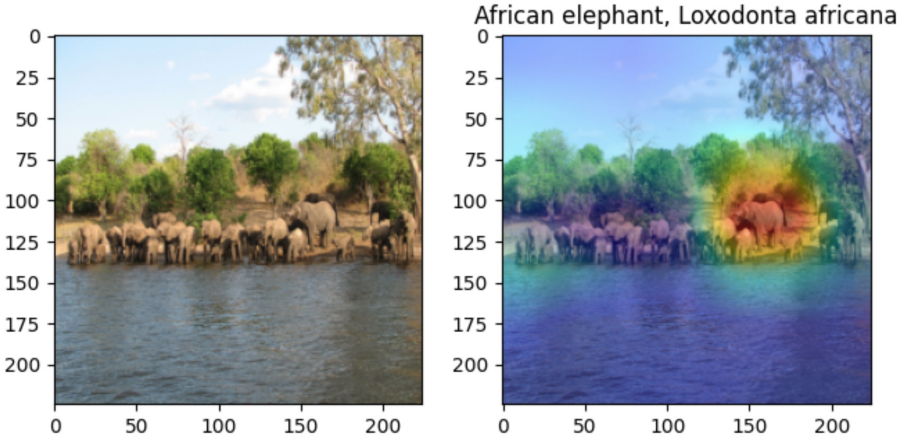
https://humanloop.com/blog/foundation-models

1.4 Explainability



Deep Learning model is Black Box Model

Doctors should explain the AI Decision to patients



GradCAM shows which parts of the image the model used for decision.

https://github.com/hyeok-jong/gradcam

1.5 Purpose of this Research

In this study

1. Developing an atherosclerosis prediction model using retinal images with various **foundation models and fine-tuning methods**.

2. To **quantify** and compare explainability, the "**average saliency intensity**" metric was proposed, measuring alignment between saliency maps and vessel regions.

2. Method

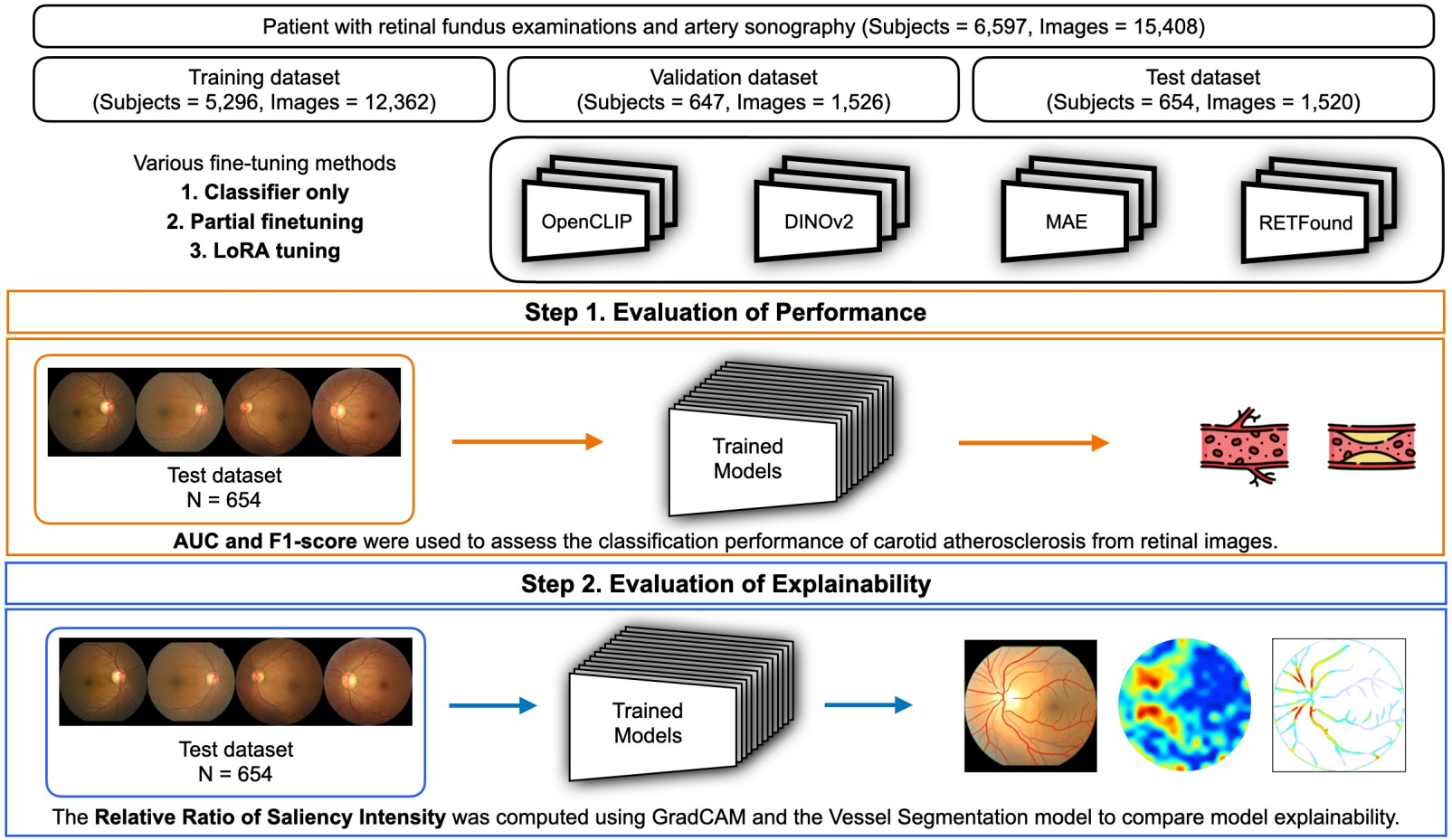
2.1 Study Design and Dataset

2.2 Foundation Models

2.3 Fine-tuning Methods

2.4 Explainability

2.1 Study Design and Dataset



- Health check-up data were collected from the **Health Promotion Center of Seoul National University Hospital** (HPC-SNUH).
- A total of **6,597 individuals** who underwent both retinal fundus photography and carotid ultrasound between January 2005 and December 2016 were included.
- **15,408 retinal fundus images** were used for analysis.

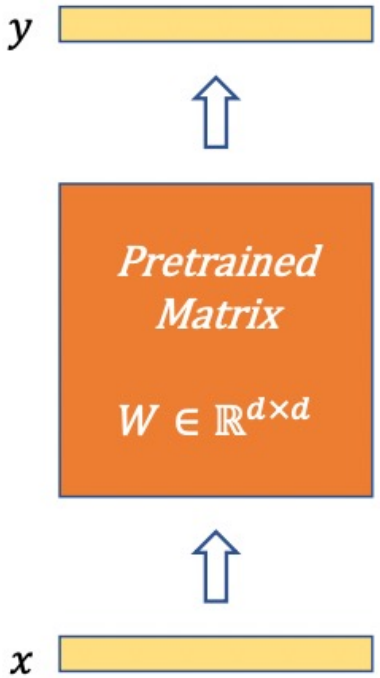
IRB No. H-1703-044-837

2.2 Foundation Models

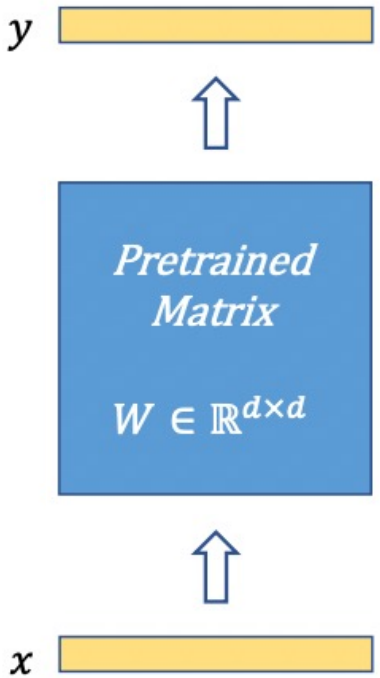
Model	Pretraining Data	SSL Strategy	Architecture	Strength
(Open) CLIP	LAION-400M (400 M image-text) LAION-2B English subset (2 B pairs)	Contrastive Learning	ViT Giant	A general-purpose vision encoder with strong multimodal understanding across vision-language tasks.
DINOv2	LVD-142M (≈142 M curated images)	Self Distillation	ViT Giant	Excels in transfer learning by providing robust global and patch-level visual representations.
RETFound	~1.6 M unlabelled retinal fundus + OCT images	Masked Auto Encoder (MAE)	ViT Large	Tailored for retinal images, demonstrating strong performance in both classification and prognosis.

2.3 Finetuning Methods

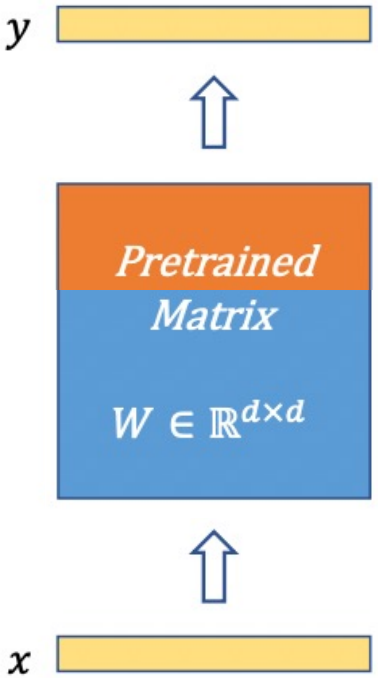
Freeze ❄️ Trainable 🔥



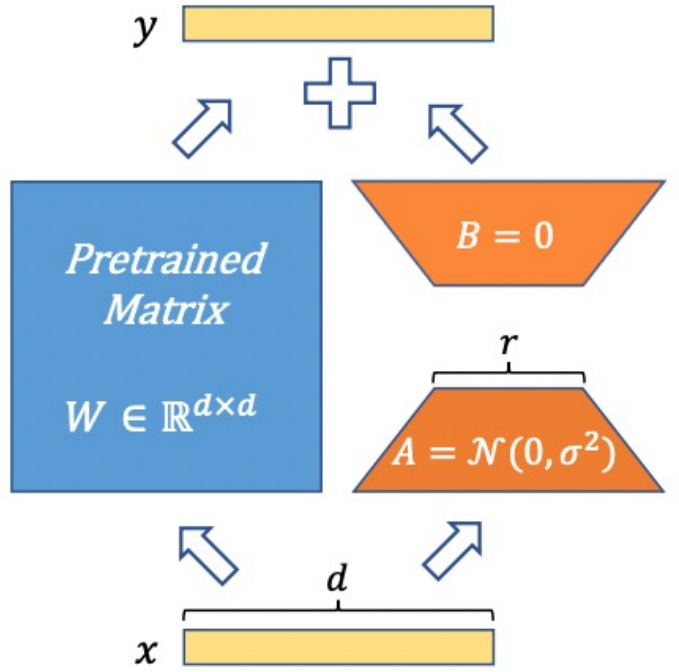
Full-finetuning



Classifier only



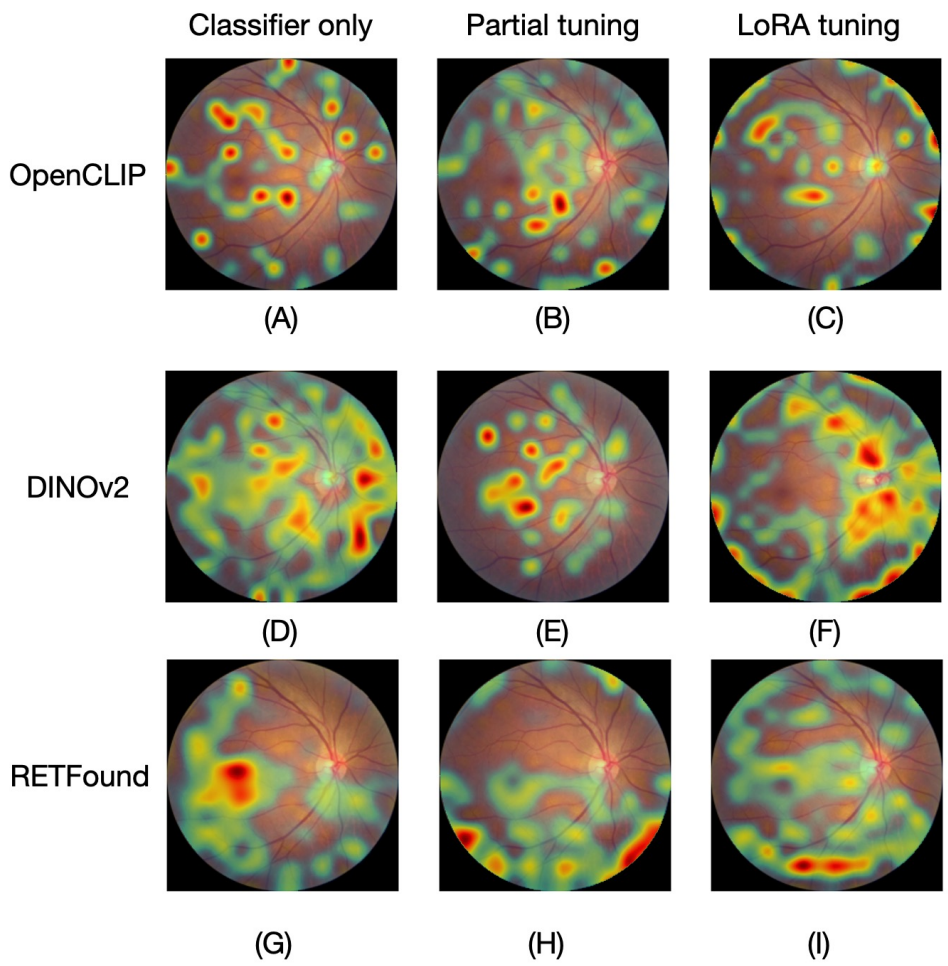
Partial finetuning



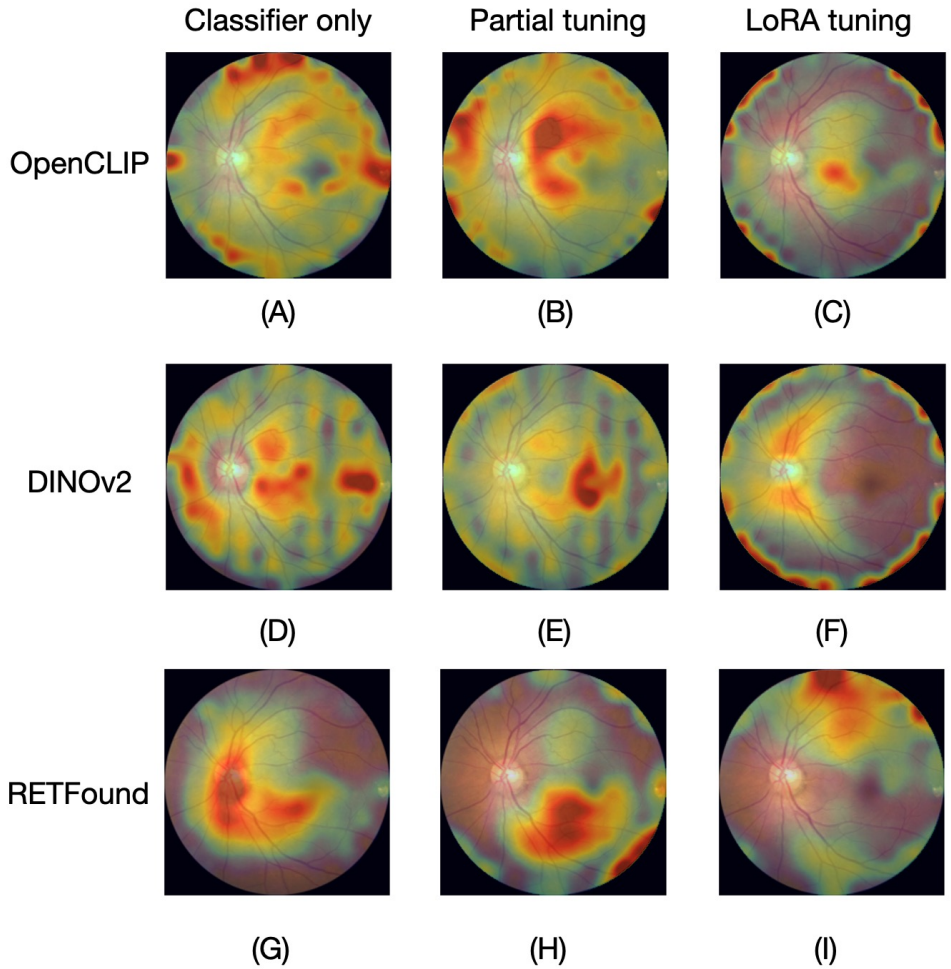
LoRA Tuning

2.4 Explainability

High Low

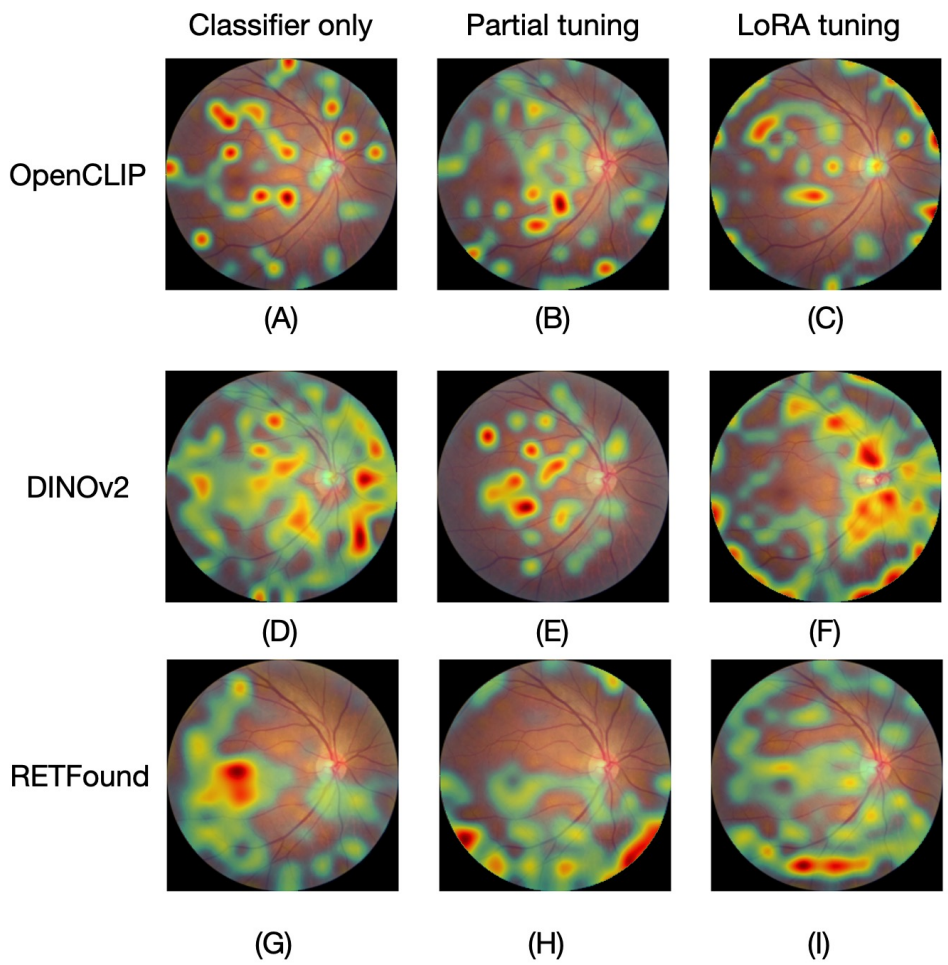


GradCAM for a single image



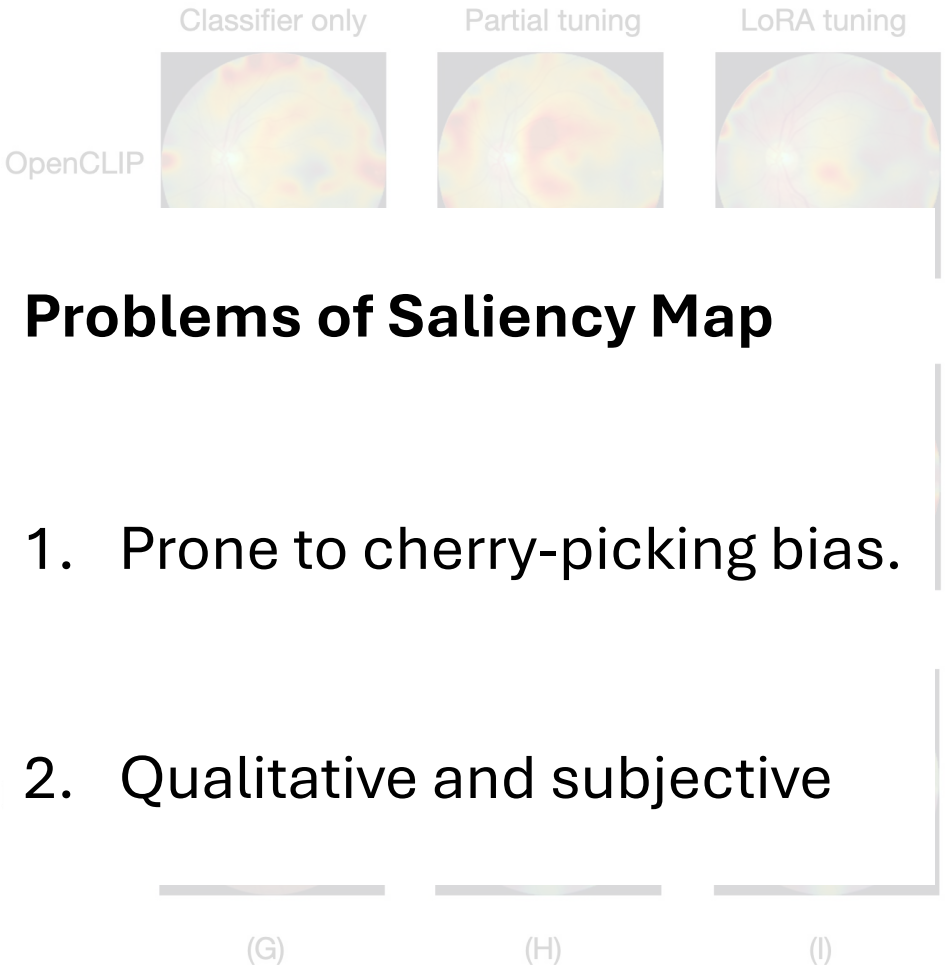
Averaged Grad-CAM across multiple images

2.4 Explainability



GradCAM for a single image

High — Low

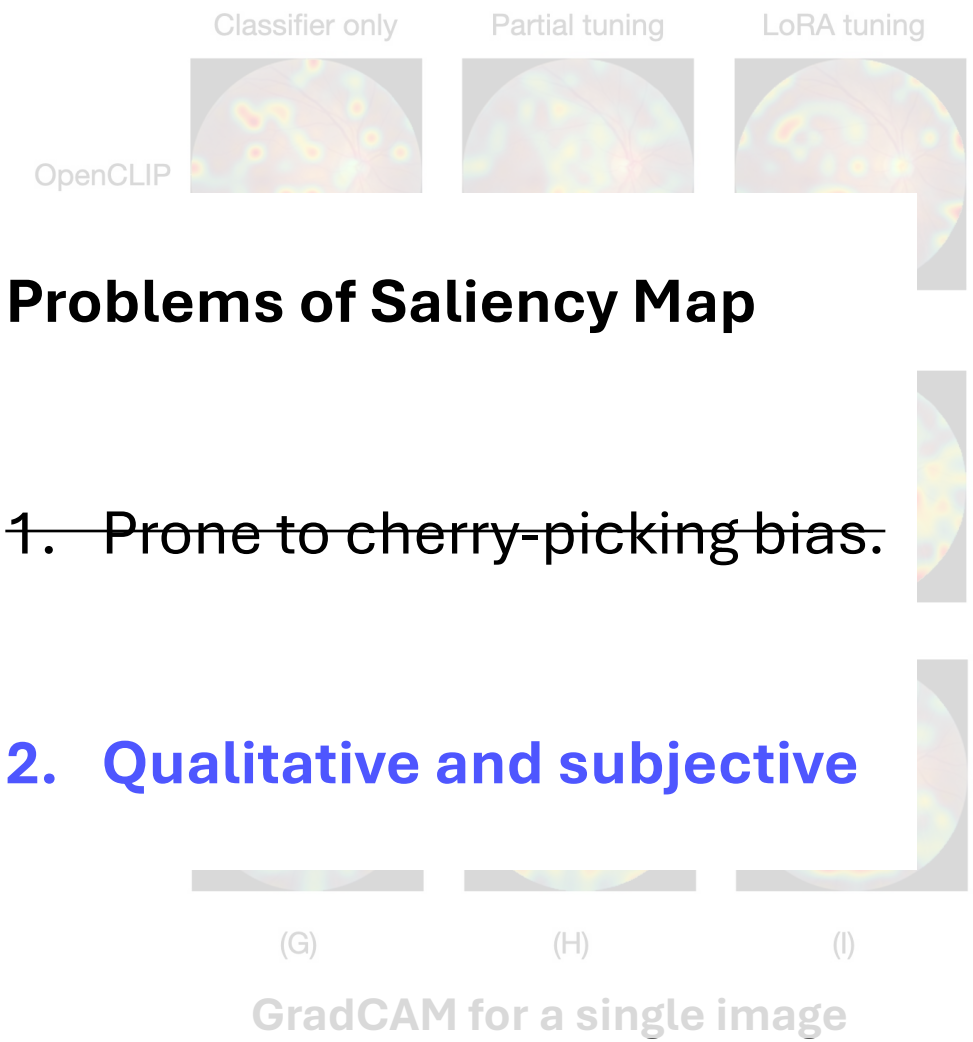


Averaged Grad-CAM across multiple images

Problems of Saliency Map

1. Prone to cherry-picking bias.
2. Qualitative and subjective

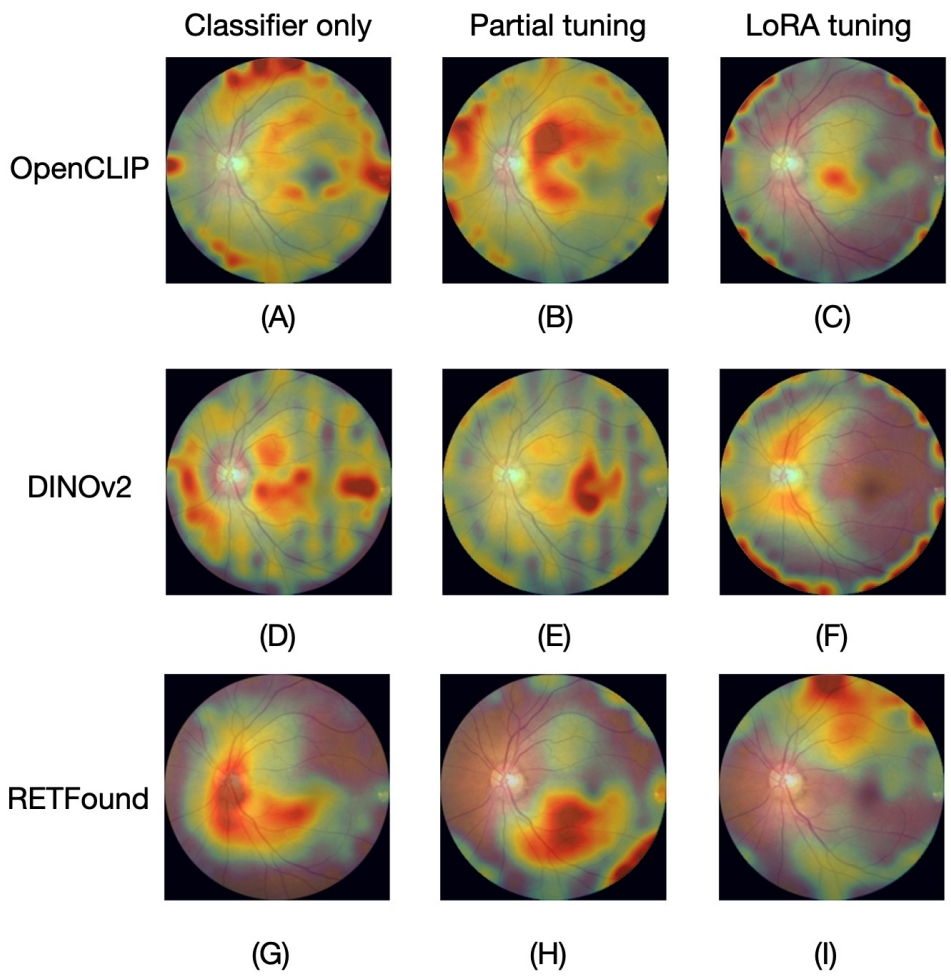
2.4 Explainability



GradCAM for a single image

Problems of Saliency Map

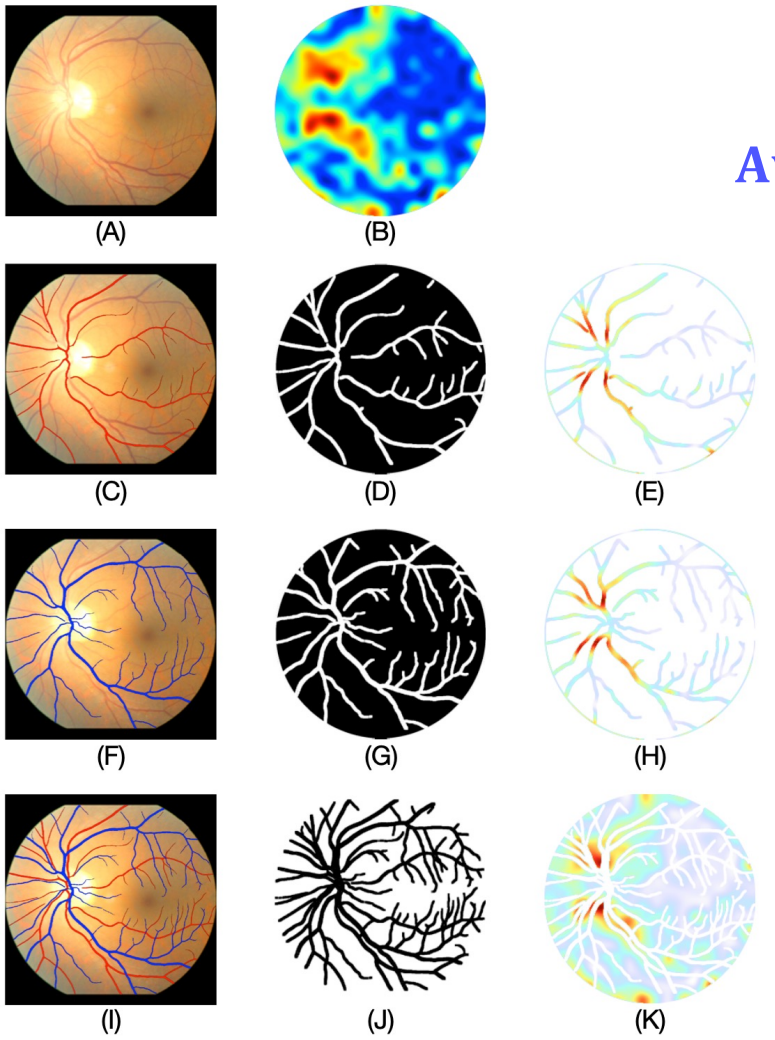
1. ~~Prone to cherry-picking bias.~~
2. **Qualitative and subjective**



Averaged Grad-CAM across multiple images

2.4 Explainability

High Low

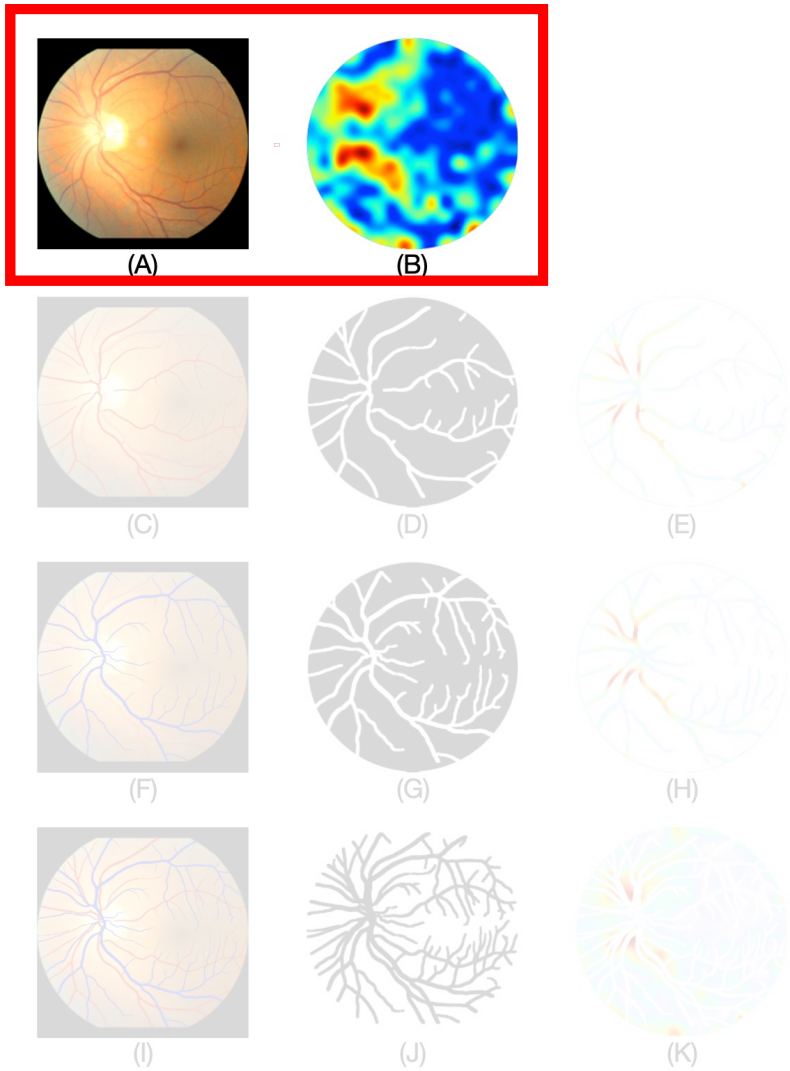


$$\text{Average saliency intensity for region of interest} = \frac{\sum_{i,j} G_{i,j} \cdot S_{i,j}}{\sum_{i,j} S_{i,j}}$$

Let Grad-CAM be denoted as $G_{i,j}$ and vessel segmentation image as $S_{i,j}$ where i, j represents the pixel coordinates of the 224×224 resolution map. For $S_{i,j}$ the predicted vessel pixel be $S_{i,j} = 1\ else\ S_{i,j} = 0$. The average saliency intensity predicting carotid atherosclerosis within a specific region can be calculated

2.4 Explainability

High Low



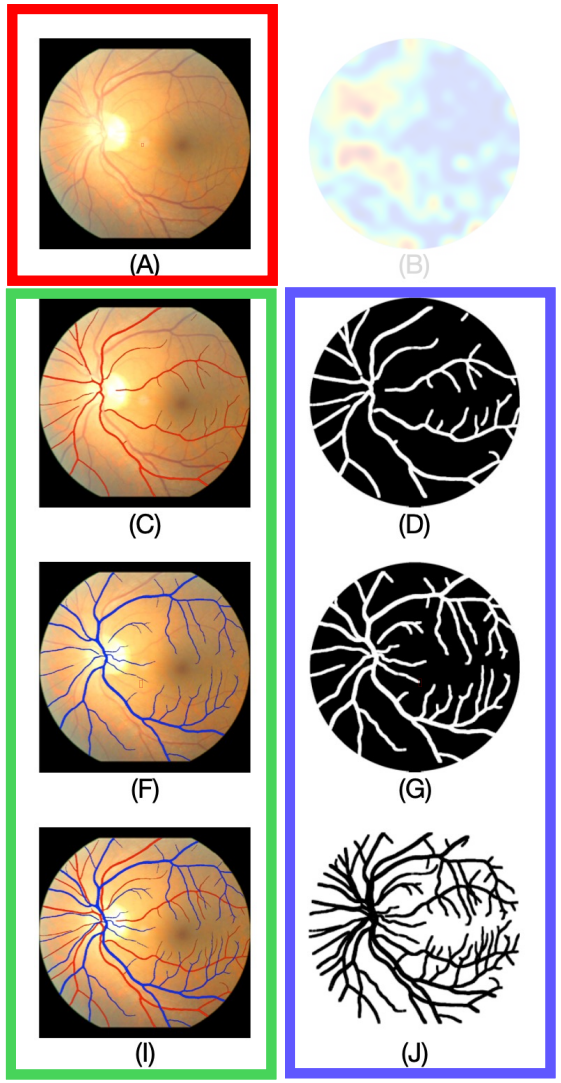
Step 1. Saliency Map

Using GradCAM, generate saliency map $G_{i,j}$

- (A)= Original Image
- (B) = GradCAM Saliency Map

2.4 Explainability

High Low



Step 2. Vessel Segmentation

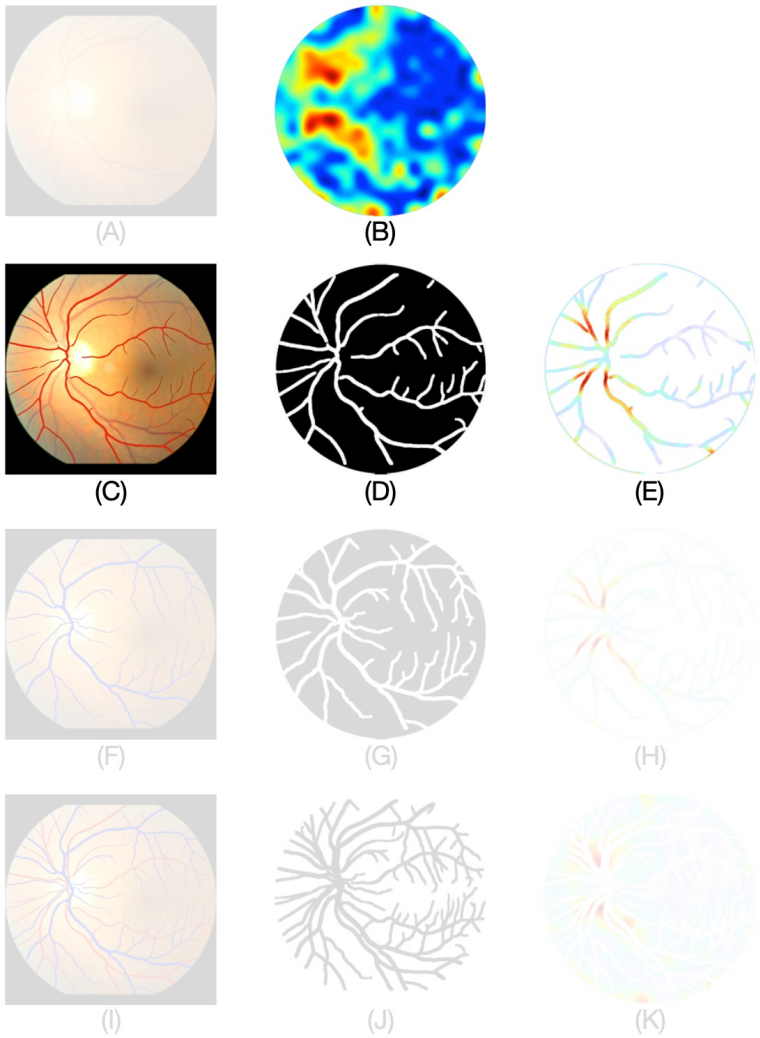
Using AutoMorph, generate vessel mask $S_{i,j}$

where, $S_{i,j} = 1\ for\ vessel\ else\ S_{i,j} = 0$

- (C, D) = Artery
- (F, G) = Vein
- (J) = (A) \ (CUF) = Non-Vessel regions

2.4 Explainability

High  Low



Step 3. Average Saliency Intensity

1. (B) ⊙ (D) => (E) : GradCAM for only artery
2. Sum(E) / Sum(D) : Average saliency intensity for artery
3. Repeat for [Vein, Vein + Artery, Non - Vessel]
4. Relative Ration : [Artery, Vein, Vein + Artery] / [Non-Vessel]

⊙ : Hadamard product (pixel-wise multiplication)

3. Results and Discussion

3.1 Performance Comparison

3.2 Explainability Comparison

3.3 Correlation between performance and explainability

3.1 Performance Comparison

Model	Architecture	Pretrained Domain	Finetuning Method	AUC (95% CI)	F1 Score (95% CI)
OpenCLIP	ViT Giant	Natural Image	Classifier Only	0.60 (0.57 - 0.63)	0.59 (0.56 - 0.62)
			Partial Finetuning	0.70 (0.67 - 0.72)	0.60 (0.57 - 0.63)
			LoRA tuning	0.71 (0.69 - 0.73)	0.61 (0.58 - 0.64)
DINOv2	ViT Giant	Natural Image	Classifier Only	0.66 (0.63 - 0.69)	0.58 (0.56 - 0.61)
			Partial Finetuning	0.66 (0.63 - 0.69)	0.59 (0.56 - 0.62)
			LoRA tuning	**0.71 (0.69 - 0.73)**	**0.62 (0.59 - 0.65)**
RETFound	ViT Large	Retinal Image	Classifier Only	0.64 (0.61 - 0.67)	0.59 (0.56 - 0.61)
			Partial Finetuning	0.70 (0.67 - 0.73)	0.61 (0.58 - 0.63)
			LoRA tuning	0.69 (0.66 - 0.72)	0.62 (0.59 - 0.64)

Discussion Point 1:
LoRA consistently improves performance across various foundation models

Discussion Point 2:
RETFound, despite being pretrained on retinal images, does not always yield optimal performance

3.2 Explainability Comparison

Model	Finetuning Method	Relative ratio of saliency intensity (95% CI)			
		Non-vessel	Artery	Vein	Vessel (Artery + Vein)
OpenCLIP	Classifier Only	1.00 (reference)	1.11 (1.10 - 1.12)	1.07 (1.06 - 1.07)	1.09 (1.09 - 1.10)
	Partial Finetuning	1.00 (reference)	1.32 (1.31 - 1.33)	1.28 (1.27 - 1.29)	1.31 (1.30 - 1.32)
	LoRA tuning	1.00 (reference)	1.24 (1.23 - 1.25)	1.20 (1.19 - 1.21)	1.22 (1.21 - 1.23)
DINOv2	Classifier Only	1.00 (reference)	1.31 (1.31 - 1.32)	1.30 (1.29 - 1.31)	1.31 (1.31 - 1.32)
	Partial Finetuning	1.00 (reference)	1.30 (1.29 - 1.32)	1.26 (1.25 - 1.28)	1.28 (1.26 - 1.29)
	LoRA tuning	1.00 (reference)	1.47 (1.46 - 1.48)	1.37 (1.36 - 1.38)	1.41 (1.40 - 1.42)
RETFound	Classifier Only	1.00 (reference)	**1.50 (1.49 - 1.52)**	**1.46 (1.44 - 1.47)**	**1.46 (1.45 - 1.48)**
	Partial Finetuning	1.00 (reference)	1.27 (1.26 - 1.28)	1.24 (1.23 - 1.25)	1.26 (1.25 - 1.27)
	LoRA tuning	1.00 (reference)	1.34 (1.33 - 1.35)	1.26 (1.25 - 1.27)	1.30 (1.29 - 1.31)

The relative ratio of saliency intensity quantifies how much more the Grad-CAM focuses on vessel regions compared to non-vessel regions.

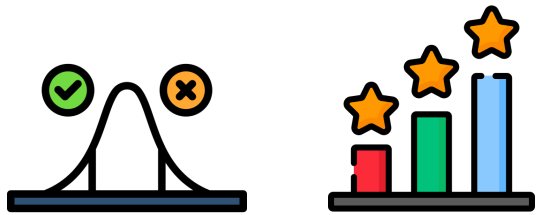
Discussion Point 3:
Although RETFound with classifier-only fine-tuning showed suboptimal performance, it demonstrated the highest level of explainability.

3.3 Correlation between performance and explainability

Discussion Point 4: The explainability ranking did not align with model performance.

Model	Architecture	Pretrained Domain	Finetuning Method	AUC (95% CI)	F1 Score (95% CI)
OpenCLIP	ViT Giant	Natural Image	Classifier Only	0.60 (0.57 - 0.63)	0.59 (0.56 - 0.62)
			Partial Finetuning	0.70 (0.67 - 0.72)	0.60 (0.57 - 0.63)
			LoRA tuning	0.71 (0.69 - 0.73)	0.61 (0.58 - 0.64)
DINOv2	ViT Giant	Natural Image	Classifier Only	0.66 (0.63 - 0.69)	0.58 (0.56 - 0.61)
			Partial Finetuning	0.66 (0.63 - 0.69)	0.59 (0.56 - 0.62)
			LoRA tuning	**0.71 (0.69 - 0.73)**	**0.62 (0.59 - 0.65)**
RETFound	ViT Large	Retinal Image	Classifier Only	0.64 (0.61 - 0.67)	0.59 (0.56 - 0.61)
			Partial Finetuning	0.70 (0.67 - 0.73)	0.61 (0.58 - 0.63)
			LoRA tuning	0.69 (0.66 - 0.72)	0.62 (0.59 - 0.64)

Model	Finetuning Method	Relative ratio of saliency intensity (95% CI)			
		Non-vessel	Artery	Vein	Vessel (Artery + Vein)
OpenCLIP	Classifier Only	1.00 (reference)	1.11 (1.10 - 1.12)	1.07 (1.06 - 1.07)	1.09 (1.09 - 1.10)
	Partial Finetuning	1.00 (reference)	1.32 (1.31 - 1.33)	1.28 (1.27 - 1.29)	1.31 (1.30 - 1.32)
	LoRA tuning	1.00 (reference)	1.24 (1.23 - 1.25)	1.20 (1.19 - 1.21)	1.22 (1.21 - 1.23)
DINOv2	Classifier Only	1.00 (reference)	1.31 (1.31 - 1.32)	1.30 (1.29 - 1.31)	1.31 (1.31 - 1.32)
	Partial Finetuning	1.00 (reference)	1.30 (1.29 - 1.32)	1.26 (1.25 - 1.28)	1.28 (1.26 - 1.29)
	LoRA tuning	1.00 (reference)	1.47 (1.46 - 1.48)	1.37 (1.36 - 1.38)	1.41 (1.40 - 1.42)
RETFound	Classifier Only	1.00 (reference)	**1.50 (1.49 - 1.52)**	**1.46 (1.44 - 1.47)**	**1.46 (1.45 - 1.48)**
	Partial Finetuning	1.00 (reference)	1.27 (1.26 - 1.28)	1.24 (1.23 - 1.25)	1.26 (1.25 - 1.27)
	LoRA tuning	1.00 (reference)	1.34 (1.33 - 1.35)	1.26 (1.25 - 1.27)	1.30 (1.29 - 1.31)



Spearman's rank Test

correlation coefficient
p = 0.343

Conclusion

In this study

1. Among all models, **DINOv2 with LoRA tuning** achieved the best classification performance for atherosclerosis prediction using retinal fundus images.

2. A new explainability metric, **average saliency intensity** for regions of interest, was proposed to quantitatively assess how much Grad-CAM focuses on clinically relevant vessel regions.

3. **Model performance and explainability did not always align.**
RETFound with classifier-only fine-tuning showed the highest explainability despite lower performance, emphasizing the need for multi-dimensional evaluation.

Thank you for listening

2025년도 1학기

[의료 데이터 기반 인공지능과 기계학습의 기초]

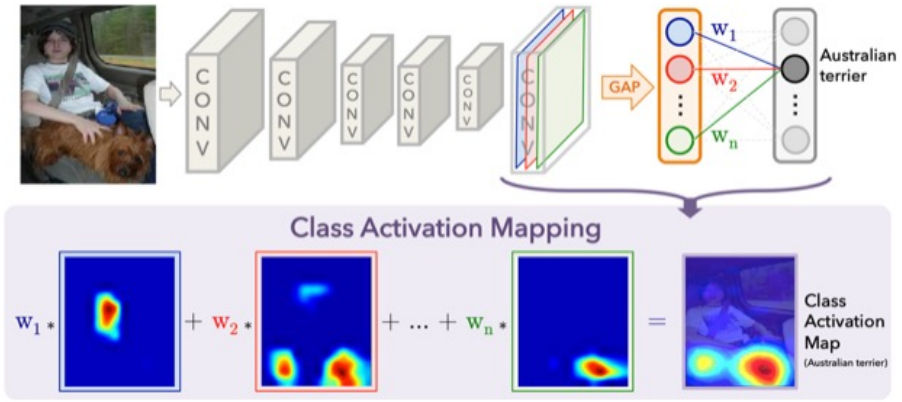
기말고사 프로젝트

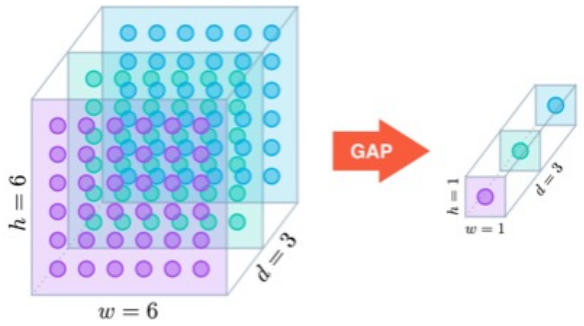
서울대학교 의과학과

2023-24885 이설하

2023-24526 이혁종

Appendix





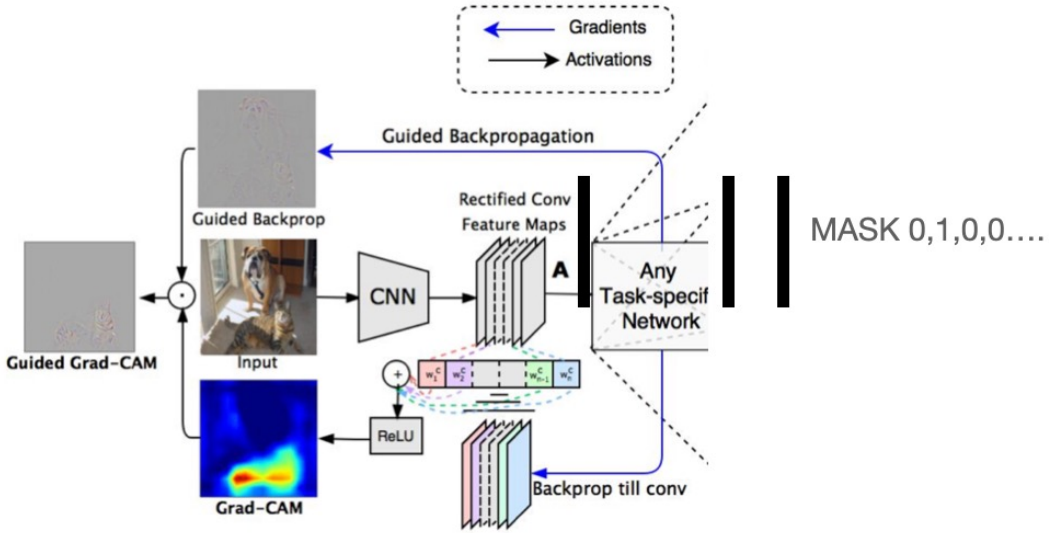
- Global Average Pooling - average all h x w

- CAM - Class Activation Maps
- $f_k(x, y)$: activation of unit k in the last convolution layer at spatial location (x, y)
- $F^k = \sum_{x,y} f_k(x, y)$: result of performing GAP. $S_c = \sum_x w_k^c F^k$ (assume that no bias).

$$S_c = \sum_k w_k^c \sum_{x,y} f_k(x, y) = \sum_{x,y} \sum_k w_k^c f_k(x, y). \quad (1) \qquad M_c(x, y) = \sum_k w_k^c f_k(x, y). \quad (2)$$

- That is, $M_c(x, y)$ directly indicates the importance of the activation at spatial grid (x, y) leading to the class c.
- Limitation : GAP

Appendix



$$\alpha_k^c = \overbrace{\frac{1}{Z}\sum_i\sum_j}^{\text{global average pooling}} \underbrace{\frac{\partial y^c}{\partial A_{ij}^k}}_{\text{gradients via backprop}} \quad (1)$$

$$L^c_{\text{Grad-CAM}} = ReLU\underbrace{\left(\sum_k \alpha_k^c A^k\right)}_{\text{linear combination}} \quad (2)$$

- Since CAM needs w_k^c which represent the weight of the k^{th} activation map, there exist architectural limitations.
- Rather then re-training for the w_k^c, GradCAM compute it by gradients in post-hoc manner.
- That is $CAM = \sum_k w_k^c f_k(x,y) = \sum_k w_k^c A^k \text{ -> } GradCAM = ReLU\sum_k \alpha_k^c A^k$: $w_k^c \rightarrow \alpha_k^c$
- α_k^c indicates the 'weight (or importance) of A^k for the model decision regarding y^c'
- The shape of GradCAM equals to activation map shape.
  (for visualization on original input image, the GradCAM is resized (interpolation))