Evaluating Visual Explanations of Attention Maps for Transformer-based Medical Imaging

Minjae Chung, Jong Bum Won, Ganghyun Kim, Yujin Kim, Utku Ozbulak

GHENT UNIVERSITY

GHENT UNIVERSITY GLOBAL CAMPUS

Marrakesh, Morocco, 6 Oct 2024

Overview

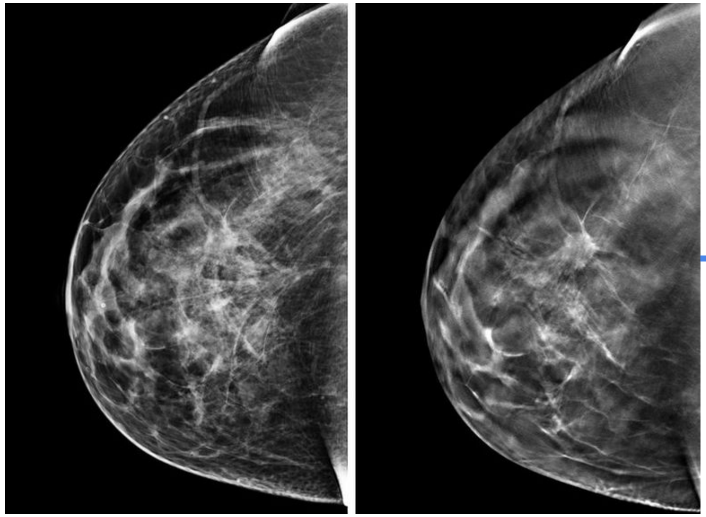
1. Interpretability for deep neural networks (classification)

2. CNN to ViT transition and what it means for interpretability

3. Experiments on ViT Interpretability for medical imaging

4. Lessons learned and open challenges

Interpretability in the context of medical image classification



DNN

There is a malignant tumor in this MRI

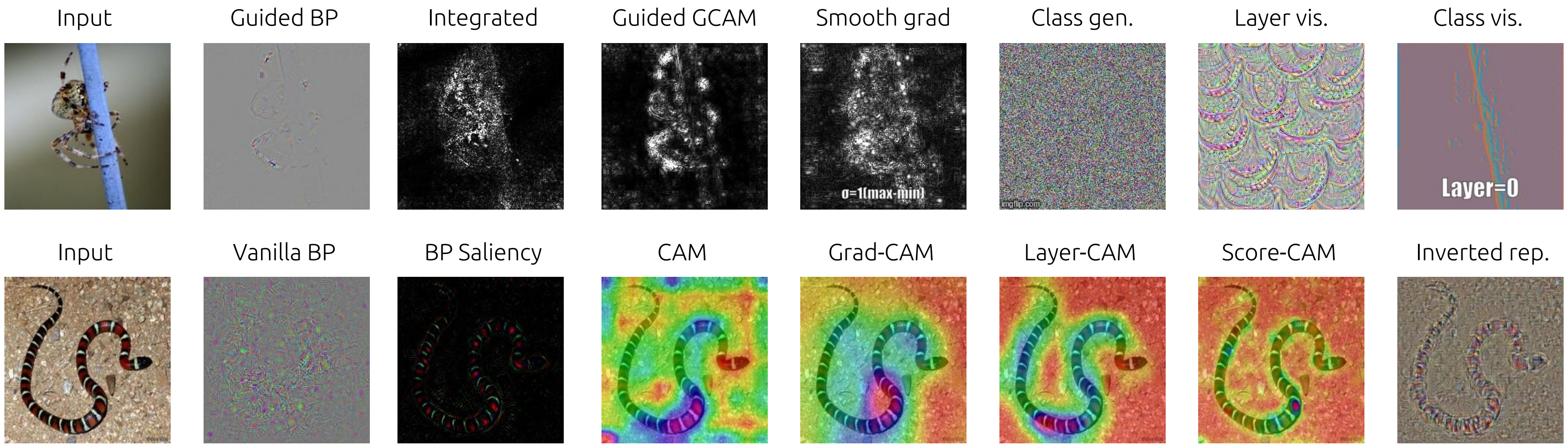
<u>The question we want answered</u>

Why did this model make this prediction? ⟷ Where is the tumor?

Interpretability methods

- There are too many interpretability methods

- Majority were proposed for CNNs

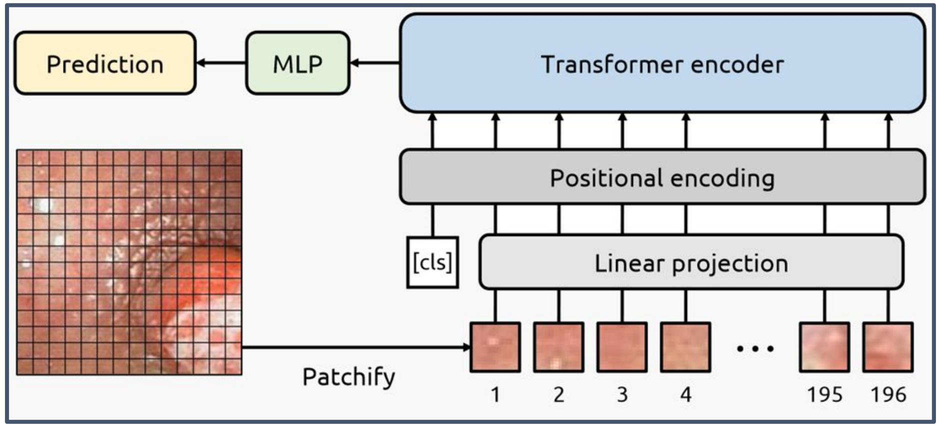


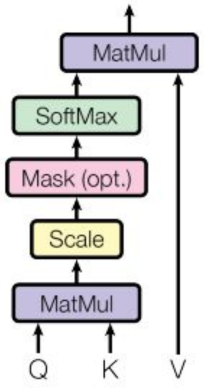
Source: https://github.com/utkuozbulak/pytorch-cnn-visualizations

Transformers and vision transformers

- A new DNN architecture that revolutionized the field

- ViTs[1] are replacing CNNs for many medical imaging problems





$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

[1] Dosovitskiy et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale

How to interpret ViTs?

- Use previously established methods

  >> GradCAM[2], Integrated Gradients[3], others

- Novel methods tailored for ViTs

  >> Attention maps, The Chefer method[4], others

[2] Selvaraju et al. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization
[3] Sundararajan et al. Axiomatic Attribution for Deep Networks
[4] Chefer et al. Transformer Interpretability Beyond Attention Visualization

ViT-specific interpretability: Attention maps vs others

- Attention maps:

  >> Good: Part of the decision-making process

  >> Bad: Only uses *q-k*, doesn't take *v* into account[4]

- The Chefer method:

  >> Good: Takes *v* into account for interpretability[4]

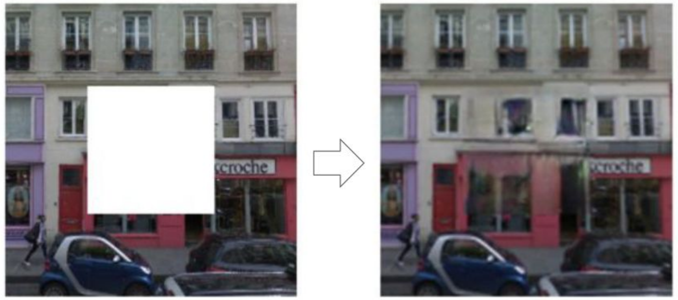
  >> Bad: Has additional calculations

[4] Chefer et al. Transformer Interpretability Beyond Attention Visualization

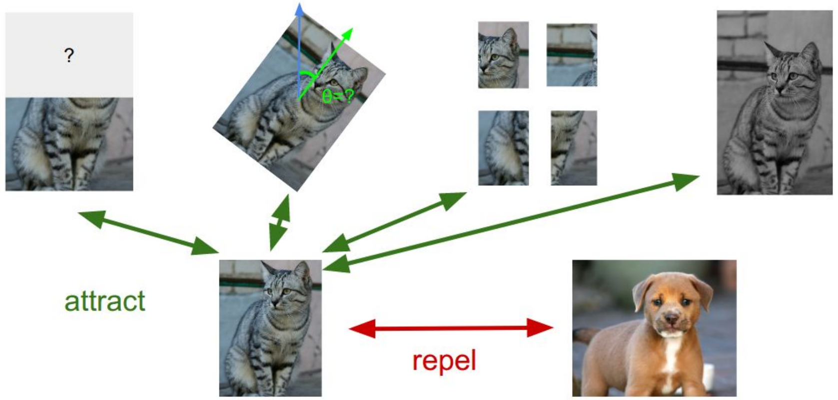
Self-supervised learning and interpretability

- Self-supervised pre-training affects the interpretability of attention maps[5,6]

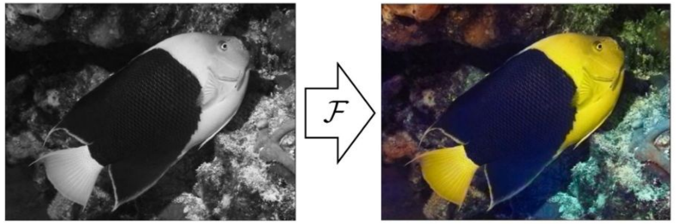
Inpainting



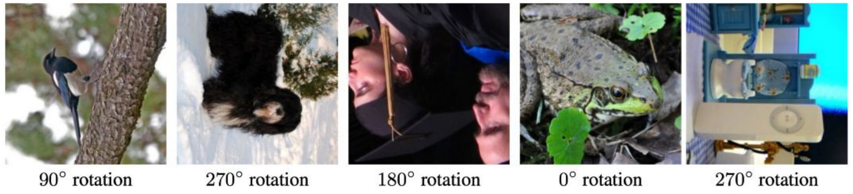
Contrastive learning



Colorization



Rotation prediction

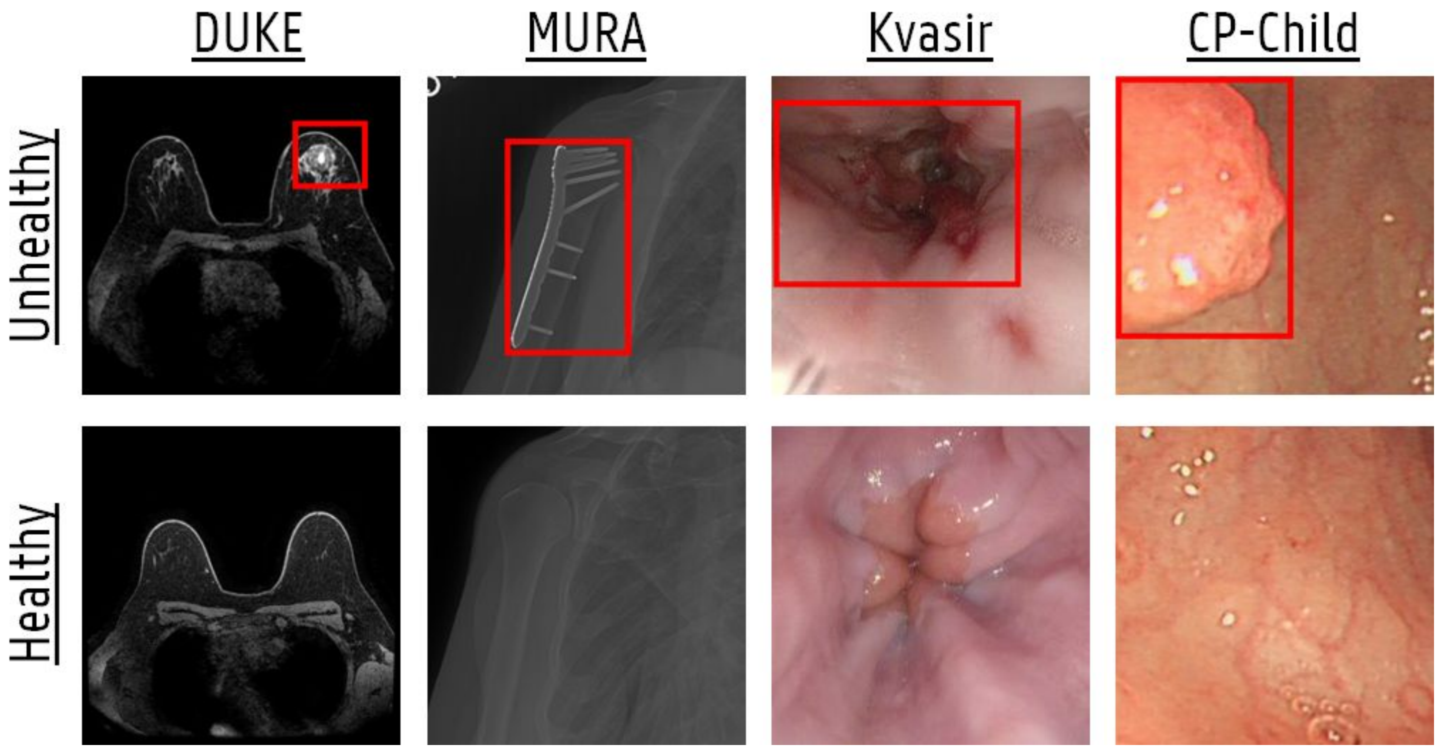


[5] Caron et al. Emerging Properties in Self-Supervised Vision Transformers
[6] Darcet et al. Vision Transformers Need Registers

Parameters for interpretability analysis - Task

1- Classification task

>> Several medical imaging datasets



Parameters for interpretability analysis - Models

2- ViT-B/16

>> Randomly initialized and (self-supervised) pre-trained

a- Randomly initialized

b- Supervised pretrained

c- Distillation with no labels (DINO)

d- Masked autoencoder (MAE)

Parameters for interpretability analysis - Interpretability

3- Interpretability methods

>>Transformer-specific and previously established methods

a- GradCAM

b- Attention maps

c- The Chefer method

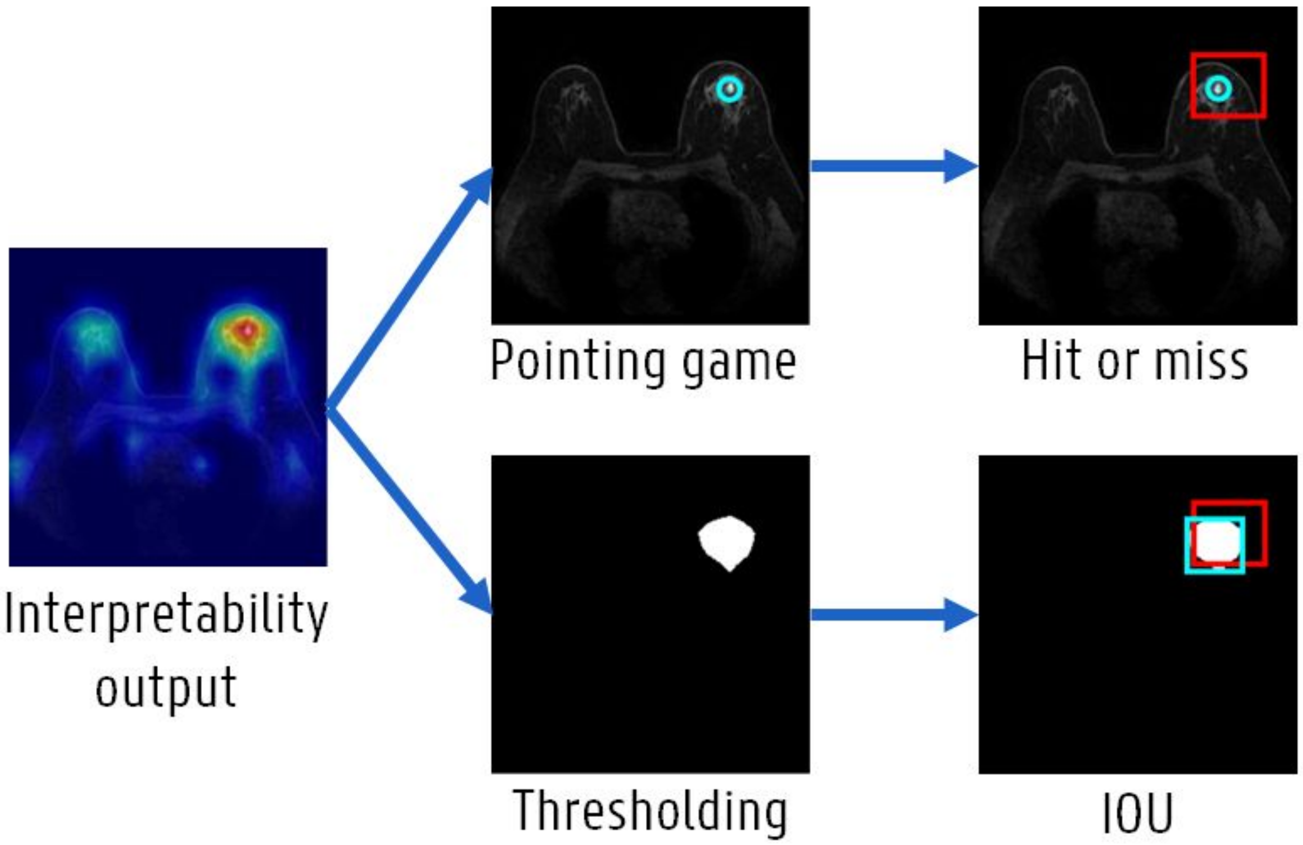
Experimental process

1- Train ViTs on medical classification tasks

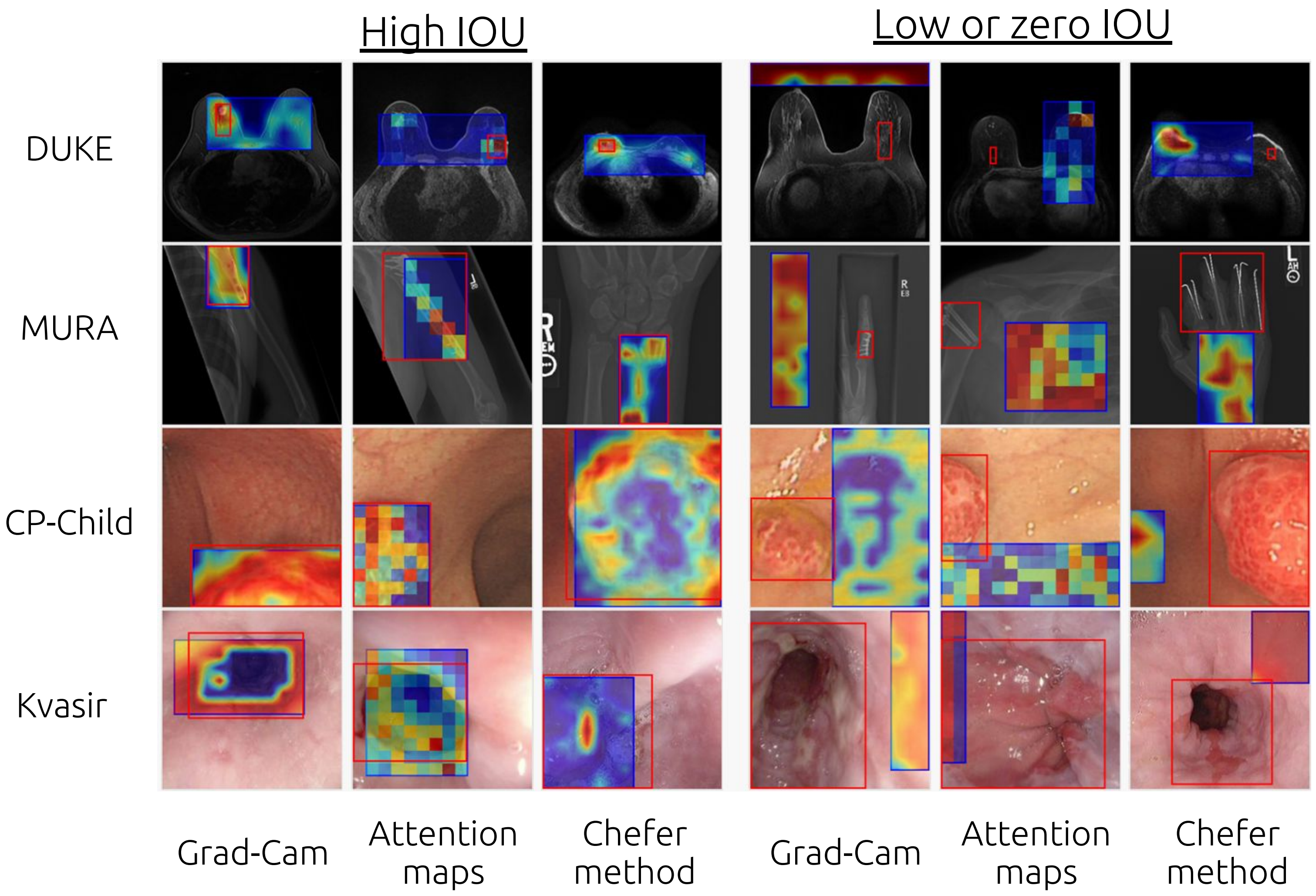
2- Extract interpretability maps

3- Evaluate interpretability results with regions of interest

Interpretability evaluation



Qualitative evaluation can be misleading



Quantitative Results

Dataset	Model init.	Pointing game			IoU		
		GradCAM	Attention	Chefer	GradCAM	Attention	Chefer
CP-Child	Random	0.64	**0.96**	0.54	0.36	<u>0.45</u>	0.33
	Supervised	0.38	0.36	**0.54**	0.27	0.42	<u>0.55</u>
	DINO	0.76	0.68	**0.86**	0.38	0.54	<u>0.63</u>
	MAE	0.56	0.96	**0.98**	0.41	0.67	<u>0.73</u>
DUKE	Random	0.04	**0.18**	0.12	0.03	<u>0.04</u>	0.03
	Supervised	0.00	**0.30**	0.28	0.00	0.04	<u>0.08</u>
	DINO	0.10	**0.52**	0.40	0.03	0.07	<u>0.08</u>
	MAE	0.12	**0.36**	**0.36**	0.01	0.07	<u>0.08</u>
Kvasir	Random	0.88	**0.98**	0.72	<u>0.54</u>	0.36	0.34
	Supervised	0.80	0.88	**0.92**	0.44	0.59	<u>0.61</u>
	DINO	0.72	**0.96**	0.90	0.32	0.43	<u>0.48</u>
	MAE	0.52	**0.82**	0.80	0.52	0.59	<u>0.62</u>
MURA	Random	0.30	**0.40**	0.34	0.18	0.19	<u>0.20</u>
	Supervised	0.56	**0.94**	**0.94**	0.36	0.47	<u>0.56</u>
	DINO	0.78	**0.90**	0.86	0.35	0.41	<u>0.52</u>
	MAE	0.30	**0.76**	0.68	0.18	0.35	<u>0.39</u>

Quantitative Results

Dataset	Model init.	Pointing game			IoU		
		GradCAM	Attention	Chefer	GradCAM	Attention	Chefer
CP-Child	Random	0.64	**0.96**	0.54	0.36	0.45	0.33
	Supervised	0.38	0.36	**0.54**	0.27	0.42	0.55
	DINO	0.76	0.68	**0.86**	0.38	0.54	0.63
	MAE	0.56	0.96	**0.98**	0.41	0.67	0.73
DUKE	Random	0.04	**0.18**	0.12	0.03	0.04	0.03
	Supervised	0.00	**0.30**	0.28	0.00	0.04	0.08
	DINO	0.10	**0.52**	0.40	0.03	0.07	0.08
	MAE	0.12	**0.36**	**0.36**	0.01	0.07	0.08
Kvasir	Random	0.88	**0.98**	0.72	0.54	0.36	0.34
	Supervised	0.80	0.88	**0.92**	0.44	0.59	0.61
	DINO	0.72	**0.96**	0.90	0.32	0.43	0.48
	MAE	0.52	**0.82**	0.80	0.52	0.59	0.62
MURA	Random	0.30	**0.40**	0.34	0.18	0.19	0.20
	Supervised	0.56	**0.94**	**0.94**	0.36	0.47	0.56
	DINO	0.78	**0.90**	0.86	0.35	0.41	0.52
	MAE	0.30	**0.76**	0.68	0.18	0.35	0.39

Takeaway Messages

1- It's easy to cherry pick good (and bad) interpretability results

2- Pre-training has (some) influence over the interpretability outcome

3- GradCAM interpretability is (generally) inadequate for ViTs

4- Attention maps show promise

Lessons learned and open challenges

- Bounding box evaluation for medical interpretability is not adequate

>> Use segmentation masks? Would it make a difference?

Lessons learned and open challenges

- Evaluation metrics are heavily influenced by the region of interest

>> Token-based evaluation? Would it be better?

Dataset	Model init.	Pointing game			IoU		
		GradCAM	Attention	Chefer	GradCAM	Attention	Chefer
DUKE	Random	0.04	**0.18**	0.12	0.03	0.04	0.03
	Supervised	0.00	**0.30**	0.28	0.00	0.04	0.08
	DINO	0.10	**0.52**	0.40	0.03	0.07	0.08
	MAE	0.12	**0.36**	**0.36**	0.01	0.07	0.08
Kvasir	Random	0.88	**0.98**	0.72	0.54	0.36	0.34
	Supervised	0.80	0.88	**0.92**	0.44	0.59	0.61
	DINO	0.72	**0.96**	0.90	0.32	0.43	0.48
	MAE	0.52	**0.82**	0.80	0.52	0.59	0.62

Lessons learned and open challenges

- Evaluation with ground-truth can be misleading

>> Take into account spurious correlations? How?

(a) Husky classified as wolf (b) Explanation

Too many parameters influence interpretability

- Architecture
- Model pretraining
- Model fine-tuning
- Dataset bias
- Spurious correlations
- Size of the regions of interest
- Interpretability evaluation
- … and more?

Too many parameters influence interpretability

- Architecture
- Model pretraining
- Model fine-tuning
- Dataset bias
- Spurious correlations
- Size of the regions of interest
- Interpretability evaluation
- … and more?

- Training optimizer
- Normalization
- Training batch size
- Operating system
- Air quality
- iMIMIC acceptance
- NVDA stock price
- … others?