

# M.Sc. in AI - Thesis Presentation

## Grad-CAM vs HiResCAM: A comparative study via quantitative evaluation metrics

June 6, 2023

# Contents

- ① XAI algorithms & Motivation
- ② Quantitative evaluation metrics
- ③ Datasets
- ④ Experimental results & Discussion

# 1. XAI algorithms & Motivation

For a CNN we denote by  $\{A^f\}_{f=1,\dots,F}$  a convolutional layer,  $A^f \in \mathbb{R}^{D_1 \times D_2}$

## Definition

For class  $m = 1, 2, \dots, M$ ,

- the Grad-CAM (2019) attribution map w.r.t.  $\{A^f\}_{f=1,\dots,F}$  is given by

$$\mathcal{A}_m^{\text{Grad-CAM}} = \text{ReLU} \left( \sum_{f=1}^F a_m^f A^f \right)$$

$$\text{where } a_m^f = \frac{1}{D_1 D_2} \sum_{i=1}^{D_1} \sum_{j=1}^{D_2} \frac{\partial s^m}{\partial A_{ij}^f} \quad (\text{Gradient Averaging step})$$

- the HiResCAM (2021) attribution map w.r.t.  $\{A^f\}_{f=1,\dots,F}$  is given by

$$\mathcal{A}_m^{\text{HiResCAM}} = \text{ReLU} \left( \sum_{f=1}^F \frac{\partial s^m}{\partial A^f} \odot A^f \right)$$

where  $\odot$  stands for the Hadamard product

- Setting? (CNN structure? Gradients wrt which Conv layer?)

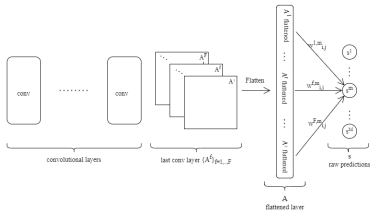
## Definition (Faithfulness)

For the purposes of this study, *an attribution map method is faithful to a CNN model* if the sum of the attribution map values reflects the class score calculation.

**Table:** Grad-CAM vs HiResCAM - Theory summary

CNN structure	Gradients wrt last Conv layer	Grad-CAM vs HiResCAM
Conv - GAP - Class Scores	yes no	Equivalent & Faithful Not equivalent & Not faithful
Conv - Flatten - Class Scores	yes no	Not equivalent. Only HiResCAM is faithful Not equivalent & Not faithful
Conv - GAP/Flatten - Dense - Class Scores	yes no	Not equivalent & Not faithful

The *setting* of this study:



**Figure:** CNN ending in one fully connected layer

⇒ If gradients wrt last conv layer, one can show that HiResCAM is faithful to the model:

$$s^m = \sum_{i,j} \left( \tilde{\mathcal{A}}_m^{\text{HiResCAM}} \right)_{i,j} + b^m$$

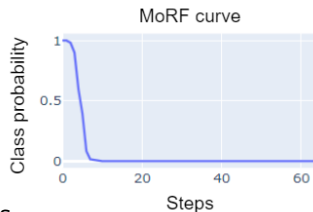
⇒ On the other hand, this is not true for Grad-CAM!

- *Goal:* In this *unique* setting, we want to quantify the quality of the Grad-CAM and HiResCAM attribution maps and examine *if faithfulness aligns with quantitative evaluation metrics results*.

## 2. Quantitative evaluation metrics

### AOPC score

- For image  $x$  and its heatmap
- $\mathcal{O}$ : heatmap regions in descending order
- Iteratively:
  - apply perturbations to most relevant regions
  - calculate difference in predicted class probabilities
- Results in *MoRF Perturbation Curve* (for image  $x$ )
- Idea: large Area Over MoRF curve
  - $\Rightarrow$  class probability decreases after *a few* steps
  - $\Rightarrow$  the heatmap can accumulate class info in *a few* regions



### Definition (AOPC, 2015)

$$AOPC = \frac{1}{L+1} \text{Avg} \left( \underbrace{\sum_{k=1}^L [f(x_{MoRF}^{(0)}) - f(x_{MoRF}^{(k)})]}_{\text{controls area over MoRF curve}} \right)_{\text{test set}}$$

## Max Sensitivity score

- Idea: Measures the degree to which the class explanation is affected by small perturbations in the test image.
- Naturally, we desire explanations with *low* sensitivity.

### Definition (Max Sensitivity, 2019)

For explanation method  $\Phi$  and radius  $r$ , we define the Max Sensitivity of explanation  $\Phi(f, x)$  as:

$$SENS_{MAX}(\Phi, f, x, r) = \max_{y: \|x-y\|_{\infty} \leq r} \|\Phi(f, x) - \Phi(f, y)\|_{Euclidean}$$

## HAAS score

- Idea: If an attribution map gives an accurate explanation then *tuning* the image pixels according to the attribution map could reduce the number of misclassifications.
- How to tune? Positive attribution values to emphasize pixel values.  
Negative attribution values to de-emphasize pixel values.

### Definition (HA image, HAAS, 2022)

- For image  $x$  and attributions  $a$  we compute the *Heatmap Assisted* (HA) image:

$$HA(x, a) = \max\{-1, \min\{1, x_{norm}(1 + a_{norm})\}\}$$

- $HAAS = \frac{\text{Accuracy over HA images}}{\text{Accuracy over original images}}$

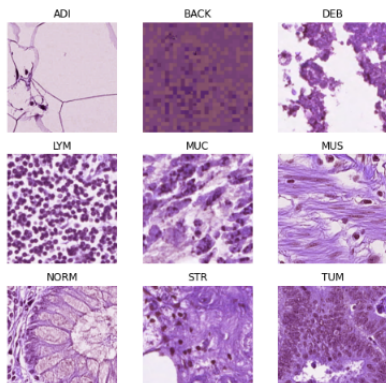
- Interpretation: if HAAS value above 1  
⇒ the HA images improve the model's accuracy  
⇒ the attribution maps explain well the pixels' importance



### 3. Datasets

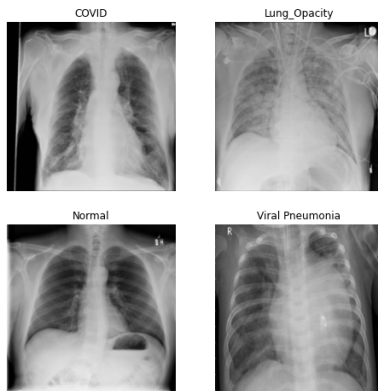
#### CRC Dataset

- Colon tissues (100K train)(7.1K test)
- ADI(10.4%), BACK(10.6%), DEB(11.5%), LYM(11.5%), MUC(8.9%), MUS(13.5%), NORM(8.8%), STR(10.5%), TUM(14.3%)



#### Covid-19 Radiography Database

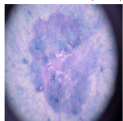
- X-Rays (21.1K)
- Covid(17%), Lung Opacity (28.4%), Normal (48.15%), Viral Pneumonia (6.35%)



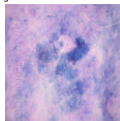
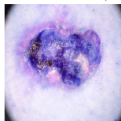
## HAM10000 Dataset

- Skin lesion images (10K)
- akiec(3.27%), bcc(5.13%), bkl(10.97%), df(1.15%), nv(66.95%), vasc(1.42%), mel(11.11%)

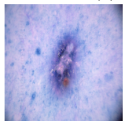
Actinic Keratoses (akiec)



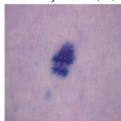
Basal Cell Carcinoma (bcc) Benign Keratosis-like Lesions (bkl)



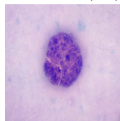
Dermatofibroma (df)



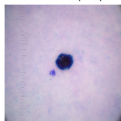
Melanocytic Nevi (nv)



Vascular lesions (vasc)



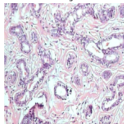
Melanoma (mel)



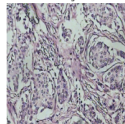
## BreakHis Dataset

- Breast tumor tissues (7.9K)
- Benign(31.4%), Malignant(68.6%)

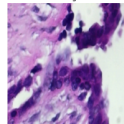
Benign



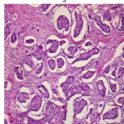
Benign



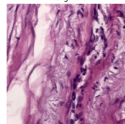
Benign



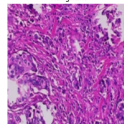
Malignant



Malignant



Malignant



**Remark:** These two datasets contain many examples of the same image at different scale. Train-Val-Test sets were constructed via an *image independent* approach.

## 4. Experimental results & Discussion

### Models

Per dataset:

- we trained one ResNet and one VGG19 model
- each customized to the *Conv - Flatten - Class scores* architecture (s.t. HiResCAM is faithful to the model when gradients are computed wrt last convolutional layer)

Table: Testing Results

	CRC		Covid-19		HAM10000		BreakHis	
	ResNet34	VGG19	ResNet34	VGG19	ResNet50	VGG19	ResNet50	VGG19
Bal. Accuracy	0.89	0.94	0.97	0.95	0.69	0.73	0.87	0.84
Mean AUC	0.993	0.997	0.995	0.992	0.934	0.938	0.942	0.932

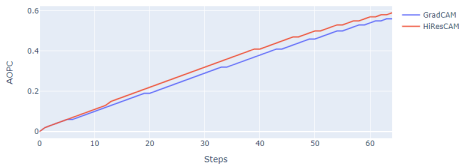
## AOPC results

- Per dataset, model and attribution method, we calculated AOPC score at heatmap sizes:  $4 \times 4$ ,  $8 \times 8$ ,  $11 \times 11$  and  $14 \times 14$ , utilizing uniform random noise.

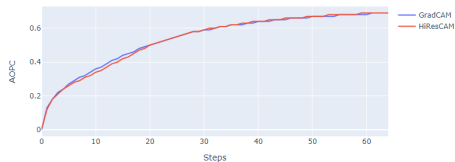
Table: AOPC Scores

			CRC		Covid-19		HAM10000		BreakHis	
			ResNet34	VGG19	ResNet34	VGG19	ResNet50	VGG19	ResNet50	VGG19
Heatmap size	$4 \times 4$	Grad-CAM	0.57	0.5	0.71	0.64	0.57	0.35	0.32	0.35
		HiResCAM	0.57	<b>0.52</b>	<b>0.73</b>	<b>0.65</b>	<b>0.58</b>	<b>0.37</b>	<b>0.34</b>	<b>0.42</b>
	$8 \times 8$	Grad-CAM	0.59	0.56	0.72	0.69	0.63	0.36	0.27	0.35
		HiResCAM	0.59	<b>0.59</b>	<b>0.73</b>	0.69	<b>0.64</b>	<b>0.37</b>	<b>0.28</b>	<b>0.45</b>
	$11 \times 11$	Grad-CAM	0.6	0.59	0.73	0.71	0.65	0.37	<b>0.28</b>	0.35
		HiResCAM	0.6	<b>0.62</b>	<b>0.74</b>	0.71	<b>0.67</b>	<b>0.39</b>	0.27	<b>0.45</b>
	$14 \times 14$	Grad-CAM	0.58	0.6	0.73	<b>0.7</b>	0.66	0.36	0.27	0.35
		HiResCAM	<b>0.6</b>	<b>0.62</b>	<b>0.74</b>	0.69	<b>0.68</b>	0.36	0.27	<b>0.48</b>

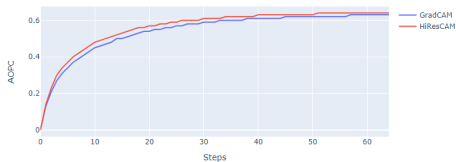
- At heatmap size  $4 \times 4$  HiResCAM prevails 7/8 cases. As size increases, the effect fades out.
- For each model and over all heatmap sizes, HiResCAM prevails 7/8 cases.



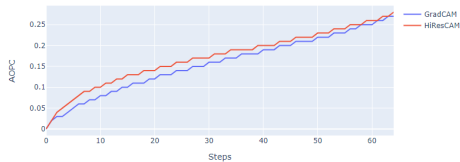
(a) CRC - VGG19



(b) Covid-19 - VGG19



(c) HAM10000 - ResNet50



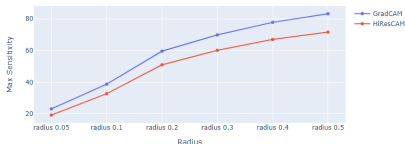
(d) BreakHis - ResNet50

Figure: AOPC Graphs for Heatmaps 8\*8 size

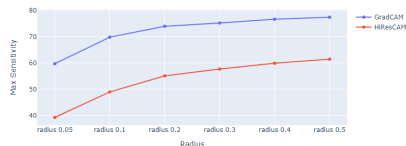
# Max Sensitivity results

Table: Experiment configurations

Radius (r)	No. of samples (y)
0.05	20
0.1	20
0.2	30
0.3	30
0.4	40
0.5	40



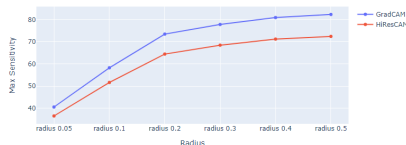
(a) CRC - VGG19



(b) Covid-19 - VGG19



(c) HAM10000 - ResNet50



(d) BreakHis - ResNet50

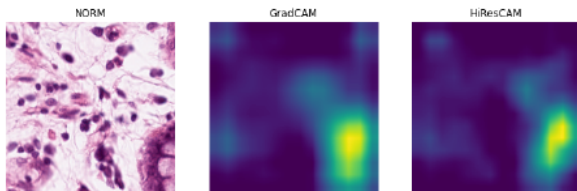
- Overall, HiResCAM prevails 8/8 cases.

## Why AOPC and Max Sensitivity favor HiResCAM?

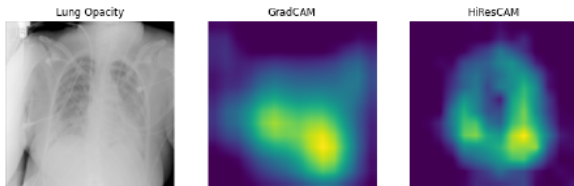
- Grad-CAM and HiResCAM treat gradients in a different way:

$$\left( \frac{1}{D_1 D_2} \sum_{i=1}^{D_1} \sum_{j=1}^{D_2} \frac{\partial s^m}{\partial A_{ij}^f} \right) A^f \quad \text{vs} \quad \frac{\partial s^m}{\partial A^f} \odot A^f$$

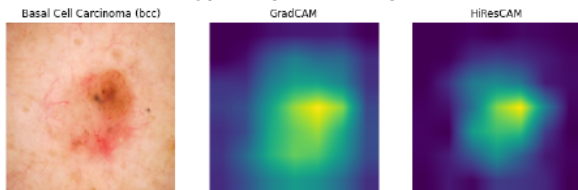
Grad-CAM	HiResCAM
Gradient values and signs are camouflaged into the average	Preserves gradient effect on pixel level and utilizes both value and sign
Larger, smoother, less-detailed attention areas	High-resolution maps with precise class localization



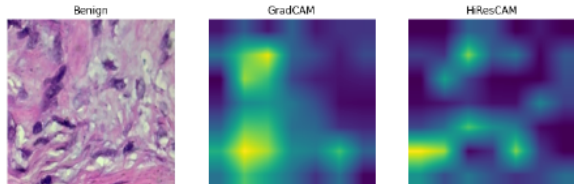
(a) CRC dataset - VGG19



**(b) Covid-19 dataset - VGG19**



**(c) HAM10000 dataset - ResNet50**



**(d) BreakHis dataset - ResNet50**



## HAAS results and discussion

Table: HAAS Scores (Medical datasets)

	CRC		Covid-19		HAM10000		BreakHis	
	ResNet34	VGG19	ResNet34	VGG19	ResNet50	VGG19	ResNet50	VGG19
Grad-CAM	0.47	0.76	<b>0.86</b>	<b>0.89</b>	<b>0.831</b>	0.714	0.927	0.985
HiResCAM	<b>0.53</b>	<b>0.8</b>	0.67	0.84	0.83	<b>0.834</b>	<b>0.936</b>	<b>1.081</b>

- HAAS doesn't acknowledge any quality in the medical models attribution maps
- Why ?

- 1st approach: *HA image pixels can have considerable value difference when compared to the original image pixels.*

ex. For  $a = \frac{1}{2}$  attribution, if pixel  $x = \frac{1}{2}$  then  $HA_x = \frac{3}{4}$   
and if pixel  $x = -\frac{1}{2}$  then  $HA_x = -\frac{3}{4}$

Thus, HA image might be *far* from the distribution that the model was trained on.

- 2nd approach: *HAAS incompatible with medical images?*

Non-Medical data tested so far	Medical data
Classes have strong shape dependency	Classes are more densely populated
Weak or No colour dependency	Could have colour dependency

As a result, for medical datasets, could changing the pixels' intensity affect the model's ability to recognize the learned pattern?

Visible color differences in the following:

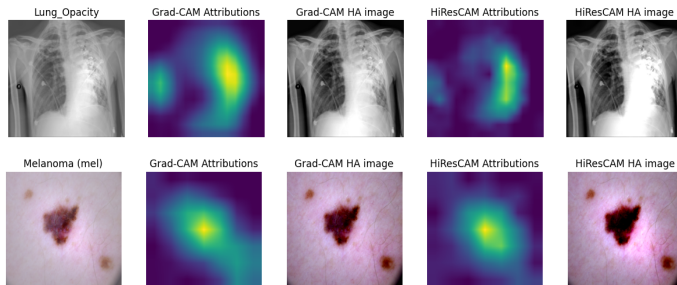


Figure: Examples of HA images

## How to test our hypothesis?

- Over Non-Medical data
- Build *random* models (of decent performance)
- Track min-max HAAS values

Table: HAAS Scores (Non Medical datasets)

HiResCAM		Cifar-10	STL-10	Imagenette
		VGG19*	VGG19*	VGG19*
	Max HAAS Score	<b>1.009</b>	<b>1.034</b>	<b>1.002</b>
	Mean AUC	0.981	0.966	0.995
	Min HAAS Score	<b>0.970</b>	<b>0.978</b>	<b>0.986</b>
	Mean AUC	0.969	0.899	0.889

Note: \* loop of 16 models for different batch size,  
learning rate, scheduler and weight decay

Thus, over non-medical datasets, we extracted *meaningful scores* with 16 random models.

## Connecting accuracy with HiResCAM metrics values

Is good Bal. Accuracy value coupled with *good* metrics results?

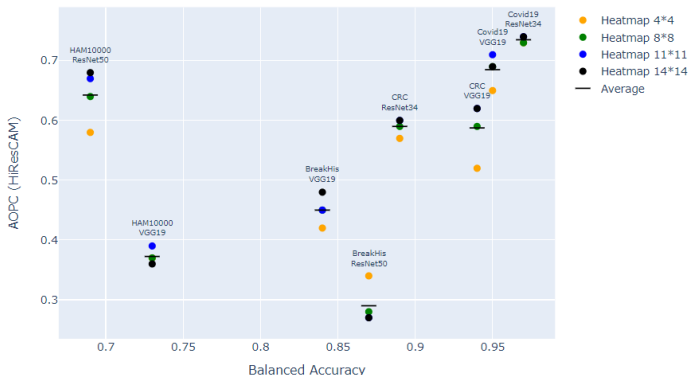


Figure: Bal. Accuracy vs AOPC

- Top 4 models wrt Bal. Accuracy: *almost* increasing pattern wrt average value
- VGG19 models: increasing pattern wrt average value

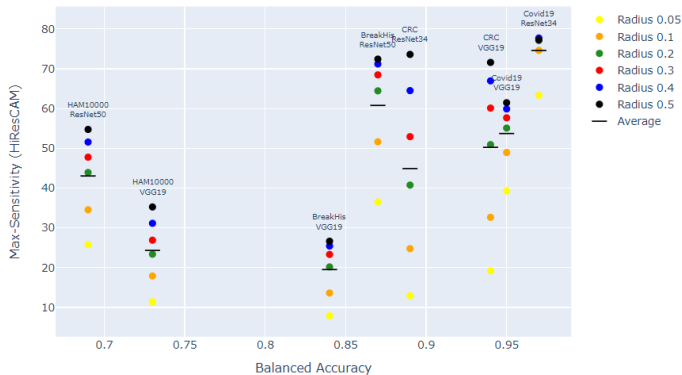


Figure: Bal. Accuracy vs Max Sensitivity

- No pattern (overall models, at dataset level, at model level)

Thank you !