From Explanation to Unsupervised Segmentation: Fusion of Multiple Explanation Maps for Vision Transformers

Prof. Darian M. Onchis (WUT)
Joint work with Prof. Adina Magda Florea and Dr. Istin Codruta
PhD students Eduard Hogea, Ana Coporan

ENFIELD Workshop on Human-Centric AI

9 Sept 2025
National University of Science and Technology POLITEHNICA Bucharest, Central Library

Outline

Motivation

Vision Transformers (ViTs)

- What are ViTs?
  - Transformer models adapted for computer vision tasks
  - ViTs process images using self-attention mechanisms

Motivation

Vision Transformers (ViTs)

- What are ViTs?
  - Transformer models adapted for computer vision tasks
  - ViTs process images using self-attention mechanisms
- Why use ViTs?
  - Good capturing of long-range dependencies
  - Superior performance on large datasets

Motivation

Vision Transformers (ViTs)

- What are ViTs?
  - Transformer models adapted for computer vision tasks
  - ViTs process images using self-attention mechanisms
- Why use ViTs?
  - Good capturing of long-range dependencies
  - Superior performance on large datasets

Explainable Artificial Intelligence (XAI)

- How does a model reach a conclusion?
  - Transparency
  - Trust

Motivation

Vision Transformers (ViTs)

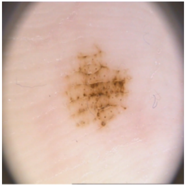
- What are ViTs?
  - Transformer models adapted for computer vision tasks
  - ViTs process images using self-attention mechanisms
- Why use ViTs?
  - Good capturing of long-range dependencies
  - Superior performance on large datasets

Explainable Artificial Intelligence (XAI)

- How does a model reach a conclusion?
  - Transparency
  - Trust
- ViT: visualization-based approaches
  - Helps highlight which image regions contribute most to a prediction

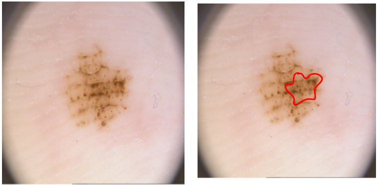
Problem

Given a ViT model and an image, identify which parts of the input image influence the classification of the ViT.



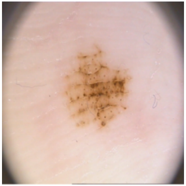
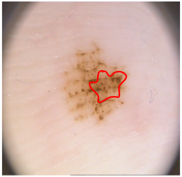
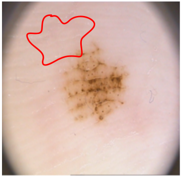
Problem

Given a ViT model and an image, identify which parts of the input image influence the classification of the ViT.



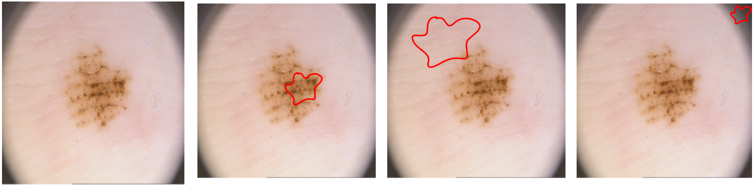
Problem

Given a ViT model and an image, identify which parts of the input image influence the classification of the ViT.

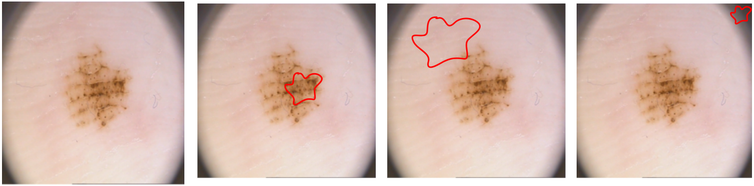
Problem

Given a ViT model and an image, identify which parts of the input image influence the classification of the ViT.



Problem

Given a ViT model and an image, identify which parts of the input image influence the classification of the ViT.



Relevant for:

- Model Validation
- Region of Interest Segmentation

Overview of existing methods

- Attention-Based Methods
- Gradient-Based Methods
- Attribution Propagation Methods
- Causal Explanations
- Hybrid Methods

Attention-Based Methods

These methods analyze how attention is distributed across layers.

- Attention Rollout
  - aggregates attention maps layer by layer
- Attention Flow
  - models information propagation using a flow-based approach
  - computationally expensive

Gradient Based Methods

💡 Compute the gradients of the model's output with respect to the input features

Gradient Based Methods

Compute the gradients of the model's output with respect to the input features

Vanilla Saliency

- maximum absolute gradient across channels

Gradient Based Methods

Compute the gradients of the model's output with respect to the input features

Vanilla Saliency

- maximum absolute gradient across channels

Gradient Class Activation Map (GradCAM)

- combine importance scores derived from gradients with
  - activation maps (CNNs)
  - attention maps (ViTs)

Attribution Propagation & Causal Explanations

Attribution Propagation Methods

Layer-wise Relevance Propagation (LRP)

- propagates relevance scores from the model's output back to the input features

Attribution Propagation & Causal Explanations

Attribution Propagation Methods

Layer-wise Relevance Propagation (LRP)

- propagates relevance scores from the model's output back to the input features

Causal Explanations Based Methods

Uncover cause-and-effect relationships between input features and model predictions

ViT-CX framework

- examine how changes in model input affect its output

Hybrid Methods

Combine attention mechanisms, gradient-based approaches, and attribution propagation

Transition Attention Maps (TAM)

- Models information flow in ViTs as a Markov process.
  - Chain states: ouput embeddings
  - State transition matrix: attention weights and residual connections
  - Explanation: combine state with gradients

Contributions

Unexplored area

Are explainability methods consistent across different data domains?

- Most techniques evaluated on standard object recognition datasets
- Explainability is key in real world applications

Contributions

Unexplored area

Are explainability methods consistent across different data domains?

- Most techniques evaluated on standard object recognition datasets
- Explainability is key in real world applications

Contributions

- Hybrid explainability approach integrating LRP, CAM, Saliency, Rollout
- Improved performance
- Consistent results: tested across general and medical datasets

Solution

Hypothesis

Combining multiple explainability methods enhances interpretability by leveraging their individual strengths.

Solution

Hypothesis

Combining multiple explainability methods enhances interpretability by leveraging their individual strengths.

Pigeonhole Principle

If n pigeons are placed in k holes, at least one hole must contain $\lceil n/k \rceil$ pigeons.

Solution

Hypothesis

Combining multiple explainability methods enhances interpretability by leveraging their individual strengths.

Pigeonhole Principle

If n pigeons are placed in k holes, at least one hole must contain $\lceil n/k \rceil$ pigeons.

- Feature attribution represented as matrices $A = [a_{ij}]$, $B = [b_{ij}]$.

Solution

Hypothesis

Combining multiple explainability methods enhances interpretability by leveraging their individual strengths.

Pigeonhole Principle

If n pigeons are placed in k holes, at least one hole must contain $\lceil n/k \rceil$ pigeons.

- Feature attribution represented as matrices $A = [a_{ij}]$, $B = [b_{ij}]$.
- Geometric mean for each pair: $\sqrt{a_{ij} \cdot b_{kl}}$

Solution

Hypothesis

Combining multiple explainability methods enhances interpretability by leveraging their individual strengths.

Pigeonhole Principle

If n pigeons are placed in k holes, at least one hole must contain $\lceil n/k \rceil$ pigeons.

- Feature attribution represented as matrices $A = [a_{ij}]$, $B = [b_{ij}]$.
- Geometric mean for each pair: $\sqrt{a_{ij} \cdot b_{kl}}$
- Total pairs: n^4, distinct mean values: V.

Solution

Hypothesis

Combining multiple explainability methods enhances interpretability by leveraging their individual strengths.

Pigeonhole Principle

If n pigeons are placed in k holes, at least one hole must contain $\lceil n/k \rceil$ pigeons.

- Feature attribution represented as matrices $A = [a_{ij}]$, $B = [b_{ij}]$.
- Geometric mean for each pair: $\sqrt{a_{ij} \cdot b_{kl}}$
- Total pairs: n^4, distinct mean values: V.
- If $n^4 > V$, by Pigeonhole Principle, at least one geometric mean appears multiple times
  - areas of interest will be highlighted by more than one method

Precision gain (Quantify)

After thresholding each attribution map into a binary mask $X_i \in \{0,1\}^{n\times n}$ on the ViT patch grid, let $R \subseteq \{1,\dots,n\}^2$ be the set of truly relevant patches. For any patch t define

$$p = \Pr[X_i(t) = 1 \mid t \in R],$$
$$q = \Pr[X_i(t) = 1 \mid t \notin R], \quad 0 < q < p < 1.$$

Assuming the masks are *conditionally independent* given R, the posterior precision of their k-way intersection $\hat{X}_k(t) = \prod_{i=1}^{k} X_i(t)$ is

$$\Pr\big[t \in R \mid \hat{X}_k(t) = 1\big] = \frac{p^k}{p^k + q^k} > \frac{p}{p+q} \tag{1}$$

where the right-hand fraction is exactly the precision obtained from a *single* explanation map ($k = 1$). Because $\frac{p^k}{p^k+q^k}$ is strictly increasing in k, each additional explainer that agrees on a pixel raises the probability that the pixel truly belongs to R, while the expected number of false positives drops geometrically with k.

Empirical link to metrics

Equation (1) predicts lower deletion-AUC and higher IoU/Dice for fused maps.
On Pascal VOC dataset the two-way fusion of LRP and Attention Rollout lowers deletion-AUC from 0.53 (best single map) to 0.43 and raises IoU by $+7.1$ points.
Similar improvements appear on ImageNet and PH^2. Hence, the theory is consistent with the observed quantitative gains.

Solution

Methodology

- Integrate 4 explainability methods in two-way and three-way combinations.
  - GradCAM
  - LRP
  - Saliency
  - Attention Rollout

Solution

Methodology

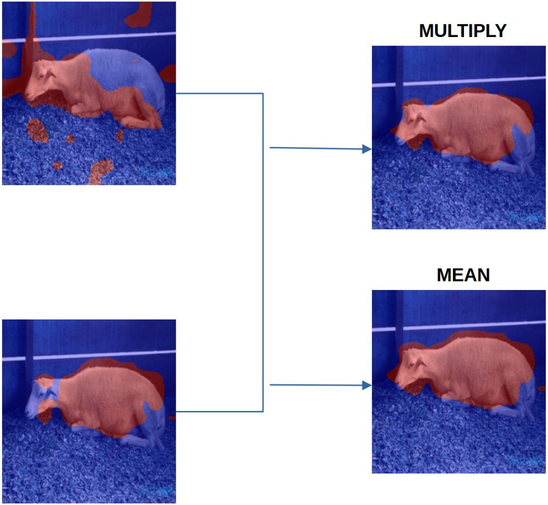
- Integrate 4 explainability methods in two-way and three-way combinations.
  - GradCAM
  - LRP
  - Saliency
  - Attention Rollout
- Fusion strategies:
  - element-wise multiplication
  - geometric mean

Solution

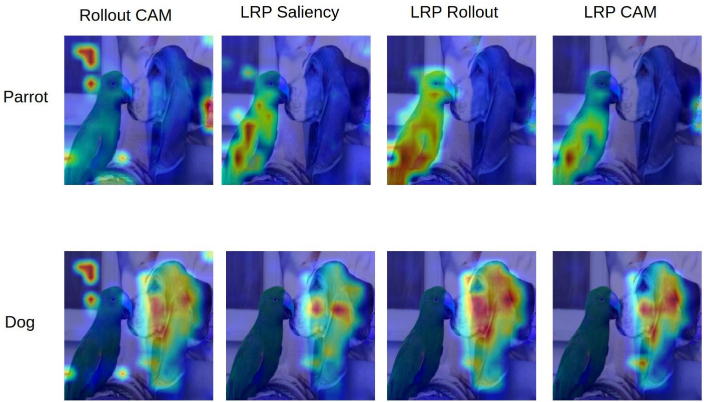
Methodology

- Integrate 4 explainability methods in two-way and three-way combinations.
  - GradCAM
  - LRP
  - Saliency
  - Attention Rollout
- Fusion strategies:
  - element-wise multiplication
  - geometric mean
- Output formats:
  - heatmaps
  - mask

Two way combinations - Masks



Two way combinations - Heatmaps



Evaluation Metrics

Segmentation-Based Metrics

IoU, F1 Score, and Pixel Accuracy measure the alignment between the predicted and ground truth masks.

Explainability Metric

Deletion AUC evaluates how much classification confidence decreases when high-attribution pixels are removed, verifying feature importance.

Results on ImageNet Subset

Table: Results using geometric mean for 2way methods. Best Results are highlighted.

Method	IoU	F1	PA	Deletion
1Way Methods				
CAM	14.33	21.23	19.40	0.40
LRP	**42.59**	**56.72**	53.32	**0.19**
Rollout	36.55	50.79	**66.32**	0.24
Saliency	7.94	12.78	13.37	0.44
2Way Methods				
LRP+CAM	21.62	31.03	27.72	0.37
LRP+Rollout	**52.33**	**65.71**	**68.71**	**0.18**
Rollout+CAM	20.63	29.01	28.95	0.39

Results on Pascal VOC

Table: Results using geometric mean for 2way methods. Best Results are highlighted. Only for images with a predicted probability above 0.85 for the main class.

Method	IoU	F1	PA	Deletion
1Way Methods				
CAM	12.43	19.13	65.94	0.25
LRP	36.41	50.19	**75.52**	**0.12**
Rollout	**43.27**	**57.90**	73.30	0.14
Saliency	11.15	17.59	64.24	0.27
2Way Methods				
LRP+CAM	22.31	32.74	69.70	0.22
LRP+Rollout	**48.65**	**62.48**	**79.09**	**0.12**
Rollout+CAM	20.45	29.68	67.55	0.23

Results on PH2

Table: Results using geometric mean for 2way methods. Best Results are highlighted. Model was finetuned on the dataset and reached an accuracy of 85%

Method	IoU	F1	PA	Deletion
1Way Methods				
CAM	34.83	45.15	39.98	0.50
LRP	52.20	66.54	53.25	**0.45**
Rollout	**53.66**	**67.43**	**67.61**	0.46
Saliency	39.09	52.62	47.22	0.49
2Way Methods				
LRP+CAM	44.10	56.63	45.50	0.38
LRP+Rollout	**64.49**	**76.66**	**67.13**	**0.32**
Rollout+CAM	46.32	57.40	55.69	0.40

We also provide a comparison between the regions highlighted by individual XAI methods and those produced by our mixed approach, ViTmiX. Notice that ViTmiX consistently emphasizes key areas that align closely with human-perceived salient regions, while the single-method heatmaps tend to be vaguer and less comprehensive. This suggests improved spatial focus on the main objects, consistent with human-marked regions.

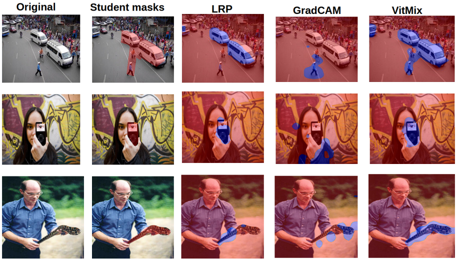


Figure: Comparison of human-perceived regions with XAI maps.

Method	Ground Truth IoU	Ground Truth F1	Ground Truth PA	Student Mask IoU	Student Mask F1	Student Mask PA	Del
1way Methods							
CAM	34.83	45.15	39.98	30.68	41.21	35.03	0.50
LRP	52.20	66.54	53.25	46.95	62.01	48.76	**0.45**
Rollout	**53.66**	**67.43**	**67.61**	**49.93**	**64.08**	**62.04**	0.46
Saliency	39.09	52.62	47.22	36.94	50.50	43.58	0.49
2way Mean							
LRP+CAM	44.10	56.63	45.50	38.40	51.40	40.25	0.38
LRP+Rollout	**64.49**	**76.66**	**67.13**	**58.15**	**71.73**	**61.53**	**0.32**
LRP+Saliency	58.18	71.62	60.74	52.45	66.89	55.59	**0.32**
Rollout+CAM	46.32	57.40	55.69	41.66	53.18	49.72	0.40
Saliency+CAM	46.10	57.93	51.53	40.95	53.28	46.04	0.39
Saliency+Rollout	55.13	69.00	66.28	52.17	66.35	61.53	0.35
3way Mean							
LRP+Rollout+CAM	43.72	56.04	44.91	37.71	50.66	39.31	0.37
LRP+Saliency+CAM	39.32	52.09	39.99	34.02	47.14	35.14	0.36
LRP+Saliency+Rollout	**56.81**	**69.75**	**59.51**	**51.69**	**65.47**	**54.52**	**0.33**
Saliency+Rollout+CAM	42.15	54.10	47.45	37.04	49.54	41.93	0.37

Conclusions

- Combining multiple explainability methods improves ViT interpretability.
- LRP and Rollout emerge as the most effective individual techniques.
- Geometric mean aggregation enhances attribution map clarity.
- Pigeonhole Principle provides theoretical proof for explainability gain.
- Approach generalizes well across datasets, including medical imaging.