Formal Abductive Latent Explanations for Prototype-Based Networks

Jules Soria, Zakaria Chihani, Julien Girard-Satabin, Alban Grastien, Romain Xu-Darme, Daniela Cancila
CEA-List, Université Paris-Saclay

We want Trustworthy & Explainable AI

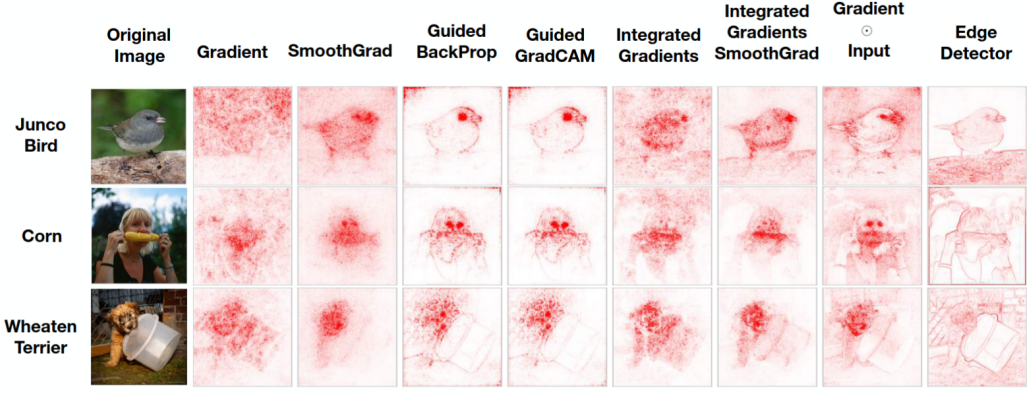
Trust through Explainability. To trust machine learning systems, we must understand how they operate.[1]

Post-hoc Explanations on Black-box Image Classifiers are Tricky

Can we *really* trust saliency maps ?

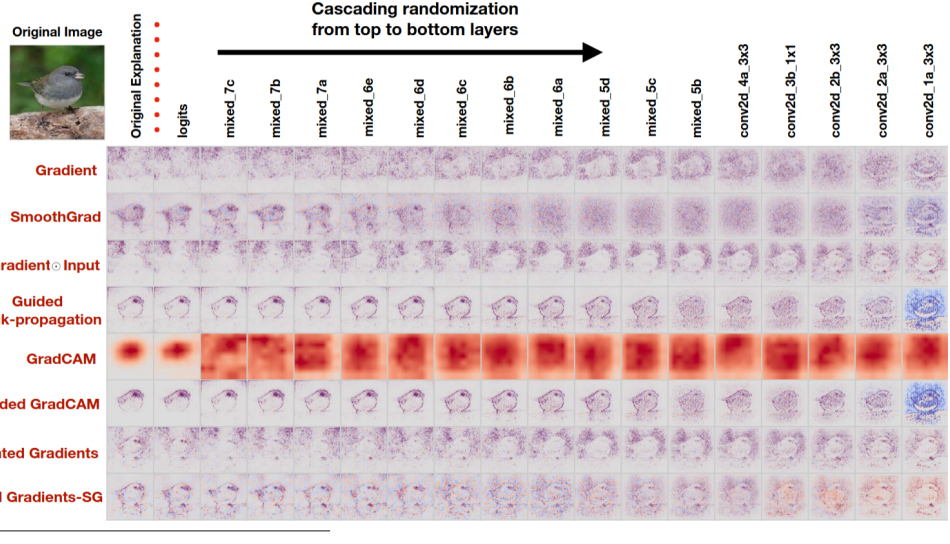
1. Marco Tulio Ribeiro, Sameer Singh et Carlos Guestrin. « "Why should i trust you ?" Explaining the predictions of any classifier ». In : *KDD*. 2016

Saliency Maps are Edge Detectors ?[2]



2. Julius Adebayo et al. « Sanity Checks for Saliency Maps ». In : *NeurIPS*. 2018

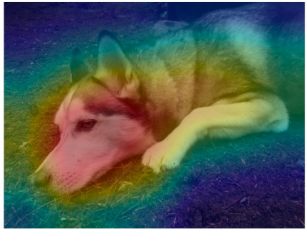
Saliency Maps are Unfaithful? [3]



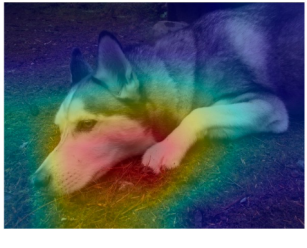
3. Julius Adebayo et al. « Sanity Checks for Saliency Maps ». In : *NeurIPS*. 2018

Saliency Maps are Class-Invariant? [4]

Evidence for Siberian Husky



Evidence for Transverse Flute



4. Cynthia Rudin. « Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead ». In : *Nature Machine Intelligence* (2019)

Use Interpretable ML models instead !

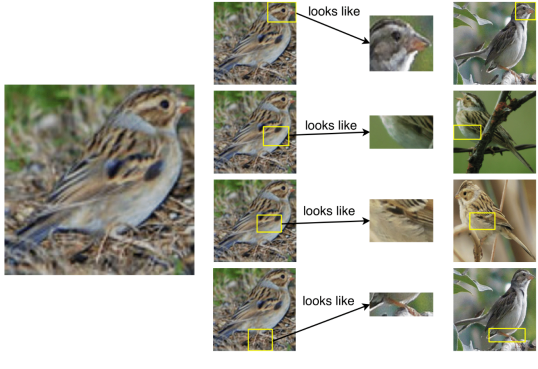
- Between two models of similar predictive capacity, we prefer the one that we can understand rather than a black-box one.[5]
- Interpretability itself is an *ill-defined* concept.[6]
  - Simple models like linear regression and decision trees are interpretable...
    ...until they are not.[7]
  - Interpretable models still need to be explained.[8]

A model for a task

What is an "*interpretable-by-design*" model for image classification ?

5. Cynthia Rudin. « Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead ». In : *Nature Machine Intelligence* (2019)
6. Zachary C. Lipton. « The Mythos of Model Interpretability ». In : *Communications of the ACM* (2018)
7. Christoph Molnar. *Interpretable Machine Learning*. 3e éd. 2025. url : https://christophm.github.io/interpretable-ml-book
8. Joao Marques-Silva et Alexey Ignatiev. « No silver bullet : interpretable ML models must be explained ». In : *Frontiers in Artificial Intelligence* (2023)

Prototypical-Parts Networks [9]

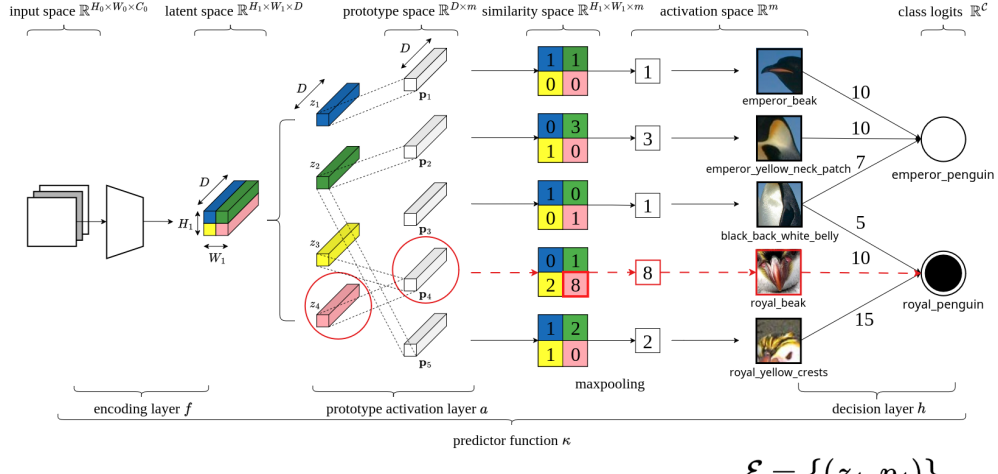


Sufficient Explanation ?

ProtoPNets give the k most activated prototypes as "evidence".
$\rightarrow k = 10$ is arbitrarily chosen.

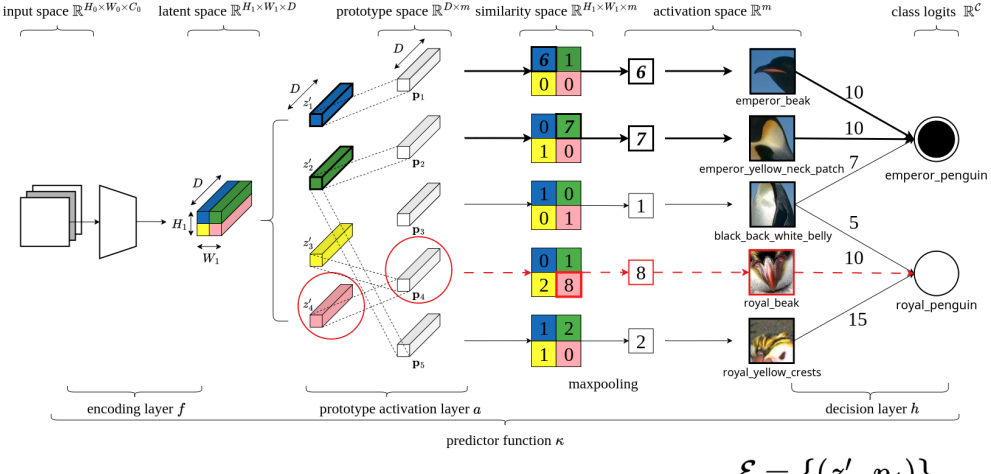
9. Chaofan Chen et al. « This Looks Like That : Deep Learning for Interpretable Image Recognition ». In : *NeurIPS*. 2019

Toy Example



$$\mathcal{E} = \{(z_4, p_4)\}$$
$$\mathcal{E} \Rightarrow \kappa(\mathbf{x}) = 2?$$

Toy Counter-Example



$$\mathcal{E} = \{(z'_4, p_4)\}$$
$$\mathcal{E} \nRightarrow \kappa(\mathbf{x}) = 2$$

Formal Explainable AI

A classification problem can be formally represented by a 4-tuple $(\mathcal{F}, \mathbb{D}, \mathcal{K}, \kappa)$.

- $\mathcal{F} = \{1, ..., m\}$ is a finite set of m features.
- $\mathbb{D} = \{\mathcal{D}_1, ..., \mathcal{D}_m\}$ is a set of domains, $\mathbb{F} = \mathcal{D}_1 \times ... \times \mathcal{D}_m$.
- $\mathcal{K} = \{c_1, ..., c_K\}$ is a finite set of classes.
- $\kappa : \mathbb{F} \to \mathcal{K}$ is a classification function.

AXp $\mathcal{E} \subset \mathcal{F}$ explains why the model predicted class $c = \kappa(v)$ for $v \in \mathbb{F}$.

Def. Abductive Explanation (AXp) [10]

$$\forall (x \in \mathbb{F}). \left[\left(\bigwedge_{i \in \mathcal{E}} (x_i = v_i) \right) \rightarrow (\kappa(x) = c) \right]$$

10. Joao Marques-Silva et Alexey Ignatiev. « Delivering Trustworthy AI Through Formal XAI ». In : *AAAI*. 2022

Formal Explainable AI for Images

- Brings formal guarantees :
  → we cannot have the same explanation for two **distinct** predictions.
- Logical conditions expressed at the input level.

Suitable for images ?

The input level is the pixel !

Forget the pixels, explain the latent representation !

- AXps on images tend to be big, they usually take up $\approx 80\%$ of the image and are not interpretable to humans.[11][12][13]
- Explanations in the latent space
  - stay *faithful* to the model's behavior.
  - *might* have human-readable semantics.

Def. Abductive Latent Explanation (ALE)

$$\forall (x \in \mathbb{F}). \left[\left(\bigwedge_{i \in \mathcal{E}} (f(x)_i = f(v)_i) \right) \rightarrow (\kappa(x) = c) \right]$$

11. Shahaf Bassan et Guy Katz. « Towards Formal XAI : Formally Approximate Minimal Explanations of Neural Networks ». In : *TACAS*. 2023
12. Min Wu, Haoze Wu et Clark Barrett. « VeriX : Towards Verified Explainability of Deep Neural Networks ». In : *NeurIPS*. 2023
13. Dorin Doncenco et al. « A Dive into Formal Explainable Attributions for Image Classification ». In : *ECAI*. 2025

Formal-izing Interpretable Explanations

We can rewrite the explanations ProtoPNet[14] and similar neural architectures *implicitly* give as follows :

Def. ProtoPNet Explanation

$$\forall (x \in \mathbb{F}). \left[\left(\bigwedge_{i \in \mathcal{E}} a(x)_i = a(v)_i \right) \wedge \left(\bigwedge_{j \notin \mathcal{E}, i \in \mathcal{E}} a(x)_j \leq a(v)_i \right) \xrightarrow{?} (\kappa(x) = c) \right]$$

Explanations are potentially misleading (or *optimistic*[15])
→ it is possible to find elements that match the explanation's condition **and** have a different classification.

14. Chaofan Chen et al. « This Looks Like That : Deep Learning for Interpretable Image Recognition ». In : *NeurIPS*. 2019
15. Alexey Ignatiev, Nina Narodytska et Joao Marques-Silva. « On Validating, Repairing and Refining Heuristic ML Explanations ». In : *arXiv :1907.02509* (2019)

Formal and Interpretable Explanations

Def. Top-k Explanation (Bounds-based)

$$\forall (x \in \mathbb{F}). \left[\left(\bigwedge_{j=1}^{m} a(x)_j \in [\underline{a}_{\mathcal{E},j}, \overline{a}_{\mathcal{E},j}] \right) \rightarrow (\kappa(x) = c) \right]$$

For $j \in \mathcal{E}$ (prototypes *in* the explanation) :

$$\underline{a}_{\mathcal{E},j} = \overline{a}_{\mathcal{E},j} = a(v)_j$$

For $j \notin \mathcal{E}$ (prototypes *not in* the explanation) :

$$\underline{a}_{\mathcal{E},j} = 0 \quad \text{and} \quad \overline{a}_{\mathcal{E},j} = \min_{i \in \mathcal{E}} a(v)_i$$

Algorithm 1 : Generating ALE

function GenerateALE(v, c)
  $\mathcal{E} = \emptyset$ *# Initialize an empty explanation*
  UnverifiedClasses $= \mathcal{K} \setminus \{c\}$ *# All classes to check against*
  $\mathcal{A} \leftarrow$ Sort($a(v)$) *# Sort prototypes by activation, high to low*
  while UnverifiedClasses $\neq \emptyset$ **do** *# Loop until $\mathcal{E}$ guarantees prediction c*
    $j =$ NextPrototype($\mathcal{E}, \mathcal{A}$) *# Get next most-activated prototype*
    $\mathcal{E} \leftarrow \mathcal{E} \cup \{j\}$ *# Add it to the explanation*
    for $c' \in$ UnverifiedClasses **do** *# Check remaining challenger classes*
      CEx $\leftarrow a^*_{\mathcal{E}}(c', c)$ *# Find worst-case activation favoring c' against c*
      if h_c(CEx) $> h_{c'}$(CEx) **then** *# If c still wins, even in worst-case…*
        UnverifiedClasses $\leftarrow$ UnverifiedClasses $\setminus \{c'\}$ *# …then c' is "beaten"*
      end if
    end for
  end while
  return $\mathcal{E}$ *# Return the cardinality-minimal sufficient top-k explanation*
end function

Experiments on CUB-200[16] (1/2)

- 200 classes
- 10 prototypes per class → $m = 2000$ prototype activations in total.
- Comparing to the default top-10 explanations provided by ProtoPNet.

Research Question 1

Does the *correctness* of the model's prediction influence the size of the sufficient top-k explanation?

	Avg. Exp. Size	Correct Exp. Size	Incorrect Exp. Size
Top-k	427	306	673

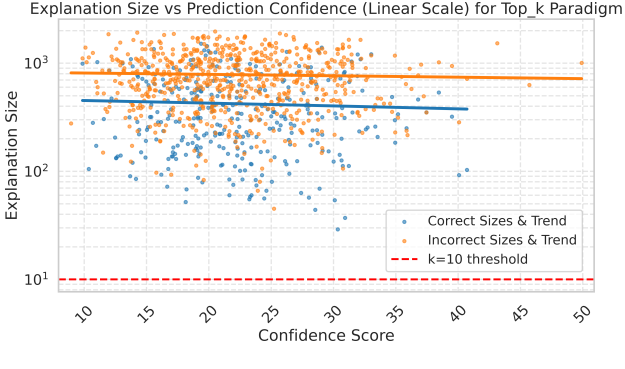
Mean Explanation Sizes

16. C. Wah et al. *The Caltech-UCSD Birds-200-2011 Dataset*. Rapp. tech. CNS-TR-2011-001. California Institute of Technology, 2011.

Experiments on CUB-200 (2/2)

Research Question 2

Does the *confidence* of the model in its prediction (class logits) influence the size of the sufficient top-k explanation ?



Key Findings

In-sufficiency of explanations

- ProtoPNet explanations are *insufficient* in all cases !
- Sufficient top-k explanations tend to be very big.
  → This means current ProtoPNet models are *not* as interpretable as they initially thought.

Answering our RQs

1. Correct and Incorrect predictions have different explanation sizes.[17]
2. Model confidence does not seem to influence the explanation size.

17. Min Wu et al. « Better Verified Explanations with Applications to Incorrectness and Out-of-Distribution Detection ». In : *arXiv :2409.03060* (2024)

Future Work

Can we extract more information from the latent representation?

Knowing that **z** looks like $\mathbf{p}_1$ may tell us that :

- **z** *also* looks like $\mathbf{p}_2$
- **z** does *not* look like $\mathbf{p}_3$

Can we train models to have smaller explanations?

- Removing similar prototypes, merging them, pruning the model?
- Incentivizing a more disentangled and transparent latent space?
- Reducing the latent space's depth D while maintaining similar predictive performances?

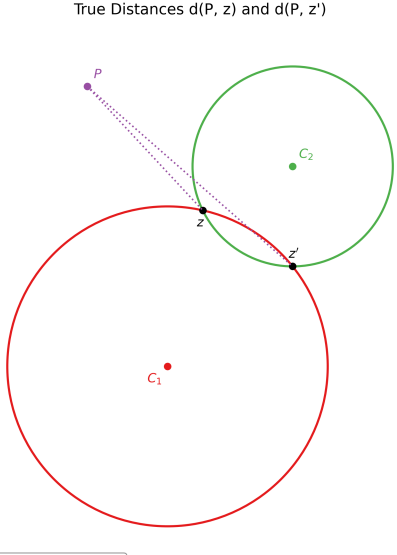
~~Future Work~~ Recent Work !

Can we extract more information from the latent representation ?

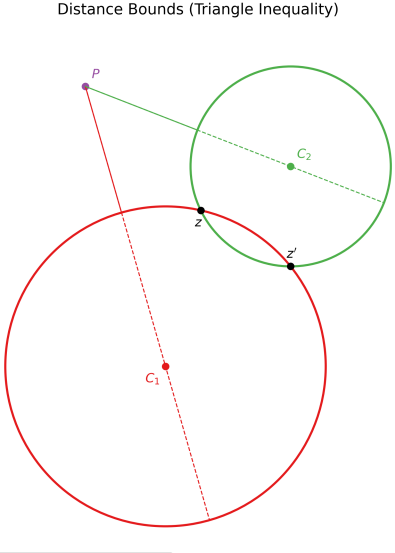
Knowing that $\mathbf{z}$ looks like $\mathbf{p}_1$ may tell us that :

- $\mathbf{z}$ *also* looks like $\mathbf{p}_2$
- $\mathbf{z}$ does *not* look like $\mathbf{p}_3$

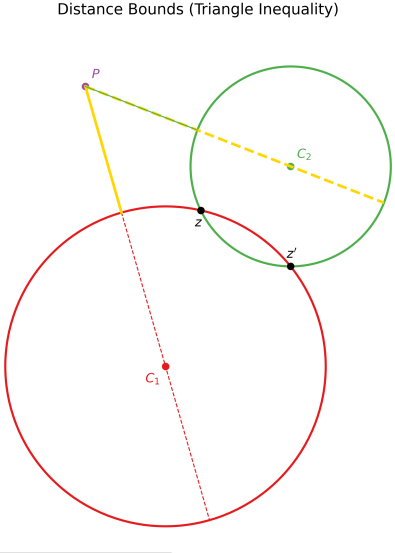
Approximating points in the latent space



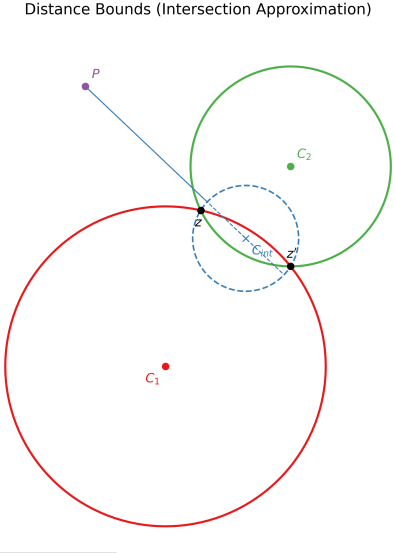
Approximating points in the latent space



Approximating points in the latent space



Approximating points in the latent space



Spatial ALE (1/3) : The Framework

Our explanation $\mathcal{E}$ is now a set of patch-prototype *pairs* (l, j). The formal definition remains the same (described on Slide 14) :

Def. Spatial Approximation-Based ALE

$$\forall (x \in \mathbb{F}). \left[\left(\bigwedge_{j=1}^{m} a(x)_j \in [\underline{a}_{\mathcal{E},j}, \overline{a}_{\mathcal{E},j}] \right) \rightarrow (\kappa(x) = c) \right]$$

The Challenge

How do we get the activation bounds $[\underline{a}_{\mathcal{E},j}, \overline{a}_{\mathcal{E},j}]$ from a set of patch-prototype pairs $\mathcal{E}$?

Answer : A two-step process using spatial approximations.

Spatial ALE (2/3) : The Two-Step Process

Step 1 : Get Patch-Prototype Similarity Bounds

We use the explanation $\mathcal{E}$ (the set of known pairs) to infer bounds $[\underline{\text{sim}}_{\mathcal{E}}, \overline{\text{sim}}_{\mathcal{E}}]$ for all *unknown* pairs.

- For pairs $(l, j) \in \mathcal{E}$: The bounds are the exact, known similarity.
- For pairs $(l, i) \notin \mathcal{E}$: The bounds are *inferred* using a spatial approximation (like Triangular Inequality or Hypersphere Intersection).

Step 2 : Get Prototype Activation Bounds (via Max Pooling)

The final activation a_j is the *max* similarity over all patches l. We apply this to our bounds :

$$\overline{a}_{\mathcal{E},j} = \max_{l \in \mathbf{L}} \overline{\text{sim}}_{\mathcal{E}}(\mathbf{z}_l, \mathbf{p}_j)$$

$$\underline{a}_{\mathcal{E},j} = \max_{l \in \mathbf{L}} \underline{\text{sim}}_{\mathcal{E}}(\mathbf{z}_l, \mathbf{p}_j)$$

Spatial ALE (3/3) : Comparing Step 1 Approximations

Method 1 : Triangular Inequality

Bounds for $(l,i) \notin \mathcal{E}$:

$$\overline{\text{sim}}_{\mathcal{E}}(\mathbf{z}_l, \mathbf{p}_i) = \min_{(l,j)\in\mathcal{E}} \sigma\left(\left| d(\mathbf{z}_l, \mathbf{p}_j) - d(\mathbf{p}_j, \mathbf{p}_i)\right|\right)$$

$$\underline{\text{sim}}_{\mathcal{E}}(\mathbf{z}_l, \mathbf{p}_i) = \max_{(l,j)\in\mathcal{E}} \sigma\left(d(\mathbf{z}_l, \mathbf{p}_j) + d(\mathbf{p}_j, \mathbf{p}_i)\right)$$

Weakness

Adding a new pair to $\mathcal{E}$ does not guarantee tighter bounds.

Method 2 : Hypersphere Intersection

Concept : Iteratively refine a bounding sphere $H_l(\mathbf{C}_l, r_l)$ for each patch $\mathbf{z}_l$.

Bounds for $(l,i) \notin \mathcal{E}$:

$$\overline{\text{sim}}_{\mathcal{E}} = \sigma\left(d(\mathbf{C}_l, \mathbf{p}_i) - r_l\right)$$

$$\underline{\text{sim}}_{\mathcal{E}} = \sigma\left(d(\mathbf{C}_l, \mathbf{p}_i) + r_l\right)$$

Weakness

Cannot parallelize the computation of new approximated spheres.

Algorithm 2 : Spatial ALE - Forward (Sufficiency)

function GenerateMinimalALE(v, c, paradigm)
    $\mathcal{E}$ = ExpInit(v, c) *# Initialize with one prototype per patch*
    UnverifiedClasses = $C \setminus \{c\}$
    # — 1. FORWARD PASS (Find a Sufficient $\mathcal{E}$) —
    while UnverifiedClasses $\neq \emptyset$ **do**
        (l, j) = NextPair$(\mathcal{E})$ *# Get next salient patch-prototype pair*
        d = ComputeDistance(v, l, j)
        $\mathcal{E} \leftarrow \mathcal{E} \cup \{(l, j)\}$ *# Add pair to explanation*
        (lb, ub) = GenerateBounds$(\mathcal{E}$, paradigm$)$ *# Compute new bounds*
        UnverifiedClasses = VerifExp(lb, ub) *# # Find remaining "challenger" classes*
    end while
    return $\mathcal{E}$ *# Return the sufficient explanation*
end function

Subset-Minimality of Spatial Approximation ALE

Comparison with top-k

- *forward* process very similar to the one described in Algorithm 1 (on Slide 15).
- Top-k ALEs have a fixed order (decreasing activations) ; Spatial Approximation ALEs have no predetermined order.
  → **subset-minimality** is not guaranteed by the forward process.

Algorithm 2 : Spatial ALE - Backward (Minimality)

function GenerateMinimalALE($v, c,$ paradigm)

...

 # — 2. BACKWARD PASS (Minimize $\mathcal{E}$) —

MarkedPairs = $\emptyset$ *# Tracks pairs proven to be necessary*

while $\mathcal{E} \neq \emptyset$ **do**

if $|\mathcal{E}|$ = |MarkedPairs| **then** *# Stop if all remaining pairs are marked*

break

end if

(l, j) = RemoveUnmarkedPair($\mathcal{E}$, MarkedPairs) *# Try removing one pair*

$\mathcal{E} \leftarrow \mathcal{E} \setminus \{(l, j)\}$

(lb, ub) = GenerateBounds($\mathcal{E}$, paradigm)

UnverifiedClasses = VerifExp(lb, ub)

if UnverifiedClasses $\neq \emptyset$ **then** *# If $\mathcal{E}$ is no longer sufficient...*

MarkedPairs ← MarkedPairs $\cup \{(l, j)\}$ *# ...mark pair as necessary*

$\mathcal{E} \leftarrow \mathcal{E} \cup \{(l, j)\}$ *# ...and add it back*

end if

end while

return $\mathcal{E}$ *# Return the subset-minimal sufficient explanation*

end function

Experiments on some CV datasets

Dataset	*Acc.*	Triangle	Hypersphere	top-k	Latent Dim.
		Avg Total / Avg Correct / Avg Incorrect			$H_1 \times W_1$
CIFAR-10	*0.83*	**22.3** / **6.6** / 100	24.3 / 8.9 / 100	41.4 / 36.1 / **62.0**	1×1
CIFAR-100	*0.62*	183.9 / 63.7 / 1000	**39.1 / 8.2 / 77.9**	896.6 / 867.6 / 940.8	1×1
MNIST	*0.98*	**6.1 / 6.1 / -**	675 / 675 / -	8.8 / 8.8 / -	4×4
Flowers	*0.72*	4946.4 / 428.1 / 16320		**287.7 / 193.6 / 525.5**	4×4
Pets	*0.82*	3755.3 / 748.9 / 18130		**77.7 / 67.9 / 122.8**	7×7
Cars	*0.90*	5072.3 / 992.1 / 31633.6		**24.9 / 12.3 / 140.6**	7×7
CUB200	*0.84*	10653.4 / 670.9 / 98000		**239.3 / 217.0 / 352.0**	7×7

Key Findings

In-sufficiency of explanations

- ProtoPNet explanations (top-10) are *insufficient* for all datasets† evaluated !

Formal Explanations Size Differences

- For datasets such that the model uses a latent space composed of a unique patch (i.e., $H_1 \times W_1 = 1 \times 1$), the spatial ALEs perform much better than their top-k counterpart.
- As the latent space grows in size, so do the spatial ALEs size. This is expected as they have to produce bounds for each latent patch.

† : except MNIST.

Issues Encountered

Explanation Size and Computation Time

- Very large explanations also take a long time to compute !
  → Backward pass has complexity $\Omega(\|\mathcal{E}^\star\|^2)$ where $\mathcal{E}^\star$ is the explanation at the end of the "forward pass".
- For Hypersphere Intersection Approximation, the **order** in which prototypes are added to approximate a patch **matters**.
  → `GenerateBounds` called at each step in the backward pass.

Mathematical Insights : High-Dimensionality

- Lévy's Lemma : All vectors are roughly at the same distance.
- J.L Lemma : many "almost orthogonal" vectors (ϵ cosine similarity).

Intuition behind failures : Triangular Inequality

Lévy's Lemma, or *The "Curse of Dimensionality"*

The distance between any two random points $\mathbf{v}_i, \mathbf{v}_j$ in $\mathbb{R}^D$ converges to the same constant value. The notions of "near" and "far" lose their meaning.

$$d(\mathbf{v}_i, \mathbf{v}_j) \approx R \quad \text{for all } i \neq j$$

Triangular Inequality Approximation

We want to bound an unknown distance $d(\mathbf{z}_l, \mathbf{p}_i)$ using a known $d(\mathbf{z}_l, \mathbf{p}_j)$ and $d(\mathbf{p}_j, \mathbf{p}_i)$.

- **Upper Bound :** $d(\mathbf{z}_l, \mathbf{p}_i) \leq d(\mathbf{z}_l, \mathbf{p}_j) + d(\mathbf{p}_j, \mathbf{p}_i)$
- **Lower Bound :** $d(\mathbf{z}_l, \mathbf{p}_i) \geq |d(\mathbf{z}_l, \mathbf{p}_j) - d(\mathbf{p}_j, \mathbf{p}_i)|$

Due to Lévy's Lemma, all distances $\approx R$. The bounds become :

- **Upper Bound :** $\approx R + R = 2R$
- **Lower Bound :** $\approx |R - R| = 0$

Intuition behind failures : Hypersphere Intersection

Johnson-Lindenstrauss Lemma, or *The "Blessing of Dimensionality"*

In $\mathbb{R}^D$, one can find a set $V = \{\mathbf{v}_1, \dots, \mathbf{v}_M\}$ where the size M is exponential in D, yet all vectors are nearly orthogonal to each other :

$$|\cos(\mathbf{v}_i, \mathbf{v}_j)| = \frac{|\langle \mathbf{v}_i, \mathbf{v}_j \rangle|}{\|\mathbf{v}_i\| \|\mathbf{v}_j\|} \leq \epsilon \quad \text{for all } i \neq j$$

Hypersphere Intersection Approximation

- When the three points $\mathbf{C}_{int}$, $\mathbf{z}_l$ and $\mathbf{p}_j$ are aligned, the new approximation is *precisely* the point $\mathbf{z}_l$.
- When the three points are orthogonal, the new approximation is as big as the previous one (smallest of the two hyperspheres that intersect).

Future Work

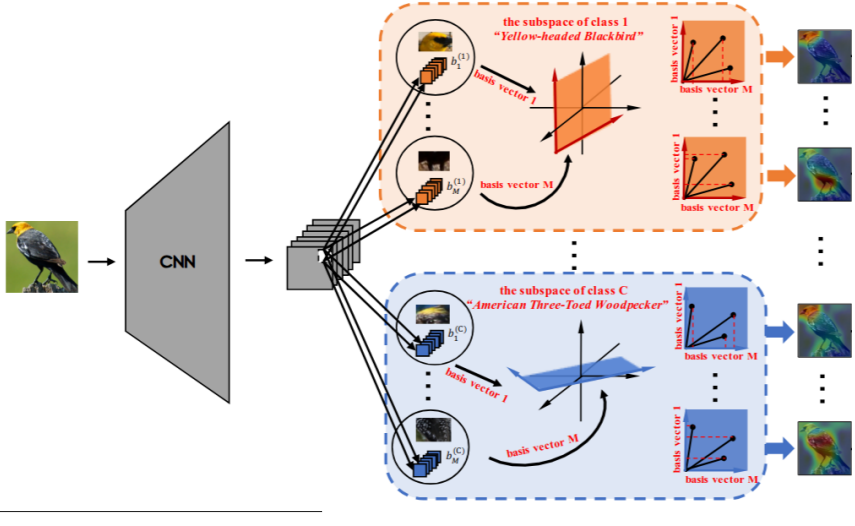
Can we extract more information from the latent representation?

- Deactivating patches that correspond to the background?

Can we train models to have smaller explanations?

- Removing similar prototypes, merging them, pruning the model
- Incentivizing a more **disentangled** and **transparent** latent space
- Reducing the latent space's depth D while maintaining similar predictive performances

Transparent Embedding Space[18]



18. Jiaqi Wang et al. « Interpretable image recognition by constructing transparent embedding space ». In : *CVPR*. 2021.

Takeaway

Key Findings

- Default ProtoPNet explanations are **insufficient**
  - Sufficiency requires many prototypes.
  - Spatial reasoning does not always lead to smaller explanations.
  - Implication : Models are not as "interpretable-by-design" as claimed.
- Explanation size correlates with model **correctness**
  - Incorrect predictions yield *much* larger explanations.
  - This could become a proxy for model *un*certainty
- The goal : Train for *provable* interpretability, not just "by design".

Contact & Resources

Reach out ! `jules.soria@cea.fr` | Also check out our framework : `CaBRNet` [19]

19. Romain Xu-Darme et al. « CaBRNet, an Open-Source Library for Developing and Evaluating Case-Based Reasoning Models ». In : *Proc. xAI-2024 Workshops*. 2024. url : https://cea.hal.science/cea-04688217

References I

[1] Marco Tulio Ribeiro, Sameer Singh et Carlos Guestrin. « "Why should i trust you?" Explaining the predictions of any classifier ». In : *KDD*. 2016.

[2] Julius Adebayo et al. « Sanity Checks for Saliency Maps ». In : *NeurIPS*. 2018.

[3] Cynthia Rudin. « Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead ». In : *Nature Machine Intelligence* (2019).

[4] Zachary C. Lipton. « The Mythos of Model Interpretability ». In : *Communications of the ACM* (2018).

[5] Christoph Molnar. *Interpretable Machine Learning*. 3e éd. 2025. url : https://christophm.github.io/interpretable-ml-book.

[6] Joao Marques-Silva et Alexey Ignatiev. « No silver bullet : interpretable ML models must be explained ». In : *Frontiers in Artificial Intelligence* (2023).

References II

[7] Chaofan Chen et al. « This Looks Like That : Deep Learning for Interpretable Image Recognition ». In : *NeurIPS*. 2019.

[8] Joao Marques-Silva et Alexey Ignatiev. « Delivering Trustworthy AI Through Formal XAI ». In : *AAAI*. 2022.

[9] Shahaf Bassan et Guy Katz. « Towards Formal XAI : Formally Approximate Minimal Explanations of Neural Networks ». In : *TACAS*. 2023.

[10] Min Wu, Haoze Wu et Clark Barrett. « VeriX : Towards Verified Explainability of Deep Neural Networks ». In : *NeurIPS*. 2023.

[11] Dorin Doncenco et al. « A Dive into Formal Explainable Attributions for Image Classification ». In : *ECAI*. 2025.

[12] Alexey Ignatiev, Nina Narodytska et Joao Marques-Silva. « On Validating, Repairing and Refining Heuristic ML Explanations ». In : *arXiv :1907.02509* (2019).

References III

[13] C. Wah et al. *The Caltech-UCSD Birds-200-2011 Dataset*. Rapp. tech. CNS-TR-2011-001. California Institute of Technology, 2011.

[14] Min Wu et al. « Better Verified Explanations with Applications to Incorrectness and Out-of-Distribution Detection ». In : *arXiv :2409.03060* (2024).

[15] Jiaqi Wang et al. « Interpretable image recognition by constructing transparent embedding space ». In : *CVPR*. 2021.

[16] Romain Xu-Darme et al. « CaBRNet, an Open-Source Library for Developing and Evaluating Case-Based Reasoning Models ». In : *Proc. xAI-2024 Workshops*. 2024. url : https://cea.hal.science/cea-04688217.

[17] Richard Tomsett et al. « Sanity Checks for Saliency Metrics ». In : *AAAI*. 2020.

[18] Romain Xu-Darme et al. « Sanity Checks for Patch Visualisation in Prototype-Based Image Classification ». In : *CVPR*. 2023.

References IV

[19] Anna Hedström et al. « A Fresh Look at Sanity Checks for Saliency Maps ». In : *xAI*. 2024.

From Sanity to Scrutiny

A vicious circle

To measure the quality of explanations, we need good metrics, and to have good metrics, we need to already have (at least an idea of) good explanations

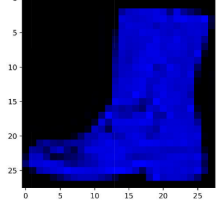
Many sanity checks have been made to re-evaluate how trustworthy image explanations are, and how trustworthy explanation metrics are.[20] [21] [22] [23]

20. Julius Adebayo et al. « Sanity Checks for Saliency Maps ». In : *NeurIPS*. 2018
21. Richard Tomsett et al. « Sanity Checks for Saliency Metrics ». In : *AAAI*. 2020
22. Romain Xu-Darme et al. « Sanity Checks for Patch Visualisation in Prototype-Based Image Classification ». In : *CVPR*. 2023
23. Anna Hedström et al. « A Fresh Look at Sanity Checks for Saliency Maps ». In : *xAI*. 2024

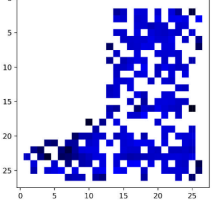
Bundle pixels into superpixels?

Def. Bundled Abductive Explanation (b-AXp) [24]

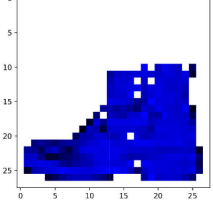
$$\forall (x \in \mathbb{F}).[(\bigwedge_{i \in \mathcal{F}_{\mathcal{B}}} (x_i = v_i)) \rightarrow (\kappa(x) = c)]$$



(a) Original Image



(b) Explanation



(c) Bundle explanation

24. Shahaf Bassan et Guy Katz. « Towards Formal XAI : Formally Approximate Minimal Explanations of Neural Networks ». In : *TACAS*. 2023

Distance-based constraints?

Def. Robust Explanation [25]

$$\forall x' \in \mathbb{F}. \left((x'_A = v_A) \wedge (\|x'_B - v_B\|_p \leq \epsilon) \right) \Rightarrow |f(v) - f(x')| \leq \delta$$

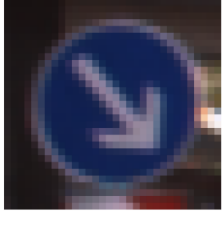
25. Min Wu, Haoze Wu et Clark Barrett. « VeriX : Towards Verified Explainability of Deep Neural Networks ». In : *NeurIPS*. 2023

Do both?

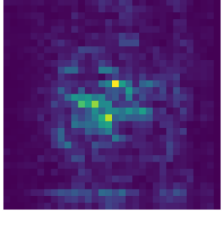
Def. Robust Bundled Abductive Explanation [26]

$$\forall x' \in \mathbb{F}. \Big((x'_{\mathcal{F}_A} = v_{\mathcal{F}_A}) \wedge (\|x'_{\mathcal{F}_B} - v_{\mathcal{F}_B}\|_p \leq \epsilon) \Big) \Rightarrow |f(v) - f(x')| \leq \delta$$

Original Image



GradCAM Sensitivity Map



Pixel-level Explanation

Superpixel-level Explanation

26. Dorin Doncenco et al. « A Dive into Formal Explainable Attributions for Image Classification ». In : *ECAI*. 2025

Thank you for your attention !

Commissariat à l'énergie atomique et aux énergies alternatives
Centre de Bruyères-le-Châtel | 91297 Arpajon Cedex
T. +33 (0)1 69 26 40 00 | F. +33 (0)1 69 26 40 00
Établissement public à caractère industriel et commercial
RCS Paris B 775 685 019