Two-Stage Holistic and Contrastive Explanation of Image Classification

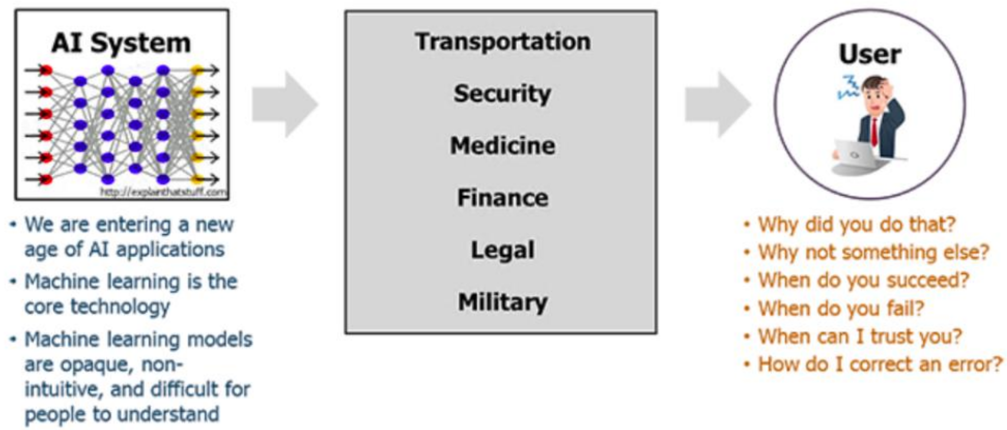
Nevin L. Zhang

CSE@HKUST

Outline

- Introduction to XAI
- UAI 2023: Explanation of Image Classification (CWOX)
  - Why holistic and contrastive?
  - Why two stages?
  - Two-stage holistic and contrastive explanations: How?
- IJCAI 2023: Causal Explanation of Vision Transformers (ViT-CX)

Introduction to XAI



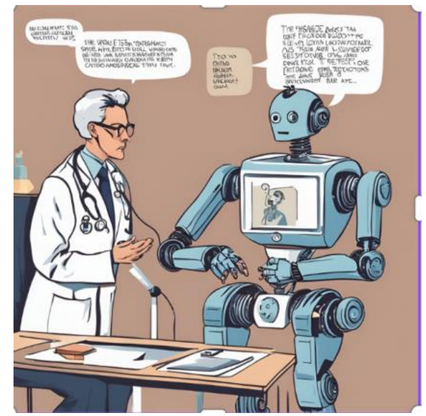
Cunning 2019

- Machine learning models are often opaque: It's hard to see why they make the predictions they do.
- Explainable AI (XAI) is the effort to make these models less of a mystery and more transparent, for both experts and general users.

The Need of XAI: End-User Perspective

Explanations are needed to foster trust

- Doctors need clear explanations before they can rely on AI’s suggestions confidently



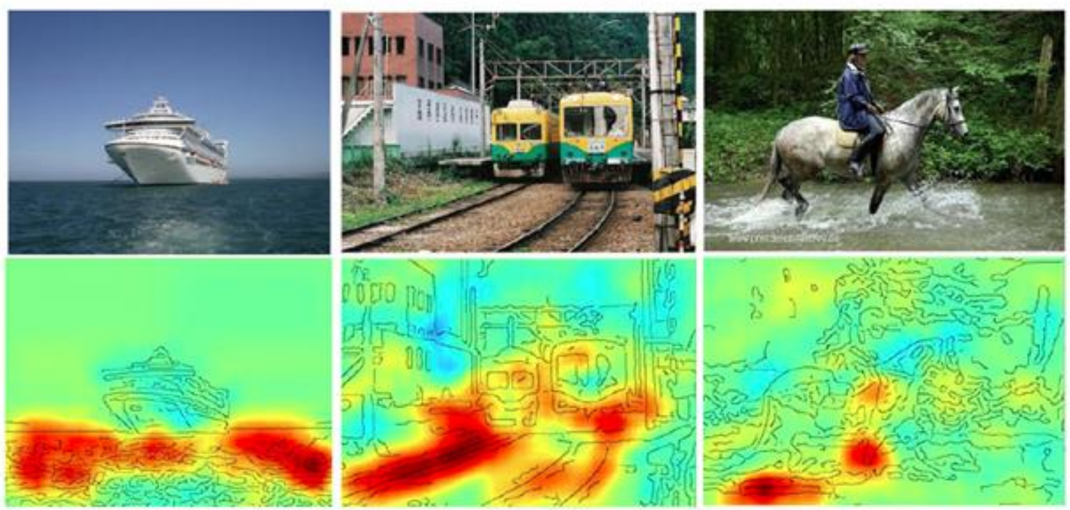
- Patients need to understand the rationale behind when receiving treatment recommendations from AI

"Honey, drink the medicine."
Man died later of poison.

The Need of XAI: ML Expert Perspective

Explanations are needed for model diagnosis

- Not sufficient to see that our models are making the right predictions.
- Need to ensure that they are **right for the right reason**



Samek (2019)

The Need of XAI: Societal Perspective

Explanations are needed for fairness and accountability

- The EU's General Data Protection Regulation (GDPR) confers a right of explanation for all individuals to obtain meaningful explanations of the logic involved for automated decision making.

Lundberg (2019): Explaining interest rate of loan.

Models to be Explained/Levels of Explanation

- Image classifiers
- Tabular data classifiers
- Large language models
- Reinforcement learning models
- Clustering algorithms
- . . .

- Pixel-Level Explanations
- Feature-Level Explanations
- Concept-Level Explanations
- Instance-Level Explanations

Types of Explanations

Local vs Global explanations:

- Local XAI: Explains one particular prediction made by a model.
- Global XAI: Explains general behaviour of a model.

Model-specific or model-agnostic:

- Model-agnostic XAI: Treats models as black-box.
- Model-specific XAI: Depends on the type of selected model

Ante Hoc. vs Post Hoc.:

- Ante Hoc. XAI: Learn models that are interpretable.
- Post Hoc. XAI: Interpret models that are not interpretable by themselves.

This talk: Pixel-level post hoc local explanation of image classification

- UAI 2023: CWOX is a meta-explainer that needs a base explainer
- IJCAI 2023: ViT-CX is model-specific,

UAI 2023

Two-Stage Holistic and Contrastive Explanation of Image Classification

Weiyan Xie [*1] Xiao-Hui Li [2] Zhi Lin [1] Leonard K. M. Poon [3] Caleb Chen Cao [†1] Nevin L. Zhang [*1]

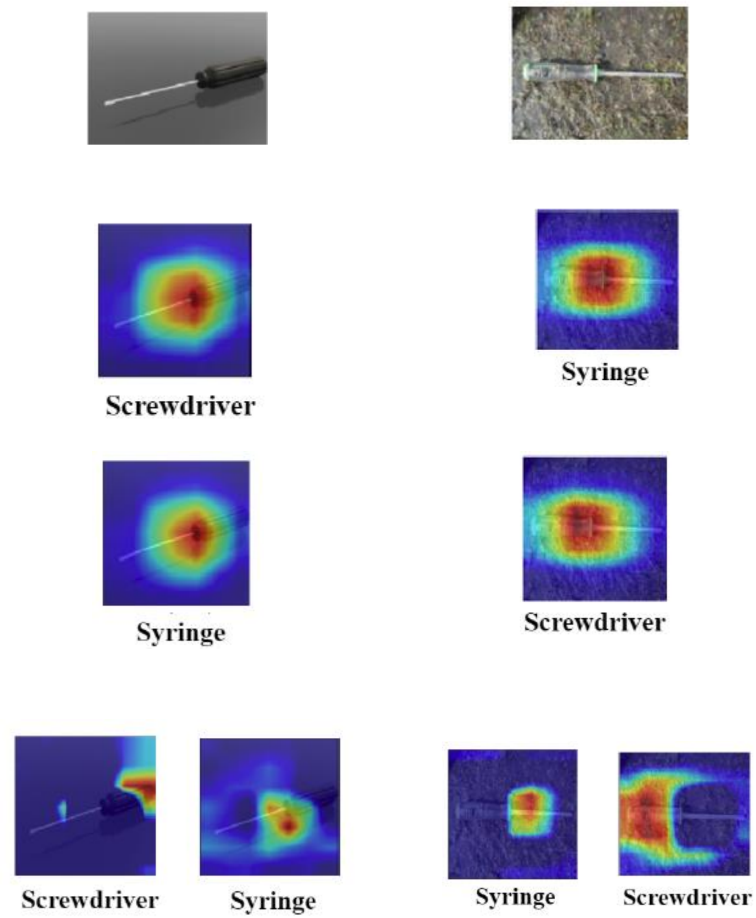
[1] The Hong Kong University of Science and Technology, Hong Kong, China
[2] Huawei Technologies Co., Ltd, Shenzhen, China
[3] The Education University of Hong Kong, Hong Kong, China[‡]

Three Modes of Explanation

- The output of an image classifier is a probability distribution over classes. Most previous XAI methods aim to explain one class in the output.
- Individual output explanation (IOX) is unable to provide users with an overall understanding of model behaviour
- Might mislead users to unjustified confidence in the explanation and the model
  - Why does ResNet give high probabilities to two different classes based on the same evidence?
- Simple whole-output explanation (SWOX) is insufficient
  - Explains all top classes one by one
- Contrastive whole-output explanation (CWOX) reveals evidence for each top class again others

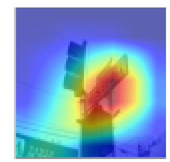
  CWOX(1, 2) = normalize(H1 – H2)

ResNet50 Explained by Grad-Cam

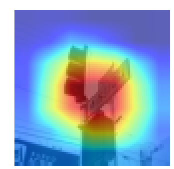


Three Modes of Explanation

ResNet50 Explained by GradCam



street sign

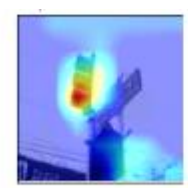


traffic light

- IOX and SWOX do not give us a good overall understanding of model behaviour.

- CWOX clearly shows why ResNet50 gives both street sign an traffic light high probabilities

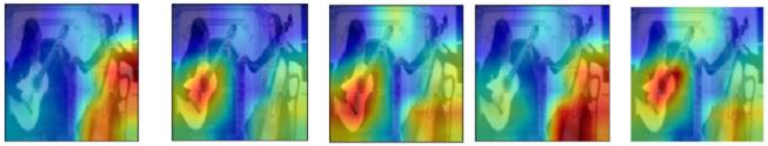
street sign-0.73



traffic light-0.27

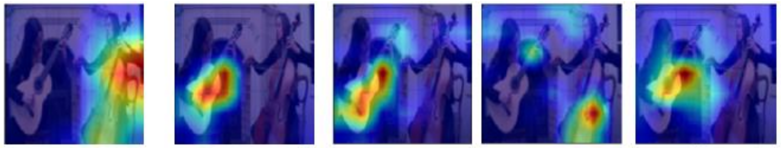
A More Complex Example

SWOX



One-stage CWOX (CWOX-1s)

CWOX(1) = normalize(H1 – H_others)

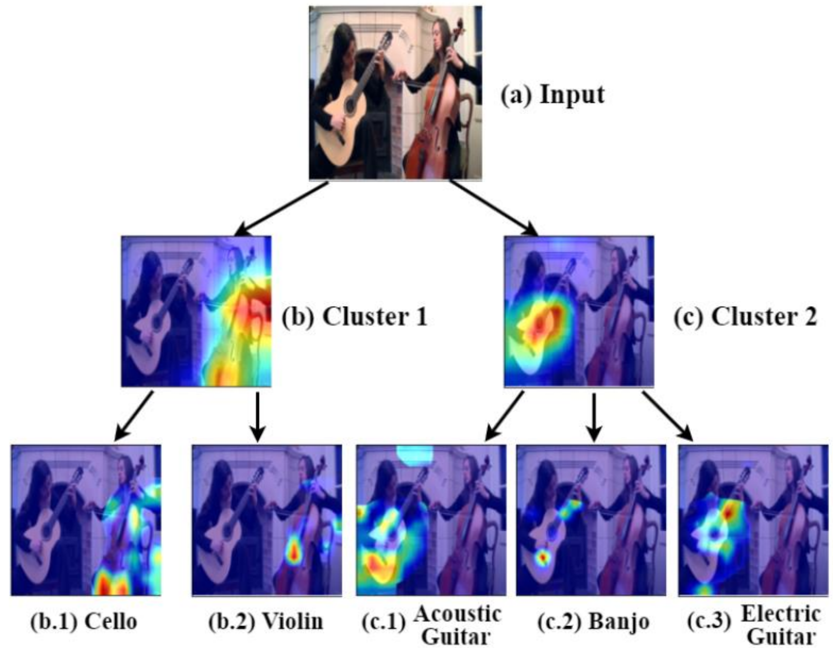


- SWOX: The evidence from a-guitar, e-guitar and bangjo are from the left side of the images.
  - But it is not clear what is the evidence for each of them.
- One-stage CWOX (CWOX-1s): Subtract the heatmap of each class by the heatmap of all others, we get the CWOX-1s heatmaps.
  - Still, it is not clear what is the evidence for each of them.

Two Stages Necessary in CWOX

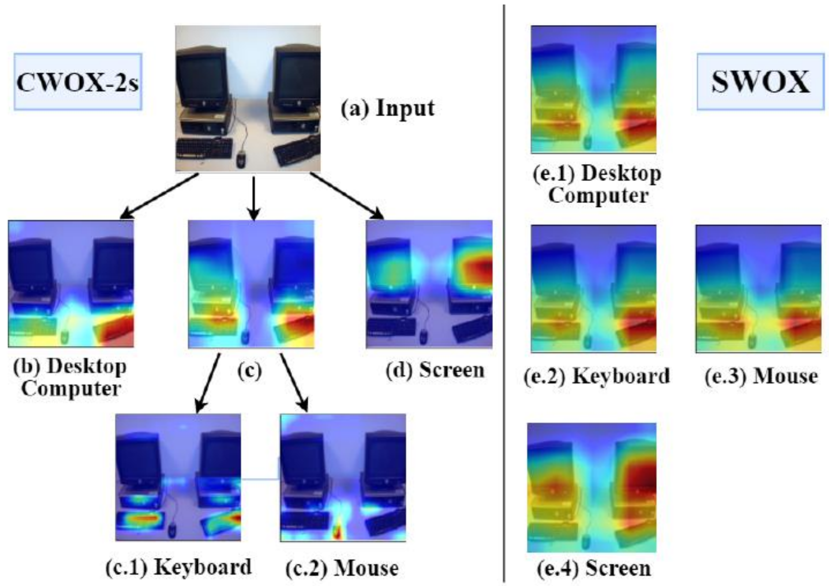
Two-stage CWOX (CWOX-2s): Divide the top classes into two clusters

1. Contrast the two clusters
2. Contrast classes within each cluster



- It is very clear that the evidence for the two clusters are from the left and right parts of the image respectively.
- Within the first cluster: The relative evidence for cello is the lower body; The relative evidence for violin is the middle section of the strings.
- Within the second cluster: The relative evidence for a-guitar is the lower body; The relative evidence for e-guitar is the strings. The relative evidence for Banjo is the bridge

Example: CWOX-2s vs SWOX



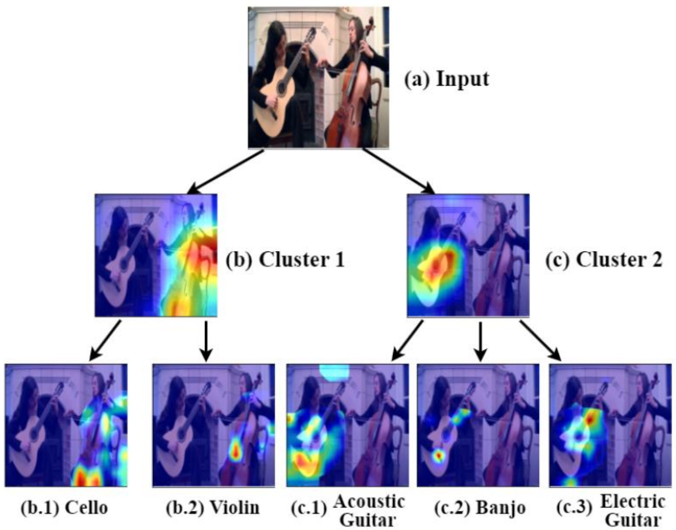
- The SWOX saliency maps are not visually discriminative.
- CWOX-2s
  - Divides the top classes into three clusters, with keyboard and mouse in one cluster, and the other two classes in two separate clusters.
  - Clearly reveals the evidence for mouse and keyboard.

How to divide the top classes into clusters?

Why two different classes can be given high probabilities?

- They are competing labels for the same object in an image (e.g. cello, violin)
  - They co-occur as top classes whenever the object is present
  - Their occurrences share a common **latent cause**, i.e., the object

  Common cause principle: When two variables are correlated, there exists a latent cause that influences both of them.



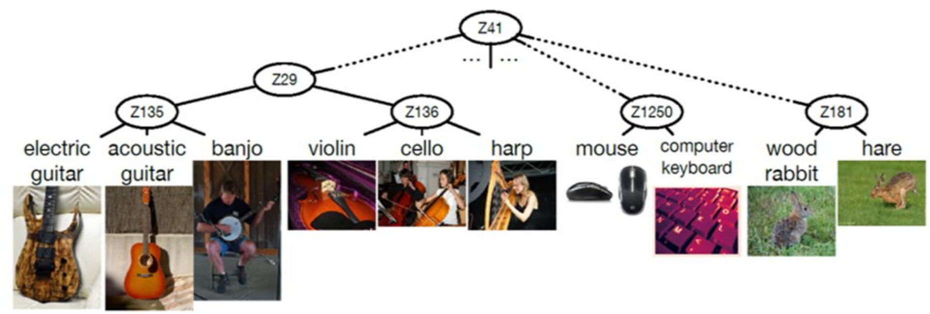
- They correspond to different objects in an image (e.g., cello, a-guitar)
  - They co-occur top classes only when both objects are present
  - Not as often as the first case

Learning the Latent Causes behind Top Classes

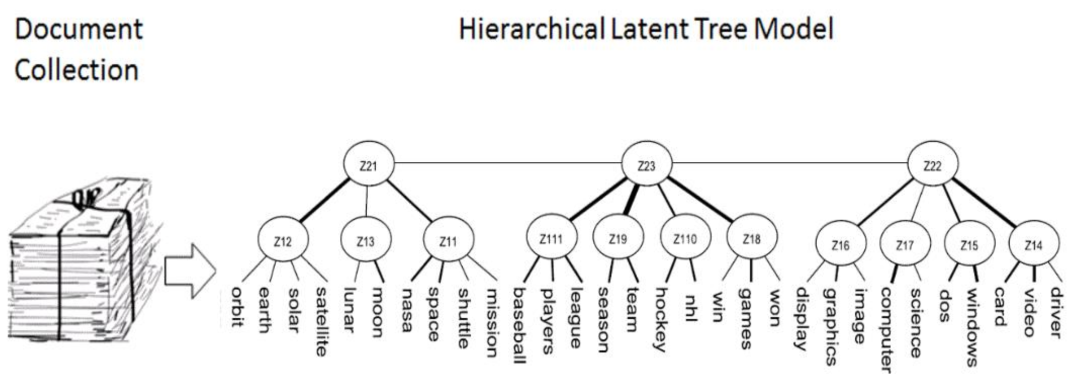
- Apply classifier to a set of images
  - A collection of short documents, each consisting of the top classes of an image

Image	Classifier	Top Classes
1	ResNet50	cello, robin, violin
3	ResNet50	cello, violin, gown, trombone
4	ResNet50	acoustic guitar, electric guitar
6	ResNet50	acoustic guitar, electric guitar, banjo
…	……	……

- Analyze the document collection to get a hierarchical latent tree model (HLTM)
  - Latent variable Z135 introduced because e-guitar, a-guitar and banjo tend to co-occur in the docs
  - Z136 introduced because violin and cello tend to co-occur in the docs
  - …



Hierarchical Latent Tree Models (HLTM)



ai-tree.pdf

- The level-1 latent variables model word co-occurrence patterns
- Those at higher levels model co-occurrences of patterns at the level below.

It is an ideal tool for our task.

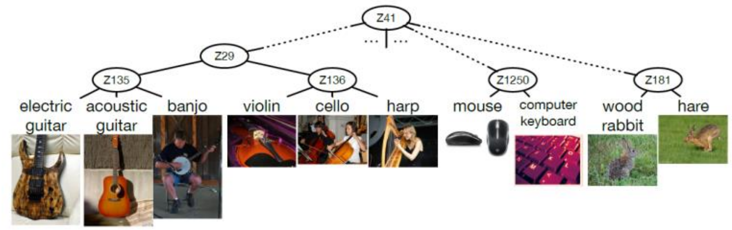
P. Chen, N.L. Zhang, et al. Latent Tree Models for Hierarchical Topic Detection. Artificial Intelligence, 250:105-124, 2017.

Creating Contrastive Explanations

ResNet50

Cello (0.839),
Acoustic Guitar (0.081),
Banjo (0.036),
Violin (0.021),
Electric Guitar (0.008)

+



restrict

cut at level 1

Confusion clusters

Cluster 1:
Cello, Violin

Cluster 2:
Acoustic Guitar (A)
Banjo (B)
Electric Guitar (E)

CWOX-2S

Input: A test example $\mathbf{x}$; a base explainer.
Do:

1: Feed $\mathbf{x}$ to m to get a list of top class labels.
2: Restrict T to those labels to get a subtree
3: Partition the labels into confusion clusters by cutting the subtree at level 1.
4: Create a heatmap to contrast each confusion cluster against other clusters using Equation (1).
5: In each cluster, create a heatmap to contrast each class in the cluster against other classes using Equation (2).

I confusion clusters $\mathbf{C} = \{\mathbf{C}_1, \ldots, \mathbf{C}_I\}$

each cluster $\mathbf{C}_i$ consists of J_i class labels $\mathbf{C}_i = \{c_{i1}, \ldots, c_{iJ_i}\}$.

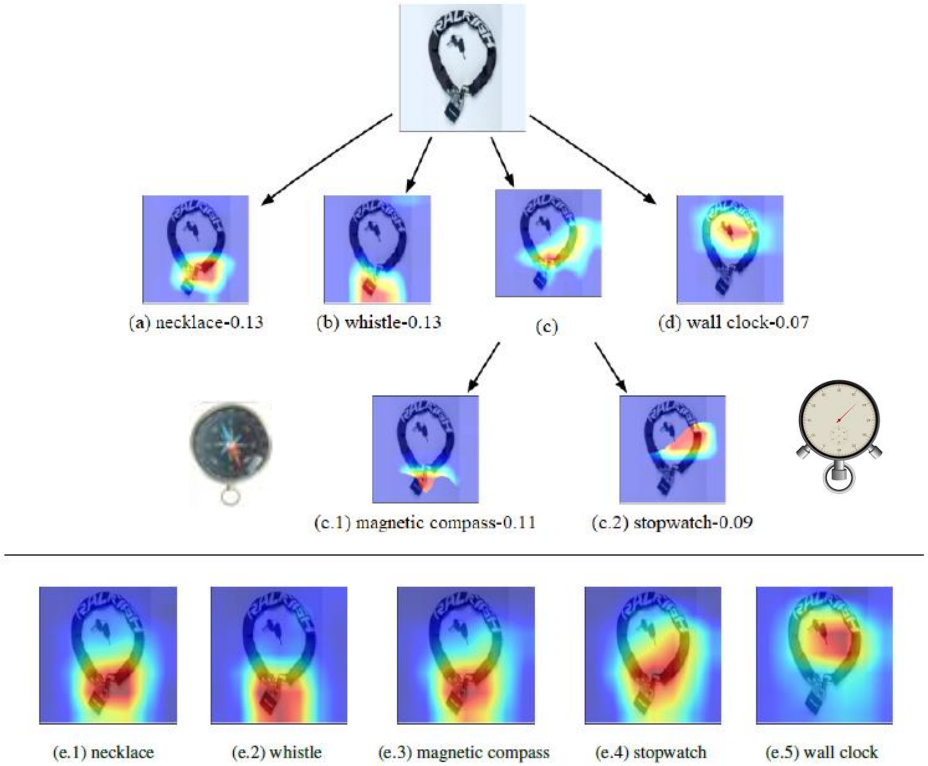
$$\hat{H}_{\mathbf{C}_i} = \begin{cases} ReLU[H_{\mathbf{C}_i} - H_{\mathbf{C}\setminus\mathbf{C}_i}] & \text{if } I > 1; \\ H_{\mathbf{C}_i} & \text{if } I = 1, \end{cases}$$

$$\hat{H}_{c_{ij}} = \begin{cases} \text{supp}(\hat{H}_{\mathbf{C}_i}) \times ReLU[H_{c_{ij}} - H_{\mathbf{C}_i \setminus c_{ij}}] & \text{if } J_i > 1; \\ \text{supp}(\hat{H}_{\mathbf{C}_i}) \times H_{c_{ij}} & \text{if } J_i = 1. \end{cases}$$

Heatmaps on the right hand sides created by a base explainer

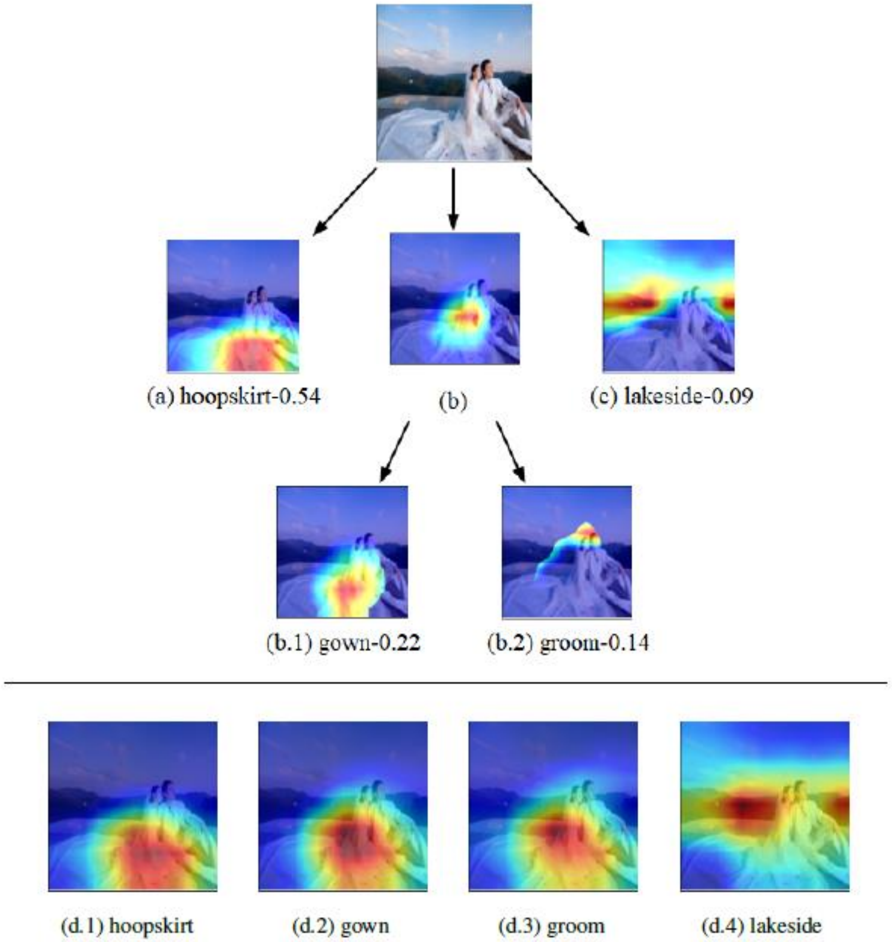
More Examples

- True label: Padlock
- CWOX-2s reveals reasonable evidence why ResNet50 got it all wrong
  - Wall clock: Two keys similar to hands on clock
  - Whistle: Similar appearance to lock
  - Necklace: The chain
  - Compass and stopwatch: Part of the ring



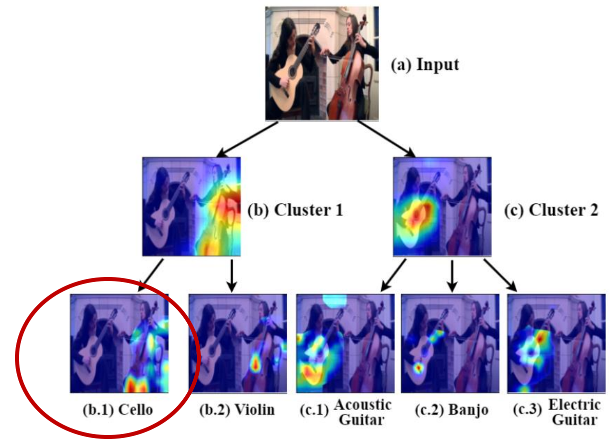
More Examples

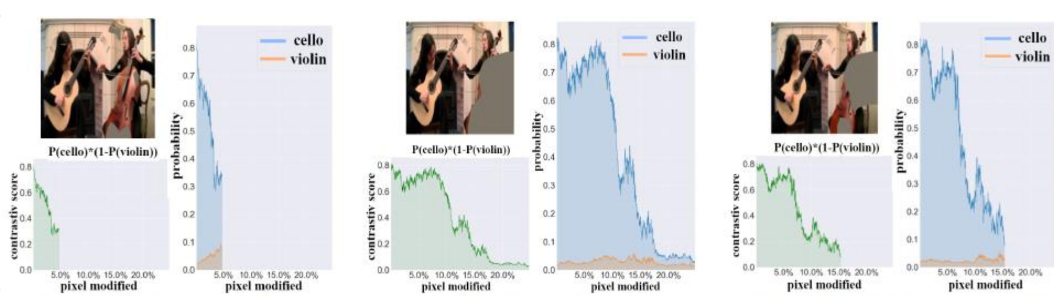
- True label: Gown
- CWOX-2s reveals reasonable evidence why ResNet50 got it wrong
  - Hoopskirt: The large skirt
  - Lakeside: The lake side
  - Gown and Groom: The couple
    - Groom: Head of male
    - Gown: Gown and female



Quantitative Evaluation

- (b.1) reveal the evidence for cello against violin. As we erasing the highlighted pixels, the probability of cell goes down and the prob of cell goes up
- Contrastive score: P(cello) (1-P(violin)).
- Contrastive AUC (CAUC): Area under the contrastive score curve
- CAUC for CWOX-2s is much smaller that those of SWOX and CWOX-1s





(i) CWOX-2s: $n_\delta = 2,377$, CAUC = 0.023, CDROP = 0.505; (ii) SWOX: $n_\delta = 12,402$, CAUC = 0.033, CDROP = 0.309; (iii) CWOX-1sA: $n_\delta = 7,686$, CAUC = 0.033, CDROP = 0.364.

Figure 6: Changes in the probabilities $P(\texttt{cello})$ and $P(\texttt{violin})$ and the contrastive score $P(\texttt{cello}) \times (1 - P(\texttt{violin}))$ as δ-salient pixels are deleted according to the order induced by: (i) the CWOX-2s heatmap in Fig. 3 (b.1); (ii) the SWOX heatmap in Fig. 2 (b.1); and (iii) the CWOX-1sA heatmap in Fig. 2 (c.1).

Quantitative Evaluation

Table 1: Average CAUC scores on the ImageNet examples (**smaller** ↓ CAUC indicates better contrastive faithfulness).

	ResNet50		GoogleNet	
	Grad-CAM	RISE	Grad-CAM	RISE
SWOX	7.54×10^{-3}	5.18×10^{-3}	5.93×10^{-3}	3.36×10^{-3}
CWOX-1sA	7.19×10^{-3}	4.65×10^{-3}	5.37×10^{-3}	3.12×10^{-3}
CWOX-1sB	7.68×10^{-3}	4.96×10^{-3}	6.12×10^{-3}	3.24×10^{-3}
CWOX-2s	$\mathbf{5.78 \times 10^{-3}}$	$\mathbf{4.08 \times 10^{-3}}$	$\mathbf{4.47 \times 10^{-3}}$	$\mathbf{2.78 \times 10^{-3}}$

- Overall performance on a subset of randomly selected 10,000 images from the ImageNet validation set.
- CWOX-2x achieves the smallest CAUC for both ResNet50 and GoogleNet, and with both Grad-CAM and RISE as the base explainer.

User Study

- Forward simulation [Hase & Bansal 2020]:
  - Given an input and an explanation, users must predict what a model would output for the given input
- Our setup:
  - Given input and heatmaps for two confusing labels, user must match the labels with the heatmaps.
- With CWOX-2s, users can do the matching with higher accuracy and confidence
  - True for both the expert and non-expert groups.

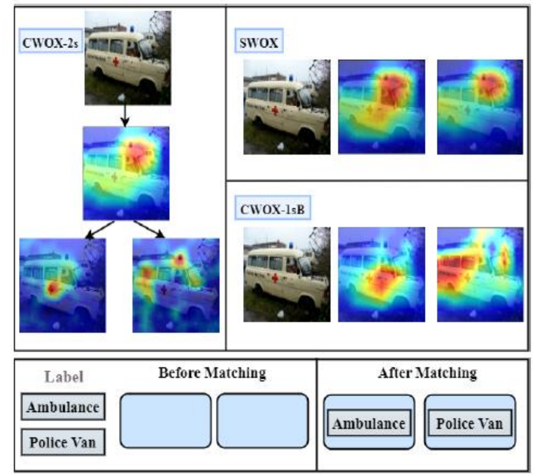


Figure 7: In the user study, heatmaps for pairs of confusing labels are displayed. A user is asked to match the labels with the heatmaps.

Table 4: Results of the user study in the expert group (± 95% confidence interval).

	SWOX	CWOX-1sB	CWOX-2s
Accuracy	0.45±0.048	0.57±0.088	**0.83**±0.092
Confidence	1.60±0.241	2.60±0.241	**3.60**±0.237

Table 5: Results of the user study in the non-expert group (± 95% confidence interval).

	SWOX	CWOX-1sB	CWOX-2s
Accuracy	0.40±0.075	0.51±0.102	**0.75**±0.119
Confidence	1.40±0.108	2.80±0.172	**3.40**±0.163

IJCAI 2023

ViT-CX: Causal Explanation of Vision Transformers

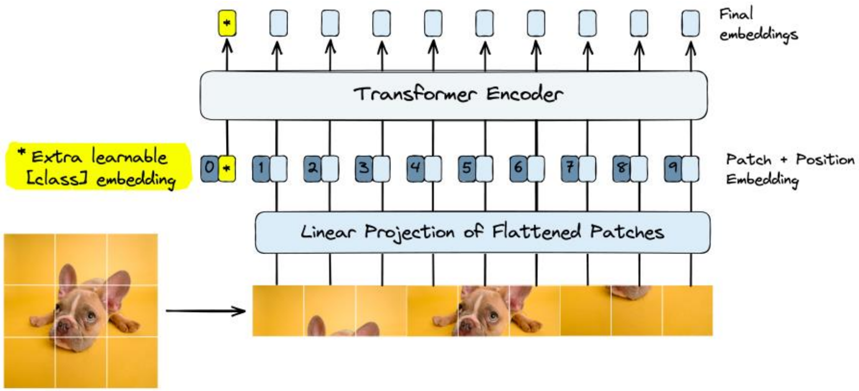
Weiyan Xie [1], **Xiao-Hui Li** [2], **Caleb Chen Cao** [1] and **Nevin L. Zhang** [1]

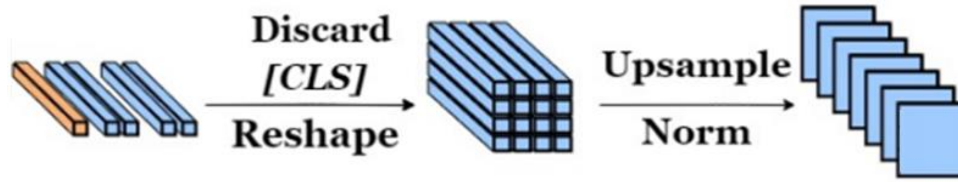
[1] The Hong Kong University of Science and Technology, China
[2] Huawei Technologies Co., Ltd, China
{wxieai, cao, lzhang}@ust.hk, {lixiaohui33}@huawei.com

ViT-CX

1. Extract ViT feature maps and use them as masks
2. Determine the causal impact of the masks on prediction
3. Aggregate the masks with their causal impact scores to create saliency maps

ViT Feature Maps and Masks





- Output layer of ViT: Embeddings of [CLS] and the patches

- Arrange patch embeddings as a 3D tensor, and use its frontal slices as feature maps

- Normalized to [0, 1] to get masks.

  0 – black, 1 – white.

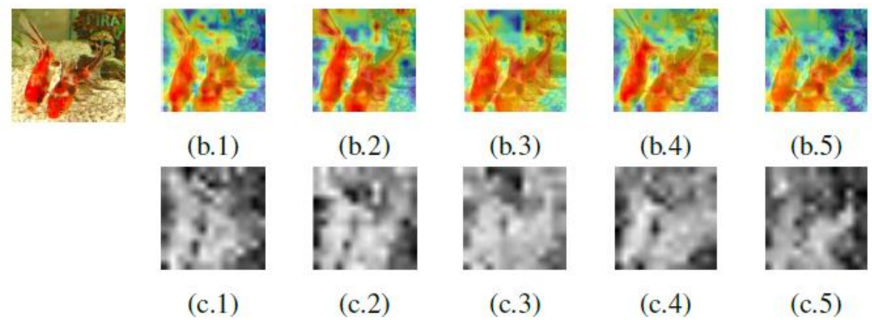
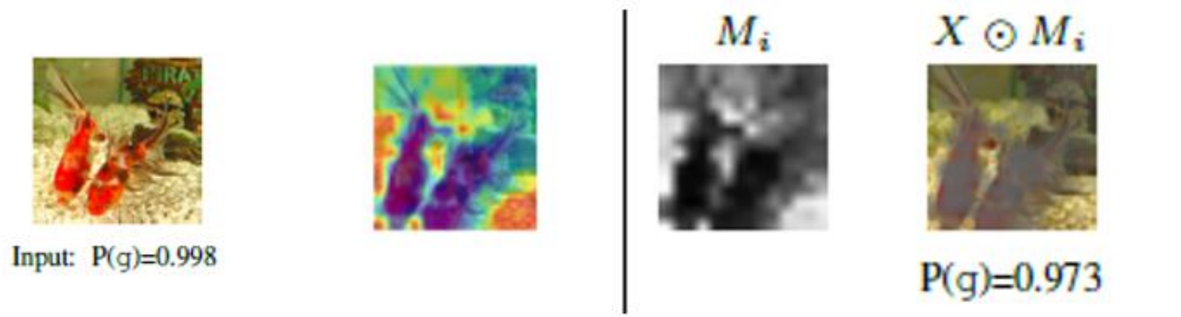


Figure 2: ViT feature maps (b.1 - b.5) are frontal slides of a 3D tensor made up of patch embedding vectors (as fibers). They are used as ViT masks (c.1 - c.5) to generate explanations.

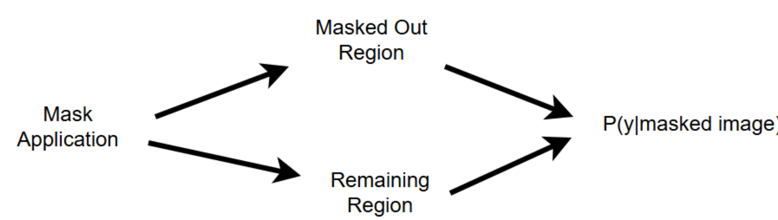
Potential outcome of Applying Mask (Treatment)

- Applying mask: pointwise product
  - Pixels with mask value close to 1 are kept
  - Pixel with mask value close to 0 are erased

X — image; M_i — a mask; $X \odot M_i$ — Masked image (pointwise product)



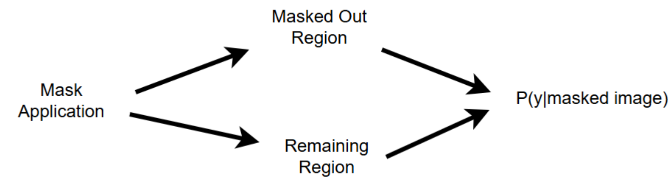
- Causal inference perspective
  - Treatment: Mask application
  - We want: Probability of gold fish of remaining region
  - Not the same as probability of gold fish of entire masked image, which has two causal paths
    - Masked-out region form silhouette of gold fish (artifact), contributing high probability



Potential outcome of Applying Mask (Treatment)

To apply **backdoor adjustment** on the masked out region, we can sample noises ϵ_{ij} $(j = 1, \ldots, J)$ for masked out pixels and estimate the potential outcome score as follows:

$$s(X, y, M_i) \approx \frac{1}{J}\sum_{j=1}^{J} P(y|X \odot M_i + \epsilon_{ij})$$



For efficiency, set $J = 1$: $s(X, y, M_i) \approx P(y|X \odot M_i + \epsilon_i)$.

To reduce variance, subtract the treatment effect of ϵ_i on the whole image:

$$s(X, y, M_i) \approx P(y|X \odot M_i + \epsilon_i) - [P(y|X + \epsilon_i) - P(y|X)]$$

Prob of gold fish without backdoor adjustment : 0.973, Corrected score: 0.012

Saliency Determination

Let us regard each mask M_i is a "team" of pixels.

$s(X, y, M_i)$ is an estimation of model score achieved by the entire "team".

We define the importance of a pixel x as the total scores achieved by all teams it is part of, weighted by its membership in each team, and the total number of teams it participates in:

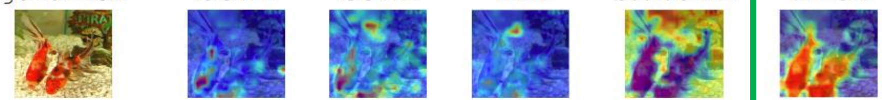
$$S(x) = \frac{1}{\rho(x)} \sum_{i=1}^{K} s(X, y, M_i) M_i(x),$$

where $M_1, \ldots, M_K$ are the masks and $\rho(x) = \sum_{i=1}^{K} M_i(x)$

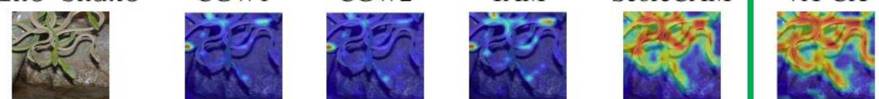
Pixel coverage bias (PCB) correction

Results

- Previous methods:
  - Designed for ViT Explanation: CGW: [Chefer et al., 2021] , TAM: [Yuan et al., 2021], etc
  - Can be adapted to ViT: Grad-CAM,ScoreCAM, RISE, etc
- ViT-CX explanations more meaningful to users than those previous methods
  - Highlighting the regions apparently important to predictions.
- ViT-CX more faithful to the model as measured by the deletion AUC (Del) insertion AUC (Ins), and point games (PG) accuracy metrics



goldfish	CGW1	CGW2	TAM	ScoreCAM	ViT-CX
Del↓	0.258	0.355	0.271	0.532	**0.202**
Ins↑	0.829	0.833	0.866	0.553	**0.879**



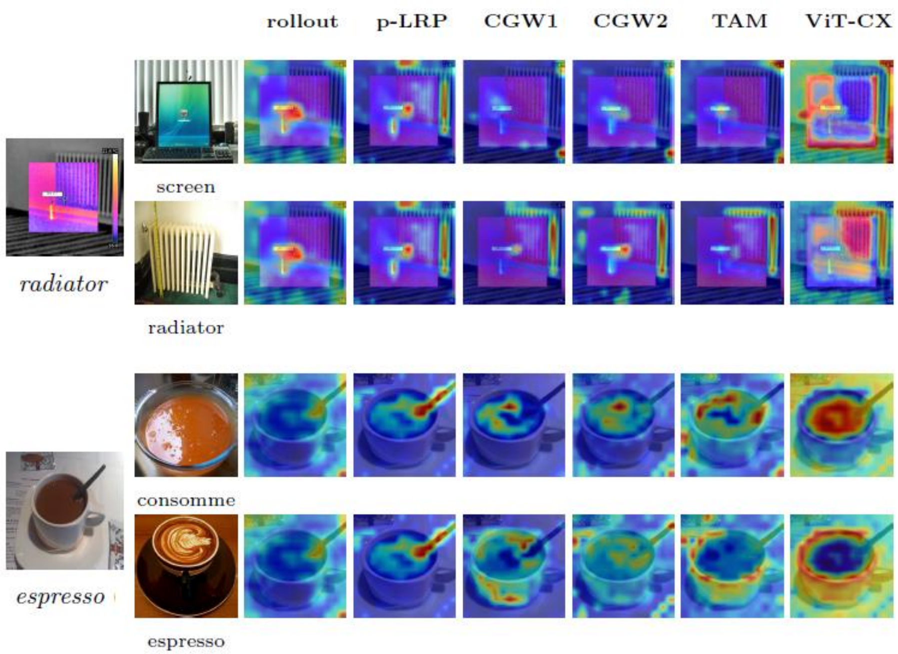
vine snake	CGW1	CGW2	TAM	ScoreCAM	ViT-CX
Del↓	0.164	0.122	0.114	0.108	**0.106**
Ins↑	0.410	0.544	0.337	0.598	**0.603**

Results on 5,000 images from ImageNet Validation Set

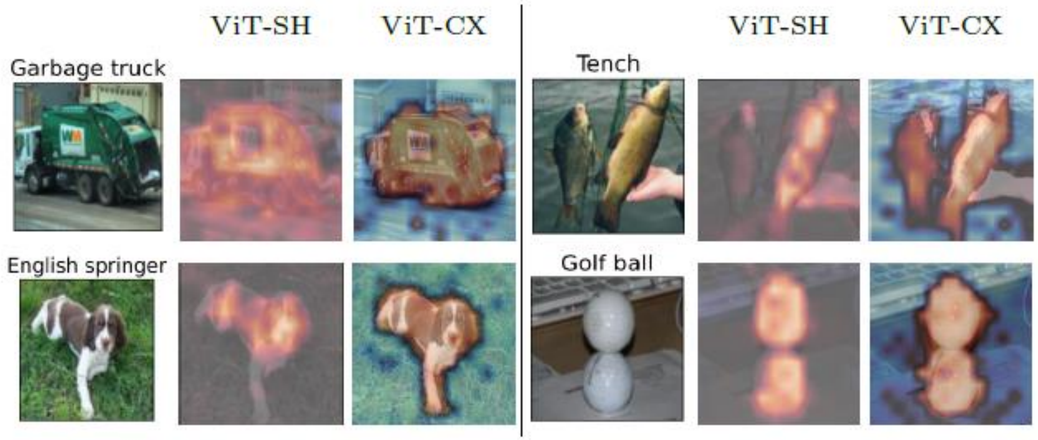
	ViT-B			DeiT-B		
	Del ↓	Ins ↑	PG Acc ↑	Del ↓	Ins ↑	PG Acc ↑
ViT-CX	**0.161**	**0.620**	**86.42%**	**0.211**	**0.802**	**86.93%**
Number of Masks	Average: 63, Std: 11			Average: 70, Std:12		
Rollout	0.251	0.517	60.91%	0.406	0.642	35.70%
Partial LRP	0.239	0.499	66.52%	0.349	0.655	61.25%
CGW1	0.201	0.542	77.14%	0.286	0.717	70.54%
CGW2	0.209	0.549	70.94%	0.271	0.736	70.54%
TAM	0.180	0.556	77.87%	0.240	0.747	75.47%
Occlusion	0.291	0.571	64.75%	0.380	0.801	59.51%
RISE	0.234	0.581	73.30%	0.366	0.759	71.84%
Score-CAM	0.291	0.471	48.89%	0.439	0.576	50.12%
Grad-CAM	0.212	0.456	50.45%	0.250	0.743	79.24 %
Integrated-Grad	0.184	0.263	10.61%	0.259	0.362	10.74%
Smooth-Grad	0.174	0.438	16.96%	0.231	0.528	31.05%

Results: Understanding Model (ViT-B) Mistakes

- ViT-CX can also help ML experts understand model mistakes better
  - It clearly reveals the evidence for
    - The predicted labels: screen, consommé
    - The correct label: radiator, expresson
- In contrast, the evidence revealed by other methods are less discriminative.



Comparisons with ViT Shapley

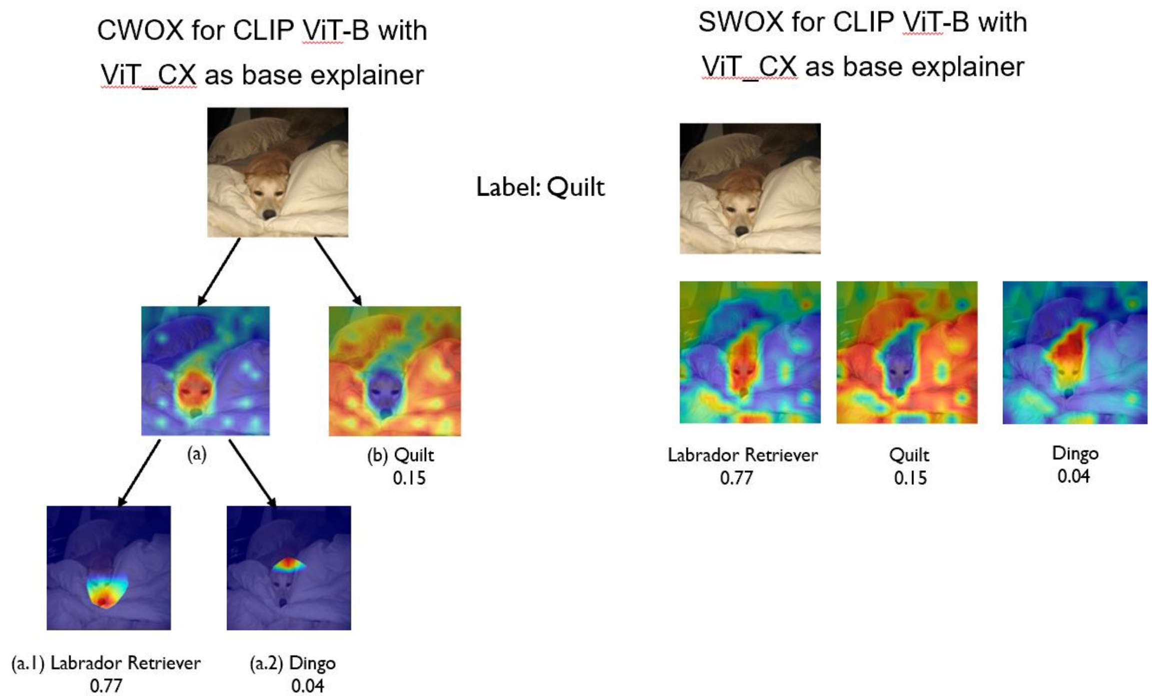


	Del AUC ↓	Ins AUC ↑
ViT-SH	0.691 (0.014)	0.985 (0.002)
ViT-CX	**0.598 (0.016)**	0.981 (0.001)

- ViT-CX outperforms ViT-SH in terms of deletion AUC, and visually better
  - In terms of insertion AUC, both methods achieved close to the best possible value 1.
- ViT-SH requires training a separate model, which is time-consuming and done only for 10 classes of ImageNet

ViT Shapley: [Covert et al. 2023] Learning to estimate shapley values with vision transformers

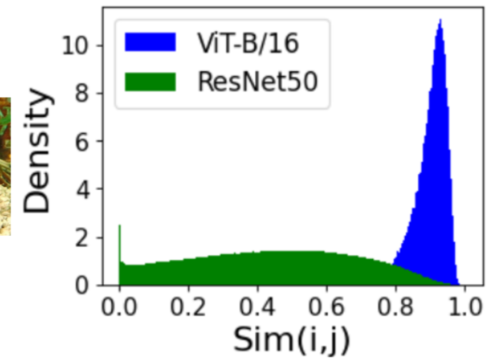
Combining CWOX and ViT-CX

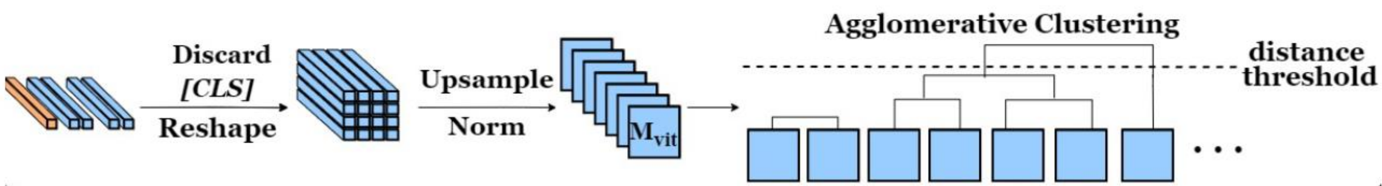


- When ViT-CX is used as the base explainer, CWOX is better than SWOX at revealing the relative evidence of classes in the same confusion cluster.

Improve Efficiency by Clustering Masks

- ViT masks are more similar to each other than CNN masks
- Clustering the ViT masks improves efficiency of explanation

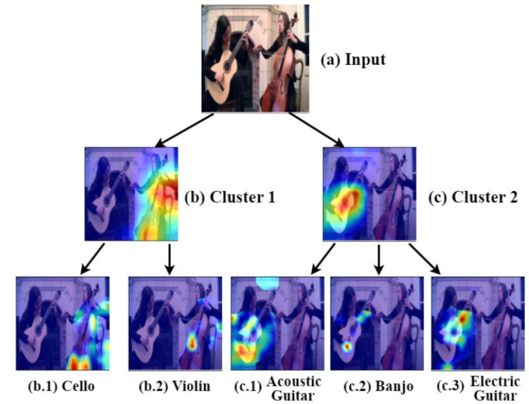




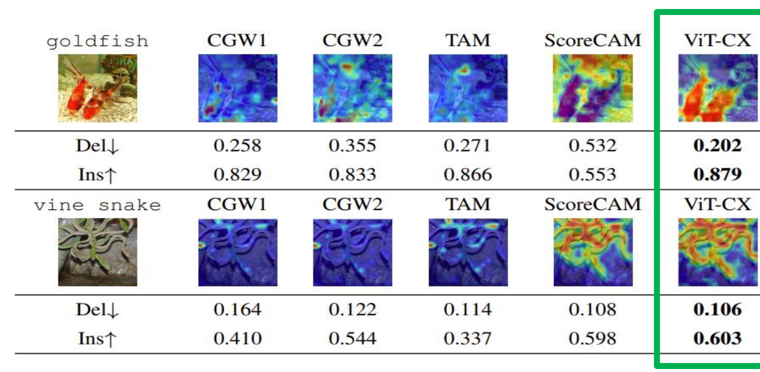
Masks	Number of Masks	Del ↓	Ins ↑	PG Acc ↑	Average Time (s)
$\mathbb{M}_{cx}$	70 ± 12	**0.211**	0.802	**86.93%**	1.15 ± 0.15
$\mathbb{M}_{vit}$	768 ± 0	0.232	**0.810**	85.52%	8.23 ± 0.03
$\mathbb{M}_{random}$	5000 ± 0	0.323	0.734	75.12%	77.78 ± 3.46

Summary

- CWOX
  - Explain all top classes in two stages
    - Contrast confusion clusters
    - Contrast classes within each cluster
  - HLTM is ideal for confusion cluster determination
  - Bears some resemblance to Argumentative XAI?



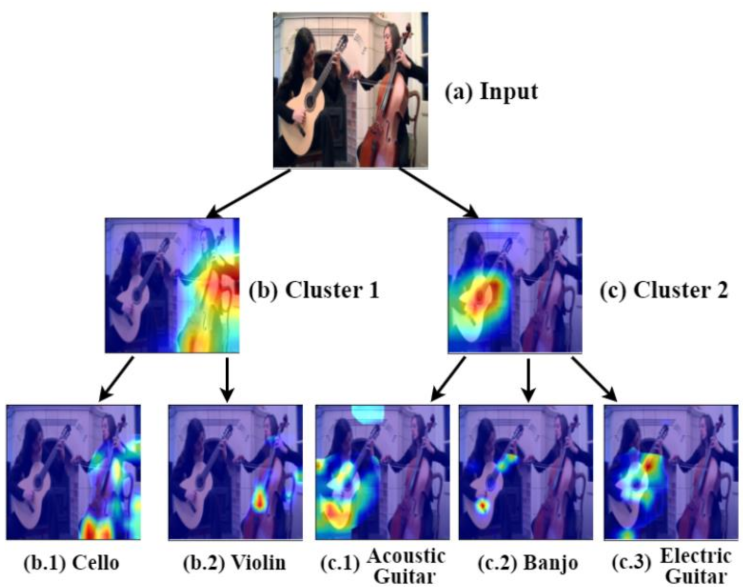
- ViT-CX
  - ViT-feature maps as masks
  - Backdoor adjustment for masked out region (artifacts).
  - Pixel coverage bias correction
  - ViT masks allow clustering to achieve high explanation efficiency



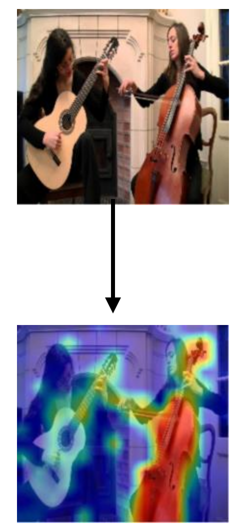
Thanks for Your Attention!

Combining CWOX and ViT-CX

CWOX for ResNet50 with
Grad-CAM as base explainer



CWOX for ViT-B with
ViT_CX as base explainer



Prediction:
[486, 'cello', 0.9729557]
[402, 'acoustic_guitar', 0.0131826075]
[889, 'violin', 0.0060735513]
[420, 'banjo', 0.0018252619]
[594, 'harp', 0.0014889699]