Cut the Black Box

Michèle Sebag
CNRS & Université Paris-Saclay

Jne 27th, 2023

Joint work: Nicolas Atienza, Roman Bresson, Cyriaque Rousselot, Philippe Caillou, Johanne Cohen, Christophe Labreuche

The AI wave faces a shock

- Why ? Lack of certification; fairness; accuracy; **explanations**.
  - Ex:
    Model (Correlation between):
        computers/books at home;
        children good grades at school
    Decision (Public policy): give computers/books to families

The dark side of AI:

C. O'Neill, 2016	Weapons of Math Destruction
Timnit Gebru, 2020	`www.technologyreview.com/2020/12/04/1013294/google-ai-ethics-research-paper-forced-out-timnit-gebru`

Explainable models

Strategies

1. Learning an explainable model from scratch Rudin 2019
2. Explaining a black-box model H (post-hoc explanation)

Position of the problem

- Option 1: requires interpretable representation / simple models;
  Throwing away existing black-box models ?
  Trade-off Explanation / Accuracy ?

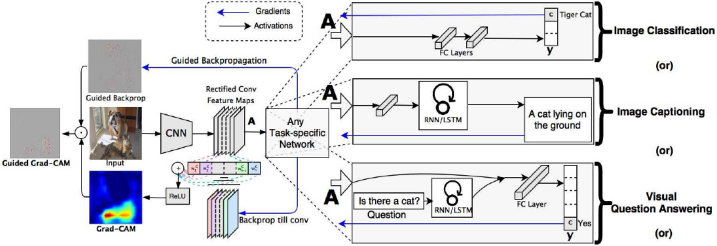
- Option 2 comes in two modes:
  * explaining $H(x)$
  * explaining H

Explaining $f(x)$

Saliency approaches

Class Activation Mapping

Gradient-based — Selvaraju et al. 17



Shapley value of attribute j wrt model H

$$\hat{\phi}_j = \frac{1}{T}\sum_{t=1}^{T}\left(H\left(x_{+j}^{t}\right) - H\left(x_{-j}^{t}\right)\right)$$

Discussion

Confirmation bias

Fig. 2 | Saliency does not explain anything except where the network is looking. We have no idea why this image is labelled as either a dog or a musical instrument when considering only saliency. The explanations look essentially the same for both classes. Credit: Chaofen Chen, Duke University

Gradients only tell where the network is looking.

Desired properties

Alvarez-Meliz et al., 2018

- **Explicitness/Intelligibility**: Are the explanations immediate and understandable?
- **Faithfulness**: Are relevance scores indicative of "true" importance?
- **Stability**: How consistent are the explanations for similar/neighboring examples?

Concept Activation Vector

Kim et al., 2018; Crabbé & v.d. Schaar 22

Input (CAV)

- a black-box $H : X \mapsto Y$
- a set of concepts
- positive/negative examples for each concept i

Method

- Learn classifier h_i for concept i in latent representation of H (noted $z(x)$)
- Assess correlations between:
  - how much x needs be changed to modify $h_i(z(x))$;
  - how much this modification would change the label $H(x)$

Overview of CBB

Building upon multi-modal NNs

Radford et al. 2021

- ϕ_i : image $\mapsto \mathbb{R}^d$
- ϕ_c : concepts $\mapsto \mathbb{R}^d$

Given concept space and its grounding w.r.t. example space X

- Dictionary $C = \{c_1, \ldots c_K\}$
- Grounding

$$\Phi : X \times C \mapsto \mathbb{R}$$

e.g. $\Phi(\text{image of zebra, striped}) = 1.$

CBB

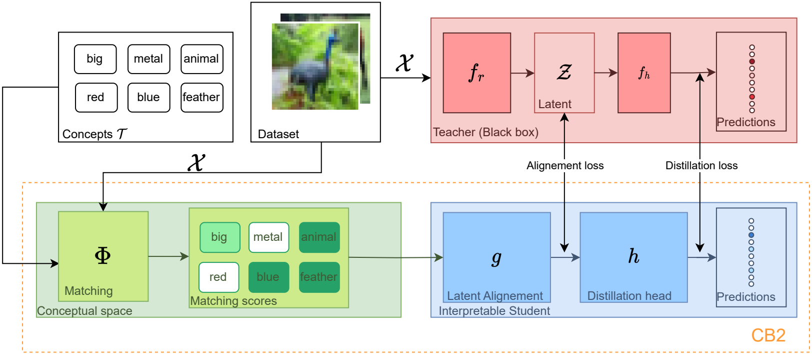
- Given a teacher H (black-box neural net)

$$H = f_h \, o \, f_r : X \mapsto Y$$

- Find explainable students, explaining:
  - The latent representation f_r
  - The classifier f_h

Kim et al. 1018

Overview of CBB



- Phase 1: explaining f_r: inspiration TCAV

  Kim et al. 18, Crabbé vd Schaar 22

- Phase 2: explaining f_h with Hierarchical Choquet integral

  Bresson et al 19, 20

Phase 1: Explaining latent representation

Given

- Sample x and conceptual representation $c(x) = (\Phi(x, c_i))_i$
- Latent representation $f_r : X \mapsto Z$

Find

$$W = \arg\min \|f_r - W.c\|_2 + \|W\|_1$$

with matrix $W = (\#C, \#Z)$

On-going experiments

- Explaining the full latent representation or each coordinate ?
- Linear student ? Non-negative W ?

Phase 2: Explaining latent classifier

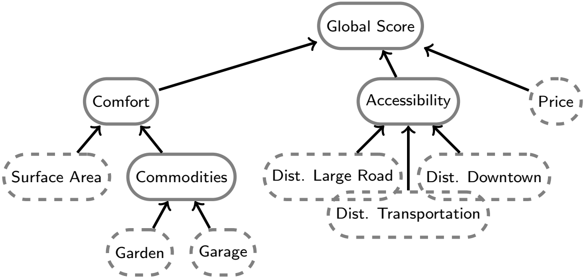
Hierarchical Choquet Integral, recap

Bresson et al. 20, 21

- Variable x_i in domain X_i
- Utility functions $u_i : X_i \mapsto \mathbb{R}$
  (continuous; monotonic, peak-shaped or valley-shape)
- Aggregation: Choquet integral

$$C_\mu(a) = \sum_{i=1}^{N} \mu(\{\tau(i), \tau(i+1), ..., \tau(n)\})(a_{\tau(i)} - a_{\tau(i-1)}) \qquad (1)$$

with τ a permutation in N s.t. $\forall i \in N$, $a_{\tau(i)} \leq a_{\tau(i+1)}$ and $a_{\tau(0)} = 0$.



- $S_{Global} = \mathcal{A}_{Global}(S_{Comfort}, S_{Accessibility}, S_{Price})$
- $S_{Comfort} = \mathcal{A}_{Comfort}(S_{Area}, S_{Commodities})$
- ...

Neural- HCI

Properties UHCIs — Grabisch & Labreuche 08

- continuous
- non-decreasing w.r.t. arguments
- piecewise linearity
- interpretable
- 1-Lipschitz

Past Results — Bresson et al. 2020, 2021

- Neur-HCI can learn HCI (HCI constraints satisfied by design)
- Identifiability in the large sample limit (with given hierarchy)
- On-going: learning the hierarchy

Phase 2: Explaining latent classifier, 2

Explain $f_c : Z \mapsto Y$

- remind: f_c: linear + softmax
- HCI: Find $h^* = \arg\min_{h \ in \ HCI}$ Distillation loss $\mathcal{L}(h, f_c)$
- MLP: Find $h^* = \arg\min_{h \ in \ MLP}$ Distillation loss $\mathcal{L}(h, f_c)$
- with Hinton et al, 215

$$\mathcal{L}(u, v) = \text{ Cross Entropy } (\sigma(u/T), \sigma(v/T))$$

σ: softmax, T a temperature parameter

The HCI case

- HCI Hierarchy = hierarchical clustering of concepts in dictionary $\mathcal{C}$ based on latent representation of samples

Experimental setting

Benchmark: CIFAR-10 Teachers

- CIFAR-10
- resnet20 and resnet32 (ResNet).
- mobilenetv2_x0.5 (MobileNet)
- repvgg_a0, vgg16_bn (VGG).

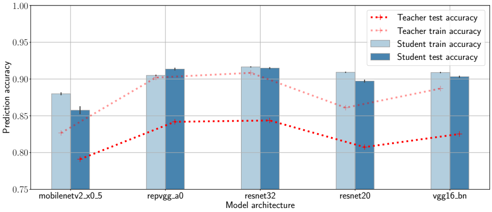
Dictionary and grounding

- use multi-modal embedding CLIP Radford et al. 2021
- $\Phi(\text{image } x, \text{ concept } c) = \text{cosine}(\phi_x(x), \phi_c(c))$
- concepts: 2096 most common English terms (filtering out class synonyms to avoid tautological explanations)

Performance indicators of students

- Accuracy wrt ground truth labels
- Faithfulness wrt teachers
- Inspecting students

Accuracy (on validation set)



No loss of accuracy wrt Teachers

On-going

- Sensitivity wrt size of Student training set.

Accuracy and Faithfulness wrt Teachers

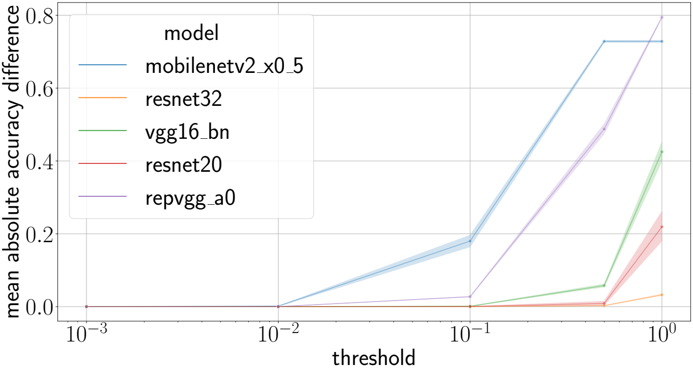
Acc.	MLP Head	Linear head
truth	91.84 $\pm$ 0.05	90.9 $\pm$ 0.02
	90.89 ± 0.01	89.66 ± 0.09
teacher	82.54 $\pm$ 0.05	82.02 $\pm$ 0.02
	78.90 ± 0.05	78.58 ± 0.04

(plain, training set; italic, test set)

Computing time: $\sim$ 50 minutes, for 30 epochs, 8 Tesla V100 16GB GPUs

Impact of sparsity on accuracy: lesion study

Removing concepts with $|\text{weight}| < x$ coordinate $\rightarrow$ loss of accuracy y coordinate



Case study

Relative sensitivity of class c_i wrt concept t_j

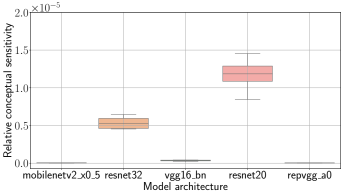
Zhou et al. 2018

where $z \sim W_g t$ and $f_h \sim W_h z$

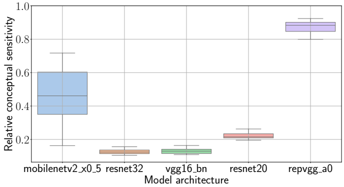
$$\text{Define } S[i,j] := \left(W_h^T W_g^T)\right)[i,j]$$

$$\text{Relative sensitivity } = \frac{exp(S[i,j]}{\sum_k exp(S[i,k])}$$

Sensitivity of 'airplane' w.r.t. 'grass'



Sensitivity of 'ship' w.r.t. 'sea'



Position of the problem

Overview of Cut the Black Box (CB2)

Experimental validation

Conclusion

Conclusion

Pros and Cons

- Students suffer no loss of accuracy
- Are they really interpretable ?
  (tells what's in z and how to pass from z to y)
- Using Shapley value to infer biases from background ('sea' for 'ship')

Compared to learning from $c(x)$?

- frugality

Perspectives

- Distill several hidden layers ?
- Impact on adversarial examples
- Automatically detect spurious inference (external sources to assess causality ?)
- Adapt/extend to opinion mining