Interpreting the Inner Workings of Deep Neural Networks through Concept-based Explanation

Cyriaque Rousselot

June 21, 2023

Our Working Group

- Nicolas Atienza
- Roman Bresson
- Philippe Caillou
- Johanne Cohen
- Christophe Labreuche
- Michele Sebag

Algorithms are everywhere 1/3

Suppose that you want to buy a house. You go to the bank to get a loan. They accept you based on an algorithm.

Figure 1: You

Figure 2: The bank

Figure 3: Your dream House

Algorithms are everywhere 2/3

Suppose that you want to get a job. You candidate. They accept you based on an algorithm.

Figure 4: You

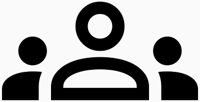

Figure 5: Recruiters

Figure 6: Your dream Job

Algorithms are everywhere 3/3

Nowadays, many people want to rely on autonomous systems to provide you with content, service, medical assistance...

Figure 7: You

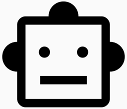

Figure 8: Automatic Systems

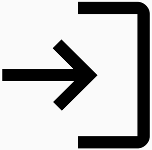

Figure 9: Access to a ressource

Examples of harm caused by algorithms

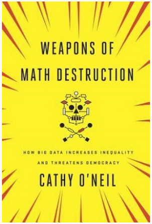

Figure 10: Danger of letting the control to algorithms and systems to take decisions

Figure 11: Adversarial Attacks (Akhtar et al. 2021)

Figure 12: Catastrophic Errors (BreezoMeter)

Black Box models

Proprietary

Figure 13: Youtube's recommendation algorithm is proprietary thus black box

Too Complex

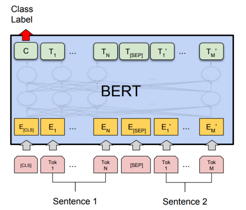

Figure 14: Even if you have access to a model, it does not mean it is interpretable

Whitening the black box

Solutions

1. Proprietary : Reconstruction attacks (Balle et al. 2022)
2. Building interpretable models from scratch (Zhang Zhu, 2018)
3. Explanation methods

Whitening the black box

Solutions

1. Proprietary : Reconstruction attacks (Balle et al. 2022)
2. Building interpretable models from scratch (Zhang Zhu, 2018)
3. **Explanation methods**

Explanations methods

Saliency maps (Looking at the activation Tensor)

$$\varphi(.) = \frac{1}{r} \cdot \sum_{k=1}^{r} x_k \cdot \left[\log_2 \left(\frac{1}{r} \cdot \sum_{k=1}^{r} x_k \right) - \frac{1}{r} \cdot \sum_{k=1}^{r} \log_2 (x_k) \right] \tag{1}$$

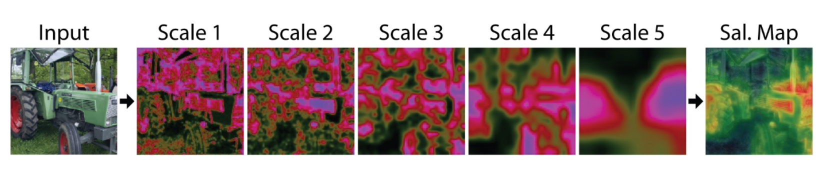

Figure 15: Multiple scale Saliency map From activation Tensor (Mundhenk et al, 2020)

Grad CAM

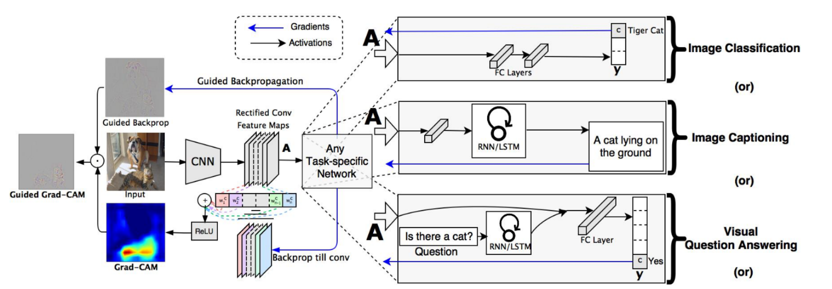

Figure 16: Schema of the GradCAM method

Shapley value

$$\hat{\phi}_j = \frac{1}{M}\sum_{m=1}^{M}\left(\hat{f}\left(x_{+j}^{m}\right) - \hat{f}\left(x_{-j}^{m}\right)\right)$$

Limits of explanation methods

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

Concept-based Explanation

Idea

- Instead of explaining the importance of each feature, intepret the decision with respect to **concepts**.

Concepts properties (Alvarez-Melis 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?

Example of Concept-based explanations 1/2

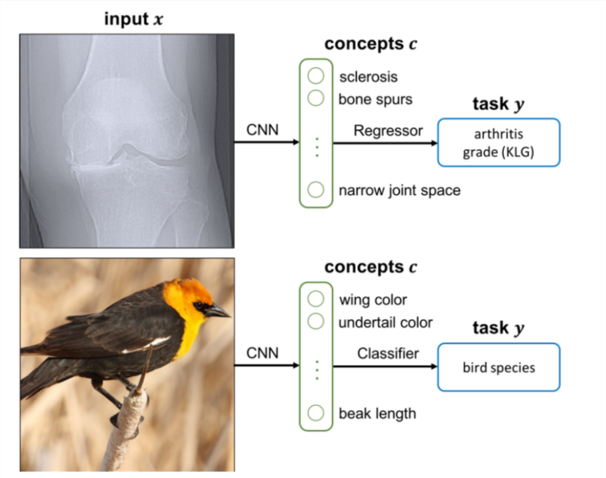

Figure 17: *Concept Bottleneck models* (Koh et al. 2020)

Example of Concept-based explanations 2/2

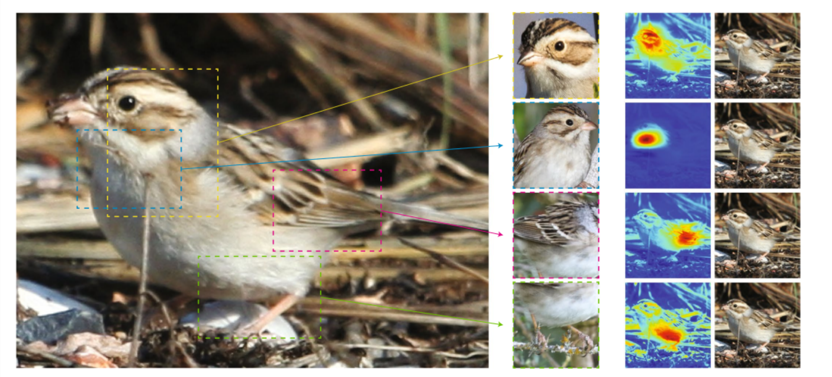

Figure 18: *This looks like that: deep learning for interpretable image recognition* (Chen et al. 2019)

Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vector (Kim et al. 2018) 1/2

Method

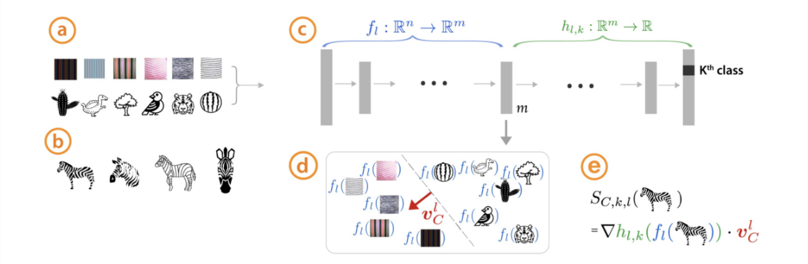

Figure 19: Schema of TCAV

$$\mathrm{TCAVQ}_{C,k,l} = \frac{|\{\boldsymbol{x} \in X_k : S_{C,k,l}(\boldsymbol{x}) > 0\}|}{|X_k|}$$

Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vector (Kim et al. 2018) 2/2

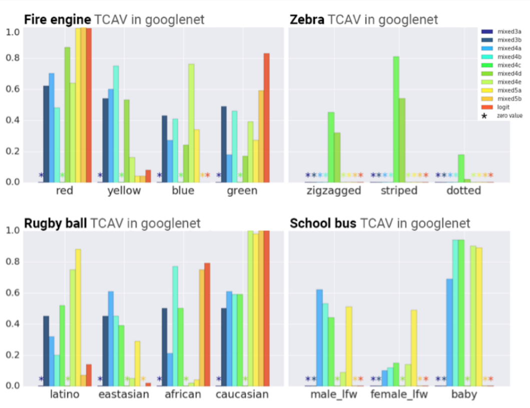

Figure 20: TCAV Importance Score for each leayer of googlenet

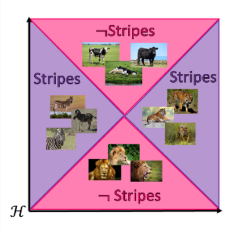

Figure 21: Non Linear Separability of concepts, see CAR (Crabbé and Scharr 2022)

Concept bottleneck models (Koh et al. 2020)

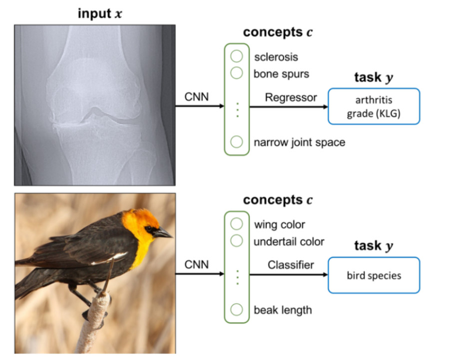

Figure 22: Schema of the approach

- Intervention of the Expert
- Not a PostHoc Method
- Requires concept labeled data

$$\hat{f}, \hat{g} = \arg\min_{f,g} \sum_i \left[L_Y\left(f\left(g\left(x^{(i)}\right)\right); y^{(i)}\right) + \sum_j \lambda L_{C_j}\left(g\left(x^{(i)}\right); c^{(i)}\right) \right]$$

Towards Automatic Concept-based Explanations (Ghorbani et al. 2018)

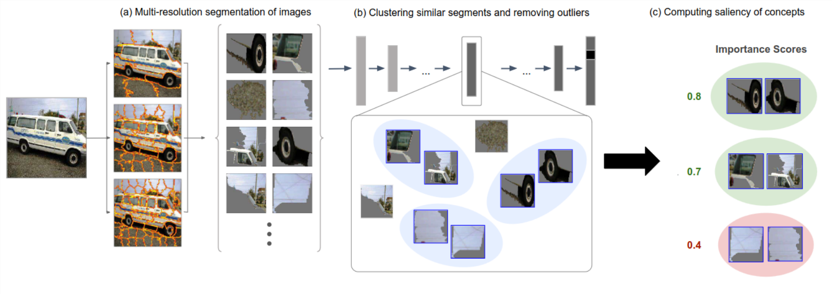

Figure 23: Schema of ACE method

- Automatic Method
- Build patches
- No Name on the concepts

Conclusion

Black Box

- Can be harmful to society (bias, failure)
- Proprietary and obscurity by complexity make a black box
- Building **trust** is necessary

Concepts

- Concepts can take multiple forms (prototype, text)
- They have to respect **Intelligibility**, **Faithfulness** and **Stability**
- It allows the dialog with the expert

Post-Hoc vs Explainable by design

- Post-Hoc methods may be unreliable
- Latent spaces of black box are not necessary separable by concepts (whitening)
- Interpretable by design suppose to retrain a model but is better