Interpretable Representation Learning

Representation Learning

Vladimir Zaigrajew

2025-04-01

Vladimir Zaigrajew - vladimir.zaigrajew.dokt@pw.edu.pl

Tymoteusz Kwieciński - tymoteuszkwiecinski@gmail.com

You can find us in Room 316, MINI, PW

Remember every information you can find on our Github Repo:

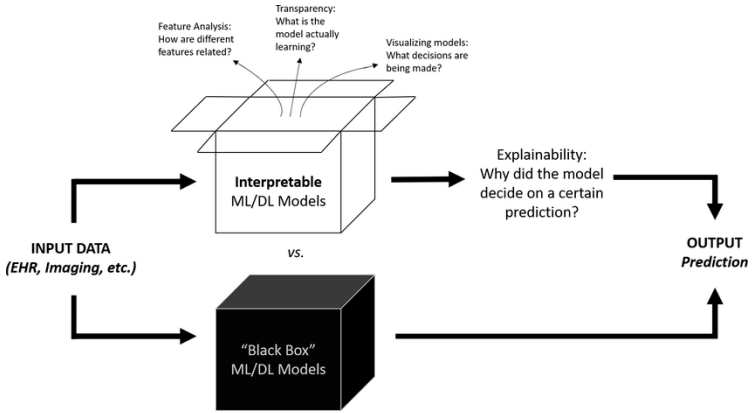
Figure 1: QR code to course Github Repo

Figure 2: QR code to our Github Repo

A Little Bit about Explainable AI

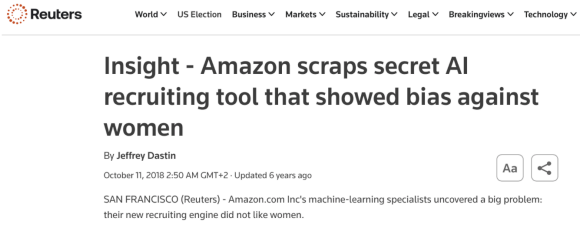
- Explainable AI (XAI) is a subfield of AI that focuses on making the decision-making process of AI systems more transparent and understandable to humans.
- It aims to provide insights into how AI models arrive at their predictions or decisions, enabling users to trust and interpret the results.
- XAI is particularly important in high-stakes domains such as healthcare, finance, and autonomous systems, where understanding the rationale behind AI decisions is crucial for safety and accountability.

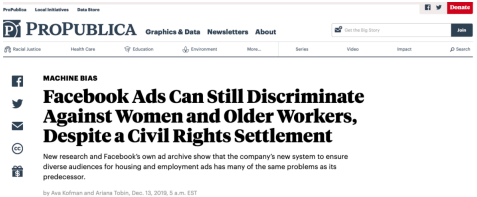
Is it hard to explain deep learning models? **Yes! those models are a black box to us due to their complexity!**



Examples When XAI is important

Read more at:
https://incidentdatabase.ai





XAI is an active area of research!

Concept-Based Explanation:

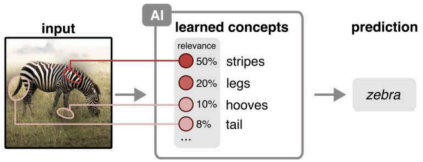
- Uses surrogate models to detect learned concepts within the main model
- Main model predicts classes while surrogate model identifies intermediate concepts

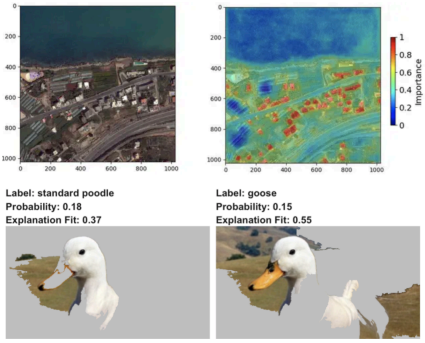
Gradient-Based Explanation:

- Highlights important input features by analyzing gradients
- Shows which pixels/regions influenced the model's decision

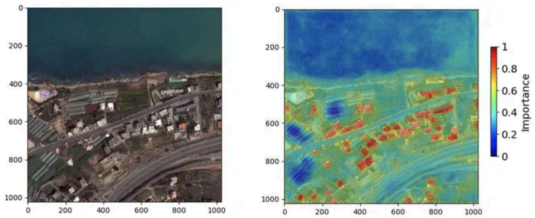
Perturbation-Based Explanation:

- Identifies critical regions by systematically masking parts of the image
- Reveals which areas must remain visible for classification





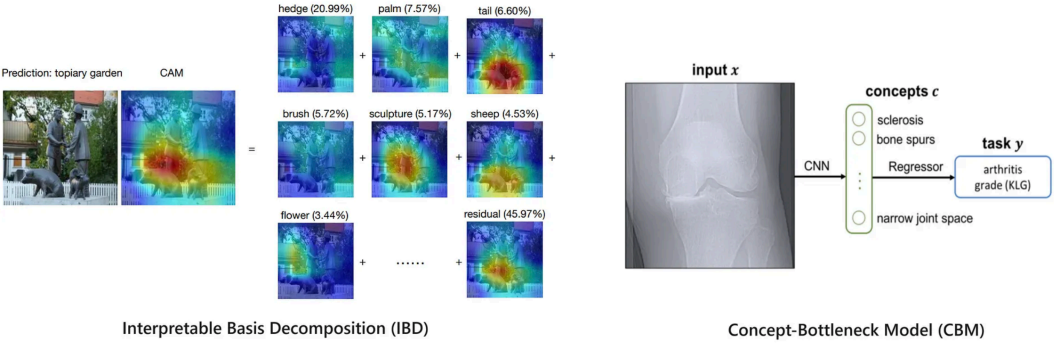
*Standard XAI methods show where the network is **looking**, but this is not sufficient to explain what it is **seeing** in a given input*



Concept based approach tries to explain the model's decision by identifying the concepts that are important for the model's prediction and are human-interpretable.

What is a concept?

"A concept can be any abstraction, such as a colour, an object, or even an idea"



Two representative examples of C-XAI methods. Left: Interpretable Basis Decomposition (IBD) provides an explanation by decomposing the decision into a set of concepts. Right: Concept Bottleneck Model (CBM) directly predicts a set of intermediate concepts that are later used to classify the input sample. Images from the papers.

[0]Gabriele Ciravegna. "C-XAI: Concept-Based Explainable AI." Medium. https://medium.com/@gabriele.ciravegna/c-xai-concept-based-explainable-ai-51dece0472f1

Post-hoc Concept-based Explanation

Post-hoc approach uses surrogate models to detect learned concepts within the main model. The main model predicts classes while surrogate model identifies intermediate concepts.

Advantages: Does not require any changes to the main model, can be applied to any model. Maintains the predictive and generalization capabilities of the model, while enhancing its interpretability.

Disadvantages: Surrogate model may not be able to capture all the concepts learned by the main model. The surrogate model may introduce additional complexity and may not be as interpretable as the main model.

Post-hoc Concept-based Explanation - Supervised

Analysis network behavior on samples representing symbolic concepts such as color, object, or idea (basically whatever you want).

Assess which concepts are learned, where (which layer), and their influence on the model.

Provide explanations as either:

- Class-concept relations (T-CAV): correlating predictions/class weights with concept projections (e.g., predicting the class of a bird based on its beak, wings, etc).
- Node-concept associations (Network Dissection): linking hidden node activations to specific concepts (e.g., identifying which neurons respond to specific features).

Key examples: T-CAV, IBD, Network Dissection

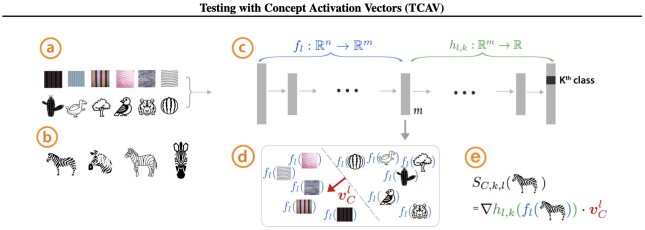


Figure 1. **Testing with Concept Activation Vectors:** Given a user-defined set of examples for a concept (e.g., 'striped'), and random examples ⓐ, labeled training-data examples for the studied class (zebras) ⓑ, and a trained network ⓒ, TCAV can quantify the model's sensitivity to the concept for that class. CAVs are learned by training a linear classifier to distinguish between the activations produced by a concept's examples and examples in any layer ⓓ. The CAV is the vector orthogonal to the classification boundary (v_C^l, red arrow). For the class of interest (zebras), TCAV uses the directional derivative $S_{C,k,l}(\boldsymbol{x})$ to quantify conceptual sensitivity ⓔ.

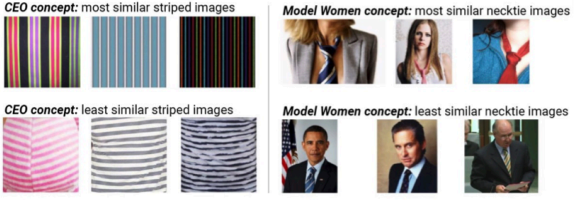


Figure 2. The most and least similar pictures of stripes using 'CEO' concept (left) and neckties using 'model women' concept (right)

Train a linear classifier to predict the concept from the activations of the last layer of the model.

Combining gradient and linear classifier to determine the influence of a concept on the model's prediction.

$$S_{C,k,l}(\boldsymbol{x}) = \lim_{\epsilon \to 0} \frac{h_{l,k}(f_l(\boldsymbol{x}) + \epsilon \boldsymbol{v}_C^l) - h_{l,k}(f_l(\boldsymbol{x}))}{\epsilon} = \nabla h_{l,k}(f_l(\boldsymbol{x})) \cdot \boldsymbol{v}_C^l, \quad (1)$$

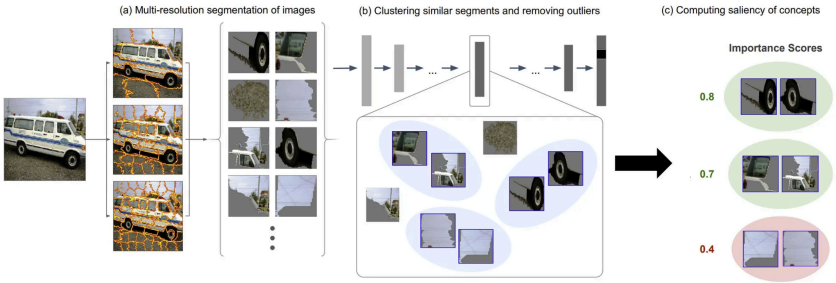
Kim, Been, et al. "Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav)." International conference on machine learning. PMLR, 2018.

Post-hoc Concept-based Explanation - Unsupervised

Identifies unsupervised clusters of samples (unsupervised concept basis) that influence predictions or classes.

Primarily provides explanations as class-concept relations.

Differ from standard XAI methods by selecting features based on their ability to represent other input samples, not just saliency or internal similarity.

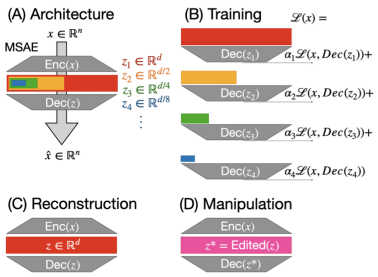


Key examples: ACE, On Completeness-aware, SAE.

Trains an sparse autoencoder to learn to disentangle polisemantic representation into interpretable monosemantic one.

Uses the concept of dictionary learning, sparse coding and autoencoders.

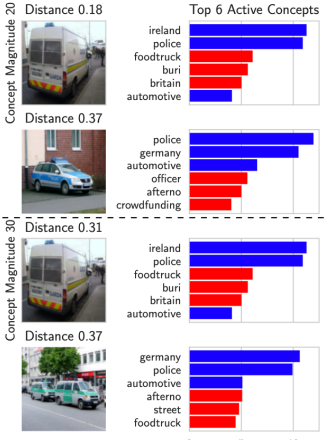
$$z = \mathrm{ReLU}\big(W_{\mathrm{enc}}\,(x - b_{\mathrm{pre}}) + b_{\mathrm{enc}}\big),$$
$$\hat{x} = W_{\mathrm{dec}} z + b_{\mathrm{pre}},$$

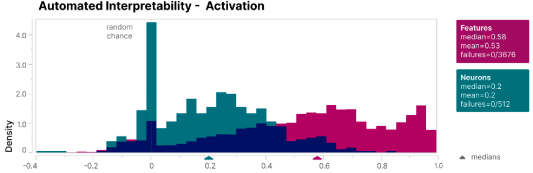


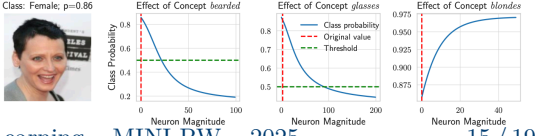
Zaigrajew, Vladimir, Hubert Baniecki, and Przemyslaw Biecek. "Interpreting CLIP with Hierarchical Sparse Autoencoders." arXiv preprint arXiv:2502.20578 (2025).

Post-hoc Concept-based Explanation - Unsupervised SAE

1. because this smile can cure the blues .
2. person with a lovely smile
3. portrait of a smiling boy
4. portrait of a senior man smiling
5. detail view of a pretty middle aged woman smiling
6. even after lots of walking ... all smiles !
7. man smiling to the camera
8. all smiles : football player has signed a contract







Explainable-by-design Concept-based Models

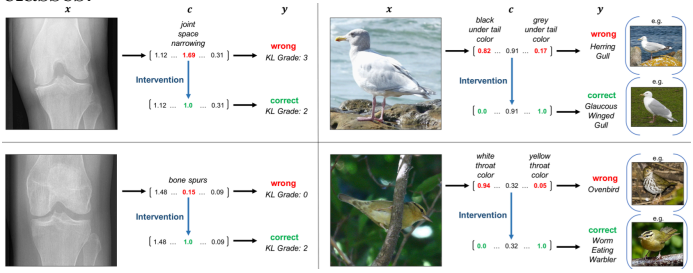
Explainable-by-design concept-based models depart from standard neural network training practices by explicitly incorporating a set of concepts within the same neural network architecture.

Concept-based Models learns two models often sequentially f and g, where g is a concept-based model $g : X \to C$ and f is a classifier $f : C \to Y$. Concepts C are learned in a supervised manner, while the classifier f is a linear probing model.

Advantages: Concept-based Models ensure that the network explicitly learns a set of concepts that are interpretable to humans. Additionally, domain experts can adjust predicted values for specific concepts and observe changes in the model's output, enabling the generation of counterfactual explanations.

Disadvantages: However, these methods can only be used when training a model from scratch is feasible, potentially tailoring it to the specific task. Furthermore, in simpler solutions, the predictive accuracy of concept-based models may be lower than that of standard black-box models.

We require the dataset to be annotated with concepts. The model is trained to predict the concepts and the classes.



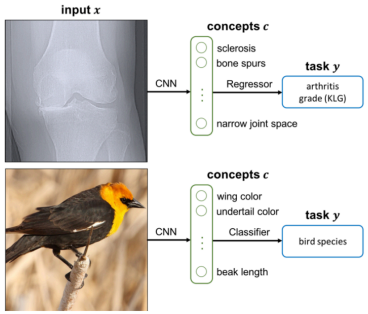


Figure 1. We study concept bottleneck models that first predict an intermediate set of human-specified concepts c, then use c to predict the final output y. We illustrate the two applications we consider: knee x-ray grading and bird identification.

In paper authors used X-ray grading (OAI) dataset and bird identification (CUB) dataset. The CUB dataset consists of images, and the concepts are the parts of the bird (e.g. beak, wings, etc.) and the classes are the species of the bird.

[0]Koh, Pang Wei, et al. "Concept bottleneck models." International conference on machine learning. PMLR, 2020.

Categories of concept-based models:

- *Supervised*: Require symbolic concept annotations. Can be jointly trained with task supervision or embedded via separate concept learning. Examples: CBM, CEM, Concept Whitening.
- *Unsupervised*: Autonomously extract concepts without predefined symbols. May use unsupervised concept basis or encode prototype-concepts representing common input samples. Examples: SENN, BotCL, ProtoPNet.
- *Hybrid*: Integrate both supervised and unsupervised concepts, enabling use when few supervised concepts are available. Examples: CBM-AUC, GlanceNets.
- *Generative*: Use external generative models to define textual concepts as numerical representations for class prediction. During testing, predict both class and most suitable descriptions. Examples: LaBO, Label-free CBM.

Closing Thoughts

XAI is an active area of research with many approaches and methods, as there is no one-size-fits-all solution in XAI.

Recently, concept-based models have gained popularity due to their ability to provide interpretable explanations.

We can leverage internal representations of the model to provide explanations or to modify the model predictions based on the concepts.

We can also train explainable-by-design concept bottleneck models that are interpretable by design. However this method also faces some limitations.

References:

- For Concept Based Methods this blog post is a great source of information.
- For a more in-depth overview of XAI methods, you can read Christopher Molnar's book or just ask Professor Biecek, as he is one of the leading experts in the field.