Towards Automatic Concept-based Explanations

Radwa El Shawi

Machine learning Interpretability

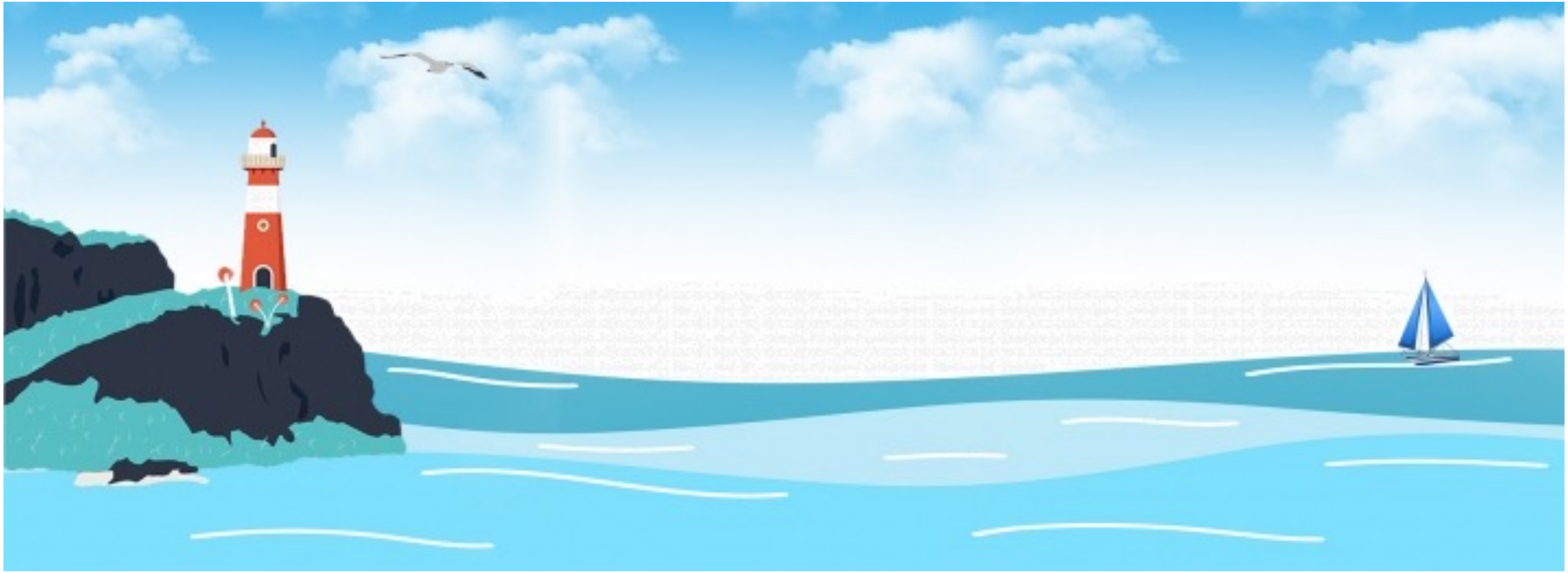
“Interpretability is the degree to which an observer can understand the cause of a decision.”

~ Miller T., 2017, Explanation in AI: Insights from the Social Sciences

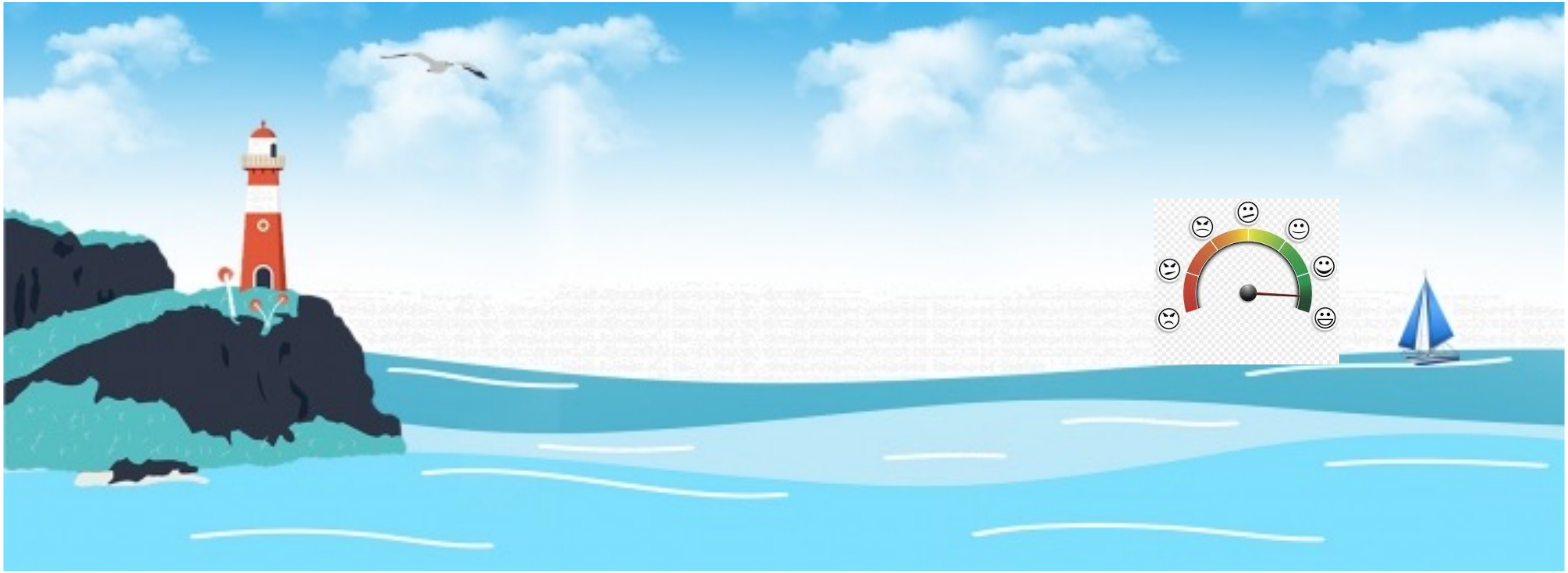
Black-box model

explantion

- Feature attributation
- Rules
- The most influential instances in the training data
- Counterfactual explanation
- High-level human understandable concepts

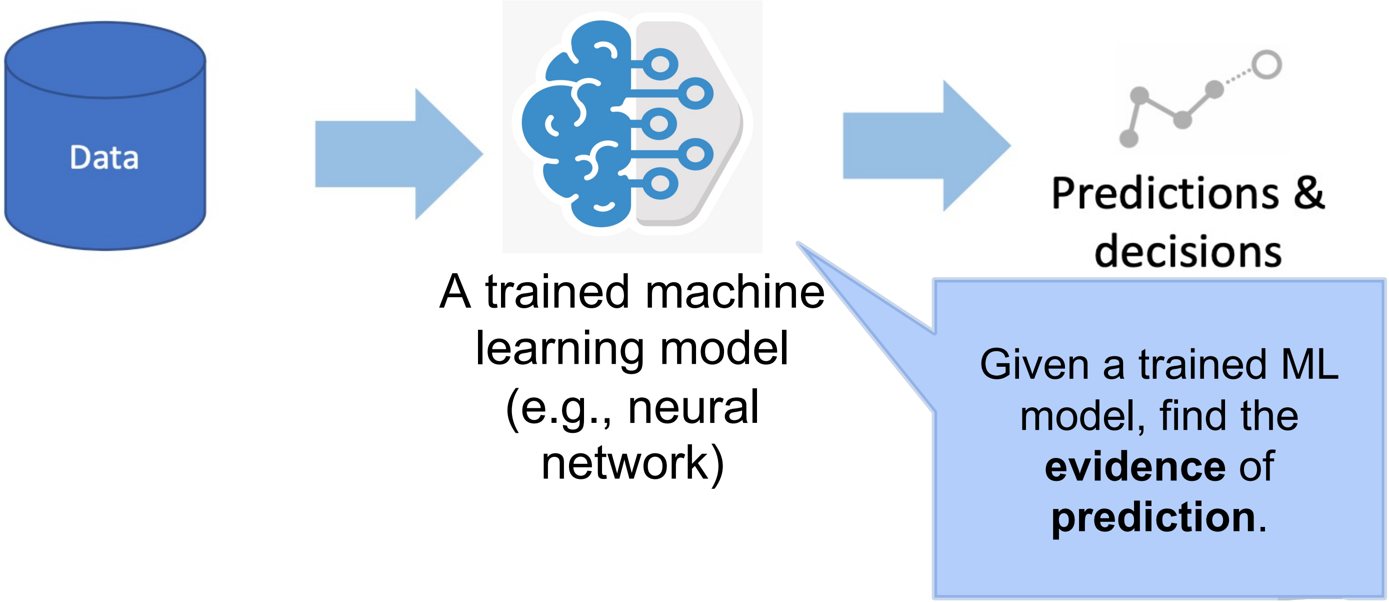


Why ML interpretability is challenging

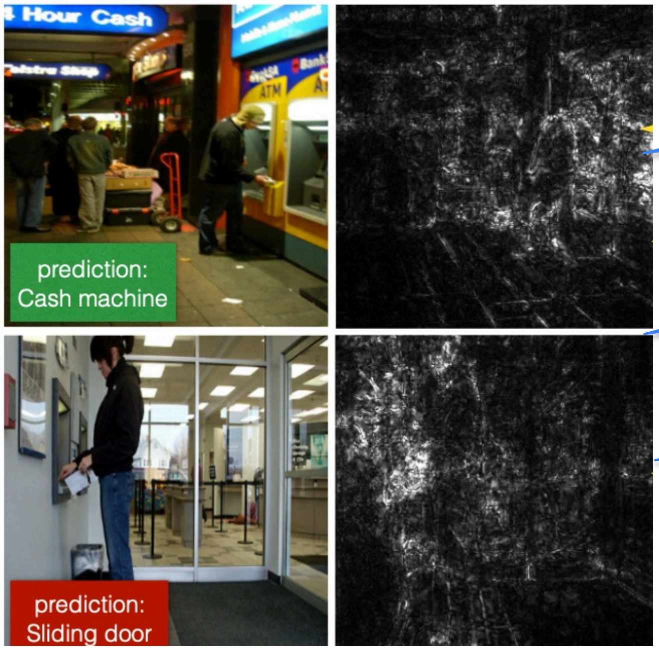


A path the may take a step toward good explanation

Investigating post-training interpretability techniques



Saliency Maps



Were there more pixels on the 'ATM' than on the 'person'?

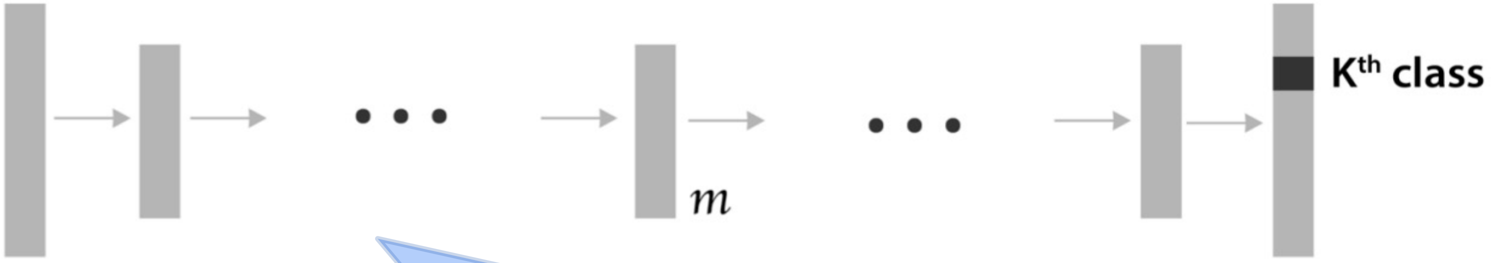
Which concept maters more for the prediction, 'glass door', 'paper'?

Which concept mattered more for the prediction?

Would not be more useful if we can quantify the imporatnce of each of **user-defined concepts**?

Selvaraju, Ramprasaath R., et al. "Grad-cam: Visual explanations from deep networks via gradient-based localization." *Proceedings of the IEEE international conference on computer vision*. 2017.

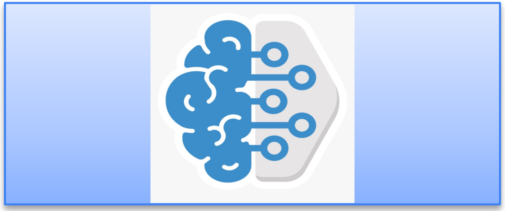
Goal of TCAV: Testing with Concept Activation Vectors



Quantitative explanation: how much a **concept** (e.g., stripes) was important for the **prediction of specific class** (e.g., zebra) in a trained model?

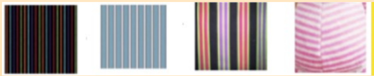
Kim, Been, et al. "Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav)." *International conference on machine learning*. PMLR, 2018.

Goal of TCAV: Testing with Concept Activation Vectors

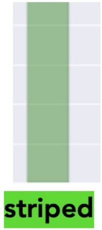


P()

Trained NN model



How important is concept stripes to the prediction of this image as zebra?

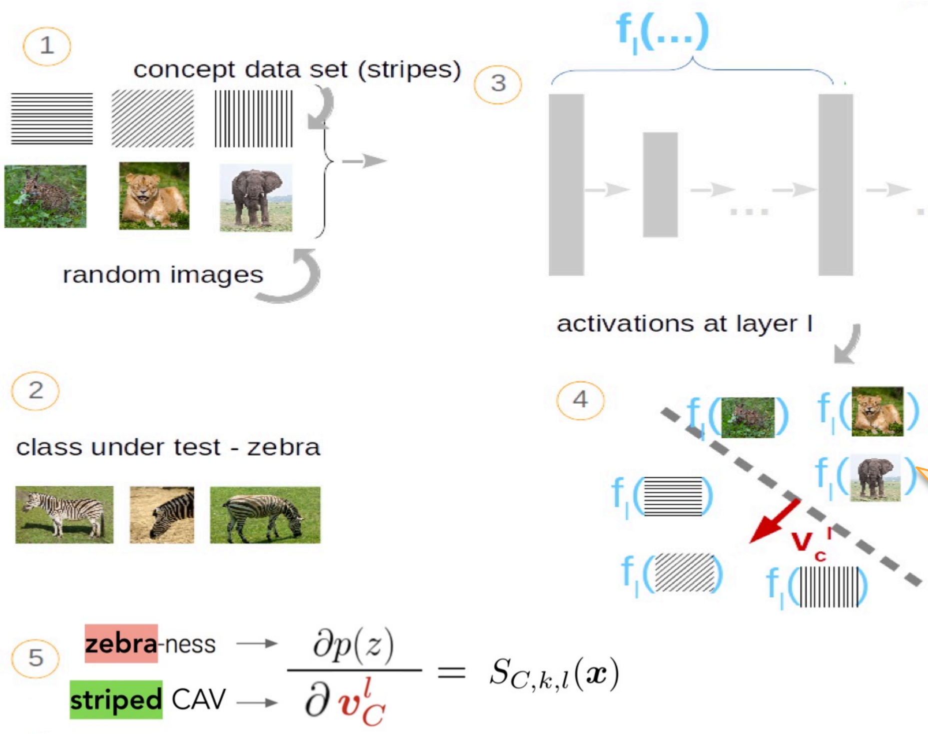


TCAV score for concept Zebra

TCAV provides quantitative importance of a concept if and only if your network learned about it.

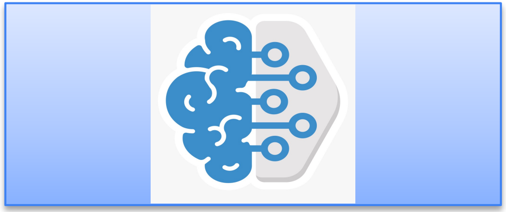
Kim, Been, et al. "Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav)." *International conference on machine learning*. PMLR, 2018.

How to define concepts



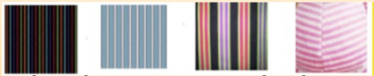
Train a linear classifier to separate activations. CAV (v'_c) is the vector orthogonal to the decision boundary.

TCAV: Testing with Concept Activation Vectors



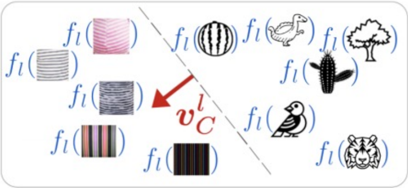
Trained NN model

P()

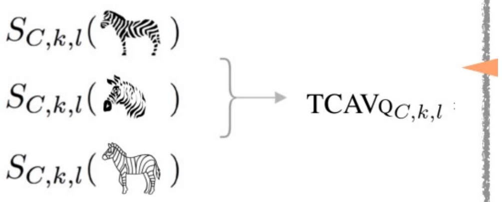


How important is concept striped to the prediction of this image as zebra?

1. Learning CAVs



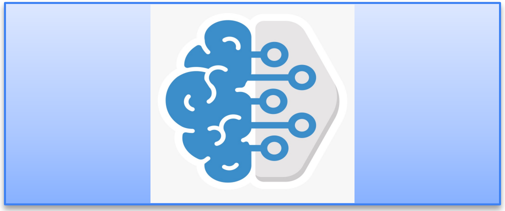
2. Getting TCAV score



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

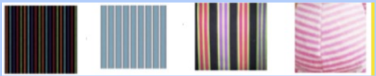
Can we do better?

TCAV: Testing with Concept Activation Vectors



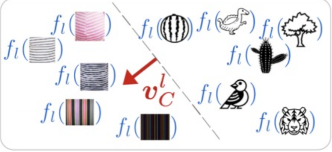
P()

Trained NN model

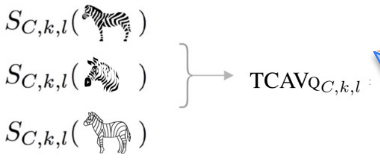


How important is concept striped to the prediction of this image as zebra?

1. Learning CAVs



2. Getting TCAV score



Would not be more useful if we can quantify the imporatnce of automatically extracted **concepts**?

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

Automated Concept-based Decision Tree Explanations for CNNs ACDTE

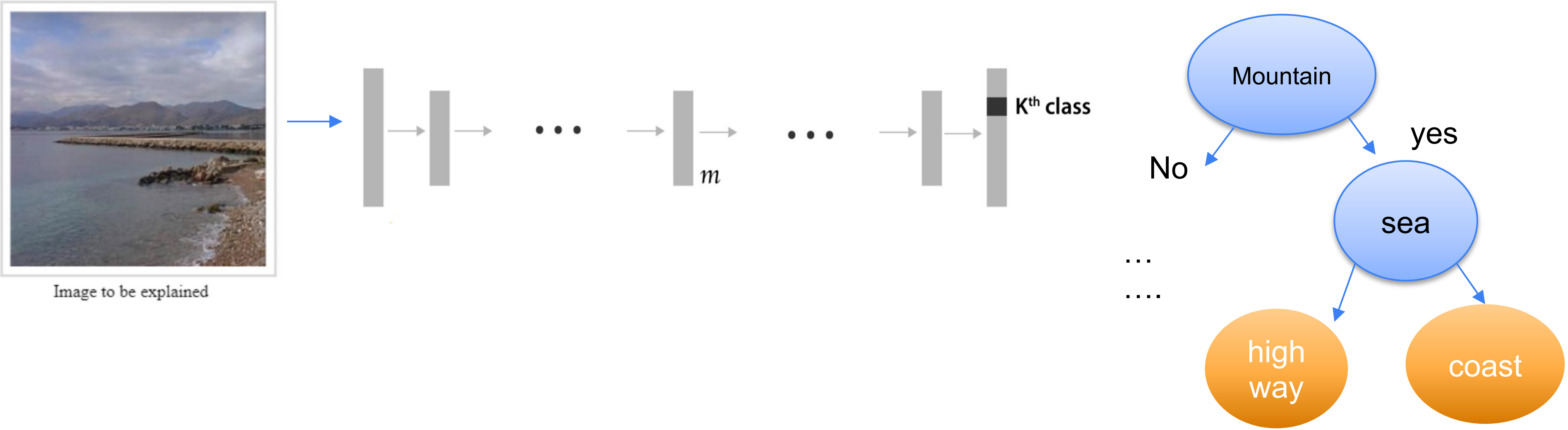
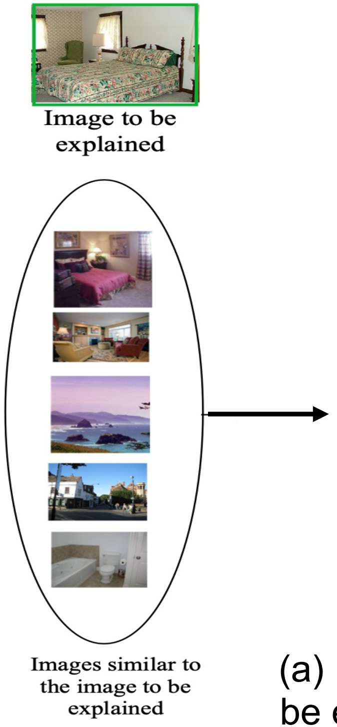


Image to be explained

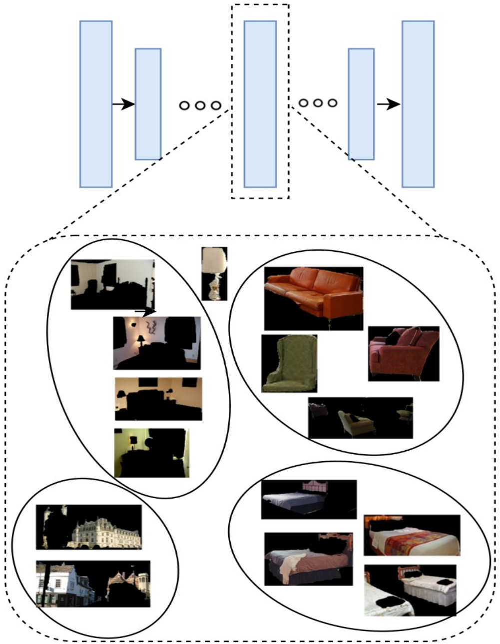
R El Shawi, Y Sherif, S Sakr . "Towards Automated Concept-based Decision Tree Explanations for CNNs". *EDBT*, 2021

ACDTE Stage1: Concept extaction



(a) Segmentation of images similar to the image to being explained

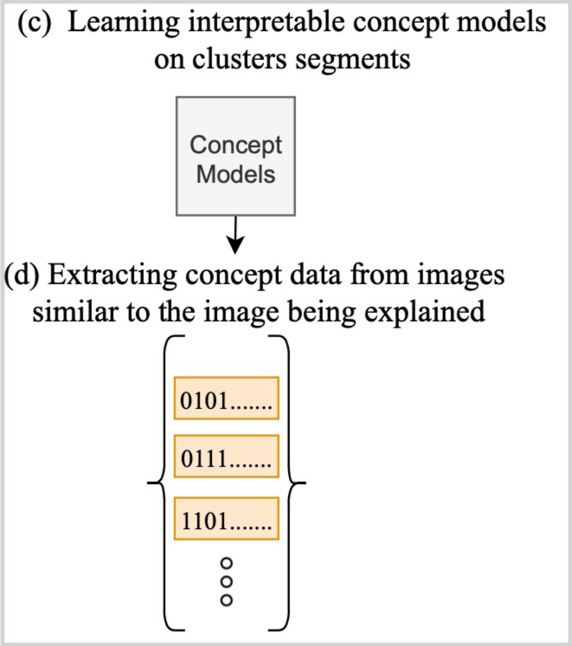
(b) Clustering similar segments and removing outliers



(a) Extract a set of similar images to the image to be explained either from the main task dataset or related dataset. Each image in the selected images is segmented.

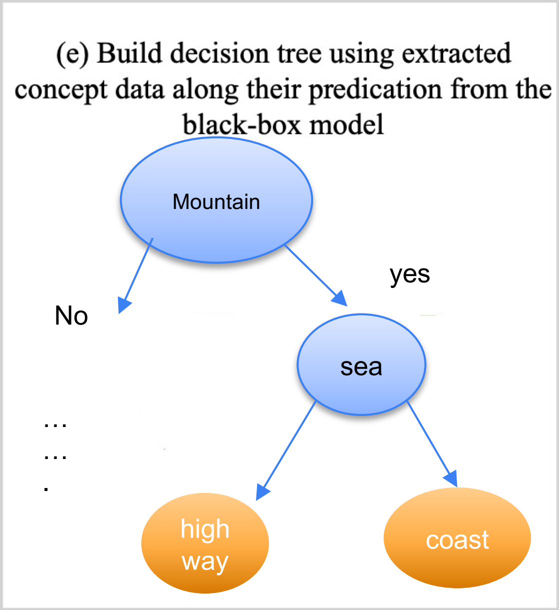
(b) Segments are clustered in the activation space and outliers are removed to form coherent clusters that represent concepts.

ACDTE Stage 2: Learning interpretable concept and extracting concept data



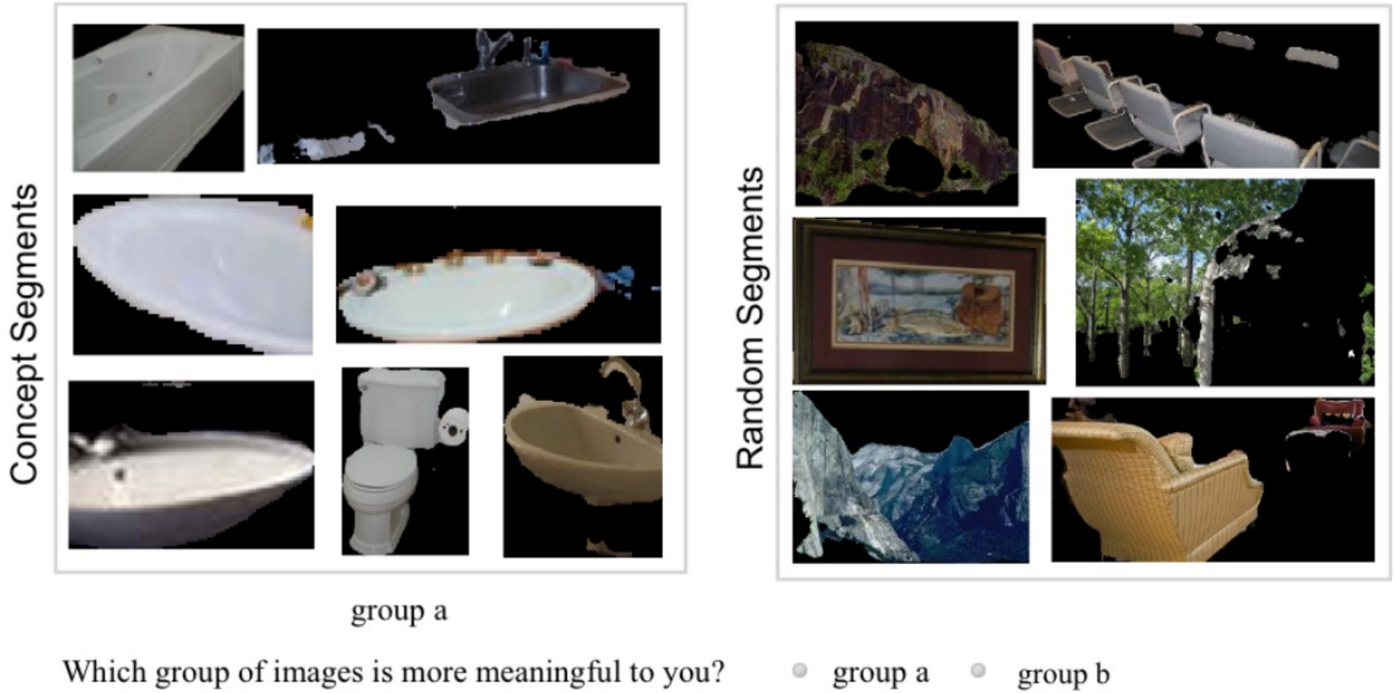
(c) Training a linear model for each concept to act as a concept detector. (d) For each image in the activation space, use concepts detectors to form a binary feature vector.

ACDTE Stage 3: Building explanation decision tree



(e) Feature vectors along with the prediction of the target network are used to train a shallow decision tree. The decision tree provides a natural explanation for the contributing concepts for the prediction, in addition to counterfactual explanation.

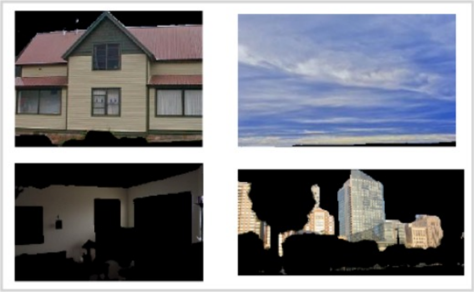
Human Evaluation of the Visual Explanations



Human Evaluation of the Visual Explanations (Cont.)

Which of the images below highly contribute to the prediction of the image above as a street?

Which of the images below highly contribute to the prediction of the image above as a park?



Open Challenges

- No single explanation can fit all users
- Rigorous, agreed upon evaluation protocols
- Human involvement (e.g. better interactive, “social” explanations)

Thank You