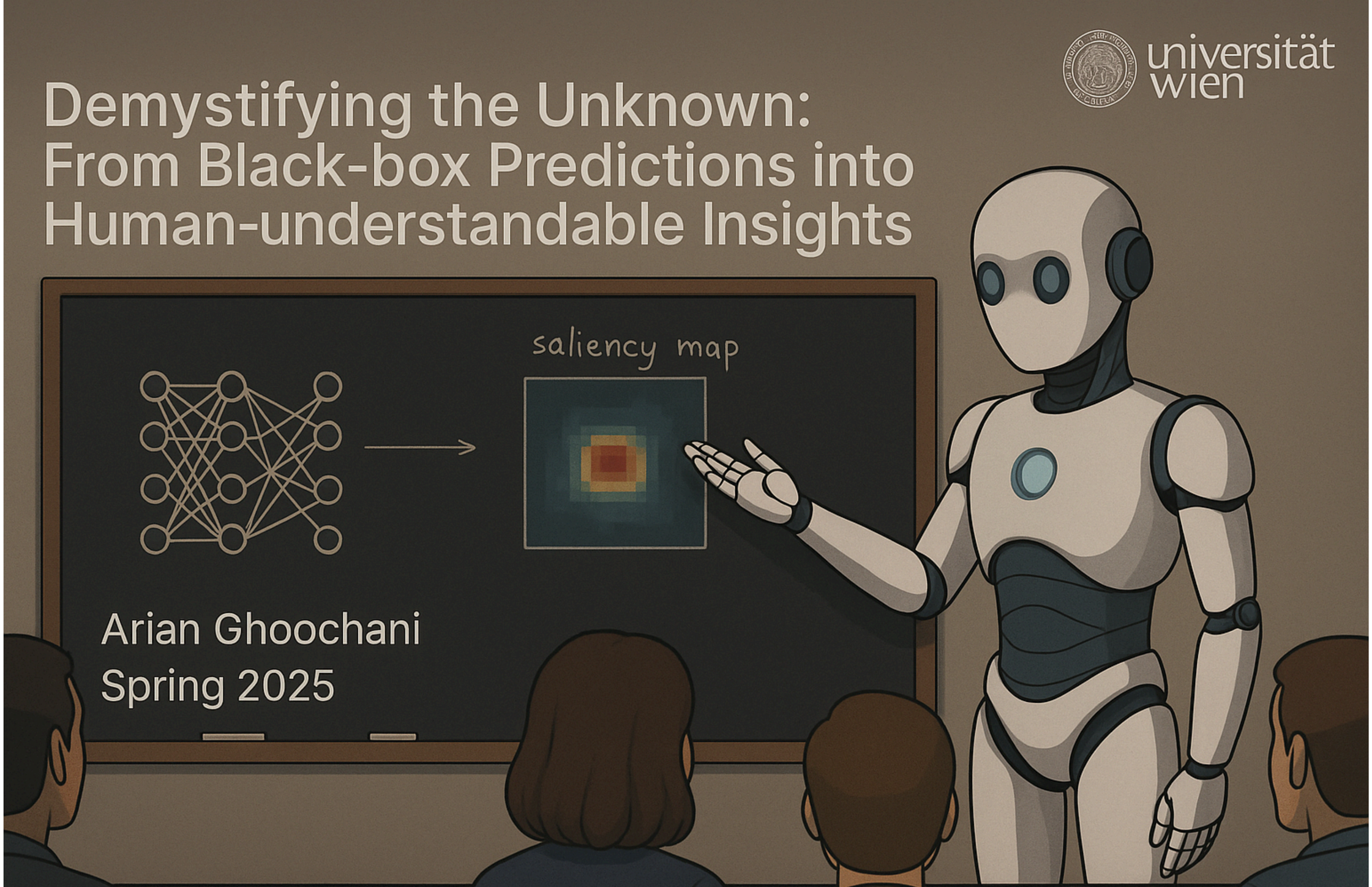
Demystifying the Unknown: From Black-box Predictions into Human-understandable Insights

universität wien



Arian Ghoochani

Spring 2025

AI in Everyday Life

Examples of Daily usages of AI

- Voice Assistants
- Facial Recognition
- Recommendation Systems
- Online rerouting feature of Maps

AI in critical domains

universität wien

AI models can be biased

Source: MIT News (2024)

Artificial intelligence models that are most accurate at predicting **race** and **gender** from X-ray images also show the biggest “**fairness gaps**.”

universität wien

When a Computer program keeps you in jail

The story

- An **inmate** in the USA, **denied** parole **despite having a perfect record of rehabilitation.**

- Because of **high score** from a computer system called **Compas.**

Compas

- A **risk assessment** software.

- Predicts how likely is to defendent: **Reoffend, Fail to appear in court or commit violent crimes in the future.**

- The algorithm is **private.**

When a Computer Program Keeps You in Jail

By Rebecca Wexler

June 13, 2017

Share full article 230

Sally Deng

Source: “When a Computer Program Keeps You in Jail,” NYT, June 13, 2017

universität wien

The Husky vs Wolf case

- A **husky** (on the left) is confused with a **wolf**, because the pixels (on the right) characterizing wolves are those of the **snowy background.**

Image source: Besse, P., Castets-Renard, C., Garivier, A., & Loubes, J.-M. (2018). "Can Everyday AI be Ethical? Machine Learning Algorithm Fairness." ResearchGate Preprint. https://doi.org/10.13140/RG.2.2.22973.31207

XAI is essential

Pitfalls AI might fall in

- Bias
- Focus on Irrelevant Features
- Learning the Wrong Patterns

Consequences

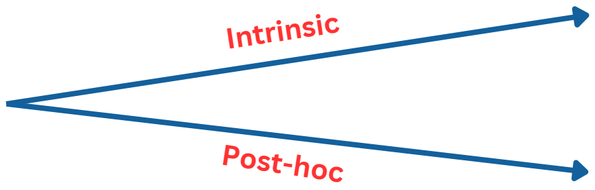
➤ Misdiagnosis
➤ Discrimination
➤ Poor generalization
➤ Loss of trust

It is essential to know how the decisions has been made.

universität wien

Categorization of Interpretability Methods

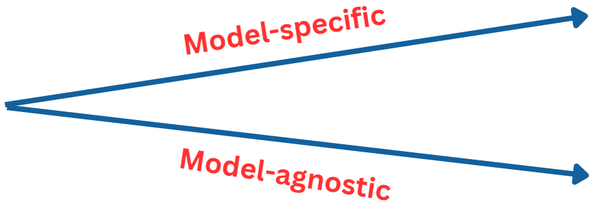
Intrinsic
vs.
post-hoc



Building the interpretability into model design

Deriving Explanations after model training

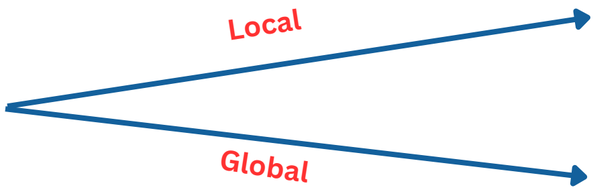
Model-specific
vs.
model-agnostic



Applicable only on certain architecrures

Treating any model as a black box

Local
vs.
global explanations



Explaining predictions indivisualy

Extracting general insights on model behavior

universität wien

Inherently Interpretable Models

Idea: Transparency by design rather than explanation after the fact.

Examples

- **Linear/Logistic Regression:** Coefficients quantify input-output relationships directly.
- **Decision Trees:** Hierarchical structure of feature-based decision rules
- **Decision Rule Models:** Flat if–then rules, often weighted or aggregated for outcomes

Limitations

- **Loss of Interpretability with Growth in model's size**
- **Limited Applicability to Unstructured Data**

Feature attribution methods

Definition: Techniques that identify and quantify the contribution of individual input features to a model's prediction.

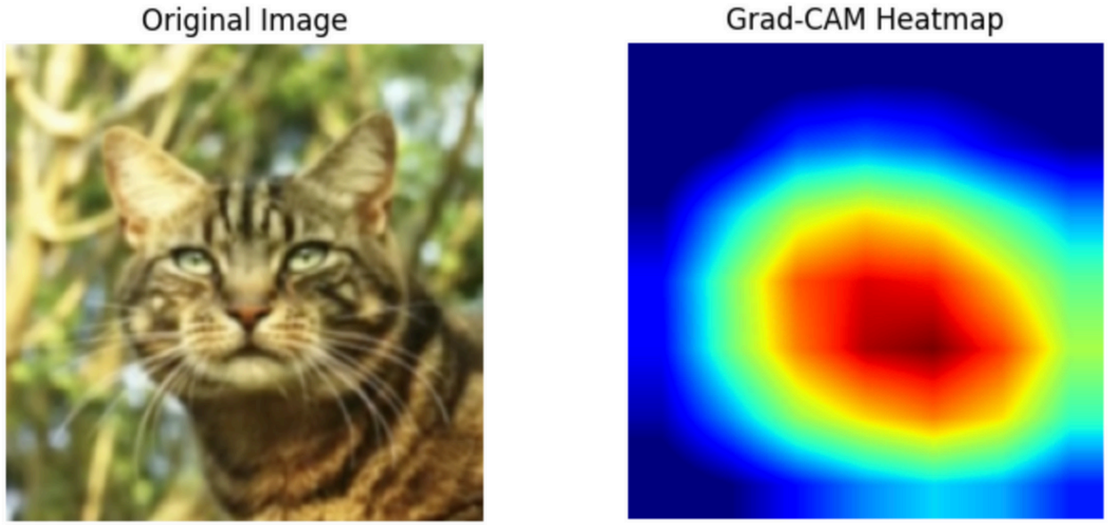
Approach	Explanation	Example
Gradient-based Methods	Using gradients of the output with respect to any internal representation (input or layer activations) to determine feature importance	Grad-CAM
Perturbation-based Methods	Altering input features to observe changes in predictions	SHAP, LIME

Grad-CAM

Background: A technique that generates visual explanations for CNNs by using class gradients to highlight key regions in an image that influenced a prediction.

Steps

1- Calculating the **importance weights**
2- Combining activation maps with weights
3- **Upsampling** heatmap to original image size



universität wien

Grad-CAM

Mathematics

$$\alpha_k^c = \overbrace{\frac{1}{Z}\sum_i\sum_j}^{\text{global average pooling}} \underbrace{\frac{\partial y^c}{\partial A_{ij}^k}}_{\text{gradients via backprop}}$$

α_k^c = Feature map k's weight for class c
Z = Number of pixels in feature map
y^c = Score of the labeled class c
A_{ij}^k = Activation at spatial location (i, j) in feature map k

- **Importance weight shows how much feature map k contributes to class's prediction.**

$$L_{\text{Grad-CAM}}^c = ReLU\underbrace{\left(\sum_k \alpha_k^c A^k\right)}_{\text{linear combination}}$$

- **ReLU keeps only positive contributions**
- **Output is a class-discriminative heatmap**

Limitations

- Low Spatial Resolution
- Local Explanation(over only one picture)
- Limited customization

universität wien

(Selvaraju et al., 2017, Grad-CAM)

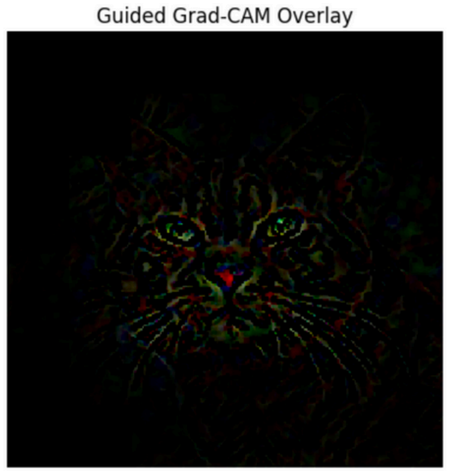
Guided Grad-CAM

Background: Combines the **localization** capability of Grad-CAM with the **fine-grained visualization** of Guided Backpropagation to produce high-resolution, class-discriminative visual explanations.

Steps

- The Grad-CAM localization map is upsampled to match the input image resolution.

- An element-wise multiplication is performed between the Grad-CAM heatmap and the Guided Backpropagation result to generate the final Guided Grad-CAM visualization.



universität wien

Concept-Based Explanations

Definition: Methods that explain model behavior through human-interpretable concepts rather than raw features or activations.

Goal: Bridging the gap between complex internal representations and meaningful ideas

Methodology: Mapping model's internal representations to recognizable concepts

CAV (Concept Activation Vectors)

Background: Measuring how sensitive a model's predictions are to userdefined concepts

Methodology

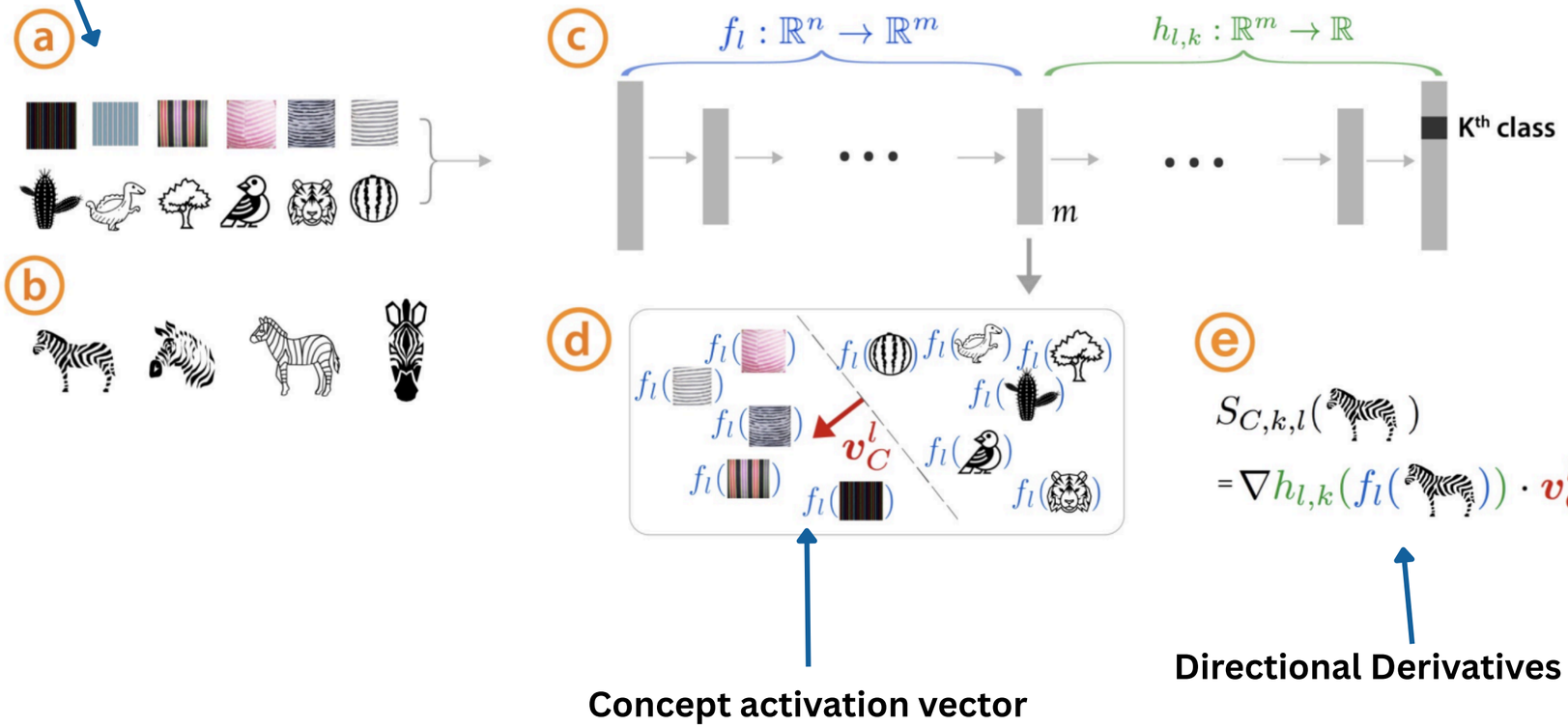
- Training linear classifier to distinguish examples with/without a concept
- Calculateing directional derivative of predictions along concept direction

Interpretability

- Measuring sensitivity of model predictions to changes in concept presence

CAV and Sensitvity Score

User-defined Concepts as Sets of Examples



Random images

Concept activation vector

Directional Derivatives and Conceptual Sensitivity

Kim et al., “Interpretability Beyond Feature Attribution,” NeurIPS 2018

TCAV

Mathematics

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

$f_l(\mathbf{x})$: Activation output at layer l for input $\mathbf{x}$.

$\mathbf{v}_C^l$: Concept Activation Vector (CAV) for concept C at layer l.

$h_{l,k}(\cdot)$: Logit score (pre-softmax output) for class k as a function of the layer-l activation.

$\nabla h_{l,k}(f_l(\mathbf{x}))$: Gradient of the class-$k$ logit with respect to the layer-l activations.

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

$\text{TCAV}_{Q_{C,k,l}}$: The TCAV score for concept C, class k, and layer l.

X_k: A set of input examples (images or data points) that belong to class k.

Kim et al., “Interpretability Beyond Feature Attribution,” NeurIPS 2018

Statistical Significance Test

What if a random set of images passed instead of meaningful concept?

- Using a randomly chosen set of images will still produce a CAV
- Testing based on such a random consept is not meaningful

To guard against this

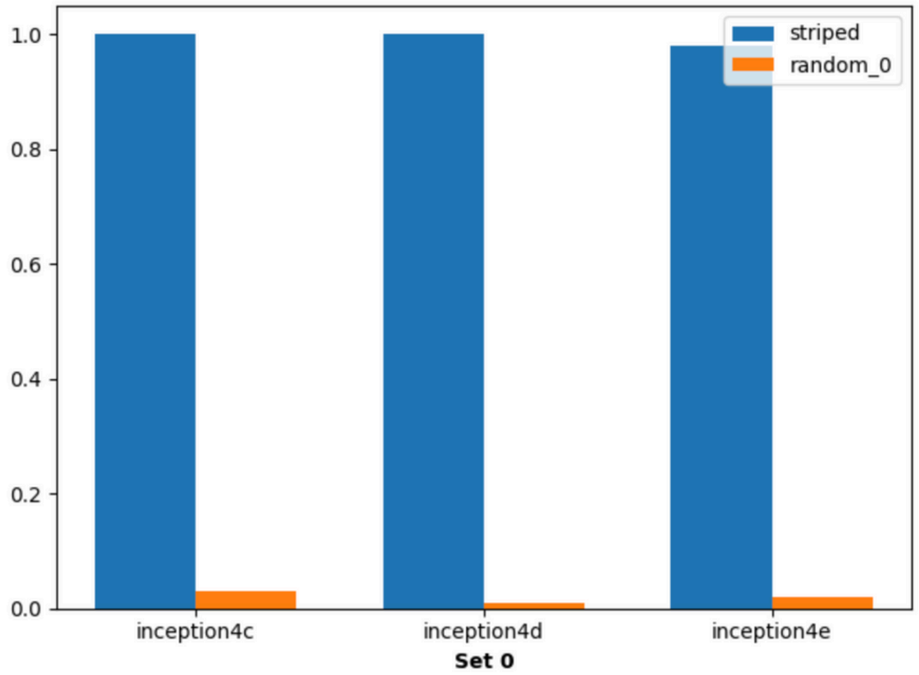
- Training the CAV **multiple times** instead of **one time training**
- Checking the consistenty of the TCAVs

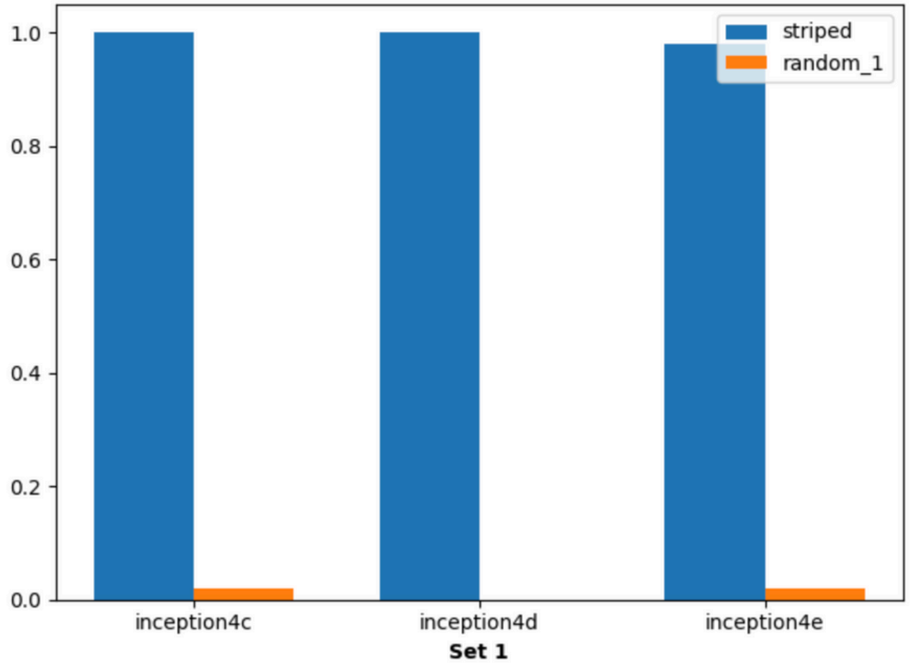
Statistical Test

- Performing **two-sided t-test** of the TCAV
- Null hypothesis of TCAV = 0.5
- If we can reject the null hypothesis,can consider the resulting concept as related to the class prediction in **a significant way.**

Example of TCAV

- Evaluating the effect of concept striped in prediction of classes with label **Zebra**
- **Googlenet** network
- 120 Striped concept images
- 120 random chosen images

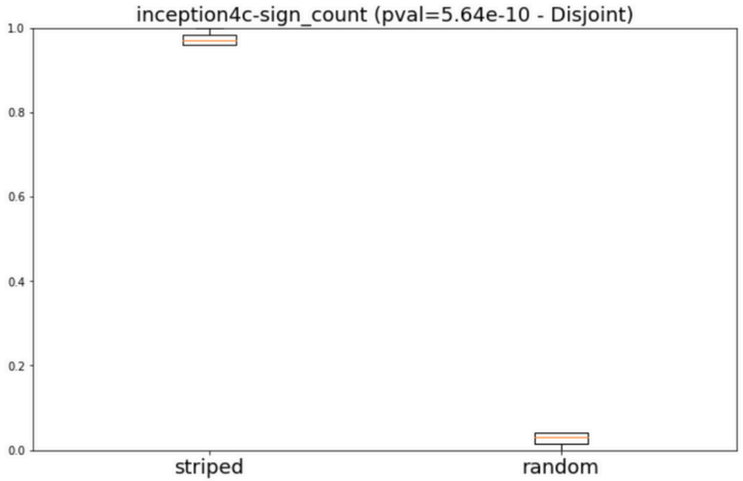


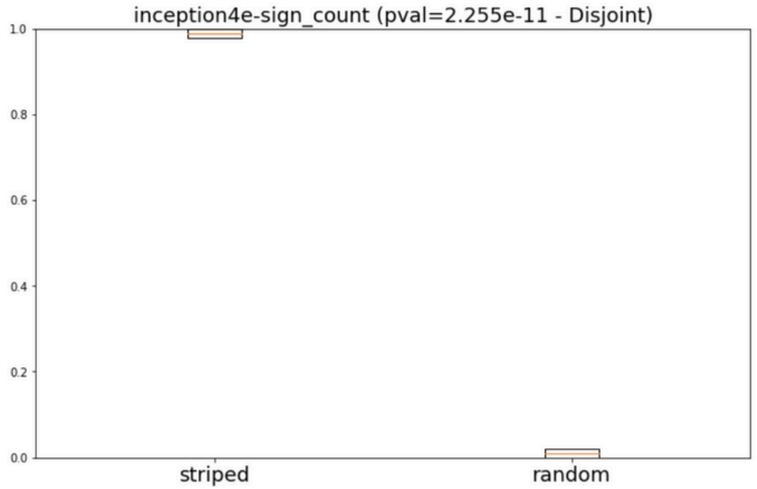


universität wien

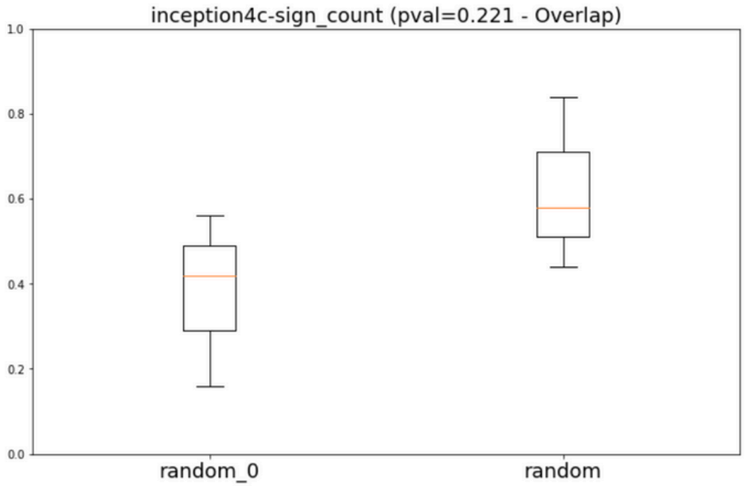
Example of TCAV

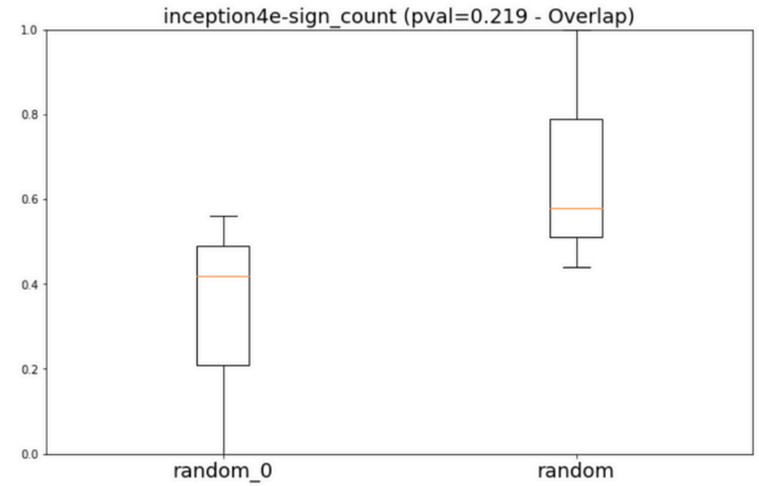
Experiment 1
(Striped vs Random)





Experiment2
(Random vs Random)





universität wien

Conclusion

In summary

- Understanding how decisions are made in black box models is crucial.
- Inherently interpretable methods are essential but often difficult to apply to unstructured data.
- Grad-CAM Offers Class-Discriminative coarse heatmaps.
- Guided Grad-CAM improves Grad-CAM's resolution by combining it with Guided Backpropagation.
- TCAV enables concept-based, user-defined interpretation of model behavior.

Challanges

- Lack of standardized benchmarks and evaluation metrics
- Providing human centered interpretations
- Rethinking Interpretability: Anticipating Black-Box models instead of chasing Them

References

- Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (pp. 618–626). https://arxiv.org/abs/1610.02391
- Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., & Sayres, R. (2018). Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV). In Advances in Neural Information Processing Systems (NeurIPS). arXiv:1711.11279.
- Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., … & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012
- MIT News. (2024, June 28). Study reveals why AI-analyzed medical images can be biased. Retrieved from
- The New York Times, When a Computer Program Keeps You in Jail. June 13, 2017. Available at: https://www.nytimes.com/2017/06/13/opinion/how-computers-are-harming-criminal-justice.html
- Besse, P., Castets-Renard, C., Garivier, A., & Loubes, J.-M. (2018). Can Everyday AI be Ethical? Machine Learning Algorithm Fairness. ResearchGate Preprint. https://doi.org/10.13140/RG.2.2.22973.31207
- Captum. (n.d.). TCAV Image Tutorial. Retrieved from https://captum.ai/tutorials/TCAV_Image

universität wien

Thanks for your attention!