PRINCETON UNIVERSITY

JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

Overlooked Factors In Concept-Based Explanations: Dataset Choice, Concept Learnability, and Human Capability

Vikram V. Ramaswamy, Sunnie S. Y. Kim, Ruth Fong, Olga Russakovsky.
Princeton University

Overview

- Concept-based explanations explain all or parts of a model in terms of human understandable semantic concepts

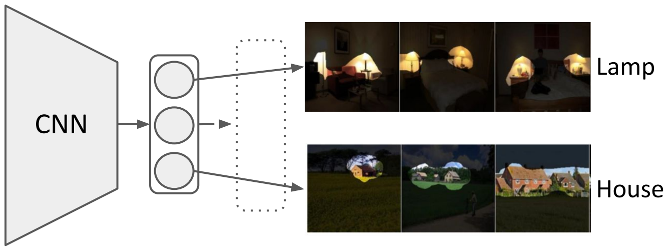
Prediction: bedroom
Explanation: + 4.2 **bed** – 1.5 *coffee-table* + 1.3 **sky** – 1.3 *sofa*
– 1.0 *drinking-glass* – 0.9 *television* – 0.9 *sconce*
– 0.8 *chair* + 0.8 *windowpane* + 0.7 *blind* + 0.7 *fan* – 0.6 *armchair*
– 0.6 *sink* – 0.6 *switch* + 0.5 *box* – 0.5 *plate* – 0.5 *ottoman* – 0.5 *paper* +0.4 *cushion* – 0.4 *tray* + …

- We investigate 3 factors of these explanations and show that they can
  1. be heavily dependent on the dataset used to learn the explanation,
  2. use concepts that are hard to learn, and
  3. be overwhelming to people due to the complexity of the explanation.

Concept-based explanations: A quick primer

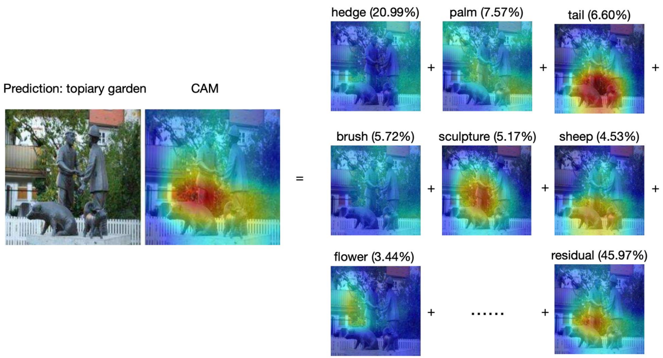
- Explain model part and/or output using semantic concepts.
- Typically trained on a “probe dataset” labelled with these concepts
  - Not necessarily the training dataset.

NetDissect



[1] David Bau*, Bolei Zhou*, et. al. Network Dissection: Quantifying Interpretability of Deep Visual Representations. CVPR, 2017

IBD



[2] Bolei Zhou, et. al. Interpretable Basis Decomposition for Visual Explanation. *ECCV* 2018

1. Effect of the probe dataset

- Using different probe datasets, we compute different kinds of concept-based explanations for a given model

- NetDissect [1]: 56% of neurons correspond to very different concepts.

Neuron	ADE20k label	ADE20k score	Pascal label	Pascal score
9	plant	0.082	potted-plant	0.194
181	plant	0.068	potted-plant	0.140
318	computer	0.079	tv	0.251
386	autobus	0.067	bus	0.200
435	runway	0.071	airplane	0.189
185	chair	0.077	horse	0.153
239	pool-table	0.069	horse	0.171
257	tent	0.042	bus	0.279
384	washer	0.043	bicycle	0.201
446	pool-table	0.193	tv	0.086

Probe dataset used can have massive impact on the explanation generated.

[1] David Bau*, Bolei Zhou*, et. al. Network Dissection: Quantifying Interpretability of Deep Visual Representations. *CVPR*, 2017

2. Learnability of the concepts used

- Compare learnability of concepts used within an explanation to the target class.
- IBD [2]: most classes are explained by at least one concept that is harder to learn.

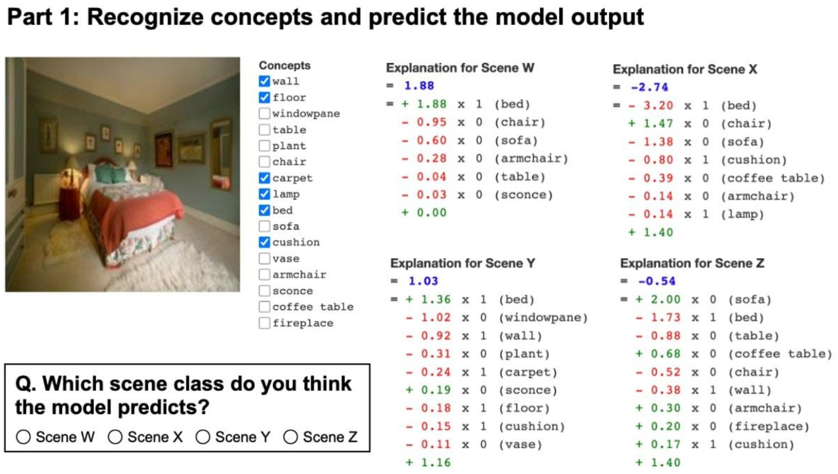
Scene	Concepts				
arena/perform 38.8	tennis court 74.0	grandstand 44.4	ice rink 40.7	*valley 19.0*	*stage 11.9*
art-gallery 27.4	binder 42.6	*drawing 10.8*	*painting 10.5*	*frame 2.5*	*sculpture 0.7*
bathroom 43.3	*toilet 39.9*	*shower 18.8*	*countertop 12.6*	*bathtub 11.1*	*screen door 9.6*
kasbah 50.2	ruins 64.3	*desert 17.3*	*arch 16.2*	*dirt track 8.9*	*bottle rack 4.2*
kitchen 33.9	*work surface 24.8*	*stove 18.2*	*cabinet 10.3*	*refrigerator 8.8*	*doorframe 2.8*
lock-chamber 36.5	water wheel 47.4	dam 43.7	*boat 16.1*	*embankment 4.8*	*footbridge 4.1*
pasture 19.2	cow 63.7	leaf 21.1	*valley 19.0*	*field 6.8*	*slope 4.1*

Concepts used within an explanation can be harder to learn than the target class.

[2] Bolei Zhou, et. al. Interpretable Basis Decomposition for Visual Explanation. *ECCV* 2018

3. Complexity of the explanation

- Want to understand how well humans can parse an explanation.
- Asked participants to predict model's output based on an explanation.
- Ask them to reason about trade-off between simplicity and correctness when varying the number of concepts.



Participants prefer explanations with fewer than 32 concepts.

Vikram V. Ramaswamy, et al. Overlooked factors in concept-based explanations: Dataset Choice, Concept Learnability, and Human Capability. *CVPR,* 2023.

Discussion: Where do we go from here?

- We show that concept-based explanations can
  - be heavily dependent on the probe dataset,
  - use concepts that are hard to learn, and
  - be more complex than people can understand.
- Some immediate suggestions:
  - Choose probe dataset with similar distribution to the training dataset, use easily learnable concepts, restrict number of concepts in explanations.
- Future work:
  - Collect more diverse and high-quality probe datasets.
  - Develop more causal explanations, which can go beyond exploiting correlations between model predictions and concept occurrences.

Goal: Understand effects of decisions made by different concept-based explanations.

Consider 3 different aspects:

- Dataset used to train the explanation
- Learnability of the concepts used
- Complexity of the explanation

Vikram V. Ramaswamy, et al. Overlooked factors in concept-based explanations: Dataset Choice, Concept Learnability, and Human Capability. *CVPR,* 2023.