ELUDE: Generating interpretable explanations via a decomposition into labelled and unlabelled features

Vikram V. Ramaswamy, Sunnie S. Y. Kim, Nicole Meister, Ruth Fong, Olga Russakovsky

Princeton University

Overview

- Problem: Want to explain decisions made by increasingly complex CNNs using labelled attributes.
- Solution: We propose **ELUDE: Explanation via Labelled and Unlabelled DEcomposition,** a method to decompose model's prediction into linear combination of labelled attributes and a few uninterpretable features.

Global explanation example:
Model predicts `cat` based on

$\boxed{1.2\ \texttt{fur} + 0.7\ \texttt{paw} - 0.6\ \texttt{tree} + \ldots}$ + $\boxed{1.1\, f_1 - 0.3\, f_2}$

labelled attributes — learned features

Local explanation example:
Model predicts `cat` for image x because:

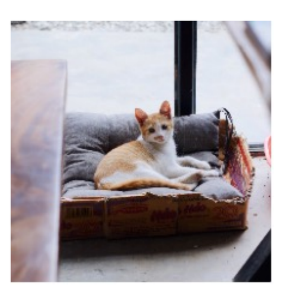
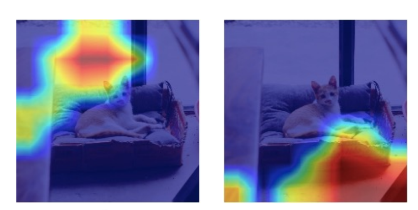


Image x

`fur`: $+1.2 \times (1)$
`paw`: $+0.7 \times (1)$
`tree`: $-0.6 \times (0)$
$= 1.9$

Presence / absence of attributes



f_1 f_2

Visualization of uninterpretable features

Method

- Given an image classification model $F = g \circ f$, such that f is linear, images $x_1, x_2, \ldots, x_N$ labelled with attributes $A(x_i) \in \{0,1\}^K$ for all $i = 1, 2, \ldots N$.
- Learn using labelled attributes: W_A to predict F as well as possible using attributes A :

$$\text{argmin}_{W_A} \sum_i CE\left(F(x_i), W_A^T A(x_i)\right) + \lambda \|W_A\|_1$$

- Learn remainder of model: low-rank uninterpretable features $U^T V$:

$$\text{argmin}_{U,V \mid rk(U)=r} \sum_i CE\left(F(x_i), W_A^T A(x_i) + (U^T V)^T f(x_i)\right)$$

Inputs:
Image x_i, Attributes $A(x_i)$,
Model output $F(x_i)$, Feature $f(x_i)$;
$i = 1, 2, \ldots N$

↓

Learn linear model W_A to predict model output $F(x)$ from attributes $A(x)$ with L1 penalty for sparsity

→

Learn remaining features as low rank space $U^T V$, such that $W_A^T A(x) + (U^T V)^T f(x) \approx F(x)$

↑

Outputs: W_A, U, V

Advantages of ELUDE.

- Simple method, can quantify fraction of model that can and cannot be explained by given attributes.
- Can identify important attributes for each class.
- Small rank implies features left to explain is simpler.

Main Takeaways

Setup

- Model to explain: Resnet18 [1] trained on Places365 [2] at 3 different resolutions (`indoor` vs `outdoor`, one of 16 scene categories and one of 365 scenes)
- Dataset with labelled attributes : ADE20k [3] labelled with Broden attributes [4]

Labelled attributes insights:

1. Fraction of the blackbox model explained using labelled attributes reduces as the model grows in complexity.

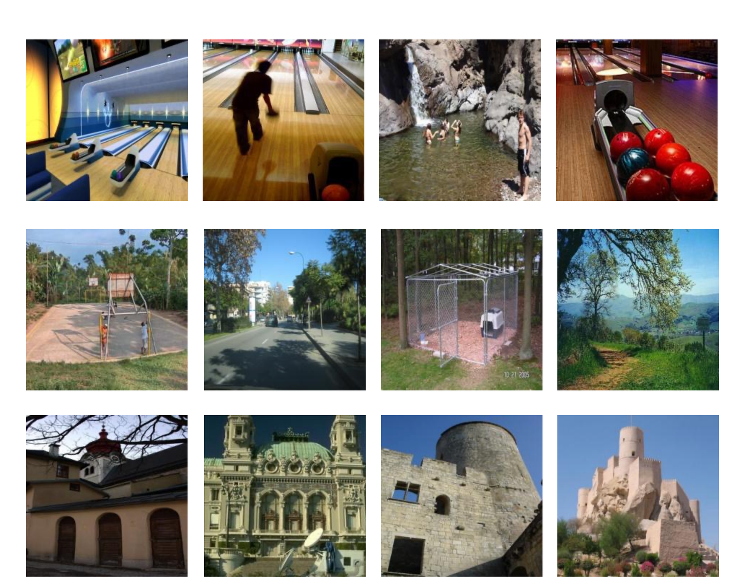
Type of model	% Explained
2-way scene classification	95.7
16-way scene classification	46.2
365-way scene classification	28.8

2. Comparison to Interpretable Basis Decomposition[5].
   - We report 6 most important attributes for 6 random scenes for 365-way classification. Attributes with positive coefficients are in `blue`, those with negative coefficients are in *`red`*.

scene name	IBD [5]	ELUDE
movie-theater /indoor	silver screen, stage, television stand, barrels, tvmonitor, seat	silver screen, microphone, stage, seat, *windowpane*, curtain
embassy	streetlight, windows, balcony, curb, mosque, slats	building, stairway, *box*, board, hedge, *floor*
jail-cell	cage, toilet, grille door, vent, ticket window, water tank	grille door, bar, *painting*, sink, bed, *sky*
auditorium	stage, seat, silver screen, barrels, stalls, grandstand	seat, stage, *painting*, piano, spotlight, *cabinet*
science-museum	case, wing, drawing, skeleton, video player, bell	pedestal, case, step, *windowpane*, ceiling, *sky*
booth/ indoor	poster, pedestal, partition, sales booth, silver screen, jacket	podium, pedestal, briefcase, spotlight, *windowpane*, person

Learned feature insights:

1. Additional concepts used by the model can be learned from the low-rank space.
   - Pictured are images that highly activate certain dimensions along with potential labels.



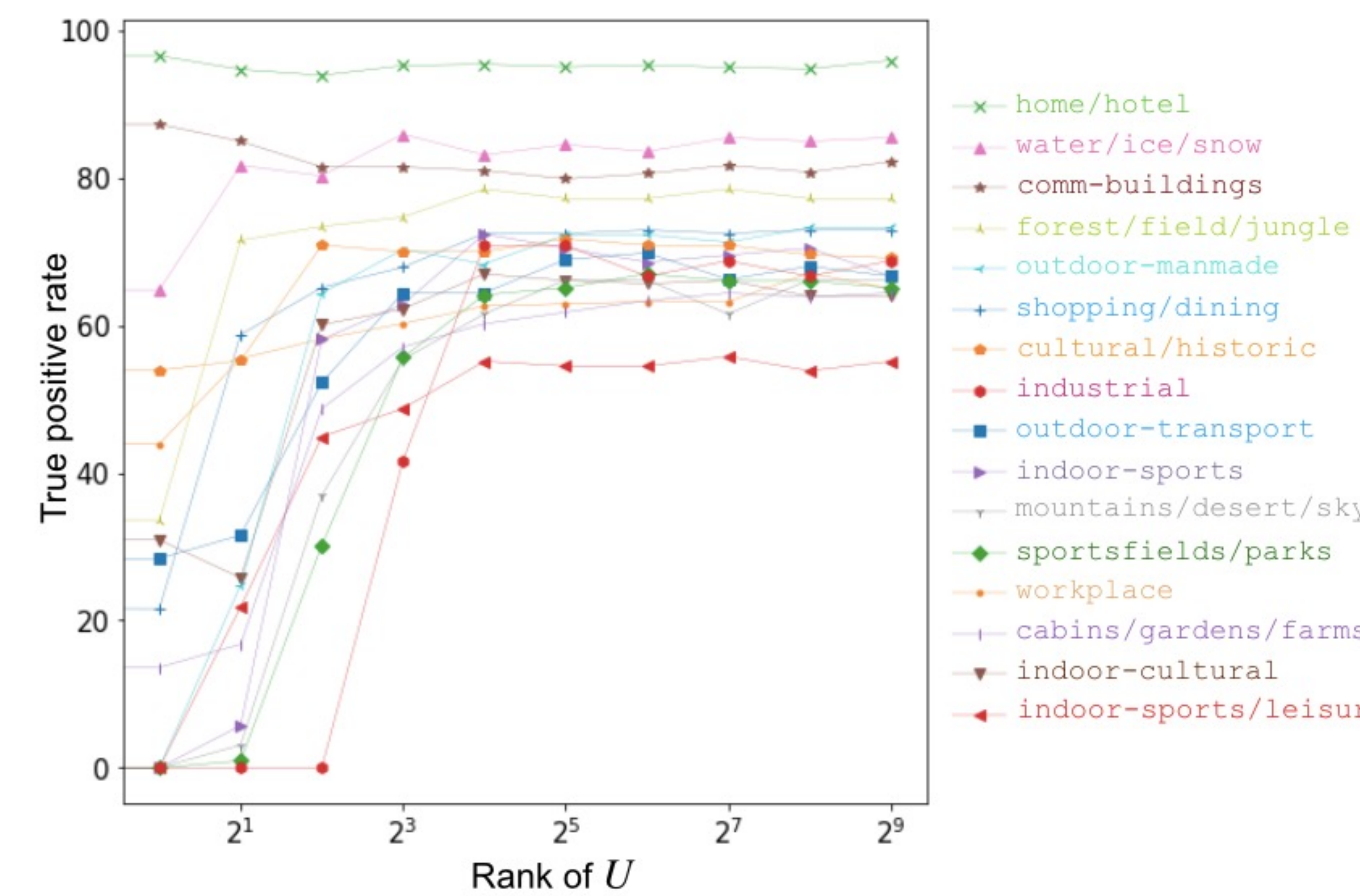
bowling alleys?

outdoor sports fields?

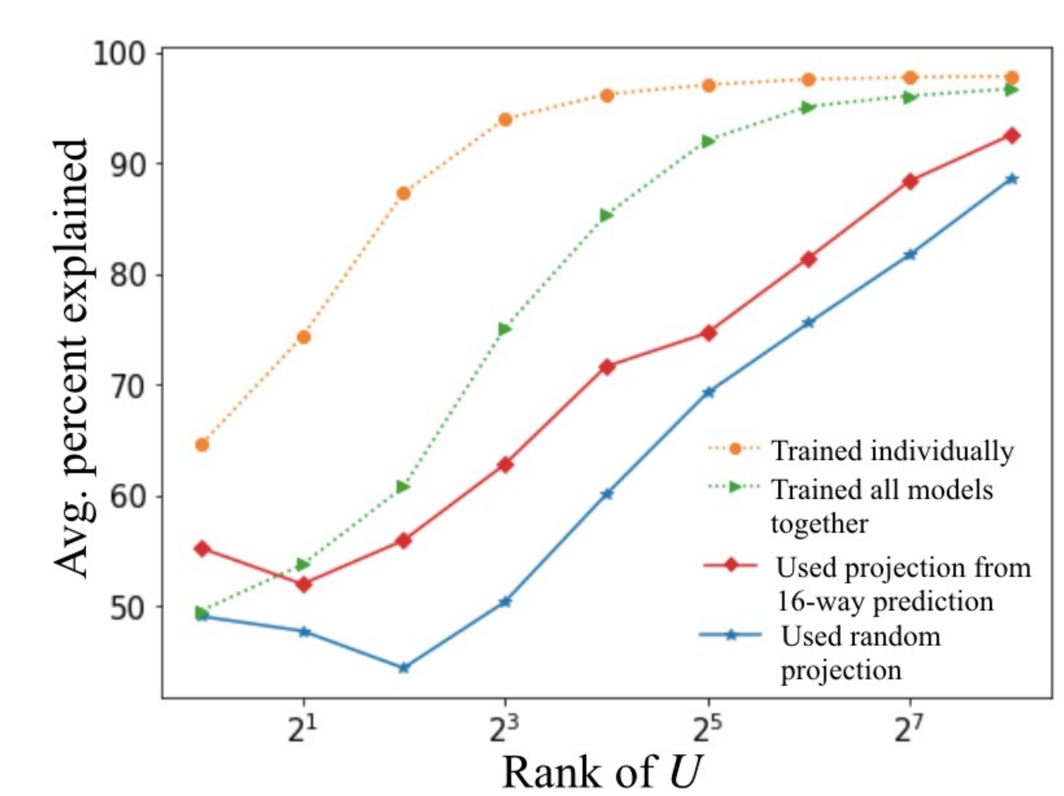
castle-like buildings?

Main Takeaways (contd.)

2. ELUDE can significantly reduce the complexity of the uninterpretable space.
   - A space of rank 8 is sufficient to explain over 75% of the model for 16-way scene classification.



3. Low-rank space U generalizes.
   - Split the original 365-way model into 16 models, each corresponding to a coarse scene category making fine-grained prediction.
   - Measure how well U trained on 16-way classification explains these models. Also measure if a common U across these models can be trained.



References

[1] He, K., Zhang, *et. al.* : Deep residual learning for image recognition. In: CVPR (2016)
[2] Zhou, B *et. al.*: A 10 million image database for scene recognition. TPAMI 40 (2017)
[3] Zhou, B. *et. al.* : Scene parsing through ade20k dataset. In: CVPR (2017)
[4] Bau, D. *et. al.*: Network dissection: Quantifying interpretability of deep visual representations. In: CVPR (2017)
[5] Zhou, B. *et al.*: Interpretable basis decomposition for visual explanation. In: ECCV (2018)

Acknowledgements. This work is sponsored by NSF grant No. 1763642, Princeton SEAS Jr Faculty award, Princeton SEAS Project X Fund and Open Philanthropy grant. We thank the authors of [5] for open-sourcing their code and models. We also thank Sharon Zhang for her initial work on this project and Moamen Elmassry, Jihoon Chung and the VisualAI lab for their feedback during the writing process.