Explainable Concept Generation through Vision-Language Preference Learning for Understanding Neural Networks' Internal Representations

Aditya Taparia, Som Sagar, and Ransalu Senanayake
ataparia@asu.edu

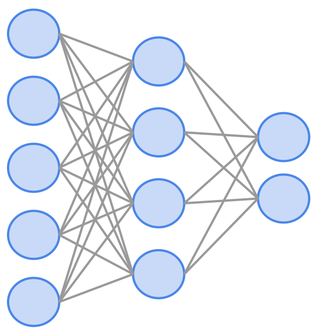
Arizona State University
LENS Lab

ICML International Conference On Machine Learning

Paper & Code

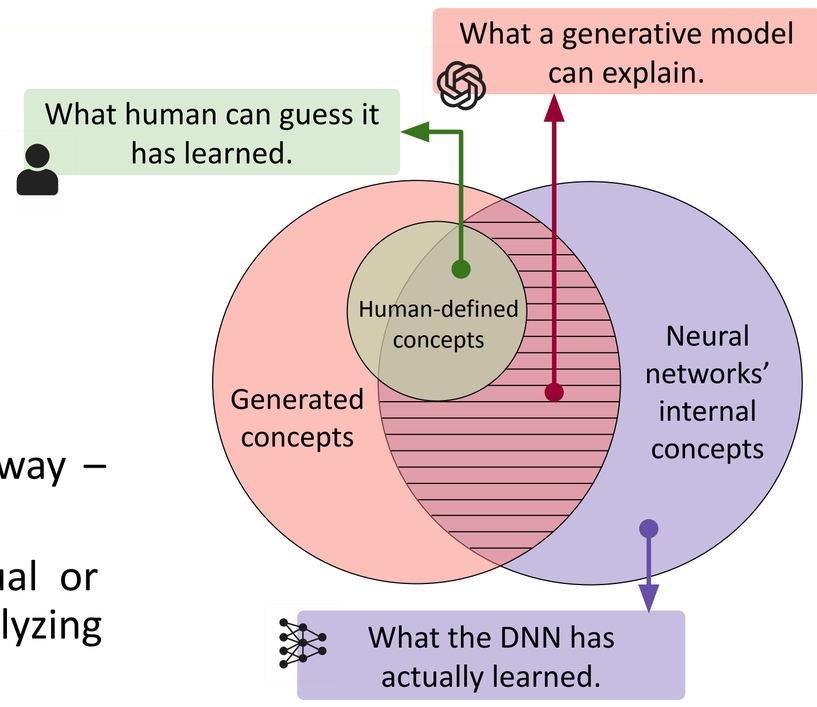
Introduction

Why did my Neural Network make that decision?



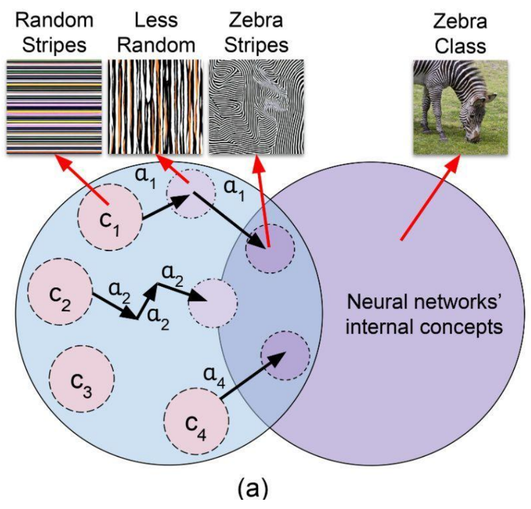
Response
"Zebra"

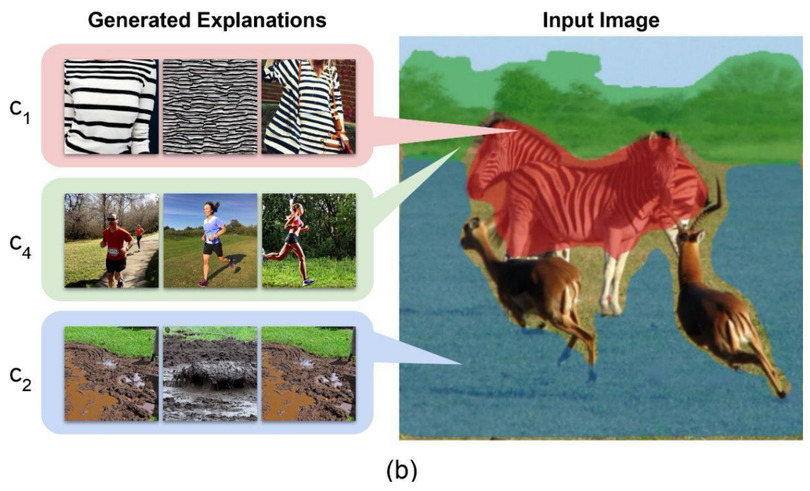
Why?
- Because there is grass.
- Because of black and white stripes.
- Random guess.

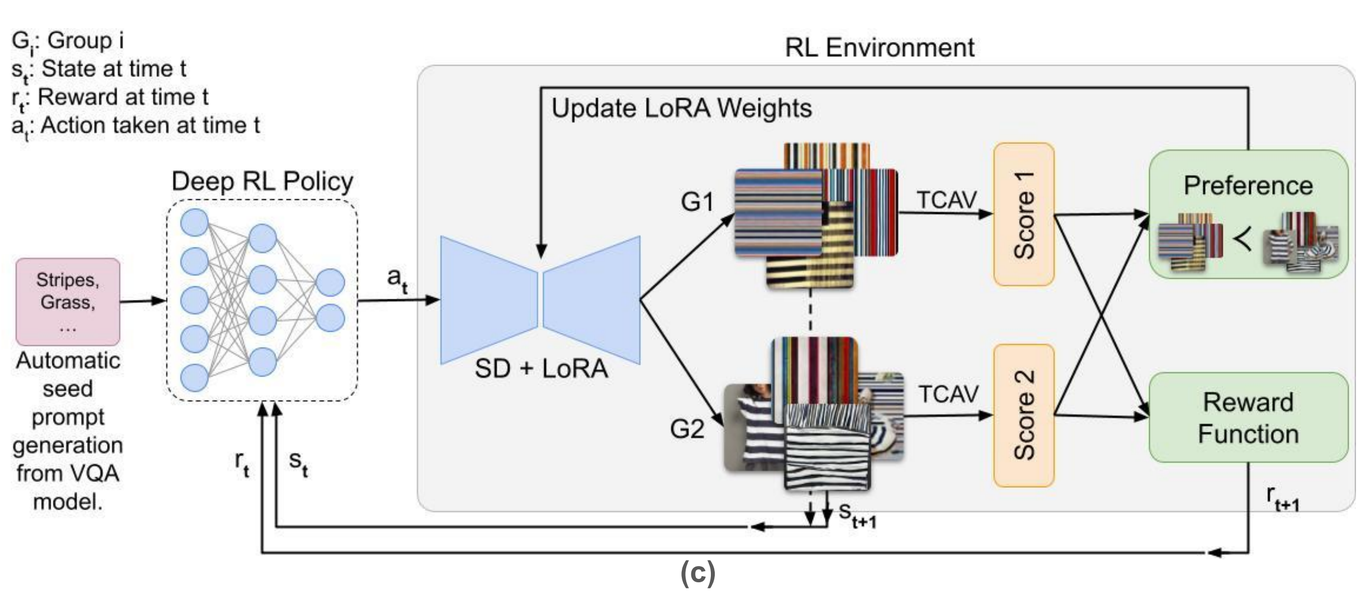


1. There could be infinite explanations when we ask why a DNN behave in a particular way – testing each individually [1] is challenging.
2. We propose a method, named RLPO, to **generate explanations**—represented as visual or natural language concepts—that truly matter to the neural network's decision by analyzing activation vectors.

Methodology







Seed Prompts: We start by questioning class images (using a VQA model) about their colors, textures, etc. After filtering, a set of top-scoring keywords is chosen as seed prompts.

Concept Generation: Using these prompts, we produce candidate concept sets.

RL-based Preference Optimization (RLPO): The RL agent selects seed prompts, guided by TCAV scores [1] as rewards.

$$TS_{c,m} = \frac{1}{|X_m|} \sum_{X_m} \mathbb{I}\left(\frac{\partial \text{output}}{\partial \text{activations}} \cdot (c \text{ direction}) > 0 \right)$$

$$= \frac{1}{|X_m|} \sum_{x_i \in X_m} \mathbb{I}\left(\frac{\partial f(x_i)}{\partial f_1(x_i)} \cdot v > 0 \right)$$

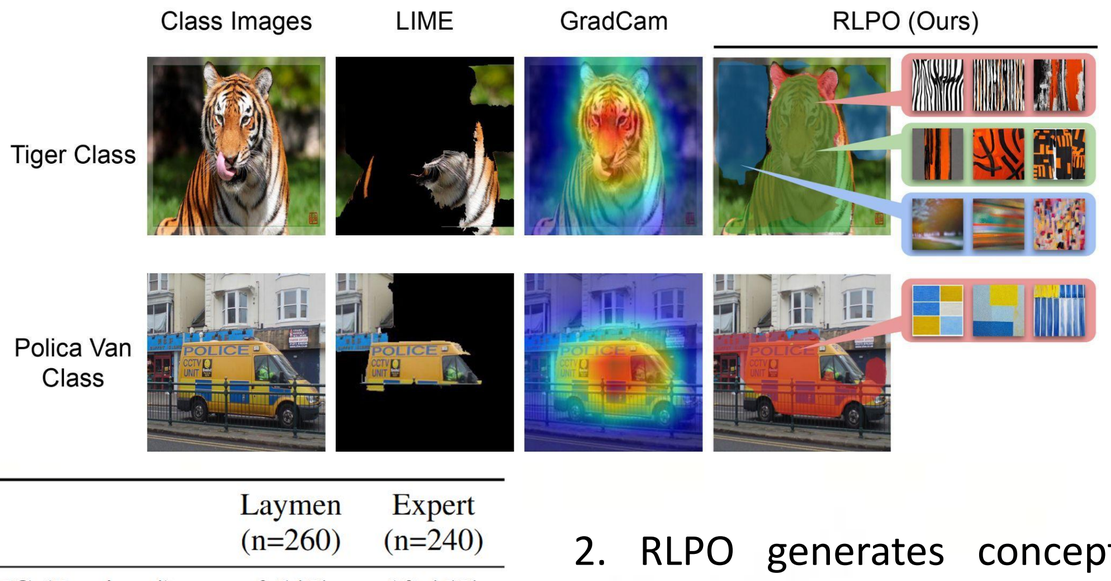
$$Q^*(s,a) = \mathbb{E}_{s' \sim P(\cdot|s,a)}[\xi_t r(s,a) + \gamma \max_{a' \in A} Q_{\text{target}}(s', a')]$$

If results are poor, preference optimization nudges the generative model toward the more promising concepts. Iteratively, the agent and preference optimization together refine the generation process, producing increasingly meaningful concepts.

Why RL? The search space is extremely large and fine-tuning is sequential and dependent on TCAV.

Experiments & Analysis

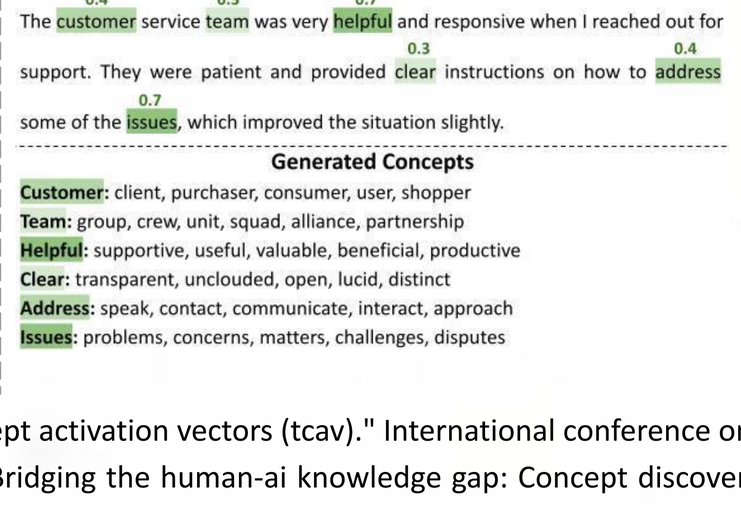
1. Our method produces **diverse, novel, and unique concepts** which triggers the DNN.



	Laymen (n=260)	Expert (n=240)
EG (Retrieval)	6.54%	10.45%
EG (Ours)	91.54%	65.45%
Odds (Retrieval)	14.29	8.57
Odds (Ours)	0.09	0.53

2. RLPO generates concepts that were **not thought by humans**.

Methods	Concepts	$TS_{c,m}(\uparrow)$	CS (↓)	ED (↑)	RCS (↓)	RED (↑)
EAC	C	1.0	0.76 ± 0.03	7.21 ± 0.63	0.67 ± 0.14	6.34 ± 2.16
Lens	C1	1.0	0.77 ± 0.02	7.17 ± 0.34	0.50 ± 0.18	9.70 ± 3.20
	C2	1.0	0.72 ± 0.04	8.02 ± 0.87	0.42 ± 0.10	10.90 ± 2.80
	C3	1.0	0.69 ± 0.05	8.45 ± 0.96	0.45 ± 0.05	11.03 ± 2.17
CRAFT	C1	1.0	0.76 ± 0.04	7.37 ± 0.62	0.57 ± 0.16	8.80 ± 3.20
	C2	1.0	0.72 ± 0.02	8.25 ± 0.39	0.50 ± 1.90	9.90 ± 3.40
	C3	1.0	0.73 ± 0.04	7.98 ± 0.79	0.44 ± 0.07	10.80 ± 1.90
RLPO (Ours)	C1	1.0	0.52 ± 0.04	10.48 ± 0.50	0.04 ± 0.01	16.80 ± 1.40
	C2	1.0	0.49 ± 0.02	10.65 ± 0.20	0.02 ± 0.02	17.20 ± 0.80
	C3	1.0	0.49 ± 0.02	10.74 ± 0.30	0.03 ± 0.01	17.60 ± 4.40



3. RLPO can reveal **influential word level concepts** in NLP tasks like sentiment analysis.

1. Kim, Been, et al. "Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav)." International conference on machine learning. PMLR, 2018.
2. Lisa Schut, Nenad Tomasev, Tom McGrath, Demis Hassabis, Ulrich Paquet, and Been Kim. Bridging the human-ai knowledge gap: Concept discovery and transfer in alphazero. arXiv preprint arXiv:2310.16410, 2023.
3. Som Sagar*, Aditya Taparia*, Harsh Mankodiya, Pranav Bidare, Yifan Zhou, and Ransalu Senanayake, "Trustworthy Conceptual Explanations for Neural Networks in Robot Decision-Making," IROS 2025.