

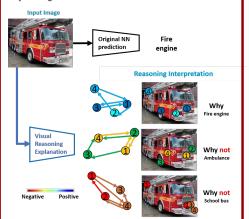
A Peek Into the Reasoning of Neural Networks: Interpreting with Structural Visual Concepts Yunhao Ge^{1,2}, Yao Xiao², Zhi Xu², Meng Zheng¹, Srikrishna Karanam¹, Terrence Chen¹, Laurent Itti², and Zivan Wu¹

¹United Imaging Intelligence, Cambridge MA ²University of Southern California, Los Angeles CA



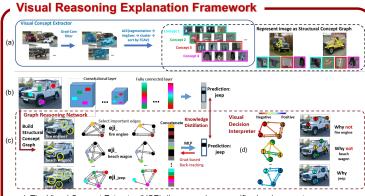
Overall

Most of the existing explanations are limited to low-level relationships and are insufficient to provide in-depth reasoning. They do not offer guidance on how to correct mistakes made by the original model.



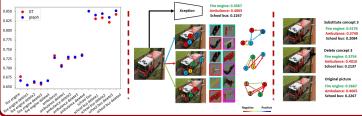
Our Contribution

- Reveal reasoning logic with the structural visual concept.
- · Answer why and why not given a prediction
- · Visual concepts + relationship between concepts
- Logical concept-level explanations
- We propose a GNN-based Graph Reasoning Network (GRN) that imitates the original NN's decision-making process with knowledge transfer and distillation.
- Our method takes a step towards diagnosing reasons for any incorrect predictions and guide improving the performance of original network.



- a) The Visual Concept Extractor (VCE) discovers class-specific visual concepts.
- b) The representation of top N concepts is distributed throughout the original NN.
- c) VRX learns the respective contributions for visual concepts and their spatial
- relationships for each potential class through distillation, to explain the network's decision.
- d) An example of why the network decides that this input is a Jeep and not others.

Logic Consistency between VRX and NN



Sensitive and Diagnosis Experiments







Fire engine: 0.18 Ambulance: 0.57 School bus: 0.25 Visual concept contribution: .39, -2.79, 0.38, 10.5

(b) Structure sensitive Experiments



10

Change good concepts

VRX guided correction

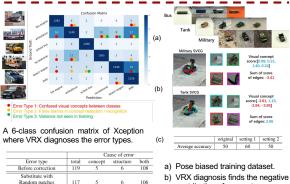
VRX-quided image editing.

Out of 119 images initially misclassified by

Xception, only 5 remain misclassified after



Fire engine: 0.71 Ambulance: 0.17 School bus: 0.12 Edges contribution [Edge_31: -0.34, Edge 32: -0.21. Edge 34: -0.1]



104

2

- contribution of concept relationships caused most incorrect predictions while visual concepts contribute positively.
- c) VRX diagnosis can guide performance improvement.