# Guaranteed Optimal Compositional Explanations for Neurons

**Biagio La Rosa and Leilani H. Gilpin**

**Looking for Faculty Positions**

University of California, Santa Cruz

Email:

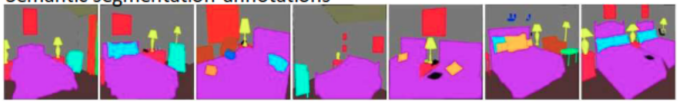
**bilarosa@ucsc.edu**

# Compositional Explanations

## 1 Guaranteed Optimal Compositional Explanations for Neurons

- Identify the logical combinations of concepts that maximize the **overlap** between the **neuron activation regions** and **human annotations**

Semantic segmentation annotations



**Human Knowledge**

**Compare**

**Activation Regions**

# Compositional Explanations

## 1 Guaranteed Optimal Compositional Explanations for Neurons

- Identify the logical combinations of concepts that maximize the **overlap** between the **neuron activation regions** and **human annotations**

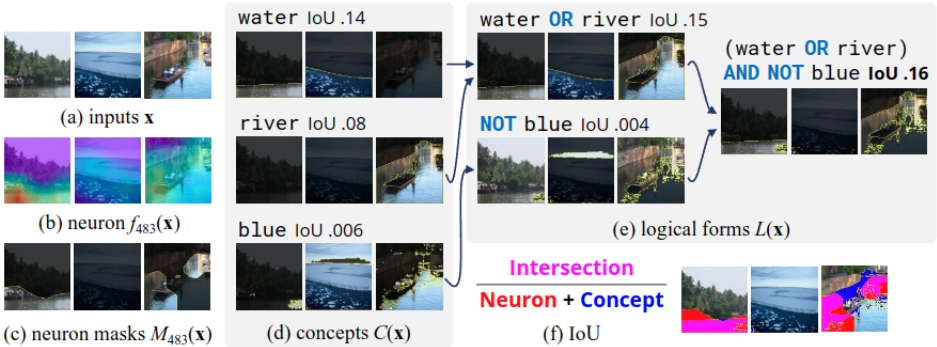
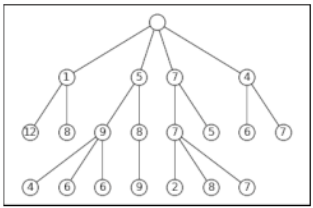



Image from *"Compositional explanations of neurons"*. Jesse Mu and Jacob Andreas. (NeurIPS 2020)

# The problem



- Computing the alignment for every possible logical combination of concepts is **infeasible in practice**
  - State space: $\sum_{k=1}^{n} n_o^{k-1} \prod_{i=0}^{k-1} (|\mathfrak{L}^1| - i)$
  - $\sim$ 400 million hours
- Solution adopted by prior work:
  - Beam search (exhaustive or informed): at every step, select only the top $b$ candidates and explore combinations involving them
- The problem:
  - Beam search **does not guarantee the optimality** of the solution



# Our Contribution

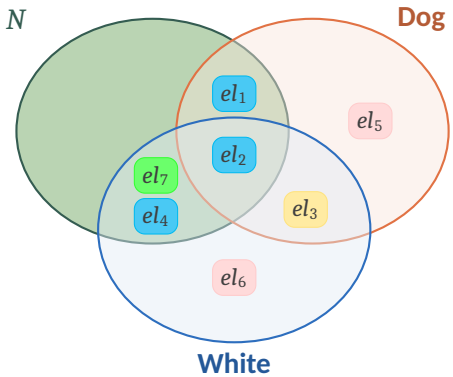
## 1 Guaranteed Optimal Compositional Explanations for Neurons

- **Decompose the alignment** (IoU) into its fundamental quantities
- **Heuristics that estimate the maximum IoU** reachable from a node by exploiting these quantities
- A heuristic-guided **algorithm** that navigates the state space and **identifies the optimal solution**

# IoU Decomposition



Combining the previously identified quantities:




- $E^U$: Unique Extras
- $E^C$: Common Extras
- $I^U$: Unique Intersection
- $I^C$: Common Intersection

$$dIoU(N, L) = \frac{|I^U(L)| + |I^C(L)|}{|N| + |E^U(L)| + |E^C(L)|}$$

# Heuristics

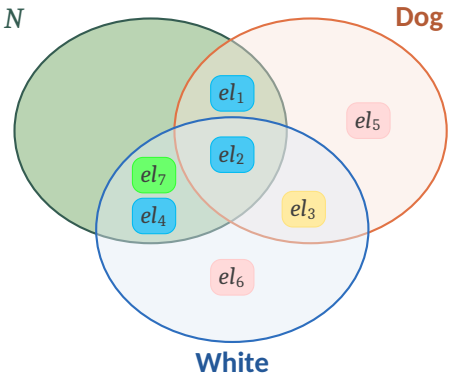


- Given a label $L$ (e.g., $((c_1 \vee c_2) \wedge c_3)$) estimate
  - Its **Label IoU**
  - Its **Path IoU**: $IoU(L^*)$, the maximum IoU reachable by chaining concepts up to a length of $n$.
- **How?**
  - Estimating the maximum and minimum values induced by operators over the identified quantities
  - Two strategies:
    - Aggregated estimation: looser but faster
    - Sample-based estimation: tighter but computationally slower

# Heuristics - An Example: the OR operator



Impact: $\vee$ **preserves** "positives" (1s) $\rightarrow$ it can only **increase** the quantities of individual concepts

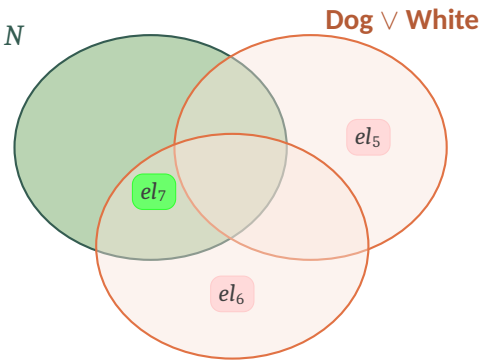



- $|E^U(Dog)| = 1$ $|E^U(White)| = 1$
- $|I^U(Dog)| = 0$ $|I^U(White)| = 1$
- $|E^C(Dog)| = 1$ $|E^C(White)| = 1$
- $|I^C(Dog)| = 2$ $|I^C(White)| = 2$

# Heuristics - An Example: the OR operator



Unique elements are naturally disjoint. The result is simply the sum.

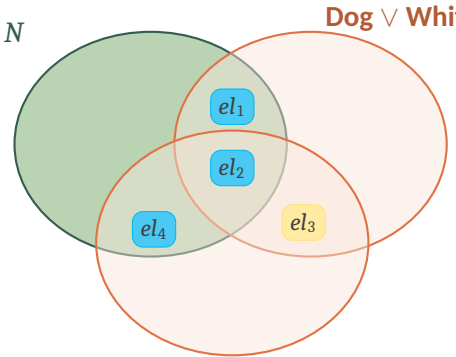



- $|E^U(Dog)| = 1$ $|E^U(White)| = 1$
  - $|E^U(Dog \vee White)| = 1 + 1 = 2$
- $|I^U(Dog)| = 0$ $|I^U(White)| = 1$
  - $|I^U(Dog \vee White)| = 0 + 1$ 1

# Heuristics - An Example: the OR operator



Common quantities may have overlap. We compute the max and min values by considering the best and worst-case scenarios.




- $|E^C(Dog)| = 1$ $|E^C(White)| = 1$
  - $|E^C_{max}(Dog \vee White)| = 1 + 1 = 2$
  - $|E^C_{min}(Dog \vee White)| = max(1, 1) = 1$
- $|I^C(Dog)| = 2$ $|I^C(White)| = 2$
  - $|I^C_{max}(Dog \vee White)| = 2 + 2 = 4$
  - $|I^C_{min}(Dog \vee White)| = max(2, 2) = 2$

# Heuristics - An Example: the OR operator



**Path IoU**: $(Dog \vee White)$ quantities + Max/Min Improvement

E.g, Unique Extras:

- Maximum Explanation Length: 3
  - $|Path_{max}(E^U, Dog \vee White, len = 3)| = 2 + 6 = 8$
  - $|Path_{min}(E^U, Dog \vee White, len = 3)| = 2 + 0 = 2$
- Maximum Explanation Length: 4
  - $|Path_{max}(E^U, Dog \vee White, len = 4)| = 2 + 11 = 13$
  - $|Path_{min}(E^U, Dog \vee White, len = 4)| = 2 + 1 = 3$

Max/Min Improvement

| Concept | $\lvert E^u \rvert$ |
|---|---|
| $C_{10}$ | 6 |
| $C_3$ | 5 |
| .. | .. |
| .. | .. |
| $C_1$ | 1 |
| $C_{14}$ | 0 |

# Algorithm: Initialize the Frontier



**Heuristic information**

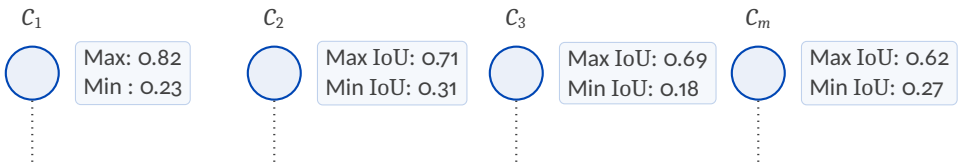
**Operator Effects**
$\vee, \wedge, \wedge\neg$

←

**Quantity Bounds**
Max/Min Improvement

←

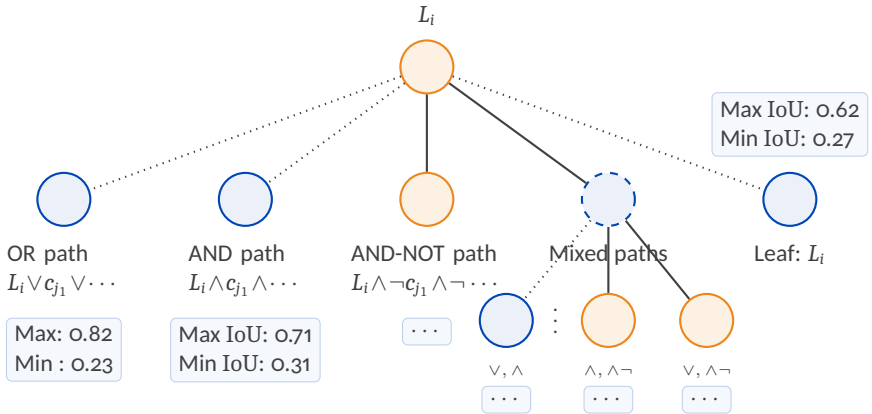
**Fundamental Quantities**
$I^U, I^C, E^U, E^C$




Initialize the frontier with single concepts. For each node, **estimate the Max and Min IoU** reachable by any **path** starting from $C_i$ by using **heuristic information**. Collect the **highest Min IoU** $\tau$ and use it to **prune** non-promising nodes.

# Algorithm: Node Expansion

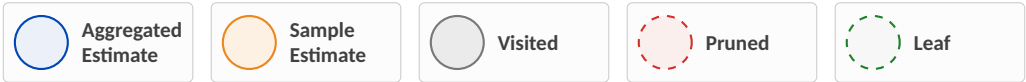


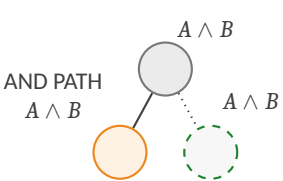


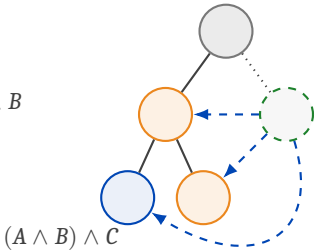

Estimate the Max and Min IoU reachable by any path starting from $L_i$. **Visit (or expand)** the most promising **sample-estimate node**. **Refine** the most promising **aggregated estimate** node.

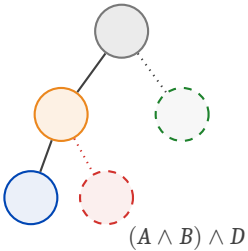
# Algorithm: Leaf Nodes, Propagation and Pruning






Compute IoU and update $\tau \leftarrow \max(\tau, \mathrm{IoU}_{exact})$

Propagate sub-label quantity information

Prune if $\widehat{IoU}_{\max} < \tau$

# Key Results: Feasibility and Performance

## 1 Guaranteed Optimal Compositional Explanations for Neurons

- Made the **problem tractable**
  - Reduced runtime from 400 million hours to 1 hour and 30 minutes
- Our heuristic-guided beam search is competitive while providing greater flexibility than existing alternatives.

| Algorithm | Visited | Exp. | Esti. | s/u |
|---|---|---|---|---|
| **Intermediate Complexity** | | | | |
| Optimal (our) | 1 | 4915 | $10^6$ | 90.57 |
| Beam | | | | |
| Our | 10 | 15 | 37956 | 11.55 |
| MMESH | 39 | 15 | 37956 | 38.42 |
| Vanilla | 37979 | 15 | - | 450 |
| **High Complexity** | | | | |
| Optimal (our) | 47 | $10^5$ | $10^8$ | 5768 |
| Beam | | | | |
| Our | 27 | 15 | 53752 | 123.33 |
| MMESH | 43 | 15 | 53752 | 102.35 |
| Vanilla | 53775 | 15 | - | 5929 |

# Key Results: Explanations Analysis

## 1 Guaranteed Optimal Compositional Explanations for Neurons

- 10-40% of **prior** compositional **explanations are suboptimal**
  - Some of them are *unverified*


| Model | Non-Optimal | Cat 1 | Cat 2 | Cat 3 |
|---|---|---|---|---|
| ResNet | 9% | 76% | 6% | 17% |
| AlexNet | 23% | 93% | 5% | 2% |
| DenseNet | 39% | 73% | 0% | 27% |
| EfficientVit | 26% | 67% | 25% | 8% |

Percentage of non-optimal explanations found by beam search and their distribution over different categories.

# Future Research

## 1 Guaranteed Optimal Compositional Explanations for Neurons

- Runtime improvements
- Optimal Explanations as a **ground-truth** reference for **future research**
  - Faster algorithms with better approximations
- Relax assumptions
  - Incrementality, non-OO preserving operators, etc.
- Better setups for non-traditional settings

# Thank You for Your Attention!

- Summary of our contribution:
  - Decomposed IoU
  - Heuristics to estimate alignment
  - Algorithm to guarantee optimality

**Looking for Faculty Positions**

Email:
**bilarosa@ucsc.edu**
Website:
**https://biagiomattialarosa.github.io**