A Broader Impact

Our proposed INVERT method significantly contributes to enhancing the transparency and safety of deep neural networks. By providing human understandable and interpretable explanations for neurons in black-box models, our approach offers valuable insights into their internal operations, improving understanding. Moreover, our method is able to identify potentially harmful representations through concept augmentation, which can be easily extended to cover a wider range of concepts. An important advantage of our method is its notable reduction in computational cost compared to previous approaches. This reduction not only improves efficiency but also minimizes the harmful environmental impact associated with excessive GPU usage.

It is important to note that we cannot make definitive claims regarding specific groups of people benefiting from or being disadvantaged by our method. The general applicability and potential implications of our approach should be explored further and with caution.

B Prior work

Let's consider a function, $g : \mathbb{D} \to \mathbb{R}^{k \times k}$, that signifies a convolutional neuron within a model that produces activation maps of dimensions $k \times k$, along with a concept $c \in \mathcal{C}$. Both Network Dissection [20] and Compositional Explanations of Neurons [22] methodologies make use of the Intersection over Union (IoU) similarity metric to measure the degree of correlation between a function and a concept. A prerequisite for these methodologies are segmentation masks pertaining to the concepts, meaning for every concept $c \in \mathcal{C}$, there exists a corresponding function $M : \mathbb{D} \to \{0,1\}^{h \times w}$, which generates a binary mask for the specific concept, of the same size as the original input.

To evaluate the correlation between function g and concept c, the multi-dimensional outputs from g are subjected to thresholding based on neuron-specific percentiles (i.e., values above chosen percentiles are converted to 1 and the remaining to 0), and upscaled to match the dimensions of the original image. We can define the resulting function that produces binary masks of the same size as the input as $G : \mathbb{D} \to \{0,1\}^{h \times w}$. The final similarity (IoU) score between g and c can be computed as the Intersection over Union score between concept masks M and function G :

$$d_{IoU}(g, c) = \frac{\sum_{x \in \mathcal{D}} \mathbf{1}\left(M(x) \cap G(x)\right)}{\sum_{x \in \mathcal{D}} \mathbf{1}\left(M(x) \cup G(x)\right)}. \tag{4}$$

Given that previous methodologies were able to procure explanations solely from convolutional neurons, we carried out a comparison with INVERT by computing INVERT explanations through the mean of the activations across the activation maps of convolutional neurons. Specifically, in section 4.2, the method of Compositional Explanations of neurons was applied using a 7x7 input map for each feature. Conversely, the INVERT approach uses a strategy that computes a scalar value by calculating the average of the input map.

C INVERT algorithm

Given a neural representation $f : \mathbb{D} \longrightarrow \mathbb{R}$, a dataset $\mathcal{D} \subset \mathbb{D}$, and a set of concepts $\mathcal{C} \in \mathbb{C}$, the INVERT approach seeks to identify a compositional concept φ^*, which is formed as a logical operation on the concepts, to optimize AUC similarity $d(f, \varphi^*(\mathcal{C}))$. For this purpose, we utilized an optimization process similar to that of the CompExpl methodology [22], employing Beam search to find the optimal compositional concept.

This method requires the configuration of certain parameters, namely the predetermined formula length $L \in \mathbb{N}$, the beam size $B \in \mathbb{N}$, and additionally, the threshold $T \in (0, 1/2)$. Beam search intends to iteratively combine concepts, starting with the atomic concepts (primitives) from $\mathcal{C}$. At every iteration of the process, the top B best-performing compositional concepts are selected, and all feasible formulas are computed with primitives. Subsequently, only the top B best-performing concepts are selected, and the process continues until the formula reaches the predetermined length.

In detail, firstly, we define a set of primitives $\bar{\Phi}$ — a set of compositional concepts that correspond to the set of concepts $\mathcal{C}$ and their negation. The set $\bar{\Phi}$ comprises $2k$ compositional concepts, with each concept corresponding to either the base concept or its negation. Next, all $2k$ concepts are evaluated in terms of AUC similarity with a given function, and the top B best performing compositional concepts, that satisfy $p(\varphi(\mathcal{C})) \geq T$ are selected, leading to the formation of the set Φ^* where $|\Phi^*| = B$, referred to as a Beam. These are the top B best-performing compositional concepts with a length of 1, satisfying the requisite condition on their positive fraction in the dataset. Subsequently, the following operations are iteratively performed until the predetermined formula length L is met:

1. Each of the B compositional concepts in the beam Φ^* is combined with all primitives using either the AND or OR operation, thereby augmenting the formula length by 1, resulting in a total of $4Bk$ new formulas.
2. All newly generated formulas are evaluated based on their similarity to the representation, and the beam Φ^* is updated to include the top B performing formulas, which satisfy the condition $p(\varphi(\mathcal{C})) \geq T$.

Upon reaching the predetermined formula length L, the Beam-Search procedure concludes by identifying the compositional concept φ^* with the highest observed AUC.

D Comparing Fuzzy Logic operators

Fuzzy logic operators [38] serve as essential instruments within the domain of fuzzy logic, a mathematical construct designed for modeling and handling data that is imprecise or vague. This contrasts with conventional logic where an element strictly either belongs to a set or not; fuzzy logic allows for the degree of membership to vary from 0 to 1, thereby allowing for partial membership.

In the present study, our objective was to contrast different categories of fuzzy logic operators and examine their behavior concerning the proposed AUC metric. To fulfill this aim, we employed four distinct deep learning image classification models: AlexNet [33], DenseNet161 [34], EfficientNet B4 [35], and ViT 16 L [36]. Each of these models was pre-trained on the ImageNet dataset. We focused on 1000 neural representations in the output logit (pre-SoftMax) layer for each model, for which we recognized the 'ground-truth' concept — the corresponding ImageNet class. For fuzzy logic operators' testing, we mapped the output of each individual representation to the set $[0, 1]$ by normalizing each representation's output using their corresponding mean and standard deviation across ImageNet dataset and applied a Sigmoid transformation. We tested four different Fuzzy logic operators, specifically Gödel, Product, Łukasiewicz, and Yager with parameter $p = 2$, as illustrated in Table 2.

For performance evaluation, we generated random compositional concepts of a given length and computed the AUC similarity between fuzzy logic norms applied to functions corresponding to these concepts. For instance, given the random compositional concept $\varphi = c_i$ OR c_j, we derive compositional representations as per each of the four examined methods (e.g., the Gödel operator produces function $h_G = \max(f_i, f_j)$). These compositional representations are then evaluated in terms of AUC similarity with the compositional concept.

We conducted the evaluation in two modes, that is, assessing the performance of the OR (T-conorm) operator and the performance of the AND (T-norm) operator. For each mode, we assembled 1000 random compositional concepts by sampling L concepts without replacement and calculated the AUC between compositional concepts. Note that for the second mode, AND (T-norm), random compositional concepts were assembled using the AND NOT operation, given the mutual exclusivity of ImageNet labels.

Figures 8 and 9 depict the mean AUC similarity between random compositional concepts of varying lengths and the corresponding compositional representations, which were assembled using four distinct fuzzy logic operators. From these figures, it becomes evident that Gödel fuzzy logic operators demonstrate the most significant robustness to the length of the formula, consistently attaining superior AUC in contrast to other operators. Consequently, we can infer that Gödel's operator emerges as the optimal choice for implementing fuzzy logic operations on these representations.

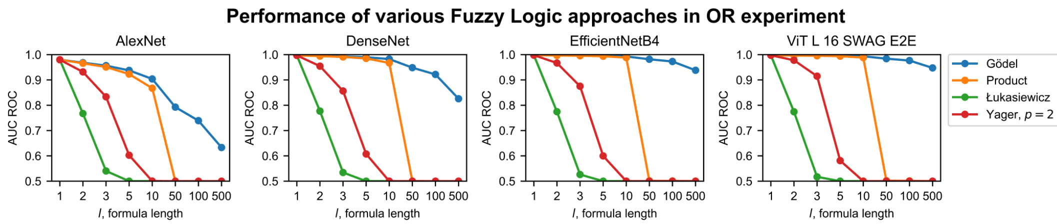

Figure 8: Average AUC similarity between random compositional OR concepts and corresponding compositional representations employing various Fuzzy logic operators (Higher is better) evaluated across four distinct models.

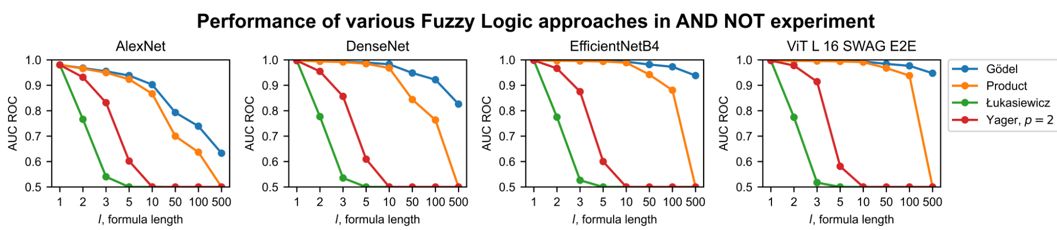

Figure 9: Average AUC similarity between random compositional AND NOT concepts and corresponding compositional representations employing various Fuzzy logic operators (Higher is better) evaluated across four distinct models.

Table 2: List of different fuzzy operators

	NOT(a)	AND(a, b) *(T-norm)*	OR(a, b) *(T-conorm)*
Gödel	$1 - a$	$\min(a, b)$	$\max(a, b)$
Product		$a \cdot b$	$a + b - a \cdot b$
Łukasiewicz		$\max(a + b - 1, 0)$	$\min(a + b, 1)$
Yager, $p = 2$		$\max(1 - ((1 - a)^2 + (1 - b)^2)^{\frac{1}{2}}, 0)$	$\min((a^2 + b^2)^{\frac{1}{2}}, 1)$