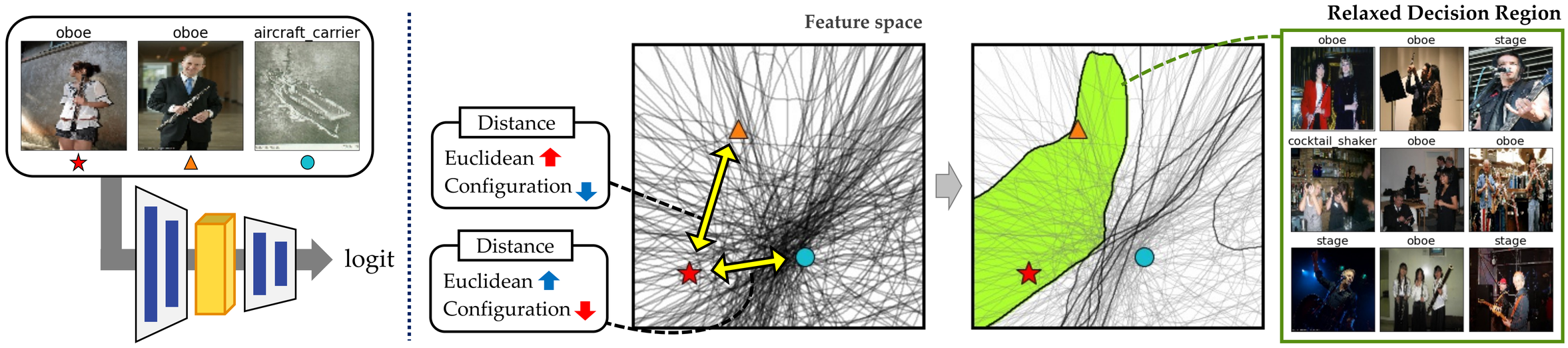
Understanding Distributed Representations of Concepts in Deep Neural Networks without Supervision

Wonjoon Chang[1*], Dahee Kwon [1*], and Jaesik Choi [1,2]

[1] Korea Advanced Institute of Science and Technology, Korea, [2] INEEJI, Korea
[*] Equally Contributed
email: {one_jj, daheekwon, jaesik.choi}@kaist.ac.kr

Relaxed Decision Region (RDR)



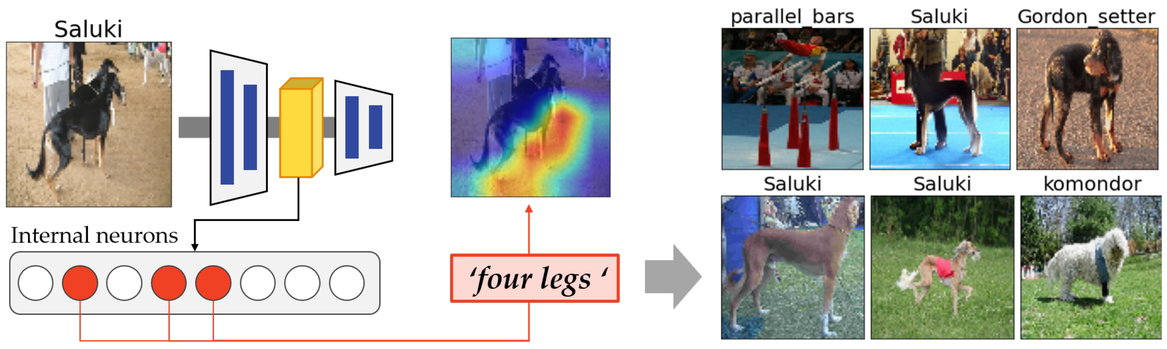
We suggest an interpretable region in the intermediate feature space where encompassing samples share coherent concepts with a target instance. In this process, we propose a novel metric, namely Configuration Distance, which computes the difference in activation states. It enables automatic evaluation of the concept similarity without supervision. Based on the Configuration distance, we can select principal neurons that construct RDR.

Problem

- **Problem:** It is crucial to identify the internal representations that DNN implicitly learned for DNN interpretability. However, existing methods often depend on human supervision, incurring significant costs. The challenges in an unsupervised approach arise from the complex feature space where numerous subregions exist in different properties.
- **Goal:** Identify the internal representations that DNN implicitly learned without supervision.
- **Approach:** Finding principal neurons that are highly related to model decisions. The chosen neurons form an interpretable region that shares learned concepts, namely the Relaxed Decision Region.

Motivation

Why do we find specific neurons to capture representations of concepts?



- Multiple neurons can represent the implicitly learned features of DNN.
- The internal space is partitioned into subregions by neuron activations.

Configuration Distance

- For the given input x, the activation states of a set of neurons N refer to **configuration**:

$$c^N(x) = [c_{N[1]}(x), ..., c_{N[|N|]}(x)] \tag{1}$$

  - Each element denotes whether each neuron is activated or not (0/1).
- Given an instance $x, \tilde{x} \in \mathcal{X}$, the **Configuration distance** for a set of neurons N is defined as follows:

$$d_C(x, \tilde{x}) = d_H(c^N(x), c^N(\tilde{x})) \tag{2}$$

where d_H denotes the Hamming distance.
  - It measures the difference in the mapping of the DNNs of two instances by comparing their internal activation states.

Criteria for Concepts

- **Learned Representations:** Representations of concepts are distributed across multiple internal neurons.
- **Coherence:** The concept is observable across multiple instances.
- **Discrimination:** The concept is distinguishable from others.

Method

Select t principal neurons to construct an internal region that exhibits strong coherence with a target instance $\mathbf{x}$, while ensuring distinctiveness from irrelevant instances.

$$\min_{\substack{\mathbf{c}_p \in \{0,1\}^t \\ N^* \subset N}} \mathbb{E}_{\mathbf{x}}[d_H(\mathbf{c}^{N^*}(\mathbf{x}), \mathbf{c}_p)] - \mathbb{E}_{\mathbf{y}}[d_H(\mathbf{c}^{N^*}(\mathbf{y}), \mathbf{c}_p)] \tag{3}$$

where $x \in S$, $y \in S_{neg}$ and $|N^*| = t$.

- positive set S: automatically collect k-nearest neighbors based on d_C
- negative set S_{neg}: sample from remaining data points

We employ a greedy approach to select the configuration that minimizes the Equation (3). In the feature space, the selected configuration creates a Relaxed Decision Region (RDR) where the included instance shares an internal representation with the target instance.

Experiments & Use Cases

- **Subclass Identification**

Our framework reveals various subclasses inherent in data without any prior knowledge. It captured shapes, crowds, composition, and the degree of flowering, as well as simple color schemes.

- **Reasoning Misclassified Cases**

(a): We leverage RDR for misclassification reasoning. In the French Bulldog image, it is misclassified as Saluki due to its long, thin legs.

- **Coherence Check with Annotated Data**

(b): To assess the coherence of captured concepts, we check whether instances in RDR share similar properties by utilizing annotated data. The provided texts illustrate the shared properties among instances in RDR.