Knowledge-Driven Interpretation of Convolutional Neural Networks

Riccardo Massidda(✉) [0000−0003−0137−7793] and Davide Bacciu[0000−0001−5213−2468]

Università di Pisa
riccardo.massidda@phd.unipi.it, davide.bacciu@unipi.it

Abstract. Since the widespread adoption of deep learning solutions in critical environments, the interpretation of artificial neural networks has become a significant issue. To this end, numerous approaches currently try to align human-level concepts with the activation patterns of artificial neurons. Nonetheless, they often understate two related aspects: the distributed nature of neural representations and the semantic relations between concepts. We explicitly tackled this interrelatedness by defining a novel semantic alignment framework to align distributed activation patterns and structured knowledge. In particular, we detailed a solution to assign to both neurons and their linear combinations one or more concepts from the WordNet semantic network. Acknowledging semantic links also enabled the clustering of neurons into semantically rich and meaningful neural circuits. Our empirical analysis of popular convolutional networks for image classification found evidence of the emergence of such neural circuits. Finally, we discovered neurons in neural circuits to be pivotal for the network to perform effectively on semantically related tasks. We also contribute by releasing the code that implements our alignment framework.

Keywords: Interpretability · Convolutional Neural Network

1 Introduction

Neural representations offer limited insights in terms of human-level interpretation. Overcoming this limitation is one of the most compelling challenges in deep learning research and is crucial when considering artificial neural networks deployed for safety- and privacy-critical tasks. Because of the opacity of their internal behavior, the literature tends to define neural networks as black boxes [10]. Nonetheless, recent research highlights how, in particular domains, some components of a neural network might instead be characterized by clear-cutting intepretations [20,8]. For this reason, both theoretical research and practical interpretability approaches require sound methods to reliably and accurately identify associations between high-level concepts and neural components.

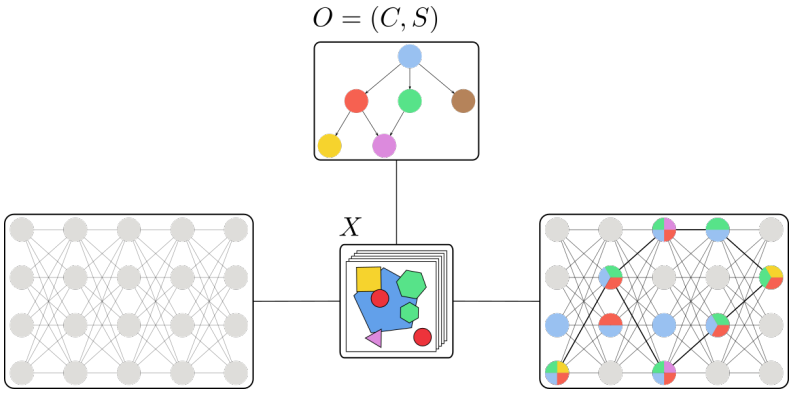


Fig. 1. Overview of the proposed methodology. A set of neural directions D is semantically aligned with an ontology O through a pixel-level annotated dataset X, whose labels are in a two-way relationship with the ontology concepts C. Semantic relations S enable the retrieval of subgraphs composed of architecturally connected and semantically related directions.

Early works tackled the alignment of human-level concepts with either distinct units [2] or directions within the output space of hidden layers [14]. Nonetheless, they considered concepts as independent entities, without adopting structured knowledge representation. In this context, we present a unified approach for the semantic alignment of neural components and visual concepts, applied to Convolutional Neural Networks (CNNs) and computer vision scenarios. Our approach considers concepts as members of a computational ontology and actively exploits their semantic relations (Figure 1). Firstly, we improve the expressiveness of the alignment by acknowledging specialization between concepts. For instance, if an artificial neuron responds to the human notion of "feline", the framework can propagate the partial alignment with the concepts of "cat" or "tiger" without the need for explicit "feline" annotations. Consequently, we tackle the identification of semantically aligned components in two complementary scenarios: by selecting concepts aligned to a given direction, and by retrieving directions aligned to a given concept. Lastly, our main original contribution leverages semantic alignment to identify meaningful subgraphs composed of architecturally connected and semantically related components within the network. We refer to these subgraphs as circuits, following the term "neural circuit" and its widespread use in neuroscience. These circuits offer new insights into the content of distributed neural representations and provide a novel instrument for network inspection and interpretation.

We validate our approach by inspecting several renowned CNN architectures for scene classification by exploiting the Broden segmented dataset [2]. As a side contribution of our empirical validation, we extend the original Broden dataset by associating its labels with WordNet synsets [17]. We consider WordNet as a simple ontology that contains a taxonomy of concepts [18]. Furthermore, we pub-

licly release the extension of Broden within the supplementary materials. While the main discussion focuses on the alignment with the Broden dataset, we also experimented with ImageNet [3], whose results we report in the supplementary materials. Empirical results highlight how our proposal yields to the emergence of meaningful neural circuits that are pivotal for the correct classification of semantically related visual categories and consitute an insightful visualization of the inner workings of the network.

2 Related Works

Zhou et al. [25] are among the first to highlight the emergence of object detectors within hidden units of CNNs trained to perform scene classification on the Places dataset [27]. Their work manually annotated such detectors by visualizing manipulated examples that maximized units activations. Olah et al. [20] approached the problem similarly by employing feature visualization techniques [21] to manually assign specific roles to individual neurons. Their contribution also highlights the role of neural units to fulfill complex tasks throughout the network. Bau et al. [2] introduced Network Dissection to automatically analyze neural activations and identify meaningful neurons in CNNs trained on the Places-365 dataset [26]. Their work introduced a pixel-level annotated image dataset called Broden, which marks portrayed objects and patterns. Zhou et al. [29] studied the role of semantically aligned units by measuring the accuracy drop when removing units aligned to a given concept. On top of Network Dissection, Mu et al. [19] discussed the consequences of analyzing compositions of visual concepts by applying logical operations to the annotations. Despite the different methodological approaches, the works discussed above analyze neural models by considering single units as meaningful artifacts as in localist networks [23]. Instead, our proposal acknowledges and investigates single units, their linear combinations, and knowledge-driven generated clusters.

More generally, other techniques aim to fulfill concept-based analysis of neural activations without restraining meaningful information to single neurons. Firstly, Fong et al. [6] expanded Network Dissection with linear combinations of hidden neurons in CNNs to identify distributed concept detectors. Similarly, Kim et al. [14] defined concept activation vectors (CAVs) as linear classifiers of visual concepts over the activations of an hidden layer. Furthermore, they proposed a measure, called TCAV, of the influence of these classifiers on specific outcomes of the network. Always using linear classifiers, Zhou et al. [28] proposed an interpretative framework based on the decomposition of hidden representations into a meaningful basis composed by such classifiers. While exploiting the expressiveness of hidden layers, these techniques considered concepts as independent entities, missing to acknowledge their semantic relations. On the contrary, we explicitly use ontological information to obtain interpretative results.

Finally, our approach might be understood in terms of ontology matching, i.e. the task of meaningfully aligning different ontologies to reduce the gap between overlapping representations [22]. Our work can be associated with extensional

based techniques, where the semantic distance between concepts from two different ontologies is estimated according to a measure of the overlapping of their extensions [4]. In comparison, our approach exploits the portrayal of visual concepts to mediate their extension and estimates the difference between an explicit ontology and concepts implicitly expressed by neural representations.

3 Ontology-Driven Semantic Alignment

Given a pre-trained CNN for computer vision, our framework estimates semantic alignment between a set of visual concepts C and a set of neural directions D. We consider directions within the output space of neural layers as a useful instrument to inquire which concepts the network is able to effectively represent and discriminate. Formally, we define a neural direction $d \in D$ as a pair

$$d = (l, v), \tag{1}$$

where $v \in \mathbb{R}^{N_l}$ is a vector weighting the N_l units at the l-th layer of the network. Given an input image x, the output of a convolutional layer l is a tensor $f^l(x) \in \mathbb{R}^{N_l \times H_l \times W_l}$, where each unit corresponds to a channel. Furthermore, we treat fully connected layers as a specialization where $H_l = 1, W_l = 1$. For an input image x, the output of a neural direction d is the activation map

$$A_d(x) = f^l(x) \cdot v, \tag{2}$$

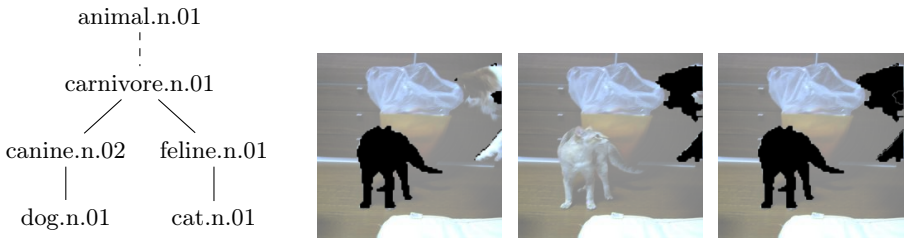
where $A_d(x) \in \mathbb{R}^{H_l \times W_l}$. Notably, when v corresponds to a vector $e^{(i)}$ from the canonical basis of $\mathbb{R}^{N_l}$, the activation map coincides with the output of the i-th neuron at layer l, i.e., the i-th channel. Furthermore, to simplify the notation, we always include a bias term β within v.

Given a segmented dataset X, for each example image $x \in X$, we require the existence of a binary mask $L_c(x)$, known as the concept mask, that marks the locations portraying the visual concept c. This requirement can be fulfilled by any dataset for object detection or semantic segmentation that provides semantic labeling of pixels. Furthermore, we require an ontology $O = (C, S)$ formalized as an extensional relational structure, where C is a set of concepts and S is a set of truth valued binary relations [9].

The presence of the specialization relation in the ontology enables the retrieval of masks for concepts which were not directly annotated in the dataset (Section 3.1). Consequently, by relating activations and concept masks, our approach computes an estimate of the alignment for direction-concept pairs (Section 3.2) and enables the retrieval of directions aligned to a given concept (Section 3.3). Finally, we exploit semantic relations between aligned concepts to cluster consecutive directions into meaningful neural circuits (Section 3.4).

3.1 High-level Concept Masks

We consider each concept as an ideal function whose argument is an object of the world and whose value is a truth-value [7]. Thus, the extension E_c of a concept c,



(a) Induced taxonomy (b) cat.n.01 (c) dog.n.01 (d) animal.n.01

Fig. 2. Example of mask generation for the higher-level concept of "animal" using taxonomical information. The induced taxonomy, built over the WordNet hypernymy (*is-a*) relation, enables the retrieval of the mask by exploiting masks annotated for "dog" and "cat" concepts from the Broden dataset, without having access to explicit annotations of the concept "animal".

is the set of all the objects of the world satisfying it. The specialization relation, also known as "is-a", is the semantic relation that expresses the inclusion between the extensions of concepts in an ontology [5]. The concept c is a specialization of c' if and only if the extension $E_{c'}$ contains E_c. Formally, $c \sqsubseteq c' \iff E_c \subseteq E_{c'}$.

Given a pixel position p in an image x, we define $Q(x_p)$ as the set of objects portrayed by that location. Consequently, a boolean concept mask $L_c(x)$ annotates for each possible pixel position p whether one of the portrayed objects by x_p pertains to the extension E_c. Consequently, if a concept c specializes c', then each location in the concept mask $L_c(x)$ implies the same location in the concept mask $L_{c'}(x)$. Formally,

$$\begin{aligned} L_c(x)_p &\iff (Q(x_p) \cap E_c) \neq \emptyset \\ &\implies (Q(x_p) \cap E_{c'}) \neq \emptyset \quad \{c \sqsubseteq c' \iff E_c \subseteq E_{c'}\} \\ &\iff L_{c'}(x)_p. \end{aligned} \tag{3}$$

Specialization induces a hierarchical taxonomy represented by a Directed Acyclic Graph (DAG). In the DAG, the node corresponding to the concept c is a child of the one corresponding to c' if and only if $c \sqsubseteq c'$. Hence, for each image, the mask of a concept at a certain level of the taxonomy can be obtained indirectly as the union of the masks of its children. The proposed approach is thus able to align higher-level concepts without explicit annotations by analyzing the concept masks of its descendants in the DAG (Figure 2).

3.2 Alignment Measure

Firstly, we address how to estimate semantic alignment for a given neural direction $d \in D$ and a concept $c \in C$. For this purpose, we define a binary classifier over the activations of the neural direction to discriminate a visual concept. Therefore, for each example x in the segmented dataset X, we threshold an

activation map $A_d(x)$ into a boolean activation mask

$$M_d(x) = A_d(x) > 0. \tag{4}$$

Typically, the activation mask $M_d(x)$ has a different shape than the input x and the concept mask $L_c(x)$. To be comparable, either the concept mask or the activation mask should be scaled to respectively match their shapes. This operation approximates the relation between a neural direction and its receptive field. For any pixel location p, the outcome of the binary classifier $M_d(x)$ in p should depend solely on x_p and ideally on each $L_c(x)_p$. This approximation discards the effects of striding and padding over the receptive field of convolutional units [1], which will be subject of future research. To ease the notation, in the following we assume that either $L_c(x)$ or $M_d(x)$ have been adequately scaled to an arbitrary shape (H, W). Furthermore, we define the operator $|K|$ to count the number of true values in an arbitrary boolean mask K.

Given this formulation, semantic alignment can be estimated by adopting an arbitrary classification performance metric. Therefore, the Jaccard similarity, also known as Intersection over Union (IoU),

$$\sigma_{\text{IoU}}(d, c) = \frac{\sum_{x \in X} |M_d(x) \wedge L_c(x)|}{\sum_{x \in X} |M_d(x) \vee L_c(x)|}, \tag{5}$$

or the Sørensen–Dice coefficient, also known as F1 score,

$$\sigma_{\text{F1}}(d, c) = \frac{\sum_{x \in X} 2|M_u(x) \wedge L_c(x)|}{\sum_{x \in X} |M_u(x)| + |L_c(x)|}, \tag{6}$$

consitute insightful measures of semantic alignment.

Furthermore, we provide an original probabilistic model of the influence of visual concepts on the output of hidden directions. We model each visual concept $c \in C$ and each direction $d \in D$ as a pair of Bernoulli random variables Y_c, Z_d. We assume that directions are conditionally independent given the concepts. Consequently, we propose a measure in terms of the maximum likelihood estimate

$$\sigma_{\mathcal{L}}(d, c) = \mathcal{L}(Y_c = 1 \mid Z_d = 1) = \frac{\sum_x |L_c(x) \wedge M_d(x)|}{\sum_x |L_c(x)|}, \tag{7}$$

of a concept being in the receptive field of a direction. This measure corresponds to the recall of the classifier, and offers different interpretative insights than other more restraining measures adopted in earlier works such as σ_{IoU} or equivalently σ_{F1}. In particular, $\sigma_{\mathcal{L}}$ is of use when the vector v of a direction $d = (l, v)$ pertains to the canonical basis of $\mathbb{R}^{N_l}$, thus it represents the output of a specific unit. Ideally, in a localist scenario, each unit of the network would activate only when stimulated by a specific visual concept. In practice, for a human observer, most units are polysemantic, i.e. they respond to multiple and possibly unrelated visual concepts. By trading off precision and recall, measures such as σ_{IoU} and σ_{F1} would ignore such concepts. On the contrary, $\sigma_{\mathcal{L}}$ can effectively highlight the partial alignment of concepts in polysemantic neurons.

3.3 Direction Learning

Other than aligning existing directions, we also address the issue of learning a vector v to determine a direction $d = (l, v)$ semantically aligned to a given concept c within the l-th layer of the network. Firstly, we consider two independents splits $X_{\text{train}}, X_{\text{val}}$ of a segmented dataset X. By solving a minimization problem, we determine the vector direction as

$$\begin{aligned} v &= \underset{v}{\operatorname{argmin}} \sum_{x \in X_{\text{train}}} \sum_{p} \ell(A_d(x)_p, L_c(x)_p) \\ &= \underset{v}{\operatorname{argmin}} \sum_{x \in X_{\text{train}}} \sum_{p} \ell((f^l(x) \cdot v)_p, L_c(x)_p) \end{aligned} \tag{8}$$

where p iterates over the locations of the activation map and of the concept mask, while ℓ is an arbitrary loss function for binary classification. Consequently, we estimate semantic alignment $\sigma(d, c)$ by computing one of the previously detailed measures on the X_{val} split of the dataset.

3.4 Neural Circuits

Given a threshold τ on the estimate $\sigma(d, c)$, we retrieve a set

$$\Psi = \{(d, c) \mid \sigma(d, c) > \tau\} \subseteq D \times C, \tag{9}$$

containing sufficiently aligned direction-concept pairs. The set Ψ offers a useful interpretative instrument by collecting the human-concepts that the network is able to sufficiently discriminate within the analyzed layers. Since we are also interested in the relation between aligned concepts, we connect alignment pairs within a directed graph $G = (\Psi, E)$ such that

$$((d, c), (d', c')) \in E \iff s(c, c') \wedge a(d, d'), \tag{10}$$

where s is a binary predicate stating the similarity of ontological concepts and a is a truth-valued function ensuring that d precedes d' in the network architecture. We detail the definition of the predicate s in the experimental setup.

Furthermore, we propose to weight the edges of the graph by estimating the influence between concept-aligned directions. The TCAV measure estimates the influence of a direction in an hidden layer towards a logit within the last layer of a classifier [14]. To compute the weight w_e of an edge $e = ((d, c), (d', c'))$, we generalize the TCAV measure to estimate the influence between two hidden directions. Firstly, we consider a function

$$h(f^l(x)) = A_{d'}(x) = f^{l'}(x) \cdot v' \tag{11}$$

that given the output of the l-th layer produces the activation map of the direction d'. Then, we are able to redefine the "conceptual sensitivity" as the direc-

tional derivative

$$g_{d,d'}(x) = \lim_{\epsilon \to 0} \frac{h(f^l(x) + \epsilon \bar{v}) - h(f^l(x))}{\epsilon} \tag{12}$$

$$= \nabla_v h(f^l(x)) \tag{13}$$

$$= v \cdot \nabla h(f^l(x)), \tag{14}$$

where $\bar{v} \in \mathbb{R}^{N_l \times H_l \times W_l}$ is a tensor obtained by repeating the vector v for each possible location of the activation map $f^l(x)$. We are interested in measuring if the direction aligned with the concept c positively influences the direction aligned with c' when a portrayal of c' is in the receptive field. To do so, we measure the fraction of inputs portraying c' that were positively influenced by the direction aligned to c. Formally,

$$w_e = \frac{\sum_{x \in X} |L_{c'}(x) \wedge (g_{d,d'}(x) > 0)|}{\sum_{x \in X} |L_{c'}(x)|} - 0.5, \tag{15}$$

where the estimate is adjusted to be either positive or negative whether the count is above or below half the inputs. Consequently, a positive value of w_e signifies a positive contribution of direction d towards d', while a negative value represents a negative contribution. As in the semantic alignment, either the concept or the sensitivity masks are scaled to match the same shape and approximate the receptive field.

Typically, because of the constraint enforced by the semantic relation s, the graph G will not be a connected graph. By extracting each non-trivial connected component, we obtain a set

$$T = \{t \mid t \subseteq \Psi, |t| > 1, G[t] \text{ is connected}\}, \tag{16}$$

where each $t \in T$ is a semantically related and architecturally connected neural circuit. Since weight estimation is a costly operation, we propose to limit the analysis to edges within neural circuits.

4 Results

We introduce an alpha version of *Bisturi*[1], a free and open source PyTorch-based library for the semantic alignment of CNNs for computer vision. *Bisturi* implements our unified framework and some of its specializations such as Network Dissection [2] and TCAV [14]. The experimental analysis focuses on the semantic alignment of neural directions with visual concepts representing concrete objects. To obtain an ontologically annotated segmented dataset, we associated each object label of the Broden dataset [2] to a member of the WordNet semantic network [17]. Since WordNet contains a taxonomy of concepts and various semantic relations, we considered it as a simple ontology [18]. As speculated,

[1] https://github.com/rmassidda/bisturi

the specialization relation automatically increased the number of alignable concepts, from the original 672 object Broden labels to 1177 distinct visual concepts. We also extensively studied semantic alignment exploiting the ILSVRC11 ImageNet dataset [3,24], by generating approximated concept masks from existing bounding-box annotations. Nonetheless, while reporting results on ImageNet in the supplementary materials, the current section focuses on Broden to directly compare with previous literature.

Overall, we report an analysis of the last layers of three popular CNN architectures trained to classify the 365 different scenes and views from the Places-365 dataset [26]. In detail, we considered:

- The last three fully connected layers and the last two convolutional layers of AlexNet [15].
- The last fully connected layer and the last two residual blocks of ResNet [11]. In each residual block, we independently analyzed the two convolutional operations and the sum after the residual connection.
- The last fully connected layer and the output of the last three dense blocks in DenseNet [12].

For replicability purposes, we adopted publicly available pre-trained models from the Places-365 project[2]. We report selected significant results, but further results and discussions are in the supplementary materials.

4.1 Unit semantic alignment

Firstly, we are interested in the semantic alignment of the directions corresponding to distinct neurons in each layer. In doing this, we wish to show how we can replicate Network Dissection [2] within our framework. Furthermore, we illustrate how, in comparison, our proposal increases the number of aligned concepts and enables semantic clustering of units. To recreate the setting introduced by Network Dissection, we consider for each layer l a set of directions

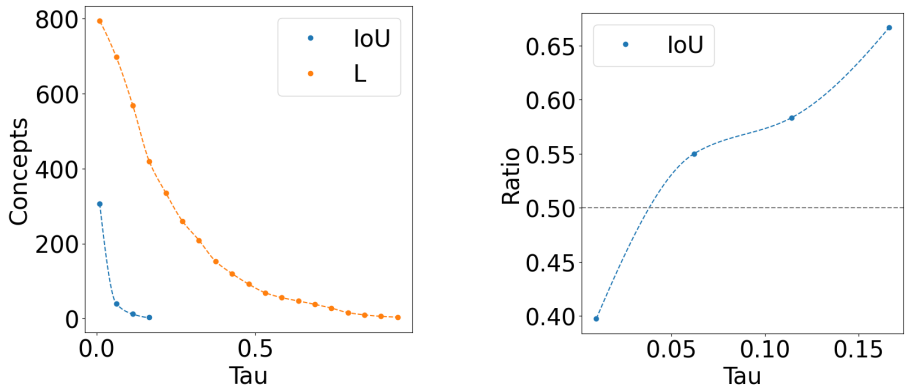
$$D_l = \{(l, [e^{(i)}; \beta_i]) \mid i \in \{0, \dots, N_l - 1\}\}, \tag{17}$$

where the bias terms β_i are concatenated to each vector $e^{(i)}$ of the canonical basis. For each direction $d = [e^{(i)}; \beta_i]$, we fix the bias term β_i such that

$$\begin{aligned} P(f^l(x)_i > \beta_i) &= P(f^l(x) \cdot e^{(i)} > \beta_i) \\ &= P(f^l(x) \cdot e^{(i)} + \beta_i - \beta_i > \beta_i) \\ &= P(A_d(x) - \beta_i > \beta_i) \\ &= 0.005. \end{aligned} \tag{18}$$

The quantity of alignable concepts is fundamental to producing numerous neural-concept associations. We found that the threshold τ highly affects the number of aligned concepts, which rapidly decays for σ_{IoU}. On the other hand,

[2] https://github.com/CSAILVision/places365



(a) Unique Concepts (b) Propagated Concepts Ratio

Fig. 3. Semantic alignment in AlexNet. For both σ_{IoU} and $\sigma_{\mathcal{L}}$, subfigure (a) plots the number of distinct concepts aligned as the threshold τ varies. Our proposal of adopting $\sigma_{\mathcal{L}}$ results in an higher number of aligned concepts. For σ_{IoU} against increasing τ, Subfigure (b) reports the fraction of aligned concepts obtained by mask propagation over the number of unique concepts. Our propagation strategy produced more than half of the concepts aligned by σ_{IoU} even for small values of τ.

since our proposal of adopting $\sigma_{\mathcal{L}}$ is less restrictive, we found an higher number of aligned concepts (Figure 3A). Furthermore, we gained empirical confirmation of the advantage in increasing the number of concepts via mask propagation. In all three target networks, we verified how a larger pool of alignable concepts effectively results in a higher number of associations. Remarkably, higher-level concepts account for a significant fraction of the concepts aligned by the IoU measure (Figure 3B). Therefore, our mask propagation strategy effectively improves the outcome of the Network Dissection approach, by providing concepts that could not have been aligned otherwise.

As expected, we consistently verified how σ_{IoU} and $\sigma_{\mathcal{L}}$ target different aspects of semantic alignment (Figure 4). The former highlights concepts that activate a unit in an exclusive way, while the latter identifies concepts producing higher than usual activations. Thus, we gained empirical confirmation that the measure $\sigma_{\mathcal{L}}$ is more apt to estimate semantic alignment when a unit responds to multiple visual concepts.

Given the aligned directions, we retrieve neural circuits by linking two concepts if their Jiang-Conrath similarity [13] overcomes a given threshold t_δ. The threshold t_δ also influences the quantity of circuits retrieved: lower values of t_δ cluster Ψ into a fully connected graph, while larger values minimize the number of connections. Furthermore, by aligning more concepts, measure $\sigma_{\mathcal{L}}$ is more apt for the retrieval of neural circuits (Figure 5).

Zhou et al. [29] tested the importance of hidden neurons in a classifier by ablating them and measuring the most affected classes. We replicate their analysis

Concept	σ_{IoU}	Concept	$\sigma_{\mathcal{L}}$
hovel.n.01	0.021	circus_tent.n.01	0.401
roof.n.03	0.025	greenhouse.n.01	0.403
building.n.01	0.031	shed.n.01	0.469
shelter.n.01	0.035	pavilion.n.01	0.568
house.n.01	0.098	bandstand.n.01	0.631

Fig. 4. Semantic alignment of unit 196 in the last residual block (`layer4.1`) of ResNet-18. We report the ten images from Broden maximally activating the unit and the top-5 aligned concepts according to respectively σ_{IoU} and $\sigma_{\mathcal{L}}$. While both measures identify visual concepts that can be found in these images, $\sigma_{\mathcal{L}}$ produces a list of more specialized concepts within the taxonomy.

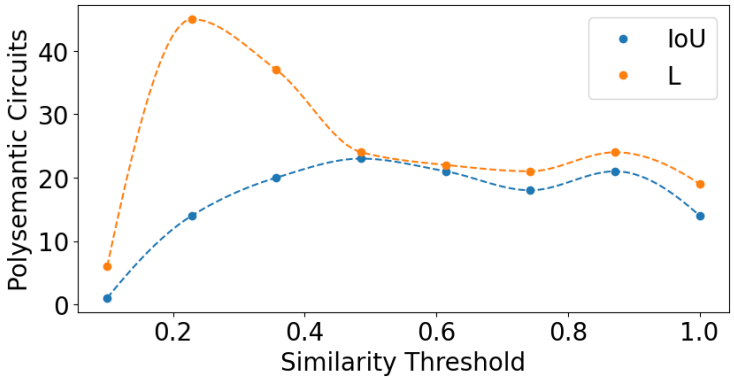
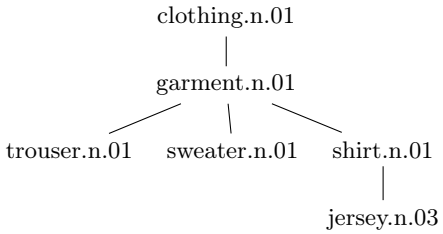
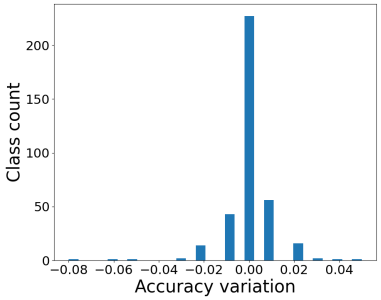


Fig. 5. Number of circuits retrieved in ResNet-18 as t_δ varies. Alignment pairs filtered according to $\tau_{\text{IoU}} = 0.04$ and $\tau_{\mathcal{L}} = 0.2$, resulting in a comparable number of aligned concepts. Our proposed measure $\sigma_{\mathcal{L}}$ produces an higher number of meaningful circuits.

by considering hidden units clustered by our circuit retrieval strategy. We measure the drop on the Top-5 classification accuracy of the 365 distinct classes from the Places-365 dataset. In general, we found that the ablation of a circuit significantly drops the accuracy of a small number of classes that are, furthermore, related to the aligned concepts (Figure 6). Targeted accuracy drop highlights how circuits cluster important units for specific tasks, resulting in a valuable instrument to understand which concepts positively affect given outcomes. As control, ablating only the units aligned to the most popular concept in a circuit results in less damaging accuracy drop.



(a) Aligned concepts



(b) Circuit Accuracy Drop

Class	Drop
/c/clothing_store	-0.08
/g/gift_shop	-0.06
/d/dressing_room	-0.05
/f/fabric_store	-0.03
/c/closet	-0.03

(c) Circuit Ablation

Class	Drop
/c/clothing_store	-0.05
/g/gift_shop	-0.05
/f/fabric_store	-0.04
/h/hardware_store	-0.03
/c/castle	-0.02

(d) `jersey.n.01` Ablation

Fig. 6. Importance analysis of a circuit found in the hidden layers of AlexNet using our proposed measure $\sigma_{\mathcal{L}}$ against $\tau_{\mathcal{L}} = 0.3$. Similarity between concepts constrained to be over $t_{\delta} = 0.2$. The circuit contains 27 distinct units aligned to 6 clothing-related visual concepts, reported according to the WordNet taxonomy in Subfigure (a). When ablating the circuit, accuracy drop significantly affects only a small number of semantically related classes, as visualized in Subfigure (b). As control, we also ablate the most popular concept in the circuit and verify how the accuracy drop is more sparse and less damaging, as in Subfigure (c). Finally, Subfigure (d) depicts the histogram of categories of the Places-365 dataset as a function of the accuracy drop (on the x-axis).

4.2 Direction learning

Unlike the TCAV [14] approach, we want to cluster hidden representations of concepts and test their reciprocal influence. To obtain aligned directions, we independently fit a neural direction $d = (v, l)$ for each concept c in each layer l of the network on a sample of the Broden dataset. As discussed in Section 3.3, a classifier on the visual concept c should be able to recreate the concept mask $L_c(x)$. We addressed the natural unbalancing of visually segmented datasets by weighting images according to the probability of extracting an example containing the concept. Consequently, we independently split the samples into a training and a validation set, with proportion $4:1$. For each example x in the training set, we applied nearest-neighbor interpolation to each concept mask $L_c(x)$ to match the shape of the activation map $A_d(x)$ and obtain the ground-truth mask. We trained the classifiers by minimizing the Focal Loss [16] between the concept mask and the activation map. Finally, we estimated their semantic alignment on the validation set using σ_{F1}.

As in the distinct-unit scenario, mask propagation significantly increased the number of aligned concepts. Furthermore, for increasing values of the threshold τ we observed that the ratio of propagated concepts over the total number of concepts increases (Figure 7). Given the higher alignment measured between concepts and directions, we consistently tested the weights of the edges within various neural circuits. We found these edges to be consistently positive, meaning that representations of similar concepts positively influence each other through the network (Table 1). As control, we also verified how randomizing the concepts of a circuit, instead, results in an average weight value of zero i.e. neither positive nor negative average influence.

Table 1. Excerpt of weighted edges between `layer4.1.conv1` and `layer4.1.conv2` in ResNet-18. Positive weight between two concepts in different layers indicate that the former positively influences the representation of the latter. Such influence is modelled after TCAV [14] and formally defined in Section 3.4. Circuit retrieved using σ_{F1} against $\tau_{\text{F1}} = 0.5$ and semantic similarity over $t_\delta = 0.7$. Overall, the circuit consists of 16 distinct learned neural directions aligned to 6 animal-related visual concepts.

	layer4.2.conv2	
	animal.n.01	placental.n.01
animal.n.01	0.114	0.210
layer4.1.conv1 vertebrate.n.01	0.358	0.368
placental.n.01	0.061	0.259

5 Conclusion

We introduced a novel framework for the semantic alignment of CNNs with a complete visual ontology. Overall, we bring three key innovative contributions.

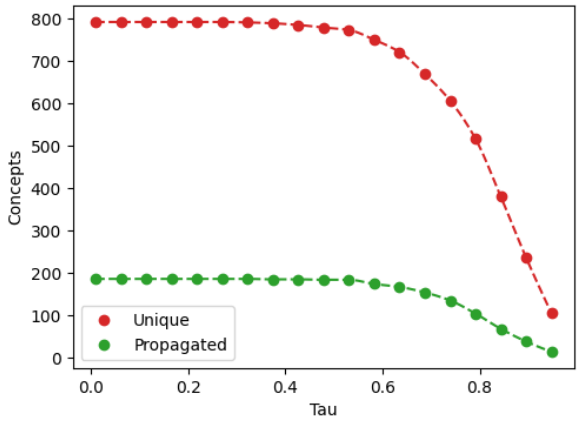


Fig. 7. For different values of the threshold τ, we measure how many of the learned directions are sufficiently semantically aligned according to the measure σ_{F1} on the validation set. In the plot, we also highlight the number of concepts obtained by concept mask propagation (in green) and the overall number of concepts (in red).

Firstly, we defined a propagation strategy to align concepts that lack an explicit annotation in the alignment dataset. Secondly, we generalized previous work on the alignment of single units and neural directions into a unified framework. Finally, we introduced an algorithm to identify connected neural circuits composed of meaningful directions. We experimentally validated our approach by aligning the WordNet ontology with three popular convolutional architectures for image classification. To this end, we considered two datasets: an original extension of the Broden dataset with ontological annotations and a bounding-box annotated subset of ImageNet. We publicly release the extended Broden dataset, the library implementing our approach, and the code used to reproduce our experiments. The experiments highlighted how our methodology can effectively capture semantic alignment. Furthermore, we assessed the emergence of semantically related neural circuits and studied their role in the overall network. This last aspect constitutes the most valuable contribution of our semantic alignment methodology: an innovative instrument to inquire about the nature of neural representations, highlighting semantically related human-interpretable features across the network and their influence towards both network outcomes and other conceptual representations.

Acknowledgments This research was partially supported by TAILOR, a project funded by EU Horizon 2020 research and innovation programme under GA No 952215.

References

1. Araujo, A., Norris, W., Sim, J.: Computing receptive fields of convolutional neural networks. Distill (2019). https://doi.org/10.23915/distill.00021, https://distill.pub/2019/computing-receptive-fields
2. Bau, D., Zhou, B., Khosla, A., Oliva, A., Torralba, A.: Network dissection: Quantifying interpretability of deep visual representations. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (Jul 2017)
3. Deng, J., Dong, W., Socher, R., Li, L., Kai Li, Li Fei-Fei: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. pp. 248–255 (2009). https://doi.org/10.1109/CVPR.2009.5206848
4. Euzenat, J., Shvaiko, P.: Classifications of Ontology Matching Techniques. In: Ontology Matching, pp. 73–84. Springer Berlin Heidelberg, Berlin, Heidelberg (2013). https://doi.org/10.1007/978-3-642-38721-0_4, http://link.springer.com/10.1007/978-3-642-38721-0_4
5. Euzenat, J., Shvaiko, P.: The Matching Problem. In: Ontology Matching, pp. 25–54. Springer Berlin Heidelberg, Berlin, Heidelberg (2013). https://doi.org/10.1007/978-3-642-38721-0_2, http://link.springer.com/10.1007/978-3-642-38721-0_2
6. Fong, R.C., Vedaldi, A.: Interpretable Explanations of Black Boxes by Meaningful Perturbation. In: 2017 IEEE International Conference on Computer Vision (ICCV). pp. 3449–3457 (Oct 2017). https://doi.org/10.1109/ICCV.2017.371, iSSN: 2380-7504
7. Frege, G.: Function und Begriff. Hermann Pohle, Jena (1891)
8. Goh, G., †, N.C., †, C.V., Carter, S., Petrov, M., Schubert, L., Radford, A., Olah, C.: Multimodal neurons in artificial neural networks. Distill (2021). https://doi.org/10.23915/distill.00030, https://distill.pub/2021/multimodal-neurons
9. Guarino, N., Oberle, D., Staab, S.: What Is an Ontology?, pp. 1–17. Springer Berlin Heidelberg, Berlin, Heidelberg (2009). https://doi.org/10.1007/978-3-540-92673-3_0, https://doi.org/10.1007/978-3-540-92673-3_0
10. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., Pedreschi, D.: A survey of methods for explaining black box models. ACM computing surveys (CSUR) **51**(5), 1–42 (2018)
11. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. CoRR **abs/1512.03385** (2015), http://arxiv.org/abs/1512.03385
12. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4700–4708 (2017)
13. Jiang, J.J., Conrath, D.W.: Semantic similarity based on corpus statistics and lexical taxonomy. In: Proceedings of the 10th Research on Computational Linguistics International Conference. pp. 19–33. The Association for Computational Linguistics and Chinese Language Processing (ACLCLP), Taipei, Taiwan (Aug 1997), https://aclanthology.org/O97-1002
14. Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., Sayres, R.: Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV). arXiv:1711.11279 [stat] (Jun 2018), http://arxiv.org/abs/1711.11279, arXiv: 1711.11279
15. Krizhevsky, A.: One weird trick for parallelizing convolutional neural networks. CoRR **abs/1404.5997** (2014), http://arxiv.org/abs/1404.5997

16. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision. pp. 2980–2988 (2017)
17. Miller, G.A.: Wordnet: a lexical database for english. Communications of the ACM **38**(11), 39–41 (1995)
18. Miller, G.A., Hristea, F.: Wordnet nouns: Classes and instances. Comput. Linguist. **32**(1), 1–3 (Mar 2006). https://doi.org/10.1162/coli.2006.32.1.1, https://doi.org/10.1162/coli.2006.32.1.1
19. Mu, J., Andreas, J.: Compositional explanations of neurons. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M.F., Lin, H. (eds.) Advances in Neural Information Processing Systems. vol. 33, pp. 17153–17163. Curran Associates, Inc. (2020), https://proceedings.neurips.cc/paper/2020/file/c74956ffb38ba48ed6ce977af6727275-Paper.pdf
20. Olah, C., Cammarata, N., Schubert, L., Goh, G., Petrov, M., Carter, S.: Zoom in: An introduction to circuits. Distill (2020). https://doi.org/10.23915/distill.00024.001, https://distill.pub/2020/circuits/zoom-in
21. Olah, C., Mordvintsev, A., Schubert, L.: Feature visualization. Distill (2017). https://doi.org/10.23915/distill.00007, https://distill.pub/2017/feature-visualization
22. Otero-Cerdeira, L., Rodríguez-Martínez, F.J., Gómez-Rodríguez, A.: Ontology matching: A literature review. Expert Systems with Applications **42**(2), 949–971 (2015)
23. Page, M.: Connectionist modelling in psychology: A localist manifesto. Behavioral and Brain Sciences **23**(4), 443–467 (2000). https://doi.org/10.1017/S0140525X00003356
24. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision (IJCV) **115**(3), 211–252 (2015). https://doi.org/10.1007/s11263-015-0816-y
25. Zhou, B., Khosla, A., Lapedriza, À., Oliva, A., Torralba, A.: Object detectors emerge in deep scene cnns. In: Bengio, Y., LeCun, Y. (eds.) 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings (2015), http://arxiv.org/abs/1412.6856
26. Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., Torralba, A.: Places: A 10 million image database for scene recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence (2017)
27. Zhou, B., Lapedriza, A., Xiao, J., Torralba, A., Oliva, A.: Learning deep features for scene recognition using places database. In: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems. vol. 27. Curran Associates, Inc. (2014), https://proceedings.neurips.cc/paper/2014/file/3fe94a002317b5f9259f82690aeea4cd-Paper.pdf
28. Zhou, B., Sun, Y., Bau, D., Torralba, A.: Interpretable basis decomposition for visual explanation. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) Computer Vision – ECCV 2018. pp. 122–138. Springer International Publishing, Cham (2018)
29. Zhou, B., Sun, Y., Bau, D., Torralba, A.: Revisiting the importance of individual units in cnns via ablation. CoRR **abs/1806.02891** (2018), http://arxiv.org/abs/1806.02891