ONTOLOGY-DRIVEN SEMANTIC ALIGNMENT OF ARTIFICIAL NEURONS AND VISUAL CONCEPTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Semantic alignment methods attempt to establish a link between human-level concepts and the units of an artificial neural network. Current approaches evaluate the emergence of such meaningful neurons by analyzing the effect of semantically annotated inputs on their activations. In doing so, they often understate two aspects that characterize neural representations and semantic concepts, namely the distributed nature of the former and the existence of semantic relationships binding the latter. In this work, we explicitly tackle this interrelatedness, both at a neural and a conceptual level, by providing a novel semantic alignment framework that builds on aligning a structured ontology with the distributed neural representations. The ontology introduces semantic relations between concepts, enabling the clustering of topologically related units into semantically rich and meaningful neural circuits. Our empirical analysis on notable convolutional models for image classification discusses the emergence of such neural circuits. It also validates their meaningfulness by studying how the selected units are pivotal for the accuracy of classes that are semantically related to the aligned concepts. We also contribute by releasing the code implementing our alignment framework.

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. Such contexts require guarantees over the behavior of a neural model, which are unachievable without a solid understanding of its inner workings. Because of the opacity of their internal behavior, the literature tends to define neural networks as black boxes (Guidotti et al., 2018). Nonetheless, recent research highlights how, in particular domains, some of the components of a neural network might instead be characterized by clear-cutting intepretations (Olah et al., 2020; Goh et al., 2021). For instance, hidden units that respond to human-level concepts autonomously emerge in various large Convolutional Neural Networks (CNNs) for image classification (Bau et al., 2017). The conditions under which this phenomenon arises still constitute an open question. For this reason, both theoretical research and practical interpretability approaches require sound methods that can reliably and accurately identify associations between high-level concepts and neural units.

In this context, the present work introduces an approach for the semantic alignment of artificial neurons with visual concepts, applied to CNN architectures and computer vision scenarios. Early works (Bau et al., 2017; 2019) considered human-level concepts as independent entities and tested their association to neural units without considering a structured representation of the semantic knowledge. Our approach, instead, regards concepts as entities of a computational ontology thus acknowledging the existence of semantic relations binding them. Figure 1 provides an intuitive depiction of our approach by hinting that an image dataset X, when pixel-wise annotated with an ontology O, enables identifying meaningful subgraphs within a set of neural units U. We refer to these subgraphs as circuits, following the term "neural circuit" and its widespread use in neuroscience.

The contribution of our approach is threefold. Firstly, we improve the expressiveness of the alignment by including a specialization semantic relation. This relation produces a more precise characterization of visual concepts and a loosening of the requirement for accurate semantic labeling of the pixels. For instance, if an artificial neuron responds to the human notion of "feline", the frame-

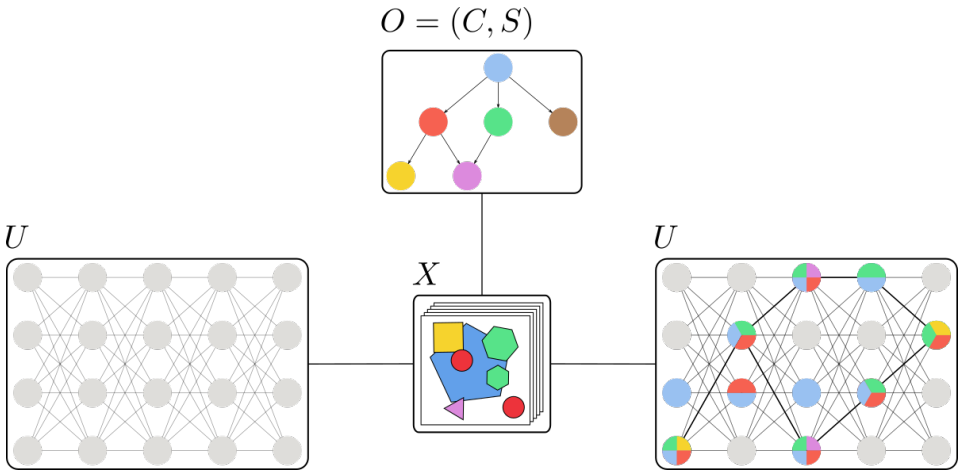


Figure 1: Overview of the proposed methodology. A set of neural units U is semantically aligned with an ontology O through a pixel-level annotated dataset X, whose labels are in a two-way relationship with the ontology concepts C. The relations S enable the retrieval of subgraphs composed of architecturally connected and semantically related units. Different colors stand for different visual concepts, while grey units account for semantically unaligned neurons.

work can propagate the partial alignment with the concepts of "cat" or "tiger" without the need for explicit "feline" annotations. Section 3.1 shows how we can align neural units with higher-level concepts via ontological information and without additional annotations. Secondly, we provide a probabilistic formulation of the interaction between visual concepts and neural activations. The model is used to define a novel measure of semantic alignment that correctly quantifies polysemantic neurons. Since being polysemantic is not an intrinsic characteristic of artificial neurons but the result of the concepts used to qualify them, this requirement is fundamental in our context (Section 3.2). The last contribution leverages the alignment of neural units with multiple concepts to propose a circuits identification algorithm. This algorithm produces meaningful subgraphs by exploiting semantic relations between aligned concepts. These subgraphs offer new insights into the content of distributed neural representations and provide a novel instrument for network inspection and interpretation (Section 3.3).

We validate our approach on the annotated dataset Broden against the state-of-the-art Network Dissection method (Bau et al., 2017) and by considering several renowned CNN architectures for image classification. As a side contribution of our empirical validation, we extend the original Broden dataset by associating its concepts with the WordNet ontology (Miller, 1995; Miller & Hristea, 2006). We publicly release this extension within the supplementary materials. While the main discussion focuses on the alignment with the Broden dataset, we also experimentally estimated the alignment with ImageNet (Deng et al., 2009), whose results we report in Appendix B. The empirical results show that our semantic alignment method yields to the emergence of meaningful neural circuits that are composed of units fundamental to predict semantically related visual categories.

2 Related Works

Zhou et al. (2015) 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 (Zhou et al., 2014). Their work manually annotated such detectors by visualizing manipulated examples that maximized units activations. The Circuits framework approached the problem similarly by employing feature visualization techniques (Olah et al., 2017) to manually assign specific roles to individual neurons, but further highlighting their contribution to the fulfillment of more complex tasks throughout the network (Olah et al., 2020). Bau et al. (2017) introduced Network Dissection to automatically analyze neural activations and identify meaningful neurons in CNNs trained on the Places-365 dataset (Zhou et al., 2017). Their work introduced a pixel-level annotated image dataset called Broden marking portrayed objects and patterns. Zhou et al. (2018) 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 & Andreas (2020) 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 (Page, 2000).

In contrast to our approach, other techniques aim to fulfill concept-based analysis of neural activations without restraining meaningful information to single neurons. Firstly, Fong & Vedaldi (2017) expanded Network Dissection with linear combinations of hidden neurons in CNNs to identify distributed concept detectors. Similarly, Kim et al. (2018) defined concept activation vectors (CAVs) as linear classifiers over the activations of an hidden layer, to identify the direction of arbitrary meaningful visual concepts and consequently estimate their importance for specific labels in a classification context. On top of CAVs, both Ghorbani et al. (2019) and Yeh et al. (2020) provided different techniques to automatically cluster visual concepts without human supervision, nonetheless practical applications require human intervention to label such concepts. While sharing the interest in concept-based analysis, our approach focuses on single hidden neurons to initially estimate their semantic role and eventually cluster them across different layers.

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 different overlapping representations (Otero-Cerdeira et al., 2015). 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 (Euzenat & Shvaiko, 2013a). In comparison, our approach exploits the portrayal of visual concepts to mediate their extensions and thus to estimate the difference between an explicit ontology and concepts implicitly expressed by individual neurons.

3 Ontology-Driven Semantic Alignment

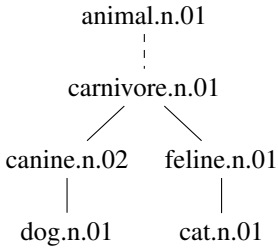
Given a pre-trained CNN architecture for computer vision, our framework assigns structured semantic roles to a subset U of its neural units. The assignment depends on the estimate of the semantic alignment between each artificial neuron and a set of visual concepts C. The alignment is estimated by analyzing neural activations over a pixel-level annotated image dataset X. In practice, for each example image $x \in X$ and concept $c \in C$, there exists a binary mask $L_c(x)$, known as the concept mask. A concept mask $L_c(x)$ has the same shape of the example image x and marks the locations portraying the visual concept c. Therefore, we are able to match the concept masks with feature maps $A_u(x)$ produced by the activation of each neural unit $u \in U$.

By iterating over the annotated dataset X and analyzing neural activations, our approach computes an estimate $\sigma(u, c) \in [0, 1]$ of the alignment for each unit-concept pair (u, c). Based on this $\sigma(u, c)$ estimate, the framework then produces, for each unit u, the subset of concepts $\psi_C(u) \subset C$ that are most significantly aligned with the neuron.

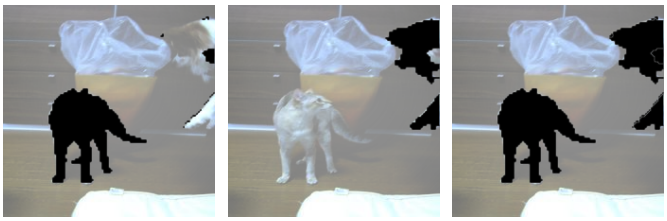
Since our proposal requires visual concepts to be structured in an ontology, we define the latter as an extensional relational structure $O = (C, S)$, where C is the set of all the possible concepts and S the set of their semantic relations (Guarino et al., 2009). Each semantic relation $s \in S$ is a truth-valued function $s\colon C \times C \rightarrow \{T, F\}$ that is true if and only if the relation between two concepts holds. The ontology is a key difference with Network Dissection which treats concepts as independent entities. In the following, we discuss the main characterizing aspects of masks generation (Section 3.1), alignment measurement (Section 3.2), and neural circuit extraction (Section 3.3).

3.1 High-level Concept Masks

As in Frege (1891), we consider each concept as an ideal function whose argument is an object of the world and whose value is a truth-value. Consequently, the extension E_c of a concept c, is the set of all the objects of the world satisfying it. For instance, the meaning of the term "dog" is the concept of dog, whose extension would be the set of all dogs. The specialization relation $\sqsubseteq$, also known as "is-a", is the semantic relation that expresses the inclusion between the extensions of concepts in an ontology (Euzenat & Shvaiko, 2013b). Formally, $c \sqsubseteq d \iff E_c \subseteq E_d$.



(a) Induced taxonomy



(b) cat.n.01 (c) dog.n.01 (d) animal.n.01

Figure 2: Example of mask generation for the higher-level concept of "animal" using ontological information. The induced taxonomy, built over the WordNet hypernymy (*is-a*) relation, enables the retrieval of the mask by exploiting the directly annotated masks for "dog" and "cat" from the Broden dataset.

Consequently, if a dataset X is correctly annotated, being $Q(x_p)$ the set of objects of the world W portrayed by an arbitrary location p of an image $x \in X$,

$$\begin{aligned} L_c(x)_p &\iff \exists o \in W.o \in Q(x_p) \wedge o \in E_c \\ &\implies \exists o \in W.o \in Q(x_p) \wedge o \in E_d \quad \{c \sqsubseteq d\} \\ &\iff L_d(x)_p. \end{aligned} \tag{1}$$

Specialization induces a hierarchical taxonomy represented by a Directed Acyclic Graph (DAG) where the node corresponding to the concept c is a child of the one corresponding to d if and only if $c \sqsubseteq d$. Hence, given a concept-annotated example image, the mask of any 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

The precise measurement of the alignment between concepts and network units is of fundamental importance for a method seeking to assign semantics to neural components. In principle, we aim to define a function $\sigma(u, c)$ such that a concept c and a unit u are perfectly aligned if and only if $\sigma(u, c) = 1$. We introduce a definition for such scoring function that relies on a probabilistic model of the influence that a set of visual concepts C exerts on the activations of a set of neural units U.

Given an image example $x \in \mathbb{R}^{W \times H}$, the output of an arbitrary convolutional unit u is a feature map $A_u(x) \in \mathbb{R}^{W_u \times H_u}$. We treat fully connected units as a special case whose feature maps have shape $(1, 1)$. Our approach focuses on higher than usual activations to highlight the most influential visual concepts. Therefore, it masks each feature map $A_u(x)$ with an arbitrarily high threshold t_u, whose determination is discussed in the experimental section. As in the Network Dissection approach, our solution scales the thresholded feature map $(A_u(x) > t_u) \in \mathbb{R}^{W_u \times H_u}$ into an activation mask $M_u(x) \in \mathbb{R}^{W \times H}$ to match the shape of the concept masks and of the example images. This operation approximates the relation between a neural unit and its receptive field, so that the thresholded activation $M_u(x)_{i,j}$ depends solely on the location (i, j) of the example image x. Consequently, we assume that $M_u(x)_{i,j}$ depends only on the portrayal of a visual concept c marked in the location $L_c(x)_{i,j}$. This approximation discards the effects of striding and padding over the receptive field of convolutional units (Araujo et al., 2019), which will be subject of future research.

Given this formulation, an annotated dataset X containing N examples of shape (W, H) produces WHN independent samples. In each sample, each visual concept $c \in C$ is modeled by a Bernoulli random variable Y_c, and each neuron $u \in U$ as a Bernoulli random variable Z_u. Our model builds on the assumption that such Z_u variables are conditionally independent given the concept random variables Y_c. Formally,

$$\forall u, u' \in U \times U . Z_u \perp\!\!\!\perp Z_{u'} \mid \{Y_c \mid c \in C\}. \tag{2}$$

Given the probabilistic formulation above, we can formulate semantic alignment in terms of the maximum likelihood estimate (MLE) of a concept being in the receptive field of a firing unit, that is

$$\mathcal{L}(Y_c = 1 \mid Z_u = 1) = \frac{\sum_x |L_c(x) \wedge M_u(x)|}{\sum_x |L_c(x)|}, \tag{3}$$

where $L_c(x) \wedge M_u(x)$ applies the logical *and* elementwise on the masks. Consequently, we define the MLE $\mathcal{L}(Y_c = 1 \mid Z_u = 1)$ as our semantic alignment measure $\sigma(u, c)$.

In an ideal scenario, each unit of the network would activate only when stimulated by a specific visual concept of the ontology. Assuming that a unit u responds solely to concept $\bar{c}$, the estimate would be

$$\sigma(u, c) = 1 \iff c = \bar{c}, \tag{4}$$

for any arbitrary concept $c \in C$. Similarly, this estimate correctly handles polysemantic neurons which ideally activate solely for a set of concepts $\bar{C} \subseteq C$. For any concept $c \in C$, the estimate straighforwardly responds

$$\sigma(u, c) = 1 \iff c \in \bar{C}. \tag{5}$$

In comparison, the IoU measure introduced by the Network Dissection approach,

$$\text{IoU}(u, c) = \frac{\sum_{x \in X} |M_u(x) \wedge L_c(x)|}{\sum_{x \in X} |M_u(x) \vee L_c(x)|}, \tag{6}$$

does not yield to a valid alignment estimate σ outside of the ideal case in which a unit responds only to a single concept. Assuming that a unit u responds to a subset of concept $\bar{C} \subset C$, it can be shown that there exists scenarios where

$$\text{IoU}(u, c) < 1 \tag{7}$$

even if $c \in \bar{C}$. Furthermore, the IoU measurement will favour visual concepts which appear more often in the annotated dataset X. Thus, it binds the estimate of the alignment of visual concepts to their popularity, failing to properly characterize polysemantic units. We report further motivations about the ideal behavior of the two measures in Appendix A.

3.3 Neural Circuits

The σ alignment function allows us to define, for each unit u, a subset $\psi_C(u) \subseteq C$ of the concepts influencing the most its activations. Having fixed an arbitrary threshold τ, standing for the tolerance towards unaligned concepts, the subset $\psi_C(u)$ can be approximated by filtering out concepts according to their alignment estimate $\sigma(u, c)$. Formally,

$$\psi_C(u) = \{c \mid \sigma(u, c) > \tau\}. \tag{8}$$

Given the alignment estimates for the unit set U, our approach exploits the semantic relations contained within an ontology to identify meaningful connected subgraphs. Circuits are systematically retrievable by identifiying non-trivial connected components in a graph composed of unit-meaning pairs (u, c) such that $c \in \psi_C(u)$: Algorithm 1 provides a procedural description of the method.

Finally, we define the *coherence* of a circuit T as the expected similarity between two random concepts. As in

$$\text{Coherence}(T) = \frac{\sum_{(c_1, c_2) \in P(T)} \delta(c_1, c_2),}{|P(T)|} \tag{9}$$

where $P(T)$ is the set of possible concept pairs in T, and δ is an arbitrary semantic similarity function between two concepts, such as the Jiang-Conrath (Jiang & Conrath, 1997) or the Lin (Lin, 1998) similarities (Appendix C). Our coherence formulation measures the semantic diversity within a circuit, highlighting the presence of distant concepts. It is worth noticing that the same concept might be aligned to many units and thus emerge multiple times in the same circuit. Consequently, it is valuable to estimate coherence with or without these repetitions. In the following, we refer to the measurement without repetitions as "unique coherence".

Algorithm 1: Algorithm to identify circuits from a set of semantically aligned units U within a neural network composed of n consecutive layers. The algorithm connects units that are aligned with semantically related concepts according to arbitrary relations $s \in S$ from the ontology. Each non-trivial connected component of the generated graph constitutes a circuit.

Function `retrieve_circuits`(U, n, C, S, ψ_C)**:**
- G = new empty graph
- $G.V = \{(u, c) \in U \times C \colon c \in \psi_C(u)\}$
- $G.E = \{\}$
- **for** $i \in \{1, \dots, n-1\}$ **do**
  - **for** $u_1 \in \{u \mid u \in U \wedge u \textit{ within } i\textit{-th layer}\}$ **do**
    - **for** $u_2 \in \{u \mid u \in U \wedge u \textit{ within } (i+1)\textit{-th layer}\}$ **do**
      - **if** $\exists c_1 \in \psi_C(u_1).\exists c_2 \in \psi_C(u_2).\exists s \in S.s(c_1, c_2)$ **then**
        - $G.E = G.E \cup ((u_1, c_1), (u_2, c_2))$
- **return** $\{T \in \text{ConnectedComponents}(G) \colon |T| > 1\}$

4 Results

The implementation of our framework is publicly available, as it is a Jupyter Notebook to replicate the whole analysis reported in this section[1]. The experimental analysis focuses on the semantic alignment of visual concepts representing concrete objects. To obtain an ontologically annotated pixel-level dataset, we associated each object label of the Broden dataset to a member of the WordNet ontology. The assignment took into consideration the explicit description of each label and a sample of corresponding annotated images. The 672 object labels within Broden produced 1177 unique concepts, of which 513 are leaves in the induced taxonomy. As speculated previously in the paper, introducing a specialization relation increased the number of distinct visual concepts. We publicly release this extension of the Broden dataset along with the code. Because of the direct comparison with the literature, the current section focuses on Broden. Nonetheless, we also extensively studied semantic alignment with the ImageNet dataset (Deng et al., 2009). The dataset provides bounding boxes selecting the portrayed objects that we exploited to construct approximated concept masks. We report and discuss the results of the semantic alignment with ImageNet in Appendix B.

To obtain activation masks, we set for each unit u a threshold t_u such that $P(A_u(x) > t_u) > 0.005$, by estimating the probability over the activations on the Broden dataset. In doing so, we follow previous works on Network Dissection to enhance comparability (Bau et al., 2017). For the same reason, when using the IoU measure, we consider a concept-unit pair to be aligned if the measure overcomes $\tau_{\text{IoU}} = 0.04$. Instead, for our probabilistic measure, we fixed the threshold as $\tau_{\text{MLE}} = 0.2$. The threshold should account for erroneous and noisy annotations in the dataset, by effectively cutting off visual concepts that are thus mistakenly aligned. Given a random selection of artificial units from different architectures, we derived τ_{MLE} in a pre-experimental phase by comparing istances of $\psi_C(u)$ and samples of images activating each unit u.

We report the analysis of the last layers of three popular CNN architectures for image classification. Firstly, we semantically aligned the last three fully connected layers and the last two convolutional layers of AlexNet (Krizhevsky, 2014). Then, we considered the last fully connected layer and the last two residual blocks of ResNet (He et al., 2015). In each residual block, we independently analyzed the two convolutional operations and the sum after the residual connection. Finally, we aligned the last fully connected layer and the output of the last three dense blocks in DenseNet (Huang et al., 2017). All networks were pre-trained to classify the 365 different scenes and views from the Places-365 dataset (Zhou et al., 2017). For replicability purposes, we adopted publicly available pre-trained models mentioned in the supplementary materials.

For the sake of compactness, Table 1 reports aggregated statistics, but we provide per-layer details in Appendix B. Firstly, we adopted the IoU measure without ontologically propagating concept masks, essentially replicating the original Network Dissection approach. Since most visual concepts in the

[1]Name and Git repository omitted to preserve anonymity. Anonymized code/data are in the supplementary materials.

	Model	AlexNet			ResNet-18			DenseNet-161		
	Units	9069			3437			509		
	σ-metric	IoU(u,c)		$\mathcal{L}(c \mid u)$	IoU(u,c)		$\mathcal{L}(c \mid u)$	IoU(u,c)		$\mathcal{L}(c \mid u)$
	τ	0.04		0.2	0.04		0.2	0.04		0.2
	Propagation	✗	✓	✓	✗	✓	✓	✗	✓	✓
Concepts	Unique	40	83	367	71	138	412	64	124	433
	Leaves	31	30	302	48	48	311	46	46	320
	Non-leaves	9	53	65	23	90	101	18	78	113
	Total	896	1892	12560	849	1967	3756	346	790	1998
$\psi_C(u)$	Size	0.10	0.21	1.38	0.25	0.57	1.09	0.68	1.55	3.93
	Non-empty size	1.14	2.33	2.25	1.37	3.01	3.00	1.48	3.39	4.79
	Depth	7.33	6.89	7.77	7.41	6.95	7.65	7.53	6.98	7.69
	$\sigma(u,c)$	0.057	0.056	0.311	0.065	0.066	0.324	0.072	0.070	0.372

Table 1: Semantic alignment of convolutional architectures. We report the number of distinct aligned concepts and we classify them according to their depth in the reference taxonomy (i.e. leaves or not). We also report the total number of aligned concepts considering repetitions between units. The set $\psi_C(u)$ contains the concepts aligned to an arbitrary unit u. We report the average size of this set, the average size when non-empty, the average depth of its members in the induced taxonomy, and the average alignment estimate σ. The IoU measurement without the automatic propagation of concept masks corresponds to the Network Dissection approach.

(a) Examples with maximal activations

Concept	$\sigma(u,c)$
hovel.n.01	0.021
roof.n.03	0.025
building.n.01	0.031
shelter.n.01	0.035
house.n.01	0.098

(b) IoU(u,c)

Concept	$\sigma(u,c)$
circus_tent.n.01	0.401
greenhouse.n.01	0.403
shed.n.01	0.469
pavilion.n.01	0.568
bandstand.n.01	0.631

(c) $\mathcal{L}(c \mid u)$

Figure 3: Semantic alignment of unit 196 in the last residual block of ResNet-18. The examples reported in (a) are the top ten images producing the maximum activations for the unit. Top-5 aligned concepts in (b) and (c) are presented in increasing order, respectively for IoU and our probabilistic measure. While IoU captures concepts that are undoubtedly representative of the unit, our score identifies visual concepts that describe the examples more precisely. As in the quantitative results, considering the average distance from the root of the retrieved concepts, the top-5 results using IoU, with depth 6.4, are more general than those obtained by our measure, with depth 8.0.

Broden dataset correspond to leaves of the WordNet ontology, units are aligned with a small number of general concepts. In comparison, our proposal aligned more visual concepts throughout all the target networks. Consequently, we retained the IoU measure and introduced ontological information to automatically propagate concept masks. Higher-level concepts introduced by the specialization relation consistently increase the number of aligned concepts, thus the expressivity of the results. Nonetheless, while increasing this number, the results highlight that in this scenario IoU mostly selects concepts nearer to root of the ontology, hence more general. This is due to the popularity of high level concepts, that accounts for the popularity of all their specializations. Differently from IoU, concept popularity does not influence our probabilistic measure $\mathcal{L}(c \mid u)$, that is therefore more apt to semantically align concepts at different levels of the ontology. Figure 3 exemplifies the different measures on a unit of ResNet-18. Overall, the results also support our intuition concerning the polysemy of neural units, since if a unit is aligned to a concept, then it is on average aligned to more than one. Finally, since the number of retrieved concepts depends on the thresholds τ_{IoU} and τ_{MLE}, we discuss their lowering in Appendix A.1.

Given the semantic alignment of the target networks, we identify circuits by exploiting different semantic relations contained in WordNet: hypernymy (*is-a*), meronymy (*part-of*) and strong semantic similarity. We consider two concepts to be sufficiently similar if their Jiang-Conrath similarity (Jiang & Conrath, 1997) overcomes a fixed threshold set to 0.4. The threshold derives from a pre-experimental phase where we empirically estimated it as a solid lower bound to construct meaningful circuits. We reserve to deepen the analysis of the effect of this parameter on the overall approach

Model	Monosemantic	Circuits	Pairs	Units	Concepts	Coh.	Unique Coh.
AlexNet	✓	175	71.308	64.634	1.874	0.949	0.535
	✗	42	175.452	147.645	4.643	0.786	0.535
ResNet-18	✓	116	29.552	20.474	2.890	0.893	0.551
	✗	43	60.605	36.116	6.093	0.710	0.551
DenseNet-161	✓	38	14.421	10.210	2.631	0.863	0.591
	✗	19	23.737	15.316	4.263	0.727	0.591

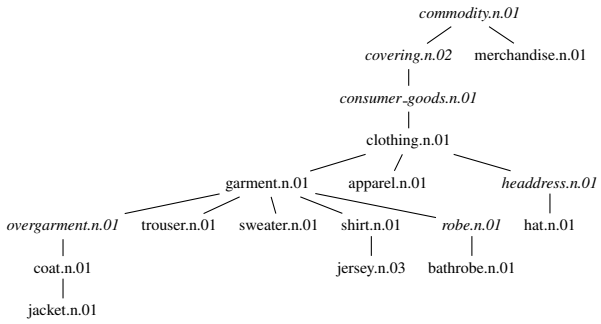
Table 2: Comparison of the results of the circuits retrieval algorithm on the three target models. We report the average number of unit-concept pairs, the number of unique units, and the number of unique concepts. Circuit coherence is presented with and without considering repetitions of the concepts. Many circuits are composed by units aligned only with the same concept. For this reason we report separately the same measures for non-monosemantic circuits, i.e. circuits containing at least two distinct concepts.

in future research. Anyhow, to highlight the independence from this particular configuration, we include in Appendix B the results of the circuits retrieval procedure with a different similarity measure. Given a circuit, its coherence is assessed using the Lin similarity (Lin, 1998). We adopt different similarity measures for neural circuits construction and for their assessment, to avoid to predetermine the coherence of a circuit. Furthermore, since Lin similarity has values on $[0, 1]$, it produces more intuitive comparisons between different neural circuits. Both similarity measures are formally defined in Appendix C. Table 2 reports an overview of the circuits identified through the target networks. The overall number of identified circuits depends on the total number of aligned concepts, thus resulting in considerable differences between the target networks. Furthermore, many circuits whose units are aligned only with the same concept emerge. For each target network, Appendix B reports a list of all the circuits aligned to at least two distinct concepts.

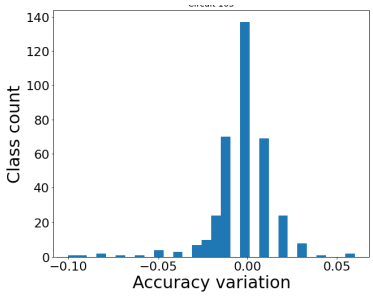
Finally, we are interested in the role of the identified neural circuits for the predictions generated by the whole network. Similarly to Zhou et al. (2018), we assess the importance of neural units by measuring the accuracy drop for specific classes in the predictive task learned by the network. More in detail, we measure the drop on the Top-5 classification accuracy of the 365 distinct classes from the Places-365 dataset. Notably, we do not ablate single units corresponding to a single concept, but an overall neural circuit. We exclude from the ablation units within the last layer of the target network, since we are interested in the characterization of hidden units within a circuit. Figure 4 shows an example of how a large circuit of highly coherent concepts is important for the correct prediction of semantically related classes. In this type of circuits, there usually exists a dominating concept that is aligned to most units. Nonetheless, by ablating only the units aligned to the most popular concept, the accuracy drop is substantially different compared to the effects of ablating the whole circuit. Therefore, the number of unique units and the coherence of a circuit are crucial in influencing the predictive accuracy of related classes. Small circuits or semantically sparse units do not result in relevant semantically related accuracy drops. For this reason, the number of circuits producing significant accuracy drop is consistently inferior in ResNet and especially in DenseNet, as a consequence of fewer identified circuits. We report additional circuit analysis in Appendix B.

5 Conclusion

We introduced what we believe to be the first framework for the semantic alignment of neural units with a complete visual ontology. Our solution builds on previous works that semantically aligned visual concepts independently without considering semantic relations between them. The introduction of semantic relations led us to three key innovative contributions. Firstly, we defined a propagation strategy to align units with concepts that lack an explicit annotation in the alignment dataset. Secondly, we defined a novel semantic alignment measure acknowledging polysemy in neural units. Finally, we introduced an algorithm to identify connected neural circuits composed of units aligned to semantically related concepts. We experimentally validated our approach by studying the semantic alignment of 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



(a)



(b)

Class	Drop
/c/closet	-0.05
/f/fabric_store	-0.05
/b/banquet_hall	-0.05
/d/dressing_room	-0.05
/d/dorm_room	-0.06
/h/hardware_store	-0.07
/c/clothing_store	-0.08
/d/department_store	-0.08
/d/drugstore	-0.09
/g/gift_shop	-0.1

(c)

Class	Drop
/b/butchers_shop	-0.03
/b/booth/indoor	-0.03
/s/slum	-0.03
/b/banquet_hall	-0.04
/d/dorm_room	-0.04
/d/department_store	-0.04
/c/closet	-0.04
/f/fabric_store	-0.06
/c/clothing_store	-0.06
/g/gift_shop	-0.09

(d)

Figure 4: Importance analysis of circuit n. 105 from AlexNet, containing 105 distinct units and 12 unique visual concepts. The circuit consists of units aligned to various concepts relative to clothing and commodities: subfigure (a) reports the aligned concepts within the WordNet taxonomy. *Italic* concepts are not directly aligned, but we include them for visualization purposes. When ablating the circuit, the accuracy drop significantly affects only a small number of classes. Subfigure (b) depicts the histogram of categories of the Places-365 dataset as a function of accuracy drop (on the x-axis). The classes most affected by the ablation are those more strictly related to the semantically aligned concepts of the circuit, as exemplified in (c). Dropping only the units aligned with the most popular concept in the circuit (d) does not account for the effects of ablating the whole circuit (c).

reproduce our experiments. The experiments highlighted how our methodology could effectively capture semantic alignment in neural units and consistently handle concepts inscribed within an ontology. Furthermore, we assessed the emergence of semantically related neural circuits and studied their role in the overall network. We found that units within sufficiently large and coherent neural circuits are pivotal to classify categories related to the aligned concepts.

This last aspect constitutes the most valuable contribution of our semantic alignment methodology. Semantically coherent neural circuits could be exploited for innovative interpretative approaches, for instance by producing explanations at different levels of detail according to circuit members and user knowledge of the context. We reserve to explore practical interpretative applications of our approach in future research. Similarly, we want to mention the main limitations affecting our proposal. Firstly, as already discussed for Network Dissection (Fong & Vedaldi, 2017), a unit might express other visual concepts in activation ranges other than the maximal. Furthermore, the idea that single neurons express human-like concepts correctly describes some units, but overall the vast majority of neurons are not immediately alignable. Nonetheless, concepts, in the sense of functions from instances to truth-values, might be a useful tool to describe artificial neurons even if not directly associated to human-like ones. This might also enable the use of tools for the automatic retrieval of visual concepts, to produce annotated datasets and consequently identify the patterns underlying unit activations without further supervision. To conclude, ontology-based neural circuits offer an innovative instrument to inquire about the nature of neural representations, highlighting semantically related human-interpretable features across the network.

Reproducibility Statement

We actively considered reproducibility issues during both research and paper preparation. For this reason, the supplementary materials contain or point out all the necessary artifacts to replicate the reported results. More in detail, we adopted publicly available datasets such as Broden and ImageNet. We release our original extension of the Broden dataset and include the code needed to preprocess images from a selection of ImageNet. Similarly, we analyzed publicly available pre-trained neural models from the Places-365 project. We implemented our methodological proposal in a Python library, publicly releasing it as free and open-source code. More importantly, supplementary materials include a Jupyter Notebook that documents step by step how to download publicly available resources, how to integrate our contributions, and how to run the experiments reported in the paper. Finally, the reader can compare its results with Appendix B, which summarizes the analysis of each model and dataset combination.

References

André Araujo, Wade Norris, and Jack Sim. Computing receptive fields of convolutional neural networks. *Distill*, 2019. doi: 10.23915/distill.00021. URL https://distill.pub/2019/computing-receptive-fields.

David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

David Bau, Jun-Yan Zhu, Hendrik Strobelt, Bolei Zhou, Joshua B. Tenenbaum, William T. Freeman, and Antonio Torralba. Gan dissection: Visualizing and understanding generative adversarial networks. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2019.

J. Deng, W. Dong, R. Socher, L. Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

Jérôme Euzenat and Pavel Shvaiko. Classifications of Ontology Matching Techniques. In *Ontology Matching*, pp. 73–84. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013a. ISBN 9783642387203 9783642387210. doi: 10.1007/978-3-642-38721-0_4. URL http://link.springer.com/10.1007/978-3-642-38721-0_4.

Jérôme Euzenat and Pavel Shvaiko. The Matching Problem. In *Ontology Matching*, pp. 25–54. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013b. ISBN 9783642387203 9783642387210. doi: 10.1007/978-3-642-38721-0_2. URL http://link.springer.com/10.1007/978-3-642-38721-0_2.

R. C. Fong and A. Vedaldi. Interpretable Explanations of Black Boxes by Meaningful Perturbation. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 3449–3457, October 2017. doi: 10.1109/ICCV.2017.371. ISSN: 2380-7504.

Gottlob Frege. *Function und Begriff*. Hermann Pohle, Jena, 1891.

Amirata Ghorbani, James Wexler, James Zou, and Been Kim. Towards Automatic Concept-based Explanations. *arXiv:1902.03129 [cs, stat]*, October 2019. URL http://arxiv.org/abs/1902.03129. arXiv: 1902.03129.

Gabriel Goh, Nick Cammarata †, Chelsea Voss †, Shan Carter, Michael Petrov, Ludwig Schubert, Alec Radford, and Chris Olah. Multimodal neurons in artificial neural networks. *Distill*, 2021. doi: 10.23915/distill.00030. URL https://distill.pub/2021/multimodal-neurons.

Nicola Guarino, Daniel Oberle, and Steffen Staab. *What Is an Ontology?*, pp. 1–17. Springer Berlin Heidelberg, Berlin, Heidelberg, 2009. ISBN 978-3-540-92673-3. doi: 10.1007/978-3-540-92673-3_0. URL https://doi.org/10.1007/978-3-540-92673-3_0.

Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5):1–42, 2018.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015. URL http://arxiv.org/abs/1512.03385.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708, 2017.

Jay J. Jiang and David W. Conrath. Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings of the 10th Research on Computational Linguistics International Conference*, pp. 19–33, Taipei, Taiwan, August 1997. The Association for Computational Linguistics and Chinese Language Processing (ACLCLP). URL https://aclanthology.org/O97-1002.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory Sayres. Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV). *arXiv:1711.11279 [stat]*, June 2018. URL http://arxiv.org/abs/1711.11279. arXiv: 1711.11279.

Alex Krizhevsky. One weird trick for parallelizing convolutional neural networks. *CoRR*, abs/1404.5997, 2014. URL http://arxiv.org/abs/1404.5997.

Dekang Lin. An information-theoretic definition of similarity. In *Proceedings of the Fifteenth International Conference on Machine Learning*, ICML '98, pp. 296–304, San Francisco, CA, USA, 1998. Morgan Kaufmann Publishers Inc. ISBN 1558605568.

Lingling Meng, Runqing Huang, and Junzhong Gu. A review of semantic similarity measures in wordnet. *International Journal of Hybrid Information Technology*, 6(1):1–12, 2013.

George A. Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11): 39–41, 1995.

George A. Miller and Florentina Hristea. Wordnet nouns: Classes and instances. *Comput. Linguist.*, 32(1):1–3, March 2006. ISSN 0891-2017. doi: 10.1162/coli.2006.32.1.1. URL https://doi.org/10.1162/coli.2006.32.1.1.

Jesse Mu and Jacob Andreas. Compositional explanations of neurons. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 17153–17163. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/c74956ffb38ba48ed6ce977af6727275-Paper.pdf.

Chris Olah, Alexander Mordvintsev, and Ludwig Schubert. Feature visualization. *Distill*, 2017. doi: 10.23915/distill.00007. URL https://distill.pub/2017/feature-visualization.

Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. Zoom in: An introduction to circuits. *Distill*, 2020. doi: 10.23915/distill.00024.001. URL https://distill.pub/2020/circuits/zoom-in.

Lorena Otero-Cerdeira, Francisco J Rodríguez-Martínez, and Alma Gómez-Rodríguez. Ontology matching: A literature review. *Expert Systems with Applications*, 42(2):949–971, 2015.

Mike Page. Connectionist modelling in psychology: A localist manifesto. *Behavioral and Brain Sciences*, 23(4):443–467, 2000. doi: 10.1017/S0140525X00003356.

Philip Resnik. Using information content to evaluate semantic similarity in a taxonomy. *arXiv preprint cmp-lg/9511007*, 1995.

Chih-Kuan Yeh, Been Kim, Sercan O. Arik, Chun-Liang Li, Tomas Pfister, and Pradeep Ravikumar. On Completeness-aware Concept-Based Explanations in Deep Neural Networks. *arXiv:1910.07969 [cs, stat]*, June 2020. URL http://arxiv.org/abs/1910.07969. arXiv: 1910.07969.

Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. Learning deep features for scene recognition using places database. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger (eds.), *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. URL https://proceedings.neurips.cc/paper/2014/file/3fe94a002317b5f9259f82690aeea4cd-Paper.pdf.

Bolei Zhou, Aditya Khosla, Àgata Lapedriza, Aude Oliva, and Antonio Torralba. Object detectors emerge in deep scene cnns. In Yoshua Bengio and Yann LeCun (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1412.6856.

Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

Bolei Zhou, Yiyou Sun, David Bau, and Antonio Torralba. Revisiting the importance of individual units in cnns via ablation. *CoRR*, abs/1806.02891, 2018. URL http://arxiv.org/abs/1806.02891.

A Alignment Measure

Given an image dataset X portraying a set of a visual concepts C, we explicitly consider the behavior of the probablistic measure introduced in Section 3.2

$$\mathcal{L}(Y_c = 1 \mid Z_u = 1) = \frac{\sum_x |L_c(x) \wedge M_u(x)|}{\sum_x |L_c(x)|} \tag{A.1}$$

and the IoU measure for semantic alignment adopted by Network Dissection (Bau et al., 2017)

$$\text{IoU}(u, c) = \frac{\sum_{x \in X} |M_u(x) \wedge L_c(x)|}{\sum_{x \in X} |M_u(x) \vee L_c(x)|}. \tag{A.2}$$

For brevity, we will refer to the former simply as $\mathcal{L}(c \mid u)$.

Furthermore, we are interested in the response of the measures to a unit u which activates solely for a non-empty subset $\psi_C(u) \subset C$ of visual concepts. Formally, assuming that the dataset X correctly annotates each visual concept,

$$c \in \psi_C(u) \iff \forall x. M_u(x) \wedge L_c(x) = L_c(x). \tag{A.3}$$

Given this formalization, $\psi_C(u)$ is closed under specialization. If $c \in \psi_C(u)$ then all the descendants of c in the specialization-induced taxonomy are also members because of the inclusion between their concept masks (Section 3.1).

For what concerns the probabilistic measure $\mathcal{L}(c \mid u)$,

$$\begin{aligned} c \in \psi_C(u) &\iff \forall x. M_u(x) \wedge L_c(x) = L_c(x) \\ &\iff \sum_x |M_u(x) \wedge L_c(x)| = \sum_x |L_c(x)| \\ &\iff \frac{\sum_x |M_u(x) \wedge L_c(x)|}{\sum_x |L_c(x)|} = 1 \\ &\iff \mathcal{L}(c \mid u) = 1. \end{aligned} \tag{A.4}$$

Instead, the $\text{IoU}(u, c)$ measure is proportional to the popularity of a visual concept c within the dataset X. In fact, if $c \in \psi_C(u)$,

$$\begin{aligned} \text{IoU}(u, c) &= \frac{\sum_{x \in X} |M_u(x) \wedge L_c(x)|}{\sum_{x \in X} |M_u(x) \vee L_c(x)|} \\ &= \frac{\sum_{x \in X} |L_c(x)|}{\sum_{x \in X} |M_u(x)|} \\ &\propto \sum_{x \in X} |L_c(x)| \end{aligned} \tag{A.5}$$

because of A.3. Furthermore, we proof that $c \in \psi_C(u) \not\Longrightarrow \text{IoU}(u, c) = 1$ by constructing a simple counter example. Suppose that a unit u responds to two visual concepts $c_1, c_2 \in \psi_C(u)$, and that there exists at least one example x portraying them differently, i.e. $L_{c_1}(x) \neq L_{c_2}(x)$. Consequently, at least one of them has a popularity $\sum_x |L_c(x)|$ inferior to the extent of the activation map $\sum_x |M_u(x)|$, therefore

$$\exists c \in \psi_C(u). \text{IoU}(u, c) < 1. \tag{A.6}$$

A.1 Alignment Threshold

In the current section, we report a practical example concerning the dependence of IoU on concept popularity and consequently on concept generality. In fact, because of the automatic propagation of concept masks, general concepts are usually significantly more popular then leaves, since they account for all their descendants in the induced taxonomy. As a remainder, Appendix B reports quantitative results on this same aspect, highlighting how IoU selects concepts on average nearer to the root of the ontology.

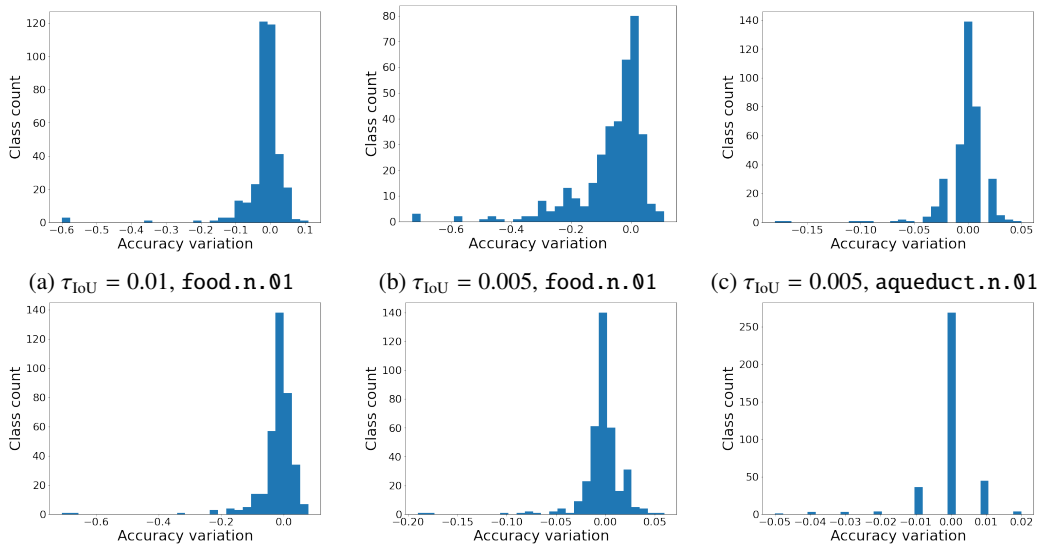
τ_{IoU}	aqueduct.n.01	food.n.01
0.04	0	3
0.02	0	21
0.01	1	165
0.005	21	602
0.002	223	1888
0	1498	7181

(a) IoU(u, c)

τ_{MLE}	aqueduct.n.01	food.n.01
0.2	56	0
0.1	317	16
0.05	797	110
0	1498	7181

(b) $\mathcal{L}(c \mid u)$

Table A.1: Each subtable reports for each measure, as the alignment threshold decreases, the number of units from the last four hidden layers of AlexNet aligned with an highly specific concept, namely `aqueduct.n.01` which has depth equal to eight, and a more general one, namely `food.n.01` with depth four.



(a) $\tau_{\text{IoU}} = 0.01$, `food.n.01` (b) $\tau_{\text{IoU}} = 0.005$, `food.n.01` (c) $\tau_{\text{IoU}} = 0.005$, `aqueduct.n.01`

(d) $\tau_{\text{MLE}} = 0.1$, `aqueduct.n.01` (e) $\tau_{\text{MLE}} = 0.01$, `aqueduct.n.01` (f) $\tau_{\text{MLE}} = 0.01$, `food.n.01`

Figure A.1: Accuracy drop histograms due to the ablation of units aligned to a given concept according to the IoU (a, b, c) and $\mathcal{L}(c \mid u)$ (d, e, f) measures. For what concerns IoU, τ_{IoU} is lowered to select sufficient units aligned to the specific concept of aqueduct. For $\mathcal{L}(c \mid u)$ instead, the lowering of τ_{MLE} enables the alignment of units to the general concept of food. The x-axis groups according to the accuracy drop extent, while the y-axis counts the number of affected classes in the given range.

Thresholds τ_{IoU} and τ_{MLE} (Section 4) fix the lower bound according to which IoU and $\mathcal{L}(c \mid u)$ consider a concept and a unit to be semantically aligned. While lowering these thresholds might increase the sensitivity of the approach, it would also acknowledge spurious concepts as semantically aligned to unrelated units. Table A.1 reports how the the number of aligned units varies with the alignment threshold in AlexNet. To identify the impact of such spurious assignments, we consider the accuracy drop effect in a classification task as a valuable indicator. Ideally, if units and concepts are perfectly aligned, the ablation of semantically coherent units should affect only a precise selection of related classes. Otherwise, we expect the accuracy drop to be more sparse across labels. In practice, we measure the accuracy drop on the Top-5 classification task on Places-365 (Zhou et al., 2017) as in the main discussion. As visualized in Figure A.1, reducing the τ_{IoU} threshold, enables the retrieval of more specialized concepts. Anyhow, this also results in a significant increase of the units aligned to more general concepts. Consequently, this results in a less ideal distribution of the accuracy drop for general concepts. Similarly, reducing τ_{MLE} to acknowledge more general concepts in $\mathcal{L}(c \mid u)$

produce an increase of the units aligned to more specific concepts. Nonetheless, the increase does not justify a significant effect on accuracy plots.

B Additional Results

The current section reports the results of the semantic alignment of the target networks with the WordNet ontology given the Broden and the ImageNet datasets. The characteristics of the Broden dataset are discussed in Section 4. Concerning ImageNet, we adopted the validation split of the ILSVRC 2011 competition. The dataset annotates the bounding boxes of 1000 different visual concepts over 50000 disting images. The induced taxonomy within WordNet corresponds to a directed acyclic graph containing 1905 distinct concepts. We directly exploited the bounding boxes to produce an approximation of the concept masks needed by our approach. Compared to Broden, the alignment resulted in a similar number of aligned concepts and consequently of retrieved circuits. Similarly, highly coherent circuits with many units result in a semantically clear accuracy drop. Nonetheless, concept masks approximation mainly affects semantic alignment in layers distant from the last. In these layers, the number of aligned concepts is significantly low when compared to the alignment obtained by pixel-level concept masks from Broden.

The following tables include additional statistics for each layer within the target networks mentioned in Section 4. More in detail, we report the number of distinct aligned concepts and we classify them according to their depth in the reference taxonomy (i.e. leaves or not). We also report the total number of aligned concepts considering repetitions between units. The set $\psi_C(u)$ contains the concepts aligned to an arbitrary unit u. We report the average size of this set, the average size when non-empty, the average depth of its members in the induced taxonomy and the average alignment estimate σ.

For each non-monosemantic circuit, i.e. aligned to at least two distinct concepts, we report the number of valid (u, c) pairs, the number of unique units and concepts and the coherence with and without considering repetitions. Furthermore, we characterize each circuit by mentioning their nearest common ancestor and the most popular concept within the aligned units. To highlight the independence from a particular similarity measure, in Table B.11 we report the results of the circuit analysis procedure on AlexNet using the Resnik similarity (Resnik, 1995) instead of the Jiang-Conrath similarity (Jiang & Conrath, 1997). Furthermore, for a selection of circuits, we report a dedicated analysis visualized as in Figure 4 from Section 4.

B.1 Broden Semantic Alignment

	Module	features.8	features.10	classifier.1	classifier.4	classifier.6	alexnet
	Units	256	256	4096	4096	365	9069
Concepts	Unique	9	36	39	73	56	83
	Leaves	4	9	13	28	22	30
	Non-leaves	5	27	26	45	34	53
	Total	9	58	263	1152	410	1892
$\psi_C(u)$	Size	3.52e-02	2.27e-01	6.42e-02	2.81e-01	1.12	2.09e-01
	Non-empty	1.50	2.64	2.01	2.32	2.59	2.33
	Depth	7.56	6.78	6.95	6.84	7.07	6.89
	$\sigma(u, c)$	4.57e-02	4.94e-02	5.19e-02	5.36e-02	6.56e-02	5.58e-02

Table B.1: Semantic alignment of AlexNet with the Broden dataset. Fixed threshold $\tau = 0.04$ over the IoU metric.

	Module	features.8	features.10	classifier.1	classifier.4	classifier.6	alexnet
	Units	256	256	4096	4096	365	9069
Concepts	Unique	8	32	265	328	255	367
	Leaves	7	28	233	277	205	302
	Non-leaves	1	4	32	51	50	65
	Total	13	50	4836	6660	1001	12560
$\psi_C(u)$	Size	5.08e-02	1.95e-01	1.18	1.63	2.74	1.38
	Non-empty	1.30	1.43	1.98	2.38	3.45	2.25
	Depth	8.12	7.75	7.86	7.76	7.64	7.77
	$\sigma(u, c)$	3.04e-01	2.70e-01	3.09e-01	3.11e-01	3.29e-01	3.11e-01

Table B.2: Semantic alignment of AlexNet with the Broden dataset. Fixed threshold $\tau = 0.2$ over the likelihood $\mathcal{L}(c \mid u)$ metric.

	Module	layer4.0.conv1	layer4.0.conv2	layer4.0	layer4.1.conv1	layer4.1.conv2	layer4.1	fc	resnet18
	Units	512	512	512	512	512	512	365	3437
Concepts	Unique	35	76	93	62	78	86	106	138
	Leaves	11	25	32	24	26	31	38	48
	Non-leaves	24	51	61	38	52	55	68	90
	Total	54	153	182	137	376	408	657	1967
$\psi_C(u)$	Size	1.05e-01	2.99e-01	3.55e-01	2.68e-01	7.34e-01	7.97e-01	1.80	5.72e-01
	Non-empty	2.35	2.64	2.64	2.85	3.16	3.14	3.17	3.01
	Depth	7.11	7.26	7.14	7.55	7.49	7.35	6.87	6.95
	$\sigma(u, c)$	5.73e-02	6.05e-02	6.42e-02	6.29e-02	6.50e-02	6.64e-02	6.80e-02	6.55e-02

Table B.3: Semantic alignment of ResNet with the Broden dataset. Fixed threshold $\tau = 0.04$ over the IoU metric.

	Module	layer4.0.conv1	layer4.0.conv2	layer4.0	layer4.1.conv1	layer4.1.conv2	layer4.1	fc	resnet18
	Units	512	512	512	512	512	512	365	3437
Concepts	Unique	31	88	128	119	225	237	368	412
	Leaves	27	72	99	98	180	188	281	311
	Non-leaves	4	16	29	21	45	49	87	101
	Total	37	154	200	213	793	820	1539	3756
$\psi_C(u)$	Size	7.23e-02	3.01e-01	3.91e-01	4.16e-01	1.55	1.60	4.22	1.09
	Non-empty	1.61	1.66	1.94	1.82	2.72	2.69	4.85	3.00
	Depth	7.77	7.57	7.46	7.71	7.75	7.79	7.65	7.65
	$\sigma(u, c)$	2.72e-01	2.75e-01	2.80e-01	2.84e-01	3.03e-01	3.04e-01	3.63e-01	3.24e-01

Table B.4: Semantic alignment of ResNet with the Broden dataset. Fixed threshold $\tau = 0.2$ over the likelihood $\mathcal{L}(c \mid u)$ metric.

	Module	DenseBlock4.22	DenseBlock4.23	DenseBlock4.24	classifier	densenet161
	Units	48	48	48	365	509
Concepts	Unique	11	9	32	123	124
	Leaves	4	2	11	45	46
	Non-leaves	7	7	21	78	78
	Total	12	13	52	713	790
$\psi_C(u)$	Size	2.50e-01	2.71e-01	1.08	1.95	1.55
	Non-empty	2.40	2.60	5.20	3.35	3.39
	Depth	8.82	7.44	7.88	6.98	6.98
	$\sigma(u, c)$	4.70e-02	4.75e-02	7.22e-02	7.08e-02	7.01e-02

Table B.5: Semantic alignment of DenseNet with the Broden dataset. Fixed threshold $\tau = 0.04$ over the IoU metric.

	Module	DenseBlock4.22	DenseBlock4.23	DenseBlock4.24	classifier	densenet161
	Units	48	48	48	365	509
Concepts	Unique	28	66	53	427	433
	Leaves	25	52	43	314	320
	Non-leaves	3	14	10	113	113
	Total	40	101	102	1755	1998
$\psi_C(u)$	Size	8.33e-01	2.10	2.12	4.81	3.93
	Non-empty	1.67	2.97	3.00	5.40	4.79
	Depth	8.32	7.82	7.79	7.70	7.69
	$\sigma(u, c)$	2.79e-01	3.16e-01	3.27e-01	3.80e-01	3.72e-01

Table B.6: Semantic alignment of DenseNet with the Broden dataset. Fixed threshold $\tau = 0.2$ over the likelihood $\mathcal{L}(c \mid u)$ metric.

B.2 ImageNet Semantic Alignment

	Module	features.8	features.10	classifier.1	classifier.4	classifier.6	alexnet
	Units	256	256	4096	4096	365	9069
Concepts	Unique	1	3	443	649	361	712
	Leaves	1	3	379	543	284	589
	Non-leaves	0	0	64	106	77	123
	Total	6	4	4129	5644	1506	11289
$\psi_C(u)$	Size	2.34e-02	1.56e-02	1.01	1.38	4.13	1.24
	Non-empty	1.00	1.33	1.96	2.33	4.56	2.32
	Depth	8.00	8.33	9.42	9.63	9.21	9.60
	$\sigma(u, c)$	3.05e-01	2.86e-01	2.67e-01	2.81e-01	3.31e-01	2.82e-01

Table B.7: Semantic alignment of AlexNet with the ImageNet dataset. Fixed threshold $\tau = 0.2$ over the likelihood $\mathcal{L}(c \mid u)$ measure.

	Module	layer4.0.conv1	layer4.0.conv2	layer4.0	layer4.1.conv1	layer4.1.conv2	layer4.1	fc	resnet18
	Units	512	512	512	512	512	512	365	3437
Concepts	Unique	3	20	24	45	194	199	482	494
	Leaves	3	20	24	41	164	170	379	390
	Non-leaves	0	0	0	4	30	29	103	104
	Total	4	24	26	48	387	399	1812	2700
$\psi_C(u)$	Size	7.81e-03	4.69e-02	5.08e-02	9.38e-02	7.56e-01	7.79e-01	4.96	7.86e-01
	Non-empty	2.00	1.71	1.62	1.50	1.99	2.05	5.30	3.40
	Depth	7.67	8.90	8.79	8.62	9.38	9.35	9.32	9.32
	$\sigma(u, c)$	2.48e-01	2.42e-01	2.40e-01	2.52e-01	2.70e-01	2.69e-01	3.63e-01	3.31e-01

Table B.8: Semantic alignment of ResNet with the ImageNet dataset. Fixed threshold $\tau = 0.2$ over the likelihood $\mathcal{L}(c \mid u)$ measure.

	Module	DenseBlock4.22	DenseBlock4.23	DenseBlock4.24	classifier	densenet161
	Units	48	48	48	365	509
Concepts	Unique	7	16	16	528	530
	Leaves	6	14	14	410	412
	Non-leaves	1	2	2	118	118
	Total	12	20	22	1661	1715
$\psi_C(u)$	Size	2.50e-01	4.17e-01	4.58e-01	4.55	3.37
	Non-empty	1.20	1.67	1.69	5.00	4.67
	Depth	8.43	8.25	9.44	9.36	9.35
	$\sigma(u, c)$	2.66e-01	2.73e-01	3.03e-01	3.75e-01	3.72e-01

Table B.9: Semantic alignment of DenseNet with the ImageNet dataset. Fixed threshold $\tau = 0.2$ over the likelihood $\mathcal{L}(c \mid u)$ measure.

B.3 Broden Neural Circuits

ID	Pairs	Units	Concepts	Highest concept	Most common concept	Coherence	Unique coherence
0	850	731	9	space.n.02	ball.n.01	0.625331	0.259294
2	5	3	2	orifice.n.01	mouth.n.01	0.964942	0.94157
3	190	189	2	structure.n.03	grid.n.01	0.989474	0
6	275	208	9	excavation.n.03	gasoline_station.n.01	0.61525	0.459214
7	60	60	2	rim.n.03	waterwheel.n.01	0.966667	0
9	331	330	3	ligament.n.02	binder.n.04	0.986998	0.2804
11	321	277	8	food.n.02	vegetable.n.01	0.635454	0.563819
13	590	511	9	tower.n.01	viaduct.n.01	0.339819	0.305501
16	111	91	6	passageway.n.01	aqueduct.n.01	0.737838	0.658568
18	7	7	2	text.n.01	text.n.01	0.714286	0
22	38	21	3	structural_member.n.01	tread.n.04	0.480597	0.243158
24	95	92	3	cloud.n.01	fog.n.01	0.995741	0.946449
25	719	559	20	porch.n.01	mosque.n.01	0.472632	0.353401
27	210	190	4	plant_organ.n.01	leaf.n.01	0.880618	0.751105
31	486	463	12	oar.n.01	cockpit.n.01	0.640215	0.302589
35	121	94	6	body_of_water.n.01	pond.n.01	0.798482	0.750965
37	94	84	4	electrical_device.n.01	dashboard.n.02	0.930582	0.732251
38	312	235	4	memory_device.n.01	videocassette.n.01	0.495736	0.421269
46	253	219	9	geological_formation.n.01	shore.n.01	0.715117	0.675303
48	160	80	2	armor_plate.n.01	helmet.n.01	0.962684	0.925835
57	365	343	8	desert.n.01	tennis_court.n.01	0.423691	0.448653
62	33	28	4	deck.n.01	aircraft_carrier.n.01	0.588415	0.397467
69	52	49	4	roof.n.02	roof.n.02	0.718415	0.552456
71	198	121	4	memorial.n.03	gravestone.n.01	0.609763	0.445555
75	396	366	5	shelter.n.01	circus_tent.n.01	0.906657	0.851022
77	76	71	3	water_faucet.n.01	steering_wheel.n.01	0.976087	0.79145
79	20	19	2	figure.n.04	dummy.n.03	0.996855	0.968546
81	16	12	2	scale.n.07	weighbridge.n.01	0.6	0
85	86	50	3	airfoil.n.01	rudder.n.01	0.829244	0.708283
90	163	115	3	workplace.n.01	vineyard.n.01	0.80421	0.621505
92	200	129	2	power_shovel.n.01	steam_shovel.n.01	0.959715	0.91353
93	128	123	3	wing.n.01	duck.n.01	0.856421	0.292789
96	54	51	2	conveyer_belt.n.01	carousel.n.01	0.893082	0
97	28	25	2	rubbish.n.01	debris.n.01	0.997945	0.989643
105	142	105	12	merchandise.n.01	jersey.n.03	0.793094	0.693743
108	97	67	2	machinery.n.01	windmill.n.01	0.916062	0.805565
117	25	25	2	pier.n.01	quay.n.01	0.92	0
123	28	28	3	hoop.n.02	tire.n.01	0.94303	0.481756
129	15	14	3	window.n.01	dormer.n.01	0.51537	0.259486
132	5	4	2	table_game.n.01	table_tennis.n.01	0.920032	0.800079
137	6	4	3	spoon.n.01	spoon.n.01	0.972869	0.956966
152	8	8	2	slot_machine.n.01	vending_machine.n.01	0.949664	0.90604

Table B.10: Circuits with more than one unique meaning retrieved in AlexNet pretrained with Places365 and aligned with Broden.

ID	Pairs	Units	Concepts	Highest concept	Most common concept	Coherence	Unique coherence
0	862	741	9	space.n.02	ball.n.01	0.626546	0.259294
2	5	3	2	orifice.n.01	mouth.n.01	0.964942	0.94157
3	189	188	2	structure.n.03	grid.n.01	0.989418	0
6	284	215	9	excavation.n.03	gasoline_station.n.01	0.617445	0.459214
7	83	83	4	barrel.n.02	waterwheel.n.01	0.567426	0.214731
9	332	331	3	ligament.n.02	binder.n.04	0.987037	0.2804
11	276	240	8	food.n.02	vegetable.n.01	0.634215	0.563819
12	157	156	2	aquarium.n.01	water_tower.n.01	0.507594	0
14	592	514	9	tower.n.01	viaduct.n.01	0.339497	0.305501
17	67	64	3	conduit.n.01	aqueduct.n.01	0.948512	0.680936
24	16	16	2	crossing.n.05	crossing.n.05	0.918874	0.750383
26	95	92	3	cloud.n.01	fog.n.01	0.995741	0.946449
27	728	565	20	porch.n.01	mosque.n.01	0.47486	0.353401
29	208	190	4	plant_organ.n.01	leaf.n.01	0.884087	0.751105
31	39	39	2	scoreboard.n.01	billboard.n.01	0.86143	0.714777
33	487	465	13	oar.n.01	cockpit.n.01	0.638221	0.316108
34	86	86	2	wheelchair.n.01	wheelchair.n.01	0.931874	0
35	152	152	3	drum.n.01	synthesizer.n.02	0.487248	0.283054
37	119	93	6	body_of_water.n.01	pond.n.01	0.796736	0.750965
38	94	91	4	dryer.n.01	sewing_machine.n.01	0.846097	0.674925
39	308	231	4	memory_device.n.01	videocassette.n.01	0.49483	0.421269
40	42	40	2	escalator.n.02	escalator.n.02	0.562137	0
46	244	217	9	geological_formation.n.01	shore.n.01	0.716514	0.675303
48	160	80	2	armor_plate.n.01	helmet.n.01	0.962684	0.925835
50	116	113	2	boot.n.01	boot.n.01	0.91978	0.731797
58	361	339	8	desert.n.01	tennis_court.n.01	0.426161	0.448653
63	33	28	4	stern.n.01	aircraft_carrier.n.01	0.582435	0.397467
65	48	44	2	conveyer_belt.n.01	carousel.n.01	0.843972	0
69	50	48	4	roof.n.02	roof.n.02	0.727762	0.552456
71	200	123	4	grave.n.02	gravestone.n.01	0.614909	0.445555
74	403	373	5	shelter.n.01	circus_tent.n.01	0.907495	0.851022
76	95	85	4	electrical_device.n.01	dashboard.n.02	0.928053	0.732251
77	77	72	3	water_faucet.n.01	steering_wheel.n.01	0.976381	0.79145
78	20	19	2	figure.n.04	dummy.n.03	0.996855	0.968546
80	20	15	2	scale.n.07	weighbridge.n.01	0.605263	0
83	138	100	11	clothing.n.01	jersey.n.03	0.810025	0.700148
85	92	53	3	airfoil.n.01	rudder.n.01	0.831804	0.708283
89	161	114	3	workplace.n.01	vineyard.n.01	0.803814	0.621505
91	202	131	2	power_shovel.n.01	steam_shovel.n.01	0.959655	0.91353
93	123	118	3	wing.n.01	duck.n.01	0.854961	0.292789
94	57	56	2	revolving_door.n.02	revolving_door.n.02	0.56015	0
95	29	26	2	rubbish.n.01	debris.n.01	0.99801	0.989643
103	99	69	2	machinery.n.01	windmill.n.01	0.917032	0.805565
106	29	29	2	crate.n.01	crate.n.01	0.91697	0.719083
109	41	30	3	passageway.n.01	tunnel.n.01	0.832241	0.712414
124	15	14	3	window.n.01	dormer.n.01	0.51537	0.259486
125	3	3	2	text.n.01	magazine.n.01	0.333333	0
127	5	4	2	table_game.n.01	table_tennis.n.01	0.920032	0.800079
132	6	4	3	spoon.n.01	spoon.n.01	0.972869	0.956966
143	7	7	2	slot_machine.n.01	vending_machine.n.01	0.955257	0.90604
145	3	3	2	merchandise.n.01	stall.n.03	0.526492	0.289738

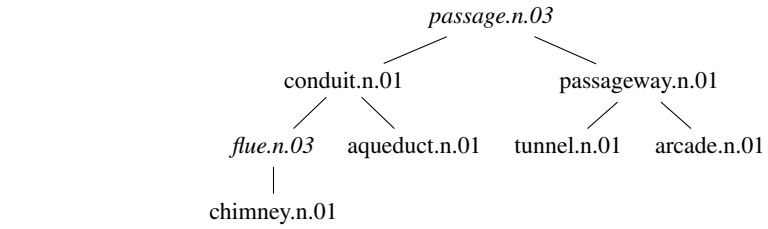
Table B.11: Circuits with more than one unique meaning retrieved in AlexNet pretrained with Places365 and aligned with Broden. Differently to the methodology reported in the main body of the paper, for this run we adopted the Resnik similarity (Resnik, 1995) tested against a threshold equal to 8.0. As already discussed, this threshold has been determined by evaluating the resulting circuits. We report results for a different similarity measure to highlight how circuits retrieval does not depend on the Jiang-Conrath measure only and how results are compatible.

ID	Pairs	Units	Concepts	Highest concept	Most common concept	Coherence	Unique coherence
1	1264	605	106	porch.n.01	beacon.n.03	0.239235	0.234927
2	17	15	3	plate.n.02	license_plate.n.01	0.44753	0.262001
3	5	4	3	text.n.01	text.n.01	0.39229	0.307634
5	45	45	2	structure.n.03	grid.n.01	0.913131	0
6	107	68	9	space.n.02	ball.n.01	0.346641	0.259294
7	104	83	8	passage.n.03	aqueduct.n.01	0.701643	0.678228
9	21	18	3	orifice.n.01	mouth.n.01	0.561412	0.313857
11	6	4	2	eye.n.01	nose.n.01	0.931428	0.871428
14	69	42	6	body_of_water.n.01	waterfall.n.01	0.780574	0.750965
15	104	33	15	merchandise.n.01	shirt.n.01	0.68488	0.571657
19	20	20	2	top.n.01	capital.n.08	0.978116	0.8845
25	30	26	3	airfoil.n.01	stabilizer.n.02	0.880132	0.708283
27	84	53	2	power_shovel.n.01	power_shovel.n.01	0.956269	0.91353
29	53	50	4	geographical_area.n.01	tennis_court.n.01	0.722598	0.602029
30	20	20	2	system.n.01	maze.n.01	0.973198	0.731977
31	52	38	2	machinery.n.01	windmill.n.01	0.921992	0.805565
34	4	2	3	memory_device.n.01	videocassette.n.01	0.326899	0.320464
35	29	29	3	area.n.01	resort_area.n.01	0.789643	0.695962
38	61	20	9	electronic_device.n.01	display_panel.n.01	0.475204	0.43386
39	20	14	3	hoof.n.01	horse.n.01	0.530977	0.372404
44	88	52	7	facility.n.01	gasoline_station.n.01	0.591917	0.50208
47	25	25	6	food.n.01	meat.n.01	0.635034	0.47772
50	12	11	2	white_goods.n.01	washer.n.03	0.983003	0.898017
51	16	12	3	activity.n.01	table_tennis.n.01	0.814476	0.549572
54	29	29	2	wing.n.01	duck.n.01	0.548711	0.0746296
56	30	19	4	electrical_device.n.01	dashboard.n.02	0.787095	0.732251
59	23	19	4	curve.n.01	arch.n.01	0.472645	0.361969
64	36	26	3	workplace.n.01	vineyard.n.01	0.737174	0.621505
65	30	18	7	writing.n.02	text.n.01	0.402414	0.296983
71	25	22	3	plant_organ.n.01	fruit.n.01	0.888738	0.851518
73	11	6	4	footwear.n.02	boot.n.01	0.865618	0.85109
74	17	13	2	land.n.02	badlands.n.01	0.897501	0.731927
78	14	13	2	booth.n.02	telephone_booth.n.01	0.975486	0.907052
79	20	16	2	pier.n.01	quay.n.01	0.557895	0
81	5	4	2	eye.n.01	eye.n.01	0.922857	0.871428
91	27	13	3	memory_device.n.01	videocassette.n.01	0.476092	0.320464
93	31	27	3	geological_formation.n.01	iceberg.n.01	0.752086	0.6328
94	7	7	2	electronic_equipment.n.01	television_camera.n.01	0.950905	0.828169
96	15	9	2	armor_plate.n.01	helmet.n.01	0.961858	0.925835
100	5	5	2	slot_machine.n.01	slot_machine.n.01	0.943624	0.90604
101	6	6	2	group.n.01	group.n.01	0.662791	0.367734
109	8	7	2	bathtub.n.01	hot_tub.n.01	0.75	0
114	11	5	3	structural_member.n.01	tread.n.04	0.429952	0.243158

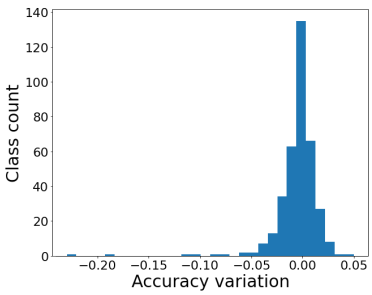
Table B.12: Circuits with more than one unique meaning retrieved in ResNet pretrained with Places365 and aligned with Broden.

ID	Pairs	Units	Concepts	Highest concept	Most common concept	Coherence	Unique coherence
1	15	9	6	musical_instrument.n.01	synthesizer.n.02	0.467058	0.580528
3	9	8	2	white_goods.n.01	washer.n.03	0.977337	0.898017
5	20	14	3	tower.n.01	tower.n.01	0.885044	0.832248
6	3	2	2	place_of_worship.n.01	temple.n.01	0.916572	0.874859
8	2	2	2	signal.n.01	signal.n.01	0.475168	0.475168
11	53	30	7	machine.n.01	steam_shovel.n.01	0.755432	0.636582
12	50	30	10	roof.n.02	bulldozer.n.01	0.326871	0.394883
15	24	15	3	airfoil.n.01	airfoil.n.01	0.805058	0.708283
16	70	37	10	space.n.02	goal.n.03	0.401247	0.299264
17	32	29	5	food.n.02	vegetable.n.01	0.711422	0.599991
19	22	20	3	bridge.n.01	covered_bridge.n.01	0.830977	0.471673
20	22	15	3	workplace.n.01	vineyard.n.01	0.735493	0.621505
21	12	12	2	state.n.02	roller_coaster.n.01	0.833333	0
22	21	19	4	wing.n.01	duck.n.01	0.794905	0.441684
23	10	8	3	lifting_device.n.01	crane.n.04	0.952759	0.941642
26	17	7	4	activity.n.01	table_tennis.n.01	0.547083	0.274786
30	2	2	2	sense_organ.n.01	sense_organ.n.01	0.984071	0.984071
34	44	13	7	fuselage.n.01	fuselage.n.01	0.640271	0.546299
36	23	19	3	geographical_area.n.01	tennis_court.n.01	0.772247	0.656816

Table B.13: Circuits with more than one unique meaning retrieved in DenseNet pretrained with Places365 and aligned with Broden.



(a)



(b)

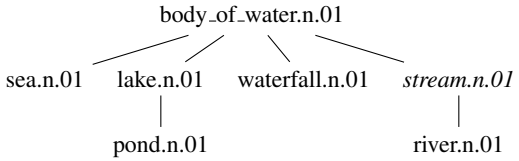
Class	Drop
/c/clothing_store	-0.05
/d/dining_room	-0.05
/g/gazebo/exterior	-0.06
/c/church/outdoor	-0.06
/a/arch	-0.08
/p/pavilion	-0.09
/r/river	-0.1
/b/bridge	-0.11
/v/viaduct	-0.19
/a/aqueduct	-0.23

(c)

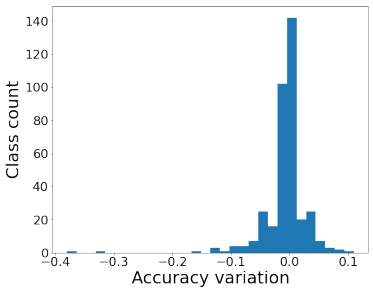
Class	Drop
/c/clothing_store	-0.05
/m/mountain	-0.05
/d/dining_room	-0.06
/c/church/outdoor	-0.07
/a/arch	-0.08
/r/river	-0.08
/p/pavilion	-0.09
/b/bridge	-0.12
/v/viaduct	-0.18
/a/aqueduct	-0.21

(d)

Figure B.1: Importance analysis of circuit n.16 from AlexNet, containing 91 distinct units and 6 unique visual concepts. The circuit is aligned to various concepts relative to passages, especially when water-related. Subfigure (a) reports the aligned concepts within the WordNet taxonomy. *Italic* concepts are not directly aligned, but we include them for visualization purposes. When ablating the circuit, the accuracy drop significantly affects only a small number of classes. Subfigure (b) depicts the histogram of categories of the Places-365 dataset as a function of accuracy drop (on the x-axis). The classes most affected by the ablation are those more strictly related to the semantically aligned concepts of the circuit, as exemplified in (c). Dropping only the units aligned with the most popular concept in the circuit (d) does not account for the effects of ablating the whole circuit (c).



(a)



(b)

Class	Drop
/p/pier	-0.1
/c/cliff	-0.1
/h/harbor	-0.1
/b/boat_deck	-0.11
/d/dam	-0.12
/b/boathouse	-0.13
/s/swimming_hole	-0.13
/p/promenade	-0.16
/w/waterfall	-0.32
/f/fountain	-0.38

(c)

Class	Drop
/f/field/cultivated	-0.06
/v/volcano	-0.06
/c/crevasse	-0.07
/a/amusement_park	-0.07
/t/ticket_booth	-0.08
/h/hot_spring	-0.09
/s/science_museum	-0.1
/d/dam	-0.15
/w/waterfall	-0.33
/f/fountain	-0.4

(d)

Figure B.2: Importance analysis of circuit n.14 from ResNet, containing 42 distinct units and 6 unique visual concepts. The circuit is aligned to various concepts relative to bodies of water. Subfigure (a) reports the aligned concepts within the WordNet taxonomy. *Italic* concepts are not directly aligned, but we include them for visualization purposes. When ablating the circuit, the accuracy drop significantly affects only a small number of classes. Subfigure (b) depicts the histogram of categories of the Places-365 dataset as a function of accuracy drop (on the x-axis). The classes most affected by the ablation are those more strictly related to the semantically aligned concepts of the circuit, as exemplified in (c). Dropping only the units aligned with the most popular concept in the circuit (d) does not account for the effects of ablating the whole circuit (c).

B.4 IMAGENET NEURAL CIRCUITS

ID	Pairs	Units	Concepts	Highest concept	Most common concept	Coherence	Unique coherence
0	291	285	4	screen.n.05	window_screen.n.01	0.952625	0
1	118	98	3	press.n.02	school_newspaper.n.01	0.756026	0.236177
3	51	32	3	bed.n.01	four-poster.n.01	0.45849	0.243415
4	81	75	4	white_goods.n.01	refrigerator.n.01	0.904113	0.870666
6	1223	694	42	ball.n.01	pickup.n.01	0.318472	0.19791
8	163	151	3	sheet.n.06	puck.n.02	0.562053	0.236684
11	102	87	2	reservoir.n.03	water_tower.n.01	0.96552	0.863904
14	89	68	6	top.n.09	bottlecap.n.01	0.322545	0.0843083
19	47	44	4	baby_bed.n.01	crib.n.01	0.974298	0.916775
20	324	169	24	nutriment.n.01	pizza.n.01	0.228595	0.189299
21	1192	766	28	housing.n.01	palace.n.04	0.314521	0.282249
22	468	371	9	memorial.n.03	triumphal_arch.n.01	0.256961	0.110752
24	250	184	6	public_transport.n.01	school_bus.n.01	0.288938	0.166539
26	916	427	33	wing.n.02	lifeboat.n.01	0.346403	0.385034
27	474	273	14	establishment.n.04	toyshop.n.01	0.216096	0.2173
29	143	102	3	locomotive.n.01	electric_locomotive.n.01	0.469812	0
31	205	143	3	slot_machine.n.01	slot.n.07	0.964002	0.935163
35	159	114	3	cabinet.n.01	medicine_chest.n.01	0.353555	0
38	79	58	3	communication.n.02	street_sign.n.01	0.726745	0.326418
40	68	35	6	duck.n.01	red-breasted_merganser.n.01	0.812208	0.457885
41	159	104	5	person.n.01	scuba_diver.n.01	0.469973	0.27022
44	7	7	2	free-reed_instrument.n.01	harmonica.n.01	0.962366	0.868279
47	385	202	15	geological_formation.n.01	alp.n.01	0.614762	0.558982
51	58	43	7	coelenterate.n.01	jellyfish.n.02	0.384755	0.190476
54	25	23	3	worm.n.01	nematode.n.01	0.975471	0.883518
56	77	65	4	blind.n.03	theater_curtain.n.01	0.483109	0.16163
59	88	77	3	seed.n.01	rapeseed.n.01	0.682281	0.289388
60	62	48	3	flower.n.01	daisy.n.01	0.778478	0.255305
61	30	24	3	bird_of_prey.n.01	kite.n.04	0.704017	0.281213
62	97	83	3	equine.n.01	sorrel.n.05	0.867603	0.645367
65	13	12	2	shoe.n.01	wing_tip.n.01	0.846154	0
70	39	39	4	lever.n.01	organ.n.05	0.420929	0.0573089
71	4	4	2	cornet.n.01	french_horn.n.01	0.955708	0.911416
72	97	88	7	pool_table.n.01	pool_table.n.01	0.293053	0.0706613
74	24	21	3	cloth_covering.n.01	band_aid.n.01	0.497121	0.263062
75	10	9	3	swimsuit.n.01	maillot.n.01	0.643598	0.320638
78	19	15	3	percoid_fish.n.01	rock_beauty.n.01	0.438596	0
81	103	97	2	gear.n.04	drilling_platform.n.01	0.889206	0
85	5	4	2	odonate.n.01	damselfly.n.01	1	1
88	58	52	2	floor_cover.n.01	prayer_rug.n.01	0.811252	0
91	35	26	3	whale.n.02	grey_whale.n.01	0.488024	0.329858
93	27	23	3	astronomical_telescope.n.01	newtonian_telescope.n.01	0.418803	0
94	37	25	6	personal_computer.n.01	hand-held_computer.n.01	0.234234	0.2
99	41	29	2	material.n.01	gravel.n.01	0.820461	0.576949
102	5	4	2	weight.n.02	barbell.n.01	0.94725	0.868126
103	3	3	2	bib.n.01	apron.n.01	0.450865	0.176297
106	25	21	5	photographic_equipment.n.01	polaroid_camera.n.01	0.300312	0.250013
107	34	18	3	shirt.n.01	polo_shirt.n.01	0.484291	0.316997
108	21	14	3	power_tool.n.01	circular_saw.n.01	0.876812	0.813926
112	30	28	4	electro-acoustic_transducer.n.01	headset.n.01	0.760106	0.461025
115	7	5	3	knife.n.01	carving_knife.n.01	0.410684	0.270726
121	66	42	8	overgarment.n.01	mess_jacket.n.01	0.301214	0.295988
123	30	20	3	citrus.n.01	orange.n.01	0.968543	0.937791
124	43	39	5	decoration.n.01	obelisk.n.01	0.681294	0.267083
127	7	6	2	food_fish.n.01	coho.n.02	0.714286	0
130	25	21	3	gymnastic_apparatus.n.01	balance_beam.n.01	0.837949	0.777162
138	21	14	3	farm_machine.n.01	harvester.n.02	0.570147	0.331539
139	14	12	2	monotreme.n.01	echidna.n.02	0.993526	0.975453
149	18	17	2	pot.n.01	coffeepot.n.01	0.987252	0.885269
156	9	8	3	bicycle.n.01	bicycle-built-for-two.n.01	0.420012	0.260037
159	41	36	3	movable_barrier.n.01	sliding_door.n.01	0.45122	0
160	15	15	3	jinrikisha.n.01	horse_cart.n.01	0.37875	0.0769099
168	8	7	2	armor_plate.n.01	pickelhaube.n.01	0.75	0
171	10	8	3	sandpiper.n.01	red-backed_sandpiper.n.01	0.311111	0
179	18	17	2	solanaceous_vegetable.n.01	mashed_potato.n.01	0.888889	0
181	17	14	2	rabbit.n.01	angora.n.03	0.691176	0
183	23	16	3	communication_system.n.01	monitor.n.05	0.639657	0.502954
188	34	18	5	elasmobranch.n.01	great_white_shark.n.01	0.229847	0.0912011
192	18	16	2	trouser.n.01	jean.n.01	0.971391	0.863211
193	9	9	3	starfish.n.01	sea_urchin.n.01	1	1
198	7	5	3	capuchin.n.02	howler_monkey.n.01	1	1
208	9	8	3	winter_squash.n.02	butternut_squash.n.02	1	1
210	8	7	2	elephant.n.01	indian_elephant.n.01	0.970095	0.880378
216	8	7	2	towel.n.01	bath_towel.n.01	0.974183	0.896731

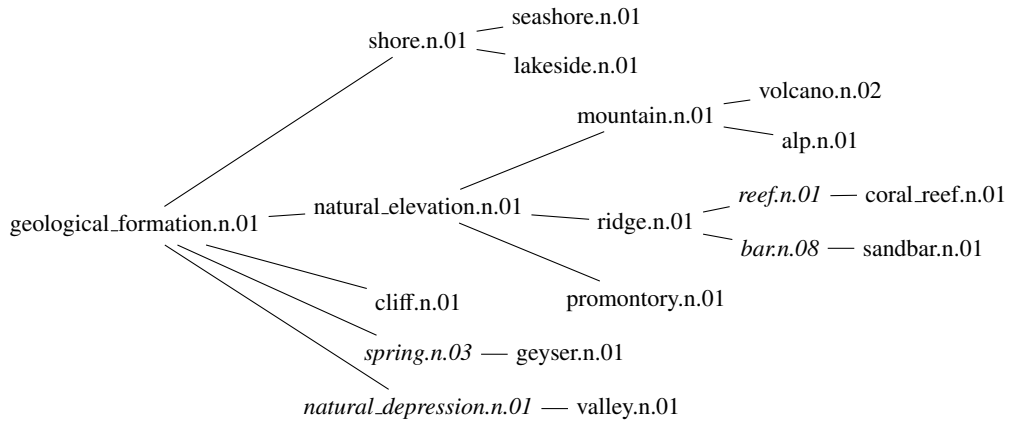
Table B.14: Circuits with more than one unique meaning retrieved in AlexNet pretrained with Places365 and aligned with ImageNet.

ID	Pairs	Units	Concepts	Highest concept	Most common concept	Coherence	Unique coherence
4	77	44	9	geological_formation.n.01	geyser.n.01	0.654843	0.671612
7	101	58	8	bridge.n.01	triumphal_arch.n.01	0.217881	0.142396
8	19	12	4	baby_bed.n.01	crib.n.01	0.957237	0.916775
18	7	5	2	material.n.01	gravel.n.01	0.798547	0.576949
19	61	41	6	public_transport.n.01	trolleybus.n.01	0.325314	0.166539
20	30	16	3	car.n.02	passenger_car.n.01	0.904969	0.86348
21	60	26	3	locomotive.n.01	electric_locomotive.n.01	0.328814	0
22	34	30	4	pool_table.n.01	pool_table.n.01	0.408058	0.152828
24	23	12	3	shirt.n.01	polo_shirt.n.01	0.472837	0.316997
26	23	14	4	condiment.n.01	carbonara.n.01	0.448014	0.398186
27	15	11	2	glass.n.02	beer_glass.n.01	0.580952	0
28	192	45	27	military_vehicle.n.01	airliner.n.01	0.370326	0.432937
29	22	17	2	reservoir.n.03	water_tower.n.01	0.949922	0.863904
31	15	10	3	swimsuit.n.01	bikini.n.02	0.512009	0.307519
33	52	29	7	overgarment.n.01	lab_coat.n.01	0.264992	0.252517
35	26	24	2	gear.n.04	drilling_platform.n.01	0.852308	0
36	73	22	12	nutriment.n.01	cheeseburger.n.01	0.405082	0.493011
39	115	44	11	establishment.n.04	confectionery.n.02	0.267614	0.276645
40	47	36	4	theater.n.01	cinema.n.02	0.427225	0.21267
41	133	71	8	housing.n.01	palace.n.04	0.407624	0.291335
42	18	14	3	cabinet.n.01	medicine_chest.n.01	0.379085	0
43	16	8	3	flower.n.01	yellow_lady	0.451232	0.255305
45	20	14	2	equine.n.01	sorrel.n.05	0.872006	0.71049
46	8	6	2	towel.n.01	bath_towel.n.01	0.955742	0.896731
47	16	12	3	blind.n.03	theater_curtain.n.01	0.598993	0.323261
48	107	32	17	grille.n.02	sports_car.n.01	0.467275	0.479944
52	36	13	3	bed.n.01	bed.n.01	0.469733	0.243415
55	17	16	2	telephone.n.01	pay-phone.n.01	0.882353	0
57	58	16	5	shorebird.n.01	redshank.n.01	0.244757	0.0983143
58	28	11	6	anseriform_bird.n.01	eider.n.01	0.57625	0.465927
59	4	3	2	rabbit.n.01	angora.n.03	0.5	0
61	67	30	4	place_of_worship.n.01	mosque.n.01	0.6712	0.43035
67	47	25	4	person.n.01	ballplayer.n.01	0.622116	0.31211
68	11	10	2	beverage.n.01	espresso.n.01	0.818182	0
69	9	4	3	spiny-finned_fish.n.01	rock_beauty.n.01	0.44387	0.331609
70	13	8	2	communication.n.02	street_sign.n.01	0.701406	0.417741
72	7	6	2	seed.n.01	rapeseed.n.01	0.714286	0
79	17	15	2	column.n.06	obelisk.n.01	0.971856	0.872413
80	15	14	2	shed.n.01	boathouse.n.01	0.975887	0.819153
83	4	3	2	decoration.n.01	necklace.n.01	0.904723	0.809447
84	7	5	2	slot_machine.n.01	slot.n.07	0.997932	0.996382
87	8	4	3	shark.n.01	hammerhead.n.03	0.321429	0
89	6	4	2	white_goods.n.01	washer.n.03	0.945609	0.898017
94	27	12	4	personal_computer.n.01	laptop.n.01	0.316239	0.166667
95	17	10	3	communication_system.n.01	monitor.n.05	0.627787	0.502954
98	9	7	3	footwear.n.02	wing_tip.n.01	0.610668	0.328013
99	14	5	6	coelenterate.n.01	stony_coral.n.01	0.43956	0.266667
101	11	5	3	bicycle.n.01	mountain_bike.n.01	0.418564	0.260037
103	5	4	2	trouser.n.01	jean.n.01	0.945284	0.863211
112	6	4	2	sheet.n.06	scoreboard.n.01	0.84536	0.710051

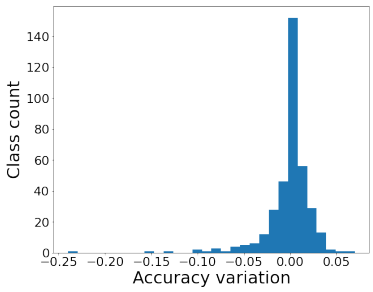
Table B.15: Circuits with more than one unique meaning retrieved in ResNet pretrained with Places365 and aligned with ImageNet.

ID	Pairs	Units	Concepts	Highest concept	Most common concept	Coherence	Unique coherence
1	23	11	3	car.n.02	freight_car.n.01	0.909144	0.86348
2	15	6	3	locomotive.n.01	electric_locomotive.n.01	0.295238	0
3	7	6	2	car.n.01	racer.n.02	0.917336	0.710677
4	7	4	3	table.n.02	pool_table.n.01	0.333333	0
5	6	5	2	seed.n.01	rapeseed.n.01	0.666667	0
6	5	3	2	flower.n.01	daisy.n.01	0.859548	0.765914
7	4	2	3	nutriment.n.01	pizza.n.01	0.768412	0.752821
8	11	9	3	overgarment.n.01	lab_coat.n.01	0.48908	0.324948
9	3	3	2	person.n.01	person.n.01	0.550697	0.326045

Table B.16: Circuits with more than one unique meaning retrieved in DenseNet pretrained with Places365 and aligned with ImageNet.



(a)



(b)

Class	Drop
/l/lagoon	-0.07
/l/lake/natural	-0.08
/m/mountain_path	-0.08
/c/coast	-0.08
/t/tundra	-0.09
/o/ocean	-0.1
/m/mountain_snowy	-0.1
/v/valley	-0.13
/h/hot_spring	-0.15
/m/mountain	-0.24

(c)

Class	Drop
/f/forest_path	-0.04
/c/castle	-0.04
/m/mansion	-0.04
/m/mountain_snowy	-0.05
/g/glacier	-0.05
/v/valley	-0.06
/d/desert/sand	-0.07
/c/chalet	-0.07
/m/mountain_path	-0.09
/m/mountain	-0.11

(d)

Figure B.3: Importance analysis of circuit n.47 from AlexNet aligned with the ImageNet dataset, containing 202 distinct units and 15 unique visual concepts. The circuit is aligned to geological formations, such as mountains. Subfigure (a) reports the aligned concepts within the WordNet taxonomy. *Italic* concepts are not directly aligned, but we include them for visualization purposes. When ablating the circuit, the accuracy drop significantly affects only a small number of classes. Subfigure (b) depicts the histogram of categories of the Places-365 dataset as a function of accuracy drop (on the x-axis). The classes most affected by the ablation are those more strictly related to the semantically aligned concepts of the circuit, as exemplified in (c). Dropping only the units aligned with the most popular concept in the circuit (d) does not account for the effects of ablating the whole circuit (c).

C Semantic Similarity

WordNet enables the estimate of semantic similarity using different measures (Meng et al., 2013). Essentially, the similarity between two concepts is estimated by studying their information content (IC) in the specialization induced taxonomy. The IC of a specific concept consists of the negative log-probability of the concept itself,

$$\mathrm{IC}(c) = -\log P(c). \tag{C.1}$$

Given a text corpus, the maximum likelihood estimate of $P(c)$ is obtained by counting the N_c occurrences of the concept c, against the total number of terms N,

$$P(c) = \frac{N_c}{N}. \tag{C.2}$$

With the current formulation, the IC of a concept monotonically increases crossing the WordNet taxonomy from the root r, which has $\mathrm{IC}(r) = 0$, to the leaves. Using the IC metric, Lin (1998) proposes to estimate the similarity s_{lin} between two arbitrary concepts c_1, c_2 as

$$s_{\mathrm{lin}}(c_1, c_2) = \frac{2\mathrm{IC}(\mathrm{LCS}(c_1, c_2))}{\mathrm{IC}(c_1) + \mathrm{IC}(c_2)}, \tag{C.3}$$

where the least common subsumer (LCS) of two concepts is defined as their nearest common ancestor in the taxonomy. Intuitively, the more the two concepts are specific, the more they are similar if their LCS is also sufficiently specific. Furthermore, given the monotonicity of IC in respect to the taxonomical order

$$\mathrm{IC}(LCS\,(c_1, c_2)) \leq \mathrm{IC}(c_1), \mathrm{IC}(c_2), \tag{C.4}$$

for any possible pair (c_1, c_2), it holds

$$s_{\mathrm{lin}}(c_1, c_2) \in [0, 1]. \tag{C.5}$$

Notably, the Lin similarity derives from the Resnisk similarity (Resnik, 1995),

$$s_{\mathrm{resnik}} = \mathrm{IC}(\mathrm{LCS}(c_1, c_2)) \tag{C.6}$$

Alternatively, Jiang & Conrath (1997) propose a similarity measure based on the notion of link strength. Given two concepts c, p, where p is an ancestor of c, the link strength can be measured as follows:

$$\begin{aligned} \mathrm{LS}(c, p) &= -\log P(c \mid p) \\ &= -\log \frac{P(c, p)}{P(p)} \\ &= -\log \frac{P(c)}{P(p)} \\ &= -\log P(c) + \log P(p) \\ &= \mathrm{IC}(c) - \mathrm{IC}(p). \end{aligned} \tag{C.7}$$

The similarity s_{jcn} is then expressed as the inverse of the sum of the link strengths between two arbitrary concepts c_1, c_2 and their LCS.

$$\begin{aligned} s_{\mathrm{jcn}}(c_1, c_2) &= \frac{1}{\mathrm{LS}(c_1, \mathrm{LCS}(c_1, c_2)) + \mathrm{LS}(c_2, \mathrm{LCS}(c_1, c_2))} \\ &= \frac{1}{\mathrm{IC}(c_1) + \mathrm{IC}(c_2) - 2\mathrm{IC}(\mathrm{LCS}(c_1, c_2))} \end{aligned} \tag{C.8}$$