Understanding Individual Neurons of ResNet Through Improved Compositional Formulas

Rafael Harth(✉)

Department of Information Security, University of Stuttgart, Stuttgart, Germany
Rafael.Harth@gmail.com

Abstract. Compositions of concepts from human-annotated datasets, e.g., "chair OR table", have been shown to approximate latent representations of image classifiers better than single concepts. In this work, we introduce the Close Algorithm, which improves performance according to the IoU metric by utilizing the non-logical connectors CLOSE-TO, WITH, and EXPAND. We consider the shortcomings of current approaches, discuss possible causes, and review a small user study we have run to collect evidence on this point. We also introduce a metric that discourages the reliance on scene-level annotations. (The code to replicate the technical results (along with additional sample images) can be accessed at https://github.com/rafaelharth/indres).

Keywords: Machine learning · Image classification · Interpretability

1 Introduction

Neural networks achieve state-of-the-art performance on many tasks but are infamously hard to understand. Existing explainability methods typically focus on local approximations [16,20], textual explanation modules [5,13,15], saliency maps [2,11,22], or pointing to examples [6,21] or influential training data [19]. In contrast, scant attention has been paid to individual neurons, and almost no work has focused on understanding and predicting neuron activations in detail.

In the context of image processing, human-annotated concepts from a densely labeled dataset provide a possible starting point for this problem [23]. Thresholding neuron activations allows computing their similarity with such concepts using metrics for the overlap between two sets, such as Intersection over Union (IoU). This method reveals that, e.g., the eleventh neuron in the final hidden layer of ResNet-18 trained on places365 is most similar to the concept "highway".

This approach can be improved by connecting several concepts into formulas [17], e.g., "(highway OR field-cultivated) AND NOT sky". In this work, we build on this approach by introducing the non-logical connectors CLOSE-TO, WITH, and EXPAND (Sect. 3). We discuss problems with IoU as a metric and suggest an alternative (Sect. 4). We show that incorporating the three new connectors improves accuracy according to either metric, observe that it is still poor, and

M. El Yacoubi et al. (Eds.): ICPRAI 2022, LNCS 13364, pp. 283–294, 2022.
https://doi.org/10.1007/978-3-031-09282-4_24

suggest possible reasons (Sect. 5). To provide evidence on this point, we review a small study in which two trained annotators were asked to take the role of the algorithm in predicting neuron activations for four different neurons (Sect. 6).

2 Related Work

In [12], Gonzalez-Garcia et al. aim to answer whether "CNNs learn semantic parts in their internal representation" using the PASCAL-Part dataset [7] as the source of ground truth. They match bounding boxes with individual neurons or combinations of neurons from AlexNet, finding that "34 out of the 105 semantic part classes emerge", which they call a "modest number".

Using Broden, a richer dataset with 1197 annotated concepts, Zhou et al. [23] find more promising results with their approach called "Network Dissection". They show that human-understandable concepts naturally emerge in hidden layers of neural networks but are lost under basis transformations of the latent representation, making them artifacts of the specific training procedure rather than an inevitable side-effect of discriminatory power. However, these results rely on an inclusive notion of emergence: the similarity between annotated concepts and neuron activations, while easily clearing the threshold for statistical significance, is usually mild.[1] Mu and Andreas [17] improve this level of similarity by matching neurons to combinations of several human-annotated concepts. This is the most direct predecessor of our work.

Fong and Vedaldi [10] introduce a method that learns a vector $\mathbf{w}_C \in \mathbb{R}^K$ for each concept C, where K is the number of neurons in a specific layer. Thus, each concept is assigned a linear combination of neurons, which improves predictive quality substantially compared to the 1:1 mapping introduced in [23].[2] Using one-hot vectors recovers single neuron mappings as a special case.

In an attempt to unify different approaches for interpreting image classifiers, [18] provides a rich set of visualizations showing how a network's latent representation evolves across layers, including visualizations that combine the representations of all neurons into a single image. [9] and [3] study individual neurons in language processing tasks. [8] provides a toolkit for analyzing individual neurons in practice.

3 Algorithmic Compositional Explanations

3.1 Setup

Following [17], we examine the 512 individual neurons (or "filters") from the last hidden layer of ResNet-18 [14] trained on places365 [24]. To predict neuron activations algorithmically, we use the Broden dataset introduced in [23].

[1] While mean IoU scores are not reported in the paper, [17] finds a mean of 0.059 for ResNet in what (as far as we can tell) is an identical setting.

[2] Even though this similarity is also measured by IoU, a quantitative comparison to our results is not possible because [10] does not examine neurons from ResNet.

In Broden, images are annotated with 1197 different classes that each belong to one of the six categories 'color', 'object', 'part', 'material', 'scene', and 'texture', where the last two categories annotate on a per-image basis, the remaining ones on the pixel level (112×112). Annotations are non-overlapping within each category but can overlap between categories, e.g., the same pixel can be annotated as 'car' (object), 'car dealership' (scene), and 'gray' (color). Broden combines several datasets, but only images from Ade20k [25] have been used in this work.

Neurons from the last hidden layer of ResNet output a 7×7 grid of numbers for each input image. This induces a division of each image into 49 cells, and we write $\mathbf{C}$ to denote the set of all cells from images in Ade20k so that each neuron n can be viewed as a function $n : \mathbf{C} \to \mathbb{R}$.

In this setting, each choice of a threshold $t_n \in \mathbb{R}$ induces a binary function $f_n : \mathbf{C} \to \{1, 0\}$, where $f_n(c) = 1$ iff $n(c) > t_n$. We call f_n the *neuron mask* for n and the value $\frac{1}{|\mathbf{C}|} \sum_{c \in \mathbf{C}} f_n(c) \in [0, 1]$ the *coverage* of f_n. Following [23] and [17], we choose thresholds t_n such that each neuron mask has coverage 0.5%.

To make human-annotated concepts comparable with neuron masks, they have been downsampled from their 112×112 resolution to 7×7 using block-based downsampling with a threshold of 100 (i.e., for each pixel-based class, each image is divided into 49 16×16 blocks corresponding to the cells in $\mathbf{C}$, and the class is considered to activate on that block if it activates on at least 100 of the 256 pixels). Image-level classes are converted to 49 or 0 activated cells, respectively. Previous approaches [4,17,23] instead rescale neuron activations to the 112×112 resolution using bilinear upsampling, but this makes the approach computationally infeasible given the connectors introduced in the upcoming Section. Given that the 7×7 neuron activations constitute the ground truth for this task, it is also unclear whether upsampling is desirable.

3.2 Connecting Annotated Concepts

In [17], human-annotated concepts are combined using the logical connectors AND, OR, and AND NOT, but many non-logical connectors are also possible.

Motivated by the observation that some neurons seem to care about two concepts appearing concurrently, we have introduced the binary connectors CLOSE TO and WITH, which restrict a given neuron mask f_n to cells where a second mask f_m activates within a two-cell radius (CLOSE TO) or anywhere in the same image (WITH). Furthermore, we found that some neurons that care about a specific concept also activate on cells adjacent to the concept. To capture this behavior, we have introduced the unitary connector EXPAND that widens the area of a single concept.

Formally, writing $\vee, \wedge, \neg$ to denote logical connectors, f_n, f_m for neuron masks, $C, D, E, F \in \mathbf{C}$ for cells, $\mathrm{im}(C)$ for the set of the 49 cells in the image corresponding to C, and $N(C)$ for the set of (at most 21) cells in the 5×5 square with corners removed around C, the three connectors can be defined as follows:

- $(f_n \text{ WITH } f_m)(C) := f_n(C) \wedge \bigvee_{D \in \mathrm{im}(C)} f_m(D)$.

- $(f_n \text{ CLOSE TO } f_m)(C) := f_n(C) \wedge \bigvee_{D \in N(C)} f_m(D)$. Furthermore, we define $(f_n \text{ CLOSE } f_m) := ((f_n \text{ CLOSE TO } f_m) \text{ OR } (f_m \text{ CLOSE TO } f_n))$. Note that $(f_n \text{ CLOSE } f_m)$ is treated as a formula of length 2.
- $\text{EXPAND}(f_n)(C) := f_n(C) \vee \bigvee_{(D,E,F) \in \text{adjacent}^3(C)} f_m(D) \wedge f_m(E) \wedge f_m(F)$, where $\text{adjacent}^3(C)$ is the set of all triples of three different cells adjacent to C (diagonal adjacency is permitted). EXPAND is applied to singular concepts only and does not increase formula length. We abbreviate it by the postfix -X, e.g., we write chair-X rather than EXPAND(chair).

One thing to keep in mind is the difference between the accuracy of an approximation (as measured by, e.g., IoU) and how much it helps a human understand the concept. For example, we will show in Sect. 5 that increasing the length of formulas increases IoU for either algorithm, but a formula of length 10 may still be less useful than one of length 3. We believe that the non-logical connectors introduced above (including the abbreviation "A CLOSE B") are intuitively simple and thus helpful for genuine understanding.

Throughout the paper, we refer to the algorithm using the new connectors as the "Close Algorithm" and the algorithm relying exclusively on AND, AND NOT, and OR as the "standard algorithm". We refer to masks produced by formulas from either algorithm as "label masks".

4 Locality and the ImRoU Metric

As mentioned in Sect. 3.1, Broden contains both pixel- and scene-level annotations. Scene-level annotations are made on a per-image basis, meaning that each image is either annotated fully or not at all. When optimizing for IoU with formula length 3, the standard algorithm finds a set of formulas in which over half of all primitive labels are scene-level.

At first glance, one may suspect that this reflects behaviors exhibited by the corresponding neuron masks. However, Fig. 1 shows that even neuron masks whose formula contains three scene-level annotations predominantly activate on small parts of their images (red graph), if they activate at all. This makes the algorithm's reliance on scene-level annotations intuitively undesirable. Furthermore, comparing them to neuron masks whose formulas contain zero scene-level annotations (blue graph) shows only marginal differences.

One way to discourage scene-level annotations is to add a penalty term that disproportionately affects false positives in images where neurons activate strongly. Given two masks $f_n, f_m : \mathbf{C} \to \{0, 1\}$ (think of one neuron-, and one label mask), let N and M be the sets of cells on which they activate, i.e., $N = \{c \in \mathbf{C} : f_n(c) = 1\}$. One can compute a "random intersection" for each image I, which is equal to the expected size of the intersection $N \cap M$ if all cells of M were chosen uniformly at random out of I (see Fig. 2 for an example). Based on this, we introduce the metric ImRoU_r (**I**ntersection **m**inus **R**andom Intersection **o**ver **U**nion), which is computed by subtracting r times the random intersection (summed over all images) before dividing by the union. Formally,

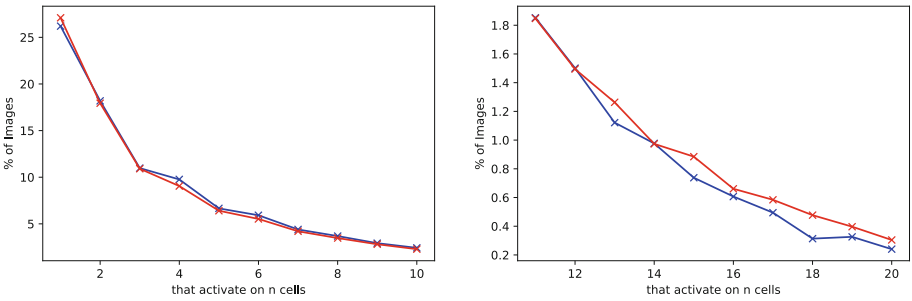

Fig. 1. Relative frequencies of images where neurons activate on exactly n cells, for different values of n, averaged across all neurons with zero (blue) and three (red) scene-level annotations in the formula of length three found by the standard algorithm. Most images (94.5% and 94.7% for the blue and red group, respectively) have no activations at all; the percentages shown are of only the set of images with nonzero activations. This was done to make the graphs readable. (Color figure online)

$$\mathrm{ImRoU}_r^{\text{not-normalized}}(N, M) = \frac{\sum_{I \in \mathbf{I}} |N \cap M \cap I| - r \cdot \frac{1}{|I|} \cdot |M \cap I| \cdot |N \cap I|}{|N \cup M|}. \tag{1}$$

where $\mathbf{I}$ is the set of all images. Normalized ImRoU_r is obtained by dividing the above expression by the maximally achievable score for a given neuron mask.[3] We write ImRoU_r to refer to normalized ImRoU_r throughout the paper.

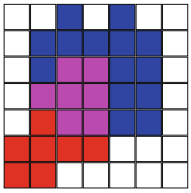

Fig. 2. An example illustrating the concept of random intersection. A neuron mask (red) and label mask (blue) intersect at 7 cells (purple). As the neuron mask covers 14 cells, choosing 21 cells in this grid at random would lead to an expected intersection of 6 cells, which means that the label mask achieves an intersection of 7 against a random intersection of 6. With $r = 0.75$, this leads to a value of 2.5 (as supposed to 6) in the respective summand in the nominator of (1). (Color figure online)

Choosing a value for r is non-trivial. With $r = 1$, every scene-level annotation achieves a score of 0 regardless of neuron behavior as the real intersection is always equal to the random intersection, which is intuitively undesirable. Thus,

[3] If the neuron mask is N, this can be computed as $(|N| - (r/|I|) \sum_{I \in \mathbf{I}} |N \cap I|^2)/|N|$.

we have used $r = .75$ for all results reported on in this paper. Figure 3 compares masks found by the Close Algorithm optimizing for $\text{ImRoU}_{.75}$ with masks found by the standard algorithm optimizing for IoU.

5 Results

5.1 Scene-Level Annotations

When the standard algorithm is optimized for IoU, it finds $(52, 146, 196, 118)$ formulas using $(0, 1, 2, 3)$ scene-level annotations, whereas the Close Algorithm finds $(120, 197, 136, 59)$. When $\text{ImRoU}_{.75}$ is optimized for instead, the numbers change to $(331, 153, 23, 5)$ and $(398, 100, 14, 0)$, respectively. While it will come as no surprise that $\text{ImRoU}_{.75}$ discourages scene-level annotations, it is also worth noting that the standard algorithm uses them more than the Close Algorithm. As shown in the upcoming Section, the Close Algorithm improves upon the accuracy of the standard algorithm according to either metric. Thus, these results may indicate that improving approximation quality will disincentivize scene-level annotations naturally.

5.2 Scores

Table 1 provides a quantitative comparison between the Close- and standard algorithm.[4] While the Close Algorithm does better, its absolute accuracy remains poor. E.g., if the neuron and label masks have the same coverage, an IoU score of 0.1 means that more than 80% of the positive predictions made by the label masks are inaccurate.

Table 1. IoU and $\text{ImRoU}_{.75}$ scores of the standard vs. Close Algorithm for formula lengths 3 and 10. Each cell shows mean/median scores in % for the respective setting.

	IoU		$\text{ImRoU}_{.75}$	
	FL3	**FL10**	**FL3**	**FL10**
Standard	8.4/7.5	9.9/9.2	6.1/5.2	7.2/6.5
Close	9.3/8.4	11.3/10.4	7.3/6.4	8.9/7.9

We can identify at least two different hypotheses to explain this:[5]

[4] Results for individual neurons differ from those in [17] because we use downsampling to make annotations comparable (see Sect. 3.1), but the difference is mild and does not systematically skew in either direction. At formula length 10, the standard algorithm achieves a mean IoU of 0.099 (rounded to 3 decimal places) in both cases.

[5] In [17], Mu and Andreas find diminishing returns of predictive performance from raising the formula length beyond 10, which is evidence against the third hypothesis that neurons frequently combine more than ten human concepts.

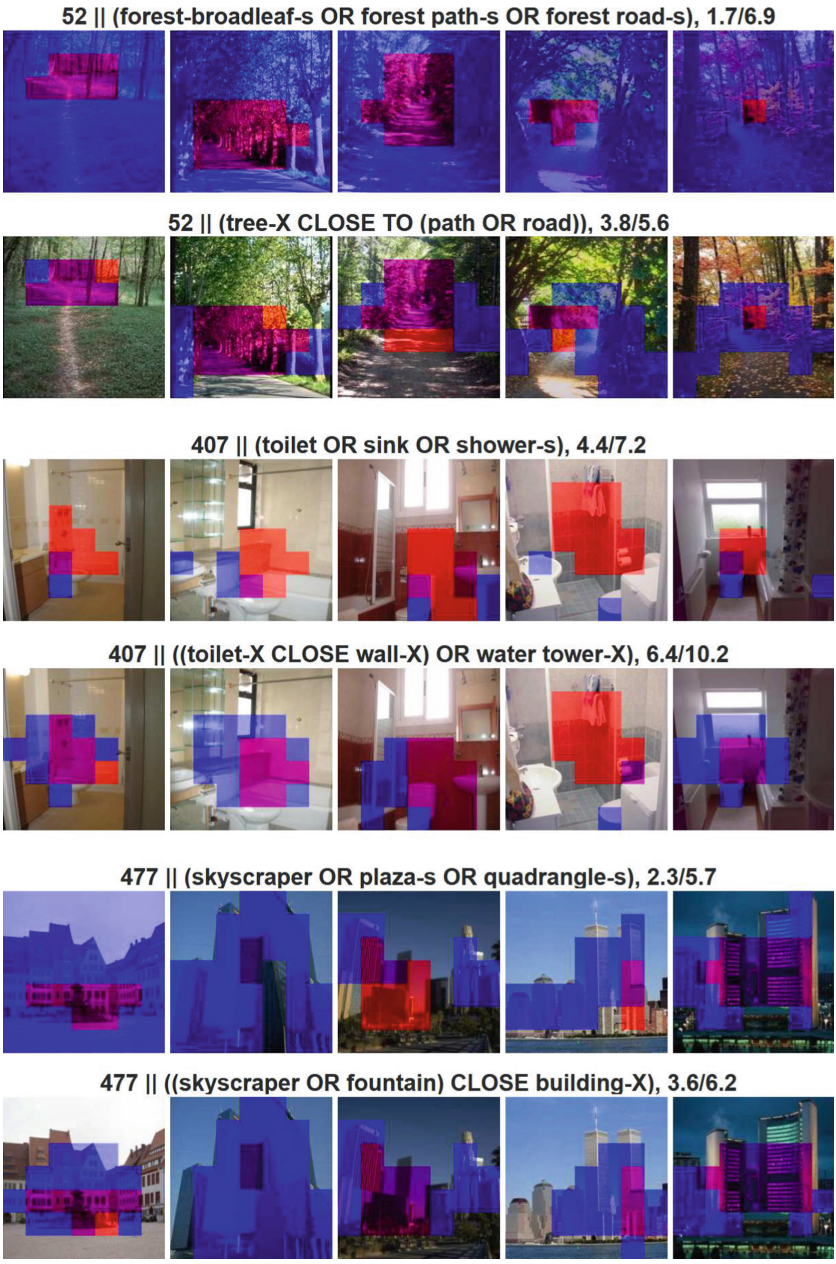

Fig. 3. Examples of masks found by the standard algorithm optimizing for IoU (first row) vs. masks found by the Close Algorithm optimizing for $\text{ImRoU}_{.75}$ (second row), at formula length 3. Here, the neuron masks are shown in red and the masks found by the algorithm in blue; intersected areas are purple. Scene-level concepts have the postfix -s. The numbers at the end of each line denote $\text{ImRoU}_{.75}$ and IoU scores in %. The three neurons shown here have been selected to be illustrative rather than at random. (Color figure online)

Hypothesis (1): neurons combine human concepts in ways that are not represented by AND, OR, AND NOT, CLOSE-TO, WITH, and EXPAND, but could be represented by different connectors.

Hypothesis (2): neurons activate on concepts that are not annotated in the Broden dataset, such as the edges of objects.

Differentiating between these hypotheses is important to guide future research as, e.g., improving scores by adding new connectors is likely to be possible insofar as problems are due to (1) rather than (2). However, the hypotheses are not mutually exclusive.

6 User Study

We have conducted a small user study in which two participants take the role of the algorithm in predicting image activations. To the extent that they outperform the algorithm, their descriptions for their neurons are evidence on the distinction between hypotheses (1) and (2) mentioned in Sect. 5.

6.1 Study Design

The two participants were selected among people who have completed a course in computer science or statistics and did well on a pilot study with a similar task. Each participant was given access to a large set of images and ground-truth neuron masks through a simple web-based application. The application allows hand-drawing masks and displays the $\text{ImRoU}_{.75}$ score averaged across those masks. After training for 40 min, participants were asked to test their understanding by drawing masks for the neuron for a disjoint random sample of 60 images. Each participant had access to the algorithmically computed label masks (but not the formula) for one of the two neurons. After the task, participants were asked to provide a brief description of their neuron (up to 2 sentences) and to describe interesting patterns they noticed (up to 10 sentences).

As mentioned in Sect. 4, neurons typically do not activate at all on over 95% of images. Furthermore, over 55% of the remaining images are still approximately false positives as the neurons only activate on 1–3 cells. For this reason, we considered it acceptable to only include images with nonzero activations, as otherwise, the task becomes tedious for the participants. The Close Algorithm was retrained on this altered data set, achieving mean and median $\text{ImRoU}_{.75}$ scores of .158 and .152, respectively (up from .089 and .079). Image subsets shown to annotators have been sampled uniformly and subsequently pruned to be disjoint for each neuron.

6.2 Results

Participants outperformed the algorithm on all four neurons. Table 2 shows the quantitative results; Table 3 contrasts the short description given by the human with the formulas found by the Close Algorithm. The participants completed

neurons 353 and 329 first, respectively. Access to label masks had a slight negative correlation with relative performance (mean performance relative to the algorithm on neurons with/without access was 1.261 and 1.277, respectively).

Table 2. $\text{ImRoU}_{.75}$ scores achieved by human annotators ("Human Score") vs. the Close Algorithm ("CA Score") on the test set (60 images, uniformly sampled), where "Ratio" denotes $\frac{\text{Human Score}}{\text{CA Score}}$

Neuron	Human score	CA score	Ratio	Access
69	0.240	0.170	1.412	No
154	0.213	0.187	1.142	Yes
329	0.072	0.063	1.142	No
353	0.220	0.159	1.379	Yes

6.3 Discussion

The descriptions given by the participants include evidence for both hypotheses – in fact, this is already true for just neuron 69. "[T]he front left pocket of a pool table" could be algorithmically annotated using a new LEFT-OF connector, whereas "rectangles on the face of things" would require a new ground-truth annotation of shapes rather than objects. Other descriptions in the latter category include "potted plants in the forefront of images" (this would require differentiating between foreground and background), "the area where you can sit and reach a table", and "where you would sit on a bed". Conversely, "vertically hung clothes" could in principle be found algorithmically, though this would be difficult.

To get a qualitative sense of this task, consider the following response to the "describe interesting patterns" question for neuron 353:

> At first I thought the filter was just selecting tables and chairs (and that seems to be what the label mask is filtering on), but there was definitely a tendency to only pick the areas that were within reach of a surface, so a chair alone may not necessarily be highlighted, but a chair with an end table would definitely be, but just the area above the chair, and the surface within reach. For something like a couch or a bed, it would only highlight the side by the table. As the images were more complex, the area highlighted tended to be smaller, so a table with 12 chairs that filled the image would have a smaller proportion highlighted than a small table with a single chair. It also tended to select the side closest to the camera, I think, though I just realized that was what it was doing after the case, and it may not really be the case, but there was a bias to a specific side.

Insofar as these observations reflect true behaviors of the neuron, the response may shed further light on why the task remains challenging to do algorithmically. Determining that a neuron cares about a particular object may fail to

Table 3. Descriptions given by human annotators vs. formulas found by the Close Algorithm (trained on the subset of only images with nonzero activations)

N#	Human description	Close algorithm formula
69	Dark windows/doors/corridors/pits; windows or rectangles on the face of things (e.g. bulletin boards/lattices); the top/middle of the first cabinet on the right, the front left pocket of a pool table	(((door-X OR pool table OR house) AND NOT kitchen-s) OR drawer OR elevator door OR elevator OR telephone booth OR hovel OR arcade machine)
154	Snowy mountains, greenhouses (especially ceilings)	((((((mountain OR ice) CLOSE TO sky-X) OR greenhouse OR tent OR iceberg OR canopy) AND NOT greenhouse-indoor-s) OR truck) AND NOT wall-X)
329	Bedsheets, curtains, shelves, people's center of mass/chest area, rough stone structures including sculptures, pool tables, potted plants in the forefront of images	(bed-X OR rock OR person OR apparel OR sofa-X OR shirt-X OR armchair OR cliff OR viaduct-X OR aqueduct-X)
353	The area where you can sit and reach a table, and the area above that table, including chairs, couches, toilets, and where you would sit on a bed. Also, vertically hung clothes and occasionally organizers with books/clothes	(table-X OR cradle-X OR chair-X OR shirt-X OR apparel-X OR pillow-X OR jacket-X OR cushion-X OR back-X OR back pillow-X)

translate into a good score if the solution misses out on subtleties of this kind. The CLOSE-TO connector can plausibly help with this, but it often remains a crude approximation, e.g., it cannot determine whether a table is within reach of a chair since the spatial distance in the scene is not strictly proportional to the cell-based distance in the image. In fact, the Close Algorithm did not choose any non-logical connectors for neuron 353 other than EXPAND, proving that the CLOSE-TO connector is not helpful for predicting this particular neuron.

One may argue that the existing metrics are unfairly harsh as they fail to take proximity into account: if the label mask predicts an incorrect cell, it makes no difference whether the predicted cell is adjacent to a cluster of correct cells or in an entirely different image. Unfortunately, a metric that cares about distance is computationally expensive, making this problem difficult to address.

Due to the small sample size, these results do not permit a more quantitative analysis. However, future experiments including more participants may make this possible.

Finally, these results show that (a) it is possible for humans to outperform the algorithm on this task, and (b), there is a substantial similarity between human descriptions and algorithmic formulas.[6] This is worth pointing out as there is precedent of interpretability tools failing comparable "sanity checks" [1].

[6] For all four neurons, there has been an overlap between the set of objects picked by the algorithm and the human description. E.g., doors, mountains, beds, tables.

References

1. Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., Kim, B.: Sanity checks for saliency maps. In: Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (eds.) Advances in Neural Information Processing Systems, vol. 31. Curran Associates, Inc. (2018). https://proceedings.neurips.cc/paper/2018/file/294a8ed24b1ad22ec2e7efea049b8737-Paper.pdf
2. Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.R., Samek, W.: On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PLOS ONE **10**(7), 1–46 (2015). https://doi.org/10.1371/journal.pone.0130140
3. Bau, A., Belinkov, Y., Sajjad, H., Durrani, N., Dalvi, F., Glass, J.: Identifying and controlling important neurons in neural machine translation. In: International Conference on Learning Representations (2019). https://openreview.net/forum?id=H1z-PsR5KX
4. Bau, D., Zhu, J.Y., Strobelt, H., Lapedriza, A., Zhou, B., Torralba, A.: Understanding the role of individual units in a deep neural network. In: Proceedings of the National Academy of Sciences (2020). https://doi.org/10.1073/pnas.1907375117. https://www.pnas.org/content/early/2020/08/31/1907375117
5. Camburu, O.M., Rocktäschel, T., Lukasiewicz, T., Blunsom, P.: e-SNLI: natural language inference with natural language explanations. In: Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (eds.) Advances in Neural Information Processing Systems, vol. 31. Curran Associates, Inc. (2018). https://proceedings.neurips.cc/paper/2018/file/4c7a167bb329bd92580a99ce422d6fa6-Paper.pdf
6. Chen, C., Li, O., Tao, C., Barnett, A.J., Su, J., Rudin, C.: This Looks like That: Deep Learning for Interpretable Image Recognition. Curran Associates Inc., Red Hook (2019)
7. Chen, X., Mottaghi, R., Liu, X., Fidler, S., Urtasun, R., Yuille, A.: Detect what you can: detecting and representing objects using holistic models and body parts. In: Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, pp. 1979–1986. IEEE Computer Society, USA (2014). https://doi.org/10.1109/CVPR.2014.254
8. Dalvi, F., et al.: NeuroX: a toolkit for analyzing individual neurons in neural networks. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 9851–9852, July 2019. https://doi.org/10.1609/aaai.v33i01.33019851
9. Durrani, N., Sajjad, H., Dalvi, F., Belinkov, Y.: Analyzing individual neurons in pre-trained language models. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 4865–4880. Association for Computational Linguistics, Online, November 2020. https://doi.org/10.18653/v1/2020.emnlp-main.395. https://aclanthology.org/2020.emnlp-main.395
10. Fong, R., Vedaldi, A.: Net2Vec: quantifying and explaining how concepts are encoded by filters in deep neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2018
11. 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 (2017). https://doi.org/10.1109/ICCV.2017.371
12. Gonzalez-Garcia, A., Modolo, D., Ferrari, V.: Do semantic parts emerge in convolutional neural networks? Int. J. Comput. Vision **126**, 476–494 (2017)

13. Hase, P., Zhang, S., Xie, H., Bansal, M.: Leakage-adjusted simulatability: can models generate non-trivial explanations of their behavior in natural language? In: Findings of the Association for Computational Linguistics: EMNLP 2020, pp. 4351–4367. Association for Computational Linguistics, Online, November 2020. https://doi.org/10.18653/v1/2020.findings-emnlp.390. https://aclanthology.org/2020.findings-emnlp.390
14. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778 (2016). https://doi.org/10.1109/CVPR.2016.90
15. Kim, J., Rohrbach, A., Darrell, T., Canny, J., Akata, Z.: Textual explanations for self-driving vehicles. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11206, pp. 577–593. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01216-8_35
16. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: Guyon, I., et al. (eds.) Advances in Neural Information Processing Systems, vol. 30. Curran Associates, Inc. (2017). https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf
17. 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
18. Olah, C., et al.: The building blocks of interpretability. Distill **3** (2018). https://doi.org/10.23915/distill.00010
19. Pruthi, G., Liu, F., Kale, S., Sundararajan, M.: Estimating training data influence by tracing gradient descent. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M.F., Lin, H. (eds.) Advances in Neural Information Processing Systems, vol. 33, pp. 19920–19930. Curran Associates, Inc. (2020). https://proceedings.neurips.cc/paper/2020/file/e6385d39ec9394f2f3a354d9d2b88eec-Paper.pdf
20. Ribeiro, M., Singh, S., Guestrin, C.: "Why should I trust you?": explaining the predictions of any classifier. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations, pp. 97–101. Association for Computational Linguistics, San Diego, June 2016. https://doi.org/10.18653/v1/N16-3020. https://aclanthology.org/N16-3020
21. Ribeiro, M.T., Singh, S., Guestrin, C.: Anchors: high-precision model-agnostic explanations. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32, no. 1, April 2018. https://ojs.aaai.org/index.php/AAAI/article/view/11491
22. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-CAM: visual explanations from deep networks via gradient-based localization. In: 2017 IEEE International Conference on Computer Vision (ICCV), pp. 618–626 (2017). https://doi.org/10.1109/ICCV.2017.74
23. Zhou, B., Bau, D., Oliva, A., Torralba, A.: Interpreting deep visual representations via network dissection. IEEE Trans. Pattern Anal. Mach. Intell. **41**(9), 2131–2145 (2019). https://doi.org/10.1109/TPAMI.2018.2858759
24. Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., Torralba, A.: Places: a 10 million image database for scene recognition. IEEE Trans. Pattern Anal. Mach. Intell. **40**, 1452–1464 (2017)
25. Zhou, B., et al.: Semantic understanding of scenes through the ADE20K dataset. Int. J. Comput. Vision **127**(3), 302–321 (2018). https://doi.org/10.1007/s11263-018-1140-0