Thanks to all for thoughtful and helpful comments, and positive feedback! Reviewers agree we have proposed a "simple and general" (**R1**) yet "imaginative and thoughtful" (**R4**) method which tackles an "extremely important" problem (**R4**) and produces significant, interesting (**R2**, **R3**), and "thought-initiat[ing]" (**R4**) results. We now address specific points:

R1: Our method requires (hand-labeled) semantic annotations. Not necessarily: in the natural language inference (NLI) experiments, we use a pretrained model to label the probe dataset with part-of-speech tags. One could also use a semantic segmentation network to generate visual concepts. Of course, these models must be trained on annotated data—at some point, any procedure for labeling neurons with concepts requires some starting source of labeled concepts.

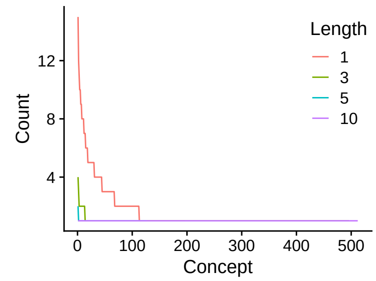

Figure S1: Number of unique concepts based on length.

R1: Heuristic generation of concepts is a limitation and hinders reproducibility. We agree that the choice of primitive concepts and compositions highly influence the discovered concepts, and that beam search only approximates the preferable, but intractable, enumeration over logical forms. However, we disagree that this hinders the *reproducibility* of our results; as **R2** notes, we have tried to precisely specify our set of inputs, concepts, and models, especially for NLI. We will also release code with the camera-ready paper, which should ease reproducibility concerns (cc **R2**).

R1: We may not find a strong relationship between interpretability and accuracy if we don't generate the right explanations. This is true, because "interpretability" is defined by the space of concepts we specify. If we find no relationship between interpretability and accuracy in a model for an explored set of concepts, it could be that we are simply using the wrong definition of "interpretability", and that alternative concept spaces could lead to more informative results. We show two tasks where we discover interpretable concepts with a noisy, but still highly significant, correlation with accuracy.

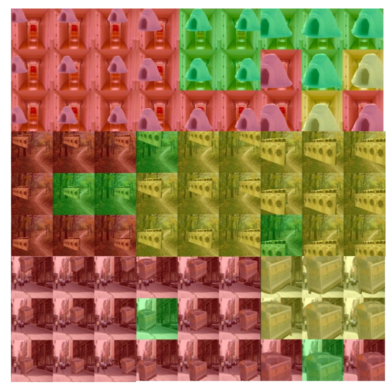

Figure S2: Varying the size and position of sub-images. Green: prediction changes to intended adversarial class; yellow: prediction changes to a different class (e.g. *aqueduct* for the middle row); red = no change.

R4: NLI isn't the best task to explore semantics, since NLI datasets are poorly built. Indeed, this is precisely why we chose NLI: since previous work has shown that NLI models learn non-robust, shallow heuristics, our experiments explore how these strategies are implemented in individual neurons.

R2: Do neurons identify similar compositional concepts? What about different models? Great questions! Figure S1 plots the counts of each concept across the 512 units of ResNet-18, by length. At length 1 (NetDissect), many concepts appear multiple times; the mean number of occurrences per concept is 2.61 (42% of concepts are unique). Uniqueness increases dramatically by length 3 (mean 1.03; 97% unique), 5 (1.01; 99%), and 10 (1.00; 100%). Our explanations thus reveal significant specialization in neuron function (vs. NetDissect). Table S1 shows some repeated concepts. We will add this to the supplement and analyze NLI as well (omitted here for space). We leave the question of different vision models for future work; our code will facilitate the necessary experiments. The adversarial examples in Figure 8 hint that some concepts (e.g. *non-blue water*) are shared across models.

R3: How sensitive are copy-paste examples to size and position? In Figure S2 we vary the size and position of subimages for the copy-paste examples (note this analysis is less straightforward for examples like *non-blue water*). Sensitivity depends on the specific example. In general, if the sub-image is too small (left), the original class prevails; otherwise, the *igloo* $\to$ *clean room* example is quite reliable, while the *street* $\to$ *fire escape* example is less so. We will add this to the supplement.

Table S1: Most common concepts by length N

N	Concept	#
1	pool table	15
	house	12
	corridor	11
3	pillow OR (bed AND bedroom)	4
	sink OR toilet OR bathtub	3
	water OR river AND (NOT blue)	2
5	auditorium OR theater OR conference center...	2

R1: Why not object detection? As noted by the original NetDissect work, for networks explicitly trained on object detection tasks, it would be less surprising that neurons specialize for object detection. This motivates us to explore a scene recognition network, and see whether or not interpretable object-level concepts emerge without explicit object-level supervision. Still, probing object detection models is an interesting and straightforward extension of our method.