Compositional Explanations of Neurons

Jesse Mu  
Stanford University  
muj@stanford.edu

Jacob Andreas  
MIT CSAIL  
jda@mit.edu

Abstract

We describe a procedure for explaining neurons in deep representations by identifying compositional logical concepts that closely approximate neuron behavior. Compared to prior work that uses atomic labels as explanations, analyzing neurons compositionally allows us to more precisely and expressively characterize their behavior. We use this procedure to answer several questions on interpretability in models for vision and natural language processing. First, we examine the kinds of abstractions learned by neurons. In image classification, we find that many neurons learn highly abstract but semantically coherent visual concepts, while other polysematic neurons detect multiple unrelated features; in natural language inference (NLI), neurons learn shallow lexical heuristics from dataset biases. Second, we see whether compositional explanations give us insight into model performance: vision neurons that detect human-interpretable concepts are positively correlated with task performance, while NLI neurons that fire for shallow heuristics are negatively correlated with task performance. Finally, we show how compositional explanations provide an accessible way for end users to produce simple “copy-paste” adversarial examples that change model behavior in predictable ways.

1 Introduction

In this paper, we describe a procedure for automatically explaining logical and perceptual abstractions encoded by individual neurons in deep networks. Prior work in neural network interpretability has found that neurons in models trained for a variety of tasks learn human-interpretable concepts, e.g. faces or parts-of-speech, often without explicit supervision [5, 10, 11, 27]. Yet many existing interpretability methods are limited to ad-hoc explanations based on manual inspection of model visualizations or inputs [10, 26, 27, 35, 38, 39]. To instead automate explanation generation, recent work [5, 11] has proposed to use labeled “probing datasets” to explain neurons by identifying concepts (e.g. dog or verb) closely aligned with neuron behavior.

However, the atomic concepts available in probing datasets may be overly simplistic explanations of neurons. A neuron might robustly respond to images of dogs without being exclusively specialized for dog detection; indeed, some have noted the presence of polysemantic neurons in vision models that detect multiple concepts [12, 27]. The extent to which these neurons have learned meaningful perceptual abstractions (versus detecting unrelated concepts) remains an open question. More generally, neurons may be more accurately characterized not just as simple detectors, but rather as operationalizing complex decision rules composed of multiple concepts (e.g. dog faces, cat bodies, and car windows). Existing tools are unable to surface such compositional concepts automatically.

We propose to generate explanations by searching for logical forms defined by a set of composition operators over primitive concepts (Figure 1). Compared to previous work [5], these explanations serve as better approximations of neuron behavior, and identify behaviors that help us answer a variety of interpretability questions across vision and natural language processing (NLP) models. First, what kind of logical concepts are learned by deep models in vision and NLP? Second, do the
quality and interpretability of these learned concepts relate to model performance? Third, can we use the logical concepts encoded by neurons to control model behavior in predictable ways? We find that:

1. Neurons learn compositional concepts: in **image classification**, we identify neurons that learn meaningful perceptual abstractions (e.g. *tall structures*) and others that fire for unrelated concepts. In natural language inference (**NLI**), we show that shallow heuristics (based on e.g. gender and lexical overlap) are not only learned, but refined in individual neurons.

2. Compositional explanations help predict model accuracy, but interpretability is not always associated with accurate classification: in **image classification**, human-interpretable abstractions are **correlated** with model performance, but in **NLI**, neurons that reflect shallower heuristics are **anticorrelated** with performance.

3. Compositional explanations allow users to predictably manipulate model behavior: we can generate crude “copy-paste” adversarial examples based on inserting words and image patches to target individual neurons, in contrast to black-box approaches [136, 37].

## 2 Generating compositional explanations

Consider a neural network model $f$ that maps inputs $x \in \mathbb{R}^d$. $f$ might be a prefix of a convolutional network trained for image classification or a sentence embedding model trained for a language processing task. Now consider an individual neuron $f_n(x) \in \mathbb{R}$ and its activation on a set of concrete inputs (e.g. ResNet-18 [15] layer 4 unit 483; Figure 1a–b). How might we explain this neuron’s behavior in human-understandable terms?

The intuition underlying our approach is shared with the NetDissect procedure of Bau et al. [5]; here we describe a generalized version. The core of this intuition is that a good explanation is a description (e.g. a named category or property) that identifies the same inputs for which $f_n$ activates. Formally, assume we have a space of pre-defined atomic concepts $C \in \mathcal{C}$ where each concept is a function $C : \mathbb{R} \to \{0, 1\}$ indicating whether $x$ is an instance of $C$. For image pixels, concepts are image segmentation masks; for the *water* concept, $C(x)$ is 1 when $x$ is an image region containing water (Figure 1c). Given some measure $\delta$ of the similarity between neuron activations and concepts, NetDissect explains the neuron $f_n$ by searching for the concept $C$ that is most similar:

$$\text{EXPLAIN-NETDISSECT}(n) = \arg \max_{C \in \mathcal{C}} \delta(n, C).$$

While $\delta$ can be arbitrary, Bau et al. [5] first threshold the continuous neuron activations $f_n(x)$ into binary masks $M_n(x) \in \{0, 1\}$ (Figure 1b). This can be done *a priori* (e.g. for post-ReLU activations, thresholding above 0), or by dynamically thresholding above a neuron-specific percentile. We can then compare binary neuron masks and concepts with the Intersection over Union score (IoU, or Jaccard similarity; Figure 1f):

$$\delta(n, C) \triangleq \text{IoU}(n, C) = \frac{\sum_x \mathbb{I}(M_n(x) \land C(x))}{\left[ \sum_x \mathbb{I}(M_n(x) \lor C(x)) \right]}.$$

---

**Figure 1**: Given a set of inputs (a) and scalar neuron activations (b) converted into binary masks (c), we generate an explanation via beam search, starting with an inventory of primitive concepts (d), then incrementally building up more complex logical forms (e). We attempt to maximize the IoU score of an explanation (f); depicted is the IoU of $M_{483}(x)$ and (water OR river) AND NOT blue.
Compositional search. The procedure described in Equation 1 can only produce explanations from the fixed, pre-defined concept inventory \( C \). Our main contribution is to combinatorially expand the set of possible explanations to include logical forms \( \mathcal{L}(C) \) defined inductively over \( C \) via composition operations such as disjunction (OR), conjunction (AND), and negation (NOT), e.g. \( (L_1 \land L_2)(x) = L_1(x) \land L_2(x) \) (Figure 1f). Formally, if \( \Omega_\eta \) is the set of \( \eta \)-ary composition functions, define \( \mathcal{L}(C) \):  

1. Every primitive concept is a logical form: \( \forall C \in C \), we have \( C \in \mathcal{L}(C) \).
2. Any composition of logical forms is a logical form: \( \forall \eta \), \( \omega \in \Omega_\eta \), \( (L_1, \ldots, L_\eta) \in \mathcal{L}(C)^\eta \), where \( \mathcal{L}(C)^\eta \) is the set of \( \eta \)-tuples of logical forms in \( \mathcal{L}(C) \), we have \( \omega(L_1, \ldots, L_\eta) \in \mathcal{L}(C) \).

Now we search for the best logical form \( L \in \mathcal{L}(C) \):

\[
\text{EXPLAIN-COMP}(n) = \arg \max_{L \in \mathcal{L}(C)} \text{IoU}(n, L). \tag{3}
\]

The \( \arg \max \) in Equation 3 ranges over a structured space of compositional expressions, and has the form of an inductive program synthesis problem [23]. Since we cannot exhaustively search \( \mathcal{L}(C) \), in practice we limit ourselves to formulas of maximum length \( N \), by iteratively constructing formulas from primitives via beam search with beam size \( B = 10 \). At each step of beam search, we take the formulas already present in our beam, compose them with new primitives, measure IoU of these new formulas, and keep the top \( B \) new formulas by IoU, as shown in Figure 1f.

3 Tasks

The procedure we have described above is model- and task-agnostic. We apply it to two tasks in vision and NLP: first, we investigate a scene recognition task explored by the original NetDissect work [5], which allows us to examine compositionality in a task where neuron behavior is known to be reasonably well-characterized by atomic labels. Second, we examine natural language inference (NLI): an example of a seemingly challenging NLP task which has recently come under scrutiny due to models’ reliance on shallow heuristics and dataset biases [13, 14, 22, 25, 30, 37]. We aim to see whether compositional explanations can uncover such undesirable behaviors in NLI models.

Image Classification. NetDissect [5] examines whether a convolutional neural network trained on a scene recognition task has learned detectors that correspond to meaningful abstractions of objects. We take the final 512-unit convolutional layer of a ResNet-18 [15] trained on the Places365 dataset [40], probing for concepts in the ADE20k scenes dataset [41] with atomic concepts \( C \) defined by annotations in the Broden dataset [5]. There are 1105 unique concepts in ADE20k, categorized by Scene, Object, Part, and Color (see Figure 2 for examples).

Broden has pixel-level annotations, so for each input image \( X \in \mathbb{R}^{H \times W} \), inputs are indexed by pixels \((i, j) : x_{i,j} \in \mathcal{X} \). Let \( f_n(x_{i,j}) \) be the activation of the \( n \)th neuron at position \((i, j)\) of the image \( X \), after the neuron’s activation map has been bilinearly upsampled from layer dimensions \( H_l \times W_l \) to the segmentation mask dimensions \( H \times W \). Following [5], we create neuron masks \( M_\eta(x) \) via dynamic thresholding: let \( T_\eta \) be the threshold such that \( P(f_n(x) > T_\eta) = 0.005 \) over all inputs \( x \in \mathcal{X} \). Then \( M_\eta(x) = 1(f_n(x) > T_\eta) \). For composition, we use operations \( \land \) (AND), \( \lor \) (OR), and \( \lnot \) (NOT), leaving more complex operations (e.g. relations like ABOVE and BELOW) for future work.

NLI. Given premise and hypothesis sentences, the task of NLI is to determine whether the premise entails the hypothesis, contradicts it, or neither (neutral). We investigate a BiLSTM baseline architecture proposed by [7]. A bidirectional RNN encodes both the premise and hypothesis to form 512-d representations. Both representations, and their elementwise product and difference, are then concatenated to form a 2048-d representation that is fed through a multilayer perceptron (MLP) with two 1024-d layers with ReLU nonlinearities and a final softmax layer. This model is trained on the Stanford Natural Language Inference (SNLI) corpus [6] which consists of 570K sentence pairs.

Neuron-level explanations of NLP models have traditionally analyzed how RNN hidden states detect word-level features as the model passes over the input sequence [3, 10], but in most NLI models, these
RNN features are learned early and are often quite distant from the final sentence representation used for prediction. Instead, we analyze the MLP component, probing the 1024 neurons of the penultimate hidden layer for sentence-level explanations, so our inputs x are premise-hypothesis pairs.

We use the SNLI validation set as our probing dataset (10K examples). As our features, we take the Penn Treebank part of speech tags (labeled by SpaCy\(^\text{1}\)) and the 2000 most common words appearing in the dataset. For each of these we create 2 concepts that indicate whether the word or part-of-speech appears in the premise or hypothesis. Additionally, to detect whether models are using lexical overlap heuristics,\(^\text{25}\), we define 4 concepts indicating that the premise and hypothesis have more than 0%, 25%, 50%, or 75% overlap, as measured by IoU between the unique words.

For our composition operators, we keep AND, OR, and NOT; in addition, to capture the idea that neurons might fire for groups of words with similar meanings, we introduce the unary NEIGHBORS operator. Given a word feature C, let the neighborhood \(N(C)\) be the set of 5 closest words \(C'\) to \(C\), as measured by their cosine distance in GloVe embedding space.\(^\text{1}\) Then, \(\text{NEIGHBORS}(C)(x) = \bigvee_{C' \in N(C)} C'(x)\) (i.e. the logical OR across all neighbors). Finally, since these are post-ReLU activations, instead of dynamically thresholding we simply define our neuron masks \(M_n(x) = \mathbb{1}(f_x(x) > 0)\). There are many “dead” neurons in the model, and some neurons fire more often than others; we limit our analysis to neurons that activate reliably across the dataset, defined as being active at least 500 times (5%) across the 10K examples probed.

4 Do neurons learn compositional concepts?

Image Classification. Figure 3(left) plots the distribution of IoU scores for the best concepts found for each neuron as we increase the maximum formula length \(N\). When \(N = 1\), we get EXPLAIN-NETDISSECT, with a mean IoU of 0.059; as \(N\) increases, IoU increases up to 0.099 at \(N = 10\), a statistically significant 68% increase \((p = 2 \times 10^{-9})\). We see diminishing returns after length 10, so we conduct the rest of our analysis with length 10 logical forms. The increased explanation quality suggests that our compositional explanations indeed detect behavior beyond simple atomic labels: Figure 4 shows an example of a bullring detector which is actually revealed to detect fields in general.

We can now answer our first question from the introduction: are neurons learning meaningful abstractions, or firing for unrelated concepts? Both happen: we manually inspected a random sample of 128 neurons in the network and their length 10 explanations, and found that 69% learned some meaningful combination of concepts, while 31% were polysemantic, firing for at least some unrelated concepts. The 88 “meaningful” neurons fell into 3 categories (examples in Figure 5; more in Appendix A):

1. 50 (57%) learn a perceptual abstraction that is also lexically coherent, in that the primitive words in the explanation are semantically related (e.g. to towers or bathrooms; Figure 5a).
2. 28 (32%) learn a perceptual abstraction that is not lexically coherent, as the primitives are not obviously semantically related. For example, cradle OR autobus OR fire escape is a vertical rails detector, but we have no annotations of vertical rails in Broden (Figure 5b).
3. 10 (12%) have the form \(L_1\) AND NOT \(L_2\), which we call specialization. They detect more specific variants of Broden concepts (e.g. (water OR river) AND NOT blue; Figure 5c).

\(\text{https://spacy.io/}\)
The observation that IoU scores do not increase substantially past length 10 corroborates the finding of [12], who also note that few neurons detect more than 10 unique concepts in a model. Our procedure, however, allows us to more precisely characterize whether these neurons detect abstractions or unrelated disjunctions of concepts, and identify more interesting cases of behavior (e.g., specialization). While composition of Broden annotations explains a majority of the abstractions learned, there is still considerable unexplained behavior. The remaining behavior could be due to noisy activations, neuron misclassifications, or detection of concepts absent from Broden.
5 Do interpretable neurons contribute to model accuracy?

A natural question to ask is whether it is empirically desirable to have more (or less) interpretable neurons, with respect to the kinds of concepts identified above. To answer this, we measure the performance of the entire model on the task of interest when the neuron is activated. In other words, for neuron \( n \), what is the model accuracy on predictions for inputs where \( M_n(x) = 1 \)? In image classification, we find that the more interpretable the neuron (by IoU), the more accurate the model is when the neuron is active (Figure 7 left; \( r = 0.31, p < 1e-13 \)); the correlation increases as the formula length increases and we are better able to explain neuron behavior. Given that we are measuring abstractions over the human-annotated features deemed relevant for scene classification, this suggests, perhaps unsurprisingly, that neurons that detect more interpretable concepts are more accurate.

However, when we apply the same analysis to the NLI model, the opposite trend occurs: neurons that we are better able to explain are less accurate (Figure 7 right; \( r = -0.60, p < 1e-08 \)). Unlike vision, most sentence-level logical descriptions recoverable by our approach are spurious by definition, as they are too simple compared to the true reasoning required for NLI. If a neuron can be accurately summarized by simple deterministic rules, this suggests the neuron is making decisions based on spurious correlations, which is reflected by the lower performance. Analogously, the more restricted our feature set (by maximum formula length), the better we capture this anticorrelation. One important takeaway is that the “interpretability” of these explanations is not \( a \ priori \) correlated with performance, but rather dependent on the concepts we are searching for: given the right concept space, our method can identify behaviors that may be correlated or anticorrelated with task performance.

6 Can we target explanations to change model behavior?

Finally, we see whether compositional explanations allow us to manipulate model behavior. In both models, we have probed the final hidden representation before a final softmax layer produces the class predictions. Thus, we can measure a neuron’s contribution to a specific class with the weight between the neuron and the class, and see whether constructing examples that activate (or inhibit) these neurons leads to corresponding changes in predictions. We call these “copy-paste” adversarial examples to differentiate them from standard adversarial examples involving imperceptible perturbations \[36\].

Image Classification. Figure 8 shows some Places365 classes along with the neurons that most contribute to the class as measured by the connection weight. In many cases, these connections are
A blond woman is holding 2 golf balls while reaching down into a golf hole.

The consistency of this misclassification suggests that models are detecting underlying biases in the training data. Other examples include a neuron contributing to clean room that also detects ice and igloos; putting an igloo in a corridor causes a prediction to shift from corridor to clean room, though this does not occur across models, suggesting that this is an artifact specific to this model.
NLIs. In NLI, we are able to trigger similar behavior by targeting spurious neurons (Figure 9): a neuron that detects the presence of nobody in the hypothesis predicts contradiction for nearly all such inputs; another neuron similarly predicts contradiction when couch appears in the hypothesis. Overall, these examples are reminiscent of the image-patch attacks of [9], adversarial NLI inputs [1, 37], and the data collection process for recent counterfactual NLI datasets [13, 22], but instead of searching among neuron visualizations or using black-box optimization to synthesize examples, we use explanations as a transparent guide for crafting perturbations by hand.

7 Related Work

Interpretability. Interpretability in deep neural networks has received considerable attention over the past few years. Our work extends existing work on generating explanations for individual neurons in deep representations [4, 5, 10–12, 27], in contrast to analysis or probing methods that operate at the level of entire representations (e.g. [2, 19, 29]). Neuron-level explanations are fundamentally limited, since they cannot detect concepts distributed across multiple neurons, but this has advantages: first, neuron-aligned concepts offer evidence for representations that are disentangled with respect to concepts of interest; second, they inspect unmodified “surface-level” neuron behavior, avoiding recent debates on how complex representation-level probing methods should be [18, 29].

Complex explanations. In generating logical explanations of model behavior, one related work is the Anchors procedure of [33], which finds conjunctions of features that “anchor” a model’s prediction in some local neighborhood in input space. Unlike Anchors, we do not explain local model behavior, but rather globally consistent behavior of neurons across an entire dataset. Additionally, we use not just conjunctions, but more complex compositions tailored to the domain of interest.

As our compositional formulas increase in complexity, they begin to resemble approaches to generating natural language explanations of model decisions [2, 8, 16, 17, 31]. These methods primarily operate at the representation level, or describe rationales for individual model predictions. One advantage of our logical explanations is that they are directly grounded in features of the data and have explicit measures of quality (i.e. IoU), in contrast to language explanations generated from black-box models that are often themselves uninterpretable: for example, [17] note that naive language explanation methods often mention evidence not directly present in the input.

Dataset biases and adversarial examples. Complex neural models are often brittle: they fail to generalize to out-of-domain data [3, 13, 22, 32] and are susceptible to adversarial attacks where inputs are subtly modified in a way that causes a model to fail catastrophically [34, 36, 37]. This may be due in part to biases in dataset collection [3, 14, 30, 32], and models fail on datasets which eliminate these biases [3, 13, 22, 32]. In this work we suggest that these artifacts are learned to the degree that we can identify specific detectors for spurious features in representation space, enabling “copy-paste” adversarial examples constructed solely based on the explanations of individual neurons.

8 Discussion

We have described a procedure for obtaining compositional explanations of neurons in deep representations. These explanations more precisely characterize the behavior learned by neurons, as shown through higher measures of explanation quality (i.e. IoU) and qualitative examples of models learning perceptual abstractions in vision and spurious correlations in NLI. Specifically, these explanations (1) identify abstractions, polysemanticity, and spurious correlations localized to specific units in the representation space of deep models; (2) can disambiguate higher versus lower quality neurons in a model with respect to downstream performance; and (3) can be targeted to create “copy-paste” adversarial examples that predictably modify model behavior.

Several unanswered questions emerge: (1) does model pruning [20] more selectively remove the “lower quality” neurons identified by this work? (2) Can we distill a deep model into a simple classifier over binary concept detectors defined by our neuron explanations? (3) Can we use neuron interpretability as a regularization signal during training, and does encouraging neurons to learn more interpretable abstractions result in better downstream task performance?
Broader Impact

Tools for model introspection and interpretation are crucial for better understanding the behavior of black-box models, especially as they make increasingly important decisions in high-stakes societal applications. We believe that the explanations generated in this paper can help unveil richer concepts that represent spurious correlations and potentially problematic biases in models, thus helping practitioners better understand the decisions made by machine learning models.

Nonetheless, we see two limitations with this method as it stands: (1) it currently requires technical expertise to implement, limiting usability by non-experts; (2) it relies on annotated datasets which may be expensive to collect, and may be biased in the kinds of features they contain (or omit). If a potential feature of interest is not present in the annotated dataset, it cannot appear in an explanation. Both of these issues can be ameliorated with future work in (1) building easier user interfaces for explainability, and (2) reducing data annotation requirements.

In high stakes cases, e.g. identifying model biases, care should also be taken to avoid relying too heavily on these explanations as causal proof that a model is encoding a concept, or assuming that the absence of an explanation is proof that the model does not encode the concept (or bias). We provide evidence that neurons exhibit surface-level behavior that is well-correlated with human-interpretable concepts, but by themselves, neuron-level explanations cannot identify the full array of concepts encoded in representations, nor establish definitive causal chains between inputs and decisions.

Acknowledgments and Disclosure of Funding

Thanks to David Bau, Alex Tamkin, Mike Wu, Eric Chu, and Noah Goodman for helpful comments and discussions. JM is supported by an NSF Graduate Research Fellowship and the Office of Naval Research Grant ONR MURI N00014-16-1-2007.

References

[1] M. Alzantot, Y. S. Sharma, A. Elgohary, B.-J. Ho, M. Srivastava, and K.-W. Chang. Generating natural language adversarial examples. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018.

[2] J. Andreas, A. Dragan, and D. Klein. Translating neuralese. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 232–242, 2017.

[3] A. Barbu, D. Mayo, J. Alverio, W. Luo, C. Wang, D. Gutfruedn, J. Tenenbaum, and B. Katz. ObjectNet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. In Advances in Neural Information Processing Systems, pages 9448–9458, 2019.

[4] A. Bau, Y. Belinkov, H. Sajjad, N. Durrani, F. Dalvi, and J. Glass. Identifying and controlling important neurons in neural machine translation. In International Conference on Learning Representations, 2019.

[5] D. Bau, B. Zhou, A. Khosla, A. Oliva, and A. Torralba. Network dissection: Quantifying interpretability of deep visual representations. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pages 6541–6549, 2017.

[6] S. Bowman, G. Angeli, C. Potts, and C. D. Manning. A large annotated corpus for learning natural language inference. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 632–642, 2015.

[7] S. Bowman, J. Gauthier, A. Rastogi, R. Gupta, C. D. Manning, and C. Potts. A fast unified model for parsing and sentence understanding. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1466–1477, 2016.

[8] O.-M. Camburu, T. Rocktäschel, T. Lukasiewicz, and P. Blunsom. e-SNLI: natural language inference with natural language explanations. In Advances in Neural Information Processing Systems, pages 9539–9549, 2018.

[9] S. Carter, Z. Armstrong, L. Schubert, I. Johnson, and C. Olah. Activation atlas. Distill, 4(3):e15, 2019.

[10] F. Dalvi, N. Durrani, H. Sajjad, Y. Belinkov, A. Bau, and J. Glass. What is one grain of sand in the desert? Analyzing individual neurons in deep NLP models. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 33, pages 6309–6317, 2019.
[11] F. Dalvi, A. Nortonsmith, A. Bau, Y. Belinkov, H. Sajjad, N. Durrani, and J. Glass. NeuroX: A toolkit for analyzing individual neurons in neural networks. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 33, pages 9851–9852, 2019.

[12] R. Fong and A. Vedaldi. 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, pages 8730–8738, 2018.

[13] M. Gardner, Y. Artzi, V. Basmove, J. Berant, B. Bogin, S. Chen, P. Dasigi, D. Dua, Y. Elazar, A. Got-tumukkala, et al. Evaluating NLP models via contrast sets. arXiv preprint arXiv:2004.02709, 2020.

[14] S. Gururangan, S. Swayamdipta, O. Levy, R. Schwartz, S. Bowman, and N. A. Smith. Annotation artifacts in natural language inference data. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), pages 107–112, 2018.

[15] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pages 770–778, 2016.

[16] L. A. Hendricks, Z. Akata, M. Rohrbach, J. Donahue, B. Schiele, and T. Darrell. Generating visual explanations. In Proceedings of the European Conference on Computer Vision, pages 3–19, 2016.

[17] L. A. Hendricks, R. Hu, T. Darrell, and Z. Akata. Grounding visual explanations. In Proceedings of the European Conference on Computer Vision, pages 264–279, 2018.

[18] J. Hewitt and P. Liang. Designing and interpreting probes with control tasks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 2733–2743, 2019.

[19] J. Hewitt and C. D. Manning. A structural probe for finding syntax in word representations. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4129–4138, 2019.

[20] G. Hinton, O. Vinyals, and J. Dean. Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531, 2015.

[21] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pages 4700–4708, 2017.

[22] D. Kaushik, E. Hovy, and Z. C. Lipton. Learning the difference that makes a difference with counterfactually-augmented data. In International Conference on Learning Representations (ICLR), 2020.

[23] E. Kitzelmann. Inductive programming: A survey of program synthesis techniques. In International workshop on approaches and applications of inductive programming, pages 50–73. Springer, 2009.

[24] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet classification with deep convolutional neural networks. In Advances in Neural Information Processing Systems, pages 1097–1105, 2012.

[25] T. McCoy, E. Pavlick, and T. Linzen. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 3428–3448, 2019.

[26] A. Nguyen, A. Dosovitskiy, J. Yosinski, T. Brox, and J. Clune. Synthesizing the preferred inputs for neurons in neural networks via deep generator networks. In Advances in Neural Information Processing Systems, pages 3387–3395, 2016.

[27] C. Olah, N. Cammarata, L. Schubert, G. Goh, M. Petrov, and S. Carter. Zoom in: An introduction to circuits. Distill, 5(3):e00024–001, 2020.

[28] J. Pennington, R. Socher, and C. D. Manning. GloVe: Global vectors for word representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, pages 1532–1543, 2014.

[29] T. Pimentel, J. Valvoda, R. H. Maudslay, R. Zmigrod, A. Williams, and R. Cotterell. Information-theoretic probing for linguistic structure. arXiv preprint arXiv:2004.03061, 2020.
[30] A. Poliak, J. Naradowsky, A. Haldar, R. Rudinger, and B. Van Durme. Hypothesis only baselines in natural language inference. In Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics, pages 180–191, 2018.

[31] N. F. Rajani, B. McCann, C. Xiong, and R. Socher. Explain yourself! Leveraging language models for commonsense reasoning. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 4932–4942, 2019.

[32] B. Recht, R. Roelofs, L. Schmidt, and V. Shankar. Do ImageNet classifiers generalize to ImageNet? In International Conference on Machine Learning, pages 5389–5400, 2019.

[33] M. T. Ribeiro, S. Singh, and C. Guestrin. Anchors: High-precision model-agnostic explanations. In Thirty-Second AAAI Conference on Artificial Intelligence, 2018.

[34] M. T. Ribeiro, S. Singh, and C. Guestrin. Semantically equivalent adversarial rules for debugging NLP models. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 856–865, 2018.

[35] K. Simonyan, A. Vedaldi, and A. Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv preprint arXiv:1312.6034, 2013.

[36] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. Intriguing properties of neural networks. In International Conference on Learning Representations, 2014.

[37] E. Wallace, S. Feng, N. Kandpal, M. Gardner, and S. Singh. Universal adversarial triggers for attacking and analyzing nlp. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 2153–2162, 2019.

[38] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In European Conference on Computer Vision, pages 818–833. Springer, 2014.

[39] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Object detectors emerge in deep scene CNNs. In International Conference on Learning Representations, 2015.

[40] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba. Places: A 10 million image database for scene recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017.

[41] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba. Scene parsing through ADE20K dataset. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pages 633–641, 2017.
A Additional image classification examples

Examples are not cherry picked; we enumerate neurons 0–39.

| Unit | Description | Length 1 | IoU | Core | Length 3 | IoU | Core |
|------|-------------|----------|-----|------|----------|-----|------|
| 0    | Lexical and perceptual: bars/surfaces | reception | 0.075 |
|      |             | (reception OR work surface) OR bar | 0.106 |
|      |             | (reception OR work surface) OR bar | 0.107 |
|      |             | (reception OR work surface) OR bar | 0.108 |
| 1    | Lexical and perceptual: windows/shelves | shop window | 0.045 |
|      |             | (shop window OR pantry) OR liquor store outdoor | 0.084 |
|      |             | (shop window OR pantry) OR liquor store outdoor | 0.103 |
| 2    | Lexical and perceptual: islands/grass/water | sea | 0.063 |
|      |             | (sea OR marsh) OR lake | 0.078 |
| 3    | Lexical and perceptual: screens | auditorium | 0.094 |
|      |             | (auditorium OR movie theater indoor) OR theater indoor procmenium | 0.177 |
|      |             | (auditorium OR movie theater indoor) OR theater indoor procmenium | 0.230 |
| 4    | Polysematic: columns/chandeliers | courthouse | 0.025 |
|      |             | (courthouse OR throne room) OR ballroom | 0.049 |
|      |             | (courthouse OR throne room) OR ballroom | 0.064 |
| 5    | Perceptual only: debris | shelf | 0.033 |
|      |             | (slum OR toilet) OR pantry | 0.062 |
|      |             | (slum OR toilet) OR pantry | 0.074 |
| 6    | Lexical and perceptual: domes | gazebo exterior | 0.049 |
|      |             | (gazebo exterior OR tent) OR dome | 0.073 |
|      |             | (gazebo exterior OR tent) OR dome | 0.098 |
| 7    | Lexical and perceptual: fields of plants | vineyard | 0.057 |
|      |             | (vineyard OR orchard) OR corn field | 0.127 |
|      |             | (vineyard OR orchard) OR corn field | 0.133 |
| Unit 16 | (polysematic: gyms/windmills/other) |
|---------|-------------------------------------|
| Length 1 | IoU 0.031 | ice skating rink indoor |
| Length 3 | IoU 0.064 | (ice skating rink indoor OR basketball court indoor OR martial arts gym) |
| Length 10 | IoU 0.112 | (((((ice skating rink indoor OR basketball court indoor) OR martial arts gym) OR windmill) OR hangar indoor) OR boxing ring) OR wrestling ring indoor) OR fire escape) OR badminton court indoor) OR subway station corridor) |

| Unit 17 | (perceptual only: flat areas) |
|---------|--------------------------------|
| Length 1 | IoU 0.068 | auditorium |
| Length 3 | IoU 0.101 | (auditorium OR conference center) OR movie theater indoor) |
| Length 10 | IoU 0.112 | (((((auditorium OR conference center) OR movie theater indoor) OR theater indoor proscenium) OR silver screen) OR courtroom) AND (NOT bench)) AND (NOT pedestal)) OR auditorium) AND (NOT swivel chair)) |

| Unit 18 | (polysematic: rocks/forests/other) |
|---------|-------------------------------------|
| Length 1 | IoU 0.047 | badlands |
| Length 3 | IoU 0.072 | (badlands OR forest needleleaf) OR slot machine) |
| Length 10 | IoU 0.081 | (((((badlands OR forest needleleaf) OR slot machine) OR junkyard) OR arcade machine) OR cave) OR semidesert ground) OR animal) OR car interior backseat) AND (NOT green)) |

| Unit 19 | (lexical and perceptual: cases) |
|---------|----------------------------------|
| Length 1 | IoU 0.144 | bakeryhop |
| Length 3 | IoU 0.173 | (bakeryhop OR case) AND (NOT supermarket) |
| Length 10 | IoU 0.188 | (((((bakeryhop OR case) OR food) AND (NOT supermarket)) OR bakery kitchen) OR butchers shop) OR ice cream parlor) OR island) AND (NOT kitchen)) AND (NOT cabinet)) |

| Unit 20 | (lexical and perceptual: houses/decks) |
|---------|----------------------------------------|
| Length 1 | IoU 0.036 | house |
| Length 3 | IoU 0.044 | (house OR motel) OR zen garden) |
| Length 10 | IoU 0.049 | (((((house OR motel) OR zen garden) OR hunting lodge outdoor) OR lido deck outdoor) AND (NOT house)) OR student residence) OR swimming pool indoor) OR barnyard) AND (NOT barn)) |

| Unit 21 | (polysematic: bookcases/fire stations) |
|---------|----------------------------------------|
| Length 1 | IoU 0.063 | bookcase |
| Length 3 | IoU 0.098 | (bookcase OR fire station) OR book) |
| Length 10 | IoU 0.116 | (((((bookcase OR fire station) OR book) OR video store) OR garage door) OR library indoor) AND (NOT archive)) OR videos) OR convenience store indoor) OR exhibitor) |

| Unit 22 | (lexical and perceptual: bridges, possibly over water) |
|---------|--------------------------------------------------------|
| Length 1 | IoU 0.035 | river |
| Length 3 | IoU 0.066 | (river OR bridge) OR rope bridge) |
| Length 10 | IoU 0.080 | (((((river OR bridge) OR rope bridge) OR creek) OR mountain path) OR aqueduct) OR gulch) OR sandbar) OR footbridge) AND (NOT canal natural)) |

| Unit 23 | (perceptual only: vertical/perspective lines) |
|---------|-----------------------------------------------|
| Length 1 | IoU 0.046 | kitchen |
| Length 3 | IoU 0.058 | (youth hostel OR stove) OR galley) |
| Length 10 | IoU 0.070 | (((((youth hostel OR stove) OR galley) OR microwave) OR work surface) OR telephone booth) OR cubicle office) OR kitchenette) OR exhaust hood) AND (NOT drawer)) |
Unit 24 (polysematic: beds/fireplaces/other)
Length 1 IoU 0.053 fireplace
Length 3 IoU 0.058 ((fireplace OR buffet) OR pulpit)
Length 10 IoU 0.060 ((((((fireplace OR buffet) OR pulpit) OR microwave) AND (NOT poolroom home)) AND (NOT wet bar)) AND (NOT pane)) AND (NOT dinette home)) AND (NOT church indoor) OR microwave)

Unit 25 (perceptual only: empty corridors)
Length 1 IoU 0.067 corridor
Length 3 IoU 0.083 ((corridor OR sauna) OR elevator)
Length 10 IoU 0.087 (((((corridor OR sauna) OR elevator) OR basement) OR fire escape) OR elevator door) OR cargo container interior) OR elevator freight elevator) AND (NOT door frame)) OR corridor

Unit 26 (lexical and perceptual: aqueducts)
Length 1 IoU 0.042 aqueduct
Length 3 IoU 0.079 ((aqueduct OR viaduct) OR cloister indoor)
Length 10 IoU 0.097 ((((((aqueduct OR viaduct) OR cloister indoor) OR bandstand) OR arch) OR aqueduct) OR viaduct) OR water tower) OR arcade) OR arcades)

Unit 27 (perceptual only: dome-like things)
Length 1 IoU 0.032 cockpit
Length 3 IoU 0.054 ((cockpit OR wave) OR viaduct)
Length 10 IoU 0.066 (((((((cockpit OR wave) OR viaduct) OR dome) OR tent) OR dome) OR dome) OR dome) OR dome)

Unit 28 (lexical and perceptual: mediterranean houses)
Length 1 IoU 0.045 alley
Length 3 IoU 0.081 ((medina OR kasbah) OR alley)
Length 10 IoU 0.092 (((((medina OR kasbah) OR alley) AND building) OR kasbah) OR medina) AND (NOT railing))

Unit 29 (lexical and perceptual: house facades)
Length 1 IoU 0.088 house
Length 3 IoU 0.092 ((house OR porch) OR town house)
Length 10 IoU 0.093 (((((house AND (NOT building facade)) OR porch) OR town house) OR inn outdoor) AND (NOT plant)) AND (NOT alley)) AND (NOT dacha)) AND (NOT stairs)) AND (NOT general store outdoor))

Unit 30 (lexical and perceptual: porches)
Length 1 IoU 0.075 balcony interior
Length 3 IoU 0.088 ((balcony interior OR dinette home) OR control tower indoor)
Length 10 IoU 0.089 (((((balcony interior OR dinette home) OR control tower indoor) AND (NOT door)) AND (NOT curtain)) AND (NOT arched))

Unit 31 (polysematic: pool tables/others)
Length 1 IoU 0.106 pool table
Length 3 IoU 0.124 ((pool table OR arcade machine) OR television camera)
Length 10 IoU 0.126 (((((pool table OR arcade machine) OR television camera) OR table tennis) AND (NOT television studio)) AND (NOT wet bar)) AND (NOT music studio))
**Additional NLI examples**

Examples are not cherry picked; we enumerate the first 25 neurons that fire reliably (i.e. at least 500 times across the validation dataset), skipping those already illustrated in the main paper.
| Unit 0 | (((NOT overlap-5%) AND pre:NN) AND (NOT hyp:VB)) AND (NOT hyp:outside) AND (NOT hyp:near)) |
|--------|-----------------------------------------------------------------------------------------------|
| IoU    | 0.355                                                                                         |
| $w_{entail}$ | -0.027                                                     $w_{neutral}$ | 0.018 | $w_{contra}$ | 0.027 |
| Pre    | a woman dressed in a blue long-sleeved shirt and wearing a hairnet.                          |
| Hyp    | the woman is naked and alone in the bathroom.                                                |
| Act    | 43.58                                                                                         |
|        | True contra                                                                                   |
|        | Pred contra                                                                                   |
| Pre    | two men are on a cherry picker proceeding to perform work at a construction site.             |
| Hyp    | two men driving in a truck down an empty highway.                                             |
| Act    | 42.50                                                                                         |
|        | True contra                                                                                   |
|        | Pred contra                                                                                   |
| Pre    | these two poodles, one black and one brown, are playing.                                     |
| Hyp    | the cats are brown and red.                                                                   |
| Act    | 41.24                                                                                         |
|        | True contra                                                                                   |
|        | Pred contra                                                                                   |

| Unit 6 | ((((NOT overlap-2%) AND pre:NN) AND (NOT hyp:people)) AND (NOT hyp:EX)) OR hyp:tall) |
|--------|-------------------------------------------------------------------------------------------|
| IoU    | 0.239                                                                                     |
| $w_{entail}$ | -0.063                                                     $w_{neutral}$ | 0.022 | $w_{contra}$ | 0.009 |
| Pre    | a man in a blue helmet jumping off of a hill on a dirt bike.                               |
| Hyp    | the man is a professional athlete.                                                         |
| Act    | 26.31                                                                                     |
|        | True neutral                                                                                |
|        | Pred neutral                                                                                |
| Pre    | a man standing in front of a class of asian students holding a picture of santa claus.      |
| Hyp    | a tall human standing                                                                       |
| Act    | 26.02                                                                                     |
|        | True neutral                                                                                |
|        | Pred neutral                                                                                |
| Pre    | a girl prepares herself for the swim meet.                                                 |
| Hyp    | the girl has swam before.                                                                  |
| Act    | 25.25                                                                                     |
|        | True entail                                                                                  |
|        | Pred neutral                                                                                |

| Unit 8 | (((hyp:for OR hyp:to) OR hyp:tall) OR hyp:their) AND (NOT hyp:next)) |
|--------|-----------------------------------------------------------------------|
| IoU    | 0.247                                                                 |
| $w_{entail}$ | -0.015                                                     $w_{neutral}$ | 0.023 | $w_{contra}$ | 0.000 |
| Pre    | a man is doing tricks on a skateboard.                                   |
| Hyp    | a tall man doing tricks                                                 |
| Act    | 29.89                                                                 |
|        | True neutral                                                             |
|        | Pred neutral                                                             |
| Pre    | a guy on inline skates with a white hat is on a yellow rail.             |
| Hyp    | the guy on inline skates is trying to impress his girlfriend.            |
| Act    | 26.10                                                                 |
|        | True neutral                                                             |
|        | Pred neutral                                                             |
| Pre    | a gentleman in a striped shirt gesturing with a stick - like object in his hand while passerby stare at him. |
| Hyp    | a gentleman in a striped shirt joyously gesturing                        |
| Act    | 24.58                                                                 |
|        | True neutral                                                             |
|        | Pred neutral                                                             |

| Unit 16 | (((NOT hyp:wearing) AND pre:NN) AND (NOT hyp:sleeping)) AND (NOT hyp:sitting)) AND (NOT hyp:eating)) |
|---------|-------------------------------------------------------------------------------------------------------------------------------|
| IoU     | 0.397                                                                                                                         |
| $w_{entail}$ | 0.022                                                     $w_{neutral}$ | 0.010 | $w_{contra}$ | 0.042 |
| Pre     | a woman wearing a red scarf raises her hand as she walks in a parade.                                                         |
| Hyp     | a woman raises her hand as she walks in a parade for st. patrick’s day.                                                        |
| Act     | 32.96                                                                                                                         |
|        | True neutral                                                             |
|        | Pred neutral                                                             |
| Pre     | a guy on inline skates with a white hat is on a yellow rail.                                                                |
| Hyp     | the guy on inline skates is trying to impress his girlfriend.                                                               |
| Act     | 29.88                                                                                                                         |
|        | True neutral                                                             |
|        | Pred neutral                                                             |
| Pre     | three men; one pedaling while playing drums, one playing piano and one both pedaling and steering, move a type of mobile band down a street. |
| Hyp     | three men are trying to attract a crowd and take them to a bar where they will be playing later                               |
| Act     | 28.13                                                                                                                         |
|        | True neutral                                                             |
|        | Pred neutral                                                             |

| Unit 70 | ((((hyp:in OR hyp:nobody) OR hyp:sitting) AND (NOT overlap-7%)) OR hyp:cat) |

17
Unit 71
(((NOT hyp:to) AND pre:NN) AND (NOT hyp:for)) AND (NOT overlap-75%) AND (NOT hyp:outdoors))

| IoU 0.366 | Neat: 0.015 | Neut: 0.015 | Contra 0.022 |
|-----------|-------------|-------------|--------------|
| Pre       | many people have painted faces at night. | True contra | P red contra |
| Hyp       | the people are swimming in the ocean at noon. | True contra | P red contra |
| Pred: NN  | a man is carrying a child while holding a red and blue umbrella . | True contra | P red contra |
| Pred: NV  | a man is swimming laps in a pool . | True contra | P red contra |
| Pred: NV  | a man with a mustache is playing ice hockey with snow in the background. | True contra | P red contra |
| Pred: NN  | people are swimming in the lake . | True contra | P red contra |

Unit 89
(((NOT overlap-50%) AND (NOT pre:NN)) AND (NOT hyp:for)) AND (NOT hyp:in) AND (NOT hyp:outdoors))

| IoU 0.251 | Neat: 0.015 | Neut: 0.015 | Contra 0.024 |
|-----------|-------------|-------------|--------------|
| Pre       | a young man smiles and points at something off . camera . while standing in front of a display . | True contra | P red contra |
| Hyp       | the young man is frowning with his hands in his pockets . | True contra | P red contra |
| Pred: NN  | a little boy in a blue shirt holding a toy . | True contra | P red contra |
| Pred: NV  | boy dressed in red lighting things on fire . | True contra | P red contra |
| Pred: NN  | a shepherd breed dog running on the beach | True contra | P red contra |
| Pred: NV  | a dog is at home sleeping | True contra | P red contra |

Unit 98
(((NOT overlap-50%) AND (hyp:in OR hyp:running)) OR hyp:swimming) OR hyp:riding)

| IoU 0.127 | Neat: 0.015 | Neut: 0.015 | Contra 0.061 |
|-----------|-------------|-------------|--------------|
| Pre       | a woman , wearing a dress , while sitting down playing a musical instrument and singing into a microphone . | True contra | P red contra |
| Hyp       | the woman is swimming in the middle of the ocean by herself . | True contra | P red contra |
| Pred: NN  | a man with a mustache is playing ice hockey with snow in the background . | True contra | P red contra |
| Pred: NV  | people are swimming in the lake . | True contra | P red contra |
| Pred: NN  | people walking through dirt . | True contra | P red contra |
| Pred: NV  | people are swimming . | True contra | P red contra |

Unit 128
(((NOT overlap-50%) AND hyp:NN) AND (NOT hyp:outside)) OR hyp:sleeping) AND (NOT hyp:near))

| IoU 0.313 | Neat: 0.001 | Neut: 0.001 | Contra 0.034 |
|-----------|-------------|-------------|--------------|
| Pre       | two men are on a cherry picker proceeding to perform work at a construction site . | True contra | P red contra |
| Hyp       | two men driving in a truck down an empty highway . | True contra | P red contra |
| Pred: NN  | a woman dressed in a blue long - sleeved shirt and wearing a hairnet . | True contra | P red contra |
| Hyp | the woman is naked and alone in the bathroom. |
|----|----------------------------------------|
| Act 44.54 | True contra | Pred contra |
| Pre | a boy in a red shirt and a boy in a yellow shirt are jumping on a trampoline outside. |
| Hyp | the boys are asleep. |
| Act 42.99 | True contra | Pred contra |

**Unit 134**

| Hyp | People walking down a busy city street in the winter. |
|----|---------------------------------------------------|
| Pre | a man in a red hat and shirt with gray shorts attempts to do the splits. |
| Hyp | the man has a blue hat. |
| Act 29.96 | True contra | Pred contra |

**Unit 157**

| Hyp | People walking down a busy city street in the summer. |
|----|-----------------------------------------------------|
| Pre | a man in a black hat and shirt with gray shorts attempts to do the splits. |
| Hyp | the man has a blue hat. |
| Act 28.37 | True contra | Pred contra |

**Unit 173**

| Hyp | a group of people sitting in a grassy area under a pink and white blossoming tree. |
|----|---------------------------------------------------------------------------------|
| Pre | three people are running around. |
| Hyp | people are running in a grassy area. |
| Act 31.63 | True contra | Pred contra |

**Unit 203**

| Hyp | a girl and two boys are playing in water. |
|----|-----------------------------------------|
| Pre | the children are eating dinner at a restaurant. |
| Hyp | while some people look in the barn, others walk on the bridge and some are enjoying cooling off in the water by the beach. |
| Hyp | the people are going in the barn to see the horse. |
| Act 37.10 | True contra | Pred contra |

**Unit 207**

| Hyp | a dog is sleeping on a blanket. |
|----|---------------------------------|
| Pre | brown dog running through shallow water. |
| Hyp | while some people look in the barn, others walk on the bridge and some are enjoying cooling off in the water by the beach. |
| Hyp | the people are going in the barn to see the horse. |
| Act 34.58 | True contra | Pred contra |

**Unit 209**

| Hyp | a police officer and his motorcycle in a crowd of people at a protest. |
|----|----------------------------------------------------------------------|
| Pre | a police officer is riding a unicorn in front of a crowd. |
| Hyp | while some people look in the barn, others walk on the bridge and some are enjoying cooling off in the water by the beach. |
| Hyp | the people are going in the barn to see the horse. |
| Act 27.77 | True contra | Pred contra |
| Unit 257 |
|--------------------------|
| (((hyp:their OR overlap-75%) AND hyp:in) OR hyp:friend) AND hyp:| |
| loU 0.218 w_\text{neutral} -0.041 w_\text{contra} 0.052 w_\text{contra} 0.003 |
| Pre | a dressed up woman walking next to a store at night. |
| Hyp | a dressed up woman is walking next to a pharmacy at night. |
| Act 30.76 | True neutral | Pred neutral |
| Pre | a man in a blue shirt, khaki shorts, ball cap and white socks and loafers walking behind a group of people walking down a stone walkway with a water bottle in his left hand. |
| Hyp | a man in a blue shirt, khaki shorts, ball cap and blue socks and loafers walking behind a group of people walking down a stone walkway with a water bottle in his left hand. |
| Act 29.69 | True contra | Pred contra |
| Pre | a man is standing in coconuts while trying to open one. |
| Hyp | a sad man is standing in coconuts while trying to open one. |
| Act 29.24 | True neutral | Pred neutral |

| Unit 265 |
|--------------------------|
| (((NOT hyp:PRPS) AND pre:) AND (NOT hyp:VB)) AND (NOT hyp:PRP) AND (NOT hyp:in)) |
| loU 0.323 w_\text{neutral} 0.028 w_\text{neutral} -0.003 w_\text{contra} 0.055 |
| Pre | a youth is kicking a soccer ball in an empty brick area. |
| Hyp | a human kicking. |
| Act 35.21 | True entail | Pred entail |
| Pre | a band of people playing brass instruments is performing outside. |
| Hyp | a group of people have instruments. |
| Act 31.93 | True entail | Pred entail |
| Pre | three hikers are hiking in a mountain filled with trees and snow. |
| Hyp | people were on grass. |
| Act 31.77 | True unknown | Pred entail |

| Unit 270 |
|--------------------------|
| (((NOT overlap-5%) AND hyp:DT) AND (NOT hyp:outside)) AND (NOT hyp:has)) AND (NOT hyp:near)) |
| loU 0.315 w_\text{neutral} -0.060 w_\text{neutral} -0.011 w_\text{contra} 0.030 |
| Pre | several people prepare their stalls that consist of fish, vegetables and fruits for the public eye. |
| Hyp | two men sit in a truck. |
| Act 38.74 | True contra | Pred contra |
| Pre | outdoors in front of a crowd, a man plays an instrument by blowing into pipes he holds up to his face. |
| Hyp | a man sitting on the couch reading a book. |
| Act 34.87 | True contra | Pred contra |
| Pre | blurry people walking in the city at night. |
| Hyp | seven people dancing in a nightclub. |
| Act 34.41 | True contra | Pred contra |

| Unit 280 |
|--------------------------|
| (((NOT hyp:for) AND pre:NN) AND (NOT hyp:VB)) AND (NOT hyp:PRP) OR overlap-75%) |
| loU 0.420 w_\text{neutral} 0.018 w_\text{neutral} -0.034 w_\text{contra} 0.022 |
| Pre | the lady in the red jacket is helping the other lady decide what to buy. |
| Hyp | there are multiple people present. |
| Act 30.85 | True entail | Pred entail |
| Pre | a sports match is taking place between one team wearing the colors red and white and another team sporting the colors black and blue. |
| Hyp | the two teams are wearing different colors. |
| Act 28.93 | True entail | Pred entail |
| Pre | two men, one with a camera and another with hair clippers are helping another man in kitchen. |
| Hyp | three men are pictured. |
| Act 28.62 | True entail | Pred entail |
### Unit 283

| (((NOT pre:and) AND hyp:IN) OR hyp:PRPS) AND hyp:) OR hyp:VB) | Act | Pre | Hyp |
|--------------------------------------------------------------|-----|-----|-----|
| 0.223            | 32.36 | a soccer game with multiple males playing. | a men’s soccer team winning the world cup. |
| 0.034            | 0.010 | Pred neutral | Pred neutral |

### Unit 284

| (((NOT hyp:J) AND overlap:25%) AND (NOT hyp:PRPS)) AND (NOT hyp:to)) OR hyp:people) | Act | Pre | Hyp |
|-----------------------------------------------------------------------------------------------------------------|-----|-----|-----|
| 0.185            | 27.62 | elegantly dressed in black, a man and woman embrace in dance. | two people are dancing. |
| 0.033            | 0.022 | Pred entail | Pred entail |

### Unit 302

| (((hyp:for OR hyp:to) OR hyp:home) OR hyp:after) OR hyp:their) | Act | Pre | Hyp |
|-----------------------------------------------------------------|-----|-----|-----|
| 0.228            | 33.80 | toddler walking along path. | toddler is walking to his mom. |
| 0.053            | 0.004 | Pred neutral | Pred neutral |

### Unit 362

| (((hyp:outdoors OR hyp:outside) OR hyp:near) OR hyp:there) OR hyp:not) | Act | Pre | Hyp |
|-----------------------------------------------------------------------|-----|-----|-----|
| 0.188            | 36.00 | man and a woman walking on the street. | there are at least two people in the picture. |
| 0.041            | 0.027 | Pred entail | Pred entail |

### Unit 375

| (((hyp:nobod) OR overlap:75%) OR hyp:not) OR hyp:no) OR hyp:one) | Act | Pre | Hyp |
|------------------------------------------------------------------|-----|-----|-----|
| 0.201            | 32.35 | a band which includes an upright bass player is playing in a tent in front of canadian flags. | the band has no bass player. |
| -0.007           | 0.038 | Pred entail | Pred entail |
Act 30.21  True contra  Pred contra
Pre a boy with a concerned look it holding up two newspapers featuring a headline about murder.
Hyp a boy is not holding anything.
Act 25.75  True contra  Pred contra
Pre a young boy wearing a red coat eats a chocolate bar.
Hyp the boy has no clothes on.
Act 20.98  True contra  Pred contra

Unit 382
(((NOT hyp:there) AND hyp:NN) AND (NOT hyp:walking)) AND (NOT hyp:standing)) OR hyp:VB
IoU 0.375  Wneutral -0.022  Wcontra 0.008
Pre a group of people wearing hats and using walking sticks are walking through a wooded area on a trail.
Hyp the tourists are being guided on their trip.
Act 40.71  True neutral  Pred neutral
Pre a gentleman in a striped shirt gesturing with a stick - like object in his hand while passersby stare at him.
Hyp a gentleman in a striped shirt joyously gesturing
Act 40.40  True neutral  Pred neutral
Pre a middle-aged man in a gray t-shirt and brown pants sitting on his bed reading a flyer - like paper.
Hyp he is reading a flyer about a new job he is interested in.
Act 40.04  True neutral  Pred neutral

Unit 386
(((hyp:IN AND overlap>75%) OR hyp:not) OR hyp:no) OR hyp:only)
IoU 0.198  Wneutral -0.075  Wcontra 0.060
Pre a man in a blue shirt, khaki shorts, ball cap and white socks and loafers walking behind a group of people walking down a stone walkway with a water bottle in his left hand.
Hyp a man in a blue shirt, khaki shorts, ball cap and blue socks and loafers walking behind a group of people walking down a stone walkway with a water bottle in his left hand.
Act 32.64  True contra  Pred contra
Pre a boy with a concerned look it holding up two newspapers featuring a headline about murder.
Hyp a boy is not holding anything.
Act 26.37  True contra  Pred contra
Pre a dressed up woman walking next to a store at night.
Hyp a dressed up woman is walking next to a pharmacy at night.
Act 25.10  True neutral  Pred contra

Unit 390
(((hyp:IN OR hyp:to) OR hyp:PRPS) AND (NOT hyp:EX)) OR hyp:NMP
IoU 0.422  Wneutral -0.045  Wcontra 0.010
Pre two women walking in an area of UNK.
Hyp two UNK workers walk down the street of the once beautiful suburban neighborhood, surveying the damage from the storm.
Act 41.84  True neutral  Pred neutral
Pre a group of kids are playing on a tire swing.
Hyp a group of dogs are chasing a duck.
Act 39.84  True contra  Pred contra
Pre a woman walks by a brick building that's covered with graffiti.
Hyp the woman's son drew some of the graffiti.
Act 37.97  True neutral  Pred neutral