
Neuron to Graph: Interpreting Language Model Neurons at Scale - Supplementary Material

Anonymous Author(s)

Affiliation

Address

email

1 A Neuron Graph Examples

2 In this appendix we explore some interesting or characteristic behaviours of the neuron graphs.
3 Polysemanticity, the phenomenon where a neuron exhibits multiple unrelated behaviours, is one of
4 the current major challenges of neuron interpretability [1]. When present, polysemanticity often
5 shows up clearly in the neuron graphs as distinct, disconnected subgraphs. For example, in Figure 1,
6 there are three separate subgraphs corresponding to three clearly distinct behaviours. The top
7 subgraph responds to a phrase in Dutch - variations on *de betrokken*, where not all tokens in *betrokken*
8 were important enough to include in the graph. The middle subgraph responds to a phrase in
9 English - variations on *a fun, over the top*. The bottom subgraph responds to a phrase in Swedish -
10 *kollegers berättigade*, with unimportant tokens not included. This natural separation of behaviours
11 into separate subgraphs could potentially make it easier to interpret polysemantic neurons, but more
12 experimentation would be needed to develop and test this further.

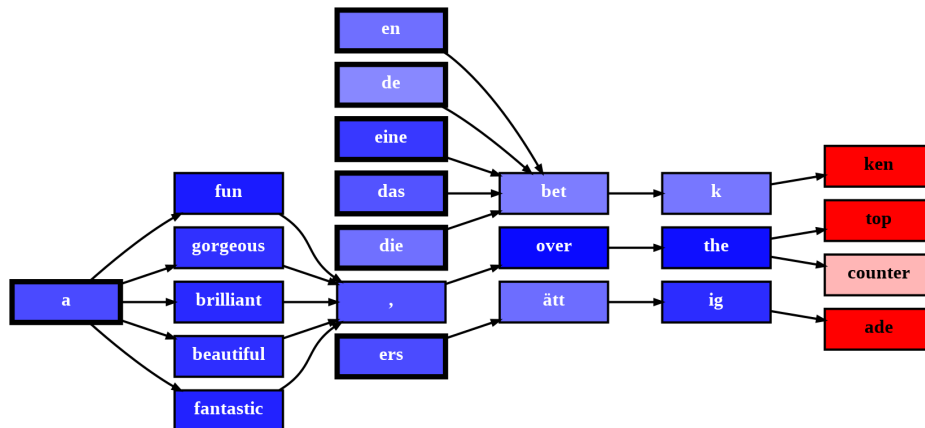


Figure 1: A neuron graph exhibiting polysemanticity, with three disconnected subgraphs each responding to a phrase in a different language.

13 Neuron graphs appear to work particularly well for “syntactic neurons” that respond to structural
14 patterns, concisely and simply capturing the structure. One rich source of syntactic text is program-
15 ming, suggesting that N2G could be particularly useful for analysing models trained to write code. We
16 use the search capability on the neuron graphs of the SoLU model to identify many neurons related
17 to import statements in various languages. Figure 2 shows three examples from different layers of
18 the model. The top and bottom left are from layer 2, and represent basic syntax in Go and Python
19 respectively. The bottom right graph is for a neuron in layer 4, and shows a neuron that responds to

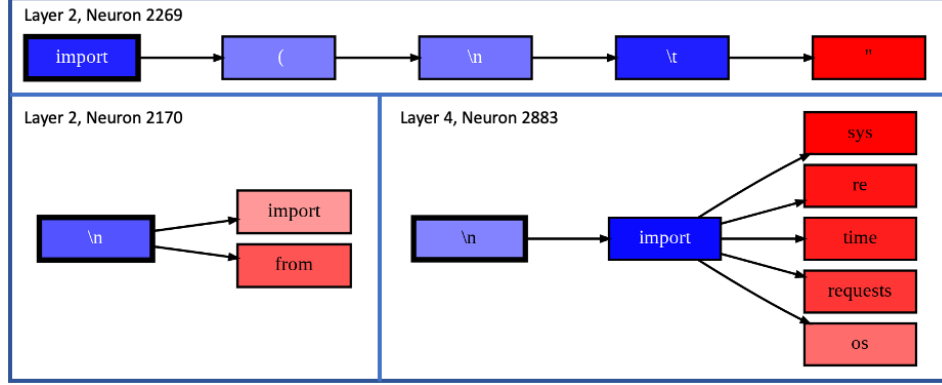


Figure 2: Neurons related to programming syntax, specifically import statements. Top - Neuron graph illustrating import syntax for the Go programming language. Bottom Left: Neuron graph showing fundamental elements of Python import syntax. Bottom Right: Neuron graph for a neuron that responds to the imports of widely-used Python packages.

imports of common Python packages. The circuits line of research [2] would suggest that later-layer neurons like the one in layer 4 may be "composed" of neurons in early layers - for example, a simple way to do this would be to union over several neurons that each respond to an import of one of the packages (i.e., the later composed neuron activates if any of the previous neurons activates). Moving from understanding individual neurons to understanding circuits of neurons is a crucial step in interpretability research. The ability to automatically identify similar neurons could be expanded to identify neurons that have a subset of another neuron's behaviour, which could provide a method for discovering simple circuits in language models. This demonstrates how the flexible representations built by N2G could help facilitate new methods for interpretability research.

B N2G Pseudocode

Algorithm 1

```

1: procedure N2G ALGORITHM
   Input: Target neuron  $n_{\ell j}$ , list of sequences  $\mathcal{X}$ 
   Prune, Saliency Identification and Augment Steps:
2: for  $x^{(i)} \in \mathcal{X}$  do
3:   Compute  $e_i \leftarrow$  pivot token index
4:   Find  $y^{(i)} \leftarrow$  minimal sub sequence with activation ratio  $\geq 0.5$ 
5:   Form  $\mathcal{Y} \leftarrow \{y^{(1)}, \dots, y^{(m)}\}$ 
6: for  $y^{(i)} \in \mathcal{Y}$  do
7:   Compute relative importance value  $\alpha_{i,k}$  for each token in  $y_k^{(i)}$ 
8:   Identify tokens with high  $\alpha_{i,k}$ 
9:   Obtain replacement tokens  $\mathcal{R}_{k,i}$  from helper model
   Construct Lattice and Optimise Steps:
10:  Combine  $y^{(i)}$  and  $\mathcal{R}_{k,i}$  to form a lattice of augmented minimal subsequences
11:  Find optimal token combination in the lattice that maximises target neuron activation  $n_{\ell j}$ 
   Output:
12:  The optimal lattice of augmented minimal subsequences representing high activation contexts
   for the target neuron  $n_{\ell j}$ 

```

C Broader Impact

It is worth mentioning that the field of mechanistic interpretability is ideal for understanding the black box nature of neural networks. This approach can help improve model comprehension and

33 enable researchers to build more transparent AI systems. However, it is essential to recognize that as
34 models become more capable, they may also be used for purposes that are not aligned with societal
35 needs and safety. This presents potential ethical concerns and underscores the need for responsible
36 development and implementation of such technologies. Our work helps illustrate in an accessible
37 manner the inner-workings of neural networks, a step towards aligning their use with responsible use.
38 Researchers and practitioners must remain vigilant in addressing these challenges and ensuring that
39 AI advances contribute positively to society. Additionally, interdisciplinary collaborations and public
40 discussions can aid in raising awareness and developing robust strategies to mitigate potential risks.

41 **References**

- 42 [1] Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna
43 Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam
44 McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. Toy
45 models of superposition, 2022.
- 46 [2] Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter.
47 Zoom in: An introduction to circuits. *Distill*, 2020. <https://distill.pub/2020/circuits/zoom-in>.