Automated Identification of Potential Feature Neurons[1]

Esben Kran
Apart Research

Michelle Lo
Alignment Jam #4

Fazl Barez
Apart Research

Abstract

This report investigates the automated identification of neurons which potentially correspond to a feature in a language model, using an initial dataset of maximum activation texts and word embeddings. This method could speed up the rate of interpretability research by flagging high potential feature neurons, and building on existing infrastructure such as Neuroscope. We show that this method is feasible for quantifying the level of semantic relatedness between maximum activating tokens on an existing dataset, performing basic interpretability analysis by comparing activations on synonyms, and generating prompt guidance for further avenues of human investigation. We also show that this method is generalisable across multiple language models and suggest areas of further exploration based on results.

Keywords: Mechanistic interpretability, ML safety

1. Introduction

The field of artificial intelligence and machine learning has seen significant advancements in recent years, and with it, an increased interest in understanding the internal workings of neural networks. One key aspect of this understanding is the identification of feature neurons, which are neurons that appear to respond to specific concepts. Feature neurons form the computational subgraphs or circuits of a neural network and are crucial in understanding how the network functions (Olah et al, 2020). However, identifying feature neurons is a time-consuming and challenging task, as neural networks can have an enormous number of neurons. When interpreting activations for individual neurons, narrow sentence datasets and human intuition can lead us to incorrect, oversimplified hypotheses about a neuron's function (Bolukbasi et al, 2021).

In this paper, we propose an automated approach to identifying feature neurons in language models. Our hypothesis is that we can automate feature neuron identification by 1) using word embeddings to detect similarities between tokens which most activate a neuron, 2) verifying the type of input which causes activation by testing the neuron on an automatically diversified dataset, and 3) generating a description of how these tokens are related.

[1] Research conducted at the Apart Research Alignment Jam #4 (Mechanistic Interpretability), 2023 (see https://itch.io/jam/mechint)

The current state of research on identifying feature neurons mainly involves examining text dataset examples and identifying prompts which produce the highest activation on each neuron (Nanda, 2022). However, this method requires qualitative exploration of prompts, which can be time-consuming and may not be representative of the full range of possible sentences (Elhage et al, 2022). Automated identification of feature neurons has the potential to make a significant impact by quantifying the exploration of similar prompts based on measures such as word embedding similarities, and by facilitating the speed of exploration. This allows researchers to focus their efforts and facilitate the speed of interpretability research.

By automating the identification of feature neurons, we aim to inform understanding of the internal workings of a language model and provide a foundation for further research on what circuits we might expect to find. Furthermore, explaining how models work is a crucial step towards achieving transparent AI and evaluating the alignment of models with their intended goals.

2. Methods

In this paper, we propose an automated approach to identifying feature neurons in language models. The methods used in our research are as follows:

Detection of neurons which activate on similar tokens:

We first scraped information from Neuroscope, where each HTML page corresponds to a neuron. We then retrieved the set of top 20 texts which activate the neuron the most, along with their activation scores. The information was processed by returning the maximum activating token in each text, along with its surrounding words (3 words before and after) for context. We then calculated the average similarity score of every token compared to every other token using FastText (an open source library for word embeddings). If the current similarity between tokens was above a certain threshold (set to 60% during experimentation), this indicated that the neuron responds to tokens in the same specific semantic category.

Verification of the type of input which causes activation on the neuron:

We then performed interpretability analysis by diversifying prompts and testing activations. For each token, we retrieved the top 5 most similar tokens using FastText. We then substituted each synonym into the original maximum activating phrase and measured the new activation score for this phrase. If the neuron activated more on this specific synonym, this supported the hypothesis that this neuron corresponds to a feature. If the new activation score was higher than the current score, we stored it and added the new score to a running total which included the original score. We calculated the average by dividing the sum of all scores by the total number of additional synonyms checked. As such, this ensured that neurons which responded also to synonyms would have a higher average score than before.

Generation of a description of the relationship between tokens which activate the neuron:

We then normalised the scores of the tokens by dividing them by the maximum activation score found, such that the scores were between 0 and 1. We connected to the OpenAI GPT-3 API and prompted it to find the common relationship between the list of tokens which activate the neuron. The tokens and their normalised scores and descriptions were then saved in a list. Finally, we displayed information about tokens which had a normalised score above 0.3. This value was set arbitrarily during experimentation, as it produced good results for potential feature neurons on manual checking of the tokens.

Overall, we tested our proposed method on two models: the SoLu 8L Pile model and GPT2-small. We selected the SoLu 8L model for analysis because it was small enough to analyse, but large enough for a significant middle layer to exist. There is evidence for the existence of feature neurons in the middle layer of SoLu models (Elhage et al, 2022). We selected the GPT2-small model for further generalised analysis to compare between models and investigate the generality of our method on a more well-known model.

Link to Google Colab notebook:
https://colab.research.google.com/drive/1F-y05Y6OYGvs0FUUAu6RurJbi1N7jGba?pli=1&authuser=1#scrollTo=-Lf4jJ6LghJz

3. Results

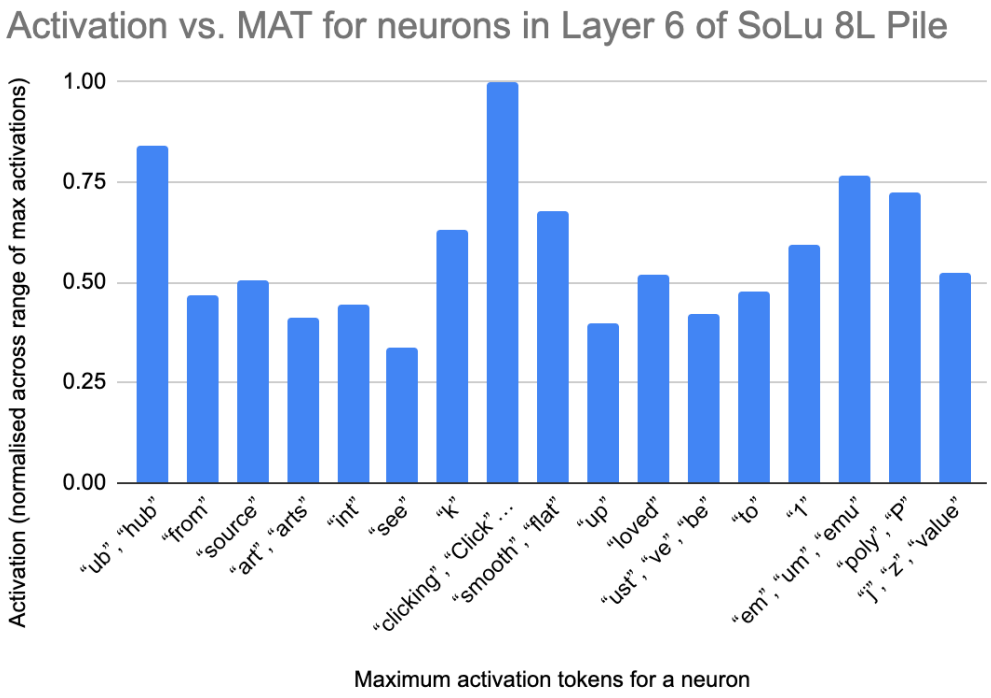
Examining the first 100 neurons of layer 6 in SoLu 8L Pile took 2 minutes 52 seconds.

Table 1. Normalised activation scores, unique tokens and auto-generated description of potential feature neurons in the first 100 neurons of layer 6 in SoLu 8L Pile.

Index	Activation (5 dp)	Tokens (unique)	Description
2	0.45408	“Ass”	They are all the word "ass."
4	0.84178	“ub”, “hub”	All of these words consist of repeated sequences of the same sounds.
5	0.46880	“from”	All the words are the word "from".
11	0.50582	“source”	They all contain the word "source".
22	0.41354	“art”, “arts”	The words "art/arts" are repeated multiple times.
33	0.44480	“int”	All of the words contain the letters "int", with the exception of the word "ent".
43	0.33564	“see”	All of the words are the same, "See."
52	0.63027	“k”	All of the words are the letter 'K'
64	1.00000	“clicking”, “Click”, “Priv”, “click”	All of the words include clicking or Click.

74	0.67768	“smooth”, “flat”	The word "smooth" is repeated multiple times.
85	0.39685	“up”	The word "up" is repeated many times.
101	0.51805	“loved”	They all contain the word "loved".
112	0.42088	“ust”, “ve”, “be”	All of the words are "ve" or "ust".
116	0.47573	“to”	Repetition of the word "to".
134	0.59282	“1”	They all contain the number 1.
135	0.76688	“em”, “um”, “emu”	The words all contain the letters "em" or "emu" at least once.
159	0.72289	“poly”, “P”	All of the words are a variation of the word "poly".
178	0.52364	“j”, “z”, “value”	All words contain the letter "j".

Fig 1. Graph showing average activation scores of neurons in layer 6 in SoLU 8L Pile.



These results identify which neurons activate to certain tokens from the original and extended dataset. The index of the neuron allows for easy navigation to the corresponding page on Neuroscope, and the normalised activation score provides guidance as to which neurons have the highest potential for further investigation. Furthermore, a quick comparison between the tokens and the auto-generated description indicates the correctness of the described relationship, and suggests avenues for further research.

Additionally, comparison between the results of automated identification and the results shown on Neuroscope suggests that this method can successfully diversify the dataset of maximum activation tokens. For instance, in the Neuroscope page for neuron 74 in layer 6 of the model (https://neuroscope.io/solu-8l-pile/6/74.html) , the only activation token

associated is “smooth”, implying that this neuron detects smoothness. However, this method reveals that the associated concept may be broader, as both “smooth” and “flat” tokens positively activate the neuron.

Next, analysing the first 100 neurons of layer 10 in GPT2-small took 3 minutes 58 seconds.

Table 2. Average activation scores of top 8 neurons on tokens in layer 10 in GPT2-small.

Index	Activation (5 dp)	Tokens (unique)	Description
172	0.73523	"each"	They are all the word "each" repeated.
13	0.75027	"\\u2022"	They are all punctuation marks.
9	0.65706	"i", "F", "-", "Y", "Ping", "Long", "an"	They all use the letter "F".
76	0.80490	"("	They all contain parentheses.
40	0.54887	"website", "group", "organisation", "contractor", "through", "aker", ")", "by"	The words group and organization are repeated multiple times.
88	1.00000	"exactly", "just", "matter"	The words all repeat multiple times, specifically "just" and "exactly" each appear seven times, and "matter" appears twice.
49	0.74613	"tell"	The word "tell" is repeated a total of 21 times.
81	0.55694	"3", "ke", "2", "25", "31", "250", "46"	The word "ke" appears in all of the words.

The table of results of GPT2-small (Table 2) shows that this method is generalisable to other language models, as neurons which a human evaluator would likely identify as potentially corresponding to a feature are successfully identified in layer 10. However, it is interesting to note that the neurons which activate the most strongly to words do not seem as closely correlated to each other as in SoLu 8L Pile (Table 1).

4. Discussion and Conclusion

The results of our research show that it is feasible to partially automate feature neuron identification by analyzing similarities between maximum activation tokens in an initial dataset and testing activations in response to synonyms. However, it is important to note that further human verification is needed to fully investigate neurons.

In terms of the field of mechanistic interpretability, our method is informative in that it has the potential to save a significant amount of time compared to human manual investigation. Based on how long it took to obtain results for 100 neurons in one layer of the SoLu 8L Pile model, we project that it would take approximately an hour to process all 4096 neurons in one layer of the same model. This implies that our method has the

potential to run through all neurons in the middle layer of a neural network and flag avenues for further investigation much more efficiently than current manual processes.

We believe that our method is new because, as far as we are aware, there is no currently existing tool which both flags potential feature neurons from existing activation data and automatically diversifies the dataset to narrow down identified neurons. Furthermore, this proof of concept tool builds on existing infrastructure such as the Neuroscope tool, and could be adapted easily to extend its functionality.

There are several limitations to this research. First, we assume that feature neurons can be identified in neurons with high activation scores and semantically similar maximum activation tokens. There may be alternative aspects to investigate which may not be suited to automation. Second, the data used in this research was scraped directly from Neuroscope. This means that our results are limited by the initial dataset examples, hence reducing the effectiveness of the proposed method, since the method may not detect neurons which process concepts that are not obvious in the starting text. Additionally, auto-generated descriptions could provide misleading prompts and misdirect researchers.

To improve upon these limitations, we suggest that we could improve on the speed of automatic processing (not possible due to time constraints). We could also further diversify the dataset to verify types of input which activate neurons. For instance, we could automate the process of replacing all words in the surrounding phrase of the maximum activation token one by one, to examine how contextual information affects activation. Further improvement and investigation is welcome to extend this tooling method.

5. References

Bolukbasi, Tolga, et al. “An Interpretability Illusion for BERT.” ArXiv:2104.07143 [Cs], 14 Apr. 2021, arxiv.org/abs/2104.07143. Accessed 22 Jan. 2023.

Elhage, et al., "Softmax Linear Units", Transformer Circuits Thread, 2022.

Nanda, Neel, "TransformerLens", GitHub, 2022, https://github.com/neelnanda-io/TransformerLens.

Olah, et al., "Zoom In: An Introduction to Circuits", Distill, 2020.