Modularity in Transformers: Investigating Neuron Separability & Specialization

Nicholas Pochinkov
Independent
Dublin, Ireland
work@nicky.pro

Thomas Jones
Independent
London, UK

Mohammed Rashidur Rahman
Independent
London, UK

Abstract

Transformer models are increasingly prevalent in various applications, yet our understanding of their internal workings remains limited. This paper investigates the modularity and task specialization of neurons within transformer architectures, focusing on both vision (ViT) and language (Mistral 7B) models. Using a combination of selective pruning and MoEfication clustering techniques, we analyze the overlap and specialization of neurons across different tasks and data subsets. Our findings reveal evidence of task-specific neuron clusters, with varying degrees of overlap between related tasks. We observe that neuron importance patterns persist to some extent even in randomly initialized models, suggesting an inherent structure that training refines. Additionally, we find that neuron clusters identified through MoEfication correspond more strongly to task-specific neurons in earlier and later layers of the models. This work contributes to a more nuanced understanding of transformer internals and offers insights into potential avenues for improving model interpretability and efficiency.

1 Introduction

As transformer neural networks continue to scale up and demonstrate improvements across various applications, their integration into more areas of technology becomes increasingly widespread. Despite these advances, our fundamental understanding of how these models operate remains limited [Hendrycks et al., 2021]. This gap in understanding poses challenges for the reliability and trustworthiness of these models, particularly in high-stakes situations.

To address this, our study draws inspiration from bottom-up circuits-based [Olsson et al., 2022] and top-down linear direction-based analyses [Zou et al., 2023], as well as from modularity research [Pfeiffer et al., 2023, Zhang et al., 2023, Filan et al., 2021]. We aim to explore the specialization of different components within neural networks in performing distinct tasks and querying different knowledge bases, defining modularity in this context as having two key traits: 1) A high degree of separability between components tasked with different functions, and 2) Positional locality of these components. Our focus primarily lies on the identification of distinct components within these networks and understanding their roles in processing different types of information.

2 Method

2.1 Pre-Trained Models

We work with a Vision Transformer, ViT-Base-224 [Dosovitskiy et al., 2021]. as well as with a causal language model, Mistral 7B instruct v0.2 [Jiang et al., 2023].

The models are accessed via the HuggingFace transformer library [Taylor et al., 2022]. For each model, we perform the analysis on both the pre-trained model (called the 'trained model') and a single seed of a randomly initialized model (called the 'random model').

The neurons of study are found post-activation in the MLP layers, that is, between the W_{in} and W_{out} weight matrices, after the activation function (see [Geva et al., 2022]).

2.2 Datasets and Tasks

For ViT, the task of study is to perform image classification on different classes of Cifar100 Krizhevsky [2009] fine-tuned model as a 'retained' baseline. For the 'forget' datasets, each super-class of Cifar100 (Cifar20) is used as a target.

For text models, we investigate next-token prediction on various categories of text as possible niches, as well as subject-split MMLU question answering Hendrycks et al. [2020]. For a broad baseline of model activations, ElutherAI's 'The Pile' Gao et al. [2020][1] is used. For the specialised datasets, use the individual subsets of the pile, as well as some some additional data sets. We use some scientific-subject specific instruction datasets on specific topics from Li et al. [2023], including 'Biology', 'Physics', 'Chemistry'. We also use 'CodeParrot GitHub Code' Tunstall et al. [2022], or 'Code' for short.

2.3 Neuron Selection and Neuron Clustering

2.3.1 Neuron selection

We select neurons using a method analogous to the Selective Pruning approach [Pochinkov and Schoots, 2024]. Given two datasets: a 'reference' dataset D_r and 'unlearned' dataset D_u, we score the neurons $S_{n_i^l}$ by getting the ratio of mean absolute deviations between the two datasets, as seen in Equation 1:

$$S_{n_i^l} := \frac{\frac{1}{|D_r|}\sum_{x_i \in D_r} |n_i^l(x_i)|}{\epsilon + \frac{1}{|D_u|}\sum_{x_j \in D_u} |n_i^l(x_j)|} \tag{1}$$

where $n_i^l(x_j)$ is the activation of neuron i in layer l for input x_j.

We then "select" a subset of neurons as being relevant to some data subset D_q by taking a percentage of neurons with the highest scores. The percentage of neurons selected for each data subset is determined by aiming for a specific drop in top1 accuracy when these neurons are removed.

For ViT, the 'retained' baseline is Cifar100 [Krizhevsky, 2009], and the 'unlearned' data subsets are those of Cifar20. For Mistral, the retained baseline is ElutherAI's 'The Pile' [Gao et al., 2020], with the subsets being the different Pile subsets and the additional specialized datasets mentioned earlier.

2.3.2 MoEfication clustering

We employ 'Mixture of Experts'-ification [Zhang et al., 2022] as a method for grouping neurons into clusters. This approach involves taking the input weights W_{in} for an MLP layer and treating its columns as a collection of vectors. We then perform balanced k-means clustering [Malinen and Fränti, 2014] on these vectors:

$$\{C_1, ..., C_k\} = \text{BalancedKMeans}(W_{in}^T, k) \tag{2}$$

where C_i represents the i-th cluster of neurons. This effectively splits each MLP layer into k balanced 'mixture of experts' clusters. The intuition is to group neurons that are often activated simultaneously, allowing for efficient expert selection during inference.

For ViT, we use $k = 96$ as in the original paper. For Mistral, $k = 128$ is used for even divisibility of layers. We perform this clustering on both pre-trained and randomly-initialized models, providing a basis for comparing the emergent structure in trained networks against random initializations.

[1]We use a subset of The Pile on HuggingFace with most copyrighted material removed.

3 Class-wise Unlearning Performance Evaluation

3.1 ViT Performance

We assess the performance of ViT across all Cifar20 classes to determine if and how the performance on other classes is affected by the deactivation of neurons associated with a particular class. We uncover that between classes with commonalities, there often seem to exist shared neural pathways.

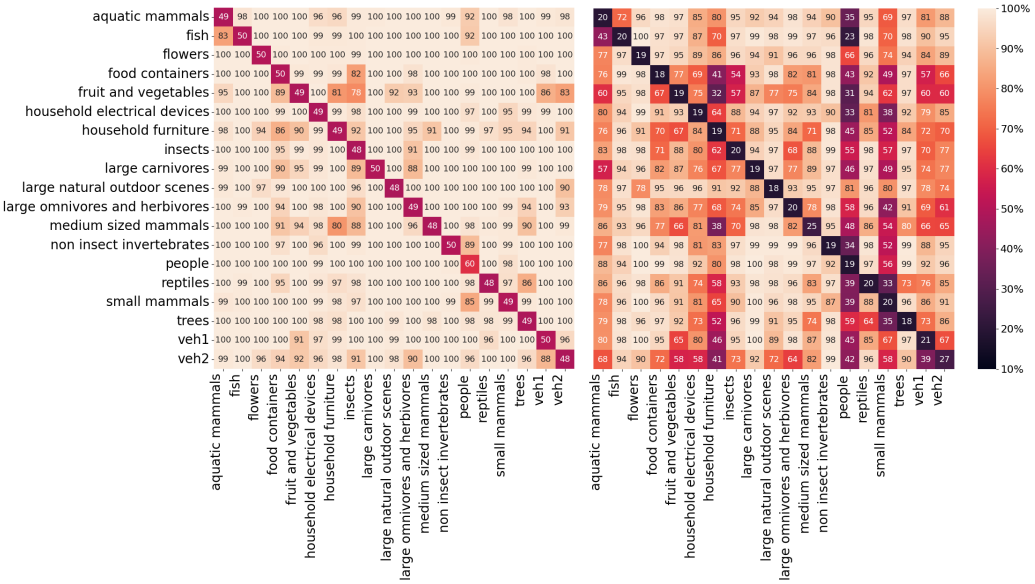


Figure 1: Relative Top1 % image classification accuracy compared to baseline for ViT in Cifar100, after selectively pruning MLP neurons to get a target drop in performance of 50% (left) and 80% (right) for the 'unlearned' class (shown as the vertical strips).

We notice that the MLP neurons selectively pruned in ViT for classes such as 'insects' and 'non-insect invertebrates' seem to have relatively high overlap, and classes such as 'fish', 'aquatic mammals', and 'reptiles' also have correlated decreases in performance. We also observe that for two classes A and B, the effect of unlearning class A on the performance of class B is not symmetric to unlearning class B affecting the performance of class A, though these effects are correlated.

As we approach attempting to get an 80% drop in top1 accuracy, the selectivity seems to significantly decrease, likely because the attention layers, which we do not modify, provide too strong a signal.

3.2 Mistral Performance

For Mistral, we investigate the degree to which a drop in performance in each class causes a correlated drop in performance across all other dataset tasks. We choose an arbitrary goal of unlearning by selective pruning until achieving a 20% drop in Top1 next-token prediction accuracy in each respective unlearned text subset.

In Figure 2, we observe that in most subsets of data that one would intuitively associate, such as unlearning 'biology' leading to a drop in MMLU 'High School biology'. Notably, we see that unlearning 'PubMed central' leads to significant drops in MMLU 'professional medicine', 'Medical Genetics', 'Clinical Knowledge', 'Anatomy', 'College Biology', and 'College Chemistry'. Additionally, we observe that unlearning 'ArXiv' leads to corresponding drops in MMLU subjects.

We also note some 'failure' cases:

- 'EuroParl' and 'FreeLaw' seem to have relatively modest effects on the accuracy of other subsets, including seemingly related MMLU topics such as 'HighSchool Government & Politics' and 'Professional Law'.

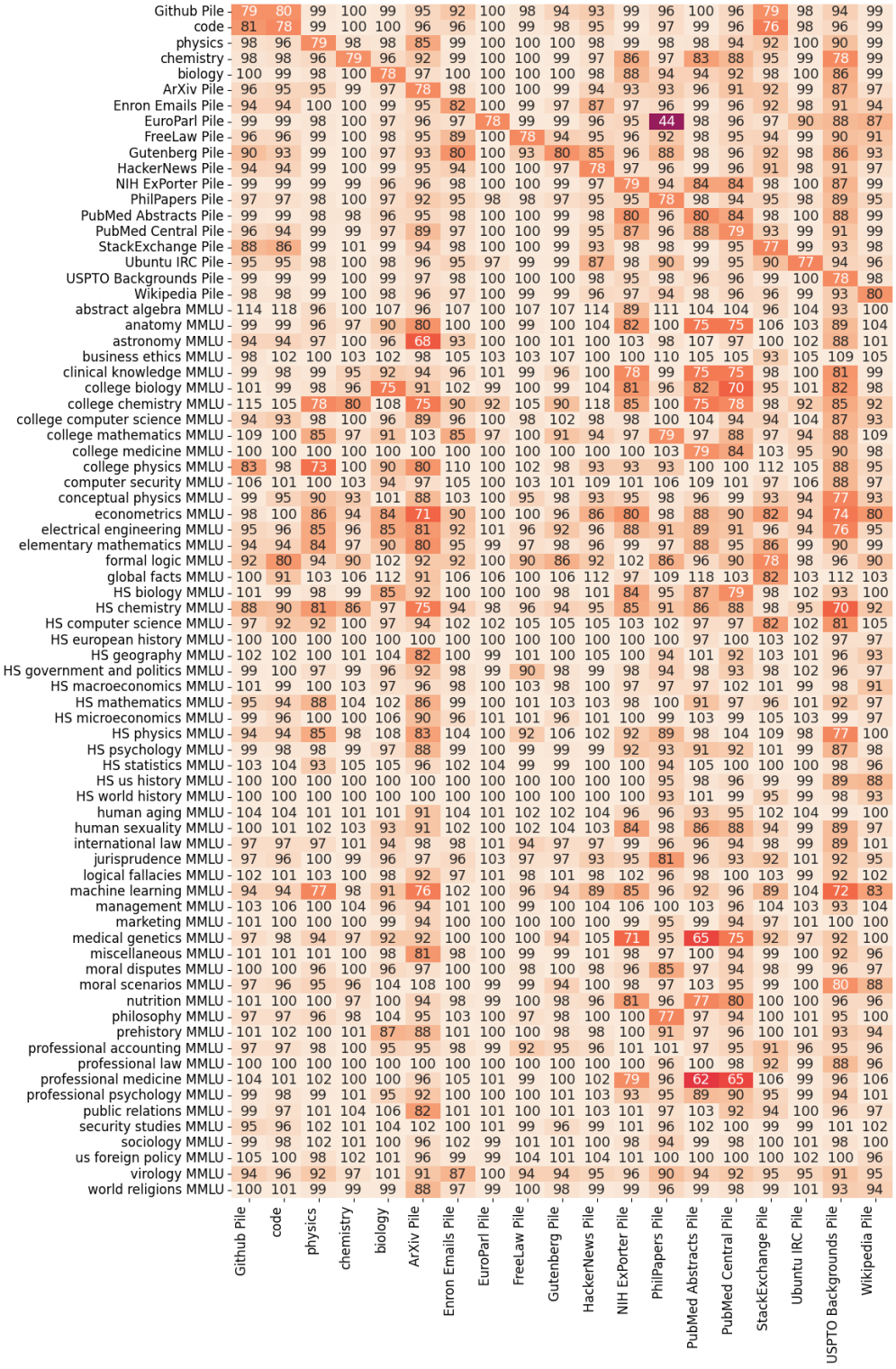


Figure 2: Drop in Top1 accuracy in Mistral 7B for next-token-prediction datasets as well as zero-shot subject-split MMLU question answering. We target a relative drop in performance of 20% for the class being 'unlearned', and each vertical strip shows the effect on the other data subsets.

- In some cases, performance seems to increase, including up to 118% of baseline in MMLU 'Abstract Algebra' after performing the unlearning of Code.
- Unlearning 'PhilPapers' seems to have an outsized negative impact on 'EuroParl' performance, significantly more so than even the 'EuroParl' unlearning has.

These results highlight that the Selective Pruning method seems to perform reasonably well at identifying relatively general class-scale features of models, despite being relatively simplistic in looking at the activations. However, there is potential room for improvement with better methods and datasets. It does, nonetheless, provide a promising glimpse of some kind of interpretable modularity.

4 Intersection Analysis

4.1 Intersection between Class Selected Neurons

We analyze the degree to which selected neurons overlap between different classes or data subsets. For some set of selected neurons N_i and N_j for classes i and j respectively, we calculate $|N_i \cap N_j|/|N_i|$.

Each class of neurons amounts to a total of $12.5 \pm 6.0\%$ of neurons in ViT, and $3.4 \pm 2.6\%$ of neurons in Mistral.

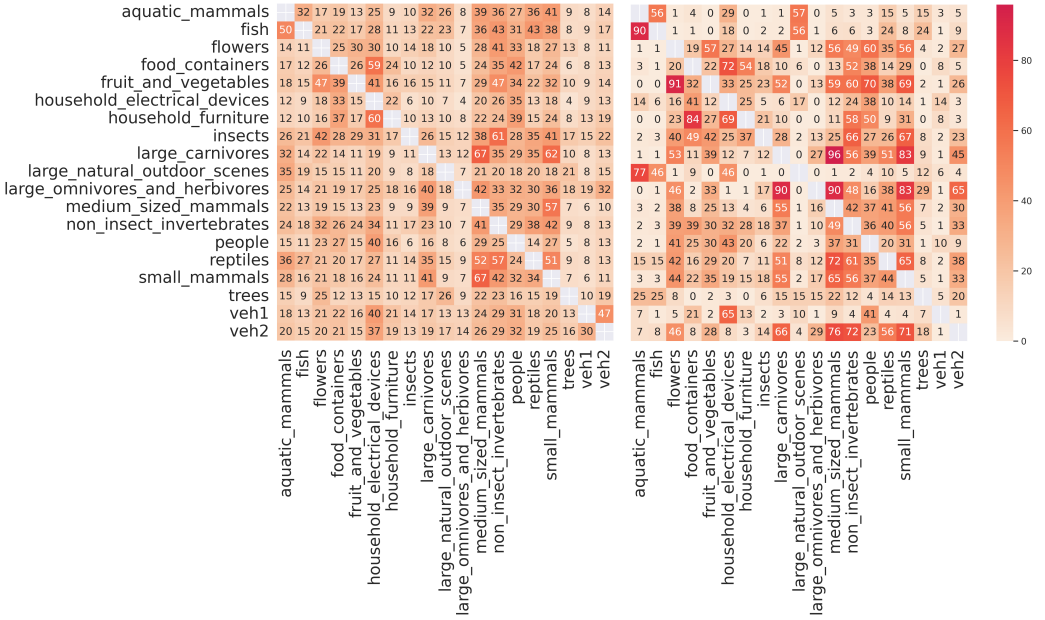


Figure 3: Intersection analysis between selected neurons ranked in the top $12.5 \pm 6.0\%$ for trained ViT model (left) and random model (right) in relative importance to Cifar20 classes. Each square shows the percentage overlap between neurons identified as being important for each class pair.

In Figure 3, we see that the selected pre-trained neurons have relatively distinct neural activations compared to the randomly initialized model. We observe that, although there are some differences, the overall pattern of overlaps seems quite similar between the models, albeit with different magnitudes.

We see three main trends in general going from a random to pre-trained model: 1) More neurons have a closer-to-median overlap. 2) Extreme overlaps were diminished, roughly in proportion to how intuitively dissimilar the classes seem to be. 3) Intuitively similar classes that had little overlap (such as 'vehicles 1' and 'vehicles 2') gain particularly larger overlaps.

The results are somewhat analogous for pre-trained and randomly initialized Mistral models seen in Figure 4.

We see again three similar trends: 1) More neurons have a closer-to-median overlap, though this time tending to be lower. 2) Most extreme overlaps were diminished, roughly in proportion to how

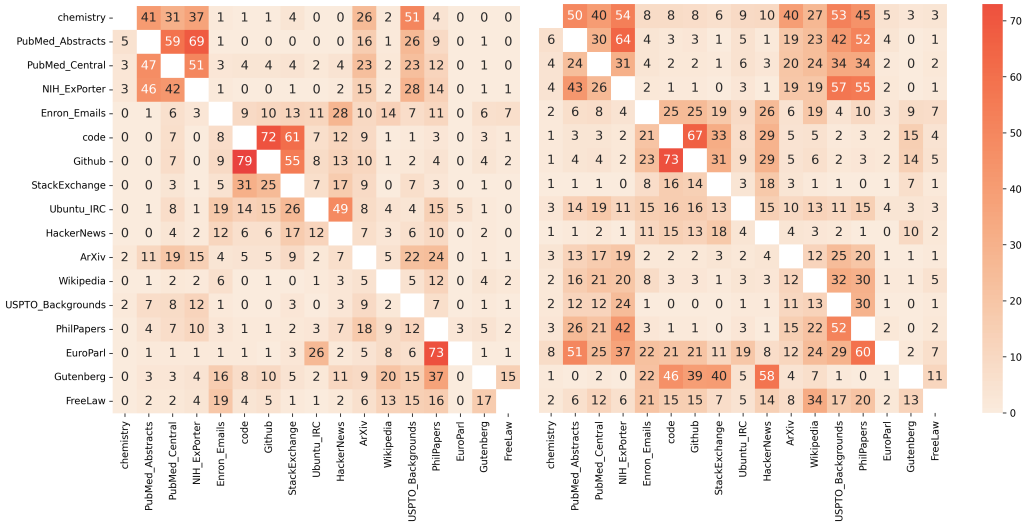


Figure 4: Intersection analysis between selected neurons ranked top 1% for Mistral 7b model that is trained (left) and randomly initialized (right) in relative importance to different text classes. Each square shows percentage overlap between neurons identified as being important for each class pair.

intuitively dissimilar the subsets seem to be. 3) Intuitively similar classes (such as PubMed Abstracts, PubMed Central and NIH ExPorter, or Ubuntu IRC and HackerNews) retain their overlaps or gain larger overlaps.

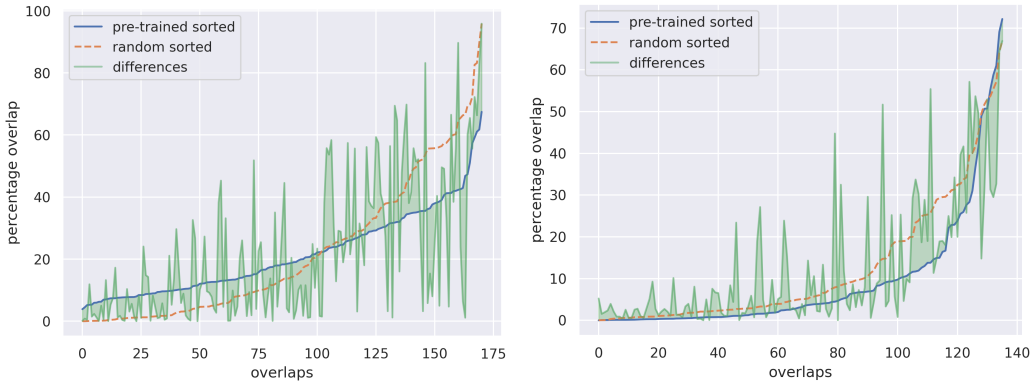


Figure 5: Plot of the overall overlaps between classes of ViT (left) and subtexts of Mistral (right) for both random and pre-trained selected neurons. We also show the difference between the random and pre-trained overlaps at each point.

In Figure 5, we see the overlaps between Cifar20 classes and text subsets sorted for both selected pre-trained and selected random neurons. We also see the difference in overlap between pre-trained and random models. We observe that there is a definite correlation between the overlaps of pre-trained and random neuron selections, but also that there are large shifts. We can more clearly see that the overlaps in ViT are more tightly bound, while in Mistral, the overlaps become overall lower.

Overall, we find the results show a surprising degree to which many of the selected neurons overlap between the randomly initialized and pre-trained models, and see this as evidence for the weak Lottery Ticket Hypothesis [Frankle and Carbin, 2018].

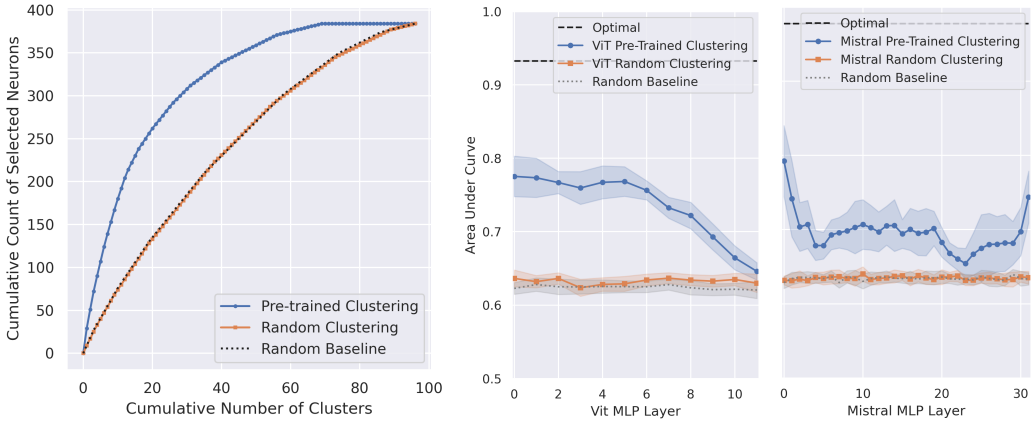
4.2 Intersection between Selected Neurons and Clusters

Inspired by the Lorentz Curve and Gini Coefficient [Gastwirth, 1971], for each layer, we measure the degree to which the selected neurons are preferentially selected into few clusters. We do this by

counting how many selected neurons are in each cluster, sorting the clusters in order from highest to lowest member count, then plotting a cumulative curve showing distribution between clusters.

That is, we create k bins, and take the indices i of the selected neurons, and count how many are assigned to each bin, $B[j]$. We then sort the bins in order of highest to lowest, giving sorted bins $\tilde{B}[j]$, and plot a cumulative distribution of counts c_n, where $c_n = c_{n-1} + B[n]$. We then find the normalized area under the curve, such that if all the neurons were in one bin, the area would be 1.0, and if the neurons were perfectly evenly distributed, it would be 0.5. For an example of such a curve, see Figure 6a.

As the MoEfication clustering of neurons is limited to single layers, we add an extra constraint such that the selection of neurons is limited to have an equal proportion per layer, as having all neurons in one layer automatically distributed neurons equally between these clusters.



(a) ViT Lorentz Curve (b) ViT (c) Mistral

Figure 6: (a) An example of a Lorentz curve in layer 0 of ViT, sorting the distribution of selected top 12.5% of Cifar20 "people" neurons between MoE clusters in trained and randomly initialized ViT models. The Area Under this Curve (AUC) is then found for all classes of Cifar20 in each layer of ViT (b). The same is done for each data subset of Mistral (c) with the top 3.4% of neurons.

We see in Figure 6b that the ViT selected pre-trained neurons correspond relatively strongly with the MoEfication clusters compared to selected random neurons, with stronger correspondence in earlier layers. In Figure 6c we also see that pre-trained neurons correspond more with MoEfication clusters, but this time we see a bi-modal correspondence with later and earlier layers corresponding more than middle layers.

This suggests that the MoEfication in layers close to the text or image seem to map relatively cleanly, but that in more central layers, the co-activation similarity between neurons is not as predictive of class.

5 Related Work

Circuits-based mechanistic interpretability seeks to uncover the exact causal mechanisms and neuron activations responsible for specific behaviours [Chan et al., 2022, Olsson et al., 2022, Wang et al., 2022, nostalgebraist, 2020]. Although this bottom-up approach offers detailed insights, the vast number of potential circuits poses a significant challenge to comprehensively understanding model behaviours.

Activation-direction-based interpretability methods, such as Activation Addition [Turner et al., 2023] and Representation Engineering [Zou et al., 2023, Burns et al., 2022, Turner et al., 2023, Gurnee and Tegmark, 2023], attempt to correlate model activations with meanings or concepts [Dar et al., 2022, Meng et al., 2022]. However, these top-down approaches often overlook finding the sources of these activation patterns.

Neuron-based analysis has also been explored, with studies suggesting that individual neurons may encode human-interpretable concepts in convolutional neural networks Meyes et al. [2020] and language models [Geva et al., 2021, 2022, Bills et al., 2023]. However, the complexity and sheer number of neurons, coupled with their potential to represent multiple features, complicate this approach [Elhage et al., 2022, Pona, 2023]. Techniques like sparse auto-encoder dictionary learning have been proposed to address these challenges, albeit at the cost of increasing the neuron count [Cunningham et al., 2023, Bricken et al., 2023].

Modularity and Clustering Recent research has explored the modular nature of neural networks and how this modularity relates to task specialization. Studies on small neural networks have revealed surprising levels of modularity [Filan et al., 2021, 2020], while work in computer vision has demonstrated specialized "branches" in convolutional networks [Voss et al., 2021, Olah et al., 2017]. In language models, concepts like "MoEfication" [Zhang et al., 2022] and contextual sparsity [Liu et al., 2023] have shown task-specific neuron activation patterns.

Extending these ideas, Meyes et al. [2020] characterized learned representations by identifying functional neuron populations in CNNs that either respond jointly to specific stimuli or have similar impacts on network performance when ablated. Their work highlights that a neuron's importance cannot be determined solely by its activation magnitude, selectivity, or individual impact on performance, emphasizing the need for a more nuanced understanding of neural network internals.

Our research builds upon these findings by investigating specialization across different text genres and image classes as a basis for clustering neurons. We adapt selective pruning techniques [Pochinkov and Schoots, 2024] from Machine Unlearning [Bourtoule et al., 2021] to assess neuron importance for specific data subsets, aiming to provide new insights into the functional organization of complex neural networks.

6 Discussion

Our analysis has shown that both the MoEfication clustering [Zhang et al., 2022] and the importance selection from Selective Pruning [Pochinkov and Schoots, 2024] seem to reveal some degree of interpretable modularity in transformer models. We believe this is a promising starting point for understanding the high-level behavior of transformer neural networks, and for identifying how different parts of a neural network might be mapped out, analogously to how regions of a human brain are mapped out in neuroscience.

Additionally, the overlap between similar categories in randomly initialized models seems to provide some evidence for a weak version of the Lottery Ticket Hypothesis [Frankle and Carbin, 2018].

Our findings have several implications:

1. The qualitatively intuitive nature with which classes seem to overlap seems to suggest some level of distinctness with how models handle these inputs.
2. The relatively small differences in neuron importance between pre-trained and random models indicate that training does indeed lead to meaningful specialization, but that some of this structure may be present even at initialization.
3. The clear but incomplete correspondence between MoEfication clusters and task-specific neurons suggests that both methods form a useful but incomplete for understanding models.

However, our study also has limitations that should be addressed in future work:

1. We focused on MLP neurons and did not analyze attention mechanisms, which play a crucial role in transformer models.
2. While we use two very different models and find similar results, our analysis was limited to a specific set of tasks and datasets, and may not generalize to all domains or model architectures.
3. The interpretation of neuron importance and specialization is still somewhat qualitative and could benefit from more rigorous statistical analysis.

Future work should look to improve the precision with which one can identify different parts of a model that are using different knowledge, as well as increasing the degree to which one can have a fine-grained hierarchical understanding of how the components of a model might be mapped.

Some potential directions include: 1) Extending the analysis to include attention mechanisms and their interaction with MLP layers. 2) Developing more sophisticated clustering techniques that can capture hierarchical relationships between neurons and tasks. 3) Investigating how neuron specialization evolves during the training process, potentially shedding light on the dynamics of learning in these models. 4) Exploring how this understanding of modularity could be leveraged to improve model interpretability, robustness, or transfer learning capabilities.

Acknowledgments and Disclosure of Funding

This study was funded by the Long-Term Future Fund, and conducted in affiliation with AI Safety Camp. The authors would like to thank the anonymous reviewers for their valuable feedback.

References

Steven Bills, Nick Cammarata, Dan Mossing, Henk Tillman, Leo Gao, Gabriel Goh, Ilya Sutskever, Jan Leike, Jeff Wu, and William Saunders. Language models can explain neurons in language models. *URL https://openaipublic. blob. core. windows. net/neuron-explainer/paper/index. html.(Date accessed: 14.05. 2023)*, 2023.

Lucas Bourtoule, Varun Chandrasekaran, Christopher A. Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. Machine unlearning. In *42nd IEEE Symposium on Security and Privacy, SP 2021, San Francisco, CA, USA, 24-27 May 2021*, pages 141–159. IEEE, 2021. doi: 10.1109/SP40001.2021.00019. URL `https://doi.org/10.1109/SP40001.2021.00019`.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, et al. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, page 2, 2023.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. *arXiv preprint arXiv:2212.03827*, 2022.

Lawrence Chan, Adrià Garriga-Alonso, Nicholas Goldowsky-Dill, Ryan Greenblatt, Jenny Nitishinskaya, Ansh Radhakrishnan, Buck Shlegeris, and Nate Thomas. Causal scrubbing: a method for rigorously testing interpretability hypotheses. AI Alignment Forum, 2022. URL `https://www.alignmentforum.org/posts/JvZhhzycHu2Yd57RN`.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600*, 2023.

Guy Dar, Mor Geva, Ankit Gupta, and Jonathan Berant. Analyzing transformers in embedding space. *arXiv preprint arXiv:2209.02535*, 2022.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL `https://openreview.net/forum?id=YicbFdNTTy`.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, et al. Toy models of superposition. *arXiv preprint arXiv:2209.10652*, 2022.

Daniel Filan, Shlomi Hod, Cody Wild, Andrew Critch, and Stuart Russell. Pruned neural networks are surprisingly modular. *arXiv preprint arXiv:2003.04881*, 2020.

Daniel Filan, Stephen Casper, Shlomi Hod, Cody Wild, Andrew Critch, and Stuart Russell. Clusterability in neural networks. *arXiv preprint arXiv:2103.03386*, 2021.

Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. *arXiv preprint arXiv:1803.03635*, 2018.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. The Pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020.

Joseph L Gastwirth. A general definition of the lorenz curve. *Econometrica: Journal of the Econometric Society*, pages 1037–1039, 1971.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 5484–5495. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.emnlp-main.446. URL https://doi.org/10.18653/v1/2021.emnlp-main.446.

Mor Geva, Avi Caciularu, Kevin Ro Wang, and Yoav Goldberg. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. *arXiv preprint arXiv:2203.14680*, 2022.

Wes Gurnee and Max Tegmark. Language models represent space and time. *arXiv preprint arXiv:2310.02207*, 2023.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

Dan Hendrycks, Nicholas Carlini, John Schulman, and Jacob Steinhardt. Unsolved problems in ml safety. *arXiv preprint arXiv:2109.13916*, 2021.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b. 2023.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, 2009.

Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. Camel: Communicative agents for" mind" exploration of large scale language model society. *arXiv preprint arXiv:2303.17760*, 2023.

Zichang Liu, Jue Wang, Tri Dao, Tianyi Zhou, Binhang Yuan, Zhao Song, Anshumali Shrivastava, Ce Zhang, Yuandong Tian, Christopher Ré, and Beidi Chen. Deja vu: Contextual sparsity for efficient llms at inference time. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 22137–22176. PMLR, 2023. URL https://proceedings.mlr.press/v202/liu23am.html.

Mikko I Malinen and Pasi Fränti. Balanced k-means for clustering. In *Structural, Syntactic, and Statistical Pattern Recognition: Joint IAPR International Workshop, S+ SSPR 2014, Joensuu, Finland, August 20-22, 2014. Proceedings*, pages 32–41. Springer, 2014.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. *Advances in Neural Information Processing Systems*, 35:17359–17372, 2022.

Richard Meyes, Constantin Waubert de Puiseau, Andres Posada-Moreno, and Tobias Meisen. Under the hood of neural networks: Characterizing learned representations by functional neuron populations and network ablations. *arXiv preprint arXiv:2004.01254*, 2020.

nostalgebraist. interpreting gpt: the logit lens. *LessWrong*, 2020. URL https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens.

Chris Olah, Alexander Mordvintsev, and Ludwig Schubert. Feature visualization. *Distill*, 2(11):e7, 2017.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, et al. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*, 2022.

Jonas Pfeiffer, Sebastian Ruder, Ivan Vulic, and Edoardo Maria Ponti. Modular deep learning. *CoRR*, abs/2302.11529, 2023. doi: 10.48550/arXiv.2302.11529. URL https://doi.org/10.48550/arXiv.2302.11529.

Nicholas Pochinkov and Nandi Schoots. Dissecting language models: Machine unlearning via selective pruning. *arXiv preprint arXiv:2403.01267*, 2024.

Edoardo Pona. Superposition and dropout. 2023. URL https://www.lesswrong.com/posts/znShPqe9RdtB6AeFr/superposition-and-dropout.

Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. Galactica: A large language model for science. *CoRR*, abs/2211.09085, 2022. doi: 10.48550/arXiv.2211.09085. URL https://doi.org/10.48550/arXiv.2211.09085.

Lewis Tunstall, Leandro Von Werra, and Thomas Wolf. *Natural language processing with transformers*. " O'Reilly Media, Inc.", 2022.

Alex Turner, Lisa Thiergart, David Udell, Gavin Leech, Ulisse Mini, and Monte MacDiarmid. Activation addition: Steering language models without optimization. *CoRR*, abs/2308.10248, 2023. doi: 10.48550/arXiv.2308.10248. URL https://doi.org/10.48550/arXiv.2308.10248.

Chelsea Voss, Gabriel Goh, Nick Cammarata, Michael Petrov, Ludwig Schubert, and Chris Olah. Branch specialization. *Distill*, 2021. doi: 10.23915/distill.00024.008. https://distill.pub/2020/circuits/branch-specialization.

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in gpt-2 small. *arXiv preprint arXiv:2211.00593*, 2022.

Zhengyan Zhang, Yankai Lin, Zhiyuan Liu, Peng Li, Maosong Sun, and Jie Zhou. Moefication: Transformer feed-forward layers are mixtures of experts. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 877–890. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.findings-acl.71. URL https://doi.org/10.18653/v1/2022.findings-acl.71.

Zhengyan Zhang, Zhiyuan Zeng, Yankai Lin, Chaojun Xiao, Xu Han, Zhiyuan Liu, Maosong Sun, and Jie Zhou. Emergent modularity in pre-trained transformers, 2023. URL https://openreview.net/forum?id=XHuQacT6sa6.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*, 2023.