Semantic Robustness of Unsupervised Approaches for Interpretability of Vision Language Models

Tamara Bíla, Adam Červenka, Igor Farkaš

Centre for Cognitive Science, Department of Applied Informatics
Comenius University Bratislava, Slovakia
Mlynská dolina, 84248 Bratislava
tamara.bila@fmph.uniba.sk

Abstract

The problem of interpreteting deep neural networks, especially large transformer-based models, is widely known and has motivated the development of diverse interpretability approaches. A dominant branch of interpretability research focuses on concept discovery in latent spaces. Sparse Autoencoders (SAEs) and clustering methods are two prominent unsupervised concept analysis approaches. However, little is known about the consistency of the structures they recover. In this work, we propose a framework for evaluating the semantic robustness of these methods by applying SAEs and clustering techniques in parallel to the same latent spaces of the Vision-Language Model (VLM) LLaVA. The proposed analysis combines layer-wise overlap measures, monosemanticity scores, and visualization of representative samples and importance heat maps. Our goal is to investigate whether the discovered concepts reflect intrinsic semantic structure or artifacts of the chosen extraction method.

1 Introduction

Modern foundation models are able to process highly rich latent representations, but their structure and the layer-wise progression of activations are hardly interpretable. A promising and popular scheme for understanding latent structure is the so-called concept analysis, aiming to map internal representations to human-understandable semantics. Within the spectrum of available concept extraction methods clustering and Sparse Autoencoders pose as two prominent unsupervised approaches. However, the associated algorithms are fundamentally different.

Clustering groups latent activations according to their similarity, implicitly assuming that semantic structure aligns with geometric proximity. SAEs, on the contrary, learn sparse, overcomplete feature dictionaries, where individual neurons or directions are interpreted as disentangled concepts. While both approaches are used widely, there is limited understanding of whether they recover consistent semantic structures or whether their outputs are method-dependent artifacts.

In our work we investigate the extent to which the discovered concepts are stable and method-invariant across these interpretability techniques. Our proposed method can be summarized in the following steps: 1) applying clustering methods and SAEs in parallel to the same latent spaces of a foundation model LLaVA Li et al. (2024); 2) evaluating layer-wise overlap between clusters and SAE features, together with MonoSemanticity scores of both approaches; 3) visualizing representative samples associated with clusters and SAE-derived concepts in order to investigate whether both methods consistently recover similar semantic structures. Note that the MonoSemanticity score for SAEs was introduced by Pach et al. (2025), whereas we have adapted it and propose an analogous version of this score for clusters in 3.

The key hypothesis of this work can be formulated as follows: If concepts extracted by clustering and SAEs converge, this suggests that the underlying semantic structure is intrinsic to the model. If they diverge, this indicates that at least part of the discovered structure may be an artifact of the chosen method.

2 Related Work

Concept-based interpretability aims to associate latent representations of neural networks with human-understandable semantic concepts. Existing approaches can broadly be divided into supervised and unsupervised methods.

Supervised methods typically rely on predefined concept annotations. Concept Bottleneck Models (CBMs) Koh et al. (2020); Rao et al. (2024) explicitly predict concepts before producing final outputs, while Concept Whitening Chen et al. (2020) aligns latent directions with semantic concepts through fine-tuning a Concept Whitening module. Another influential post hoc approach is TCAV Kim et al. (2018); Schrouff et al. (2022), which estimates concept activation directions in latent space and measures their importance with respect to model predictions. Since TCAV assumes approximate linear separability of concepts, Crabbé and van der Schaar (2022) proposed nonlinear support vec-

tor classifiers to better capture complex latent geometries. Although effective, supervised approaches generally require a datasets annotated with predefined concepts, restraining scalability and generalization.

These limitations motivated increasing interest in unsupervised concept discovery methods. Sparse Autoencoders (SAEs) Huben et al. (2024) learn sparse overcomplete feature dictionaries and are commonly interpreted through the superposition hypothesis Elhage et al. (2022), where individual neurons correspond to disentangled semantic directions. However, SAE training often requires very large datasets, and interpretation of the learned features frequently depends on external language or vision-language models Lieberum et al. (2024); Zhang et al. (2025).

A complementary line of work employs clustering techniques to uncover semantic organization in latent spaces. Prior work demonstrated the effectiveness of unsupervised clustering approaches for discovering meaningful latent structures in both vision and language models Kolouri et al. (2017); Ghorbani et al. (2019); Dalvi et al. (2021). In particular, O'Mahony et al. (2023) showed that hierarchical clustering can disentangle multiple semantic meanings hidden within polysemantic neuron activations. Fel et al. (2023) further proposed a unifying framework interpreting concept extraction as a form of dictionary learning, while Hawasly et al. (2024) analyzed quality–efficiency trade-offs of clustering-based concept discovery methods.

Despite the growing popularity of both SAEs and clustering approaches, relatively little work investigates whether the semantic structures recovered by these methods are consistent across extraction paradigms. Our work addresses this gap by directly comparing clustering-based and SAE-derived concepts within the same latent spaces and by evaluating their robustness under varying data conditions.

3 Methodology

In order to test the robustness of the unsupervised concept extraction approaches, we state a proposition. Particularly, if concepts are encoded in the latent space – manifold – of the target model, they should be discoverable irrespective of the method employed, assuming the used method is faithful to the latent representations. To this end, we implement and test two interpretability approaches, namely clustering-based and SAE concept learning. Specifically, we chose and apply both methods to subset of deep layers in a pretrained Vision-Language Model, such as LLaVA.

Then, we compare the semantic similarity between clusters and SAE features measured by their pointwise overlap, given as the weighted cosine similarity. For the normalized feature activations of the k-th SAE neuron, $\tilde{a}_n^k \in [0,1]$ and the smoothed cluster membership $\tilde{c}_n^j \in [0,1]$ given by the normalized angular distance to j-th centroid the overlap becomes

$$O_{jk} := \frac{\sum_{n=1}^{N} \tilde{c}_n^j \tilde{a}_n^k}{\sqrt{\sum_n (\tilde{c}_n^j)^2 \sum_n (\tilde{a}_n^k)^2}} \tag{1}$$

Note that both SAE feature activations and the cluster memberhip of latent activations refer to the same layer ℓ from which they are extracted. In addition, all activations come from a common dataset $\mathcal{J} = \{\mathbf{x}_n \in \mathbb{X}\}_{n=1}^{N}$, which should span a wide variety of images to include enough concept of diverse granularity. Alternatively, a probability based measure can be defined using the Jensen-Shannon divergence (JSD), which is a symmetrized version of the Kullback-Leibler divergence for two distributions by transforming the normalized activations into probabilities $p_k(n) = \tilde{a}_n^k / \sum_m \tilde{a}_n^k$.

A high overlap between a pair of cluster and SAE neuron indicates a substantial semantic closeness.

The validity of semantic coherence of each SAE feature and cluster can also be tested individually using the MonoSemanticity score introduced by Pach et al. (2025). For the SAE features the score of the k-th neuron becomes

$$\mathrm{MS}_{\mathrm{sae}}^k = \frac{\sum_{1 \le n < m \le N} r_{nm}^k s_{nm}}{\sum_{1 \le n < m \le N} r_{nm}^k} \tag{2}$$

where the relevance matrix $r_{nm}^k = \tilde{a}_n^k \tilde{a}_m^k$ quantifies the joint neuron activation of each image pair.s_{nm} is the pairwise similarity between images computed as the cosine distance of their embeddings from an auxilliary vision encoder E, i.e.

$$s_{nm} = \frac{E(\mathbf{x}_n) \cdot E(\mathbf{x}_m)}{|E(\mathbf{x}_n)||E(\mathbf{x}_m)|}. \tag{3}$$

We formulate an analogous MonoSemanticity metric adapted to clusters, which replaces neuron activations $\tilde{a}_n^k$ by the normalized distance of each activation $\mathbf{a}_n$ in the embedding space from the cluster centroid $\boldsymbol{\mu}_j$:

$$\tilde{d}_n^j = \frac{1 + \cos(\theta)}{2} \quad \text{with} \quad \cos(\theta) = \frac{\mathbf{a}_n \cdot \boldsymbol{\mu}_j}{|\mathbf{a}_n||\boldsymbol{\mu}_j|}. \tag{4}$$

Then the relevance matrix is computed analogously as for the SAE metric, that is, $\rho_{nm}^j = \tilde{d}_n^j \tilde{d}_m^j$. Hence, applying the same logic to construct the MonoSemnaticity score for clusters yields

$$\mathrm{MS}_{\mathrm{cl}}^j = \frac{\sum_{1 \le n < m \le N} \rho_{nm}^j s_{nm}}{\sum_{1 \le n < m \le N} \rho_{nm}^j}. \tag{5}$$

Computing the aforementioned metrics allows a quantitative comparison of qualitatively distinct concept extraction methods. On the one hand, the SAEs are trained in a self-supervised fashion, disentangling the latent signal into an overcomplete dictionary of features using an encoder-decoder like architecture. On

the other side, the clustering-based approaches decompose the latent activations into groups (clusters) according to their pairwise (agglomerative clustering) or point to centroid (K-means clustering) distances. Therefore, finding matching concepts among the SAE features and the clusters strengthens the validity of the individual results and affirms the robustness of both present concepts extraction approaches.

4 Visual Inspection Techniques

Once the overlap between SAE features and clusters is established, the feature-cluster pairs with overlap higher than a set threshold $O_{jk} \leq \tau$ will be selected for introspection and concept description. Presumably, clusters and features whose activations overlap significantly share common input samples and will thus constitute the same (or similar) semantic meaning – concept. Particularly, samples whose activations fire for both a feature and a cluster simultaneously can be considered as the core representatives for the sought-for concept.

Conversely, examining the semantic content of clusters and SAE features independently may reveal discrepancies between the two approaches, thereby exposing inherent differences in the corresponding extraction methods. These differences constitute an important object of study in their own right.

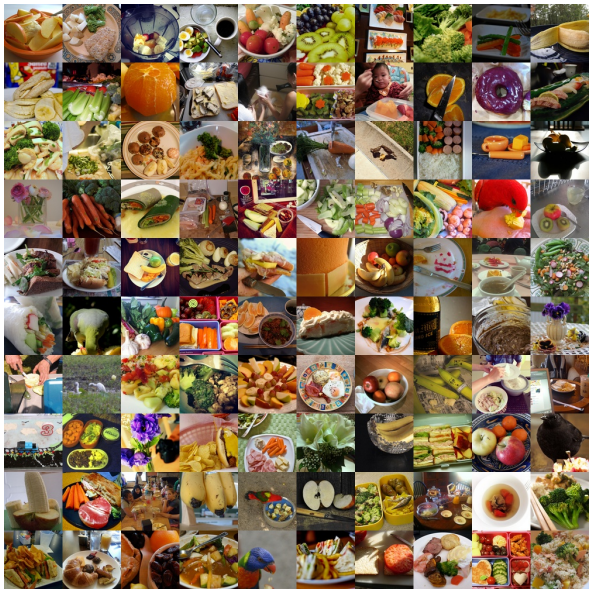

Fig. 1: A collage plot of 100 input images, which yield the highest activation of the 41045th neuron in the SAE layer. The apparent concept common concept among these images is *food* dominated by fruits and vegetables.

To interpret these features, it is necessary to determine which semantic concepts they embody and how these concepts can be mapped to human language. At present, there exist two primary approaches for inspecting and describing their semantic content in natural language. The first, most intuitive yet labor-intensive approach relies on human visual inspection of the corresponding input images. To illustrate this method, we provide an example below. Another possibility is to employ an external Vision-Language Model (VLM) capable of generating textual descriptions for sets of images associated with a given cluster or SAE feature. In this work, we do not further explore the latter approach, although it may serve as an interesting direction for future analysis.

Here, we briefly present the process of visually inspecting the input images associated with the highest activations a_n^k of the k-th neuron in the SAE. Similarly, the semantic content of a cluster can be analyzed by displaying the input samples whose activation vectors $\mathbf{a}_n$ lie closest to the centroid $\boldsymbol{\mu}_j$ of a given cluster. Furthermore, importance heat maps are constructed for a subset of representative samples in order to highlight the input patches that contribute most strongly to the activation of the corresponding neuron.

A composite visualization of the 100 most representative samples associated with the 41 045-th SAE feature, trained on the 18-th layer of LLaVA 1.5 Liu et al. (2024), is shown in Figure 1.

Another useful approach, mentioned above, is the construction of importance maps or heatmaps. Figure 2 presents four heatmaps highlighting the input patches that contribute the most strongly to the activation of the same 41 045-th neuron. Interestingly, the boundaries of the depicted objects, predominantly fruits and vegetables, consistently exhibit the highest contributions to the feature activation.

Fig. 2: Importance heatmap revealing image patches with highest contribution to the SAE feature activation. The four presented images belong to the set of samples with overall highest activation of the 41045th neuron.

5 Conclusion and Future Work

To conclude, we have presented a framework together with quantitative measures for evaluating the robustness of extracted concepts with respect to the choice of unsupervised extraction method. The proposed methodology consists of three main stages. First, Sparse Autoencoders are trained on selected layers of a foundation model, while activations from the same layers are simultaneously analyzed using clustering techniques. Next, the pairwise overlap between SAE features and clusters is evaluated layer-wise, together with MonoSemanticity scores computed for the top N activating samples and the top N samples closest to the corresponding cluster centroids. Finally, representative samples associated with both SAE features and clusters are visualized using collage plots, complemented by importance heat maps highlighting the input regions contributing most strongly to feature activation.

The primary direction of future work is the implementation and large-scale evaluation of the proposed experimental framework. The expected contribution lies in providing empirical evidence regarding the semantic robustness of unsupervised concept discovery methods. In particular, the proposed analysis aims to clarify the relationship between geometric, clustering-based representations and sparse SAE-derived representations. Quantitative overlap and monosemanticity measures may provide practical insight into the reliability of these methods and help determine in which settings one approach may be preferable over the other.

More broadly, improving our understanding of the stability and consistency of unsupervised interpretability techniques may contribute to more reliable and trustworthy deployment of foundation models.

Acknowledgement

This work was supported by the TRAIL project (TRAnsparent, InterpretabLe Robots), funded by Horizon Europe (HORIZON) as a Marie Skłodowska-Curie Action Doctoral Network (MSCA-DN) under the grant agreement No. 101072488.

References

Chen, Z., Bei, Y. and Rudin, C. (2020). Concept whitening for interpretable image recognition. *Nature Machine Intelligence*, 2:1–11.

Crabbé, J. and van der Schaar, M. (2022). Concept activation regions: A generalized framework for concept-based explanations. In *Neural Information Processing Systems*.

Dalvi, F., Khan, A. R., Alam, F., Durrani, N., Xu, J. and Sajjad, H. (2021). Discovering Latent Concepts Learned in BERT. In *International Conference on Learning Representations*, Virtual Event, Austria. Poster at International Conference on Learning Representations (ICLR) 2021.

Elhage, N., Hume, T., Olsson, C., Schiefer, N., Henighan, T., Kravec, S., Hatfield-Dodds, Z., Lasenby, R., Drain, D., Chen, C., Grosse, R., McCandlish, S., Kaplan, J., Amodei, D., Wattenberg, M. and Olah, C. (2022). Toy models of superposition.

Fel, T., Boutin, V., Moayeri, M., Cadène, R., Bethune, L., Andéol, L., Chalvidal, M. and Serre, T. (2023). A holistic approach to unifying automatic concept extraction and concept importance estimation. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Ghorbani, A., Wexler, J., Zou, J. Y. and Kim, B. (2019). Towards Automatic Concept-based Explanations. In *Advances in Neural Information Processing Systems*, vol. 32. Curran Associates, Inc.

Hawasly, M., Dalvi, F. and Durrani, N. (2024). Scaling up Discovery of Latent Concepts in Deep NLP Models. Graham, Y. and Purver, M. (ed.), In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 793–806, St. Julian's, Malta. Association for Computational Linguistics.

Huben, R., Cunningham, H., Smith, L. R., Ewart, A. and Sharkey, L. (2024). Sparse Autoencoders Find Highly Interpretable Features in Language Models. In *The Twelfth International Conference on Learning Representations*, pp. 7827–7845, Vienna, Austria. ICLR.

Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F. and Sayres, R. (2018). Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In *International Conference on Machine Learning*.

Koh, P. W., Nguyen, T., Tang, Y. S., Mussmann, S., Pierson, E., Kim, B. and Liang, P. (2020). Concept Bottleneck Models.

Kolouri, S., Martin, C. E. and Hoffmann, H. (2017). Explaining Distributed Neural Activations via Unsupervised Learning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1670–1678, Honolulu, HI, USA. IEEE.

Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang, H., Zhang, K., Zhang, P., Li, Y., Liu, Z. and Li, C. (2024). LLaVA-OneVision: Easy Visual Task Transfer.

Lieberum, T., Rajamanoharan, S., Conmy, A., Smith, L., Sonnerat, N., Varma, V., Kramar, J., Dragan, A., Shah, R. and Nanda, N. (2024). Gemma scope: Open sparse autoencoders everywhere all at once on gemma 2. Belinkov, Y., Kim, N., Jumelet, J., Mohebbi, H., Mueller, A. and Chen, H. (ed.), In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP of Association for Computational Linguistics,*, pp. 278–300, Miami, Florida, US. Association for Computational Linguistics.

Liu, H., Li, C., Li, Y. and Lee, Y. J. (2024). Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26296–26306.

O'Mahony, L., Andrearczyk, V., Müller, H. and Graziani, M. (2023). Disentangling neuron representations with concept vectors. In *Conference on Computer Vision and Pattern Recognition*, pp. 3770–3775. IEEE.

Pach, M., Karthik, S., Bouniot, Q., Belongie, S. and Akata, Z. (2025). Sparse Autoencoders Learn Monosemantic Features in Vision-Language Models.

Rao, S., Mahajan, S., Böhle, M. and Schiele, B. (2024). Discover-then-name: Task-agnostic concept bottlenecks via automated concept discovery. Leonardis, A., Ricci, E., Roth, S., Russakovsky, O., Sattler, T. and Varol, G. (ed.), In *Computer Vision – ECCV 2024*, pp. 444–461, Cham. Springer Nature Switzerland.

Schrouff, J., Baur, S., Hou, S., Mincu, D., Loreaux, E., Blanes, R., Wexler, J., Karthikesalingam, A. and Kim, B. (2022). Best of both worlds: local and global explanations with human-understandable concepts. arXiv:2106.08641.

Zhang, K., Shen, Y., Li, B. and Liu, Z. (2025). Large multi-modal models can interpret features in large multi-modal models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3650–3661, Honolulu, Hawaii, US. IEEE.