The Multi-faceted Monosemanticity in Multimodal Representations

Hanqi Yan[1], **Yulan He**[1,2], **Yifei Wang**[3*]
[1]King's College London,[2]The Alan Turing Institute
[3]MIT CSAIL
{hanqi.yan,yulan.he}@kcl.ac.uk
yifei_w@mit.edu

Abstract

In this paper, we leverage recent advancements in feature monosemanticity to extract interpretable features from deep multi-modal models, offering a data-driven understanding of modality gaps. Specifically, we investigate CLIP (Contrastive Language-Image Pretraining) (Radford et al., 2021), a prominent visual-language representation model. Building upon interpretability tools developed for single-modal models, we extend these methodologies to assess the *multi-modal interpretability* of CLIP's features. Additionally, we introduce the Modality Dominance Score (MDS) to attribute the interpretability of each feature to its respective modality. Next, we transform CLIP's features into a more interpretable space, enabling us to categorize them into three distinct classes: vision features, language features (both single-modal), and visual-language features (cross-modal). Our findings reveal that this categorization aligns closely with human cognitive understandings of different modalities. These results indicate that large-scale multi-modal models, equipped with advanced interpretability tools, offer valuable insights into the key connections and distinctions between different data modalities. This work not only bridges the gap between cognitive science and machine learning but also introduces new data-driven tools to advancing both fields.

1 Introduction

Multi-modal models have become foundational in the development of artificial intelligence systems, enabling the processing and understanding of information from multiple data modalities, such as vision and language (Radford et al., 2021; Kim et al., 2021; Lu et al., 2019). These models are built on the premise that different data modalities share common, or cross-modal, features that can be jointly learned (Ngiam et al., 2011). However, it is widely acknowledged that certain features are modality-specific; for example, some emotions are difficult to visualize, while certain visual experiences cannot be accurately described through language (Paivio, 1991).

The exploration of modality commonality and gaps has long been a focus in cognitive science, where researchers have investigated how humans integrate and differentiate information across sensory modalities (Spence, 2011). However, these studies are often human-centric and may not directly translate to artificial systems due to fundamental differences in how information is processed and represented (Calvert et al., 2004). Meanwhile, recent advances in interpretability methods, particularly in the area of monosemantic features, providing a promising path towards a more detailed understanding of deep models (Elhage et al.; Bills et al., 2023; Gurnee et al., 2023; Yan et al., 2024). Monosemantic features/neurons refer to model components that correspond to a single, interpretable

*Corresponding author

Workshop on Responsibly Building the Next Generation of Multimodal Foundational Models
38th Conference on Neural Information Processing Systems (NeurIPS 2024).

concept or feature. By leveraging these methods, we can extract monosemantic, interpretable features from deep learning models, providing a data-driven approach to exploring modality gaps.

In this paper, we focus on CLIP (Contrastive Language-Image Pretraining) (Radford et al., 2021), a visual-language representation model trained on massive image-text pairs. We investigate the modality association of features extracted from CLIP by introducing a modality metric that categorizes these interpretable features into: vision, language and visual-language features.

Our study reveals that single-modal features align well with human cognition and highlight diverse aspects of the visual-language modality gap. We find that visual-language features capture modality-aligned semantics. These findings suggest that interpretability tools can enable deep models to provide a systematic understanding of the similarities and distinctions between different modalities.

2 Towards Multi-modal Monosemanticity

In this section, we build a pipeline to extract monosemantic multi-modal features and evaluate interpretability the these features. We also characterize the **modality relevance** in extracted features with the proposed Monosemantic Relevance Score (MRS).

We consider two CLIP models, i.e., canonical ViT-B-32 CLIP model by OpenAI (Radford et al., 2021) and a popular CLIP variant, DeCLIP (Li et al., 2022). Beyond multi-modal supervision (image-text pairs), DeCLIP also integrates single-modal self-supervision (image-image pairs and text-text pairs) for more efficient joint learning. We hypothesize that, with the incorporation of self-supervision tasks, DeCLIP is able to extract more single-modal features from the data, enhancing its interpretability and alignment with modality-specific characteristics.

2.1 Interpretability Tools for Multi-modal Monosemantic Feature Extraction

The features in deep models are observed to be quite *polysemantic* (Olah et al., 2020), in the sense that activating samples along each feature dimension often contain multiple unrelated semantics. Therefore, we first need to disentangle the CLIP features to have *monosemantic features*. Borrowing from the recent progress in monosemanticity in self-supervised models, we study the two methods to attain better multi-modal monosemanticity.

Multi-modal SAE. Sparse Autoencoders (SAEs) (Cunningham et al., 2023) are a new scalable interpretability method, demonstrating success in multiple large language models (LLMs) (Templeton, 2024; Gao et al., 2024; Lieberum et al., 2024). Here, we train a *multi-modal SAE (MSAE)* g^+ by training **one** SAE model to reconstruct both image and text representations. Specifically, we adopt a top-K SAE model (Makhzani & Frey, 2013; Gao et al., 2024), and train it with a *multi-modal reconstruction objective*. In this way, the sparse latent feature $z \in \mathbb{R}^n$ can encode multi-modal representations from both modalities.

Multi-modal NCL. Inspired by the interpretable self-supervised loss with non-negative constraint (NCL) proposed by (Wang et al., 2024) to extract sparse features, we adapt it to enhance multi-modal interpretability. A shared MLP network (of similar size to SAE) on top of the encoder outputs is trained by a *Multi-modal NCL* loss.

2.2 Measures for Multi-modal Interpretability

Existing quantitative interpretability measures (Bills et al., 2023) often require access to high pricing models (like GPT-4o) and suffer from poor scalability and poor precision (Gao et al., 2024), damping the progress of open science. It motivates us to propose scalable measures as below.

Embedding-based Similarity. We propose a scalable measure based on embedding models that work for both image and text.[2] For each image/text feature z, we select the top m activated image/text samples on this dimension, denoting their embeddings as $Z_+ \in \mathbb{R}^{m\times d}$; similarly, K random samples are encoded into $Z_- \in \mathbb{R}^{m\times d}$ as the baseline. Then, we calculate the inter-sample similarity between the selected samples, $S_+ = Z_+ Z_+^\top \in \mathbb{R}^{m\times m}$ and $S_- = Z_- Z_-^\top \in \mathbb{R}^{m\times m}$. Then we measure the monosemanticity degree of z by calculating the relative difference between the two

[2]We use Vision Transformer (ViT-B-16-224-in21k) for image embeddings and Sentence Transformer (all-MiniLM-L6-v2) for text embeddings.

similarity scores: $I(z) = \frac{1}{m(m-1)} \sum_{i \neq j} \frac{(S_+)_{ij} - (S_-)_{ij}}{(S_-)_{ij}}$. A larger score indicates that the extracted features have more consistent semantics on average.

WinRate. Since the representations obtained from different embedding models (e.g., vision and text) are not directly comparable, we propose similarity WinRate, a binary version of the relative similarity score, by counting the percentage that the elements in S_+ is larger than that in S_-: $W(z) = \frac{1}{m(m-1)} \sum_{i \neq j} \mathbf{1}_{[(S_+)_{ij} > (S_-)_{ij}]}$.

Model	Similarity		WinRate		
	Image	Text	Image	Text	$\lvert\Delta\rvert$(img − txt)
CLIP	0.113	0.451	0.652	0.594	0.058
DeCLIP	0.058	-0.073	0.615	0.457	**0.158**
CLIP+NCL	**0.161**	**0.592**	**0.727**	**0.608**	**0.119**
CLIP+SAE	**0.120**	0.244	**0.667**	0.540	**0.127**

Table 1: The average interpretability scores for features extracted from the four models. A larger $|\Delta|$ represents that the features are more aligned with a single modality.

Results. From the results of interpretability distribution of the features extracted in Table 1, we observe: (1) Features in NCL have the overall best monosemanticity (2) Compared to CLIP, all other models have more single-modality aligned features. We evaluate the average interpretability, regardless of their predominant modality. Next, we split the neurons into different groups and study the modality-specific features.

2.3 Grouping Modality in Multi-modal Representations

Modality Dominance Score (MDS). We propose a metric to assert the predominant modality of each neuron. Specially, we feed m input-output pairs to CLIP and obtain the image features $Z_I \in \mathbf{R}^{m \times n}$ and text features $Z_T \in \mathbf{R}^{m \times n}$. For each feature $k \in [1, \ldots, n]$, we calculate the relative activation between image and text features over the m inputs, i.e., $R(k) = \frac{1}{m} \sum_{i=1}^{m} \frac{(Z_I)_{ik}}{(Z_I)_{ik} + (Z_T)_{ik}}$. The ratio $R(k)$ reflects the ratio of the kth feature k being activated in the image modality. Based on this value, we split all n features into three groups according to their dominant modality, i.e., $sigma$ of the distribution: `ImgD (Image Dominant)`:$r_i > \mu + \sigma$; `TextD (Text Dominant)`: $r_i < \mu - \sigma$; `CrossM (Cross Modality)`: $\mu - \sigma < r_i < \mu + \sigma$.

We anticipate that `ImgD` features are mostly activated by images and `TextD` features by text, while `CrossM` features are *simulatenously* activated by both image and text when paired.

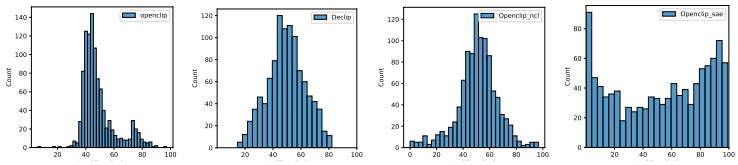

Figure 1: MDS distributions for different Language-Vision Models. Left to right: CLIP, DeCLIP, CLIP+NCL, CLIP+SAE.

Results of MDS in Fig 1. Interestingly, we find that CLIP, which is only trained on image-text paired contrastive learning objective, also contains a spectrum of features with different modality dominance. DeCLIP features are less skewed and less centered, showing better coverage of both image-dominant, text-dominant, and cross-model features. Therefore, it demonstrates from a mechanistic interpretability perspective that self-supervision extract more modality-specific features that might be overlooked by pure visual-language contrastive models like CLIP. The extracted features from NCL and SAE are less skewed. SAE has a more balanced distribution, indicating its better capability of extracting diverse monosemantic features.

3 Understanding Multi-modal Features in Different Modality Groups

With the protocol developed above, we have separated the the neurons into three groups, which allows for a deeper quantitative and qualitative understanding of the connections and gaps between different modalities in data-driven approach.

3.1 Multi-modal Interpretability

Modality	CLIP	DeCLIP	CLIP+NCL	CLIP+SAE
Image	0.118	0.070	**0.197**	0.135
Text	-0.07	-0.059	0.132	**0.439**

Table 2: The visual and textual monosemanticity.

A major implication of modality domain is its influence on feature interpretability at different modalities. Ideally, when fed with image samples, `ImgD` neurons should be more effective at capturing concrete and consistent features than `TextD` neurons. Similar for input text samples.

Therefore, we measure both visual and textual monosemanticity. Specially, for image inputs, we calculate the *visual monosemanticity* by comparing the interpretability between `ImagD` and `TextD`, i.e., EmbedSimi(`ImgD`)-EmbedSimi(`TextD`); for text inputs, we calculate *textual monosemanticity* via EmbedSimi(`TextD`)-EmbedSimi(`ImgD`). We have the following observations from Table2: (1) On image input, except for CLIP, all the other three models demonstrate positive visually monosemanticity than the other two types of neurons. (2) On text input, both NCL and SAE capture better monosemantic textual features than the other two models. (3) SAE is the best in capturing both visual and textual monosemantic features.

3.2 Case Studies

In addition to quantification of the interpretability of neurons dominating different modalities, we look closer to a few examples of captured features. **ImgD neurons capture both coarse and fine-grained visual features.** Among `ImgD` neurons, we randomly select two neurons and display the top5 activated images in Figure 2. The activated images can be concrete concept, i.e., the inner living space, also the patterns and textures that are the basic patterns learnt in the lower layers by image-only model. Moreover, Neuron668 is a monosemantic neuron to blue color. **TextD neurons capture abstract and semantic information that is less prevalent in visual representations.** We randomly select three neurons and display the top4 activated sentences in Table 3. Neuron45 focuses on strong emotions by highlighting the text with **?, !** and strong emotion words. Neuron932 focuses a happy and warm atmosphere, even though there are no common visual concepts.

Figure 2: Activated images by `ImgD` neurons. Top to bottom: Patterns and textures; Water and aquatic themes.

`Neuron 45: Strong affection`
sinkhole,**most terrifying thing I have ever seen.**
Alligators: what's in my bag**?** gloves I need **!**
i never have to paint a mural again ! =)
`Neuron 932: Moments of joy, warmth`
Couple **kissing** in a gazebo.
Man with red jumper stand by a **Christmas tree**.
Funny summer background with the little girl.

Table 3: Activated sentences by `TextD` neurons.

Cross-Modality neurons capture common features between image and text. Different from the `TextD` and `ImgD`, whose activated samples tend to contain modality-exclusive features, `CrossM` neurons are more capable in capturing common features shared by the two modalities. We randomly select two `CrossM` neurons and display their top activated images and texts, shown in Figure 3 and Table 4. It is clear that Neuron6 activate male action related concepts in both modality and Neuron47 activate outdoor scenes.

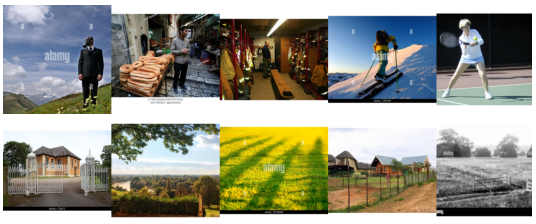

Figure 3: Activated images by `CrossM` neurons. Top to bottom: individuals engaged in various activities; outdoor scenes.

`Neuron 6: Actions performed by males`
Young man working on invention in a warehouse
Cricket player checks his bat in a training
Handsome man walks on ruins
`Neuron 47: Outdoor scenes`
A private chapel in the **grounds**
An open gate in a **meadow**
People look out over loch in the **village**

Table 4: Activated samples by the same set of `CrossM` neurons.

4 Conclusion

In this study, we explored the monosemanticity of features within the CLIP model to elucidate the commonalities and distinctions across visual and linguistic modalities. We successfully categorized interpretable features according to their predominant modality, which demonstrate close correspondence to human cognitive interpretations. Future work may extend these methodologies to other multi-modal architectures and investigate their implications for cognitive science, ultimately fostering the development of more interpretable and cognitively aligned AI systems.

Acknowledgements

This work was supported in part by the UK Engineering and Physical Sciences Research Council through a Turing AI Fellowship (grant no. EP/V020579/1, EP/V020579/2). YW was funded by Office of Naval Research grant N00014-20-1-2023 (MURI ML-SCOPE), NSF Award CCF-2112665 (TILOS AI Institute), and an Alexander von Humboldt Professorship.

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)*, 2, 2023.

Gemma Calvert, Charles Spence, and Barry E Stein (eds.). *The Handbook of Multisensory Processes*. MIT Press, 2004.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, R. Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. *International Conference on Learning Representations*, 2023. doi: 10.48550/arXiv.2309.08600.

Nelson Elhage, Neel Nanda, Catherine Olsson, and Others. A mathematical framework for transformer circuits. *Transformer Circuits Thread (2022)*. URL `https://transformer-circuits.pub/2022/solu/index.html`.

Leo Gao, Tom Dupré la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders. *arXiv preprint arXiv:2406.04093*, 2024.

Wes Gurnee, Neel Nanda, Matthew Pauly, Katherine Harvey, Dmitrii Troitskii, and Dimitris Bertsimas. Finding neurons in a haystack: Case studies with sparse probing. *ArXiv*, abs/2305.01610, 2023. URL `https://api.semanticscholar.org/CorpusID:258437237`.

Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. *arXiv preprint arXiv:2102.03334*, 2021.

Yangguang Li, Feng Liang, Lichen Zhao, Yufeng Cui, Wanli Ouyang, Jing Shao, Fengwei Yu, and Junjie Yan. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. In *International Conference on Learning Representations*, 2022. URL `https://openreview.net/forum?id=zq1iJkNk3uN`.

Tom Lieberum, Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Nicolas Sonnerat, Vikrant Varma, János Kramár, Anca Dragan, Rohin Shah, and Neel Nanda. Gemma scope: Open sparse autoencoders everywhere all at once on gemma 2. *arXiv preprint arXiv:2408.05147*, 2024.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, pp. 13–23, 2019.

Alireza Makhzani and Brendan Frey. K-sparse autoencoders. *arXiv preprint arXiv:1312.5663*, 2013.

Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. Multimodal deep learning. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pp. 689–696, 2011.

Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. Zoom in: An introduction to circuits. *Distill*, 5(3):e00024–001, 2020.

Allan Paivio. Dual coding theory: Retrospect and current status. *Canadian Journal of Psychology/Revue canadienne de psychologie*, 45(3):255, 1991.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021.

Charles Spence. Crossmodal correspondences: A tutorial review. *Attention, Perception, & Psychophysics*, 73(4):971–995, 2011.

Adly Templeton. *Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet*. Anthropic, 2024.

Yifei Wang, Qi Zhang, Yaoyu Guo, and Yisen Wang. Non-negative contrastive learning. *ICLR*, 2024.

Hanqi Yan, Yanzheng Xiang, Guangyi Chen, Yifei Wang, Lin Gui, and Yulan He. Encourage or inhibit monosemanticity? revisit monosemanticity from a feature decorrelation perspective. *ArXiv*, abs/2406.17969, 2024. URL `https://api.semanticscholar.org/CorpusID:270737676`.