REPRESENTATIONAL DIFFERENCE CLUSTERING

Neehar Kondapaneni[1] **Emily Gu**[1] **Oisin Mac Aodha**[2] **Pietro Perona**[1]
[1]Caltech [2]University of Edinburgh

ABSTRACT

Current computer vision models are approaching superhuman performance on visual categorization tasks in domains such as ecology, radiology, etc. Explainable AI (XAI) methods aim to explain how such models make decisions. Unfortunately, in order to make explanations that are human-friendly, XAI methods can often simplify model behavior to the point that critical information is lost. For humans to learn how model's achieve superhuman performance, we must work towards understanding these nuances. In this work, we consider the challenging task of visually explaining the differences between two representations. By nature, this task forces XAI methods to discard coarse-grained, obvious aspects of a model's representation to focus on nuances that make a model unique. To this end, we propose a clustering method that is able to isolate neighborhoods of images that are close together in one representation, but distant in the other. These discovered clusters represent concepts that are present in only one of the two representations. We use our method to compare different model representations and discover semantically meaningful clusters.

1 INTRODUCTION

Explaining model decisions is incredibly important for the safe deployment and use of AI models (Kop, 2021). Explainable methods for AI (XAI) aim to organize, simplify, and visualize a model's reasoning process so that humans can interpret it. Recent evaluations have shown that XAI methods improve human understanding, but there is still significant room for improvement (Achtibat et al., 2023; Fel et al., 2023b; Colin et al., 2022; Shen & Huang, 2020; Nguyen et al., 2021; Sixt et al., 2022; Kim et al., 2022). Broadly speaking, post-hoc XAI methods are faced with a trade-off: simplifying model behavior helps humans understand a model's predictions, but reduces faithfulness to the original model (Fel et al., 2023b; Kondapaneni et al., 2024; Cunningham et al., 2024). To understand how AI models achieve superhuman performance in critical tasks (e.g., in medical image analysis), we must be able to explain the intricacies of model behavior. We consider the problem of explaining the difference between two vision model representations. Solving this task necessitates the development of methods that can attend to model differences. This approach highlights nuanced, fine-grained aspects of a model's representation, since obvious concepts are likely to be shared by both models.

Representational Similarity. Our task is closely related to measuring "representational similarity", in which methods provide a single score to quantify similarity (Hotelling, 1936; Kornblith et al., 2019; Raghu et al., 2017; Li et al., 2015; Huh et al., 2024). These scores provide a coarse insight into differences between architectures, pre-training methods, etc. (Nguyen et al., 2021; Raghu et al., 2021; Xie et al., 2023; Neyshabur et al., 2020; Park et al., 2024; Zhang et al., 2020). Somewhat also related is work that aims to discover similar neurons across models and visualize the features they encode (Dravid et al., 2023). However, none of these methods provide *fine-grained* explanations for the nuanced differences between a pair of models.

Explainable AI (XAI). XAI methods in vision can be broadly grouped into local, global, and glocal methods. Local explanations provide users with attribution maps that identify regions of the input image that the model uses in its decision (Selvaraju et al., 2020; Ribeiro et al., 2016; Lundberg & Lee, 2017). Global explanations provide users with a collage of images that represent a learned semantic concept Kim et al. (2018); Ghorbani et al. (2019); Zhang et al. (2021); Fel et al. (2023a); Kowal et al. (2024); Poeta et al. (2023); Bau et al. (2020). Glocal methods fuse both approaches, both localizing *where* and *what* a model is using to make its decision Schrouff et al. (2021); Fel et al. (2023b); Achtibat et al. (2023); Kondapaneni et al. (2024). However, these methods are not explicitly designed to compare two models.

Explainable Representational Similarity. Recently, Kondapaneni et al. (2025) proposed a method called RSVC in which they generate concept-based explanations for two models and compare the explanations. While RSVC can surface some concepts that are unique to a single model, they find that many concepts are partially related, making it challenging to interpret model differences. In this work, we consider an approach that uses information from *both* models to isolate clusters of images that exist in one model, but do not exist in the other. This approach makes it easier to clearly isolate and visualize model differences. We visualize our clusters using a representative image grid, which have been shown to be human-friendly and interpretable (Fel et al., 2023b; Kim et al., 2018; Ghorbani et al., 2019). Our approach is also related to DiSC (Sristi et al., 2022), a method that discovers clusters of features that differentiate data collected from two experimental conditions but share a feature space. In contrast to DiSC, our approach is concerned with identifying clusters of images which can come from different models with different feature spaces. In addition, DiSC directly modifies the spectral clustering objective to discover differentiating feature clusters. In contrast, our approach is based on defining an affinity matrix and thus can be used flexibly with different clustering algorithms.

2 METHOD

As input we take two embedding matrices obtained from two different models, $\boldsymbol{X} \in \mathbb{R}^{n \times d_X}$ and $\boldsymbol{Y} \in \mathbb{R}^{n \times d_Y}$, where d_X and d_Y are the embedding dimensions for models X and Y respectively, and n is the total number of images. $\boldsymbol{X}$ and $\boldsymbol{Y}$ contain embeddings over the same set of images, where each row corresponds to the same image. We propose a method to identify clusters present in $\boldsymbol{X}$ but are absent in $\boldsymbol{Y}$. To do so, we construct an affinity matrix that assigns high affinity to images that are close in $\boldsymbol{X}$ but distant in $\boldsymbol{Y}$, and perform clustering to reveal the distinctive structure in $\boldsymbol{X}$.

At a high-level, our Representational Difference Clustering (RDC) approach performs the following steps: (1) compute the pairwise distances between images in $\boldsymbol{X}$ and $\boldsymbol{Y}$ to build fully connected graphs, $\boldsymbol{G}_X$ and $\boldsymbol{G}_Y$, (2) compute a *normalized* difference between the graphs to form $\boldsymbol{G}_\Delta$, and (3) convert the difference into an affinity graph ($\boldsymbol{G}_A$) and apply a clustering algorithm. Intuitively, negative edges in $\boldsymbol{G}_\Delta$ indicate that the corresponding pair of images were closer together in $\boldsymbol{X}$ than they were in $\boldsymbol{Y}$. We provide details in the following sections.

2.1 GRAPH CONSTRUCTION

We compute the pairwise Euclidean distances for each embedding matrix, $\boldsymbol{D}_X \in R^{n \times n}$ and $\boldsymbol{D}_Y \in R^{n \times n}$. We define $\boldsymbol{G}_X$ to be the graph with adjacency matrix $\boldsymbol{D}_X$. To construct the normalized difference graph $\boldsymbol{G}_\Delta$ from $\boldsymbol{G}_X$ and $\boldsymbol{G}_Y$, we must make the distances in each representation comparable. Since our final goal is to visualize an image collage representing a cluster, we must preserve the relative positions of images. A natural choice is to use a scale-invariant neighborhood ranking where the edge weight between i and j is the nearest neighbor rank of j. Specifically, we define the neighbor ranking graph as:

$$\boldsymbol{N}_X^{i,:} = \text{argsort}(\text{argsort}(\boldsymbol{D}_X^{i,:})) + 1, \tag{1}$$

where $i,:$ indicates the i^{th} row of $\boldsymbol{D}_x$. When using the neighborhood ranking, vertices in $\boldsymbol{G}_X$ and $\boldsymbol{G}_Y$ are considered similar if they have similar neighborhood rankings. We prioritize differences in nearby neighbors by dividing by the minimum neighbor rank. This ensures that large differences in distant neighbors are ignored, but large differences in nearby neighbors are emphasized. To avoid exponential growth in our difference function when a neighbor rank is small, we apply a tanh function to normalize the outputs:

$$\boldsymbol{G}_\Delta^{ij} = \tanh(\gamma \cdot (\boldsymbol{N}_X^{ij} - \boldsymbol{N}_Y^{ij})/(\min(\boldsymbol{N}_X^{ij}, \boldsymbol{N}_Y^{ij})). \tag{2}$$

The γ parameter controls how quickly the function saturates. Given an image i, this function will output negative values when the neighbor rank between i and j is smaller in $\boldsymbol{X}$ than in $\boldsymbol{Y}$. Thus, negative values in this matrix indicate images that are closer in $\boldsymbol{X}$ than in $\boldsymbol{Y}$. Finally, we convert $\boldsymbol{G}_\Delta$ into affinities with:

$$\boldsymbol{G}_A = \exp(-\beta \cdot \boldsymbol{G}_\Delta). \tag{3}$$

Since neighborhood rankings are not symmetric, we symmetrize the affinity graph adjacency matrix by averaging with its transpose.

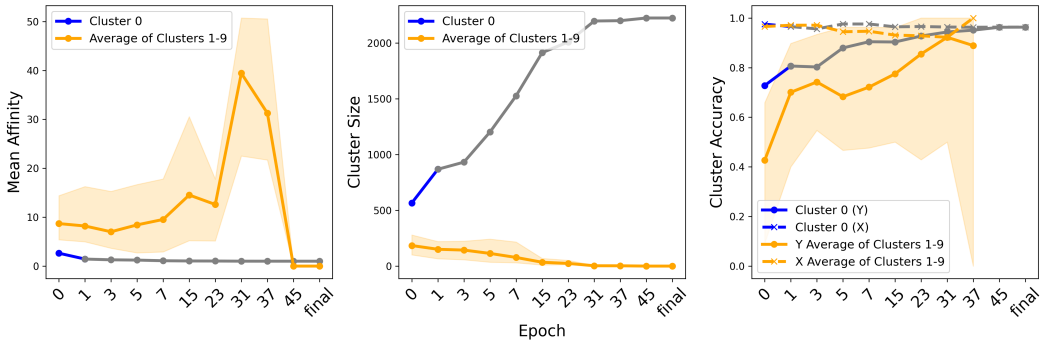

Figure 1: **Difference clusters when comparing model checkpoints**. Here we compare checkpoints from different points in time from a ResNet-18 trained on the Butterflies dataset against the final representation obtained. Cluster 0 is a "special" cluster since it contains the images that change the least between two representations. We compare cluster 0 to the average of clusters 1-9 and visualize the minimum and maximum value with the shaded region. *(Left)* We see that cluster 0 decreases in mean affinity over time and transitions to a "no-change" cluster (gray) at checkpoint 1 (C1). In contrast, clusters 1-9 increase in mean affinity until checkpoint 45 (C45), when the cluster size for 1-9 becomes zero. *(Middle)* The cluster size increases steadily for cluster 0 as the representations become more similar. *(Right)* Both the training checkpoints and final checkpoint are more accurate on images in cluster 0. Performance within clusters 1-9 is much more variable, indicating that the clusters capture images that the model is less certain about.

2.2 Clustering

Normalized cuts (N-Cut) seek out partitions of a graph that minimizes the sum of the cut edges, while balancing the size of the partition (Shi & Malik, 2000). Spectral clustering solves a relaxed version of the N-Cut problem (Von Luxburg, 2007) given an affinity matrix. Since, edges in $\boldsymbol{G}_A$ are large when images are closer in $\boldsymbol{X}$ than they are in $\boldsymbol{Y}$, the clustering algorithm is biased to finding partitions in which images are close together in $\boldsymbol{X}$, but far apart in $\boldsymbol{Y}$. In practice, if representational similarity between $\boldsymbol{X}$ and $\boldsymbol{Y}$ is high, edges in $\boldsymbol{G}_A$ will be close to 1, since the value in $\boldsymbol{G}_\Delta$ will be near zero. In this setting, spectral clustering will tend to return some clusters with average affinities close to 1. We apply a simple post-processing step, such that clusters with an average affinity below 1.5 are considered part of a "no-change" group. Finally, we order the clusters by their mean within-cluster affinity (low to high).

3 Results

We train a randomly initialized ResNet-18 model for 50 epochs on the butterflies dataset from Mac Aodha et al. (2018) and save checkpoints during training. This dataset contains five butterfly species, in which three are challenging for humans to distinguish. We use RDC to discover 10 clusters that exist in the final checkpoint (CF), but do not exist in the compared checkpoint (CY). Since RDC is not symmetric, we set $\boldsymbol{X}$ to be the embedding matrix for CF and $\boldsymbol{Y}$ to be the output of CY. Our probe set is the entire training set ($\sim$ 2000 images). When comparing training checkpoints to their final representation, we expect that the no-change cluster becomes larger as the two representations become more similar. Additionally, as the model improves, we expect to discover clusters of images for which the model is more likely to make errors.

In Fig. 1, we explore the mean affinity, cluster size, and cluster accuracy of cluster 0 vs. the average of clusters 1-9. At checkpoint 1 (C1), the mean affinity of cluster 0 drops below 1.5 and it becomes the no-change cluster (gray color). In the middle panel, we can see that cluster 0 grows in size over time, indicating that our RDC method correctly detects that the two representations are becoming more similar. It also appears that clusters 1-9 have lower average accuracy than cluster 0. This indicates that clusters are forming on images in which the model is still refining its predictions. In sum, we find that our method accurately identifies the growing similarity between the checkpoints and the final representation and that clusters are localized to regions of greater uncertainty.

In Fig. 2, we take a closer look at the differences between the checkpoints. We highlight C0 vs. CF, C5 vs. CF, and C23 vs. CF. For each comparison, we visualize the model representations using a

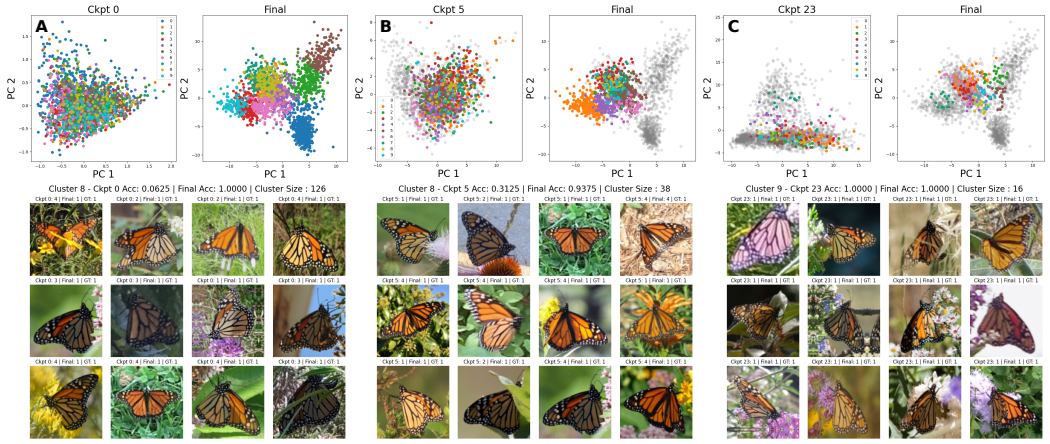

Figure 2: **Cluster visualizations**. We compare checkpoint 0 (C0), C5, and C23 against the final representation (CF). We project the high-dimensional representations into a PCA space and visualize clusters discovered by RDC. We seek clusters that exist in CF, but not in CY. When comparing PCA plots, CY tends to have more distant images within a cluster, whereas the same images will be tightly grouped in CF. *(A)* At C0, there is a large difference between the two representations and the lowest affinity cluster, cluster 0, is still above the "no-change" threshold. We visualize a cluster that shows diverse images of Monarch butterflies. In the cluster title, we show that C0 has low accuracy on these images. *(B)* At C5, the model has greater similarity to CF and many images belong to the no-change cluster (gray). We visualize a cluster that shows C5 is still weak at classifying Monarch butterflies. *(C)* At C23, the representations are quite similar and most images belong to the no-change cluster. We visualize a cluster of Monarch butterflies that are classified well by both models. In this case, our method is identifying a semantic grouping in CF that does not impact overall performance. This collage seem to have images of Monarchs that are more occluded or blurry.

PCA projection and a subset of images from one discovered cluster. In each panel, the left subplot contains CY, the right subplot contains CF and the bottom contains the image collage. We color points according to the difference cluster discovered by RDC. We can see that points within a cluster are closer together in CF than they are in CY, indicating that RDC is successfully identifying regions of high similarity in CF that have low similarity in CY. Additionally, we can see that cluster 0 transitions from a difference cluster (blue) to a no-change cluster (gray). The no-change cluster grows in size from C5 to C23 indicating that representation is becoming more similar. Finally, clusters in C5 form around images that are from classes with lower accuracy. C5 accuracies for the five classes are 0.962, 0.709, 0.954, 0.955, 0.803 respectively and 99% of the images in clusters 1-9 belong to class 1, 2, or 4. which are known to be the most commonly confused (Mac Aodha et al., 2018). We visualize clusters from each comparison that are semantically similar. We find that these clusters correspond to Monarch butterfly images that are close together in CF but far apart in CY. This is sensible, since, C0 and C5 are much worse at identifying Monarch butterflies than CF. Finally, in panel C, we show that when models are highly similar (C23), RDC is able to capture more fine-grained, nuanced clusters. In this cluster we identify a semantic grouping of images in CF that appears to organize Monarch images that are blurry or occluded. Interestingly, C23 has no trouble classifying these images correctly, even though they are not embedded as near each other.

4 Discussion

We propose Representational Difference Clustering (RDC), a new approach for identifying clusters that are present in one representation, but not the other. These clusters can be visualized through human-friendly image collages that convey differences in model behavior (Fig. 2 A,B). Additionally, by comparing model representations over training, we can focus our explanations on increasingly nuanced and complex patterns (Fig. 2 C). In the future, this approach can be used to contrast representations learned from human similarity judgments against superhuman AI models. This may help us discover patterns/groupings that have been learned by the model that humans do not observe. By teaching these patterns back to human subjects, humans can better understand models and improve their own understanding of the visual world.

REFERENCES

Reduan Achtibat, Maximilian Dreyer, Ilona Eisenbraun, Sebastian Bosse, Thomas Wiegand, Wojciech Samek, and Sebastian Lapuschkin. From attribution maps to human-understandable explanations through concept relevance propagation. *Nature Machine Intelligence*, 2023.

David Bau, Jun-Yan Zhu, Hendrik Strobelt, Agata Lapedriza, Bolei Zhou, and Antonio Torralba. Understanding the role of individual units in a deep neural network. *PNAS*, 2020.

Julien Colin, Thomas Fel, Rémi Cadène, and Thomas Serre. What i cannot predict, i do not understand: A human-centered evaluation framework for explainability methods. *NeurIPS*, 2022.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *ICLR*, 2024.

Amil Dravid, Yossi Gandelsman, Alexei A Efros, and Assaf Shocher. Rosetta neurons: Mining the common units in a model zoo. In *ICCV*, 2023.

Thomas Fel, Victor Boutin, Louis Béthune, Rémi Cadène, Mazda Moayeri, Léo Andéol, Mathieu Chalvidal, and Thomas Serre. A holistic approach to unifying automatic concept extraction and concept importance estimation. *NeurIPS*, 2023a.

Thomas Fel, Agustin Picard, Louis Bethune, Thibaut Boissin, David Vigouroux, Julien Colin, Rémi Cadène, and Thomas Serre. CRAFT: Concept recursive activation factorization for explainability. In *CVPR*, 2023b.

Amirata Ghorbani, James Wexler, James Y Zou, and Been Kim. Towards automatic concept-based explanations. *NeurIPS*, 2019.

Harold Hotelling. Relations between two sets of variates. In *Biometrika*, 1936.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The platonic representation hypothesis. In *ICML*, 2024.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TACV). In *ICML*, 2018.

Sunnie SY Kim, Nicole Meister, Vikram V Ramaswamy, Ruth Fong, and Olga Russakovsky. Hive: Evaluating the human interpretability of visual explanations. In *ECCV*, 2022.

Neehar Kondapaneni, Markus Marks, Oisin Mac Aodha, and Pietro Perona. Less is more: Discovering concise network explanations. In *ICLR Workshop on Representational Alignment*, 2024.

Neehar Kondapaneni, Oisin Mac Aodha, and Pietro Perona. Representational similarity via interpretable visual concepts. In *ICLR*, 2025.

Mauritz Kop. Eu artificial intelligence act: the european approach to ai. Stanford-Vienna Transatlantic Technology Law Forum, Transatlantic Antitrus and IPR Developments, 2021.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *ICML*, 2019.

Matthew Kowal, Achal Dave, Rares Ambrus, Adrien Gaidon, Konstantinos G Derpanis, and Pavel Tokmakov. Understanding video transformers via universal concept discovery. In *CVPR*, 2024.

Yixuan Li, Jason Yosinski, Jeff Clune, Hod Lipson, and John Hopcroft. Convergent learning: Do different neural networks learn the same representations? In *International Workshop on Feature Extraction: Modern Questions and Challenges at NeurIPS*, 2015.

Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *NeurIPS*, 2017.

Oisin Mac Aodha, Shihan Su, Yuxin Chen, Pietro Perona, and Yisong Yue. Teaching categories to human learners with visual explanations. In *CVPR*, 2018.

Behnam Neyshabur, Hanie Sedghi, and Chiyuan Zhang. What is being transferred in transfer learning? *NeurIPS*, 2020.

Thao Nguyen, Maithra Raghu, and Simon Kornblith. Do wide and deep networks learn the same things? uncovering how neural network representations vary with width and depth. In *ICLR*, 2021.

Young-Jin Park, Hao Wang, Shervin Ardeshir, and Navid Azizan. Quantifying representation reliability in self-supervised learning models. In *UAI*, 2024.

Eleonora Poeta, Gabriele Ciravegna, Eliana Pastor, Tania Cerquitelli, and Elena Baralis. Concept-based explainable artificial intelligence: A survey. *arXiv:2312.12936*, 2023.

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. *NeurIPS*, 2017.

Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks? *NeurIPS*, 2021.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why should I trust you?" Explaining the predictions of any classifier. In *KDD*, 2016.

Jessica Schrouff, Sebastien Baur, Shaobo Hou, Diana Mincu, Eric Loreaux, Ralph Blanes, James Wexler, Alan Karthikesalingam, and Been Kim. Best of both worlds: local and global explanations with human-understandable concepts. *arXiv:2106.08641*, 2021.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: visual explanations from deep networks via gradient-based localization. *IJCV*, 2020.

Hua Shen and Ting-Hao Huang. How useful are the machine-generated interpretations to general users? a human evaluation on guessing the incorrectly predicted labels. In *AAAI Conference on Human Computation and Crowdsourcing*, 2020.

Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *TPAMI*, 2000.

Leon Sixt, Martin Schuessler, Oana-Iuliana Popescu, Philipp Weiß, and Tim Landgraf. Do users benefit from interpretable vision? a user study, baseline, and dataset. In *ICLR*, 2022.

Ram Dyuthi Sristi, Gal Mishne, and Ariel Jaffe. Disc: Differential spectral clustering of features. *NeurIPS*, 2022.

Ulrike Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 2007.

Zhenda Xie, Zigang Geng, Jingcheng Hu, Zheng Zhang, Han Hu, and Yue Cao. Revealing the dark secrets of masked image modeling. In *CVPR*, 2023.

Ruihan Zhang, Prashan Madumal, Tim Miller, Krista A Ehinger, and Benjamin IP Rubinstein. Invertible concept-based explanations for cnn models with non-negative concept activation vectors. In *AAAI*, 2021.

Wentao Zhang, Jiawei Jiang, Yingxia Shao, and Bin Cui. Efficient diversity-driven ensemble for deep neural networks. In *ICDE*, 2020.