Selected Papers of #AoIR2025:
The 26th Annual Conference of the
Association of Internet Researchers
Niterói, Brazil / 15 – 18 Oct 2025

SEMANTIC CLUSTERING FOR VISUAL DATA

Luigi Arminio
IT University of Copenhagen

Matteo Magnani
Uppsala University

Matias Piqueras
Uppsala University

Luca Rossi
IT University of Copenhagen

Alexandra Segerberg
Uppsala University

Clustering visual data, state of the art and limitations

Social media have gone through an overall visual turn. From the text first nature of former Twitter or the first era of Facebook, we now face platforms that are visual first, if not visual only both in terms of design and usage. This poses new challenges for researchers that aim at understanding this growing amount of data from a computational or quantitative perspective. The methods that were originally developed for textual data in domains such NLP are suddenly ineffective and when researchers turned to the domain of computer vision (Zhang & Peng 2024) they have often realised that these methods, while powerful, have been developed for task (e.g., object recognition, of image segmentation) that are of not always of immediate use in research dealing with users or social practices (WIlliams, Casas & Wilkerson 2020). This is especially evident in the case of image clustering. Image clustering aims at grouping together in *clusters* images that are similar. Within the context of computer vision two images are similar (thus should be clustered together) if they represent the same object or collection of objects. This has traditionally been achieved with the use of Convolutional Neural Networks (CNN) that learn and generalize from repeating visual features in images. While very good at recognizing objects this strategy often results in clusters based on the denotative nature of the image (what is represented) rather than

Suggested Citation (APA): Arminio, L., Magnani, M.,Piqueras, M., Rossi, L., & Segerberg, A. (2025, October). *Semantic Clustering For Visual Data*. Paper presented at AoIR2025: The 26th Annual Conference of the Association of Internet Researchers. Niterói, Brazil: AoIR. Retrieved from http://spir.aoir.org.

being able to deal with its connotative meaning (Cai et al. 2023). CNN-based method's struggle to capture the social or symbolic meaning of images is especially relevant in the context of highly intertextual visual objects like memes or coordinated visual campaings where the interplay between the text, the visual and the context is central to the correct interpretation (Wiggins, Bradley & Bowers 2015).

Leveraging Visual LLMs for Visual clustering

To address the limitations shown by current CNN-based approaches, we propose a Visual LLM-based semantic clustering methodology that can capture subtle social and cultural meanings within images, going beyond mere visual or spatial similarities. Following a variety of different strategies VLLMs are trained on pairs of visual and text that, often taken from online sources. This paired-training translates into VLLMs being able to see in context. This not only results in VLLMs outperforming more traditional CNNs approaches in zero-shot classification problems (Saha, Van Horn and Maji 2024) but opens the possibility of achieving visual clustering that takes into consideration the actual semantic of the visual content.

Our VLLM-based methodology involves a multi-step process that we tested on a visual dataset (in our case, a set of 11873 images used for communication around climate debate on social media that have been previously used in work by Zhang and Peng (2024) on visual clustering for social sciences).
First, we input the images into GPT-4 (Achiam et al. 2023) to generate textual descriptions. Subsequently, we convert the generated texts into vector representations through a BERT-based embedding model. Then, we reduce the dimensionality of these vectors using UMAP (McInnes, Healy & Melville 2018). Following these, we perform clustering on the resulting vectors through HDBSCAN (McInnes, Healy & Astels 2017). After applying our approach to the abovementioned visual dataset, we benchmarked it against the CNN-based algorithm VGG16 (Simonyan & Zisserman, 2014), a state-of the-art approach for large-scale image recognition and clustering.
We explored two key research questions: (1) whether a VLLM-based image segmentation approach can improve the connotative validity of clusters with respect to the traditional CNN-based methods, and (2) whether the clusters obtained from the VLLM-based methodology are interpretable by end-users.
We interpret the connotative quality of the obtained clusters as a proxy for their semantic coherence. We measure it by adapting a cluster quality metric introduced by Grimmer and King (2011) to measure both the connotative and denotative quality of clusters. We randomly selected 500 image pairs that were ranked by 2 human coders on a 3-point scale both based on their connotative and denotative similarity (Krippendorff's α = 0.81 for denotative scores and 0.71 for connotative scores). As for the second research question, the interpretability of the VLLM-generated clusters was assessed by generating textual labels for the highest TF-IDF scoring terms in the image descriptions. Then, human evaluators matched random image sets taken from a cluster to the cluster description based on these textual labels, allowing us to calculate precision and recall per cluster.

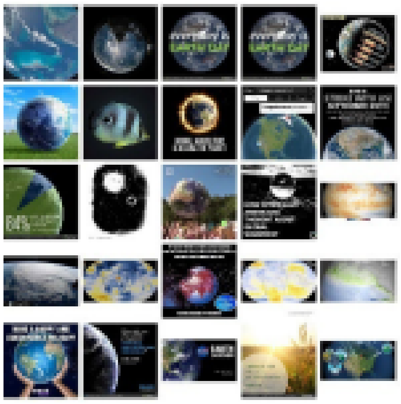

Figure 1a: One of the clusters obtained with the CNN-based method

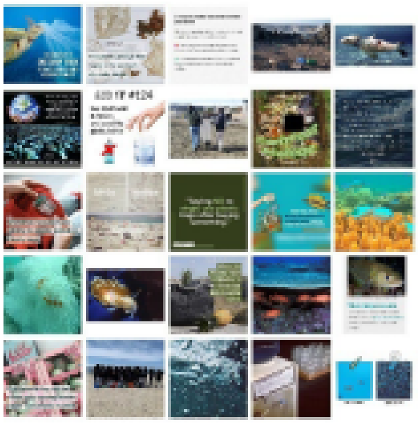

Figure 1b: One of the clusters obtained with the VLLM-based method.

Results

Figure 1 shows an example of the clusters obtained by the two methods. On the one side (Figure 1a) the CNN-based method shows a clear reliance on structural visual features, clustering together all images that show rounded shapes regardless of what they represent (the Earth, Maps or a fish). On the other side the VLLM based approach clusters together images and complex visual objects that focus plastic and its impact on sea life.

Figure 2 shows the results of the cluster-quality evaluation. The VLLM method outperforms the VGG-16 methods in denotative quality and offers very comparable results in connotative quality. These results are particularly clear when clusters are allowed to be as small as 50 images. Increasing the minimum number of images in the clusters affects the performance of the VLLM- based method more than the VGG-16. When investigating the interpretability of clusters via the TF-IDF labels we obtained an average precision and recall over all 32 clusters of, respectively, 0.83 and 0.83, with an overall accuracy of 0.83. As a point of reference, the expected precision and recall for a cluster assuming equal cluster sizes and random assignments is .03.

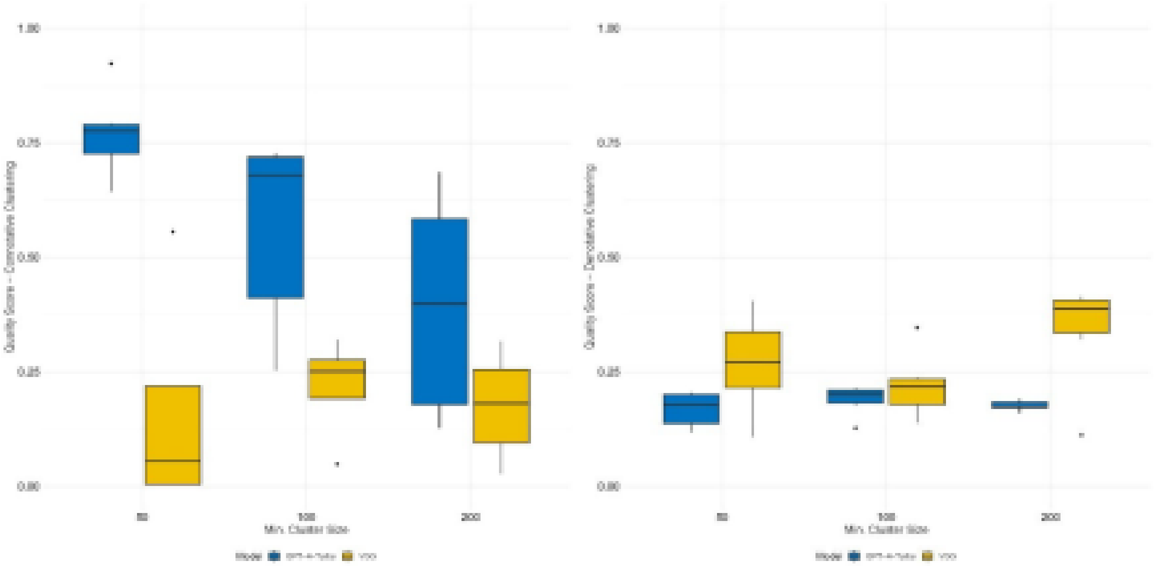

Figure 2: The quality score obtained for connotative clustering (left) and denotative clustering (right).

In conclusion, we showed how an approach to visual clustering that leverages the textual descriptions offered by VLLMs greatly enhances the connotative quality of the resulting clusters and facilitates their interpretability. This approach outperforms current methods for visual clustering, opens new and exciting opportunities to study the complex visual culture and processes of contemporary social media.

References

Arminio, L., Magnani, M., Piqueras, M., Rossi, L., & Segerberg, A. (2024). *Leveraging VLLMs for Visual Clustering: Image-to-text mapping shows increased semantic capabilities and interpretability*.

Cohn, N. (2013). Visual narrative structure. *Cognitive science*, *37*(3), 413-452.

Forchtner, B. (2019). Climate change and the far right. *Wiley Interdisciplinary Reviews: Climate Change*, *10*(5), e604.

Krange, O., Kaltenborn, B. P., & Hultman, M. (2021). “Don’t confuse me with facts”—how right wing populism affects trust in agencies advocating anthropogenic climate change as a reality. *Humanities and Social Sciences Communications*, *8*(1).

Page, R. (2015). The narrative dimensions of social media storytelling: Options for linearity and tellership. *The handbook of narrative analysis*, 329-347.

Wang, S., Corner, A., Chapman, D., & Markowitz, E. (2018). Public engagement with climate imagery in a changing digital landscape. *Wiley Interdisciplinary Reviews: Climate Change*, *9*(2), e509.