Enhancing Image Clustering with Captions

Yuanyuan Cai, Satoshi Kosugi, Kotaro Funakoshi, Manabu Okumura

Proceedings of the 38th Pacific Asia Conference on
Language, Information and Computation (PACLIC 38)

Nathaniel Oco, Shirley N. Dita, Ariane Macalinga Borlongan, Jong-Bok Kim (eds.)

2024

Enhancing Image Clustering with Captions

Yuanyuan Cai, Satoshi Kosugi, Kotaro Funakoshi, Manabu Okumura
Institute of Science Tokyo
{cai, kosugi, funakoshi, oku}@lr.pi.titech.ac.jp

Abstract

The limitations of traditional image clustering methods arise from their reliance on single-modal image representations, which impedes their ability to capture complex relationships within datasets and lacks interpretability of clustering results. In this work, we introduce a novel approach by incorporating captions directly generated from images and integrating image and caption embeddings to enhance image clustering performance. This method utilizes generated captions from images, thereby eliminating the need for human-labeled annotations. Experiments on five datasets validate the effectiveness of our approach, demonstrating notable improvements in clustering performance compared to methods that rely solely on visual or textual information. By fusing multimodal information from images and captions, we significantly improve clustering stability and accuracy, with enhancements ranging from 0.003 to 0.129 in the ACC, NMI, and ARI metrics for more challenging image datasets. In addition, we improve the interpretability of the cluster by employing advanced language models to generate a concise summary for each cluster. The summaries produced by ChatGPT enhance the comprehension of clustered data by effectively encapsulating the distinctive features of images within each cluster, thereby improving the accessibility and interpretability of the clustering results more nuancedly. Overall, this research paves the way for a new approach to image clustering by leveraging multimodal representations that integrate images with generated captions.

1 Introduction

Image clustering is a foundational technique in data analysis and machine learning, crucial for organizing data into meaningful groups based on similarity. Traditional methods often rely on single-modal data representations, which can limit their ability to capture the full complexity of datasets. The advent of vision-language models such as CLIP (Radford et al., 2021), BLIP (Li et al., 2022), and BLIP2 (Li et al., 2023a) has transformed clustering by integrating both visual and textual information, offering promising avenues for enhanced performance.

This research explores the integration of image and caption embeddings to enhance clustering performance. The images convey detailed visual information, while the captions provide contextual summaries, enriching the overall data representation. Our approach introduces a novel clustering methodology that directly utilizes generated captions from images, thus eliminating the requirement for human-labeled annotations. By embedding images and captions using advanced vision-language models into a unified multimodal space, our method aims to improve clustering accuracy and stability significantly.

The major contributions of this work can be summarized as follows:

1. We introduce an approach that improves image clustering by incorporating generated captions, reducing the reliance on manual annotations and leading to a more practical and cost-effective method.

2. Advanced language models generate concise sentence-type summaries for clusters, improving the interpretability of clustering results and revealing underlying data patterns.

3. Experiments validate that our multimodal clustering approach significantly improves over traditional unimodal methods for most datasets. This highlights the role of multimodal fusion in enhancing clustering performance.

2 Related Work

In this section, we review some recently published image clustering methods and briefly introduce the combination of text and image information methods.

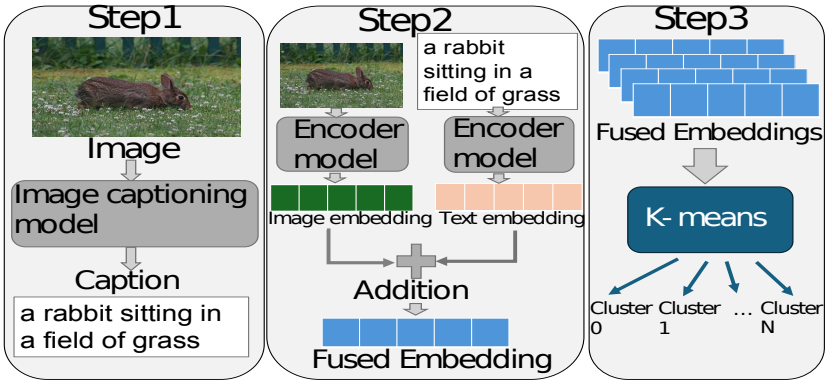


Figure 1: Overview of our method. Step 1: The image captioning model generates descriptive captions from the input images. Step 2: The encoder model encodes both the image and text into their respective embeddings, which are subsequently integrated into a single fused embedding. Step 3: These fused embeddings are clustered using K-means, enhancing the representation of the data and improving clustering performance.

2.1 Modern Image Clustering

Recent image clustering methods have improved significantly due to advanced deep learning-based representation techniques, particularly through contrastive learning (Li et al., 2021; Shen et al., 2021; Zhong et al., 2021). These advancements have enhanced the ability to map similar images closer together in feature spaces, improving the effectiveness of clustering algorithms in capturing semantic similarities.

In addition to these advances, externally guided image clustering methods, particularly those guided by text, enhance performance by incorporating additional information. TAC (Li et al., 2023b) uses WordNet textual semantics to improve feature discriminability and distill neighborhood information between text and images. The Text-Guided Image Clustering method (Stephan et al., 2024) generates text using image captioning and visual question-answering (VQA) models to inject task- or domain-specific knowledge and then utilizes only text to cluster images. The IC | TC methodology (Kwon et al., 2024) leverages modern vision language and large language models to group images based on user-specified text criteria, representing a new paradigm in image grouping.

Additionally, leveraging textual knowledge not only enables the meaningful and accurate clustering of images based on semantic meanings but also provides text explanations that are easily understandable for humans. Methods often employ interpretable features like semantic tags (Sambaturu et al., 2020; Davidson et al., 2018), particularly when aiming for textual explainability. For instance, the method of Zhang and Davidson (2021) uses integer linear programming to assign tags to clusters. The Text-Guided Image Clustering method introduces an approach that enriches cluster descriptions with keyword-based explanations.

In our method, as shown in Figure 1, we leverage vision-language models (VLMs) to generate image descriptions, thus introducing additional textual information. Subsequently, we employ contrastive learning-based deep learning models to encode both images and descriptions. Unlike previous research by Stephan et al. (2024), we do not rely solely on text to cluster images. Clustering based solely on text can lead to unstable results. Instead, we fuse both image and text embeddings, enhancing clustering results' stability and accuracy. Furthermore, we generate sentence-type textual explanations for the clusters by summarizing the image descriptions within each cluster, making them more understandable compared to using just a few keywords as explanations.

2.2 Text And Image Combination

In recent years, there has been considerable focus on developing VLMs due to their impressive performance in multimodal representation learning from large datasets of image-text pairs. These models learn joint representations from both images and text, capturing the interplay between visual and linguistic information (Al-Tameemi et al., 2023; Bakkali et al., 2020; Do et al., 2020). The emergence of CLIP (Radford et al., 2021), BLIP (Li et al., 2022), and BLIP2 (Li et al., 2023a) demonstrated robust zero-shot performance across various benchmarks, solidifying VLMs as a leading approach in visual recognition. In Menon and

Vondrick (2023) study, they utilized GPT-3 as a large language model (LLM) to generate textual descriptions of category names. They then used CLIP for image embeddings and text description embeddings to compare similarities for image classification. The combination of external linguistic knowledge and images enhanced interpretability in model decisions and improved performance in recognition tasks. In Do et al. (2020) study, images and their associated human-labeled text descriptions are fused into a unified, information-enriched image, and they demonstrated the effectiveness in the image-text pairs clustering task.

Some studies suggest that integrating textual and image information across various tasks enhances performance compared to utilizing unimodal data alone. Techniques such as concatenation, addition, multiplication of diverse embeddings, and training fusion models illustrate improved accuracy and other advantageous attributes (Zhao et al., 2023; Tembhurne and Diwan, 2021). Each modality contributes complementary insights, enriching the holistic representation and mitigating ambiguities in data interpretation.

Our method also combines text and image information. However, unlike existing approaches that use pre-existing human-labeled text descriptions, we generate descriptions automatically based on images and then fuse the information by adding the embeddings of the descriptions and the images.

3 Methodology

This section presents a simple yet effective clustering method in Figure 1. In brief, this approach involves generating textual descriptions for images and leveraging VLMs to embed both the image and caption. Subsequently, these embeddings are fused into multimodal embeddings used for k-means (MacQueen et al., 1967) clustering. Our method capitalizes on the zero-shot capabilities inherent in large-scale vision-language models, thereby obviating the need for model training, rendering our approach both cost-effective and influential.

3.1 Image Information

Image embedding is the process of transforming images into high-dimensional vector representations that encapsulate the essential features and characteristics of the images.

There are various advanced methods for extracting salient information from images. In this study, we employ two state-of-the-art models, CLIP (Radford et al., 2021) and BLIP (Li et al., 2022), for image embedding, leveraging their robust zero-shot learning capabilities without any further training or fine-tuning. These models possess a comprehensive understanding of images' content and context, enabling them to generate rich, semantically meaningful embeddings. In the subsequent experiment section, we also compare the performance of these two models on clustering tasks.

3.2 Caption Information

We experiment with BLIP (Li et al., 2022), BLIP2 (Li et al., 2023a), and ClipCap (Mokady et al., 2021) models to generate image captions. Despite these models achieving state-of-the-art results in image captioning tasks, we employ the CLIPscore (Hessel et al., 2021) model to assess the quality of the generated captions. Due to superior scoring performance, we opt to use the BLIP and BLIP2 models for caption generation. Subsequently, we utilize the BLIP and CLIP (Radford et al., 2021) models to embed these captions, as both models have achieved state-of-the-art results in various text embedding tasks.

3.3 Modality Fusion

Modality fusion involves integrating data from diverse modalities, such as text, images, and audio, to improve machine learning model performance. In the context of fusing image and caption embeddings, concatenation, addition, and multiplication are frequently used methods that do not necessitate additional training. Our study chose addition due to its simplicity and effectiveness in preserving the information from both modalities while maintaining computational efficiency relative to concatenation and multiplication approaches.

3.4 Clustering Method

We employ the K-means (MacQueen et al., 1967) algorithm as our clustering method, renowned for its popularity and widespread use in partitioning datasets into clusters. K-means clustering groups similar data points to uncover patterns by iteratively assigning each point to the nearest cluster centroid and updating centroids based on assigned points' means until convergence. K-means clustering endeavors to divide n data points into N clusters, In our study, N was defined based on the number of categories present in each dataset.

3.5 Clustering Summary

We use captions generated by BLIP (Li et al., 2022) model, then summarize these captions into 30-word descriptions for each cluster using the ChatGPT (OpenAI, 2023) and T5 (Raffel et al., 2020) models. The purpose of these summaries is to provide an easily understandable explanation for each clustered group of images, serving as folder names for each cluster. This offers a general description of the images without requiring detailed visual inspection of numerous images in each cluster, allowing for a quick overview of the cluster contents. The summaries are condensed to 30 words for direct visibility and easy checking in Windows system folder names, ensuring key information is quickly accessible and readable at a glance.

4 Experiments

This section assesses the proposed method across two widely-used and three more challenging image clustering datasets. A series of quantitative and qualitative comparisons and analyses are carried out to investigate the method's effectiveness and robustness.

4.1 Experimental Setup

In this subsection, we outline the datasets and metrics employed for evaluation and then detail the implementation of our method.

4.1.1 Datasets

To evaluate the performance of our method, we initially apply it to two widely-used image clustering datasets: ImageNet-10-train and ImageNet-10-val (Deng et al., 2009). Additionally, we assess this method on three more complex datasets: DTD (Cimpoi et al., 2014), WEAPD (Xiao et al., 2021), and Food-101-tiny-val (Bossard et al., 2014), which are characterized by a larger number of categories or more challenging image compositions. DTD is a dataset for texture recognition, WEAPD comprises 11 categories of weather phenomena for climate recognition, and Food-101-tiny-val is a subset for food recognition. Table 1 summarizes concise details of all datasets used in our evaluation.

4.1.2 Evaluation Metrics

To evaluate the clustering performance, we utilize three widely-used clustering metrics, including NMI (Vinh et al., 2010), ACC (Yang et al., 2010), and ARI (Hubert and Arabie, 1985). Higher values of these metrics collectively indicate superior clustering performance, providing a robust and comprehensive evaluation of the clustering results.

Dataset	Used Split	#Used Split	#Classes
ImageNet10	Train	13,000	10
ImageNet10	Val	500	10
DTD	Train+Val	5,640	47
WEAPD	Train+Val	6,862	11
Food101tiny	Val	500	10

Table 1: Dataset Splits and Sizes

4.1.3 Implementation Details

In our experimental setup, we compare clustering based on different data representations: solely keywords, solely captions, solely images, and fused image captions. Following the previous works (Stephan et al., 2024), we utilize the BLIP2 model (Li et al., 2023a) with the blip2-flan-t5-xxl variant to generate keywords using the prompt: "Which keywords describe the image?" For caption generation, we employ the BLIP model (Li et al., 2022) with the base-coco configuration and BLIP2 model (Li et al., 2023a) using blip2-flan-t5-xl and the ClipCap model (Mokady et al., 2021) using clip-ViT-B-32 to generate one caption for each image. Caption quality is evaluated using the CLIPscore metric (Hessel et al., 2021), as shown in Table 2, with the best scores highlighted in bold. CLIPscore is a reference-free metric with a strong correlation to human judgment and outperforms existing reference-based metrics. Since the performance of the BLIP and BLIP2 models is comparable, in subsequent experiments, we aim to evaluate the effectiveness of a single caption and compare it with previous studies that suggest multiple captions may be more effective. For this purpose, we use BLIP to generate one caption for each image and BLIP2 to generate six captions for each image.

Subsequently, we use the BLIP model with the blip-image-captioning-base configuration and the CLIP model (Radford et al., 2021) with clip-ViT-B-32 to embed images, as well as the generated single caption and keywords. To facilitate comparison with the previous study, we also use SBERT to embed six captions. The image and caption embeddings were then fused through additive combination. Finally, we apply k-means clustering (MacQueen et al., 1967) with a random state of 42 to ensure that the k-means algorithm produces consistent and reproducible results by fixing the seed for random initialization. Subsequently, we

Dataset	Used Split	BLIP	BLIP2	ClipCap
ImageNet10	Val	0.775	**0.788**	0.739
DTD	Train +Val	**0.782**	0.777	0.717

Table 2: CLIPscore of different captions

set the number of clusters to correspond with the number of classes listed in Table 1.

4.2 Main Results

In this study, we test our proposed method on both a widely-used and a challenging image clustering dataset. Additionally, we present the performance outcomes on three other datasets, followed by an in-depth analysis of the results.

4.2.1 Text Clustering

Prior research performed clustering using texts generated from images. However, the generated texts, say, captions, prompts, or keywords, can vary significantly according to the model or prompt they used, which greatly affects the text information. As a result, the clustering target can change, and thus, the clustering results can also be greatly influenced.

In Table 3, we present examples from the ImageNet10 (Deng et al., 2009) and Food101-tiny (Bossard et al., 2014) datasets, illustrating significant variations in the information provided by keywords and captions. It is apparent that keywords are less descriptive and lack the context and detail that captions provide, as seen with "dessert, plate, strawberry" versus "a piece of cake on a plate with chocolate sauce and berries." Besides, keywords can sometimes be ambiguous or unrelated, like "yelp" in the ImageNet10 example, leading to potential confusion. Moreover, identical keywords can correspond to different classes, necessitating more detailed captions for accurate class differentiation.

Table 4 compares the performance of different models on clustering tasks using various types of input. Our observations reveal that using only keywords or a single caption for clustering with the embedding models BLIP and CLIP resulted in low accuracy and unstable outcomes. Keywords perform worse than single captions and images, indicating that keywords alone do not capture sufficient information for effective clustering. One single caption significantly outperforms keywords but remains less effective than images. Furthermore, with different embedding models, the metrics show substantial variability, approaching differences of 0.4, highlighting the instability of clustering results based on captions. This suggests that while one single caption provides more context than keywords, it still lacks some of the visual details necessary for accurate and stable clustering. Using images yields the most stable and highest-quality clustering results. However, images alone do not provide a textual explanation of the clusters, which can be a limitation for interpretability.

4.2.2 Image Clustering with Captions

Texts provide coarse-grained information, while images provide fine-grained details. This difference arises because texts are concise and constrained by space, leading to general descriptions. Language abstracts information, as seen in captions like "A man riding a bicycle," which omit specific details such as the bicycle's color, the man's clothing, or the background. Texts highlight the main subject or action, offering a broad overview rather than detailed information.

Integrating textual information with image data enhances clustering accuracy and stability, as shown in Table 4. For single caption, with the embedding models BLIP and CLIP, regardless of the embedding model or dataset used, the combined use of images and captions consistently yields the best overall clustering performance. Besides, because captions outperform keywords, we used captions as text information, combined with image information, experimented on five datasets, and compared the clustering results on caption embeddings, image embeddings, and fused embeddings.

As demonstrated in Table 5, the instability of clustering results based solely on captions is evident once again. The best results for each dataset are highlighted in bold. For single caption, with the embedding models BLIP and CLIP, regardless of whether the dataset is widely used, like ImageNet (Deng et al., 2009), or more challenging, the combination of images and captions consistently outperforms using either image or caption data alone. Additionally, performance varies between CLIP (Radford et al., 2021) and BLIP (Li et al., 2022) models depending on the dataset, indicating no universal model superiority. Furthermore, for ImageNet10-train and ImageNet10-val, despite being from the same dataset, differences in data volume or the specific images included can lead to variations in clustering metrics.

However, the situation changed when multiple captions with SBERT embeddings were used. We compare our method, which utilizes an image with

Dataset	**Imagenet10**		**Food101-tiny**	
Image Example				
Ground Truth	wood	tiramisu	apple pie	cannoli
Keywords	yelp	dessert, plate, strawberry	dessert, plate, strawberry	dessert, plate, strawberry
Captions	a group of people standing around a wooden structure	a piece of cake on a plate with chocolate sauce and berries	a plate of food with strawberries on it	a white plate topped with a dessert covered in chocolate

Table 3: Keywords generated by BLIP2 and captions generated by BLIP

Dataset	**Encoder model**	**Keywords**	**Caption**	**Image**	**Image + Caption**
Food101tiny-Val		dessert, plate, strawberry	a white plate topped with a dessert covered in chocolate strawberry		 a white plate topped with a dessert covered in chocolate strawberry
		ACC NMI ARI	ACC NMI ARI	ACC NMI ARI	ACC NMI ARI
	BLIP	0.482 0.492 0.280	0.271 0.376 0.145	0.846 0.819 0.741	**0.930 0.875 0.853**
	CLIP	0.474 0.480 0.278	0.610 0.614 0.462	0.916 0866 0.832	0.924 0.863 0.838
Imagenet10-Val		grass, field, rabbit	a rabbit sitting in a field of grass		 a rabbit sitting in a field of grass
		ACC NMI ARI	ACC NMI ARI	ACC NMI ARI	ACC NMI ARI
	BLIP	0.476 0.349 0.226	0.480 0.402 0.222	0.906 0.898 0.845	0.932 0.925 0.878
	CLIP	0.448 0.353 0.211	0.832 0.804 0.716	0.910 0904 0.855	**0.946 0.927 0.897**

Table 4: Clustering with different inputs

one single caption generated by the BLIP model and an image with six captions generated by the BLIP2 model for clustering, with the previous study Stephan et al. (2024) that uses SBERT to embed six captions generated by the BLIP2 model. For consistency, we refer to some experimental setups from prior research: we use the BLIP2 model using blip2-flan-t5-xl to generate six captions, and the same captions were used for comparison. The results are shown in Table 6.

In Table 6, for the entire Imagenet10Train+Val dataset, we observe that when using BLIP for embedding, the results of combining an image and six caption embeddings outperform those combining an image and a single caption. However, SBERT's performance with only six caption embeddings still surpasses our method. This may be attributed to SBERT's specialization for textual representations and the BLIP2 model's pre-training on the Imagenet dataset, which enables it to generate high-quality captions for Imagenet10. Additionally, we observe that for the DTD, WEAPD, and Food101tiny-Val datasets, even when using embeddings from the fusion of an image and a single caption generated by the BLIP model, our method performs better than the previous study. In these cases, the captions generated by BLIP or BLIP2 might not capture the nuances of images. However, BLIP's ability to create strong image embeddings compensates for this, making the image+1 caption embeddings more powerful than SBERT's embeddings of potentially weaker captions from these datasets.

In our opinion, the combination of image and text modalities is effective for clustering when the quality of generated captions is not sufficiently high, and the reasons for this effectiveness are as follows: First, images capture fine-grained visual

Representation	DTD			Imagenet10-Train			Imagenet10-Val			WEAPD			Food101tiny-Val		
Image	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
BLIP	0.498	0.567	0.309	**0.925**	0.885	0.852	0.906	0.898	0.845	0.712	0.722	0.592	0.846	0.819	0.741
CLIP	0.476	0.548	0.296	0.903	0.878	0.837	0.910	0.904	0.855	0.790	0.731	0.619	0.916	0.866	0.832
Caption	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
BLIP	0.206	0.223	0.057	0.413	0.275	0.168	0.480	0.402	0.222	0.354	0.251	0.161	0.271	0.376	0.145
CLIP	0.358	0.404	0.174	0.693	0.623	0.516	0.832	0.804	0.716	0.628	0.585	0.416	0.610	0.614	0.462
Image+Caption	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
BLIP	**0.523**	**0.586**	**0.338**	0.919	0.881	0.845	0.932	0.925	0.878	0.735	0.723	0.580	**0.930**	**0.875**	**0.853**
CLIP	0.511	0.578	0.330	0.911	**0.897**	**0.857**	**0.946**	**0.927**	**0.897**	**0.806**	**0.753**	**0.642**	0.924	0.863	0.838

Table 5: Clustering Results on Other Datasets

Representation	DTD			WEAPD			Food101tiny-Val			Imagenet10TrainVal		
1 Caption (BLIP)	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
BLIP	0.206	0.223	0.057	0.354	0.251	0.161	0.271	0.376	0.145	0.413	0.275	0.168
CLIP	0.358	0.404	0.174	0.628	0.585	0.416	0.610	0.614	0.462			
Image+1Caption (BLIP)	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
BLIP	**0.523**	**0.586**	**0.338**	0.735	0.723	0.580	**0.930**	**0.875**	**0.853**	0.919	0.881	0.845
CLIP	0.511	0.578	0.330	**0.806**	**0.753**	**0.642**	0.924	0.863	0.838			
Image+6Captions (BLIP2)	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
BLIP										0.946	0.902	0.886
6 Captions (BLIP2)	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
SBERT	0.467	0.522	0.265	0.730	0.712	0.577	0.826	0.812	0.724	**0.969**	**0.933**	**0.933**

Table 6: Comparison with Previous Research

details, while captions provide a high-level summary, highlighting aspects or context not immediately obvious from visual data alone. Second, visual information reduces ambiguity in textual descriptions, and captions clarify important objects or actions in the image. Third, multi-modal integration creates a more comprehensive representation of the content, leveraging the strengths of both modalities for better clustering performance. However, to optimize the integration of image and text information, we should consider employing more effective embedding models. Besides, although generating multiple captions requires more time and cost compared to a single caption, it has the potential to enhance the clustering results.

4.3 Cluster Explainability

In this study, we initially employ BLIP (Li et al., 2022) to generate one caption for each image. These captions, corresponding to images grouped within the same cluster, are then processed by ChatGPT (OpenAI, 2023) and T5 (Raffel et al., 2020) models to create 30-word summaries for each cluster, aiming to identify the common characteristics of images within the same cluster. Examples from the ImageNet10-val (Deng et al., 2009) and DTD (Cimpoi et al., 2014) datasets are shown in Table 7.

From the generated summaries, it is evident that the summaries produced by ChatGPT more effectively encapsulate the features of images within each cluster. In contrast, the summaries generated by the T5 model often fail to form coherent sentences and include repeated words. This discrepancy may be attributed to ChatGPT's capability to embed a larger number of words in a single instance, allowing us to input the image captions in one go and generate a summary. On the other hand, the T5 model can embed a limited number of words at a time, necessitating multiple inputs of captions and subsequent summarization, which might lead to less coherent outputs.

5 Conclusion

In this study, we present a novel clustering method that enhances image clustering by incorporating generated captions directly from images, bypassing the need for human-labeled annotations. Our approach leverages advanced models like CLIP (Radford et al., 2021), BLIP (Li et al., 2022), and BLIP2 (Li et al., 2023a) to generate captions and embed both textual descriptions and images without additional training, ensuring practicality and cost-effectiveness. By fusing these embeddings into multimodal representations, we exploit the complementary strengths of image and text modalities.

Our experimental results, conducted on a variety of datasets ranging from widely-used datasets to more challenging collections, demonstrate that our multimodal fusion significantly enhances cluster-

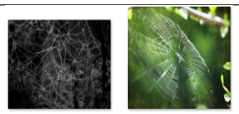
Dataset	**Imagenet10-Val**		**DTD**	
Cluster	Cluster1	Cluster3	Cluster0	Cluster9
Image Example				
Summary by GPT-3.5	images feature various aspects of violins and other musical instruments: close-ups of violins, people playing violins, instruments on display, and scenes of musicians in different settings.	assorted decorative pillows featuring various designs such as trees, owls, trucks, and patchwork. Colors range from pink and black to green and gold, adding vibrancy to beds, couches, and chairs.	various spider webs, some with water droplets, on different backgrounds like a blue sky, green surface, and black background. Close-ups and details of webs covered in dew or illuminated at night.	various striped patterns and designs: black and white, green, brown and tan, red and white, pink, rainbow, purple, orange, multicolored, and more on wallpaper, fabric, and clothing.
Summary by T5-base	a violin and strings a violin and strings a violin and strings a violin and strings a violin and strings a violin	a pillow with a picture of a truck on it a pillow with a picture of a truck on it a	on a tree a spider web with water drops on it a on a fence a spider web with water drops on it	a striped wallpaper pattern with vertical stripes a purple background with vertical stripes a purple and white striped wallpaper with vertical stripes

Table 7: Cluster Summarization

ing performance compared to using either modality independently on more challenging collections. This fusion captures detailed visual features alongside high-level textual summaries, reducing ambiguity and improving feature richness for more stable and accurate clustering outcomes.

Furthermore, we address cluster interpretability by employing advanced language models to generate concise summaries for each cluster. These summaries facilitate a better understanding of the clustered data, thereby making the clustering results more accessible and interpretable. Overall, our study underscores the significance of multimodal data fusion in clustering tasks when the quality of generated captions is not sufficiently high, also demonstrating that generated textual information can enhance interpretability for clustering.

References

IK Salman Al-Tameemi, Mohammad-Reza Feizi-Derakhshi, Saeed Pashazadeh, and Mohammad Asadpour. 2023. Multi-model fusion framework using deep learning for visual-textual sentiment classification. *Computers, Materials & Continua*, 76(2):2145–2177.

Souhail Bakkali, Zuheng Ming, Mickaël Coustaty, and Marçal Rusiñol. 2020. Visual and textual deep feature fusion for document image classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 562–563.

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. 2014. Food-101–mining discriminative components with random forests. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part VI 13*, pages 446–461. Springer.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. 2014. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3606–3613.

Ian Davidson, Antoine Gourru, and S Ravi. 2018. The cluster description problem-complexity results, formulations and approximations. *Advances in Neural Information Processing Systems*, 31.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

Truong Dong Do, Kyungbaek Kim, Hyukro Park, and Hyung-Jeong Yang. 2020. Image and encoded text fusion for deep multi-modal clustering. In *The 9th International Conference on Smart Media and Applications*, pages 308–312.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*.

Lawrence Hubert and Phipps Arabie. 1985. Comparing partitions. *Journal of classification*, 2:193–218.

Sehyun Kwon, Jaeseung Park, Minkyu Kim, Jaewoong Cho, Ernest K. Ryu, and Kangwook Lee. 2024. Image clustering conditioned on text criteria. *International Conference on Learning Representations*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR.

Yunfan Li, Peng Hu, Zitao Liu, Dezhong Peng, Joey Tianyi Zhou, and Xi Peng. 2021. Contrastive clustering. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 8547–8555.

Yunfan Li, Peng Hu, Dezhong Peng, Jiancheng Lv, Jianping Fan, and Xi Peng. 2023b. Image clustering with external guidance. *arXiv preprint arXiv:2310.11989*.

James MacQueen et al. 1967. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA.

Sachit Menon and Carl Vondrick. 2023. Visual classification via description from large language models. *ICLR*.

Ron Mokady, Amir Hertz, and Amit H Bermano. 2021. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*.

Chuang Niu, Hongming Shan, and Ge Wang. 2022. Spice: Semantic pseudo-labeling for image clustering. *IEEE Transactions on Image Processing*, 31:7264–7278.

OpenAI. 2023. Chatgpt (mar 14 version) [large language model]. Retrieved from https://chat.openai.com/chat.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Prathyush Sambaturu, Aparna Gupta, Ian Davidson, SS Ravi, Anil Vullikanti, and Andrew Warren. 2020. Efficient algorithms for generating provably near-optimal cluster descriptors for explainability. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 1636–1643.

Yuming Shen, Ziyi Shen, Menghan Wang, Jie Qin, Philip Torr, and Ling Shao. 2021. You never cluster alone. *Advances in Neural Information Processing Systems*, 34:27734–27746.

Karen Sparck Jones. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 28(1):11–21.

Andreas Stephan, Lukas Miklautz, Kevin Sidak, Jan Philip Wahle, Bela Gipp, Claudia Plant, and Benjamin Roth. 2024. Text-guided image clustering. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2960–2976, St. Julian's, Malta. Association for Computational Linguistics.

Jitendra V Tembhurne and Tausif Diwan. 2021. Sentiment analysis in textual, visual and multimodal inputs using recurrent neural networks. *Multimedia Tools and Applications*, 80(5):6871–6910.

Nguyen Xuan Vinh, Julien Epps, and James Bailey. 2010. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *Journal of Machine Learning Research*, 11(95):2837–2854.

Haixia Xiao, Feng Zhang, Zhongping Shen, Kun Wu, and Jinglin Zhang. 2021. Classification of weather phenomenon from images by using deep convolutional neural network. *Earth and Space Science*, 8(5):e2020EA001604.

Yi Yang, Dong Xu, Feiping Nie, Shuicheng Yan, and Yueting Zhuang. 2010. Image clustering using local discriminant models and global integration. *IEEE Transactions on Image Processing*, 19(10):2761–2773.

Hongjing Zhang and Ian Davidson. 2021. Deep descriptive clustering. *arXiv preprint arXiv:2105.11549*.

Qihui Zhao, Tianhan Gao, and Nan Guo. 2023. Tsvfn: Two-stage visual fusion network for multimodal relation extraction. *Information Processing & Management*, 60(3):103264.

Huasong Zhong, Jianlong Wu, Chong Chen, Jianqiang Huang, Minghua Deng, Liqiang Nie, Zhouchen Lin, and Xian-Sheng Hua. 2021. Graph contrastive clustering. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9224–9233.