NEURAL INFORMATION PROCESSING SYSTEMS

Customized Multiple Clustering via Multi-Modal Subspace Proxy Learning

Jiawei Yao[1], Qi Qian[2], Juhua Hu[1]

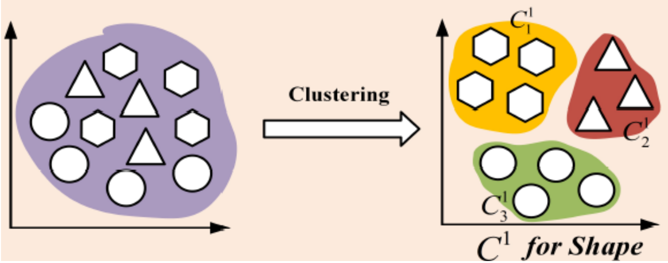
[1]School of Engineering and Technology, University of Washington, Tacoma, WA 98402, USA
[2]Zoom Video Communications

{jwyao, juhuah} @uw.edu qi.qian@zoom.us

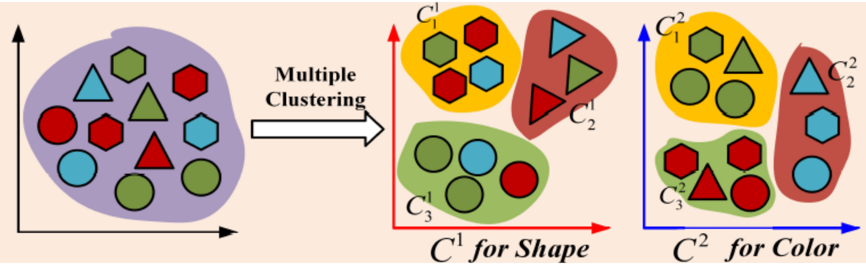
UNIVERSITY of WASHINGTON

Background

- **Traditional clustering**: only discover **one** clustering structure C^1 about shape



- **Multiple clustering**: reveal **two or more** distinct clustering (i.e., C^1 for shape and C^2 for color)



[1]

[1]. Guoxian Yu, Liangrui Ren, Jun Wang, Carlotta Domeniconi, and Xiangliang Zhang. "Multiple clusterings: Recent advances and perspectives." Computer Science Review 52 (2024): 100621.

Background

> Existing methods:

- ENRC[1]: auto-encoder based
- iMClust[2]: auto-encoder and multi-head attention
- MCV[3]: pre-trained feature extractors
- MSC[4]: maximizing the Laplacian eigenmap
- AugDMC[5]: data augmentation based

> **Major challenge**:

- Users often do not need **all the clusterings** that algorithms generate
- Users have to **manually** check them one by one to choose those ones of interest

[1].Lukas Miklautz, Dominik Mautz, Muzaffer Can Altinigneli, Christian Böhm, and Claudia Plant. "Deep embedded non-redundant clustering." In Proceedings of the AAAI conference on artificial intelligence, vol. 34, no. 04, pp. 5174-5181. 2020.

[2]. Liangrui Ren, Guoxian Yu, Jun Wang, Lei Liu, Carlotta Domeniconi, and Xiangliang Zhang. "A diversified attention model for interpretable multiple clusterings." IEEE Transactions on Knowledge and Data Engineering (2022).

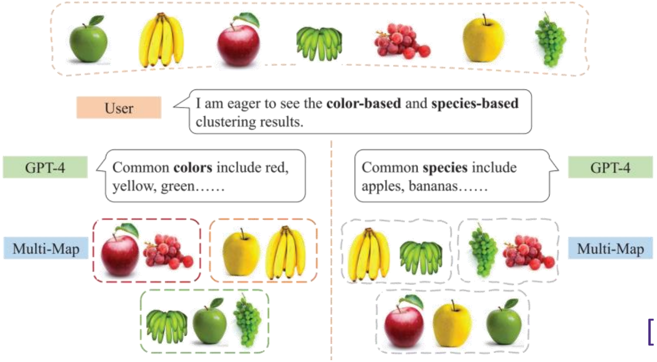
[3]. Joris Guérin, and Byron Boots. "Improving image clustering with multiple pretrained cnn feature extractors." arXiv preprint arXiv:1807.07760 (2018).

[4]. Juhua Hu, Qi Qian, Jian Pei, Rong Jin, and Shenghuo Zhu. "Finding multiple stable clusterings." Knowledge and Information Systems 51 (2017): 991-1021.

[5]. Jiawei Yao, Enbei Liu, Maham Rashid, and Juhua Hu. "Augdmc: Data augmentation guided deep multiple clustering." Procedia Computer Science 222 (2023): 571-580.

Recent work

- How to align user’s interest with different visual components precisely? - **CLIP**[1] **could help!**
  - **Multi-MaP**[2] first use a user's high-level concept to trigger the corresponding feature extraction from the pre-trained encoders from CLIP in multiple clustering

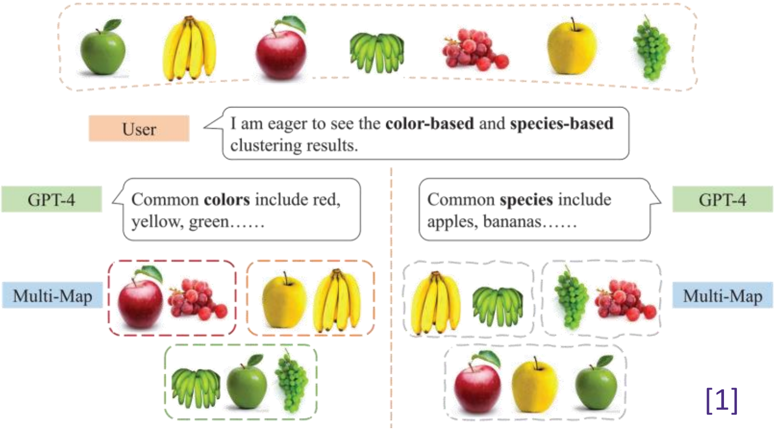

User
I am eager to see the color-based and species-based clustering results.
GPT-4
Common colors include red, yellow, green……
GPT-4
Common species include apples, bananas……
Multi-Map
Multi-Map

[2]

[1]. Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry et al. "Learning transferable visual models from natural language supervision." In International conference on machine learning, pp. 8748-8763. PMLR, 2021.
[2]. Jiawei Yao, Qi Qian, and Juhua Hu. "Multi-modal proxy learning towards personalized visual multiple clustering." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14066-14075. 2024.

Challenges

- Require the user to provide a **contrastive concept**
- Representation learning and clustering method in **a separate stage**

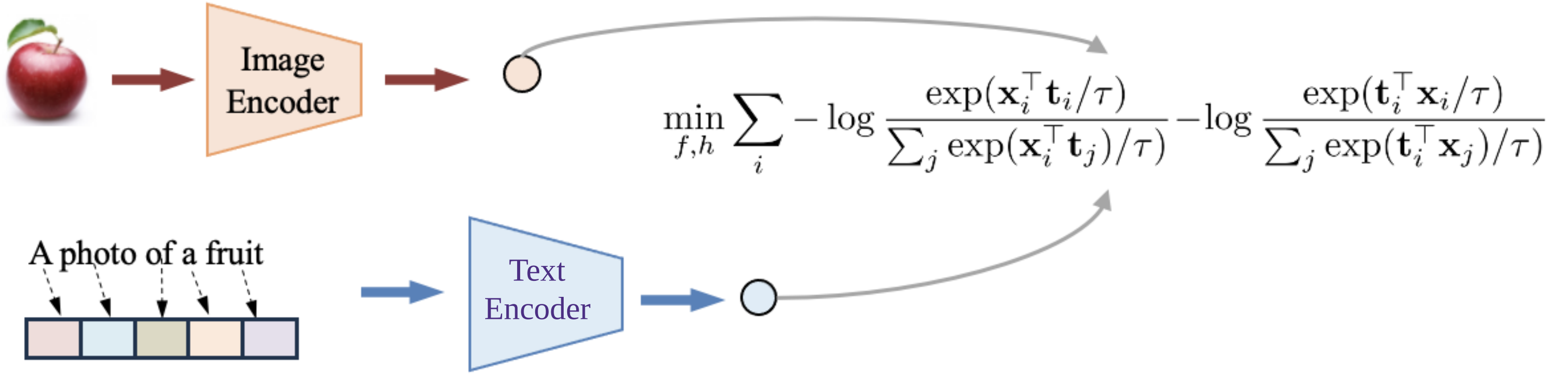

User
I am eager to see the **color-based** and **species-based** clustering results.
GPT-4
Common **colors** include red, yellow, green……
Common **species** include apples, bananas……
GPT-4
Multi-Map
Multi-Map

[1]

[1]. Jiawei Yao, Qi Qian, and Juhua Hu. "Multi-modal proxy learning towards personalized visual multiple clustering." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14066-14075. 2024.

Preliminary

Multi-modal Pre-training in CLIP

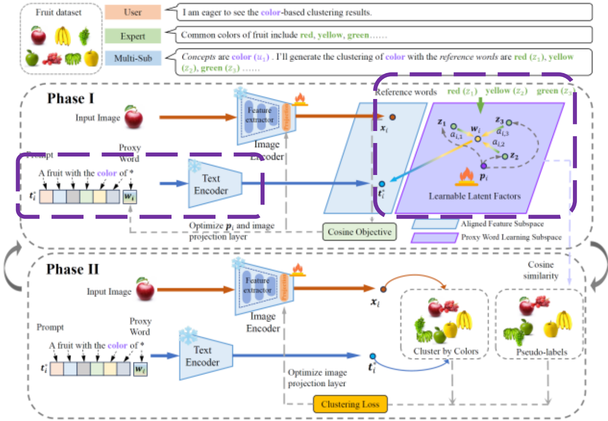


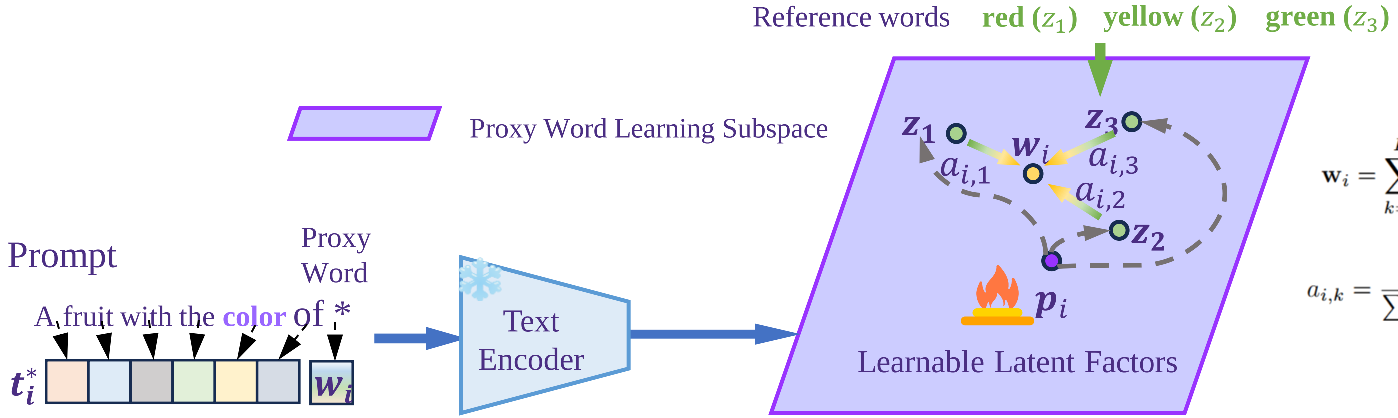
$$\min_{f,h}\sum_{i} -\log\frac{\exp(\mathbf{x}_i^\top\mathbf{t}_i/\tau)}{\sum_j\exp(\mathbf{x}_i^\top\mathbf{t}_j)/\tau)} - \log\frac{\exp(\mathbf{t}_i^\top\mathbf{x}_i/\tau)}{\sum_j\exp(\mathbf{t}_i^\top\mathbf{x}_j)/\tau)}$$

- Train vision and text encoders using a set of image-text pairs to obtain **representations**

Methodology

Subspace Proxy Word Representation

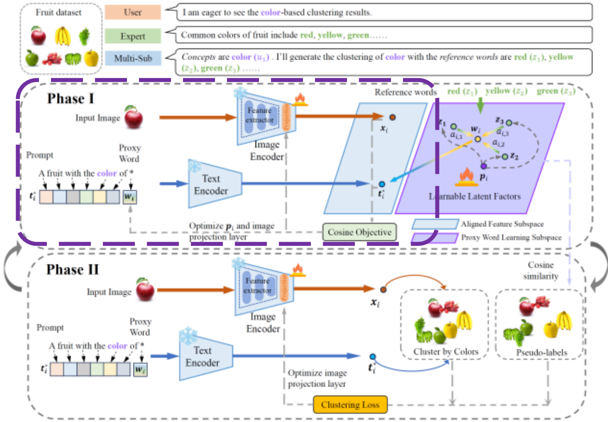




$$\mathbf{w}_i = \sum_{k=1}^{K} a_{i,k}\phi(z_k)$$

$$a_{i,k} = \frac{\exp(\mathbf{p}_i \mathbf{z}_k)}{\sum_j \exp(\mathbf{p}_i \mathbf{z}_j)}$$

- Learning a proxy word embedding based on user-preferred aspects, and representing proxy words as a linear combination of reference words corresponding to common categories under the concept



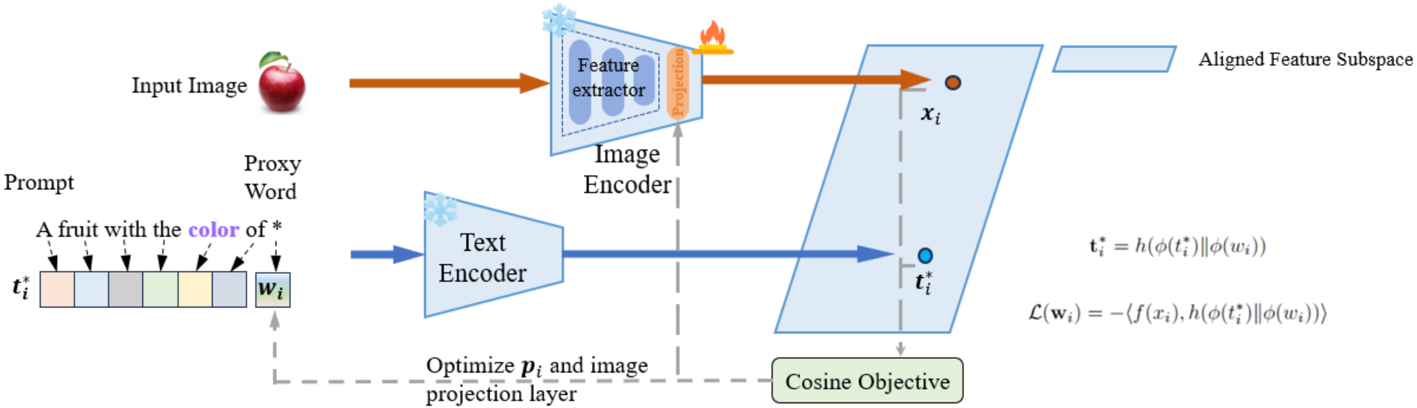
I am eager to see the color-based clustering results.

Common colors of fruit include red, yellow, green……

Concepts are color (u_1) . I'll generate the clustering of color with the *reference words* are red (z_1), yellow (z_2), green (z_3) ……

Methodology

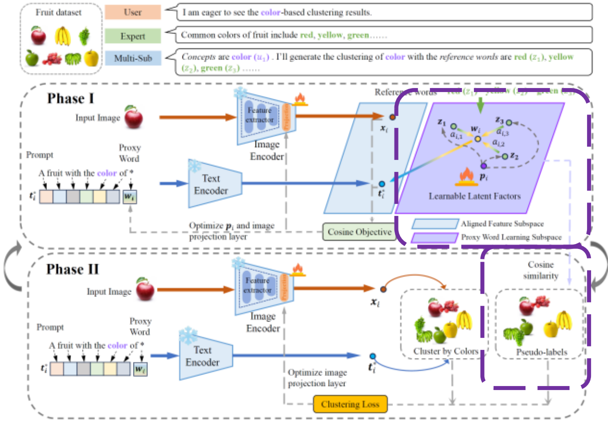
Multi-Modal Subspace Proxy Learning



$$\mathbf{t}_i^* = h(\phi(t_i^*)\|\phi(w_i))$$

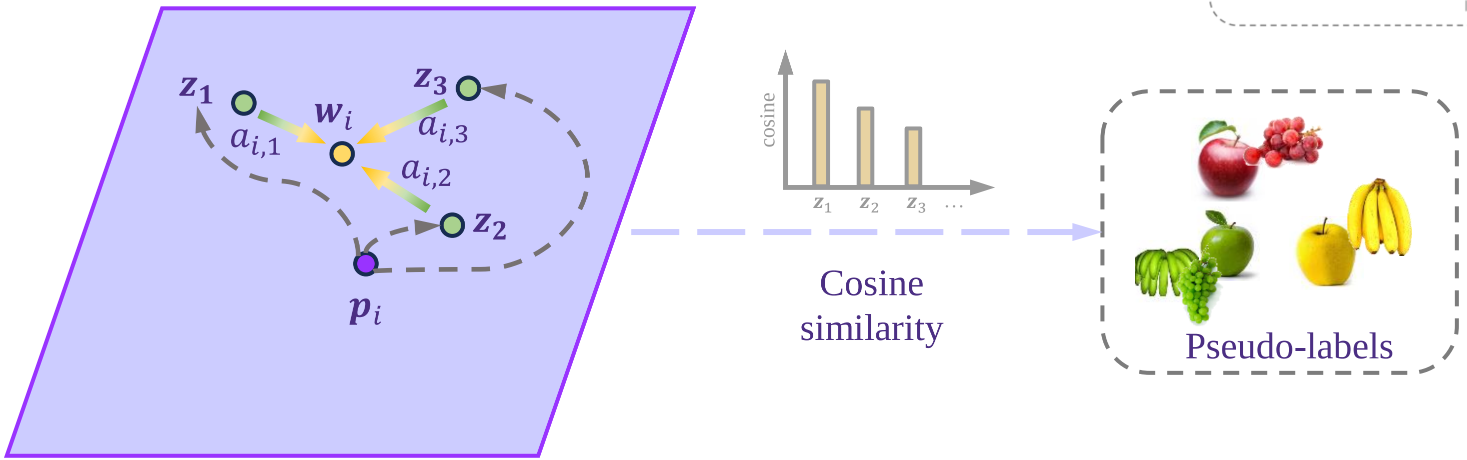
$$\mathcal{L}(\mathbf{w}_i) = -\langle f(x_i), h(\phi(t_i^*)\|\phi(w_i))\rangle$$

- Adjust trainable latent factors to align user-specific image representations with corresponding proxy word embeddings, optimizing the alignment through cosine similarity

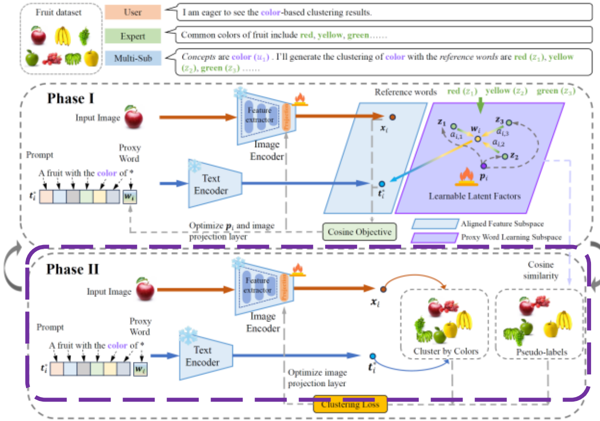

Fruit dataset
User
I am eager to see the color-based clustering results.
Expert
Common colors of fruit include red, yellow, green……
Multi-Sub
Concepts are color (u_1) . I'll generate the clustering of color with the reference words are red (z_1), yellow (z_2), green (z_3) ……
Phase I
Input Image
Feature extractor
Projection
Image Encoder
Prompt
Proxy Word
A fruit with the color of *
t_i^*
w_i
Text Encoder
x_i
t_i^*
Reference words red (z_1) yellow (z_2) green (z_3)
z_1
z_2
z_3
w_i
$a_{i,1}$
$a_{i,2}$
$a_{i,3}$
p_i
Learnable Latent Factors
Optimize p_i and image projection layer
Cosine Objective
Aligned Feature Subspace
Proxy Word Learning Subspace
Phase II
Input Image
Feature extractor
Projection
Image Encoder
Prompt
Proxy Word
A fruit with the color of *
t_i^*
w_i
Text Encoder
x_i
t_i^*
Optimize image projection layer
Cluster by Colors
Cosine similarity
Pseudo-labels
Clustering Loss

Methodology

Pseudo-labels generation



- Obtain the pseudo-labels using the highest cosine similarity between the currently learned proxy word embeddings and the reference word embeddings.



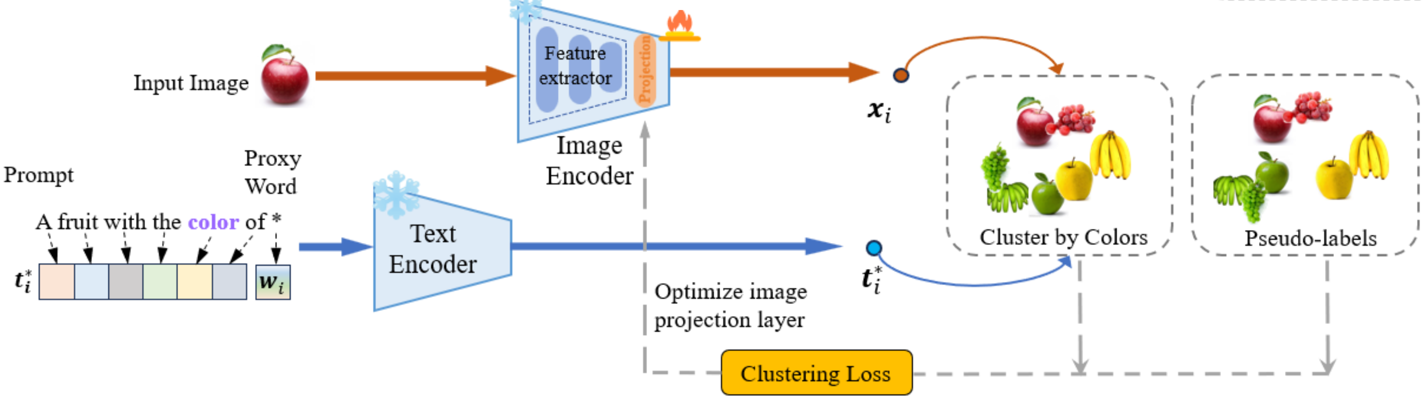
I am eager to see the color-based clustering results.

Common colors of fruit include red, yellow, green……

Concepts are color (u_1) . I'll generate the clustering of color with the *reference words* are red (z_1), yellow (z_2), green (z_3) ……

Methodology

Clustering Loss



- Given the pseudo-labels, the image embeddings can be then optimized through a clustering loss that combines intra-cluster loss to encourage compactness and inter-cluster loss to enhance separability

Datasets

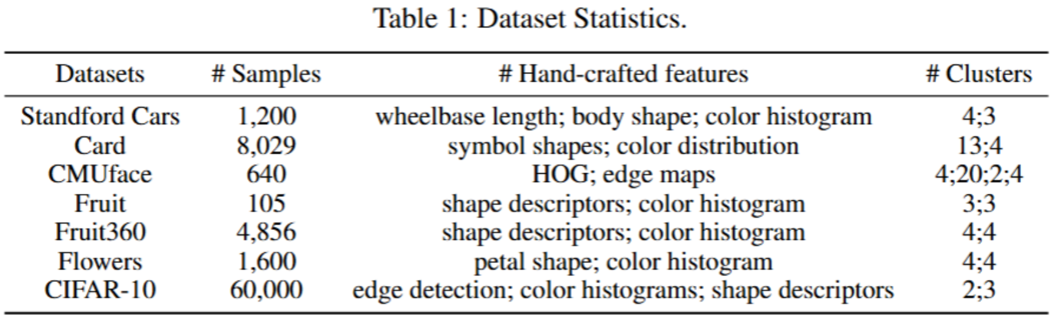


Table 1: Dataset Statistics.

Datasets	# Samples	# Hand-crafted features	# Clusters
Standford Cars	1,200	wheelbase length; body shape; color histogram	4;3
Card	8,029	symbol shapes; color distribution	13;4
CMUface	640	HOG; edge maps	4;20;2;4
Fruit	105	shape descriptors; color histogram	3;3
Fruit360	4,856	shape descriptors; color histogram	4;4
Flowers	1,600	petal shape; color histogram	4;4
CIFAR-10	60,000	edge detection; color histograms; shape descriptors	2;3

- ❑ **CIFAR-10**[1]: 60,000 images
  - ❑ Two clusterings: **type** (transportation and animals) and **environment** (land, air and water)
- ❑ **CMUface**[2]: 640 gray images
  - ❑ Four clusterings: **Pose** (left, right, straight and up), **identity** (20 individuals), **glass** (with or without glass), and **emotions** (angry, happy, neutral and sad).
- ❑ **Fruit**[4]: 105 images
  - ❑ Two clusterings: **Specie** (apple, banana, and grape) and **Color** (green, red, and yellow).
- ❑ **Fruit360**[5]: 4,856 images
  - ❑ Two clusterings: Specie (apple, banana, cherry, and grape) and Color (red, green, yellow, and maroon).
- ❑ **Card**[6]: 8,029 images
  - ❑ Two clusterings: Rank (Ace, King, Queen, …) and suits (clubs, diamonds, hearts and spades)
- ❑ **Stanford Cars**[3]: 1200 images
  - ❑ Two clusterings: color (white, grey, black, and red) and type (truck, sedan, and SUV)
- ❑ **Flowers**[7]: 1600 images
  - ❑ Two clusterings: color (yellow, pink, red, and white) and species (iris, aster, lotus, and tulip)

CIFAR-10

[1]. Alex Krizhevsky, and Geoffrey Hinton. "Learning multiple layers of features from tiny images." (2009): 7.

[2]. Stephan Günnemann, Ines Färber, Matthias Rüdiger, and Thomas Seidl. "Smvc: semi-supervised multi-view clustering in subspace projections." In Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 253-262. 2014.

[3]. Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. "3d object representations for fine-grained categorization." In Proceedings of the IEEE international conference on computer vision workshops, pp. 554-561. 2013.

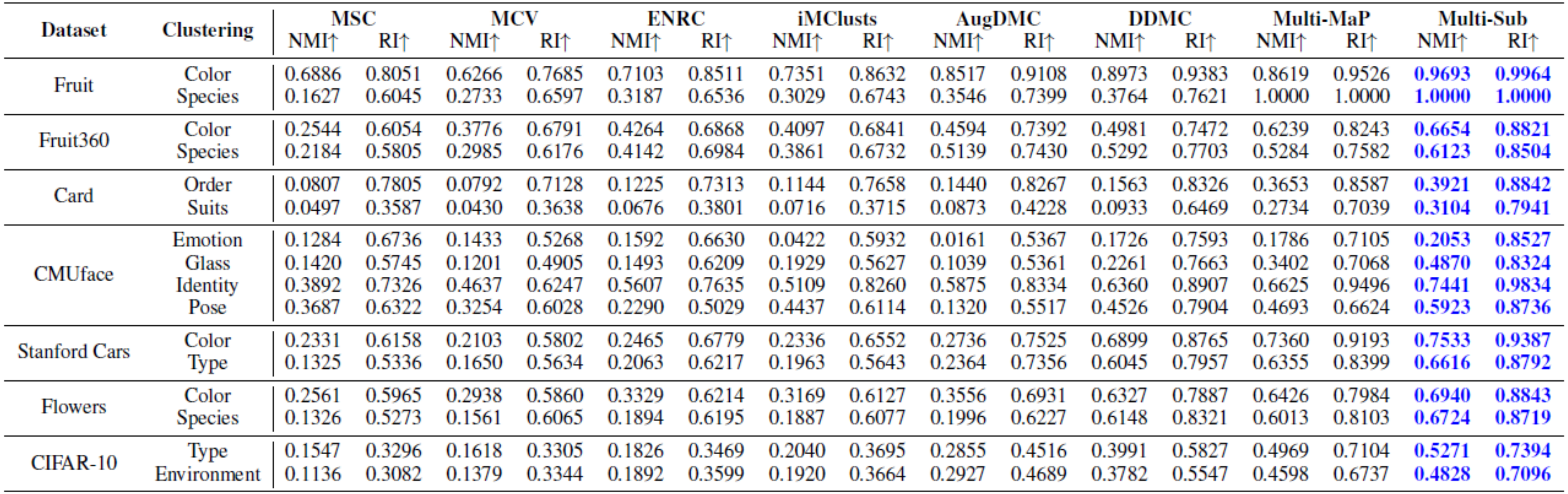
[4]. Juhua Hu, Qi Qian, Jian Pei, Rong Jin, and Shenghuo Zhu. "Finding multiple stable clusterings." Knowledge and Information Systems 51 (2017): 991-1021.

[5]. https://www.kaggle.com/moltean/fruits

[6]. https://www.kaggle.com/datasets/gpiosenka/cards-image-datasetclassification

[7]. Maria-Elena Nilsback, and Andrew Zisserman. "Automated flower classification over a large number of classes." In 2008 Sixth Indian conference on computer vision, graphics & image processing, pp. 722-729. IEEE, 2008.

Quantitative comparison



Dataset	Clustering	MSC		MCV		ENRC		iMClusts		AugDMC		DDMC		Multi-MaP		Multi-Sub	
		NMI↑	RI↑	NMI↑	RI↑	NMI↑	RI↑	NMI↑	RI↑	NMI↑	RI↑	NMI↑	RI↑	NMI↑	RI↑	NMI↑	RI↑
Fruit	Color	0.6886	0.8051	0.6266	0.7685	0.7103	0.8511	0.7351	0.8632	0.8517	0.9108	0.8973	0.9383	0.8619	0.9526	**0.9693**	**0.9964**
	Species	0.1627	0.6045	0.2733	0.6597	0.3187	0.6536	0.3029	0.6743	0.3546	0.7399	0.3764	0.7621	1.0000	1.0000	**1.0000**	**1.0000**
Fruit360	Color	0.2544	0.6054	0.3776	0.6791	0.4264	0.6868	0.4097	0.6841	0.4594	0.7392	0.4981	0.7472	0.6239	0.8243	**0.6654**	**0.8821**
	Species	0.2184	0.5805	0.2985	0.6176	0.4142	0.6984	0.3861	0.6732	0.5139	0.7430	0.5292	0.7703	0.5284	0.7582	**0.6123**	**0.8504**
Card	Order	0.0807	0.7805	0.0792	0.7128	0.1225	0.7313	0.1144	0.7658	0.1440	0.8267	0.1563	0.8326	0.3653	0.8587	**0.3921**	**0.8842**
	Suits	0.0497	0.3587	0.0430	0.3638	0.0676	0.3801	0.0716	0.3715	0.0873	0.4228	0.0933	0.6469	0.2734	0.7039	**0.3104**	**0.7941**
CMUface	Emotion	0.1284	0.6736	0.1433	0.5268	0.1592	0.6630	0.0422	0.5932	0.0161	0.5367	0.1726	0.7593	0.1786	0.7105	**0.2053**	**0.8527**
	Glass	0.1420	0.5745	0.1201	0.4905	0.1493	0.6209	0.1929	0.5627	0.1039	0.5361	0.2261	0.7663	0.3402	0.7068	**0.4870**	**0.8324**
	Identity	0.3892	0.7326	0.4637	0.6247	0.5607	0.7635	0.5109	0.8260	0.5875	0.8334	0.6360	0.8907	0.6625	0.9496	**0.7441**	**0.9834**
	Pose	0.3687	0.6322	0.3254	0.6028	0.2290	0.5029	0.4437	0.6114	0.1320	0.5517	0.4526	0.7904	0.4693	0.6624	**0.5923**	**0.8736**
Stanford Cars	Color	0.2331	0.6158	0.2103	0.5802	0.2465	0.6779	0.2336	0.6552	0.2736	0.7525	0.6899	0.8765	0.7360	0.9193	**0.7533**	**0.9387**
	Type	0.1325	0.5336	0.1650	0.5634	0.2063	0.6217	0.1963	0.5643	0.2364	0.7356	0.6045	0.7957	0.6355	0.8399	**0.6616**	**0.8792**
Flowers	Color	0.2561	0.5965	0.2938	0.5860	0.3329	0.6214	0.3169	0.6127	0.3556	0.6931	0.6327	0.7887	0.6426	0.7984	**0.6940**	**0.8843**
	Species	0.1326	0.5273	0.1561	0.6065	0.1894	0.6195	0.1887	0.6077	0.1996	0.6227	0.6148	0.8321	0.6013	0.8103	**0.6724**	**0.8719**
CIFAR-10	Type	0.1547	0.3296	0.1618	0.3305	0.1826	0.3469	0.2040	0.3695	0.2855	0.4516	0.3991	0.5827	0.4969	0.7104	**0.5271**	**0.7394**
	Environment	0.1136	0.3082	0.1379	0.3344	0.1892	0.3599	0.1920	0.3664	0.2927	0.4689	0.3782	0.5547	0.4598	0.6737	**0.4828**	**0.7096**

Conclusion and limitations

> **We propose a novel multiple clustering method (Multi-Sub) that:**

- Explicitly capture a user’s clustering interest by aligning the textual interest with the visual features of images.
- Learn the desired clustering proxy in the subspace spanned by the common categories under a user's interest
- Obtain both the desired representations and clustering simultaneously, which can significantly improve the clustering performance and efficiency

> **Limitations**

- More extensive datasets are needed

Acknowledgement

Yao's research was funded in part by J.P. Morgan Chase & Co and Advata Gift funding. Any views or opinions expressed herein are solely those of the authors listed, and may differ from the views and opinions expressed by J.P. Morgan Chase & Co. or its affiliates.

Friday, Dec 13, 2024, 11:00AM - 2:00PM
in Poster Session 5

Our Paper