Inspiration Retrieval for Visual Exploration

Nihal Jain[*]
Carnegie Mellon University
nihalj@cs.cmu.edu

Praneetha Vaddamanu
Adobe Research
vaddaman@adobe.com

Paridhi Maheshwari[*]
Stanford University
paridhi@stanford.edu

Vishwa Vinay
Adobe Research
vinay@adobe.com

Kuldeep Kulkarni
Adobe Research
kulkulka@adobe.com

1 Motivation

In creative workflows, designers compile a collection (or "moodboard") of inspirational assets for ideation. They may use this as reference for finding additional assets. In this work, we aim to stimulate creative ideation by making suggestions that take cues from the designer's moodboard but also *carefully* diverge from it. A collection of images may have rich information along different axes (e.g. color, composition, or style) – these axes or *channels* can be used to model relevance or divergence of any image from a query collection. So, we develop a self-supervised model that can extract channel-specific representations from collections of images. We propose a search algorithm that uses these representations to obtain results that satisfy the collection's intent along some channels but diverge from the query along others. We show that this allows for effective exploration of the creative space of possibilities. Finally, we demonstrate a mix-and-match visual querying mechanism that allows us to combine channels from different collections of inspirational content, thus facilitating ease in creative expression.

2 Method and Results

We begin with input representations $\mathbf{x_{mi}} \in \mathbb{R}^{d_m}$, where $m \in \mathcal{M}$ is the set of channels. In the current paper, $\mathcal{M} = \{\text{objects,style,color}\}$. We use the ResNet-152 model [3] for the *object* channel, the outputs of the ALADIN architecture [5] for the *style* channel, and a histogram over discretized LAB bins [4] as the *color* channel. These provide the input representation $\mathbf{x_{mi}}$ for each data point i and every channel m. Our model consists of three neural networks: (1) $\mathcal{F}^p_m$ produces channel-specific representations $\mathbf{z^p_{mi}} = \mathcal{F}^p_m(\mathbf{x_{mi}})$; (2) $\mathcal{F}^a_m$ provides channel-aligned representations $\mathbf{z^a_{mi}} = \mathcal{F}^a_m(\mathbf{u})$, where $\mathbf{u} = \mathcal{F}^u([\mathbf{x_{*i}}])$; (3) $\mathcal{F}^r_m$ reconstructs a channel as $\bar{\mathbf{x}}_{\mathbf{mi}} = \mathcal{F}^r_m([\mathbf{z^p_{mi}}; \mathbf{z^a_{mi}}])$. $\mathcal{F}^*_m$ and $\mathcal{F}^u$ are two-layer and one-layer feed-forward networks respectively with $ReLU$ activation between layers. For ease of notation, we vertically stack the individual feature vectors to form matrices $\mathcal{X}_m = [\mathbf{x_{mi}}]$ (containing input representations), $\mathcal{Z}^p_m = [\mathbf{z^p_{mi}}]$ (containing representations private to a channel), $\mathcal{Z}^a_m = [\mathbf{z^a_{mi}}]$ (having the aligned representations) and $\bar{\mathcal{X}}_m = [\bar{\mathbf{x}}_{\mathbf{mi}}]$ (reconstructed per-channel representations). The model is trained in a self-supervised manner, given a training set of images. The parameters for the proposed model are estimated by minimizing the following loss function: $\mathcal{L} = \lambda_1 \cdot \mathcal{L}_{ali} + \lambda_2 \cdot \mathcal{L}_{spc} + \lambda_3 \cdot \mathcal{L}_{inf} + \lambda_4 \cdot \mathcal{L}_{rec}$. Each of the loss terms capture an aspect of the desired behaviour, and take their motivations from related literature about Mutual Information Maximization [7] and balance it with Redundancy Minimization [10].

Loss Component	Definition
Inter-channel Orthogonalization	$\mathcal{L}_{spc} = \sum_{(m,m')} \left\| \mathcal{Z}^p_m * \mathcal{Z}^p_{m'} \right\|_2$
Intra-channel Information Transfer	$\mathcal{L}_{inf} = \sum_m \left(1 - trace(\mathcal{X}_m * \mathcal{Z}^p_m)\right)$
Inter-channel Alignment	$\mathcal{L}_{ali} = \sum_{(m,m')} \left(1 - trace(\mathcal{Z}^a_m * \mathcal{Z}^a_{m'})\right)$
Intra-channel Reconstruction	$\mathcal{L}_{rec} = \sum_m \left\| \bar{\mathcal{X}}_m - \mathcal{X}_m \right\|_2$

[*]Work done when authors were at Adobe Research

35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia.

Given a collection $\mathcal{C}$ of new images, we compute the orthogonalized per-channel representations $\hat{\mathbf{c}}^{\mathbf{p}}_{\mathbf{mi}}$ for $i \in \mathcal{C}$. The collection-level representation for a channel m ($\mathbf{C}^{\mathbf{p}}_{\mathbf{m}}$) is taken to be the mean of the channel-specific representations over all images in $\mathcal{C}$. To obtain the intent of a collection, we first compute the average pairwise cosine similarity between images in the collection along a given channel m: $\hat{\beta}_m = \frac{1}{N\times(N-1)} \sum_{(i,j)} dot(\hat{\mathbf{c}}^{\mathbf{p}}_{\mathbf{mi}}, \hat{\mathbf{c}}^{\mathbf{p}}_{\mathbf{mj}})$. To ensure that these raw intent weights are comparable across channels, we standardize the raw channel weights as: $\beta_m = \frac{\hat{\beta}_m - \mu_m}{\sigma_m}$, where μ_m and σ_m are the mean and standard deviation of the pairwise similarities between any two different images from the dataset, measured in the embedding space for channel m. Finally, we normalize across channels so that the intent weights sum to 1: $\alpha_m = \frac{exp(\beta_m)}{\sum_{m'} exp(\beta_{m'})}$.

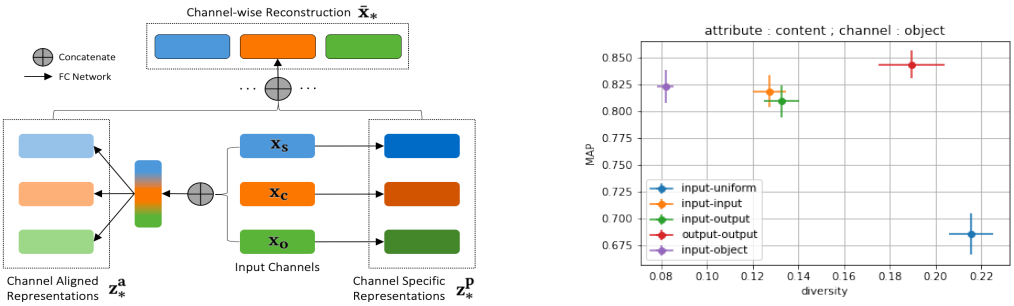

Figure 1: Left: Our proposed model that captures aspects of Multi-view Learning [8, 1], and the learning of Factorized [6] and Disentangled [2] Representations. Right: MAP-diversity trade-off with $\lambda_1 = 0.001$, $\lambda_2 = 0.05$, $\lambda_3 = 0.0001$ and $\lambda_4 = 0.0001$.

For collection $\mathcal{C}$, we rank a corpus of images $\mathcal{D}$ in decreasing order of relevance to the intent of the collection. From our model, we obtain channel-specific representations – $\mathbf{C_m}$ for the collection, and $\mathbf{d_m}$ for a test image $d \in \mathcal{D}$. We then compute a weighted similarity score for d as $s_{\mathcal{C},d} = \sum_m \alpha_m \cdot sim\,(\mathbf{C_m}, \mathbf{d_m})$. Here, $sim\,(\mathbf{a}, \mathbf{b})$ is the appropriate measure of similarity between $\mathbf{a}$ and $\mathbf{b}$.

Figure 1 (Right) shows the diversity (computed as $1/\beta_m$) along the object channel on $x - axis$, and the Mean Average Precision on the $y - axis$. Each point corresponds to a pair of representation choices – for the retrieval similarity computation and for the intent computation– indicated by *similarity-intent* in the legend. The baseline of using only one channel (here, object) is the left most point, indicating least diversity. Uniformly weighting all channels (blue point) provides the highest diversity, but trades in ranking relevance. The *output-output* combination from our model provides the best benefits of both relevance and diversity. Note that visual exploration requires a thorough exploration of the creative design space, therefore making diversity a core requirement.

Deriving disentangled channel representations from a collection of assets also enables the novel use-case of composing multiple moodboards to search for inspirational assets. By selectively picking channel representations from different existing collections, we can create a composite representation for a new collection that can be used for the expansion task. Figure 2 shows collections of 'bicycle' and 'vector art' images. By selecting the object representations from the former, style representations from the latter, and averaging out the color representations between the two, we composite representations for a hypothetical collection that has the object features of the former and style features of the latter. By ensuring that representations along different channels are de-correlated, these dimensions can be mixed and matched, allowing for a powerful visual querying mechanism.

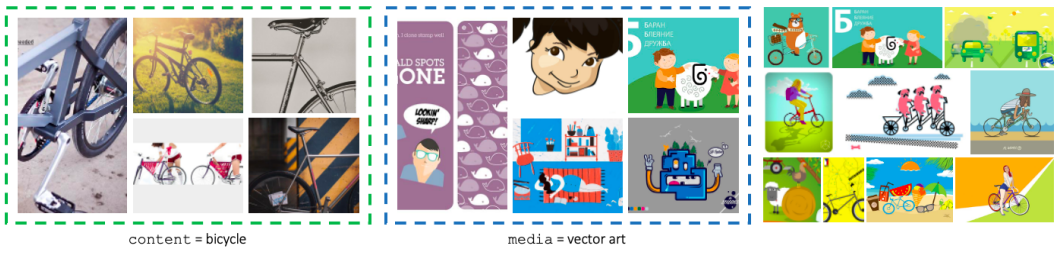

Figure 2: Left – moodboard with *object* intent ('bicycles'), center – *style* intent ('vector art'). Images of bicycles styled in vector art form are retrieved on the right. All images from the BAM dataset [9].

References

[1] Learning robust representations via multi-view information bottleneck. *arXiv preprint arXiv:2002.07017*, 2020.

[2] Cian Eastwood and Christopher KI Williams. A framework for the quantitative evaluation of disentangled representations. In *International Conference on Learning Representations*, 2018.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[4] Paridhi Maheshwari, Manoj Ghuhan, and Vishwa Vinay. Learning colour representations of search queries. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1389–1398, 2020.

[5] Dan Ruta, Saeid Motiian, Baldo Faieta, Zhe Lin, Hailin Jin, Alex Filipkowski, Andrew Gilbert, and John Collomosse. Aladin: All layer adaptive instance normalization for fine-grained style similarity. In *Proceedings of the International Conference on Computer Vision*, 2021.

[6] Yao-Hung Hubert Tsai, Paul Pu Liang, Amir Zadeh, Louis-Philippe Morency, and Ruslan Salakhutdinov. Learning factorized multimodal representations. *arXiv preprint arXiv:1806.06176*, 2018.

[7] Michael Tschannen, Josip Djolonga, Paul K Rubenstein, Sylvain Gelly, and Mario Lucic. On mutual information maximization for representation learning. *arXiv preprint arXiv:1907.13625*, 2019.

[8] Weiran Wang, Raman Arora, Karen Livescu, and Jeff Bilmes. On deep multi-view representation learning. In *International conference on machine learning*, pages 1083–1092. PMLR, 2015.

[9] Michael J Wilber, Chen Fang, Hailin Jin, Aaron Hertzmann, John Collomosse, and Serge Belongie. Bam! the behance artistic media dataset for recognition beyond photography. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1202–1211, 2017.

[10] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. *arXiv preprint arXiv:2103.03230*, 2021.