Probabilistic Compositional Embeddings for Multimodal Image Retrieval (Supplementary Materials)

Andrei Neculai[1], Yanbei Chen[1], Zeynep Akata[1,2,3]
[1]University of Tübingen [2]MPI for Informatics [3]MPI for Intelligent Systems
neculai.andrei0@gmail.com, {yanbei.chen, zeynep.akata}@uni-tuebingen.de

A. Additional Details in Benchmark Setup

Dataset and benchmark setup. In order to construct our benchmark, we start from the 118,287 images in the training split of the MS-COCO dataset [6]. We split the images with a 4:1:1 ratio for the train/validation/test sets. Our benchmark is designed for the task of composing the concepts (which are object categories in this work) specified in the input queries for image retrieval. For this aim, we need to generate compositions of individual concepts that are well presented in a sufficient amount of images to serve for the retrieval. We choose 8:2:2 as the splitting thresholds for the train/validation/test splits. This means that a composition of concepts is valid only if it is present in at least 8 of the images in the training set, and 2 of the images in the validation and testing sets respectively. These values ensure that there is enough diversity in the training phase and also multiple correct answers (i.e. more than 1) in the evaluation phase. Using these thresholds allows us to find 1000 viable compositions for using either 2, 3 or 4 input queries. Algorithm A details the process of generating a compositions of k inputs. We will release our benchmark upon acceptance.

Training. During training, for every concept (i.e. category) in the compositional inputs, we randomly choose an image id containing that concept, and also randomly pick either the visual or textual the modality to represent it. The target image is chosen randomly from the images that contain all the k concepts specified in the inputs, but the modality is fixed to visual as we aim to solve an image retrieval task.

Evaluation. At test time, we evaluate for composing k inputs (with k fixed for an evaluation setup). The test queries are described by k concepts and we generate all the test queries randomly to obtain a roughly equal mix of modality combinations. For example, for $k = 2$, we will have around 25% cases for each modality mixture in {*image-image*, *image-text*, *text-image*, *text-text*}.

We also establish evaluation setups to test the model's ability of recognizing unseen compositions for retrieval, as well as testing the model's ability of identifying feasible/infeasible compositions. For the case of testing unseen compositions, we use $k = 2$ and generate 100 compositions of different concepts for training and 500 new compositions of different concepts for testing, where the concepts at test time are seen during training but their compositions are unseen. We choose to use only 100 compositions for the training phase but 500 unseen compositions to simulate a more challenging evaluation setup. For testing the model's ability of identifying feasible/infeasible compositions, we build the evaluation setup upon the data generated for the seen compositions scenario for $k = 2$. On top of the 1000 compositions, we generate an additional 250 unseen compositions and 250 infeasible compositions. An infeasible composition is defined as infeasible given that it is not found in any image across all the images in the dataset.

B. Numerical Results

We present the numerical results used to generate Figure 4 in the main paper. Figure 4 shows the model performance when being trained to compose two inputs but tested on composing a varying number of inputs (i.e. 2, 3 and 4). We present these numerical results in Table A.

C. Additional Evaluation

Training with more inputs. In the main paper, we explored the scenarios where we train our model to compose two and three query inputs. However, our model formulation is not limited to a fixed number of inputs. Hence, we further explore the scenario where we train the models to compose four query inputs here. Table B contains the results for the top performing models when trained to compose four query inputs and be evaluated on a seen composition setup. As we can see, our model MPC achieves the best performance on the composition of multimodal and text-only inputs, obtaining a R@5 of 5.98%/8.46% vs 4.76%/7.40% by the best competitor MRN. Although MPC does not surpass TIRG on the compositions of image-only inputs, its performance is quite close to TIRG, obtaining a R@5 of 2.76% vs 3.40%. Another interesting observation is that while the other probabilistic model PCME+addition performs rel-

Algorithm A Algorithm for generating a compositions of k concepts (i.e. categories)

```
compositions_set ← empty set                          ▷ initialize a set to store the compositions of k concepts
num_of_compositions ← 0                               ▷ record the number of found compositions in dataset
target_num_of_compositions ← 1000                     ▷ expected number of compositions to be found
threshold_{train/val/test} ← 8 : 2 : 2                ▷ minimal numbers of compositions in the train/val/test splits
while num_of_compositions < target_num_of_compositions do
    cur_composition ← sample k categories w/out replacement   ▷ get images for the current composition of k concepts
    images_{train/val/test} ← get {train/val/test} images that contain "cur_composition"
    if len(images_{train/val/test) ≥ threshold_{train/val/test} then
        if cur_composition not in compositions_set then
            compositions_set.insert(cur_composition)          ▷ add a found composition to the set
            num_of_compositions ++
        end if
    end if
end while
```

Input modalities	images only			multimodal			texts only		
Number of inputs	2	3	4	2	3	4	2	3	4
MRN [4]	28.17	6.55	0.97	34.84	7.88	1.49	42.48	8.01	2.57
TIRG [7]	31.58	7.52	0.16	26.85	5.60	0.03	51.43	4.94	0.00
PCME [2] + addition	21.85	10.19	4.87	29.41	15.65	7.95	44.43	20.34	11.48
MPC	**36.52**	**18.93**	**6.17**	**48.23**	**27.73**	**11.85**	**69.42**	**41.02**	**14.35**

Table A. Numerical results of model performance trained for composing 2 inputs but evaluated for composing 2, 3, and 4 inputs. Results are accompanied to Figure 4 in the paper. Metrics: R@10 (%).

Method	images only			multimodal			texts only		
	R@5	R@10	R_P	R@5	R@10	R_P	R@5	R@10	R_P
MRN [4]	2.11	5.19	0.50	4.76	8.44	1.03	7.40	11.78	1.51
TIRG	**3.40**	**6.98**	**0.63**	4.73	7.98	0.97	7.10	9.67	1.21
PCME + addition	0.00	0.16	0.01	0.13	0.19	0.31	0.00	0.00	0.00
MPC	2.76	5.52	0.59	**5.98**	**10.03**	**1.30**	**8.46**	**16.77**	**1.95**

Table B. Evaluation of composing ***four*** query inputs for image retrieval on a ***seen*** composition setup.

atively well when composing two or three inputs, its performance degrades when composing four inputs. This suggests that composition with addition is weaker at preserving the information from more inputs, while our MPC model formulation with probabilistic composer can better capture the additive information from the increasingly more inputs.

Evaluation on Fashion200k. In the main paper, we showed how our model is capable of composing queries of arbitrary sizes and modalities. In this section we want to showcase the performance of our model on the problem of image retrieval using image and attribute-based text feedback [1, 5, 7]. This is a slightly different setting than our original problem, as in this case the aim is to model the interactions between the image and text queries instead of adding them together. Fashion200k [3] is a dataset that contains around 200k images of fashion products. Each image is tagged with a set of attributes. Using these attributes, [7] generated pairs of products that differ by only one attribute. The text modifications are generated using the attribute that is different between the 2 products. Table C contains the results on the Fashion200k dataset. We limited our comparisons to methods that are capable of modeling the original task in the paper. Other methods [1, 5] have better performance, but use complex network architectures, specialized on this task, that can not be used to model arbitrary queries. Because of this, we chose to not add them to the comparison. As we can see, our model MPC achieves the best performance among the compared methods. We can also see that using simple additions of probabilistic embeddings (i.e. PCME + addition) is not capable of modeling the interactions between the image and text parts of the query, leading to poor results. These results show that our MPC also generalizes well on the other benchmark dataset Fashion200k.

Method	R@1	R@10	R@50
Relationship	13.0	40.5	62.4
Film	12.9	39.5	61.9
MRN	12.3	39.4	60.9
TIRG	14.1	42.5	63.8
PCME + addition	1.8	11.4	27.3
MPC	**14.6**	**45.4**	**66.0**

Table C. Retrieval results on the Fashion200k dataset.

References

[1] Yanbei Chen, Shaogang Gong, and Loris Bazzani. Image search with text feedback by visiolinguistic attention learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 2

[2] Sanghyuk Chun, Seong Joon Oh, Rafael Sampaio De Rezende, Yannis Kalantidis, and Diane Larlus. Probabilistic embeddings for cross-modal retrieval. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[3] Xintong Han, Zuxuan Wu, Phoenix X Huang, Xiao Zhang, Menglong Zhu, Yuan Li, Yang Zhao, and Larry S Davis. Automatic spatially-aware fashion concept discovery. In *IEEE International Conference on Computer Vision*, 2017. 2

[4] Jin-Hwa Kim, Sang-Woo Lee, Donghyun Kwak, Min-Oh Heo, Jeonghee Kim, Jung-Woo Ha, and Byoung-Tak Zhang. Multimodal residual learning for visual qa. In *Advances in Neural Information Processing Systems*, 2016. 2

[5] Seungmin Lee, Dongwan Kim, and Bohyung Han. Cosmo: Content-style modulation for image retrieval with text feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 802–812, June 2021. 2

[6] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015. 1

[7] Nam Vo, Lu Jiang, Chen Sun, Kevin Murphy, Li-Jia Li, Li Fei-Fei, and James Hays. Composing text and image for image retrieval - an empirical odyssey. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 2