CoSMo: Content-Style Modulation for Image Retrieval with Text Feedback

Supplementary Document

Seungmin Lee* Dongwan Kim* Bohyung Han
Seoul National University
{dltmdals14, dongwan123, bhhan}@snu.ac.kr

Appendix A. Identity of λ_i

This section uncovers the identity of λ_i, which is introduced in Eq. (13). From Eq. (12) and Eq. (10), we can derive the following equation:

$$\begin{aligned}\omega(\mathbf{z}_i^r, \mathbf{z}_j^r, \mathbf{t}) &= \psi(\mathbf{q}_i^T \mathbf{k}_j) \\ &= \psi\left(A_j + B_j\right) \\ &= \frac{\exp\left(A_j + B_j\right)}{\sum_{s\in\Omega} \exp\left(A_s + B_s\right)} \\ &= \frac{\exp\left(A_j + B_j\right)}{C_i},\end{aligned} \tag{a-1}$$

where $C_i = \sum_{s\in\Omega} \exp\left(A_s + B_s\right)$. Note that we use subscript i in C_i to reflect that C_i is dependent on the value of $\mathbf{z}_i^r$. We can further develop the above equation as follows:

$$\begin{aligned}\omega(\mathbf{z}_i^r, \mathbf{z}_j^r, \mathbf{t}) &= \frac{\exp\left(A_j + B_j\right)}{C_i} \\ &= \frac{1}{C_i} \exp\left(A_j\right) \exp\left(B_j\right) \\ &= \frac{\sum_{s\in\Omega} \exp\left(A_s\right) \cdot \sum_{s\in\Omega} \exp\left(B_s\right)}{C_i} \frac{\exp\left(A_j\right)}{\sum_{s\in\Omega} \exp\left(A_s\right)} \frac{\exp\left(B_j\right)}{\sum_{s\in\Omega} \exp\left(B_s\right)} \\ &= \lambda_i \psi\left(A_j\right) \cdot \psi\left(B_j\right),\end{aligned} \tag{a-2}$$

where $\lambda_i = \frac{1}{C_i} \sum_{s\in\Omega} \exp\left(A_s\right) \cdot \sum_{s\in\Omega} \exp\left(B_s\right)$. As seen in the above equations, λ_i is simply a normalization constant.

Appendix B. FashionIQ Results

As mentioned in the main paper, we follow the evaluation scheme of VAL [1] for the FashionIQ dataset. While this is done for fair comparison, we realize that there is a no need for future works to follow this scheme as well. Thus, we provide a second version of our FashionIQ results, using the *standard* evaluation scheme, *i.e*. the originally intended scheme proposed by the FashionIQ authors. However, since we were unable to fully reproduce the results of VAL (using the VAL evaluation method), we only compare with some TIRG [2], as well as some basic baselines such as *Image Only* and *Concat*.

Method	Dress		Toptee		Shirt		Average	
	R@10	R@50	R@10	R@50	R@10	R@50	R@10	R@50
Image Only	4.46	13.19	5.46	13.21	6.13	13.64	5.35	13.35
Concat	14.92	34.95	14.28	34.73	12.71	30.08	13.92	33.25
TIRG [2]	14.13	34.61	14.79	34.37	13.10	30.91	14.01	33.30
Ours	**21.39** $\pm$ 0.29	**44.45** $\pm$ 0.32	**21.32** $\pm$ 0.17	**46.02** $\pm$ 0.10	**16.90** $\pm$ 0.28	**37.49** $\pm$ 0.17	**19.87**	**42.62**

Table A-1. Results on FashionIQ using ResNet-50 and Original Validation Sets.

Appendix C. Qualitative Results

Here we display some qualitative results of our model on the FashionIQ Dress and Shoes datasets. The very leftside contains the reference image and the modifier text (in the case of FashionIQ, there are two modifier texts), followed by top 6 retrieved samples by our model. The green bounding box indicates a correct prediction, while the red one indicates an incorrect result.

Figure A-1. Some visualized examples from the FashionIQ Dress dataset where our model made correct predictions (retrieval).

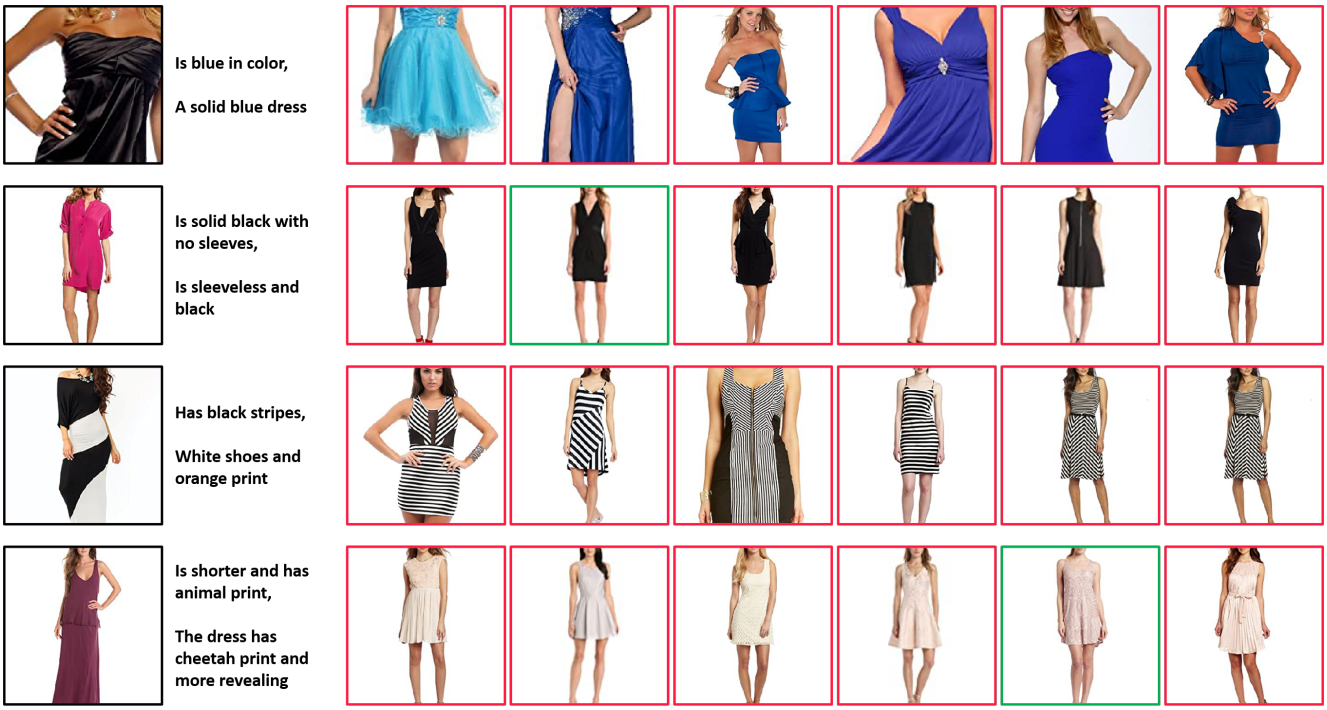

Figure A-2. Some visualized examples from the FashionIQ Dress dataset where our model made incorrect predictions (retrieval). Note that, although these predictions are all considered incorrect, our model retrieves some reasonable samples.

Figure A-3. Some visualized examples of the Shoes dataset.

References

[1] Yanbei Chen, Shaogang Gong, and Loris Bazzani. Image search with text feedback by visiolinguistic attention learning. In *CVPR*, 2020.

[2] 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 *CVPR*, 2019.