Image Search with Text Feedback by Visiolinguistic Attention Learning

CVPR SEATTLE WASHINGTON JUNE 16-18 2020

Yanbei Chen[1] Shaogang Gong[1] Loris Bazzani[2]

yanbei.chen@qmul.ac.uk s.gong@qmul.ac.uk bazzanil@amazon.com

[1]Queen Mary University of London [2]Amazon

Queen Mary University of London

amazon

Introduction

Problem

- ❑ Given a *reference image* and *user text* as input query, we consider to retrieve new images that resemble the reference image while changing certain aspects as specified by text.
- ❑ The text can be given as *attribute* or *natural language*.

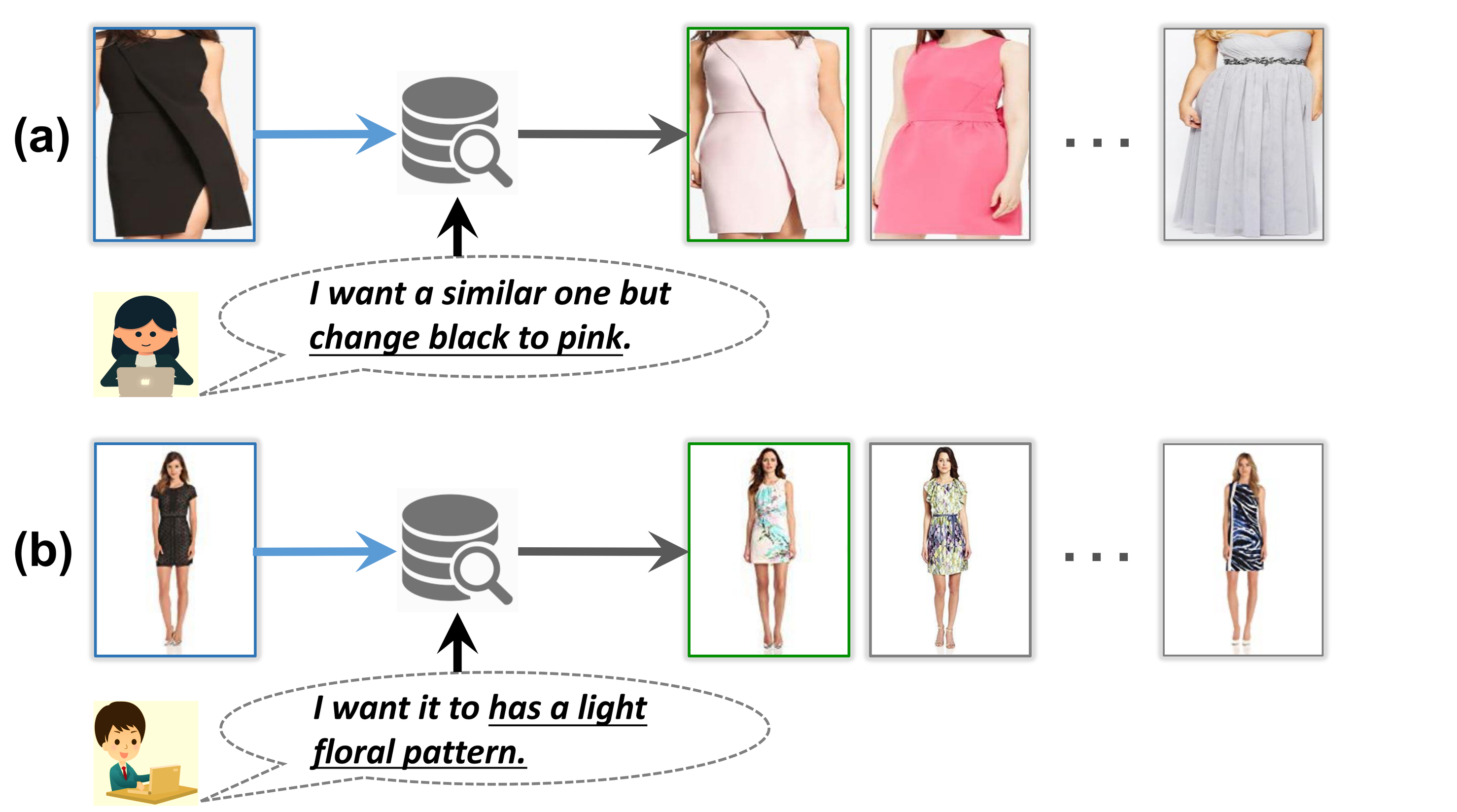

Figure 1. Problem illustration of image search with text feedback. The text describes the visual content to refine in the reference image, ranging from **(a)** a concrete attribute to **(b)** more abstract visual properties such as fashion style.

Main Challenges

- ❑ simultaneously *preserve* and *transform* the visual content in accordance with the text feedback
- ❑ learn a composite representation that jointly encapsulate visual and textual contents from *coarse* to *fine-grain*

Main Idea

➢ **Visiolinguistic Attention Learning (VAL)**

❖ **model architecture**

– composite transformers at multi-level

- *attentional transformation* and *preservation*
- fuse vision and language features via attention learning at varying representation depths.

❖ **learning objective**

– hierarchical matching

- align with the target visual and textual representations in a two-level hierarchical space

Methodology

Proposed Approach

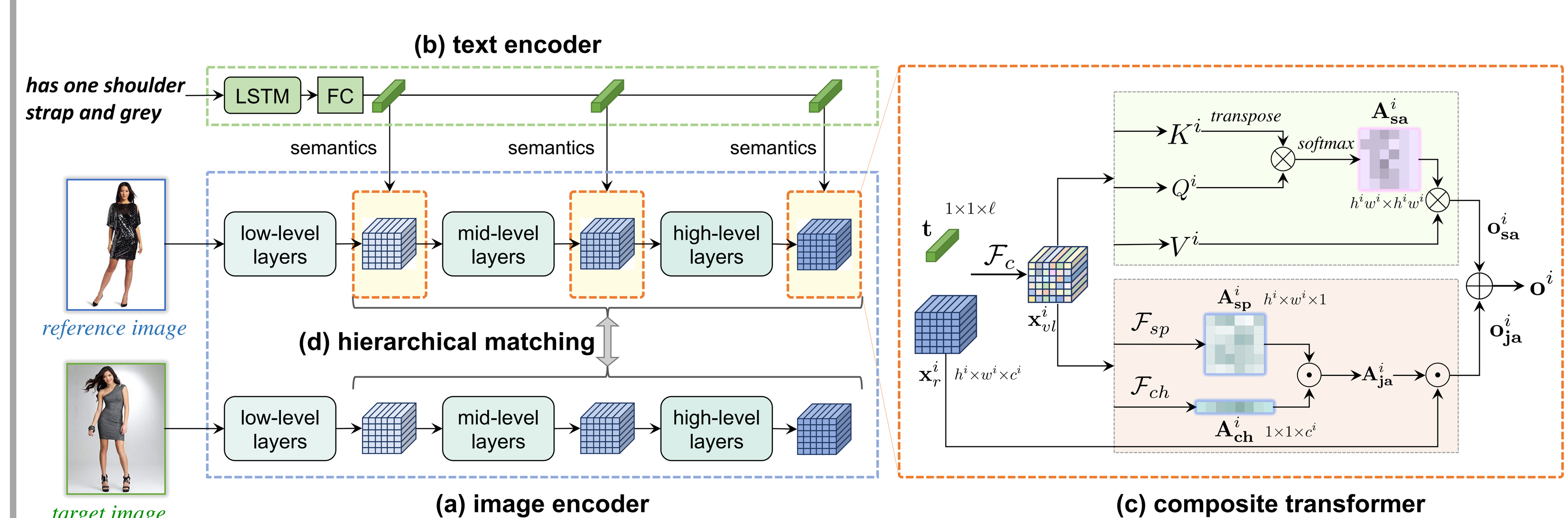

Figure 2. An **overview** of our Visiolinguistic Attention Learning (VAL) framework

❑ **Composite transformers** at multi-level inside the CNN

- visiolinguistic representation

$$\mathbf{x}^i_{vl} = \mathcal{F}_c([\mathbf{x}^i_r, \mathbf{t}])$$

- self-attentional transformation

$$Q^i = \mathcal{F}_Q(\mathbf{x}^i_{vl}),\ K^i = \mathcal{F}_K(\mathbf{x}^i_{vl}),\ V^i = \mathcal{F}_V(\mathbf{x}^i_{vl})$$

$$\mathbf{A}^i_{\mathbf{sa}} = \mathrm{softmax}(\frac{Q^i {K^i}^T}{\sqrt{c}}) \qquad \mathbf{o}^i_{\mathbf{sa}} = \mathcal{F}_{sa}(\mathbf{A}^i_{\mathbf{sa}} V)$$

- joint-attentional preservation (spatial-wise & channel-wise)

$$\mathbf{A}^i_{\mathbf{sp}} = \mathrm{sigmoid}(\mathcal{F}_{sp}(\frac{1}{c^i}\sum_j^{c^i} \mathbf{x}^i_{vl}(:,:,j))),\ \mathbf{A}^i_{\mathbf{ch}} = \mathrm{sigmoid}(\mathcal{F}_{ch}(\frac{1}{h^i \times w^i}\sum_j^{h^i}\sum_k^{w^i} \mathbf{x}^i_{vl}(j,k,:)))$$

$$\mathbf{A}^i_{\mathbf{ja}} = \mathbf{A}^i_{\mathbf{sp}} \odot \mathbf{A}^i_{\mathbf{ch}} \qquad \mathbf{o}^i_{\mathbf{ja}} = \mathbf{A}^i_{\mathbf{ja}} \odot \mathbf{x}^i_r$$

- composite representation of image and text

$$\mathbf{o}^i = w_{sa}\mathbf{o}^i_{\mathbf{sa}} + w_{ja}\mathbf{o}^i_{\mathbf{ja}}$$

❑ **Hierarchical Matching**

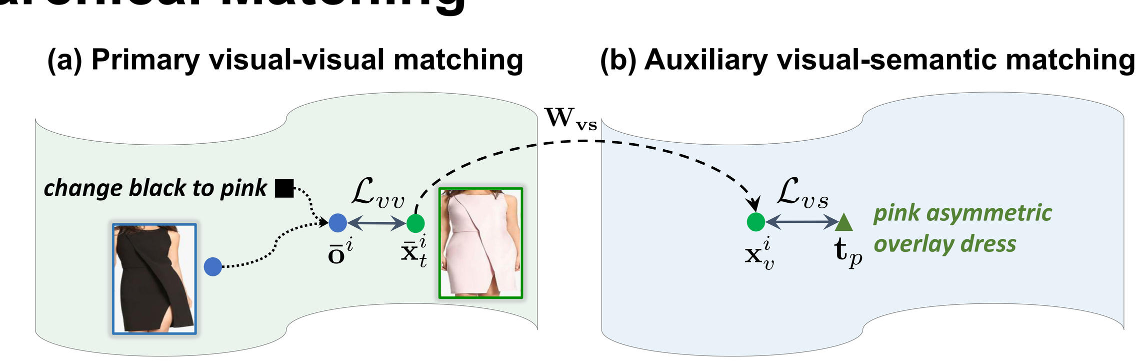

Figure 3. Discriminative feature learning in a two-level hierarchical space

a) visual space (tie with the target image $\bar{\mathbf{x}}^i_t$)

$$\mathcal{L}_i(\bar{\mathbf{o}}, \bar{\mathbf{x}}^i_t) = \max(0, d(\bar{\mathbf{o}}^i, \bar{\mathbf{x}}^i_t) - d(\bar{\mathbf{o}}^i, \bar{\mathbf{x}}^i_n) + m)$$

b) semantic space (tie with the target tagged text $\mathbf{t}_p$)

$$\mathcal{L}_i(\mathbf{x}^i_v, \mathbf{t}_p) = \max(0, d(\mathbf{x}^i_v, \mathbf{t}_p) - d(\mathbf{x}^i_v, \mathbf{t}_n) + m)$$

Experiments

Experiments on three benchmark datasets

❑ **Qualitative results**

Figure 4. Image search with *attribute-like* text feedback on Fashion200k

Figure 5. Image search with *natural language* feedback on FashionIQ/Shoes

❑ **Quantitative results**

Method	R@1	R@10	R@50
Han et al. [18]	6.3	19.9	38.3
Show and Tell [65]	12.3	40.2	61.8
Param Hashing [43]	12.2	40.0	61.7
FiLM [47]	12.9	39.5	61.9
Relationship [53]	13.0	40.5	62.4
MRN [25]	13.4	40.0	61.9
TIRG [66]	14.1	42.5	63.8
MRN	14.2	43.6	63.8
TIRG	14.8	43.7	64.1
VAL ($\mathcal{L}_{vv}$)	**21.2**	**49.0**	**68.8**
VAL ($\mathcal{L}_{vv} + \mathcal{L}_{vs}$)	**21.5**	**53.8**	**73.3**
VAL (GloVe)	**22.9**	**50.8**	**72.7**

Table 1. Fashion200k

Method	R@1	R@10	R@50
FiLM	10.19	38.89	68.30
MRN	11.74	41.70	67.01
Relationship	12.31	45.10	71.45
TIRG	12.60	45.45	69.39
VAL ($\mathcal{L}_{vv}$)	**16.49**	**49.12**	**73.53**
VAL ($\mathcal{L}_{vv} + \mathcal{L}_{vs}$)	**16.98**	**49.83**	**73.91**
VAL (GloVe)	**17.18**	**51.52**	**75.83**

Table 2. Shoes

Method	Dress		Shirt		Toptee		Avg	
	R@10	R@50	R@10	R@50	R@10	R@50	R@10	R@50
TIRG	8.10	23.27	11.06	28.08	7.71	23.44	8.96	24.93
Image+Text Concatenation	10.52	28.98	13.44	34.60	11.36	30.42	11.77	31.33
Side Information [17]	11.24	32.39	13.73	37.03	13.52	34.73	12.82	34.72
MRN	12.32	32.18	15.88	34.33	18.11	36.33	15.44	34.28
FiLM	14.23	33.34	15.04	34.09	17.30	37.68	15.52	35.04
TIRG	14.87	34.66	18.26	37.89	19.08	39.62	17.40	37.39
Relationship	15.44	38.08	18.33	38.63	21.10	44.77	18.29	40.49
VAL ($\mathcal{L}_{vv}$)	**21.12**	**42.19**	**21.03**	**43.44**	**25.64**	**49.49**	**22.60**	**45.04**
VAL ($\mathcal{L}_{vv} + \mathcal{L}_{vs}$)	**21.47**	**43.83**	**21.03**	**42.75**	**26.71**	**51.81**	**23.07**	**46.13**
VAL (GloVe)	**22.53**	**44.00**	**22.38**	**44.15**	**27.53**	**51.68**	**24.15**	**46.61**

Table 3. FashionIQ

Reference

[1] Xiaoxiao Guo, Hui Wu, Yupeng Gao, Steven Rennie, and Rogerio Feris. The fashion iq dataset: Retrieving images by combining side information and relative natural language feedback. ICCVW19'

[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. CVPR19'