Image Search with Text Feedback by Visiolinguistic Attention Learning (Supplementary Material)

Yanbei Chen
Queen Mary University of London
yanbei.chen@qmul.ac.uk

Shaogang Gong
Queen Mary University of London
s.gong@qmul.ac.uk

Loris Bazzani
Amazon
bazzanil@amazon.com

1. Architecture Details

Part Name	Layer Description
composite function	$\mathcal{F}_c$: conv(K-1×1, N-c^i, ReLU)
SA stream	$\mathcal{F}_Q, \mathcal{F}_K, \mathcal{F}_V$: conv(K-1×1, N-$\bar{c}^i$) $\mathcal{F}_{sa}$: conv(K-1×1, N-c^i, ReLU)
JA stream	$\mathcal{F}_{sp}$: conv(K-$h^i \times w^i$, N-1, sigmoid) $\mathcal{F}_{ch}$: conv(K-1×1, N-c^i, sigmoid)

Table 1. Architecture of composite transformer. SA: self-attention. JA: joint-attention. conv(K,N): stands for convolution layer, where K: filter size, N: number of filters. $\bar{c}^i = \frac{c^i}{\text{num\_heads}}$, num_heads = 2.

Training. Table 1 details the architecture of our composite transformer. To learn the transformation at varying depths, we plug three composite transformers into the CNN at the low, mid, high-level. In ResNet-50, the low, mid, high-level feature maps are from the last three residual blocks, which give feature tensors of size $16 \times 16 \times 512$, $8 \times 8 \times 1024$, $8 \times 8 \times 2048$. In MobileNet, the low, mid, high-level are set as the $6, 11, 13$th layer, which give feature tensors of size $16 \times 16 \times 512$, $16 \times 16 \times 512$, $8 \times 8 \times 1024$.

Testing. At inference time, outputs from three composite transformers are average-pooled and concatenated to derive the *composite feature*. The *test image feature* is simply the concatenation of average-pooled features at the low, mid, high-level. For retrieval, the *composite feature* is compared with *test image features* by measuring their pairwise similarities, formally computed as L2 distance.

Computational costs. At test time, the computational costs are decided by **(1)** model complexity (FLOPs); **(2)** matching and ranking. On **(1)**, our composite transformers have FLOPs (8.10×10^7) vs. ResNet50 (3.80×10^9), which bring small computational cost - an additional FLOPs of 2.13%. On **(2)**, the complexity is $\mathcal{O}(QN)$, $\mathcal{O}(QNlogN)$ for similarity matching, ranking – Q, N are the size of query, test set. We implemented similarity matching on GPU, which yields lower computational cost than CPU implementation.

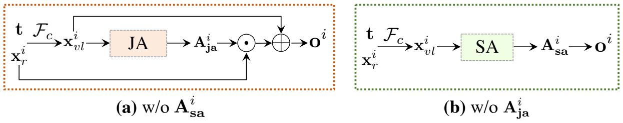

Figure 1. A schematic illustration of two baselines.

Ablative Baselines. As aforementioned, we test our VAL in comparison to two ablative baselines: **(a)** w/o self-attention (w/o $\mathbf{A}_{\mathbf{sa}}^i$), and **(b)** w/o joint-attention (w/o $\mathbf{A}_{\mathbf{ja}}^i$). We show a graphical illustration of these two baselines in Fig. 1.

2. Dataset and Training Details

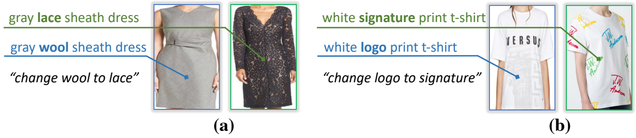
gray **lace** sheath dress
gray **wool** sheath dress
"change wool to lace"
(a)
white **signature** print t-shirt
white **logo** print t-shirt
"change logo to signature"
(b)

Figure 2. Example image pairs on Fashion200k. For each pair, we show the tagged attribute-like product descriptions of the *reference image* and *target image*; while the *user text* (in quotation marks) describes the difference between two images in *attributes*.

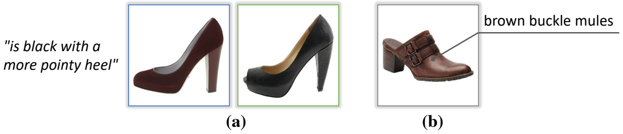

Figure 3. Examples on Shoes. **(a)** Image pair with *relative caption* in natural language. **(b)** Example image with tagged description.

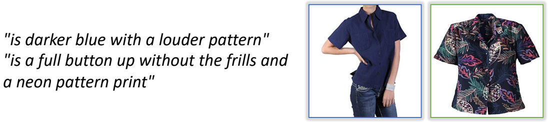
"is darker blue with a louder pattern"
"is a full button up without the frills and a neon pattern print"

Figure 4. Image pair tagged with *relative captions* on FashionIQ.

We present illustration of different datasets in Fig. 2, 3 and 4. Below, we detail how each dataset is used in training.

Fashion200k. In training, we utilise the samples that can find their corresponding pairs with one word difference by comparing the tagged attribute-like product descriptions, as shown in Fig. 2. In VAL ($\mathcal{L}_{vv} + \mathcal{L}_{vs}$), we exploit the tagged

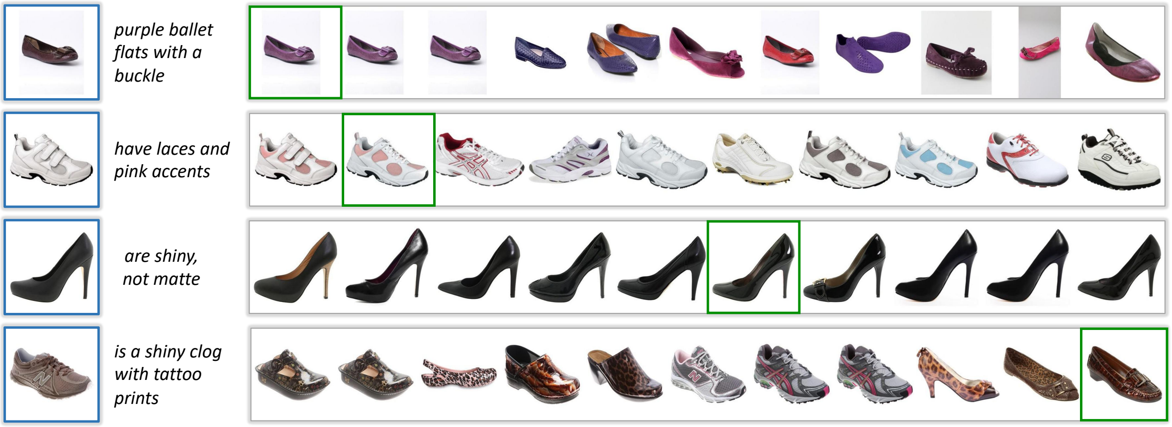
purple ballet flats with a buckle
have laces and pink accents
are shiny, not matte
is a shiny clog with tattoo prints

Figure 5. Qualitative results on Shoes. First two examples: "***success***" cases with small R@K (i.e. R@1, R@2); Last two examples: "***failure***" cases with relatively larger R@K (i.e. R@6, R@10). blue/green boxes: reference/target images.

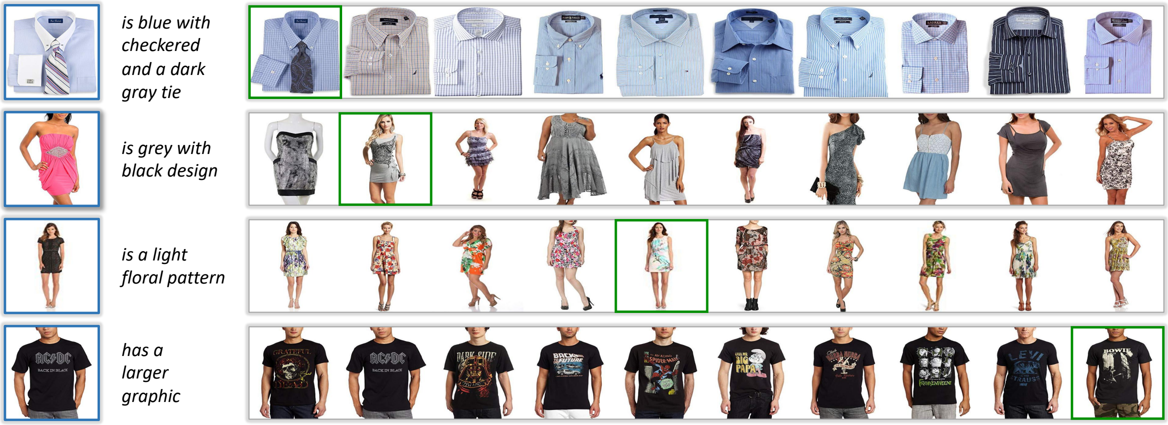
is blue with checkered and a dark gray tie
is grey with black design
is a light floral pattern
has a larger graphic

Figure 6. Qualitative results on FashionIQ. First two examples: "***success***" cases with small R@K (i.e. R@1, R@2); Last two examples: "***failure***" cases with relatively larger R@K (i.e. R@5, R@10). blue/green boxes: reference/target images.

descriptions as side information, which serve as auxiliary supervision to train our VAL via a *joint-training* objective as $\mathcal{L}_{vv} + \mathcal{L}_{vs}$. We train for 160k iterations on Fashion200k.

Shoes. In training, we use 17,954 image pairs with relative captions (Fig. 3(**a**)). In VAL ($\mathcal{L}_{vv} + \mathcal{L}_{vs}$), we use the tagged descriptions (Fig. 3(**b**)) of 3,000 samples as side information for *pre-training* via $\mathcal{L}_{vs}$; then we fine-tune with the primary objective $\mathcal{L}_{vv}$ for 30 iterations.

FashionIQ. In training, we use all training pairs with tagged relative captions (Fig 4). In VAL ($\mathcal{L}_{vv} + \mathcal{L}_{vs}$), due to missing narrative descriptive texts tagged for each sample, we exploit Fashion200k to provide auxiliary supervision ($\mathcal{L}_{vs}$) via *pre-training*, and fine-tune with the primary objective $\mathcal{L}_{vv}$. We train for 50k iterations on FashionIQ.

3. Additional Qualitative Results

We provide additional qualitative results on Shoes and FashionIQ (Fig. 5, 6) to further give an in-depth analysis when using *natural language* based text feedback.

Further Analysis. Unlike attribute-like text feedback that generally describes a concrete visual concept, *natural language* text feedback may be highly *abstract*, thus likely to be *ambiguous* and indicate *multiple possibilities*. As Fig. 5, 6 show, there are multiple "*failure*" cases, which show the model does properly return the "desired" images that resemble the reference images whilst reflecting changes specified in the input texts. For instance, in the 3rd example in Fig. 6, there are more than one dress that contains "`a light floral pattern`" in the top retrieved items; but the target image is only R@5, mostly because the input text does indicate multiple possible desired outcomes.

Future Work. Overall, these results suggest that natural language based text feedback could sometimes be *ambiguous*, and thus indicate *multiple possible* desired items rather than a single one. To further examine or address this issue, we consider there are several potential research directions: (1) propose *new evaluation metrics* to quantify visual similarities among the top retrieved items; and (2) conduct *human studies* to test the interactive image retrieval performance in practical use.