Image Search with Text Feedback by Visiolinguistic Attention Learning

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

Abstract

Image search with text feedback has promising impacts in various real-world applications, such as e-commerce and internet search. Given a reference image and text feedback from user, the goal is to retrieve images that not only resemble the input image, but also change certain aspects in accordance with the given text. This is a challenging task as it requires the synergistic understanding of both image and text. In this work, we tackle this task by a novel Visiolinguistic Attention Learning (VAL) framework. Specifically, we propose a composite transformer that can be seamlessly plugged in a CNN to selectively preserve and transform the visual features conditioned on language semantics. By inserting multiple composite transformers at varying depths, VAL is incentive to encapsulate the multi-granular visiolinguistic information, thus yielding an expressive representation for effective image search. We conduct comprehensive evaluation on three datasets: Fashion200k, Shoes and FashionIQ. Extensive experiments show our model exceeds existing approaches on all datasets, demonstrating consistent superiority in coping with various text feedbacks, including attribute-like and natural language descriptions.

1. Introduction

Image search is a fundamental task in computer vision. It has been serving as the cornerstone in a wide range of application domains, such as internet search [42], fashion retrieval [34], face recognition [57] and product identification [44]. The most prevalent paradigms in image search take either image or text as the input query to search for items of interest, commonly known as image-to-image [15] and text-to-image matching [12]. However, an intrinsic downside of these paradigms lies in the infeasibility to refine the retrieved items tailored to users’ intentions, especially when users cannot precisely describe their intentions by a single image or with all the keywords.

To overcome the aforementioned limitation, different user interactive signals have been explored over the past two decades [61]. The basic idea is to incorporate user feedback to refine or discover image items retrieved by the system [52, 81, 70, 11, 45, 28, 27, 18, 79, 2, 39, 16, 48, 75, 17]. Most of these interactions are delivered in the form of text, describing certain attributes [18, 79, 2] or relative attributes [45, 28, 75] to refine or modify upon a reference image. More recently, natural language feedback [17] is introduced as a more flexible way to convey user’s intention for interactive image search. Despite having great potential value in practice, incorporating various types of text feedback for image search still remains understudied.

In this work, we investigate the task of image search with text feedback, which entitles user to interact with the system by selecting a reference image and providing additional text to refine or modify the retrieval results. Unlike the prior works that mostly focus on one type of text feedback, we consider the more general form of text, which can be either attribute-like description, or natural language expression. This poses a more challenging multimodal learning problem that requires the synergistic understanding of both visual and linguistic contents at different granularities – the given text may convey multi-granular semantics, ranging from a concrete attribute to highly abstract visual properties (Fig. 1). As a task lying at the intersection of vision and language, image search with text feedback, however, differs...
greatly from other extensively studied vision-and-language tasks, such as image-text matching [12, 73, 67], image captioning [24, 60], and visual question answering [4, 60]. This is because, it uniquely entails learning a composite representation that can jointly capture visual cues and linguistic information to match the target image of interest.

One intrinsic challenge is the difficulty to simultaneously preserve and transform the visual content in accordance with the given text. For instance, when a text snippet specifies the colour to modify (Fig. 1(a)), it means the other visual cues such as silhouette, pattern, trim should all be preserved in the retrieved items, with only the colour transformed to the desired one. Another challenge is to learn a composite representation that can jointly encapsulate visual and linguistic contents from coarse to fine-grain. Since the text feedback may convey multi-level semantics (Fig. 1), the composite representation is also expected to capture the multi-granular visiolinguistic information. To address these challenges in a unified solution, we propose a novel Visiolinguistic Attention Learning (VAL) framework, which fuses vision and language features via attention learning at varying representation depths.

Briefly, VAL is featured with multiple composite transformers plugged at multi-level inside a CNN to compose visual features and language semantics. Our core idea is to learn the attentional transformation and preservation concurrently, such that the composite features not only preserve the unaltered visual content in image, but also transform certain content as specified by text. To train our VAL, we devise a hierarchical matching objective, which incentivises exclusive alignments to the desired visual and semantic features for discriminative feature learning.

To summarise, our contribution is two-fold:

• We tackle the challenging task of image search with text feedback by a novel Visiolinguistic Attention Learning (VAL) framework. VAL is characterised by multiple composite transformers that compose multi-level visual features and language semantics via attention learning. Through a hierarchical matching objective, VAL is incentive to encapsulate visual and linguistic contents as composite representations for effective image search.

• We set a new state-of-the-art on three datasets: Fashion200k, Shoes, and FashionIQ. Remarkably, VAL performs consistently well in coping with various types of text feedback, demonstrating a greater potential in practical use. We also present an insightful ablation study to analyse the underlying attentions learnt by VAL.

2. Related Work

Interactive image search aims to incorporate user feedback as an interactive signal to navigate the visual search. In general, the user interaction can be given in various formats, including relative attribute [45, 28, 75], attribute [79, 18, 2], attribute-like modification text [66], natural language [16, 17], spatial layout [37], and sketch [76, 74, 14]. As text is the most pervasive interaction between human and computer in contemporary search engines, it naturally serves to convey concrete information that elaborates user’s intricate specification for image search. In this work, we investigate various text feedbacks for image search. Thanks to the rich annotations released recently on several fashion benchmark datasets [18, 17], we present the first attempt to consider richer forms of text feedback in one-turn interactive search, including attribute-like and natural language expression.

Attention mechanism is widely adopted as an important ingredient in various vision-and-language tasks, which aims to mimic human’s capability of attending to salient sensory information [7]. To steer where to fixate in images, spatial attention is commonly used to assign importance weights on image regions. This helps to select informative regions for captioning [65, 3], or locate relevant visual content for question answering [72, 82]. For attention learning in vision and language domains, co-attention [36, 41] is generally adopted to fuse visual and textual contents by generating attention weights on image regions and question words. Recently, several self-attention mechanisms are proposed for VQA [77, 13, 23, 35], which builds upon transformer [64] to learn the inter-modal or intra-modal latent attention. Inspired by this line of works, we propose a generic visiolinguistic attention learning scheme, which learns the attentional interactions upon the visiolinguistic features. Unlike previous works that rely heavily on off-the-shelf Faster R-CNN [51] to extract image region features, our approach avoids the dependency on a pre-trained object detector, and thus generalises well to fine-grained visual search, especially when the imagery data does not share the common objects as those in the object detection datasets.

Composition learning is deemed as an essential functionality to build intelligent machine [29, 30]. The general aim is to learn a feature encoding that encompasses multiple primitives [38, 40, 62, 49, 69]. Although convolutional neural networks (CNNs) inherently learn the composition of visual parts [78, 5, 31], they do not explicitly tie visual representation and language semantics in a compositional way. Recently, several concurrent works [59, 56, 35] extend the pre-training strategies from BERT [9] to learn the latent compositional representations, which jointly represent images and descriptive texts for solving VQA, captioning, or image-text matching. However, these works mostly fix the image representation pre-extracted from a detection [51] or recognition [71] model. This not only limits their applicability to certain imagery domain, but also leads to an overall complex, heavy modelling framework. We propose a remedy by injecting language semantics at varying depths inside a CNN. This effectively yields a more powerful composite representation with simpler, lighter modelling.
3. Visiolinguistic Attention Learning

Fig. 2 presents an overview of our Visiolinguistic Attention Learning (VAL) framework. Given a reference image and user text as input query, the ultimate aim of VAL is to learn a composite representation that aligns exclusively to the target image representation. VAL contains three major components: (a) an image encoder, (b) a text encoder (Sec. 3.1), (c) composite transformers (Sec. 3.2) that are plugged into different convolution layers to compose the vision and language contents. All components are synergistically optimised by (d) hierarchical matching (Sec. 3.3). Symbols of $\odot$, $\otimes$, $+$ stand for the Hadamard product, matrix multiplication and element-wise addition, respectively.

### 3.1. Representing Images and Texts

**Image Representation.** To encapsulate the visual contents into discriminative representations, we employ an image encoder, i.e., a standard CNN, for image representation learning. As CNNs inherently learn visual concepts of increasing abstraction in a compositional, hierarchical order [5, 31, 78], we conjecture that image features from a single convolution layer do not capture the visual information of different granularities. Thus, we extract the feature maps from multiple convolution layers to construct a build-in feature pyramid [33] for more expressive representation learning. Concretely, the feature pyramid $F$ is obtained from three different levels inside the CNN $\theta_{\text{CNN}}$:

- $F_r = \{x^L_r, x^M_r, x^H_r\} = \theta_{\text{CNN}}(I_r)$
- $F_t = \{x^L_t, x^M_t, x^H_t\} = \theta_{\text{CNN}}(I_t)$

Here, $I_r, I_t$ refer to the reference image and target image; $F_r, F_t$ are their corresponding feature pyramids, with each containing multi-level feature maps $x^L, x^M, x^H$ extracted from the Low, Mid, High-level convolution layers$^1$.

**Text Representation.** To represent the semantics of texts, we utilise a text encoder to map the user text $T$ into a vectorised text representation. Formally, the text encoder is implemented as an LSTM, followed by max-pooling and a linear projection layer. In brief, we first apply basic tokenising on text, then feed the token sequence into the text encoder to obtain the final text representation: $t \in \mathbb{R}^\ell$.

### 3.2. Composite Transformer

To jointly represent images and texts, we propose to transform and preserve the visual features conditioned on language semantics. Inspired by the superiority of transformer [64] in multimodal learning [23, 35], we devise a composite transformer plugged at multi-level inside a CNN. Our key idea is to learn a composite representation of image and text through attentional transformation and preservation learnt upon the visiolinguistic features (Fig. 2(c)), with the ultimate aim to capsule the essential visual and linguistic contents for visual search, which we describe next.

**Visiolinguistic Representation.** To digest the information flows from vision and language domains, the reference image feature $F_r$, text feature $t$ are first fused to obtain the visiolinguistic representation. Formally, for feature maps $x^i_t$ (where $i=\text{L, M, H}$ is the level in feature pyramid), multimodal fusion is performed by concatenation with the text feature $t$, followed by a composite function $F_c$ to learn the fused visiolinguistic feature $x^c_{vl}$:

$$x^c_{vl} = F_c([x^i_t, t])$$

---

$^1$Refer to Supplementary Material for more architecture details.
learnt upon the visiolinguistic feature \( x_{vl}^i \) to recalibrate the strength of preservation on \( x_{vl}^i \). This is motivated that different feature maps encode different semantics, e.g. colors, materials, parts [80]. Thus, to selectively suppress and highlight the visual content in \( I_v \), attentional preservation is introduced to selectively reuse the reference image feature \( x_{vl}^i \). Formally, a lightweight joint-attention is learnt upon the visiolinguistic feature \( x_{vl}^i \) in a squeeze-and-excite manner [22] to obtain the selective activation on \( x_{vl}^i \):

\[
A_{sp}^i = \text{sigmoid}(\mathcal{F}_{sp}(\frac{1}{c} \sum_j c x_{vl}(j; :))) \quad A_{ch}^i = \text{sigmoid}(\mathcal{F}_{ch}(\frac{1}{h' \times w'} \sum_j h' \times w' x_{vl}(j, k; :)))
\]

\[
A_{ja}^i = A_{sp}^i \odot A_{ch}^i
\]

where \( A_{sp}^i \in \mathbb{R}^{h' \times w' \times 1} \), \( A_{ch}^i \in \mathbb{R}^{1 \times 1 \times c'} \), \( A_{ja}^i \in \mathbb{R}^{h' \times w' \times c'} \); \( \mathcal{F}_{sp}, \mathcal{F}_{ch} \) are implemented as \( h' \times w' \), \( 1 \times 1 \) convolutions to learn the spatial, channel attentions \( A_{sp}^i, A_{ch}^i \). \( A_{ja}^i \) is the joint-attention matrix derived from \( A_{sp}^i, A_{ch}^i \), which dynamically modulates the intensity to preserve the reference image feature \( x_{vl}^i \):

\[
o_{ja}^i = A_{ja}^i \odot x_{vl}^i
\] (3)

where \( o_{ja}^i \in \mathbb{R}^{h' \times w' \times c'} \). The final output of the composite transformer is the weighted sum of outputs \( o_{sa}^i, o_{ja}^i \) from two complementary attentional streams:

\[
o^i = w_{sa} o_{sa}^i + w_{ja} o_{ja}^i
\] (4)

where \( w_{sa}, w_{ja} \) are learnable scalars to control the relative importance of two streams. The composite output of VAL is denoted as \( F_o = \{ \mathbf{o}^L, \mathbf{o}^M, \mathbf{o}^H \} \) – a feature pyramid with each level derived from one composite transformer. The final composite feature used for image retrieval is simply the concatenation of multi-level outputs after average-pooling.

### 3.3. Hierarchical Matching

As our ultimate aim is to align the composite output \( F_o \) and the target image representation \( F_i \) exclusively, we formulate a hierarchical matching objective, with two losses formed in a two-level hierarchy to match with the desired visual and semantic features (Fig. 3), as detailed next.

**Primary visual-visual matching.** We introduce visual-visual matching as our primary objective to ensure the composite feature match the target feature with high similarity. Formally, with similarity measured by L2 distance \( d \), a bi-directional triplet ranking loss [10] is imposed to align the multi-level feature maps in two feature pyramids \( F_o, F_i \):

\[
\mathcal{L}_{uv} = \sum_i L_i(\mathbf{\hat{o}}^i, \mathbf{\hat{x}}_v^i) + L_i(\mathbf{\hat{x}}_v^i, \mathbf{\hat{o}}^i)
\]

\[
\text{rank} \hat{o} \quad \text{rank} \hat{o} \quad \text{rank} \hat{o}
\]

with \( L_i(\mathbf{\hat{o}}, \mathbf{\hat{x}}_v^i) = \max(0, d(\mathbf{\hat{o}}, \mathbf{\hat{x}}_v^i) - d(\mathbf{\hat{o}}, \mathbf{\hat{x}}_v^i) + m) \)
Among the above methods, TIRG is proposed 

Datasets. To validate the model’s generalisability to various datasets, we evaluate on three datasets, including (1) Fashion200k using attribute-like description, (2) Shoes and FashionIQ using natural language expression. We details these datasets in Sec. 4.2, Sec. 4.3 and Sec. 4.4.

Compared Methods. To validate the efficacy of our approach in image search with text feedback, we compare with four representative multimodal learning methods:

- Relationship [53]: A relation reasoning module. It takes in feature maps extracted from the final layer of a CNN and text feature from an RNN, followed by concatenation and an MLP to learn the cross-modal relationships. The pairwise relationships are simply summed and processed through another MLP to get the final output.

- FiLM [47]: A Feature-wise Linear Modulation component. It contains a stack of three FiLM layers cascaded after a CNN. The text information is represented by the text feature extracted from an RNN to modulate each feature map by affine transformation.

- MRN [25]: A Multimodal Residual Learning component. It learns multimodal representations by fusing visual and textual features from a CNN and an RNN. The cross-modal features are obtained through three blocks of element-wise multiplication and residual learning.

- TIRG [66]: An image-text composition approach for image retrieval. It composes visual and textual features by concatenation, followed by learning a gating connection and a residual connection for cross-modal fusion.

Discussion. Among the above methods, TIRG is proposed for image search with attribute-like text feedback; whilst others are originally used in VQA. However, unlike existing methods that stack transformation layers after a CNN, VAL uniquely plugs the composite transformers at multi-level inside a CNN to capture multi-granular visiolinguistic information. In addition, VAL is specially featured with two attentional streams that operate upon the visiolinguistic features to selectively transform and preserve the visual features conditioned on the language semantics. For a fair comparison, we implement existing methods using the same CNN, RNN trained by a bi-directional ranking loss.

Ablative baselines. Besides comparing with existing methods, we conduct several ablative tests on our model:

- VAL (L_{vv}): VAL optimised with the primary objective (Eq. 5), i.e. auxiliary regulariser (Eq. 6) is not used.

- VAL (L_{vv} + L_{vs}): VAL trained by hierarchical matching, using side information by joint training or pre-training.

- VAL (GloVe): It shares the same structure as VAL (L_{vv} + L_{vs}), with word vectors initialised from GloVe [46]. The latter two tests endow our VAL model with prior linguistic knowledge from side information and GloVe.

Implementation Details. We conduct all the experiments in Tensorflow [1]. We initialise the CNNs pre-trained from ImageNet [8], and integrate the composite transformers into ResNet-50 [19] on Shoes, FashionIQ, and MobileNet [21] on Fashion200k. In the self-attentional stream, we set the number of heads to 2. The LSTM [20] is one-layer with 1024 hidden units, followed by a linear projection layer that maps the max-pooled LSTM feature to the text feature of 512 dimension. We use Adam [26] optimiser with a constant learning rate of $2 \times 10^{-4}$ and $\alpha, \beta$ of 0.999, $1 \times 10^{-8}$. The batch size is set to 32. The margin $m$ in Eq. 5, Eq. 6 is set to 0.2. More network architecture and training details are given in Supplementary Material due to space limit.

Evaluation Metric. We adopt the standard evaluation metric in retrieval, i.e. Recall@K, denoted as R@K for short.
Figure 4. Qualitative results of image search with attribute-like text feedback on Fashion200k. blue/green boxes: reference/target images.

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

Table 1. Quantitative results of image search with text feedback on Fashion200k. Rows in colours indicate results obtained with the same networks and data. Overall 1st/2nd best in red/blue.

4.2. Fashion200k

Fashion200k [18] is a large-scale fashion dataset crawled from multiple online shopping websites. It contains more than 200k fashion images collected for attribute-based product retrieval. It also covers a diverse range of fashion concepts, with a total vocabulary size of 5,590. Each image is tagged with descriptive texts as product description, such as “white logo print t-shirt”, which is exploited as side information for auxiliary supervision via joint training. Following [66], we use the training split of around 172k images for training and the test set of 33,480 test queries for evaluation. During training, pairwise images with attribute-like modification texts are generated by comparing their product descriptions (see Supplementary Material).

Table 1 shows our comparison with existing methods. We reproduce the best competitors with the same networks and optimiser for a like-to-like fair comparison. As can be seen, our model demonstrates compelling results compared to all other alternatives, e.g. VAL ($L_{vv}$) outperforms the best competitor TIRG with an improved margin of 6.4% in R@1. We also observe that (1) VAL ($L_{vv}+L_{vs}$) performs better than VAL ($L_{vv}$), which indicates the advantage of introducing auxiliary supervision to match with additional semantics; (2) VAL (GloVe) performs on par with VAL ($L_{vv}+L_{vs}$), suggesting using GloVe word vectors is not so vital when using attribute-like text feedback.

Fig. 4 shows our qualitative results on Fashion200k. We notice our model is able to retrieve new images that resemble the reference image, while changing certain attributes conditioned on text feedback, e.g. colour, material and trim.

4.3. Shoes

Shoes [6] is a dataset originally crawled from like.com. It is further tagged with relative captions in natural language for dialog-based interactive retrieval [16]. Following [16], we use 10,000 training samples for training and 4,658 test samples for evaluation. Besides relative captions, there are 3,000 images tagged with descriptive texts, such as “brown buckle mules”, which are used as auxiliary supervision (Eq. 6) for pre-training in VAL ($L_{vv}+L_{vs}$). Due to missing results of state-of-the-art methods in composing image and text for image search, we provide a new benchmark on this dataset by performing experiments under the same networks and optimiser for a comprehensive comparison.

Table 4 shows the clear superiority of our model compared to other alternatives. For instance, VAL ($L_{vv}$) surpasses the best competitor TIRG by 3.89% in R@1. We also notice the clear advantages of utilising prior linguistic knowledge in VAL ($L_{vv}+L_{vs}$) and VAL (GloVe), as compared to not using such knowledge in VAL ($L_{vv}$).

Fig. 5 further shows our qualitative results on Shoes. It suggests our model is capable of ingesting multiple visual attributes and properties in the natural language text feedback to search for the desired target images. More qualitative results are given in Supplementary Material.
are red with a woven top pattern
have no buckle or wedge heel
have fur on the outside

Figure 5. Qualitative results of image search with natural language text feedback on Shoes. blue/green boxes: reference/target images.

<table>
<thead>
<tr>
<th>Method</th>
<th>Dress</th>
<th>Shirt</th>
<th>Toptee</th>
<th>Avg</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R@10</td>
<td>R@50</td>
<td>R@10</td>
<td>R@50</td>
</tr>
<tr>
<td>TIRG</td>
<td>8.10</td>
<td>23.27</td>
<td>11.06</td>
<td>28.08</td>
</tr>
<tr>
<td>Image+Text Concatenation</td>
<td>10.52</td>
<td>28.98</td>
<td>13.44</td>
<td>34.60</td>
</tr>
<tr>
<td>Side Information [17]</td>
<td>11.24</td>
<td>32.39</td>
<td>13.73</td>
<td>37.03</td>
</tr>
<tr>
<td>MRN</td>
<td>12.32</td>
<td>32.18</td>
<td>15.88</td>
<td>34.33</td>
</tr>
<tr>
<td>FiLM</td>
<td>14.23</td>
<td>33.34</td>
<td>15.04</td>
<td>34.09</td>
</tr>
<tr>
<td>TIRG</td>
<td>14.87</td>
<td>34.66</td>
<td>18.26</td>
<td>37.89</td>
</tr>
<tr>
<td>Relationship</td>
<td>15.44</td>
<td>38.08</td>
<td>18.33</td>
<td>38.63</td>
</tr>
<tr>
<td>VAL (Lvv)</td>
<td>21.12</td>
<td>42.19</td>
<td>21.03</td>
<td>43.44</td>
</tr>
<tr>
<td>VAL (Lvv + Lvs)</td>
<td>21.47</td>
<td>43.83</td>
<td>21.03</td>
<td>42.75</td>
</tr>
<tr>
<td>VAL (GloVe)</td>
<td>22.53</td>
<td>44.00</td>
<td>22.38</td>
<td>44.15</td>
</tr>
</tbody>
</table>

Table 3. Quantitative results of image search with text feedback on FashionIQ. Avg: averaged R@10/50 computed over three categories. Rows in colour indicate results obtained with the same backbone networks (i.e. CNN, LSTM) and data. Overall 1st/2nd best in red/blue.

leopard print with deeper neck
is purple in color with playing cards graphic
is yellow with fringe

4.4. FashionIQ

FashionIQ [17] is a natural language based interactive fashion product retrieval dataset. It contains 77,684 images crawled from Amazon.com, covering three categories: Dresses, Tops&Tees and Shirts. Among the 46,609 training images, there are 18,000 image pairs, with each pair accompanied with around two natural language sentences that describe one or multiple visual properties to modify in the reference image, such as “is darker” and “has short sleeves and is longer and more flowing”. We use the side information from Fashion200k as auxiliary supervision for pre-training in VAL (Lvv + Lvs). Following the same evaluation protocol of composing image and text for retrieval [17], we use the same training split and evaluate on the validation set³. We report results on individual category, as well as the averaged results over three categories⁴.

Table 3 shows our model outperforms other competitors substantially, e.g. VAL (Lvv) surpasses Relationship with an overall margin of 4.31% in R@10. We also notice the performance boosts in VAL (Lvv + Lvs) and VAL (GloVe), as compared to VAL (Lvv). This again indicates the benefit of using prior linguistic knowledge from auxiliary semantics and GloVe when using natural language text feedback.

Fig. 6 presents our qualitative results on FashionIQ. It shows that given multiple semantic concepts within a sentence snippet, our model captures both concrete and abstract semantics, including various fashion elements [63].

³The groundtruth of test set in FashionIQ has not been released yet.
⁴The unpublished state-of-the-art uses an ensemble of diverse models.
like colour, silhouette, printing, etc. We also observe that our model can jointly comprehend the global appearance (e.g. overall colours, patterns), as well as local fine-grained details (e.g. a specific logo and trim) for image search.

### 4.5. Ablation Study

In this section, we conduct analysis to give an insight of the key ingredient in VAL (i.e. composite transformers). We perform experiments with the primary objective (Eq. 5) to exclude the effect of auxiliary regulariser.

**Effect of self-attention and joint-attention.** To analyse the synergistic effect of self-attentional transformation (SA) and joint-attention preservation (JA), we compare our composite transformer with two baselines: (a) remove SA stream (i.e. “w/o SA”); (b) remove JA stream (i.e. “w/o JA”) – see a graphical illustration in Supplementary Material. For each baseline, we remove one attentional stream to study its effect. Table 4 shows the comparison on FashionIQ and Shoes. It can be seen that our VAL does profit substantially from the complementary benefits of SA and JA. This verifies our rationale of composing visual features and language semantics through attentional transformation and preservation learnt upon the visiolinguistic features.

**Attention visualisation.** To further interpret the attentions learnt by VAL at varying representation depths (i.e. low, mid, high level), we visualise the attended regions by joint-attention and self-attention in Fig. 7. From Fig. 7(b), we notice that the spatially attended region varies across different levels. This indicates the joint-attention stream picks up different visual cues to preserve across varying depths. From Fig. 7(c), we observe that the multi-level self-attention triggers various attended regions for learning the transformation, e.g. in the dress example, the low-level self-attention highlights the overall silhouette, while the mid, high-level self-attentions pick up the thigh area to focus on.

Overall, Fig. 7 shows our model captures visual cues at different granularities to selectively preserve and transform the reference image features according to language semantics. This suggests that VAL learns to capture the essential multi-granular visiolinguistic contents for image search.

**Effect of composition at multi-level.** We test how composition at multi-level aids in representation learning by comparing VAL (high+mid+low) to two baselines: (a) high, (b) high+mid, which perform composition at high or high+mid level. Fig. 8 shows composition at multi-level improves the overall performance. This verifies the efficacy of employing composite transformers at varying depths to capture the multi-granular information, which also accords with the fact that CNNs learn visual features of increasing abstraction from lower to higher layers [58]. While focusing on multi-modal representation learning, our model can also be integrated with a dialogue manager [16] for interactive search.

### 5. Conclusion

We introduced VAL, a novel approach to tackle the challenging task of image search with text feedback. VAL is featured with multiple composite transformers that selectively preserve and transform multi-level visual features conditioned on semantics to derive an expressive composite representation. We validate the efficacy of VAL on three datasets, and demonstrate its consistent superiority in handling various text feedbacks, including attribute-like description and natural language expression. We also explore auxiliary semantics to further boost the model performance. Overall, this work provides a novel approach along with a comprehensive evaluation, which collectively advance the research in interactive visual search using text feedback.

**Acknowledgement:** We would like to thank Maksim Lapin, Michael Donoser, Bojan Pepik, and Sabine Sternig for their helpful discussions.
References


[38] Ishan Misra, Abhinav Gupta, and Martial Hebert. From red wine to red tomato: Composition with context. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 2


[66] 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, 5, 6


[75] Aron Yu and Kristen Grauman. Thinking outside the pool: Active training image creation for relative attributes. In IEEE Conference on Computer Vision and Pattern Recognition, 2019. 1, 2

[76] Qian Yu, Feng Liu, Yi-Zhe Song, Tao Xiang, Timothy M Hospedales, and Chen-Change Loy. Sketch me that shoe. In IEEE Conference on Computer Vision and Pattern Recognition, 2016. 2


[82] Chen Zhu, Yanpeng Zhao, Shuaiyi Huang, Kewei Tu, and Yi Ma. Structured attentions for visual question answering. In IEEE Conference on Computer Vision and Pattern Recognition, 2017. 2
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

<table>
<thead>
<tr>
<th>Part Name</th>
<th>Layer Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>composite function</td>
<td>( \mathcal{F}_c: \text{conv}(K-1 \times 1, N-c^i, \text{ReLU}) )</td>
</tr>
</tbody>
</table>
| SA stream       | \( \mathcal{F}_Q, \mathcal{F}_K, \mathcal{F}_V: \text{conv}(K-1 \times 1, N-c^i) \)  
|                 | \( \mathcal{F}_{sa}: \text{conv}(K-1 \times 1, N-c^i, \text{ReLU}) \)             |
| JA stream       | \( \mathcal{F}_{sp}: \text{conv}(K-h^i \times w^i, N-1, \text{sigmoid}) \)     
|                 | \( \mathcal{F}_{ch}: \text{conv}(K-1 \times 1, N-c^i, \text{sigmoid}) \)          |

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

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, 13th 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 \( 8.10 \times 10^7 \) vs. ResNet50 \( 3.80 \times 10^9 \), which bring small computational cost - an additional FLOPs of 2.13%. On (2), the complexity is \( O(QN), O(QN\log N) \) 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 \( A_{sa}^i \)), and (b) w/o joint-attention (w/o \( A_{ja}^i \)). We show a graphical illustration of these two baselines in Fig. 1.

2. Dataset and Training Details

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.

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 \( (L_{uv} + L_{vs}) \), we exploit the tagged
descriptions as side information, which serve as auxiliary supervision to train our VAL via a joint-training objective as $L_{vv} + L_{vs}$. We train for 160k iterations on Fashion200k. 

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

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