Interpretable Multimodal Retrieval for Fashion Products

Lizi Liao¹, Xiangnan He¹, Bo Zhao², Chong-Wah Ngo³, Tat-Seng Chua¹
¹National University of Singapore, ²University of British Columbia, ³City University of Hong Kong
{liaolizi.llz, xiangnanhe, zhaobo.cs}@gmail.com, cscwngo@cityu.edu.hk, chuats@comp.nus.edu.sg

ABSTRACT

Deep learning methods have been successfully applied to fashion retrieval. However, the latent meaning of learned feature vectors hinders the explanation of retrieval results and integration of user feedback. Fortunately, there are many online shopping websites organizing fashion items into hierarchical structures based on product taxonomy and domain knowledge. Such structures help to reveal how human perceive the relatedness among fashion products. Nevertheless, incorporating structural knowledge for deep learning remains a challenging problem. This paper presents techniques for organizing and utilizing the fashion hierarchies in deep learning to facilitate the reasoning of search results and user intent.

The novelty of our work originates from the development of an EI (Exclusive & Independent) tree that can cooperate with deep models for end-to-end multimodal learning. EI tree organizes the fashion concepts into multiple semantic levels and augments the tree structure with as well as independent constraints. It describes the different relationships among sibling concepts and guides the end-to-end learning of multi-level fashion semantics. From EI tree, we learn an explicit hierarchical similarity function to characterize the semantic similarities among fashion products. It facilitates the interpretable retrieval scheme that can integrate the concept-level feedback. Experiment results on two large fashion datasets show that the proposed approach can characterize the semantic similarities among fashion items accurately and capture user’s search intent precisely, leading to more accurate search results as compared to the state-of-the-art methods.

KEYWORDS

Multimodal fashion retrieval, EI tree, attribute manipulation

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

ACM Reference Format:

1 INTRODUCTION

As evidenced by Black Friday’s record-high of $5.03 billion online sales in U.S. and Alibaba’s $25 billion Singles Day sales in 2017, the modern e-commerce traffic volume is growing fast. At the same time, consumers have become very exigent. For instance, they may have in mind a specific fashion item in a particular color or style, and want to find it online without much effort [28]. Therefore, making the retrieval procedure explainable as in Figure 1 and being able to leverage user feedback become essential requirements.

Fashion search by text has been widely used (e.g. search engines, shopping apps) to fulfill such requirements [32], owing to its natural way of expression and flexibility of description [45]. However, such freedom also leads to rather diverse textual descriptions of fashion items, making the retrieval results unsatisfactory. More importantly, there are many visual traits of fashion items that are not easily translated into words. Meanwhile, with the growing volume of online images, Content Based Image Retrieval (CBIR) [49] comes into play and allows users to simply upload a query image. Items are then retrieved based on their visual similarities to the query. A major challenge to such methods is the well-known semantic gap between the low-level visual cues and the high-level semantic features (e.g., neckline, sleeve length) that interpret users’ search intent. Therefore, considering the strengths and weaknesses of both methods, it is natural to combine the textual and image modalities. Indeed, many efforts linking image and text have shown promising results and can be applied to fashion retrieval, such as the visual-semantic embedding [22] and multimodal correlation learning [4]. Typically, such models take in image-text pairs and optimize a similarity based distance based loss function (e.g., CCA loss, contrastive ranking loss) to discover a shared feature space [26]. However, the learned feature vectors are usually opaque, making it difficult to explain the retrieved results, incorporate user feedback and further improve the search performance. Thus, a major research question is: can we develop a solution that takes advantage of multi-modalities and is able to perform interpretable fashion retrieval?

Fortunately, the abundant resources of taxonomies for fashion items in online shops and the domain specific knowledge shed...
some light on this question. Many general e-commerce sites (e.g. Amazon and Taobao) as well as fashion specific sites (e.g. esos.com and polyvore.com) have similar ways of organizing fashion products. These organization schemes describe the taxonomy of fashion items and give clues to human perception of fashion product similarity. For example, as shown in Figure 1, upper-garments associate with concepts like neckline and sleeve length, which are absent in bottom-garments such as skirts or pants. Such structure encodes fashion knowledge by imposing certain concepts conditioning on others. As a result, learning a tree structure of concept dependency has potential to achieve better performance [20, 31, 50]. More importantly, there exists many exclusive and independent relationships among these fashion concepts that can be leveraged to boost performance [43]. For instance, a single item may only belong to one of the category concepts such as coat and shirt, which are mutually exclusive. Meanwhile, concepts like sleeve length and neckline seem to be independent of each other. Therefore, a tree structure augmented with such constraints becomes a viable way to integrate human perception to the modeling process. Note that it differs from the And-Or graph [5] which models logical AND or OR relationships between siblings. EI tree models the exclusive (choose only one) and independent (choose freely) relationships and guide the end-to-end learning of multi-level fashion semantics.

Figure 2 presents an overview of the proposed framework, which is composed of two parts. In the offline part, we first map the clothing images and text descriptions into a joint visual semantic embedding space. We then apply the EI tree to guide the learning procedure and obtain meaningful representations where each dimension corresponds to a concrete fashion concept. Meanwhile, the EI loss is propagated back through the network to update feature learning. After the end-to-end training, we leverage the learned EI tree weights to localize fashion concepts, which provides a straightforward way to visualize the validity of EI tree. In the online part, a given query image or text description is first processed by the EI model to generate its vector representation. Similar items are then retrieved from the collection according to their similarities to the query. Supported by the learned representation and explicit hierarchical similarity function, we enable a direct channel to help user express search intent through providing feedback on fashion concepts. It offers a clearer semantic description of search intent. For example, by viewing the searched results, users can specify that they prefer ‘short sleeve, rather than long sleeve’. Based on the feedback, the model can manipulate the query representation by assigning a 1 to the feature dimension corresponding to short sleeve while setting that of long sleeve to 0.

The main contributions of this paper are as follows:

• We propose an EI Tree to guide the end-to-end deep learning. It bridges the gap between opaque deep features and meaningful fashion concepts.
• We learn an explicit hierarchical similarity function to accurately characterize the semantic affinities among fashion items. A direct feedback mechanism is then proposed to collect user feedback and capture the search intent precisely.
• We design an interpretable multimodal fashion retrieval scheme based on EI tree and demonstrate its effectiveness in facilitating explainability with superior search performance over the state-of-the-art approaches.

2 RELATED WORK

2.1 Fashion retrieval
Interest in fashion retrieval has increased recently. While text retrieval looks for repetitions of query words in text descriptions or product titles, newer latent semantic models [2, 34] use more powerful distributed representations [11]. On the other hand, deep convolutional networks have been used to learn visual representations and achieved superior performance [24] in image classification. However, the generated features are largely uninterpretable.

As mid-level representations that describe semantic properties, semantic attributes or concepts [27] have been applied effectively to object categorization [39] and fine-grained recognition [25]. Inspired by these results, researchers in fashion domain have annotated clothes with semantic attributes [3, 19, 30] (e.g., material, pattern) as intermediate representations or supervisory signals to bridge the semantic gap. For instance, [3] automatically generated a list of nameable attributes for clothing on human body. [40] learned visually relevant semantic subspaces using a multiquery triplet network. [30] proposed FashionNet to jointly predict attributes and landmarks of the clothing image. As another direction, attributes conditioned on object parts have achieved good performance in fine-grained recognition [29, 46, 51]. However, these methods are limited by their ability to accurately parse the human body in images. In contrast, we propose to integrate domain knowledge for more effective learning of fashion attributes.

2.2 Attribute Manipulation
Regarding fashion query formulation and manipulation, WhittleSearch [23] allows users to upload a query image with text descriptions. However, only the relative attributes were considered. More recently, Generative Visual Manipulation (GVM) model [55] was proposed to directly edit image and generate new query image using GAN [12] for search. Generally, the retrieval results relied highly on the quality of a generated image. More importantly, GVM is limited in depicting certain concepts, such as style or pattern. Instead of editing the image, AMNet [52] resorts to communicate additional concept description to the search engine. Memory network was leveraged to manipulate image representation at the concept level. However, the quality of extracted prototype concept representations largely affected the manipulation results. Also, the relationships between concepts were largely ignored, which has been demonstrated to be important to model [33, 48]. In our work, as the semantics and relationships are captured by EI tree, we can explicitly modify the corresponding dimension in the learned concept vector to encode user feedback on attributes.

2.3 Semantic Hierarchy
Taxonomy or ontology based semantic hierarchies such as WordNet [9], ImageNet [7], and LSCOM [35] have been successfully applied for knowledge inferencing. Particularly, the organization of semantic concepts from general to specific provides reasoning capability to boost recognition and retrieval [6, 8, 10]. Most recent works exploiting semantic hierarchy focused on designing new similarity metrics that embeds semantics and hierarchical information. For example, [8] proposed to find visually nearest neighbors for two images and then compute their semantic distance based on the concepts of their neighbors. Meanwhile, [6] developed a hierarchical bilinear similarity function directly and achieved the state-of-the-art
performance of image retrieval on ImageNet. Toward the direction of refinement, [41] proposed to associate separated visual similarity metrics for every concept in a hierarchy, and [49] augmented semantic hierarchy with a pool of attributes. Different from these existing works, our work builds fashion domain specific hierarchy – EI tree, which not only helps to guide the end-to-end learning of multi-level fashion semantics, but also modifies the similarity metric as the proximity of two surrogate-EI trees.

3 THE PROPOSED FRAMEWORK

The proposed framework as shown in Figure 2 consists of an offline model training part and an online retrieval part. In offline, EI tree helps to bridge the gap between opaque deep features and interpretable fashion concepts. It guides the model to obtain meaningful representations. Differing from the implicit feature vectors learned by existing deep models, our representations have the following traits: 1) each dimension corresponds to a concrete fashion concept, which enables the interpretability of search queries and results; 2) the representation can be recovered to a surrogate-EI tree where concept relations are captured; and 3) the spatial regions for each concept can be identified via the learned EI weights. In online retrieval, an explicit hierarchical similarity function is learned to compute the semantic similarities among fashion items. Based on it, an interpretable multimodal fashion retrieval scheme is proposed to facilitate concept-level user feedback. This section describes the major components of the offline learning part.

3.1 EI Tree

Deep models have shown superior performance in extracting features for various applications. However, the opacity of these feature vectors hinders the explainability of results. To map the implicit features to interpretable concepts, we may apply traditional multi-class or multi-label classification techniques. Suppose we have concepts $C = \{\text{up\_cloth}, \text{bottom\_cloth}, \text{color}, \text{rise}, \text{fry}\}$, multi-class classification associates an item with a single concept $c \in C$ such as $\text{color}$. As only one concept label can be generated, it works like assuming exclusive relations among all concepts. In contrast, multi-label classification associates a finite set of labels $C' \subset C$ such as $\{\text{up\_cloth}, \text{bottom\_cloth}, \text{color}\}$. Each concept corresponds to a binary classifier, which is similar to assuming independent relations among concepts. In fashion domain, neither one of these interpretation is complete. For example, an upper body cloth may belong only to the $\text{up\_cloth}$ category but not the $\text{bottom\_cloth}$. It thus does not have details such as $\text{rise}$ or $\text{fry}$. Also, details such as $\text{sleeve length}$ and $\text{color}$ are independent of each other.
can refer to a product category (e.g. coat) or an attribute (e.g. color).

Generally speaking, EI tree organizes semantic concepts from general to specific, where exclusive and independent relationships are integrated among siblings. In general, sibling concepts involving product categories usually share exclusive relationships, while sibling concepts involving attributes are often characterized by independent relationships. To generate an EI tree for fashion concepts, we crawled product hierarchies from 40 e-commerce sites such as amazon.com, asos.com and polyvore.com. We next applied the Bayesian Decision approach developed in [38] to obtain an unified hierarchy, and then extracted exclusive and independent relationships manually by a fashion expert. Finally, we obtained an EI tree with 334 concept nodes (excluding the root) organized into six levels. Figure 3 shows part of the resulting fashion EI tree with top level concepts such as up, bottom, color etc.

In next subsections, we will elaborate the application of EI tree in end-to-end learning and concept localization after introducing the image and text pipelines.

### 3.2 Representation Learning with EI Tree

#### 3.2.1 Image & Text Pipelines

Following numerous prior works showing the effectiveness of CNN in extracting image features, we use ResNet-50 [14] pre-trained on ImageNet as the base networks before conducting fine-tuning for visual feature learning. Given an input image $I$ of size $224 \times 224$, a forward pass of a base network produces a feature vector $f_I \in \mathbb{R}^{2048}$. The forward pass process of ResNet-50 is a non-linear function which we denote as $F_{resnet}()$.

As shown in Figure 2, the pipeline takes an anchor image $I_a$ and a negative image $I_n$ as input, and generates the feature vectors for the two images as:

$$f_{I_a} = F_{resnet}(I_a), \quad f_{I_n} = F_{resnet}(I_n).$$

To establish the inter-modal relationships, we represent the words in text descriptions in the same embedding space that the images occupy. The simplest approach might be to project every word individually into this embedding space. However, it does not consider any ordering and word context information. A possible extension might be to integrate dependency tree relations among these words. However, it requires the use of Dependency Tree Parsers trained on text corpora unrelated to fashion domain.

Encouraged by the good performance in [17, 54], we use Bidirectional Long Short-Term Memory units (BLSTM) to compute the text representations. The BLSTM takes a sequence of $T$ words $S = \{x_1, x_2, \cdots, x_T\}$ and transforms the sequence into $\mathbb{R}^{2048}$ vector space. Using the index $t = 1, \cdots, T$ to denote the position of a word in a sentence, the hidden state of basic LSTM unit is calculated by:

$$\overrightarrow{h_t} = LST\text{M}(W_{emb}x_t, \overrightarrow{h}_{t-1}).$$

where $x_t$ is the 1-of-V representation of word $x_t$, $W_{emb}$ is the word embedding matrix initialized with word2vec [34] weights learned from text descriptions and set to be trainable for later training stage. BLSTM consists of two independent streams of LSTM processing, one moving from left to right $\overrightarrow{h_t}$ and the other from right to left $\overleftarrow{h_t}$. We use element-wise sum to combine the two direction outputs $h_t = \overrightarrow{h_t} + \overleftarrow{h_t}$. The final text representation $f_t$ is generated via max-pooling over $[h_t | t = 1 \cdots T]$. We denote the BLSTM forward pass process as a non-linear function $F_{blstm}()$. The dual path text pipeline takes an anchor text $S_a$ and a negative text $S_n$ as input, and generates feature vectors for them accordingly:

$$f_{S_a} = F_{blstm}(S_a), \quad f_{S_n} = F_{blstm}(S_n).$$

As evidenced by the superior performance in linking textual and image modalities [18, 36, 42], we adopt the bi-directional ranking loss as a regularizer to integrate the two modalities for boosting multimodal fashion retrieval. By denoting the cosine similarity measure as $\cos(\cdot, \cdot)$, the bi-directional ranking loss is expressed as:

$$L_{rank} = \frac{1}{N} \sum_{a=1}^{N} \left( \max\{0, m - (\cos(f_{I_a}, f_{S_a}) - \cos(f_{I_a}, f_{S_n}))\} \\
+ \max\{0, m - (\cos(f_{S_n}, f_{I_a}) - \cos(f_{S_n}, f_{I_a}))\} \right),$$

where $m$ is a margin and $N$ is the number of training instances.

As we have mapped image and text features into the same vector space and have assumed similarity relations between them, we naturally sum up the feature representations for the item as:

$$f = f_{I_a} + f_{S_a},$$

which also showed better performance in preliminary experiments.

#### 3.2.2 Interpretation of EI Tree

During the end-to-end model training procedure, EI tree is used to map the implicit deep features $f$ to explicit fashion concept probability vector $p$. Each concept is traced from the root to itself along the EI tree and a probability is generated based on the tracing path, which mimics the general to specific recognition procedure, e.g., high level concepts such as bottom_cloth will have larger probability than lower ones such as trouser_fry. A high probability on bottom_cloth indicates low probabilities of its exclusive siblings such as up_cloth.

![Figure 4: An illustration of the EI tree converter calculation.](image)
where \( c_n \) denotes the one hot vector for node \( c_n \), \( \sigma(\cdot) \) denotes the sigmoid function.

For example, for the EI tree in Figure 4, the probability of \( \text{neckline} \):
\[
p(\text{neckline}|\text{root} \rightarrow \text{neckline}, f, W_{EI}) = \frac{\exp(f^T W_{EI} c_{\text{up}})}{\exp(f^T W_{EI} c_{\text{up}}) + \exp(f^T W_{EI} c_{\text{bottom}})} \cdot \sigma(f^T W_{EI} c_{\text{neckline}}).
\]

The process is intuitive: a softmax constraint is put among the \( \text{up}_\text{cloth} \) and \( \text{bottom}_\text{cloth} \) category, forcing the model to choose only one of them; the independent siblings material and color do not affect this choice. Also, after choosing the \( \text{up}_\text{cloth} \) category, the \( \text{neckline} \) and \( \text{sleeve} \) are decided independently.

To fulfill the whole training procedure, we define a loss function \( L_{EI} \) for the EI converter. Suppose the ‘true’ label vector of concepts is \( y \), the EI loss resumes the cross-entropy loss for \( N \) samples as:
\[
L_{EI} = -\frac{1}{N} \sum_{a=1}^{N} [y_a \log(p_a) + (1 - y_a) \log(1 - p_a)]
\] (4)

To sum up, we integrate the EI tree enhanced cross-entropy loss and bi-directional ranking loss together via a weighted combination. The bi-directional ranking loss is cast as a regularizer:
\[
L = L_{EI} + \lambda L_{rank}
\] (5)

where \( \lambda \) is the weight to adjust the proportion of regularization. We optimize Equation 5 using the Adaptive Moment Estimation (Adam) [21], which adapts the learning rate for each parameter by taking into account both the past and recent gradients.

3.3 Concept Localization with EI Tree

The learned concept weight matrix \( W_{EI} \) can be used to localize concept regions in fashion images, which enables explainable concept predictions. Each column of \( W_{EI} \) is a weight vector for the corresponding concept. Similar to [13, 53], we use upsampled concept activation map to localize the image regions that are most relevant to the particular concept, which provides a direct way to validate the multi-level semantics in EI tree. More detailed results will be shown in Subsection 5.4.

4 INTERPRETABLE FASHION RETRIEVAL

We now describe the details of our proposed fashion retrieval scheme. An explicit hierarchical similarity function is learned to characterize fashion item proximity. Concept manipulation is also supported, which facilitates interactive retrieval.

4.1 Explicit Similarity Measure

Through the end-to-end neural network introduced above, we can generate a representation of a fashion item as \( v = [p, f] \), where \( p \) refers to the concept probability vector learned in subsection 3.2.2, \( f \) refers to the embedding vector as in Equation 3, and \([\cdot]\) denotes vector concatenation. Specifically, since the EI tree structure is predefined, a hierarchical tree representation of \( p \) can be recovered—each concept corresponds to a node in the tree. Therefore, we formulate an explicit hierarchical similarity function to characterize fashion item proximity by aggregating their local proximity among the fashion concepts.

Formally, we define the explicit distance between two items as:
\[
d(v_i, v_j) = \sqrt{(v_i - v_j)^T D (v_i - v_j)},
\] (6)

where \( D \) is a positive semi-definite diagonal matrix. In general, we need to ensure that the items with similar styles to be close and items with rather different fashion concepts to be separated with a large margin. Thus, for each item \( i \), we require its distance to its K-nearest neighbors to be small, and the distance should be smaller than that between \( i \) and any other item \( l \) which is rather different to \( i \). We denote \( i \sim j \) as such a neighborhood. We can have a set of training triplets as \( T = \{ (i,j,l) : i \sim j, i \sim l \} \). Therefore, the metric learning objective can be formulated as follows:

\[
\begin{align*}
\text{minimize} & \quad \frac{1}{|T|} \sum_{i,j,l \in T} d^2(v_i, v_j) + \mu \sum_{(i,j,l) \in T} \xi_{ijl} \\
\text{subject to} & \quad \forall (i,j,l) \in T, \quad \xi_{ijl} \geq 0, \\
& \quad d_{aa} \geq 0 \quad \text{and} \quad D_{ab} = 0 \quad \text{if} \quad a \neq b,
\end{align*}
\] (7)

where \( \mu > 0 \) is a regularization constant. The problem can be solved rather efficiently by employing the LMNN solver [44] modified with the above sampled triplets \( T \) and a diagonal matrix requirement.

4.2 Integration of User Feedback

Based on the meaningful representation \( p \) and explicit similarity function as in Equation 6, user feedback at concept-level can be easily incorporated to achieve interpretable fashion retrieval. In particular, we allow a user to give “yes”/“no” feedback on fashion concepts to state which concepts are in or not in her search intent. Suppose we are at the \( t \)-th feedback iteration. The system records the “yes” concepts as \( R_+ \) and the “no” concepts as \( R_- \). Therefore, the item representation \( p \) can be updated as:
\[
\forall c \in C, \quad p_{t+1}[c] = \begin{cases} 
1 & c \in R_+, \\
0 & c \in R_-, \\
\hat{p}_t[c] & \text{others}.
\end{cases}
\] (8)

The \( p_{t+1} \) is then integrated into \( v_{t+1} \) to form a new query. The corresponding dimensions in \( D \) for concepts in \( R_+ \) and their parent nodes can be increased to emphasize the user intent.

5 EXPERIMENTS

In this section, we systematically evaluate the proposed method, termed as EITree, in multimodal fashion retrieval. The experiments are carried out to answer the research questions as follows. RQ1: Can the proposed EI tree structure help deep models to learn interpretable representations and achieve explainable results? RQ2: Does the EITree method improve the retrieval performance? What are the key reasons behind? RQ3: Does the EITree method manage to integrate user feedback to accurately infer search intent and further boost search performance?

5.1 Experimental Setup

5.1.1 Datasets. Although there exist several clothing datasets [3, 19, 29, 30, 46], the majority of these datasets only contain a limited number of images or lack attribute concept annotations. In this work, we initially crawled 200 clothing categories from Amazon, resulting in 1.66 million instances. After filtering based
on the quality of text and product images, each instance now has meaningful textual information, visual image and product category path. We then sent all the images to a commercial tagging tool and only kept those instances where all tagging scores are above the average of each tag. Through further manual correction and selective validation, we obtained the AMAZON dataset with 489K instances and over 95% validation accuracy. Similarly, for the DARN dataset [16] which has no product category path (thus no results for the prodTree method in Figure 5 and 6) and the texts of which are product description frames from Taobao, we re-processed it by tagging the images using the tool. Finally, we obtained the DARN dataset with 100K instances and over 93% validation accuracy.

5.1.2 Comparing Methods. We compare with the following four representative solutions, including two popular image based methods and two cross-modal approaches. a) Vebay [47] performs end-to-end visual search in eBay. The product categories are separated from other attributes during the training procedure. b) AMNet [52] retrieves image and manipulates image representation at the attribute level. c) DSP [42] learns joint embeddings of images and texts using a two-branch neural network for image-to-text and text-to-image retrieval. d) prodTree is a variant of our framework by replacing EI tree with a product tree (constructed by product category path). Different from EI tree, the product tree encodes only exclusive relationship of concepts. Thus, it organizes cloth category concepts into a tree and other concepts in a flat organization. It serves to verify the contribution of EI tree when the same deep model and learning procedure are employed. To analyze the effect of information modalities, we also compared with another two variations of our method: e) txtEI which only uses text descriptions and f) imgEI which only uses the product images.

5.1.3 Training Setups. For product images, we trained a MultiBox model [37] to detect and crop clothing items. For text descriptions of products, we pre-processed all the sentences with WordNet’s lemmatizer [1] and removed stop words. We then applied word2vec [34] on text descriptions to learn the embedding weights for each word. Regarding the base network for visual modality, we chose ResNet-50 with pretrained weights on ImageNet for our method as well as comparing methods. For the proposed EI tree methods, we set the margin $m = 1.0$ in Equation 2, and the weight $\lambda = 0.01$ in Equation 5. The learning rate of Adam optimizer was initialized to 0.001. The batch size was set to 16.

5.1.4 Evaluation Protocols. In the fashion concept prediction task, we grouped fashion concepts into several groups following [30] and evaluated the performance within each group. We employed the top-1 accuracy score as our metric. It was obtained by ranking the concept classification scores within each group and determined how many concepts have been predicted accurately. To further evaluate the performance of each concept, we treated the prediction of each concept as binary classification. Finally, we adopted AP (Area under precision and recall curve) for evaluation.

Regarding the automatic fashion retrieval task, building a representative query sets with corresponding ground truth is essential for evaluation. We first divided the items with similar ground truth to ensure that items within the same group are in the same style. For ease of manual correction, we randomly sampled 50,000 items as our retrieval database to arrive at about 2,000 queries with ground truth answers. Following numerous retrieval works, Recall@K was adopted for evaluation.

To test whether the EI Tree method can handle user feedback and further boost search performance, we performed simulation of interactive search in the following way: we extend the image groups to contain images in the similar style but with certain different attributes. We then use the difference of items’ attribute concepts as concept feedback. For example, in a group with two items, a red dress and a blue dress in the same style, we can use the one with red attribute as query and the blue one as the ground truth answer. We report the results for retrieval after adding ‘red, +blue’ as attribute concept feedback. To evaluate the performance extensively, for each query, we conducted two feedback iterations with two attribute concept feedback per iteration.

5.2 Evaluations of Concept Prediction (RQ1)

![Figure 5: Performance of fashion concept prediction (RQ1).](https://www.visenze.com/solutions-overview)
five hand-picked representative groups and the overall performance are shown. The key observations are as follows:

1) Our EITree method achieves the best performance. Notably, compared to the pure image-based methods Vebay and AMNet, EITree performs significantly better for concepts that exhibit relatively large intra-concept visual variance but are easy to describe in words. For instance, we observe large performance improvements in concept groups such as “skirt shape” and “sleeve style”. This demonstrates the usefulness of text modality in accurately predicting fashion concepts. Moreover, we observe performance drops of txtEI and imgEI in which only a single modality is exploited. It shows the importance of multimodal information modeling as well as its potential in boosting the performance of retrieval systems.

2) Focusing on methods that account for both visual and textual modalities, we find that incorporating domain knowledge constraints plays a pivotal role. Firstly, although DSP jointly models the two modalities, it only focuses on embedding the visual image and text into the same vector space by leveraging the cross-modal ranking constraints. Therefore, it generally performs worse than other methods. Secondly, we observe a moderate performance improvement of the prodTree method on “cloth category”, which is supported by the introduced product category tree structure. However, the tree structure ignores other concepts and only exclusive relation can be found among siblings. Thus, we only observe slight variations in performance on other concept groups. In EITree method, we not only capture all these concepts into a tree structure, but also incorporate different relations among sibling concepts. The average 6.63% performance improvement of EITree method over prodTree demonstrates the necessity of introducing such fashion domain knowledge into our end-to-end model.

5.3 Evaluations of Retrieval

5.3.1 Automatic Fashion Retrieval (RQ2). Figure 6 illustrates the performance comparison between the proposed EITree and the other retrieval methods. We observe that EITree achieves the best retrieval performance in terms of Recall@K at all the top K results as compared to the other methods. The performance improvements of EITree over the other methods are significant. For example, in terms of Recall@10, EITree improves the performance of image query by 4.6%, 5.3%, 7.8%, and 9.3% as compared to the prodTree, DSP, AMNet, and Vebay methods, respectively. For text query, the performance improvements of the EITree method are 6.9% and 9.0% as compared to the prodTree and DSP methods on average. The similar performance improvements in terms of image&text query are also observed. Generally speaking, the proposed EITree method is trained on both visual and textual modalities which helps to achieve superior performance on single modality queries. Moreover, when two modalities are both available, the performance improvements of EITree method demonstrate its effectiveness in automatic fashion retrieval due to the following aspects: a) EITree models the semantics of fashion items in the form of a hierarchical semantic representation consisting of multiple levels of concepts, the different relationships between them are also incorporated in the tree structure. Such hierarchical semantic representation provides a more precise interpretation of fashion semantics and guides the end-to-end multi-level learning procedure. b) The explicit similarity function in EITree more accurately characterizes the semantic similarities among items by ensembling different contributions of concepts and features. Note that during the design and implementation of the model, we did not emphasize on the efficiency. When a query product arrives, the total time for processing it through our trained model to get a representation and calculating its similarity to others amounts to about 0.26 s in NVIDIA Titan X GPU.

5.3.2 Fashion Retrieval with Concept Feedbacks (RQ3). In this experiment, concept level manipulations are incorporated to characterize search intent. From the results presented in Figure 7, we observe substantial performance improvements for both EITree method and AMNet with feedback iterations as compared to their automatic retrieval version. Also, the results with two feedback iterations (EITree-2, AMNet-2) generally work better than those with
one feedback iteration (EITree-1, AMNet-1). It validates the usefulness of involving feedback in the loop. However, we also observe that the performance of EITree supersedes that of AMNet by a large margin, which is attributed to the explicit retrieval scheme of the proposed method. Equipped with explicit representation, EITree facilitates the adding and removing operation of fashion concepts by directly increasing or decreasing the corresponding dimensions, while AMNet can only add in concepts via matrix interpolations and no removing operations are allowed. Moreover, because EITree organizes fashion concepts into multiple levels and relationships between concepts are captured, specific concept manipulation only affects other concepts subtly. However, the direct matrix interpolation operation on feature vector does not have such insulation effect. Thus, we even observe performance decrease of Recall@1 and Recall@10 scores for AMNet-2, which might be due to the accumulation of noise introduced by the two feedback iterations.

5.4 Qualitative Analysis

5.4.1 Concept Localization Examples (RQ1). To validate the learning of multi-level concepts in EIT tree, we visualize concepts as in subsection 3.3. Figure 8 shows the up-sampled concept activation maps over the original item images. We observe that the concepts are mapped to appropriate spatial regions. For example, neckline is most likely to occur in the upper part of cloth images, while sleeve often occurs on two sides of cloth images, and big-graphic is usually around the center region of a cloth. For concepts whose location coverage is relatively large and flexible, like floral, texture and colors, their activation maps span large portions of images.

More importantly, we discover certain relationships between the activation maps corresponding to their concept relations as depicted in the EIT tree. First, if two concepts are under the same parent node (or say they are siblings), they describe similar spatial part of a cloth, e.g., peplum skirt and pencil skirt, or v-neck and o-neck. Thus, their spatial information are also similar. Second, we can discover general to specific spatial regions corresponding to their concept relations. For example, we observe that the activation map of T-shirt includes that of cloth parts such as short-sleeve and big-graphic, and the activation map of coat includes that of cloth parts such as fur-neckline and flare-style.

5.4.2 Concept Manipulation Examples (RQ3). In this subsection, we give some examples in Figure 9 for fashion retrieval with concept manipulations by the proposed EITree method. It can be seen that the method is capable of accurately capturing user feedback on fashion concepts. For example, the concepts such as color, sleeve length and skirt length in these four examples are all correctly changed to the user provided ones. Moreover, we observe that modifying several concepts at the same time does not seem to deteriorate the performance (except when the changes made by users conflict with each other or the dataset does not contain such items). As discussed in subsection 5.3, this is because the proposed EITree method encourages concept insulation via explicit representation and explicit similarity.

6 CONCLUSIONS

In order to take advantage of multi-modalities and be able to perform interpretable fashion retrieval, we proposed the EITree which organizes the fashion concepts into multiple semantic levels and augments the tree structure with exclusive as well as independent relations. It captures fashion domain knowledge and guides our end-to-end learning framework. An explicit hierarchical similarity function is then learned to calculate the semantic similarities among fashion products. Based on the proposed EITree, we developed a fashion retrieval scheme supporting both automatic retrieval and retrieval with fashion concept feedback. We systematically evaluated the proposed method on two large fashion datasets. Experimental results demonstrated the effectiveness of EITree in characterizing fashion items and capturing search intent precisely, leading to more accurate results as compared to the state-of-the-art approaches.

In future, we will continue our work in two directions. First, we will study how to build or refine EITree automatically by mining concepts and relations online. Second, with EITree structure, we may also support personalized fashion recommendation [15].

ACKNOWLEDGMENT

This research is part of NExT++ project, supported by the National Research Foundation, Prime Minister’s Office, Singapore under its IRC@Singapore Funding Initiative.