Conversational Image Search

Liqiang Nie*, *Senior Member, IEEE,* Fangkai Jiao, Wenjie Wang, Yinglong Wang, and Qi Tian, *Fellow, IEEE*

***Abstract*—Conversational image search, a revolutionary search mode, is able to interactively induce the user response to clarify their intents step by step. Several efforts have been dedicated to the conversation part, namely automatically asking the right question at the right time for user preference elicitation, while few studies focus on the image search part given the well-prepared conversational query. In this paper, we work towards conversational image search, which is much difficult compared to the traditional image search task, due to the following challenges: 1) understanding complex user intents from a multimodal conversational query; 2) utilizing multiform knowledge associated images from a memory network; and 3) enhancing the image representation with distilled knowledge. To address these problems, in this paper, we present a novel contextuaL imAge seaRch sCHeme (LARCH for short), consisting of three components. In the first component, we design a multimodal hierarchical graph-based neural network, which learns the conversational query embedding for better user intent understanding. As to the second one, we devise a multi-form knowledge embedding memory network to unify heterogeneous knowledge structures into a homogeneous base that greatly facilitates relevant knowledge retrieval. In the third component, we learn the knowledge-enhanced image representation via a novel gated neural network, which selects the useful knowledge from retrieved relevant one. Extensive experiments have shown that our LARCH yields significant performance over an extended benchmark dataset. As a side contribution, we have released the data, codes, and parameter settings to facilitate other researchers in the conversational image search community.**

***Index Terms*—Conversational Image Search, Multiform Knowledge Modeling, Knowledge-enhanced Image Representation.**

I. Introduction

In the traditional image search services, such as Flickr[1] and Bing[2], users seek the desired images by issuing a simple keyword-based query, *i.e.,* a string of unconnected and usually ungrammatical terms. After that, similarity matching [1], [2] is performed between the keywords and the surrounding metadata of images, including captioning, title and description. The search results can be further refined by reordering the images via exploring the visual cues manifested in the initial ranking list or other external sources [3], [4], [5]. The traditional keyword-based image search has achieved considerable advancement over the past decades, yet users' information need of images is quickly becoming more complicated. For instance, in an E-commerce scenario, a customer is keen to buy a T-shirt, but does not express his/her intent exactly at once due to some practical reasons, like the customer does not know what kind of T-shirts the shop precisely has. In light of this, devising a revolutionary search mode, capable of interactively guiding users to clarify their intents step by step, deserves our attention.

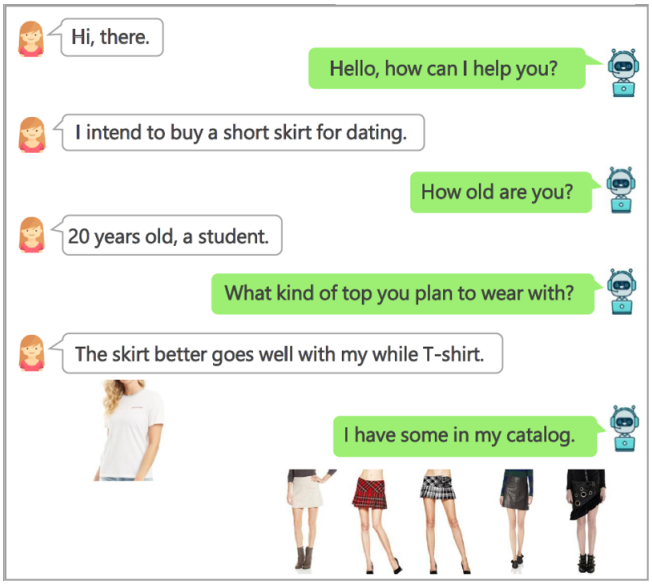


Fig. 1. A typical example of a conversational image search, in the scenario of fashion products seeking.

To cope with the aforementioned problem, conversational image search has attracted increasing interest in recent years. A typical example of a conversational image search is demonstrated in Fig. 1, which encourages user to interact with the system regarding products or services of their interest. This is accomplished by allowing the system to ask more specific and personalized questions, and inducing users' responses, until the system clearly comprehends users' requirements. In contrast to the traditional keyword-based image search, users easily obtain the expected images quickly and efficiently, while the system narrows down their search space dramatically by collecting a set of constraints. Considering its big commercial potential, a long track of research efforts have been dedicated to the conversational image search [6], [7]. Our investigation shows that existing studies usually concentrate on the "conversation" part [8], [9], [10], [11], [12], [13], namely developing preference elicitation paradigm relying on natural language processing techniques to determine which question to ask at each time, so that the system can quickly understand the user need with fewer

Liqiang Nie and Fangkai Jiao are with the Department of Computer Science and Technology, Shandong University, Shandong, China, (e-mail: nieliqiang@gmail.com; jiaofangkai@hotmail.com). *: Liqiang Nie is the corresponding author.

Wenjie Wang is with the School of Computing, National University of Singapore, (e-mail: wenjiewang96@gmail.com).

Yinglong Wang is with Qilu University of Technology (Shandong Academy of Sciences), China (e-mail: wangyl@sdas.org.)

Qi Tian is with Cloud & AI, Huawei Technologies (e-mail: tian.qi1@huawei.com).

[1]https://www.flickr.com.

[2]www.bing.com.

conversational rounds. However, the "search" part is largely unexplored. In general, the current methods only superficially estimate the textual similarity between the conversational query and the metadata of image candidates.

In this paper, we work towards the image search part given the conversational query, namely, users have well-clarified their intents. As a novel problem, it is non-trivial regarding the following research challenges. 1) Understanding the complex user intent expressed in a conversational query. In the conversational image search setting, user intents are revealed in a multimodal and hierarchical conversational session via multi-round dialog pairs. Each dialog pair contains two utterances in the form of ⟨question, response⟩, whereby each utterance may be composed of multiple sentences, and a sentence contains syntactically dependent words and semantically correlated images sometimes. Therefore, how to effectively learn the structural and multimodal query embedding is the first research challenge, which is critical in understanding the user intent. 2) Utilizing multi-form knowledge associated with images. Contexts associated with images are able to emphasize different aspects that provide complementary knowledge for image understanding. Considering the fashion product images in E-commerce websites as an example, they are often described by rich attributes, style tips and product popularity. These contexts can be organized into various forms, spanning from a graph, table and matrix to a digital vector. In this work, we name such structured and useful contexts as knowledge. Unifying multi-form knowledge into a homogeneous base to benefit the relevant knowledge retrieval is another key challenge we face. And 3) learning knowledge-enhanced image representation. In addition to vision, the inextricable knowledge of each image also plays a vital role in describing each image, especially the ones highly correlated to the visual content or the given conversational query. Based upon the unified knowledge base, we are able to retrieve related knowledge via the vision representation and conversational query. However, not all the related knowledge is helpful. Taking product image search for instance, a conversational query may care color, and hence the material- and brand-related knowledge may bring in noise. Thereby, learning enhanced image representation with only the useful knowledge is highly desired.

To address these research problems, we devise a novel contextuaL imAge seaRch sCHeme (LARCH for short). As illustrated in Fig. 2, LARCH comprises three components: query representation learning, multi-form knowledge modeling, and image representation learning. In the first component, we devise a multimodal hierarchical graph-based neural network to capture the user intent. In this network, each word, image, sentence, utterance, dialog pair and the entire session is treated as a node of the graph as shown in Fig. 3. One edge links two nodes if they have dependency correlation or subordination relationship. The session node representation is finally used to represent the conversational query, and hence denotes the user intent. For unifying multi-form knowledge, we present a multi-form knowledge embedding memory network. It separately embeds multi-form knowledge structures into knowledge entries, stored in the ⟨key, value⟩ pairs. We then project the key (value) of each knowledge entry into the same semantic key (value) space, and stack them into a key (value) memory. In this way, we obtain a homogeneous knowledge base from heterogeneous knowledge in various forms. Thereafter, we use the conversational query (visual representation) to retrieve the relevant knowledge from this homogeneous base. Specifically, we obtain an attentive vector via dot product between the conversational query (visual representation) and each key embedding in the key memory. The weighted fusion of all the value embeddings in the value memory with respect to the attentive vector is ultimately viewed as the query-aware (vision-aware) knowledge representation. For each retrieval, we indeed only calculate the attentive vector to learn the knowledge representation, instead of re-modeling the knowledge, which is hence very fast. In the third component, we learn the image representation by fusing the visual representation and the prior retrieved relevant knowledge. This is accomplished via a gated neural network, which further filters the useful knowledge to strengthen image representation. Extensive experiments over publicly accessible benchmark dataset have verified the superiority of our LARCH model over several state-of-the-art baselines. As a side product, we have released all the data, codes, and parameter settings to facilitate other researchers in this community[3].

In summary, the contributions of this work are in three-fold:

- To learn the conversational query embedding, we develop a multimodal hierarchical graph-based neural network, which is capable of characterizing the session structure and multimodal context for better user intent understanding.
- To facilitate relevant knowledge retrieval, we devise a multi-form knowledge embedding memory network, which unifies heterogeneous knowledge into a homogeneous base. As far as we know, this is the first work on modeling multi-form knowledge towards conversational image search.
- We present a novel gated neural network to learn the knowledge-enhanced image representation, which further filters useful knowledge from the retrieved relevant one.

The rest of this paper is structured as follows. In Section II, we briefly review the related literature. In Section III, we detail our proposed LARCH scheme. Section IV and V respectively introduces our extended dataset and analyzes the experimental results, followed by conclusion and future work in Section VI.

II. Related Work

In this section, we review the existing approaches of traditional image search and conversational search, which are highly related to our work.

A. Traditional Image Search

Traditional image search is a sophisticated process of finding the desired images, involving query understanding, indexing, unimodal matching, and reranking. Considering that

[3]https://github.com/SparkJiao/LARCH.

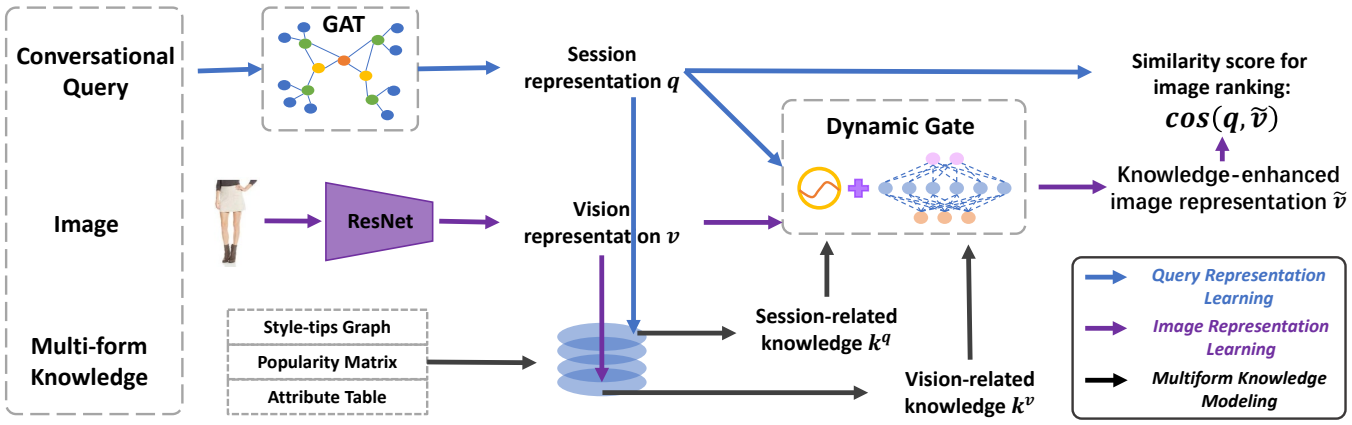


Fig. 2. Schematic illustration of our proposed LARCH model. It comprises three components, namely query representation learning, multi-form knowledge modeling, and image representation learning.

reranking is usually positioned as the key component to improve the search results [14], we only introduced this part, which can be roughly grouped into four categories: self-reranking, example-based reranking, crowd-reranking, and interactive reranking.

Self-reranking approaches hypothesize that the top-ranked images in the initial ranking list returned by the text-only match are much more relevant to the given query. In light of this, existing studies mine relevant patterns from the initial ranked list and rerank the results via different techniques, such as clustering [15], [16], pseudo-relevance feedback [17], [18], [19], object recognition [20], and graph-based learning [21], [22]. Although straightforward, self-reranking approaches may yield suboptimal performance since the initial ranking results are not reliable, caused by noisy or missing surrounding texts.

To overcome the problem of self-reranking, along with the content-based image search, the paradigm of example-based reranking emerges, whereby the query is composed by both an image example and the textual description. Many technical methods are designed to boost the ranking performance by leveraging the user-provided visual examples, such as objects in images, to discover the relevant patterns regarding the given query. Typical methods include linear multimodal fusion [23], geometric verification [24], and query expansion [25], [26]. Despite its scientific value, example-based search is not commonly used in our daily life, let alone the corresponding reranking approaches.

Compared to the above two categories, crowd-reranking methods generate the final result list by exploring multiple online image resources. For instance, Liu *et al.* [27] mined common patterns from results returned by multiple image search engines. Another example is presented in [28], whereby authors augmented queries from the image collection on the Web. The philosophy of crowd-reranking is that different search results are capable of reinforcing or complementing the relevant visual information, besides, there may exist common visual patterns across different search results regarding the same query. However, the biggest challenge is the noisy nature of Web knowledge.

Approaches in the interactive reranking category require a user in the loop to provide complementary requirements or annotate results, which consistently outperform the aforementioned three reranking methods by a large margin. Researchers in [29] leveraged relevance feedback to identify the relevant clusters for improving browsing efficiency. In particular, they first employed clustering techniques to cluster the top image search results and then asked the users to label the relevance of those clusters. The work in [30] presents an image search system via the color map, which enables users to specify color distributions in the desired images. This system provides a way to enable users to indicate their visual expectation. In a sense, designing a friendly interface to maintain the user experience while minimizing the interaction time is critical in interactive reranking.

Conversational image search, although close to the interactive image reranking in principle, still exhibits major difference. In the former, the model plays a proactive role in guiding users to clarify their intents; whereas, the model in the latter one is reactive and waiting for users' feedback.

B. Conversational Search

Conversational search has attracted great attention in recent years. Back to 2016, a preference elicitation framework is introduced in [31], which identifies the questions asked to users for quickly learning their preferences in seeking restaurants. The framework is based on the model of probabilistic matrix factorization and improves personalized search over a static model remarkablely. One year later in 2017, Kenter and Maarten [32] argued that the conversation can be framed as a machine reading task and introduced an attentive memory network with a hierarchical input encoder towards machine reading. In the same year, Radlinski and Craswell [6], presented a theoretical framework for the basic design and evaluation of conversational information retrieval systems. In 2018, more literature on conversational search appears. Sun and Zhang [33] successfully searched the right item(s) for a user by analyzing the user conversation in the current session, via interactively inducing the user to clarify the purchase requirement, and making personalized search, based on the current session and user's purchase history. Zhang *et al.* [34] believed that conversational search can actively ask appropriate questions so as to understand the user needs.

They presented a multi-memory network architecture as well as its personalized version for conversational search, jointly integrating sequential modeling and attention mechanisms. In [7], Feng et al. proposed a unified implicit dialog framework by enabling dialog interactions with domain data. It makes existing development and chat data reusable and adaptable to new domains. Most recently in 2019 [35], Qu and his colleagues studied shallow models with a rich set of hand-crafted features and deep models incorporating context information without feature engineering, to accurately detect and predict user intents in the conversation.

Beyond the data-driven methods aforementioned, some researchers explored knowledge to boost the conversational search performance. Agarwal *et al.* [36] focused on the task of generating textual responses conditioned on the previous multimodal conversational history. Towards this goal, they designed a knowledge-grounded multimodal conversational model, whereby an encoded knowledge base representation is appended to the decoder input. Liu *et al.* [37] proposed a deep neural matching network to leverage external knowledge for response ranking in the conversational setting. Specifically, they incorporated external knowledge into deep neural models with pseudo-relevance feedback and QA correspondence knowledge distillation. In [38], to produce more contentful responses, the authors presented a knowledge-grounded neural conversation model, which generalizes the sequence-to-sequence approach by conditioning responses on both conversation history and external "facts".

It is worth mentioning that the increasing research interests in conversational search have resulted in a consequent growth in recordings of spoken search interactions [39]. That is out of the research scope of this work. Although great success has been achieved by the conversational search, we observed that existing studies usually work on the "conversation" part, namely asking the right question at the right time, so that the system can better capture the user need. As a complement, we will focus on the "search" part, *i.e.,* seeking the relevant images given the conversational query.

III. Our Proposed LARCH Model

Our proposed LARCH model comprises three key components: 1) query representation learning; 2) multi-form knowledge modeling; and 3) image representation learning. In particular, we first understand the user intent by embedding the conversational query into a vector. After that, we encode the auxiliary multi-form knowledge of the image into a homogeneous key-value base, and then retrieve the relevant knowledge via both the conversational query and image representation for each search. We then fuse the visual representation and useful knowledge towards the knowledge-enriched representation. Ultimately, we leverage the query representation and knowledge-enhanced image representation to estimate the ranking scores for all the image candidates.

During the phase of inference, the representations of all images and the vision-related multiform knowledge are cached for less latency. Given a conversational query, it will be first encoded through the query representation learning module

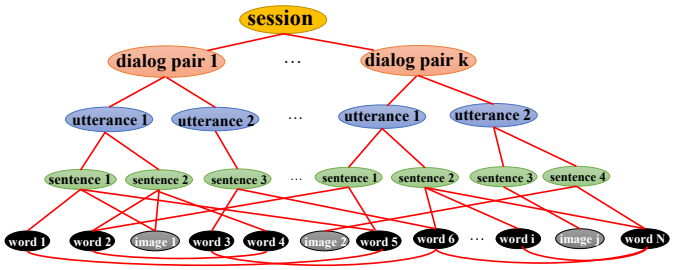


Fig. 3. Illustration of the proposed multimodal hierarchical graph-based neural network. The edges are constructed by the dependency correlation between the words within one sentence, the subordination relationship in a session, or the cross-modal relation. The representation of the session node is used to indicate the user interest.

(blue arrows in Fig. 2), and then adopted to retrieve the relevant knowledge from the homogeneous key-value base via multi-form knowledge modeling. Finally, LARCH ranks all image candidates via the similarity between their knowledge-enhanced visual representations and the query representation, and returns the top-ranked images.

In this section, we will detail each component of LARCH.

A. Query Representation Learning

To characterize the user intent conveyed in the given conversational query, we propose a multimodal hierarchical graph-based neural network. To be more specific, the units in the hierarchical context are treated as different nodes, including words, images, sentences, utterances, dialog pairs, and the entire session. Edges in the graph are built by one of the following three approaches: 1) Subordination relationships. In a conversational query, it has natural subordination relationships. Taking the case in Fig. 3 as an example, the *sentence 1* node is composed of *word 1* and *word 5* and hence these two word-level nodes are linked to *sentence 1* node via two edges. In addition, we build short-cuts across layers to facilitate information propagation, *e.g.,* the node of *sentence 1* is connected with the *session* node directly, and the *word 2* node is also connected with *dialog pair 1* node. For brevity, we omit the short-cuts from the Fig. 3. 2) Cross-modality relationships. To incorporate the visual information into the query representation in a fine-grained manner, the visual features of images are viewed as separate nodes at the same level with the words, and associated with the sentences mentioning it. The short-cuts connecting the image nodes are also included for better multimodal message passing. And 3) dependency correlation. The semantic information at the word level has not been fully explored thus far. We thus build the semantic dependency between different words in the graph. Taking the sentence "I intend to buy a short skirt for dating" for an example, the noun *skirt* is the object of the verb *buy*, and the adjective *short* modifies it, therefore there are two edges between each two words to indicate the relations. In this work, we leverage an efficient neural dependency parser proposed by Chen *et al.* [40] to analyze the grammatical structure of the sentences. It should be noted that the edges defined above are undirected.

In the multimodal graph, the node embeddings are propagated via the edges defined above. We initialize the image nodes with the features extracted by the pre-trained ResNet [41] and the word nodes via the pre-trained GloVe [42] word vectors. For the upper level nodes, the features are initialized as the average of their children embeddings, whereas the image features are excluded during the average process since the vision information lies in spaces different from that of the textual ones. Due to the different hidden size of image and textual features, we use two linear layers to project the hidden states of image nodes and the textual nodes into the same dimension size, respectively.

Inspired by [43], we utilize the simple but effective Graph Attention Networks (GAT) to propagate the node features via edges in each session graph. Supposing that we have N nodes in a graph and their embeddings are denoted as $\mathcal{H} = \{\mathbf{h_1}, \mathbf{h_2}, ..., \mathbf{h_N}\}$, we update the node features by a multi-head attention mechanism [44]. For $\mathbf{h_i} \in \mathbb{R}^H$, the attention coefficients of its neighbor nodes for each head are calculated by

$$\alpha_{ij} = \frac{\exp(\varphi(\mathbf{a}^\top[\mathbf{W}\mathbf{h}_i||\mathbf{W}\mathbf{h}_j]))}{\sum_{k\in\mathcal{N}_i}\exp(\varphi(\mathbf{a}^\top[\mathbf{W}\mathbf{h}_i||\mathbf{W}\mathbf{h}_k]))}, \tag{1}$$

where α_{ij} indicates the importance of node j to node i, $\mathcal{N}_i$ denotes the neighbor set of node i, φ is the LeakyReLU activation function, $||$ refers to the concatenation operation, $\mathbf{a} \in \mathbb{R}^{H'}$ and $\mathbf{W} \in \mathbb{R}^{H\times H'}$ are the learned parameters. Note that $\mathbf{W}$ projects node embedding into a new vector space with dimension H'. Formally, the updated node embedding $\mathbf{h'_i}$ is obtained by

$$\mathbf{h'_i} = \|_{k=1}^{K}\phi(\sum_{j\in\mathcal{N}_i}\alpha_{ij}^k\mathbf{W}^k\mathbf{h_j}), \tag{2}$$

where K is the total number of attention heads and ϕ is the Exponential Linear Unit (ELU) activation function [45]. We adopt L GAT layers with the last layer as the output layer, where the single head attention is used and the ELU activation function is removed, inspired by the setting in [43]. Finally, the session node embedding is regarded as the query representation $\mathbf{q} \in \mathbb{R}^D$, since it aggregates the information of the entire conversational query.

B. Multi-form Knowledge Modeling

Knowledge associated with images is usually presented in various structures, including but not limited to table, matrix, and graph. Taking the fashion product image as an example, the heterogeneous knowledge contains style tips, popularity, and attributes. Style tips, advising users on compatibility between two fashion items such as "white T-shirt going well with black pants", are usually organized into a graph where a node presents a fashion product and an edge links two matched products. Product popularity in celebrities, stands for the preference distribution of celebrities over different kinds of products, for instance, some celebrities favor white sneakers. Product popularity can be presented as a matrix, whereby each row and column denotes a celebrity and a fashion product, respectively. The entry in a matrix indicates the preference score of a celebrity to a product. As to product attributes,

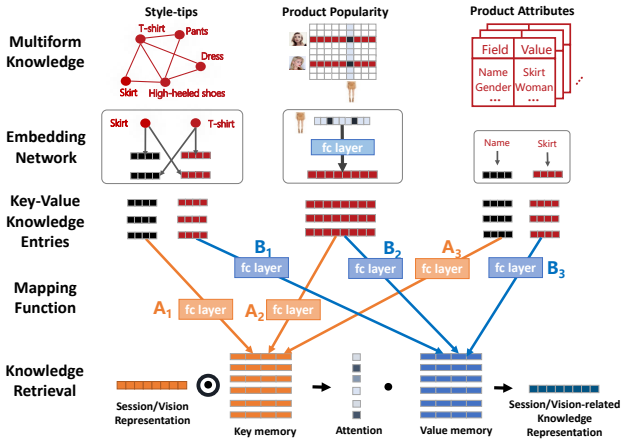


Fig. 4. Illustration of multi-form knowledge modeling. We use "session" and "vision" to represent "conversational query" and "visual part" of an image, respectively.

they are organized in a field-value table to record the common attributes of products, such as material, color and size. In this work, we only consider three forms of knowledge: table, matrix and graph, since they cover most of the commonly seen structures.

To encode such heterogeneous knowledge structures into a homogeneous base that facilitates relevant knowledge retrieval, we devise a multiform knowledge modeling memory network as illustrated in Fig. 4. We first extract the raw knowledge related to the image through keyword matching to compress the search space. After that, we adopt various embedding networks for specific knowledge structure embedding:

- **Style-tips graph.** We embed each word in the product name into a vector, initialized by the pre-trained GloVe embeddings. Since the product name (*i.e.,* one node) may contain multiple words, we take the average of the word embeddings to initialize the node embedding. Each edge in the style-tips graph connects two matched items, and both embedded items are viewed as key and value entries.
- **Popularity matrix.** The matrix structure is regarded as a set of column vectors, whose elements are the preference scores of all celebrities to one product, and a fully-connected layer is applied to embed each column vector into a knowledge entry.
- **Attribute table.** The embeddings of each field and its value are initialized as the average of their words. The filed and value representations are treated as the key and value entries, respectively.

After that, the knowledge entries extracted from different structures are further projected into a knowledge base, *i.e.,* a homogeneous key-value memory. To accomplish this, we adopt two distinct linear mapping functions for each knowledge entry type, denoted as $\mathbf{A}_i$ and $\mathbf{B}_i$ to project the knowledge entries to key memories and value memories, respectively [46]. It is worth noting that the key and value knowledge entries of the matrix are the same, whereas the key and value memories for graph and table are encoded from the corresponding key and value entries. From the unified knowledge base, taking the given conversational query representation $\mathbf{q}$, *i.e.,* session node embedding, as an example,

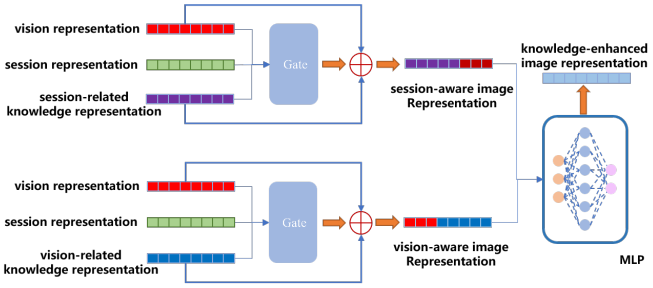


Fig. 5. Illustration of knowledge-enhanced image representation learning via a dynamic gate mechanism.

the retrieved knowledge of the session $\mathbf{q}$ is fused based upon an attentive mechanism, formulated as below

$$\begin{cases} \beta_i = \dfrac{\exp(\mathbf{q}^\top \mathbf{m}_i^k)}{\sum_j^M \exp(\mathbf{q}^\top \mathbf{m}_j^k)}, \\ \mathbf{k}^q = \sum_i^M \beta_i \cdot \mathbf{m}_i^v, \end{cases} \tag{3}$$

where M is the number of knowledge entries, $\mathbf{m}_i^k$ and $\mathbf{m}_i^v$ are the key and value for the i-th knowledge entry, respectively, and $\mathbf{m}_i^k, \mathbf{m}_i^v \in \mathbb{R}^D$. In this way, we obtain the session-related knowledge representation $\mathbf{k}^q$. Similarly, we are able to use the vision representation of the product, denoted as v, to retrieve vision-related knowledge representation $\mathbf{k}^v$.

C. Image Representation Learning

For a given conversational query and any image candidates in our dataset, we actually have two knowledge vectors retrieved from the unified knowledge base: the session-related knowledge representation and the vision-related one. The retrieved knowledge greatly affects the search performance, since it may be highly relevant to the given query but not appears in the vision image, or it encodes the consistent information highly correlated to the visual image. We thus strengthen the visual image features with such two knowledge representations.

Despite its relevance to the given conversational query or the visual image, not all the retrieved knowledge is helpful to the search task. We take two examples to intuitively explain the reasons: 1) For the first conversational query "... buy a white T-shirt matching my black pants like this ...", it contains a target product *white T-shirt* and a side product *black pants*. The retrieved knowledge by the given query may contains the black pants. Assuming that an image candidate is a black T-shirt, it may be treated as positive. This is because the image candidate contains the black pants after knowledge incorporation, which well-matches the conversational query. And 2) for the second example, we search the knowledge base with an image candidate of a white T-shirt with long sleeves and V-neckline, where the retrieved knowledge may include both styles. The conversational query, however, may only care about the materials and sleeves. The above two examples demonstrate the fact that we retrieve the knowledge by the vision and session representation separately, and ignore the correlations among them.

Towards this end, we devise a novel gated neural network to further select the useful knowledge from the relevant one to enhance the image representation learning. In particular, we apply a dynamic gate mechanism conditioned on the conversational query $\mathbf{q}$, vision embedding $\mathbf{v}$, and retrieved knowledge $\mathbf{k}^q$ ($\mathbf{k}^v$), formulated as

$$\begin{cases} \mathbf{g}^v = \sigma(\mathbf{W}_g^v[\mathbf{q}||\mathbf{v}||\mathbf{k}^v]), \\ \mathbf{g}^q = \sigma(\mathbf{W}_g^q[\mathbf{q}||\mathbf{v}||\mathbf{k}^q]), \end{cases} \tag{4}$$

where $\mathbf{W}_g^v \in \mathbb{R}^{D\times 3D}$ and $\mathbf{W}_g^q \in \mathbb{R}^{D\times 3D}$ are the learnable parameters, and σ is the sigmoid function. Meanwhile, we use $\mathbf{g}^v$ and $\mathbf{g}^q$ to denote the controller vectors, whose elements are between 0 and 1. The controller is to determine which knowledge should be incorporated to enhance the image representation, written as

$$\begin{cases} \mathbf{v}^v = \mathbf{g}^v \odot \mathbf{k}^v + (\mathbf{1} - \mathbf{g}^v) \odot \mathbf{v}, \\ \mathbf{v}^q = \mathbf{g}^q \odot \mathbf{k}^q + (\mathbf{1} - \mathbf{g}^q) \odot \mathbf{v}, \end{cases} \tag{5}$$

where $\odot$ is the element-wise product operation. In this way, we obtain the query-aware and vision-aware image representation, respectively. We finally adopt a two-layer Multilayer Perceptron (MLP) with ReLU activation to fuse these two kinds of image representations as,

$$\tilde{\mathbf{v}} = \mathbf{W}_1(f(\mathbf{W}_2[\mathbf{v}^v||\mathbf{v}^q] + \mathbf{b}_2) + \mathbf{b}_1), \tag{6}$$

where $f(\cdot)$ is the ReLU function. Other symbols like $\mathbf{W}_1$, $\mathbf{W}_2$, $\mathbf{b}_1$ and $\mathbf{b}_2$ are the learnable parameters.

D. Loss Function

Given a conversational query $\mathbf{q}$, assuming that we have a set of positive images $\mathcal{I}^+$ and negative ones $\mathcal{I}^-$, we leverage the cosine similarity function to estimate the similarity scores of each query-image pair. We then turn to a margin ranking loss to optimize all trainable parameters in our proposed LARCH model. In particular, for each query representation $\mathbf{q}$, its one positive image representation is denoted as $\tilde{\mathbf{v}}_i$ and the negative one is $\tilde{\mathbf{v}}^j$. Then the objective function is written as

$$\mathcal{L} = \max(0, \cos(\mathbf{q}, \tilde{\mathbf{v}}_i) - \cos(\mathbf{q}, \tilde{\mathbf{v}}_j) + 1), \tag{7}$$

where $i \in \mathcal{I}^+$ and $j \in \mathcal{I}^-$. During the phase of inference, the image candidates are ranked based upon the similarity score for each conversational query.

IV. Dataset

To justify the effectiveness of our proposed LARCH model and fairly compare it with the baselines, we constructed a new dataset, namely MMD 2.0, based upon the widely-used MMD benchmark dataset [47]. MMD comprises more than 150 thousands conversational sessions between users and the dialog robot in the retail domain, and each session on average consists of approximately 40 multimodal utterances. Over 1 million fashion product images with abundant related knowledge are crawled from the well-known online retailing

TABLE I
STATISTICS OF THE ORIGINAL MMD DATASET AND OUR AUGMENTED MMD 2.0 DATASET.

		MMD			MMD 2.0		
		Train	Dev	Test	Train	Dev	Test
#Sessions		105,439	22,595	22,595	105,439	22,595	22,595
#Positive	*max.*	10	10	10	10	10	10
Images	*min.*	1	1	1	1	1	1
per Session	*avg.*	4	4	4	4	4	4
#Negative	*max.*	88	91	88	1,050	1,050	999
Samples	*min.*	1	1	1	1	1	995
per Session	*avg.*	26	26	26	966	964	990

websites, such as Amazon[4], Jabong[5], and Abof[6]. More details of MMD dataset are shown in Table I.

Given a specific conversational query, the dialog robot is to select the correct (positive) product images from a set of image candidates, including some unmatched (negative) images. For example, as shown in Fig. 1, the dialog robot is requested to recommend a skirt going well with the white T-shirt, and the correct images can be the skirts presented in the conversation. However, the candidates also include other incorrect products, such as the skirt unsuitable for the white T-shirt, or even a pair of pants.

As displayed in Table I, the number of the negative images is too small in the original MMD dataset compared with that of the positive ones. This is inconsistent with the real application scenarios, where the model retrieves correct images from millions or even billions of image candidates.

Prior to the introduction of our augmented dataset, we first explain the term *category*. The *category* of a image is the type of the corresponding product, *e.g.,* the *category* of *a pair of silver lambskin high-heel shoes* is *high-heel shoes*. Based on this, we divided the negative samples into two types. With the given query requesting a skirt going well with the white T-shirt, a skirt unmatched with the T-shirt is viewed as a negative image within the same category (*skirt*). Whereas a pair of pants is viewed as a negative image with different category (*pants*). Generally, discriminating the positive images from the negatives sharing the same categories is much more difficult than from the ones of different categories, since the former process requires a more detailed understanding of the user intents, such as the preferred material or style.

To better simulate the real-world image search environment and increase the difficulty of the dataset, we added more negative images that are in the same category with the positive ones but contain incorrect attributes. As shown in Fig. 6, the details of our augmentation algorithm can be summarized as follows:

step 1 Given a labeled conversational query and a product image pair, we extracted the *name* and *fields* from the session annotations, and the *taxonomy* as well as *fields* from the image annotations, which have been provided by [47]. Note that *name* and *taxonomy* are used to describe the *category* information for the conversational query and product, respectively. Meanwhile, the *fields* is a key-value set indicating the product attributes, *e.g.,* $\langle gender, male \rangle$.

step 2 If the *name* and *taxonomy* are different, the product image is a negative sample in a different category with the correct images. Then the image is discarded and the algorithm ends. Otherwise, it moves to the next step.

step 3 If the *fields* annotations of the session and the image are all the same, *i.e.,* the user preferred attributes and image are consistent, the image is judged as a positive sample and filtered out. Otherwise, the algorithm moves to the final step.

step 4 If the image is in the same category with the correct ones but has some different attributes, it is used to augment the dataset as a negative sample.

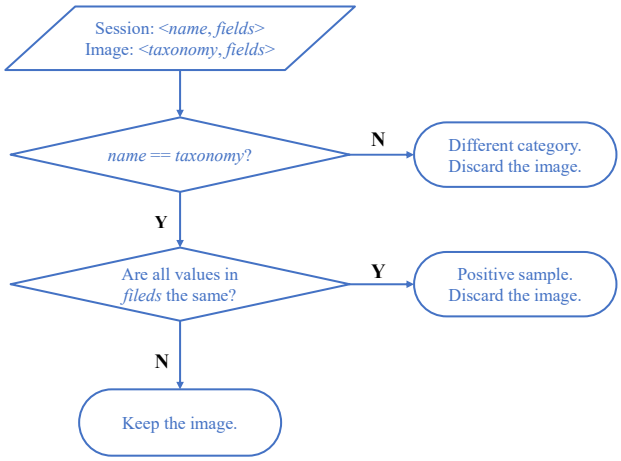


Fig. 6. The flow diagram of our data augmentation algorithm based upon the MMD dataset.

In our experiments, we adopted the *material* and *gender* as the keys for comparison in the *fields* annotations since they are the most distinguishable properties. For example, given the query "I want a white T-shirt for women", a men's T-shirt is viewed as a negative sample due to the inconsistent *gender* field. After that, we followed the original split for training, validation and testing set, and chose up to 1,000 negative samples selected by the algorithm to augment each conversational query. All of the supplemented negative images are sampled from the original image set. We named the augmented dataset as MMD 2.0 and the statistics are also shown in Table I.

V. EXPERIMENTS

In this section, we first detail the experimental settings including hyper-parameters and evaluation metrics. Following that, we conduct experiments to justify the effectiveness of our model and each of its components. We finally present some representative cases and error analyses.

[4]https://www.amazon.com/.
[5]https://www.jabong.com/.
[6]https://www.abof.com/.

TABLE II
PERFORMANCE COMPARISON BETWEEN OUR MODEL AND SEVERAL STATE-OF-THE-ART BASELINES OVER OUR AUGMENTED DATASET MMD 2.0. THE SYMBOL * MEANS STATISTICALLY SIGNIFICANT IMPROVEMENT OVER THE STRONGEST BASELINE WITH $p < 0.05$.

	K = 5			K = 10			K = 20		
	Precision@5	Recall@5	NDCG@5	Precision@10	Recall@10	NDCG@10	Precision@20	Recall@20	NDCG@20
CMR	0.1306	0.1595	0.1851	0.0851	0.2076	0.2088	0.0532	0.2572	0.2286
MHRED	0.1112	0.1450	0.2643	0.1088	0.2175	0.2976	0.1027	0.4106	0.3496
UMD	0.3422	0.4036	0.3662	0.2134	0.5036	0.4087	0.1264	0.5969	0.4415
MAGIC	0.4711	0.5642	0.4806	0.2688	0.6414	0.5148	0.1467	0.7005	0.5361
LARCH (Ours)	**0.5501***	**0.6582***	**0.6829***	**0.2999***	**0.7161***	**0.7121***	**0.1595***	**0.7620***	**0.7302***

A. Experimental Settings

1) Hyper-parameter Settings: We adopted the Adam optimizer with the initial learning rate as 1e-5 in all experiments. The batch size is 200 and the dropout rate is set to 0.1. For query representation learning, we used a graph attention network with $L = 5$, where the first four layers employ the multi-head attention [44] with 16 attention heads. The hidden size for each head is set to 64 and the output dimension of the final layer is 512. For multi-form knowledge modeling and image representation learning, the extracted image features via ResNet18 and various knowledge embeddings are projected to 512 by a fully-connected layer for further computation. The ratio of positive and negative image is 1:1 during the training. And the models are trained for 35,000 steps for around 41 hours.

2) Experiment Environment: We implemented our model and baselines with the help of Pytorch[7] and conducted all the experiments over single RTX 2080Ti with CUDA 10.1. The operation system of the server is Ubuntu 18.04.05.

3) Evaluation Metrics: We chose the model with the best performance measured by the loss value on the validation set for testing. During testing, the model ranks all image candidates based on the cosine similarity and the top-K images are selected for evaluation, where $K \in \{5, 10, 20\}$. Following the previous work [47], we adopted three metrics to measure the performance:

- **Precision@K** measures the fraction of the number of the relevant images that have been selected over K.
- **Recall@K** measures the fraction of the number of the selected relevant images over the number of total relevant ones.
- **NDCG@K**: The Normalized Discounted Cumulative Gain (NDCG) measures the ranking quality. It is the fraction of the Discounted Cumulative Gain (DCG) over the Ideal Discounted Cumulative Gain (IDCG), where the former is the sum of the relative scores scaled by their rank position and the latter is the maximum possible DCG in an ideal setting.

B. Overall Performance Comparison

In this subsection, we compared our proposed LARCH model with several state-of-the-art baselines:

[7]https://pytorch.org/.

- **CRM** [48]: It takes a belief tracker to generate and update the semi-structured query with facet-value pairs to represent the user conversation history. And the concatenated learned facet vector representations are taken as input features to a 2-way factorization machine [49], which predict the rating score for each query-image pair.
- **MHRED** [47]: It leverages a hierarchical multimodal encoder where the image features extracted from VGGNet [50] are concatenated with the textual ones at the dialog level to generate the multimodal representation. Finally, the cosine similarity between the image embedding and the multimodal conversational query representation is calculated for ranking.
- **UMD** [9]: It leverages a bidirectional recurrent neural network to model the dialog at the high level, and it uses a multimodal encoder with attention mechanism and a taxonomy-attribute combined tree to encode the multimodal utterances at the low level. After that, the cosine similarity is used for ranking, similar to [47].
- **MAGIC** [8]: It is the cutting-edge method for contextual image search, which is the first work to consider the knowledge of images (*i.e.,* attributes) for conversational image search. In general, MAGIC ranks the product image candidates via the cosine similarity between the conversational query representation and the knowledge-enhanced image representation.

Table II summarizes the overall performance of our LARCH model compared with the aforementioned baselines. From this table, we have the following observations: 1) Our proposed LARCH model outperforms all the baselines consistently and significantly with a large margin, which demonstrates the superiority of our method. 2) Incorporating the knowledge into the image representation greatly enhances the overall performance of conversational image searching, since both MAGIC and LARCH obtain large improvement over the other methods. 3) LARCH exceeds the best method, MAGIC, verifying the advantages of LARCH on utilizing knowledge for image representation learning. One reason may be that MAGIC encodes only the *attribute* knowledge to strengthen the image representation, whereas ours encodes the multi-form knowledge in a unified manner for image representation learning. And 4) when $K \in \{5, 10\}$, the LARCH model largely outperforms MAGIC and the other baselines on all the metrics. As for $K = 20$, the gap of improvement over

TABLE III
EFFECT OF DIFFERENT MODEL COMPONENTS OVER OUR AUGMENTED DATASET MMD 2.0. THE SYMBOL * MEANS STATISTICALLY SIGNIFICANT IMPROVEMENT OVER OTHER RESULTS WITH $p < 0.05$.

	K = 5			K = 10			K = 20		
	Precision@5	Recall@5	NDCG@5	Precision@10	Recall@10	NDCG@10	Precision@20	Recall@20	NDCG@20
LARCH	**0.5501***	**0.6582***	**0.6829***	**0.2999***	**0.7161***	**0.7121***	0.1595	**0.7620**	**0.7302**
w/o Vision-related Knowledge	0.5273	0.6306	0.6398	0.2988	0.7111	0.6813	**0.1598**	0.7607	0.7010
w/o Session-related Knowledge	0.5154	0.6169	0.6453	0.2841	0.6781	0.6763	0.1538	0.7348	0.6986
w/o Gate	0.4981	0.5980	0.6235	0.2847	0.6795	0.6654	0.1545	0.7389	0.6890
w/o GAT	0.5189	0.6194	0.6399	0.2862	0.6811	0.6714	0.1541	0.7329	0.6921
w. VGG	0.4953	0.5934	0.6208	0.2776	0.6624	0.6559	0.1516	0.7235	0.6800
w. Weighted Avg.	0.5413	0.6494	0.6728	0.2963	0.7081	0.7028	0.1575	0.7534	0.7206

TABLE IV
EFFECT OF DIFFERENT KNOWLEDGE FORMS OVER OUR AUGMENTED DATASET MMD 2.0. THE SYMBOL * MEANS STATISTICALLY SIGNIFICANT IMPROVEMENT OVER OTHER RESULTS WITH $p < 0.05$.

	K = 5			K = 10			K = 20		
	Precision@5	Recall@5	NDCG@5	Precision@10	Recall@10	NDCG@10	Precision@20	Recall@20	NDCG@20
LARCH	**0.5501***	**0.6582***	**0.6829***	0.2999	0.7161	**0.7121**	0.1595	0.7620	**0.7302**
w/o Attribute	0.5259(-4.40%)	0.6321(-3.97%)	0.6462(-5.37%)	0.2954(-1.50%)	0.7060(-1.41%)	0.6843(-3.90%)	0.1583(-0.75%)	0.7570(-0.66%)	0.7043(-3.55%)
w/o Style tip	0.5498(-0.05%)	0.6578(-0.06%)	0.6778(-0.75%)	**0.3005**(+0.2%)	**0.7171**(+0.14%)	0.7079(-0.59%)	**0.1598**(+0.19%)	**0.7637**(-0.22%)	0.7261(-0.56%)
w/o Celebrity	0.5176(-5.91%)	0.6213(-5.61%)	0.6458(-5.43%)	0.2891(-3.60%)	0.6905(-3.57%)	0.6812(-4.34%)	0.1568(-1.69%)	0.7493(-1.67%)	0.7045(-3.52%)
w/o All	0.4961(-9.82%)	0.5969(-9.31%)	0.6072(-11.09%)	0.2862(-4.57%)	0.6835(-4.55%)	0.6522(-8.41%)	0.1550(-2.82%)	0.7411(-2.74%)	0.6749(-7.57%)

precision and *recall* drops. However, the gain over *NDCG@20* keeps stable. This is possibly because when more image candidates are selected, the amount of correct images is also increasing for baseline models, however, our LARCH model can rank more positive images at a higher position, exhibiting the better ranking quality of our model.

C. Component-wise Evaluation

In this section, we worked towards answering these two research questions: 1) how much does each component in our proposed LARCH model contribute to the overall performance? And 2) which knowledge form is more important to the overall performance?

1) Ablation Study: We conducted this set of experiments by disabling a module at each time while fixing the others.

- **w/o vision-related knowledge**: We removed the branch of vision-related knowledge and left others untouched.
- **w/o session-related knowledge**: Similarly, we removed the branch of session-related knowledge and kept others unchanged.
- **w/o gate**: In this variant, the gate module defined in Eqn. (4), (5) and (6) is replaced with a simple 3-layer MLP, whose input is the concatenation of the vision embedding $\mathbf{q}$ together with the retrieved knowledge $\mathbf{k}^q$ and $\mathbf{k}^v$. This is to verify the necessity of the dynamic gate mechanism.
- **w/o GAT**: To show the importance of our proposed conversational query representation learning, we replaced the GAT layers with the LSTM-based [51] multimodal hierarchical encoder, which is used by MAGIC.
- **w. VGG**: To clarify the influence brought by different pre-trained backbones and keep fair comparison with MHRED, we replaced the ResNet18 with VGG16 to extract the image features.
- **w. Weighted Avg.**: In order to show the impact of different initialization methods, we initialized the embedding of upper level nodes in the multimodal graph as the **weighted** average of their children embeddings, in which the weights are controlled by a learnable self-attention mechanism.

Table III displays the impact of each module of LARCH regarding three metrics with varying depths. It can be observed that 1) the multi-form knowledge retrieved by both vision and session representation plays an important role in finding the relevant images, especially when K is restricted. 2) The performance of the variant without session-related knowledge drops by around 3.5% to 4% over all the metrics when $K = 5$, showing that using the session query to retrieve the relevant knowledge is essential. This is because the knowledge of the images is complex and the session query only reaches part of it, *e.g.,* several attributes, so that session-related knowledge reduces the incorporation of noises. 3) The dynamic gate mechanism is vital on injecting the multi-form knowledge into image representation. The overall performance drops significantly without this module. The dominant reason may be that the knowledge retrieved through the session or vision representation often has its own bias, thus introducing much useless side information. 4) Our proposed GAT based query representation learning component also contributes to the ranking quality greatly, verifying the necessity of the fine-grained multimodal fusion. One possible reason of the superiority of GAT over the simple multimodal LSTM is that the interaction among nodes applies different weight over the information from different modalities, while the LSTM based method simply fuse them into a shared space. Besides, the

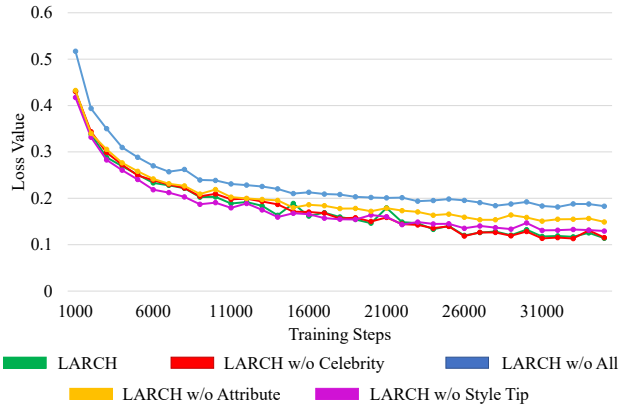


Fig. 7. The curves of loss values on the validation set of our augmented MMD 2.0 dataset.

short-cuts across different hierarchies may also help to reduce the lost of key semantic information. And 5) our LARCH model with VGG16 as the backbone still outperforms all the strong baselines in Table II, demonstrating the superiority of our method. Besides, when using the weighted average for nodes initialization, the performance of image search slightly drops, proving that the simple average initialization can reach a better accuracy.

2) Effect of Multiform Knowledge: To measure the contribution of different knowledge on the overall ranking quality, we conducted many ablation experiments, whereby the different forms of knowledge entries extracted from specific knowledge base in Eqn. (3) are removed one by one. It should be noted that, if all the three types of knowledge are removed, the projected image embedding and the conversational query $\mathbf{q}$ are directly used for ranking.

Table IV summarizes the results of LARCH and its variants of removing one specific form of knowledge. From the table, we have the following findings: 1) the model with all the knowledge (*i.e.,* LARCH) achieves the best performance under the setting of $K = 5$, while the model without any knowledge (*i.e.,* LARCH w/o all) performs the worst, verifying the importance of encoding multiform knowledge into image representation learning. Moreover, LARCH without considering knowledge still outperforms MAGIC, which validates the superiority of our proposed query representation learning method. 2) LARCH without the *style tip* slightly ties with LARCH, indicating that the *style tip* is possibly useless for many queries, yet introduces much noise instead. 3) The removal of the *celebrity* knowledge leads to the largest performance degradation, followed by that of the *attribute* knowledge. And 4) the performance gain brought by introducing knowledge drops as K increases, over both *precision* and *recall*, while the improvement over the *NDCG* is still stable. In general, the results show that with the help of multiform knowledge, most positive image candidates have been selected at higher ranks, especially when few top-ranked images are selected for evaluation.

Fig. 7 displays the curves of loss values of LARCH and its variants on our augmented MMD 2.0 dataset. All the models converge after around 35,000 steps. It can be concluded that the models introduced with specific form of knowledge often achieve lower loss value than that with no knowledge on the development set. However, the curve of loss cannot fully reflect the relative performance, which possibly is due to the discrepancy between the distribution of development set and test set.

(a) Case 1

(b) Case 2

Fig. 8. Two cases on the test set of our LARCH model and the baselines.

D. Hyper-parameter

In this section, we mainly investigated the effect of the number of GAT layers on query representation learning. Here, we explored the performance of LARCH when $L \in \{3, 4, 5, 6\}$.

As shown in Table V, LARCH with $L = 5$ achieves the best performance. When $L \in \{3, 4\}$, the ranking quality drops greatly, indicating that a 5-layer GAT model is able to sufficiently represent the complex conversational queries. Besides, the performance of LARCH with L=6 drops slightly in most cases, proving that more parameters might bring overfitting, and thus hurt the ranking quality.

E. Case Study

In this paper, we proposed a novel knowledge-enhanced image representation learning method. To intuitively understand the contribution of multiform knowledge, we randomly selected two cases with the top-10 ranked images recommended by the baselines and LARCH from the testing set of MMD 2.0.

TABLE V
EFFECT OF THE LAYER DEPTH OF GAT OVER OUR AUGMENTED DATASET MMD 2.0. THE SYMBOL * MEANS STATISTICALLY SIGNIFICANT IMPROVEMENT OVER OTHER RESULTS WITH $p < 0.05$.

	K = 5			K = 10			K = 20		
	Precision@5	Recall@5	NDCG@5	Precision@10	Recall@10	NDCG@10	Precision@20	Recall@20	NDCG@20
LARCH (5 layers)	**0.5501***	**0.6582***	**0.6829***	**0.2999***	**0.7161***	**0.7121***	0.1595	0.7620	**0.7302**
3 layers	0.4918	0.5911	0.6191	0.2816	0.6733	0.6610	0.1543	0.7370	0.6864
4 layers	0.5251	0.6292	0.6564	0.2910	0.6950	0.6898	0.1565	0.7470	0.7105
6 layers	0.5420	0.6503	0.6750	0.2980	0.7130	0.7068	**0.1596**	**0.7641**	0.7270

For the first case in Fig. 8(a), given the query "Can you show me a few heels which is of ten brand that I might like", the models with no knowledge (*i.e.,* UMD and LARCH w/o Knowledge) fail to predict the correct images since the *brand* information contained in the *attribute* knowledge is not available. On the contrary, the models with the *attribute* incorporated (*i.e.,* LARCH and MAGIC) are able to return the correct images.

As for the second case in Fig. 8(b), the baseline models focus more on either the vision features due to the absence of knowledge (*i.e.,* MAGIC, UMD and LARCH w/o Knowledge), or the semantic information (*i.e.,* LARCH w/o GAT), causing that either the correct images are not been predicted, or ranked at lower positions.

VI. CONCLUSION AND FUTURE WORK

In this paper, we present a novel contextual image search scheme, LARCH for short. It comprises three components, a multimodal hierarchical graph-based neural network for conversational query representation learning, an embedding memory network for multiform knowledge modeling, and a novel gated neural network for knowledge-enhanced image representation learning. To justify the effectiveness of our proposed LARCH model, we have constructed a new dataset towards conversational image search based upon the widely-used MMD benchmark dataset by adding more challenging image candidates for each query, named as MMD 2.0. Extensive experiments over this dataset have demonstrated the superiority of our LARCH model and each of its components.

The proposed LARCH takes the first step to verify the potential of knowledge-enhanced contextual image search on MMD 2.0. In the future, the performance of LARCH can be further improved from the following aspects: 1) incorporating unstructured knowledge or rules to boost image search performance, and 2) considering the attribute manipulation within images, such as changing the long sleeves into short ones.

REFERENCES

[1] Y. Gao, M. Wang, Z. Zha, J. Shen, X. Li, and X. Wu, "Visual-textual joint relevance learning for tag-based social image search," *IEEE Transactions on Image Processing*, vol. 22, no. 1, pp. 363–376, 2013.

[2] J. Cai, Z. Zha, M. Wang, S. Zhang, and Q. Tian, "An attribute-assisted reranking model for web image search," *IEEE Transactions on Image Processing*, vol. 24, no. 1, pp. 261–272, 2015.

[3] Y. Zhang, X. Yang, and T. Mei, "Image search reranking with query-dependent click-based relevance feedback," *IEEE Transactions on Image Processing*, vol. 23, no. 10, pp. 4448–4459, 2014.

[4] Z. Ji, Y. Pang, and X. Li, "Relevance preserving projection and ranking for web image search reranking," *IEEE Transactions on Image Processing*, vol. 24, no. 11, pp. 4137–4147, 2015.

[5] X. Yang, T. Mei, Y. Zhang, J. Liu, and S. Satoh, "Web image search re-ranking with click-based similarity and typicality," *IEEE Transactions on Image Processing*, vol. 25, no. 10, pp. 4617–4630, 2016.

[6] F. Radlinski and N. Craswell, "A theoretical framework for conversational search," in *Proceedings of the Conference on Human Information Interaction and Retrieval*. ACM, 2017, pp. 117–126.

[7] S. Feng, R. C. Gunasekara, S. Shashidhara, K. P. Fadnis, and L. C. Polymenakos, "A unified implicit dialog framework for conversational search," *CoRR*, vol. abs/1802.04358, 2018.

[8] L. Nie, W. Wang, R. Hong, M. Wang, and Q. Tian, "Multimodal dialog system: Generating responses via adaptive decoders," in *Proceedings of the International Conference on Multimedia*. ACM, 2019, pp. 1098–1106.

[9] C. Cui, W. Wang, X. Song, M. Huang, X. Xu, and L. Nie, "User attention-guided multimodal dialog systems," in *Proceedings of the International Conference on Research and Development in Information Retrieval*. ACM, 2019, pp. 445–454.

[10] W. Lei, G. Zhang, X. He, Y. Miao, X. Wang, L. Chen, and T. Chua, "Interactive path reasoning on graph for conversational recommendation," in *the Conference on Knowledge Discovery and Data Mining*. ACM, 2020, pp. 2073–2083.

[11] W. Lei, X. He, M. de Rijke, and T. Chua, "Conversational recommendation: Formulation, methods, and evaluation," in *Proceedings of the International Conference on Research and Development in Information Retrieval*. ACM, 2020, pp. 2425–2428.

[12] H. Zhang, M. Liu, Z. Gao, X. Lei, Y. Wang, and L. Nie, "Multimodal dialog system: Relational graph-based context-aware question understanding," in *Proceedings of the International Conference on Multimedia*. ACM, 2021.

[13] W. Wang, M. Huang, X. Xu, F. Shen, and L. Nie, "Chat more: Deepening and widening the chatting topic via A deep model," in *The International Conference on Research & Development in Information Retrieval*. ACM, 2018, pp. 255–264.

[14] T. Mei, Y. Rui, S. Li, and Q. Tian, "Multimedia search reranking: A literature survey," *ACM Computing Surveys*, vol. 46, no. 3, pp. 38:1–38:38, 2014.

[15] S. Wang, F. Jing, J. He, Q. Du, and L. Zhang, "Igroup: Presenting web image search results in semantic clusters," in *Proceedings of the Conference on Human Factors in Computing Systems*. ACM, 2007, pp. 587–596.

[16] J. Tang, Z. Li, M. Wang, and R. Zhao, "Neighborhood discriminant hashing for large-scale image retrieval," *IEEE Transactions on Image Processing*, vol. 24, no. 9, pp. 2827–2840, 2015.

[17] Y. Liu and T. Mei, "Optimizing visual search reranking via pairwise learning," *IEEE Transactions on Multimedia*, vol. 13, no. 2, pp. 280–291, 2011.

[18] Z. Li, J. Tang, and T. Mei, "Deep collaborative embedding for social image understanding," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 9, pp. 2070–2083, 2019.

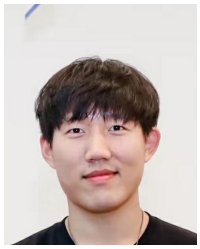
[19] Z. Li, J. Tang, L. Zhang, and J. Yang, "Weakly-supervised semantic guided hashing for social image retrieval," *International Journal of Computer Vision*, vol. 128, no. 8, pp. 2265–2278, 2020.

[20] X. Liu, Z. Li, Z. Shi, and Z. Shi, "Filter object categories: Employing visual consistency and semisupervised approach," in *Proceedings of the International Conference on Multimedia and Expo*. IEEE, 2009, pp. 678–681.

[21] L. Nie, M. Wang, Z. Zha, and T. Chua, "Oracle in image search: A content-based approach to performance prediction," *ACM Transactions on Information Systems*, vol. 30, no. 2, pp. 13:1–13:23, 2012.

[22] M. Wang, H. Li, D. Tao, K. Lu, and X. Wu, "Multimodal graph-based reranking for web image search," *IEEE Transactions on Image Processing*, vol. 21, no. 11, pp. 4649–4661, 2012. [Online]. Available: https://doi.org/10.1109/TIP.2012.2207397
[23] A. Natsev, A. Haubold, J. Tesic, L. Xie, and R. Yan, "Semantic concept-based query expansion and re-ranking for multimedia retrieval," in *Proceedings of the International Conference on Multimedia*. ACM, 2007, pp. 991–1000.
[24] H. Jégou, M. Douze, and C. Schmid, "Improving bag-of-features for large scale image search," *International Journal of Computer Vision*, vol. 87, no. 3, pp. 316–336, 2010.
[25] O. Chum, A. Mikulík, M. Perdoch, and J. Matas, "Total recall II: query expansion revisited," in *Proceedings of the Conference on Computer Vision and Pattern Recognition*. IEEE, 2011, pp. 889–896.
[26] X. Liu, L. Huang, C. Deng, B. Lang, and D. Tao, "Query-adaptive hash code ranking for large-scale multi-view visual search," *IEEE Transactions on Image Processing*, vol. 25, no. 10, pp. 4514–4524, 2016. [Online]. Available: https://doi.org/10.1109/TIP.2016.2593344
[27] Y. Liu, T. Mei, and X. Hua, "Crowdreranking: Exploring multiple search engines for visual search reranking," in *Proceedings of the International Conference on Research and Development in Information Retrieval*. ACM, 2009, pp. 500–507.
[28] Z. Zha, L. Yang, T. Mei, M. Wang, Z. Wang, T. Chua, and X. Hua, "Visual query suggestion: Towards capturing user intent in internet image search," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 6, no. 3, pp. 13:1–13:19, 2010.
[29] X. Tian, D. Tao, X. Hua, and X. Wu, "Active reranking for web image search," *IEEE Transactions on Image Processing*, vol. 19, no. 3, pp. 805–820, 2010.
[30] J. Wang and X. Hua, "Interactive image search by color map," *ACM Transactions on Intelligent Systems and Technology*, vol. 3, no. 1, pp. 12:1–12:23, 2011.
[31] K. Christakopoulou, F. Radlinski, and K. Hofmann, "Towards conversational recommender systems," in *Proceedings of the International Conference on Knowledge Discovery and Data Mining*. ACM, 2016, pp. 815–824.
[32] T. Kenter and M. de Rijke, "Attentive memory networks: Efficient machine reading for conversational search," *CoRR*, vol. 1712.07229, 2017.
[33] Y. Sun and Y. Zhang, "Conversational recommender system," in *Proceedings of the International Conference on Research and Development in Information Retrieval*. ACM, 2018, pp. 235–244.
[34] Y. Zhang, X. Chen, Q. Ai, L. Yang, and W. B. Croft, "Towards conversational search and recommendation: System ask, user respond," in *Proceedings of the International Conference on Information and Knowledge Management*. ACM, 2018, pp. 177–186.
[35] C. Qu, L. Yang, W. B. Croft, Y. Zhang, J. R. Trippas, and M. Qiu, "User intent prediction in information-seeking conversations," in *Proceedings of the Conference on Human Information Interaction and Retrieval*. ACM, 2019, pp. 25–33.
[36] S. Agarwal, O. Dusek, I. Konstas, and V. Rieser, "A knowledge-grounded multimodal search-based conversational agent," in *Proceedings of the International Workshop on Search-Oriented Conversational AI*. ACL, 2018, pp. 59–66.
[37] L. Yang, M. Qiu, C. Qu, J. Guo, Y. Zhang, W. B. Croft, J. Huang, and H. Chen, "Response ranking with deep matching networks and external knowledge in information-seeking conversation systems," in *Proceedings of the Conference on Research and Development in Information Retrieval*. ACM, 2018, pp. 245–254.
[38] M. Ghazvininejad, C. Brockett, M. Chang, B. Dolan, J. Gao, W. Yih, and M. Galley, "A knowledge-grounded neural conversation model," in *Proceedings of the Conference on Artificial Intelligence*. AAAI, 2018, pp. 5110–5117.
[39] J. R. Trippas, D. Spina, L. Cavedon, and M. Sanderson, "A conversational search transcription protocol and analysis," in *Proceedings of the International Workshop on Conversational Approaches to Information Retrieval*. ACM, 2017.
[40] D. Chen and C. D. Manning, "A fast and accurate dependency parser using neural networks," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. ACL, 2014, pp. 740–750.
[41] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the conference on computer vision and pattern recognition*. IEEE, 2016, pp. 770–778.
[42] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. ACL, 2014, pp. 1532–1543.
[43] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph attention networks," in *International Conference on Learning Representations*, 2018.
[44] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Annual Conference on Neural Information Processing Systems*, 2017, pp. 5998–6008.
[45] D. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and accurate deep network learning by exponential linear units (elus)," in *International Conference on Learning Representations*, Y. Bengio and Y. LeCun, Eds., 2016.
[46] A. H. Miller, A. Fisch, J. Dodge, A. Karimi, A. Bordes, and J. Weston, "Key-value memory networks for directly reading documents," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. ACL, 2016, pp. 1400–1409.
[47] A. Saha, M. M. Khapra, and K. Sankaranarayanan, "Towards building large scale multimodal domain-aware conversation systems," in *Proceedings of the Conference on Artificial Intelligence*. AAAI, 2018, pp. 696–704.
[48] Y. Sun and Y. Zhang, "Conversational recommender system," in *International Conference on Research and Development in Information Retrieval*. ACM, 2018, pp. 235–244.
[49] S. Rendle, "Factorization machines," in *International Conference on Data Mining*. IEEE, 2010, pp. 995–1000.
[50] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations*, 2015.
[51] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

Liqiang Nie (Senior Member, IEEE) is currently a professor with Shandong University and the dean with the Shandong AI institute. He received his B.Eng. and Ph.D. degree from Xi'an Jiaotong University and National University of Singapore (NUS), respectively. After PhD, Dr. Nie continued his research in NUS as a research follow for three years. His research interests lie primarily in multimedia computing and information retrieval. Dr. Nie has co-/authored more than 200 papers and 4 books, received more than 11,000 Google Scholar citations as of Aug 2021. He is an AE of IEEE TKDE, IEEE TMM, ACM ToMM, and Information Science. Meanwhile, he is an area chair of ACM MM 2018-2021. He has received many awards, like ACM MM and SIGIR best paper honorable mention in 2019, SIGMM rising star in 2020, TR35 China 2020, DAMO Academy Young Fellow in 2020 and SIGIR best student paper in 2021.



Fangkai Jiao is a Master student in the School of Computer Science and Technology, Shandong University. He received the B.E. degree from the School of Software, Shandong University, in 2019. His research interests include pre-training, question answering and dialogue system.

Wenjie Wang is a Ph.D. student in the School of Computing, National University of Singapore. He received the B.E. degree from the School of Computer Science and Technology, Shandong University in 2019. His research interests cover recommendation, data mining, and multimedia. He has around 10 publications appeared in several top conferences such as SIGIR, KDD, and ACM MM. Moreover, he has been served as the PC member and reviewer for several top conferences and journals including TKDE, TOIS, SIGIR, AAAI, MM, WSDM, ECML/PKDD.

Yinglong Wang is a full professor with Qilu University of Technology (Shandong Academy of Sciences). He is the president of the Shandong Internet of Things Association, and the director of the China-Australia International Health Technology Joint Laboratory. Dr. Wang Yinglong's main research areas are medical artificial intelligence and high-performance computing. In recent years, he has taken charge of more than 20 national, provincial, and ministerial projects. The scientific research projects led by him won 2 First Prizes, 4 Second Prizes, and 2 Third Prizes of Shandong Science and Technology Progress Award. He has published over 60 top academic papers and owns more than 20 authorized invention patents. Moreover, he organized the compilation of three volumes of national standards.

Qi Tian (Fellow, IEEE) received the B.E. degree in electronic engineering from Tsinghua University, the M.S. degree in ECE from Drexel University, and the Ph.D. degree in ECE from the University of Illinois at Urbana-Champaign (UIUC). He is currently a Chief Scientist in artificial intelligence at Cloud BU, Huawei. From 2018 to 2020, he was a Chief Scientist in computer vision at Huawei Noah's Ark Lab. He was also a Full Professor with the Department of Computer Science, The University of Texas at San Antonio (UTSA) from 2002 to 2019. His research interests include computer vision, multimedia information retrieval, and machine learning, and he has published over 550 refereed journal articles and conference papers.