Semantic Hilbert Space for Interactive Image Retrieval

Amit Kumar Jaiswal
University of Bedfordshire
Luton, United Kingdom
amit.jaiswal@study.beds.ac.uk
amitkumarj441@gmail.com

Haiming Liu
University of Bedfordshire
Luton, United Kingdom
haiming.liu@beds.ac.uk

Ingo Frommholz
University of Wolverhampton
Wolverhampton, United Kingdom
ifrommholz@acm.org

ABSTRACT

The paper introduces a model for interactive image retrieval utilising the geometrical framework of information retrieval (IR). We tackle the problem of image retrieval based on an expressive user information need in form of a textual-visual query, where a user is attempting to find an image similar to the picture in their mind during querying. The user information need is expressed using guided visual feedback based on Information Foraging which lets the user perception embed within the model via semantic Hilbert space (SHS). This framework is based on the mathematical formalism of quantum probabilities and aims to understand the relationship between user textual and image input, where the image in the input is considered a form of visual feedback. We propose SHS, a quantum-inspired approach where the textual-visual query is regarded analogously to a physical system that allows for modelling different system states and their dynamic changes thereof based on observations (such as queries, relevance judgements). We will be able to learn the input multimodal representation and relationships between textual-image queries for retrieving images. Our experiments are conducted on the MIT States and Fashion200k datasets that demonstrate the effectiveness of finding particular images autocratically when the user inputs are semantically expressive.

CCS CONCEPTS

• **Information systems** → **Users and interactive retrieval**; *Personalization*; **Image search**.

KEYWORDS

Image Search, Information Retrieval, Quantum Theory

ACM Reference Format:
Amit Kumar Jaiswal, Haiming Liu, and Ingo Frommholz. 2021. Semantic Hilbert Space for Interactive Image Retrieval. In *Proceedings of the 2021 ACM SIGIR International Conference on the Theory of Information Retrieval (ICTIR '21), July 11, 2021, Virtual Event, Canada.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3471158.3472253

ICTIR '21, July 11, 2021, Virtual Event, Canada

ACM ISBN 978-1-4503-8611-1/21/07…$15.00
https://doi.org/10.1145/3471158.3472253

1 INTRODUCTION

A Web searcher is the primary actor in the process of interaction with a search system. A specific challenge lies on how the search system caters the underlying user information need. Great progress has been made in modelling user information needs [4, 7, 16, 19, 23, 25, 29, 30, 41]. However, most of these previous works rely on textual and visual representations of information needs, and corresponding user information needs are either dynamic, evolving or fixed. The challenge of an image retrieval task is to delineate a user information need (as expressed by the query) in a way that captures the user's perception at a granular level.

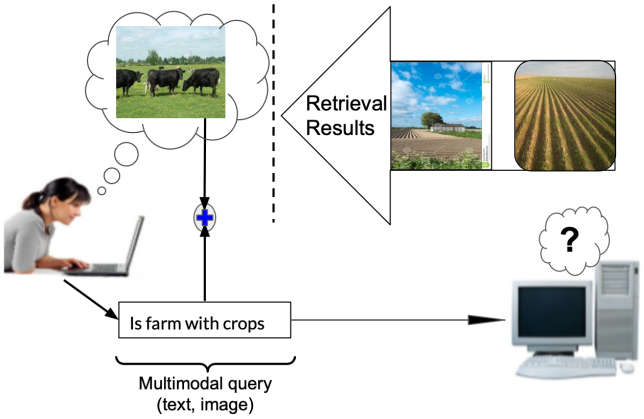

Figure 1: An illustration of the expressive user information need in an image retrieval scenario. The input consists of a textual query which expresses the required context for the input still image query. The output is a set of retrieved results that are very similar to the subjected input query. The retrieval example is from the MIT States [14] dataset.

A typical instance of an image retrieval system takes a reference (or query) image and then performs a kind of matching to return a list of candidate images identical to the input query. In general, such kind of image retrieval system tends to limit support for users in expressing their information need because it demands users to manifest without vagueness what they have in mind via a single image query. Practically, it is difficult in most of the cases for a user to convey their perception or intention via an image query.

In this paper, we address an interactive image retrieval task inspired by [36]. In our task setting, the input query consists of a text and an image described as intra-feedback[1]. *Intra-feedback* refers to the grounded context of a still image, so called, *guided visual*

[1]Intra-feedback is different from intra-query [40]

feedback as opposed to [40] which considers the entire image. The motivation for intra-feedback stems from [23], which allows to encode the cognitive aspect of a user's information need using a still image and the guided visual feedback follows a semantic matching at patch-level[2]. The guided visual feedback approach uses Information Foraging Theory [27] to encode contextual explanations via the patch selection mechanism [15]. An overview of our task is shown in Fig. 1.

Our work is inspired by the recent trends in deep complex networks [21, 33] referred to as complex-valued neural networks. These deep neural nets (DNN) have shown to have better representational capability than real-valued DNNs. [37] developed a complex-valued DNN-based framework that shows the phase components to be of importance for words that combine to form a sentence. Such representation frameworks are based on the existing embark [2, 3, 5] in understanding user thought processes by means of the Hilbert space formalism [34], which forms the basis of our work. We investigate the question whether an expressive multimodal representation of user's information need benefits from such approaches in a more common setting for an image retrieval task. Specifically, we seek to understand if we can learn textual-visual representations based on a complex-valued vector embedding as opposed to the standard joint vector embedding [39]. We tackle this by presenting a complex-valued embedding approach for learning the multimodal query. The input multimodal representation consists of text and image features extracted using pre-trained BERT [6] and ResNet [11] models. We assume based on [38] that combining or joining different modality feature vectors can not be performed directly due to their different statistical properties of extracted text and image features from varied deep neural networks. Also, their representation spaces are different as opposed to joint text-image embeddings [39]. Therefore, we do not combine the input text-image query during training; our model performs a projection of image feature vectors to textual feature vectors via an inner product and learns this multimodal representation.

To this end, we propose a new framework called *Semantic Hilbert space* (SHS). The SHS represents a common complex-valued feature space, which inherently maps the input image feature (driven by intra-feedback) to the target image feature given both the source and target image features are in a common Hilbert space. We assume that the input query image and target image are *projective transformations* of each other in a complex-valued Hilbert space. The transformation is symmetric i.e., the textual features which assess the transformation among the source image and target image and the target image features with the textual feature can re-assess the input image features. This kind of linear transformation makes the user information need more expressive as it integrates the textual information pertaining to the input image in the semantic Hilbert space. It leads to an importance of textual features in such a linear transformation and therefore, we enhance the learning metric of such projective transformation (PT) among different modalities features via PT symmetry. The PT symmetry allows to generate the representation alike input query image features from the product of complex conjugate of the input query features with the target image features. Moreover, the proposed semantic Hilbert space is a generalisation of classical approaches inspired by quantum probabilistic frameworks [28, 29, 32, 34], which treats the cognitive and complex aspects of user's information need components (text and image query) [7, 13, 20, 37]. We have performed experiments on the Fashion200k [10] and MIT States [14] datasets to evaluate our proposed SHS model, including comparisons with several baselines. Our proposed method is superior to the state-of-the-art cross-retrieval model [36] which demonstrates the effectiveness of the SHS.

Our main contributions are as follows:

(1) We propose a representation framework, so called an *semantic Hilbert space* to model multimodal (text-image) information need in an interactive image retrieval task.
(2) We present a projective transformation that inherently maps the input image feature space to target image feature space via complex-valued text encoding.
(3) The proposed method SHS is evaluated on two benchmark datasets i.e., MIT States and Fashion200k and is superior to other state-of-the-art methods.

2 RELATED WORK

This section reviews various existing works and methods relevant to image retrieval and multimodal representation learning. Also, we highlight the differences between these methods and tasks.

2.1 Image Retrieval

The image retrieval task varies based on the type of input query. However, several modern image retrieval systems [18] follow the ongoing trend of techniques that use deep learning. A general image retrieval system allow users to query it via a single image. This type of retrieval system belongs to content-based image retrieval (CBIR) and its application has been explored in the task of understanding user interaction using Information Foraging [22]. Studies in interactive image search [18] adopted attribute feedback such as "skateboard is blue" empowered by effective feedback techniques but the model exhibits uncertainty in the search system. Another similar line of research [19] developed a interactive lifelog search engine that uses text/image as a query and supports user feedback to find images. However, user feedback still has its limit shown in the performance of the image retrieval system due to the prior indexing and feature extraction. In such kind of typical image retrieval system, the users' flexibility is limited and they find it difficult to express their information need in order for the system to retrieve effectively. To support such information needs (queries), systems need to add more contextual components (such as image with text or vice-versa). An expressible information need can be explored in a multimodal representation learning due to the typical unambiguous text query which tends to be semantic and heuristic in the way of natural languages and the need of implicit and contextual information [40] can be of much significant role.

2.2 Multimodal Representation Learning

Multimodal representation approaches in image retrieval have gained momentum due to their applicability to more practical scenarios such as product retrieval [10]. Multimodal representation learning approaches learn the latent embeddings that associate

[2]Here, patch (from Information Foraging Theory) refers to image regions

each (text, image) pair to a dense vector, which represents a point in a d-dimensional space. However, such a point fails to explicitly encode uncertainty, for instance, a word "date" can have several meanings i.e., *to eat a date*, *out on a date*, and *event date*. A few recent works have attempted to capture word or word-meaning polarity based on complex/probabilistic word embeddings [21, 37]. Our work addresses a similar scenario to capture the contextual information from an input image so as to encode the implicit information via text, but in an image retrieval task. Also, a multimodal representation has been explored in web image search [40], where the authors developed a context-aware re-ranking model for one-to-one mapping of textual queries and image queries to a dense vector representation in an embedding space to model user preferences. [39] presented a joint-space embedding approach that learns both the textual and image features in a common space of identical dimension. This work demonstrates its method on two tasks i.e., image retrieval and sentence retrieval. Recently [36] developed a multimodal learning approach for image retrieval and it stands as the first work for such kind of method. However, we think their method exploits the importance of image features in different spaces and less the importance to textual features. We present these differences in the analysis part of this paper. Another work addressing a cross-modal retrieval task [9] proposes FashionBERT, which performs multitask learning. This method presents a patch-level alignment i.e., image patches as tokens extracted from fashion images. A few joint representation approaches [1, 17, 39] adopted a learning mechanism of visual-semantic embedding by measuring the distance among an input linguistic query and a target image within a common feature space. However, these approaches are not suitable enough for retrieving very generic images, because users either may not have the complete picture in mind already or find it difficult to express their information need sufficiently enough via a single query.

Our work investigates how to learn a textual-visual query that can coherently express users' information needs based on the quantum probabilistic framework [29, 34], in particular, the Hilbert space formalism. The input representation is mapped to the target image representation via the textual input which encodes the contextual information of the input still image; this encoding process is complex in nature. Therefore, we restrict the requirement of a linear transformation as contrast to affine transformation [42] due to the fact that the performance of the former transformation is superior over the latter method. We also analyse the effectiveness of the PT symmetry. The contextual encoding process from an still image follows the patch-level and text-level correspondence delineated by Information Foraging Theory.

3 MATHEMATICAL PRELIMINARIES

Our idea is inspired by the view of an IR system as analogous to a quantum system used in physics. In this section, we overview the mathematical formalism [34] behind the quantum-theoretic models [8, 28] of IR. In contrast to set-based classical probability Theory, in quantum probabilities, the probabilistic space is geometrically defined, and its representation becomes an infinite set of angles and distances in a finite or infinite dimensional Hilbert Space [32] denoted by $\mathbb{H}$. A vector space with complete inner product forms a Hilbert space. Each and every event is a subspace of the Hilbert space. To represent the n-dimensional vectors that compose a Hilbert space, the Dirac notation is widely adopted, using *ket* and *bra* nomenclatures. More concretely, this means representing one given vector ψ as $|\psi\rangle$ and its transposed view, ψ^T as $\langle\psi|$. Also, the vectors under consideration in a Hilbert space are called *state vectors*, which are unit vectors. The squared length of the orthogonal projection of a state vector onto subspaces representing events (denoted by the projector $|\psi\rangle\langle\psi|$) induces a probability distribution over these quantum events. In a quantum system there can be more than one state vector; a probability distribution over state vectors (which is different from the probability distributions induced by every state vector) reflects the uncertainty about the state the system might be in. The unit vectors-induced probability distributions over events (subspaces) can be represented by a so-called *density matrix*, denoted by ρ. Each state vector is a linear combination of the eigenvectors spanning the subspace. Each quantum event has its quantum probabilities defined by density matrices. In linear transformation, a general geometrical transformation method that maps an n-dimensional feature space to m-dimensional feature space given $n \geq m$, where points are homogeneous complex coordinates. Such linear transformation is known as projective transformation $P_{\mathbb{C}}$. In this paper, we deal with a Hilbert space defined over the complex field i.e., a complex plane consists of complex numbers. The input image query and target image are in a common Hilbert space. Also, the projective transformation allows the textual query as part of the input query to be in the same common Hilbert space as it is used to determine the angle of rotation.

4 SEMANTIC HILBERT SPACE

We present a framework to uniquely learn to represent the input textual feature and visual feature space, and map the source image feature space to the target image feature space via textual features in a common complex-valued space, referred to as *Semantic Hilbert space* $\mathcal{H}$. This framework depicts representations to capture high-level "semantic interaction" and serve discriminative features of multimodal components (text and image) equally in a common space i.e., Hilbert space. An overview of the framework is shown in Fig. 2.

4.1 Problem Constructs and Multimodal Representation

We employ the standard mathematical notations described in Section 3. Let a set of textual queries $\mathcal{Q} = \{|q_1\rangle, |q_2\rangle, \ldots, |q_n\rangle\}$, a set of image queries $\mathcal{I} = \{|i_1\rangle, |i_2\rangle, \ldots, |i_n\rangle\}$ and a set of target image queries $\mathcal{R} = \{|r_1\rangle, |r_2\rangle, \ldots, |r_n\rangle\}$, where $|q_i\rangle$, $|i_i\rangle$, and $|r_i\rangle$ depicts the state vector of input textual query, input visual query, and retrieved target image. These state vectors[3] can be treated as a *ray* in the semantic Hilbert space i.e., $|Q\rangle, |I\rangle, |R\rangle \in \mathcal{H}$. We use pre-trained embedding models for feature extraction i.e., BERT embedding model [6] for textual features and ResNet-34 [11] for image features. An image pre-trained embedding model $\mathcal{E}_p(.)$ extracts

[3]The state vector representations are for conciseness i.e., $|Q\rangle$ can be written as a column vector $\vec{Q}$. State vectors notation follows Dirac nomenclature throughout in this paper.

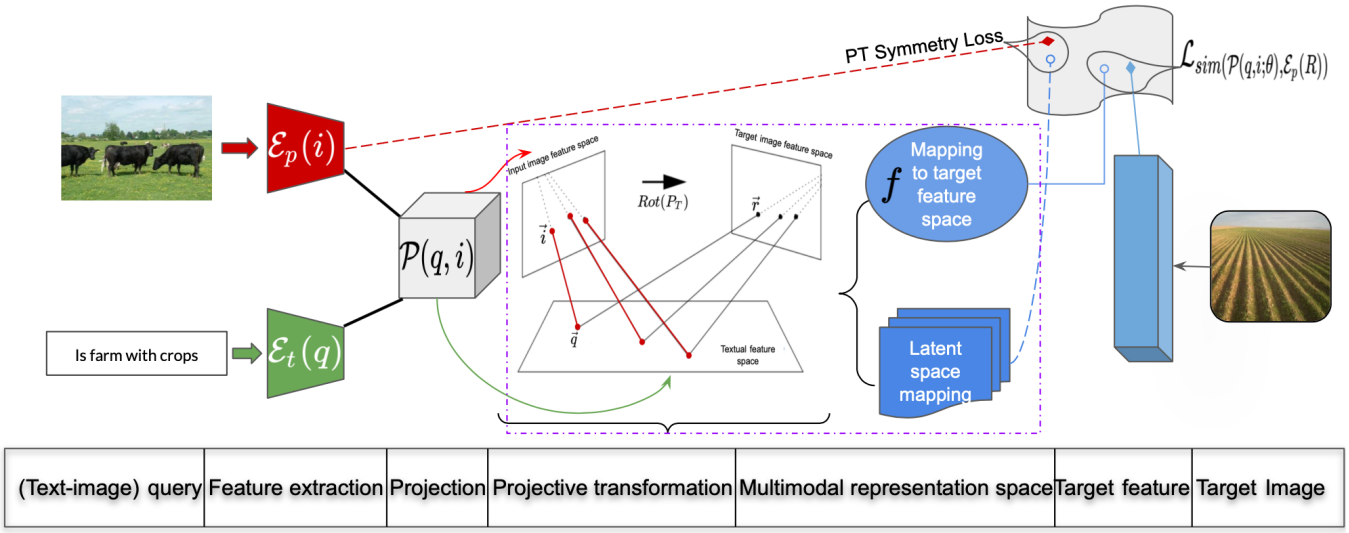

Figure 2: The general framework of the proposed semantic Hilbert space. The input text and image query are fed into a pre-trained BERT embedding model [6] $\mathcal{E}_p(q)$ and a pre-trained ResNet34 [11] $\mathcal{E}_p(i)$ one, respectively. Then, a modality distribution is realised via the projection between the image features and text features. Finally, a projective transformation of the input image to the target image features in a common complex-valued Hilbert space is performed to prepare the multimodal representation, which we map in two stages. Firstly, a mapping function (f) associates the input image features to the target feature space. Secondly, the visual guided feedback maps the multimodal feature space to the target feature space, and this step is referred to as *latent space mapping*. Then, the training loss is estimated via PT symmetry and similarity measure learning. These two mapping functions inherently capture the shared multimodal representation, which leads to retrieve the most closest image.

image features in a k-dimensional space. We use these embeddings to generate projectors in a common semantic space i.e., the projection operation $\mathcal{P}(q,i)$ of the input image $|i\rangle$ onto $|q\rangle$ is the inner product between these two and can be equated as

$$\mathcal{P}(q,i) = \langle q|i\rangle\,.$$

Then, the first step is to maximise the below objective function in which the projector learns the textual-visual query representation formulated as

$$\max_{\theta} sim(\mathcal{P}(q,i;\theta),\mathcal{E}_p(R)). \tag{1}$$

where sim is a measure of similarity and θ depicts the learnable parameters. A better input representation (i.e., information need representation) enhances the learning task and in Eq. 1, the maximisation of the similarity measure is among the projection function output of the multimodal input query and the features of the target image. The similarity measure is not restricted to model only the multimodal input query but also the target image in a common Hilbert space. This way of modelling gives importance to both the text and image query. This is in contrast to [36] which defines the learning mechanism between query image and target image, and relaxes the textual features in the learning process. Our assumption to maximise the objective function (Eq. 1) are as follows:

- The image features of the input image query and target image are in a common space.
- The transformation (from input image query to target image) encoding uses textual features in the same common space.

Based on the following assumption, we argue that the input image query and target image are projective transformations of each other in a complex-valued Hilbert space. The intuition behind choosing such kind of linear transformation is based on [42]. The textual features are used to compute the projective transformation. The need of complex space is for the PT to be symmetric; the textual features, when applying complex conjugate on them, can make such linear transformation symmetric. We arrive at

$$P_T(\overrightarrow{i}) \underset{\overrightarrow{q}}{\longrightarrow} \overrightarrow{r} \implies P_T(\overrightarrow{r}) \underset{\overrightarrow{\overline{q}}}{\longrightarrow} \overrightarrow{i} \tag{2}$$

where P_T represents the projective transformation and $\overrightarrow{\overline{q}}$ represents the complex conjugate of the textual query. Eq. 2 describes the projective transformation symmetry in which an input image query on such transformation from learned textual features results in the target image. Conversely, applying the complex conjugate on the textual features can transform the source image back to the original input query image. We describe the PT symmetry below in the training steps of our multimodal input representation. Also, we analyse the effectiveness of this learning metric in different settings.

The PT symmetry learning metric benefits from the textual feature as it help determine the angle of such linear transformation. Such metric as a regularization can be an advantageous in terms of generalisation [24, 33].

4.2 Model

We have prepared the learning constructs based on different feature extraction techniques for textual-visual queries. We now incorporate them into complex-valued embedding approach, which is an extension of [21, 37].

4.2.1 Feature Embedding. We generate the feature vector $|i_f\rangle$ of an input image query i via $\mathcal{E}_p(.)$ in a k-dimensional space i.e., $\mathcal{E}_p(i) = |i_f\rangle \in \mathbb{R}^k$. We extract text features through the BERT [6] embedding model $\mathcal{E}_t(.)$ to generate text feature vectors $|q_f\rangle$ in an l-dimensional space i.e., $\mathcal{E}_t(q) = |q_f\rangle \in \mathbb{R}^l$.

Both text feature vector and image feature vector are extracted from different embedding models and exhibit different statistical properties. Therefore, we perform a projection operation $\mathcal{P}(q, i)$ from an input image query to the text query. The next step is letting text features encode the input image information for its transformation to the target image. To implement this, we follow our assumption in Eq. 2 based on the projective transformation. In particular, the projection of an input image query (features) to a target image in the complex space is done by a linear transformation. The text query features elicit the required information for the projective transformation of the input query image to the target image. This mapping of the query image and target image in a common space[4] during training is learned via $|i_f\rangle$ and $|q_f\rangle$ in $\mathbb{C}^d$. Also, the angle of rotation during the projective transformation is learned via the textual features. It can be formulated as:

$$\begin{aligned} M : \mathbb{R}^d &\longrightarrow M_D \in \mathbb{R}^{dxd} \\ Rot(P_T) &= e^{\imath M(|q_f\rangle)} \end{aligned} \tag{3}$$

where M is a mapping function constructed using two fully connected layers with non-linear activation, M_D is a matrix diagonal, $Rot(P_T)$ depicts the rotation operation of projective transformation and $\imath = \sqrt{-1}$ depicts the imaginary number *iota*. Then, the network learns the mapping of input image features (i_f) to the complex-valued space. This mapping function is also implemented like M with fully connected layers.

$$\begin{aligned} M_I : \mathbb{R}^k &\longrightarrow \mathbb{C}^d \\ I_M &= Rot(P_T) M_I(|i_f\rangle) \end{aligned} \tag{4}$$

where M_I is the image mapping function and I_M represents the multimodal representation of the input query.

Then, from Eq. 1, the goal is to maximise the measure of similarity among the input multimodal features and features of the target image. Therefore, the network requires the mapping function to learn from $f : \mathbb{C}^d \longrightarrow \mathbb{R}^k$, where f is implemented using fully connected layers with non-linear activation. The mapping function associates the features from the complex space to the k-dimensional real-valued space where the extracted features of target image reside. We construct another mapping function f_l for *visual guided feedback* that is made up of two fully connected layers including a single convolutional layer, which is to capture the textual-visual similarity pattern from the data. It allows to benefit learning from local features distributed across different modalities of features. The mapping functions I_M and f_l also make use of the extracted textual (q_f) and visual features (i_f) as an input. The importance of visual guided feedback can be for queries that are evolving [8, 22], e.g. a user seeks for a jacket with a specific logo on the left-front.

Consider a function $g(i_f, q_f)$ to delineate the input multimodal representation that learns both the textual and visual features for image retrieval. This function can be formulated as:

$$g(i_f, q_f) = \alpha f(I_M) + \beta f_l(I_M, i_f, q_f) \tag{5}$$

where α and β are learnable parameters.

The modelling strategy for learning multimodal representation given in Eq. 5 follows [21, 32], which is a complex-valued embedding network. We extend these networks for learning a multimodal representation for our image retrieval task. Having the extracted features of text and image based on the pre-trained text encoder BERT and image encoder ResNet, we use these features to learn a multimodal representation in a d-dimensional complex space. A complex-valued convolutional neural network (CNN) is built to learn these representations, where the network consists of an encoder and a decoder. The complex-valued encoder network learns the multimodal representation (Eq. 5) and the complex-valued decoder generates the extracted textual ($|q_f\rangle$) and visual feature ($|i_f\rangle$) vectors back from the representation function $g(i_f, q_f)$. We adopt such a strategy to promote the image-pixel level features and to incorporate the learning of contextual information embedding within a still image. The other benefits are generalisation [24, 33] and a better representational power [37].

The complex-valued CNN network based on the above formulation contains four components which can be described as follows:

Input layer The input layer consist of a (text, image) feature vector i.e., (q_f, i_f). Each one of the features (q_f or i_f) will be represented by a state vector.

Encoder network The encoder network for learning the multimodal representation function $g(i_f, q_f)$ (Eq. 5) extracts characteristic features to semantically map the input image to the most similar images (target image). For each text-image pair, the convolution can be defined based on [33], where an input is $\mathbb{I} = I + \imath Q$. The $\mathbb{I}$ depicts the multimodal input matrix in which the image is realised as the real-part and the word is initialised as a complex entity. The reason behind considering words as part of complex entity is that it encodes the meaning when multiple word combined and refer to a new meaning, and it has been used earlier in text classification tasks [21, 37]. Another reason is that encoding the context of still images and textual queries as complex entity not only enriches the contextual information from images but also encodes such information in locating similar image for the image retrieval task. The complex-valued representation follows Eq. 2. Then, a matrix for convolutional filter is $W = A + iB$, where A and B are real-valued matrices and W represents the size of a convolutional kernel. The main idea behind this encoder layer is to simulate a complex entity via real-valued entities. Now, the convolution layer

[4] d-dimensional complex space with $d = k + l$

can be written as

$$\sum_{f=1}^{F}(I_f.W_m) = \sum_{f=1}^{F}\mathrm{Re}(I_f).\mathrm{Re}(W_m) - \mathrm{Im}(I_f).\mathrm{Im}(W_m)) + j\sum_{f=1}^{F}(\mathrm{Re}(I_f).\mathrm{Im}(W_m) + \mathrm{Im}(I_f).\mathrm{Re}(W_m))$$

where Re, Im are real-part and imaginary part of the complex entity, W_m depicts the matrix form of convolution kernel and '.' depicts the convolutional operator. Our convolutional layer is of 3×3 convolution and 64 convolutional filters. The encoder architecture is made of a fully connected layer.

Decoder network The decoder network (D_i and D_q for image and text decoder) generates back the original extracted features of text and image from the multimodal representation function (Eq. 5). The decoder layer first performs up-pooling and then deconvolution. The up-pooling step maps the encoded features by using the location information generated from the pooling process. The second step of deconvolution re-generates the actual input text and image features from the generated sparse representation of both text and image using up-pooling.

Output The output ranks images by similarity.

4.2.2 Model Training. MIT States [14] is used to train our proposed model. We use triplet ranking loss based on [39], which is defined as:

$$\mathcal{L}_{triplet} = \frac{1}{S \times n_{triplet}} \sum_{tr=1}^{n_{triplet}} \sum_{s=1}^{S} \log\left(1 + e^{\mathcal{P}(g(i_f,q_f),\mathcal{E}_p(\widetilde{R}_{tr,s})) - \mathcal{P}(g(i_f,q_f),\mathcal{E}_p(R_s))}\right) \quad (6)$$

where S represents the batch size for training sample s, the value in the exponent term represents a pair of multimodal features and dissimilar image feature (shown in $\mathcal{E}(\widetilde{R}_{tr,s})$), and similarly, the second term with leading negative depicts a pair of multimodal features and the target image features. Each training sample s requires a fixed number of triplets $n_{triplet}$. We select the same number of triplets i.e., 3 as given in [36].

We employ the softmax loss similar to one used in [36] for training the model on Fashion200k [10]. This loss function can be defined as:

$$\mathcal{L}_{softmax} = \frac{1}{S}\sum_{s=1}^{S} -log\left\{\frac{e^{\mathcal{P}(g(i_f,q_f),\mathcal{E}_p(R_s))}}{\sum_{b=1}^{S} e^{\mathcal{P}(g(i_f,q_f),\mathcal{E}_p(R_b))}}\right\}. \quad (7)$$

This batch classification based loss function performs normalisation among the multimodal features ($g(i_f, q_f)$) and the features of the target image divided by the summation of similarities between $g(i_f, q_f)$ and the sample image collection in a batch (j).

The loss function for training the complex-valued CNN based decoder is a L2 regularizer and it can be defined as $\mathcal{L}_Q$ and $\mathcal{L}_I$ for corresponding query reconstruction and image reconstruction

$$\mathcal{L}_{CE} = \frac{1}{S}\sum_{s=1}^{S} \|q_{f_s} - \hat{q}_{f_s}\|_2$$
$$\mathcal{L}_{CD} = \frac{1}{S}\sum_{s=1}^{S} \|i_{f_s} - \hat{i}_{f_s}\|_2 \quad (8)$$

where $\|.\|_2$ depicts the L2 norm, $\hat{q}_{f_s} = D_q(g(i_f, q_f))$ and $\hat{i}_{f_s} = D_i(g(i_f, q_f))$. The reconstruction loss is crafted as L2 norm.

4.2.3 Projective Transformation (PT) Symmetry. We present the projective transformation of an input image query to the target image in a common space, where the angle of rotation for this transformation is estimated via textual features (Eq. 3) formulated in Section 4.2.

Similarly, for symmetry, we solicit if the product of target visual features with the complex conjugate of textual features is similar to the image query features. To do so, applying the operation of complex conjugate on the textual features in a complex space can be formulated as $\overline{Rot(P_T)} = e^{-\iota M(q_f)}$. Then, this joint features in the complex space form a multimodal representation for target images which are given by $\hat{I}_M = \overline{Rot(P_T)} M_I \mathcal{E}_p(R)$. This can be crafted to compute the multimodal representation formulated as

$$g(\mathcal{E}_p(R), q_f) = \alpha f(\widetilde{I}_M) + \beta f_l(\widetilde{I}_M, \mathcal{E}_p(R), q_f). \quad (9)$$

The above multimodal representation and the input image features from Eq. 5 refers to formulate the PT symmetry. The PT symmetry aims to maximise the similarity function $sim(g(\mathcal{E}_p(R), q_f), i_f)$

The training loss function using the PT learning metric ($\mathcal{L}_{PT}$) for the Fashion200k dataset is computed by replacing the similarity function given in softmax loss in Eq. 7 by Eq. 9. It can hence be equated as

$$\mathcal{L}_{PT_{Fashion200k}} = \frac{1}{S}\sum_{s=1}^{S} -log\left\{\frac{e^{\mathcal{P}(g(\mathcal{E}_p(R),q_f),i_{f_s})}}{\sum_{b=1}^{S} e^{\mathcal{P}(g(\mathcal{E}_p(R),q_f),i_{f_b})}}\right\}. \quad (10)$$

Similarly, for the MIT States dataset, the training loss function follows from Eq. 6 where replacing the similarity function by Equation 9 gives us

$$\mathcal{L}_{PT_{MITStates}} = \frac{1}{S \times n_{triplet}} \sum_{tr=1}^{n_{triplet}} \sum_{s=1}^{S} \log\left(1 + e^{\mathcal{P}(g(\mathcal{E}_p(R),q_f),\widetilde{i_{f_{tr,s}}}) - \mathcal{P}(g(\mathcal{E}_p(R),q_f),i_{f_s})}\right). \quad (11)$$

where $\widetilde{i_{f_{tr,s}}}$ represents the dissimilar image features from the training batch size of $n_triplet$.

The corresponding PT symmetry loss for the MIT States and Fashion200k datasets is a uniform loss function employed in training the proposed model, provided it depends on the datasets it trained upon. We choose triplet ranking loss due to its nature of output being probabilistic, whereas softmax loss caters identifying the output image classes, as well as it can assign probabilities to those classes.

5 EXPERIMENT

5.1 Dataset

We conduct experiments on the MIT States and Fashion200k datasets to validate the effectiveness of the proposed model.

MIT States [14] consists of images collected from Bing search. It contains 63,440 images representing 245 nouns altered by approximately 115 adjectives. This collection of images are human-annotated for the sake of data quality. There are certain ambiguous and mislabeled images. The number of images in the training and test set are 43,207 and 10,546, respectively, after removal of mislabeled and unclear images. It contains 82,732 textual queries. Fashion200k [10] is a large collection of five different clothes categories such as dresses, jackets, pants, skirts and tops. It contains of 201,838 images. Also, this image collection is human-annotated. The training and test set contains 172,049 images and 29,789 images, respectively. The number of unique textual queries in the training set is 53,099 and the test set contains 10,332 queries.

5.2 Evaluation

For an interactive image retrieval task, we follow the evaluation metric used in [36, 43] to ensure comparability. We adopt *Recall@K* (R@K) to evaluate the retrieval performance. It is the fraction of queries for which the correct image is retrieved among the closest K points to the query.

Training Setting: We adopt most of the training settings from [36] except the pre-trained image feature extractor, where we employ ResNet34 [11]. The dimension of textual feature vectors are of 768 which is extracted using BERT embedding model [6]. The batch size for training is 64.

6 RESULTS

We report a comparison performance of our proposed model among several existing and created baselines in Table 1. Our model outperforms all other methods and baselines except for the composed query [12] model at R@1 which stand best among all of the models. We also compute the test accuracy of our SHS model which is 93.44%.

Model	MIT States	Fashion200k
SHS	48.2	55.6
(+) Real space	46.0	49.2
(-) $\mathcal{L}_{PT}$	47.9	52.4
(-) visual guided feedback (f_l)	46.9	53.0
(-) image mapping (f)	45.4	52.6

Table 2: Ablation test on both MIT States and Fashion200k datasets. (+) and (-) depicts with and without the respective components to the proposed SHS model.

The loss function of PT symmetry for the corresponding datasets follows Eq. 10 and Eq. 11.

Ablation Test: We analyse the effect and impact of different sub-components of the SHS model. The ablation test is reported in Table 2. Based on the analysis of different components, we found that the SHS model achieves better performance among all other models. This analysis verifies that modelling both text and image query via a modality distribution realised via projection delineates the implicit contextual information of a still image as opposed to the fusing/concatenating these two modalities. The representation power of our approach leads to effective image retrieval. We can observe in Table 2 that without including the loss function of projective transformation ($\mathcal{L}_{PT}$), the performance slightly decreases, which tells that the complex-valued representation captures the discriminant features of both modalities, and it leads to a very minor difference among the best performed model. Another instance of the analysis is that we concatenate the real space with SHS in a complex space, and it decreases the performance significantly. We can say that mapping the extracted textual and visual features into a common complex-valued space leads to a better capturing of information and encoding as opposed to the typical concatenation in the real space.

Qualitative Examples: We also report a set of image retrieval instances to validate our proposed model via qualitative examples. Fig. 3 and Fig. 4 are the reported qualitative examples from both the MIT States and Fashion200k datasets. The input consists of a query image and textual query and the outputs are a set of retrieved images from the test set.

7 CONCLUSION

In this paper, we propose our SHS approach, which could capture the implicit contextual information between an image and text (describing a multimodal information need), and effectuate the image retrieval task. The proposed model extends the classical way of representing query and images by means of modality distribution i.e., via a projection, which leverages the projective transformation in a complex-valued Hilbert space to delineate the encoding of still images via textual features. The main idea is to realise a multimodal representation via a complex-valued CNN to enhance the image retrieval task. The experimental results on both MIT States and Fashion200k show that our approach outperforms a number of state-of-the-art cross-modal retrieval methods, and also proves the importance of multimodal representations to represent users' information needs.

ACKNOWLEDGMENTS

This work was carried out in the context of Quantum Access and Retrieval Theory (QUARTZ) project, which has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 721321.

REFERENCES

[1] Omri Abend, Tom Kwiatkowski, Nathaniel J Smith, Sharon Goldwater, and Mark Steedman. 2017. Bootstrapping language acquisition. *Cognition* 164 (2017), 116–143.

[2] Diederik Aerts. 2014. Quantum theory and human perception of the macro-world. *Frontiers in Psychology* 5 (2014), 554.

[3] Diederik Aerts, Liane Gabora, and Sandro Sozzo. 2013. Concepts and their dynamics: A quantum-theoretic modeling of human thought. *Topics in Cognitive Science* 5, 4 (2013), 737–772.

Model \ Dataset	MIT States			Fashion200k		
	Metrics - Recall@K					
	K=1	K=5	K=10	K=1	K=10	K=50
Show and Tell [35]	$11.9^{\pm 0.2}$	$31.0^{\pm 0.5}$	$42.0^{\pm 0.8}$	$12.3^{\pm 1.1}$	$40.2^{\pm 1.7}$	$61.8^{\pm 0.9}$
Relation Network [31]	$12.3^{\pm 0.5}$	$31.9^{\pm 0.7}$	$42.9^{\pm 0.9}$	$13.0^{\pm 0.6}$	$40.5^{\pm 0.7}$	$62.4^{\pm 0.6}$
Film [26]	$10.1^{\pm 0.3}$	$27.7^{\pm 0.7}$	$42.9^{\pm 0.9}$	$12.9^{\pm 0.7}$	$39.5^{\pm 2.1}$	$61.9^{\pm 1.9}$
TIRG [36]	$12.2^{\pm 0.4}$	$31.9^{\pm 0.3}$	$41.3^{\pm 0.3}$	$14.01^{\pm 0.6}$	$42.5^{\pm 0.7}$	$63.8^{\pm 0.8}$
(+) BERT	$12.6^{\pm 1.0}$	$31.6^{\pm 1.0}$	$43.1^{\pm 0.3}$	$15.2^{\pm 0.4}$	$43.4^{\pm 0.2}$	$63.8^{\pm 1.2}$
Composed Query [12]	$14.29^{\pm 0.6}$	$34.67^{\pm 0.7}$	$46.6^{\pm 0.6}$	$16.26^{\pm 0.6}$	$46.90^{\pm 0.3}$	$71.73^{\pm 0.6}$
SHS (Ours)	$14.2^{\pm 0.6}$	$\mathbf{36.4}^{\pm 0.1}$	$\mathbf{48.2}^{\pm 0.3}$	$\mathbf{23.2}^{\pm 0.4}$	$\mathbf{55.6}^{\pm 1.0}$	$\mathbf{74.2}^{\pm 0.6}$

Table 1: Performance comparison with baselines. (+) depicts the BERT model added to the TIRG for a new baseline. The best performances are in boldface.

Figure 3: Some qualitative examples of our proposed model. The examples are from MIT States dataset.

Figure 4: Some qualitative examples of our proposed model. The examples are from Fashion200k dataset.

[4] Gerd Berget. 2020. " Information Needs of the End Users Have Never Been Discussed" An Investigation of the User-intermediary Interaction of People with Intellectual Impairments. In *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval*. 93–102.

[5] Jerome R Busemeyer and Peter D Bruza. 2012. *Quantum models of cognition and decision*. Cambridge University Press.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 4171–4186.

[7] Ingo Frommholz, Birger Larsen, Benjamin Piwowarski, Mounia Lalmas, Peter Ingwersen, and Keith Van Rijsbergen. 2010. Supporting polyrepresentation in a quantum-inspired geometrical retrieval framework. In *Proceedings of the third symposium on Information interaction in context*. 115–124.

[8] Ingo Frommholz, Benjamin Piwowarski, Mounia Lalmas, and Keith Van Rijsbergen. 2011. Processing queries in session in a quantum-inspired IR framework. In *European Conference on Information Retrieval*. Springer, 751–754.

[9] Dehong Gao, Linbo Jin, Ben Chen, Minghui Qiu, Peng Li, Yi Wei, Yi Hu, and Hao Wang. 2020. Fashionbert: Text and image matching with adaptive loss for cross-modal retrieval. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2251–2260.

[10] Xintong Han, Zuxuan Wu, Phoenix X Huang, Xiao Zhang, Menglong Zhu, Yuan Li, Yang Zhao, and Larry S Davis. 2017. Automatic spatially-aware fashion concept discovery. In *Proceedings of the IEEE International Conference on Computer Vision*. 1463–1471.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[12] Mehrdad Hosseinzadeh and Yang Wang. 2020. Composed Query Image Retrieval Using Locally Bounded Features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3596–3605.

[13] Peter Ingwersen. 1996. Cognitive perspectives of information retrieval interaction: elements of a cognitive IR theory. *Journal of documentation* 52, 1 (1996), 3–50.

[14] Phillip Isola, Joseph J Lim, and Edward H Adelson. 2015. Discovering states and transformations in image collections. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1383–1391.

[15] Amit Kumar Jaiswal, Haiming Liu, and Ingo Frommholz. 2020. Utilising information foraging theory for user interaction with image query auto-completion. In *European Conference on Information Retrieval*. Springer, 666–680.

[16] Xiaoran Jin, Marc Sloan, and Jun Wang. 2013. Interactive exploratory search for multi page search results. In *Proceedings of the 22nd international conference on World Wide Web*. 655–666.

[17] Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. 2014. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539* (2014).

[18] Adriana Kovashka, Devi Parikh, and Kristen Grauman. 2015. Whittlesearch: Interactive image search with relative attribute feedback. *International Journal of Computer Vision* 115, 2 (2015), 185–210.

[19] Jiayu Li, Min Zhang, Weizhi Ma, Yiqun Liu, and Shaoping Ma. 2020. A Multi-level Interactive Lifelog Search Engine with User Feedback. In *Proceedings of the Third Annual Workshop on Lifelog Search Challenge*. 29–35.

[20] Qiuchi Li, Jingfei Li, Peng Zhang, and Dawei Song. 2015. Modeling multi-query retrieval tasks using density matrix transformation. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 871–874.

[21] Qiuchi Li, Sagar Uprety, Benyou Wang, and Dawei Song. 2018. Quantum-Inspired Complex Word Embedding. In *Proceedings of The Third Workshop on Representation Learning for NLP*. 50–57.

[22] Haiming Liu, Paul Mulholland, Dawei Song, Victoria Uren, and Stefan Rüger. 2010. Applying information foraging theory to understand user interaction with content-based image retrieval. In *Proceedings of the third symposium on Information interaction in context*. 135–144.

[23] Yashar Moshfeghi and Joemon M Jose. 2013. On cognition, emotion, and interaction aspects of search tasks with different search intentions. In *Proceedings of the 22nd international conference on World Wide Web*. 931–942.

[24] Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nati Srebro. 2017. Exploring generalization in deep learning. In *Advances in neural information processing systems*. 5947–5956.

[25] Vicki L O'Day and Robin Jeffries. 1993. Orienteering in an information landscape: how information seekers get from here to there. In *Proceedings of the INTERACT'93 and CHI'93 conference on Human factors in computing systems*. 438–445.

[26] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. 2018. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.

[27] Peter Pirolli and Stuart Card. 1999. Information foraging. *Psychological review* 106, 4 (1999), 643.

[28] Benjamin Piwowarski, Ingo Frommholz, Mounia Lalmas, and Keith Van Rijsbergen. 2010. What can quantum theory bring to information retrieval. In *Proceedings of the 19th ACM international conference on Information and knowledge management*. 59–68.

[29] Benjamin Piwowarski and Mounia Lalmas. 2009. A quantum-based model for interactive information retrieval. In *Conference on the Theory of Information Retrieval*. Springer, 224–231.

[30] Daniel E Rose and Danny Levinson. 2004. Understanding user goals in web search. In *Proceedings of the 13th international conference on World Wide Web*. 13–19.

[31] Adam Santoro, David Raposo, David G Barrett, Mateusz Malinowski, Razvan Pascanu, Peter Battaglia, and Timothy Lillicrap. 2017. A simple neural network module for relational reasoning. In *Advances in neural information processing systems*. 4967–4976.

[32] Alessandro Sordoni, Jian-Yun Nie, and Yoshua Bengio. 2013. Modeling term dependencies with quantum language models for IR. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*. 653–662.

[33] Chiheb Trabelsi, Olexa Bilaniuk, Ying Zhang, Dmitriy Serdyuk, Sandeep Subramanian, Joao Felipe Santos, Soroush Mehri, Negar Rostamzadeh, Yoshua Bengio, and Christopher J Pal. 2018. Deep Complex Networks. In *International Conference on Learning Representations*. https://openreview.net/forum?id=H1T2hmZAb

[34] Cornelis Joost Van Rijsbergen. 2004. *The geometry of information retrieval*. Cambridge University Press.

[35] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3156–3164.

[36] Nam Vo, Lu Jiang, Chen Sun, Kevin Murphy, Li-Jia Li, Li Fei-Fei, and James Hays. 2019. Composing text and image for image retrieval-an empirical odyssey. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6439–6448.

[37] Benyou Wang, Qiuchi Li, Massimo Melucci, and Dawei Song. 2019. Semantic Hilbert space for text representation learning. In *The World Wide Web Conference*. 3293–3299.

[38] Bokun Wang, Yang Yang, Xing Xu, Alan Hanjalic, and Heng Tao Shen. 2017. Adversarial cross-modal retrieval. In *Proceedings of the 25th ACM international conference on Multimedia*. 154–162.

[39] Liwei Wang, Yin Li, and Svetlana Lazebnik. 2016. Learning deep structure-preserving image-text embeddings. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 5005–5013.

[40] Xiaohui Xie, Jiaxin Mao, Yiqun Liu, Maarten de Rijke, Qingyao Ai, Yufei Huang, Min Zhang, and Shaoping Ma. 2019. Improving Web Image Search with Contextual Information. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*. 1683–1692.

[41] Oleg Zendel, Anna Shtok, Fiana Raiber, Oren Kurland, and J Shane Culpepper. 2019. Information needs, queries, and query performance prediction. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 395–404.

[42] Liheng Zhang, Guo-Jun Qi, Liqiang Wang, and Jiebo Luo. 2019. Aet vs. aed: Unsupervised representation learning by auto-encoding transformations rather than data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2547–2555.

[43] Liangli Zhen, Peng Hu, Xu Wang, and Dezhong Peng. 2019. Deep supervised cross-modal retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 10394–10403.