Semantic Hilbert Space for Interactive Image Retrieval

Amit Kumar Jaiswal, Haiming Liu, Ingo Frommholz

University of Bedfordshire

UNIVERSITY OF WOLVERHAMPTON

QUARTZ
Quantum Information Access and Retrieval Theory

Outline

- What and Why?
- Background & Motivation
- Multimodal Information Need
    - Projective Transformation
- Multimodal Information Need in Image Retrieval
- Complex-valued CNN
- The Framework - Semantic Hilbert Space
- Experiment
- Conclusion and Future Work

What and Why?

What are we trying to do?

- Construct a method to explicate the user's multimodal information need (IN) into a complex-valued vector space (Hilbert space) for an image retrieval task.
- The modelling method should be capable of performing transformations based on modality features.

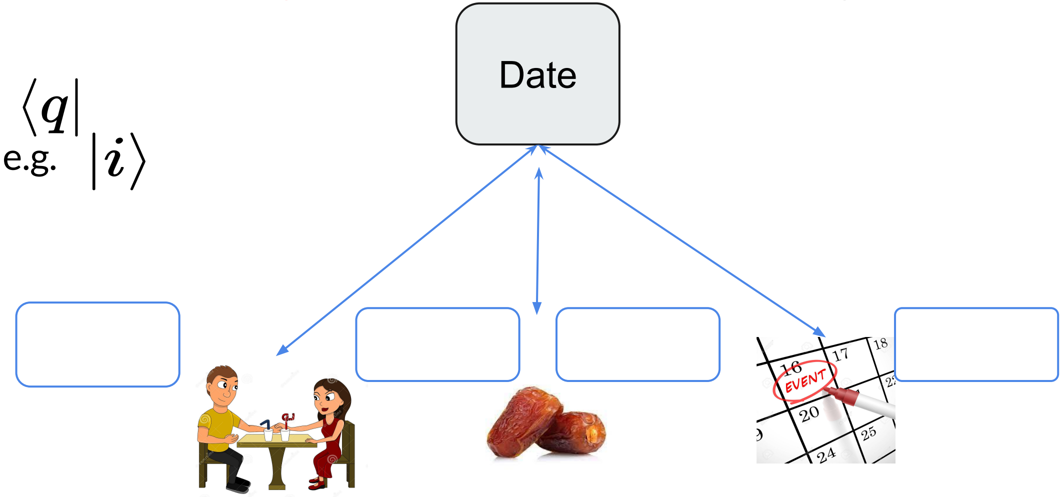
Why are we doing this?

- Such an expressive user's information need can characterise their cognitive aspects.
- The complex-valued vector space can be leveraged by various techniques (neural and deep networks) to accomplish several IR and NLP tasks, such as
    - Text Representation [1]
    - Session Search [2]
    - Text Matching [3]
    - Multimodal Fusion [4]

1. Wang, B., Li, Q., Melucci, M., & Song, D. (2019, May). Semantic Hilbert space for text representation learning. In *The World Wide Web Conference*.
2. Li, Q., Li, J., Zhang, P., & Song, D. (2015, August). Modeling multi-query retrieval tasks using density matrix transformation. In *ACM SIGIR*
3. Li, Q., Wang, B., & Melucci, M. (2019, June). CNM: An Interpretable Complex-valued Network for Matching. In *NACCL*
4. Li, Q., Gkoumas, D., Lioma, C., & Melucci, M. (2021). Quantum-inspired multimodal fusion for video sentiment analysis. *Information Fusion*, *65*, 58-71.

Background & Motivation

A single word can have multiple meanings



Preliminaries - Dirac Notations

- Bra-ket
  - Bra: $\langle . |$ as a row vector, e.g. $\langle q|$
  - Ket: $| . \rangle$ as a column vector, e.g. $|i\rangle$
- Inner Product (or Projection)

$$\langle q|i\rangle$$

$$|Date\rangle = a|i_1\rangle + b|i_2\rangle + c|i_3\rangle + \cdots + d|?\rangle$$

Conversely, a still image manifest multiple words distinctively

Multimodal Information Need

- Input query consists of
  - Textual query $\Longrightarrow$ $|\boldsymbol{q}\rangle$
  - Image query $\Longrightarrow$ $|\boldsymbol{i}\rangle$

- Corresponding embedding vectors
  - Textual features uses BERT: $\mathcal{E}_t(q) = |q_f\rangle$
  - Visual features uses ResNet-34: $\mathcal{E}_p(i) = |i_f\rangle$

- Generate projector to learn textual-visual representation
  - Projection operation

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

- Based on our hypothesis:
  - The textual features are used to encode the transformation process of image features (the input part) and target image features in a common space
- Designates an objective function

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

Projective Transformation (PT)

- We infer that input image query and target image are projective transformations of each other in a complex-valued vector space.
- The input image, textual query and target image are represented as a state vector

  $\overrightarrow{i}$, $\overrightarrow{q}$, $\overrightarrow{r}$

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

- The projective transformation symmetry in above can transform the target image to the input image through the complex conjugate on the textual features.

Multimodal Information Need

- Learn textual features as PT of input image features via a mapping function
  - $$M : \mathbb{R}^d \longrightarrow M_D \in \mathbb{R}^{d \times d}$$
  - Projective transformation

$$Rot(P_T) = e^{\imath M(|q_f\rangle)}$$

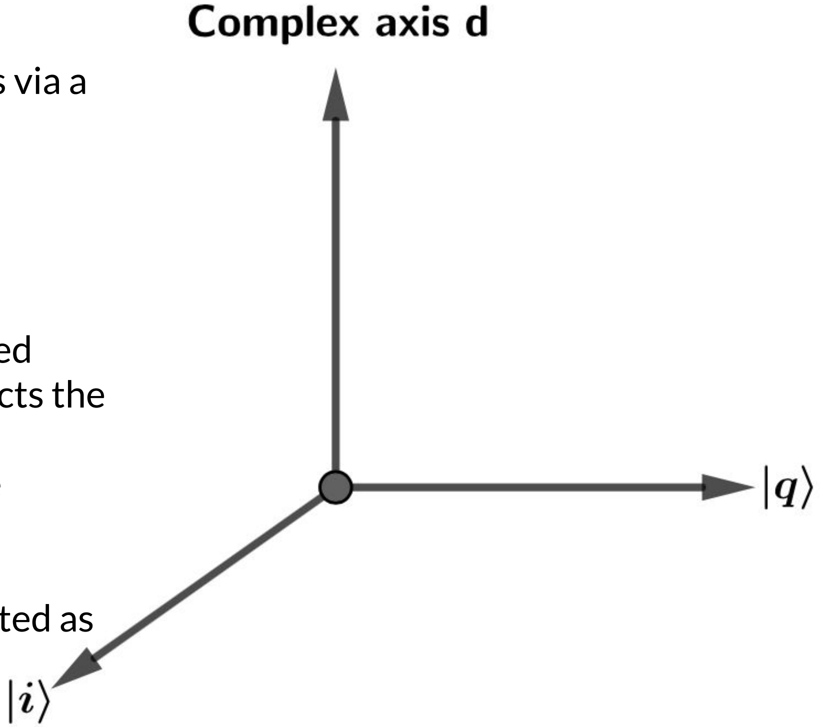
  - Mapping function (M) uses two fully-connected layers with non-linear activation and M_D depicts the matrix diagonal
- Mapping of input image feature ($\mathcal{E}_p(i) = |i_f\rangle$) to the complex space

$$M_I : \mathbb{R}^k \longrightarrow \mathbb{C}^d$$

  - The multimodal information need is represented as

$$I_M = Rot(P_T) M_I(|i_f\rangle)$$

  - M_I is an image mapping function implemented using two-fully connected layers



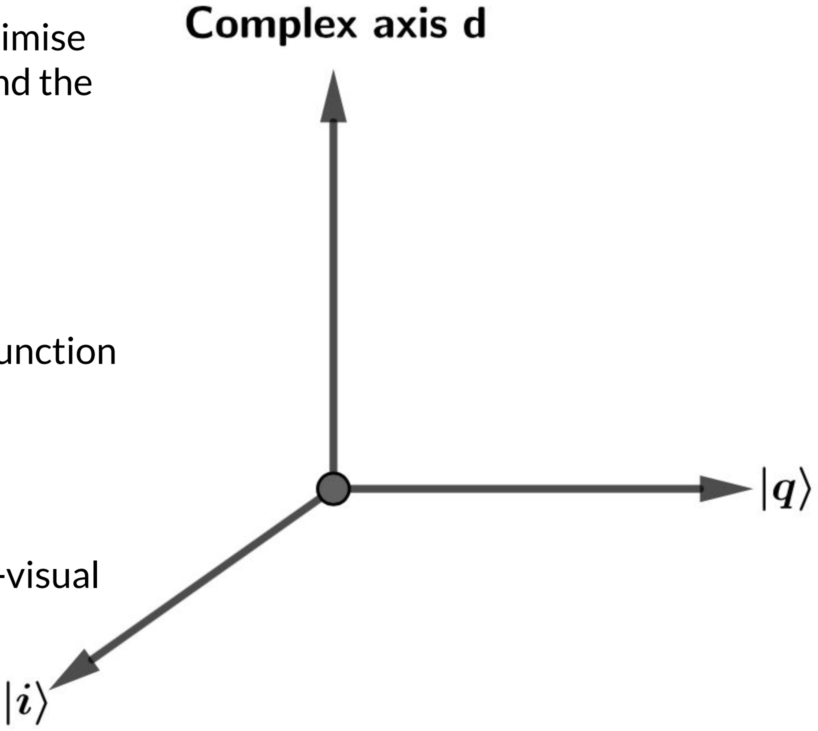
Multimodal Information Need

- Based on the objective (similarity) function, we maximise the similarity between features of multimodal IN and the target image features.
    - Learn the mapping function from

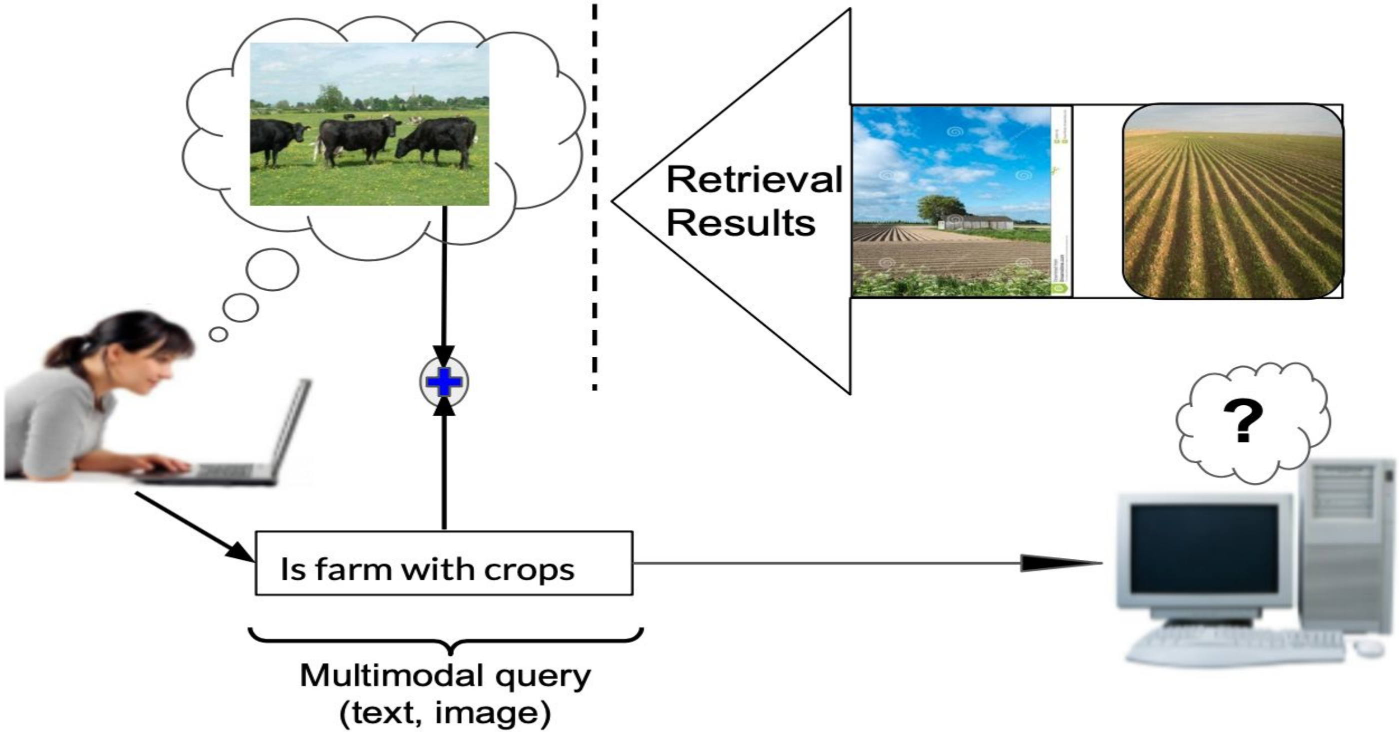
$$f : \mathbb{C}^d \longrightarrow \mathbb{R}^k$$

- To enrich the local features distributed across textual-visual features, we learn another mapping function f_l as below
    - Visual guided feedback
        - Two fully connected layers with a single convolution layer
- The overall representation after composing textual-visual features

$$g(i_f, q_f) = \alpha f(I_M) + \beta f_l(I_M, i_f, q_f)$$



Multimodal IN in Image Retrieval



An instance of our image retrieval task

Complex-valued CNN

To learn the multimodal information need, a complex-valued based CNN is constructed based on [5, 6, 7] which contains:

- Encoder network
  - A fully connected network, where convolutional layer

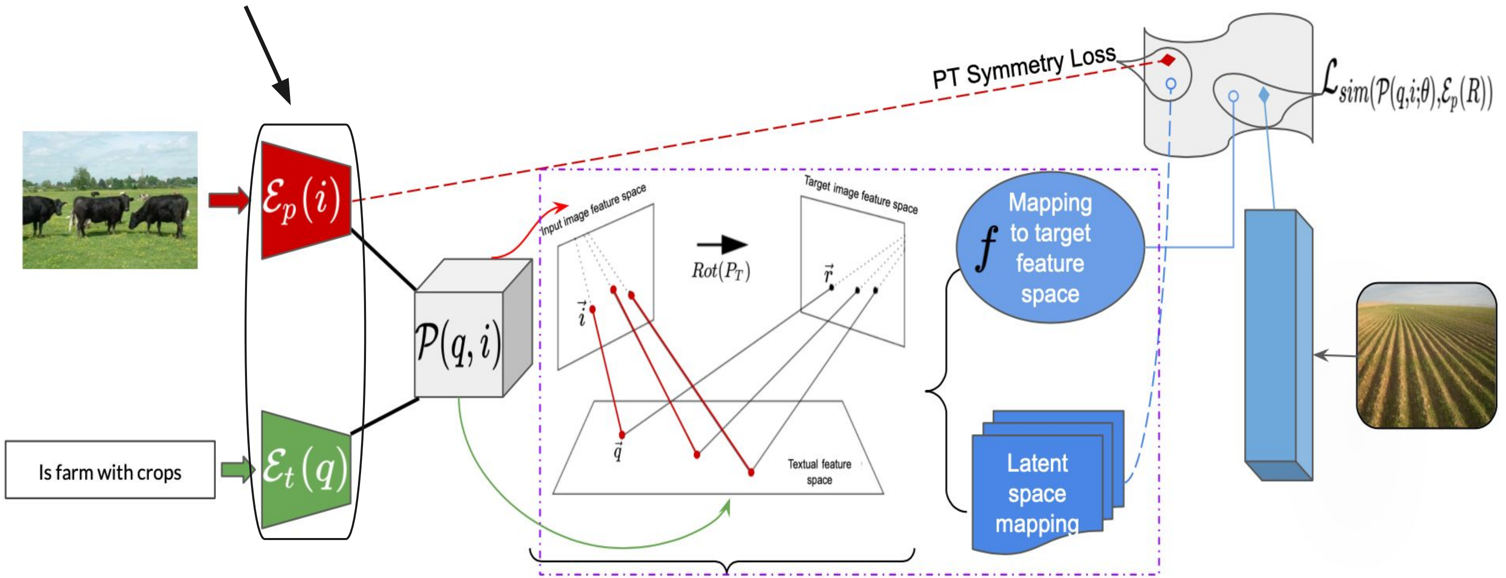
$$\sum_{f=1}^{F}(I_f . W_m) = \sum_{f=1}^{F} \mathrm{Re}(I_f).\mathrm{Re}(W_m) - \mathrm{Im}(I_f).\mathrm{Im}(W_m))$$
$$+ j\sum_{f=1}^{F}(\mathrm{Re}(I_f).\mathrm{Im}(W_m) + \mathrm{Im}(I_f).\mathrm{Re}(W_m))$$

- Decoder network
  - To generate the original extracted textual features and image features
  - Image and text decoder are depicted as D_i and D_q

5. Li, Q. et. al. (2018, July). Quantum-Inspired Complex Word Embedding. In *Proceedings of The Third Workshop on Representation Learning for NLP*, ACL
6. Sordoni, A., Nie, J. Y., & Bengio, Y. (2013, July). 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*.
7. Trabelsi, Chiheb, et al. "Deep Complex Networks." *International Conference on Learning Representations*. 2018.

The Framework - Semantic Hilbert Space

Pre-trained embedding vectors



(Text-image) query	Feature extraction	Projection	Projective transformation	Multimodal representation space	Target feature	Target Image

Experiment

Model \ Dataset	MIT States			Fashion200k		
	Metrics - Recall@K					
	K=1	K=5	K=10	K=1	K=10	K=50
Show and Tell [192]	$11.9^{\pm 0.2}$	$31.0^{\pm 0.5}$	$42.0^{\pm 0.8}$	$12.3^{\pm 1.1}$	$40.2^{\pm 1.7}$	$61.8^{\pm 0.9}$
Relation Network [166]	$12.3^{\pm 0.5}$	$31.9^{\pm 0.7}$	$42.9^{\pm 0.9}$	$13.0^{\pm 0.6}$	$40.5^{\pm 0.7}$	$62.4^{\pm 0.6}$
Film [145]	$10.1^{\pm 0.3}$	$27.7^{\pm 0.7}$	$42.9^{\pm 0.9}$	$12.9^{\pm 0.7}$	$39.5^{\pm 2.1}$	$61.9^{\pm 1.9}$
TIRG [193]	$12.2^{\pm 0.4}$	$31.9^{\pm 0.3}$	$41.3^{\pm 0.3}$	$14.01^{\pm 0.6}$	$42.5^{\pm 0.7}$	$63.8^{\pm 0.8}$
(+) 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 [75]	$14.29^{\pm 0.6}$	$34.67^{\pm 0.7}$	$46.6^{\pm 0.6}$	$16.26^{\pm 0.6}$	$46.90^{\pm 0.3}$	$71.73^{\pm 0.6}$
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: Models on the MIT States and Fashion200k datasets. Models tagged with (+) indicate that they are extended by ourselves.

Ablation Test

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

Table 2: Ablation study of the proposed model on the MIT States and Fashion200k datasets. Components tagged with (+) and (-) indicate the presence and absence in the proposed SHS model.

Qualitative Examples

Instance from MIT States dataset

Instance from Fashion200k dataset

Conclusion and Future Work

Our proposed model is a learning based, but generalised approach that uses Hilbert space formalism

- The model captures the implicit contextual information among an image and textual query to enhance the image retrieval.
- The model generalises in a classical manner by representing textual and image queries via modality distribution (projective transformation).
- Further explore the potential of semantic Hilbert space for Conversational image retrieval.

Thank you!