Semantic Hilbert Space for Interactive Image Retrieval

Amit Kumar Jaiswal* Haiming Liu* Ingo Frommholz†

*University of Bedfordshire † University of Wolverhampton

ICTIR 2021

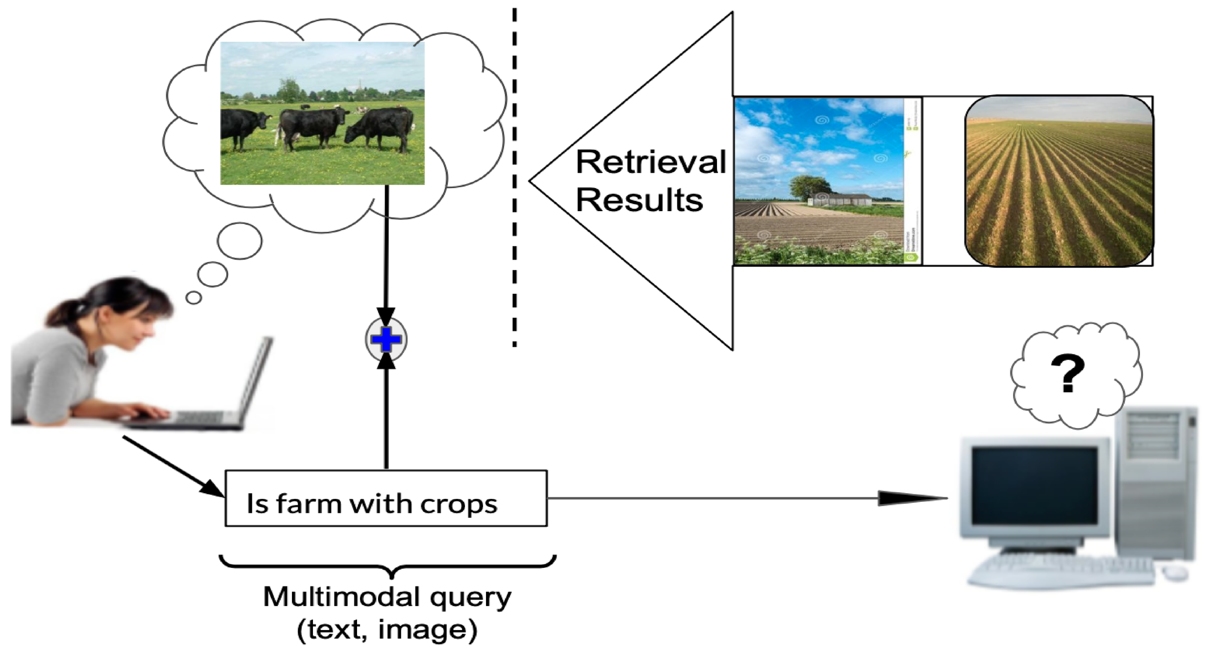
July 11, 2021, Virtual Event, Canada

Problem Formulation and Contribution

Goal: To model the user expressive information needs that delineates their cognitive aspects in an image retrieval task.

Problem Settings: The input consists of a state vectors of image $\vec{i}$ and a textual query $\vec{q}$. These textual-visual query uses projection operation $\mathcal{P}(q, i)$ as $\langle q|i\rangle$. The feature space for both modalities are generated using pre-trained embedding models ($\mathcal{E}_p(.)$). The state vector of retrieved target image ($\vec{r}$). The task is framed as a maximisation problem to learn the multimodal representation

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



Contributions: A quantum-inspired modelling approach [1] to the user multi-semantic information needs that

- present a unified framework - SEMANTIC HILBERT SPACE (SHS), to characterise textual-visual (multimodal) information need in an interactive image retrieval task.
- magnifies model capability using a projective transformation strategy that inherently maps the feature space of input image to the target image feature space via complex-valued text encoding.
- significantly improves on MIT States and Fashion200k datasets over the existing deep networks baseline.

Methodology

Feature Embedding:

- Input Image Feature - $\mathcal{E}_p(i) = i_f \in \mathbb{R}^k$
- Input Textual Feature - $\mathcal{E}_t(q) = q_f \in \mathbb{R}^l$
- Mapping of query image to the target image in a complex-valued space

$$M : \mathbb{R}^d \longrightarrow M_D \in \mathbb{R}^{dxd}$$

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

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

$$I_M = Rot(P_T)M_I(i_f)$$

Information Need Function:

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

Projective Transformation:

$$P_T(\vec{i}) \xrightarrow[\vec{q}]{} \vec{r} \implies P_T(\vec{r}) \xrightarrow[\vec{q}]{} \vec{i}$$

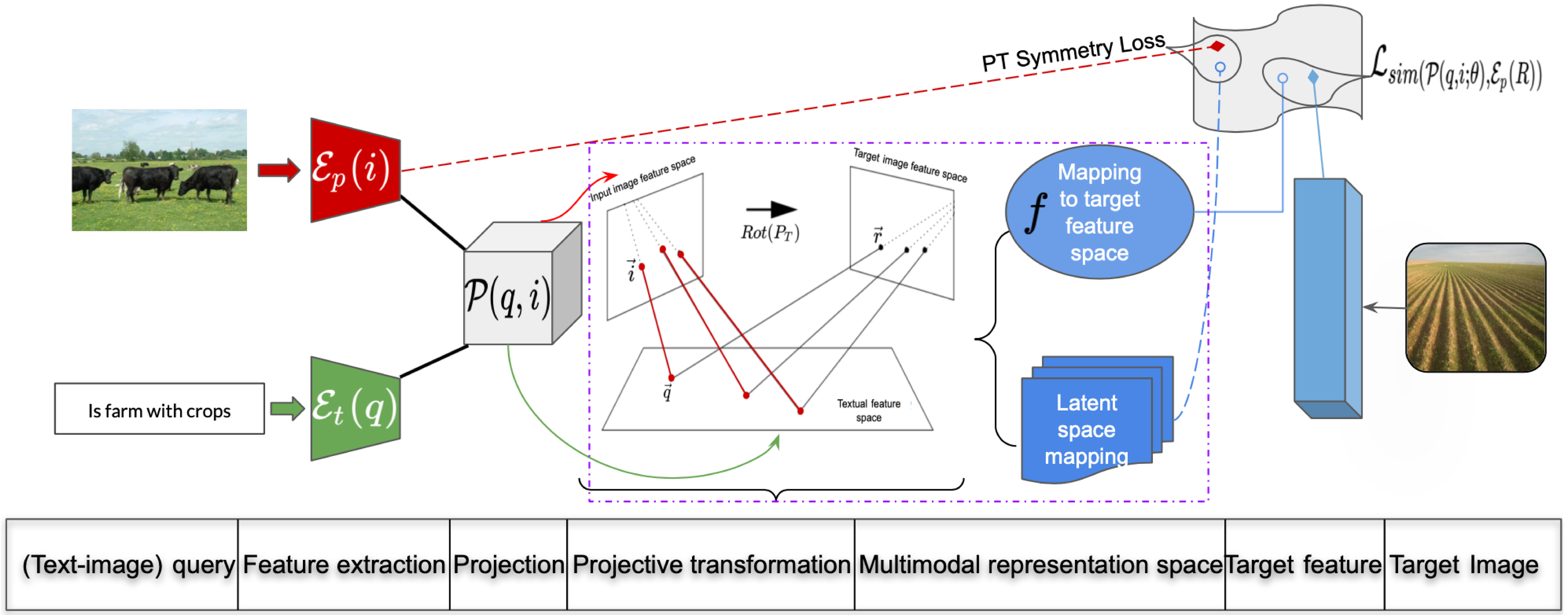
Projective Transformation Symmetry Loss Function:

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

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

Model & Experiments

Semantic Hilbert Space Model



Qualitative Examples & Results

(a) MIT States dataset

(b) Fashion200k dataset

Model \ Dataset	MIT States			Fashion200k		
	Metrics - Recall@K					
	K=1	K=5	K=10	K=1	K=10	K=50
Show and Tell [192]	$11.9^{\pm0.2}$	$31.0^{\pm0.5}$	$42.0^{\pm0.8}$	$12.3^{\pm1.1}$	$40.2^{\pm1.7}$	$61.8^{\pm0.9}$
Relation Network [166]	$12.3^{\pm0.5}$	$31.9^{\pm0.7}$	$42.9^{\pm0.9}$	$13.0^{\pm0.6}$	$40.5^{\pm0.7}$	$62.4^{\pm0.6}$
Film [145]	$10.1^{\pm0.3}$	$27.7^{\pm0.7}$	$42.9^{\pm0.9}$	$12.9^{\pm0.7}$	$39.5^{\pm2.1}$	$61.9^{\pm1.9}$
TIRG [193]	$12.2^{\pm0.4}$	$31.9^{\pm0.3}$	$41.3^{\pm0.3}$	$14.01^{\pm0.6}$	$42.5^{\pm0.7}$	$63.8^{\pm0.8}$
(+) BERT	$12.6^{\pm1.0}$	$31.6^{\pm1.0}$	$43.1^{\pm0.3}$	$15.2^{\pm0.4}$	$43.4^{\pm0.2}$	$63.8^{\pm1.2}$
Composed Query [75]	$14.29^{\pm0.6}$	$34.67^{\pm0.7}$	$46.6^{\pm0.6}$	$16.26^{\pm0.6}$	$46.90^{\pm0.3}$	$71.73^{\pm0.6}$
SHS (Ours)	$14.2^{\pm0.6}$	$\mathbf{36.4}^{\pm0.1}$	$\mathbf{48.2}^{\pm0.3}$	$\mathbf{23.2}^{\pm0.4}$	$\mathbf{55.6}^{\pm1.0}$	$\mathbf{74.2}^{\pm0.6}$

Results

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

Ablation Test

Conclusion:

Our proposed model is a learning-based, but generalised approach that uses Hilbert space formalism [1].

- Our model captures the implicit contextual information among an image and textual query to enhance the image retrieval.
- The proposed model generalises in a classical manner by representing textual and image queries via modality distribution (projective transformation).
- Outperforms strong image retrieval methods [2] on benchmark datasets.

References:

[1] Piwowarski *et al.* (2010, October). What can quantum theory bring to information retrieval. In Proceedings of the 19th ACM international conference on Information and knowledge management (pp. 59-68).

[2] Hosseinzadeh, M., & Wang, Y. (2020). Composed query image retrieval using locally bounded features. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 3596-3605).