COFAR: Commonsense and Factual Reasoning in Image Search

Prajwal Gatti[1], **Abhirama Subramanyam Penamakuri**[1], **Revant Teotia**[2,*],
Anand Mishra[1], **Shubhashis Sengupta**[3], **Roshni Ramnani**[3]

[1]Indian Institute of Technology Jodhpur, [2]Columbia University, [3]Accenture Labs

{pgatti, penamakuri.1, mishra}@iitj.ac.in, rt2819@columbia.edu

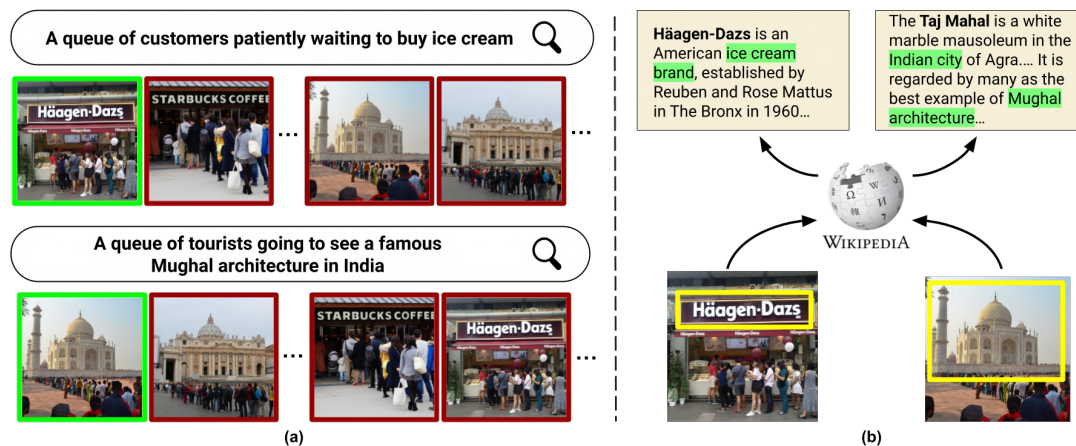{shubhashis.sengupta, roshni.r.ramnani}@accenture.com

Figure 1: Consider the following two natural language queries shown in (a). Retrieving images relevant to these queries (shown using a green bounding box) requires a model that has the ability to interpret images beyond just what is visually apparent, such as interpreting – who are customers vs. who are tourists? Who are waiting to buy vs. who are going to see? in other words, visual commonsense. Additionally, the model would need to interpret facts or world knowledge, such as Häagen-Dazs is an ice cream brand and the Taj Mahal in India is an example of Mughal architecture. This can be enabled by linking visual entities in the image to an encyclopedic knowledge source such as Wikipedia. Our work presents such a model, namely KRAMT.

Abstract

One characteristic that makes humans superior to modern artificially intelligent models is the ability to interpret images beyond what is visually apparent. Consider the following two natural language search queries – (i) "a queue of customers patiently waiting to buy ice cream" and (ii) "a queue of tourists going to see a famous Mughal architecture in India." Interpreting these queries requires one to reason with (i) **Commonsense** such as interpreting people as customers or tourists, actions as waiting to buy or going to see; and (ii) **Fact** or world knowledge associated with named visual entities, for example, whether the store in the image sells ice cream or whether the landmark in the image is a Mughal architecture located in India. Such reasoning goes beyond just visual recognition. To enable both commonsense and factual reasoning in the image search, we present a unified framework, namely

Knowledge Retrieval-Augmented Multimodal Transformer (KRAMT), that treats the named visual entities in an image as a gateway to encyclopedic knowledge and leverages them along with natural language query to ground relevant knowledge. Further, KRAMT seamlessly integrates visual content and grounded knowledge to learn alignment between images and search queries. This unified framework is then used to perform image search requiring commonsense and factual reasoning. The retrieval performance of KRAMT is evaluated and compared with related approaches on a new dataset we introduce – namely COFAR. We make our code and dataset available at https://vl2g.github.io/projects/cofar.

1 Introduction

Retrieving relevant images for a natural language query has been an exciting field of research in the vision-and-language community (Johnson et al., 2015; Wang et al., 2016a, 2020). Most of the available literature focuses on querying visually-evident

---

*This work was done while Revant Teotia was affiliated with Indian Institute of Technology Jodhpur.

aspects in the images, such as searching for objects or their interactions in natural scenes. However, as illustrated in Figure 1, users often require an image search engine that can perform commonsense reasoning and leverage facts (world knowledge) about the image content. To fill this gap, we propose a novel image search task requiring commonsense and factual reasoning associated with named visual entities.

To study this problem, a suitable dataset is required. While many text-to-image search datasets are publicly available (Lin et al., 2014; Young et al., 2014; Sidorov et al., 2020), they have not been explicitly created to study our proposed task. Few of the recently introduced knowledge-enabled VQA datasets such as OK-VQA (Marino et al., 2019), KVQA (Shah et al., 2019), text-KVQA (Singh et al., 2019), FVQA (Wang et al., 2017) require either factual or commonsense or a combination of both. However, they may not be well-suited for studying the "image search" task we are interested in. Note that in the conventional VQA task, a query (question) is evaluated against a single image which is often directly relevant to the query; whereas, in image search, a query needs to be evaluated against several thousands of images, including distractors and then needs to rank the relevant image as the top result. Moreover, to our knowledge, there is no dataset available that includes natural scene images containing a diverse set of visual named entities (such as business brands, celebrities, and world landmarks), visual details of the natural scene along with annotations that demands commonsense and factual reasoning associated with the images. To meet these requirements, we present COFAR, which contains manually annotated English language queries for natural scenes containing named visual entities.

A plausible approach to addressing our image search problem on COFAR is large-scale vision-language pretraining (Radford et al., 2021; Lu et al., 2020) and learning the associations between commonsense-factual concepts and images. This can be successful in learning popular associations, e.g., Starbucks to Coffee, Eiffel tower to Paris if it has seen such samples during training. However, such methods often require large data and generalize poorly to unseen or rare entities. In contrast, we take a distinct path in this work and ground external knowledge associated with entities in the images to perform commonsense and fac-

tual reasoning. To this end, we present a unified model, namely Knowledge Retrieval-Augmented Multimodal Transformer (KRAMT), that retrieves relevant knowledge from Wikipedia by performing query-knowledge similarity-guided visual entity linking. It then encodes the retrieved knowledge, query and visual features, and learns image-query alignment using a multimodal transformer to perform knowledge-aware image search.

Contributions of this paper: (i) We study the problem of image search requiring both commonsense and factual reasoning associated with named visual named entities such as business brands, celebrities, and world landmarks for the first time and introduce a novel dataset, viz. COFAR for this task. We firmly believe that the proposed task, accompanying dataset, and benchmarks presented in this paper will open up future research avenues. (Section 3) (ii) We introduce a knowledge retrieval augmented multimodal transformer (KRAMT) – a unified framework that learns to align queries with the relevant images by performing visual entity linking, retrieving relevant knowledge, and seamlessly integrating it with visual content. The experimental results demonstrate that KRAMT, besides visual reasoning, can perform commonsense and factual reasoning (Section 4 and Section 5).

2 Related Work

2.1 Image Search by Visio-lingual alignment

The performance of image search using natural language query has been significantly improved in the last few years. Typically, the methods in this space learn the semantic visio-lingual (V-L) alignment; during retrieval, rank the images according to the learned similarity function. Early works (Faghri et al., 2018; Wang et al., 2016b) learn to project image representations and text embeddings into a joint space. Recently, multimodal transformers have become a de facto model for V-L tasks. Their different avatars (Zhang et al., 2021; Lu et al., 2019) tackle multiple V-L tasks jointly by using multi-headed self-attention to encode word tokens and visual objects and are the current state of the art for text-to-image retrieval. However, these methods focus only on the visual cues to represent images and do not encode any external knowledge in their framework. Consequently, any explicit crucial information associated with the image is also ignored.

(a) *Query*: Two people getting married in front of a tower in Paris.
Commonsense: Two people in white gown and suit holding hands leads to the commonsense that they are getting married.
Visual named entity: The Eiffel Tower
Fact: The landmark is Eiffel Tower, which is located in Paris, France.

(b) *Query*: The captain of the Argentina national football team celebrating after scoring a goal.
Commonsense: The person is running cheerfully next to a goalpost leads to commonsense that they are celebrating after scoring a goal.
Visual named entity: Lionel Messi
Fact: Lionel Messi is the captain of the Argentina national football team.

(c) *Query*: Two people showing an interest to purchase a watch.
Commonsense: People looking into the display of a watch store implies they could be interested to purchase a watch there.
Visual named entity: Rolex
Fact: The store Rolex sells watches.

Figure 2: A selection of examples from COFAR showing query, relevant image, associated visual named entity, commonsense and fact.

2.2 Commonsense and Factual Reasoning

Bringing commonsense in vision and language tasks is one of the exciting areas of research. The works in this area primarily address: (i) tasks where commonsense reasoning is purely visio-lingual data-driven (Yin et al., 2021; Park et al., 2020; Zellers et al., 2019; Xing et al., 2021) and (ii) tasks where commonsense is enabled by associating the images with external knowledge (Wang et al., 2017; Marino et al., 2019, 2021; Shah et al., 2019; Singh et al., 2019; Wu et al., 2016). Our proposed task falls in the latter category. However, it is distinctly different from others as none of these works address *image search* requiring detailed visual, commonsense as well as factual reasoning *associated to a diverse set of named entities appearing in the image* including business brands, celebrities, and landmarks. Concerning using named visual entities and associated factual reasoning, the only works closest to ours are (Shah et al., 2019; Singh et al., 2019). However, compared to ours, these works restrict themselves to only celebrities or business brands and have weaker annotations for visual and commonsense reasoning. Despite its importance and many real-world applications on the Web such as news-search, named visual entity linking and its utility towards downstream tasks have been under-explored in the literature. We aim to fill this gap.

3 COFAR: Dataset for Image Search requiring <u>CO</u>mmonsense and <u>FA</u>ctual Reasoning

We introduce COFAR, a dataset for studying the novel problem of image search that requires commonsense and factual reasoning. A detailed com-

COFAR in brief:

Number of queries	40,757
Number of images	25,297
Number of unique named entities	5,060
Source of images	text-KVQA (Singh et al., 2019), Celebrity in Places (Zhong et al., 2016), Google Landmarks (Weyand et al., 2020).
External knowledge source	Wikipedia
Average query length (words)	10.5
Average knowledge length (words)	43.7

Table 1: COFAR dataset statistics.

parison with related datasets is made in Table 2. COFAR contains images of natural scenes that include visual named entities of business brands, celebrities, and world landmarks. We provide annotations created to query commonsense and factual knowledge pertaining to named entities present in images. We use Wikipedia articles as the external knowledge source for the visual named entities. The dataset contains 40,757 manually annotated English language search queries for 25,297 natural images covering a diverse set of 5,060 named entities. We further provide external knowledge sources for each visual entity. COFAR is made publicly available for download: `https://vl2g.github.io/projects/cofar`.

3.1 Image collection:

We begin our dataset creation process by collecting images containing one of the three popular named visual entity types: business brands, famous personalities, and landmarks across the globe. To this end, we first started collecting images from different publicly available sources, i.e., we obtain natural scene images containing business brands, personalities, and landmarks using text-KVQA (Singh et al., 2019), VGG-celebrity in places (Zhong et al.,

Dataset	#Images	Visual Reasoning	Commonsense Reasoning	Factual Reasoning	Contains Named Entities	External Knowledge
VQA datasets						
FVQA (Wang et al., 2017)	2.1K	Minimal	Not a major focus	Yes*	✗	Conceptnet
KVQA (Shah et al., 2019)	24K	Minimal	Not a major focus	Yes	✓	Wikidata
text-KVQA (Singh et al., 2019)	257K	Minimal	Not a major focus	Yes	✓	Wikidata
OK-VQA (Marino et al., 2019)	14K	Minimal	Not a major focus	Yes*	✗	Wikipedia
VCR (Zellers et al., 2019)	110k	Detailed	Major Focus	No	✗	✗
GD-VCR (Yin et al., 2021)	328	Detailed	Major Focus (geo-diverse)	No	✗	✗
Image search datasets						
MS-COCO (Lin et al., 2014)	120K	Detailed	Not a major focus	No	✗	✗
Flickr30k (Young et al., 2014)	30K	Detailed	Not a major focus	No	✗	✗
COFAR (This work)	**25K**	**Detailed**	**Major focus**	**Major Focus**	**✓**	**Wikipedia**

Table 2: Comparison of COFAR with other related datasets. Examples of Minimal vs. Detailed visual reasoning: 'How many chromosomes does the creature in this image have?' (Source: OK-VQA) vs. '**A lady wearing a blue t-shirt** going home after purchasing groceries' (Source: COFAR). Further, Yes* under the factual reasoning column indicates that though these datasets require factual reasoning, their facts are about common objects (such as Orange is a citric fruit) and not about named entities (such as Lionel Messi is an Argentine professional footballer). Besides detailed visual reasoning, commonsense and factual reasoning associated with *visual named entities* appearing in the image are unique aspects of COFAR that distinguish it from other related datasets.

2016) and the Google landmarks (Weyand et al., 2020) respectively.[2] Note that these sources do not provide any natural language queries relevant to the images and, therefore are not directly usable for our task. We then associate each of these images with the Wikipedia page of the entity it contains. Note that during training, this association is assumed to be known, but during testing, we perform visual entity linking. Some of the example entities in our dataset are *Rolex*, *Lionel Messi*, and the *Eiffel Tower*. As shown in Figure 3 the distribution of visual named entities in the images of our dataset is geographically diverse. Further, we also illustrate the diversity in the category-wise distribution of COFAR in Figure 4. We refer the reader to the Appendix for further details on COFAR.

Figure 3: Distribution of named entities in COFAR on the world map. COFAR contains named entities from a diverse list of countries, with a slight unintentional bias towards countries such as the United States of America and Canada. Darker color indicates more entities.

3.2 Manual annotation:

The images, along with their associated Wikipedia summary texts, were given to three hired human annotators with the task of annotating queries. These annotators were from geographically diverse locations and had proficiency in written English. In particular, they were instructed to create queries that include (i) factual information of the entity present in the image, for example, *captain of the Argentina national football team, landmark located in Paris*, as well as (ii) commonsense knowledge about events, activities, people, what is going to happen in the scene, or what might have just occurred, for example, *celebrating after scoring a goal, people in the image are getting married*. An-

notators have also been given the option to discard those images where it is very hard to associate visual commonsense, for example, just a frontal view image of a landmark or a signboard of a business brand or an image without any interesting visual activity around. The entire process of manually coming up with queries that require commonsense and factual reasoning, followed by a manual quality check of the data, took approximately 800 person-hours by three annotators. At the end of this stage, we obtained 25K images and 40K queries involving commonsense and factual information about the image. Table 1 summarizes the dataset statistics of COFAR.

A selection of examples from COFAR are shown in Figure 2. An image search model relying exclusively on visual cues would find it challenging to retrieve the relevant images for the queries

---

[2]Restricted by the budget, instead of choosing entire celebrity in places and the Google landmarks, we choose a reasonably large subset.
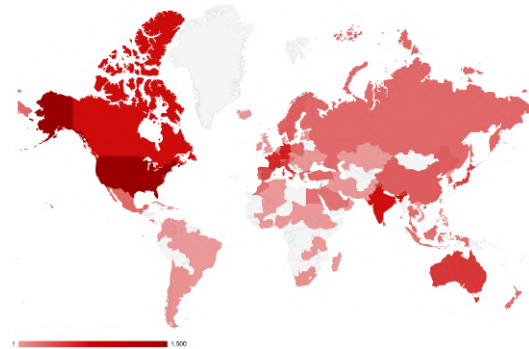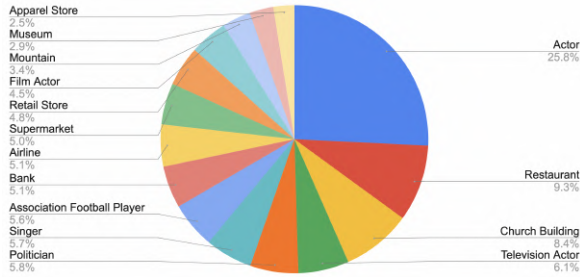
Figure 4: Distribution of the top fifteen categories of named entities present in COFAR.

in COFAR. Consider search query-(c) shown in the figure i.e., two people showing interest in purchasing a watch.. In this image, two people are looking at a display in a Rolex store that sells watches (world knowledge). Therefore, even though detecting watches in this image may be hard for vision models, the matching image shown at the top of this query is relevant. The use of visual entity recognition to associate encyclopedic knowledge and commonsense and factual reasoning are some of the salient features that make COFAR distinctly different from existing text-to-image retrieval datasets.

3.3 Train and Gallery Split:

Based on categories of named entities present, dataset is grouped into COFAR (landmark), CO-FAR (celeb), and COFAR (brand). All the baselines and our proposed method are evaluated on them separately as well together. Further, we split the dataset into (i) **Train set:** Used for learning image-query alignment, this set contains 12,120 images and 33,800 queries. (ii) **Small and large gallery sets:** We show retrieval on two gallery sets containing 1K and 5K images for COFAR. We use 2,800, and 9,800 natural language queries in all for 1K and 5K image galleries, respectively. Please note that retrieval on the test galleries is performed with images containing *entities that are unseen* during training.

4 Knowledge Retrieval-Augmented Multimodal Transformer (KRAMT)

Given a natural language query and a large gallery of images each containing a visual named entity, our goal is to retrieve relevant images. To this end, we present Knowledge Retrieval-Augmented Multimodal Transformer (KRAMT) – an unified framework that contains two major modules: (i) visual entity and query-aware knowledge retrieval

and (ii) knowledge-infused multimodal transformer as illustrated in Figure 5.

4.1 Visual Entity and Query-Aware Knowledge Retrieval:

We posit that visual entities appearing in the image act as a gateway to the encyclopedic knowledge, and its integration to an image retrieval system has the potential to bring commonsense and factual reasoning ability. Therefore, to associate visual entities appearing in the given image to their corresponding Wikipedia page, we perform *visual entity linking* or Image Wikification which is an analogous task to Wikification (Shnayderman et al., 2019) of text corpora, i.e. linking entity mentions in text documents to their corresponding Wikipedia page. More formally, given an image, a set of m candidate entities $\mathcal{E} = \{e_1, e_2, \cdots, e_m\}$ containing business brands, celebrities, and world landmarks, and associated knowledge text (obtained from Wikipedia articles of these entities) $\mathcal{K} = \{k_1, k_2, \cdots, k_m\}$; Image Wikification aims to rank these entities with respect to their image wikification likelihood (s_{iw}). Here, for an image, s_{iw}^u denotes likelihood of uth entity in that image. We obtain these likelihood scores by using off-the-shelf approaches such as CRAFT+CRNN (Baek et al., 2019; Shi et al., 2017) for detecting and recognizing business brand mentions in the image, VGG face (Parkhi et al., 2015) for comparing celebrity faces appearing in the images against a set of reference faces, and landmark recognition (Weyand et al., 2020) for recognizing world landmarks.

If we link images to only that entity which corresponds to the highest likelihood score, linking may be incorrect (especially due to look-alike faces or similar world landmarks or noisy text recognition). This is also evident from the experiment, which clearly shows the gap between top-1 and top-K performance of visual entity linking (Refer to Table 5). To resolve any error in visual entity linking and subsequently retrieving relevant knowledge, we further leverage the natural language query. To this end, we compute the similarity between query and knowledge text associated with top-K entities using a trainable BERT model f and denote these similarity scores as s_{qk} where s_{qk}^u denotes the similarity between query and knowledge text corresponding to uth entity. Further, relevance of each entity with respect to image and given query is computed as follows: $s = \Psi(\alpha s_{iw} + \beta s_{qk})$, here
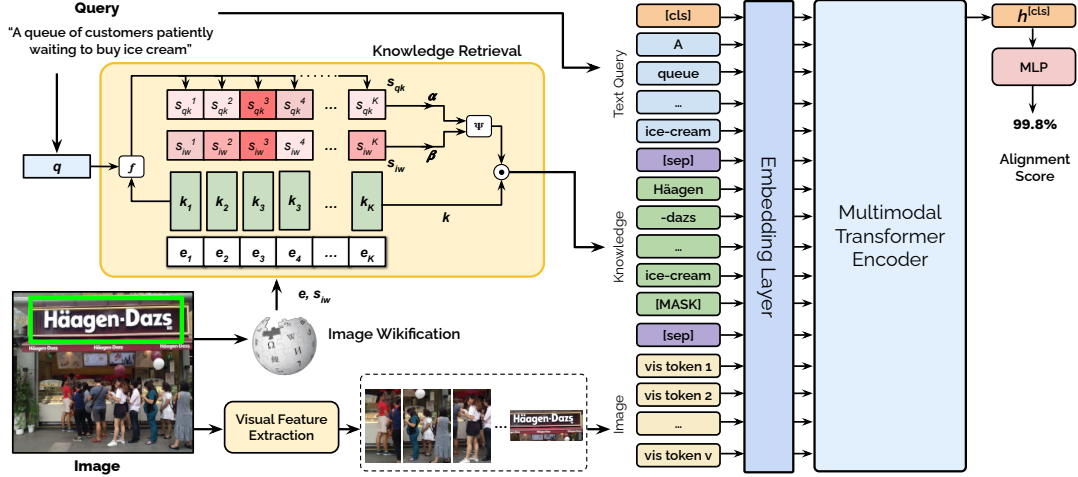
Figure 5: **Overview of proposed Knowledge Retrieval Augmented Multimodal Transformer (KRAMT):** Given a query and a ranked list of visual entities identified in the image, KRAMT grounds the relevant knowledge. This grounded knowledge, along with visual objects and natural query, is fed to a multimodal transformer that learns to align query and relevant image. Please refer Section 4 for more details. **[Best viewed in color]**.

Ψ is argmax. The choice of argmax over softmax is intuitive as only one knowledge text is relevant for a given query and image in our task. Once we obtain s, we perform element-wise multiplication to $\mathcal{K} = \{k_1, k_2 \cdots k_K\}$ and feed this knowledge to a multimodal transfer as described next.

4.2 Knowledge-infused Multimodal Transformer:

Once we obtain relevant knowledge from our knowledge retrieval module, we use Knowledge-infused Multimodal Transformer – a simple and effective architecture to learn alignment between natural language search queries and images along with their associated external knowledge. KRAMT seamlessly integrates these three input modalities in a unified end-to-end trainable architecture. To achieve this, we first encode the query text, knowledge text, and visual regions as three sequences of features. We then project these features to a shared embedding space before using them as input to the KRAMT. These features then attend to each other through multiple self-attention layers (Vaswani et al., 2017). The output of a special class token from the final layer's output is then used to predict the alignment between the query and image along with its knowledge text.

4.3 Pretraining:

We learn a strong vision-language grounding capability in KRAMT through pretraining on MS-COCO (Lin et al., 2014) with the objective tasks of masked language modelling (MLM) and image text matching (ITM).

4.4 Query and Knowledge Encoder:

We fine-tune pretrained BERT (Devlin et al., 2019) to encode the text of the query and external knowledge. For a given search query Q containing L words and a given knowledge k_i containing M words, we embed them into sequences of d-dimensional BERT feature vectors $\{q_l\}_{l=1}^L$ and $\{k_{ij}\}_{j=1}^M$ respectively.

4.5 Image Encoder:

Given an image, we detect a fixed set of N visual objects using Faster R-CNN (Ren et al., 2015) pretrained on Visual Genome (Krishna et al., 2017). Each image I is represented as an unordered sequence of the N object proposals $\{R_i\}_{i=1}^N$ where each R_i is represented as (R_i^{cnn}, R_i^{bbox}), which denote *2048*-dimensional region feature and *4*-dimensional spatial feature, respectively.

We project regional feature R_i^{cnn} and spatial feature R_i^{bbox} into the same d-dimensional space as the search query and the knowledge text using two different learnable transformation matrices $\mathbf{W}_{cnn}$ and $\mathbf{W}_{bbox}$. We apply layer normalization $L(\cdot)$ (Ba et al., 2016) to each transformed feature, and add them to get the final visual object feature F_{R_i}.

$$F_{R_i} = L(\mathbf{W}_{cnn}R_i^{cnn}) + L(\mathbf{W}_{bbox}R_i^{bbox}). \quad (1)$$

1190

4.6 Query-Image Alignment Learning:

Besides learning d-dimensional embeddings for the three inputs, we also learn it for three special tokens, namely $[SEP]$ to separate the input modalities, $[CLS]$ to calculate the final alignment score and $[MASK]$ to replace the text tokens during MLM. We then allow all the $L + M + N + 3$ input token features to attend to each other through T transformer encoder layers to obtain a joint representation.

As the final step, a multi-layer perceptron that takes d-dimensional $[CLS]$ output feature and produces an alignment score $Out^{[CLS]}$ indicating if the given pair of a search query and the image with associated knowledge are aligned or not, is used. During training, we create positive pairs by selecting images and their corresponding queries from the dataset and negative pairs by randomly changing either the image or the query of the selected pair with another random choice in the dataset. We train the model using binary classification loss. Further, to make the image-query alignment robust, we also train the model with the MLM objective wherein each iteration of training, we replace text input tokens at random with a special token $[MASK]$ with a probability of 0.15 and predict the masked tokens based on the context of image, query, and knowledge. During retrieval, for a given query, we rank all the images in the gallery based on the predicted alignment scores. Further implementation details of KRAMT are provided in the Appendix.

5 Experiments and Results

We group image retrieval baseline approaches into three categories: (i) Knowledge-only, (ii) Vision-only, and (iii) Knowledge-aware vision and language (V-L) models to investigate the following questions respectively:

- How much impact does external knowledge have? Can it alone drive performance in CO-FAR without any visual cues?
- Is there a need for integrating external knowledge in COFAR?
- How do other knowledge-aware baselines perform on COFAR?

Under **Knowledge-only**, we utilize BERT (Devlin et al., 2019) to perform query-knowledge sentence-matching. In **VL models**, we use modern text-to-image retrieval methods, namely VSE++ (Faghri et al., 2018), and competitive vision-and-language transformers such as Visual-BERT (Li et al., 2020), ViLBERT (Lu et al., 2019), and VinVL (Zhang et al., 2021). **Knowledge-aware VL models:** As there are no directly comparable knowledge-aware image-retrieval methods in current literature, we implement a few knowledge-aware visual question answering-based models with appropriate modifications to make them compatible for our task: **(i) Modified Memory Network:** Memory networks, and their variations have shown to yield state-of-the-art performance on knowledge-aware VQA benchmarks (Shah et al., 2019; Su et al., 2018). We implement this baseline by using top-K knowledge texts. These texts are scored with a query, and the weighted sum of this representation, CNN features of the image, and query representation are passed to a binary classifier that classifies if the image is relevant to the query. **(ii) KRISP-inspired model:** KRISP (Marino et al., 2021) addresses open knowledge-based VQA using implicit and symbolic knowledge stored in a graph data structure. In our setting, we use unstructured knowledge text in place of symbolic knowledge. We model implicit knowledge using MM-BERT, similar to KRISP, and for unstructured text, we use BERT embedding of the knowledge text. The output of these representations along with BERT-based query representation is fed to an MLP for learning alignment. **(iii) KQIA**: Here, knowledge text, along with queries and images, are encoded using gated recurrent units and CNN, respectively, and are then projected into a common space to learn alignment. All baselines are pretrained on the COCO dataset unless mentioned otherwise.

5.1 Ablations:

To evaluate the effect of different components of KRAMT, we present the following ablations: **KRAMT (w/o Knowledge):** where knowledge text is omitted, **KRAMT (w/o vision):** where only query and retrieved knowledge is used, and **KRAMT (Oracle)** that assumes ground-truth knowledge is available to the model.

5.2 Results and Discussions

We quantitatively evaluate KRAMT on COFAR and compare it against related approaches in Table 3. We report recall (R1, R5 and, R10) and median rank (MdR) averaged over all the test queries. Note that higher values for recall and lower values for median rank are desired. The poor perfor-

Method	COFAR (Unified)				COFAR (Brand)				COFAR (Celeb)				COFAR (Landmark)			
	R1	R5	R10	MdR	R1	R5	R10	MdR	R1	R5	R10	MdR	R1	R5	R10	MdR
1K Gallery																
Knowledge-only																
Sentence similarity	3.1	8.7	19.0	84	2.4	9.3	18.8	68	3.0	8.2	16.9	143	4.2	9.1	19.3	97
Vision-only																
VSE++ (Faghri et al., 2018)	7.4	19.2	23.8	68	6.9	19.5	27.6	60	6.0	25.1	38.5	27	21.8	48.0	59.0	9
VisualBERT (Li et al., 2020)	22.7	50.0	62.5	5	24.0	50.9	63.3	5	8.0	29.3	37.3	22	32.4	64.5	70.0	4
ViLBERT (Lu et al., 2019)	29.8	57.9	71.0	5	28.1	55.4	68.6	4	16.5	34.4	42.0	15	36.0	66.9	74.0	4
VinVL (Zhang et al., 2021)	30.5	62.1	74.3	4	31.2	64.8	75.7	4	18.3	38.9	46.5	10	38.7	68.0	76.3	3
Knowledge-aware V-L Models																
Modified Memory Network	15.2	35.0	50.3	5	14.4	34.9	48.6	18	6.1	26.8	39.4	23	24.5	51.1	60.3	5
KQIA	22.0	52.4	64.5	5	19.9	48.2	57.5	9	10.1	29.2	40.5	19	31.9	57.8	67.0	5
KRISP-inspired model	28.1	53.8	69.0	4	26.8	51.5	67.6	5	13.6	32.5	39.8	17	34.3	65.9	74.2	3
Ours																
KRAMT (w/o Vision)	1.9	6.6	12.6	57	1.1	7.4	12.4	35	2.6	6.6	17.1	164	2.7	10.9	14.5	100
KRAMT (w/o Knowledge)	19.8	39.1	49.8	14	19.4	38.3	49	15	11.8	26.3	35.5	25	35.5	67.3	74.5	2
KRAMT	**31.6**	**64.4**	**76.2**	**3**	**32.9**	**66.5**	**78.6**	**3**	**19.7**	**44.7**	**51.3**	**8**	**40.0**	**69.1**	**80.0**	**2**
KRAMT (Oracle)	40.0	73.2	84.5	2	38.5	72.0	83.3	2	26.3	48.7	61.8	6	42.7	76.4	87.3	2
5K Gallery																
Vision-only																
VSE++ (Faghri et al., 2018)	4.7	11.2	18.0	119	3.9	9.2	17.4	128	2.9	9.1	12.5	274	8.8	20.4	33.6	49
VisualBERT (Li et al., 2020)	11.4	28.6	40.0	19	11.1	28.0	38.8	20	6.7	13.3	20.0	95	13.6	31.0	40.1	18
ViLBERT (Lu et al., 2019)	13.6	31.7	43.5	12	13.0	30.8	41.5	10	9.1	15.8	25.0	67	12.2	43.6	54.0	8
VinVL (Zhang et al., 2021)	15.9	35.6	49.2	10	14.9	33.6	44.5	9	11.2	17.7	30.4	31	14.2	44.9	58.0	6
Knowledge-aware V-L Models																
Modified Memory Network	7.3	21.8	34.6	40	6.8	19.9	30.1	46	3.8	10.1	14.6	143	9.3	26.8	37.9	38
KQIA	9.8	25.3	36.2	21	9.1	24.9	35.4	24	7.7	14.9	20.8	79	10.8	28.1	37.4	28
KRISP-inspired model	14.1	36.6	45.9	10	13.3	32.4	43.7	10	8.8	14.1	23.9	61	12.0	41.4	53.7	7
Ours																
KRAMT	**17.1**	**42.9**	**57.2**	**8**	**16.7**	**42.2**	**56.5**	**8**	**11.8**	**18.4**	**34.2**	**28**	**12.7**	**45.5**	**58.2**	**6**
KRAMT (Oracle)	18.9	45.8	59.9	8	18.5	45.0	58.9	7	15.8	25	38.2	18	18.2	52.7	65.5	5

Table 3: Comparison of retrieval performance on COFAR (with 1K and 5K gallery each) with baselines and ablations. We report mean recall (R) at top 1, 5, and, 10 retrievals and median rank (MdR) over all the test queries.

(a) "Lady handling a financial transaction inside a bank"   *(b):* "People protesting outside the world's most visited museum"

Figure 6: Top-3 retrieved images using proposed KRAMT(w/o Knowledge) and KRAMT on COFAR-1K for two queries. We see that models without access to external knowledge often fail to interpret commonsense such as a financial transaction or protest, and factual information, such as the world's most visited museum, present in the query. On the contrary, KRAMT retrieves semantically more coherent images. Here green colored bounding box indicates the ground truth image.

mance of knowledge-only models confirms that image search in COFAR is non-trivial and external knowledge about the entities in images alone is insufficient. Further, we observe that the vision-only models such as VisualBERT, ViLBERT, and VinVL, without access to external knowledge, do reasonably well solely through visual reasoning. However, it falls short to KRAMT. By virtue of its seamless integration of search query, visual

content, and unstructured knowledge, KRAMT clearly outperforms other baselines, including other Knowledge-aware V-L baselines. These results show the effectiveness of transformer-based methods in COFAR task. The results of ablations are also reported in Table 3. Here, we observe that KRAMT that leverages harvested knowledge for enabling commonsense and factual reasoning is significantly superior to KRAMT (w/o knowledge).

	# of	COFAR-1K			
Method	Pre-train Images	R1	R5	R10	MdR
CLIP (Radford et al., 2021)	400M	26.4	58.1	72.8	6
12-in-1 (Lu et al., 2020)	6.3M	30.2	59.9	74.3	4
KRAMT	125K	**31.6**	**64.4**	**76.2**	**3**

Table 4: Using external knowledge over very large-scale pretraining on COFAR 1K.

COFAR Category	Top 1 (%)	Top 5 (%)
Brand	60.8	79.6
Landmark	63.5	70.2
Celeb	80.1	83.0

Table 5: Results of Image Wikification (visual entity linking) on different categories of COFAR test data.

5.3 Models Pretrained on large-scale datasets

We note it may not be fair to compare our model with those which use very-large-scale datasets for pretraining due to significant differences in size of training data. Moreover, there is possibility of overlap of images in their train sets and CO-FAR-test set; for the sake of a comprehensive comparison, we compare KRAMT with two modern transformer-based models namely CLIP (Radford et al., 2021) and 12-in-1 (Lu et al., 2020) in Table 4. Please note that they use 400M and 6.3M images, respectively, for pretraining as compared to 125K images (COCO) in our model. We see KRAMT surpasses CLIP and 12-in-1 despite being a smaller model.

We show a selection of visual results for top-3 retrievals for two queries in Figure 6. The retrieved images by KRAMT (w/o knowledge) may contain the relevant image, but often ranked lower due to their inability to recognize the entities and perform factual reasoning. On the contrary, the proposed KRAMT consistently retrieves relevant images, confirming our hypothesis.

5.4 Limitations and Future Scope

We observe the following limitations of our work: (i) for the introduction of COFAR, we have chosen natural scenes that contain only one visual named entity. This may not be the case in a real-world setting, (ii) restricted by the budget, current version of COFAR contains only 25K images of 5K named entities in all. However, in an open-set scenario, a much larger and diverse set of visual named entities can be considered, and Image Wikification can be a promising research challenge. In fact a contemporary work (Zheng et al., 2022) poses this as a stand-alone task, and (iii) explicit external knowl-

edge associated with common objects has not been leveraged. We leave addressing these limitations as a future work of this paper.

6 Conclusion

In Information Retrieval and NLP community, knowledge bases are instrumental in enabling commonsense and semantic search. However, their utility in semantic image search has not been extensively explored in the literature. We have drawn the attention of the vision and language community towards this issue through our work and presented a novel multimodal transformer namely KRAMT which seamlessly combines image, query, and knowledge encoding to learn alignment between the image with associated knowledge and query. We firmly believe that image search requiring commonsense and factual reasoning and the new dataset viz. COFAR introduced in this work will open up several future research avenues.

7 Ethical Considerations

One caveat of COFAR is that the images have been collected from various publicly available sources that may contain geographical bias inherently present in them that were undetected in this work. This problem is common with many public vision benchmarks. A more rigorous inspection is indeed required before deploying the proposed model for real-world applications.

References

Lei Jimmy Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. 2016. Layer normalization. *CoRR*, abs/1607.06450.

Youngmin Baek, Bado Lee, Dongyoon Han, Sangdoo Yun, and Hwalsuk Lee. 2019. Character region awareness for text detection. In *CVPR*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*.

Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. 2018. Vse++: Improving visual-semantic embeddings with hard negatives. In *BMVC*.

Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David Shamma, Michael Bernstein, and Li Fei-Fei. 2015. Image retrieval using scene graphs. In *CVPR*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. 2017. Visual Genome: Connecting language and vision using crowdsourced dense image annotations. *Int. J. Comput. Vis.*, 123(1):32–73.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2020. What does BERT with vision look at? In *ACL*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *ECCV*.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *NeurIPS*.

Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 2020. 12-in-1: Multi-task vision and language representation learning. In *CVPR*.

Kenneth Marino, Xinlei Chen, Devi Parikh, Abhinav Gupta, and Marcus Rohrbach. 2021. Krisp: Integrating implicit and symbolic knowledge for open-domain knowledge-based vqa. In *CVPR*.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. OK-VQA: A visual question answering benchmark requiring external knowledge. In *CVPR*.

Jae Sung Park, Chandra Bhagavatula, Roozbeh Mottaghi, Ali Farhadi, and Yejin Choi. 2020. Visual-COMET: Reasoning about the dynamic context of a still image. In *ECCV*.

Omkar M. Parkhi, Andrea Vedaldi, and Andrew Zisserman. 2015. Deep face recognition. In *BMVC*.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *ICML*.

Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. 2015. Faster R-CNN: towards real-time object detection with region proposal networks. In *NeurIPS*.

Sanket Shah, Anand Mishra, Naganand Yadati, and Partha Pratim Talukdar. 2019. KVQA: Knowledge-aware visual question answering. In *AAAI*.

B. Shi, X. Bai, and C. Yao. 2017. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(11):2298–2304.

Ilya Shnayderman, Liat Ein-Dor, Yosi Mass, Alon Halfon, Benjamin Sznajder, Artem Spector, Yoav Katz, Dafna Sheinwald, Ranit Aharonov, and Noam Slonim. 2019. Fast end-to-end wikification. *CoRR*, abs/1908.06785.

Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. 2020. Textcaps: A dataset for image captioning with reading comprehension. In *ECCV*.

Ajeet Kumar Singh, Anand Mishra, Shashank Shekhar, and Anirban Chakraborty. 2019. From strings to things: Knowledge-enabled VQA model that can read and reason. In *ICCV*.

Zhou Su, Chen Zhu, Yinpeng Dong, Dongqi Cai, Yurong Chen, and Jianguo Li. 2018. Learning visual knowledge memory networks for visual question answering. In *CVPR*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NeurIPS*.

Liwei Wang, Yin Li, and Svetlana Lazebnik. 2016a. Learning deep structure-preserving image-text embeddings. In *CVPR*.

Liwei Wang, Yin Li, and Svetlana Lazebnik. 2016b. Learning deep structure-preserving image-text embeddings. In *CVPR*.

Peng Wang, Qi Wu, Chunhua Shen, Anthony Dick, and Anton Van Den Hengel. 2017. Fvqa: Fact-based visual question answering. *IEEE transactions on pattern analysis and machine intelligence*, 40(10):2413–2427.

Sijin Wang, Ruiping Wang, Ziwei Yao, Shiguang Shan, and Xilin Chen. 2020. Cross-modal scene graph matching for relationship-aware image-text retrieval. In *WACV*.

T. Weyand, A. Araujo, B. Cao, and J. Sim. 2020. Google Landmarks Dataset v2 - A Large-Scale Benchmark for Instance-Level Recognition and Retrieval. In *CVPR*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *EMNLP: System Demonstrations*.

Qi Wu, Peng Wang, Chunhua Shen, Anthony R. Dick, and Anton van den Hengel. 2016. Ask me anything: Free-form visual question answering based on knowledge from external sources. In *CVPR*.

Yiran Xing, Zai Shi, Zhao Meng, Gerhard Lakemeyer, Yunpu Ma, and Roger Wattenhofer. 2021. KM-BART: Knowledge enhanced multimodal BART for visual commonsense generation. In *ACL*.

Da Yin, Liunian Harold Li, Ziniu Hu, Nanyun Peng, and Kai-Wei Chang. 2021. Broaden the vision: Geo-diverse visual commonsense reasoning. In *EMNLP*.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. From recognition to cognition: Visual commonsense reasoning. In *CVPR*.

Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. 2021. Vinvl: Making visual representations matter in vision-language models. *CVPR*.

Qiushuo Zheng, Hao Wen, Meng Wang, and Guilin Qi. 2022. Visual entity linking via multi-modal learning. *Data Intell.*, 4(1):1–19.

Yujie Zhong, Relja Arandjelović, and Andrew Zisserman. 2016. Faces in places: Compound query retrieval. In *BMVC*.

Appendix

KRAMT Pre-training

To train our full KRAMT model, we initially pre-train on the COCO captions dataset (Lin et al., 2014) for the objective task of image-caption alignment and masked language modelling. COCO presents a huge diversity of visual content and serves as a good dataset for improving visual reasoning abilities in KRAMT. Further, the model is finetuned on the trainset of COFAR.

KRAMT Implementation Details

We implement the code in PyTorch (Paszke et al., 2019). The transformer layers of KRAMT are implemented using Hugging Face's transformers library (Wolf et al., 2020). We use three transformer encoder layers, with 8 attention heads. The hidden dimension of each block of the transformer layer, as well as the input token feature dimension, is the same as the standard BERT (Devlin et al., 2019) model's hidden dimension of 768.

To encode the query, we use pretrained BERT ('bert-base-uncased') provided by Hugging Face. We keep the sequence length of query text to 40, by truncating the longer sequences and padding the shorter ones. To encode knowledge text, we use the same pretrained BERT, however, this time we keep the sequence length to 80 to accommodate the Wikipedia summary of a page (typically at most 70 words long). This BERT is further fine-tuned during the training of KRAMT with 0.1 times smaller learning rate than that of the KRAMT layers.

To encode images, we extract visual objects using Faster R-CNN (Ren et al., 2015) pretrained on Visual Genome (Krishna et al., 2017). We use top-50 most confident visual object proposals for each image, and represent the visual object's appearance features using Faster R-CNN's 'fc6' features of 2048 dimensions. For spatial features, we use 4-dimensional normalized bounding box representation as mentioned in our approach in the main paper. To represent special tokens $[CLS]$ and $[SEP]$ we learn 768-dimensional embedding for each of them during training.

To get alignment scores from the output embedding of the $[CLS]$ token, we learn a multi-layer-perceptron (MLP) with one hidden layer of size 512 and a ReLU activation. For pretraining on COCO, the knowledge text input is masked and trained for 42 epochs using Adam (Kingma and Ba, 2014) optimizer, with a constant learning rate

Figure 7: Knowledge word cloud

of 1e-4. Before we finetune KRAMT on COFAR for the task of query-image alignment, we finetune KRAMT on text of COFAR with just masked language modelling objective for 10 epochs using Adam (Kingma and Ba, 2014) optimizer, with a constant learning rate of 5e-5. Finally, we finetune KRAMT on COFAR with the task of query-image alignment for 15 epochs using Adam (Kingma and Ba, 2014) optimizer, with a constant learning rate of 0.00002. The model is trained with the binary cross-entropy loss for query-image alignment task, and cross-entropy loss over vocabulary for masked language modelling task. The model was trained using two Nvidia RTX 5000 GPUs (each having 16GB of GPU memory) with a batch size of 64 while training and 128 while testing. KRAMT pre-training takes approximately four days on the two GPUs, whereas KRAMT finetuning on COFAR takes lesser time.

Further details of the implementation can be found in the code which we provide in the project page.
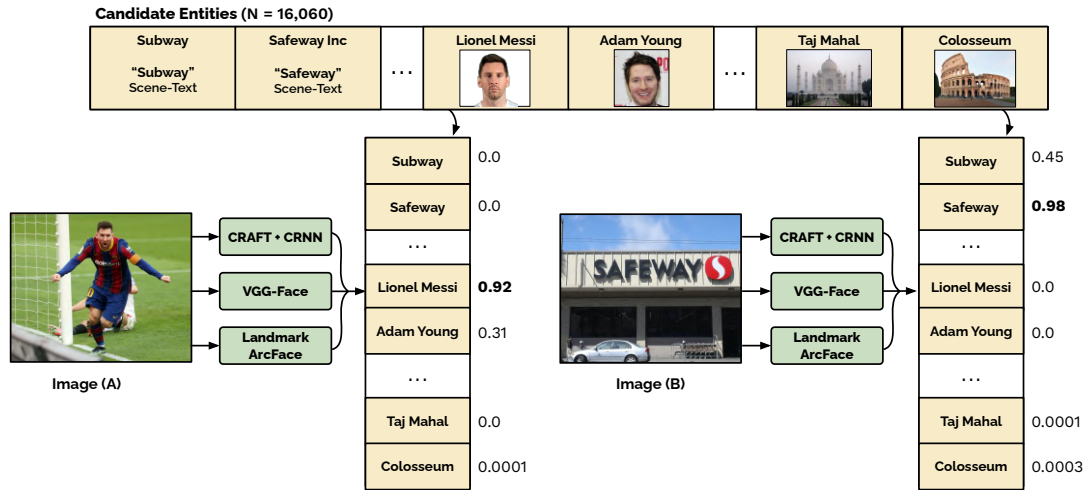
Figure 8: **Overview of Image Wikification (visual entity linking) method in KRAMT**. To recognize named visual entities in images, we use available methods such as CRAFT+CRNN, VGG-Face, and Landmark ArcFace for brands, celebrities, and landmarks respectively. Using these experts, we measure similarity against several thousands of reference entities to obtain a set of high ranking candidates. This open-set recognition approaches allow for addition or removal of any number of reference entities without a need to re-train.
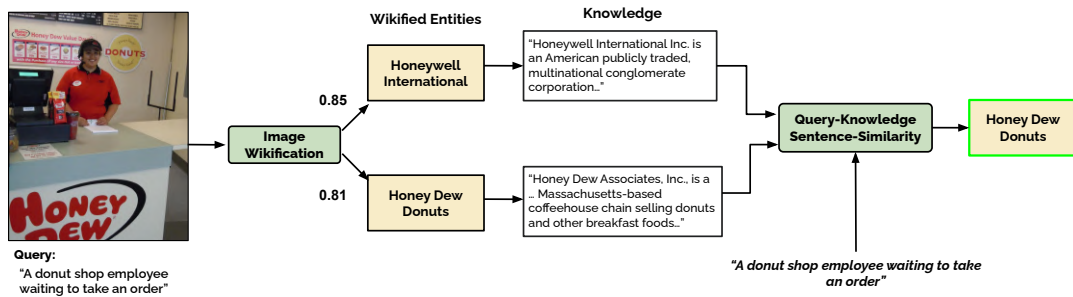

Figure 9: **Using query-based guidance in knowledge-retrieval for KRAMT.** Taking the set of top-ranked candidate entities, we use the search query to select the most appropriate entity by measuring sentence-similarity between the query and entity's knowledge text.

Query: Visitors standing in rain admiring a temple dedicated to the Greece goddess Athena

Visual Named Entity: Parthenon

Knowledge Text: The Parthenon is a former temple on the Athenian Acropolis, Greece, dedicated to the goddess Athena, whom the people of Athens considered their patroness.

Query: A young fan asking the author of the Harry Potter series for an autograph

Visual Named Entity: J. K. Rowling

Knowledge Text: Joanne Rowling (born 31 July 1965), also known by her pen name J. K. Rowling, is a British author and philanthropist. She wrote a seven-volume children's fantasy series, Harry Potter, published from 1997 to 2007.

Query: A white truck parked outside a grocery store waiting to pick up orders

Visual Named Entity: Walmart

Knowledge Text: Walmart Inc. is an American multinational retail corporation that operates a chain of hypermarkets (also called supercenters), discount department stores, and grocery stores from the United States, headquartered in Bentonville, Arkansas.

Figure 10: **A selection of examples from COFAR** along with the ground truth visual named entities present in the images and the associated knowledge texts extracted from their respective Wikipedia articles.

Named Entity Category	# Entities	Belongs to	Examples
Actor	660	Celebrity	Sean Connery, Kim Hyun-joong
Restaurant	237	Business Brand	Panda Express, KFC
Church	215	Landmark	Wolvendaal Church, Innvik Church
Television actor	157	Celebrity	Simon Cowell, Whitney Port
Politician	149	Celebrity	Boris Johnson, Barack Obama
Singer	146	Celebrity	Seun Kuti, Shreya Ghoshal
Football Player	143	Celebrity	Marco Reus, James Milner
Bank	130	Business Brand	DBS Bank, Lloyds Bank
Airline	130	Business Brand	Air Tahiti, Zambezi Airlines
Supermarket	128	Business Brand	Mercadona, Piggly Wiggly
Retail Store	124	Business Brand	Spencer's Retail, Conad
Film Actor	116	Celebrity	Paul Rudd, Anil Kapoor
Mountain	88	Landmark	Mount Majura, Mount Uhud
Museum	74	Landmark	Louvre Museum, Bapu Museum
Apparel Store	65	Business Brand	Quiksilver, Zara
Singer-songwriter	59	Celebrity	Joey Tempest, Tuomas Holopainen
Lake	49	Landmark	Lough Key, Qinghai Lake
Model	47	Celebrity	Lily Cole, Tyson Beckford
Mosque	47	Landmark	The Fatih Mosque, Ahl Fas Mosque
Castle	46	Landmark	Dunsany Castle, Egeskov Castle
Park	45	Landmark	Cove Island Park, Baishamen Park
Auto showroom	38	Business Brand	Honda, Volkswagen
Petrol Station	35	Business Brand	Petrobras, Petro-Canada
Comedian	34	Celebrity	Kapil Sharma, Ken Jeong
Building	33	Landmark	De Bazel, ASEM Tower

Table 6: Distribution of the top 25 most frequent categories of named entities present in the COFAR dataset.

Type	Number of Named Entities	Avg. Length of Knowledge (words)	Avg. Length of Queries (words)	Number of Countries	Number of Entity types
Brand	1060	44.2	11.7	79	39
Celeb	2000	39.0	14.0	92	150
Landmark	2000	41.7	13.6	40	463

Table 7: Statistics about the three categories of data in COFAR.

	COFAR-1K (Unseen entities)				COFAR-1K (Seen entities)			
Method	R1	R5	R10	MdR	R1	R5	R10	MdR
KRAMT	31.6	64.4	76.2	3	**35.1**	**72.6**	**88.6**	**3**

Table 8: Performance of KRAMT on two COFAR-1K versions comprising of entities previously unseen during training and entities seen during training. We observe that performance of KRAMT is higher for already-seen entities.

Query: A person taking home groceries after shopping at a supermarket

Query: A grey car waiting to refuel at a gas station

Query: Celebration at a prehistoric monument known for a ring of standing stones

Query: A crowd of people posing for pictures near a tower famously known for its unstable foundation

Query: The 44th President of the United States of America celebrating his birthday

Query: Kids learning to play the game of chess from a former World Champion

Figure 11: **A selection examples from COFAR**