



**HAL**  
open science

## Explicit knowledge integration for knowledge-aware visual question answering about named entities

Omar Adjali, Paul Grimal, Olivier Ferret, Sahar Ghannay, Hervé Le Borgne

► **To cite this version:**

Omar Adjali, Paul Grimal, Olivier Ferret, Sahar Ghannay, Hervé Le Borgne. Explicit knowledge integration for knowledge-aware visual question answering about named entities. ICRM'23 - ACM International Conference on Multimedia Retrieval, ACM, Jun 2023, Thessalonique, Greece. pp.29-38, 10.1145/3591106.3592227 . cea-04172061

**HAL Id: cea-04172061**

**<https://cea.hal.science/cea-04172061>**

Submitted on 27 Jul 2023

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

# Explicit Knowledge Integration for Knowledge-Aware Visual Question Answering about Named Entities

Omar Adjali  
Université Paris-Saclay, CEA, List  
F-91120, Palaiseau, France  
omar.adjali@cea.fr

Paul Grimal  
Université Paris-Saclay, CEA, List  
F-91120, Palaiseau, France  
paul.grimal@cea.fr

Olivier Ferret  
Université Paris-Saclay, CEA, List  
F-91120, Palaiseau, France  
olivier.ferret@cea.fr

Sahar Ghannay  
Université Paris-Saclay, CNRS, LISN  
France  
sahar.ghannay@lisn.upsaclay.fr

Hervé Le Borgne  
Université Paris-Saclay, CEA, List  
F-91120, Palaiseau, France  
herve.le-borgne@cea.fr

## ABSTRACT

Recent years have shown unprecedented growth of interest in Vision-Language related tasks, with the need to address the inherent challenges of integrating linguistic and visual information to solve real-world applications. Such a typical task is Visual Question Answering (VQA), which aims to answer questions about visual content. The limitations of the VQA task in terms of question redundancy and poor linguistic variability encouraged researchers to propose Knowledge-aware Visual Question Answering tasks as a natural extension of VQA. In this paper, we tackle the KVQAE (Knowledge-based Visual Question Answering about named Entities) task, which proposes to answer questions about named entities defined in a knowledge base and grounded in visual content. In particular, besides the textual and visual information, we propose to leverage the structural information extracted from syntactic dependency trees and external knowledge graphs to help answer questions about a large spectrum of entities of various types. Thus, by combining contextual and graph-based representations using Graph Convolutional Networks (GCNs), we are able to learn meaningful embeddings for Information Retrieval tasks. Experiments on the ViQuAE public dataset show how our approach improves the state-of-the-art baselines while demonstrating the interest of injecting external knowledge to enhance multimodal information retrieval.

## CCS CONCEPTS

• **Information systems** → **Question answering; Test collections; Multimedia and multimodal retrieval.**

## KEYWORDS

Multimedia retrieval, Knowledge injection

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

ICMR '23, June 12–15, 2023, Thessaloniki, Greece

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 979-8-4007-0178-8/23/06...\$15.00

<https://doi.org/10.1145/3591106.3592227>

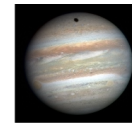
## ACM Reference Format:

Omar Adjali, Paul Grimal, Olivier Ferret, Sahar Ghannay, and Hervé Le Borgne. 2023. Explicit Knowledge Integration for Knowledge-Aware Visual Question Answering about Named Entities. In *International Conference on Multimedia Retrieval (ICMR '23)*, June 12–15, 2023, Thessaloniki, Greece. ACM, New York, NY, USA, 10 pages. <https://doi.org/10.1145/3591106.3592227>

## 1 INTRODUCTION



What planet in our solar system is closest in size to this one?



Jupiter [SEP] Jupiter is the fifth planet from the Sun and the largest in the Solar System. It is a giant planet with a mass one-thousandth that of the Sun, but two-and-a-half times that of all the other planets in the Solar System [...].



What community-support focused branch of Christianity grew from the work of this cleric?



United Methodist Church [SEP] The United Methodist Church (UMC) is a mainline Protestant denomination and a major part of Methodism. In the 19th century, its main predecessor, the Methodist Episcopal Church, was a leader in [...].



What is the all-seated capacity of this defunct sports venue?



Wembley Stadium (1923) [SEP] On 26 May 1975, in front of 90,000 people, Evel Knievel crashed while trying to land a jump over 13 single decker city buses, an accident which resulted in his initial retirement from his daredevil life. In 1992, the World Wrestling Federation [...].



Which college was founded by this man in 1440?



Eton College [SEP] Eton College () is a 13–18 independent boarding school and sixth form for boys in the parish of Eton, near Windsor in Berkshire, England. It was founded in 1440 by King Henry VI as [...].

**Figure 1: Examples of Question and Passage Text+Image Pairs in the ViQuAE dataset and Knowledge Base.**

Active research effort has been spent on methods that relate linguistic and visual information to efficiently solve several downstream tasks such as visual question answering and image captioning. Multimodal extensions of traditional Natural Language Processing (NLP) tasks, e.g. Multimodal Machine Translation (MMT) [19], Multimodal Named Entity Recognition (MNER) [27, 57], and Multimodal Entity Linking (MEL) [1, 2], can also benefit from vision-language integration. However, understanding the interactions that exist between language and vision poses several challenges. In particular, recent deep learning approaches require multimodal representations to capture the existing alignments between language and vision. The Visual Question Answering (VQA) task has been at the forefront of benchmarks for evaluating such multimodal integration. In its early definition, VQA aim to answer simple questions given an input image, which allows assessing the reasoning abilities of models on visual and language understanding. More recent work [26, 37, 47, 48] proposed a Knowledge-based VQA task formulation that requires reasoning abilities that go beyond counting objects and color recognition. In contrast, Knowledge-based VQA datasets include questions that require external commonsense knowledge to be correctly answered, as the image content does embed only partial information. To further expand the scope of possible improvements, recent work [18, 38] pointed out the limitations of the knowledge-based VQA datasets, which restrict the reasoning to commonsense knowledge, arguing with the necessity to answer questions that require knowledge about named entities. Following [38], who had previously introduced questions focusing on named entities with text and image, [18] more specifically proposed the Knowledge-based Visual Question Answering about named Entities (KVQAE) task, a VQA problem formulation where answering questions requires knowledge about named entities defined within a knowledge base (KB). They proposed the ViQuAE dataset, which covers hundreds of entity types whereas the KVQA dataset [38] is limited to person entity types.

In this context, we classically tackle the KVQAE task as a two-step process: a first Information Retrieval (IR) step followed by a Reading Comprehension (RC) step. As illustrated by Figure 1, the IR step starts from a (question, image) pair and aims to retrieve a restricted set of (passage, image) pairs from the reference KB. This KB is assumed to have unstructured content, made of texts and images. The RC step aims to extract and rank answers from the passages retrieved by the first step. In this work, we more particularly focus on the IR step and propose two main contributions.

The first and main contribution enhances the IR step by integrating structural information about named entities under the form of various relations from a knowledge graph. This integration is more specifically performed by combining a dual encoder architecture with Graph Convolutional Networks (GCNs). Besides contextual information, GCNs allow exploiting structural information to learn richer question and passage representations that help retrieve the relevant passages.

The second contribution has a more indirect benefit. As demonstrated in [18], the KVQAE task is both conceptually and computationally challenging because the IR is done on a KB with millions of passages. Thus, we propose a smaller version of the ViQuAE dataset in which the KB is drastically reduced. We show experimentally the interest of such a reduced version for fast prototyping, i.e. the

performance obtained on the reduced version is a good proxy of the performance on the large dataset with a positive impact in terms of computational resources.

## 2 RELATED WORK

### Visual Question Answering

VQA focuses on answering questions conditioned on visual input. It has been a longstanding benchmark for vision and language integration/reasoning. Early research effort [3, 24] has been spent on developing many datasets that include questions about object names, colors, and attributes. With the rapid improvement of image understanding techniques, answering such questions boils down to a visual recognition task since images contain all the necessary knowledge to answer questions. Similar to Knowledge-based Textual Question Answering [23, 40, 54–56], Knowledge-based VQA received attention [29, 30, 47, 48, 51] to further explore commonsense/visual reasoning abilities. Several datasets have been developed, such as FVQA [47] and KB-VQA [48], where external knowledge is leveraged to answer questions whose image does not carry all the required knowledge. However, such small-scale datasets comprised only trivial questions that required knowledge about common nouns, involving simple reasoning schema such as KB retrieval. Alternatively, [26] proposed OK-VQA, a dataset for visual reasoning with open knowledge, i.e., the external knowledge is not restricted to a predefined (closed) Knowledge Graph (KG). Moreover, the required commonsense knowledge is not involved in the dataset building process, resulting in unbiased questions. Likewise, [37] proposed A-OKVQA, an open-domain KB-based VQA dataset that improves the previous one with more qualitative questions and varied commonsense knowledge that alleviate the single retrieval problem. Thus, questions require multiple steps of reasoning to be correctly answered. Different from knowledge-based VQA, KVQAE questions rely on knowledge about named entities defined in a KB rather than commonsense knowledge. [38] first built the KVQA dataset with entities being limited to named persons from the Wikidata KG. However, the lack of diversity of entity types leads VQA systems to rely too much on face recognition, ignoring the unstructured knowledge about entities. In this context, we tackle the KVQAE task by focusing on the ViQuAE dataset [18], which offers more entity type diversity and being more challenging.

### Vision-Language Pretrained Models

Joint representations learned from text/image pairs are used in various multimodal downstream tasks such as VQA, image captioning, and cross-modal retrieval [3, 43, 46]. Recently, Vision-Language Pretrained Models demonstrated significant performance gains, which led to a proliferation of architectures and pretraining objectives. Taking inspiration from the advances made in studies about attention, these models mainly rely on transformers [45]. Inspired by BERT [6], they are pretrained on a large amount of aligned text/image pairs to address text+image matching and Visual-Language Mask-based modeling objectives. Specifically, LXMERT [42] and ViLBERT [22] employ two separate encoders pretrained on Masked multimodal learning and multi-modal alignment. Contrastive learning is another Vision-Language pretraining strategy

illustrated in the recent work on Contrastive Language-Image Pre-training (CLIP) [33], ALIGN [11], and CoCa [58]. Contrastive representation learning consists in learning a mapping function that projects similar inputs in close regions in the embedding space according to a distance metric such as the Cosine distance or the Euclidean distance. CLIP has demonstrated good zero-shot performance on image-text retrieval tasks and few-shot abilities on some multimodal tasks [39]. Thus, fine-tuning CLIP features on the KVQAE task might be relevant, although it fails to perform well on several other language-vision tasks [16].

## Knowledge Injection into Language Models

After pretraining a language model (LM), it undoubtedly acquires linguistic knowledge but also factual (facts about entities), relational (relation between concepts/entities), and commonsense knowledge [32, 35]. However, they also tend to suffer from some issues such as (1) catastrophic forgetting after fine-tuning on a downstream task despite any type of regularization, (2) the hallucination problem, for example in dialogue systems (pretrained LMs generate factually incorrect statements) [8, 34], and (3) the fact that they rely on memorization during pretraining, which makes them struggle with unseen entities [4, 21]. To alleviate these issues, a certain number of recent approaches focused on solutions to better implicitly incorporate knowledge in Pretrained Language Models (PLMs). For example, [15] proposed a label-aware masked language model to solve the sentiment classification task. Using the same principle, [41] employed phrase-level and entity-level masking strategies to learn knowledge-enhanced language models. LUKE [52] distinguishes tokens and entities by pretraining a LM using a new entity-aware self-attention mechanism on two different tasks, a standard Masked Language Modeling (MLM) objective for words and an entity-level MLM for named entities. In contrast, explicit knowledge injection approaches learn how to directly leverage knowledge (e.g., unstructured texts, structured knowledge bases) from an external source. While such approaches do not integrate knowledge into the model, they are more suitable for contrastive-based information retrieval tasks [53].

## 3 PROBLEM FORMULATION AND DATASET

### 3.1 Task definition

Our work addresses the KVQAE task originally defined in [38] for person entities but adopts its extension by [18] for many more types of entities. Hence, we consider 980 types of entities rather than one (person) only and we apply our approach to the ViQuAE dataset [18]. The target task can be seen as a multimodal information retrieval (IR) problem where the objective is to retrieve relevant textual passages given input questions and their visual content. Specifically, text passages are part of unstructured texts from Wikipedia articles; thus each passage can be mapped to a Wikidata entity. Formally, given a text-image pair  $(Q_T, Q_I)$  representing a natural language question associated with visual content and a knowledge base  $\mathcal{KB} = \{(P_T, P_I)\}$  of text-image passage pairs, the goal is to retrieve the  $k$  most relevant passage pairs  $\{(P_T, P_I)_1 \cdots (P_T, P_I)_k\}$  with respect to the query question pair.

The ViQuAE dataset is very large, with 1.5 million entities. As each article is divided into passages of 100 words, it results in

12 million passages. Computing new representations with a deep model is thus very costly in terms of computational resources. We propose to create a reduced version of ViQuAE on which the experiments can be conducted much faster while remaining a good proxy for the performance on ViQuAE itself.

### 3.2 Knowledge Base Reduction

We identified 2.4k unique entities in the questions and corresponding answers. Many of the answers are directly present in the entity Wikipedia article. Therefore, we matched the entities with their Wikipedia articles thanks to the key *Wikidata ID*. However, all the entities do not have their paired article in the KB, which means that the answers are present in other articles. Indeed, answers can be found in the entity article and/or in other linked articles (e.g. some answers about the entity *James Bond* can be found in the articles dedicated to the movies about James Bond). Such cases were removed, which brings the number of entities in the dataset to 2,337 (against 2,397 in ViQuAE) and 3,618 questions (against 3,697 in ViQuAE). After matching the articles, we have 2.4k articles to build the KB. Then we add 2,663 randomly chosen non-relevant articles to add some noise while maintaining the proportion of entity type with regard to the original KB (human and non-human). After dividing the retained articles into passages, the new KB size is 170k passages, which is 1.4% of the original size (12M passages). The computation time is reduced accordingly, as verified in Section 4.3. We keep the same train/val/test split as in ViQuAE.

To assert whether the resulting miniViQuAE benchmark is a good proxy of the larger one, we reproduce all the IR experiments of the ViQuAE paper<sup>1</sup>. From the DPR model pretrained on TriviaQA [13] (filtered from the same questions as ViQuAE), we fine-tuned the model on the reduced dataset and passages and computed the embeddings of the dataset and passages with the best fine-tuned model. We also extract features with BM25 from this textual data. For visual features, we extracted embeddings with CLIP, a Resnet-50 pretrained on ImageNet and ArcFace [5].

We compute the retrieval score  $s_{method}$  for each feature, then apply a late fusion scheme [28] similar to [18]:

$$P = r_m s_{bm25} + (1 - r_m) s_{dpr} + F \alpha_1 s_{arcface} + (1 - F) \alpha_2 s_{cnn} + \alpha_3 s_{clip} \quad (1)$$

where  $r_m$  is a binary variable to choose whether to use BM25 or DPR for the textual modality,  $F$  takes a binary value at inference depending on face detection, and  $\sum_i \alpha_i = 1$ . All hyperparameters are determined on the validation set.

The results of the evaluation are reported in Table 2 for text-only and multimodal settings and can be compared to those of [18] also reported in Table 3. Although the performances are higher on miniViQuAE, which is easily understood since the dataset is smaller and less noisy, one finds the same relative order of the methods, showing the relevance of our reduced benchmark as a proxy.

## 4 METHOD

The Dense Passage Retrieval (DPR) dual encoder proposed by [14] is a state-of-the-art approach to retrieve passages from a large corpus (see Figure 2). However, it only relies on the information that is

<sup>1</sup>Code will be released at: [https://github.com/OA256864/MEERQAT\\_Entity](https://github.com/OA256864/MEERQAT_Entity)

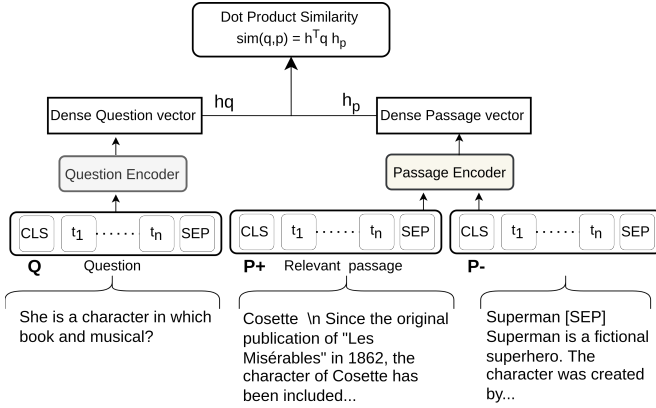


Figure 2: Dual Encoder Architecture Overview.

available in the corpus, which is quite limiting for practical applications that need “commonsense” or external knowledge that is not explicit in the corpus. Usually, such knowledge can be available in the form of a knowledge graph in an external source. The approach we propose leverages such structural information from external sources (see Figure 3) and includes it in the passage representation. In practice, to address the task defined in Section 3.1, we propose question/passage knowledge-enhanced encoders  $E_Q^{TK}(\cdot)$  and  $E_P^{TK}(\cdot)$  that leverage contextual, syntactical, and structural information<sup>2</sup>.

### Text Encoder

We consider baseline textual encoders  $E_Q^T(\cdot)$  and  $E_P^T(\cdot)$  to get dense vector representations for respectively the input question and passage texts. Specifically,  $E_Q^T(\cdot)$  and  $E_P^T(\cdot)$  are two pretrained BERT-based models similar to those used in [18] with two pretraining stages: the original MLM and next sentence prediction pretraining [6] followed by a Question Answering fine-tuning on the TriviaQA [13]. Given an input sequence, its contextual representation  $H_i^T$  is the 768-dimensional vector of the [CLS] wordpiece token of the last hidden layer of BERT.

### Graph Encoder

Using Graph Convolution Networks [17], we leverage the structural information of a knowledge graph by encoding the local graph of neighbor nodes for each node of interest, i.e., the nodes associated with the entities extracted from the passages (see Figure 3). GCNs have been widely used to encode undirected graphs and solve graph-related tasks such as node classification and link prediction but also downstream tasks such as VQA [29]. During training, node representations are learned by aggregating information from the local neighboring nodes of each node, which allows capturing the inherent structural information of the graph. Formally, GCNs encode a graph  $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$  defined by a set of  $N$  nodes  $\mathcal{V}$  and a set of edges  $\mathcal{E}$  using the adjacency matrix  $A \in \mathbb{R}^{N \times N}$ , that reflects the graph structure. Specifically, the entry  $A_{ij} = 1$  if an edge exists between the  $i^{th}$  and  $j^{th}$  nodes; otherwise,  $A_{ij} = 0$ . A deep

<sup>2</sup>The code will be released at [https://github.com/OA256864/MEERQAT\\_Entity](https://github.com/OA256864/MEERQAT_Entity)

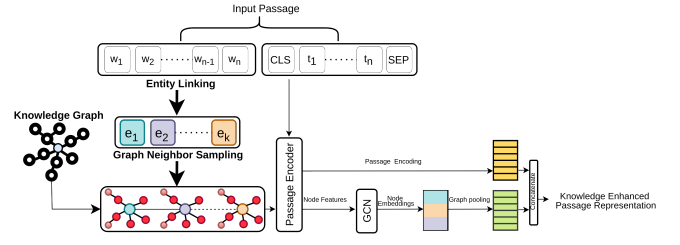


Figure 3: Enhanced Contextual Passage Representation with Graph Structural Information Using GCNs.

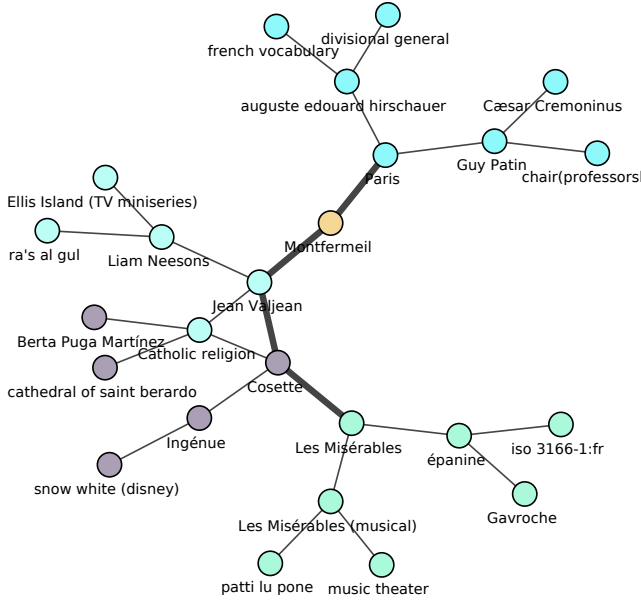
graph convolution network stacks  $L$  hidden layers and iteratively propagates information following the rule:

$$H^{(l+1)} = \sigma(\tilde{D}^{-\frac{1}{2}} \tilde{A} \tilde{D}^{-\frac{1}{2}} H^l W^l) \quad (2)$$

where  $\tilde{A} = A + I_N$  is the adjacency matrix augmented with self-connections,  $I_N$  being the identity matrix;  $\tilde{D}$  is the degree matrix of  $A$  used for normalization purposes where  $\tilde{D}_{ii} = \sum_j \tilde{A}_{ij}$ ;  $W^l$  is a trainable weight matrix at the  $l$ -th layer and  $\sigma(\cdot)$  denotes a non-linear activation function (RELU in our case). Let  $X = H^0 \in \mathbb{R}^{N \times D}$  be the input node feature matrix with  $D=768$ , as we apply the passage encoder on the entity descriptions to initialize node features. We feed  $X$  to the  $L$ -layer GCN, which performs  $L$  hops of message passing and aggregation over the  $N$  nodes of  $\mathcal{G}$  to obtain the final node embeddings  $H^L \in \mathbb{R}^{N \times D}$ .

### 4.1 Knowledge-Enhanced Representation

Our retrieval approach aims to enhance textual representation with external knowledge thus expecting a better alignment between questions and relevant passages. Specifically, given an input sequence, we apply named entity recognition and disambiguation using a state-of-the-art zero-shot Entity Linking (EL) system, BLINK [20]. It allows linking the detected mentions to the 5.9 million article entities in English Wikipedia. While its dual+cross-encoder architecture makes the 12M of passages annotation quite computationally expensive, it guarantees fewer entity annotation errors compared to faster EL systems. Thus, we obtain for a given passage  $P_i$  a set  $M_i = \{m_1 \dots m_n\}$  of mention spans and their corresponding set  $E_i = \{e_1 \dots e_n\}$  of linked entities. This allows relating entity-level information of each passage to external knowledge resources. For this last, we leverage Wikidata5M (W5M) [49], a large-scale knowledge graph (KG) with nearly 5 million nodes and 20 million edges (subject-relation-object triples). It was built upon the July 2019 dump of Wikidata and Wikipedia where each entity in Wikidata is aligned to its Wikipedia page while entities with no pages or with descriptions of fewer than 5 words are discarded. We first map each linked entity  $e_i \in E_i$  to a node  $n_i$  in the W5M KG, resulting with the corresponding node set  $N_i$ . Ideally, in order to leverage KG structural information, we would have built for



**Figure 4: Example of a local neighbor subgraph built from a relevant passage. Edges between Linked entities nodes are in bold. The local neighbor graph for each entity node, extracted from the knowledge graph, is highlighted with the same color. The illustrated subgraph is built with  $k_{nb} = 2$  and  $d = 2$ .**

each passage  $P_i$  the induced W5M subgraph that connects either directly the corresponding nodes in  $N_i$  or through all the intermediate nodes in the KG. Besides the hard computational complexity of subgraph extraction, our preliminary empirical analysis showed that two node entities in  $N_i$  can be very far (hundreds of intermediate nodes) in the KG, leading to large extracted subgraphs whose encoding is unfeasible. Instead, we build a heterogeneous subgraph  $\mathcal{G}_i$  where nodes in  $N_i$  are connected sequentially following the order in the original passage  $P_i$ , which allows the propagation of entity-level information. In addition, to leverage KG structural information, we build for each node in  $N_i$  a local neighborhood graph extracted from W5M KG (see Figure 4 for an example). Given a W5M KG node  $n_i$ , we randomly sample  $k_{nb}$  neighbor nodes with a depth  $d$  using a Depth-First Search (DFS) approach. The hyperparameters  $k_{nb}$  and  $d$  that control the size of  $\mathcal{G}_i$  are determined on the validation set.

After applying an L-layer GCN following the update rule in Eq.2 over the subgraph  $\mathcal{G}_i$  of a passage  $P_i$ , we obtain the node representation matrix  $H_i^{gcn} \in \mathbb{R}^{N_{\mathcal{G}_i} \times D}$ , where  $N_{\mathcal{G}_i}$  is the number of nodes in the subgraph  $\mathcal{G}_i$ . Node features are initialized by encoding the textual description of their corresponding Wikipedia entities using  $E_p^T(\cdot)$ , enabling the GCN to fine-tune BERT layers during training. The number of GCN layers L determines how many information aggregation hops it performs on the local subgraph. Therefore, we set L equals to  $d$ . In our experiments, we also found that encoding edge directions and types using relational GCNs (R-GCNs) [36] did not improve performance, likely due to over-parameterization [25, 44]. Indeed, the number of relation types (more than 800 in

WD5M) requires a reasonable number of training examples for each type. Table 1 shows statistics about the number of linked entities in question and passage texts. One can observe that questions in the validation set have poor entity-level information and a great proportion of questions do not exhibit any named entity. Indeed, the KVQAE is a challenging task that requires leveraging visual content. To alleviate this challenge, we propose to enhance question representations with incorporating syntactic information instead. We use Spacy<sup>3</sup> to annotate question sequences with dependency parsing labels and build a syntactic dependency graph we encode using GCNs as in [25] (see Figure 5).

**Table 1: Statistics about the Number of Linked Entities in Input Sequences**

KB/Dataset	mean	median	std	min	max
<i>passage</i>	6.44	6	3.31	1	101
<i>question<sub>validation</sub></i>	0.495	0	0.738	0	6

In order to obtain a graph-level representation, we apply a graph pooling (max,mean) function that maps the  $H_i^{gcn} \in \mathbb{R}^{N_{\mathcal{G}_i} \times D}$  node representation matrix to a vector  $H_i^K \in \mathbb{R}^{N_{\mathcal{G}_i} \times D}$ . Lastly, the final knowledge-enhanced representation  $H_i^{TK}$  is obtained using concatenation in order to preserve the contextual information during retrieval such that:  $H_i^{TK} = \text{Concat}(H_i^T, H_i^K)$

## Learning Objective

Like standard DPR training, given an input query question  $Q_i$ , a relevant passage  $P_i^+$ , and an irrelevant passage  $P_i^-$ , which represents a hard negative we mined using BM25, we compute the dot product similarity scores  $\text{sim}(Q_i, P_i^+)$  and  $\text{sim}(Q_i, P_i^-)$  as:

$$\text{sim}(Q_i, P_i^+) = E_Q^{\text{TK}}(Q_i)^T \cdot E_P^{\text{TK}}(P_i^+). \quad (3)$$

During training, the encoders learn to project relevant passage vectors closer to question vectors in the embedding space while maximizing the distance with irrelevant passage vectors. Formally, the objective is to minimize the contrastive log-likelihood loss function  $\mathcal{L}$  as follows:

$$\mathcal{L} = -\log \frac{\exp(\text{sim}(Q_i, P_i^+))}{\exp(\text{sim}(Q_i, P_i^+)) + \sum_{j=1}^M \exp(\text{sim}(Q_i, P_{ij}^-))} \quad (4)$$

where M is the number of negative (irrelevant passages) examples per question. This loss also takes advantage of in-batch negatives to increase the number of training examples without any additional computational cost. Indeed, for example, consider B questions in a mini-batch, B relevant passages and  $B \times M$  BM25 mined irrelevant passages; thus, each question is trained on  $(B \times M) + (B-1)$  negative examples since relevant+irrelevant passages for other questions are considered hard negatives for a given question.

<sup>3</sup><https://spacy.io/>

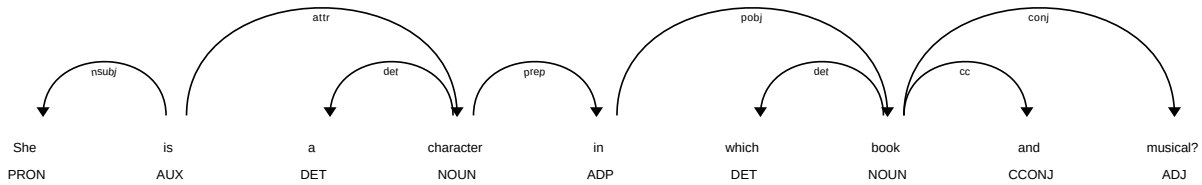


Figure 5: Dependency Parsing Example of a Question in the ViQuAE Dataset.

## 4.2 Image Retrieval

As mentioned earlier, the ViQuAE knowledge base exhibits various entity types, but with a prevalence of named person entities. Similar to [18], we distinguish images with faces using MTCNN [59], an image face detector, from images without faces, assuming that images with faces are more likely to refer to person entities. Thus, images with faces are encoded using ArcFace [5] while images without detected face are represented with ImageNet-ResNet [10] and CLIP, both with a ResNet-50 backbone (see [18] for more details).

## 4.3 Inference

In this work, we easily embed question texts using  $E_Q^{TK}(\cdot)$  while embedding all the KB passages with  $E_P^{TK}(\cdot)$  is computationally expensive. It takes on average 25 hours to embed the 12M passages in the KB using the textual encoder  $E_P^T(\cdot)$ . The same computation using the proposed knowledge-enhanced encoder  $E_P^{TK}(\cdot)$  is 3 times longer, due to our online local neighborhood graph sampling. Although this step can be parallelized, it motivated us to build the reduced KB as explained in Section 3.2. Hence, on miniViquAE, the embedding computation is reduced to 22 minutes rather than 25 hours, allowing to conduct more experiments. At retrieval time, top-100 passage embeddings are retrieved based on their cosine similarity scores with question embeddings.

## 4.4 Late Fusion and Indexing

Following [18], we adopt a late fusion information retrieval approach since our work focuses on enhancing text retrieval encoders with the ability to integrate external knowledge. Indeed, late fusion involves a single modality search at a time before fusing the resulting scores. We thus encode question and passage texts using the proposed knowledge-enhanced text retrieval encoder  $E_{TK}$  and then, using a visual encoder  $E_I$ , we compute a representation for all questions and passage associated images. Once texts and images for each question and passages are mapped to a dense vector representation, the most relevant passages are retrieved according to a similarity measure such that their vector representations are the closest to question ones in the embedding space.

Given the multimodal representations of all questions in the dataset and all passages in the KB, we index them using the Faiss [12] library and perform a dense similarity search to retrieve the top-100 closest passage vector representations to the question ones using the maximum inner product. The resulting scores for each modality are normalized to zero mean and unit variance in order to have comparable distributions [14, 18]. Finally, the scores are fused according to Eq.1.

Similarly to the experiment in Section 3.2, a grid search is applied on the validation set in order to fix the interpolation hyperparameters  $\alpha_j$ , retaining those that maximize the Mean Reciprocal Rank (MRR) metric.

## 5 EXPERIMENTS

We fine-tuned the proposed knowledge-enhanced dual encoder on the ViQuAE dataset following the official split: train (1,190), validation (1,250), and test (1,257). The ViQuAE KB comprises 12M passages whose maximum length equals 100 words. The associated Wikipedia article title is added to each passage header. Note that a single visual content is associated with the passages extracted from the same Wikipedia article while the ViQuAE KB comprises on average 8 passages per article. Hence, the visual content is not helpful in discriminating between passages originating from the same article.

Table 2: Overall IR Results on the Reduced KB

Model	MRR@100	P@1	P@20	Hit Rate@20
DPR	0.354	0.246	0.141	0.669
DPR + GCN	<b>0.367</b>	<b>0.264</b>	<b>0.146</b>	0.666
resnet	0.030	0.022	0.012	0.055
clip-RN50	0.044	0.034	0.021	0.080
arcface	0.169	0.136	0.059	0.255
fusion	0.409	0.307	0.164	0.686
Our fusion	<b>0.414</b>	<b>0.313</b>	0.159	<b>0.696</b>

### 5.1 Experiment Settings










Experiments are performed on a multi-GPU setup (4xGPUs), which is favorable for in-batch negative training. Given 1 relevant + 1 BM25 mined irrelevant passages for each question, with a batch size of 16 per GPU, the total batch size across GPUs equals  $4 \times (1 + 1) \times 16 = 128$ . Questions and passages are truncated to a maximum of 256 tokens and entity descriptions to 16. We performed grid search on the graph sampling hyperparameters and obtained the best performance with a graph depth  $d = 2$  and a number of neighbors  $k_{nb} = 2$ . Similarly, we optimized the interpolation hyperparameters and found:  $\alpha_{dpr} = 0.3$ ,  $\alpha_{resnet} = 0.1$ ,  $\alpha_{arcface} = 0.4$  and  $\alpha_{clip} = 0.2$ .

Loss optimization is performed using Adam over 40 epochs with a learning rate of  $4.10^{-5}$  and a linear schedule. We perform model selection according to the best Mean Reciprocal Rank score on the validation set. The implementation relies on PyTorch [31], Transformers [50], and PyTorch Geometric [9] for graph modeling.

**Table 3: Overall effectiveness of the models. The best results are highlighted in boldface. Superscripts denote significant differences in paired Student’s t-test with  $p \leq 0.01$ . BL denotes previously published baseline results [18].**

#	Model	MRR@100	P@1	P@5	P@20	Hit Rate@5	Hit Rate@20
BL	BM25, text-only	0.190	0.131	0.87	0.59	0.239	0.395
BL	DPR text-only	0.328	0.228	0.200	0.164	0.436	0.612
BL	fusion	0.379	0.278	0.225	0.175	0.495	0.657
a	Ours (text+graph)	<b>0.336<sup>bcd</sup></b>	<b>0.241<sup>bcd</sup></b>	<b>0.207<sup>bcd</sup></b>	<b>0.166<sup>bcd</sup></b>	0.434 <sup>bcd</sup>	0.599 <sup>bcd</sup>
b	resnet	0.019	0.012	0.009	0.009	0.022	0.044
c	clip-RN50	0.036 <sup>b</sup>	0.025 <sup>b</sup>	0.018 <sup>b</sup>	0.017 <sup>b</sup>	0.041 <sup>b</sup>	0.081 <sup>b</sup>
d	arcface	0.145 <sup>bc</sup>	0.111 <sup>bc</sup>	0.074 <sup>bc</sup>	0.052 <sup>bc</sup>	0.188 <sup>bc</sup>	0.222 <sup>bc</sup>
e	Our fusion	<b>0.383<sup>abcd</sup></b>	<b>0.290<sup>abcd</sup></b>	<b>0.223<sup>abcd</sup></b>	<b>0.171<sup>bcd</sup></b>	<b>0.478<sup>abcd</sup></b>	<b>0.644<sup>abcd</sup></b>

**Table 4: Examples of top-1 retrieved passages where only the knowledge-enhanced representations allowed to select relevant passages containing the correct answer**

$Q_I$	$Q_T$	$P_I$	$P_T$
	Which film opens with this fictional universe performing a bungee jump from a dam?		Canton of Ticino [SEP] The opening of the Gotthard Railway in 1882 led to the establishment of a sizeable tourist industry mostly catering to German-speakers, although since the early 2000s the industry has suffered from the competition of more distant destinations. In 2011, 1,728,888 overnight stays were recorded. The mild climate throughout the year makes the canton a popular destination for hikers. The Verzasca Dam, known for the opening scene of the 1995 film "GoldenEye", is popular with bungee jumpers.
	Who was the principal villain in the 2004 movie with this character?	SPIDER-MAN 2	Doctor Octopus [SEP] Doctor Octopus has begun wearing a full-body armor suit due to a crippling illness caused by the amount of punishment he has sustained over the years, made even worse by the fact that his ability to "take" damage is still at a human norm, even if he can deliver a superhuman level of punishment; he relies completely on his arms to prevent opponents with superhuman strength getting in close enough to damage his relatively unfit physical form even before his illness.
	In which city would you find this high-rise building?		Royal Park, Melbourne [SEP] While the move attracted some opposition from green groups, it was promoted with the claim that there would be no net loss of parkland at Royal Park. The basis of area calculations used to support this claim is unclear. Relocation of the Hospital's helipad onto the roof of the new building has removed this impact from the park, but after completion of construction it appears that the hospital buildings now occupy a substantially larger area than previously, and parkland has been lost.
	On which island is this city?		Sicily [SEP] After taking Carthage the Vandals personally led by King Gaiseric laid siege to Palermo in 440 as the opening act in an attempt to wrest the island from Roman rule personally. The Vandals made another attempt to take the island one year after the sack of Rome in 455, at Agrigento, but were defeated decisively by Ricimir in a naval victory off Corsica in 456. The island remained under Roman rule until 469.
	On Which castle was the last stronghold of this monarch?		Powis Castle [SEP] Unlike the case of castles at Conwy, Caernarfon, Harlech and nearby Montgomery which were all built by the English to subdue the Welsh, Powis castle was built by a Welsh prince in the thirteenth century. Following the end of the Welsh Wars (1282) and for his loyalty to Edward I, the King permitted Gruffydd ap Gwenwynwyn to begin building Powis Castle circa 1283.

## 5.2 Quantitative results

In order to analyze the impact of knowledge enhancement on text and multimodal retrieval, we conduct performance evaluation with Precision@K (P@K), Mean Reciprocal Rank (MRR@k), and Hit Rate metrics. Retrieval systems are evaluated on their ability to retrieve the relevant passages i.e., passages containing the ground truth strings. Note that, given a Wikipedia entity, all its aliases are considered ground truth. Table 3 reports the performance of our approach against state-of-the-art baselines, namely BM25 and DPR dual-encoder, for the aforementioned metrics. The Fisher’s randomization test is used for statistical significance tests.

Without surprise, BM25-based sparse retrieval has a worse performance compared to dense retrieval systems, which have the ability to capture more semantic information than traditional IR approaches. In the text-only retrieval setting, our approach performs better than DPR on all metrics, showing that our knowledge-enhanced encoders provide additional discriminative power with richer question and passage representations. It is worth mentioning that our experiments were carried out using a batch size per GPU of 16 compared to 32 used in [18] due to GPU memory limitations. This suggests that some implementation optimizations could potentially lead to greater performance gains, as contrastive learning benefits from more in-batch hard negatives during training. We



also observe that our approach performs best at retrieving (+1.3pt P@1) the top-1 relevant passages, which confirms the benefit of incorporating external knowledge with contextual information.

In the multimodal setting, as pointed out in [18], visual content provides substantial improvement for both BM25 and DPR due to the entity type bias in the dataset. Indeed, their fusion analysis showed that image encoders greatly help to retrieve relevant passages for questions about person entities compared to non-person entity types. In particular, a subset of the ViQuAE KB images overlaps with the MS-Celeb dataset used to pretrain the ArcFace encoder, which allows a straightforward alignment between question and passage image representations. Thus, our approach naturally benefits from this bias, but also further achieves improvement over the DPR-based fusion baseline. This suggests that our knowledge-enhanced approach effectively captures additional information and enriches contextualized representations with structural information. The evaluation results on the reduced KB reported in Table 2 follow the same trend as for the full KB, confirming that the reduced KB is a good proxy.

### 5.3 Qualitative Analysis

Table 4 shows some qualitative examples of top-1 passages correctly retrieved using our approach (with only knowledge-enhanced representations) whereas DPR and image features failed to. These examples illustrate typical cases where question and passage images depict heterogeneous contents. For instance, the first question is about a film in which a character is performing a bungee jump from a dam. It is associated with the photography of the *007 museum* whereas the image of the relevant passage is the flag of a Switzerland canton. More generally, an entity can admit a variety of illustrations, e.g. statues, logos, maps, etc., making visual retrieval difficult when a question and a passage are illustrated very differently. Visual encoders naturally project those image representations in subspaces far from each other, resulting in a visual miss-alignment. DPR also failed to retrieve relevant passages despite its ability to capture lexical variations and contextual information. This is likely due to the high lexical and semantic overlap between the question and many passages in the knowledge base, including the relevant ones. For the first example, the DPR wrongly retrieved a passage about the *Niagara Falls*, which has been a featured location for several movies. Keywords like “film” and “dam” mislead the search for passages that include them. By combining contextual and graph-based representations, our approach can learn additional signals between question vectors enhanced with syntactic information and knowledge-enhanced passage vectors, which helps to discriminate passages with lexical and semantic overlaps. We empirically observe that our approach can better handle entity type variety, which is beneficial for the ViQuAE task.

### 5.4 Limitations

In our approach, we fine-tune the proposed knowledge-enhanced dual encoder on a relatively small dataset, which can rapidly lead to overfitting. A pre-training stage on large QA datasets and knowledge graphs would potentially help to produce richer representations. Moreover, our approach is agnostic to the type of entity

a passage is related to. Some work [7] on comparable tasks suggests that the explicit injection of such knowledge may improve IR performance. Another limitation of our work is the random-based neighbor graph sampling, which is likely to inject noisy information into representations. Future work includes investigating more deterministic neighbor graph-building strategies beyond random sampling, for example by exploring more relevant paths in the knowledge graph that better integrate meaningful information depending on the target passage.

## 6 CONCLUSION AND PERSPECTIVES

We presented an explicit knowledge integration approach for information retrieval on the KVQAE task. We proposed to leverage external resources such as knowledge graphs in order to enhance dense contextual vector representations. In this work, questions being mainly visual with poor entity-level information, we proposed to enhance their representation using syntactic information in the form of dependency parsing trees. On the other hand, entity linking is performed on passages, which allows the building of local subgraphs using an external knowledge graph. Altogether, syntactic dependencies and local knowledge subgraphs are encoded using graph convolutional networks. By combining contextual and graph-based representations, we demonstrated through experiments the benefit of such integration while improving the state-of-the-art IR on the ViQuAE dataset. The proposed method is orthogonal to existing approaches and can be integrated with various architectures. Finally, experiments being computationally expensive, we proposed a reduced version of the original KB that stands as a good proxy that facilitates the experiments.

In this work, we have considered the integration of knowledge at the passage level, which was the most obvious level for such integration due to the number of named entities that can be found in them. One direct extension of the work would be to consider the integration of knowledge at the level of questions as well. The number of named entities in questions is much lower than in passages but questions also include references to two “ghost” entities that we can exploit through their type: the target of the question and the entity of the image associated with the question. The knowledge brought by a KG can also be exploited more indirectly for taking into account the fact that an entity, especially abstract entities, can be represented visually in many different ways, as for the example about a film in Section 5.3. More precisely, the relations between entities in a KG can be used for enlarging selectively the set of images associated with an entity, which can be viewed as a form of predictive visual semantic expansion.

## ACKNOWLEDGMENTS

This work was supported by the ANR-19-CE23-0028 MEERQAT project. This work was granted access to the HPC resources of IDRIS under the allocation 2022-AD011013719 made by GENCI. It also relied on the use of the FactoryIA cluster, financially supported by the Ile-de-France Regional Council.

## REFERENCES

- [1] Omar Adjali, Romaric Besançon, Olivier Ferret, Hervé Le Borgne, and Brigitte Grau. 2020. Building a Multimodal Entity Linking Dataset From Tweets. In

- International Conference on Language Resources and Evaluation (LREC)*. European Language Resources Association, Marseille, France.
- [2] Omar Adjali, Romaric Besançon, Olivier Ferret, Herve Le Borgne, and Brigitte Grau. 2020. Multimodal Entity Linking for Tweets. In *European Conference on Information Retrieval (ECIR)*. Springer, Lisbon, Portugal.
  - [3] Stanislaw Antol, Aishwarya Agrawal, Jiaseen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. In *IEEE International Conference on Computer Vision, ICCV December 7-13*. IEEE Computer Society, Santiago, Chile, 2425–2433.
  - [4] Prajwal Bhargava and Vincent Ng. 2022. Commonsense Knowledge Reasoning and Generation with Pre-trained Language Models: A Survey. *arXiv preprint arXiv:2201.12438 abs/2201.12438* (2022).
  - [5] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. 2019. ArcFace: Additive Angular Margin Loss for Deep Face Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR June 16-20*. Computer Vision Foundation / IEEE, Long Beach, CA, USA, 4690–4699. <https://doi.org/10.1109/CVPR.2019.00482>
  - [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)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. <https://doi.org/10.18653/v1/N19-1423>
  - [7] Xiangyu Dong, Wenhao Yu, Chenguang Zhu, and Meng Jiang. 2020. Injecting entity types into entity-guided text generation. *arXiv:2009.13401* (2020).
  - [8] Nouha Dziri, Sivan Milton, Mo Yu, Osmar Zaiane, and Siva Reddy. 2022. On the Origin of Hallucinations in Conversational Models: Is it the Datasets or the Models?. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Seattle, United States, 5271–5285. <https://doi.org/10.18653/v1/2022.naacl-main.387>
  - [9] Matthias Fey and Jan E. Lenssen. 2019. Fast Graph Representation Learning with PyTorch Geometric. In *ICLR Workshop on Representation Learning on Graphs and Manifolds*. MIT Press, New Orleans, LA, USA.
  - [10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR, June 27-30*. IEEE Computer Society, Las Vegas, NV, USA, 770–778. <https://doi.org/10.1109/CVPR.2016.90>
  - [11] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event (Proceedings of Machine Learning Research, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). PMLR, 4904–4916. <http://proceedings.mlr.press/v139/jia21b.html>
  - [12] Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data* 7, 3 (2019), 535–547.
  - [13] Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Vancouver, Canada, 1601–1611. <https://doi.org/10.18653/v1/P17-1147>
  - [14] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense Passage Retrieval for Open-Domain Question Answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Online, 6769–6781. <https://doi.org/10.18653/v1/2020.emnlp-main.550>
  - [15] Pei Ke, Haozhe Ji, Siyang Liu, Xiaoyan Zhu, and Minlie Huang. 2020. SentiLARE: Sentiment-Aware Language Representation Learning with Linguistic Knowledge. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Online, 6975–6988. <https://doi.org/10.18653/v1/2020.emnlp-main.567>
  - [16] Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July (Proceedings of Machine Learning Research, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). PMLR, Virtual Event, 5583–5594. <http://proceedings.mlr.press/v139/kim21k.html>
  - [17] Thomas N. Kipf and Max Welling. 2017. Semi-Supervised Classification with Graph Convolutional Networks. In *5th International Conference on Learning Representations ICLR April 24-26, Conference Track Proceedings*. OpenReview.net, Toulon, France. <https://openreview.net/forum?id=SJU4ayYgl>
  - [18] Paul Lerner, Olivier Ferret, Camille Guinaudeau, Hervé Le Borgne, Romaric Besançon, José G Moreno, and Jesús Lovón Melgarejo. 2022. ViQuAE, a dataset for knowledge-based visual question answering about named entities. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. Association for Computing Machinery, Madrid, Spain, 3108–3120.
  - [19] Bei Li, Chuanhao Lv, Zefan Zhou, Tao Zhou, Tong Xiao, Anxiang Ma, and JingBo Zhu. 2022. On Vision Features in Multimodal Machine Translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Dublin, Ireland, 6327–6337. <https://doi.org/10.18653/v1/2022.acl-long.438>
  - [20] Belinda Z. Li, Sewon Min, Srinivasan Iyer, Yashar Mehdad, and Wen-tau Yih. 2020. Efficient One-Pass End-to-End Entity Linking for Questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Online, 6433–6441. <https://doi.org/10.18653/v1/2020.emnlp-main.522>
  - [21] Robert Logan, Nelson F. Liu, Matthew E. Peters, Matt Gardner, and Sameer Singh. 2019. Barack’s Wife Hillary: Using Knowledge Graphs for Fact-Aware Language Modeling. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 5962–5971. <https://doi.org/10.18653/v1/P19-1598>
  - [22] Jiaseen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems (December 8-14)*, Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d’Alché-Buc, Emily B. Fox, and Roman Garnett (Eds.). Neural Information Processing Systems Foundation, Inc., Vancouver, BC, Canada, 13–23.
  - [23] Kaixin Ma, Hao Cheng, Xiaodong Liu, Eric Nyberg, and Jianfeng Gao. 2022. Open Domain Question Answering with A Unified Knowledge Interface. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Dublin, Ireland, 1605–1620. <https://doi.org/10.18653/v1/2022.acl-long.113>
  - [24] Mateusz Malinowski and Mario Fritz. 2014. A Multi-World Approach to Question Answering about Real-World Scenes based on Uncertain Input. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13*, Zoubin Ghahramani, Max Welling, Corinna Cortes, Neil D. Lawrence, and Kilian Q. Weinberger (Eds.). MIT Press, Montreal, Quebec, Canada, 1682–1690. <https://proceedings.neurips.cc/paper/2014/hash/d516b13671a4179d9b7b458a6ebdeb92-Abstract.html>
  - [25] Diego Marcheggiani and Ivan Titov. 2017. Encoding Sentences with Graph Convolutional Networks for Semantic Role Labeling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Copenhagen, Denmark, 1506–1515. <https://doi.org/10.18653/v1/D17-1159>
  - [26] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. OK-VQA: A Visual Question Answering Benchmark Requiring External Knowledge. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Computer Vision Foundation / IEEE, 3195–3204. <https://doi.org/10.1109/CVPR.2019.00331>
  - [27] Seungwhan Moon, Leonardo Neves, and Vitor Carvalho. 2018. Multimodal Named Entity Recognition for Short Social Media Posts. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics, New Orleans, Louisiana, 852–860. <https://doi.org/10.18653/v1/N18-1078>
  - [28] Débora Myoupo, Adrian Popescu, Hervé Le Borgne, and Pierre-Alain Moëllic. 2010. Multimodal image retrieval over a large database. In *Proceedings of the 10th international conference on Cross-language evaluation forum: multimedia experiments (Lecture Notes in Computer Science)*, Carol Peters, Barbara Caputo, Julio Gonzalo, Gareth J.F. Jones, and Jayashree Kalpathy-Cramer (Eds.). Springer Berlin / Heidelberg, Berlin, Heidelberg, 177–184.
  - [29] Medhini Narasimhan, Svetlana Lazebnik, and Alexander G. Schwing. 2018. Out of the Box: Reasoning with Graph Convolution Nets for Factual Visual Question Answering. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8*, Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett (Eds.). MIT Press, Montréal, Canada, 2659–2670.
  - [30] Medhini Narasimhan and Alexander G Schwing. 2018. Straight to the facts: Learning knowledge base retrieval for factual visual question answering. In *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, Munich, Germany, 451–468.
  - [31] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in pytorch. In *NIPS 2017 Workshop on Autodiff*. MIT Press, Long Beach, CA, USA.
  - [32] Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language Models as Knowledge Bases?. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 2463–2473. <https://doi.org/10.18653/v1/D19-1250>

- [33] 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 *Proceedings of the 38th International Conference on Machine Learning, ICML (18-24 July) (Proceedings of Machine Learning Research, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). PMLR, Virtual Event, 8748–8763. <http://proceedings.mlr.press/v139/radford21a.html>
- [34] Hannah Rashkin, David Reitter, Gaurav Singh Tomar, and Dipanjan Das. 2021. Increasing Faithfulness in Knowledge-Grounded Dialogue with Controllable Features. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, Online, 704–718. <https://doi.org/10.18653/v1/2021.acl-long.58>
- [35] Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How Much Knowledge Can You Pack Into the Parameters of a Language Model?. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Online, 5418–5426. <https://doi.org/10.18653/v1/2020.emnlp-main.437>
- [36] Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne Van Den Berg, Ivan Titov, and Max Welling. 2018. Modeling relational data with graph convolutional networks. In *The Semantic Web: 15th International Conference, ESWC 2018, Heraklion, Crete, Greece, June 3–7, 2018, Proceedings 15*. Springer, 593–607.
- [37] Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. A-OKVQA: A Benchmark for Visual Question Answering using World Knowledge. *arXiv preprint arXiv:2206.01718* abs/2206.01718 (2022). <https://arxiv.org/abs/2206.01718>
- [38] Sanket Shah, Anand Mishra, Naganand Yadati, and Partha Pratim Talukdar. 2019. KVQA: Knowledge-Aware Visual Question Answering. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI (January 27 - February 1)*. AAAI Press, Honolulu, Hawaii, USA, 8876–8884.
- [39] Haoyu Song, Li Dong, Weinan Zhang, Ting Liu, and Furu Wei. 2022. CLIP Models are Few-Shot Learners: Empirical Studies on VQA and Visual Entailment. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Dublin, Ireland, 6088–6100. <https://doi.org/10.18653/v1/2022.acl-long.421>
- [40] Haitian Sun, Bhuvan Dhingra, Manzil Zaheer, Kathryn Mazaitis, Ruslan Salakhutdinov, and William Cohen. 2018. Open Domain Question Answering Using Early Fusion of Knowledge Bases and Text. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Brussels, Belgium, 4231–4242. <https://doi.org/10.18653/v1/D18-1455>
- [41] Yu Sun, Shuohuan Wang, Yu-Kun Li, Shikun Feng, Hao Tian, Hua Wu, and Haifeng Wang. 2020. ERNIE 2.0: A Continual Pre-Training Framework for Language Understanding. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*. AAAI Press, 8968–8975.
- [42] Hao Tan and Mohit Bansal. 2019. LXMERT: Learning Cross-Modality Encoder Representations from Transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 5100–5111. <https://doi.org/10.18653/v1/D19-1514>
- [43] Thi Quynh Nhi Tran, Hervé Le Borgne, and Michel Crucianu. 2016. Aggregating Image and Text Quantized Correlated Components. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Las Vegas, USA.
- [44] Shikhar Vashishth, Soumya Sanyal, Vikram Nitin, and Partha Talukdar. 2019. Composition-based multi-relational graph convolutional networks. *arXiv preprint arXiv:1911.03082* (2019).
- [45] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (Eds.)*. Long Beach, CA, USA, 5998–6008.
- [46] Liwei Wang, Yin Li, and Svetlana Lazebnik. 2016. Learning Deep Structure-Preserving Image-Text Embeddings. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 5005–5013. <https://doi.org/10.1109/CVPR.2016.541>
- [47] 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 (2017), 2413–2427.
- [48] Peng Wang, Qi Wu, Chunhua Shen, Anthony R. Dick, and Anton van den Hengel. 2017. Explicit Knowledge-based Reasoning for Visual Question Answering. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI, August 19-25, Carles Sierra (Ed.)*. ijcai.org, Melbourne, Australia, 1290–1296. <https://doi.org/10.24963/ijcai.2017/179>
- [49] Xiaozhi Wang, Tianyu Gao, Zhaocheng Zhu, Zhengyan Zhang, Zhiyuan Liu, Juanzi Li, and Jian Tang. 2021. KEPLER: A Unified Model for Knowledge Embedding and Pre-trained Language Representation. *Transactions of the Association for Computational Linguistics* 9 (2021), 176–194. [https://doi.org/10.1162/tacl\\_a\\_00360](https://doi.org/10.1162/tacl_a_00360)
- [50] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi 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 Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, Online, 38–45. <https://doi.org/10.18653/v1/2020.emnlp-demos.6>
- [51] 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 *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, 4622–4630. <https://doi.org/10.1109/CVPR.2016.500>
- [52] Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto. 2020. LUKE: Deep Contextualized Entity Representations with Entity-aware Self-attention. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Online, 6442–6454. <https://doi.org/10.18653/v1/2020.emnlp-main.523>
- [53] Jian Yang, Gang Xiao, Yulong Shen, Wei Jiang, Xinyu Hu, Ying Zhang, and Jinghui Peng. 2021. A survey of knowledge enhanced pre-trained models. *arXiv:2110.00269* (2021).
- [54] Wei Yang, Yuqing Xie, Aileen Lin, Xingyu Li, Luchen Tan, Kun Xiong, Ming Li, and Jimmy Lin. 2019. End-to-End Open-Domain Question Answering with BERTserini. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*. Association for Computational Linguistics, Minneapolis, Minnesota, 72–77. <https://doi.org/10.18653/v1/N19-4013>
- [55] Yi Yang, Wen-tau Yih, and Christopher Meek. 2015. WikiQA: A Challenge Dataset for Open-Domain Question Answering. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Lisbon, Portugal, 2013–2018. <https://doi.org/10.18653/v1/D15-1237>
- [56] Xuchen Yao and Benjamin Van Durme. 2014. Information Extraction over Structured Data: Question Answering with Freebase. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Baltimore, Maryland, 956–966. <https://doi.org/10.3115/v1/P14-1090>
- [57] Jianfei Yu, Jing Jiang, Li Yang, and Rui Xia. 2020. Improving Multimodal Named Entity Recognition via Entity Span Detection with Unified Multimodal Transformer. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 3342–3352. <https://doi.org/10.18653/v1/2020.acl-main.306>
- [58] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. 2022. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917* abs/2205.01917 (2022).
- [59] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. 2016. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE signal processing letters* 23, 10 (2016), 1499–1503.