Knowledge Aware Question Answering: A Survey

Manas Jhalani, Annervaz. K.M, Pushpak Bhattacharyya
{manasj, pb}@cse.iitb.ac.in, annervaz@gmail.com

Abstract

Question answering involves models responding to natural language questions, while knowledge-infused question answering expands on this by incorporating external knowledge from a predefined source. Another multimodal task, Visual question answering, deals with questions grounded in visual content, necessitating models to access external context from diverse sources. Knowledge-based visual question answering further extends this by integrating external knowledge and images to generate responses. Knowledge graphs are valuable sources for extracting structured information and consolidating data from various sources. This survey paper overviews different question-answering frameworks leveraging knowledge graphs as an external context. It also discusses potential models for constructing such frameworks and outlines previously proposed datasets. Finally, we highlight gaps in current research that could be addressed to advance this field.

1 Introduction

Question answering is a very essential task in natural language processing which aims to provide correct answers to a question base. It aims to use NLP technologies to generate a corresponding answer to a given question based on the massive unstructured corpus.QA is a traditional research direction that has been proposed half a century ago. People hope to help with everyday life by teaching the program how to answer questions like a real person. This paper delves into two primary domains of QA: Textual Question Answering and Visual Question Answering.

In Textual Question Answering, models generate answers by leveraging external contexts from various sources like knowledge graphs, and Wikipedia sentences, among others. These questions can be tackled using conventional large language models such as GPT3.5, LLAVA, T5, etc. Recently, Multimodal LLMs like LLAVA and GPT4, incorporating images alongside text, have emerged to enhance task-specific performance.

Visual Question Answering addresses natural language questions grounded in visual content. A specific variant, Knowledge-based VQA, not only utilizes visual information extracted from images but also integrates supportive facts to facilitate accurate reasoning and answer prediction. KB-VQA questions are categorized into closed-domain KBVQA, requiring knowledge from predefined knowledge bases like Wikidata or Freebase, and open-domain KBVQA, where no fixed knowledge base is mandated, and questions may demand varying degrees of knowledge.

Recent methodologies demonstrate the efficacy of external knowledge retrieval coupled with extensive filtering to increase accuracy significantly. Prior research has explored diverse knowledge sources and filtering mechanisms. This paper provides an overview of these approaches and identifies future research avenues. Additionally, it discusses open-source datasets vital for training state-of-the-art models in these domains.

2 Motivation

Large language models like ChatGPT and GPT-4 indeed store a vast amount of information within their parameters, enabling them to answer knowledge-aware questions effectively. However, these models have limitations. They are large in terms of parameters and may lack domain-specific knowledge. Additionally, some models are paid, restricting their accessibility.

To address this, external knowledge sources become crucial. Knowledge graphs such as Wikidata, Freebase, and WordNet are valuable because they are frequently updated and freely available. These external resources complement models with fewer parameters, allowing them to provide accurate answers to a wider range of questions.

Regarding visual question-answering models like GPT-4 and Gemini, they face challenges when handling user-centric questions. For example, questions like "Who is the person in the middle of the image?" or "What is the age of the person shown in the image?" require specific answers related to named entities. Generic responses like "man" or "I can't guess the age" are insufficient. To improve performance, fetching external context from knowledge bases becomes essential, especially in real-time applications where user-centric data matters.

3 Knowledge Graphs

3.1 KG Embeddings

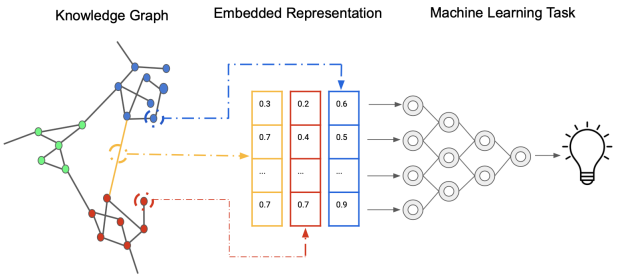

Figure 1: KG Embeddings are trained with the objective of link prediction, Using KGE in a deep learning network allows the DL model to do multihop reasoning over graph

Knowledge graph embeddings are a type of representation learning that maps entities and relations in a knowledge graph to a continuous vector space. This allows us to represent the meaning of entities and relations in a way that can be used by machine learning algorithms. There are two main types of knowledge graph embeddings: translation-based and tensor factorization-based.

1. **Translation-based embeddings** use a scoring function that measures the distance between the embedding of the head entity, the embedding of the relation, and the embedding of the tail entity. The goal is to minimize the distance between the head entity and the tail entity when they are connected by the relation. Translation-based embeddings are relatively simple to train and can be effective for a variety of tasks. However, they can be sensitive to noise in the data and may not be able to capture complex relationships between entities.

2. **Tensor factorization-based embeddings** use a tensor factorization model to learn the embedding of entities and relations. The tensor factorization model is trained to predict the existence of a relation between two entities. Tensor factorization-based embeddings are more complex to train than translation-based embeddings, but they can be more effective for capturing complex relationships between entities. However, they can be more computationally expensive to train and may not be as effective for tasks that require real-time inference.

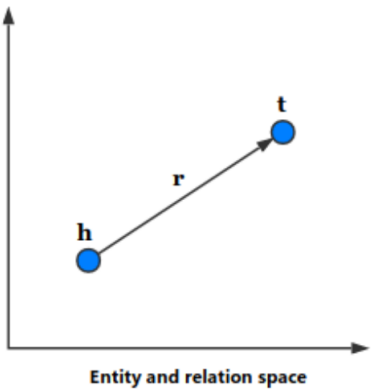

Figure 2: Entities and Relations are represented as vectors, relation is a translation from one entity to another. (Dai et al., 2020)

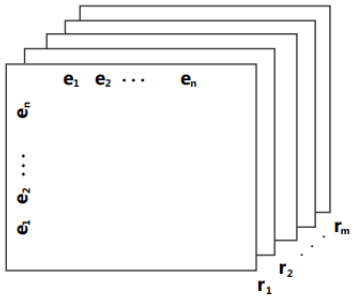

Figure 3: Knowledge Graph represented as tensor, where entities are two of the dimensions and the third dimension is relation (Dai et al., 2020)

Some popular translation-based embeddings are TransE (Bordes et al., 2013),TransH (Wang et al., 2014), TransR (Lin et al., 2015), and TransD (Ji et al., 2015). Some popular tensor factorization-based embeddings are RESCAL (Nickel et al., 2011b), DistMult (Nickel et al., 2011a), and ComplEx (Trouillon et al., 2016). Knowledge graph embeddings are effective for a variety of tasks, including link prediction and question answering.

3.1.1 Translation Based Models

Translation-based Knowledge Graph Embedding models represent relation as a translation from a head entity to a tail entity in an embedding space. The proposed translation-based models vary in the space these elements are projected. The following are some translation-based models:

1. TransE: TransE is the first Knowledge Graph Embedding method which translates head by the relation to reach the tail entity keeping both head and tail entities in the same dimension i.e. k=d. The figure 4 shows the translation from head to tail using a relation.
   The scoring function of TransE is f_r(h,t) = $||\mathbf{h} + \mathbf{r} + \mathbf{t}||_{l_1/l_2}$. The model complexity of TransE is $\mathcal{O}(N_e d + N_r k)(d = k)$. TransE fails to model the one-to-many, many-to-one and many-to-many relationship between entities. It can't model symmetric and reflexive relations.

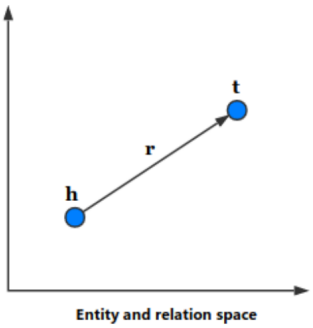

Figure 4: TransE: Translation over hyperplane, here **r** is translation from **h** to **t**.

2. TransH: TransH projects the head and tail entities onto a relation-specific hyperplane and then uses translation on the hyperplane to translate the projected head to obtain the projected tail entity. Since multiple entities can be projected to the same point on the hyperplane, it can model one-to-many, many-to-one, and many-to-many relations.
   The scoring function of TransH is f_r(h,t) = $||(\mathbf{h} - \mathbf{w}_r^T \mathbf{h} \mathbf{w}_r) + \mathbf{d}_r - (\mathbf{t} - \mathbf{w}_r^T \mathbf{t} \mathbf{w}_r)||_2^2$. The model complexity of TransH is $\mathcal{O}(N_e d + N_r k)(d = k)$. The ability to model n-ary relationship is limited due to the projection onto a hyperplane in the same space as entities.

3. TransR: TransR projects the head and tail entities from a k-dimensional space to a d-dimensional space. Projection into a separate space allows TransR to model more n-ary relations. The projection is performed using a relation-specific projection matrix.
   The scoring function of TransR is f_r(h,t) = $||(\mathbf{M}_r \mathbf{h} + \mathbf{r} - \mathbf{M}_r \mathbf{t}||_2^2$. The model complexity of TransR is $\mathcal{O}(N_e d + N_r dk)$. TransR increased the complexity of parameters due to the projection matrix $\mathbf{M}_r$.

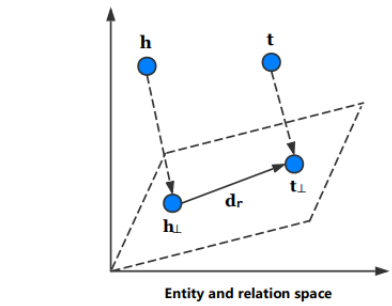

Figure 5: TransH: Entities are first projected onto hyperplane and then translated over it.

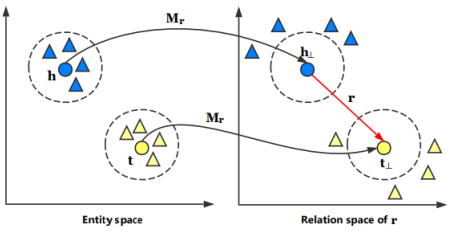

Figure 6: TransR: Projection matrices are used to project entities to relation space.

4. TransD: TransD projects the head and tail entities from a k-dimensional space to a d-dimensional space. Unlike TransR, TransD uses a separate projection matrix for head and tail entities. TransD reduces the number of parameters by obtaining the projection matrices using vector multiplication.
   The projection matrix for head entity is obtained using vectors $\mathbf{r}_p$ $\mathbf{h}_p^T$ for tail using vectors $\mathbf{r}_p$ and $\mathbf{t}_p^T$. The scoring function of TransD is f_r(h,t) = $||(\mathbf{r}_p \mathbf{h}_p^T + \mathbf{I})\mathbf{h} + \mathbf{r} - (\mathbf{r}_p \mathbf{t}_p^T + \mathbf{I})\mathbf{t}||_2^2$. The model complexity of TransD is $\mathcal{O}(N_e d + N_r k)$.

3.1.2 Tensor Factorization Based Models

Tensor factorization-based methods represent entities and the relation between them as a tensor, as shown in figure 8. The tensor representing the

Datasets	WN18				FB15K			
Metric	Mean Rank		HITS@10(%)		Mean Rank		Hits@10(%)	
	Raw	Filter	Raw	Filter	Raw	Filter	Raw	Filter
TransE	263	251	75.4	89.2	243	125	34.9	47.1
TransH	401	388	73.0	82.3	212	87	45.7	64.4
TransR	238	225	79.8	92.0	198	77	48.2	68.7
TransD	224	212	79.6	92.2	194	91	53.4	77.3
RESCAL	1180	1163	37.2	52.8	828	683	28.4	44.1
DistMult -	-	-	-	94.2	-	-	-	58.5
HOLE	-	-	-	94.9	-	-	-	73.9
ComplEx	-	-	-	94.7	-	-	-	84.0

Table 1: Evaluation of Knowledge Graph Embedding Models on WN18 and FB15K datasets

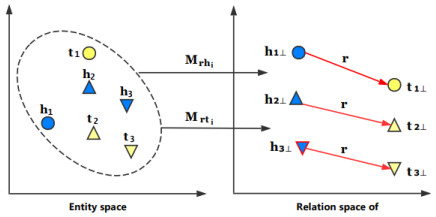

Figure 7: TransD: Sparse projection matrices are used to project entities to relation space. The Sparse projection matrices can be represented as the multiplication of two vectors forming a projection matrix.

knowledge graph is a three-dimensional binary matrix X ϵ $R^{n,n,m}$ where n is the number of entities and m is the number of relations. Tensor factorization-based methods decompose the KG tensor into a multiplication of factors as entities and relations. The following are some Tensor-factorization-based methods:

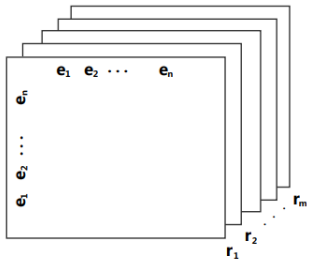

Figure 8: Entities and relation represented as a three-dimensional tensor

1. RESCAL: RESCAL expresses the tensor into representation of head, relation and tail.
   The scoring function of RESCAL is f_r(h,t) = $\mathbf{h}^T\mathbf{M}_r\mathbf{t}$. The model complexity of RESCAL is $\mathcal{O}(N_e d + N_r k^2)(d = k)$. The model complexity of RESCAL is quadratic in k.

2. DistMult: DistMult restricts the relation matrix to a diagonal matrix, hence reducing the number of parameters required by a relation. It allows symmetric relations between head and tail entities.
   The scoring function of DistMult is f_r(h,t) = $\mathbf{h}^T$diag(**r**)**t**. The model complexity of DistMult is $\mathcal{O}(N_e d + N_r k)(d = k)$.

3. HolE: HolE simplifies the tensor product by introducing circular correlation. The circular correlation operation is indicated using $\star$. For two entities, it can be calculated as $[\mathbf{h}\star\mathbf{t}]_k = \sum_{i=0}^{d-1} h_i t_{(k+i) mod d}$. The operation is asymmetric which allows the model to represent asymmetric relations. Also, the complexity can be further improved using fast fourier transform(FFT) using $h \star t = \mathcal{F}^{-1}(\overline{\mathcal{F}(\mathbf{h})} \odot \mathcal{F}(\mathbf{t}))$.
   The scoring function of HolE is f_r(h,t) = $\mathbf{r}^T(\mathbf{h}\star\mathbf{t})$. The model complexity of HolE is $\mathcal{O}(N_e d + N_r k)(d = k)$.

4. ComplEx: ComplEx uses embeddings from complex space i.e. C^d. Because of the complex space, along with symmetric, asymmetric relations can be modeled.
   The scoring function of ComplEx is f_r(h,t) = Re($\mathbf{h}^T$diag(**r**)$\bar{\mathbf{t}}$) where $\bar{\mathbf{t}}$ is complex conjugate of t and Re() returns the real part of a complex value. The model complexity of ComplEx is $\mathcal{O}(N_e d + N_r k)(d = k)$.

3.2 Benchmark Knowledge Graphs

A knowledge graph (KG) is a structured collection of data that represents entities and their relationships. KGs are used to store and organize information about the real world, and they can be used

for a variety of tasks, such as question-answering, natural language processing, and machine learning. In this section, we will discuss popularly used KGs for experiments from different domains and provide statistics on their sizes.

3.2.1 WordNet

WordNet (Silva, 2019) is a lexical database that organizes words into sets of synonyms called synsets, each representing a distinct concept. Synsets are interlinked using conceptual-semantic and lexical relations, such as "is-a" (hypernymy) and "part-of" (meronymy). Hypernymy and hyponymy are two of the most important semantic relations in WordNet. Hypernymy is a "is-a" relationship, and it indicates that one concept is a more general concept than another. For example, the word "apartment" is a hyponym of the word "dwelling," because an apartment is a type of dwelling. Hyponymy is the opposite of hypernymy, and it indicates that one concept is a more specific concept than another. WordNet can be viewed as a knowledge graph (KG), which is a structured collection of information about entities and their relationships. In WordNet, the entities are words, and the relationships are the semantic relations between them. For example, the semantic relation "is-a" can be viewed as a relationship between two entities, where one entity is a more general concept and the other entity is a more specific concept. WN18 is a benchmark dataset for evaluating the performance of systems that work with knowledge graphs. WN18 contains 18 relations, 40K entities and 151K triples. The triples in WN18 are extracted from WordNet, and they represent the semantic relations between words in WordNet. Following are a few examples from WN18: The head and tail entities are in the format entity(dot)part of speech(dot)sense. Let us understand the meaning of the triple <future.n.01, hypernym, time.n.05>. The head entity, future.n.01, is a noun with the sense of "the time yet to come." The tail entity, time.n.05, is a noun with the sense of "the continuum of experience in which events pass from the future through the present to the past." The relation between the two entities is hypernym, which means that future.n.01 is a type of time.n.05. In other words, the future is a part of time. The sense means of each entity can be understood using Stanford's WordNet Search

3.2.2 Wikidata

Wikidata is a free and open knowledge base that anyone can edit. It contains information about people, places, things, and events, and is used by a wide variety of applications, including search engines, virtual assistants, and educational tools. Wikidata5m is a subset of Wikidata that contains 5 million entities and their associated properties. It was created by the MilaGraph team at the University of Montreal and is used for research in knowledge graph embedding and natural language processing. Wikidata5M has 822 relations and 20 million KG triples.The knowledge graph is stored in the triplet list format, where each line corresponds to a triple of entity, relation, and value. For example, the triple <Q22686, P39, Q11696> corresponds to the triple Donald Trump position held as President of the United States. The corpus is a collection of documents, indexed by entity ID. Each document describes the entity. For example, the document for Donald Trump is Q22686 Donald John Trump (born June 14, 1946) is the 45th and current president of the United States .... The aliases file lists the aliases for entities and relations. For example, the line Q22686 Donnie Trump 45th President of the united states Donald John Trump ... lists the aliases for Donald Trump.

3.2.3 Freebase

Freebase (Bollacker et al., 2008) was a large-scale knowledge base that was acquired by Google in 2010. Freebase contained information about a wide variety of topics, including people, places, things, and events. Freebase is no longer available, but its data has been used to create other KG datasets, such as Google Knowledge Graph. FB15k and FB15k-237 are knowledge graph datasets that are based on the Freebase knowledge base. They are commonly used as benchmarks for evaluating the performance of knowledge graph embedding models. FB15k contains 592,213 triples with 14,951 entities and 1,345 relationships. FB15k-237 is a subset of FB15k that contains 237 relationships. This was done to reduce the number of inverse relations in the dataset, as it was found that a large number of test triplets could be obtained by inverting triplets in the training set.

4 Visual Question Answering

Visual Question Answering (VQA) is a challenging task that combines computer vision and natural

language processing. In VQA, a system is tasked with answering questions related to an image based on its content. These systems find practical applications in assisting visually impaired individuals and improving image search capabilities for IoT devices like smart hubs. There are two main types of questions in VQA:

Image-Only Questions: These questions can be answered using only the features extracted from the image itself. For example, a question like "Who is to the right of R. Madhavan?" falls into this category.

Knowledge-Based Questions: These questions require external knowledge beyond just understanding image features. For instance, questions like "In which country was the person in the image born?" fall into this category.

State-of-the-art multimodal language models are adept at accurately answering Image-only questions, even without relying on external information. However, knowledge-based questions necessitate additional context beyond what the image provides. Several open-source datasets are available for knowledge-based Visual Question Answering (VQA). These datasets play a crucial role in advancing research in this field as shown in Table.

5 Knowledge Aware Visual Question Answering

Knowledge-aware visual question answering (VQA) seeks to answer questions that require external knowledge beyond what images alone can provide. In general, works in this field fetch relevant knowledge from external sources to answer questions. Knowledge can be obtained from various sources, such as knowledge graphs, fixed multimodal knowledge bases and many more. Additionally, recent approaches leverage the parameters of current large language models (LLMs), utilizing prompting techniques to extract relevant knowledge for answering questions. Broadly, KB-VQA methods can be categorized into two main types:

5.1 Open Domain KB-VQA

Open-domain KB-VQA require knowledge from an open knowledge base instead of a fixed knowledge base i.e. the knowledge required to answer these questions is not confined to a particular knowledge base. The datasets released for this task are shown in Table 3. This section will discuss some of the latest works for the OD-KBVQA task.

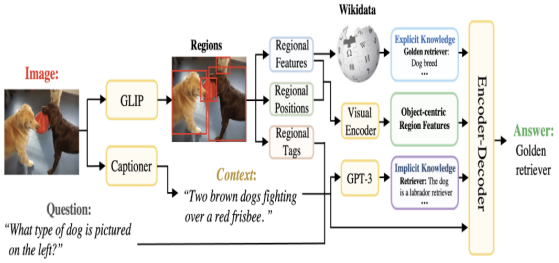

Figure 9: Splitting the image into four patches to extract relevant triples.

REVIVE (Lin et al., 2024): Previous works have focused on extracting external information relevant to objects in the image. These approaches typically use either the whole image or a sliding window to retrieve knowledge. In this paper, the authors revisit visual representation in knowledge-based Visual Question Answering (VQA) and argue that the information from object regions and their relationships should be considered and utilized in a more dedicated way. Therefore, they propose REVIVE to better utilize Regional Visual Representation for knowledge-based VQA. REVIVE not only exploits detailed regional information for enhanced knowledge retrieval but also integrates this regional visual representation into the final answering model.

The architecture of their approach is illustrated in the figure and is divided into three parts:

Regional Feature Extraction Module: The authors employ the GLIP model (Li et al., 2022b) to extract object-specific coordinates in the image. They then use the CLIP model (Li et al., 2022b) to find the most similar tag for each object and compute captions using the VinVL (Zhang et al., 2021) model to determine relationships among the objects.

Object-Centric Knowledge Retrieval Module: This stage involves both explicit and implicit knowledge retrieval. For explicit knowledge retrieval, the authors extract the top-K relevant entries from the WikiData knowledge base. For implicit knowledge retrieval, they use context-aware prompts with regional descriptions, providing overall context and regional tags to generate comprehensive descriptions.

Transformer Encoder-Decoder Model: The regional features and object-centric knowledge retrieval results are fed into this model to produce the final output.

Dataset	Answer-Type	Size	Domain	Evaluate Ability
CLEVR	Open-ended	853K	3D CG	Reasoning
RecipeQA	Multi-Choice	36K	Cooking Recipes	Procedural
CRIC	Open-ended	494K	Visual Genome	Scene Reasoning
DocVQA	Open-ended	50,000	Document	Recognition
FVQA	Open-ended	5,826	Open Domain	Knowledge
Visual Genome	Open-ended	1,445,322	Open Domain	Recognition
VCR	Multi-choice	290K	Movie	Reasoning
GQA	Open-ended/Yes/No	22M	Visual Genome	Reasoning
HowMany-QA	Number	106,356	VG/VQA2.0	Counting
TallyQA	Number	287,907	VG/COCO	Counting
TDIUC	Open-ended	1.6M	VG/COCO	Multiple
TextVQA	Open-ended	45,336	Open Domain	Text Recognition
VCOPA	Multi-choice	380	Open Domain	Causality
Visual7W	Open-ended/Multi-choice	327,939	VG	Reasoning
VizWiz	Open-ended	31,000	Photo	Recognition
VQA2.0	Open-ended	1.11M	COCO	Recognition
KVQA	Open-ended	183,007	Wikipedia	Knowledge
OK-VQA	Open-ended	14,000	Open Domain	Knowledge
R-VQA	Open-ended	478,287	VG	Reasoning
KB-VQA	Open-ended	2,402	COCO/ImageNet	Knowledge
WebQA	Open-ended	25K	Wikipedia	Multi-hop
AQUA	Open-ended	79,848	Art	Knowledge
IndiFoodVQA	Multi-Choice	16,716	Food	Knowledge

Table 2: **Statistics of VQA datasets. RC means Reading Comprehension, MC means Multi-choice.**

Generate Then Select (Fu et al., 2023): Previous works discuss retrieving from knowledge graphs with the results being input to an answer generation model. Motivated by PLMs such as GPT3 more recent approaches PiCA (Yang et al., 2022) and KAT (Gui et al., 2022) propose to retrieve from GPT-3 and achieve better performance for their neat and high-quality knowledge. While these methods achieve SOTA the two models suffer from low knowledge coverage caused by PLM bias, the tendency to generate certain tokens over other tokens despite the prompt changes and the performance depends on PLM quality.

The authors solved this problem by proposing a two-stage pipeline to generate the answers.

Multiple choice generation: The authors draw on (Yang et al., 2022) and (Gui et al., 2022) for the methodology, using few-shot in-context learning with a frozen PLM to generate answer choices for image-question pairs. Each image is converted to a textual context via a captioning model and tags from Microsoft Azure. Authors created prompts with context and few-shot examples, using CLIP embeddings to select 16-shot examples. The PLM generates outputs which are combined to form the final answer choices for each pair.

Answer Selection: To train the model for selecting an answer, authors first generate Chain-of-Thought

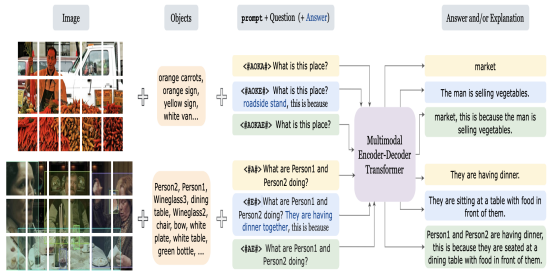

Figure 10: Splitting the image into four patches to extract relevant triples.

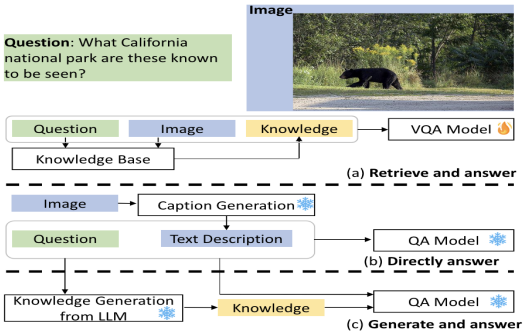

Figure 11: Splitting the image into four patches to extract relevant triples.

(CoT) rationales to guide the selection process. Using a fixed prompt, they created CoT rationales as per (Wei et al., 2023) and (Schwenk et al., 2022). Then the input for the answer selection model by concatenating the question, the image representation or its CLIP embedding, the CoT rationale, and the generated answer choices, formatted with sentinel tokens.

UMAE (Whitehouse et al., 2023): The authors proposed a multitask learning approach for multimodal transformer-based encoder-decoder models, towards a United Model for Answer and Explanation generation (UMAE). Previous approaches have a separate answer prediction and explanation module based on answers, authors add the capability of jointly generating answers and explanations together. The authors divided the answer prediction and explanation into two modules:

Multitask Learning with Artificial Prompt: The authors propose three generation settings for VQA: answer prediction (Q→A), explanation generation (QA→E), and joint answer-explanation generation (Q→AE). Using a pre-trained multimodal transformer (OFA), enhance it with object and attribute extraction for improved open-domain VQA performance. Authors also incorporate artificial prompts to signal tasks and mix training instances to the ground, aligning generated answers and explanations.

Perplexity as Multiple Choice Metric: Instead of loosely matching predictions to multiple-choice options using embedding similarity methods like GloVe, authors evaluate each option as a text-generation task. By providing the model with the same information used for generating answers, they calculate the likelihood of each option's tokens being generated. Then compared their approach's performance, measured by perplexity, against the GloVe embedding similarity method for A-OKVQA (Marino et al., 2019).

5.2 Closed Domain KB-VQA

Closed Domain KB-VQA involves questions that depend on information from a fixed knowledge base. These questions can be answered using the knowledge contained within a static knowledge base, such as knowledge graphs or other similar sources.

Cross-modal Retrieval for Knowledge-based Visual Question Answering (Lerner et al., 2024): The paper focussed on KB-VQA about named Entities (KVQAE) in Multimodal Information Retrieval. It addresses the challenge of answering questions about named entities using a visual context, where images represent the entities and multimodal interactions between text and images are complex. The authors propose a multimodal dual encoder, CLIP, for mono-modal and cross-modal retrieval. They demonstrate the complementarity of both retrieval approaches and compare their performance on several datasets. Additionally, they explore different strategies for fine-tuning the model in this context: mono-modal, cross-modal, or joint training.

$$-log\frac{exp(s(i_q,t_p^{(+)},i_p^{(+)})e^{\tau})}{exp(s(i_q,t_p^{(+)},i_p^{(+)})e^{\tau})+\sum_j exp(s(i_q,t_p^{(j)},i_p^{(j)})e^{\tau}}$$

Given an image of the question iq and a collection of entities (t_p,i_q), where t_p denotes the name of the entity and i_p its reference image. The authors define the similarity function as:

$s(i_q, t_p, i_p) = \alpha_I s_I(i_q, i_p) + \alpha_C s_C(i_q, t_p)$

$\alpha_{I,C}$ weigh each similarity

$s_C(i_q, t_p) = cos(CLIP_V(iq), CLIP_T(tp))$

Where CLIP denotes clip embedding To implement this approach, authors jointly trained $s_I(i_q, i_p)$

and $s_C(i_q, t_p)$ for each i_q image of the batch by minimizing the following objective, given the temperature τ:

6 Zero-Shot Visual Question Answering

Zero-shot VQA methodologies enable models to predict answers without requiring additional training. With the advent of large language models like ChatGPT, training these models becomes unnecessary, as they possess extensive knowledge embedded within their parameters. Utilizing these models without extra training has proven beneficial for numerous tasks. Recently, several innovative zero-shot methods for VQA have been proposed. In this section, we will explore some of the latest and most unique approaches for VQA that do not require any extra training.

Plug-and-Play VQA (Tiong et al., 2023): In this work, authors proposed a zero-shot vqa approach where instead of providing a caption for the whole image authors provide question-guided informative image captions, and pass the captions to a PLM as context for question answering. The architecture is divided into three modules:

Image Question Matching Module: Authors generated heatmaps where the heatmap is dense in the parts where the image is related to the question. The heatmaps are generated using GRADCAM (Selvaraju et al., 2017). This is done to get captions relevant to questions and the context is more related and useful.

Image Captioning Module: Even with relevant image regions, descriptions can vary, with some containing the desired answer and others not. To cover possible answers, we generate diverse captions using BLIP's (Li et al., 2022a) image captioning network and stochastic top-k sampling, which avoids repetitive outputs. We sample image patches based on relevance and use a prompt, generating multiple captions to ensure coverage of visual content, and retaining only non-repetitive captions.

Prediction Module: The authors provide questions and diverse captions generated above to get the predictions.

LAnguage MOdel guided Captioning (LAMOC) (Du et al., 2023): In this paper, the authors leverage the guidance and feedback of the prediction model to improve the capability of the captioning model. In this way, the captioning model can become aware of the task goal and information needed from the PLM. The captions will be used as the context for answer prediction. Formally we predict the answer as:
$p(y|x_i, x_q) = \sum_{z \in Z} p(z|x_i, x_q; \theta_C).p(y|x_q, z; \theta_P)$
The process of answer prediction is divided into two steps:

- Captioning adaptation aims to adjust θ_C to produce informative captions that are suitable for θ_P.
- Feedback-based learning aims to optimize θ_C according to task-specific feedback from θ_P.

Once the captioning model is well trained, we employ the prediction model for predicting the final answer as in Eq. (1), based on the captions provided by the captioning model.

K-GEN-VQA (Cao and Jiang, 2024): In this paper, authors propose a knowledge generation-based K-VQA method, which first generates knowledge from an LLM and then incorporates the generated knowledge for K-VQA in a zero-shot manner. There are two modules for question-answering:

Knowledge Generation: It involves two steps. First, the authors generate a single knowledge statement for each (image, question) pair in the K-VQA test dataset. Then, we perform a self-supervised knowledge diversification step, using a diverse set of these initial statements as in-context demonstrations. This second round of knowledge generation aims to produce multiple knowledge statements per (image, question) pair, leveraging the diversity to increase the likelihood of covering different aspects and improving the chances of obtaining the correct answer.

Knowledge Integration: The final set of T knowledge statements generated for each (image, question) pair, authors combine them with the image captions and the question and pass them to a pre-trained text-based QA model for answer generation.

7 Evaluation Metrics

Performance is evaluated using several metrics for Question Answering (QA) and Information Retrieval systems. Exact Match (EM) checks if the predicted answer matches the ground truth exactly. Semantic Match focuses on the meaning alignment between predicted and ground truth answers. Precision measures accuracy, while Recall assesses the system's ability to retrieve relevant answers. Mean

Rank evaluates the average rank of the correct answer, and Mean Reciprocal Rank (MRR) calculates the average reciprocal rank of the correct answer. Hits measure how often the correct answer appears within the top-N ranked answers. These metrics collectively assess QA system performance.

7.1 Exact Match and Semantic Match

Exact match and semantic match are two different ways of matching text. While exact match checks that the strings are the same, semantic match checks if the meaning is the same. Both metrics are explained below.

7.1.1 Exact Match

An exact match is when the text in the query matches exactly the text in the document. Exact match is a simple and straightforward way to match text, but it can be limited in its ability to find relevant documents. For example, if the query is "What is the capital of France?", an exact match would only return documents that contain the exact phrase "What is the capital of France?". However, there may be other documents that contain the same information, but in different words, such as "Paris is the capital of France". In the scenario with multiple gold answers, (Rajpurkar et al., 2016) defines Exact Match as the percentage of predictions that match any of the ground truth answers exactly.

7.1.2 Semantic Match

Semantic match is when the text in the query has the same meaning as the text in the document, even if the words are not exactly the same. Semantic match is a more sophisticated way to match text that can overcome some of the limitations of an exact match. Semantic match uses natural language processing (NLP) techniques to understand the meaning of the text in the query and the document. This allows the semantic match to find relevant documents that do not use exactly the same words as the query.

7.2 Precision, Recall and F1

Precision, recall, and F1 score are commonly used evaluation metrics in Question Answering (QA) systems to measure their performance. These metrics help assess the accuracy and completeness of the generated answers. Let's delve into each metric, provide their formulas, and offer examples to illustrate their calculation.

7.2.1 Precision

Precision is a measure of how accurate a system is. It is calculated as the number of correct answers divided by the total number of answers returned by the system. For example, if a system returns 10 answers and 8 of them are correct, then the precision is 0.8.

$$Precision = \frac{TP}{TP+FP}$$

where TP is true positive and FP is false positive. A high precision indicates that the system is returning a lot of correct answers.
Precision@k is a variant of precision used to evaluate the performance of ranking systems. It is calculated by counting the number of relevant documents in the top k positions of a ranked list, divided by the total number of documents in the ranked list. A higher precision@k indicates better performance. The formula for precision@k is as follows:

$$Precision@k = \frac{topk}{total}$$

where k is the number of positions considered, topk is the number of documents in the top k positions that are relevant to the query and total is the total number of documents in the ranked list. For example, if there are 10 documents in a ranked list and 5 of them are relevant to the query, then the precision@5 would be 0.5. Precision@k is a useful metric for evaluating the performance of ranking systems, as it considers the position of relevant documents in the ranked list. However, it is important to note that precision@k is sensitive to the number of relevant documents in the ranked list. A ranking system with high precision@k on a dataset with few relevant documents may not perform as well on a dataset with many relevant documents.

7.2.2 Recall

Recall is a measure of how complete a system is. It is calculated as the number of correct answers divided by the total number of correct answers. For example, if there are 10 correct answers and a system returns 8, then the recall is 0.8.

$$Recall = \frac{TP}{TP+FN}$$

where TP is true positive, and FN is false negative. A high recall indicates that the system is returning a lot of the correct answers.

7.2.3 F1-Score

The F1 score is a measure of both precision and recall. It is calculated as the harmonic mean of precision and recall. The harmonic mean is a more

sensitive measure of performance than the arithmetic mean because it gives more weight to low values. For example, if the precision is 0.8 and the recall is 0.6, then the F1 score is 0.72.

$$F1 = \frac{2*(Precision+Recall)}{(Precision+Recall)}$$

A high F1 score indicates that the system returns a good balance of correct answers and recall.

7.3 Mean Rank, MRR, Hits@k

Mean Rank, Mean Reciprocal Rank (MRR) and Hits are three common metrics used to evaluate the performance of ranking algorithms.

7.3.1 Mean Rank

Mean rank is the average rank of all relevant documents in a ranked list. A lower mean rank indicates better performance. For example, a mean rank of 1 indicates that all relevant documents are at the top of the ranked list, while a mean rank of 10 indicates that all relevant documents are at the bottom of the ranked list. The formula for mean rank is:

$$Meanrank = \frac{1}{N} \sum Rank$$

where rank is the rank of each relevant document and n is the number of relevant documents. While MR is simple, it can be sensitive to the number of relevant documents in a ranked list. Let us say that a few relevant documents are ranked very high due to errors in the ranking system. It will pull down the average. Also, it gives importance to lower ranks. Many a time, for a system 20th rank, would be equally bad as the 100th rank. At the same time, ranks 1,3,5,7, and 10, even though they are close, would make a huge difference while evaluating a ranking system.

7.3.2 Mean Reciprocal Rank(MRR)

Mean reciprocal rank is the average of the reciprocal ranks of all relevant documents in a ranked list. A higher mean reciprocal rank indicates better performance. For example, a mean reciprocal rank of 1 indicates that all relevant documents are at the top of the ranked list, while a mean reciprocal rank of 0.5 indicates that the relevant documents are ranked in the middle of the ranked list. The formula for MRR is:

$$MRR = \frac{1}{N} \sum \frac{1}{Rank}$$

where rank is the rank of each relevant document and n is the number of relevant documents. MRR gives higher importance to the top ranks and lower importance to the bottom ranks. In the case of MRR, rank 1 and rank 2 has a difference of 0.5 while rank 10 and 100 have a difference of 0.09. MRR is very useful in QA systems as the correct answers to a question are only a few and when ranked correctly we should ignore ranks away from the top ranks.

7.3.3 Hits@k

Hits@k is the percentage of queries for which at least one relevant document is ranked in the top k positions. A higher hits@k indicates better performance. For example, a hits@k of 1 indicates that all queries returned at least one relevant document in the top k positions, while a hits@k of 0.5 indicates that half of the queries returned at least one relevant document in the top k positions. The formula for hits@k is:

$$Hits@k = \frac{1}{N} \sum [1 \; if \; rank \leq k \; else \; 0]$$

where rank is the rank of each relevant document and n is the number of queries. For example, if there are 5 queries and 2 relevant documents are ranked in the top 3 positions, then the hits@3 would be 2/5 = 0.4. Hits@k is useful in QA systems when we want the answer to a question at top-k ranks.
When a QA model predicts multiple answers, a gold answer might be at 2nd rank, but 1st rank answer could be another gold answer. In such situations, we don't want to penalize our model. Usually when evaluating a system, hits are calculated with different values of k and can be helpful in identifying how many top results should be shown to ensure that the user gets the answer to its question most of the time with less lookup through the k results.

8 Summary

This paper provides a comprehensive overview of the various approaches employed in knowledge-based visual question answering (KB-VQA). We delve into the fundamental concept of knowledge graphs, exploring how they can serve as an external knowledge source to enhance visual question-answering systems. Additionally, we discuss a range of question-answering datasets pertinent to both visual and textual domains.
The KB-VQA task is categorized into two primary types: closed-domain and open-domain. In a closed-domain KB-VQA task, the system is restricted to a predefined set of knowledge, whereas an open-domain KB-VQA task requires the system to leverage a broader and potentially unlimited range of external knowledge. Throughout the

paper, we review and analyze various methodologies that have been previously employed to tackle these tasks. This includes a discussion on the integration of knowledge graphs with visual question-answering systems, highlighting the strengths and limitations of each approach.

Furthermore, the paper provides a detailed explanation of the different evaluation metrics used to assess the performance of QA and retrieval systems. These metrics are crucial for understanding the effectiveness of various approaches and for benchmarking progress in the field.

In summary, this paper not only outlines the current state of KB-VQA research but also provides insights into the challenges and potential future directions for improving knowledge-based visual question-answering systems.

9 Conclusion and Future Work

This paper examines various approaches for knowledge-based visual question answering (KB-VQA) tasks. Our findings indicate that incorporating real-time external knowledge as an additional knowledge vector significantly enhances accuracy. In open-domain visual question answering, accuracy improves substantially with the increase in external knowledge sources. For closed-domain KB-VQA tasks, effective filtering and the provision of relevant knowledge as an additional vector markedly boost accuracy. Future work could focus on developing an end-to-end model that filters knowledge and predicts answers based on the given question and knowledge vector.

References

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*, SIGMOD '08, page 1247–1250, New York, NY, USA. Association for Computing Machinery.

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

Rui Cao and Jing Jiang. 2024. Knowledge generation for zero-shot knowledge-based VQA. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 533–549, St. Julian's, Malta. Association for Computational Linguistics.

Yuanfei Dai, Shiping Wang, Neal N. Xiong, and Wenzhong Guo. 2020. A survey on knowledge graph embedding: Approaches, applications and benchmarks. *Electronics*, 9(5).

Yifan Du, Junyi Li, Tianyi Tang, Wayne Xin Zhao, and Ji-Rong Wen. 2023. Zero-shot visual question answering with language model feedback.

Xingyu Fu, Sheng Zhang, Gukyeong Kwon, Pramuditha Perera, Henghui Zhu, Yuhao Zhang, Alexander Hanbo Li, William Yang Wang, Zhiguo Wang, Vittorio Castelli, Patrick Ng, Dan Roth, and Bing Xiang. 2023. Generate then select: Open-ended visual question answering guided by world knowledge. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2333–2346, Toronto, Canada. Association for Computational Linguistics.

Liangke Gui, Borui Wang, Qiuyuan Huang, Alexander Hauptmann, Yonatan Bisk, and Jianfeng Gao. 2022. KAT: A knowledge augmented transformer for vision-and-language. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 956–968, Seattle, United States. Association for Computational Linguistics.

Guoliang Ji, Shizhu He, Liheng Xu, Kang Liu, and Jun Zhao. 2015. Knowledge graph embedding via dynamic mapping matrix. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 687–696, Beijing, China. Association for Computational Linguistics.

Paul Lerner, Olivier Ferret, and Camille Guinaudeau. 2024. Cross-modal retrieval for knowledge-based visual question answering.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022a. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation.

Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. 2022b. Grounded language-image pre-training.

Yankai Lin, Zhiyuan Liu, Maosong Sun, Yang Liu, and Xuan Zhu. 2015. Learning entity and relation embeddings for knowledge graph completion. *Proceedings of the AAAI Conference on Artificial Intelligence*, 29(1).

Yuanze Lin, Yujia Xie, Dongdong Chen, Yichong Xu, Chenguang Zhu, and Lu Yuan. 2024. Revive: regional visual representation matters in knowledge-based visual question answering. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge.

Maximilian Nickel, Volker Tresp, and Hans-Peter Kriegel. 2011a. A three-way model for collective learning on multi-relational data. In *International Conference on Machine Learning*.

Maximilian Nickel, Volker Tresp, and Peer Kröger. 2011b. A three-way model for collective learning on multi-relational data. pages 809–816.

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.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. A-okvqa: A benchmark for visual question answering using world knowledge.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626.

Vivian Silva. 2019. WordNet Knowledge Graph.

Anthony Meng Huat Tiong, Junnan Li, Boyang Li, Silvio Savarese, and Steven C. H. Hoi. 2023. Plug-and-play vqa: Zero-shot vqa by conjoining large pretrained models with zero training.

Théo Trouillon, Johannes Welbl, Sebastian Riedel, Eric Gaussier, and Guillaume Bouchard. 2016. Complex embeddings for simple link prediction. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 2071–2080, New York, New York, USA. PMLR.

Zhen Wang, Jianwen Zhang, Jianlin Feng, and Zheng Chen. 2014. Knowledge graph embedding by translating on hyperplanes. *Proceedings of the AAAI Conference on Artificial Intelligence*, 28(1).

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models.

Chenxi Whitehouse, Tillman Weyde, and Pranava Madhyastha. 2023. Towards a unified model for generating answers and explanations in visual question answering.

Bishan Yang, Wen tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. 2015. Embedding entities and relations for learning and inference in knowledge bases.

Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Yumao Lu, Zicheng Liu, and Lijuan Wang. 2022. An empirical study of gpt-3 for few-shot knowledge-based vqa.

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

Dataset	Answer-Type	Size	Domain	Evaluate Ability
ARC	Multi-Choice	7,787	Science	Reasoning
BoolQ	Bool	16K	Wikipedia	Reasoning
BioASQ	Span	282	Biomedical Articles	Indexing
CaseHOLD	Multi-Choice	53,137	Law	Pre-training
bABi	Bool/Entity	40K	Open Domain	Reasoning
CBT	Entity	20K	Children's Book	Model Memory
CliCR	Entity	105K	Medical Domain	Knowledge
CNN and Daily Mail	Entity	311K	News	Text Summarization
CODAH	Multi-choice	4,149	Open Domain	Commonsense
CommonsenseQA	Multi-choice	12,247	ConceptNet	Commonsense
ComplexWebQuestions	Entity	34,689	Freebase	Multi-hop
ConditionalQA	Entity/Span	9983	Public Policy	Multi-hop
COPA	Multi-choice	1000	Commonsense	Reasoning
CoQA	Entity	127K	Open Domain	Conversation
DROP	Span	96K	Wikipedia	Multi-hop
FinQA	Number/Span	8,281	Finance	Multi-hop
HotpotQA	Entity	113K	Wikipedia	Multi-hop
JD Production QA	Generation	469,953	E-commerce	Domain Knowledge
LogiQA	Multi-choice	8,678	Exam	Reasoning
MCTest	Multi-choice	2,000	Fictional Story	Reading Comprehension
Mathematics Dataset	Numeric	2.1×10^6	Mathematics	Calculate
MS MARCO	Generation	1,010,916	Web pages	Search
NewsQA	Span	100,000	CNN news	Reading Comprehension
OpenBookQA	Multi-choice	6000	Science	Facts Reasoning
PIQA	Multi-choice	21,000	Physical	Physical
PubMedQA	Multi-choice	1K	Medical	Summarization
RACE	Multi-choice	100,000	Exam	Reading Comprehension
ReClor	Multi-choice	6138	Exam	Logical
SCDE	Exam	6K	Exam	Reading Comprehension
SimpleQuestions	Entity	100K	Freebase	Knowledge
Squad	Span	130,319	Wikipedia	Reading Comprehension
TriviaQA	Span	650K	Open Domain	Reading Comprehension
TweetQA	Generation	13,757	Tweet	Reading Comprehension
WikiHop	Multi-choice	51,318	Wikipedia	Multi-hop
WikiQA	Sentence	3,047	Wikipedia	Reading Comprehension

Table 3: **Statistics of textual QA datasets**