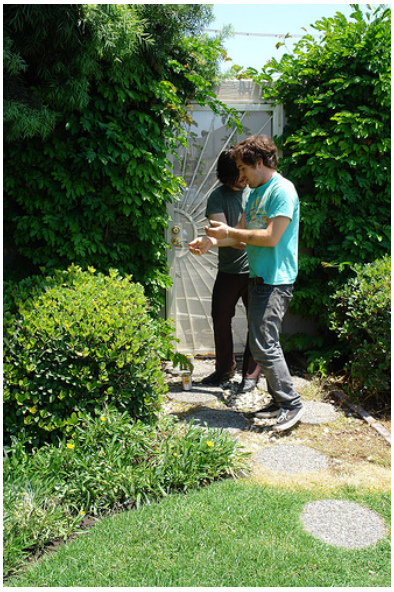
VQG & Image Captioning Resources Digest

Datasets

Flickr

Description: Sentence-based image description.
Source: https://www.kaggle.com/datasets/hsankesara/flickr-image-dataset
Format: 1 image directory, 1 csv file (containing 5 descriptions per image_id)
Comments: -
Example:

Image	Descriptions (5)	Image ID
	Two young guys with shaggy hair look at their hands while hanging out in the yard. Two young , White males are outside near many bushes. Two men in green shirts are standing in a yard. A man in a blue shirt standing in a garden. A man in a blue shirt standing in a garden.	1000092795.jpg

MSCOCO Captions(Microsoft Common Objects in Context)

Description: COCO Captions contains over one and a half million captions describing over 330,000 images. For the training and validation images, five independent human generated captions are provided for each image.
Source(s):
https://paperswithcode.com/dataset/coco-captions
https://github.com/tylin/coco-caption

Format:
There would be 5 of these sets for each image_id:
{"image_id": 404464, "caption": "black and white photo of a man standing in front of a building"}

Comments: This is a **SUBSET** of the original, larger MSCOCO dataset.
Example:

Image	Descriptions	Image ID
	rows of cars parked on the side of a city street going uphill. several cars parked along a no parking area. there are plenty of cars parked on the side of the street. several cars parked along the side of a street next to a street sign. cars parked on the side of a street.	https://cocodataset.org/#explore?id=**227511**

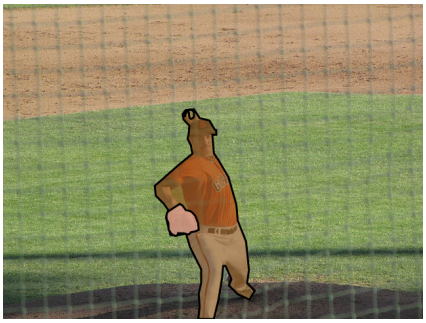
Visual Question Generation (VQG) Dataset

Description:
Includes 3 datasets (subsets of Flickr, MSCOCO, with a new Bing dataset) with a total of 75,000 questions, which range from object- to event-centric images.
Source:
https://www.microsoft.com/en-us/download/details.aspx?id=53670
First introduced in this paper: https://arxiv.org/pdf/1603.06059v3.pdf
Format:
Varies between the 3 datasets, but typically json file where each row:
{image_id, image source/link, questions, captions}
Comments:
NOT specifically Knowledge-based questions. It makes **no distinction** between them.
The bing dataset was made by querying a search engine (bing) with 1,200 event-centric query terms.
Example:

Image	Descriptions	Questions
 ID: a3d69dd7-2e07-49af-b478-b5a43344b336	A black Subaru sports sedan is parked. A black car A small black car sitting in the desert A black four door car parked in a lot in the desert A sporty black car with an air foil on the back in front of desert hills	Is this car up for sale? What is the top speed of this vehicle? Have you seen this car drift? What model type of vehicle is this? What kind of car is that?

Visual Question Answering (VQA) Dataset

Description:
Contains open-ended questions about images. These questions require an understanding of vision, language and commonsense knowledge to answer.
265,016 images (COCO and abstract scenes)
Source:
https://visualqa.org/
https://paperswithcode.com/dataset/visual-question-answering
Format:
Image files, with name as image id.
Questions Json file similar to the following:
{ "image_id": 458752, "question": "What is this photo taken looking through?", "question_id": 458752000}
Comments:
Part of these images are a subset of the COCO dataset.
Are **NOT** specifically knowledge-based questions.
Example:

Image	Questions	Image ID
	Is this man a professional baseball player? What color is the players shirt? What position is this man playing? What is this photo taken looking through?	458752

Visual Genome Dataset

Description: Consists of 101,174 images from MSCOCO with 1.7 million QA pairs.
Also presents 108K images with densely annotated objects, attributes and relationships.
Source:
https://homes.cs.washington.edu/~ranjay/visualgenome/api.html
Format:
Image Data Json:
{"width": 1024, "url": "https://cs.stanford.edu/people/rak248/VG_100K/61525.jpg", "height": 683, "image_id": 61525, "coco_id": 322472, "flickr_id": 2219749782}
Question Data Json:
{"id": 61525, "qas": [{"a_objects": [], "question": "Who is wearing a glove?", "image_id": 61525, "qa_id": 991791, "answer": "The player.", "q_objects": []}, …….]}

Comments: Compared to VQA, is a more balanced distribution over these question types: What, Where, When, Who, Why and How. Are **NOT** specifically knowledge-based questions. Around 60-70 questions for this 1 sample image
Example:

Image	Questions	Image ID
	Why is the woman's hat floppy? What brand of shirt is the man in white to the left? What are the two women in the middle watching? "Why does the player have a bat?	61525

K-VQG: Knowledge-aware Visual Question Generation for Common-sense Acquisition Dataset

Description:
Novel knowledge-aware VQG dataset called K-VQG. This is the first large, humanly annotated dataset in which questions regarding images are tied to structured knowledge
Source:
https://uehara-mech.github.io/kvqg
Format:
{ "target_object": "container", "question": "what is the back of the truck shown used for?", "knowledge": "container, UsedFor, store all the collection", "answer": "store all the collection", "image_id": 713614, "question_id": 0, "query_knowledge": ["container", "UsedFor", "[MASK]"], "img_fname": "713614.jpg", "object_id": 1586409, "x1": 103, "y1":421, "x2":572, "y2":03}

Comments:
Images are from the **Visual Genome** Dataset. These questions **ARE** knowledge-based.
Only 1 Question per image.
Example:

Image	Question	Image ID
	what is the back of the truck shown used for?	713614