|  |  |  |
| --- | --- | --- |
| **Aspect** | **Visual Genome** | **VQA v2.0** |
| Primary Purpose | General **image understanding, including object detection, relationships, attributes, and scene graphs**. | Visual Question Answering with a focus on **generating answers from images**. |
| Number of Images | ~ **108,077 images** | ~ **204,721 images** |
| Question-Answer Pairs | ~ **1.7 million QA pairs** | **~ 1.1 million QA pairs** |
| Size on Disk | ~ **40 GB** | **~ 20 GB** |
| Additional Features | Includes scene graphs (object relationships, attributes), region descriptions, and synsets. | Primarily focused on QA pairs; no additional scene information or region descriptions. |
| QA Types | Open-ended and multiple-choice QA pairs (e.g., What is the person doing?). | Open-ended and multiple-choice QA pairs (e.g., Yes/No, "What," "How many"). |
| Object and Attribute Annotations | **Highly detailed** annotations for objects, relationships, and attributes in images. | No object or attribute annotations; focused solely on questions and answers. |
| Scene Graphs | includes scene graphs describing object-object relationships (e.g., "man holding cup"). | No scene graphs. |
| Image Sources | COCO images (like VQA), but includes additional Flickr images. | COCO dataset images |
| Complexity of QA Pairs | QA pairs are more varied, with complex, descriptive questions (e.g., reasoning-based questions). | A mix of simple and complex questions, but generally focused on answering from visual content. |
| Language Annotations | Object names, attributes, and relationships annotated in addition to QA pairs. | Only question and answer annotations in the dataset. |
| Use Cases | Scene graph generation, object recognition, visual understanding, QA tasks, visual reasoning. | Primarily for training Visual Question Answering models, understanding image-based question reasoning. |