**Scene graph generation**

24 Fall ECE549 project proposal

**Group members**

Weijie Liang weijiel4@illinois.edu

Zhihua Gong zhihuag2@illiinois.edu

**Project description and goals**

**Project Description:**

Scene graph generation is a key task in computer vision that transforms an image into a structured representation. In this process, objects in the image are identified as nodes, while the relationships or interactions between these objects are represented as edges, forming a graph. Each node contains detailed information about the object it represents, such as its class (e.g., "dog," "car"), while the edges define how the objects interact with each other (e.g., "on," "next to") and add semantic context to the relationships.

**Project Goal:**

The primary goal of this project is to develop a system containing two models that can accurately generate a scene graph from a given image. The first model will be able to do the object detection part to identify the objects from a given image and create the “nodes” for the graph. The second model will find out the relationship between nodes and add edges among nodes to construct the final graph. By doing so, we will be able to offer a semantically rich and structured understanding of the visual scene, enabling more advanced reasoning about the image content. This structured graph can be leveraged for various applications, such as image captioning, visual question answering, and enhanced scene understanding.

**Member roles**

Zhihua Gong will implement object detection module and Weijie Liang will implement relationship identification module.

**Resources**

**Dataset:**

For this project, which focuses on generating scene graphs to represent visual content, the Visual Genome dataset is an ideal choice due to its extensive annotations and rich variety of data. The dataset provides over 108,000 images with detailed labels, including objects, attributes, and relationships—exactly the elements needed to construct accurate scene graphs. By using these annotations, the project can effectively map objects (nodes) and their interactions (edges) in a structured and semantically meaningful way.

In addition, Visual Genome's balanced distribution of question types (What, Where, When, Who, Why, and How) makes it especially suited for tasks that require higher-level reasoning and understanding of image content. The 1.7 million question-answer pairs provided by the dataset can also support the training and evaluation of models that require context-based reasoning, a crucial aspect of scene graph generation.

A computer desk with a keyboard and a monitor

Description automatically generatedA group of people walking on the sidewalk

Description automatically generated

Pictures from Visual Genome Dataset

url**:** [**https://paperswithcode.com/dataset/visual-genome**](https://paperswithcode.com/dataset/visual-genome)

**Implementation platform:**

Personal laptop with GPU RTX 3070

**Coding:**

Since the project will consist of machine learning knowledge and skills that we are not familiar with, we plan to first compare the performance of several object detection algorithms whose source code are available on Github. Then we are going to implement the best algorithm with code from Github or Huggingface and make modifications so that it will fit our dataset. We will do the same thing for the relationship identification part: compare the existing algorithms, find the best one and use it in our project with open-source code.

**Reservations**

1. Since we do not have access to adequate computational resources and the visual genome dataset is quite large (with more than 100k images), we may only use part of it as the training set. We may also use a smaller category set for objects as we only want to focus on the main objects in each image. Therefore, the performance of our project may not be good enough to identify and find relationships for every single object of the given image.

2. We found out that the object detection part has large amounts of algorithms, each performing well in different cases, so we do not worry about this part. However, the relationship identification part is much more challenging. As far as I know, the existing algorithms and theories are mostly focusing on certain specific cases, making it difficult for us to generalize them to our dataset as well as making them compatible with our object detection part. Therefore, the “minimum” goal of this project is to do the object detection part and recognizing if there exists connection between nodes, we may not guarantee that the exact relationship can be identified very well.

**Relationship to your background**

After studying computer vision, we became deeply interested in this field, especially in how it can be applied to understand complex visual scenes. Our experience with classical computer vision techniques made us aware of their limitations in handling tasks with high complexity, such as recognizing subtle relationships between objects or dealing with large-scale datasets. These classical methods, while useful, often struggle with the level of generalization and precision required for more challenging real-world applications.

To address these challenges, we are eager to explore and apply more advanced tools of deep learning. However, deep learning is relatively new to us and we understand that frameworks like PyTorch are highly popular and powerful for implementing deep learning models, so we see this project as an opportunity to further our knowledge by applying PyTorch to develop a scene graph generation model.

While this project aligns with our existing interests in computer vision, it also presents a new learning opportunity. We will be exploring advanced techniques such as graph neural networks (GNNs) and deep object detection models, which we have not previously worked with in depth. This makes the project both a natural extension of our background and a valuable chance to gain hands-on experience with state-of-the-art deep learning techniques in computer vision.