Composite Relationship Fields with Transformers for Scene Graph Generation

George Adaimi, David Mizrahi, Alexandre Alahi

Scene graph generation methods aim to extract a structured semantic representation of a scene by detecting the objects and their relationships. This representation can be used for different downstream tasks, such as intent prediction for safer autonomous transportation. While most recent methods focus on improving top-down approaches, which build a scene graph based on predicted objects from an off-the-shelf object detector, there is a limited amount of work on bottom-up approaches, which directly predict objects and their relationships in a single stage.

In this work, we present a novel bottom-up scene graph generation (SGG) approach by representing relationships using *Composite Relationship Fields* (CoRF). CoRF turns relationship detection into a dense regression and classification task, where each cell of the feature map has to identify surrounding objects and the relationships between them. Furthermore, we propose a refinement head that leverages Transformers for global scene reasoning, resulting in more meaningful relationship predictions between all objects in the scene. By combining CoRF with our Transformer-based refinement head, our method outperforms previous bottom-up methods on the Visual Genome dataset by 30% while preserving real-time performance.