First thought of the architecture, in actual it will improve a lot.

A close-up of a black and white text

Description automatically generated

One use case : - **Visual question answering using SGG**,

So a lot of images are captured of a scene let’s say for a day, and now I want to answer reasoning based question answering. For example, Who are all rich employees smoke cigarettes? – employees who owns expensive cars based on morning and evening when they came and when they go for smoke in lunch…..

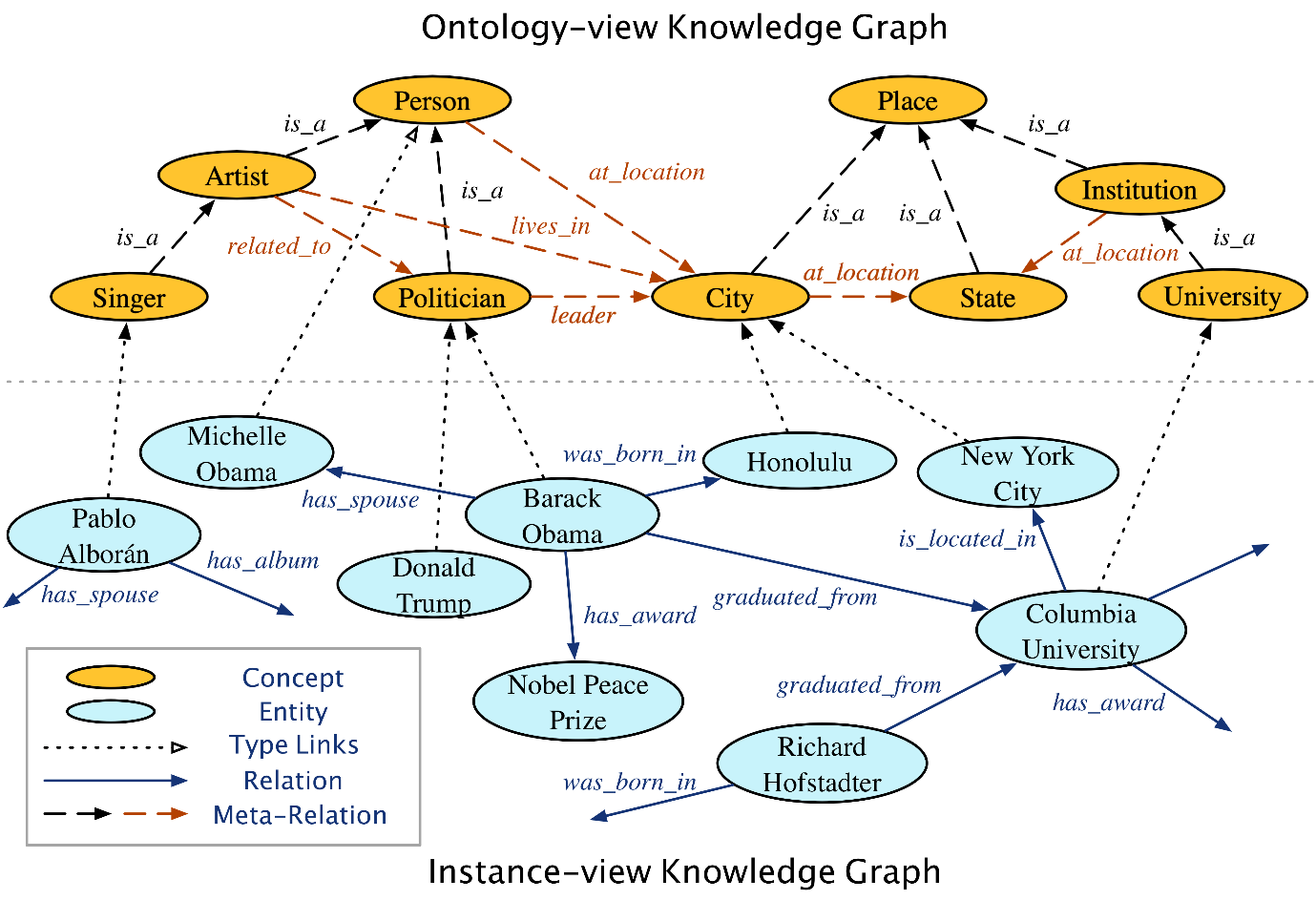
We can also use like knowledge graph based link prediction to further enhance this by predicting missing links.

There are 2 different things here, i) generate a scene graph for an image or a video, and ii) generate a scene graph for a broader time duration like maybe a day , so for a day u will get a whole big KG

**Knowledge graphs analogy**:

Data Model / Data Schema / Data Ontology == Image Ontology

Instance Model / Actual Graph Data == Scene Graphs



A complete scene graph can represent the detailed semantics of a dataset of scenes, but not a single image or a video; moreover, it contains pow erful representations that encode 2D/3D images [6], [7] and videos [8], [9] into their abstract semantic elements without restricting either the types and attributes of objects or the relationships between objects.

So for a domain, if a specified ontology is given, that can be used as a prior information to generate the Scene Graphs.

Or maybe even learning these two things side by side,

SG provides concrete data for ontology refinement to a given specific domain, **While** , Ontology guides scene graphs annotation and interpretation.

So as an input to the system (training system) – we will give a general ontology and training image set – which as an output generates the imprived refined specific domain ontology and a model capable of generating scene graphs KGs for all scene all together

Two applications:

1. Advanced Image Search
2. Advanced Image Description

To build a SGG, we need to:

1. Identify objects (Nodes)
2. Identify objects metadata (Nodes attributes)
3. Identify objects relations (Edges)

Major problem in image searching: - **Semantic Gap**

A diagram of a new level

Description automatically generated

The semantic gap in image searching represents the disparity between the low-level visual features extracted from images, such as colors, textures, and shapes and the high-level semantic concepts that humans associate with those images

Visual Genome provides a vast dataset of **images**, each accompanied by **detailed annotations** that describe **objects, relationships, and attributes** within the image

**Research Opportunities:**

1. Video based scene graph extraction **(Visual Genome)**

Using videos to generate SGGs -- according to surveys papers, there are still very few papers on this.

These efforts also include defining and detecting comprehensive sets of visual relationships that require temporal observation and understanding how these relationships can be temporally correlated or change over time​​.

Basically which image points should be considered for temporal interactions, it can be center of mass, or downsampled version of pixels so as to retain only relevant points. Or can be just bounding boxes of detected objects (but it has its own issues) or there are several researches to extract keypoint representation of an object

1. Zero-Shot domain (Never seen previously interactions in scene, unseen domain) Never seen previously objects altogether doesn’t makes sense hehehe, so zero-shot can be done still on zero-shot scene settings (all not connected objects usually has relative position info) also how to deal with unseen relations altogether

--Unseen scene domain, unseen object interactions, – probably will be solved by learning causal effects from a video, basically how two objects are behaving across a time frame, w.r.t each other. So even if there are two entirely different objects , but the way they interact with each other, is same as one of the pair of objects in training set – we are good. Example dog riding bycycle? – fairly possible but might not be in the datasets.

--Unseen relationships overall, Long tailed relationships (a big challenge in the domain)

--What should be a fair evaluation metrics --- this in itself could be a good research topic

1. Unsupervised (Without the need of detailed annotations like given in Visual Genome)

Idea would be that using generated scene graph and historical time series , future time series would be correctly predicted. Thus internally learning scene graph.

Having a proper labelled annotated data is a big challenge. There are very very less papers in this way

1. SGG enhanced with causal relationships (very very few papers)
2. Visual Question answering, KG QA based on SGG, KG reasoning, KG link prediction (Just an applications) – This is the most industry thing I guess, but it would be helpful in cases like okay u have whole day cctv footage, ask questions based on that, not just based on single images, infact reasoning based on whole days
3. how to reduce the computational complexity as much as possible while ensuring deep interaction between the triplet’s different features --- as in a way to not try predicting the relationship between each pair of objects? Or maybe matrix/adjacency matrix, graph way of doing things instead of predicting each separately ? something like locality sensitive hashing attention? This is also a huge problem as usually complexity is O(NXNXR) – it scales very bad with increase in number of objects in the scene (less papers)
4. Relationships Detection Between Distant Objects (survey didn’t mention any papers)

Just an idea to further improve the models, probabilistic model in the end combining several priors (language, statistical and knowledge graphs—mentioned 3 types of priors in the survey paper)to generate the final graph. There are also good amount of efforts in the are to how to use priors to improve the SGG

**ChatGPT references (some general discussion with chatgpt, so basically I added 3 literature surveys papers to chatgpt and then asked this new chtgpt to answer us):**

what are some new research opportunites in scene graph generation domain?



**Research paper analysis**

The document outlines several promising research opportunities in the scene graph generation domain:

1. **Building Large and Accurate Datasets**: Developing extensive datasets with fine-grained labels and accurate annotations is highlighted as both necessary and significant. Such datasets, ideally created by computer vision experts, should encompass a broad array of scenes to enhance model performance in visual semantics and foster a deeper understanding of the visual world.
2. **Designing Models for Specific Scenarios**: To address the imbalance in relationship distribution, the document suggests designing models tailored to specific scenarios. This approach involves segmenting scenes differently based on the varying categories and distributions of visual relationships and object types encountered across scenarios. Employing ensemble learning methods is recommended to advance scene graph generation applications.
3. **Exploring 3D Scene Graphs**: Another avenue for future work is the development of 3D scene graphs. This involves defining an effective and unified 3D scene graph structure and determining the information it should encapsulate. The document points out the challenge of defining spatial semantic relationships in 3D scenes without relying on a specific viewpoint, as is common in 2D image analysis. It mentions efforts by Armeni et al. to incorporate essential 3D information into the scene graph structure, extending it into 3D space to ground semantic information there. However, it notes that the proposed structure lacks expansibility and generality. Additionally, the document highlights technological challenges in presenting and extracting 3D semantic information due to the diverse storage formats of 3D information, such as RGB-D, point clouds, 3D mesh/CAD models, etc​​.

User

**You**

what about following Research Opportunities: i) Video based scene graph extraction ii) Zero-Shot interaction (Never seen previously interactions in scene) Never seen previously objects doesn’t makes sense hehehe, so zero-shot can be done still on zero-shot scene settings iii) Unsupervised (Without the need of detailed annotations like given in Visual Genome)



**Research paper analysis**

The document outlines the following research opportunities in the scene graph generation domain:

1. **Video-Based Scene Graph Extraction**: The dynamic nature of videos provides a rich set of features for detecting visual relations, including dynamic interactions between objects and temporal changes in relationships. Videos enable the modeling of a comprehensive set of visual relationships that require temporal observations, such as actions happening over time or relationships that change throughout the video. This opens up research opportunities in defining and detecting spatio-temporal relationships in videos, leading to the concept of Spatio-Temporal Scene Graph Generation (ST-SGG), where the aim is to detect entire visual relation instances with bounding box trajectories throughout a video​​.
2. **Zero-Shot Interaction**: While the document did not explicitly mention zero-shot interaction in the context of never-before-seen interactions in a scene, zero-shot learning is a significant area of research in scene graph generation. It involves the ability to recognize or infer relationships or interactions between objects that the model has not explicitly seen during training. This capability is crucial for understanding novel scenes and interactions, potentially leading to advancements in how models generalize from seen to unseen object relationships.
3. **Unsupervised Scene Graph Generation**: The exploration of unsupervised methods for scene graph generation, where detailed annotations like those in Visual Genome are not available, represents a significant research direction. This involves developing techniques that can learn to identify and label objects and their relationships without relying on extensive hand-labeled datasets. Such approaches could leverage weak supervision, unsupervised feature learning, or transfer learning from related tasks to construct meaningful scene graphs from unlabeled visual data.

These areas are essential for advancing the capabilities of scene graph generation models, enabling them to understand more complex, dynamic, and previously unseen visual scenes. They also highlight the ongoing need for novel methodologies that can reduce reliance on heavily annotated datasets and improve the ability to generalize to new situations.

User

**You**

Are there already enough researches in Spatio-Temporal Scene Graph Generation (ST-SGG)



**Research paper analysis**

The area of Spatio-Temporal Scene Graph Generation (ST-SGG) is relatively new, and while some research has been initiated, it appears there is substantial room for further exploration and development. The documents reviewed discuss the inception and early efforts in this domain, indicating an emerging interest rather than a fully matured field of study.

Spatio-Temporal Scene Graph Generation extends the traditional scene graph concept from static images to videos, emphasizing the importance of temporal dynamics alongside spatial relationships between objects. This extension allows for modeling more comprehensive sets of visual relationships that incorporate temporal observations, such as actions or interactions unfolding over time, and relationships that change or evolve throughout the duration of a video​​.

Initial efforts in ST-SGG have focused on adapting existing relationship detection models from still images to videos, highlighting the unique challenges videos present, such as dynamic interactions and the need for temporal reasoning. These efforts include defining and detecting comprehensive sets of visual relationships that require temporal observation and understanding how these relationships can be temporally correlated or change over time​​.

Given the nascent stage of ST-SGG research, coupled with the complexity and richness of video data compared to static images, it's clear that the domain offers a broad spectrum of research opportunities. The challenges inherent in accurately capturing and modeling dynamic relationships between objects in videos provide fertile ground for future work. This includes the development of new models that can more effectively harness the temporal dimension of videos, as well as efforts to create datasets and evaluation frameworks tailored to the unique demands of ST-SGG.

In summary, while initial steps have been taken toward understanding and generating spatio-temporal scene graphs, the research community is just beginning to scratch the surface of what appears to be a deeply rich and complex area of study. Further exploration, innovation, and development are needed to fully realize the potential of ST-SGG in enhancing machine understanding of video content.

User

**You**

what about generating causal relationships between objects interacting , will that be same as scene graphs



**Research paper analysis**

Generating causal relationships between objects in a scene graph involves understanding the cause and effect dynamics between entities, which can add a layer of complexity beyond what traditional scene graphs capture. Scene graphs primarily focus on depicting the entities (objects), their attributes, and the relationships (interactions) between them within a visual scene. These relationships are often descriptive or spatial (e.g., "next to," "holding," "above"), capturing the static state of a scene without necessarily implying causality.

Incorporating causal relationships would extend the scene graph's utility by providing insights into why certain interactions are happening, potentially leading to richer interpretations of the scene and more accurate predictions about future states or unseen parts of the scene. For instance, understanding that "Person A is throwing a ball to Person B" involves not just recognizing the action and the actors but also inferring the intention (cause) behind the action and its anticipated effect.

However, identifying and representing causal relationships requires a deeper level of understanding and inference, which may involve:

1. **Temporal Reasoning**: Many causal relationships unfold over time, requiring the analysis of sequences of events or states. This goes beyond the static representation in traditional scene graphs and may require temporal scene graphs or video-based scene graph generation.
2. **Intention and Effect Inference**: Determining the cause behind an action often involves inferring the intention of the actors and the effects of their actions. This might require integrating additional knowledge sources or leveraging models trained on action recognition and intention prediction tasks.
3. **Complex Modeling**: Causal relationships can be more complex to model and predict, as they may not be directly observable from the visual features alone. Incorporating causal reasoning might require combining visual cues with common-sense knowledge or external datasets that provide annotations for causal interactions.
4. **Richer Annotations**: Existing scene graph datasets may not provide annotations for causal relationships, necessitating richer, more detailed annotations that capture not just what is present in the scene but also why actions are taking place.

While scene graphs and causal relationship modeling share the goal of enhancing the understanding of visual scenes, integrating causality into scene graphs represents an advanced and somewhat different endeavor. It necessitates methodologies that can infer beyond visible interactions, possibly requiring more sophisticated models that combine visual analysis with knowledge-based reasoning or predictive capabilities to infer cause and effect.

User

**You**

In Spatio-Temporal Scene Graph Generation (ST-SGG), just tracking the four object bounding box, would it be enough, or is there a need to track the actual boundaries of object? what are the cases in which actual object boundary tracking would be needed?



**Research paper analysis**

The documents reviewed do not specifically discuss the necessity of tracking the actual boundaries of objects versus merely tracking the four corners of object bounding boxes within the context of Spatio-Temporal Scene Graph Generation (ST-SGG). However, based on foundational knowledge in computer vision and scene understanding, the need to track actual object boundaries as opposed to just bounding boxes can depend on several factors:

1. **Complexity of the Scene**: In highly dynamic and complex scenes, where objects frequently overlap or occlude each other, tracking the precise boundaries of objects can be crucial for accurate relationship detection. Precise boundary tracking can help in understanding the depth and spatial arrangement of objects, which is essential for accurately modeling their interactions over time.
2. **Detail Required for Application**: For applications requiring a high level of detail in understanding object interactions, such as fine-grained action recognition or detailed spatial reasoning, tracking actual object boundaries may provide additional cues that are not captured by bounding boxes alone.
3. **Temporal Consistency**: In videos, maintaining temporal consistency of object identities and their interactions can be challenging, especially in scenes with rapid motion or appearance changes. Precise boundary tracking might offer more robust features for ensuring consistent identification and tracking of objects across frames.
4. **Nature of Relationships and Interactions**: If the application requires understanding of detailed interactions (e.g., "holding" vs. "touching"), the actual contact points or interaction regions between objects may only be accurately determined through detailed boundary analysis.
5. **3D Scene Understanding and Augmentation**: For applications that require or benefit from 3D scene reconstruction or augmentation based on 2D video input, understanding the actual shape and size of objects through boundary tracking can be crucial for accurate 3D modeling.

In summary, while bounding box tracking provides a simpler and computationally cheaper way to locate and identify objects in a scene, there are scenarios where detailed boundary tracking is necessary to achieve the desired level of understanding and accuracy in depicting object interactions and relationships in Spatio-Temporal Scene Graph Generation.

User

**You**

what about the research opportunities in the field on visual question answering? Have there been any papers, which deals with using scene graphs as database for answering questions based on set of images, performing knowledge graph reasoning, missing link prediction on final knowledge graphs



**Research paper analysis**

The field of Visual Question Answering (VQA) presents several research opportunities, particularly with the use of scene graphs as a foundational structure for answering questions based on a set of images. Scene graphs, which contain structured semantic information about an image including knowledge of present objects, their attributes, and pairwise relationships, offer a beneficial prior for VQA tasks​​.

Research in this area explores several innovative approaches:

1. **Graph Network Encoding for VQA**: Utilizing scene graphs derived from images to abstractly represent an image as a graph. This method involves adapting a graph network to encode the scene graph and perform structured reasoning according to the input question, potentially outperforming state-of-the-art VQA algorithms by leveraging the essential information captured by scene graphs​​.
2. **Natural Language Explanations and Visual Attention**: Generating natural language explanations for VQA using annotations of entities in an image from the scene graph, combined with an attention map generated by a VQA model. This approach integrates visual and textual information to support the answers to questions asked of an image​​.
3. **Multimodal Fusion and Visual Relationship Detection**: Implementing multimodal fusion techniques, such as the block-superdiagonal tensor decomposition, for VQA and visual relationship detection tasks. This method focuses on representing fine interactions between modalities while preserving powerful mono-modal representations and leveraging end-to-end learnable architectures​​.
4. **Scene Graph Convolutional Network (Scene GCN)**: Designing Scene GCNs to jointly reason about object properties and relational semantics for the VQA task. This involves building a visual relationship encoder to yield discriminative and type-aware visual relationship embeddings, constrained by both the visual context and language priors. The Scene GCN aims to reason about visual clues for the correct answer, guided by the question​​.

These approaches highlight the potential of using scene graphs not just as a database for answering questions but also for performing knowledge graph reasoning and missing link prediction on final knowledge graphs. The integration of scene graphs into VQA represents a promising direction for research, offering possibilities for more nuanced and accurate question answering systems that leverage deep understanding of visual content.

User

**You**

Have there been already papers, which deals with using scene graphs as database for answering questions based on set of images, performing knowledge graph reasoning, missing link prediction on final knowledge graphs? Which of these I can research on as they don't have any prior papers on them



**Research paper analysis**

Based on the documents reviewed, there have been several papers that explore the use of scene graphs for visual question answering (VQA), incorporating knowledge graph reasoning, and addressing missing link prediction on final knowledge graphs:

1. **Visual Question Answering**: Scene graphs are leveraged to abstractly represent images as graphs, with nodes corresponding to object entities and edges to object relationships. This representation aids in encoding the scene graph and performing structured reasoning in response to input questions, potentially outperforming state-of-the-art VQA algorithms by leveraging the essential information captured by scene graphs. Methods like BLOCK and Scene Graph Convolutional Network (Scene GCN) have been developed to enhance VQA tasks through multimodal fusion based on block-superdiagonal tensor decomposition and by jointly reasoning about object properties and relational semantics​​.
2. **Knowledge Graph Reasoning**: GB-Net proposes constructing scene graphs and knowledge graphs into a unified framework, where the scene graph serves as the image-conditioned instantiation of the commonsense knowledge graph. This approach redefines the generation of scene graphs as a bridge mapping between scene graphs and common sense graphs, aiming to address the challenges posed by noise and missing annotations in existing label datasets​​.
3. **Missing Link Prediction**: While the documents specifically highlight applications in VQA and integration with knowledge graphs, they imply the potential for research into missing link prediction within the constructed knowledge graphs. By identifying and exploiting patterns in the scene and knowledge graphs, there could be opportunities to predict missing relationships or attributes that were not explicitly captured in the original scene graph generation process.

These areas represent significant opportunities for further research, especially in developing methodologies that can enhance the accuracy and utility of scene graphs for complex reasoning and inference tasks across diverse application domains. Given the novelty and complexity of these challenges, they remain ripe for exploration, with potential for groundbreaking contributions in how visual and relational data are processed and understood.

User

**You**

what does Visual genome actually contains as inputs?



**Research paper analysis**

The Visual Genome dataset is a richly annotated dataset designed to connect language and vision to enable detailed image understanding. It contains images paired with detailed annotations that go beyond simple object recognition to include attributes of objects and the relationships between objects. Here's what the Visual Genome dataset specifically contains as inputs:

1. **Objects and Their Attributes**: Each image in Visual Genome includes annotations for objects present in the image. These objects are annotated with various attributes that describe them in more detail. Attributes can include color, size, state, and more, providing a richer understanding of the objects beyond their identity.
2. **Relationships Between Objects**: Apart from identifying objects and their attributes, Visual Genome also annotates the relationships between different objects in an image. These relationships describe how objects interact with each other within the scene. Examples of relationships can be “sitting on,” “holding,” “next to,” etc., which help in understanding the dynamics and context of the scene.
3. **Question-Answer Pairs**: Visual Genome includes question-answer pairs for each image, aimed at visual question answering (VQA) tasks. These pairs are designed to test the ability to retrieve detailed information from the image, requiring an understanding of objects, their attributes, and their relationships.
4. **Region Descriptions**: The dataset provides detailed region descriptions that narrate what is happening in different parts of the image. These descriptions are more granular than image-wide captions and focus on specific regions within the image, providing context and detail that support complex scene understanding.
5. **Scene Graphs**: Visual Genome is known for its scene graphs, which are structured representations of an image's content. A scene graph for an image includes nodes representing objects and edges representing relationships between these objects. These graphs provide a structured way to represent the semantic content of an image, facilitating tasks like image retrieval, image generation, and enhanced image captioning.

The combination of these detailed annotations makes the Visual Genome dataset a comprehensive resource for advancing research in computer vision and natural language processing, specifically in areas that require deep understanding of visual scenes, such as visual question answering, image captioning, and the generation of detailed scene descriptions.

User

**You**

Does visual genome contains scenes of a specific or list of specific domain?

2 / 2



**Research paper analysis**

Stopped searching knowledge

User

**You**

How Zero-Shot Interaction is a problem in SGG if Visual Genome has already covered a wide variety of domains of images



**Research paper analysis**

Zero-shot interaction remains a challenge in Scene Graph Generation (SGG) even though the Visual Genome dataset covers a wide variety of image domains. This is primarily because zero-shot learning involves recognizing or inferring relationships between objects that have not been explicitly seen during training. The challenge arises in several ways:

1. **Scale of Diversity**: While the Visual Genome dataset is extensive and diverse, it cannot possibly cover every potential object interaction or relationship that exists in the real world. The infinite variety of objects and their possible interactions means there will always be novel combinations not represented in the dataset.
2. **Generalization to Unseen Interactions**: Zero-shot learning in SGG aims to generalize from seen to unseen interactions. This requires models to infer relationships based on learned patterns, semantic understanding, or attributes of objects, rather than relying on direct examples from the training data. Even with a broad dataset like Visual Genome, models must learn to extrapolate and predict interactions that have never been explicitly labeled in the dataset.
3. **Semantic Gap**: There's a semantic gap between the descriptive nature of relationships in datasets and the capability of models to understand and generalize these descriptions to new, unseen contexts. The complexity of natural language and the subtlety of visual cues make it challenging for models to bridge this gap without specific examples.
4. **Domain-Specific Challenges**: Certain domains or scenarios may be underrepresented in even comprehensive datasets like the Visual Genome, leading to challenges in zero-shot learning for these specific contexts. The diversity within the dataset may not equally cover all areas of interest, especially niche or highly specific interactions.
5. **Long Tail of Rare Interactions**: Many datasets, including the Visual Genome, exhibit a long-tail distribution where common objects and interactions are well-represented, but rare or unique interactions are underrepresented. This imbalance makes zero-shot learning critical for effectively recognizing and interpreting these rare interactions.

In summary, the breadth of the Visual Genome dataset does not eliminate the challenge of zero-shot interaction in SGG. Instead, it highlights the need for advanced modeling techniques that can leverage the dataset's diversity while also generalizing beyond it to understand and predict unseen object interactions.

User

**You**

Explain how current scene graph generation works from a single images?



**Research paper analysis**

User

**You**

Is AT (Analogies Transfer) same as zero shot interaction, if not what's different



**Research paper analysis**

Analogies Transfer (AT) and Zero-Shot Learning (ZSL) are related but distinct concepts in the context of machine learning and, more specifically, in tasks like scene graph generation or visual relationship detection.

**Analogies Transfer (AT)**

* **Concept**: Analogies Transfer typically involves leveraging known relationships or patterns from one domain to apply or infer similar relationships in a different, but related, domain. It's about drawing parallels that can help in understanding or solving problems in a new context by transferring knowledge from a familiar context.
* **Application**: In scene graph generation or other AI tasks, AT might be used to improve model performance or understanding by transferring insights or learned patterns from one set of data (or tasks) to another. This could involve using analogies to bridge gaps in data, enhance creativity, or provide novel solutions based on existing knowledge.
* **Focus**: The focus is on the transfer of knowledge or patterns across different but analogous situations or domains.

**Zero-Shot Learning (ZSL)**

* **Concept**: Zero-Shot Learning, on the other hand, is specifically concerned with the ability of models to correctly perform tasks (like classification, detection, or relationship prediction) on data they have never seen during training. It's about enabling models to infer or generalize knowledge to unseen classes or interactions based on learned attributes or relationships.
* **Application**: In the context of scene graph generation, ZSL would allow a model to predict relationships or interactions between objects that were not part of the training dataset. For example, if a model trained on images of "person riding a bicycle" and "person riding a horse" can infer "person riding an elephant" without having seen such an image, it's demonstrating zero-shot learning.
* **Focus**: The focus is on generalization to unseen instances or categories without direct prior exposure.

**Key Differences**

* **Scope of Application**: AT is broader and can apply to knowledge transfer across domains, while ZSL is specifically about handling unseen data instances within a domain.
* **Objective**: The objective of AT is to utilize similarities across domains to solve new problems, whereas ZSL aims to generalize learning to unseen classes or relationships within the same domain.
* **Mechanism**: AT involves analogy-based reasoning or knowledge transfer techniques, while ZSL often relies on attributes, semantic embeddings, or other forms of abstract representation that allow for inference about unseen categories.

In summary, while both Analogies Transfer and Zero-Shot Learning deal with the application of learned knowledge to new or unseen contexts, they approach this challenge from different angles and with different mechanisms. AT focuses on cross-domain knowledge transfer, and ZSL focuses on within-domain generalization to unseen instances.