Motivation & Definition

- **Scene graphs** capture image semantics.
- Graphs have a variety of powerful downstream applications such as image retrieval and generation.
- No neural pipeline solution for generating scene graphs from paragraph-level natural language.

Problem Statement

- How can we build semantically rich scene graphs from a paragraph-level description?
- Input: A natural language description of a scene.
- Output: \(G = \{V, E, A\}\): a set of node multisets, relationship edges, and node attributes.
- Isomorphism: many valid ways to represent \(G\).

Data

- **Visual Genome**: a collection of dense image annotations, including scene graphs and image descriptions.
- **Preprocessing**: translated objects, relationships, and attributes to synsets (collection of synonyms).
- Training pairs: (paragraph: \{objects, relationships\}).

Approaches

- Experimented with various novel architectures/models.
- **One-shot Multiset Prediction**: avoid set isomorphism.
- **Multitask Learning**: predict nodes and set cardinality.
- **Dependency Parsing**: parse input into dependency graphs and align predicted nodes and edges.
- **Neural Decoders**: use sequence models to build graph.

End-to-End Neural Pipeline

- Multitask Architecture
- Graphs have a variety of powerful downstream applications such as image retrieval and generation.
- No neural pipeline solution for generating scene graphs from paragraph-level natural language.

Conclusions and Future Work

- Presented novel architecture for scene graph generation from natural language.
- Computational limitations prevented full architecture test.
- Future Work:
  - Completely implement proposed GRU with message passing.
  - Explore algorithms for merging sentence-level scene graphs into paragraph-level scene graphs.