Introduction

- Scene graph generation (SGGen)
  - Object Detection + Visual Relation Detection
- Discrepancy
  - Scene graph is a holistic representation
  - Standard crowdsourced supervisions focus on local information

Our supervision:

- Scene graph generation (SGGen)

Approach

- Inference
  - Generates scene graphs without help from texts, i.e. SGGen
- Training pipeline

Our aim:

- Using image captions as supervision, to learn SGGen model

Cost is low

Captions capture global context

Disambiguating entities shared in triplets

Captions are linguistic constructs

Utilizing sequential patterns (common sense)

Contributions

- Phrasal (cross-triplet) context
  - Message passing on the text graph
  - Enables more accurate representations of concepts
- Sequential (within-triplet) context
  - Encodes language patterns in the RNN model

Approach

- Inference
  - Generates scene graphs without help from texts, i.e. SGGen
- Training pipeline

Experiments

- Metrics
  - Recall@50, 100 on Xu et al.’s and Zareian et al.’s VG test splits
  - Requiring three labels and two boxes to be correct
- Learning tasks
  - GT-Graph setting
    - Uses ground-truth text graphs
    - Isolates the effects of noise in captions
  - Cap-Graph settings
    - COCO and VG captions
    - Uses noisy text graphs parsed from real captions
    - Domain gap between the settings (following chart)
      - Some ground-truth relations rarely appear in text descriptions
      - Some prepositions from captions are rarely denoted as relations

- Results (middle-bottom)
  - All components have positive effects
  - Outperforms weakly supervised baselines (see underlined results)
  - On Zareian et al.’s VGSET: IMP (R@50=3.44, R@100=4.24)
  - On Xu et al.’s VG: IMP (R@50=3.10, R@100=3.50)

Conclusions

- We introduced a method that leverages caption supervision for SGGen
- We proposed to capture common sense relations (phrasal and sequential contexts)
- We iteratively refine detection scores