Graph R-CNN for Scene Graph Generation

Keegan Lensink, Xiang Liu, Zicong (Alex) Fan

THE UNIVERSITY
OF BRITISH COLUMBIA

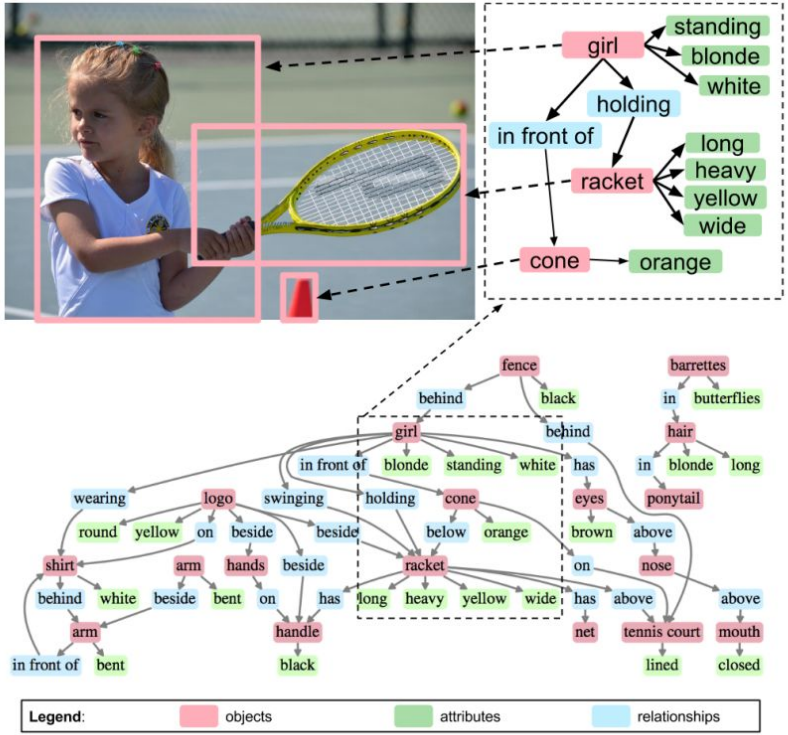
Table of Contents

Table of Contents

Introduction

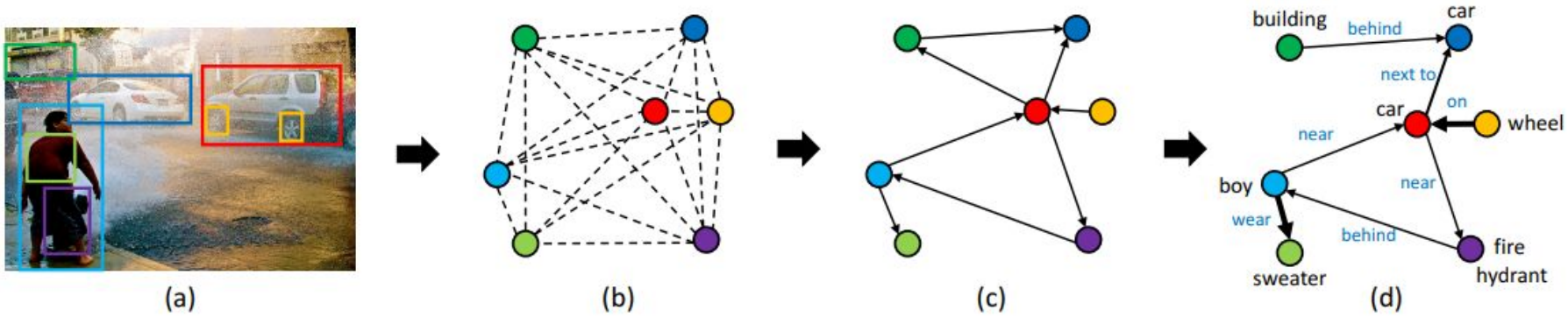
- Scene graphs form a structured representation of the image.

- Difficult to generate from a given image.



Johnson, J., Krishna, R., Stark, M., Li, L.J., Shamma, D.A., Bernstein, M., Fei-Fei, L.: Image retrieval using scene graphs. In: CVPR (2015)

Introduction



- Given n detected objects, a fully connected graph has n^2 edges.
- Many of these edges are probably between unrelated objects.
- How can we prune this graph?

Table of Contents

Contributions

- RePN
  - Relation Proposal Network
- aGCN
  - attentional Graph Convolutional Networks.
- SGGEN+
  - Modified evaluation metric for scene graphs that gives more realistic results.
- Graph RCNN framework
  - Generates scene graph for a given image.
  - Provides node features for downstream tasks.

Table of Contents

Architecture

$$P(\mathcal{S}|\boldsymbol{I}) = \overbrace{P(\boldsymbol{V}|\boldsymbol{I})}^{\text{Object Region Proposal}} \underbrace{P(\boldsymbol{E}|\boldsymbol{V},\boldsymbol{I})}_{\text{Relationship Proposal}} \overbrace{P(\boldsymbol{R},\boldsymbol{O}|\boldsymbol{V},\boldsymbol{E},\boldsymbol{I})}^{\text{Graph Labeling}}$$

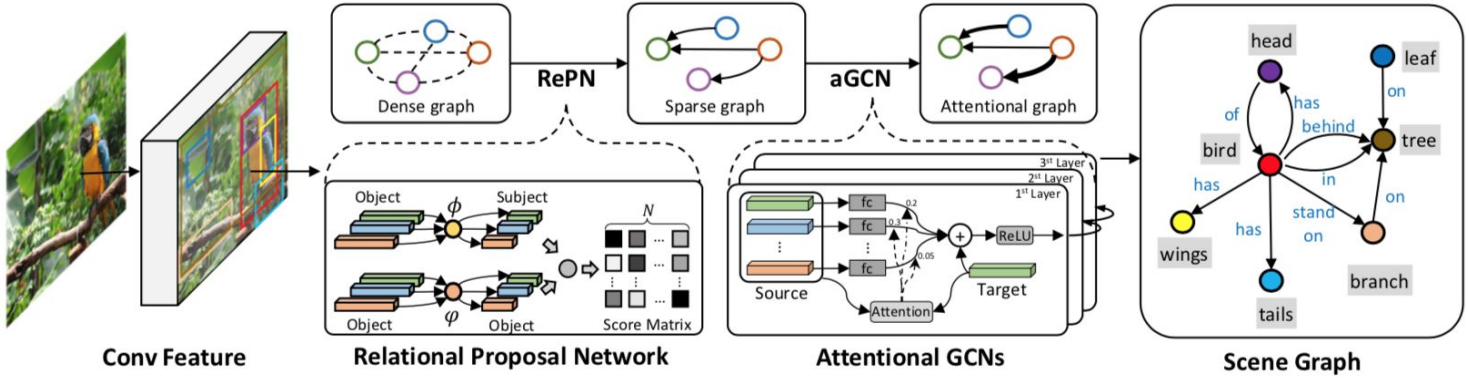
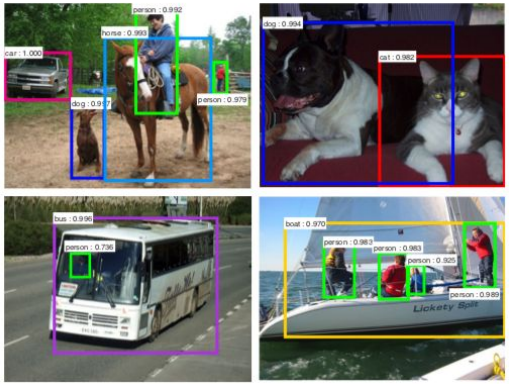


Table of Contents

Object Region Proposal

Given an image *I*, extract the following quantities with Faster-RCNN:

- Bounding boxes: $R^o \in \mathbb{R}^{n \times 4}$
- Feature vectors: $X^o \in \mathbb{R}^{n \times d}$
- Label distributions: $P^o \in \mathbb{R}^{n \times |C|}$



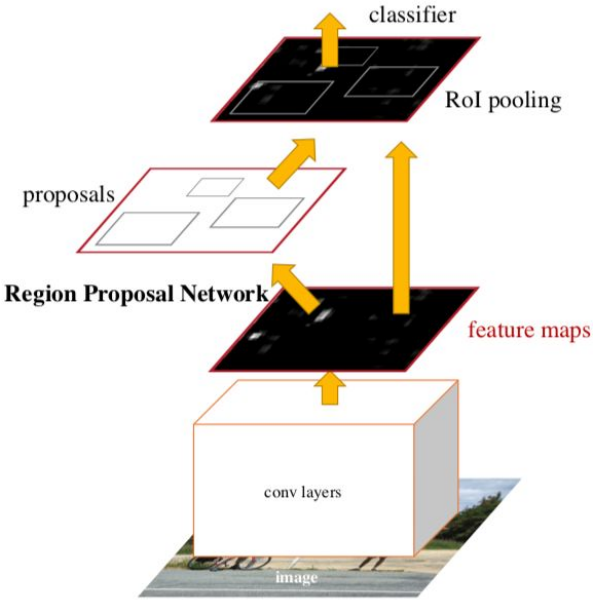


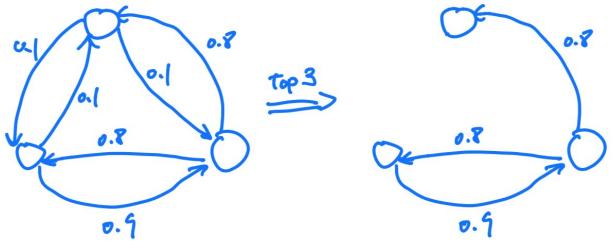
Table of Contents

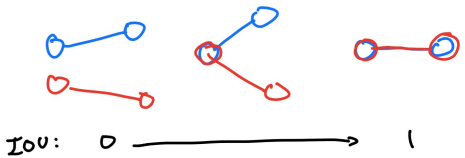
Relationship Proposal (RePN)

Learn a relatedness function:

$$f(\boldsymbol{p}_i^o, \boldsymbol{p}_j^o) = \langle \Phi(\boldsymbol{p}_i^o), \Psi(\boldsymbol{p}_j^o) \rangle, i \neq j$$

- Binary classification:
  - Score of relatedness: [0, 1]
  - {edge, not_edge}
- Non-maximal suppression
  - k top-scored edges
  - m edges with least overlaps with others.



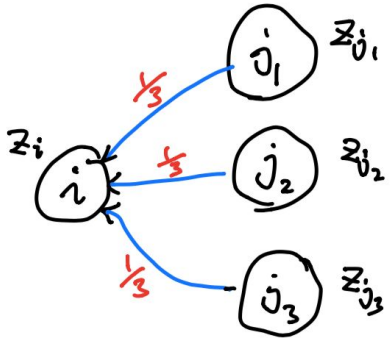


$$IoU(\{u, v\}, \{p, q\}) = \frac{I(r_u^o, r_p^o) + I(r_v^o, r_q^o)}{U(r_u^o, r_p^o) + U(r_v^o, r_q^o)}$$

Table of Contents

Conventional GCN

Recap on Graph Convolutional Network (GCN):



- Given a graph represented by:
  - Feature matrix: Z in N x D_L.
  - Adjacency matrix: alpha in N x N.
  - Dimension map: W
- Propagation rule:

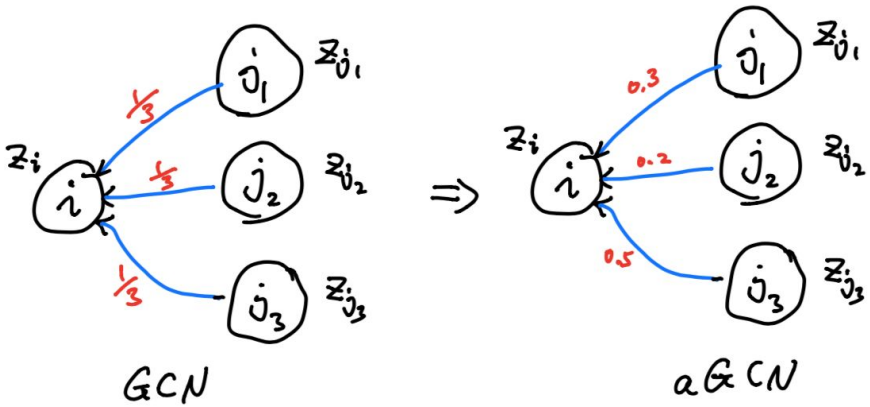
$$z_i^{(l+1)} = \sigma\left(z_i^{(l)} + \sum_{j \in \mathcal{N}(i)} \alpha_{ij} W z_j^{(l)}\right)$$

$$z_i^{(l+1)} = \sigma\left(W Z^{(l)} \boldsymbol{\alpha}_i\right)$$

- Interpretation:
  - Representation of each node at the next layer is aggregated by the equally-weighted average of its neighbours and itself.

Attentional GCN (aGCN)

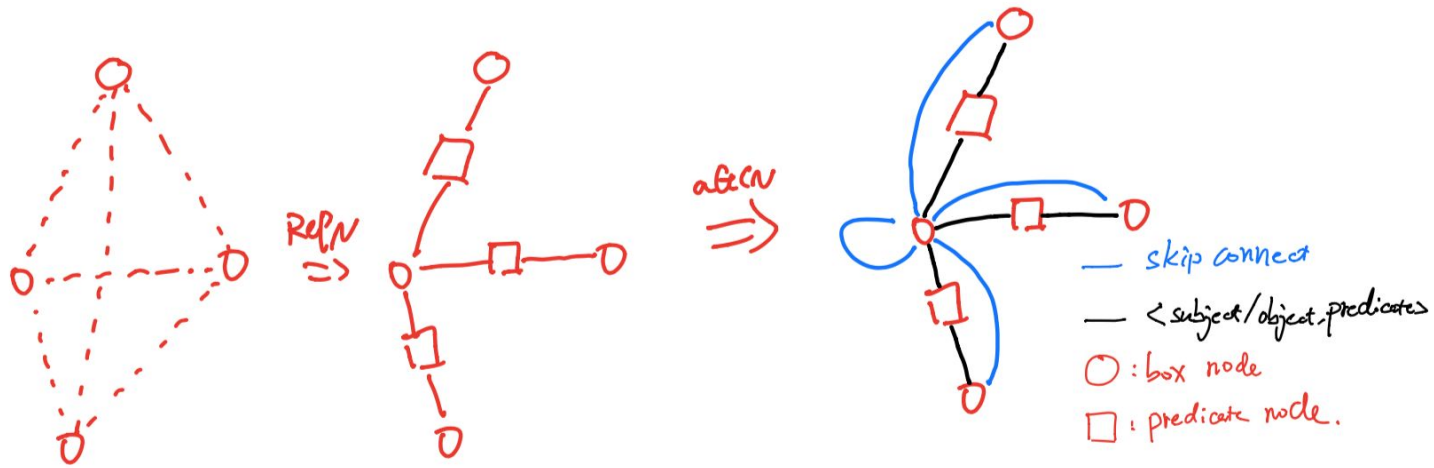
aGCN: ".. ~~equally~~ weighted average of its neighbours"



$$z_i^{(l+1)} = \sigma\left(z_i^{(l)} + \sum_{j \in \mathcal{N}(i)} \alpha_{ij} W z_j^{(l)}\right)$$

$$u_{ij} = w_h^T \sigma(W_a[z_i^{(l)}, z_j^{(l)}])$$
$$\boldsymbol{\alpha}_i = \text{softmax}(\boldsymbol{u}_i),$$
$$\boldsymbol{\alpha}_{ii} = 1 \text{ and } \boldsymbol{\alpha}_{ij} = 0 \; \forall j \notin \mathcal{N}(i).$$

Graph Formation

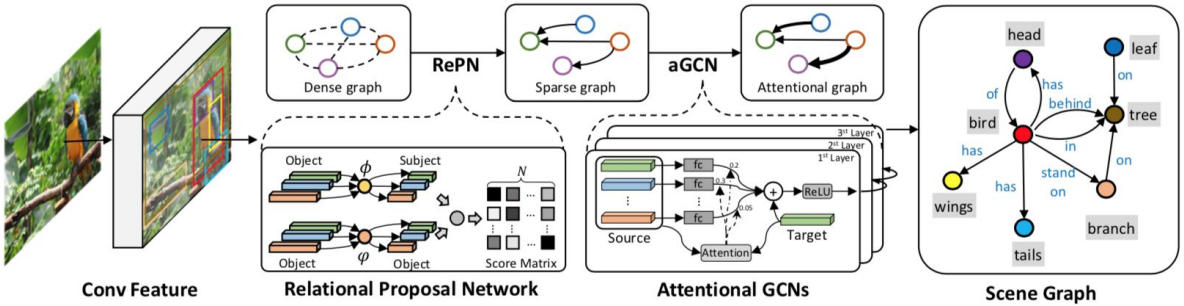


$$z_i^o = \sigma(\overbrace{W^{\texttt{skip}} Z^o \boldsymbol{\alpha}^{\texttt{skip}}}^{\text{Message from Other Objects}} + \overbrace{W^{sr} Z^r \boldsymbol{\alpha}^{sr} + W^{or} Z^r \boldsymbol{\alpha}^{or}}^{\text{Messages from Neighboring Relationships}})$$

$$z_i^r = \sigma(z_i^r + \underbrace{W^{rs} Z^o \boldsymbol{\alpha}^{rs} + W^{ro} Z^o \boldsymbol{\alpha}^{ro}}_{\text{Messages from Neighboring Objects}}).$$

Graph Labeling

“Two multi-class cross entropy losses are used for object classification and predicate classification.”



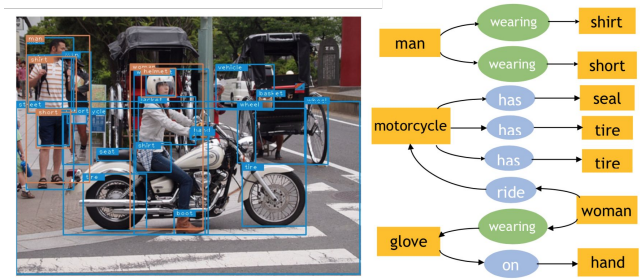


Table of Contents

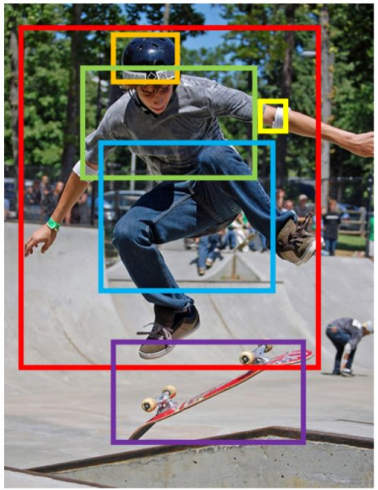
Metric: SGGen

The ground truth scene graph: <subject, relationship, object>

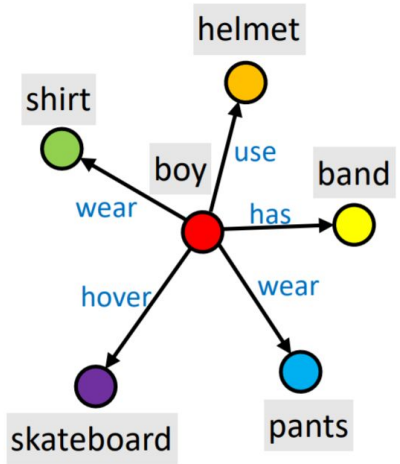
E.g.: <boy, wear, shirt>, <boy, use, helmet>, <boy, has, band> …

SGGen counts one match when:

1. all three elements have been correctly labeled
2. both object and subject nodes have been properly localized (i.e., bounding box IoU > 0.5).



(a)

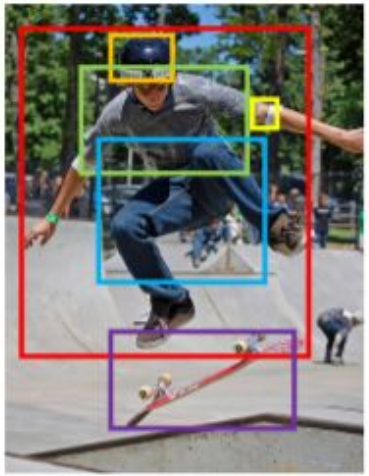


SGGen = 5 SGGen+ = 16

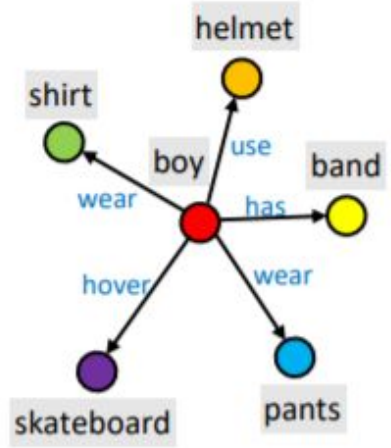
(b)

Problem of SGGen

It only counts exact matches.

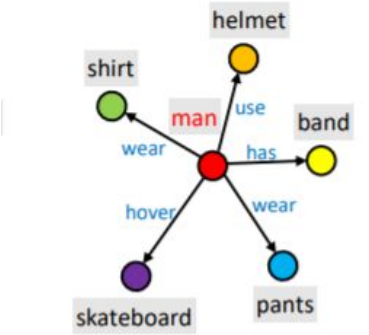


(a)



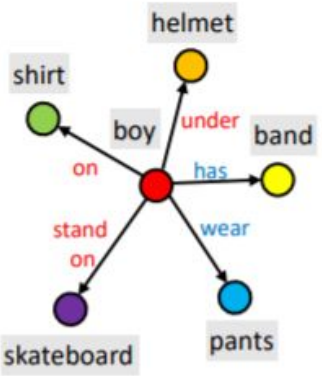
SGGen = 5 SGGen+ = 16

(b)



SGGen = 0 SGGen+ = 10

(d)



SGGen = 2 SGGen+ = 9

(e)

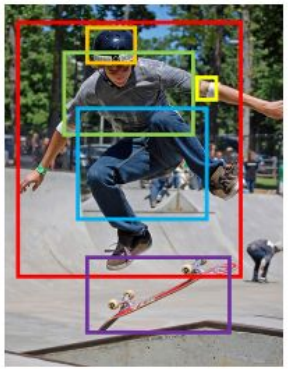
“A More Comprehensive Metric”: SGGen+

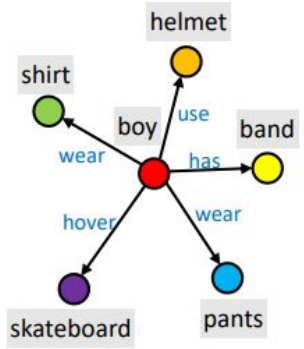
SGGen+ = C(O) + C(P) + C(T)

C(O): the number of object nodes correctly localized and recognized

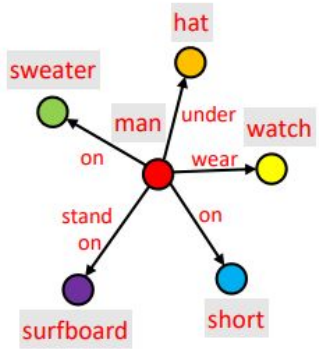
C(P): the number of predicates correctly localized and recognized

C(T): the number of matched triples, which is SGGen

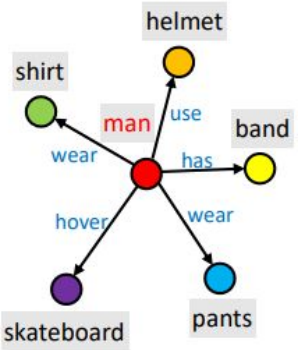




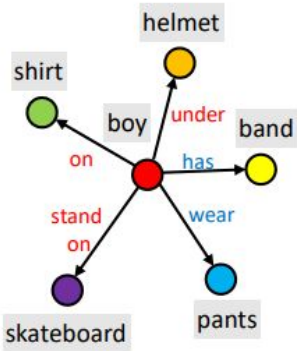
SGGen = 5 SGGen+ = 16



SGGen = 0 SGGen+ = 0



SGGen = 0 SGGen+ = 10



SGGen = 2 SGGen+ = 9

Comparing of SGGen and SGGen+

Assign random incorrect labels to objects perturbing objects

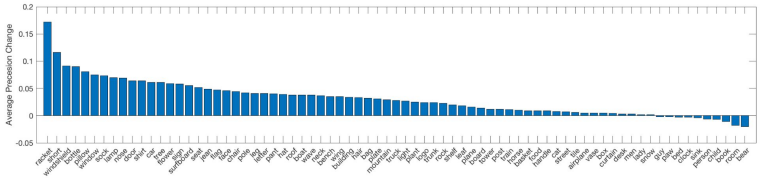
Perturb Type	none	w/o relationship			w/ relationship			both		
Perturb Ratio	0%	20%	50%	100%	20%	50%	100%	20%	50%	100%
SGGen	100.0	100.0	100.0	100.0	54.1	22.1	0.0	62.2	24.2	0.0
SGGen+	100.0	94.5	89.1	76.8	84.3	69.6	47.9	80.1	56.6	22.8

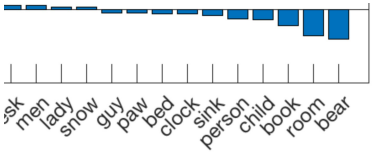
Table 1. Comparisons between **SGGen** and **SGGen+** under different perturbations.

SGGen is completely insensitive to the perturbation of objects without relationships.
SGGen is overly sensitive to label errors on objects with relationships.
100% perturbation: the object localizations and relationships are still correct such that SGGen+ provides a non-zero score.

Table of Contents

Discussion

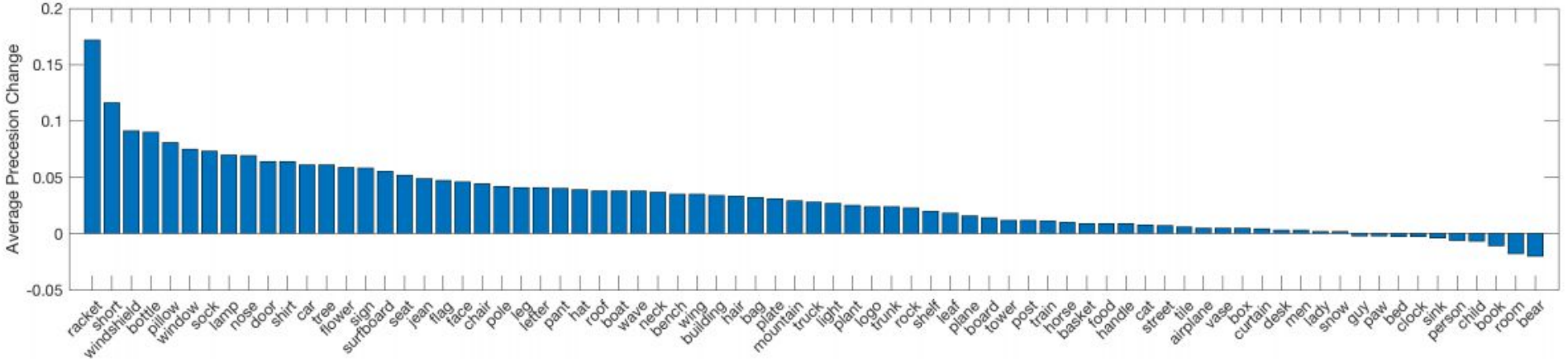




- Less robust to objects that are used in varied contexts (person), trading performance for small objects that are used in relatively few contexts (racket).
- Use P to learn relatedness between two boxes.
  - Better solution:
    - learn feature representation of a class s.t. closer labels are closer in feature space.
    - e.g. D(car, wheel) < D(car, noodles)
    - Use this distance as a score instead.
- The choice of the new metric might be biased toward their model.
- Choosing *m* for remaining object pairs is not clear, and depends on the image.
- Only top 150 classes and top 50 relations; real world has more labels.
- Use similarity instead of count, e.g. Similarity(boy, man) = 0.8

References

- Yang, Jianwei, et al. "Graph r-cnn for scene graph generation." ECCV, 2018.
- Kipf, Thomas N., and Max Welling. "Semi-supervised classification with graph convolutional networks." ICLR, 2017.
- Ren, Shaoqing, et al. "Faster r-cnn: Towards real-time object detection with region proposal networks." NeurIPS, 2015.


Average Precesion Change
0.2
0.15
0.1
0.05
0
-0.05
racket
short
windshield
bottle
pillow
window
sock
lamp
nose
door
shirt
car
tree
flower
sign
surfboard
seat
jean
flag
face
chair
pole
leg
letter
pant
hat
roof
boat
wave
neck
bench
wing
building
hair
bag
plate
mountain
truck
light
plant
logo
trunk
rock
shelf
leaf
plane
board
tower
post
train
horse
basket
food
handle
cat
street
tile
airplane
vase
box
curtain
desk
men
lady
snow
guy
paw
bed
clock
sink
person
child
book
room
bear