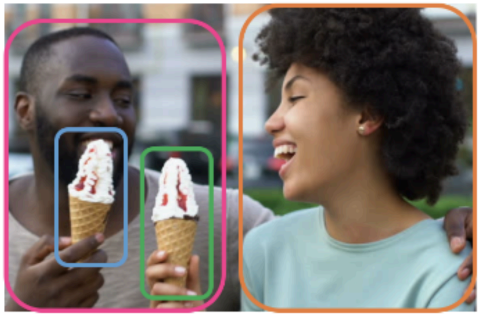
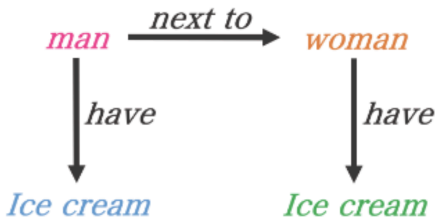
Exploring and Exploiting the Hierarchical Structure of a Scene for Scene Graph Generation

Ikuto Kurosawa, Tetsunori Kobayashi, Yoshihiko Hayashi
Waseda University

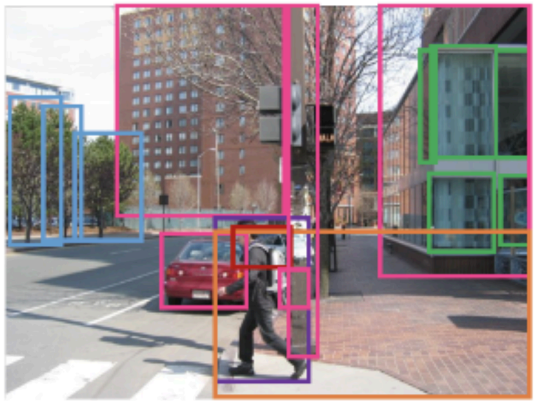
Abstract

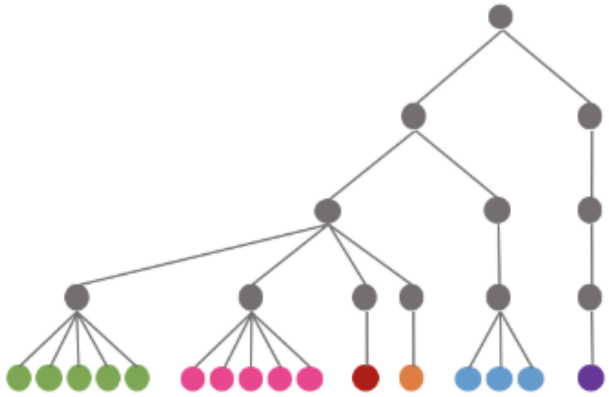
We propose novel neural network models for generating scene graphs that maintain global consistency.





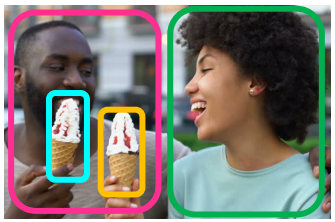
Our models build a hierarchical structure for message passing, whose leaf nodes correspond to object instances.

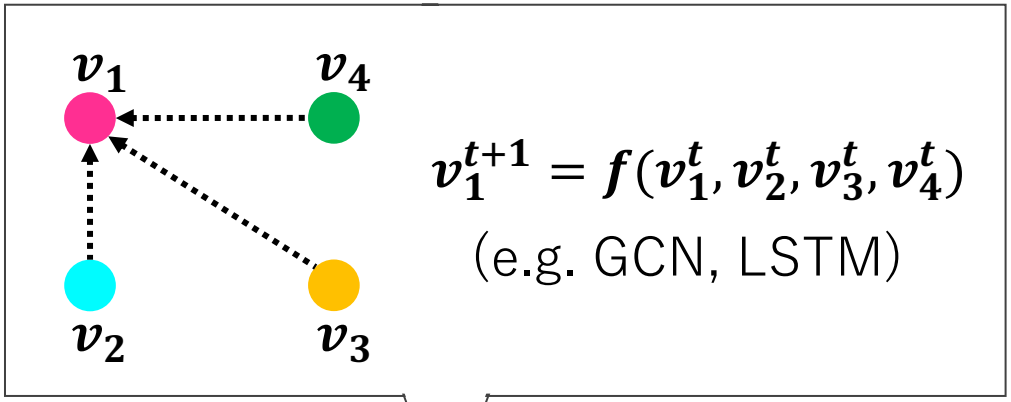


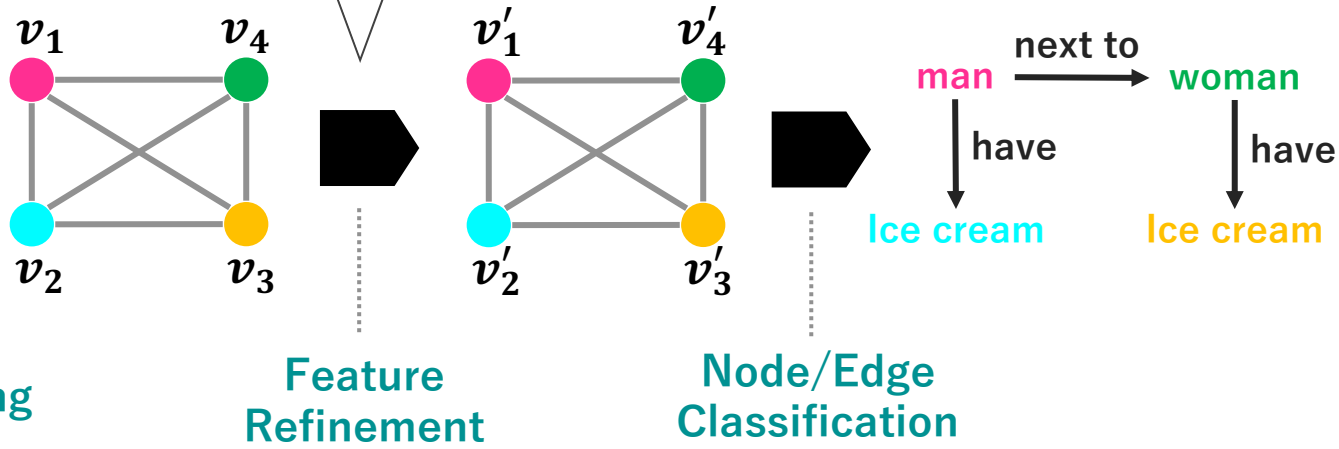


Background

Many existing methods contain a feature refinement step for maintaining global consistency.



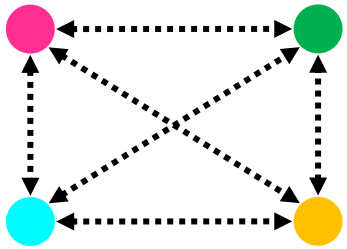




Motivation

The existing message passings.

If all nodes are connected with each other...(e.g. GCN)



✓ Permutable.

✗ Unrelated object pairs can be connected.

✗ It is unclear how global context is acquired.

If nodes are connected on a line... (e.g. LSTM)

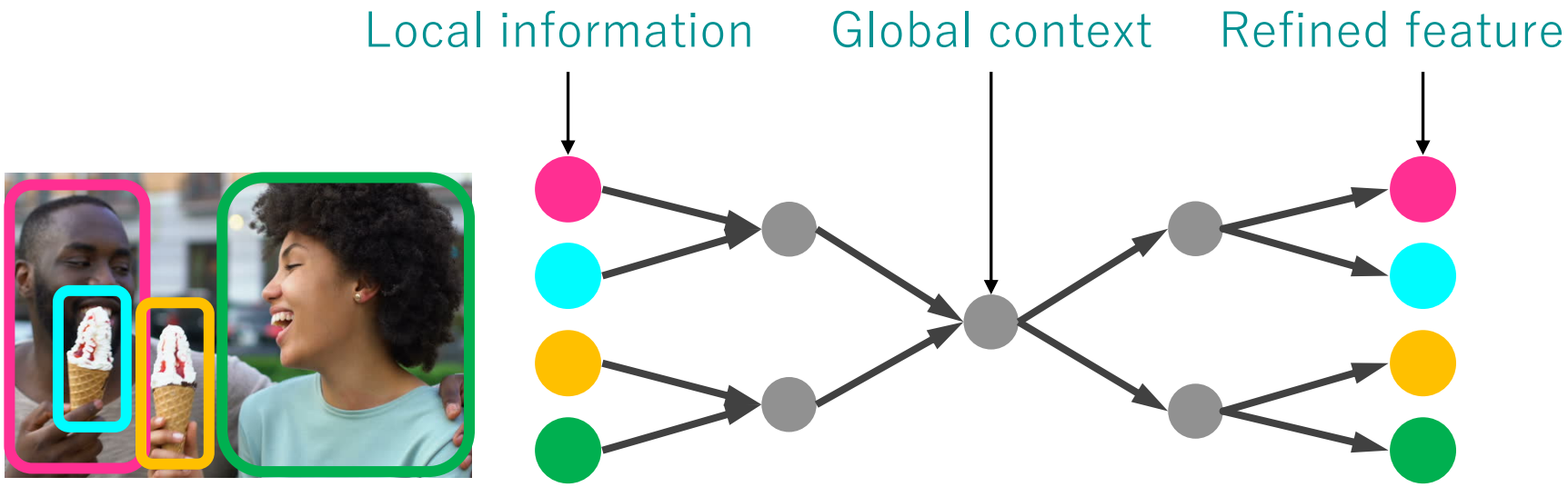
✓ The flow of global context is clear.

✗ Not permutable.

✗ Optimal ordering has not found.

Approach

We add some parent nodes for constructing a hierarchical structure.



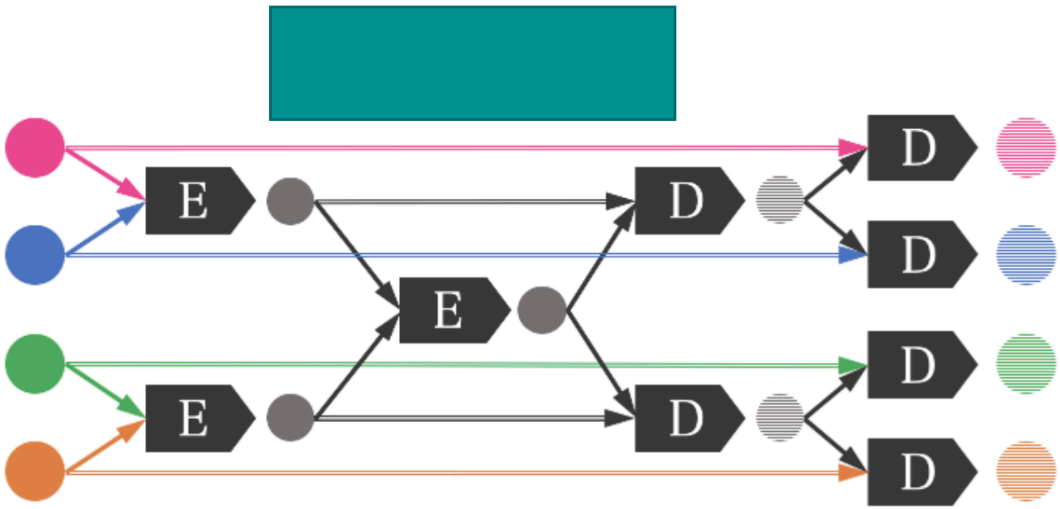
✓ It is clear where global context is maintained.

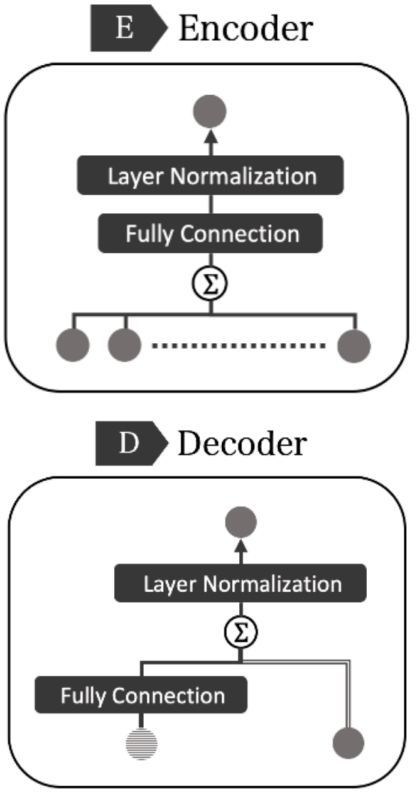
✓ No need to consider node ordering.

Proposal (Message Passing)

Encoding : combine features of child nodes into their parent nodes.

Decoding : backpropagate features of parent node to their child nodes.





Proposal (Structure Construction)

Joint Training

End-to-end training of hierarchical structure construction and scene graph prediction.

✓ Generated structures are expected to improve the scene graph generation performances.

✗ Training is harder.

Separated Training

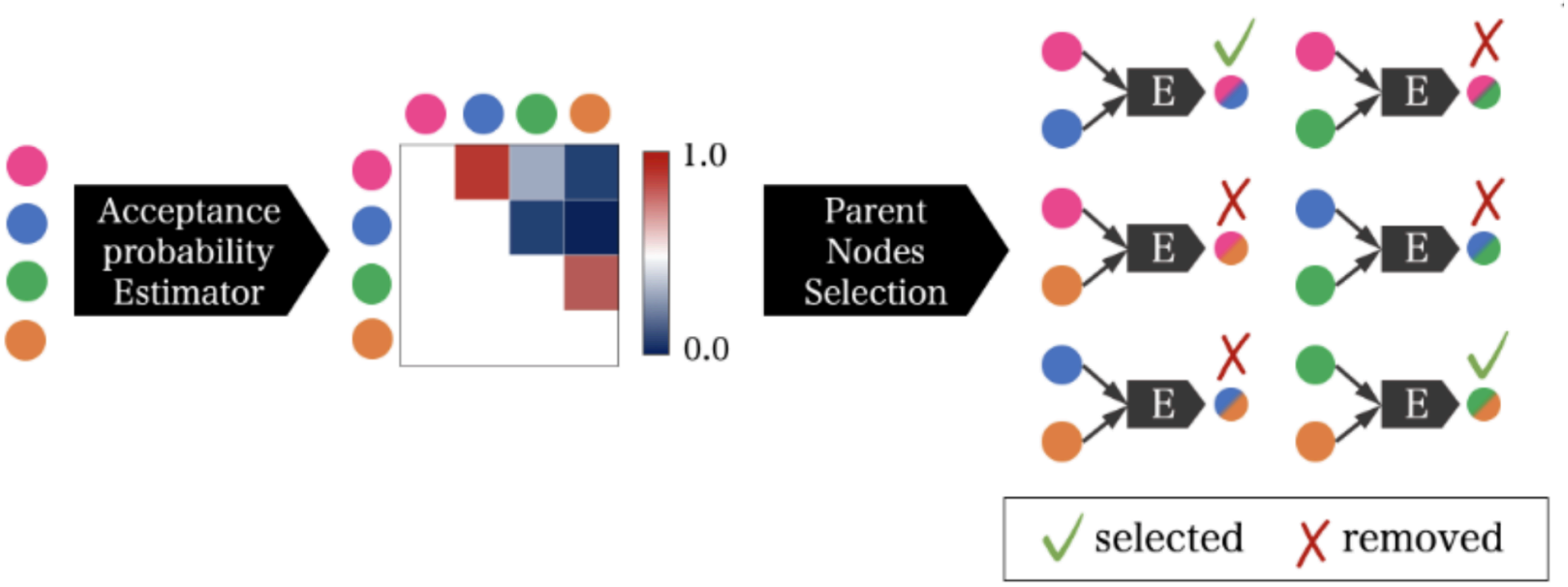
First train a structure construction module and then train a scene graph prediction module.

✓ More algorithms are adoptable to build hierarchical structures.

✗ Hierarchical structure are constructed regardless of the scene graph generation accuracy.

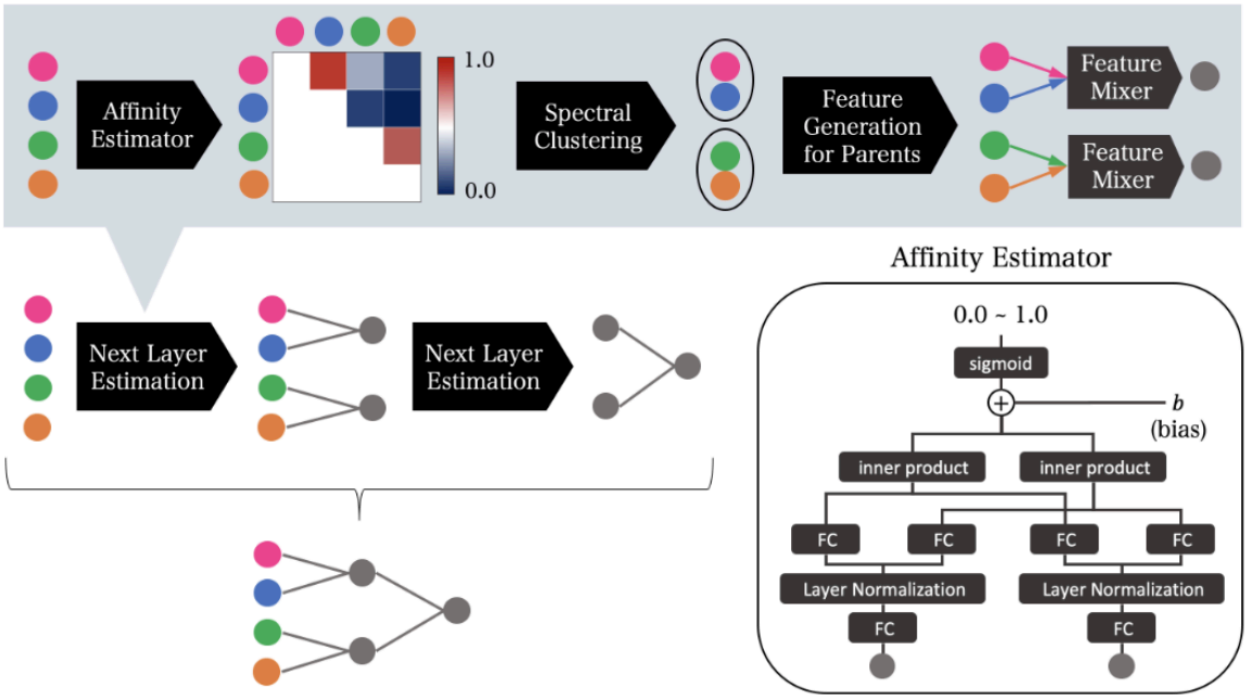
Joint Training

1. Generate parent nodes for every pair of nodes and calculate their confidence scores.
2. Delete low-confident parent nodes while leaving a certain number of nodes.
3. Repeat steps 1&2 until a hierarchical structure with predefined depth is built.



Separated Training

1. Generate a certain number of parent nodes by spectral clustering with calculated scores.
2. Provide features from child nodes to parent nodes.
3. Repeat steps 1&2 until a hierarchical structure with predefined depth is built.



Separated Training

The Affinity estimator is trained in unsupervised learning.

Data : Image pairs that are cropped by bounding boxes.

Label : 1 (if the image pair is extracted from the same image)

0 (otherwise)

Experimental setting

Dataset : VisualGenome

- (train) 70K images / (test) 30K images
- 150 object categories / 50 predicate categories

Evaluation metrics

- Recall@K (for scene graph generation performances)
- Accuracy (for object recognition performances)

Model

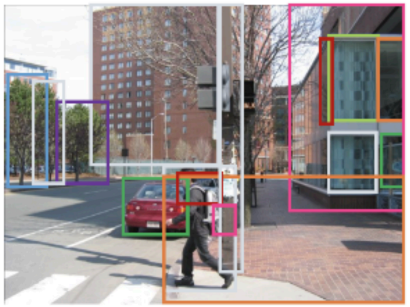
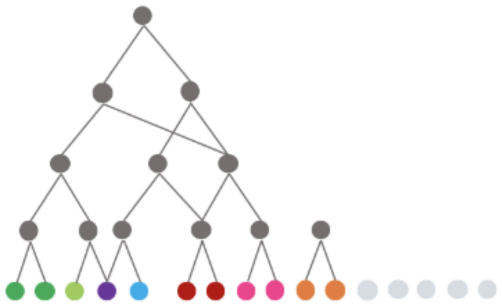
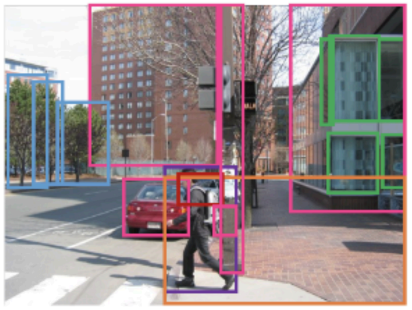
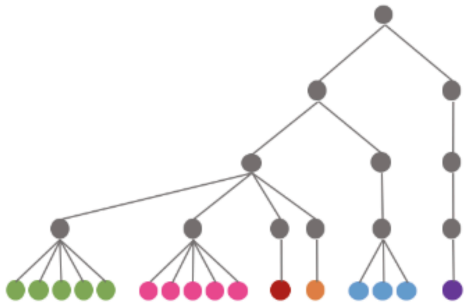
- Backbone: VGG 16-layers (pretrained on ImageNet)
- The number of hierarchical layers: 3
- The number of parent nodes: N/3 + 1

(N = the number of child nodes)

Experimental Results

We compared our proposed models and the existing models.

	Object Classification	Scene Graph Classification		Scene Graph Detection	
Model	Accuracy	R@50	R@100	R@50	R@100
CNN only	64.7	-	-	-	-
MOTIFNET [11]	66.8	35.8	36.5	27.2	30.3
Graph R-CNN [8]	65.9	35.3	36.0	11.4	13.7
VCTREE [6]	-	35.9	36.6	27.1	31.3
LinkNet [23]	67.0	36.0	36.7	28.2	32.1
Ours (Joint)	66.7	36.0	36.5	27.8	31.8
Ours (Separated)	**67.1**	**36.6**	**37.2**	**28.9**	**32.4**

Summary

We proposed scene graph generation models that build intuitive hierarchical structures to perform message passing.

Our proposed models are trained in two different ways: Separated training and the Joint training.

Our models that are trained in the Separated training achieved higher accuracy than the other models including the existing models on scene graph generation tasks.

The experimental result shows that a good hierarchical structure is very effective for generating scene graphs, and structure construction module can be trained on tasks unrelated to scene graph generation.