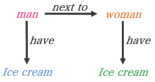
Exploring and Exploiting the Hierarchical Structure of a Scene for Scene Graph Generation

Ikuto Kurosawa, Tetsunori Kobayashi, Yoshihiko Hayashi
(Waseda University)

Overview

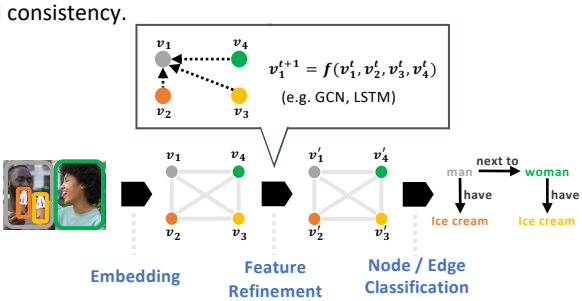
We propose novel neural network models for generating scene graphs that maintain global consistency.



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

Background

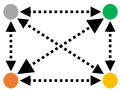
Many existing methods contain a feature refinement step for maintain global consistency.



Motivation

The message passing flow should be well managed.

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



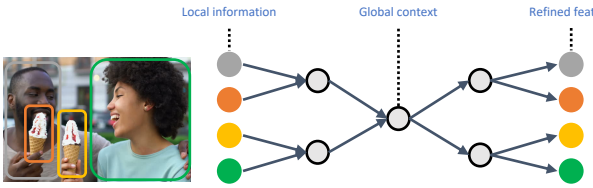
✓ Permutable.
✗ Unrelated object pairs can be connected.
✗ It is unclear how global context is acquired.

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.

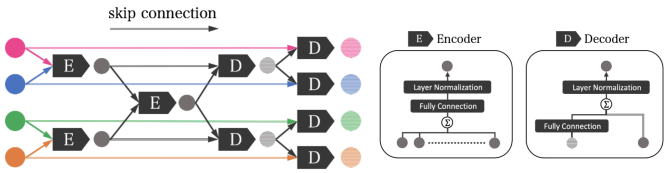


✓ It is clear where global context is held.
✓ Permutable for node ordering

Proposal

Message Passing Module

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

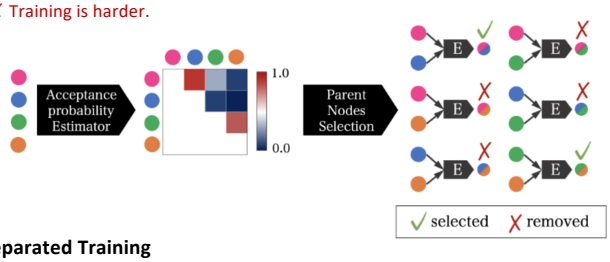


Structure Construction Module

Joint Training

simultaneous training of two modules in end-to-end.

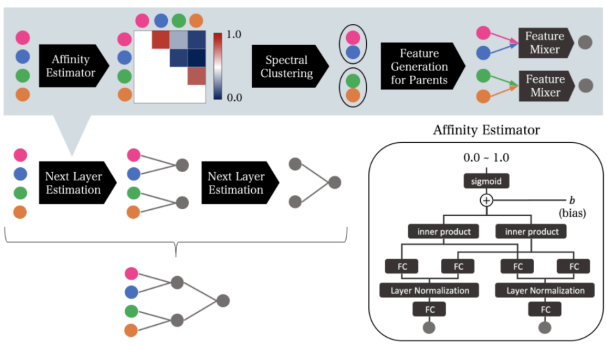
✓ Built structures might improve the scene graph generation performances.
✗ Training is harder.



Separated Training

Sequential training of two modules in separate tasks.

✓ More algorithms are adoptable to build hierarchical structures.
✗ The structure are constructed regardless of the scene graph generation accuracy.



Affinity estimator is trained in another task.

Data

Image pairs cropped by bounding boxes.

Label

1 (if extracted from the same image)
0 (otherwise)

Experiments

Dataset : VisualGenome

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

Evaluation metrics

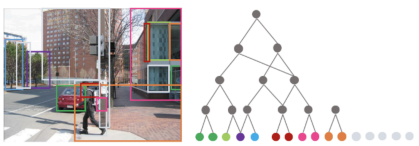
- Recall@K (for scene graph generation performances)
- Accuracy (for object recognition performances)

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)

	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**

Joint Training



Separated Training