Image as a single label

“king crab”

Image Source: ImageNet
Image as an object set

Image Source: ImageNet
Image as a scene graph

Relationships:

“Woman look at box”

“Man hold king crab”

“Woman wear coat”

“Man embrace woman”

Image Source: ImageNet
Image as a scene graph

Attributes:
“Red king crab”
“Transparent box”
“Blue coat”
“Smiling woman”
“Smiling Man”

Relationships:
“Woman look at box”
“Man hold king crab”
“Woman wear coat”
“Man embrace woman”

Image Source: ImageNet
Why we need scene graph?

Distinguish images more accurately


Left: https://cals.ncsu.edu/wp-content/uploads/2016/08/horse-1500x931.png
Why we need scene graph?

Describe images more grounding


Left: https://cals.ncsu.edu/wp-content/uploads/2016/08/horse-1500x931.png
Why we need scene graph?

Answer question more precisely

Q: What is the man walking with?
A: A horse

Q: Is the man feeding a horse?
A: Yes


Left: https://cals.ncsu.edu/wp-content/uploads/2016/08/horse-1500x931.png
Why we need scene graph?

Generate questions more grounding

Q: What animal is the man walking with?

Q: What is the man doting with the horse?

Visual System
Scene Graph generator

Communication

Human
Visual Question Answering

Answer Questions

Visual System
Scene Graph generator

Human
Visual Question Answering

Answer Questions

Ask Questions

Visual System
Scene Graph generator

Human

Visual Question Generation
Visual Question Answering

Human

Visual Question Generation

Ask Questions

Answer Questions

Visual System
Scene Graph generator
Skeleton Model
Skeleton Model

Input
Skeleton Model
Skeleton Model

Input

Region Proposals

RPN

ROI Pooling

ROI Pooling

Object Features

Relationship Features
Skeleton Model
Skeleton Model

Input → RPN → Region Proposals

ROI Pooling

Object Features → Object Scores

Relationship Features → Relationship Scores

Diagram with objects and relationships:
- Cup
- On
- Dog
- Hold
- In
- Person
- Book
- Watch
- TV
- Left of
- Right of
- Cat
Iterative Message Passing (IMP)

Scene Graph Generation by Iterative Message Passing. Xu et al. CVPR 2017
Multi-level Scene Description Network (MSDN)

Scene Graph Generation from Objects, Phrases and Region Captions. Li et al. ICCV 2017
Neural Motif Network

Neural Motifs: Scene Graph Parsing with Global Context. Zellers et al. CVPR 2018
Graph R-CNN (Our work)

Input

Region Proposals

RPN

Feature Updating

ROI Pooling

Object Features

Message Passing

Object Scores

Relationship Features

Score Updating

Relationship Scores

Neural Motifs: Scene Graph Parsing with Global Context. Zellers et al. CVPR 2018
Motivations
Motivations

1. Objects in a scene usually have relationships with others;
Motivations

1. Objects in a scene usually have relationships with others;

2. Not all object pairs have relationships, the scene graph is usually sparse;
Motivations

1. Objects in a scene usually have relationships with others;

2. Not all object pairs have relationships, the scene graph is usually sparse;

3. Existence of relationships highly depends on the object categories, and type of relationships highly depends on the context.
Framework

Conv Feature
Framework

Conv Feature

Dense graph
1. Relation proposal network (RePN) to learn to prune the densely connected scene graph;
1. Relation proposal network (RePN) to learn to prune the densely connected scene graph;
2. Attentional graph convolutional networks (aGCN) to incorporate the contextual information.
1. Relation proposal network (RePN) to learn to prune the densely connected scene graph;
2. Attentional graph convolutional networks (aGCN) to incorporate the contextual information.
Framework

\[ P(S|I) \]

I: Input Image; S: Scene graph
V: Scene graph vertices (object)
E: Scene graph edges (relationship)
O: Scene graph object labels
R: Scene graph relationship labels
Framework

Conv Feature → Relational Proposal Network → aGCN

Scene Graph

Region Proposal

\[ P(S|I) = P(V|I) \]
Framework

I: Input Image; S: Scene graph
V: Scene graph vertices (object)
E: Scene graph edges (relationship)
O: Scene graph object labels
R: Scene graph relationship labels

Relation Proposal

\[ P(S|I) = P(V|I) \cdot P(E|V, I) \]
**Framework**

I: Input Image; S: Scene graph  
V: Scene graph vertices (object)  
E: Scene graph edges (relationship)  
O: Scene graph object labels  
R: Scene graph relationship labels

\[
P(S|I) = P(V|I) \cdot P(E|V,I) \cdot P(R,O|V,E,I)
\]

Relation Proposal

Graph Labeling
Training

\[ P(V|I) \cdot P(E|V,I) \cdot P(R,O|V,E,I) = P(S|I) \]
Training

Region Proposal Network

Binary Cross Entropy Loss

\[
P(V|I) \cdot P(E|V, I) \cdot P(R, O|V, E, I) = P(S|I)
\]
Training

\[
P(V|I) \quad P(E|V,I) \quad P(R,O|V,E,I) = P(S|I)
\]
Training

Binary Cross Entropy Loss
Region Proposal Network

$P(V|I)$  $P(E|V,I)$  $P(R,O|V,E,I) = P(S|I)$

Two Cross Entropy Losses, one for node and one for edge
Graph Labeling Network

Binary Cross Entropy Loss
Relation Proposal Network

Binary Cross Entropy Loss
Metrics

[1]. Scene Graph Generation by Iterative Message Passing. Xu et al. CVPR 2017
Metrics

Assume there are $N$ objects extracted from an image, then $N \times (N - 1)$ edges

[1]. Scene Graph Generation by Iterative Message Passing. Xu et al. CVPR 2017
Metrics

Assume there are $N$ objects extracted from an image, then $N \times (N - 1)$ edges.

Step 1: Take maximum for object scores and predicate scores, excluding background class.

[1]. Scene Graph Generation by Iterative Message Passing. Xu et al. CVPR 2017
Metrics

Assume there are $N$ objects extracted from an image, then $N \times (N - 1)$ edges

Step 1: Take maximum for object scores and predicate scores, excluding background class.

Step 2: Compute relationship scores: $Rel(i, j) = Subj(i) \times Obj(j) \times Pred(i, j)$

[1]. Scene Graph Generation by Iterative Message Passing. Xu et al. CVPR 2017
Metrics

Assume there are $N$ objects extracted from an image, then $N \times (N - 1)$ edges

Step 1: Take maximum for object scores and predicate scores, excluding background class.

Step 2: Compute relationship scores: $Rel(i,j) = Subj(i) \times Obj(j) \times Pred(i,j)$

Step 3: Sort the relationship triplets in a descending order.

[1]. Scene Graph Generation by Iterative Message Passing. Xu et al. CVPR 2017
Metrics

Assume there are $N$ objects extracted from an image, then $N \times (N - 1)$ edges

Step 1: Take maximum for object scores and predicate scores, excluding background class.

Step 2: Compute relationship scores: $Rel(i, j) = Subj(i) \times Obj(j) \times Pred(i, j)$

Step 3: Sort the relationship triplets in a descending order.

Step 4: Compute the triplet recalls ($Recall@50$, $Recall@100$) based on the ground-truth

\[
SGGen: \text{Recall} = \frac{C(T_{\text{pred and } T_{\text{gt}}})}{N(T_{\text{gt}})} \quad \text{IoU} > 0.5
\]

[1]. Scene Graph Generation by Iterative Message Passing. Xu et al. CVPR 2017
Metrics

Assume there are $N$ objects extracted from an image, then $N \times (N - 1)$ edges

Step 1: Take maximum for object scores and predicate scores, excluding background class.

Step 2: Compute relationship scores: $\text{Rel}(i, j) = \text{Subj}(i) \times \text{Obj}(j) \times \text{Pred}(i, j)$

Step 3: Sort the relationship triplets in a descending order.

Step 4: Compute the triplet recalls (Recall@50, Recall@100) based on the ground-truth

\[
\text{SGGen: Recall} = \frac{C(T_{pred \text{ and } T_{gt}})}{N(T_{gt})} \quad \text{IoU} > 0.5
\]

\textbf{PhrCls: all object locations are known} \hspace{1cm} \textbf{PredCls: all object locations and labels are known}

[1]. Scene Graph Generation by Iterative Message Passing. Xu et al. CVPR 2017
Experiments

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Backbone</th>
<th>#objects</th>
<th>#predicates</th>
<th>Metrics</th>
</tr>
</thead>
</table>

Comparing with Previous Work

Recall@100

Our model has over four point improvement on SGGen, and two point on SGGen+

[1] Scene Graph Generation by Iterative Message Passing. Xu et al. CVPR 2017
[2] Scene Graph Generations from Objects, Phrases and Captions. Li et al. ICCV 2017
Qualitative Results

short
on
on
surfboard
in
water
man
ride
wave
has
arm

bear
has
near
flower
leg
ear

behind
branch
of
head
bird
in
tree
behind
has
wing
behind
leaf
Ablation Study

- PredCls
- PhrCls
- SGGen
- SGGen+
- mAP

Graph showing Recall@50 for different models:
- Base
- Base+RePN
- Base+RePN+GCN
- Base+RePN+aGCN
Ablation Study

RePN improves SGGen, SGGen+ and mAP
Ablation Study

GCN/aGCN improves PredCls and PhrCls

RePN improves SGGGen, SGGGen+ and mAP
Ablation Study

GCN/aGCN improves PredCls and PhrCls

RePN improves SGGGen, SGGGen+ and mAP
A new codebase for scene graph generation

https://github.com/jwyang/graph-rcnn.pytorch

This project is a set of reimplemented representative scene graph generation models based on Pytorch 1.0, including:

- **Graph R-CNN for Scene Graph Generation**, our own. ECCV 2018.
- **Scene Graph Generation by Iterative Message Passing**, Xu et al. CVPR 2017
- **Scene Graph Generation from Objects, Phrases and Region Captions**, Li et al. ICCV 2017
- **Neural Motifs: Scene Graph Parsing with Global Context**, Zellers et al. CVPR 2018
- **Graphical Contrastive Losses for Scene Graph Generation**, Zhang et al, CVPR 2019

The goal of gathering all these representative methods into a single repo is to establish a more fair comparison across different methods under the same settings.

**Welcome to contribute!**
Summary

• Takeaways:
  • Introducing a general base model for scene graph generation
  • Pruning the fully-connected graph is important for scene graph generation
  • Exploiting the context across objects and predicates is crucial
  • Scene graph generation helps to improve object detection

• Challenges:
  • The dataset is noisy (incomplete and inconsistent annotations)
  • Relationships need more fine-grained categorizing (spatial, semantic, etc)
  • Rare/novel relationship is hard to detect
Visual Question Answering

Visual System
Scene Graph generator

Human

Answer Questions

Ask Questions

Visual Question Generation
Visual Question Answering

Visual Question Answering is a challenging task that involves fully visual understanding, language understanding and reasoning.
Methodology: Graph Reasoning Machine

- **Visual Understanding**: Scene Graph Generator (extract object, attribute and relationship between objects).
- **Language Understanding**: Program Generator (extract logic reasoning chain in the question)
- **Reasoning**: Learnable Neural-Symbolic Executor (execute programs on s.g.)

**Pros:**
- Make the VQA model more interpretable;
- Easy to diagnose and analyze the model predictions;
- Modules are disentangled from each other;
- Introduces no or few language priors, (probably) better generalization ability;

*Joint work with Chuang Gan et al*
Compositional Reasoning VQA Dataset

Pattern: What|Which <type> {do you think} <is> <dobject>, <attr> or <decoy>?
Program: Select:<dobject> → Choose <type>:{attr}|<decoy>
Reference: The food on the red object left of the small girl that is holding a hamburger
Decoy: brown

What color is the food on the red object left of the small girl that is holding a hamburger, yellow or brown?
Select: hamburger → Relate: girl, holding → Filter size: small → Relate: object, left → Filter color: red → Relate: food, on → Choose color: yellow | brown

Graph Normalization
- Ontology construction
- Edge Pruning
- Object Augmentation
- Global Properties

Question Generation
- Pattern Collection
- Compositional References
- Decoy Selection
- Probabilistic Generation

Sampling and Balancing
- Distribution Balancing
- Type-Based Sampling
- Deduplication

Entailment Relations
- Functional Programs
- Entailment Relations
- Recursive Reachability

New Metrics
- Consistency
- Validity & Plausibility
- Distribution
- Grounding

The shorts have what color? [p]gray
What color is the frisbee? [p]yellow
Does the man to the right of the frisbee wear glasses? [p]no
Are there both fences and frisbees in the picture? [p]no
Is the logo different in color than the roof? [p]'I do not know'
Who wears shorts? [p]man
Q: The shorts have what color?

P: Filter( shorts ) -> Query( color )

The shorts have what color? [ p ] gray

What color is the frisbee? [ p ] yellow

Does the man to the right of the frisbee wear glasses? [ p ] no

Are there both fences and frisbees in the picture? [ p ] no

Is the logo different in color than the roof? [ p ] no

Who wears shorts? [ p ] man
Q: The shorts have what color?

P: Filter(shorts) -> Query(color)

The shorts have what color?  yes [p]gray
What color is the frisbee?  yes [p]yellow
Does the man to the right of the frisbee wear glasses?  no [p]no
Are there both fences and frisbees in the picture?  no [p]no
Is the logo different in color than the roof?  yes [p]"I do not know"
Who wears shorts?  yes [p]man
Q: The shorts have what color?
A: gray

SG:
- shorts: 0.54
- gray: 0.47
- brown: 0.19
Q: What color is the frisbee?

P: Filter(frisbee) -> Query(color)

A: yellow
Q: Who wears shorts?

A: Man

The shorts have what color? [p] gray
What color is the frisbee? [p] yellow
Does the man to the right of the frisbee wear glasses? [p] no
Are there both fences and frisbees in the picture? [p] no
Is the logo different in color than the roof? [p] no
Who wears shorts? [p] man
Visual System
Scene Graph generator

Human

Visual Question Answering
Answer Questions

Visual Question Generation
Ask Questions

Ask Questions

Visual Question Answering
The Open-World Recognition Problem
The Open-World Recognition Problem
The Open-World Recognition Problem

What is the black object on the top of the table at left side?

That’s a coffee bottle.

Human (Oracle)
The Open-World Recognition Problem

How to train an agent to ask questions about its unknowns based on its knowns to improve its visual understanding capabilities?
Visual Question Generation: Visual Curiosity

Agent Architecture

Image

Visual System

Graph Memory

Question Generator

Answer Digestor

Same Content

Oracle/Human

Question

Answer

Agent

Slide Credit: Stefan Lee
Bottom-up Update

Visual Graph

Graph Memory
Bottom-up Update

Visual Memory
Bottom-Up Update

Visual Graph

Graph Memory
Top-down Update

Visual Memory

Memory Graph

Q: What is the color of the right most object?
A: Orange
Top-down Update

Visual Memory Top-Down Update

Q: What is the color of the right most object?

A: Orange

Graph Memory
Question Generator

Target: 1
Attribute: Shape
Reference: None
Question: What is the shape of the front most large red object?

3
Color

What is the color of the metal cube on the left side of a small object?

5
Material

What is the material of object at left side of metal cube?
Training Objective: Visual System

**Loss:** cross-entropy loss between the graph memory and the visual predictions over all images, objects, and attributes

\[ \theta_v^* = \arg \min \sum_{l_i \in I} \sum_{k=1}^{K_i} \log \left( \begin{pmatrix} \vdots \end{pmatrix} \right) \]

Update visual system after a piece of dialogs with Oracle/Human.
Training Objective: Questioner Policy

Oracle Graph

Reward: \[ r_t^i = S(G_t^i, G_i^*) - S(G_t^{i-1}, G_i^*) \]

Optimal Policy: \[ \theta_{\pi}^* = \arg \max E_{V} E_{I \sim \mathcal{E}} E_{\pi_q} \left[ \sum_{i=1}^{n} \sum_{t=1}^{T} r_t^i (q_t^i \sim \pi_q (h_t^i; \theta_{\pi})) \right] \]

Use A2C and update policy after each episode based on all rounds

Can be improved by:
- Asking unambiguous, informative questions (top-down)
- Improving the visual system quickly (bottom-up)
Experiments: Environments

**Synthesized Dataset**
Different shapes, colors, materials and sizes. Extended from CLEVR dataset [1]


**Realistic Dataset**
Various real indoor scenes. Annotated based on the ARID dataset [2]

Experiment: Standard Training + Standard Testing

Consistent improvements over heuristic baselines especially over longer dialogs

Graph Recovery
Experiments: Novel Object Environments

**Novel**
- New colors and shapes
- 600 images for test (12 episodes)

**Mixed**
- Mix of novel and standard colors and shapes
- 600 images for test (12 episodes)

**Realistic**
- 51 categories, 11 colors, 6 materials
- 1200 images for test (24 episodes)
Experiments: Standard Train – New Test Environments

- Std-Std
- Std-Novel
- Std-Mixed
- Std-Realistic

Graph Recovery

Practically no loss of performance in synthetic settings and small reductions for realistic (many more categories)

- R@10: 42.1, 43.3, 42.9, 35.6
- R@20: 59.1, 58.4, 60.1, 53.4
- R@50: 89.3, 88.9, 90.3, 86.2
Experiments: Visual System Performance

Standard

![Accuracy vs Number of Images for Standard](image)

Novel

![Accuracy vs Number of Images for Novel](image)

Mixed

![Accuracy vs Number of Images for Mixed](image)

Realistic

![Accuracy vs Number of Images for Realistic](image)
What material is the leftmost thing? food

There is a leftmost object; what is it? potato

The leftmost object is what color? brown

What is the closest thing that is in front of the yellow plastic ball made of? paper

What is the closest thing that is in front of the yellow plastic ball? cereal
Summary for this part

• Takeaways:
  • Scene graph can be used as a comprehensive semantic abstraction of image
  • Scene graph provides grounding information for language-based interaction with human, especially visual question answering and generation
  • Scene graph gives it a chance to make models more interpretable and explainable

• Potential Directions:
  • Leverage scene graph for explicit and effective reasoning on more vision-language tasks, such as expression coreference
  • Language context dependent scene graph generation
  • Combine scene graph and knowledge graph for common sense reasoning
Summary for all

Leveraging external "knowledge" when interpreting images
Specifically, using richer vision, language models/data to improve vision+language models

Representing internal structure in images
Specifically, scene graphs: generating, evaluating and using them for vision+language
Thanks!