Lecture 17:

Scene Graphs and Graph Convolutions

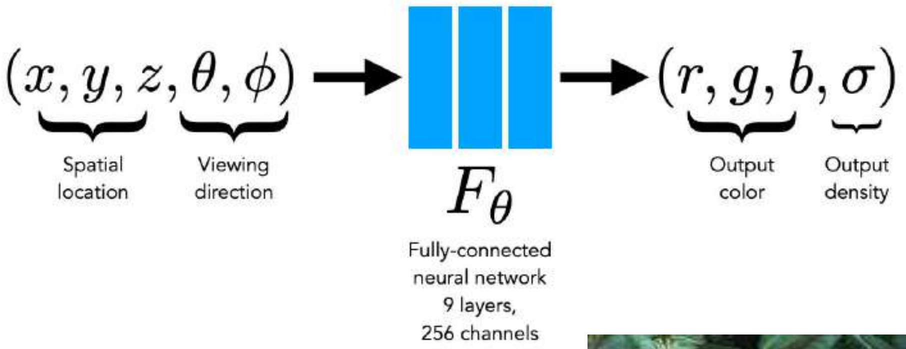
Administrative

- A3 grades will be released next week

- Final project reports due June 3rd
  - Final project video due June 4th
  - No late days for final project

- 2 guest lectures next week:
  - First on multimodal learning combining vision and sound
  - Second on combining vision with action

Last time: NeRF



Today's agenda

- Beyond objects
- Scene Graphs
- Scene Graph Generation
- Graph Convolutional Networks

Computer vision was focused on disconnected objects

Image Classification

Object Detection

Instance Segmentation

Shilane et al, 2004; Fei-Fei et al, 2004; Griffin et al, 2006; Russell et al, IJCV 2007; Torralba et al, TPAMI 2008; Chen et al, SIGGRAPH 2009; Quattoni and Torralba, CVPR 2009; Deng et al, CVPR 2009; Xiao et al, CVPR 2010; Everingham, IJCV 2010; Silberman et al, ECCV 2012; Xiao et al, ICCV 2013; Lim et al, ICCV 2013; Lin et al, ECCV 2014; Zhou et al, NIPS 2014; Russakovsky et al, IJCV 2015; Chen et al, arXiv 2015; Chang et al, 2015 image #1

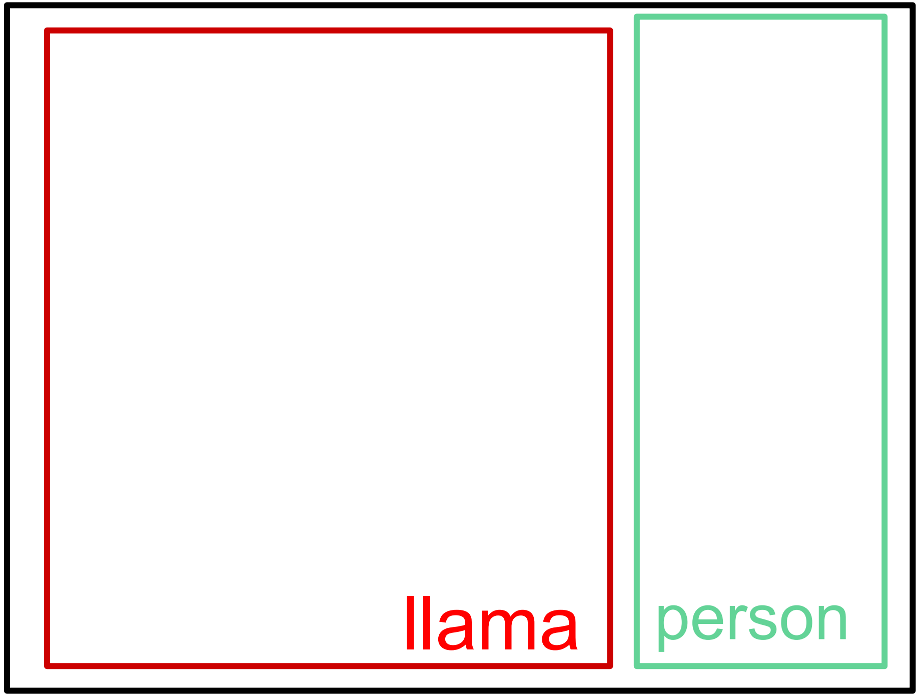
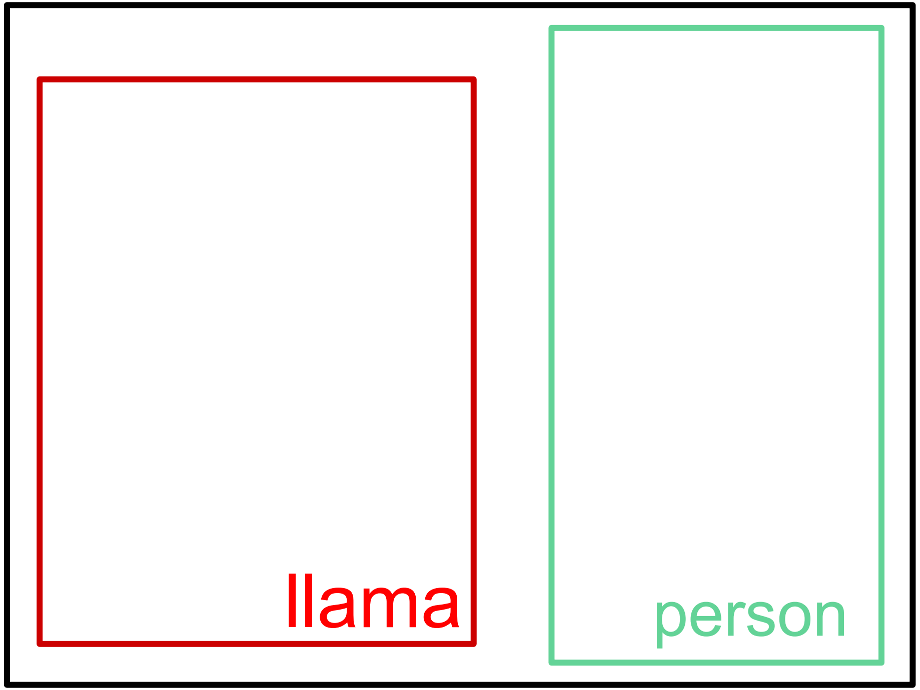
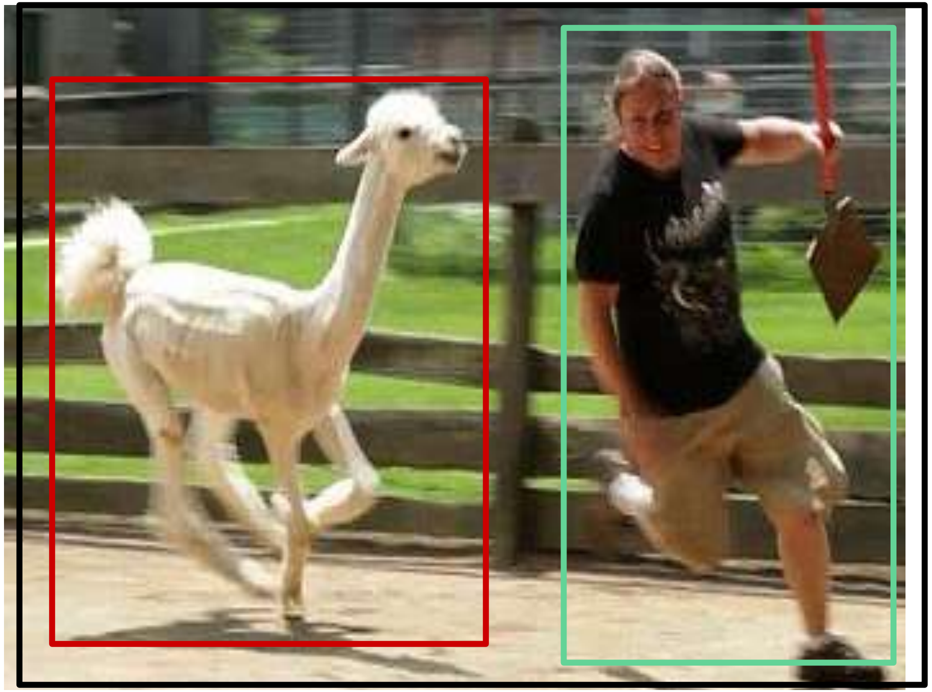


image #2



Fei-Fei et al, 2004; Griffin et al, 2006; Torralba et al, TPAMI 2008; Quattoni and Torralba, CVPR 2009; Deng et al, CVPR 2009; Xiao et al, CVPR 2010; Zhou et al, NIPS 2014; Russakovsky et al, IJCV 2015

next to



chasing

Fei-Fei et al, 2004; Griffin et al, 2006; Torralba et al, TPAMI 2008; Quattoni and Torralba, CVPR 2009; Deng et al, CVPR 2009; Xiao et al, CVPR 2010; Zhou et al, NIPS 2014; Russakovsky et al, IJCV 2015

Can image captioning models capture this information?

A man walking a dog

- Wrong! Not a dog
- Wrong! Not walking
- Missed ribbon held by person
- Missed any descriptions of the llama (the model could have said that they are next to one another or that they are in front of the wall).

Lin et al, ECCV 2014
Chen et al, arXiv 2015

What information would people convey if asked to caption?

A llama standing next to a person

White llama in front of a blue wall

A huacaya alpaca held by a person who is holding a big ribbon

What information would people convey if asked to caption?

Objects

A llama standing next to a person

White llama in front of a blue wall

A huacaya alpaca held by a person who is holding a big ribbon

What information would people convey if asked to caption?

Objects Attributes

A llama standing next to a person

White llama in front of a blue wall

A huacaya alpaca held by a person who is holding a big ribbon

What information would people convey if asked to caption?

Objects Attributes Relationships

A llama standing next to a person

White llama in front of a blue wall

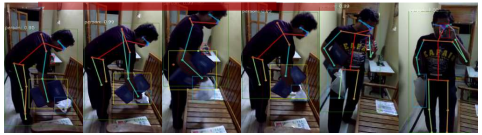
A huacaya alpaca held by a person who is holding a big ribbon

Many Vision tasks share a similar underlying structure

Action classification

action: drinking from a cup

action: take notebook from somewhere



Grounding objects

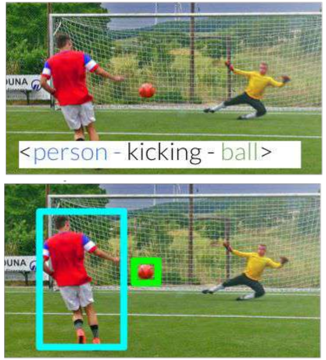
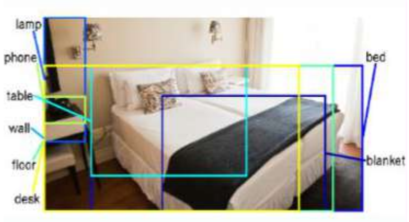


Image retrieval

Black phone is on top of white, wooden desk. The desk is next to a clean white bed that has a black blanket and is next to a white table. The lamp is on a tan wall. The table is by the bed, which is next to the phone. The floor is under the bed, table, lamp and blanket.



Question answering

how many types of vegetables are there?

is the food in the foreground prickly?

how many types of fruits are there?

is the food healthy?

how many people are in the photo?

is this a busy street?

how many skateboards are there?

is the man wearing a hat?

Agrawal, et al. Vqa: Visual question answering, ICCV 2015
Swets et al. Using discriminant eigenfeatures for image retrieval, TPAMI 1996
Yu et al. Modeling context in referring expressions, ECCV 2016
Simonyan et al. Two-stream convolutional networks for action recognition in videos ,NeurIPS 2014

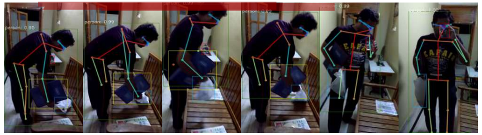
Ji, Krishna et al. Action Genome: Actions as Compositions of Spatio-Temporal Scene Graphs, CVPR 2020
Krishna et al. Information Maximizing Visual Question Answering, CVPR 2019
Krishna et al. Referring Relationships, CVPR 2018
Johnson, Krishna et al. Image Retrieval with Scene Graphs, CVPR 2015

Many Vision tasks share a similar underlying structure

Action classification

action: drinking from a cup

action: take notebook from somewhere



Grounding objects

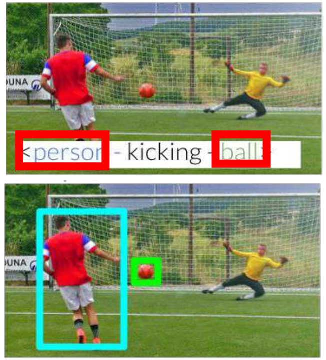
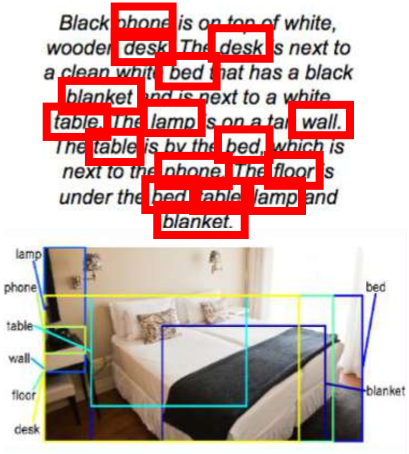


Image retrieval



Question answering

Agrawal, et al. Vqa: Visual question answering, ICCV 2015
Swets et al. Using discriminant eigenfeatures for image retrieval, TPAMI 1996
Yu et al. Modeling context in referring expressions, ECCV 2016
Simonyan et al. Two-stream convolutional networks for action recognition in videos ,NeurIPS 2014

Ji, Krishna et al. Action Genome: Actions as Compositions of Spatio-Temporal Scene Graphs, CVPR 2020
Krishna et al. Information Maximizing Visual Question Answering, CVPR 2019
Krishna et al. Referring Relationships, CVPR 2018
Johnson, Krishna et al. Image Retrieval with Scene Graphs, CVPR 2015

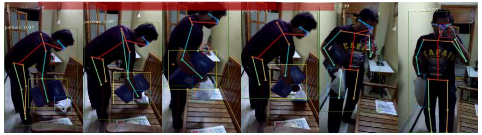
Many Vision tasks share a similar underlying structure

objects
attributes

Action classification

action: drinking from a cup

action: take notebook from somewhere



Grounding objects

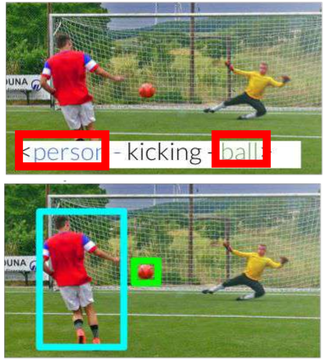
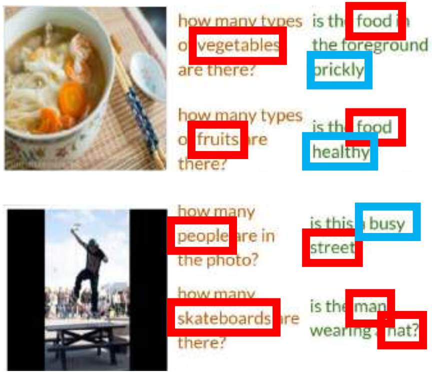


Image retrieval

Question answering



Agrawal, et al. Vqa: Visual question answering, ICCV 2015
Swets et al. Using discriminant eigenfeatures for image retrieval, TPAMI 1996
Yu et al. Modeling context in referring expressions, ECCV 2016
Simonyan et al. Two-stream convolutional networks for action recognition in videos ,NeurIPS 2014

Ji, Krishna et al. Action Genome: Actions as Compositions of Spatio-Temporal Scene Graphs, CVPR 2020
Krishna et al. Information Maximizing Visual Question Answering, CVPR 2019
Krishna et al. Referring Relationships, CVPR 2018
Johnson, Krishna et al. Image Retrieval with Scene Graphs, CVPR 2015

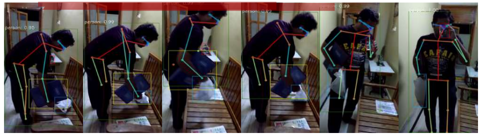
Many Vision tasks share a similar underlying structure

objects
attributes
relationships

Action classification

action: drinking from a cup

action: take notebook from somewhere



Grounding objects

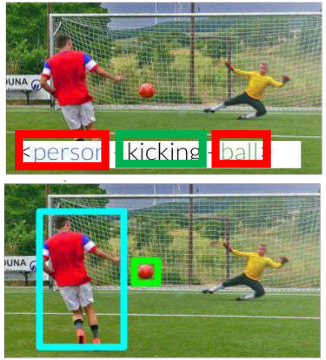
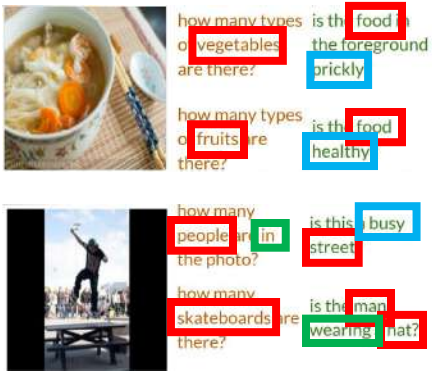


Image retrieval

Question answering

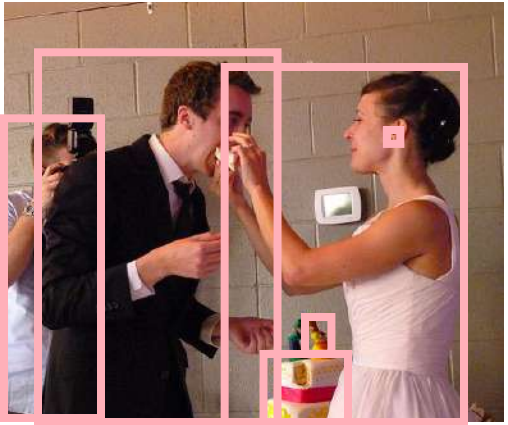


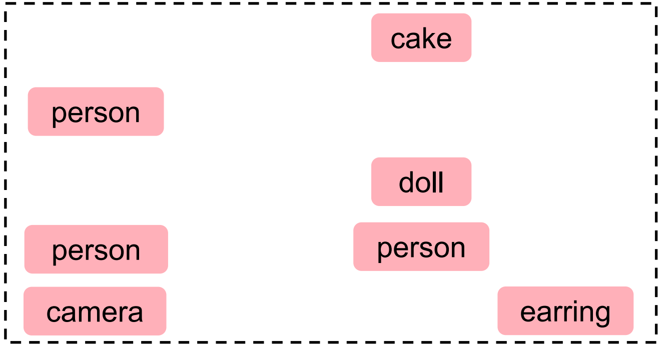
Agrawal, et al. Vqa: Visual question answering, ICCV 2015
Swets et al. Using discriminant eigenfeatures for image retrieval, TPAMI 1996
Yu et al. Modeling context in referring expressions, ECCV 2016
Simonyan et al. Two-stream convolutional networks for action recognition in videos ,NeurIPS 2014
Ji, Krishna et al. Action Genome: Actions as Compositions of Spatio-Temporal Scene Graphs, CVPR 2020
Krishna et al. Information Maximizing Visual Question Answering, CVPR 2019
Krishna et al. Referring Relationships, CVPR 2018
Johnson, Krishna et al. Image Retrieval with Scene Graphs, CVPR 2015

The scene graph representation

Krishna et al., Visual Genome: Connecting Vision and Language using Crowdsourced Image Annotations, IJCV 2017

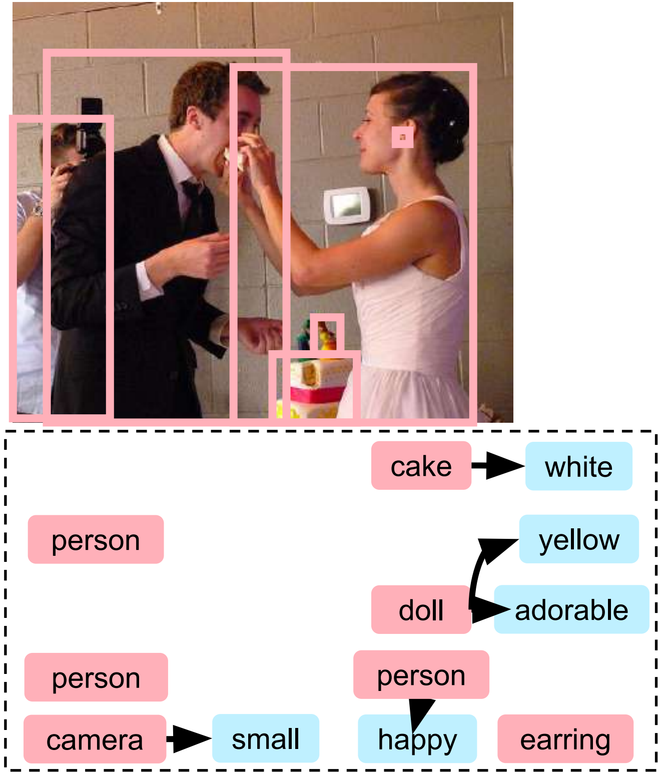
The scene graph representation





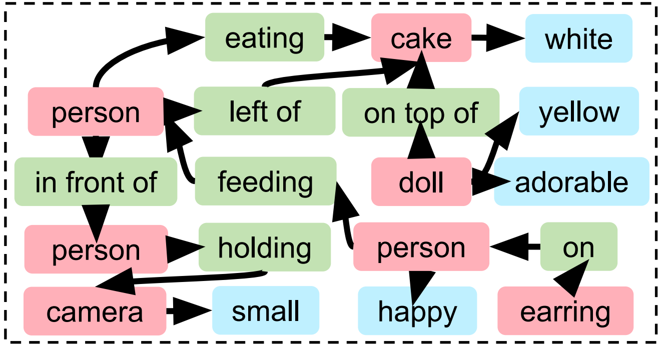
Krishna et al., Visual Genome: Connecting Vision and Language using Crowdsourced Image Annotations, IJCV 2017

The scene graph representation



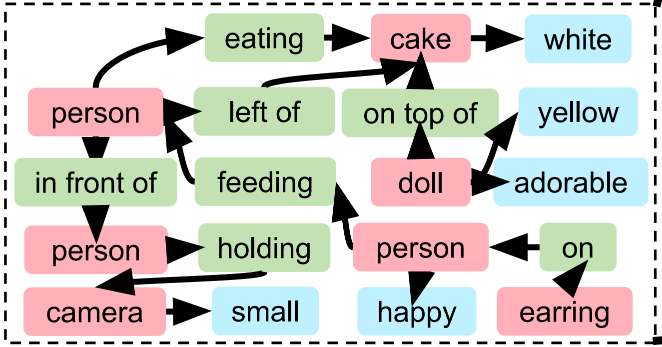
Krishna et al., Visual Genome: Connecting Vision and Language using Crowdsourced Image Annotations, IJCV 2017

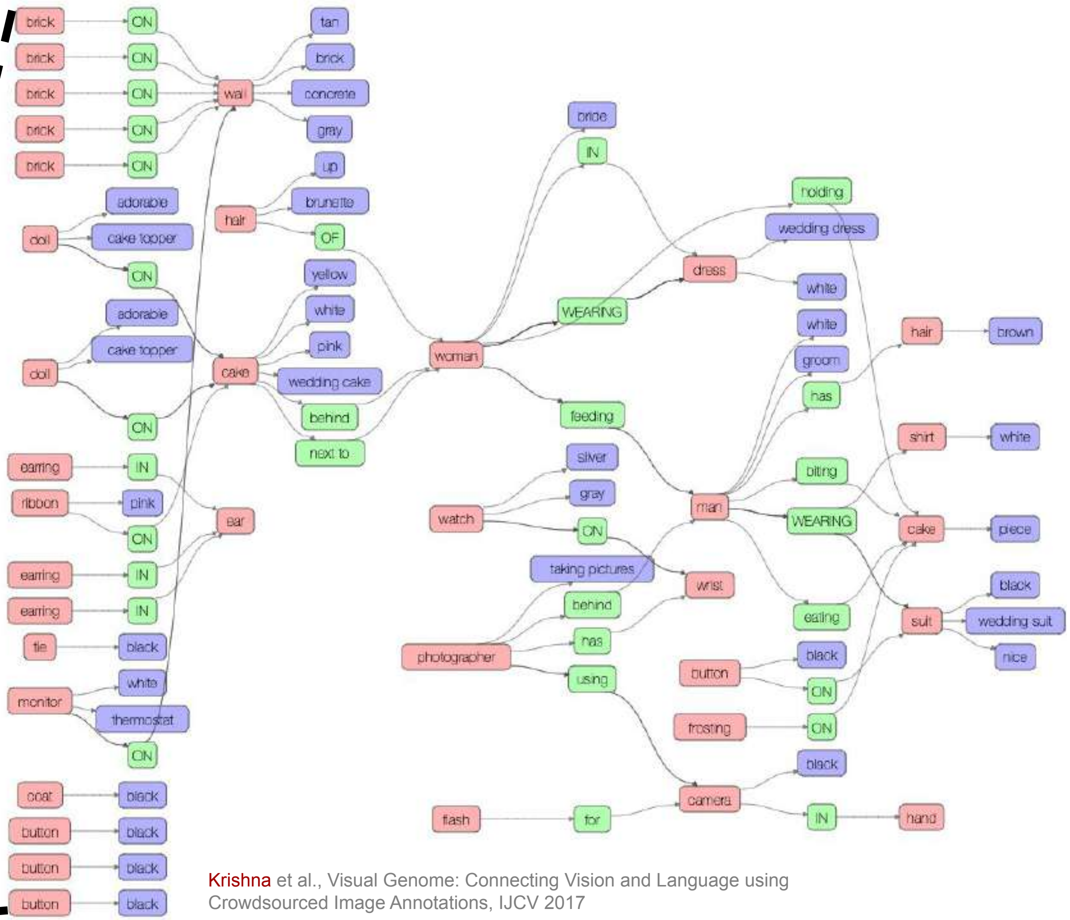
The scene graph representation



Krishna et al., Visual Genome: Connecting Vision and Language using Crowdsourced Image Annotations, IJCV 2017

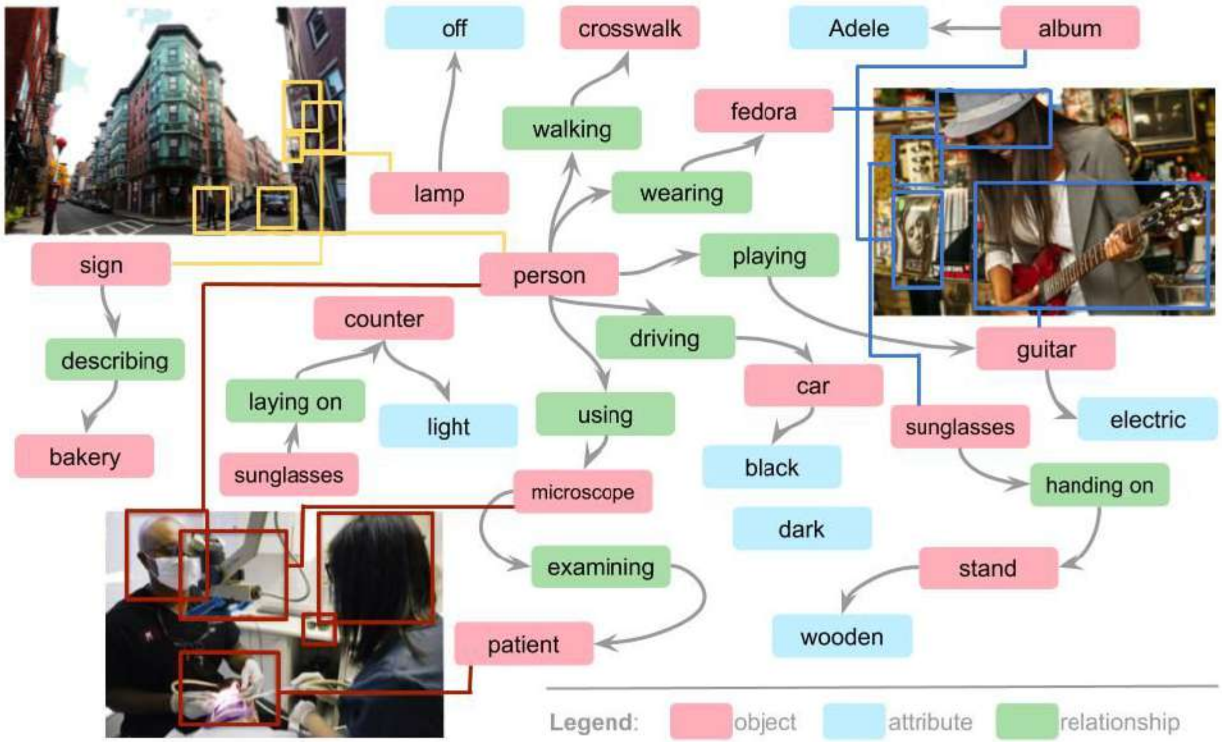
The scene graph representation





Krishna et al., Visual Genome: Connecting Vision and Language using Crowdsourced Image Annotations, IJCV 2017

Visual Genome – connects images together with scene graphs



108K images
3.8 Million Objects
2.8 Million Attributes
2.3 Million Relationships
1.7 Million question answers
5.4 Millions descriptions

Everything Mapped to Wordnet Synsets

Code and dataset available:
http://visualgenome.org
Visualization code:
https://github.com/ranjaykrishna/graphviz

Krishna et al., Visual Genome: Connecting Vision and Language using Crowdsourced Image Annotations, IJCV 2017

But why is scene graph the right representation?

Try and remember all these images

Do you remember seeing this image?

a

b

c

d

The difficulty with the appealing idea that we remember the gist of a scene is that there is no consensus about the contents of a 'gist'. Intuition suggests that an inventory of some of the objects in the scene should be at least a part of the gist.

Wolfe, *Visual Memory: What do you know about what you saw?* Biology, 1998

to one object than to spatially equivalent properties spread over two or more objects [17–19].

needed to show that changes in the gist were necessary and sufficient for efficient change detection. Given the

Where does this leave us in the s… picture? Evidence from visual se… gests that objects can be identified… per second [20]. It is possible that t… any sort of stable memory runs at th… 'rapid serial visual presentation' exp… example). In either case, a relativel… a scene would allow several object… passed to memory. Is that list the g… that a list of N objects would be su… scene, but a series of thought expe… gist is more than a list. Some… objects must be coded into the gist… being poured from a carton into a gl… picture of milk being poured from… next to a glass, even if all of the… Moreover, even if all the prope… between objects remain the same, s…

…is an available preattentively. And… of objects passed by attention thro… memory, and you might just have t… represented in this way, it will not t… a brain can remember thousands of… nition experiment. Another program…

Visual memory: **What do you know about what you saw?**

Jeremy M. Wolfe

Recent studies of visual perception are bringing us closer to an understanding of what we remember – and what we forget – when we recall a scene.

Address: Center for Ophthalmic Research, Brigham and Women's Hospital, 221 Longwood Avenue, Boston, Massachusetts 02115, USA.

Current Biology 1998, 8:R303–R304
http://biomednet.com/elecref/09609822008R0303

…d is your memory? One line of research starting …years ago shows that your memory for visually …l material is quite remarkably good [1,2]. In a …icture recognition study, subjects are shown a …of scenes — such as images cut from a glossy …gazine — each of which is presented for a second …n the test phase, subjects are shown a second set …, half of them from the first set and the other half …d for the first time. The task is to identify …of the second set as old or new. Subjects …very well on such a task, even when thousands of …re shown [3,4].

…the subjects of such a study remembering? …sense tells us that our memory for a picture is …sort of highly detailed neural photocopy. Indeed, …ls of the image are not well remembered [5]. …trying to distinguish new from old from among a …different pictures of the stacks in the university …flurry of recent research has shown how bad we …cognizing differences between similar scenes or between two versions of the same scene, a phe… that has been referred to as 'change blindness' …jects might be shown a picture of an airplane on…, and in the second view the engines would be …from the plane. The two images, one with and …without engines, alternate on the screen every …ds, with a blank screen presented in between to mask luminance transients. It can take a surprisingly long time to notice this change. Subjects also fail to notice changes made during an eye movement [8–10]. This is exploited in movies, where cuts between views render subjects insensitive to changes in clothing, props or even the identity of actors [11].

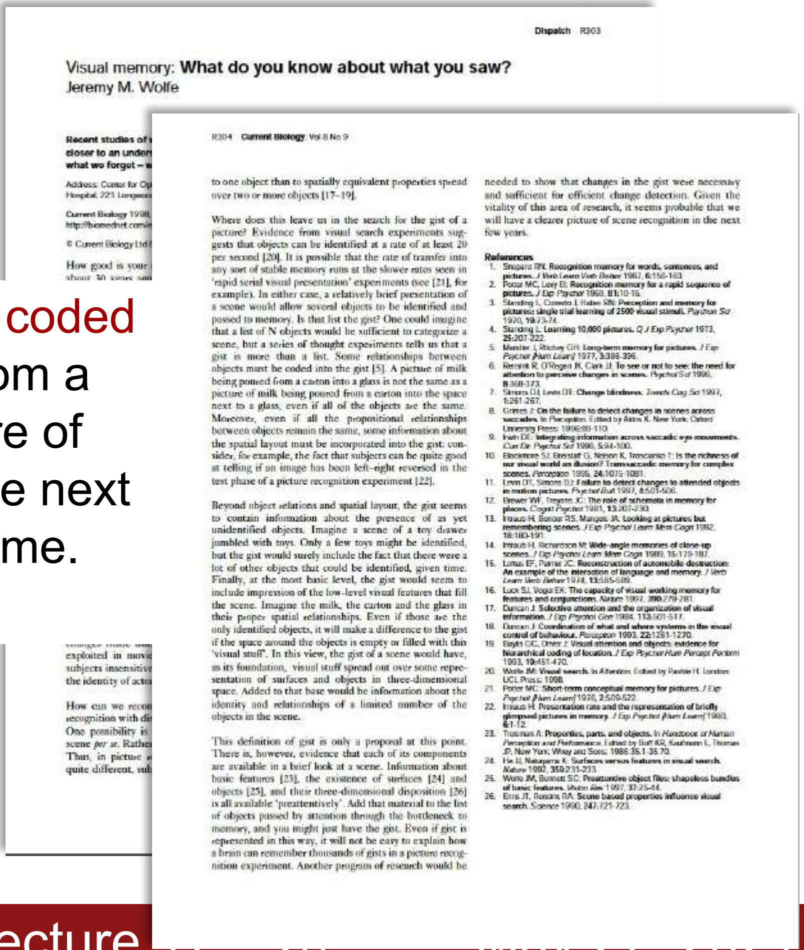
How can we reconcile excellent performance on picture recognition with dismal performance on change detection? One possibility is that observers do not remember the scene *per se*. Rather, they remember the gist of the scene. Thus, in picture recognition, where all the pictures are quite different, subjects can say to themselves, "Ah yes, I have seen a picture of a burning house; no, I didn't see a picture of a cat in the bathtub". By this account, change blindness occurs because the change does not alter the gist. A conversation between two women remains a conversation between two women, even if the clothing or the props change. In support of this idea, there is strong evidence that the meaning of a scene can influence memory for that scene. For instance, Brewer and Treyans [12] had subjects wait in an office, and then questioned them about the contents of the office. Subjects routinely reported books in the office, not because books were present — they were not — but because books are part of the schema for what should be in an office. People routinely remember seeing more of a scene than was presented [13,14]. On a more sinister note, memory for scenes can be colored by the biases of the observer [15].

The difficulty with the appealing idea that we remember the gist of a scene is that there is no consensus about the contents of a 'gist'. Intuition suggests that an inventory of some of the objects in the scene should be at least a part of the gist. If you asked someone to describe a scene, you would be surprised if the description named no objects but relied only on a description of features, such as color or size. A recent experiment by Luck and Vogel [16] seems to show this coding into memory for objects, rather than simple features. They performed a variation of a change-detection experiment. Two arrays of items were presented to subjects; on half the trials, the second array contained one item that was changed. If one-to-three colored squares were presented, subjects could perfectly detect the change; performance fell off with larger set sizes. These results suggest that subjects can keep track of four colors. Now, suppose that each item on the screen could vary in color, orientation, size and the presence or absence of a gap. Would subjects be able to keep track of just four individual features, or would they be able to keep track of up to four objects with all of their associated features? The answer, in a variety of versions of this experiment, is that subjects kept track of objects. They could detect any single feature change in any of up to four objects, even though that meant keeping track of more than a dozen individual features.

There is a bottleneck between vision and memory. If you close your eyes, you will immediately lose access to many of the details that were obvious a moment ago. The results of Luck and Vogel [16] show that it is objects and not raw features, that move through that bottleneck. The selection of objects is governed by attention. There is copious evidence that it is easier to attend to properties belonging

We encode more than objects

Some relationships between objects must be coded into the gist. A picture of milk being poured from a carton into a glass is not the same as a picture of milk being poured from a carton into the space next to a glass, even if all of the objects are the same.


Dispatch R303

Visual memory: **What do you know about what you saw?**
Jeremy M. Wolfe

R304 **Current Biology**, Vol 8 No 9

to one object than to spatially equivalent properties spread over two or more objects [17–19].

Where does this leave us in the search for the gist of a picture? Evidence from visual search experiments suggests that objects can be identified at a rate of at least 20 per second [20]. It is possible that the rate of transfer into any sort of stable memory runs at the slower rates seen in 'rapid serial visual presentation' experiments (see [21], for example). In either case, a relatively brief presentation of a scene would allow several objects to be identified and passed to memory. Is that list the gist? One could imagine that a list of N objects would be sufficient to categorize a scene, but a series of thought experiments tells us that a gist is more than a list. Some relationships between objects must be coded into the gist [5]. A picture of milk being poured from a carton into a glass is not the same as a picture of milk being poured from a carton into the space next to a glass, even if all of the objects are the same. Moreover, even if all the propositional relationships between objects remain the same, some information about the spatial layout must be incorporated into the gist: consider, for example, the fact that subjects can be quite good at telling if an image has been left–right reversed in the test phase of a picture recognition experiment [22].

Beyond object relations and spatial layout, the gist seems to contain information about the presence of as yet unidentified objects. Imagine a scene of a toy drawer jumbled with toys. Only a few toys might be identified, but the gist would surely include the fact that there were a lot of other objects that could be identified, given time. Finally, at the most basic level, the gist would seem to include impression of the low-level visual features that fill the scene. Imagine the milk, the carton and the glass in their proper spatial relationships. Even if those are the only identified objects, it will make a difference to the gist if the space around the objects is empty or filled with this 'visual stuff'. In this view, the gist of a scene would have, as its foundation, visual stuff spread out over some representation of surfaces and objects in three-dimensional space. Added to that base would be information about the identity and relationships of a limited number of the objects in the scene.

This definition of gist is only a proposal at this point. There is, however, evidence that each of its components are available in a brief look at a scene. Information about basic features [23], the existence of surfaces [24] and objects [25], and their three-dimensional disposition [26] is all available 'preattentively'. Add that material to the list of objects passed by attention through the bottleneck to memory, and you might just have the gist. Even if gist is represented in this way, it will not be easy to explain how a brain can remember thousands of gists in a picture recognition experiment. Another program of research would be needed to show that changes in the gist were necessary and sufficient for efficient change detection. Given the vitality of this area of research, it seems probable that we will have a clearer picture of scene recognition in the next few years.

References

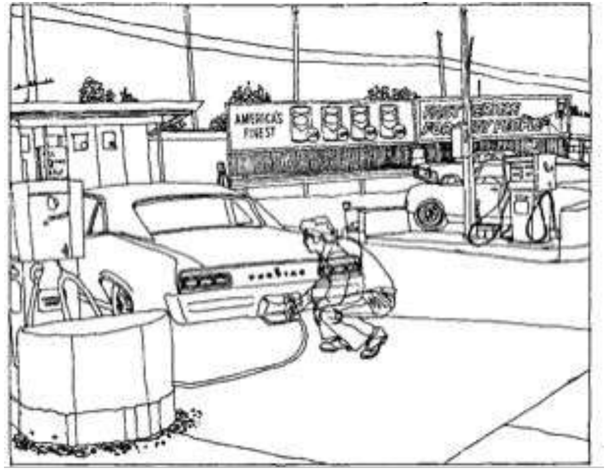
1. Shepard RN: Recognition memory for words, sentences, and pictures. J Verb Learn Verb Behav 1967, 6:156-163.
2. Potter MC, Levy EI: Recognition memory for a rapid sequence of pictures. J Exp Psychol 1969, 81:10-15.
3. Standing L, Conezio J, Haber RN: Perception and memory for pictures: single trial learning of 2500 visual stimuli. Psychon Sci 1970, 19:73-74.
4. Standing L: Learning 10,000 pictures. Q J Exp Psychol 1973, 25:207-222.
5. Mandler J, Ritchey GH: Long-term memory for pictures. J Exp Psychol [Hum Learn] 1977, 3:386-396.
6. Rensink R, O'Regan JK, Clark JJ: To see or not to see: the need for attention to perceive changes in scenes. Psychol Sci 1996, 8:368-373.
7. Simons DJ, Levin DT: Change blindness. Trends Cog Sci 1997, 1:261-267.
8. Grimes J: On the failure to detect changes in scenes across saccades. In Perception. Edited by Akins K. New York: Oxford University Press; 1996:89-110.
9. Irwin DE: Integrating information across saccadic eye movements. Curr Dir Psychol Sci 1996, 5:94-100.
10. Blackmore SJ, Brelstaff G, Nelson K, Troscianko T: Is the richness of our visual world an illusion? Transsaccadic memory for complex scenes. Perception 1995, 24:1075-1081.
11. Levin DT, Simons DJ: Failure to detect changes to attended objects in motion pictures. Psychol Bull 1997, 4:501-506.
12. Brewer WF, Treyens JC: The role of schemata in memory for places. Cognit Psychol 1981, 13:207-230.
13. Intraub H, Bender RS, Mangels JA: Looking at pictures but remembering scenes. J Exp Psychol Learn Mem Cogn 1992, 18:180-191.
14. Intraub H, Richardson M: Wide-angle memories of close-up scenes. J Exp Psychol Learn Mem Cogn 1989, 15:179-187.
15. Loftus EF, Palmer JC: Reconstruction of automobile destruction: An example of the interaction of language and memory. J Verb Learn Verb Behav 1974, 13:585-589.
16. Luck SJ, Vogel EK: The capacity of visual working memory for features and conjunctions. Nature 1997, 390:279-281.
17. Duncan J: Selective attention and the organization of visual information. J Exp Psychol Gen 1984, 113:501-517.
18. Duncan J: Coordination of what and where systems in the visual control of behaviour. Perception 1993, 22:1261-1270.
19. Baylis GC, Driver J: Visual attention and objects: evidence for hierarchical coding of location. J Exp Psychol Hum Percept Perform 1993, 19:451-470.
20. Wolfe JM: Visual search. In Attention. Edited by Pashler H. London: UCL Press; 1998.
21. Potter MC: Short-term conceptual memory for pictures. J Exp Psychol [Hum Learn] 1976, 2:509-522.
22. Intraub H: Presentation rate and the representation of briefly glimpsed pictures in memory. J Exp Psychol [Hum Learn] 1980, 6:1-12.
23. Treisman A: Properties, parts, and objects. In Handbook of Human Perception and Performance. Edited by Boff KR, Kaufmann L, Thomas JP. New York: Wiley and Sons; 1986:35.1-35.70.
24. He ZJ, Nakayama K: Surfaces versus features in visual search. Nature 1992, 359:231-233.
25. Wolfe JM, Bennett SC: Preattentive object files: shapeless bundles of basic features. Vision Res 1997, 37:25-44.
26. Enns JT, Rensink RA: Scene based properties influence visual search. Science 1990, 247:721-723.

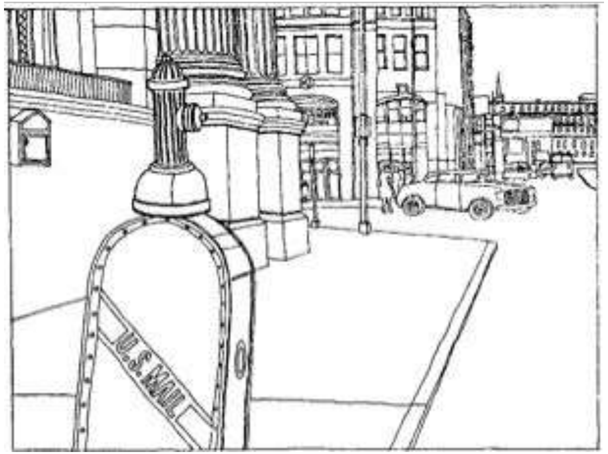
Wolfe, *Visual Memory: What do you know about what you saw?* Biology, 1998

Attributes and Relationships are processed independent of Objects

Attribute and relationship violations are noticed within 150ms.

Relationship violations slow down object identification.





Biederman, *Visual Memory: What do you know about what you saw?* Cognitive Psychology, 1982

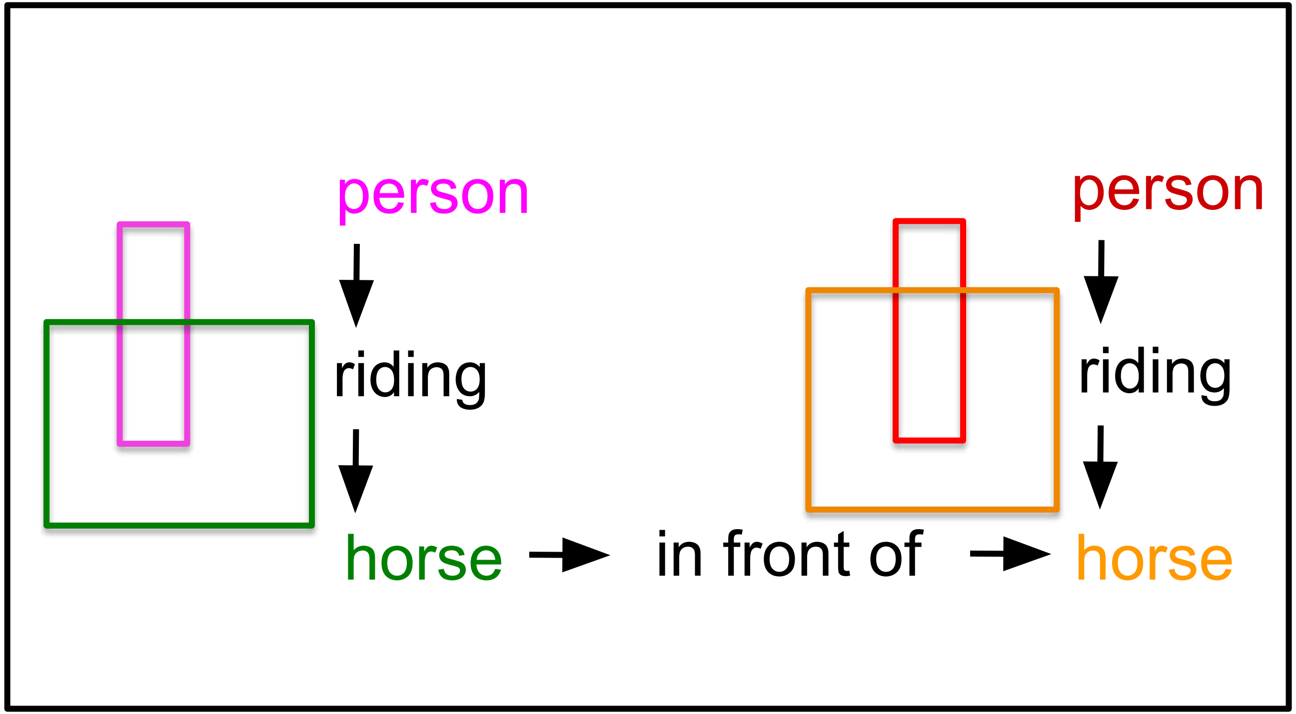
Scene Graph Generation - Problem formulation

Input
(image only)

Lu, Krishna et al., Visual Relationship Detection with Language Priors, ECCV 2016

Scene Graph Generation - Problem formulation

Input
(image only)

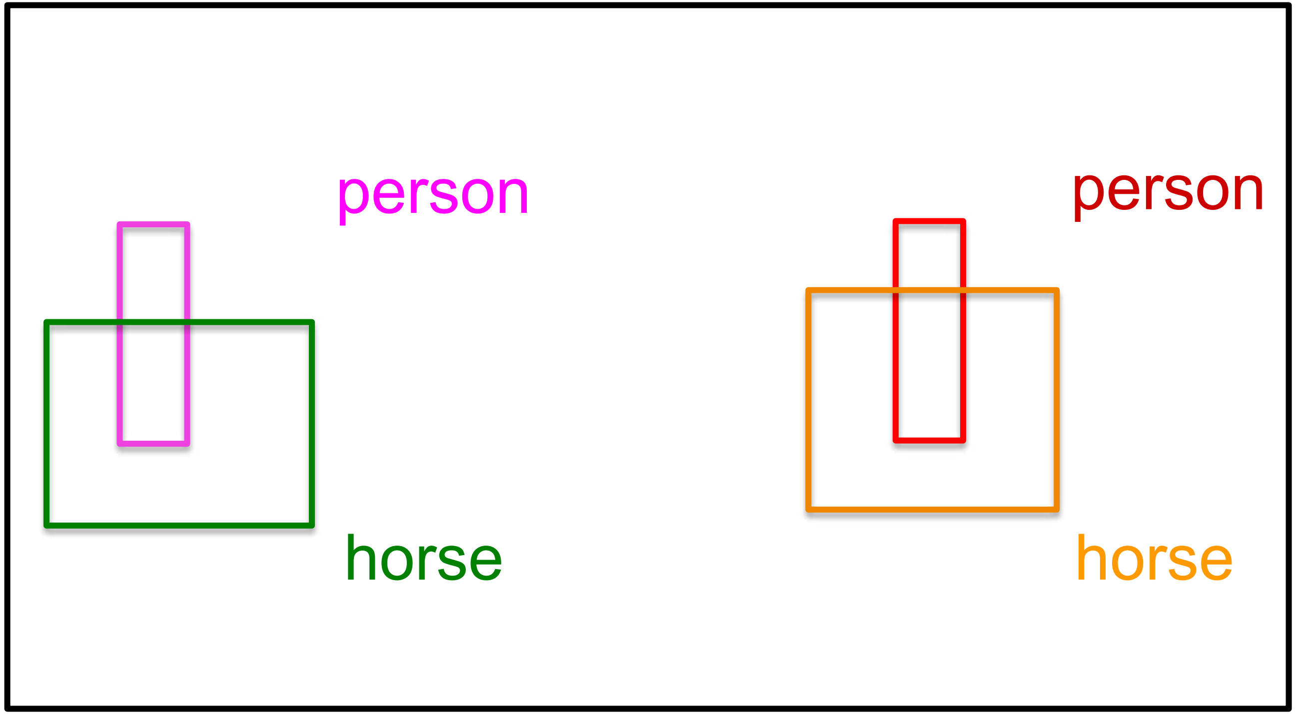


Output

Lu, Krishna et al., Visual Relationship Detection with Language Priors, ECCV 2016

Scene Graph Generation - Problem formulation

Input
(image only)

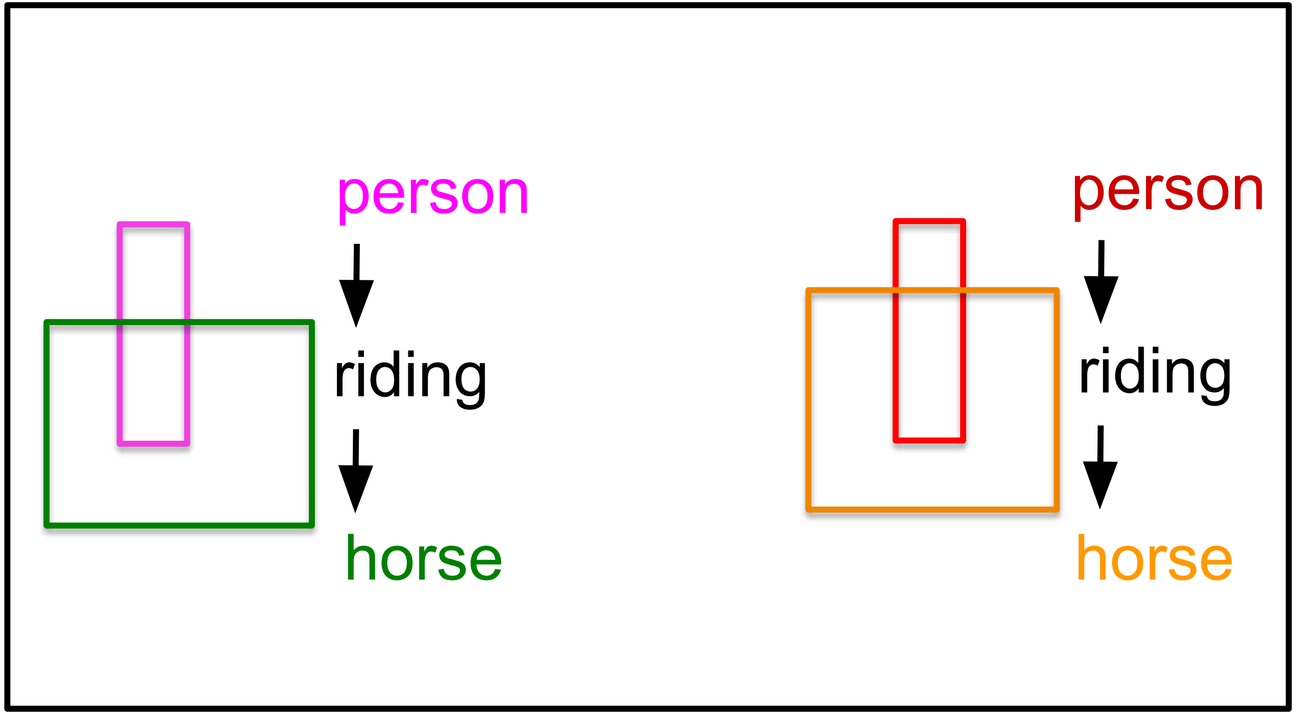


Output

Lu, Krishna et al., Visual Relationship Detection with Language Priors, ECCV 2016

Scene Graph Generation - Problem formulation

Input
(image only)

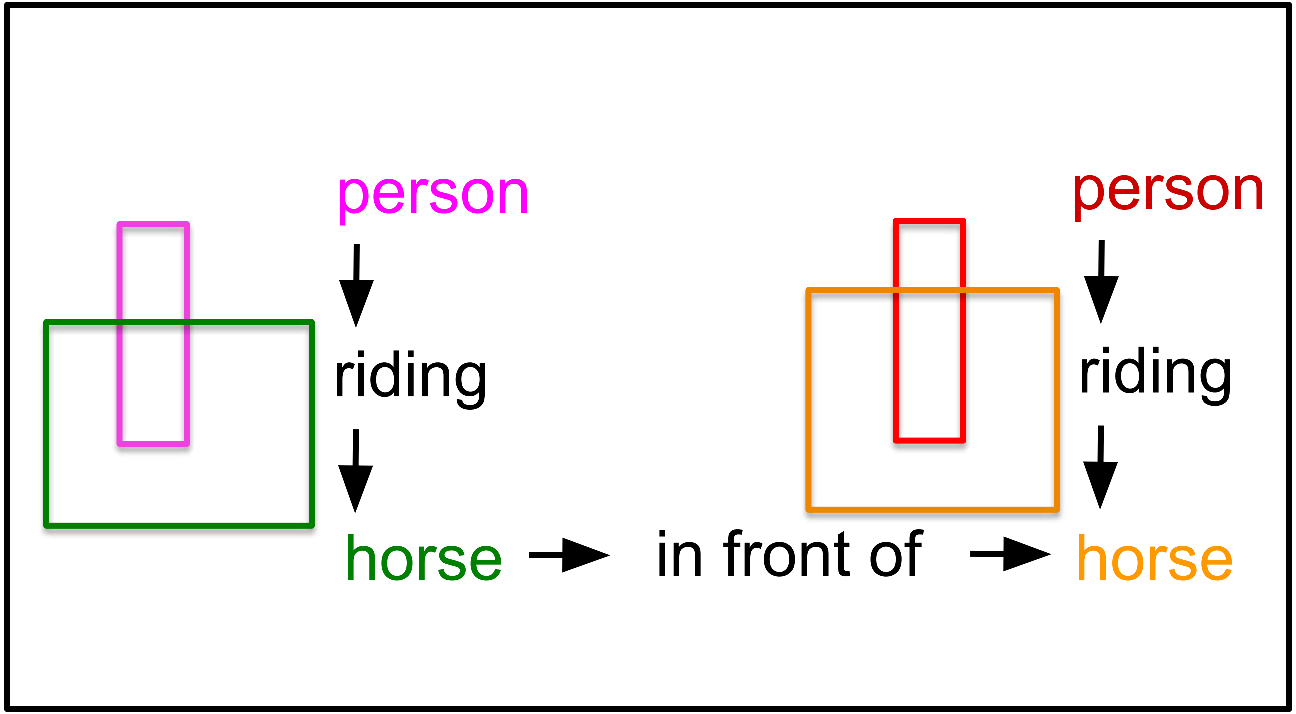


Output

Lu, Krishna et al., Visual Relationship Detection with Language Priors, ECCV 2016

Scene Graph Generation - Problem formulation

Input
(image only)



Output

Challenge 1:

Quadratic explosion of
- N objects,
- K relationships

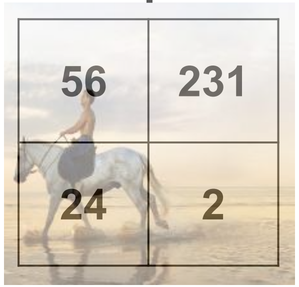
leading to N^2K detectors

Visual Genome dataset
N = 100
K = 50

Lu, Krishna et al., Visual Relationship Detection with Language Priors, ECCV 2016

Recall the algebraic interpretation of linear models:

Features from the last layer



Input image

$$
\begin{bmatrix} 0.2 & -0.5 & 0.1 & 2.0 \\ 1.5 & 1.3 & 2.1 & 0.0 \\ 0 & 0.25 & 0.2 & -0.3 \\ \vdots & & & \end{bmatrix}
\begin{bmatrix} 56 \\ 231 \\ 24 \\ 2 \end{bmatrix}
+
\begin{bmatrix} 1.1 \\ 3.2 \\ -1.2 \\ \vdots \end{bmatrix}
=
\begin{bmatrix} -96.8 \\ 437.9 \\ 61.95 \\ \vdots \end{bmatrix}
\begin{matrix} \text{person - riding - horse} \\ \text{person - driving - car} \\ \text{dog - holding - frisbee} \\ \vdots \end{matrix}
$$

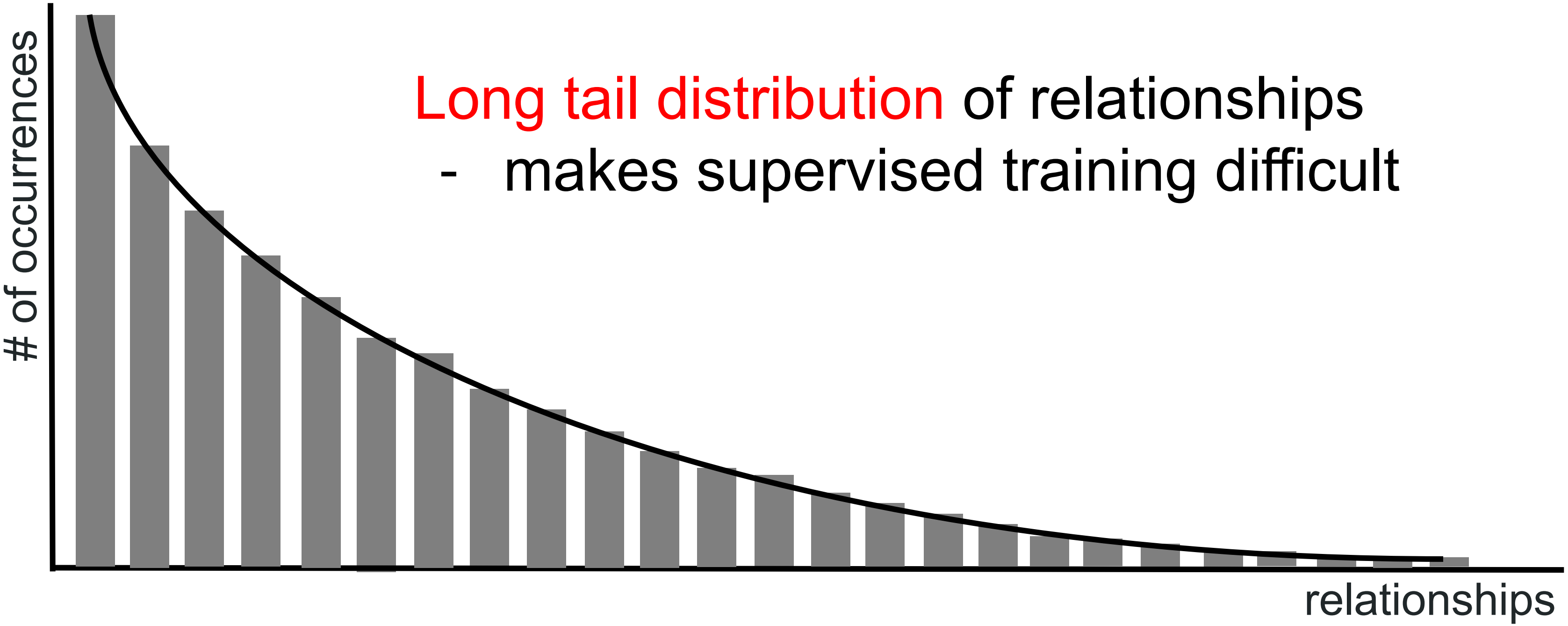
N^2K rows!!

W

b

Scores for each relationship

Challenge #2

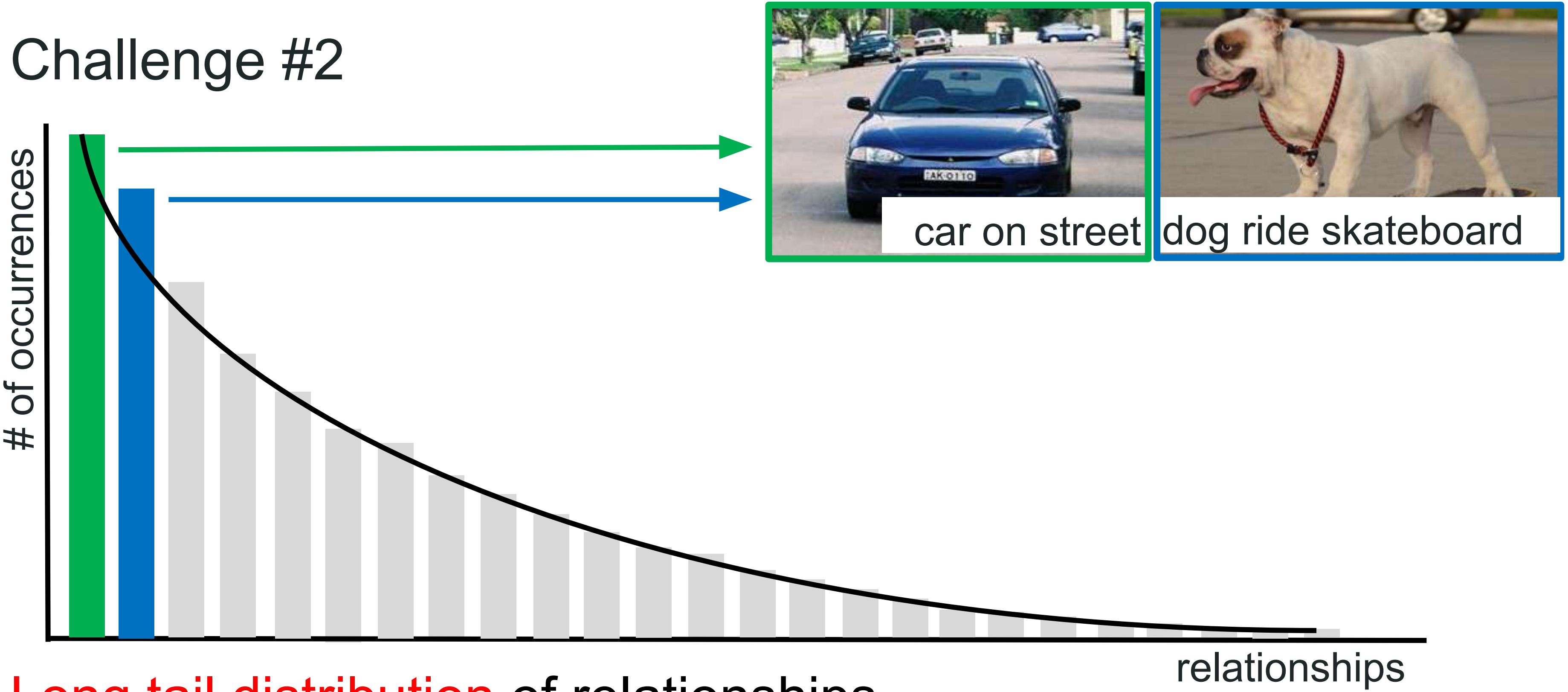


Long tail distribution of relationships

- makes supervised training difficult

Lu, Krishna et al., Visual Relationship Detection with Language Priors, ECCV 2016

Challenge #2



Long tail distribution of relationships

- makes supervised training difficult

Lu, Krishna et al., Visual Relationship Detection with Language Priors, ECCV 2016

Challenge #2

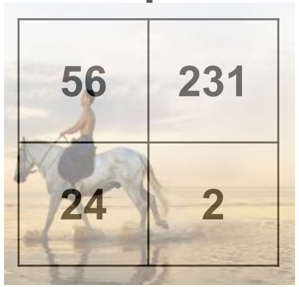
Long tail distribution of relationships

- makes supervised training difficult

Lu, Krishna et al., Visual Relationship Detection with Language Priors, ECCV 2016

Intuition: Compose visual relationships from objects and predicates

Features from the last layer



Input image

0.2	-0.5	0.1	2.0
1.5	1.3	2.1	0.0
0	0.25	0.2	-0.3

⋮

W

56
231
24
2

+

1.1
3.2
-1.2

⋮

b

=

-96.8	person
437.9	horse
61.95	cat

⋮

17.2	riding
10.9	holding
2.95	driving

N+K rows only

Visual module

Tackles:
Quadratic explosion of N^2K detectors

$\odot$

Language module

Tackles:
Long tail distribution of relationships

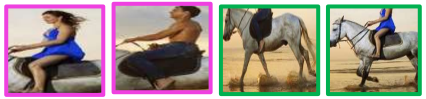
Visual module

Input

Definitions:

Visual module

Proposals:



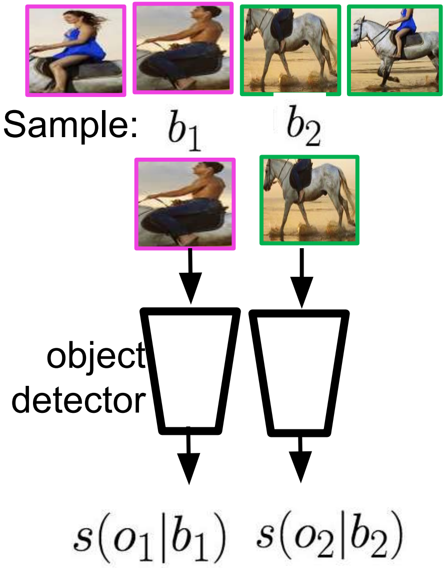
Input

Definitions:

b_1, b_2 are object proposals

Visual module

Proposals:



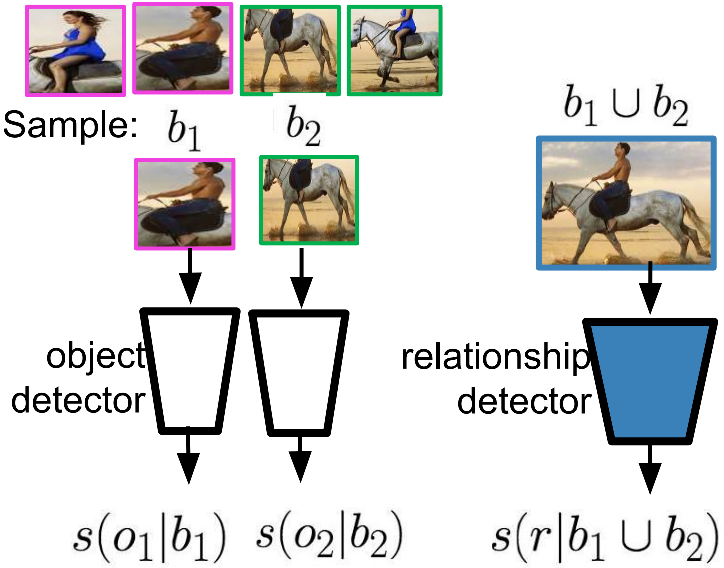
Definitions:

b_1, b_2 are object proposals

o_1, o_2 ∈ [person, horse, …]

Visual module

Proposals:

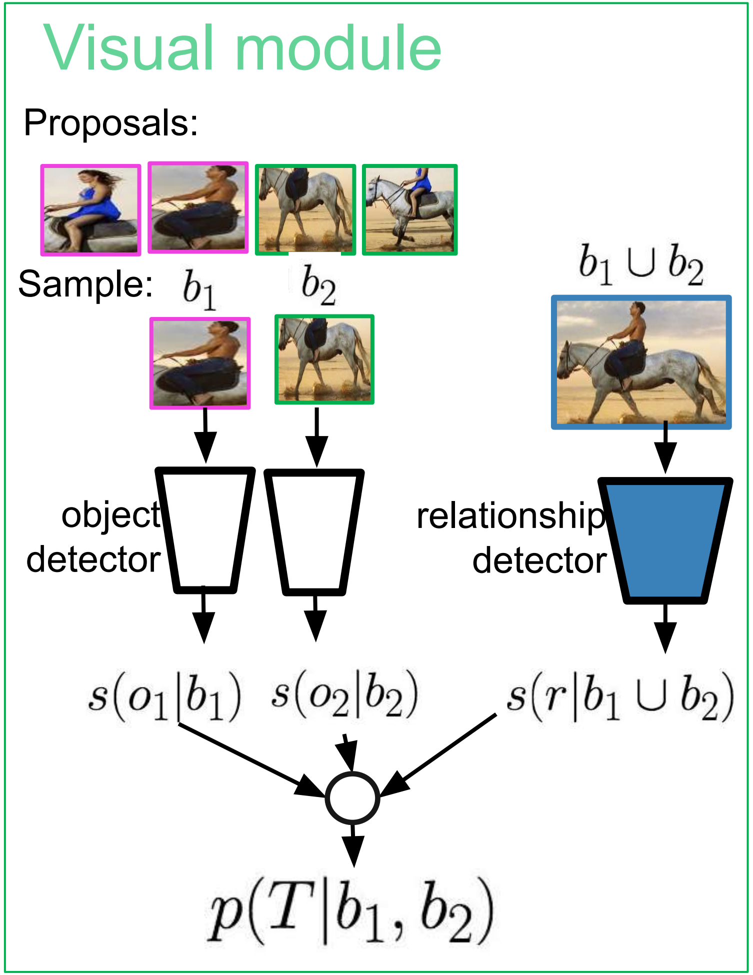


Definitions:

b_1, b_2 are object proposals

o_1, o_2 $\in$ [person, horse, …]

r $\in$ [on, in, ride, front of, …]



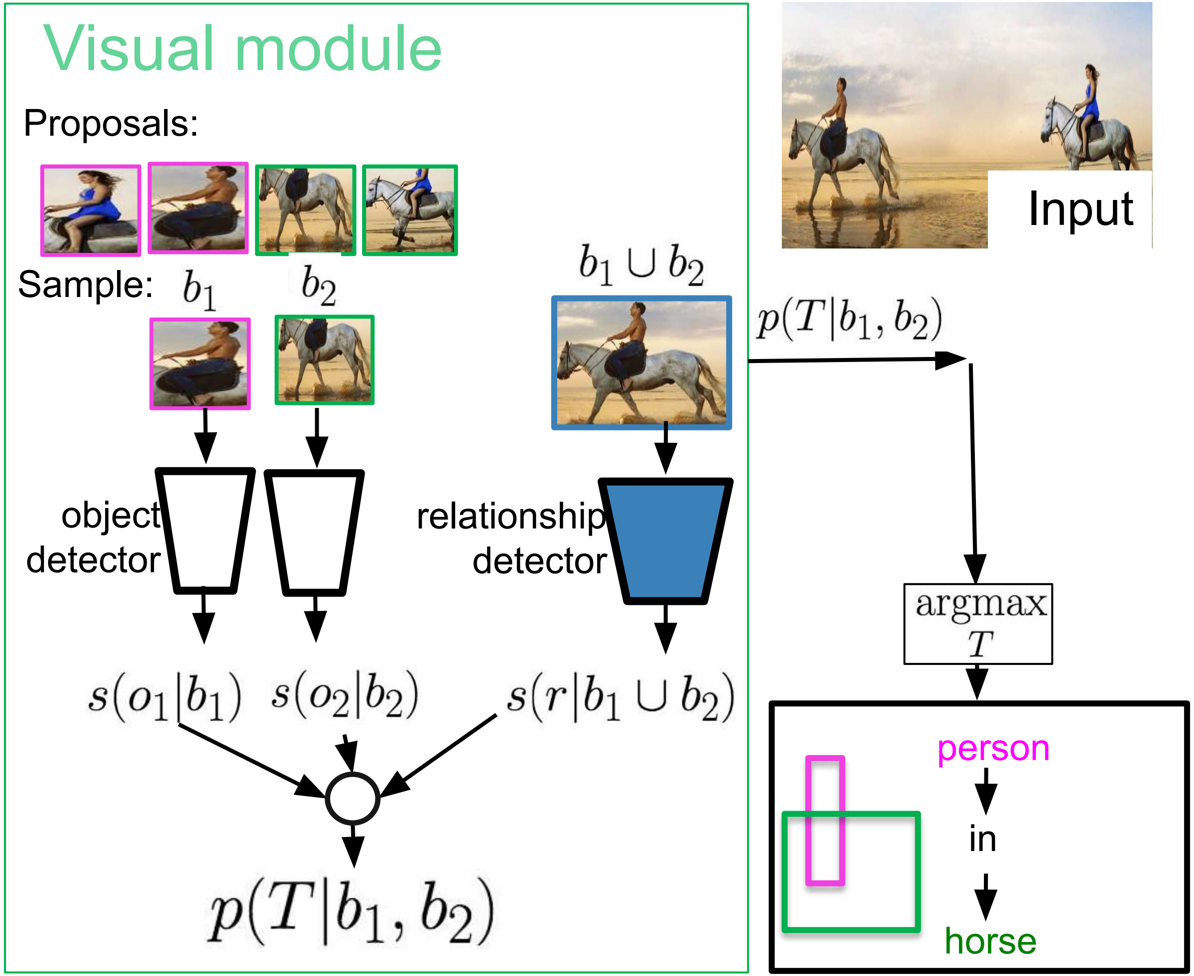
Definitions:

b_1, b_2 are object proposals

o_1, o_2 ∈ [person, horse, …]

r ∈ [on, in, ride, front of, …]

T is a <o_1, r, o_2> triple



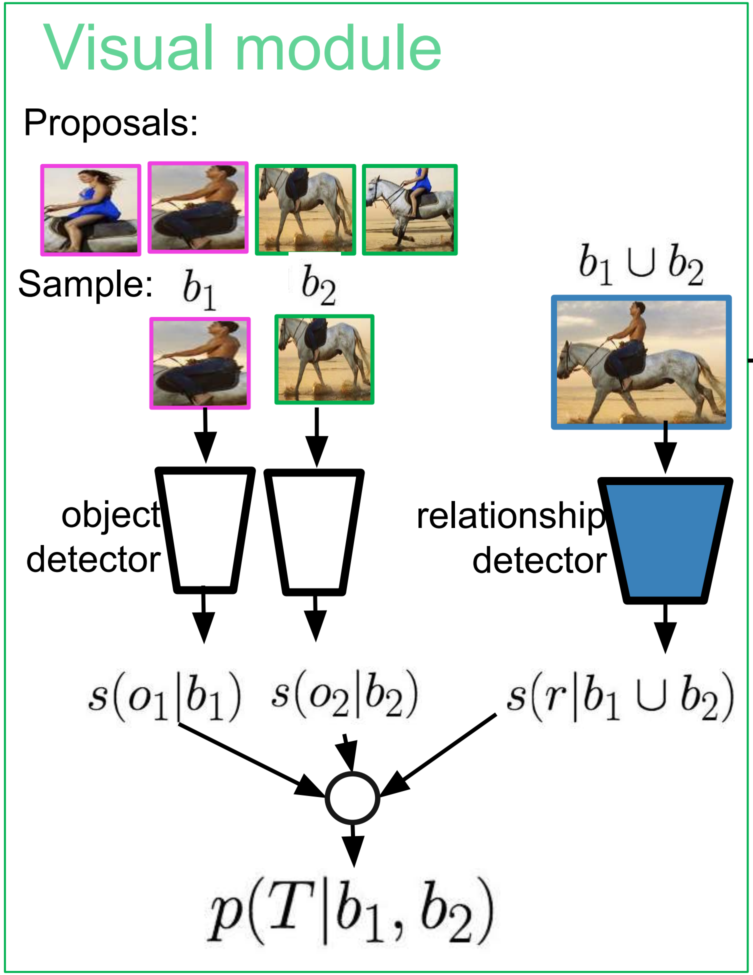
Definitions:

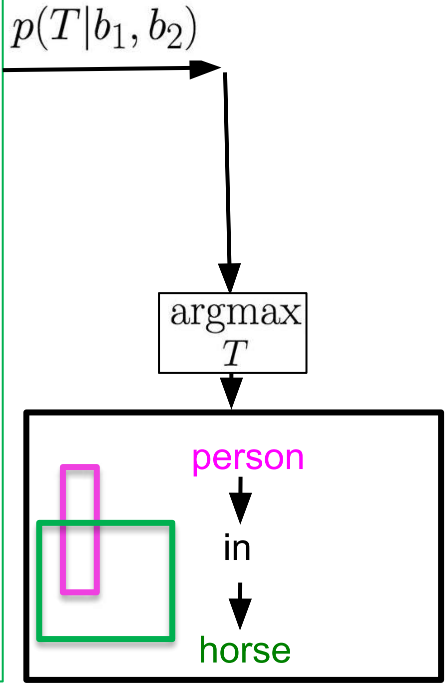
b_1, b_2 are object proposals

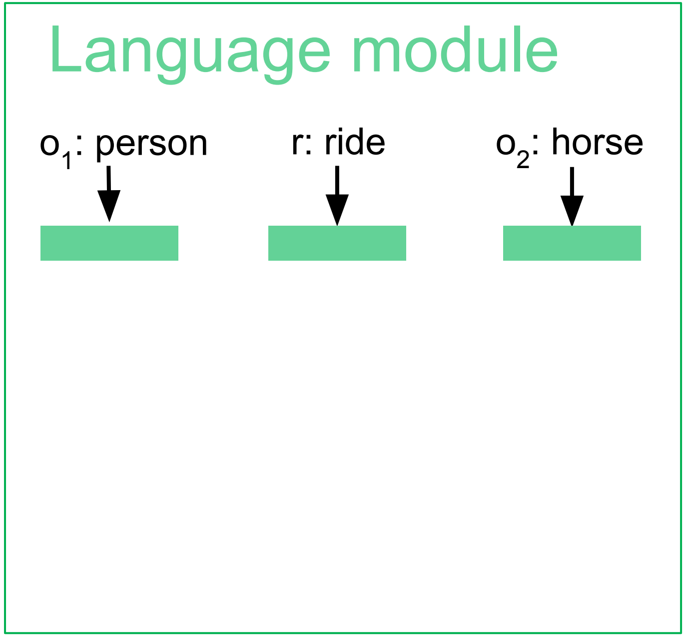
o_1, o_2 $\in$ [person, horse, …]

r $\in$ [on, in, ride, front of, …]

T is a <o_1, r, o_2> triple







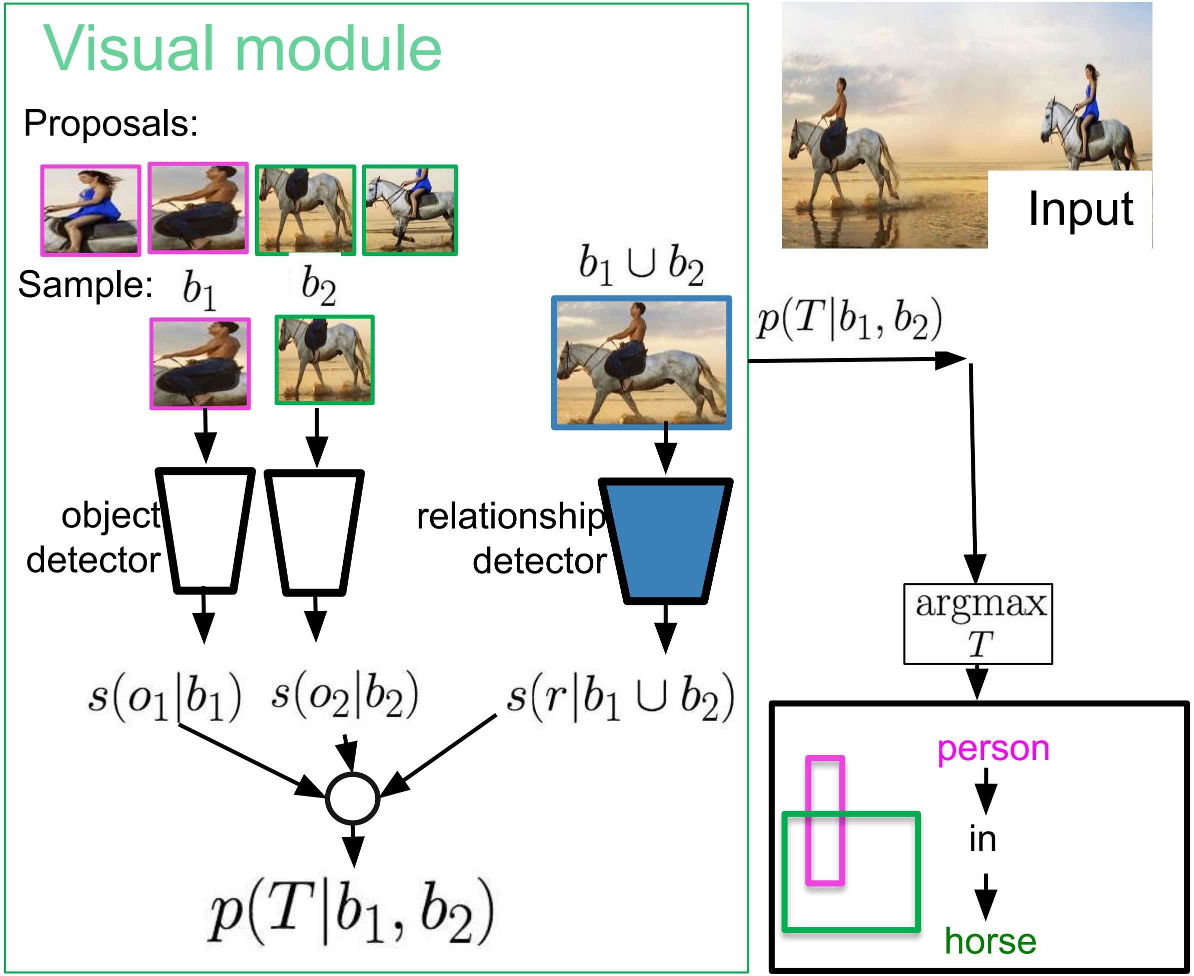
Definitions:

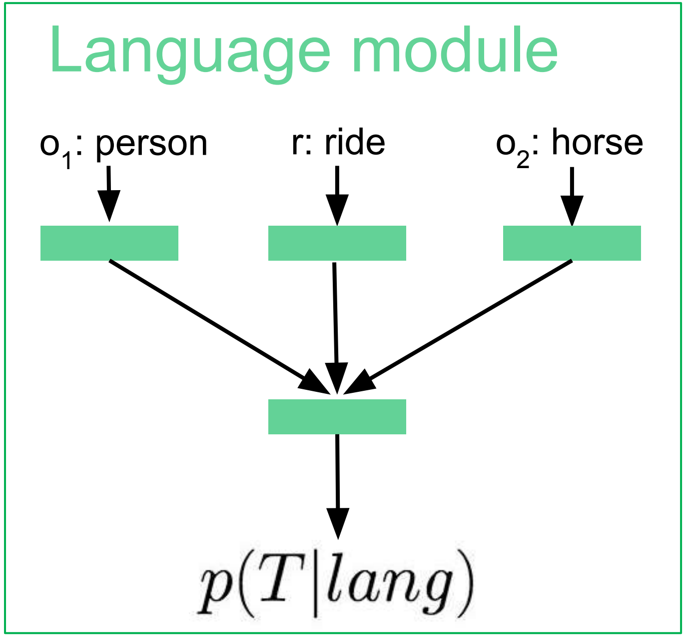
b_1, b_2 are object proposals

o_1, o_2 ∈ [person, horse, …]

r ∈ [on, in, ride, front of, …]

T is a <o_1, r, o_2> triple





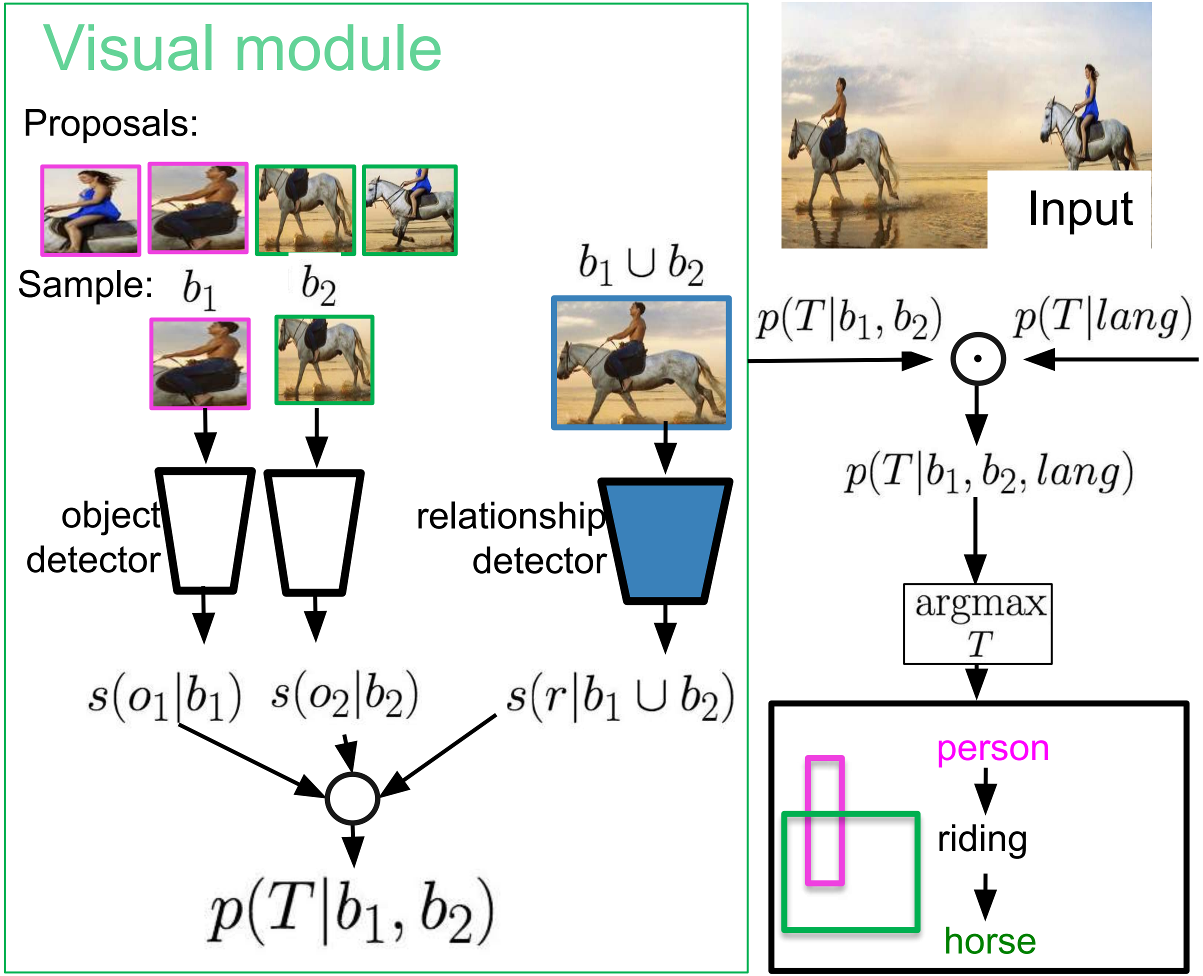
Definitions:

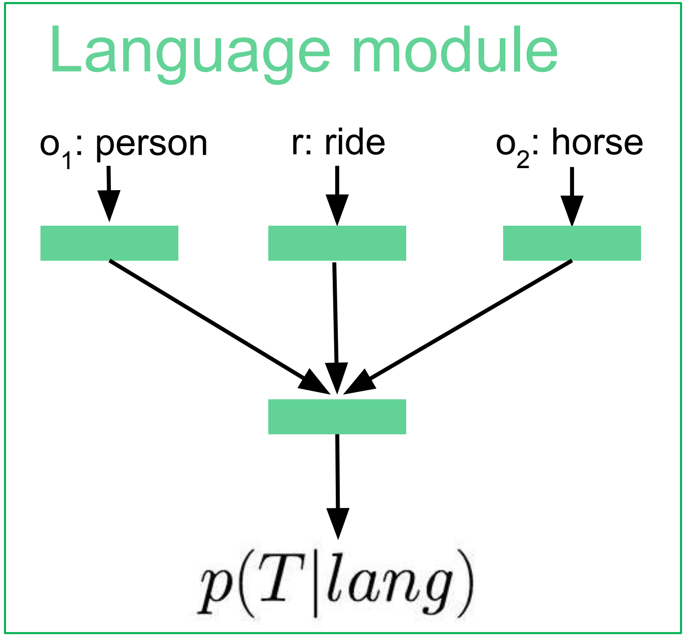
b_1, b_2 are object proposals

o_1, o_2 $\in$ [person, horse, …]

r $\in$ [on, in, ride, front of, …]

T is a <o_1, r, o_2> triple





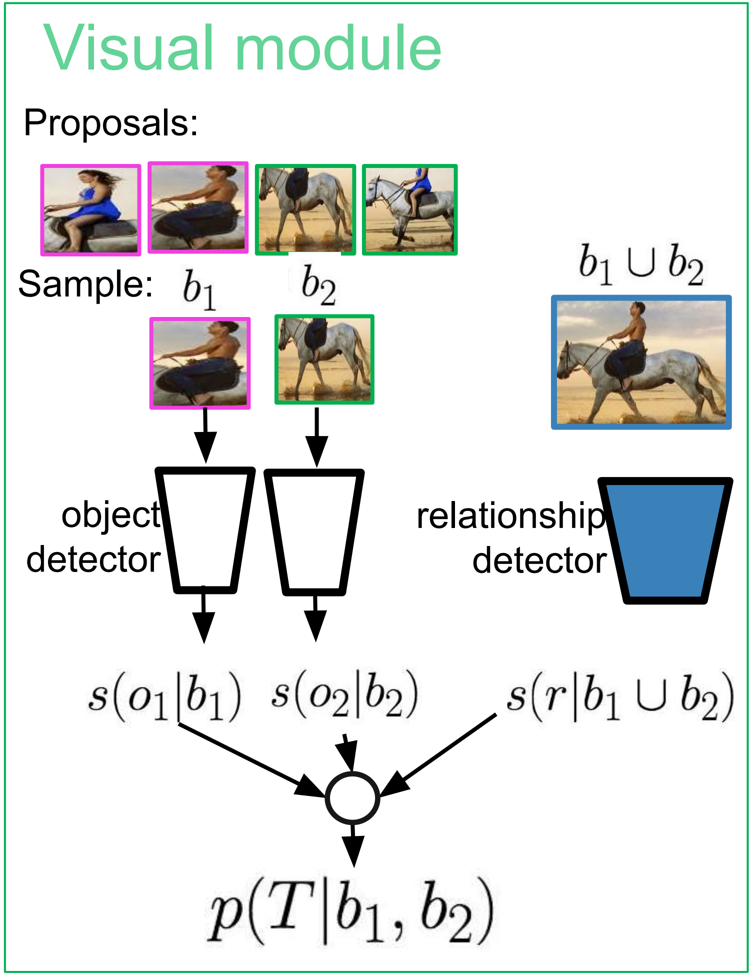
Definitions:

b_1, b_2 are object proposals

o_1, o_2 ∈ [person, horse, …]

r ∈ [on, in, ride, front of, …]

T is a <o_1, r, o_2> triple



Tackles:

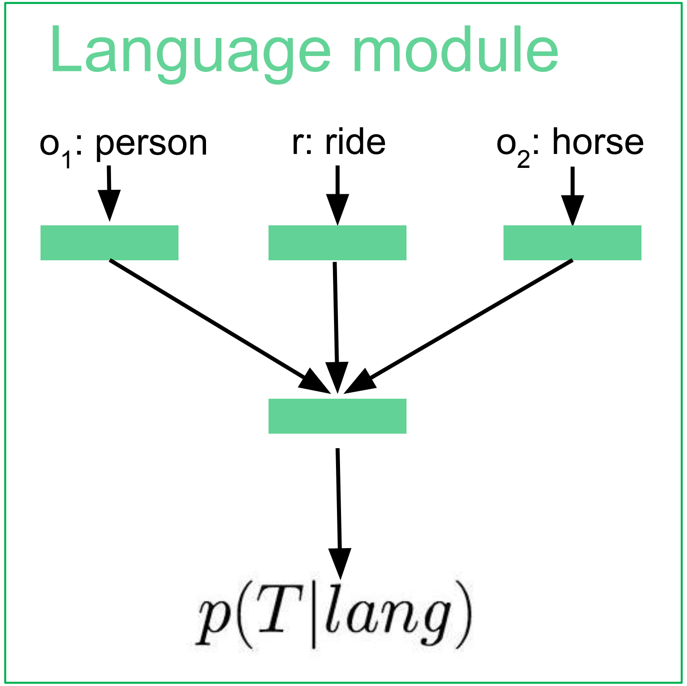
Quadratic explosion

only requires N+K detectors

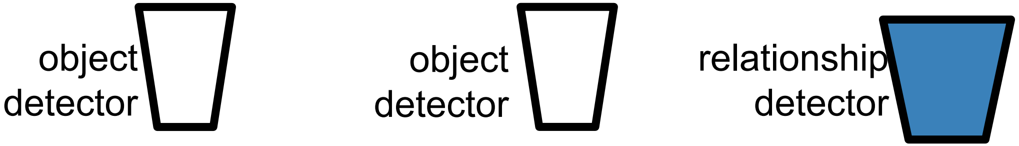
Tackles:

Long tail distribution

can predict rare relationships



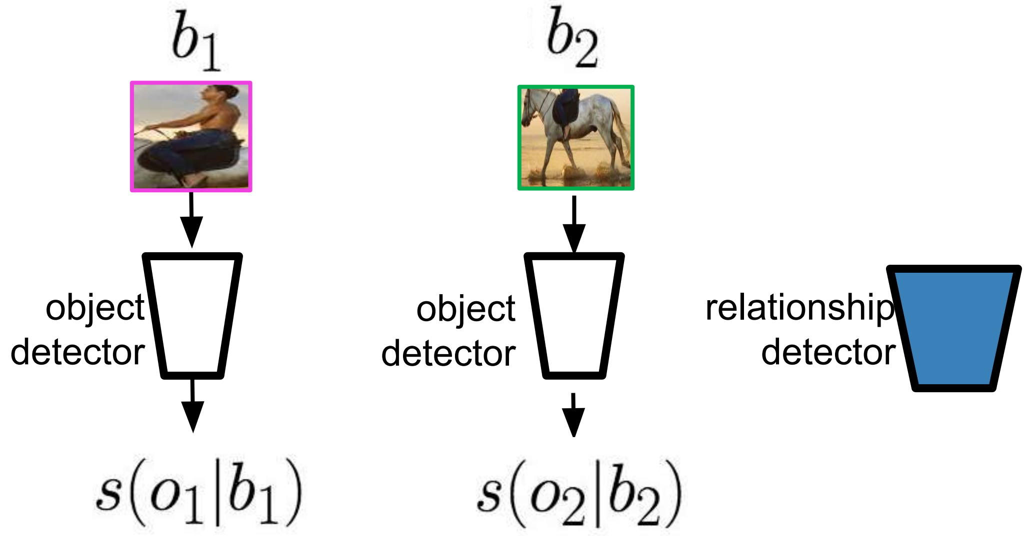
Training the visual module



1. Pre-train using ImageNet

Definitions:

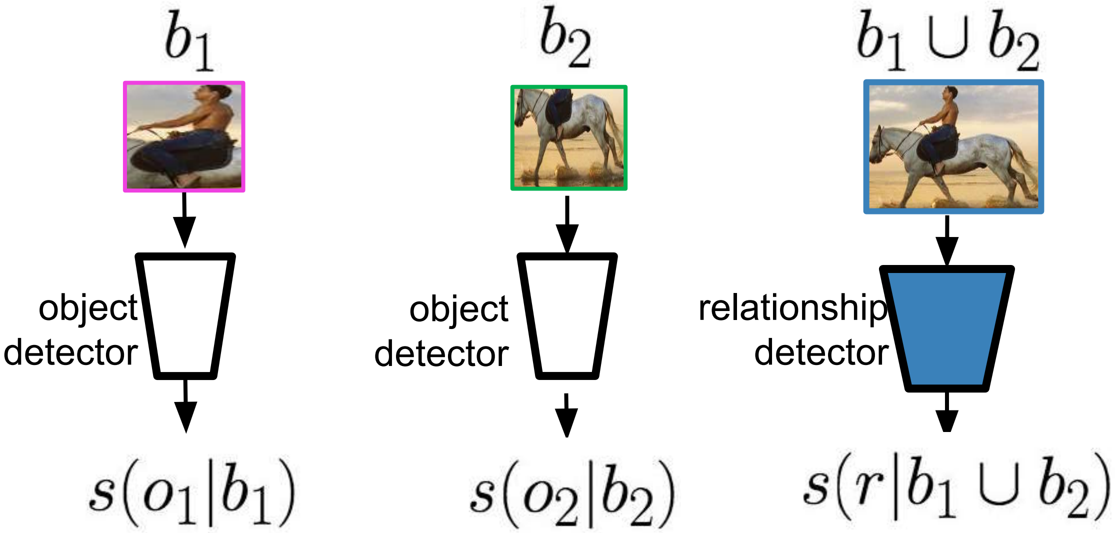
Training the visual module



1. Pre-train using ImageNet
2. Train object detector

Definitions:
b_1, b_2 are object proposals
o_1, o_2 $\in$ [person, horse, …]

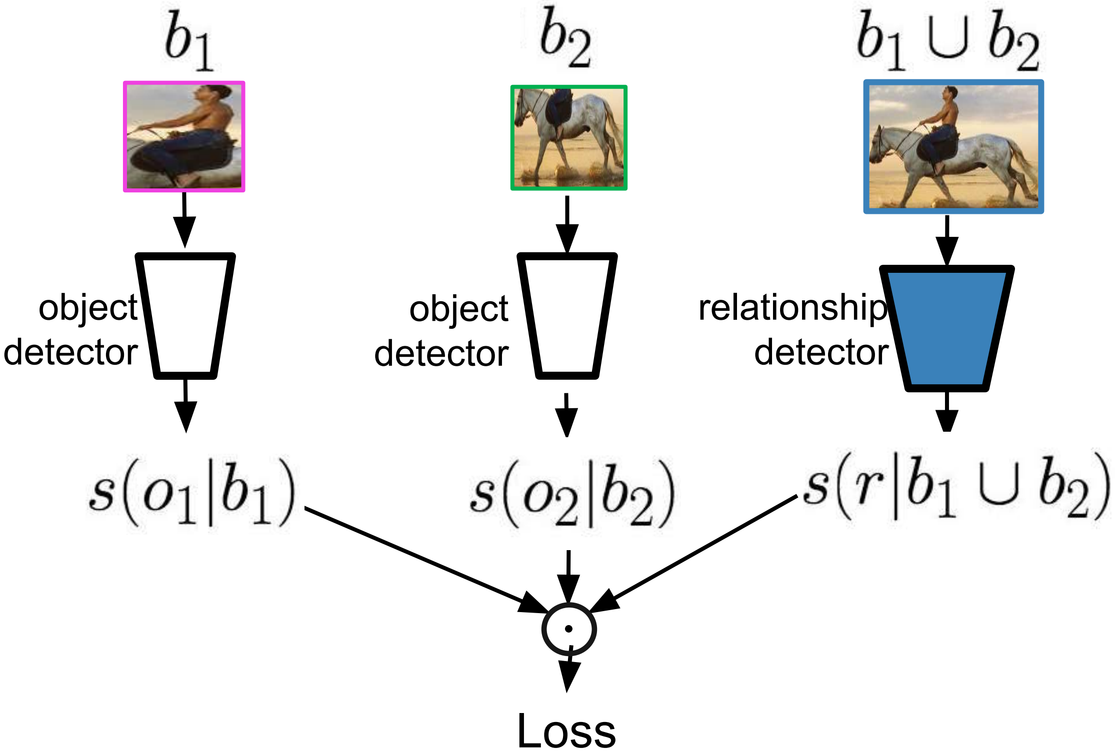
Training the visual module



1. Pre-train using ImageNet
2. Train object detector
3. Train relationship detector

Definitions:
b_1, b_2 are object proposals
o_1, o_2 ∈ [person, horse, …]
r ∈ [on, in, ride, front of, …]

Training the visual module



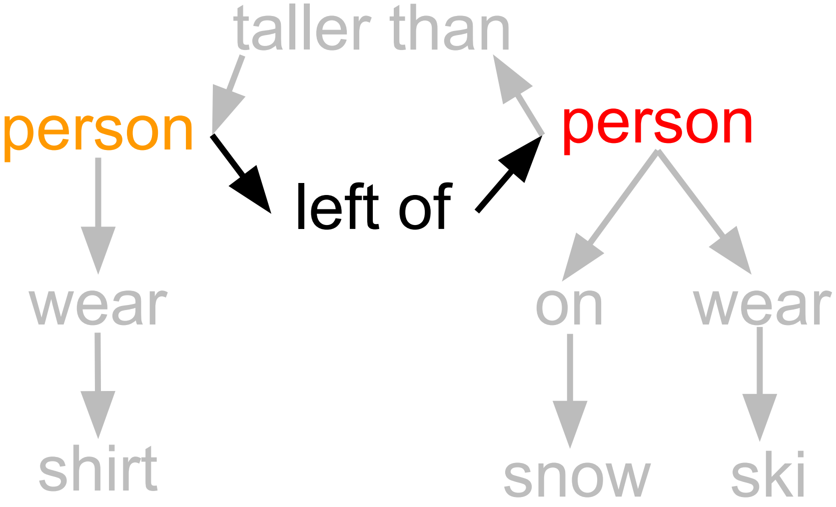
1. Pre-train using ImageNet
2. Train object detector
3. Train relationship detector
4. Fine-tune both jointly

Definitions:
b_1, b_2 are object proposals
o_1, o_2 ∈ [person, horse, …]
r ∈ [on, in, ride, front of, …]

Our results:

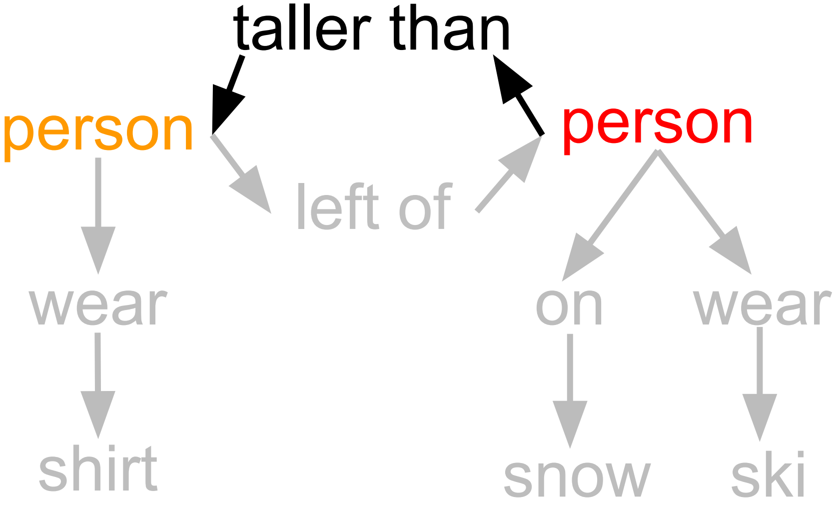
Our results:

spatial, comparative, asymmetrical, verb, prepositional



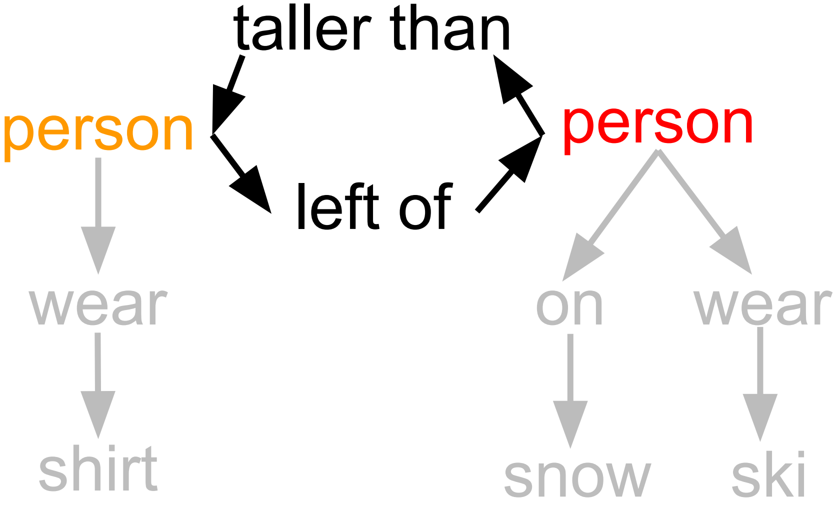
Our results:

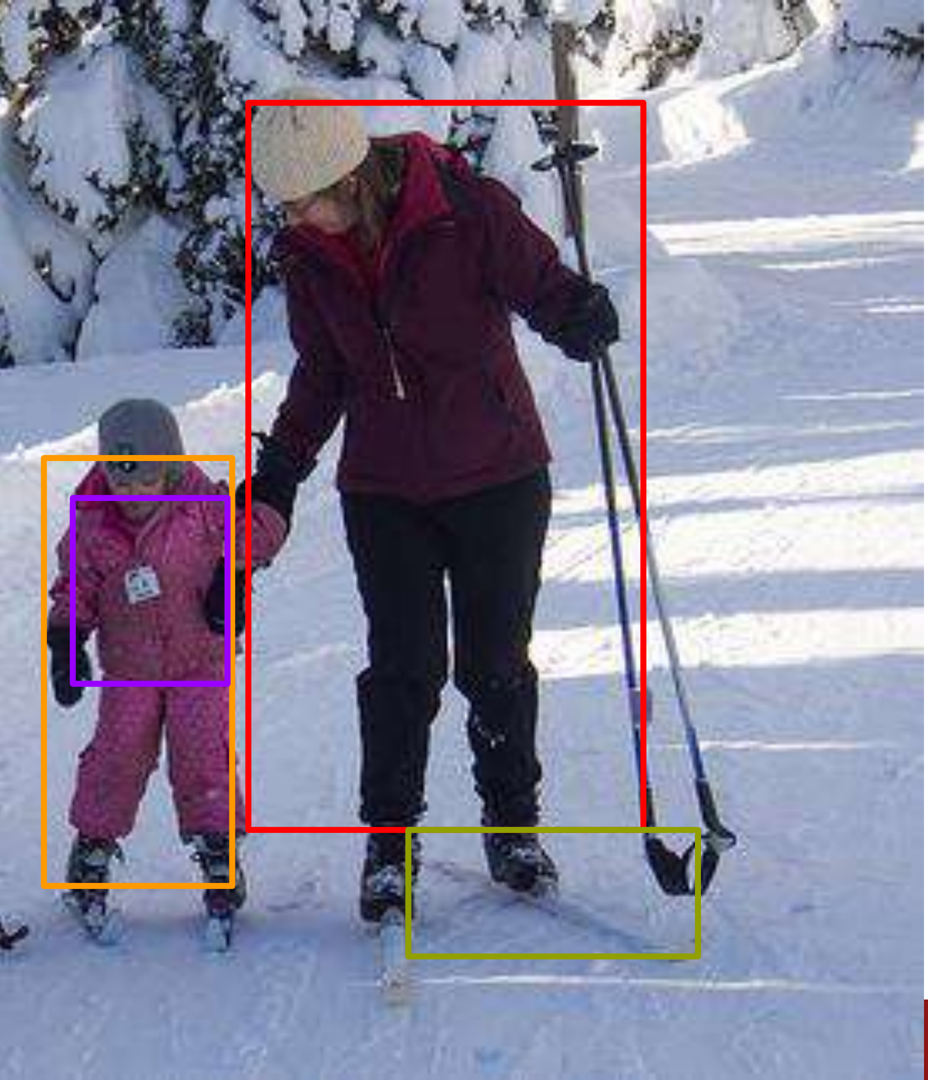
spatial, **comparative**, asymmetrical, verb, prepositional



Our results:

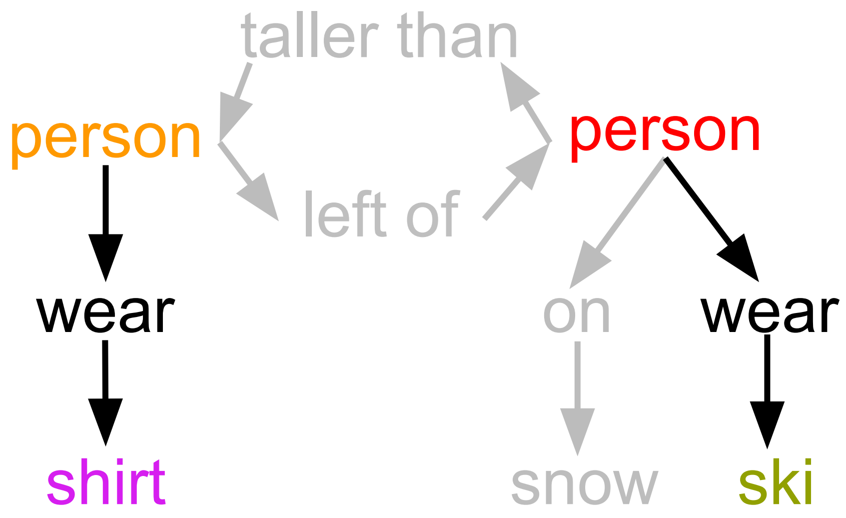
spatial, comparative, asymmetrical, verb, prepositional





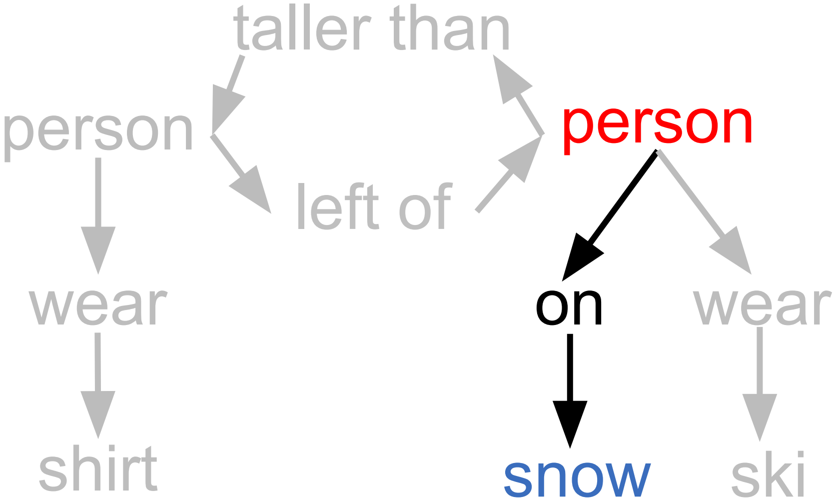
Our results:

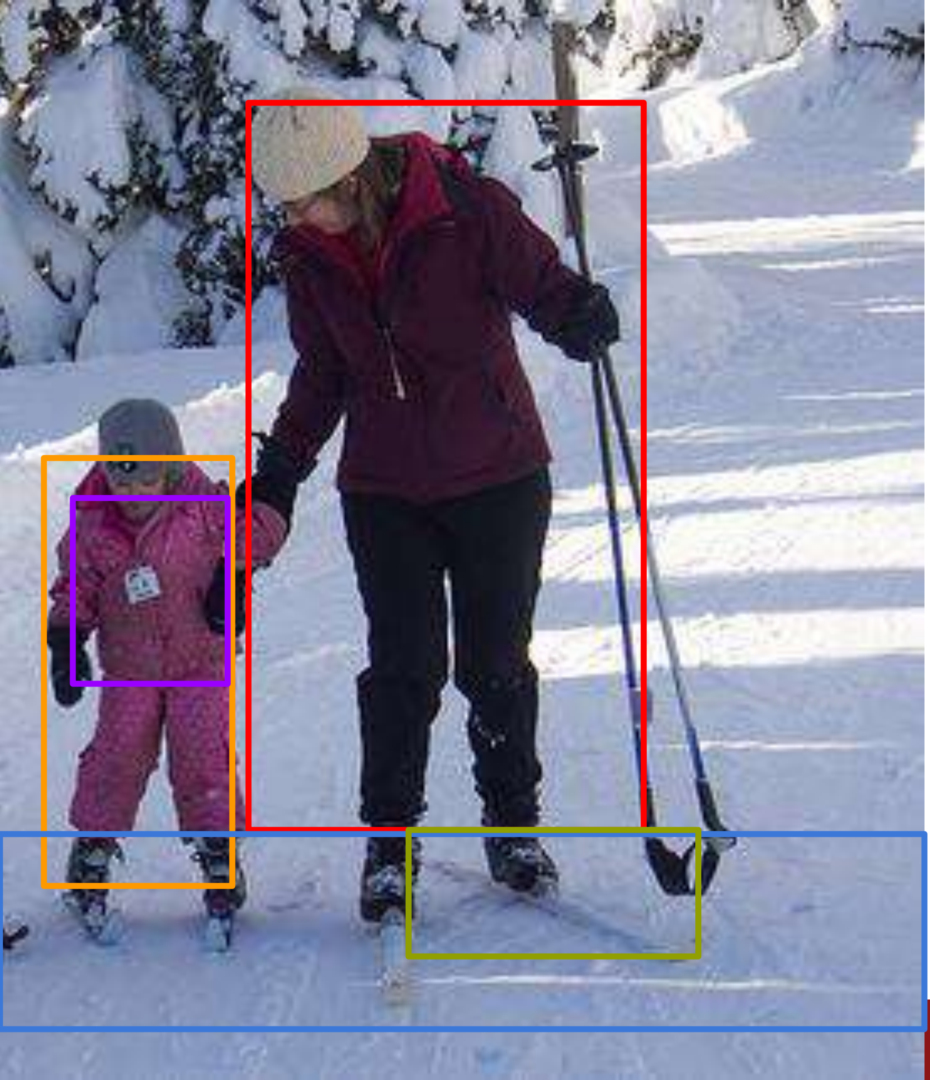
spatial, comparative, asymmetrical, verb, prepositional



Relationship types:

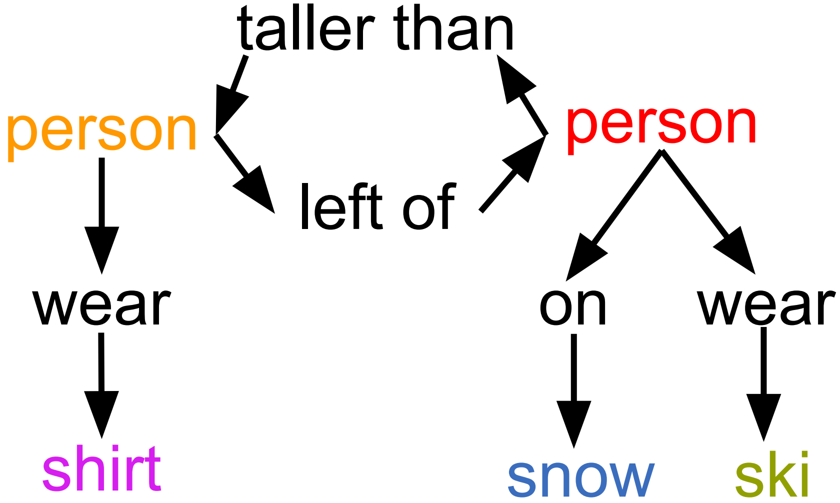
spatial, comparative, asymmetrical, verb, prepositional



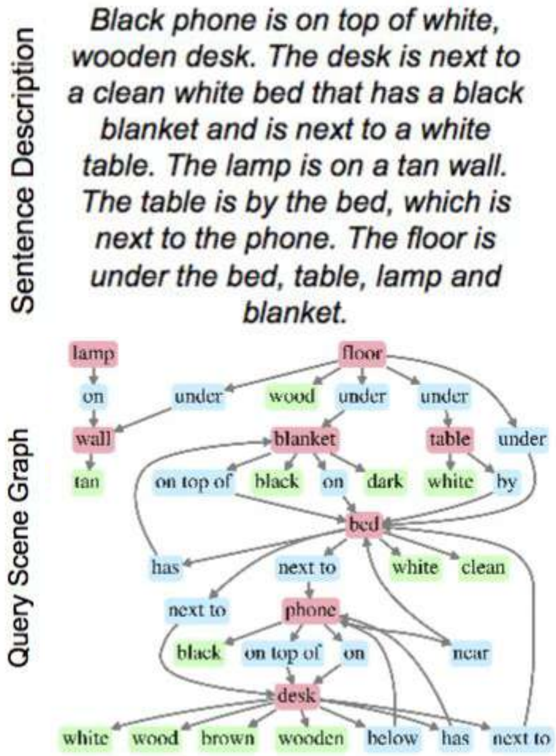


Our results:

spatial, comparative, asymmetrical, verb, prepositional



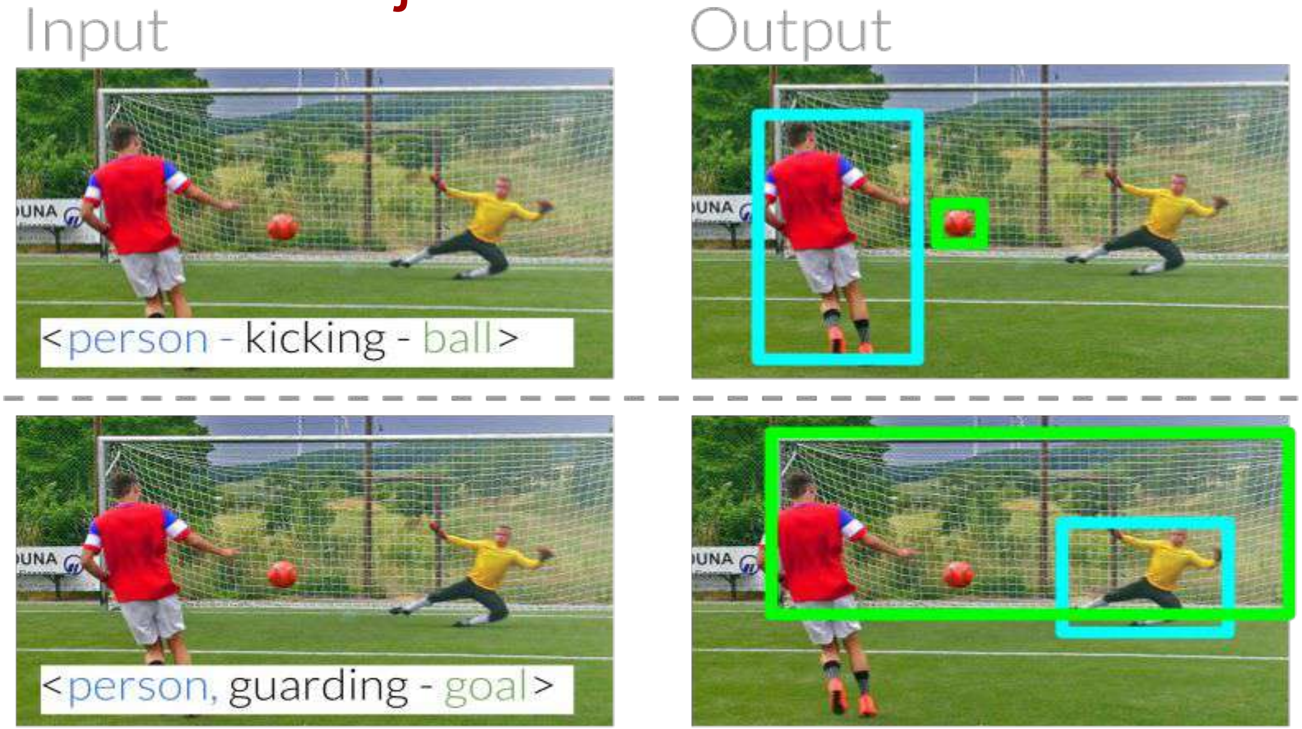
Scene graphs can improve image retrieval



Johnson, Krishna et al., Image Retrieval using Scene Graphs CVPR, 2015

Schuster, Krishna, et al., Generating Semantically Precise Scene Graphs from Textual Descriptions for Improved Image Retrieval, EMNLP 2015 workshop

Modeling relationships can improve existing vision tasks like object localization



Krishna et al., Referring Relationships CVPR, 2018

Zero shot detection

person sit chair
948 training examples

+

hydrant on ground
29 training examples

Zero shot detection


person sit chair
948 training examples


hydrant on ground
29 training examples

person sit hydrant
0 training examples

Zero shot detection


person ride horse
578 training examples

+


person wear hat
1023 training examples

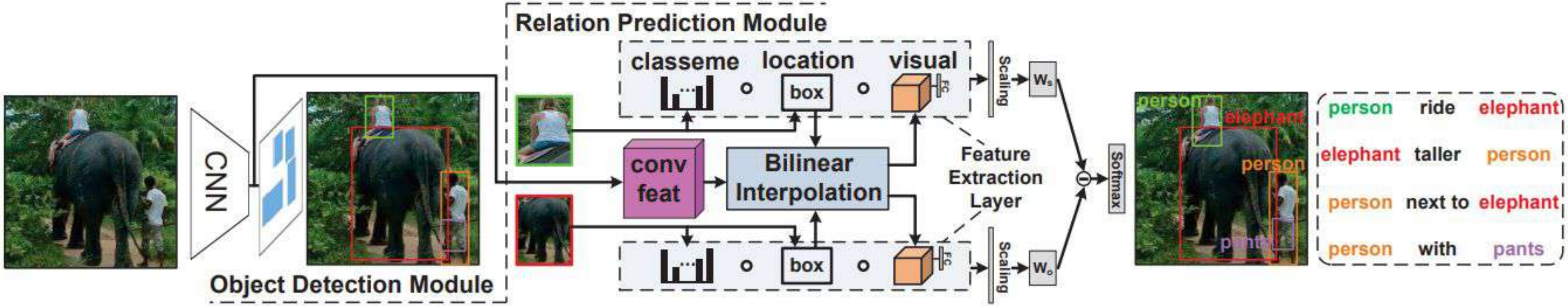
Zero shot detection

person ride horse
578 training examples

person wear hat
1023 training examples

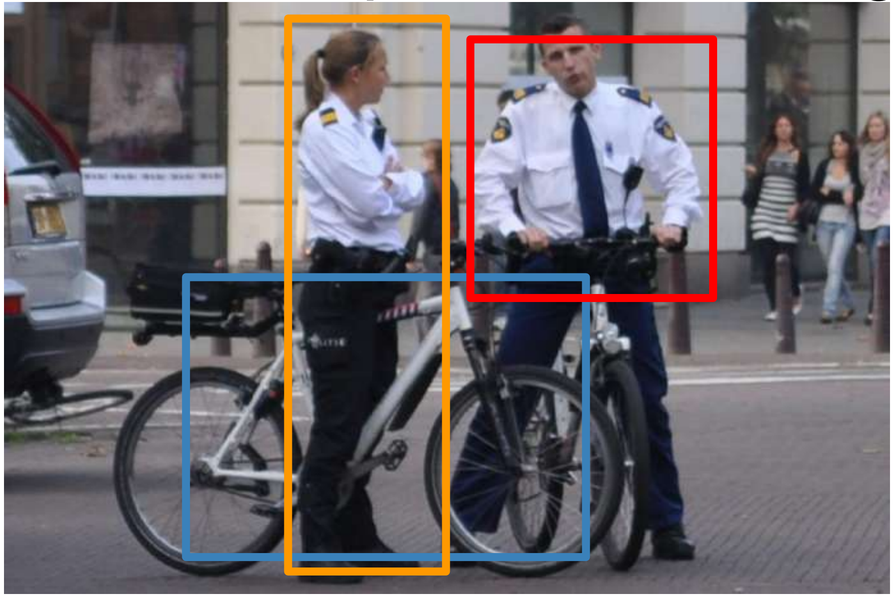
horse wear hat
0 training examples

Incorporating spatial features and classemes



Zhang, Hanwang, et al. "Visual translation embedding network for visual relation detection CVPR 2017

Problem with current method: Doesn't consider other relationships when making predictions

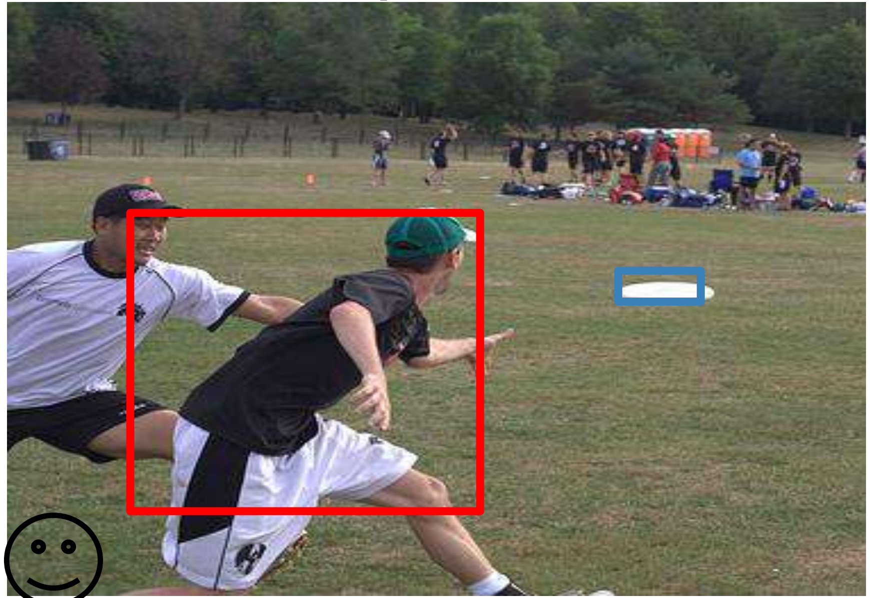


person ride bicycle

person ride bicycle

☹

Problem with current method: Doesn't consider other relationships when making predictions

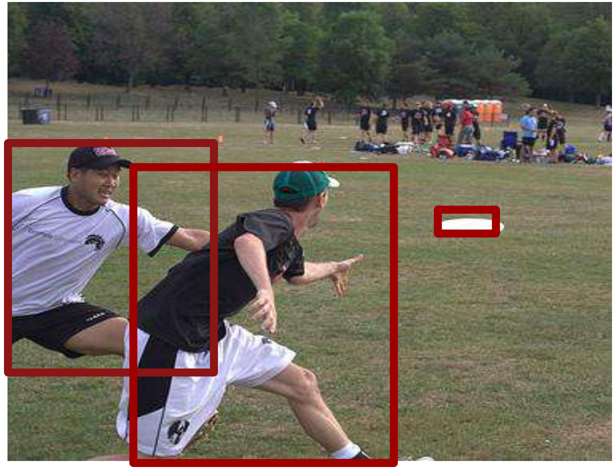


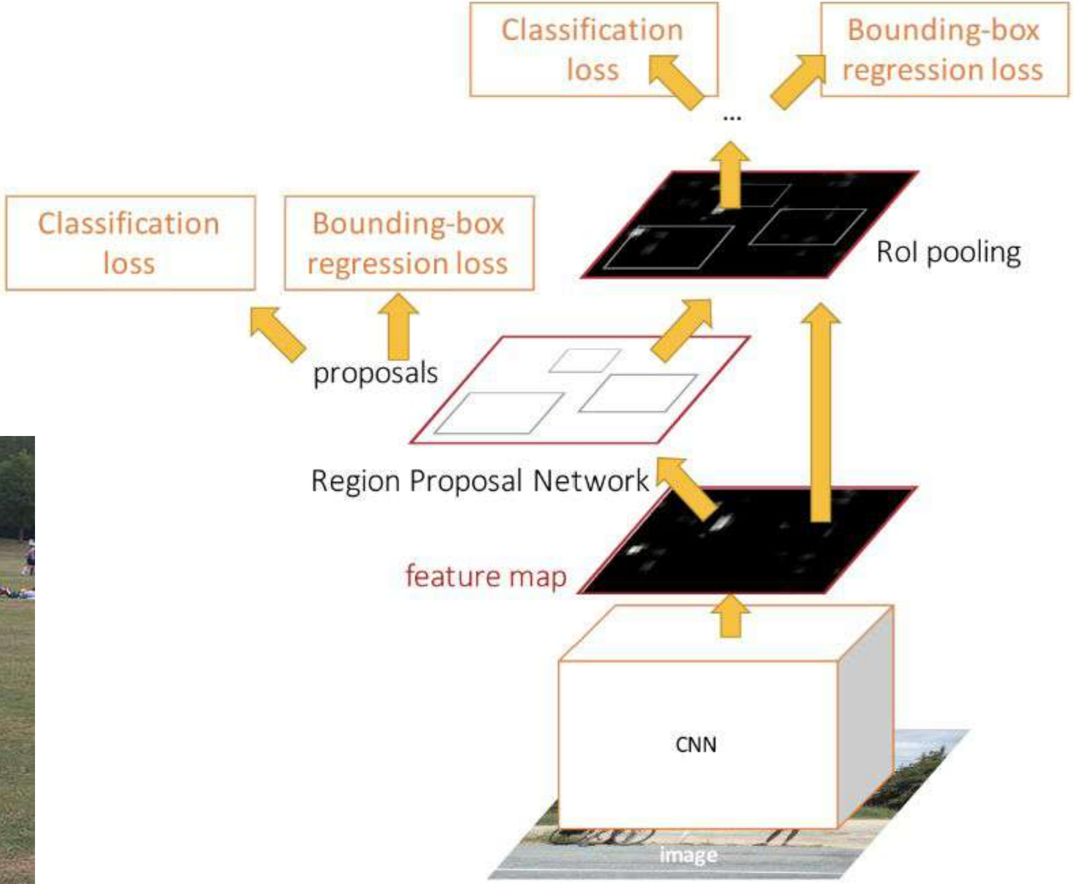
person throw frisbee

person throw frisbee

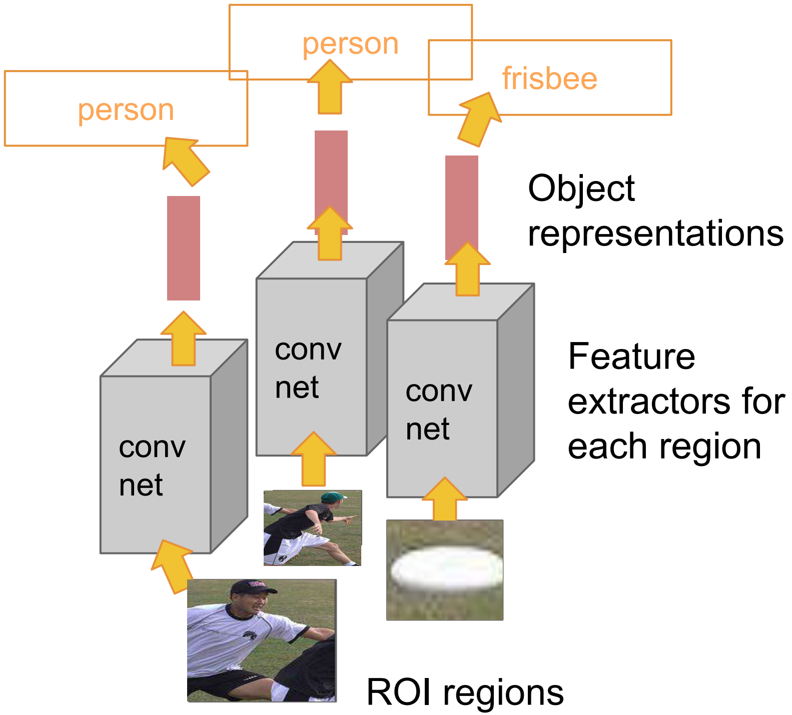
How do we model the other relationships in the image when making a prediction for a given relationship?

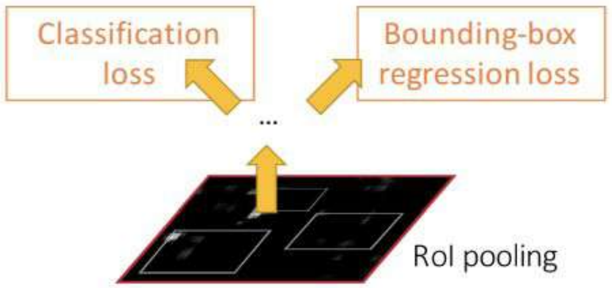
Recall Faster RCNN

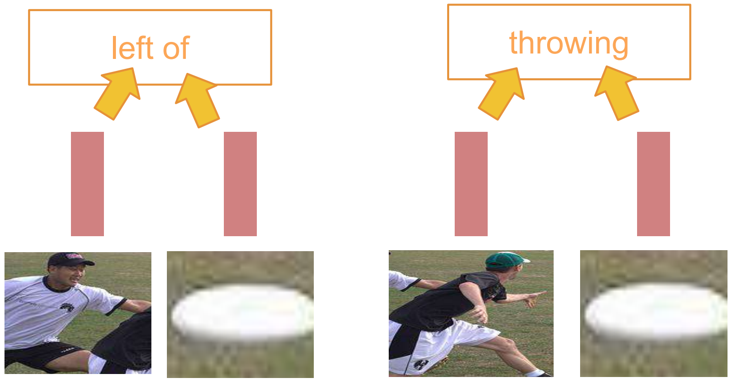




Each prediction in isolation



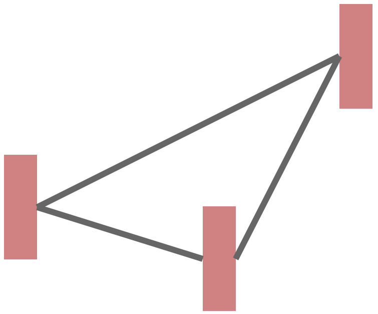




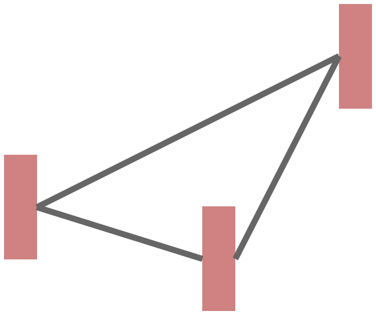
Representing objects as a graph with pairwise connections

But this graph doesn't encode the different kinds of relationships

Perform some operation that allows each node to encode what else is in the image.

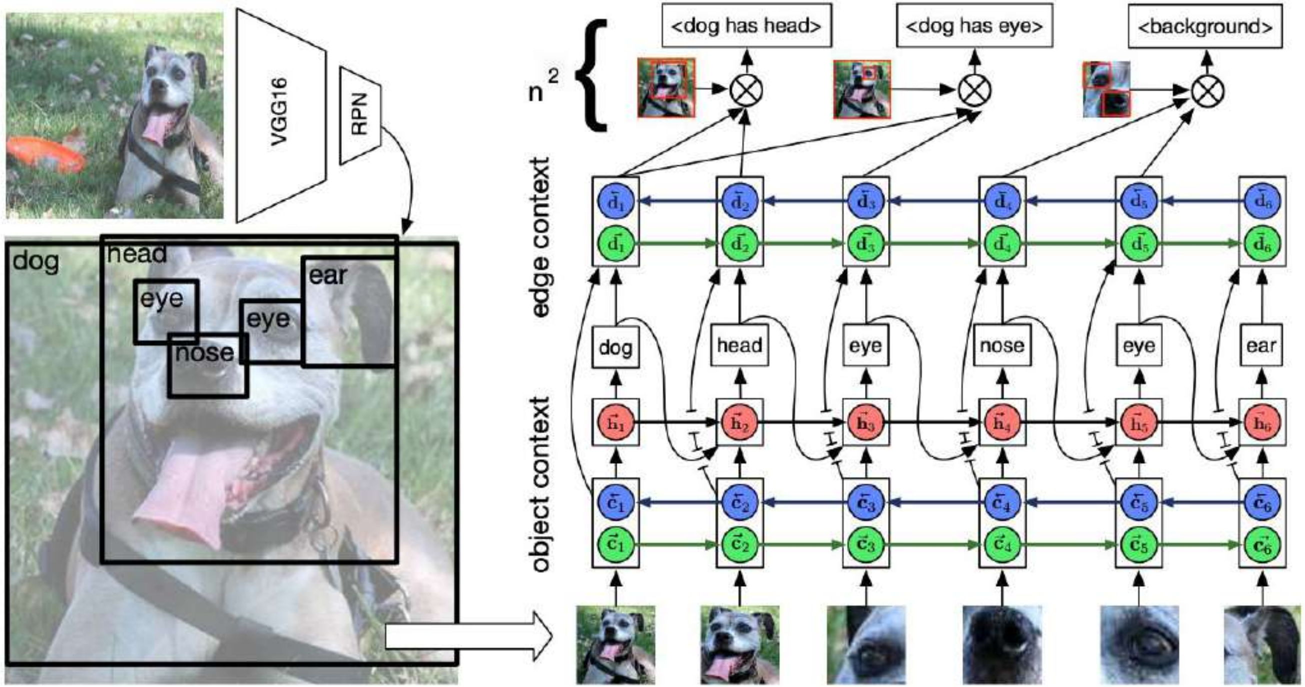


Each node contains features from individual regions



Each node now also contains features from all other regions

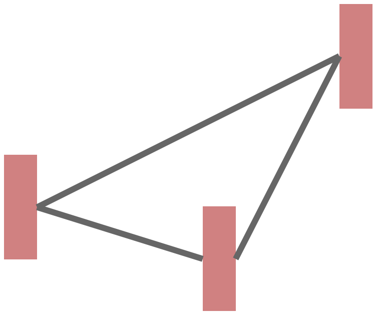
Use an RNN to collect information? But order of objects impacts predictions



Zellers et al. "Neural motifs: Scene graph parsing with global context." CVPR 2018

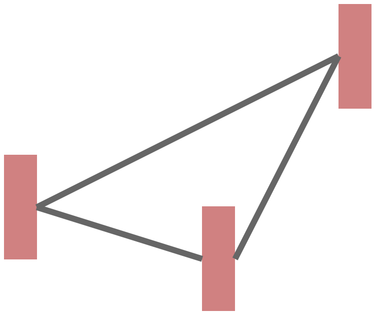
Representing objects as a graph with pairwise connections

But this graph doesn't encode the different kinds of relationships



Each node contains features from individual regions

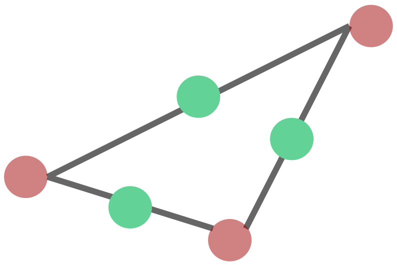
Perform some operation that allows each node to encode what else is in the image.



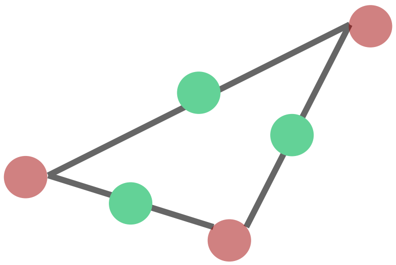
Each node now also contains features from all other regions

Graph representation with relationships included as nodes

Perform some operation that allows each node to encode what else is in the image.



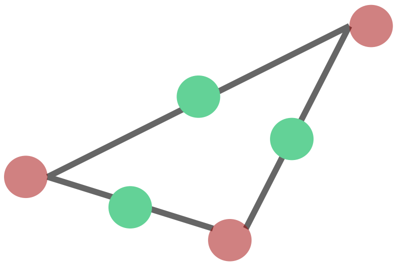
Each node contains features from individual regions



Each node now also contains features from all other regions

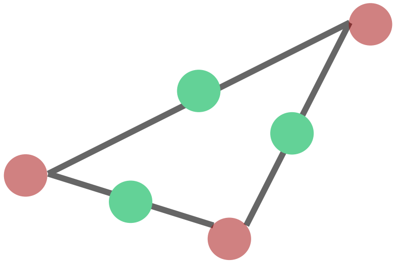
Graph representation with edges included as nodes

What operation have we already seen that updates features in a graph?



Each node contains features from individual regions

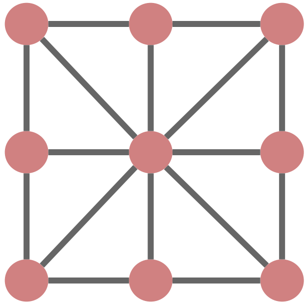
Perform some operation that allows each node to encode what else is in the image.

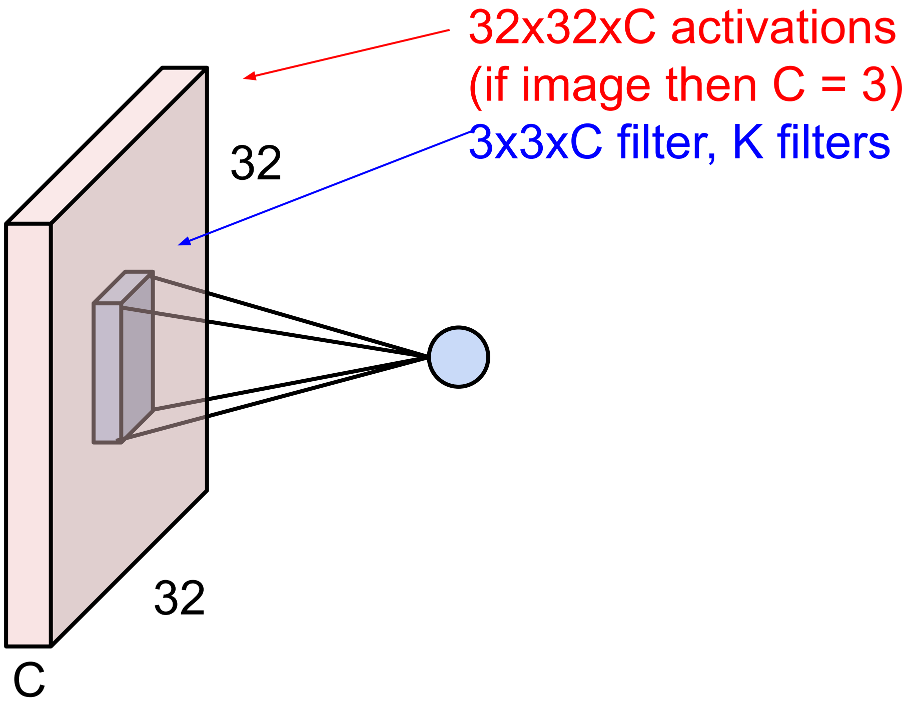


Each node now also contains features from all other regions

Recall Convolutions

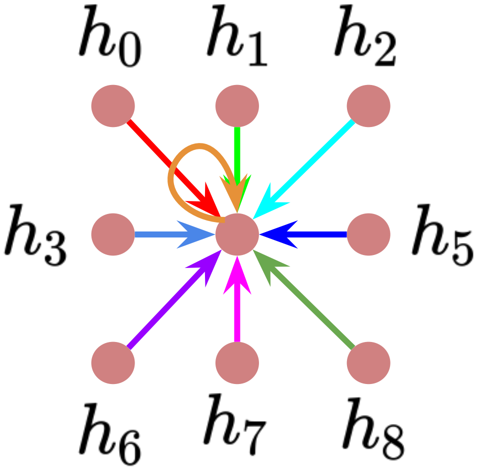
Images are a structured graph of pixels!



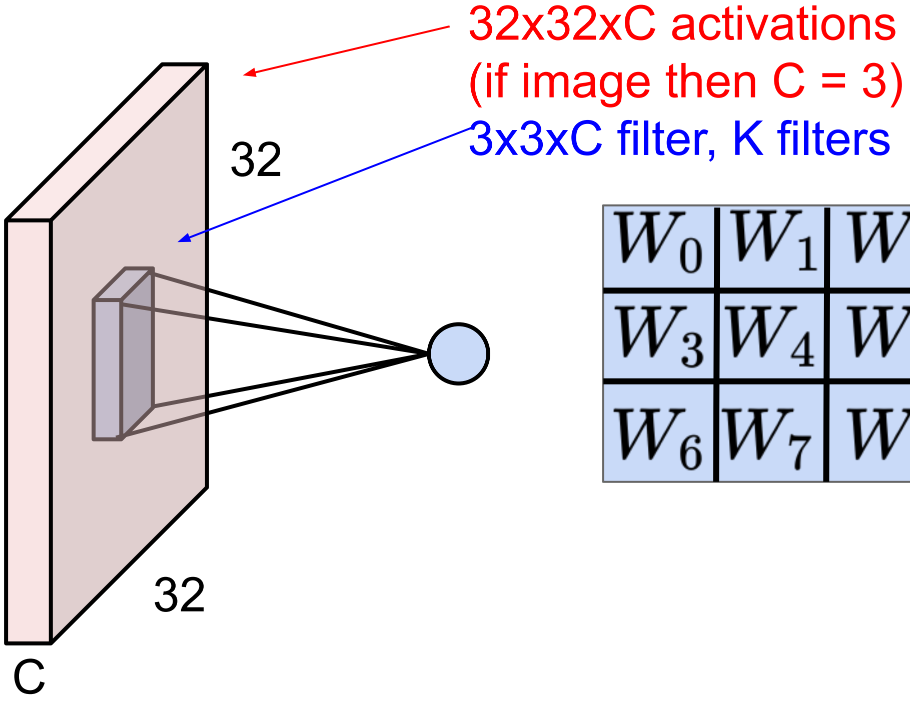


Recall Convolutions

Images are a structured graph of pixels!



$h_i \in \mathbb{R}^C$

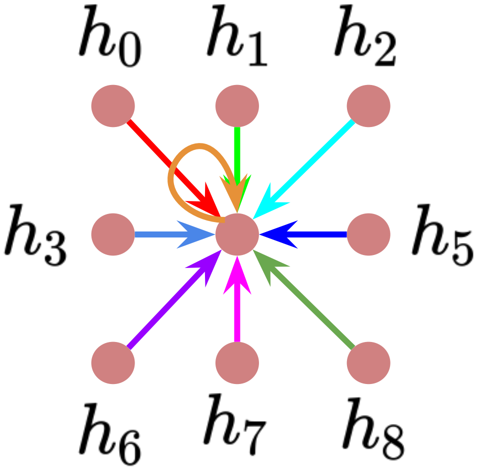


W_0	W_1	W_2
W_3	W_4	W_5
W_6	W_7	W_8

$W_i \in \mathbb{R}^{K \times C}$

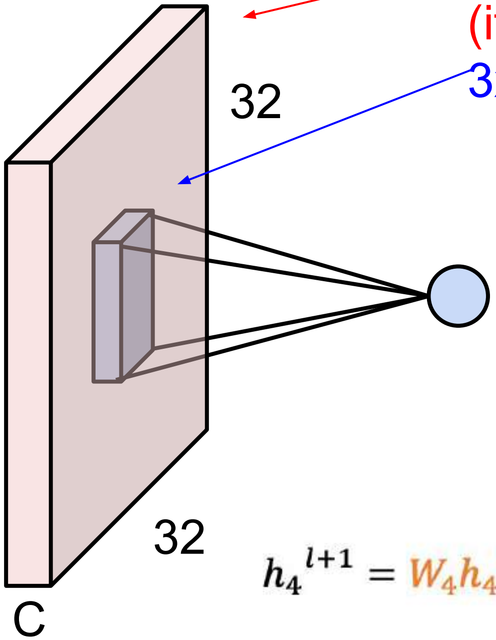
Recall Convolutions

Images are a structured graph of pixels!
Convolutions are local operations



$h_i \in \mathbb{R}^C$

32x32xC activations
(if image then C = 3)
3x3xC filter, K filters

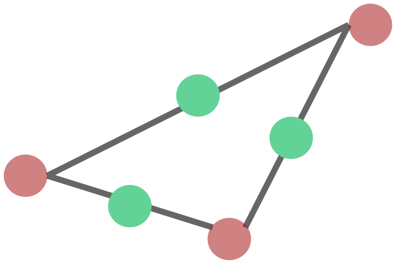


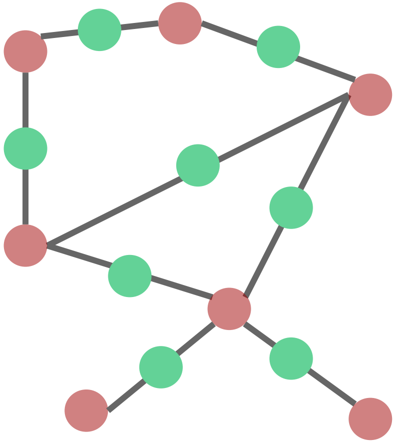
W_0	W_1	W_2
W_3	W_4	W_5
W_6	W_7	W_8

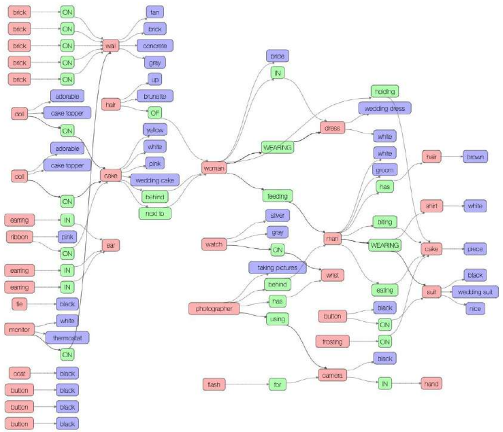
$W_i \in \mathbb{R}^{K \times C}$

$$h_4^{l+1} = W_4 h_4^{\,l} + W_0 h_0^{\,l} + W_1 h_1^{\,l} + W_2 h_2^{\,l} + \cdots + W_8 h_8^{\,l}$$

In comparison, scene graphs are not uniformly structured

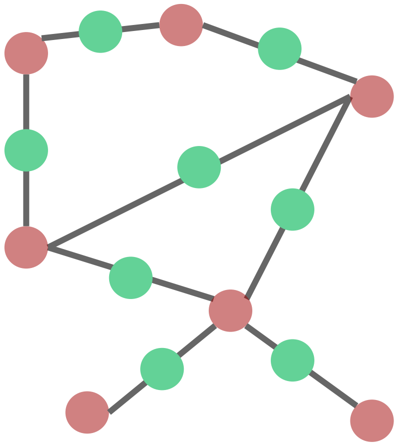


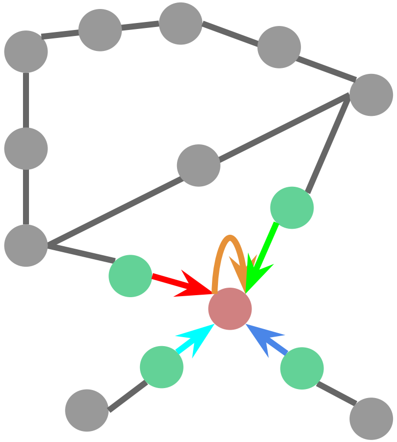




Objects have have varying number of relationships

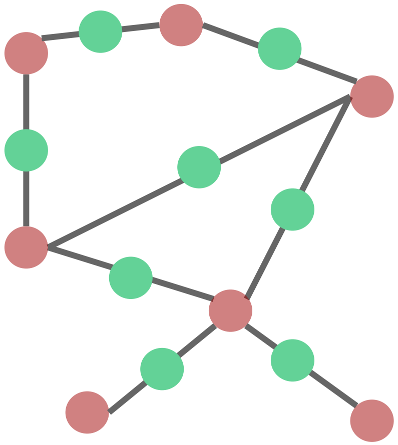
Generalizing 2D convolutions to Graph Convolutions

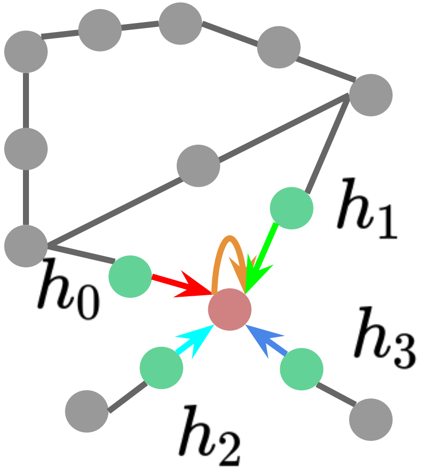




- Graph convolutions involve similar local operations on nodes.
- The ordering of neighbors should not matter.
- The number of neighbors should not matter.

Generalizing 2D convolutions to Graph Convolutions



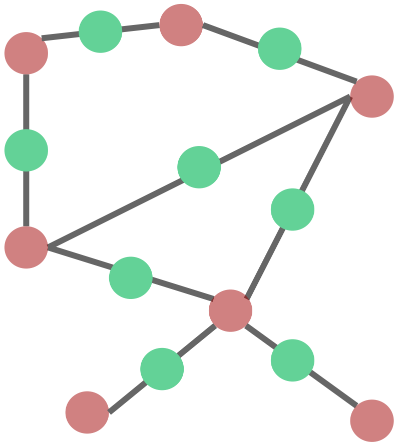


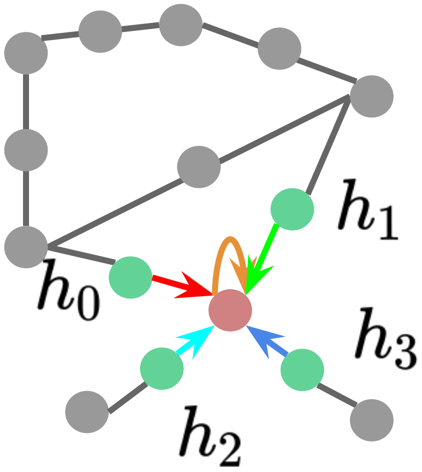
- Graph convolutions involve similar local operations on nodes.
- The ordering of neighbors should not matter.
- The number of neighbors should not matter.

$$h_4^{\,l+1} = W_4 h_4^{\,l} + W_0 h_0^{\,l} + W_1 h_1^{\,l} + W_2 h_2^{\,l} + W_3 h_3^{\,l}$$

But in this formulation the ordering matters

Generalizing 2D convolutions to Graph Convolutions



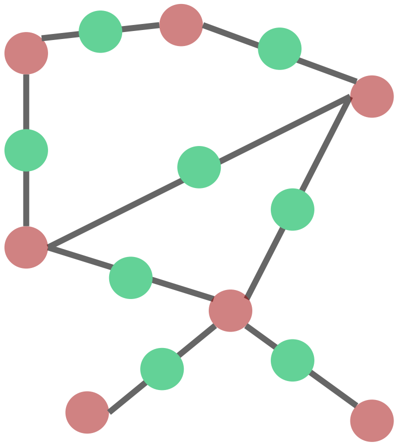


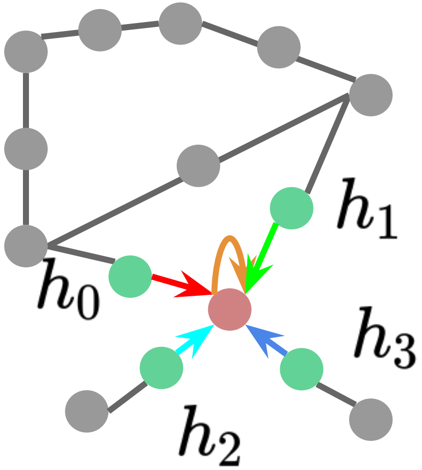
- Graph convolutions involve similar local operations on nodes.
- Nodes are now object representations and not activations
- The ordering of neighbors should not matter.
- The number of neighbors should not matter.
- N(i) are the neighbors of node i
- c_{ij} is a normalization constant

$$\cancel{h_4^{\,l+1} = W_4 h_4^{\,l} + W_0 h_0^{\,l} + W_1 h_1^{\,l} + W_2 h_2^{\,l} + W_3 h_3^{\,l}}$$

$$\boxed{h_i^{l+1} = W_{self}\, h_i^l + \sum_{j \in N(i)} \frac{1}{c_{ij}} W_{other} h_j^l}$$

Generalizing 2D convolutions to Graph Convolutions





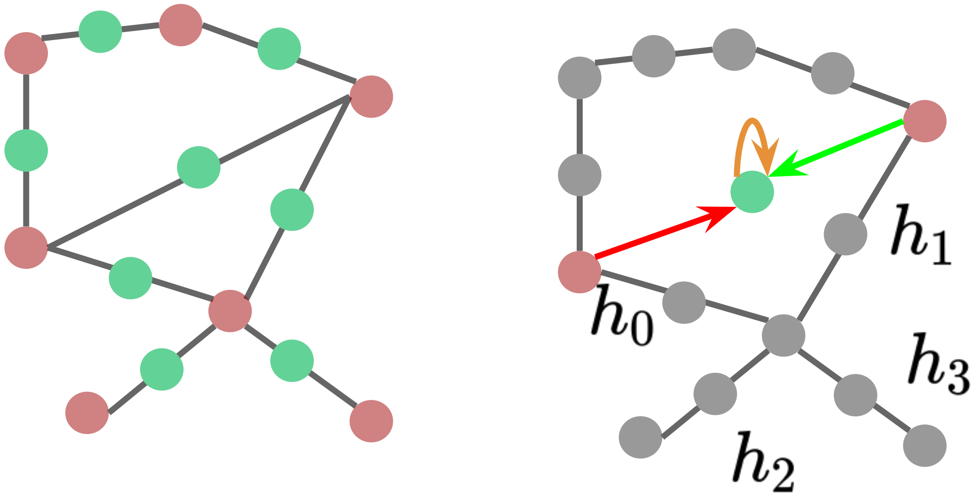
- Graph convolutions involve similar local operations on nodes.
- Nodes are now object representations and not activations
- The ordering of neighbors should not matter.
- The number of neighbors should not matter.
- N(i) are the neighbors of node i
- c_{ij} is a normalization constant

$$\cancel{h_4^{l+1} = W_4 h_4^{l} + W_0 h_0^{l} + W_1 h_1^{l} + W_2 h_2^{l} + W_3 h_3^{l}}$$

$$h_i^{l+1} = W_{self}\, h_i^l + \sum_{j \in N(i)} \frac{1}{c_{ij}} W_{other} h_j^l$$

$$\boxed{h_i^{l+1} = W\ \ h_i^l + \sum_{j \in N(i)} \frac{1}{c_{ij}} W\ \ h_j^l}$$

Generalizing 2D convolutions to Graph Convolutions

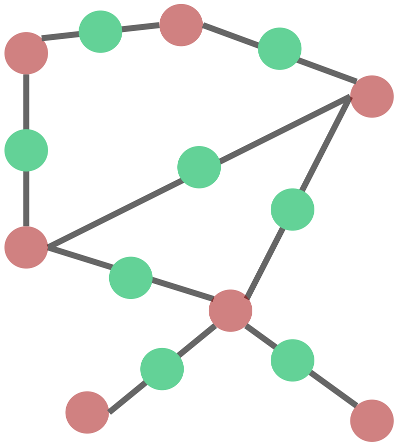


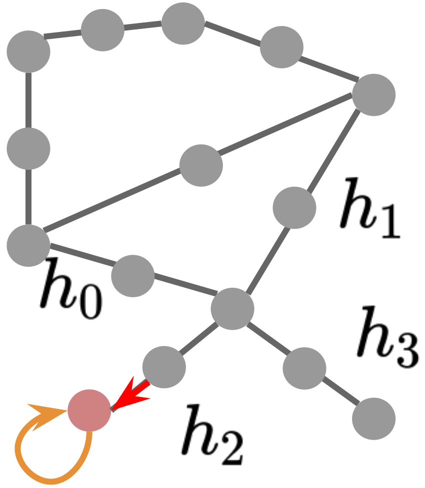
- Graph convolutions involve similar local operations on nodes.
- Nodes are now object representations and not activations
- The ordering of neighbors should not matter.
- The number of neighbors should not matter.
- N(i) are the neighbors of node i

$$h_i^{l+1} = W\ h_i^l + \sum_{j \in N(i)} \frac{1}{c_{ij}} W\ h_j^l$$

Kipf & Welling (ICLR 2017)

Generalizing 2D convolutions to Graph Convolutions



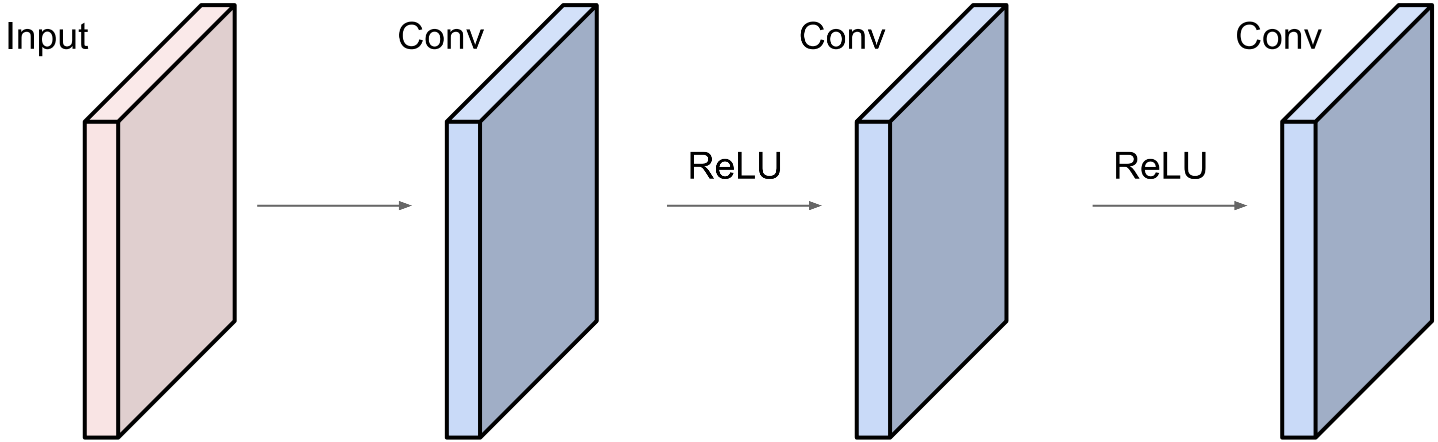


- Graph convolutions involve similar local operations on nodes.
- Nodes are now object representations and not activations
- The ordering of neighbors should not matter.
- The number of neighbors should not matter.
- N(i) are the neighbors of node i

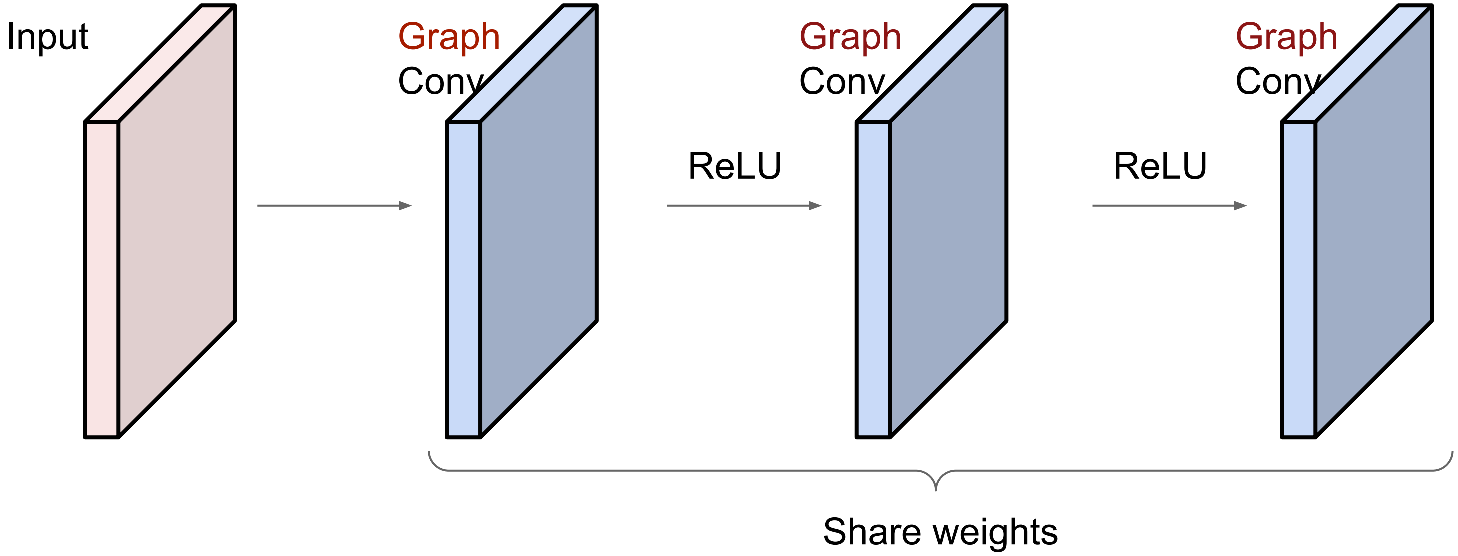
$$h_i^{l+1} = W \; h_i^l + \sum_{j \in N(i)} \frac{1}{c_{ij}} W \; h_j^l$$

Kipf & Welling (ICLR 2017)

To increase receptive field of CNNs: increase depth



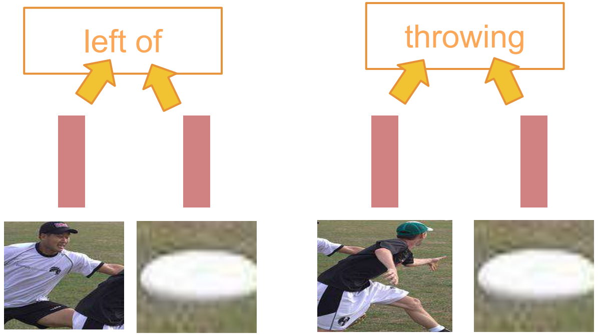
To increase receptive field of GCNs: increase depth



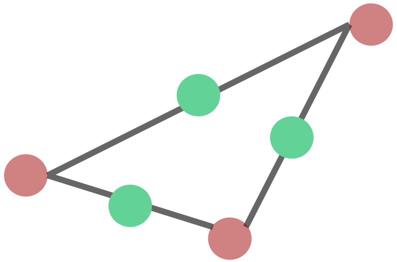
GCNs: Graph Convolutional Networks

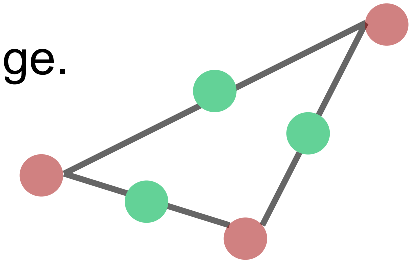
Kipf & Welling (ICLR 2017)

Graph representation with edges included as nodes



Perform Graph Convolutions,
which allows each node to
encode what else is in the image.

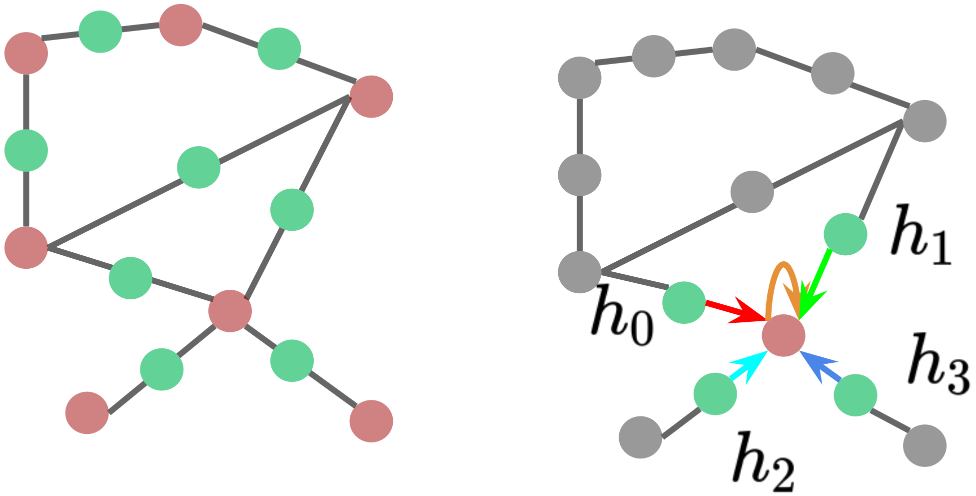




Each node contains features from individual regions

Each node now also contains features from all other regions

Graph Convolutions with Attention



- Updates from some neighbors can be more important than others.
- Attention over neighbors allows graph convolutions to focus on specific neighbors
- σ is a non-linearity, usually ReLU or LeakyReLU.

Without attention: $$h_i^{l+1} = W \ h_i^l + \sum_{j \in N(i)} \frac{1}{c_{ij}} W \ h_j^l$$

With attention: $$h_i^{l+1} = W \ h_i^l + \sum_{j \in N(i)} \frac{1}{c_{ij}} \boxed{\alpha_{ij}} W \ h_j^l$$

where $$\alpha_{ij} = \frac{e^{\sigma(a^T[Wh_i || Wh_j])}}{\sum_{k \in \mathcal{N}(i)} e^{\sigma(a^T[Wh_i || Wh_k])}}$$

How is it actually implemented?

For loops iterating over all the neighbors is expensive

$$h_i^{l+1} = W \; h_i^l + \sum_{j \in N(i)} \frac{1}{c_{ij}} W \; h_j^l$$

Formalizing a graph representation

For loops iterating over all the neighbors is expensive

$$h_i^{l+1} = W \; h_i^l + \sum_{j \in N(i)} \frac{1}{c_{ij}} W \; h_j^l$$

Let's define a graph with nodes and edges: $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with N nodes

Formalizing a graph representation

For loops iterating over all the neighbors is expensive

$$h_i^{l+1} = W \; h_i^l + \sum_{j \in N(i)} \frac{1}{c_{ij}} W \; h_j^l$$

Let's define a graph with nodes and edges: $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with N nodes

Let's define the adjacency matrix of a graph as: $A \in \mathbb{R}^{N \times N}$ $A_{ij} = \begin{cases} 1 & \text{if } e_{ij} \in \mathcal{E} \\ 0 & \text{otherwise} \end{cases}$

Formalizing a graph representation

For loops iterating over all the neighbors is expensive

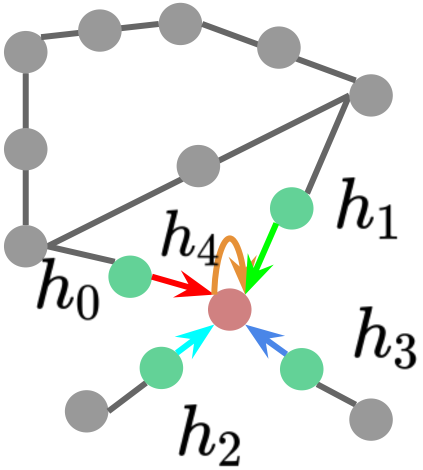
$$h_i^{l+1} = W \ h_i^l + \sum_{j \in N(i)} \frac{1}{c_{ij}} W \ h_j^l$$

Let's define a graph with nodes and edges: $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with N nodes

Let's define the adjacency matrix of a graph as: $A \in \mathbb{R}^{N \times N}$ $A_{ij} = \begin{cases} 1 & \text{if } e_{ij} \in \mathcal{E} \\ 0 & \text{otherwise} \end{cases}$

Finally, let's define the degree matrix: $D \in \mathbb{R}^{N \times N}$ $D_{ij} = \begin{cases} \mathcal{N}(i) & \text{if } i = j \\ 0 & \text{otherwise} \end{cases}$

Vectorized graph convolution



For loops iterating over all the neighbors is expensive

$$h_i^{l+1} = W\ h_i^l + \sum_{j \in N(i)} \frac{1}{c_{ij}} W\ h_j^l$$

Examples:

$$D_{00} = 2$$
$$D_{44} = 4$$

$$A \in \mathbb{R}^{N \times N} \qquad A_{ij} = \begin{cases} 1 & \text{if } e_{ij} \in \mathcal{E} \\ 0 & \text{otherwise} \end{cases}$$

$$A_{04} = A_{40} = 1$$
$$A_{01} = A_{10} = 0$$

$$D \in \mathbb{R}^{N \times N} \qquad D_{ij} = \begin{cases} \mathcal{N}(i) & \text{if } i = j \\ 0 & \text{otherwise} \end{cases}$$

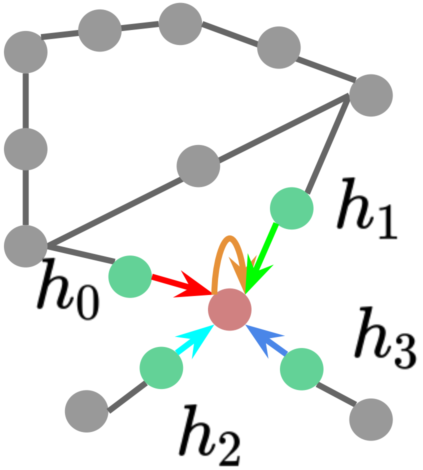
Vectorized graph convolution

First, let's stack all the node representations in a matrix H:

$$H^l \in \mathbb{R}^{N \times C}$$

Such that every row is a node:

$$h_i \in \mathbb{R}^C$$



For loops iterating over all the neighbors is expensive

$$h_i^{l+1} = W\ h_i^l + \sum_{j \in N(i)} \frac{1}{c_{ij}} W\ h_j^l$$

$$A \in \mathbb{R}^{N \times N} \qquad A_{ij} = \begin{cases} 1 & \text{if } e_{ij} \in \mathcal{E} \\ 0 & \text{otherwise} \end{cases}$$

$$D \in \mathbb{R}^{N \times N} \qquad D_{ij} = \begin{cases} \mathcal{N}(i) & \text{if } i = j \\ 0 & \text{otherwise} \end{cases}$$

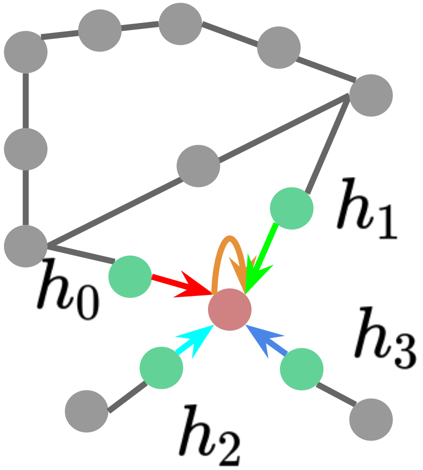
Vectorized graph convolution

First, let's stack all the node representations in a matrix H:

$$H^l \in \mathbb{R}^{N \times C}$$

Such that every row is a node:

$$h_i \in \mathbb{R}^C$$



For loops iterating over all the neighbors is expensive

$$h_i^{l+1} = W\ h_i^l + \sum_{j \in N(i)} \frac{1}{c_{ij}} W\ h_j^l$$

$$A \in \mathbb{R}^{N \times N} \qquad A_{ij} = \begin{cases} 1 & \text{if } e_{ij} \in \mathcal{E} \\ 0 & \text{otherwise} \end{cases}$$

$$D \in \mathbb{R}^{N \times N} \qquad D_{ij} = \begin{cases} \mathcal{N}(i) & \text{if } i = j \\ 0 & \text{otherwise} \end{cases}$$

The vectorized computation of graph convolution is:

$$H^{l+1} = D^{-1/2} \hat{A} D^{-1/2} H^l W \qquad \hat{A} = A + I$$

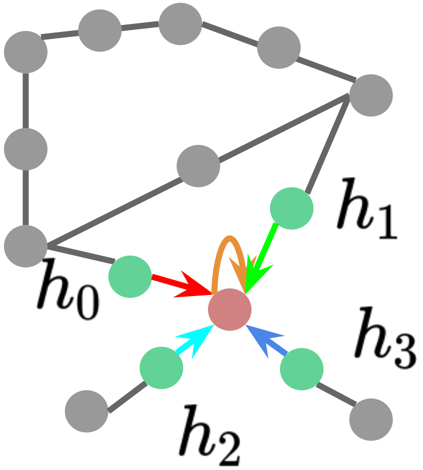
Vectorized graph convolution

First, let's stack all the node representations in a matrix H:

$$H^l \in \mathbb{R}^{N \times C}$$

Such that every row is a node:

$$h_i \in \mathbb{R}^C$$



For loops iterating over all the neighbors is expensive

$$h_i^{l+1} = W \ h_i^l + \sum_{j \in N(i)} \frac{1}{c_{ij}} W \ h_j^l$$

$$A \in \mathbb{R}^{N \times N} \qquad A_{ij} = \begin{cases} 1 & \text{if } e_{ij} \in \mathcal{E} \\ 0 & \text{otherwise} \end{cases}$$

$$D \in \mathbb{R}^{N \times N} \qquad D_{ij} = \begin{cases} \mathcal{N}(i) & \text{if } i = j \\ 0 & \text{otherwise} \end{cases}$$

Can be pre-calculated once per graph:

$$H^{l+1} = \boxed{D^{-1/2} \hat{A} D^{-1/2}} H^l \boxed{W}$$

Linear layer weights

$$\hat{A} = A + I$$

Aside: Grounding to spectral convolutions with graph laplacian

Convolutions in the spectral domain:

$$W * h = U \operatorname{diag}(W)\, U^T h$$

Where U is the eigenvectors of the graph laplacian:

$$L = I + D^{-1/2} A D^{-1/2} = U \Lambda U^T$$

You can approximate spectral graph convolutions as 1st order Chebyshev polynomials to get:

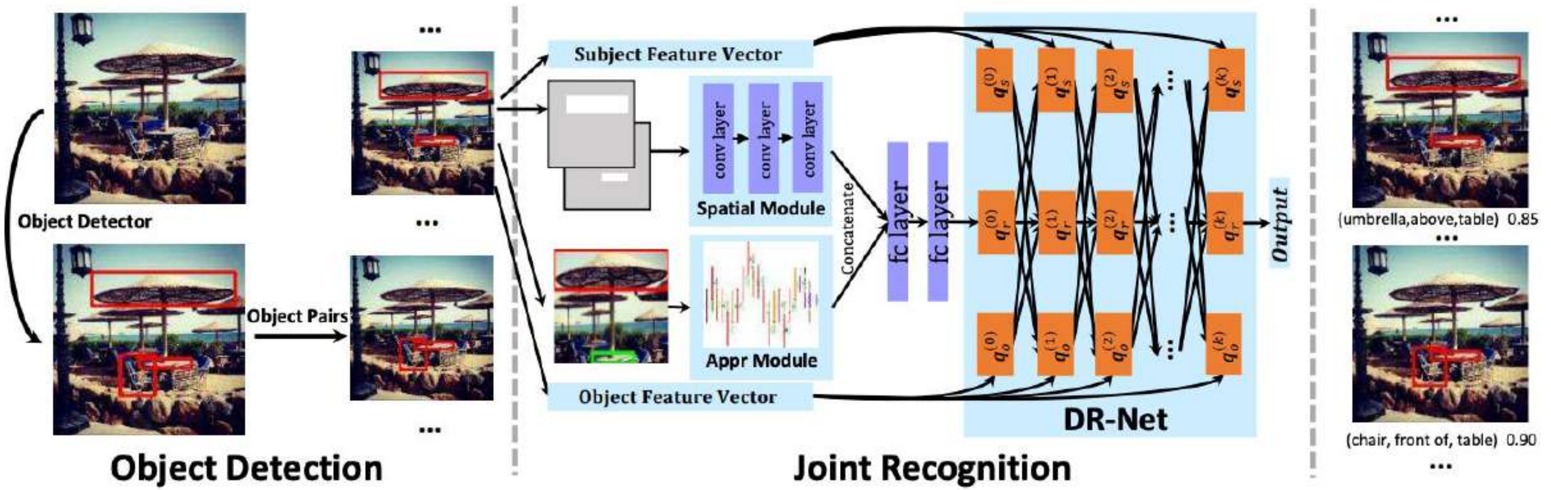
$$W * h = W(I + D^{-\frac{1}{2}} A D^{-\frac{1}{2}})h$$

Renormalize the weights to get our spatial graph convolutions:

$$I + D^{-1/2} A D^{-1/2} \rightarrow D^{-1/2} \hat{A} D^{-1/2}$$

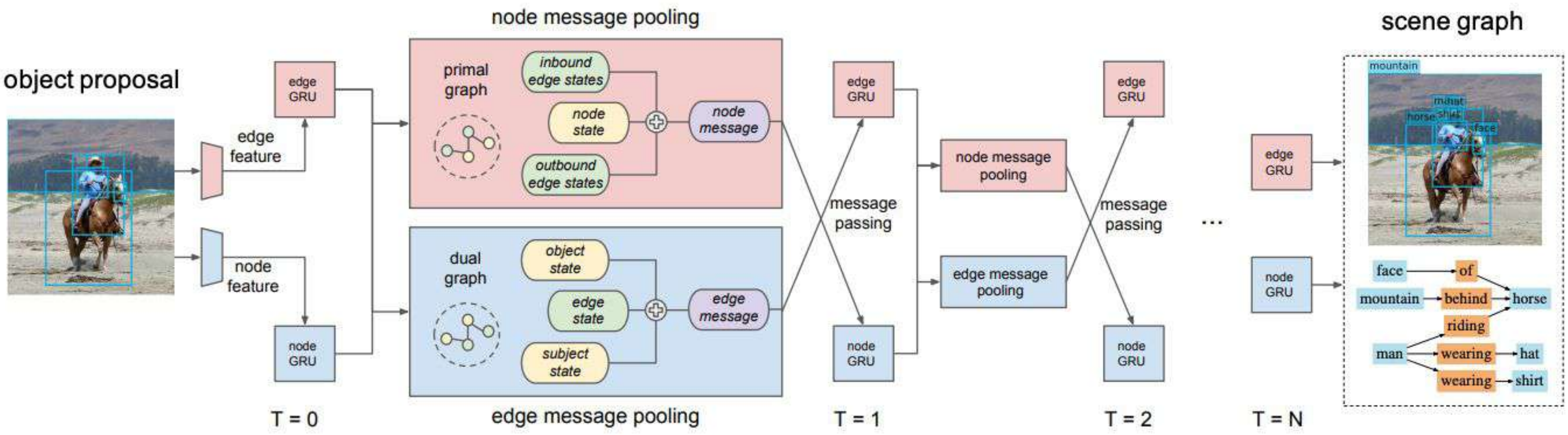
$$H^{l+1} = D^{-1/2} \hat{A} D^{-1/2} H^l W \qquad \hat{A} = A + I$$

Scene Graph Generation with Graph Convolution methods



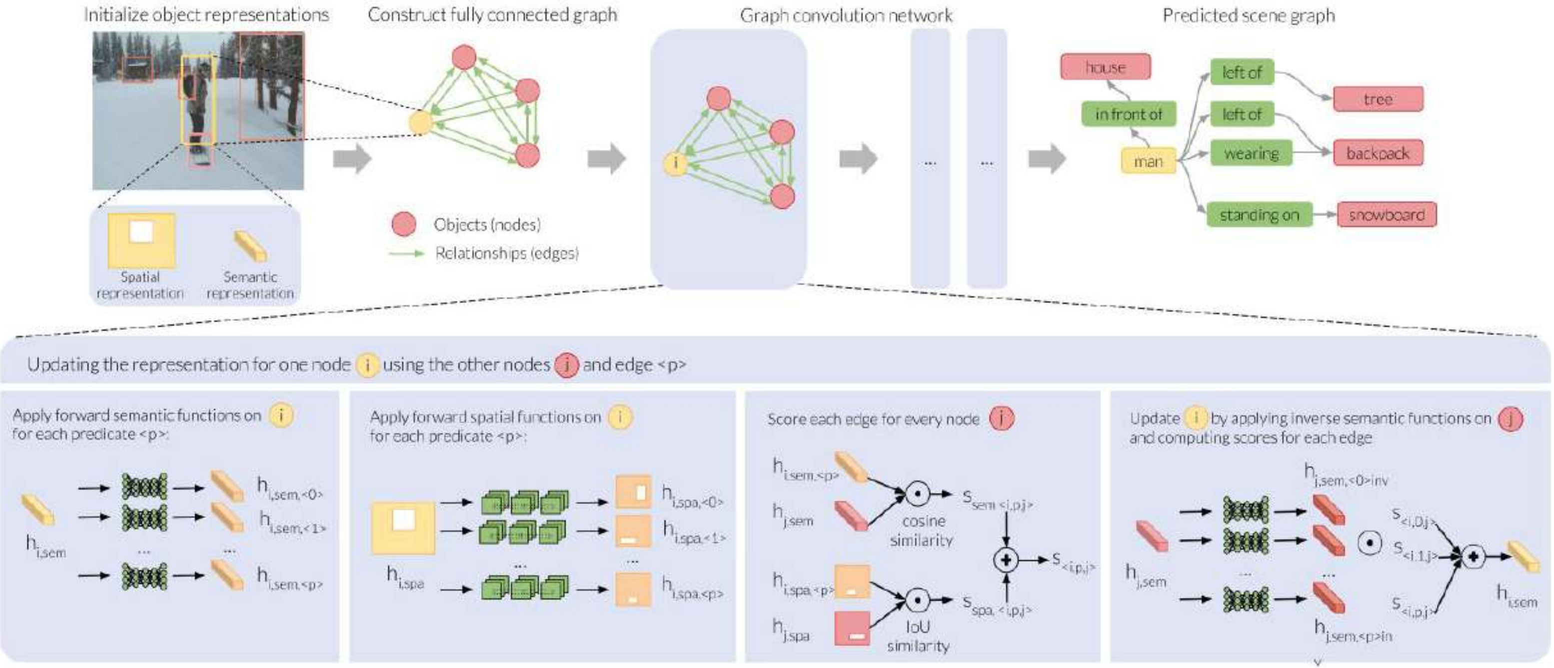
Liang et al. Deep variation-structured reinforcement learning for visual relationship and attribute detection, CVPR 2017

Scene Graph Generation with node and edge Graph Convolution methods



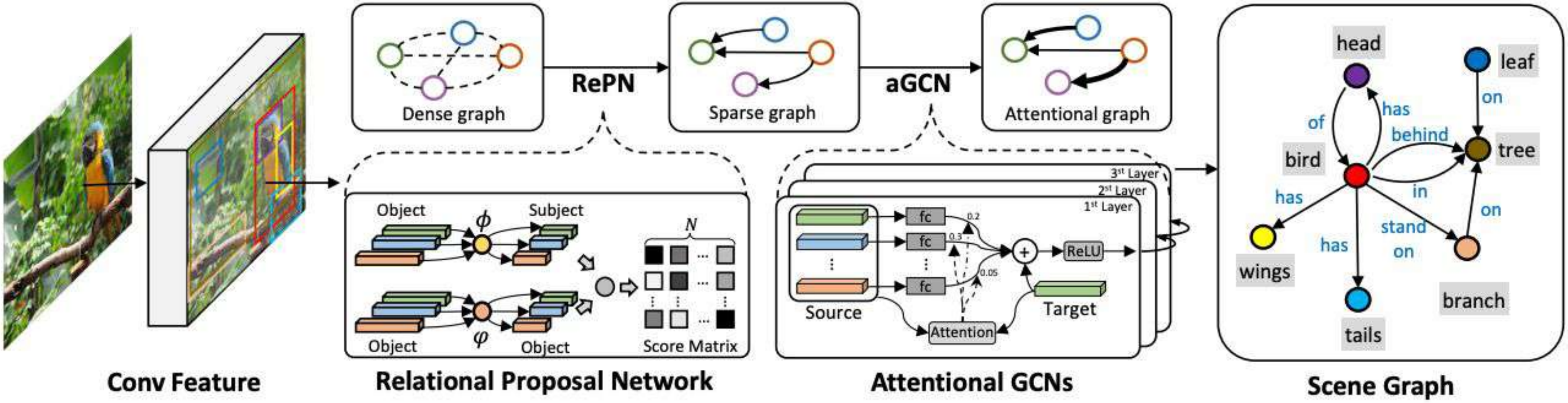
Xu et al. "Scene graph generation by iterative message passing, CVPR 2017

Few shot scene graph generation with graph convolution methods



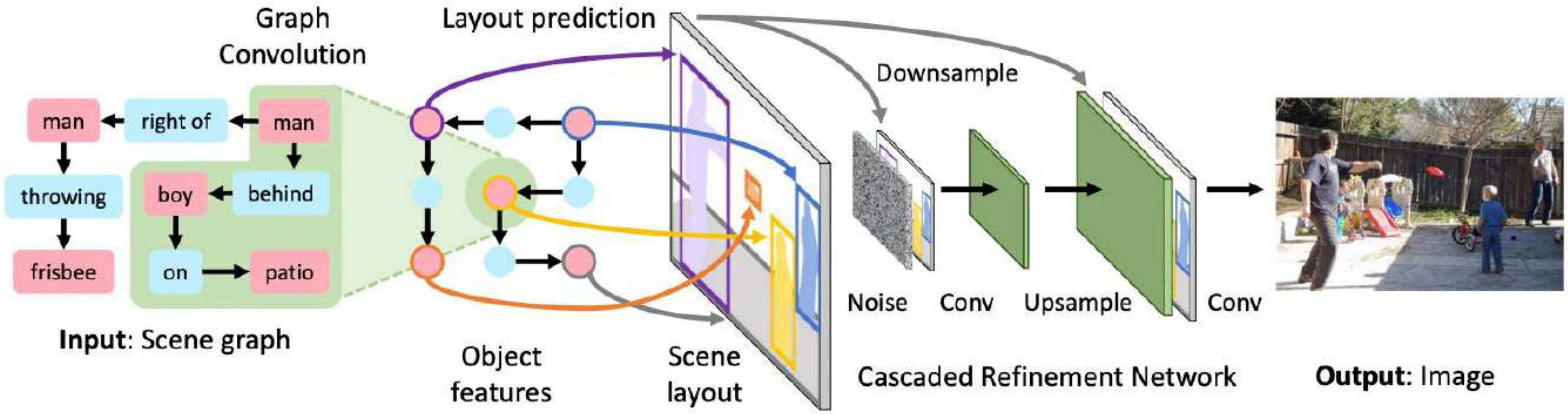
Dornadula, Narcomey, Krishna, et al. "Visual Relationships as Functions: Enabling Few-Shot Scene Graph Prediction." Proceedings of the IEEE International Conference on Computer Vision Workshops. 2019

Scene graph generation with graph attention convolutions



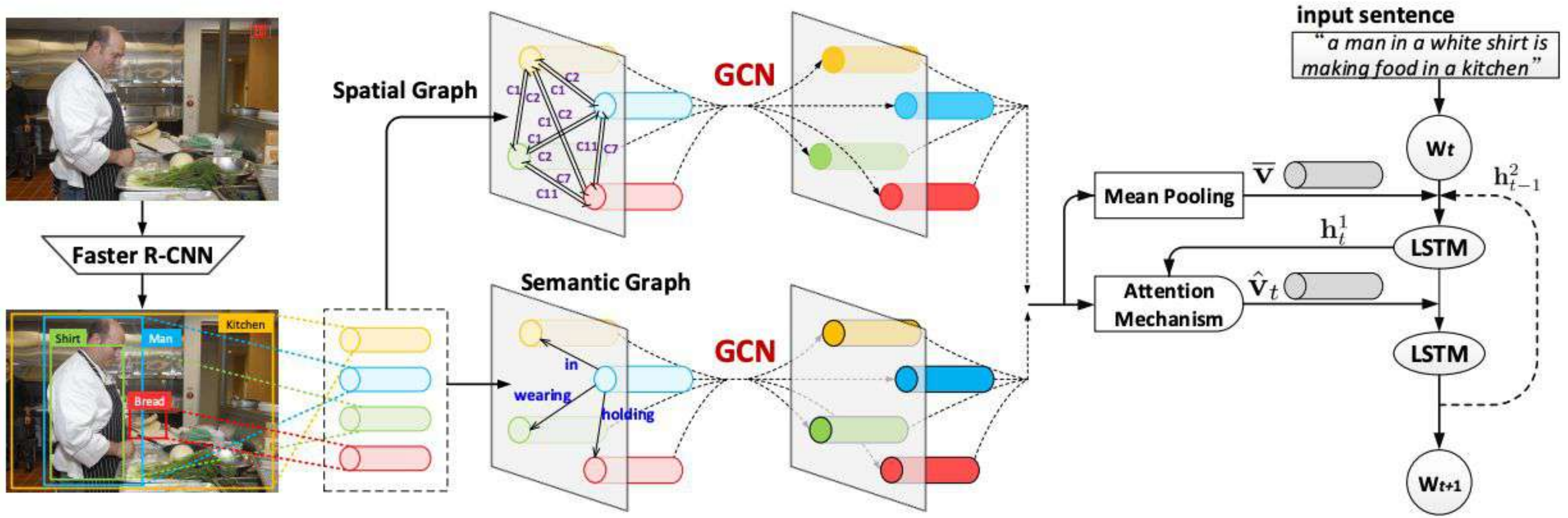
Yang, et al. "Graph r-cnn for scene graph generation." Proceedings of the European conference on computer vision ECCV 2018

Image generation from scene graphs



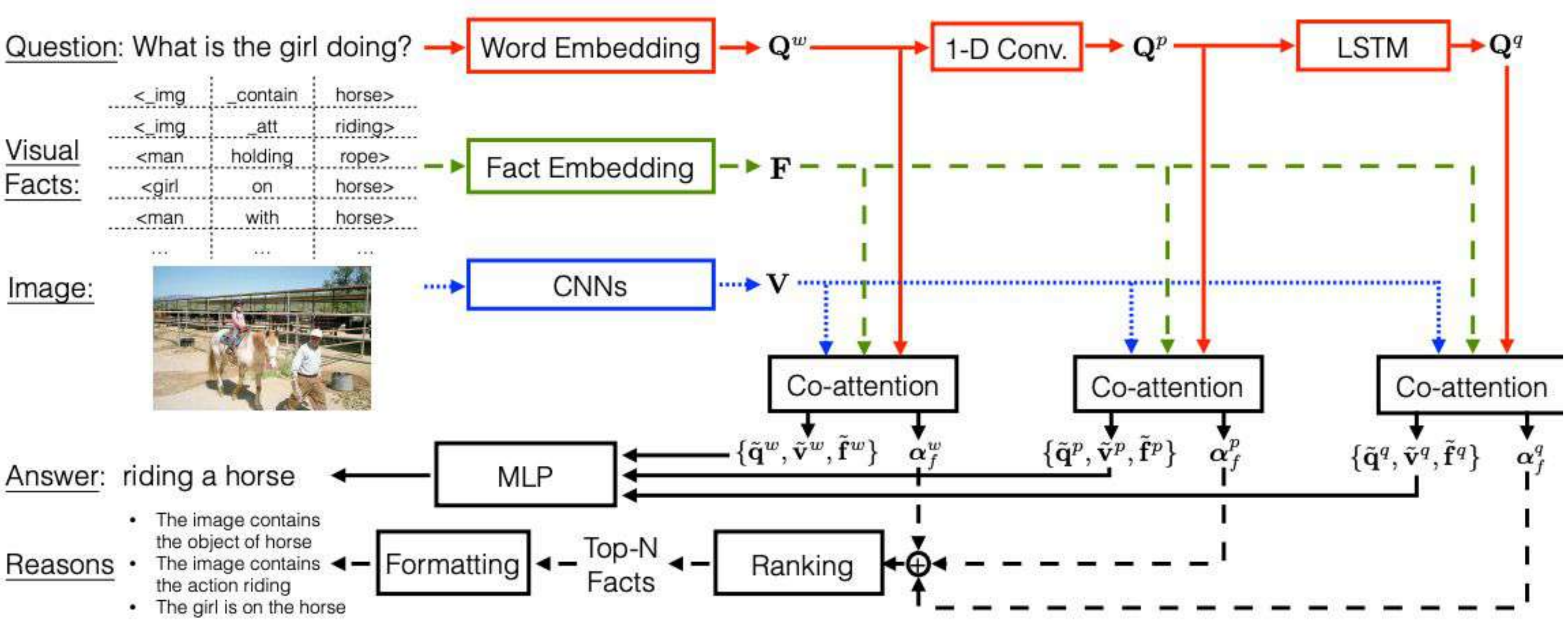
Johnson et al. Image generation from scene graphs, CVPR 2019

Scene graphs as intermediate representation for image captioning



Yao et al. Exploring Visual Relationship for Image Captioning, ECCV 2018

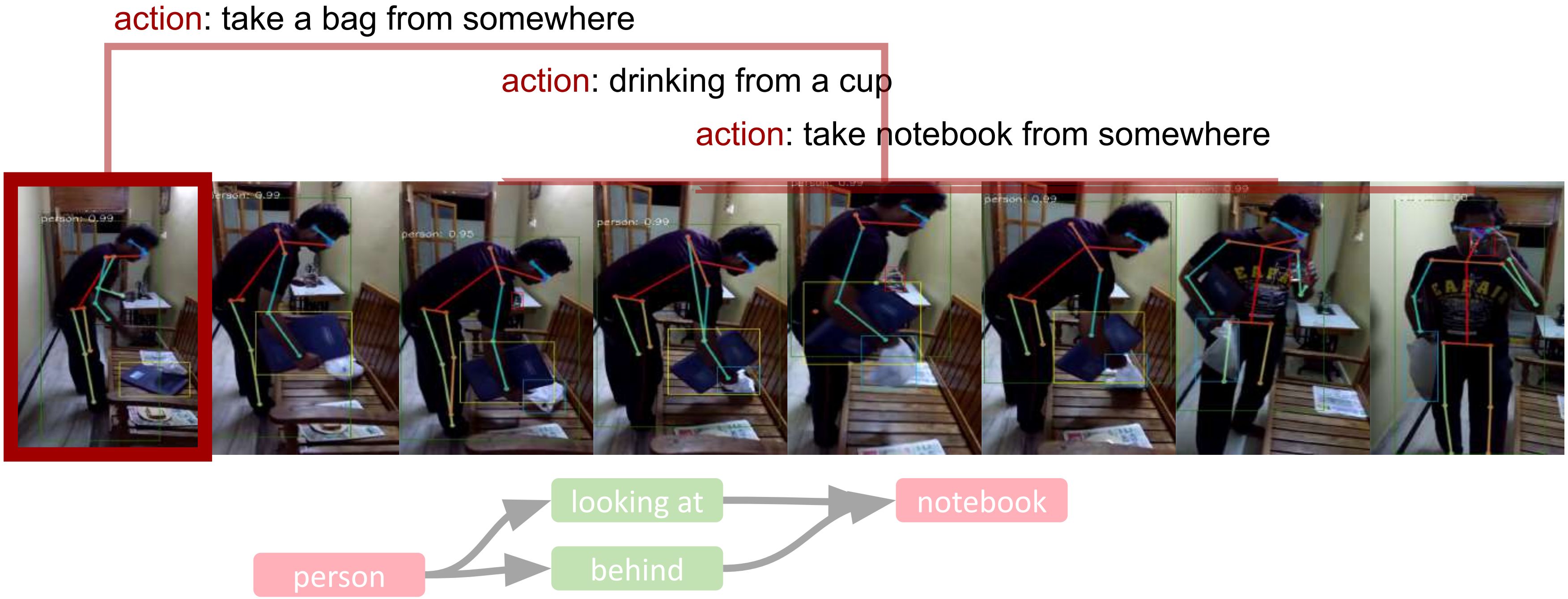
Scene graphs as intermediate representation for visual question answering



Wang et al. The vqa-machine: Learning how to use existing vision algorithms to answer new questions CVPR 2017

So what’s next for scene graphs?

Action Genome: Understanding Actions with Spatio-Temporal Scene Graphs



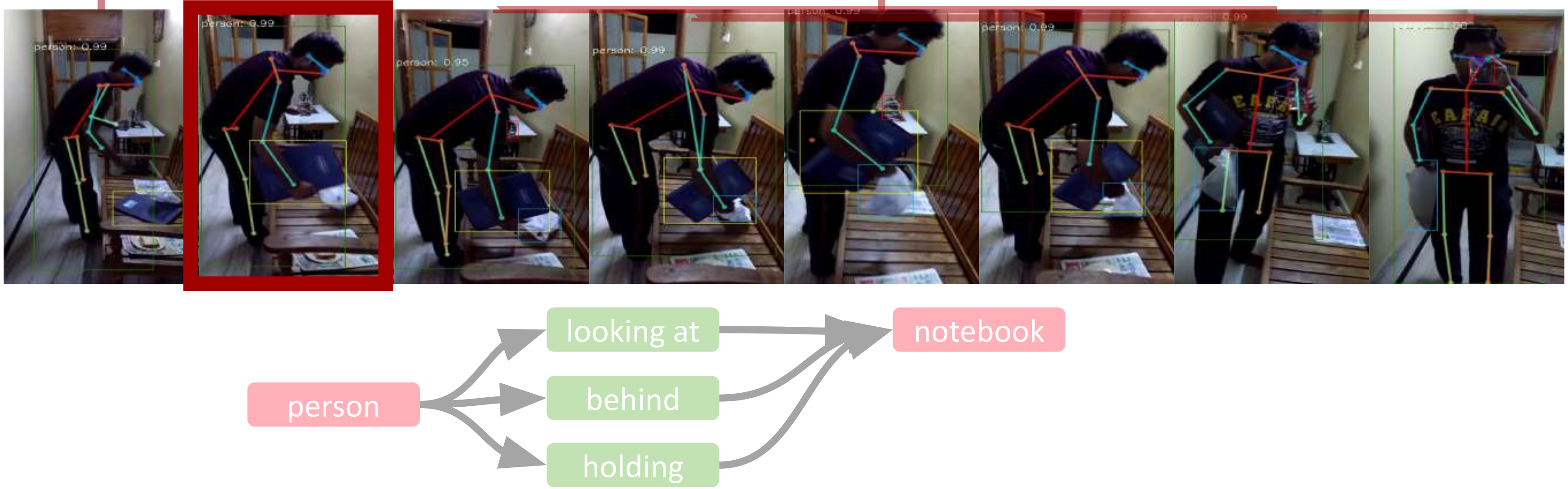
Krishna et a. Dense Captioning Events in Videos, CVPR 2017
Ji, Krishna et al. Action Genome: Actions as Compositions of Spatio-Temporal Scene Graphs, CVPR 2020

Action Genome: Understanding Action with Spatio-Temporal Scene Graphs

action: take a bag from somewhere

action: drinking from a cup

action: take notebook from somewhere



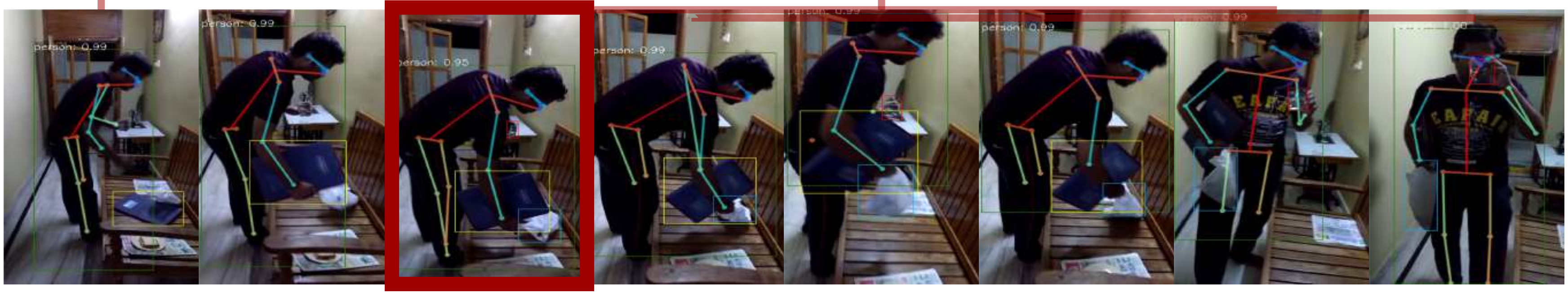
Krishna et a. Dense Captioning Events in Videos, CVPR 2017
Ji, Krishna et al. Action Genome: Actions as Compositions of Spatio-Temporal Scene Graphs, CVPR 2020

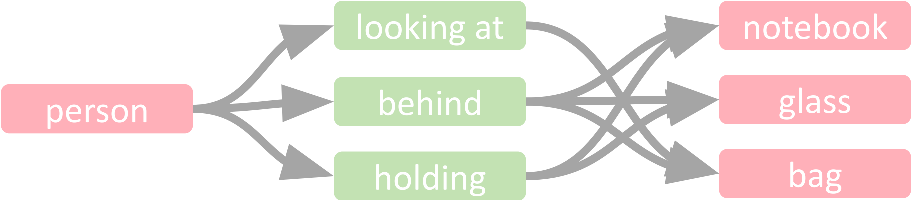
Action Genome: Understanding Action with Spatio-Temporal Scene Graphs

action: take a bag from somewhere

action: drinking from a cup

action: take notebook from somewhere





Krishna et a. Dense Captioning Events in Videos, CVPR 2017

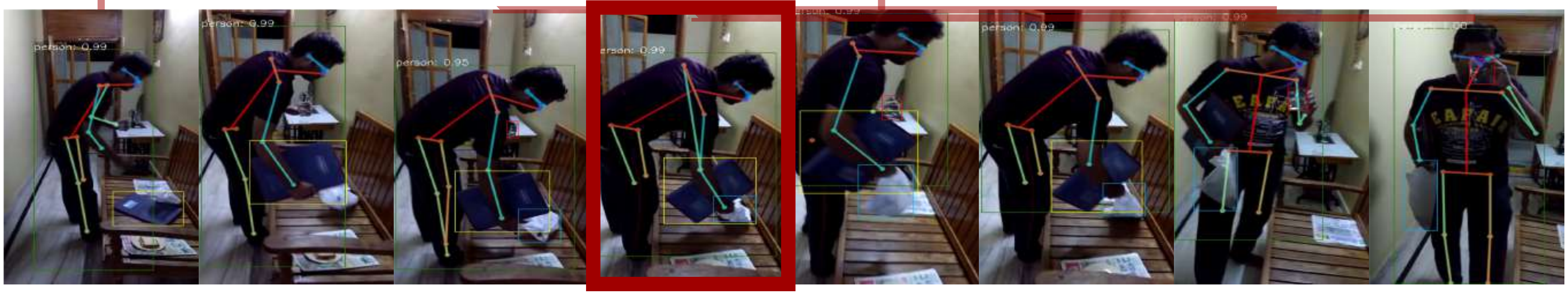
Ji, Krishna et al. Action Genome: Actions as Compositions of Spatio-Temporal Scene Graphs, CVPR 2020

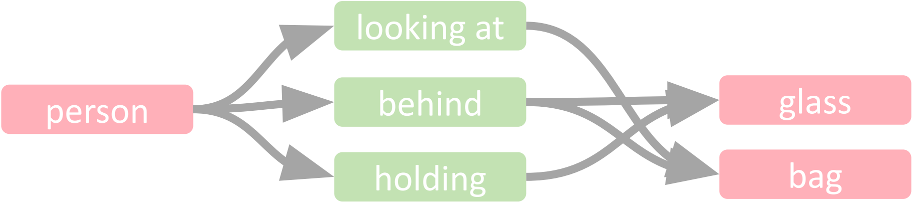
Action Genome: Understanding Action with Spatio-Temporal Scene Graphs

action: take a bag from somewhere

action: drinking from a cup

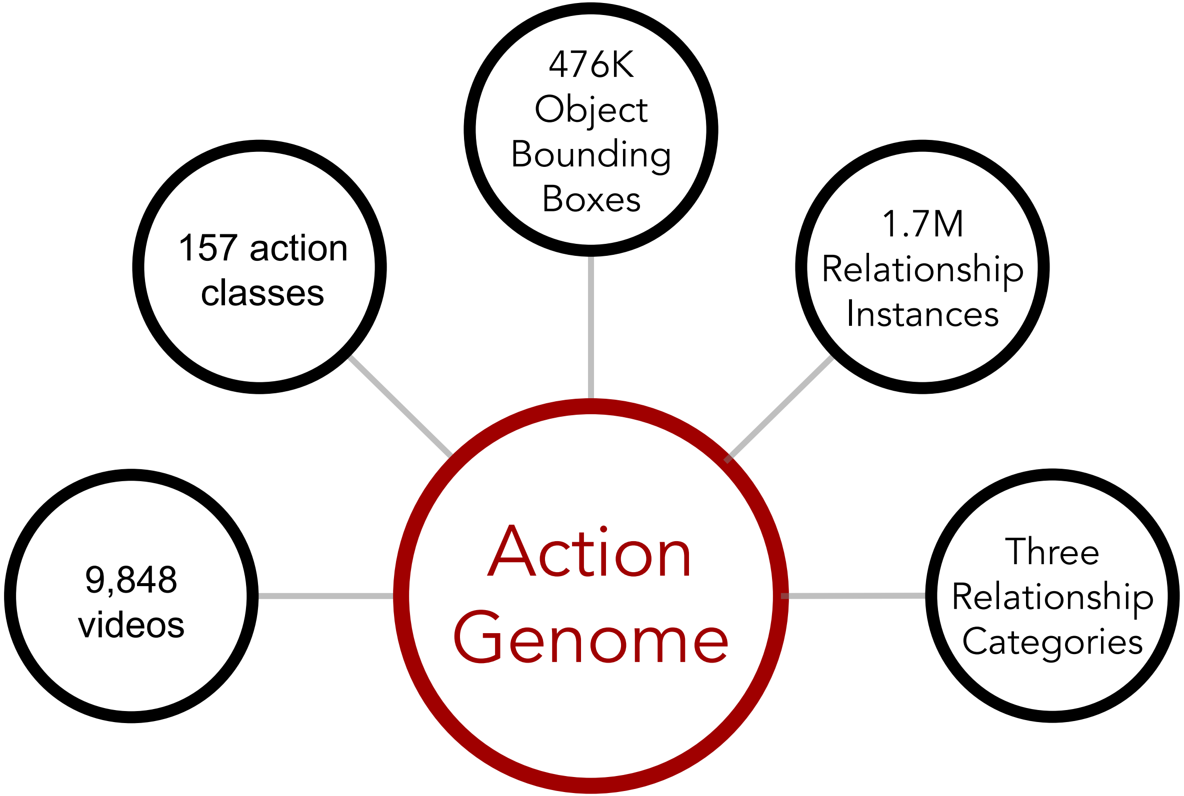
action: take notebook from somewhere





Krishna et a. Dense Captioning Events in Videos, CVPR 2017

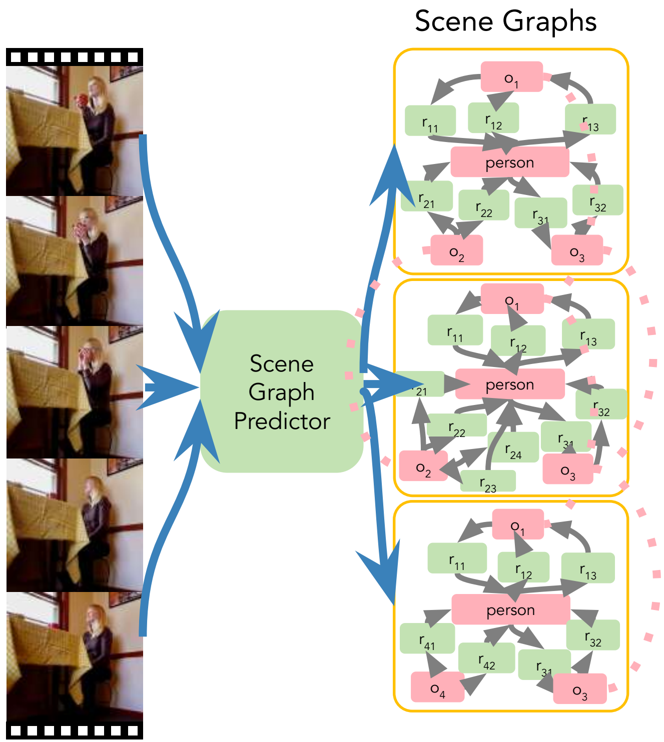
Ji, Krishna et al. Action Genome: Actions as Compositions of Spatio-Temporal Scene Graphs, CVPR 2020



Code and dataset available: http://actiongenome.org

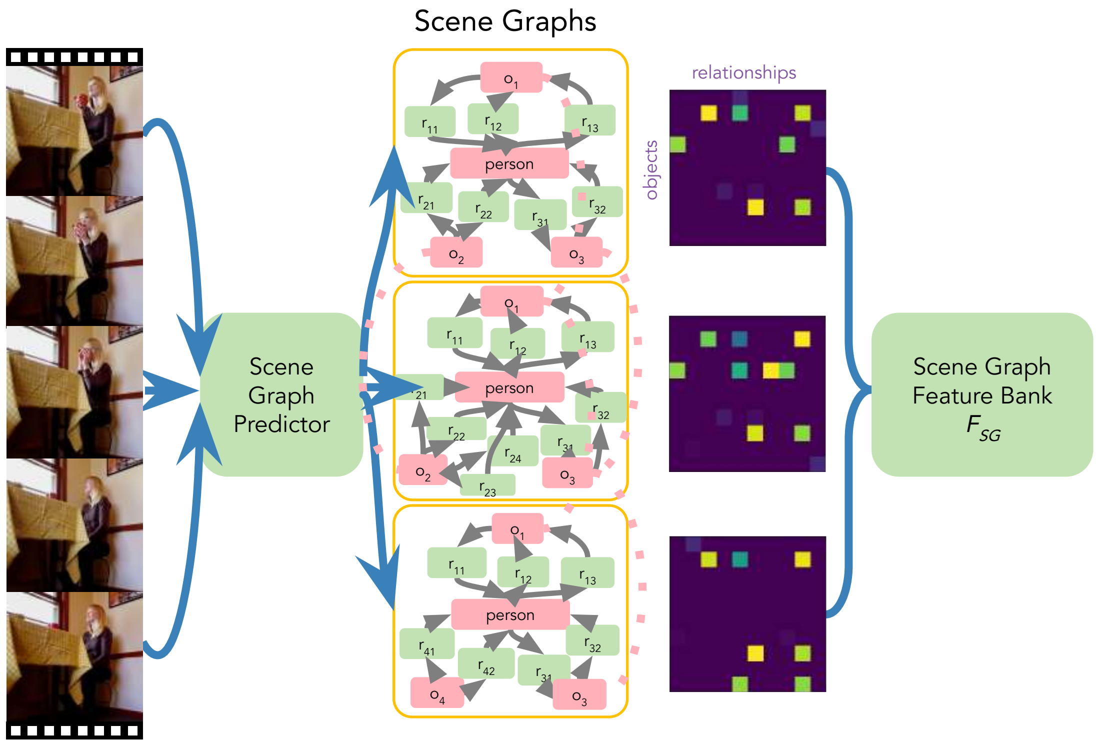
Ji, Krishna et al. Action Genome: Actions as Compositions of Spatio-Temporal Scene Graphs, CVPR 2020

Spatio-temporal Scene Graph Feature Banks (SGFB) for Action Recognition



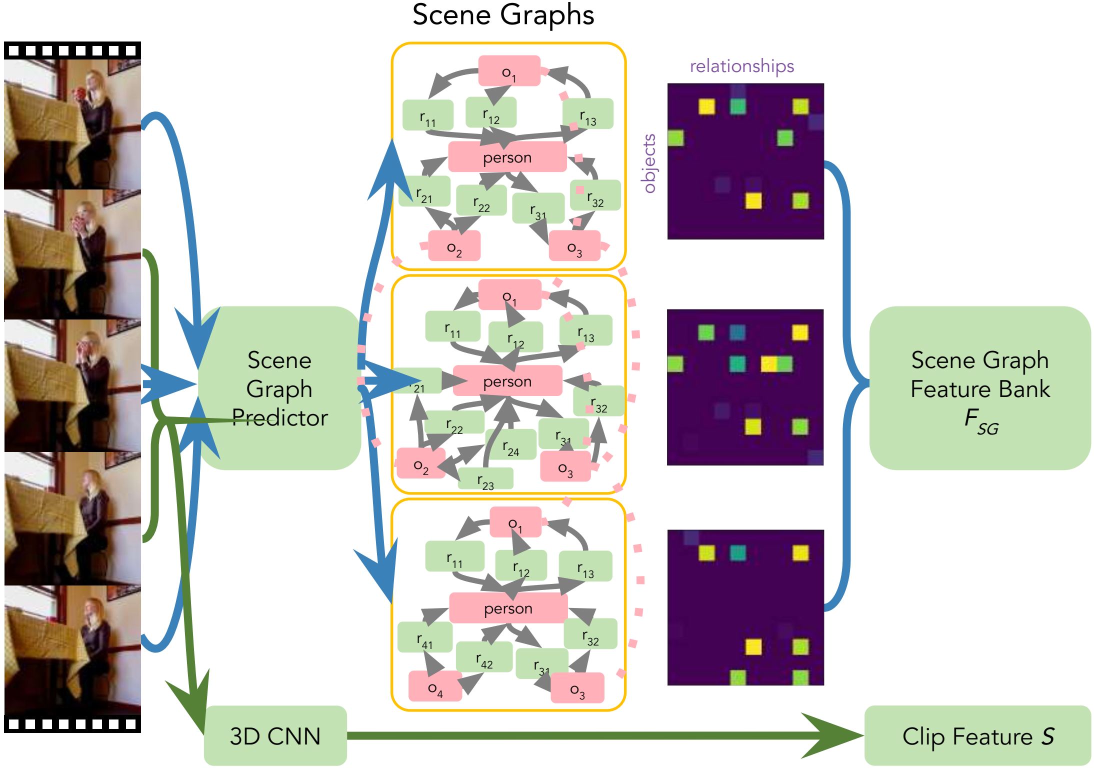
Ji, Krishna et al. Action Genome: Actions as Compositions of Spatio-Temporal Scene Graphs, CVPR 2020

Spatio-temporal Scene Graph Feature Banks (SGFB) for Action Recognition



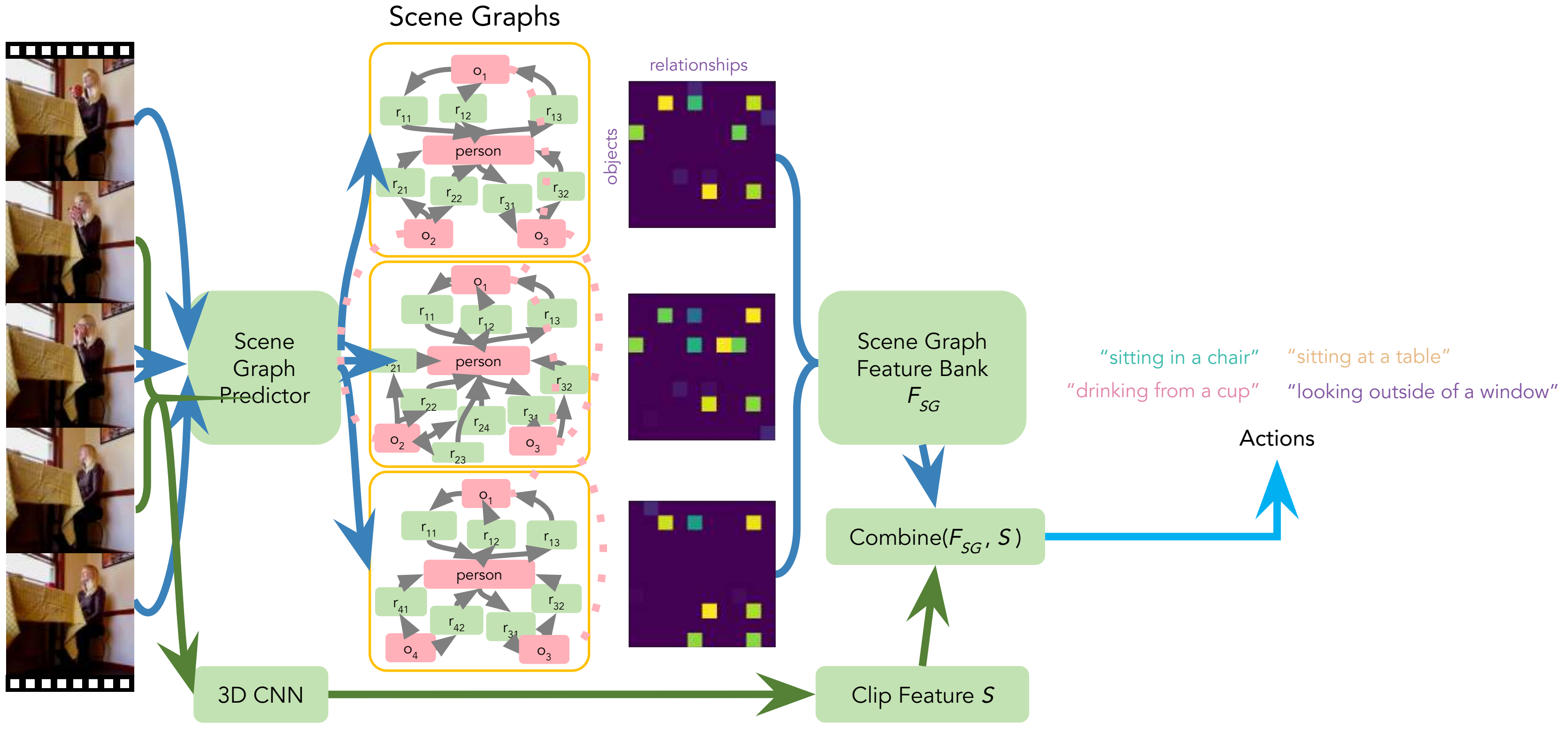
Ji, Krishna et al. Action Genome: Actions as Compositions of Spatio-Temporal Scene Graphs, CVPR 2020

Spatio-temporal Scene Graph Feature Banks (SGFB) for Action Recognition



Ji, Krishna et al. Action Genome: Actions as Compositions of Spatio-Temporal Scene Graphs, CVPR 2020

Spatio-temporal Scene Graph Feature Banks (SGFB) for Action Recognition



Ji, Krishna et al. Action Genome: Actions as Compositions of Spatio-Temporal Scene Graphs, CVPR 2020

From Scene Graphs to Action Recognition

Ground truth action labels:

Lying on a bed,

Awakening in bed,

Holding a pillow

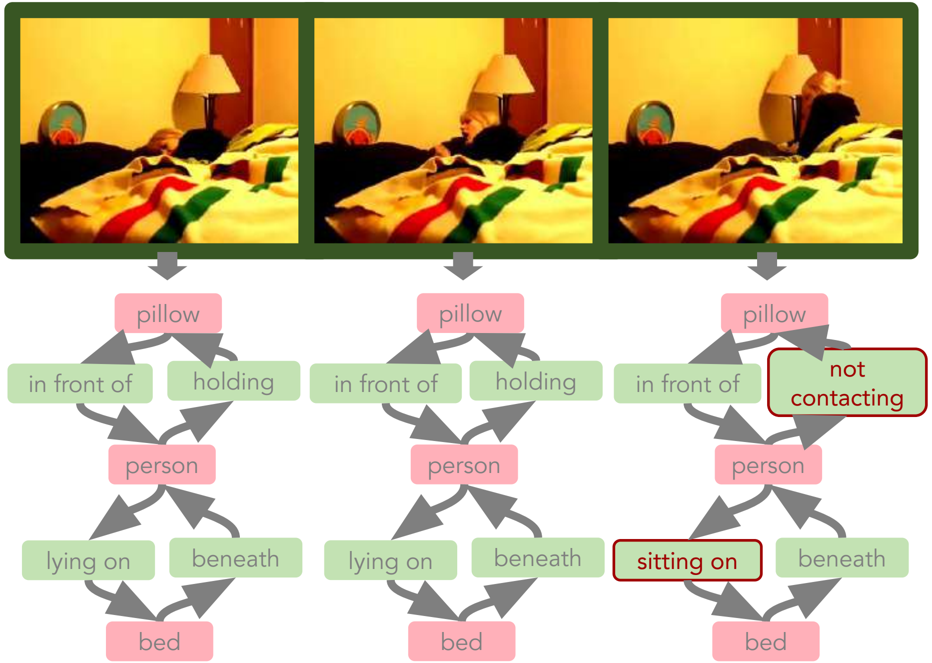
Baselines rely heavily on training set priors

Ground truth:
Lying on a bed,
Awakening in bed,
Holding a pillow

Baseline (LFB) predictions:
Lying on a bed,
Watching television,
Holding a pillow

Wu et al. Long-term feature banks for detailed video understanding, CVPR 2019

Modeling temporal changes in relationships lead to improved inference



Ground truth:
Lying on a bed,
Awakening in bed,
Holding a pillow

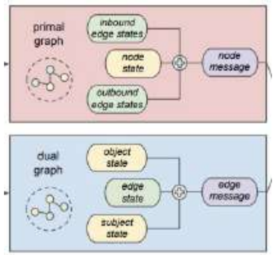
Baseline (LFB) predictions:
Lying on a bed,
Watching television,
Holding a pillow

Our top-3 predictions:
Lying on a bed,
Awakening in bed,
Holding a pillow

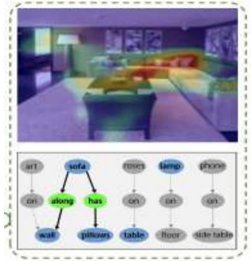
Wu et al. Long-term feature banks for detailed video understanding, CVPR 2019
Ji, Krishna et al. Action Genome: Actions as Compositions of Spatio-Temporal Scene Graphs, CVPR 2020

The community has published hundreds of scene graph papers

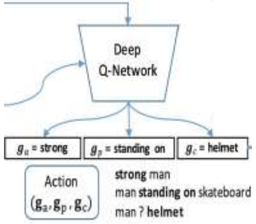
Message passing



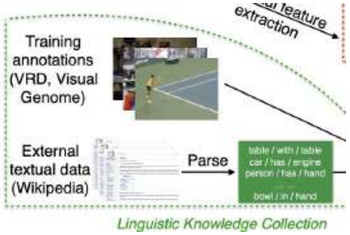
Attention



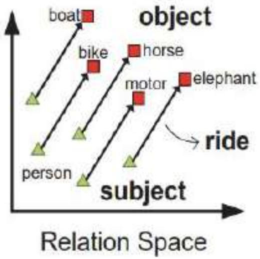
Reinforce



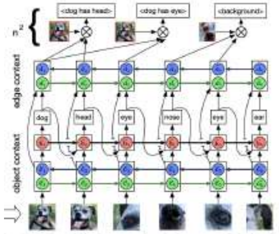
External knowledge



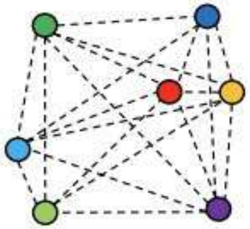
Transformations



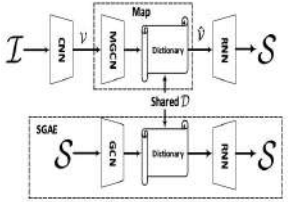
Recurrent networks



Graph Convolutions



Auto-encoders

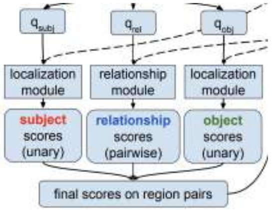


Zellers et al. Neural motifs: Scene graph parsing with global context CVPR. 2018
Yang et al. Graph r-cnn for scene graph generation ECCV 2018
Yang et al. Shuffle-then-assemble: Learning object-agnostic visual relationship features ECCV 2018
Zhang et al. Visual translation embedding network for visual relation detection CVPR 2017
Liang et al. Deep variation-structured reinforcement learning for visual relationship and attribute detection CVPR 2017
Dornadula et al. Visual Relationships as Functions: Enabling Few Shot Scene Graph Generation ICCV SGRL 2019
Xu et al. Scene graph generation by iterative message passing CVPR 2017
Yu et al. Visual relationship detection with internal and external linguistic knowledge distillation ICCV 2017

Scene graphs have achieved state of the art in many tasks

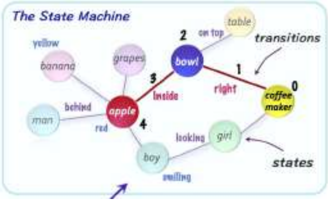
3D scene graphs

Explainable AI

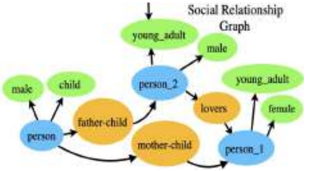


Human intentions

VQA

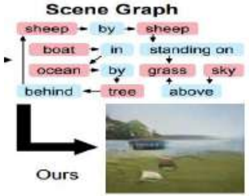


Social relationships



Fashion

Image generation



Program synthesis

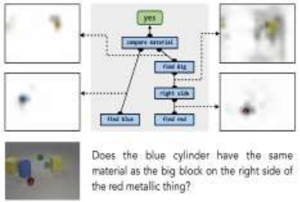
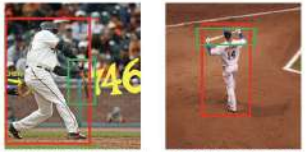
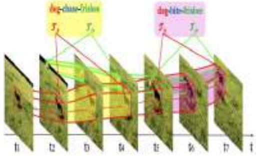


Image captioning



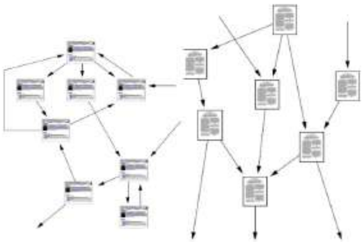
a baseball player swinging a bat   a baseball player holding a bat

Video understanding

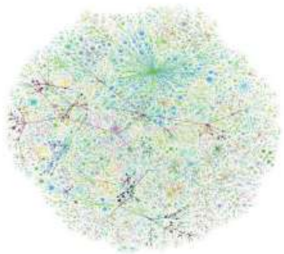


Armeni et al. 3D Scene Graph: A Structure for Unified Semantics, 3D Space, and Camera ICCV 2019
Hu et al. Modeling relationships in referential expressions with compositional modular networks, CVPR 2017
Xu et al. Interact as you intend: Intention-driven human-object interaction detection, Transactions on Multimedia 2019
Hudson et al. Neural State Machine, NeurIPS 2019
Hu, Ronghang, et al. Learning to reason: End-to-end module networks for visual question answering *ICCV* 2017
Johnson et al. Image generation from scene graphs CVPR 2018
Yu et al. Layout-graph reasoning for fashion landmark detection CVPR 2019
Goel et al. An End-to-End Network for Generating Social Relationship Graphs CVPR 2019
Kim et al. Dense relational captioning: Triple-stream networks for relationship-based captioning CVPR 2019
Tsai et al. Video relationship reasoning using gated spatio-temporal energy graph CVPR 2019

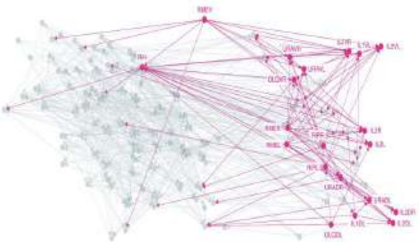
Graphs are everywhere – in numerous fields



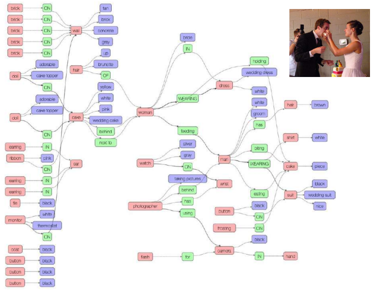
Information networks:
Web & citations



Internet

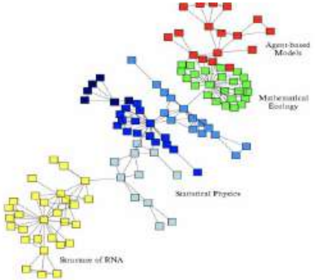


Networks of neurons

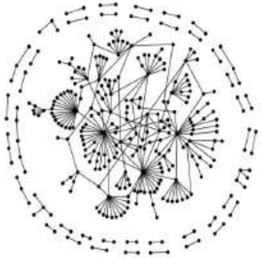


Scene Graphs

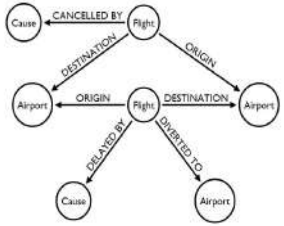
Social networks



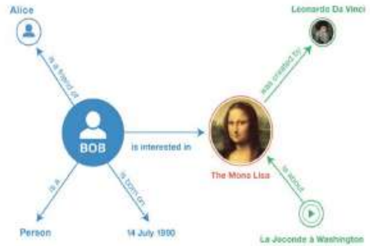
Economic networks



Communication networks



Event Graphs



Knowledge Graphs

Summary

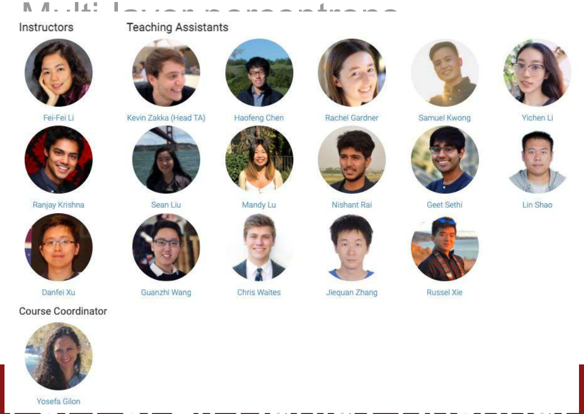
- Scene graphs are a symbolic, compositional, knowledge representation inspired by Cognitive Science and is a common underlying structure in many Computer Vision tasks.
- The task of Scene Graph Generation requires more complex structured prediction models
- GCNs are a generalization of the CNNs you have already learned about.
  - Use them when you work with graph-related data
- This is a relatively new sub-field and there is a lot of work left to do and a lot of promise for future research.

What have we learned this quarter?

Neural Network Fundamentals → Convolutional Neural Networks → Computer Vision Applications

Neural Network Fundamentals

Data-driven learning
Linear classification & kNN
Loss functions
Optimization
Backpropagation
Multi-layer perceptrons



Convolutional Neural Networks

Convolutions
Pytorch 1.4 / Tensorflow 2.0
Activation functions
Batch normalization
Transfer learning
Data augmentation
Momentum / RMSProp / Adam
Architecture design

Computer Vision Applications

RNNs / LSTMs
Attention & Transformers
Image captioning
Interpreting neural networks
Style transfer
Adversarial examples
NeRF
Scene graphs
Graph Convolutions
Self-supervised learning
Multimodal learning
Perception & Action

Instructors

Teaching Assistants

Course Coordinator


Fei-Fei Li


Kevin Zakka (Head TA)


Haofeng Chen


Rachel Gardner


Samuel Kwong


Yichen Li


Yosefa Gilon


Ranjay Krishna


Sean Liu


Mandy Lu


Nishant Rai


Geet Sethi


Lin Shao


Danfei Xu


Guanzhi Wang


Chris Waites


Jiequan Zhang


Russel Xie