RECENT ADVANCES IN IMAGE CAPTIONING, IMAGE-TEXT RETRIEVAL AND VISUAL QUESTION ANSWERING USING SCENE GRAPH PARSING, WHAT NEXT?

HAMID PALANGI

DEEP LEARNING GROUP, MICROSOFT RESEARCH AI

July 09, 2019 at MSR AI Seminar, Redmond, US

Acknowledgements: Xi (Stephen) Chen, Kuang-Huei Lee, Houdong Hu, Jianfeng Gao, Lei Zhang, Qiuyuan Huang, Alex Polozov, Pengchuan Zhang, Asli Celikyilmaz, Xiujun Li, Kenneth Tran, Prithvijit Chattopadhyay, Yichen Huang

- [2012]

Building High-level Features
Using Large Scale Unsupervised Learning

Quoc V. Le QUOCLE@CS.STANFORD.EDU
Marc'Aurelio Ranzato RANZATO@GOOGLE.COM
Rajat Monga RAJATMONGA@GOOGLE.COM
Matthieu Devin MDEVIN@GOOGLE.COM
Kai Chen KAICHEN@GOOGLE.COM
Greg S. Corrado GCORRADO@GOOGLE.COM
Jeff Dean JEFF@GOOGLE.COM
Andrew Y. Ng ANG@CS.STANFORD.EDU

2012

From https://arxiv.org/abs/1112.6209

From https://www.wired.com/2012/06/google-x-neural-network/

2,000 CPUs (16,000 cores) – 600 kWatts - $5,000,000

- [2013]

3 GPUs (18,432 cores) – 4 kWatts - $33,000

DANIELA HERNANDEZ BUSINESS 06.17.13 6:30 AM

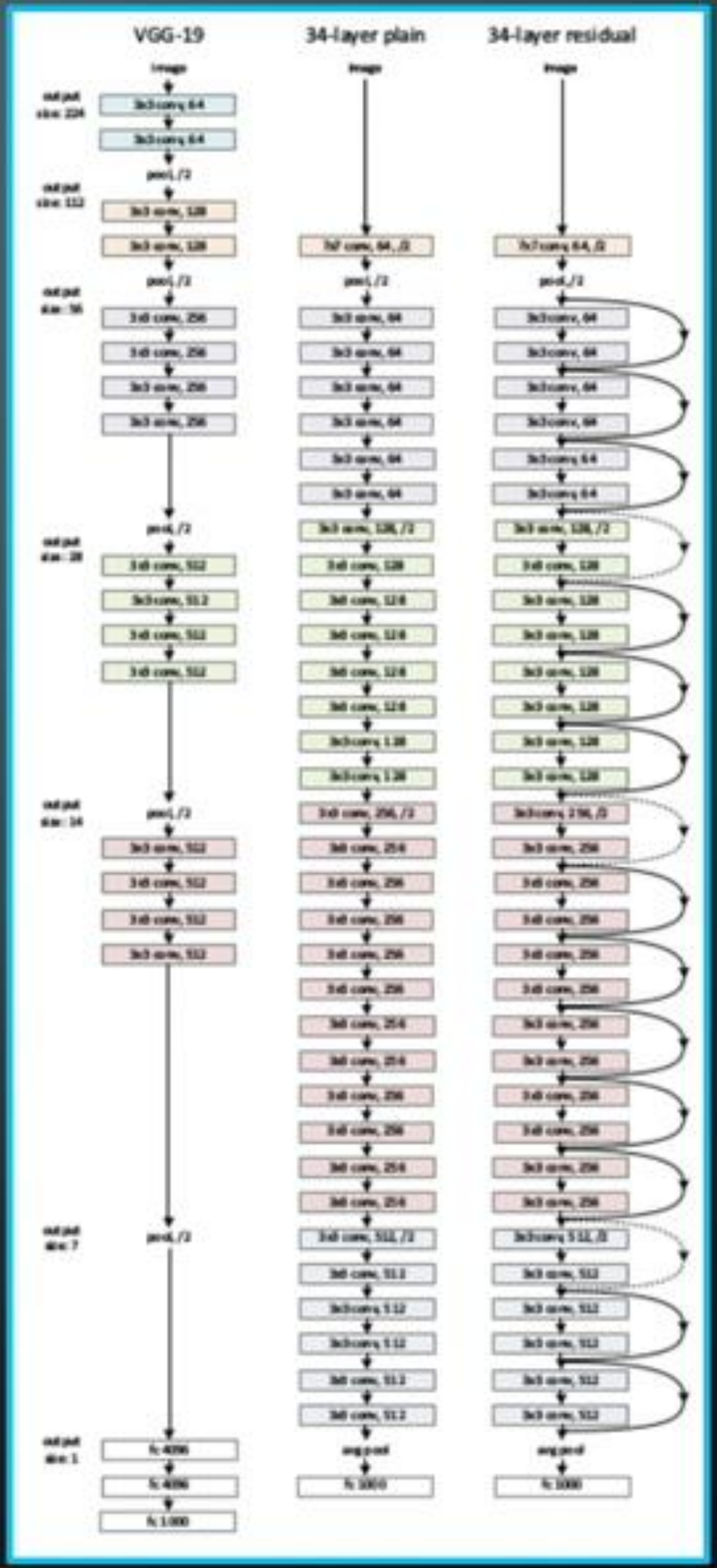
NOW YOU CAN BUILD GOOGLE'S ARTIFICIAL BRAIN ON THE CHEAP

Andrew Ng. *Photo: Ariel Zambelich/Wired*

From https://www.wired.com/2013/06/andrew_ng/

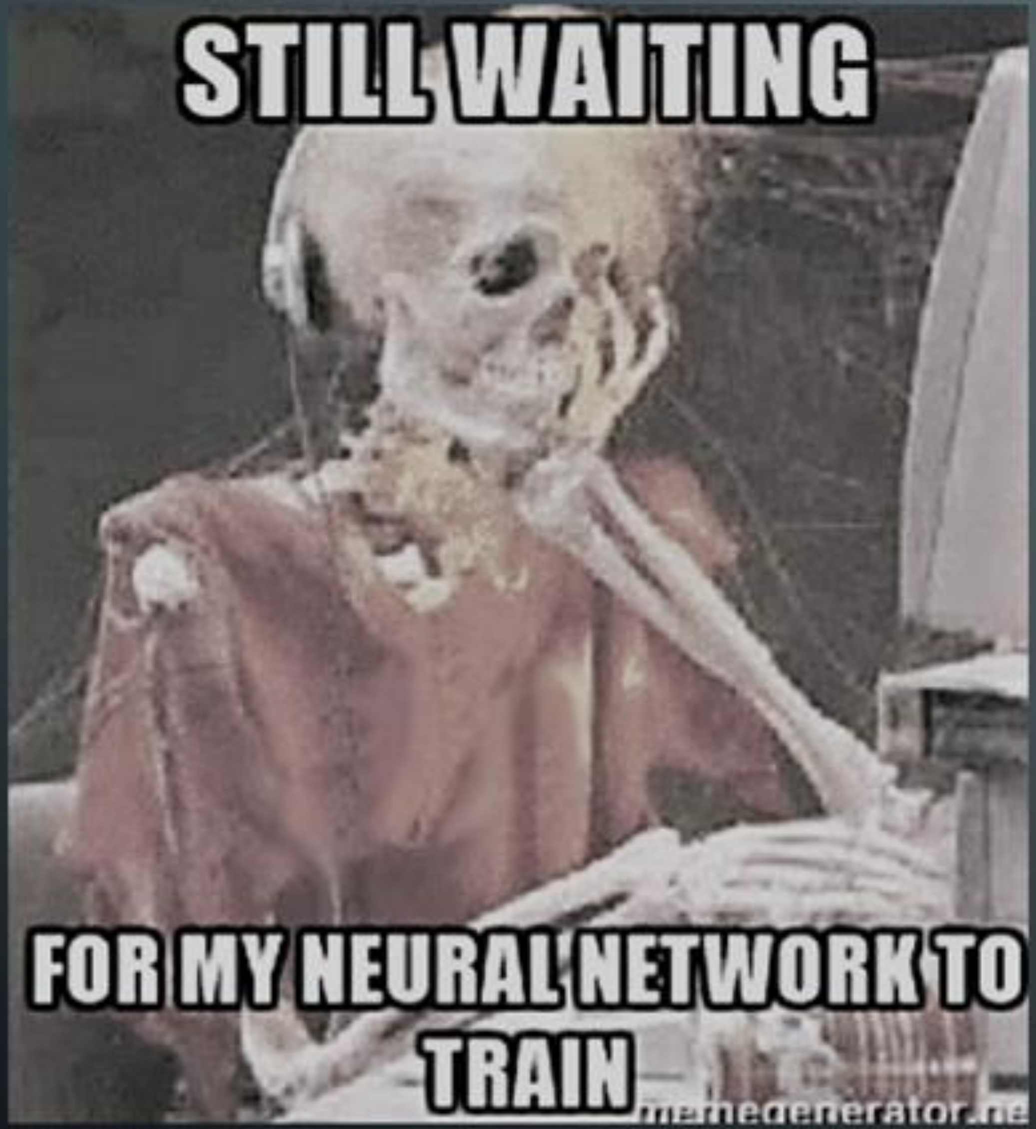
LENET TO ALEXNET TO RESNET

Picture from https://www.d2l.ai/chapter_convolutional-modern/alexnet.html



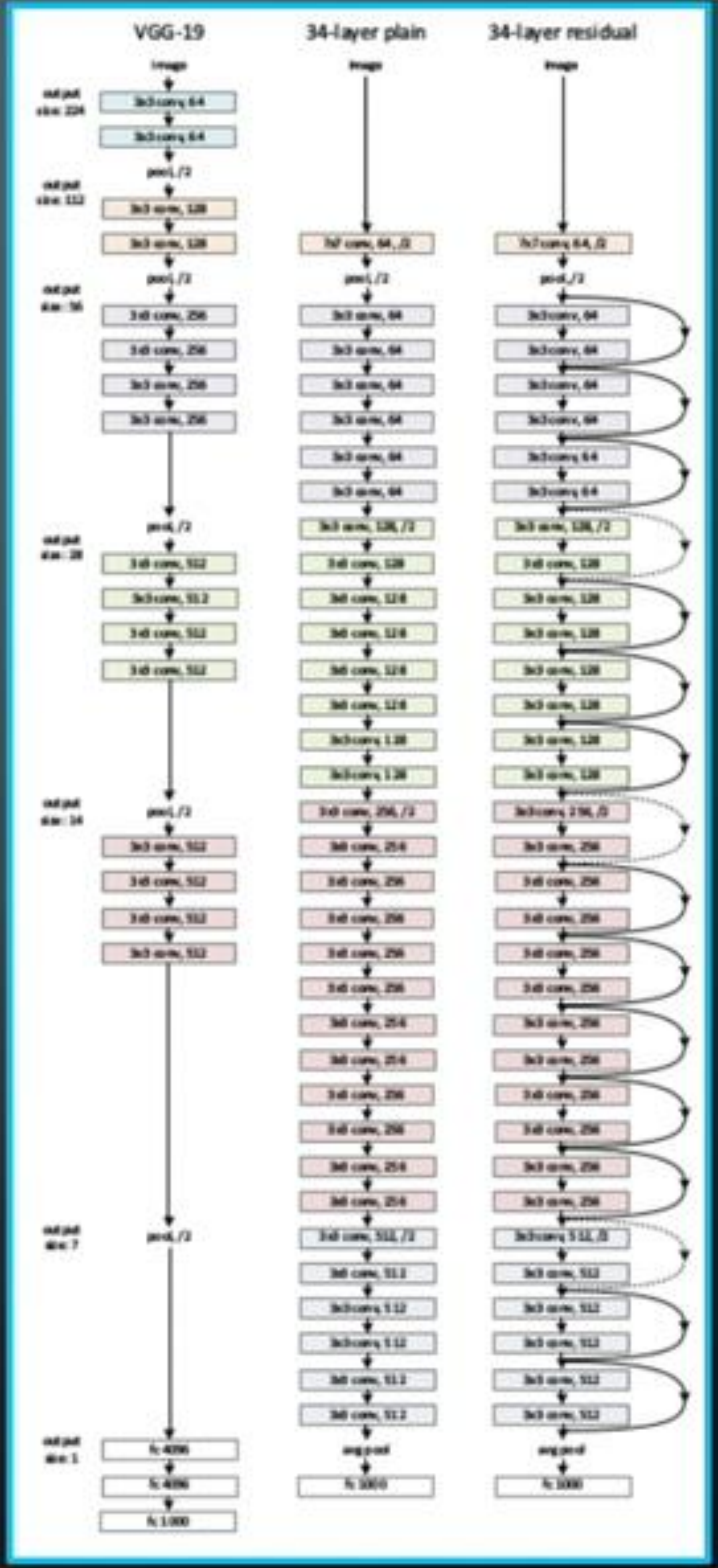
Picture from https://medium.com/@14prakash/understanding-and-implementing-architectures-of-resnet-and-resnext-for-state-of-the-art-image-cf51669e1624

LENET TO ALEXNET TO RESNET



Picture from https://www.analyticsvidhya.com/blog/2017/05/gpus-necessary-for-deep-learning/

Picture from https://www.d2l.ai/chapter_convolutional-modern/alexnet.html



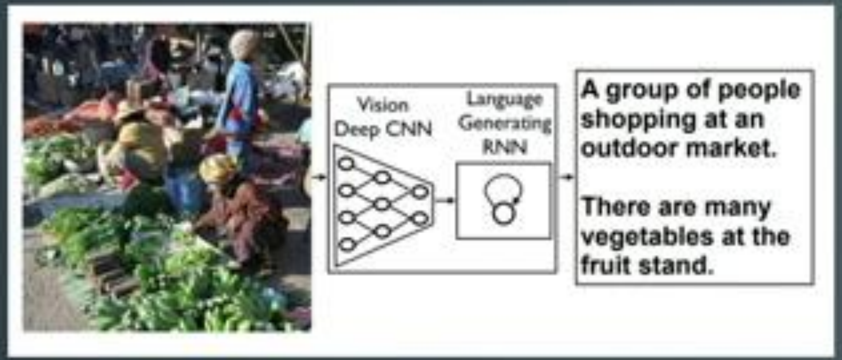
Picture from https://medium.com/@14prakash/understanding-and-implementing-architectures-of-resnet-and-resnext-for-state-of-the-art-image-cf51669e1624

Task: Captioning

GOOGLE,2014

Show and Tell: A Neural Image Caption Generator

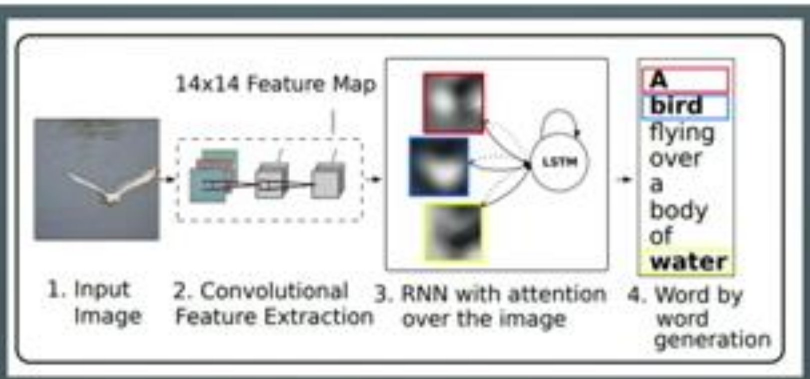
Oriol Vinyals, Google, vinyals@google.com
Alexander Toshev, Google, toshev@google.com
Samy Bengio, Google, bengio@google.com
Dumitru Erhan, Google, dumitru@google.com



U MONTREAL & U of T,2015

Show, Attend and Tell: Neural Image Caption Generation with Visual Attention

Kelvin Xu KELVIN.XU@UMONTREAL.CA
Jimmy Lei Ba JIMMY@PSI.UTORONTO.CA
Ryan Kiros RKIROS@CS.TORONTO.EDU
Kyunghyun Cho KYUNGHYUN.CHO@UMONTREAL.CA
Aaron Courville AARON.COURVILLE@UMONTREAL.CA
Ruslan Salakhutdinov RSALAKHU@CS.TORONTO.EDU
Richard S. Zemel ZEMEL@CS.TORONTO.EDU
Yoshua Bengio FIND-ME@THE.WEB



MSFT,2014

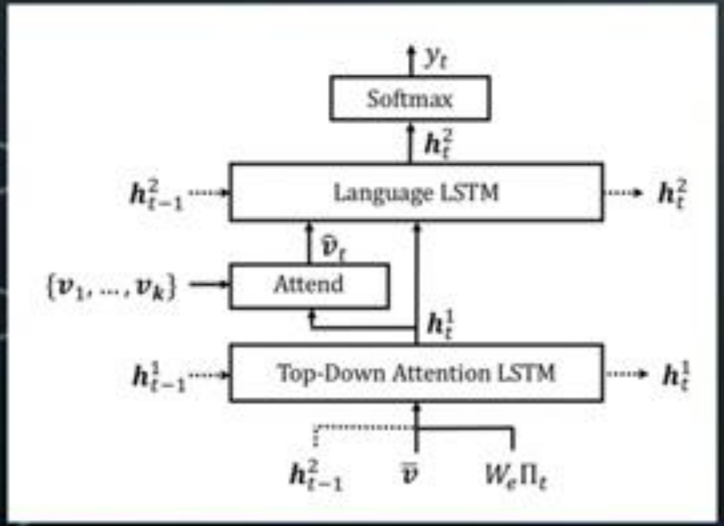
From Captions to Visual Concepts and Back

Hao Fang* Saurabh Gupta* Forrest Iandola* Rupesh K. Srivastava*
Li Deng Piotr Dollár‡ Jianfeng Gao Xiaodong He
Margaret Mitchell John C. Platt‡ C. Lawrence Zitnick Geoffrey Zweig
Microsoft Research

MSFT,2017

Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering

Peter Anderson[1]* Xiaodong He[2] Chris Buehler[3] Damien Teney[4]
Mark Johnson[5] Stephen Gould[1] Lei Zhang[3]
[1]Australian National University [2]JD AI Research [3]Microsoft Research [4]University of Adelaide [5]Macquarie University
[1]firstname.lastname@anu.edu.au, [2]xiaodong.he@jd.com, [3]{chris.buehler,leizhang}@microsoft.com
[4]damien.teney@adelaide.edu.au, [5]mark.johnson@mq.edu.au



In this effort we were (are) planning to cover four directions:

1. What is a good way to represent images to get better IR and Captioning performance (e.g., using scene graphs or other structured representations)?
2. How to design an effective alignment model that can capture the relevance between image and text (e.g., using attention over relations in the scene graph)?
3. How huge weakly supervised data from a search engine like Bing can help to improve the structured representation of image (e.g., to design/train better scene graph generation methods/models). Here weak supervision means clickthrough data, a user uploads an image and clicks on a webpage (image, webpage title pair), or a user inserts a text query and clicks on an image (query, image pair).
4. Visual Grounding and Reasoning

Step 0: Covers 1 and 2 above

Step 1: Will cover 3 above

Step 2: Will cover 4 above, visual grounding deserves to spend some time to design specific models for it so that the models do not only rely on simple statistics of the given (usually limited) dataset.

OUTLINE

- Scene Graph Generation (SGG)
- Image-Text Retrieval
- Image Captioning
- Weakly supervised SGG
- Visual Question Answering
- Challenges and Opportunities

STEP 0

STEP 1

STEP 2 & beyond

OUTLINE

- Scene Graph Generation (SGG)
- Image-Text Retrieval
- Image Captioning

STEP 0

Exploring Visual Relations for Image-Text Matching

Kuang-Huei Lee * Hamid Palangi * Xi Chen Houdong Hu Jianfeng Gao

Microsoft AI and Research

Step 0: Exploring Visual Relations for Image-Text Matching

Task:

Scene Graph Generation (SGG):

1. **PredCLS:** Predicate classification given (source,target) objs

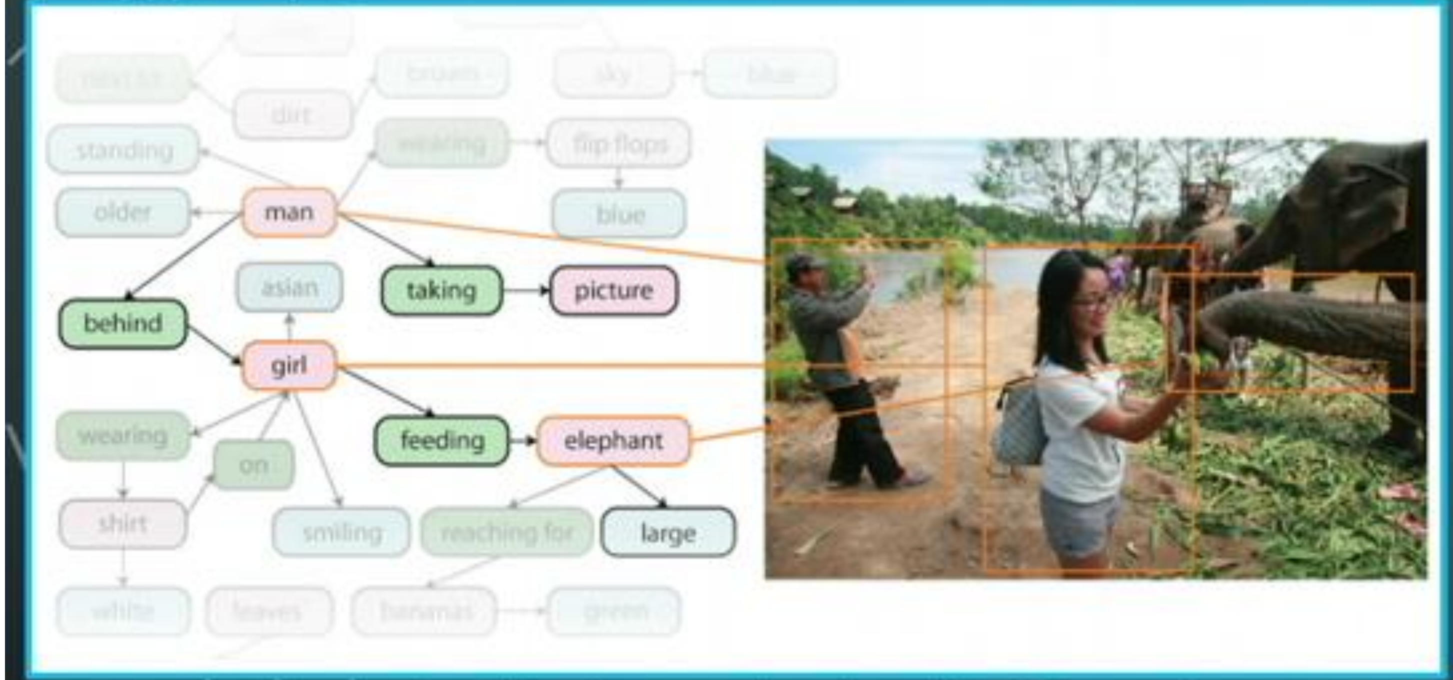


Figure from https://visualgenome.org/static/paper/Visual_Genome.pdf

Step 0: Exploring Visual Relations for Image-Text Matching

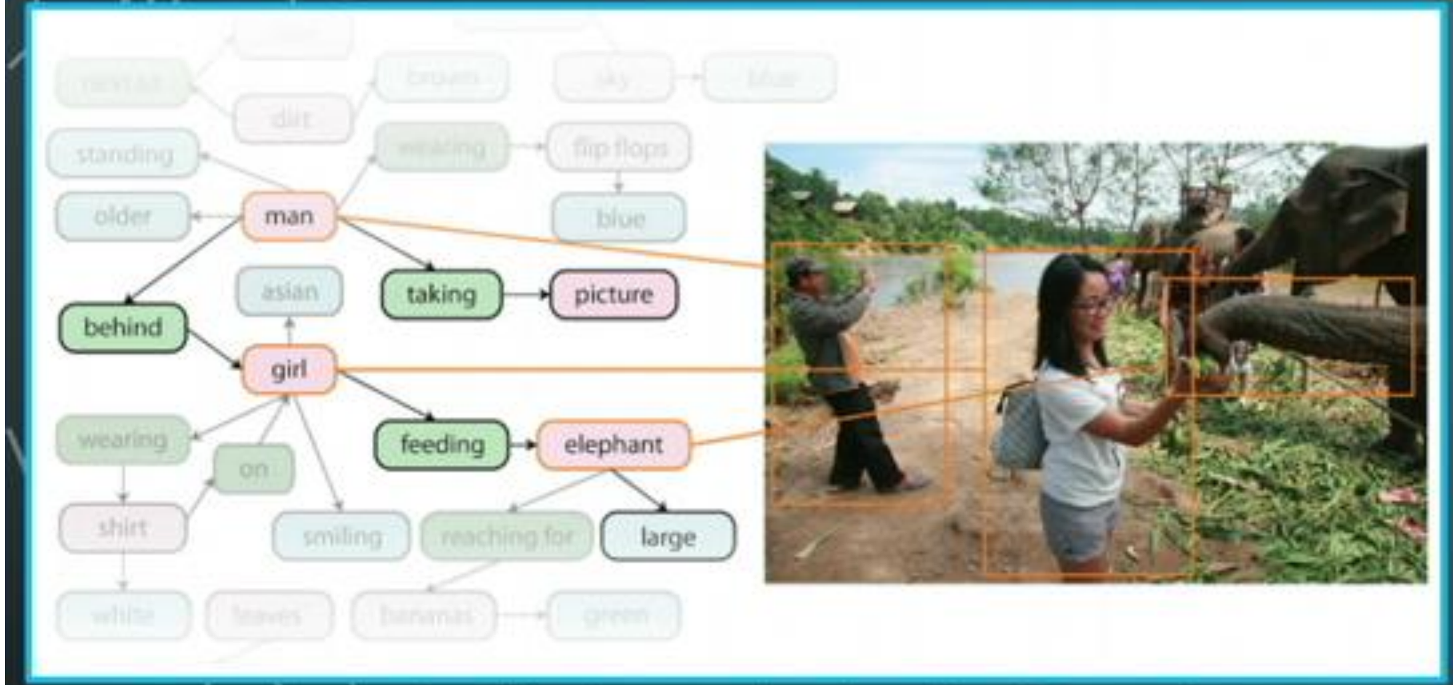


Figure from https://visualgenome.org/static/paper/Visual_Genome.pdf

Task:

Scene Graph Generation (SGG):

1. **PredCLS:** Predicate classification given (source,target) objs
2. **SgCLS:** Both obj classification and predicate classification "given" the ground truth bounding boxes

Step 0: Exploring Visual Relations for Image-Text Matching

Task:

Scene Graph Generation (SGG):

1. **PredCLS:** Predicate classification given (source,target) objs
2. **SgCLS:** Both obj classification and predicate classification "given" the ground truth bounding boxes
3. **SgDET:** Detecting bboxes using a backend (e.g., Faster R-CNN), predicting obj classes and predicate classes

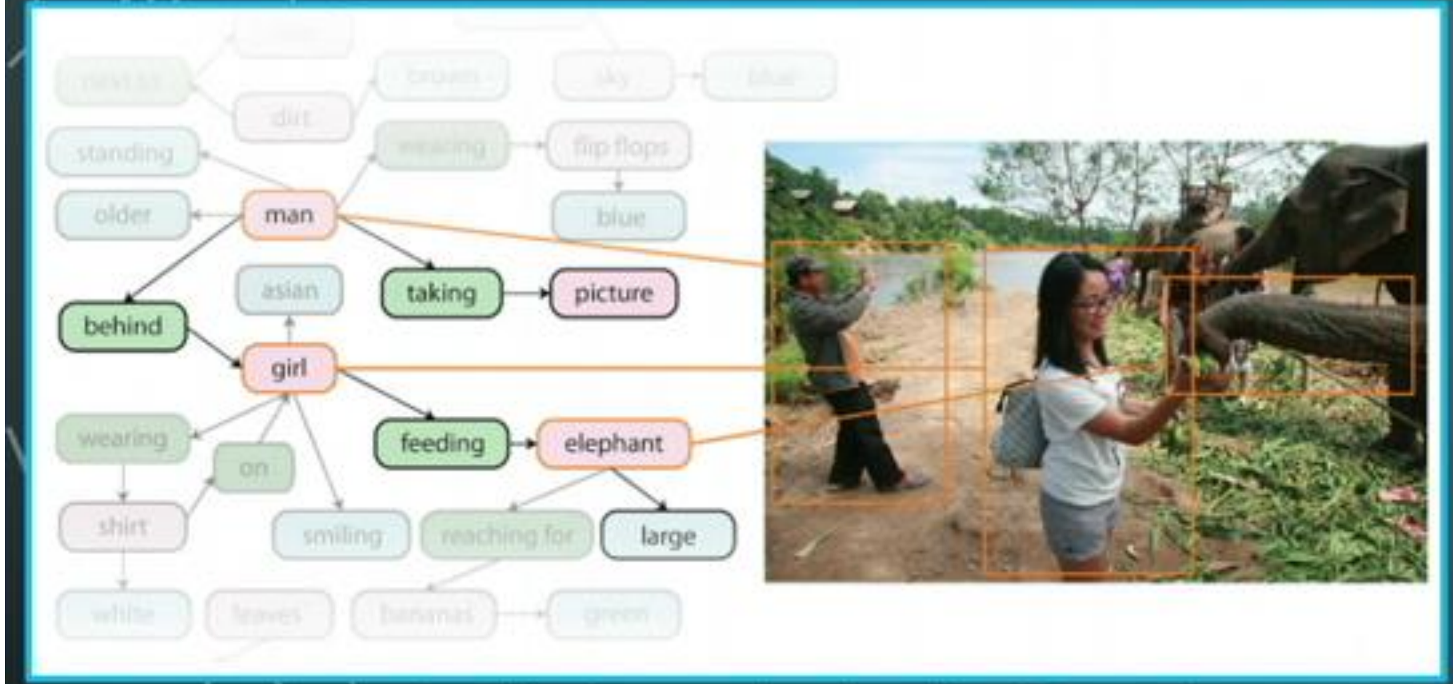


Figure from https://visualgenome.org/static/paper/Visual_Genome.pdf

Step 0: Exploring Visual Relations for Image-Text Matching

Figure from https://visualgenome.org/static/paper/Visual_Genome.pdf

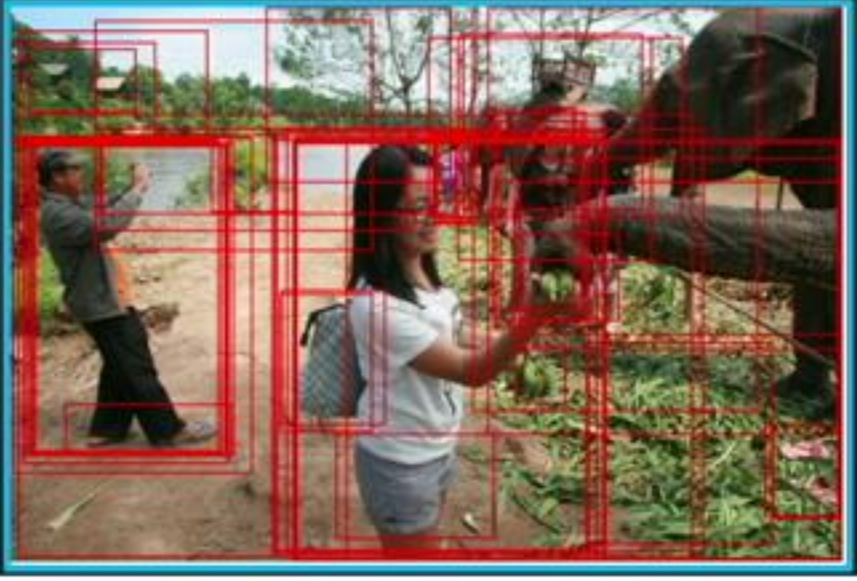
Task:

Scene Graph Generation (SGG):

1. **PredCLS:** Predicate classification given (source,target) objs
2. **SgCLS:** Both obj classification and predicate classification "given" the ground truth bounding boxes
3. **SgDET:** Detecting bboxes using a backend (e.g.,Faster R-CNN), predicting obj classes and predicate classes

Datasets:

Several datasets to address some of above tasks individually, the most popular one is Visual Genome.



Step 0: Exploring Visual Relations for Image-Text Matching

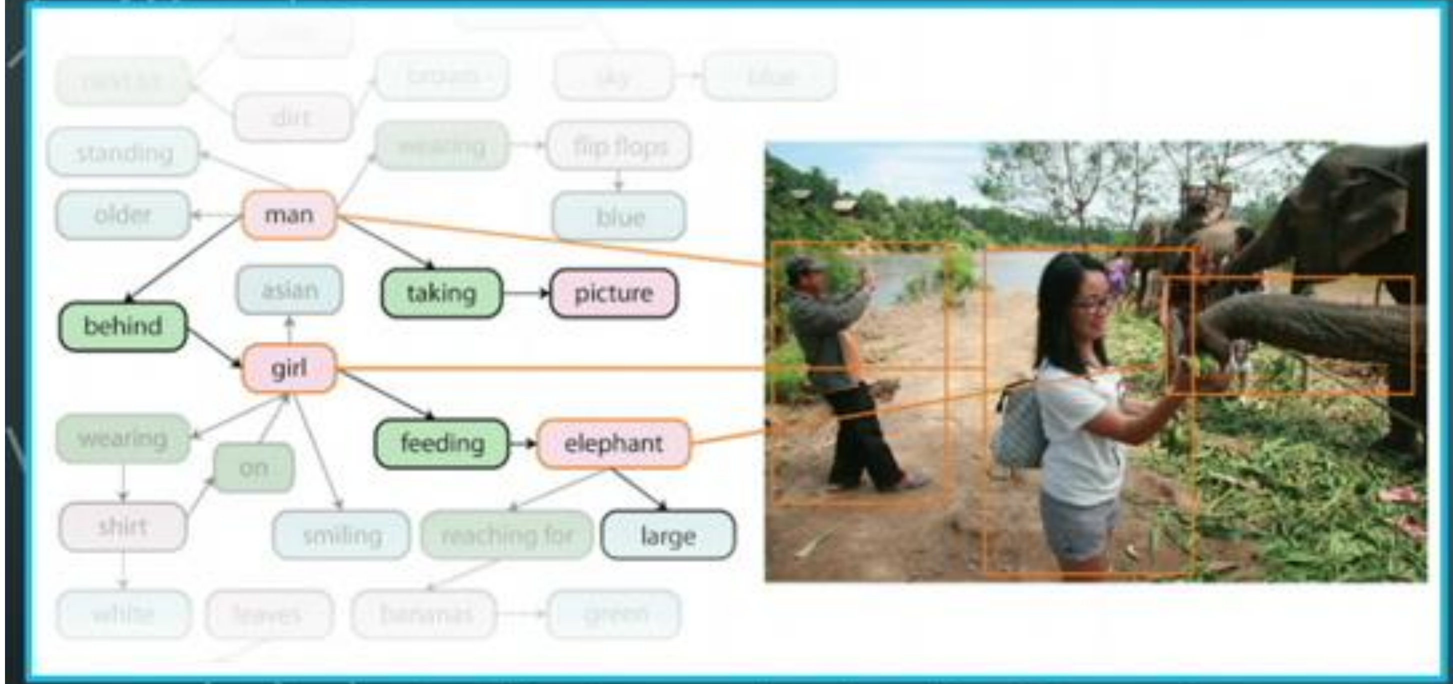


Figure from https://visualgenome.org/static/paper/Visual_Genome.pdf

Task:

Scene Graph Generation (SGG):

1. **PredCLS:** Predicate classification given (source,target) objs
2. **SgCLS:** Both obj classification and predicate classification "given" the ground truth bounding boxes
3. **SgDET:** Detecting bboxes using a backend (e.g.,Faster R-CNN), predicting obj classes and predicate classes

Datasets:

Several datasets to address each of above tasks, the most popular one is visual genome.

Methods:

Various methods proposed including iterative message passing from Stanford, Neural Motifs from UW, etc (A complete up to date list http://picdataset.com/challenge/paper_list/)

Step 0: Exploring Visual Relations for Image-Text Matching

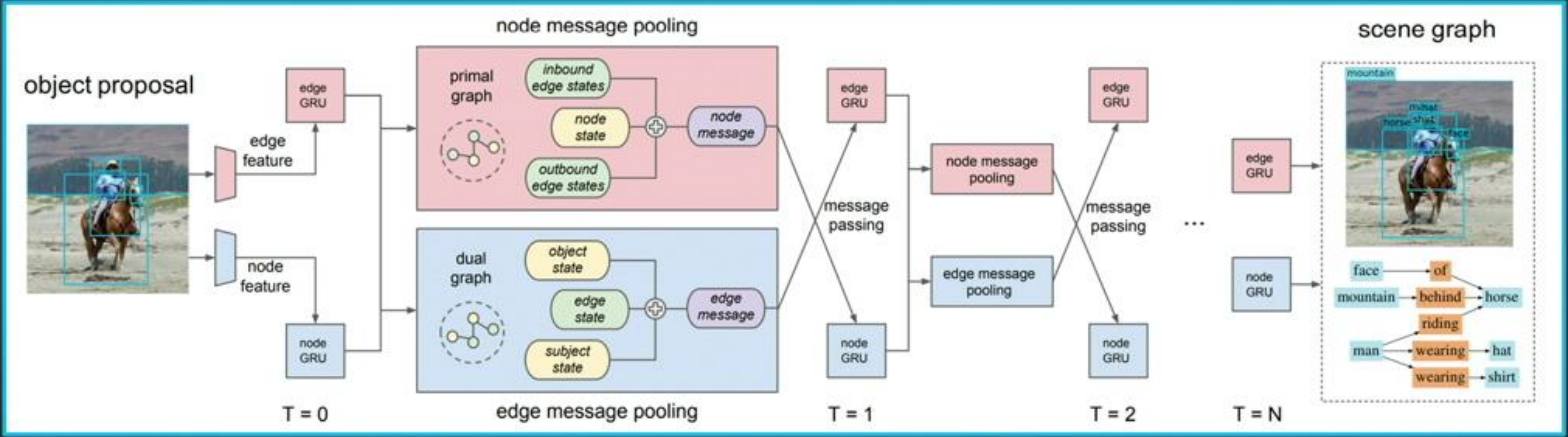


Figure from https://arxiv.org/pdf/1701.02426.pdf

Step 0: Exploring Visual Relations for Image-Text Matching

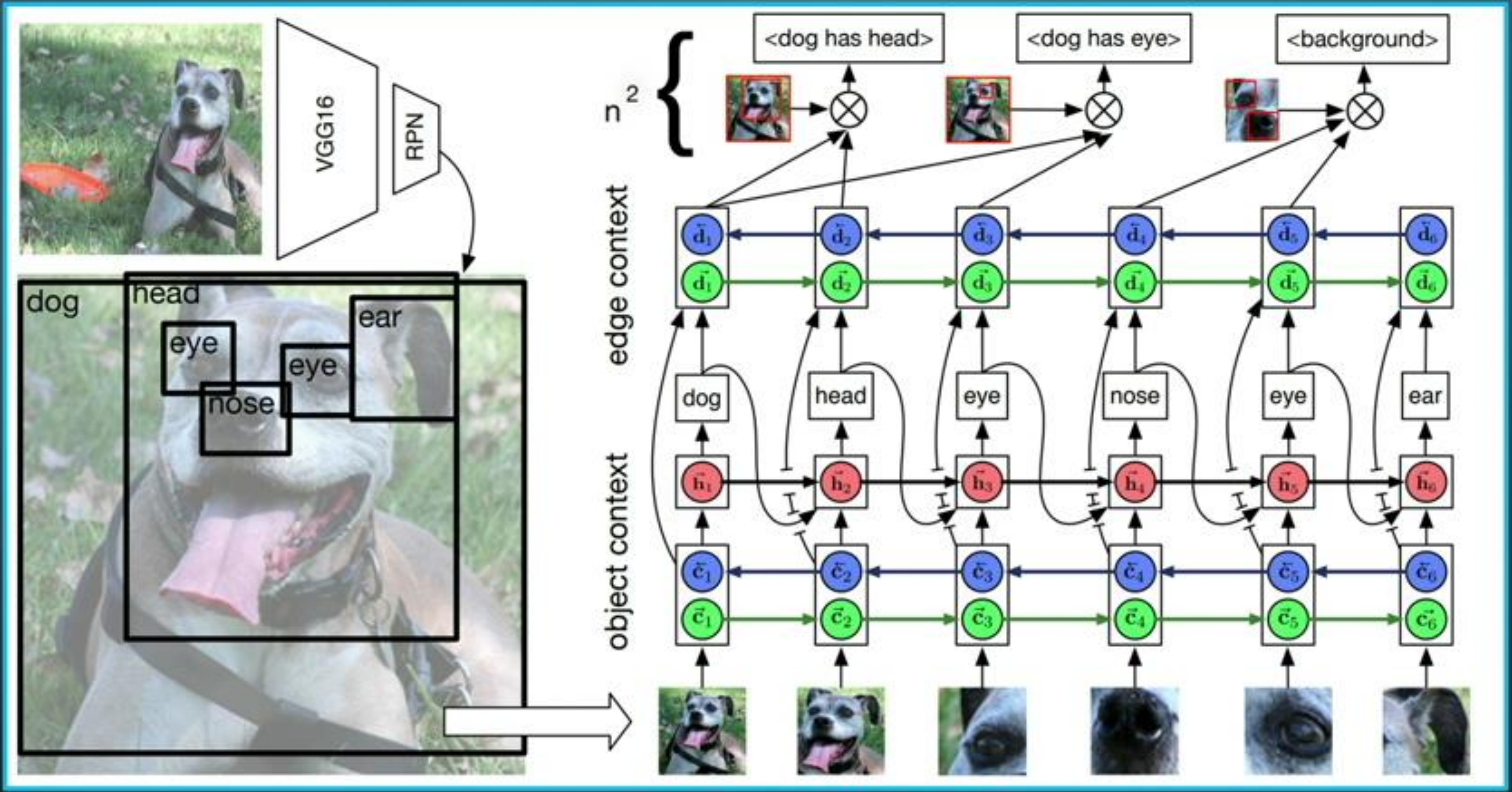
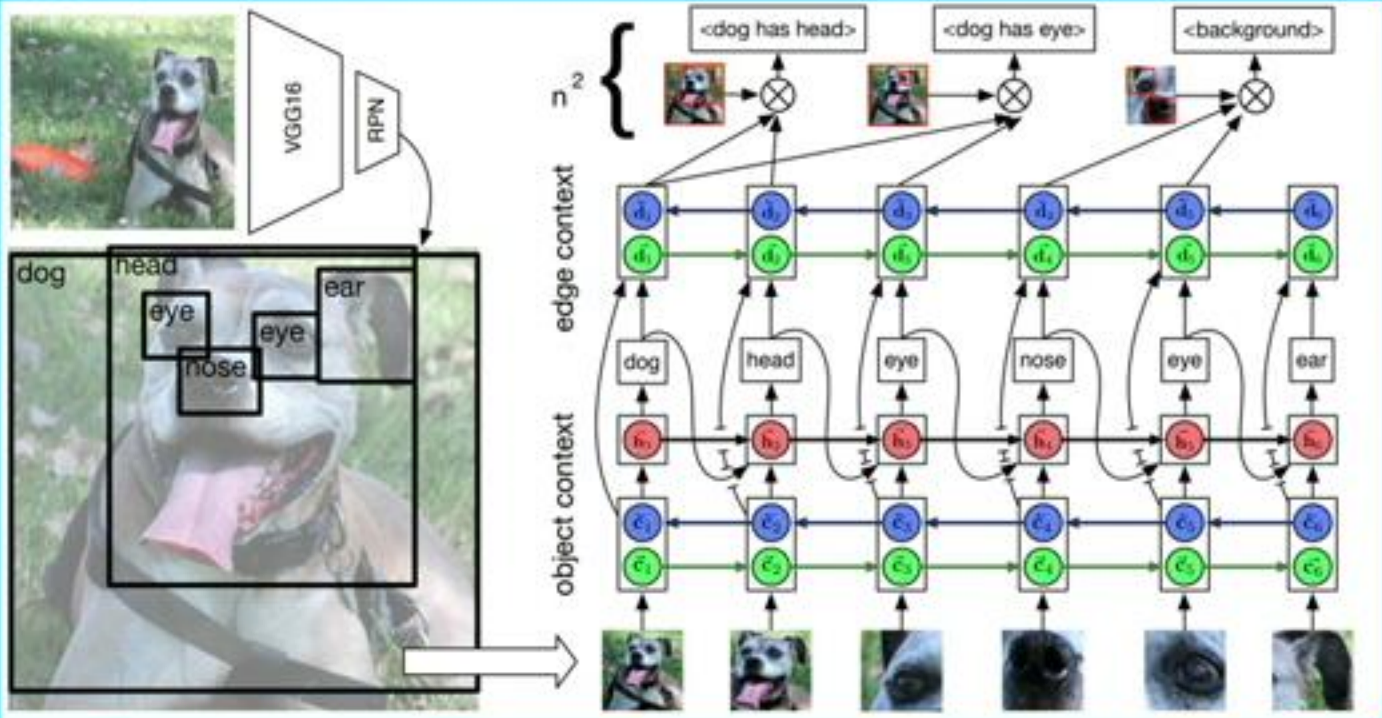


Figure from https://arxiv.org/pdf/1711.06640.pdf

Step 0: Exploring Visual Relations for Image-Text Matching



	Model	Scene Graph Detection R@20	Scene Graph Detection R@50	Scene Graph Detection R@100	Scene Graph Classification R@20	Scene Graph Classification R@50	Scene Graph Classification R@100	Predicate Classification R@20	Predicate Classification R@50	Predicate Classification R@100	Mean
models	VRD [29]		0.3	0.5		11.8	14.1		27.9	35.0	14.9
models	MESSAGE PASSING [47]		3.4	4.2		21.7	24.4		44.8	53.0	25.3
models	MESSAGE PASSING+	14.6	20.7	24.5	31.7	34.6	35.4	52.7	59.3	61.3	39.3
models	ASSOC EMBED [31]*	6.5	8.1	8.2	18.2	21.8	22.6	47.9	54.1	55.4	28.3
models	FREQ	17.7	23.5	27.6	27.7	32.4	34.0	49.4	59.9	64.1	40.2
models	FREQ+OVERLAP	20.1	26.2	30.1	29.3	32.3	32.9	53.6	60.6	62.2	40.7
models	MOTIFNET-LEFTRIGHT	21.4	27.2	30.3	**32.9**	**35.8**	**36.5**	**58.5**	**65.2**	**67.1**	**43.6**
ablations	MOTIFNET-NOCONTEXT	21.0	26.2	29.0	31.9	34.8	35.5	57.0	63.7	65.6	42.4
ablations	MOTIFNET-CONFIDENCE	**21.7**	**27.3**	**30.5**	32.6	35.4	36.1	58.2	65.1	67.0	43.5
ablations	MOTIFNET-SIZE	21.6	**27.3**	30.4	32.2	35.0	35.7	58.0	64.9	66.8	43.3
ablations	MOTIFNET-RANDOM	21.6	**27.3**	30.4	32.5	35.5	36.2	58.1	65.1	66.9	43.5

Type	Examples	Classes	Instances
Entities			
Part	arm, tail, wheel	32	200k (25.2%)
Artifact	basket, fork, towel	34	126k (16.0%)
Person	boy, kid, woman	13	113k (14.3%)
Clothes	cap, jean, sneaker	16	91k (11.5%)
Vehicle	airplane, bike, truck,	12	44k (5.6%)
Flora	flower, plant, tree	3	44k (5.5%)
Location	beach, room, sidewalk	11	39k (4.9%)
Furniture	bed, desk, table	9	37k (4.7%)
Animal	bear, giraffe, zebra	11	30k (3.8%)
Structure	fence, post, sign	3	30k (3.8%)
Building	building, house	2	24k (3.1%)
Food	banana, orange, pizza	6	13k (1.6%)
Relations			
Geometric	above, behind, under	15	228k (50.0%)
Possessive	has, part of, wearing	8	186k (40.9%)
Semantic	carrying, eating, using	24	39k (8.7%)
Misc	for, from, made of	3	2k (0.3%)

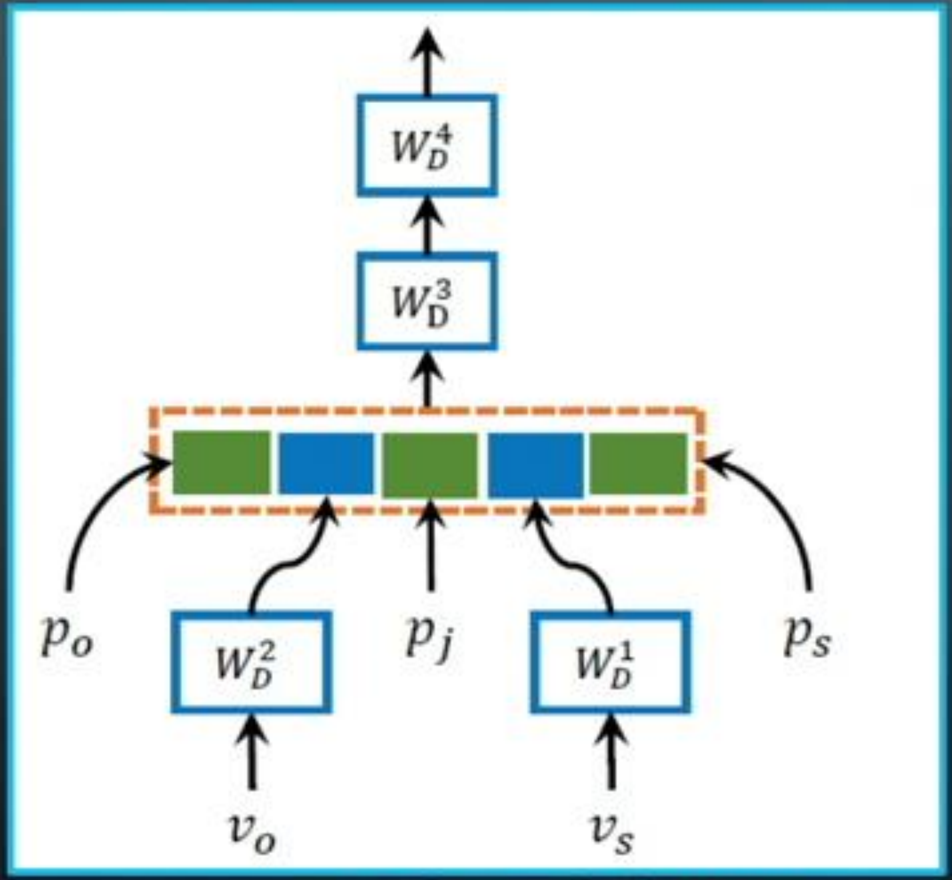
Table 1. Object and relation types in Visual Genome, organized by super-type. Most, 25.2% of entities are parts and 90.9% of relations are geometric or possessive.

Figures from https://arxiv.org/pdf/1711.06640.pdf

VG150

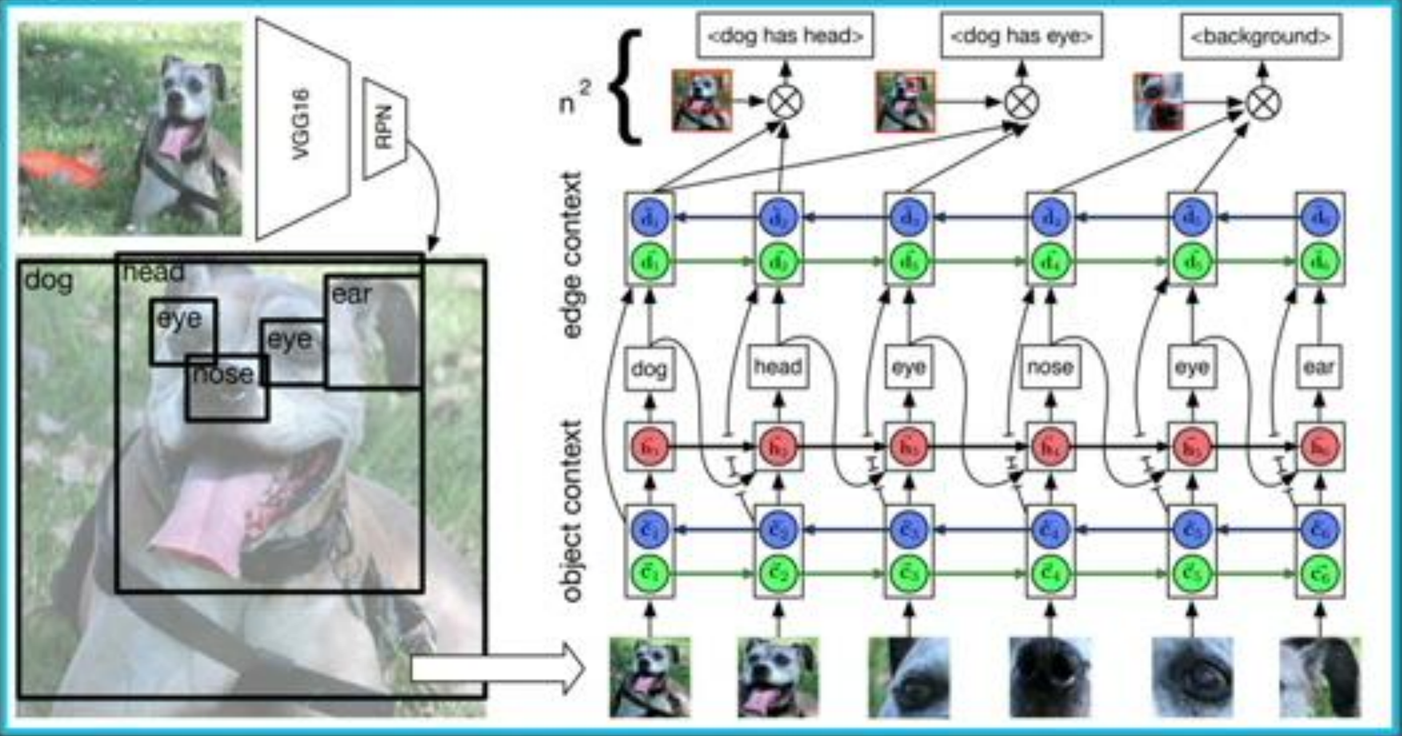
Step 0: Exploring Visual Relations for Image-Text Matching

- Discarding relationships classified with high confidence using the simple prior net.
- Top 1600 objs/500 rels
  - Show each predicate by Glove, run clustering to remove duplicates, e.g., “wears” and “is wearing a” ➔ 180 rels
  - Run VrR, remove rels that can be predicated with > 50% accuracy ➔ 117 rels
  - 58,983 images



Figures from https://arxiv.org/abs/1902.00313

Step 0: Exploring Visual Relations for Image-Text Matching



	Model	Scene Graph Detection R@20	Scene Graph Detection R@50	Scene Graph Detection R@100	Scene Graph Classification R@20	Scene Graph Classification R@50	Scene Graph Classification R@100	Predicate Classification R@20	Predicate Classification R@50	Predicate Classification R@100	Mean
models	VRD [29]		0.3	0.5		11.8	14.1		27.9	35.0	14.9
models	MESSAGE PASSING [47]		3.4	4.2		21.7	24.4		44.8	53.0	25.3
models	MESSAGE PASSING+	14.6	20.7	24.5	31.7	34.6	35.4	52.7	59.3	61.3	39.3
models	ASSOC EMBED [31]⋆	6.5	8.1	8.2	18.2	21.8	22.6	47.9	54.1	55.4	28.3
models	FREQ	17.7	23.5	27.6	27.7	32.4	34.0	49.4	59.9	64.1	40.2
models	FREQ+OVERLAP	20.1	26.2	30.1	29.3	32.3	32.9	53.6	60.6	62.2	40.7
models	MOTIFNET-LEFTRIGHT	21.4	27.2	30.3	**32.9**	**35.8**	**36.5**	**58.5**	**65.2**	**67.1**	**43.6**
ablations	MOTIFNET-NOCONTEXT	21.0	26.2	29.0	31.9	34.8	35.5	57.0	63.7	65.6	42.4
ablations	MOTIFNET-CONFIDENCE	**21.7**	**27.3**	**30.5**	32.6	35.4	36.1	58.2	65.1	67.0	43.5
ablations	MOTIFNET-SIZE	21.6	**27.3**	30.4	32.2	35.0	35.7	58.0	64.9	66.8	43.3
ablations	MOTIFNET-RANDOM	21.6	**27.3**	30.4	32.5	35.5	36.2	58.1	65.1	66.9	43.5

Type	Examples	Classes	Instances
Entities			
Part	arm, tail, wheel	32	200k (25.2%)
Artifact	basket, fork, towel	34	126k (16.0%)
Person	boy, kid, woman	13	113k (14.3%)
Clothes	cap, jean, sneaker	16	91k (11.5%)
Vehicle	airplane, bike, truck,	12	44k (5.6%)
Flora	flower, plant, tree	3	44k (5.5%)
Location	beach, room, sidewalk	11	39k (4.9%)
Furniture	bed, desk, table	9	37k (4.7%)
Animal	bear, giraffe, zebra	11	30k (3.8%)
Structure	fence, post, sign	3	30k (3.8%)
Building	building, house	2	24k (3.1%)
Food	banana, orange, pizza	6	13k (1.6%)
Relations			
Geometric	above, behind, under	15	228k (50.0%)
Possessive	has, part of, wearing	8	186k (40.9%)
Semantic	carrying, eating, using	24	39k (8.7%)
Misc	for, from, made of	3	2k (0.3%)

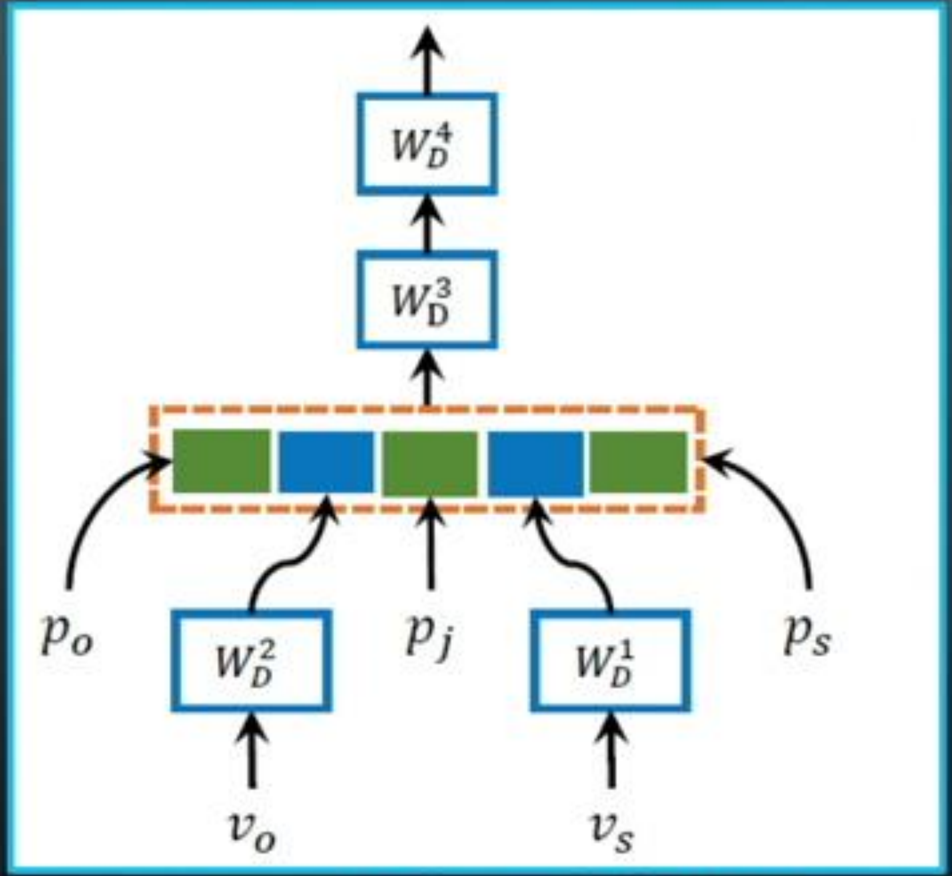
Table 1. Object and relation types in Visual Genome, organized by super-type. Most, 25.2% of entities are parts and 90.9% of relations are geometric or possessive.

Figures from https://arxiv.org/pdf/1711.06640.pdf

VG150

Step 0: Exploring Visual Relations for Image-Text Matching

- Discarding relationships classified with high confidence using the simple prior net.
- Top 1600 objs/500 rels
  - Show each predicate by Glove, run clustering to remove duplicates, e.g., "wears" and "is wearing a" ➔ 180 rels
  - Run VrR, remove rels that can be predicated with > 50% accuracy ➔ 117 rels
  - 58,983 images



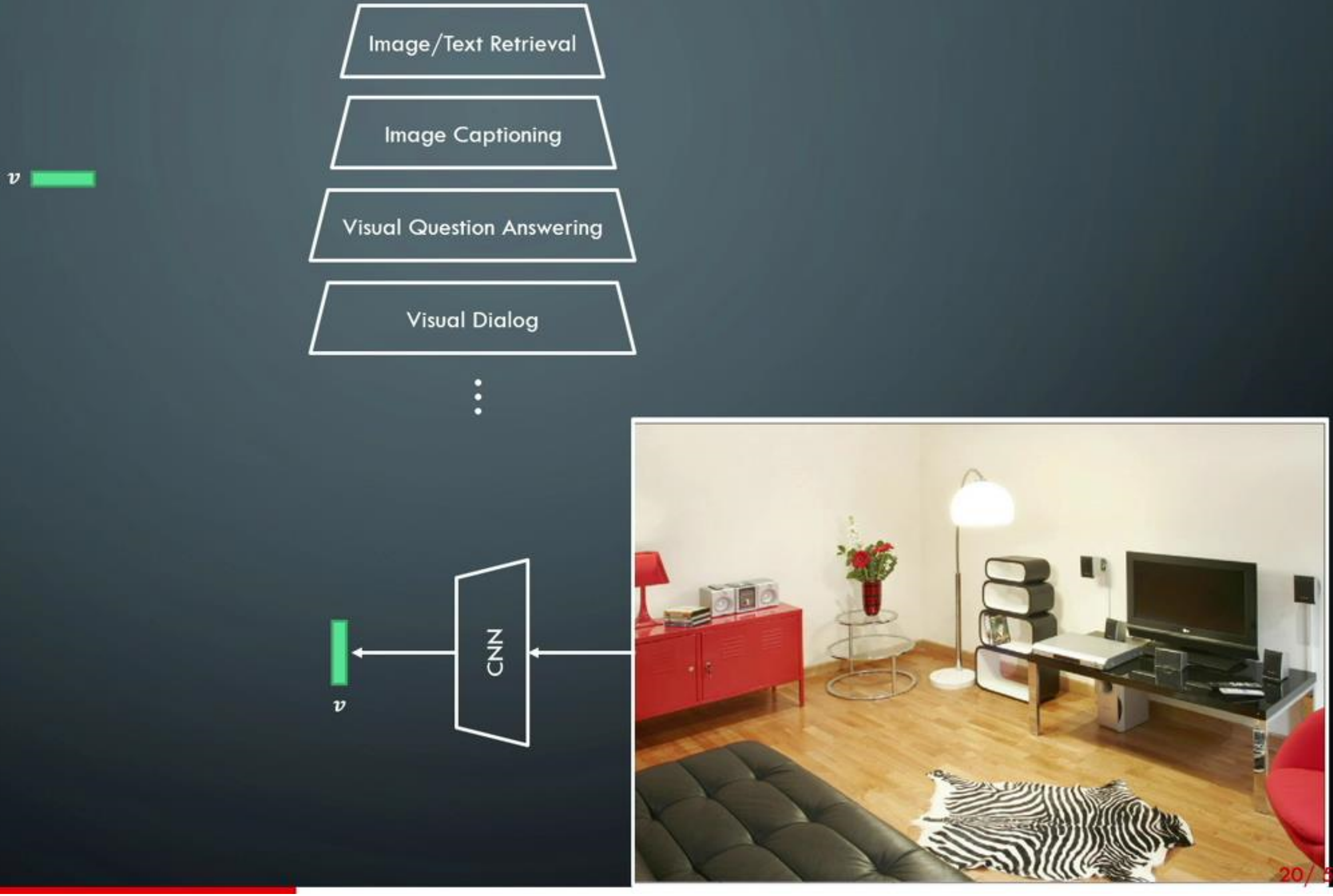
Figures from https://arxiv.org/abs/1902.00313

Step 0: Exploring Visual Relations for Image-Text Matching

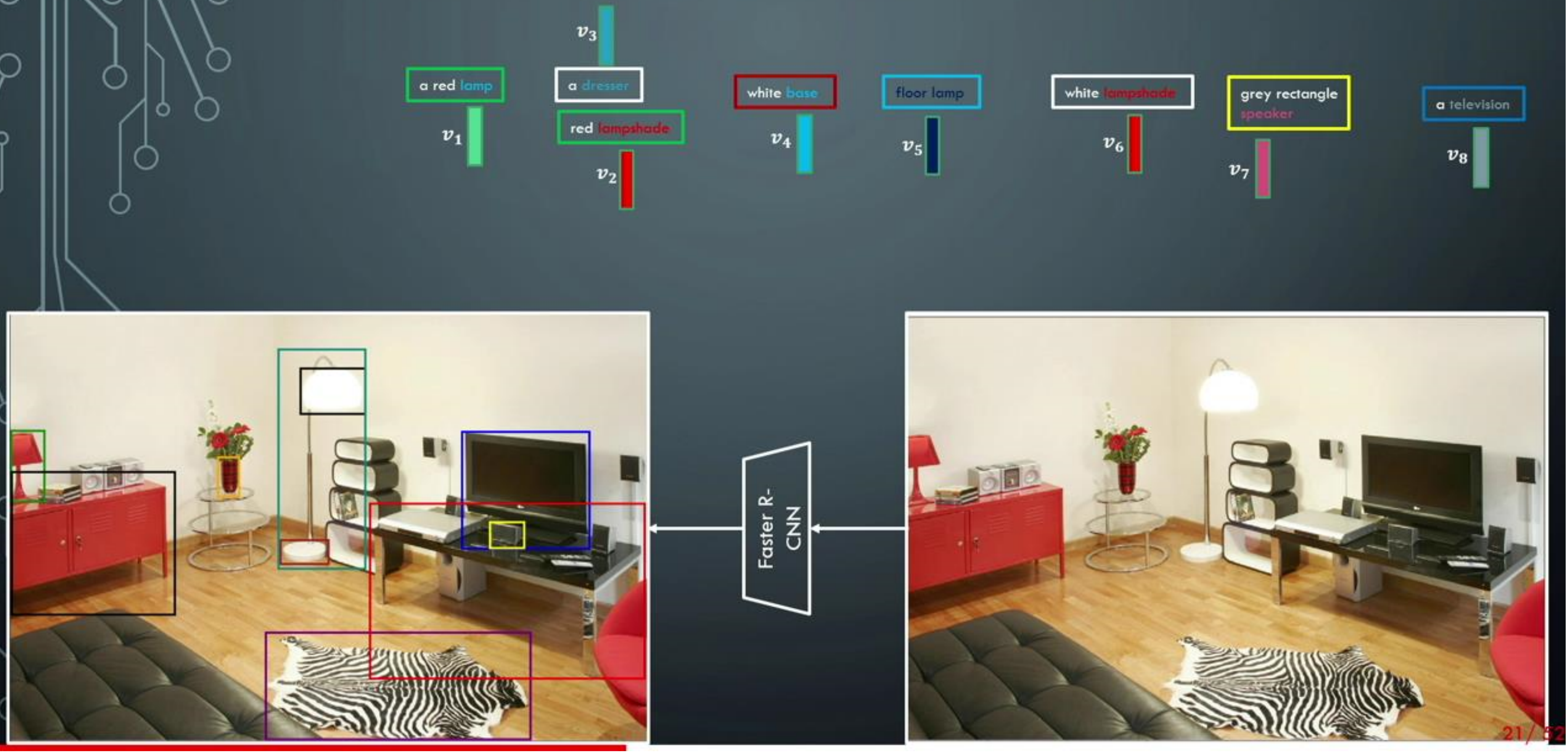
Methods	Datasets							
	Method specific VG splits				VrR-VG			
	Metrics	SGDet	SGCls	PredCls	Metrics	SGDet	SGCls	PredCls
MSDN [12]	R50	11.7	20.9	42.3	R50	3.59	-	-
	R100	14.0	24.0	48.2	R100	4.36	-	-
	R-gap	2.3	3.1	5.9	R-gap	0.77	-	-
Vtrans [31]	R50	5.52	-	61.2	R50	0.83	-	44.69
	R100	6.04	-	61.4	R100	1.08	-	44.84
	R-gap	0.52	-	0.26	R-gap	0.25	-	0.15
	VG150				VrR-VG			
	Metrics	SGDet	SGCls	PredCls	Metrics	SGDet	SGCls	PredCls
Neural-Motifs [30]	R50	27.2	35.8	65.2	R50	14.8	16.5	46.7
	R100	30.3	36.5	67.1	R100	17.4	19.2	52.5
	R-gap	3.1	0.7	1.9	R-gap	2.6	2.7	5.8
Message Passing [25]	R50	20.7	34.6	59.3	R50	8.46	12.1	29.7
	R100	24.5	35.4	61.3	R100	9.78	13.7	34.3
	R-gap	3.8	0.8	2.0	R-gap	1.3	1.6	4.6

Figures from https://arxiv.org/abs/1902.00313

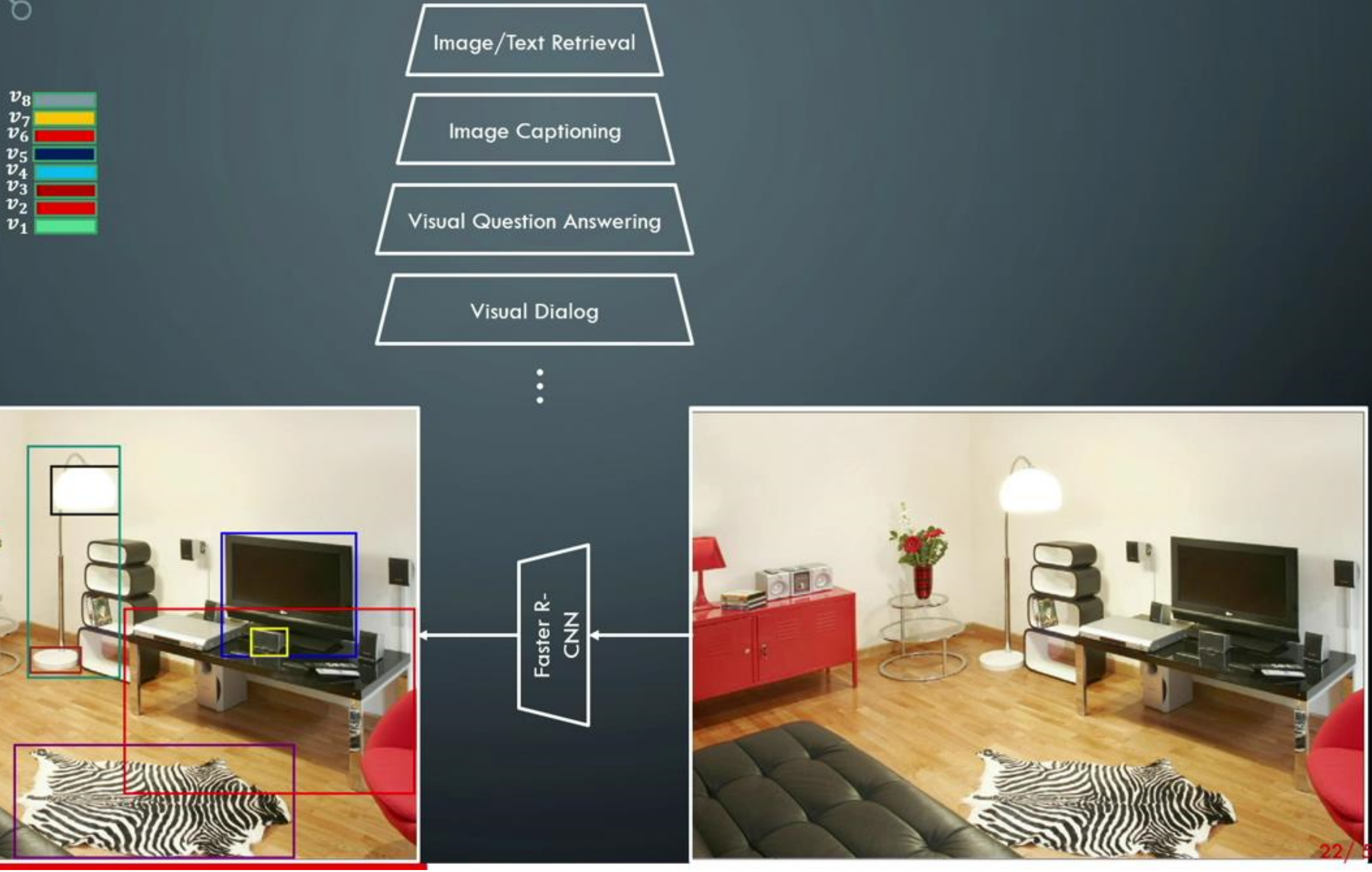
Step 0: Exploring Visual Relations for Image-Text Matching



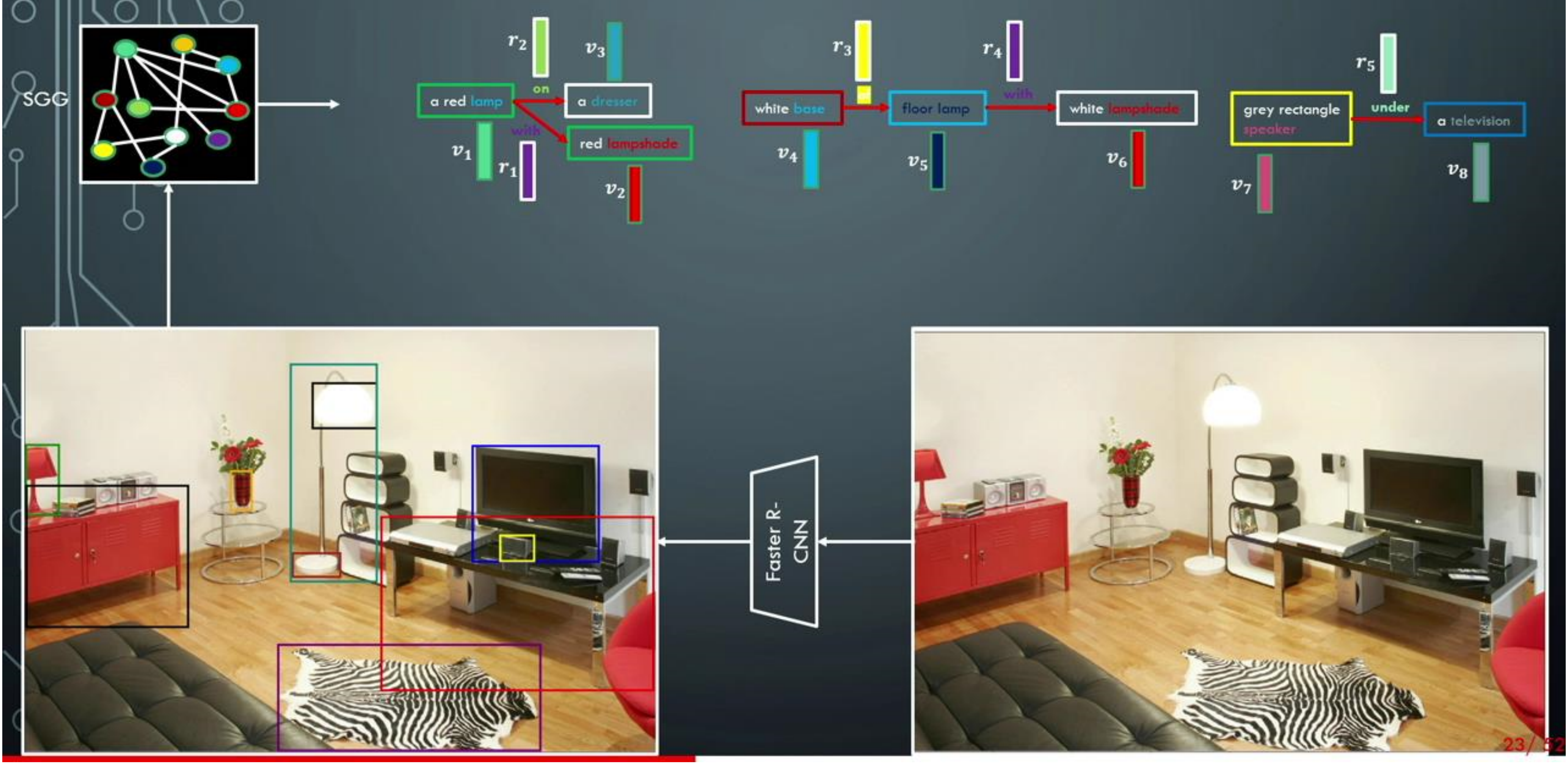
Step 0: Exploring Visual Relations for Image-Text Matching



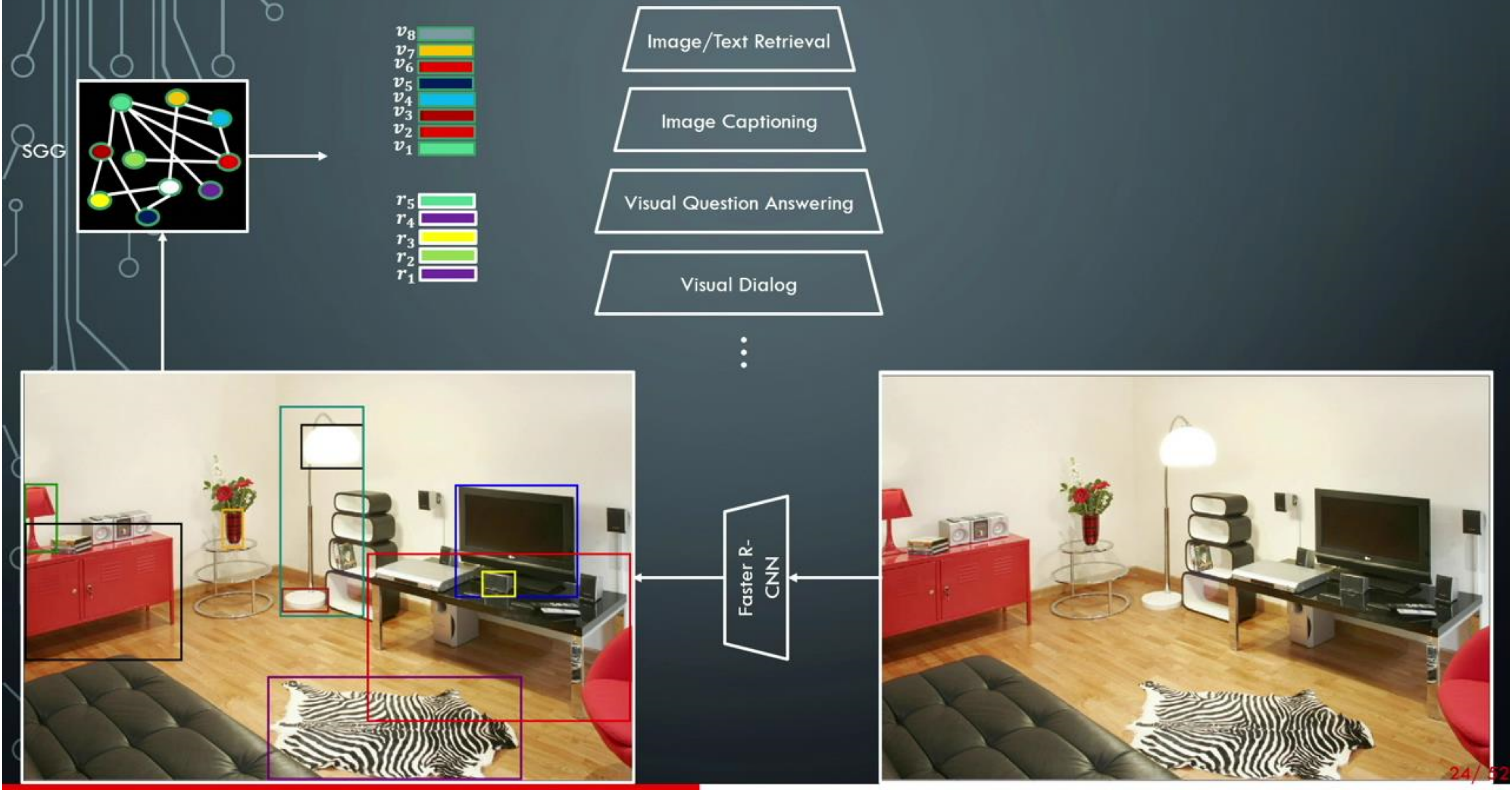
Step 0: Exploring Visual Relations for Image-Text Matching



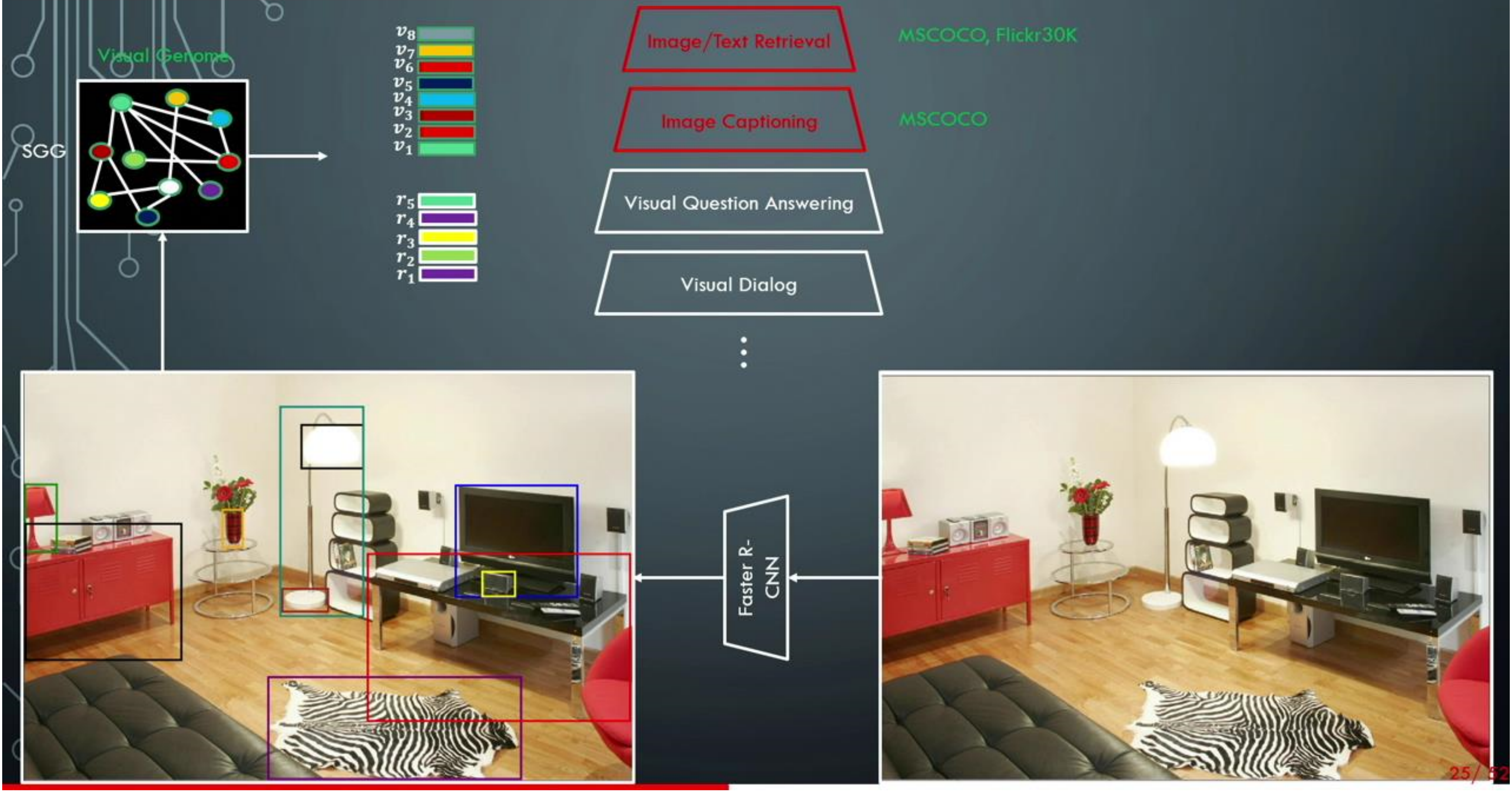
Step 0: Exploring Visual Relations for Image-Text Matching



Step 0: Exploring Visual Relations for Image-Text Matching



Step 0: Exploring Visual Relations for Image-Text Matching



Step 0: Exploring Visual Relations for Image-Text Matching

Model: IR

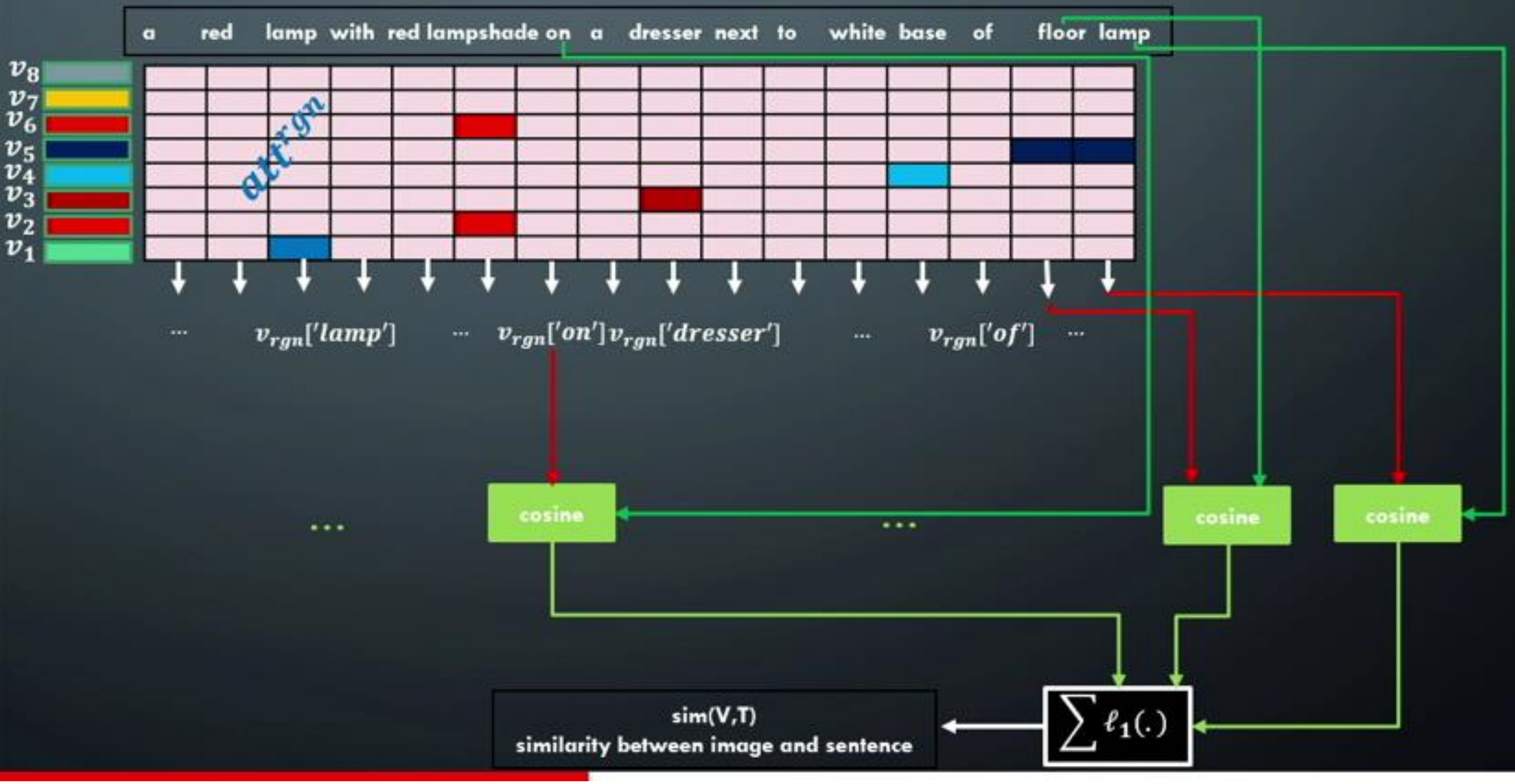
$$v_{rgn}['on'] = att^{rgn}['on'][v_1] \times \underset{v_1}{\blacksquare} + att^{rgn}['on'][v_2] \times \underset{v_2}{\blacksquare} + \ldots$$

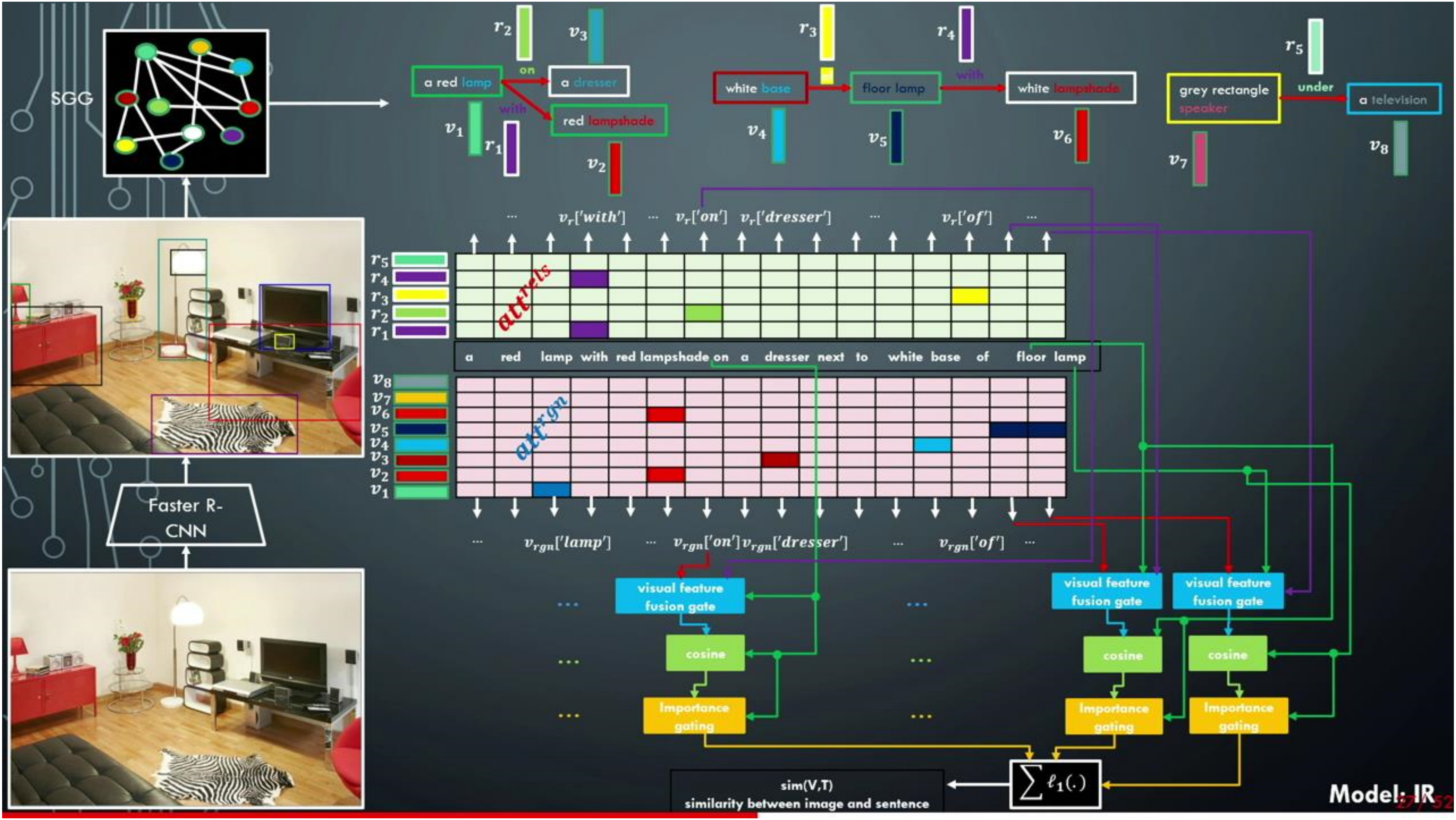
$$l(V,T) = [\alpha - sim(V,T) + sim(V,T^-)]_+ + [\alpha - sim(V,T) + sim(V^-,T)]_+$$

$$att_{ij}^{rgn} = \frac{exp(\lambda^{rgn}\hat{s}_{ij}^{rgn})}{\sum_{i=1}^{k} exp(\lambda^{rgn}\hat{s}_{ij}^{rgn})}$$

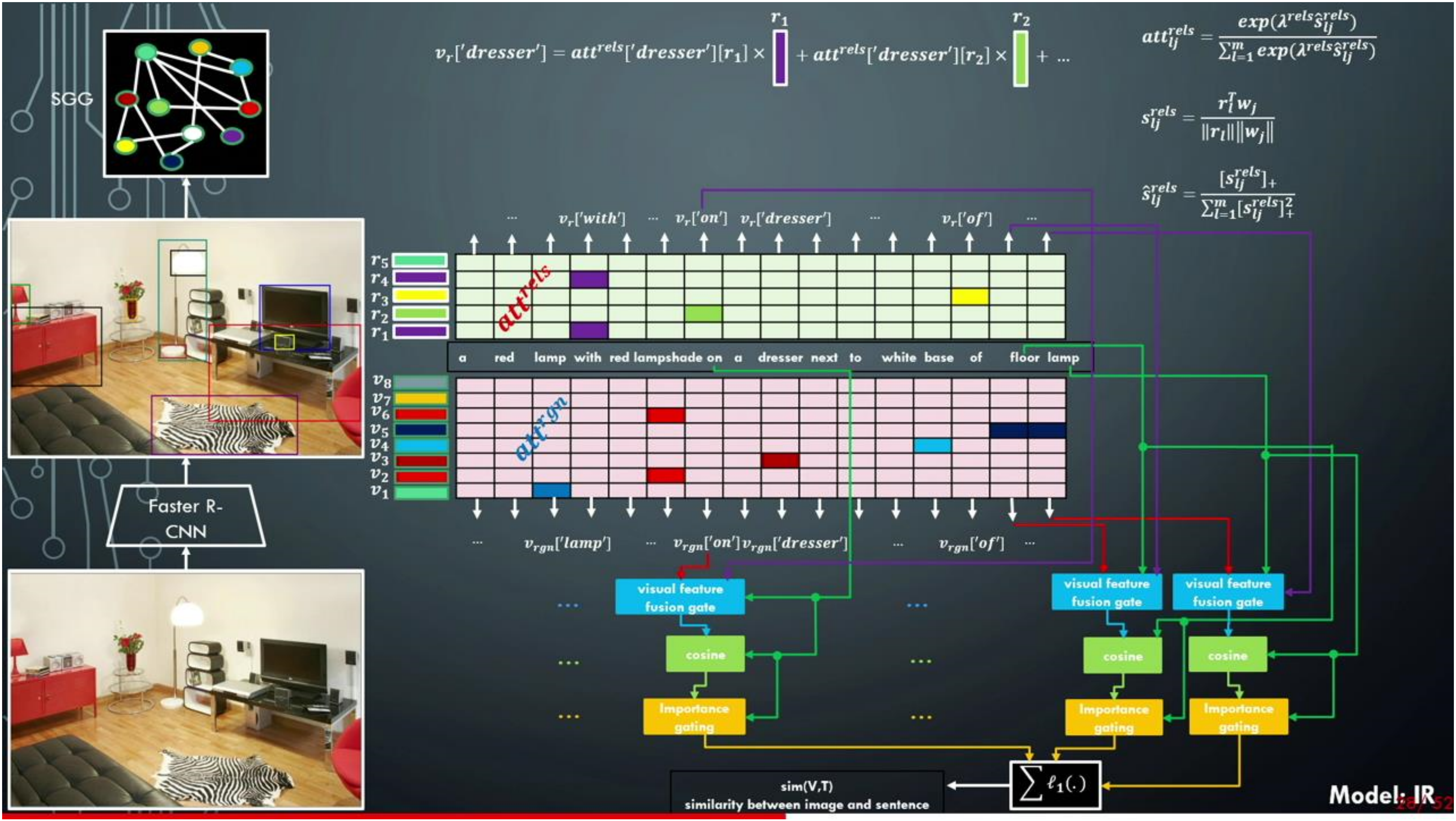
$$s_{ij}^{rgn} = \frac{v_i^T w_j}{\|v_i\|\|w_j\|}$$

$$\hat{s}_{ij}^{rgn} = \frac{[s_{ij}^{rgn}]_+}{\sum_{i=1}^{k}[s_{ij}^{rgn}]_+^2}$$




SGG
Faster R-CNN
a red lamp
on
a dresser
with
red lampshade
white base
floor lamp
with
white lampshade
grey rectangle speaker
under
a television
r_1 r_2 r_3 r_4 r_5
v_1 v_2 v_3 v_4 v_5 v_6 v_7 v_8
... $v_r['with']$ ... $v_r['on']$ $v_r['dresser']$ ... $v_r['of']$ ...
att^{rels}
a red lamp with red lampshade on a dresser next to white base of floor lamp
att^{rgn}
... $v_{rgn}['lamp']$ ... $v_{rgn}['on']$ $v_{rgn}['dresser']$ ... $v_{rgn}['of']$ ...
visual feature fusion gate
cosine
Importance gating
$\sum \ell_1(.)$
sim(V,T)
similarity between image and sentence

Model: IR



SGG
Faster R-CNN
$v_r['dresser'] = att^{rels}['dresser'][r_1] \times + att^{rels}['dresser'][r_2] \times + \dots$
r_1
r_2
$att^{rels}_{lj} = \frac{exp(\lambda^{rels}\hat{s}^{rels}_{lj})}{\sum_{l=1}^{m} exp(\lambda^{rels}\hat{s}^{rels}_{lj})}$
$s^{rels}_{lj} = \frac{r_l^T w_j}{\|r_l\|\|w_j\|}$
$\hat{s}^{rels}_{lj} = \frac{[s^{rels}_{lj}]_+}{\sum_{l=1}^{m}[s^{rels}_{lj}]^2_+}$
... $v_r['with']$ ... $v_r['on']$ $v_r['dresser']$ ... $v_r['of']$ ...
r_5 r_4 r_3 r_2 r_1
att^{rels}
a red lamp with red lampshade on a dresser next to white base of floor lamp
v_8 v_7 v_6 v_5 v_4 v_3 v_2 v_1
att^{rgn}
... $v_{rgn}['lamp']$ ... $v_{rgn}['on']$ $v_{rgn}['dresser']$ ... $v_{rgn}['of']$ ...
visual feature fusion gate
cosine
Importance gating
$\sum \ell_1(.)$
sim(V,T)
similarity between image and sentence
Model: IR

Step 0: Exploring Visual Relations for Image-Text Matching

	Flickr30K 1K Test Images						MSCOCO 5-fold 1K Test Images					
	text-to-image			image-to-text			text-to-image			image-to-text		
Method	r@1	r@5	r@10	r@1	r@5	r@10	r@1	r@5	r@10	r@1	r@5	r@10
UVS [21]	16.8	42.0	56.5	23.0	50.7	62.9	-	-	-	-	-	-
DVSA [18]	15.2	37.7	50.5	22.2	48.2	61.4	27.4	60.2	74.8	38.4	69.9	80.5
HM-LSTM [34]	27.7	-	68.8	38.1	-	76.5	36.1	-	86.7	43.9	-	87.8
DAN [32]	39.4	69.2	79.1	55.0	81.8	89.0	-	-	-	-	-	-
VSE++ [9]	39.6	70.1	79.5	52.9	80.5	87.2	52.0	84.3	92.0	64.6	90.0	95.7
Picturebook [20]	-	-	-	-	-	-	55.2	87.2	94.4	63.4	90.3	96.5
GXN [13]	41.5	-	80.1	56.8	-	89.6	56.6	-	94.5	68.5	-	97.9
SCO [16]	41.1	70.5	80.1	55.5	82.0	89.3	56.7	87.5	**94.8**	69.9	92.9	97.5
SCAN:												
SCAN ensemble† [24]	*48.6*	*77.7*	*85.2*	*67.4*	*90.3*	*95.8*	*58.8*	*88.4*	*94.8*	*72.7*	*94.8*	*98.4*
SCAN i-t AVG [24]	44.0	74.2	82.6	67.7	88.9	94.0	54.4	86.0	93.6	69.2	93.2	97.5
SCAN t-i AVG [24]	45.8	74.4	83.0	61.8	87.5	93.7	56.4	87.0	93.9	70.9	**94.5**	97.8
Ours:												
R-SCAN-VrRVG	**51.4**	**77.8**	**84.9**	66.3	90.6	**96.0**	57.6	87.3	93.7	70.3	**94.5**	**98.1**
R-SCAN-VG1500	51.1	**77.8**	84.7	**68.4**	**91.5**	95.3	**58.2**	**87.5**	93.8	**71.3**	94.0	98.0

	MSCOCO 5K Test Images					
	text-to-image			image-to-text		
Method	r@1	r@5	r@10	r@1	r@5	r@10
DVSA [18]	10.7	29.6	42.2	16.5	39.2	52.0
VSE++ [9]	30.3	59.4	72.4	41.3	71.1	81.2
GXN [13]	31.7	-	74.6	42.0	-	84.7
SCO [16]	33.1	62.9	75.5	42.8	72.3	83.0
SCAN:						
SCAN ens† [24]	*38.6*	*69.3*	*80.4*	*50.4*	*82.2*	*90.0*
SCAN t-i AVG [24]	34.4	63.7	75.7	46.4	77.4	87.2
Ours:						
R-SCAN-VrRVG	36.2	**65.5**	76.7	45.4	**77.9**	**87.9**
R-SCAN-VG1500	**36.4**	65.3	**77.0**	**46.7**	77.7	87.8

Model: IR

Step 0: Exploring Visual Relations for Image-Text Matching

- **R-COCO**
  - A subset of MS-COCO Karapathy's 5K test split
  - Focus on evaluating image-text matching on the pairs with semantic visual relations
  - 117 relations in VG are recognized as semantic relations using the prior detection network
  - These 117 relations can be mapped back to 259 relations in original VG
  - VG relations can be grouped in four categories
    - Geometric: e.g., above, behind
    - Possesive: e.g., has, part of

      majority of relations in VG, easy to predict
    - Semantic: e.g., carrying, eating ➔ more challenging to predict
    - Miscellaneous: e.g., for, from
  - Out of 259 relations we identified 164 of them as semantic relations
  - R-COCO focuses on semantic relations
    - 3,403 images from MS-COCO Karapathy's 5K test split where each image has at least 1 ground truth caption with one of above 164 relations
    - 1 ground truth caption that has semantic relation is selected per image, so 1 caption per image

Step 0: Exploring Visual Relations for Image-Text Matching

- **R-COCO**
  - A subset of MS-COCO Karapathy's 5K test split
  - Focus on evaluating image-text matching on the pairs with semantic visual relations
  - 117 relations in VG are recognized as semantic relations using the prior detection network
  - These 117 relations can be mapped back to 259 relations in original VG
  - VG relations can be grouped in four categories
    - Geometric: e.g., above, behind
    - Possesive: e.g., has, part of

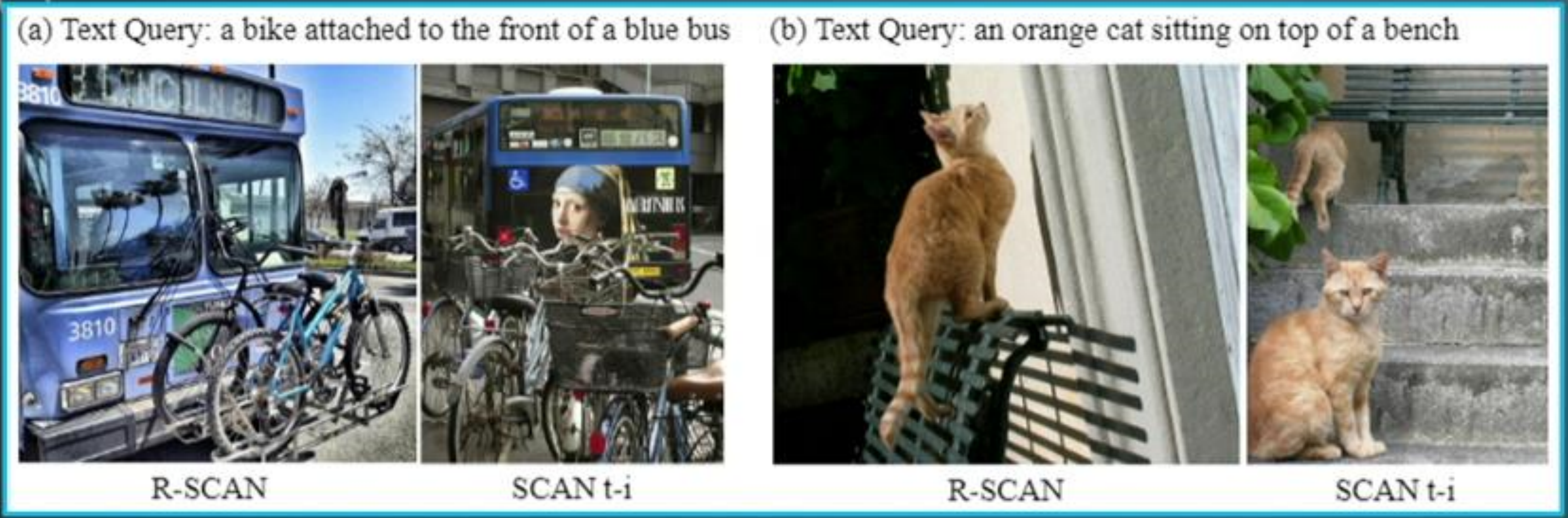
      majority of relations in VG, easy to predict
    - Semantic: e.g., carrying, eating ➔ more challenging to predict
    - Miscellaneous: e.g., for, from
  - Out of 259 relations we identified 164 of them as semantic relations
  - R-COCO focuses on semantic relations
    - 3,403 images from MS-COCO Karapathy's 5K test split where each image has at least 1 ground truth caption with one of above 164 relations
    - 1 ground truth caption that has semantic relation is selected per image, so 1 caption per image

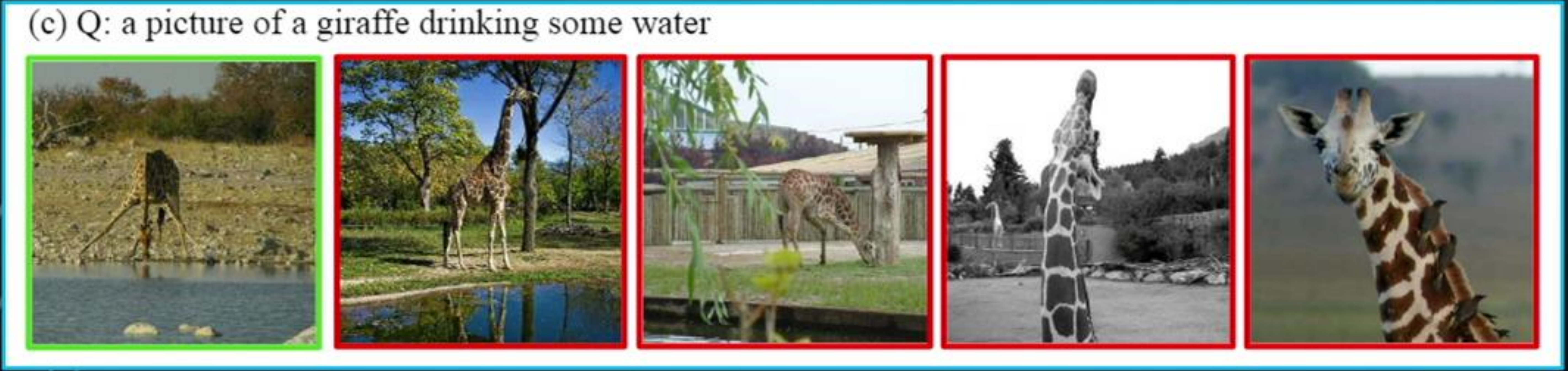
Model	text-to-image r@1	text-to-image r@5	text-to-image r@10	image-to-text r@1	image-to-text r@5	image-to-text r@10
SCAN t-i AVG	37.9	69.4	80.8	38.5	70.7	82.5
R-SCAN-VG150	39.8	70.6	82.0	38.1	71.0	83.5
R-SCAN-VrRVG	40.1	70.5	81.8	39.6	72.7	83.7
R-SCAN-VG1500	40.5	70.9	82.0	40.7	73.0	84.1

Table 1. Comparison of the cross-model retrieval results in terms of recall@K (r@K) on R-COCO. 'text-to-image' denotes image retrieval given text query. 'image-to-text' denotes text retrieval given image query.

Step 0: Exploring Visual Relations for Image-Text Matching

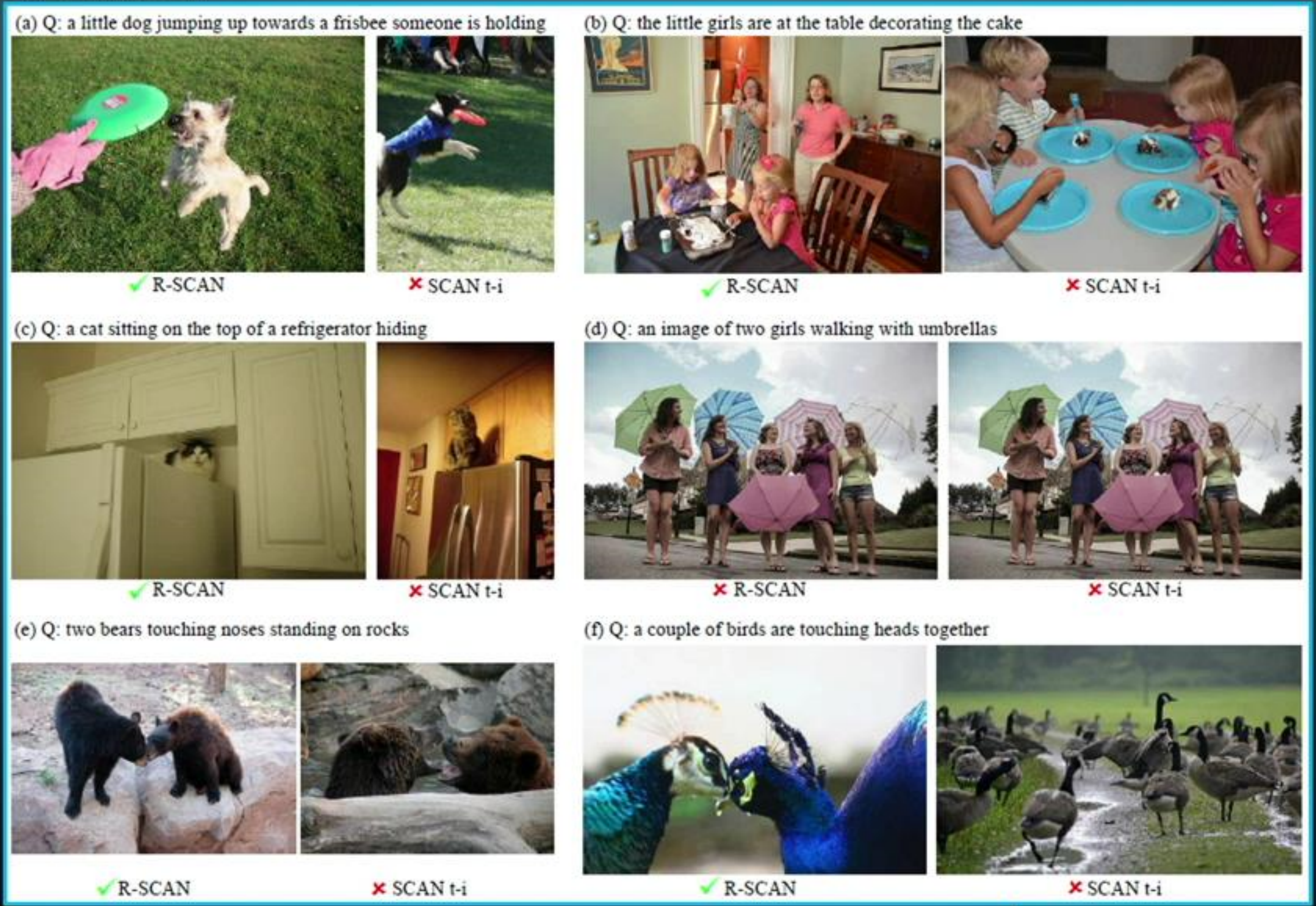
Samples: Image Retrieval





Step 0: Exploring Visual Relations for Image-Text Matching

Samples: Image Retrieval



Step 0: Exploring Visual Relations for Image-Text Matching

Samples: Text Retrieval

✓ **R-SCAN**
A dog playing with a toy in a grassy yard

✗ **SCAN**
A dog sitting on the grass with something in it 's mouth

✓ **R-SCAN**
A bus driving down a street on a road

✗ **SCAN**
A bus , cars and a motorcycle driving in busy traffic on the street

✗ **R-SCAN**
A woman is talking on her mobile phone angrily

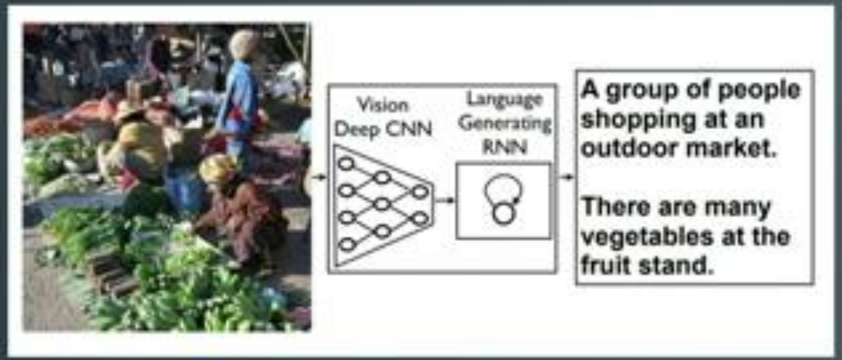
✗ **SCAN**
A woman is talking on her mobile phone angrily

Task: Captioning

GOOGLE,2014

Show and Tell: A Neural Image Caption Generator

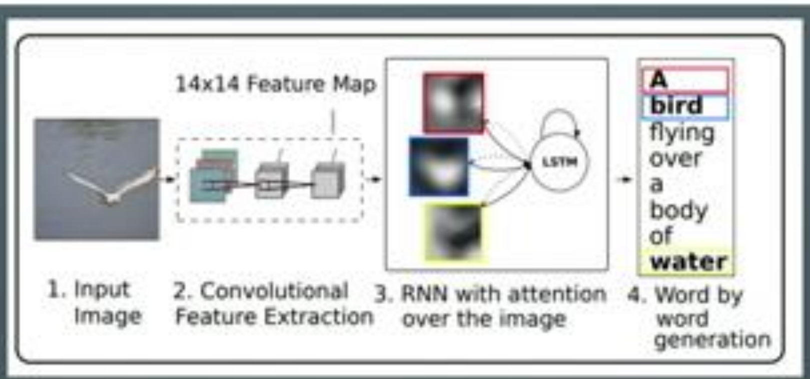
Oriol Vinyals, Google, vinyals@google.com
Alexander Toshev, Google, toshev@google.com
Samy Bengio, Google, bengio@google.com
Dumitru Erhan, Google, dumitru@google.com



U MONTREAL & U of T,2015

Show, Attend and Tell: Neural Image Caption Generation with Visual Attention

Kelvin Xu KELVIN.XU@UMONTREAL.CA
Jimmy Lei Ba JIMMY@PSI.UTORONTO.CA
Ryan Kiros RKIROS@CS.TORONTO.EDU
Kyunghyun Cho KYUNGHYUN.CHO@UMONTREAL.CA
Aaron Courville AARON.COURVILLE@UMONTREAL.CA
Ruslan Salakhutdinov RSALAKHU@CS.TORONTO.EDU
Richard S. Zemel ZEMEL@CS.TORONTO.EDU
Yoshua Bengio FIND-ME@THE.WEB



MSFT,2014

From Captions to Visual Concepts and Back

Hao Fang*, Saurabh Gupta*, Forrest Iandola*, Rupesh K. Srivastava*, Li Deng, Piotr Dollár[‡], Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C. Platt[‡], C. Lawrence Zitnick, Geoffrey Zweig

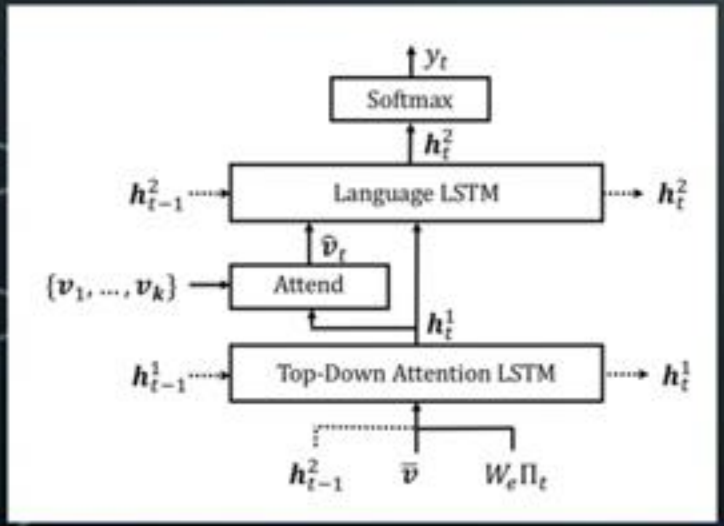
Microsoft Research

MSFT,2017

Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering

Peter Anderson[1]*, Xiaodong He[2], Chris Buehler[3], Damien Teney[4], Mark Johnson[5], Stephen Gould[1], Lei Zhang[3]

[1]Australian National University [2]JD AI Research [3]Microsoft Research [4]University of Adelaide [5]Macquarie University

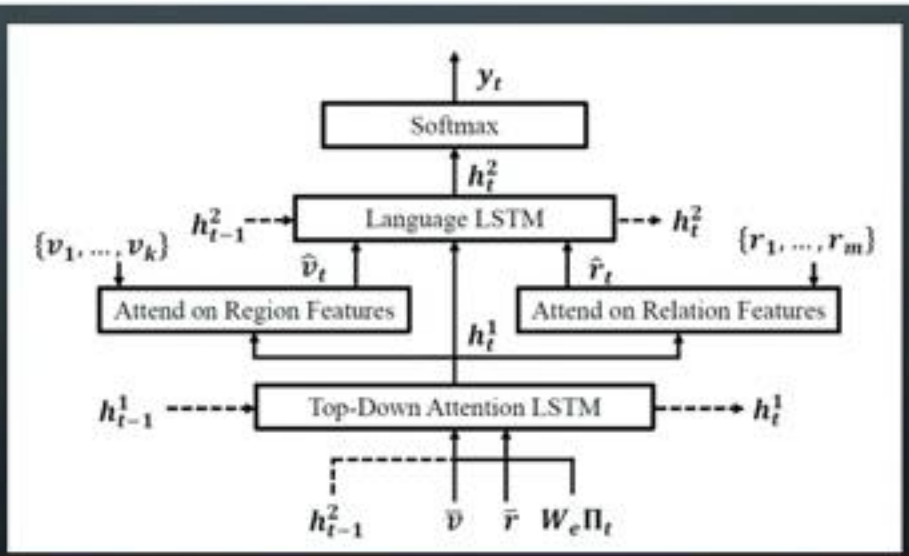


MSFT,2019

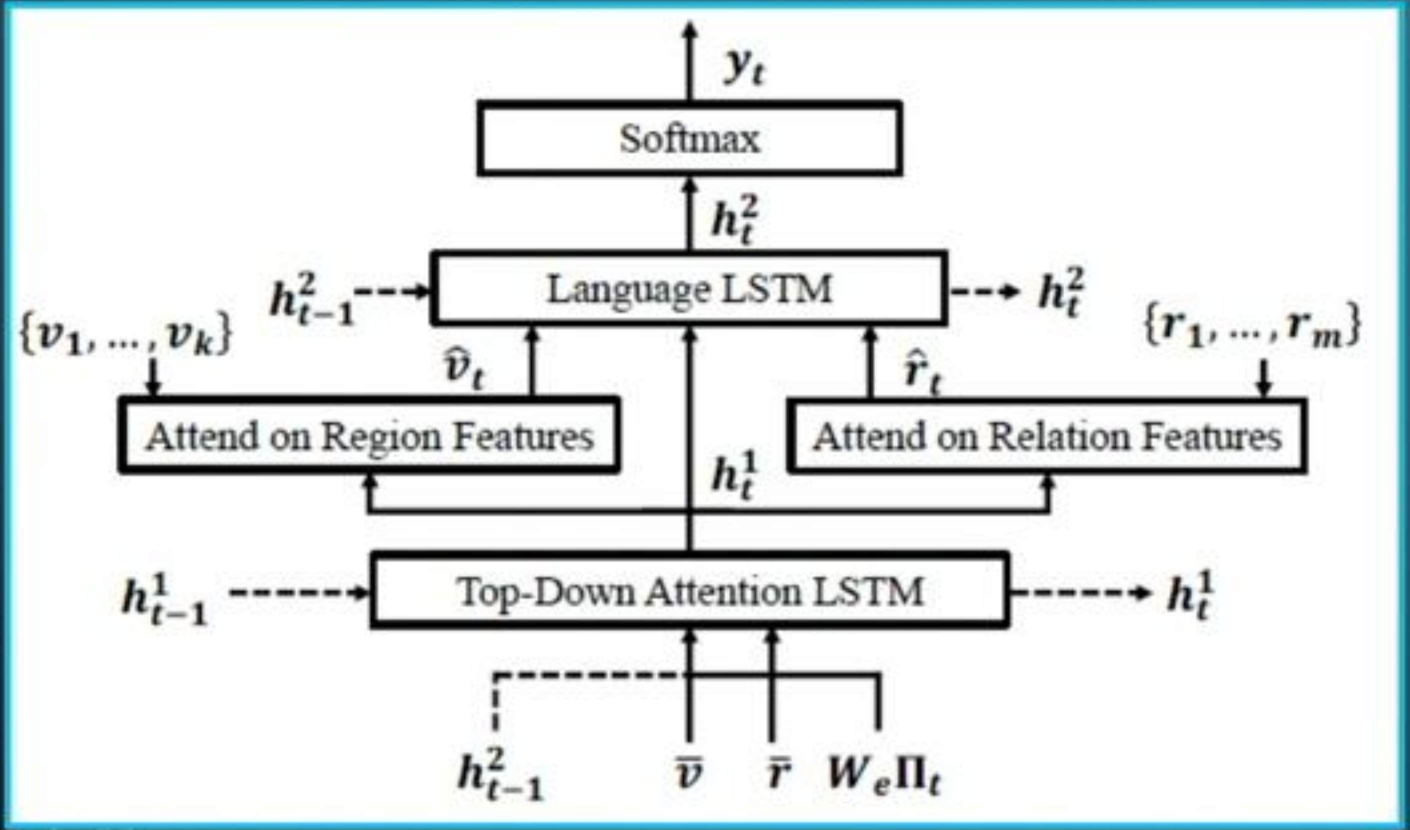
Exploring Visual Relations for Image-Text Matching

Kuang-Huei Lee *, Hamid Palangi *, Xi Chen, Houdong Hu, Jianfeng Gao

Microsoft AI and Research



Step 0: Exploring Visual Relations for Image-Text Matching



Bottom-up baseline
a woman standing on a sidewalk talking on a cell phone

Ours using Visual Relations
a woman standing on a sidewalk looking at her cell phone

Bottom-up baseline
a man holding a nintendo wii game controller

Ours using Visual Relations
a man sitting on a couch holding a wii remote

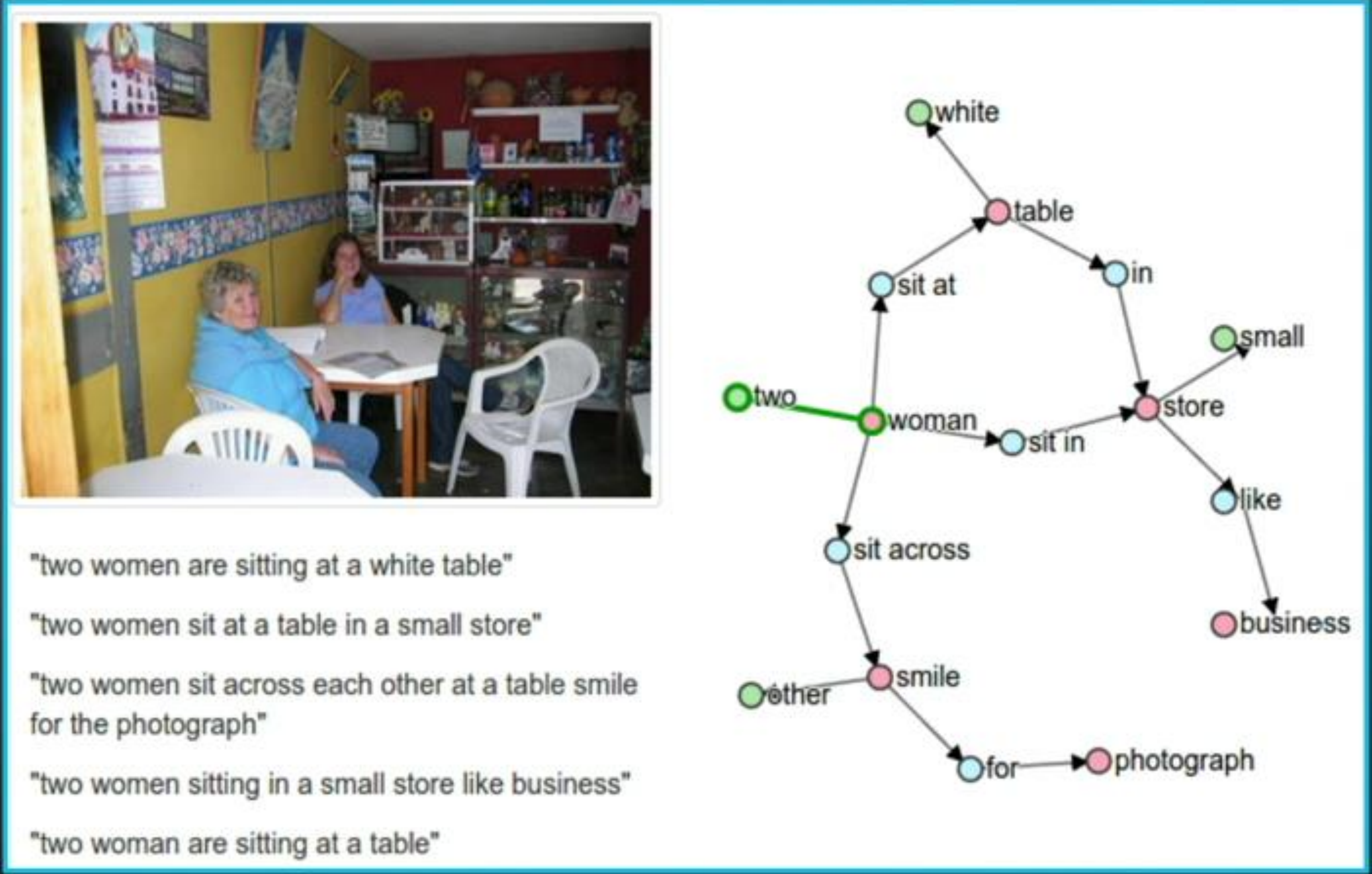
Bottom-up baseline
a couple of men standing next to each other

Ours using Visual Relations
a couple of men sitting next to each other

Bottom-up baseline
a man standing on the side of a road

Ours using Visual Relations
a man repairing a traffic light at an intersection

Step 1: Weakly supervised Scene Graph Generation



Figures from https://arxiv.org/pdf/1607.08822.pdf

Step 2: Visual Reasoning, Grounding, and beyond

VQA1, 2015

VQA: Visual Question Answering

www.visualqa.org

Aishwarya Agrawal*, Jiasen Lu*, Stanislaw Antol*,
Margaret Mitchell, C. Lawrence Zitnick, Dhruv Batra, Devi Parikh

What color are her eyes?
What is the mustache made of?

How many slices of pizza are there?
Is this a vegetarian pizza?

Is this person expecting company?
What is just under the tree?

Does it appear to be rainy?
Does this person have 20/20 vision?

Figures from https://arxiv.org/pdf/1505.00468.pdf

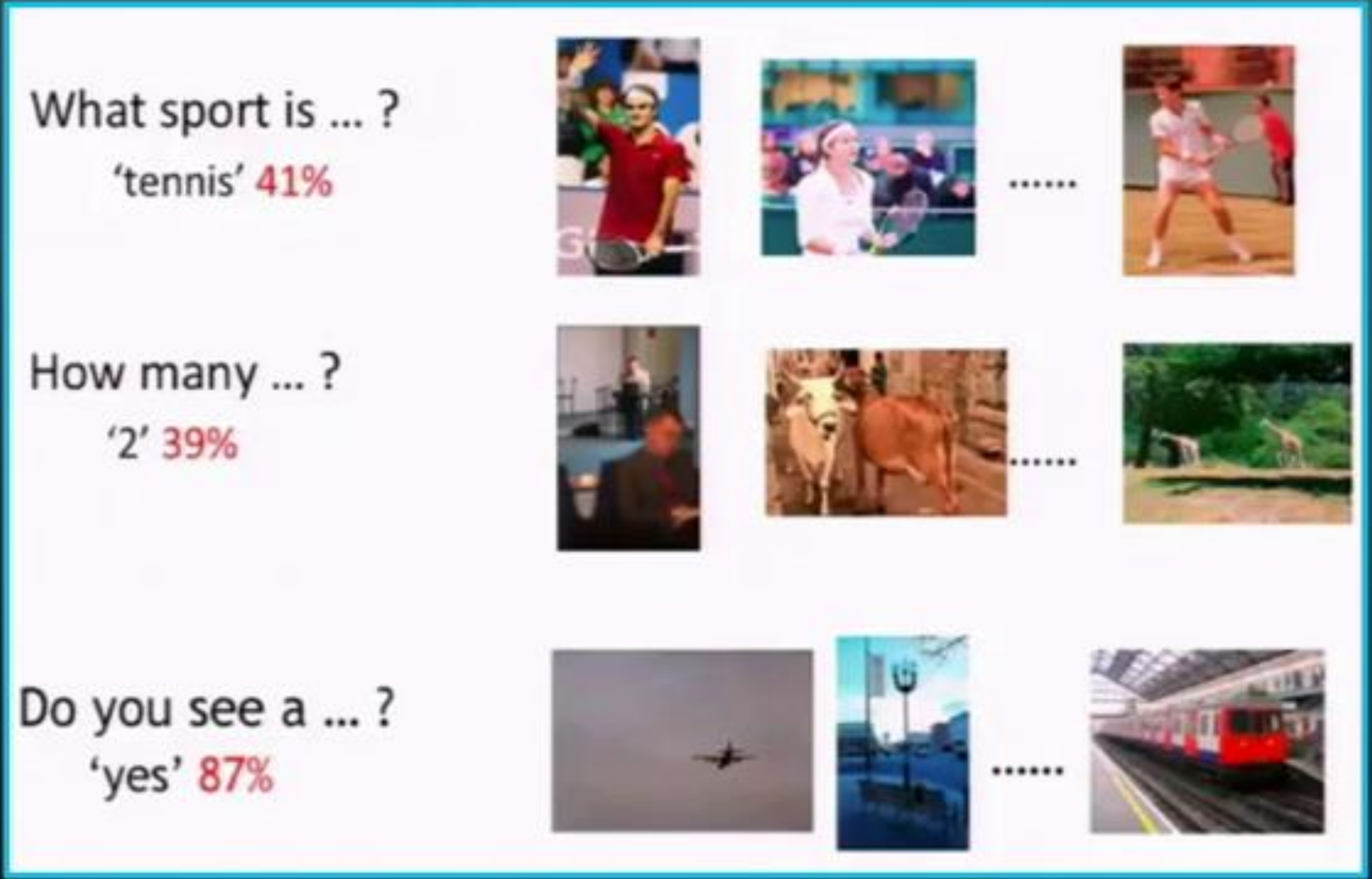
Step 2: Visual Reasoning, Grounding, and beyond

VQA1, 2015

VQA: Visual Question Answering

www.visualqa.org

Aishwarya Agrawal*, Jiasen Lu*, Stanislaw Antol*,
Margaret Mitchell, C. Lawrence Zitnick, Dhruv Batra, Devi Parikh



Figures from https://arxiv.org/pdf/1505.00468.pdf

Step 2: Visual Reasoning, Grounding, and beyond

VQA1, 2015

VQA2, 2017

Making the V in VQA Matter:
Elevating the Role of Image Understanding in Visual Question Answering

Yash Goyal*[1] Tejas Khot*[1] Douglas Summers-Stay[2] Dhruv Batra[3] Devi Parikh[3]
[1]Virginia Tech [2]Army Research Laboratory [3]Georgia Institute of Technology
[1]{ygoyal, tjskhot}@vt.edu [2]douglas.a.summers-stay.civ@mail.mil [3]{dbatra, parikh}@gatech.edu

A pair of similar images that result in two different answers to the same question

Figures from https://arxiv.org/pdf/1612.00837.pdf

Step 2: Visual Reasoning, Grounding, and beyond

VQA1, 2015   VQA2, 2017

- Not reading the whole question before picking an answer

Q: Are A: military

Q: Are they A: yes

Q: Are they playing A: yes

Q: Are they playing a A: yes

Q: Are they playing a game? A: yes

GT Answer: yes

Q: What A: umbrella

Q: What season A: summer

Q: What season of A: summer

Q: What season of year A: summer

Q: What season of year was A: summer

Q: What season of year was this A: summer

Q: What season of year was this photo A: summer

Q: What season of year was this photo taken A: summer

Q: What season of year was this photo taken in? A: summer

GT Answer: spring

Q: How A: no

Q: How many A: 2

Q: How many horses A: 2

Q: How many horses are A: 2

Q: How many horses are on A: 2

Q: How many horses are on the A: 2

Q: How many horses are on the beach? A: 2

GT Answer: 6

Q: Is A: kitchen

Q: Is the A: outside

Q: Is the bench A: no

Q: Is the bench made A: no

Q: Is the bench made of A: no

Q: Is the bench made of metal? A: no

GT Answer: yes

Figures from https://arxiv.org/abs/1606.07356

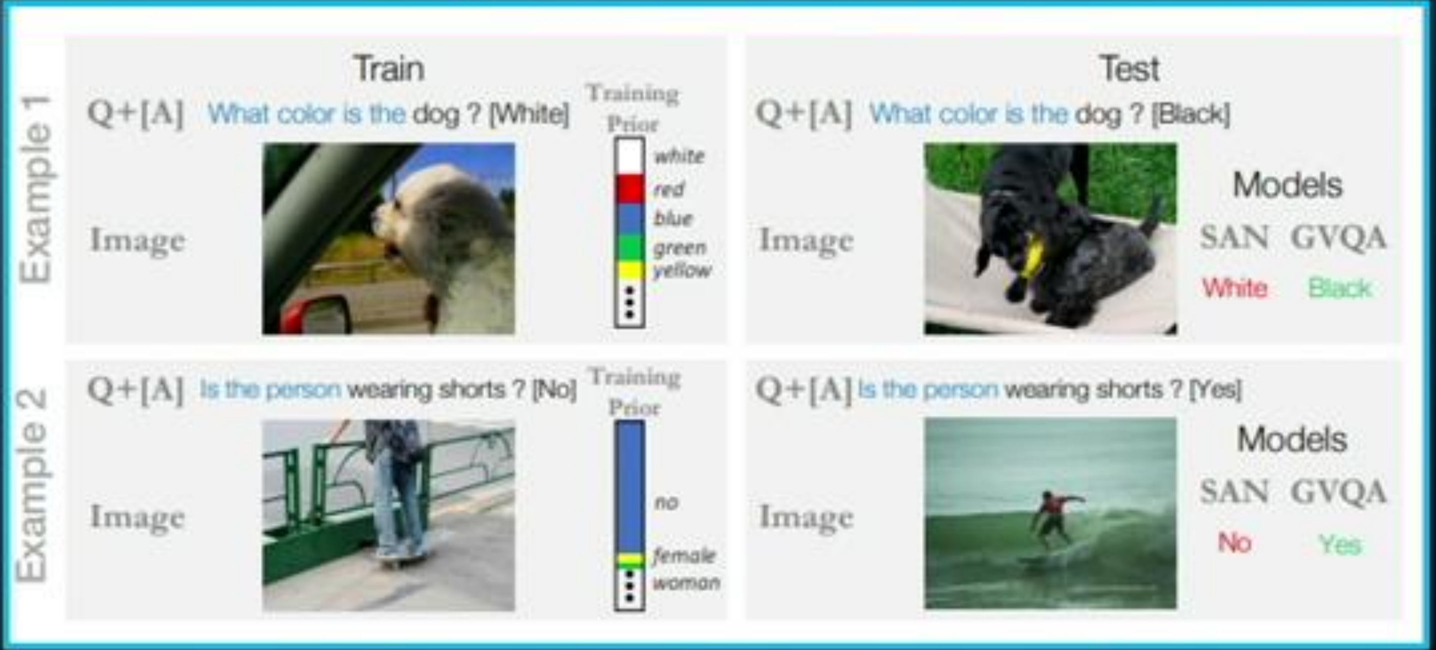
Step 2: Visual Reasoning, Grounding, and beyond

VQA1,2015 VQA2,2017 VQA-CP,2018

- Not reading the whole question before picking an answer
- Ignoring the context (image) and relying on language priors

Don't Just Assume; Look and Answer:
Overcoming Priors for Visual Question Answering

Aishwarya Agrawal[1*], Dhruv Batra[1,2], Devi Parikh[1,2], Aniruddha Kembhavi[3]
[1]Georgia Institute of Technology, [2]Facebook AI Research, [3]Allen Institute for Artificial Intelligence
{aishwarya, dbatra, parikh}@gatech.edu, anik@allenai.org



Figures from https://arxiv.org/abs/1712.00377

Step 2: Visual Reasoning, Grounding, and beyond

VQA1,2015

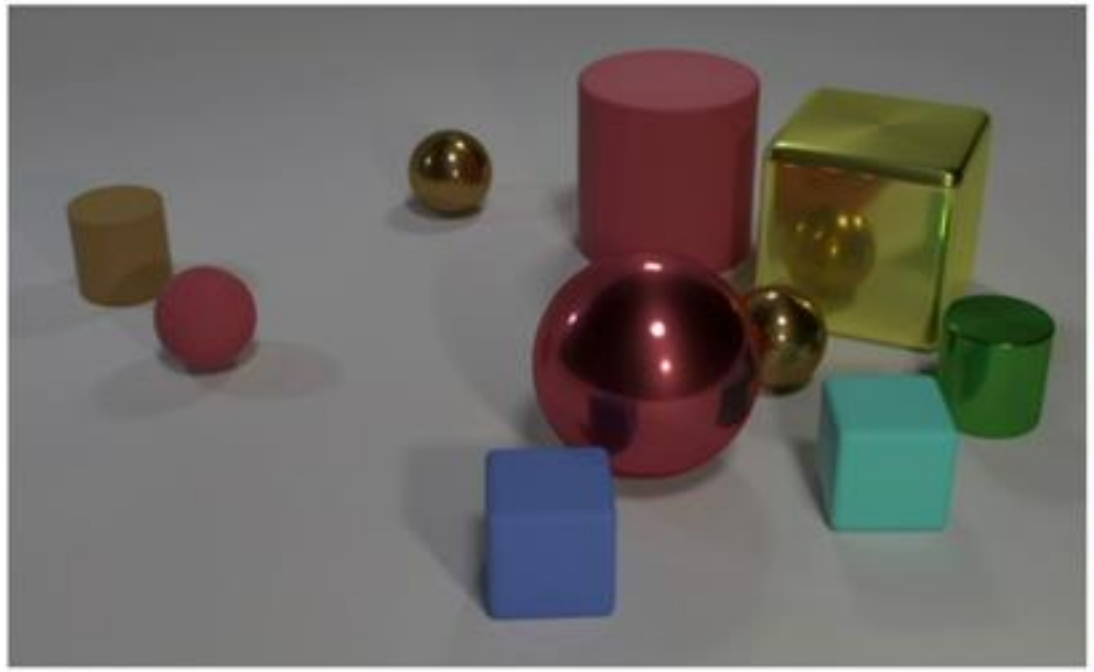
VQA2,2017

VQA-CP,2018

CLEVR,2017

CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning

Justin Johnson[1,2*] Bharath Hariharan[2] Laurens van der Maaten[2]
Li Fei-Fei[1] C. Lawrence Zitnick[2] Ross Girshick[2]
[1]Stanford University [2]Facebook AI Research



Q: Are there an equal number of large things and metal spheres? **Q:** What size is the cylinder that is left of the brown metal thing that is left of the big sphere? **Q:** There is a sphere with the same size as the metal cube; is it made of the same material as the small red sphere? **Q:** How many objects are either small cylinders or metal things?

Figures from https://arxiv.org/pdf/1612.06890.pdf

Step 2: Visual Reasoning, Grounding, and beyond

VQA1,2015

VQA2,2017

VQA-CP,2018

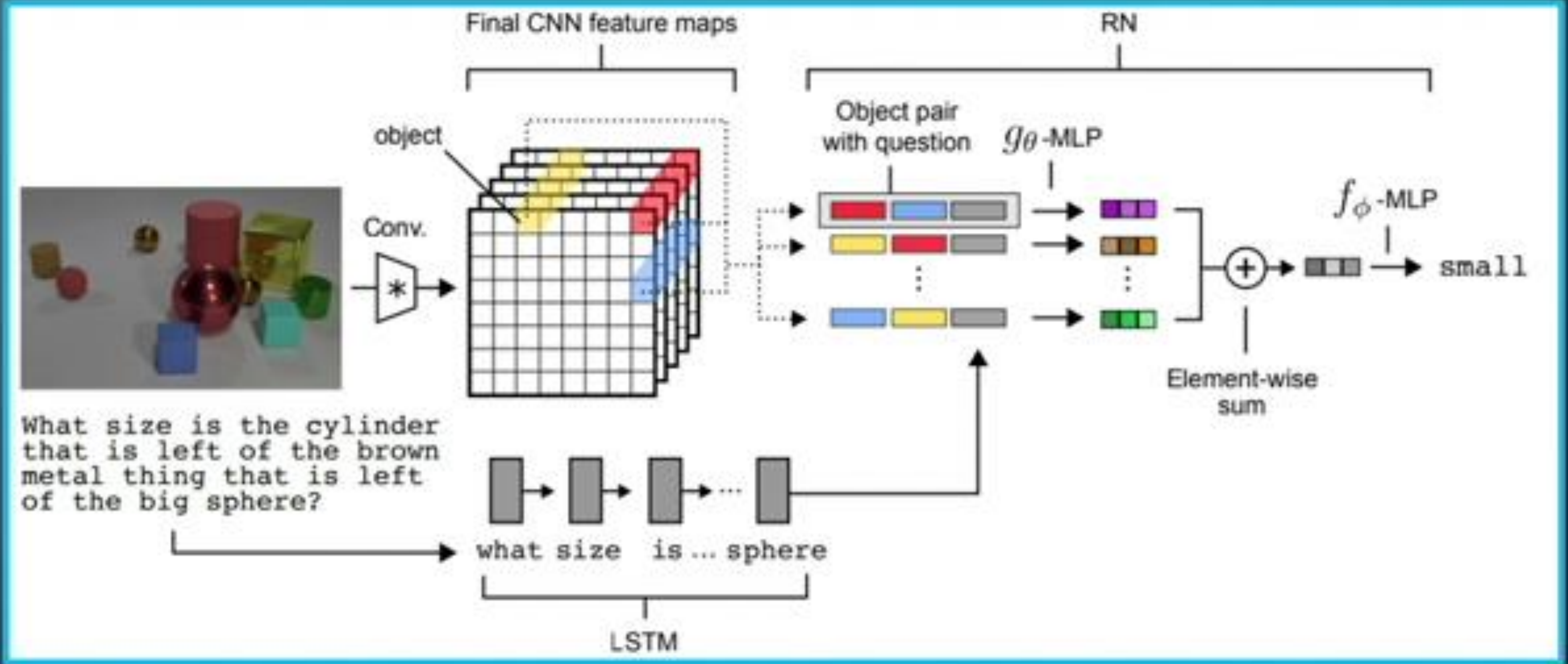
CLEVR,2017

A simple neural network module for relational reasoning

Adam Santoro*, David Raposo*, David G.T. Barrett, Mateusz Malinowski, Razvan Pascanu, Peter Battaglia, Timothy Lillicrap

adamsantoro@, draposo@, barrettdavid@, mateuszm@, razp@, peterbattaglia@, countzero@google.com

DeepMind
London, United Kingdom



Model	Overall	Count	Exist	Compare Numbers	Query Attribute	Compare Attribute
Human	92.6	86.7	96.6	86.5	95.0	96.0
Q-type baseline	41.8	34.6	50.2	51.0	36.0	51.3
LSTM	46.8	41.7	61.1	69.8	36.8	51.8
CNN+LSTM	52.3	43.7	65.2	67.1	49.3	53.0
CNN+LSTM+SA	68.5	52.2	71.1	73.5	85.3	52.3
CNN+LSTM+SA*	76.6	64.4	82.7	77.4	82.6	75.4
CNN+LSTM+RN	**95.5**	**90.1**	**97.8**	**93.6**	**97.9**	**97.1**

Figures from https://arxiv.org/pdf/1706.01427.pdf

Step 2: Visual Reasoning, Grounding, and beyond

VQA1,2015

VQA2,2017

VQA-CP,2018

CLEVR,2017

GQA,2019

GQA: A New Dataset for Real-World Visual Reasoning and Compositional Question Answering

visualreasoning.net

Drew A. Hudson
Stanford University
353 Serra Mall, Stanford, CA 94305
dorarad@cs.stanford.edu

Christopher D. Manning
Stanford University
353 Serra Mall, Stanford, CA 94305
manning@cs.stanford.edu

Figure 1: Examples from the new GQA dataset for visual reasoning and compositional question answering:
*Is the **bowl** to the right of the **green apple**?*
*What type of **fruit** in the image is **round**?*
*What color is the **fruit** on the right side, red or **green**?*
*Is there any **milk** in the **bowl** to the left of the **apple**?*

Figures from https://arxiv.org/pdf/1902.09506.pdf

Step 2: Visual Reasoning, Grounding, and beyond

VQA1,2015

VQA2,2017

VQA-CP,2018

CLEVR,2017

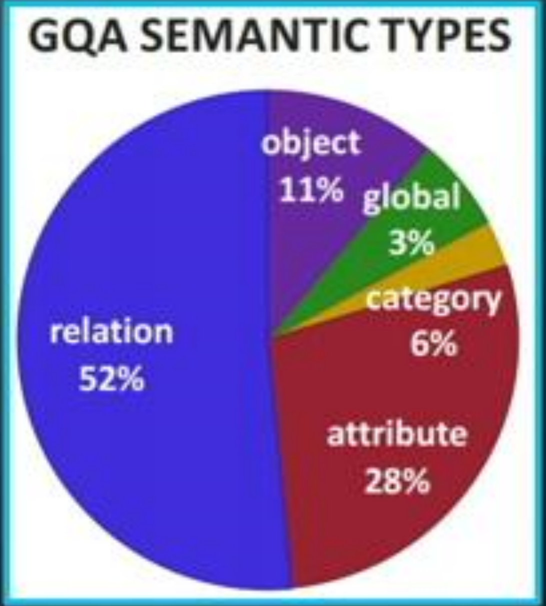
GQA,2019

GQA
1. Are there any coats? yes
2. Do you see a red coat in the image? no
3. Is the person that is to the left of the man exiting a truck? no
4. Which place is this? road

VQA
1. Where is the bus driver?
2. Why is the man in front of the bus?
3. What numbers are repeated in the bus number?

GQA
1. What is in front of the green fence? gate
2. Of which color is the gate? silver
3. Where is this? street
4. What color is the fence behind the gate? green
5. Is the fence behind the gate both brown and metallic? no

VQA
1. What are the yellow lines called?
2. Why don't the trees have leaves?
3. Where is the stop sign?



Figures from https://arxiv.org/pdf/1902.09506.pdf

Step 2: Visual Reasoning, Grounding, and beyond

VQA1,2015

VQA2,2017

VQA-CP,2018

CLEVR,2017

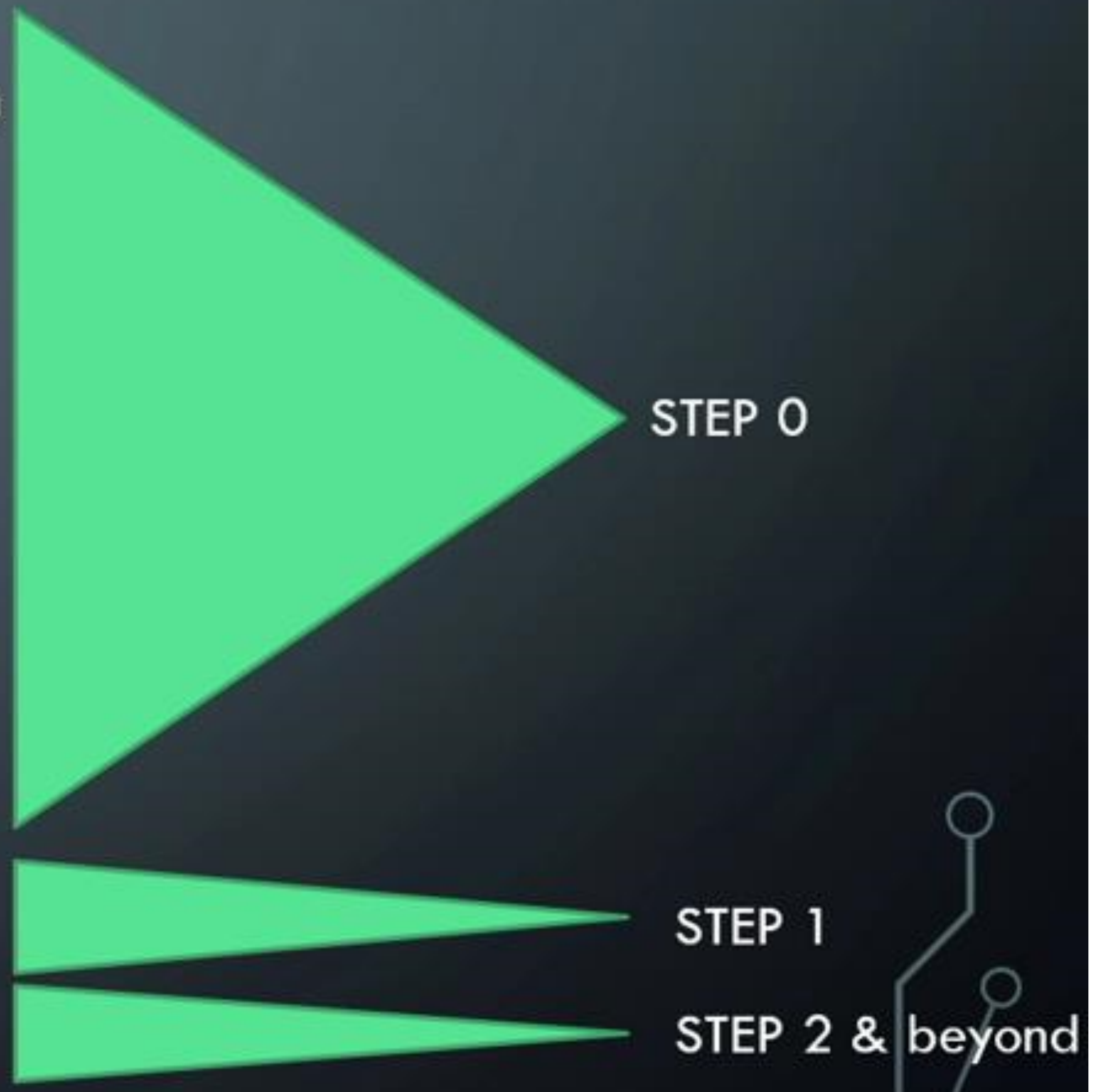
GQA,2019

- **Evaluation**
  - Accuracy
  - Consistency: model's consistency across entailed questions
    - Is there a red apple to the left of the white plate?
      - Is the plate to the right of the apple?
      - Is there a red fruit to the left of the plate?
      - What is the white thing to the right of the apple?
  - Plausibility: are the answers plausible in real world?
    - Example 1: For a question about color of an apple, green and red are plausible but attributes like blue are not
    - Example 2: For a question about relation r between s and t, the existence of triplet (s,r,t) is checked across dataset
  - Grounding
    - Total attention weights for the answer on object or relation which will be grounding score
  - ...

Figures from https://arxiv.org/pdf/1902.09506.pdf

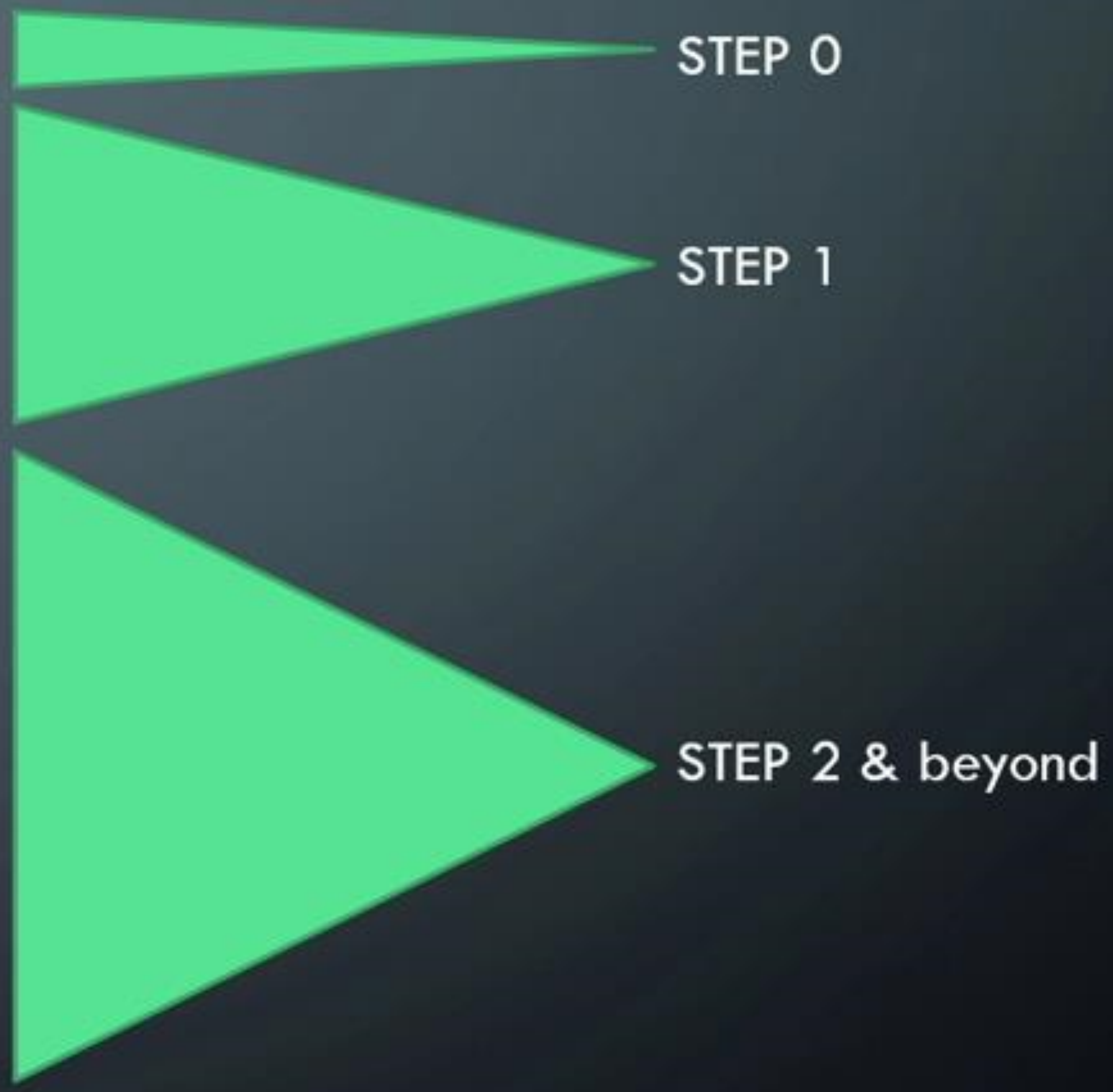
CONCLUSIONS

- Scene Graph Generation (SGG)
  - Pretraining SGGs that result in semantically rich features is challenging
  - Release of VG1500 consisting of 1500 objects and 500 relations
- Image-Text Retrieval
  - R-SCAN for image-to-text and text-to-image retrieval with significant gains compared to previous SOTA
  - Shipping in progress in BING
  - Release of R-COCO evaluation set
- Image Captioning
  - Completing the missing edge piece in current captioning systems
- Weakly supervised SGG
- Reasoning and beyond



CONCLUSIONS

- Weakly supervised SGG
  - How to exploit large scale click data?
  - How to expand VG's ontology?
  - What are useful approaches for SGG pretraining using imperfect "often" noisy labels from click data?
- Reasoning and beyond
  - Given a query how to reason over scene graphs?
  - How to overcome biases in VQA systems?
    Is it easier to control bias using a structured representation?
  - Is GQA enough?
    - No counting questions
    - What is the performance of current systems on golden scene graphs?
    - Is it a testbed for a good vision backend or it can measure the reasoning capability of the model?
  - Can we embed common sense into SGGs? How to define visual common sense?
    - Are datasets like VCR (Visual Commonsense Reasoning) enough?



Thanks!

Step 0: Exploring Visual Relations for Image-Text Matching

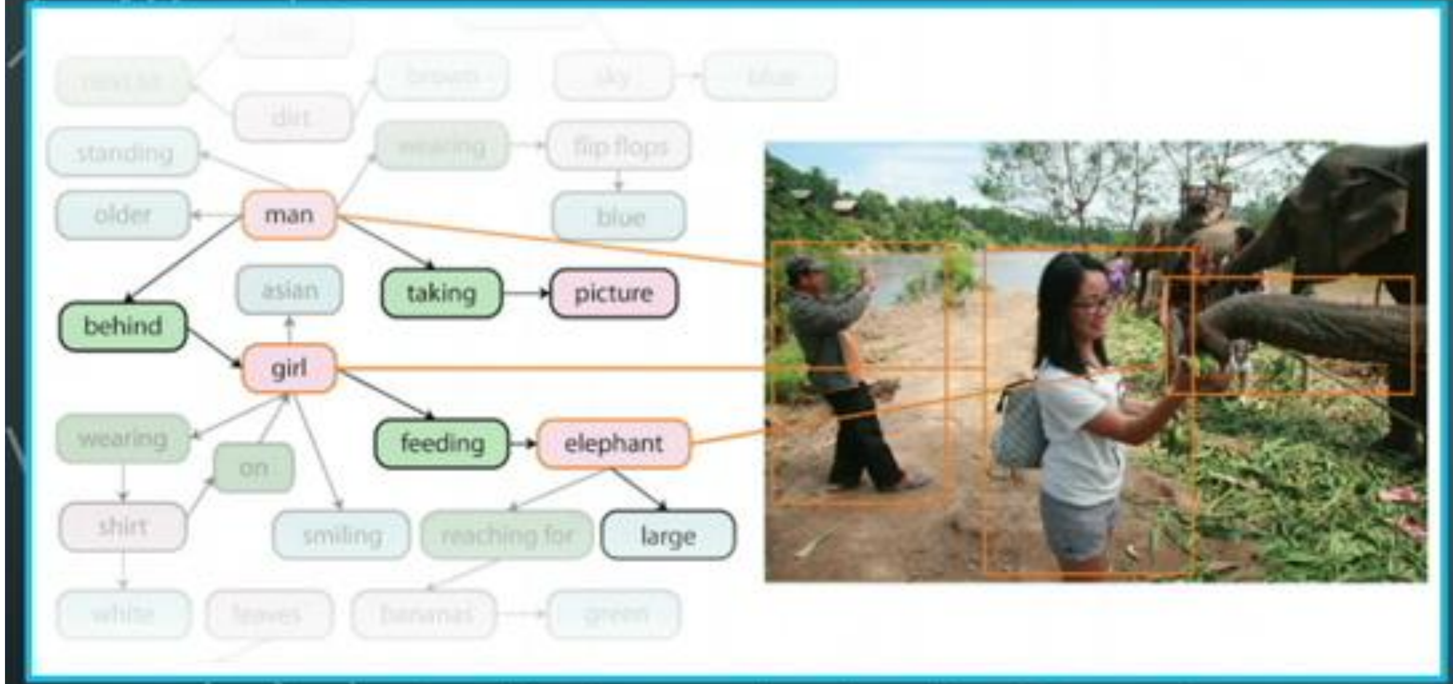


Figure from https://visualgenome.org/static/paper/Visual_Genome.pdf

Task:

Scene Graph Generation (SGG):

1. **PredCLS:** Predicate classification given (source,target) objs
2. **SgCLS:** Both obj classification and predicate classification "given" the ground truth bounding boxes
3. **SgDET:** Detecting bboxes using a backend (e.g.,Faster R-CNN), predicting obj classes and predicate classes

Datasets:

Several datasets to address each of above tasks, the most popular one is visual genome.

Methods:

Various methods proposed including iterative message passing from Stanford, Neural Motifs from UW, etc (A complete up to date list http://picdataset.com/challenge/paper_list/)