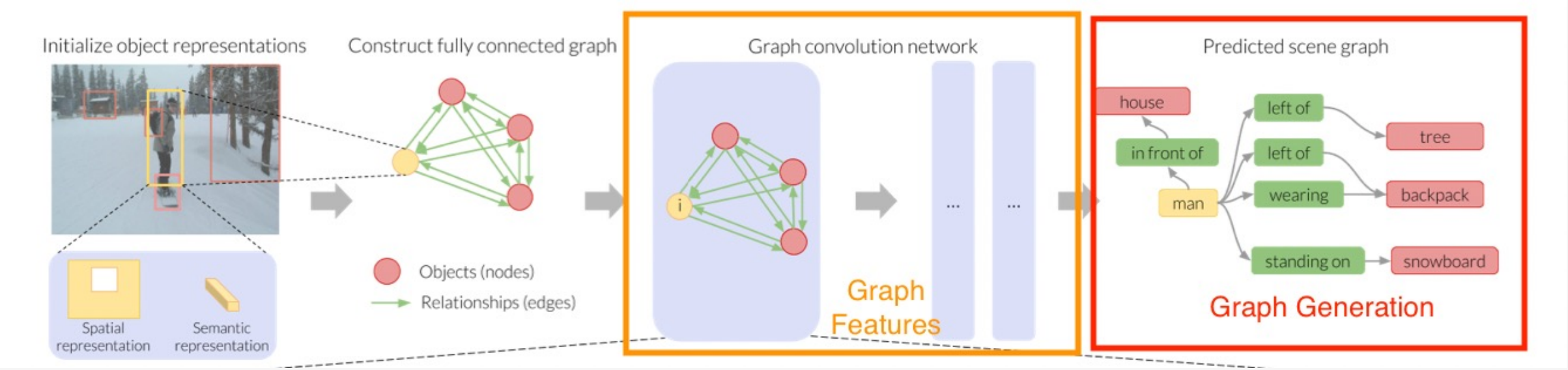
Scene Graph Generation

Efficient Graph Generation

汇报人：王健

问题介绍



Efficient Graph Generation：准确率+速度

Methods

- 准确率
  - Two-stage
    - [ECCV 2016] Visual Relationship Detection with Language Priors
    - [IEEE 2017] Visual Relationship Detection with Internal and External Linguistic Knowledge Distillation
    - [CVPR 2018] Neural Motifs:Scene graph parsing with global context
    - [CVPR 2019] Learning to Compose Dynamic Tree Structures for Visual Contexts
  - One-stage
    - [CVPR 2017]Scene graph generation by iterative message passing
- 效率
  - Two-stage
    - [ECCV 2018] Graph R-CNN for Scene Graph Generation
    - [ECCV 2018] Factorizable Net: An Efficient Subgraph-based Framework for Scene Graph Generation
  - One-stage
    - [CVPR 2021] Fully Convolutional Scene Graph Generation

Visual Relationship Detection with Language Priors

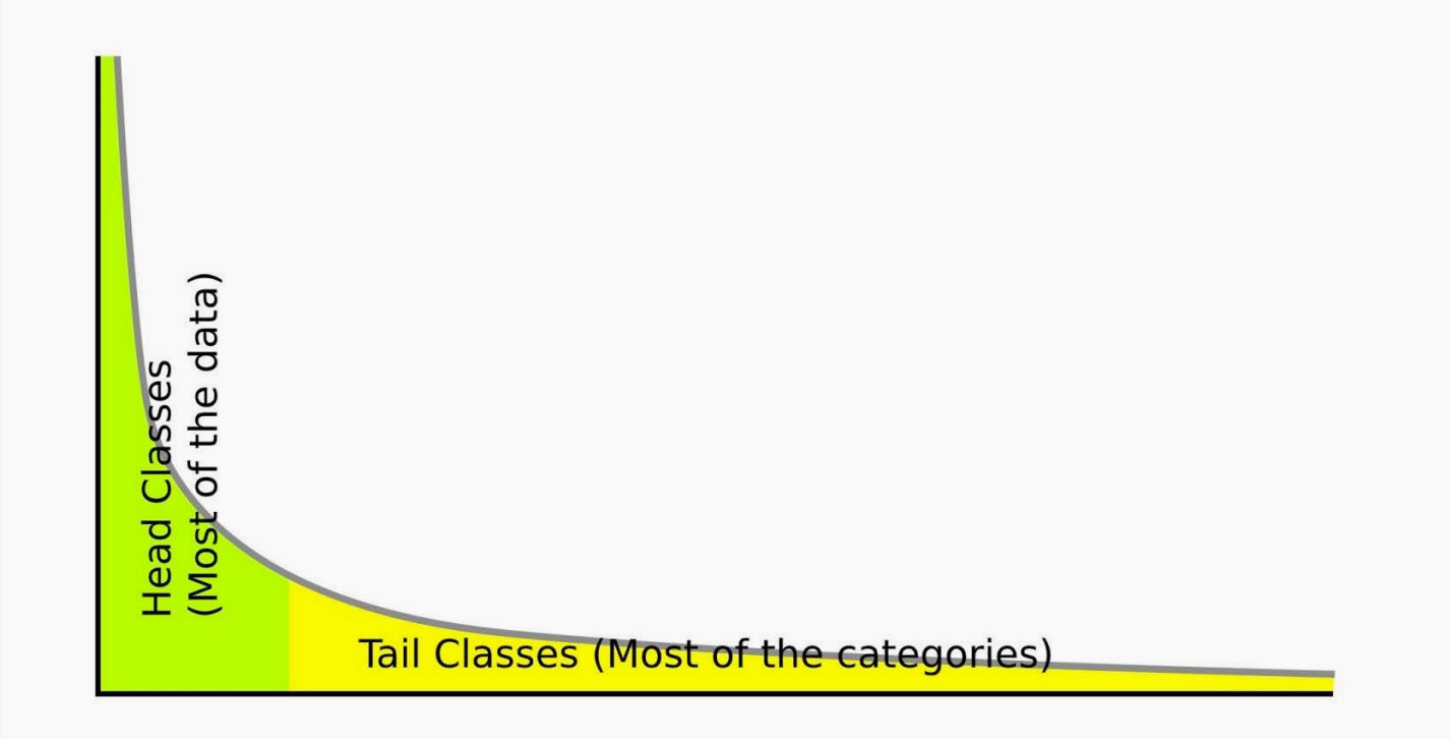
Cewu Lu, Ranjay Krishna, Michael Bernstein, Li Fei-Fei

Stanford University

ECCV 2016

Visual Relationship Detection with Language Priors [ECCV 2016] - Problem

- 可能关系的语义空间的大小远远大于对象的语义空间
- long-tail分布问题

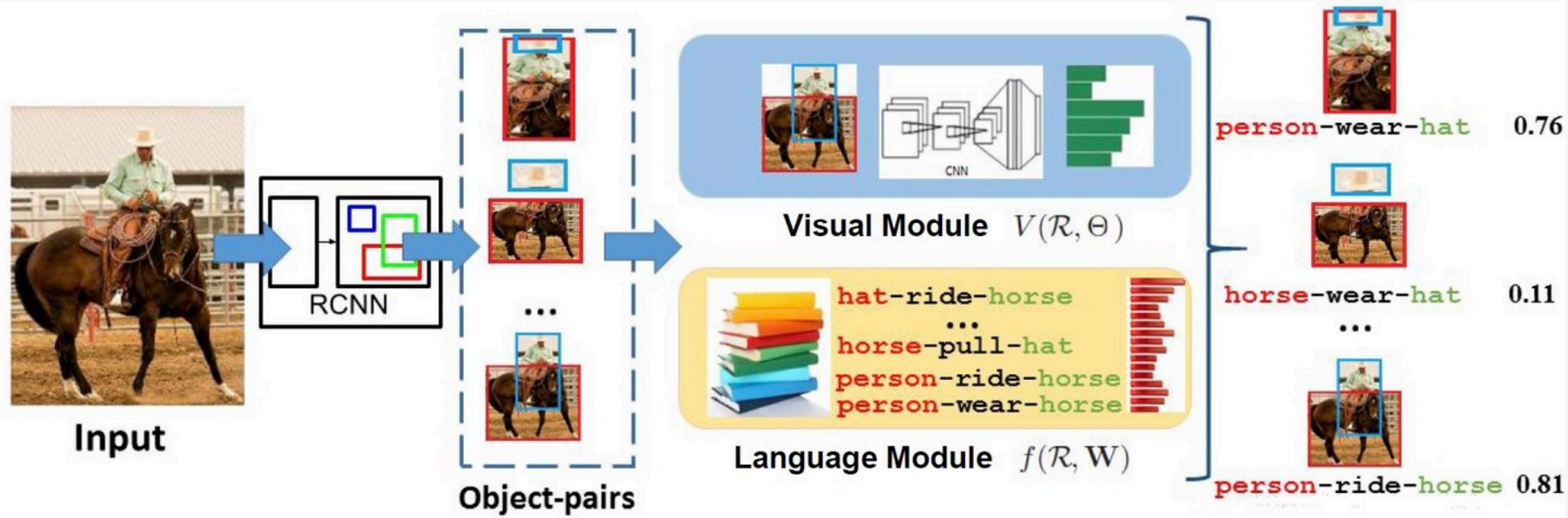


Visual Relationship Detection with Language Priors [ECCV 2016] - Approach

不常见的<S-P-O>关系推断可以通过物体和谓词的协助完成

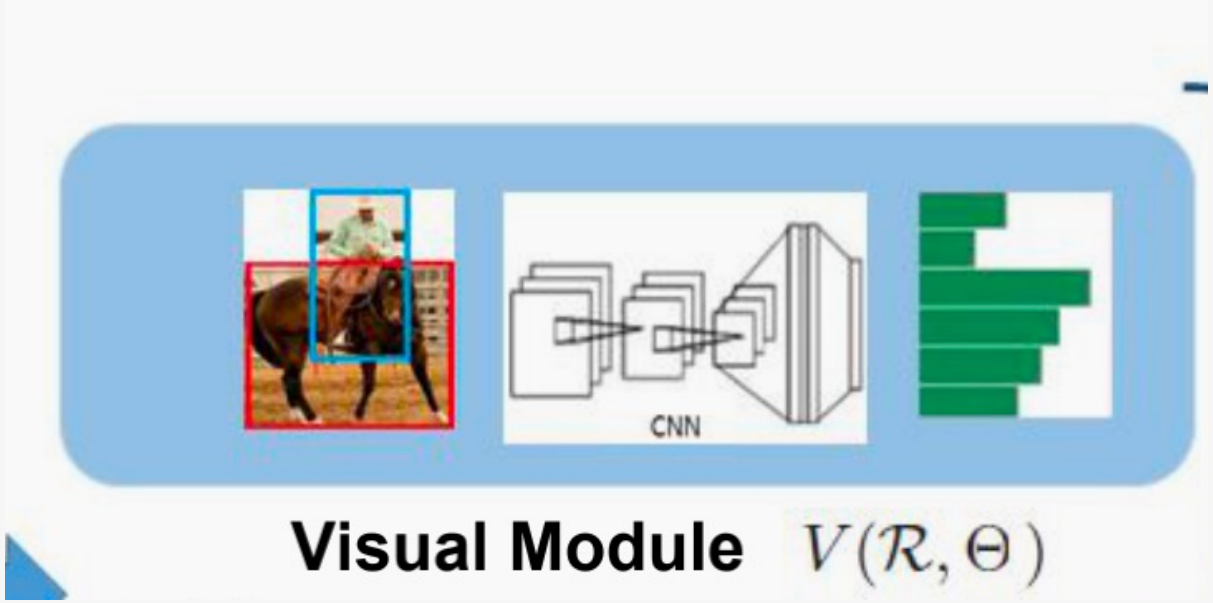
relationship之间有semantic的关联

Visual Relationship Detection with Language Priors [ECCV 2016] - Approach



- Faster R-CNN
- Visual Appearance Module+Language Module
- Objective function

Visual Relationship Detection with Language Priors [ECCV 2016] - Visual Appearance Module



K是谓词类别

z_k^T , s_k 是将CNN特征转换成关系可能性的参数

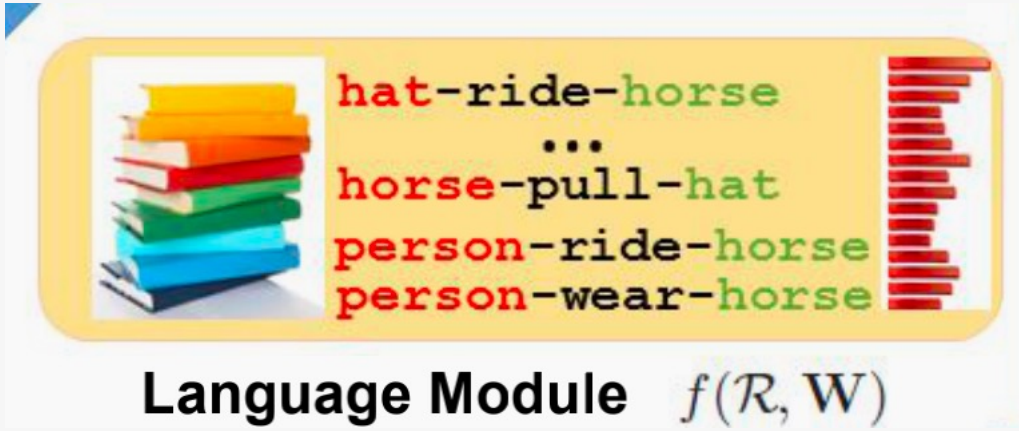
$P_i(O_1)$ 是CNN将盒子1归类为对象类别i的可能性

$CNN(O_1, O_2)$

是从1,2边界框的联合中提取的谓词CNN特征

$$V(R_{\langle i,k,j\rangle}, \Theta|\langle O_1, O_2\rangle) = P_i(O_1)(\mathbf{z}_k^T \mathrm{CNN}(O_1, O_2) + s_k)P_j(O_2)$$

Visual Relationship Detection with Language Priors [ECCV 2016] - Language Module

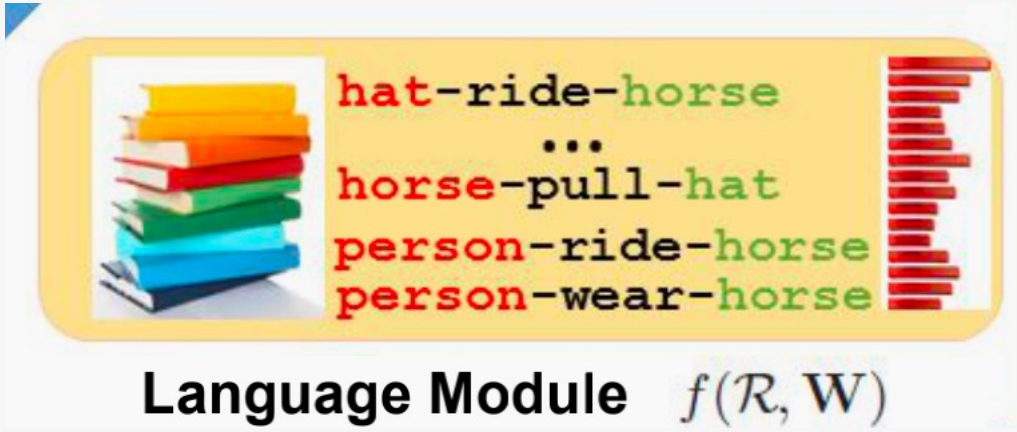


t_i 是第i个对象类别文本

公式得到的是数值而不是向量

$$f(\mathcal{R}_{\langle i,k,j\rangle}, \mathbf{W}) = \mathbf{w}_k^T[word2vec(t_i), word2vec(t_j)] + b_k$$

Visual Relationship Detection with Language Priors [ECCV 2016] - Language Module--trainning



$$f(\mathcal{R}_{\langle i,k,j \rangle}, \mathbf{W}) = \mathbf{w}_k^T [word2vec(t_i), word2vec(t_j)] + b_k$$

$$\frac{[f(\mathcal{R}, \mathbf{W}) - f(\mathcal{R}', \mathbf{W})]^2}{d(\mathcal{R}, \mathcal{R}')} = constant, \quad \forall \mathcal{R}, \mathcal{R}'$$

关系发生的可能性（发现1）：

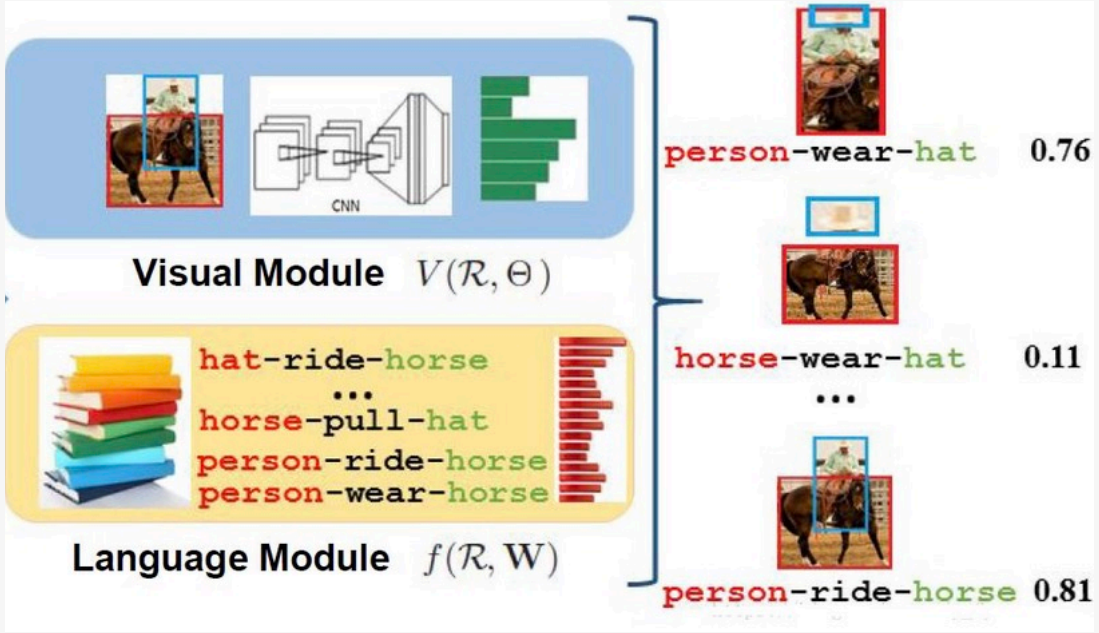
$$L(\mathbf{W}) = \sum_{\{\mathcal{R},\mathcal{R}'\}} \max\{f(\mathcal{R}', \mathbf{W}) - f(\mathcal{R}, \mathbf{W}) + 1, 0\}$$

使相似的关系预测的更加紧密（发现2）：

$$K(\mathbf{W}) = var(\{\frac{[f(\mathcal{R}, \mathbf{W}) - f(\mathcal{R}', \mathbf{W})]^2}{d(\mathcal{R}, \mathcal{R}')} \quad \forall \mathcal{R}, \mathcal{R}'\})$$

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}},$$

Visual Relationship Detection with Language Priors [ECCV 2016] - Objective function



$$C(\Theta, \mathbf{W}) = \sum_{\langle O_1 O_2 \rangle, \mathcal{R}} \max\{1 - V(\mathcal{R}, \Theta | \langle O_1, O_2 \rangle) f(\mathcal{R}, \mathbf{W}) + \max_{\langle O_1', O_2' \rangle \neq \langle O_1, O_2 \rangle, \mathcal{R}' \neq \mathcal{R}} V(\mathcal{R}', \Theta | \langle O_1', O_2' \rangle) f(\mathcal{R}', \mathbf{W}), 0\}$$

$$\min_{\Theta, \mathbf{W}} \{C(\Theta, \mathbf{W}) + \lambda_1 L(\mathbf{W}) + \lambda_2 K(\mathbf{W})\}$$

$$L(\mathbf{W}) = \sum_{\{\mathcal{R}, \mathcal{R}'\}} \max\{f(\mathcal{R}', \mathbf{W}) - f(\mathcal{R}, \mathbf{W}) + 1, 0\}$$

$$K(\mathbf{W}) = var(\{\frac{[f(\mathcal{R}, \mathbf{W}) - f(\mathcal{R}', \mathbf{W})]^2}{d(\mathcal{R}, \mathcal{R}')} \quad \forall \mathcal{R}, \mathcal{R}'\})$$

Test:

$$\mathcal{R}^* = \arg\max_{\mathcal{R}} V(\mathcal{R}, \Theta | \langle O_1, O_2 \rangle) f(\mathcal{R}, \mathbf{W})$$

Visual Relationship Detection with Language Priors [ECCV 2016] - Objective function

	Phrase Det.		Relationship Det.		Predicate Det.	
	R@100	R@50	R@100	R@50	R@100	R@50
Ours - V only	1.12	0.95	0.78	0.67	3.52	3.52
Ours - L only	0.01	0.00	0.01	0.00	5.09	5.09
Ours - V + L only	2.56	2.43	2.66	2.27	6.11	6.11
Ours - V + L + K	**3.75**	**3.36**	**3.52**	**3.13**	**8.45**	**8.45**

	Recall @ 1	Recall @ 5	Recall @ 10	Median Rank
GIST [46]	0.00	5.60	8.70	68
SIFT [47]	0.70	6.10	10.3	54
CNN [44]	3.15	7.70	11.5	20
Visual Phrases [6]	8.72	18.12	28.04	12
Our Model	**10.82**	**30.02**	**47.00**	**4**

Visual Relationship Detection with Internal and External Linguistic Knowledge Distillation

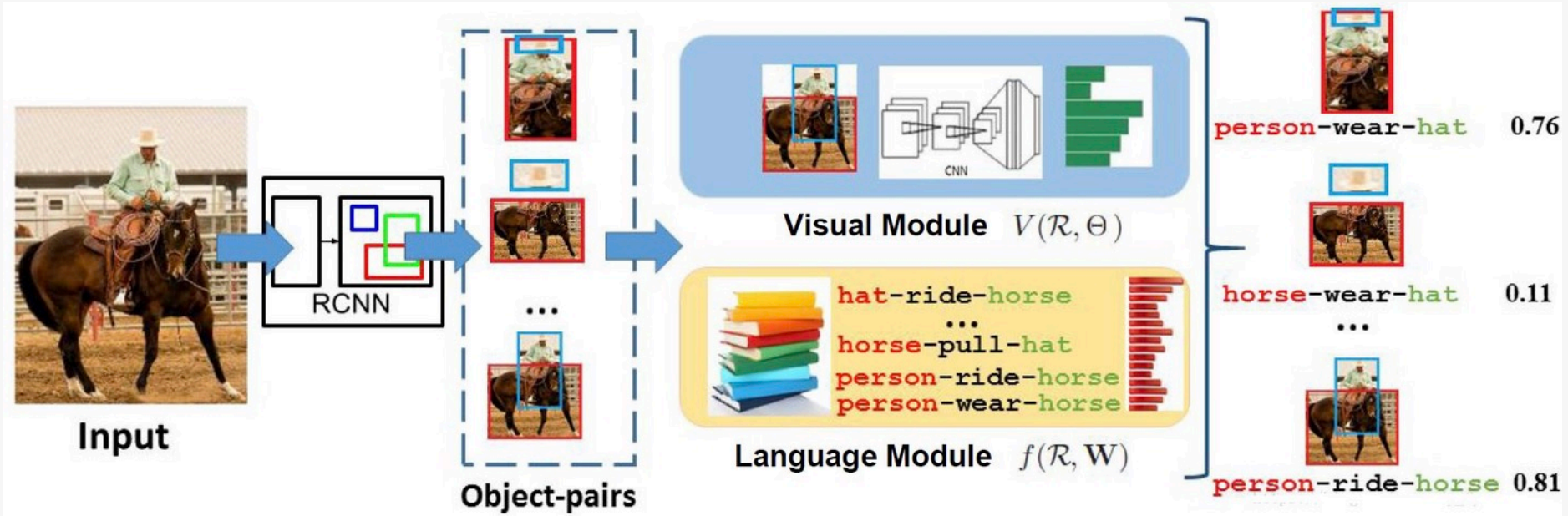
Ruichi Yu, Ang Li, Vlad I. Morariu, Larry S. Davis

University of Maryland, College Park

EEE 2017

Detection with Internal and External Linguistic Knowledge Distillation - Problem

- 准确率
  - Detection with Language Priors没由此充分利用视觉信息。
- 效率
  - 大量关系三元组，模型参数空间大。训练不能仅仅依靠标签数据



Detection with Internal and External Linguistic Knowledge Distillation - Approach

Detection with Language Priors没由此充分利用视觉信息

$$P(R|I) = P(pred|I_{\text{union}}, subj, obj)P(subj)P(obj).$$

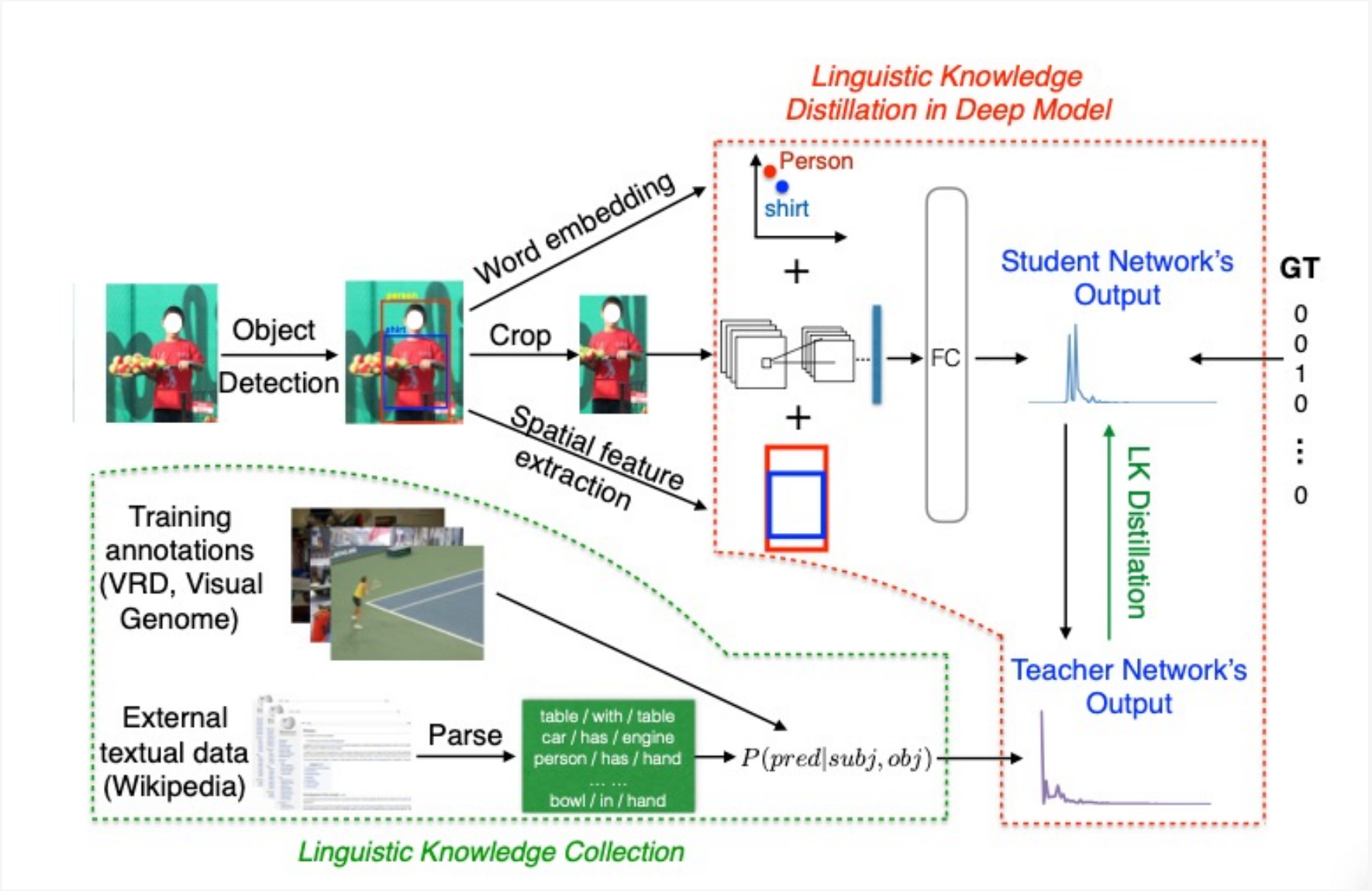
大量关系三元组，模型参数空间大。训练不能仅仅依靠标签数据

- internal statistic (e.g. statistics of the VRD dataset)
- external linguistic knowledge (e.g. from Wikipedia)

- student network
- teacher network

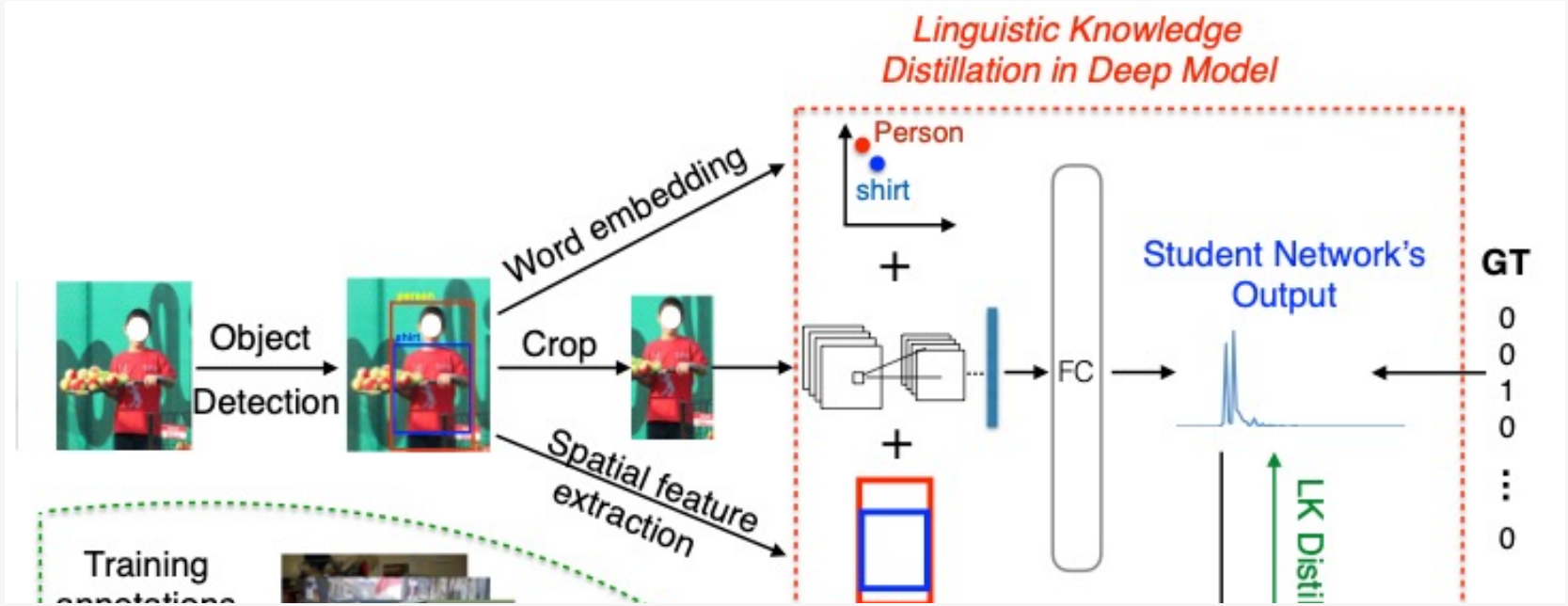
$$P(pred|subj, obj)$$

Detection with Internal and External Linguistic Knowledge Distillation - Approach



- student network
- teacher network

Detection with Internal and External Linguistic Knowledge Distillation - student network



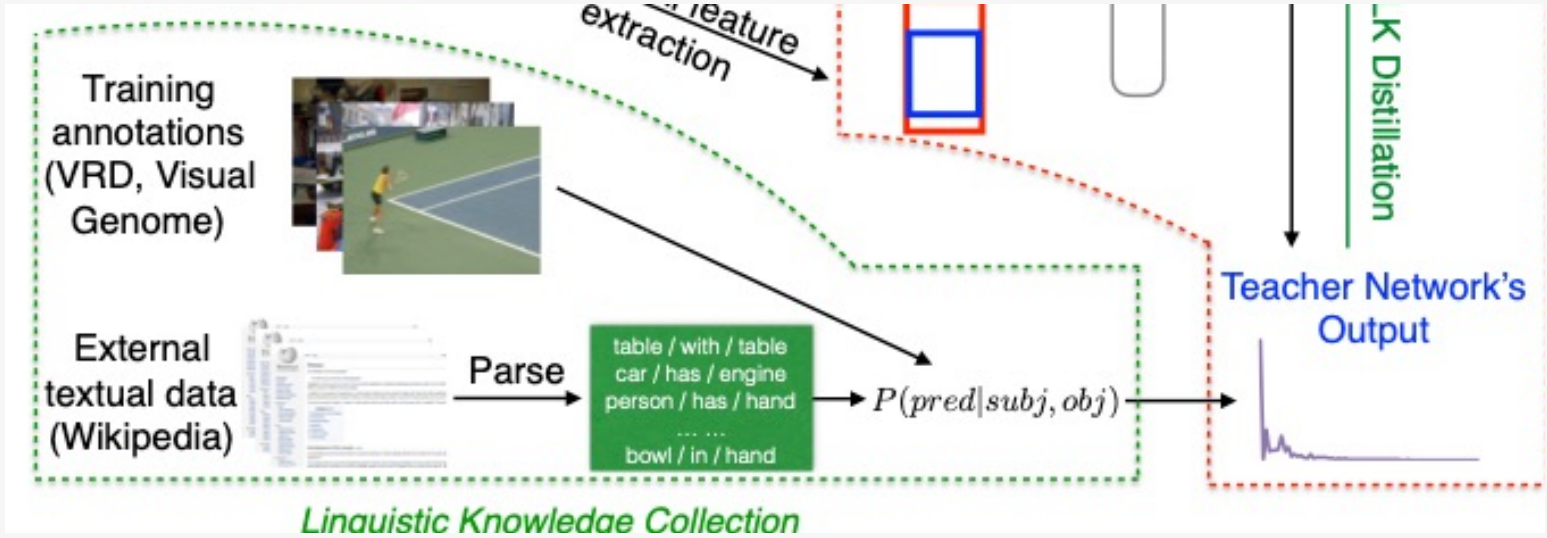
student network输入：

- 两个检测区域的并集BB-Union
- 物体对象语意表示
- 一对边界盒中提取的空间特征

$$\left[\frac{x_{min}}{W}, \frac{y_{min}}{H}, \frac{x_{max}}{W}, \frac{y_{max}}{H}, \frac{A}{A_{img}}\right],$$

$$P(R|I) = P(pred|subj, obj, B_s, B_o, I) \cdot P(subj, B_s|I)P(obj, B_o|I),$$

Detection with Internal and External Linguistic Knowledge Distillation - Approach



- t（Y）和ϕ（Y | X）是教师和学生网络的预测结果
- C是一个平衡项
- ϕ是学生网络的参数集
- L（X，Y）是一个一般的约束函数，它具有很高的值来奖励满足约束的预测，惩罚其他的预测
- KL度量教师和学生预测分布的KL-divergence

teacher network是通过优化以下标准来构建:

$$\min_{t \in T} \mathrm{KL}(t(Y) || s_\phi(Y|X)) - C\mathbb{E}_t[L(X,Y)],$$

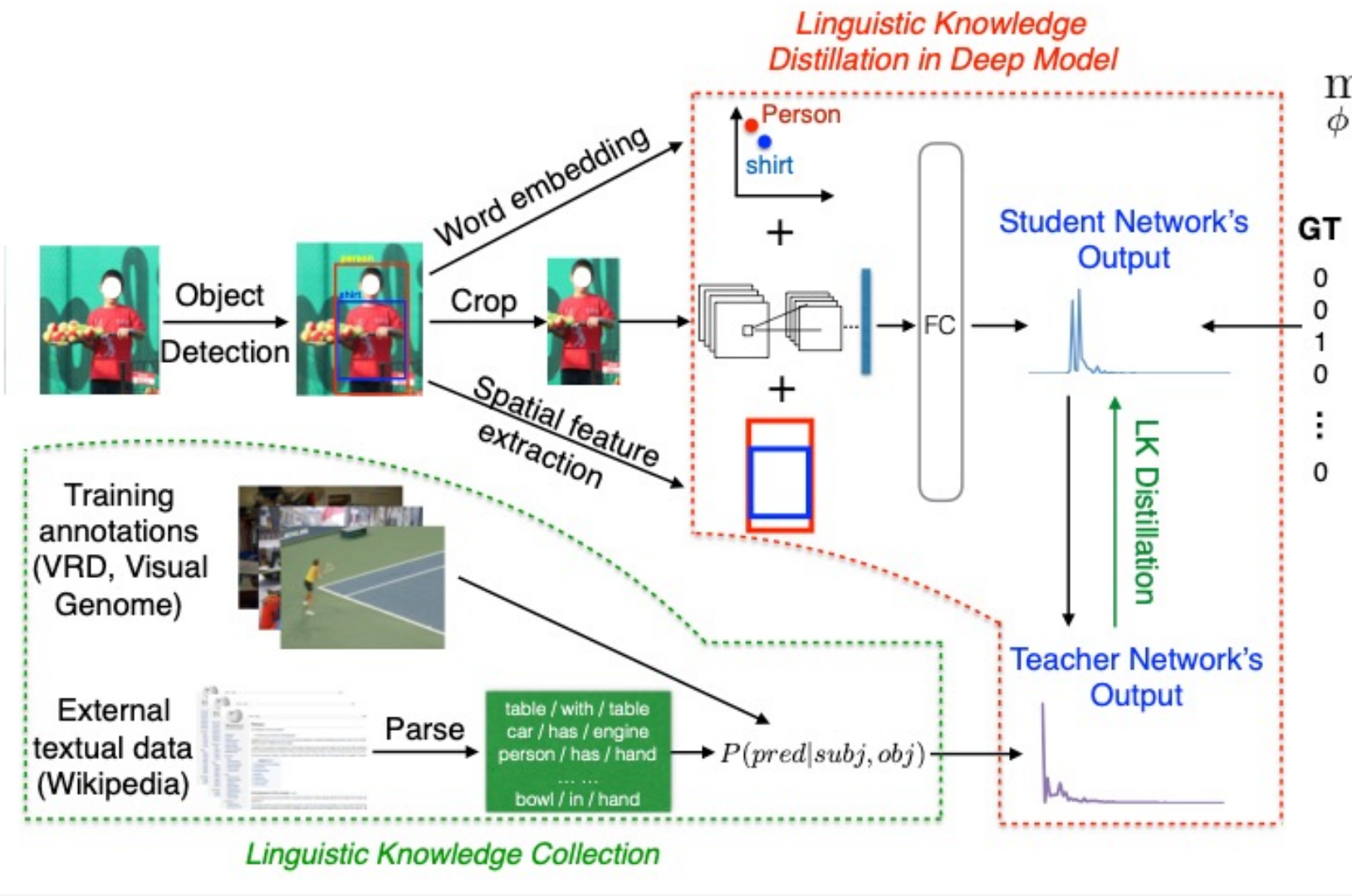
$$d(P,Q) = D_{KL}(P||Q) = \sum_i P(i) \log(\frac{P(i)}{Q(i)})$$

优化问题的闭式解是:

$$t(Y) \propto s(Y|X)\exp(CL(X,Y)) \ .$$

$$L(X,Y) = \log P(pred|subj, obj),$$

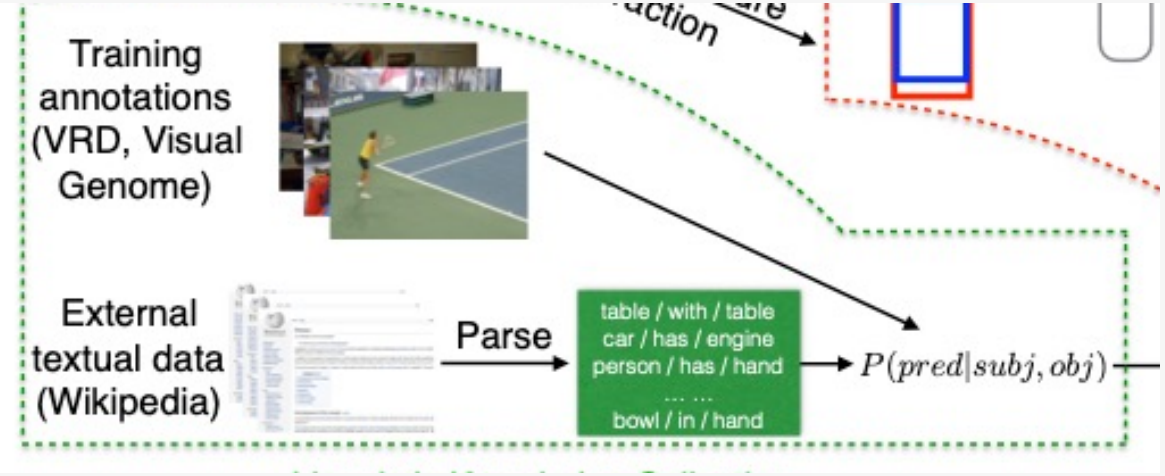
Detection with Internal and External Linguistic Knowledge Distillation - Approach



$$\min_{\phi \in \Phi} \frac{1}{n} \sum_{i=1}^{n} \alpha l(s_i, y_i) + (1-\alpha) l(s_i, t_i)$$

- s_i, t_i是学生和老师对样本i的预测，
- y_i是样本i的ground truth
- α 是ground truth和老师网络输出之间的平衡项
- l 是loss function

Detection with Internal and External Linguistic Knowledge Distillation - Knowledge Collection



Filename	Language	Date of the dump	Filesize
wikipedia.txt.dump.20140616-de.SZTAKI.torrent	de	16/06/2014	2.74 GiB
wikipedia.txt.dump.20140615-en.SZTAKI.torrent	en	15/06/2014	6.76 GiB
wikipedia.txt.dump.20140608-hu.SZTAKI.torrent	hu	08/06/2014	398.41 MiB
wikipedia.txt.dump.20140608-fr.SZTAKI.torrent	fr	08/06/2014	1.79 GiB
wikipedia.txt.dump.20130425-de.SZTAKI.torrent	de	25/04/2013	2.45 GiB
wikipedia.txt.dump.20130421-fr.SZTAKI.torrent	fr	21/04/2013	1.6 GiB
wikipedia.txt.dump.20130413-hu.SZTAKI.torrent	hu	13/04/2013	355.76 MiB
wikipedia.txt.dump.20130405-en.SZTAKI.torrent	en	05/04/2013	6.12 GiB
wikipedia.txt.dump.20120722-ar.SZTAKI.torrent	ar	22/07/2012	189.67 MiB
wikipedia.txt.dump.20120117-fr.SZTAKI.torrent	fr	17/01/2012	1.36 GiB
wikipedia.txt.dump.20120117-de.SZTAKI.torrent	de	17/01/2012	2.08 GiB
wikipedia.txt.dump.20120106-hu.SZTAKI.torrent	hu	06/01/2012	302.07 MiB
wikipedia.txt.dump.20120105-en.SZTAKI.torrent	en	05/01/2012	5.34 GiB
wikipedia.txt.dump.20111231-fr.SZTAKI.torrent	fr	31/12/2011	1.35 GiB
wikipedia.txt.dump.20111230-de.SZTAKI.torrent	de	30/12/2011	2.07 GiB
wikipedia.txt.dump.20111205-hu.SZTAKI.torrent	hu	05/12/2011	302.41 MiB

Detection with Internal and External Linguistic Knowledge Distillation - Result

Table 2. Phrase and Relationship Detection: Distillation of Linguistic Knowledge. We use the same notations as in Table 1.

	Phrase Detection						Relationship Detection					
	R@100, k=1	R@50, k=1	R@100, k=10	R@50, k=10	R@100, k=70	R@50, k=70	R@100, k=1	R@50, k=1	R@100, k=10	R@50, k=10	R@100, k=70	R@50, k=70
Part 1: Training images VRD only												
Visual Phrases [26]	0.07	0.04	-	-	-	-	-	-	-	-	-	
Joint CNN [6]	0.09	0.07	-	-	-	-	0.09	0.07	-	-	-	-
VRD - V only [19]	2.61	2.24	-	-	-	-	1.85	1.58	-	-	-	-
VRD - Full [19]	17.03	16.17	25.52	20.42	24.90	20.04	14.70	13.86	22.03	17.43	21.51	17.35
Linguistic Cues [25]	-	-	20.70	16.89	–	–	–	–	18.37	15.08	–	–
VIP-CNN [17]	**27.91**	22.78	-	-	–	–	20.01	17.32	-	-	–	–
VRL [18]	22.60	21.37	-	-	–	–	20.79	18.19	-	-	–	–
U+W+SF+L: S	19.98	19.15	25.16	22.95	25.54	22.59	17.69	16.57	27.98	19.92	28.94	20.12
U+W+SF+L: T	23.57	22.46	29.14	25.96	29.09	25.86	20.61	18.56	29.41	21.92	31.13	21.98
U+W+SF+L: T+S	24.03	**23.14**	**29.76**	**26.47**	**29.43**	**26.32**	**21.34**	**19.17**	**29.89**	**22.56**	**31.89**	**22.68**
Part 2: Training images VRD + VG												
U+W+SF+L: S	20.32	19.96	25.71	23.34	25.97	22.83	18.32	16.98	28.24	20.15	29.85	21.88
U+W+SF+L: T	23.89	22.92	29.82	26.34	29.97	26.15	20.94	18.93	29.95	22.62	31.78	22.65
U+W+SF+L: T+S	24.42	23.51	30.13	26.73	30.01	26.58	21.72	19.68	30.45	22.84	32.56	23.18

Scene graph generation by iterative message passing

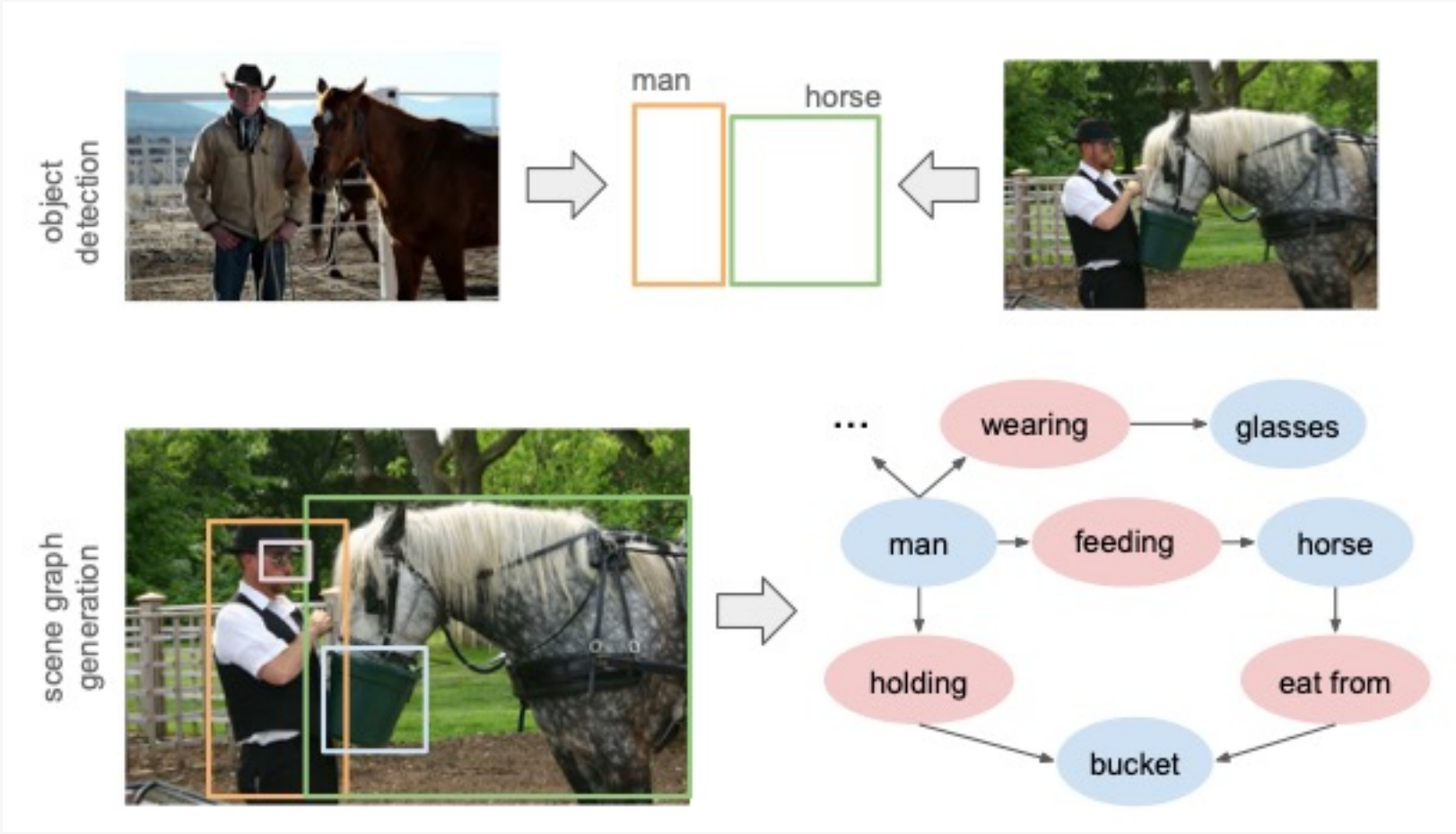
Danfei Xu, Yuke Zhu, Christopher B. Choy, Li Fei-Fei,

Department of Computer Science, Stanford University

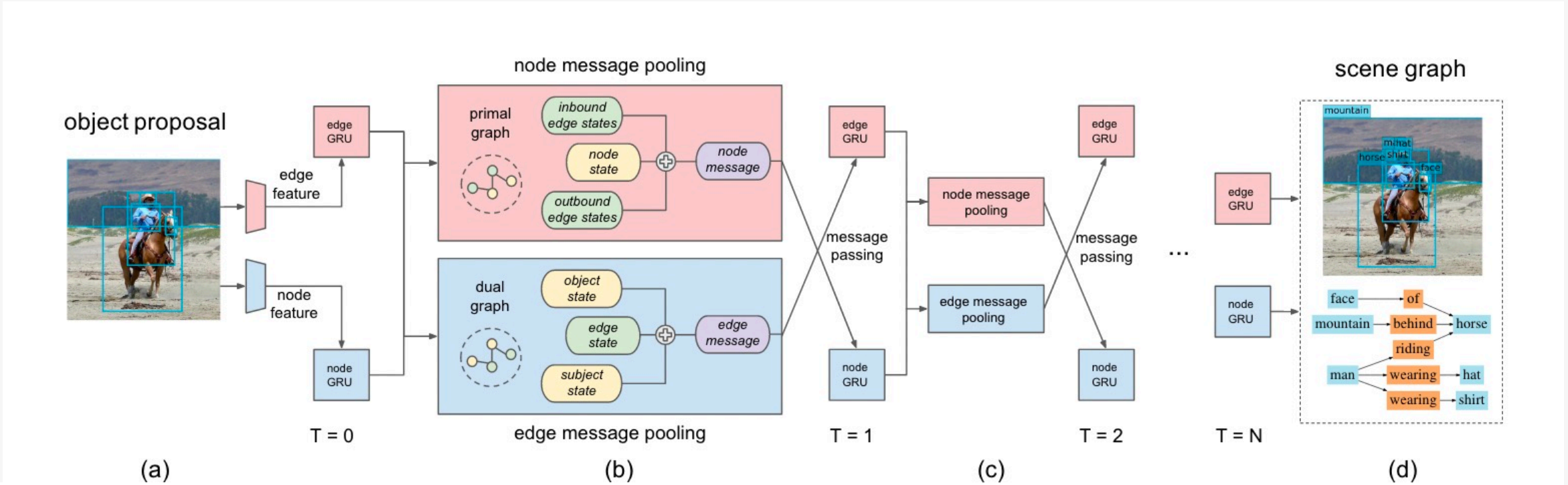
Department of Electrical Engineering, Stanford University

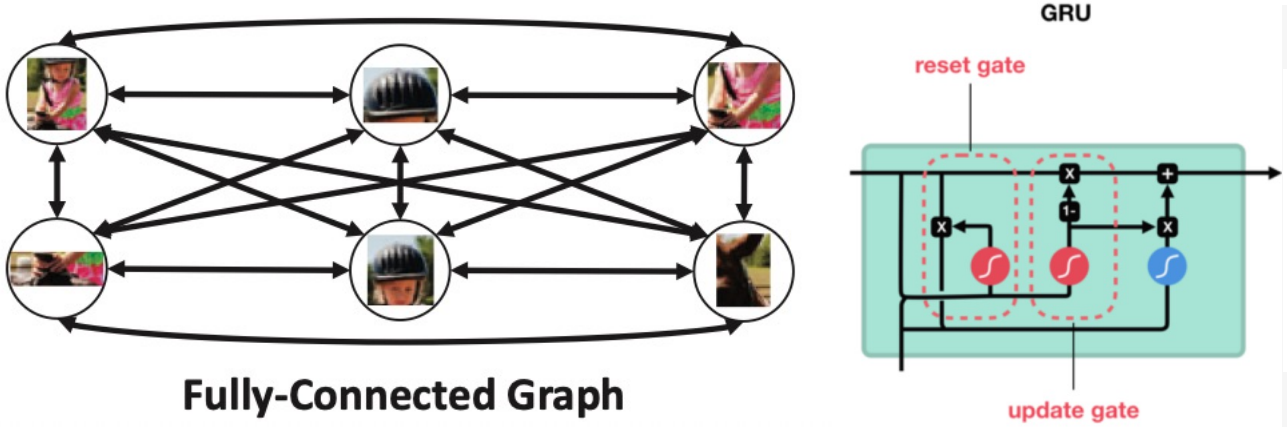
CVPR 2017

Scene graph generation by iterative message passing - Problem



Scene graph generation by iterative message passing - Approach

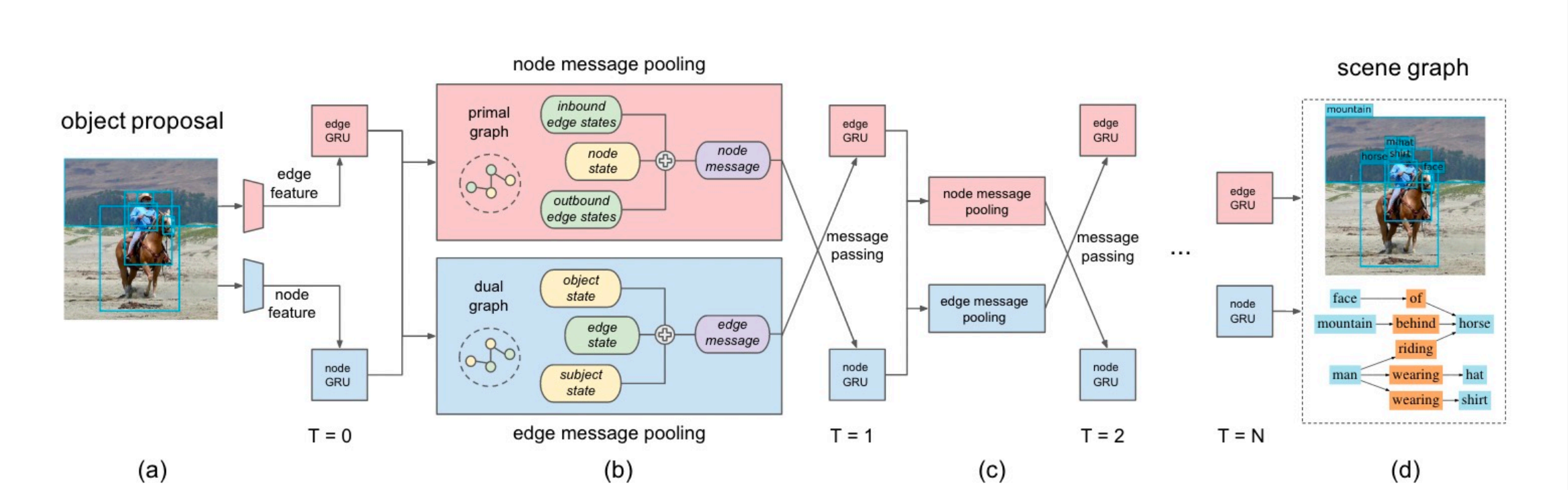




$$m_i = \sum_{j:i\to j} \sigma(\mathbf{v}_1^T[h_i, h_{i\to j}])h_{i\to j} + \sum_{j:j\to i} \sigma(\mathbf{v}_2^T[h_i, h_{j\to i}])h_{j\to i} \quad (3)$$

$$m_{i\to j} = \sigma(\mathbf{w}_1^T[h_i, h_{i\to j}])h_i + \sigma(\mathbf{w}_2^T[h_j, h_{i\to j}])h_j \quad (4)$$

Scene graph generation by iterative message passing - Approach



$$MAE = \frac{\sum_{n=1}^{n} | f(x_i) - y_i |}{n}$$

Neural Motifs: Scene graph parsing with global context

Rowan Zellers, Mark Yatskar, Sam Thomson, Yejin Choi Paul G. Allen

School of Computer Science & Engineering,

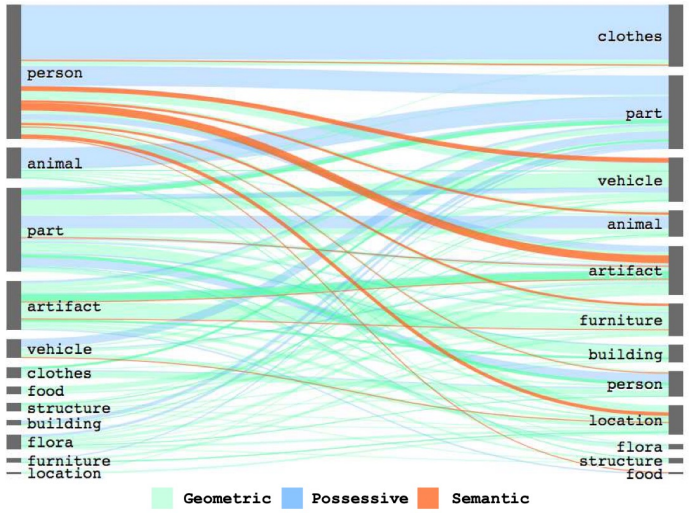
University of Washington Allen Institute for Artificial Intelligence

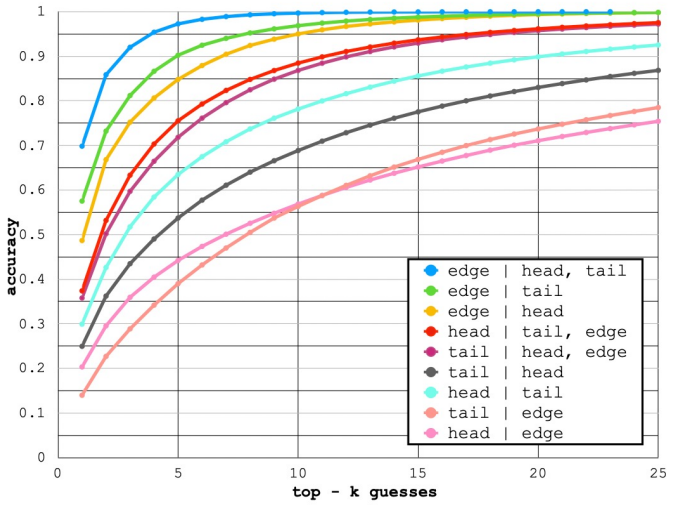
School of Computer Science, Carnegie Mellon University

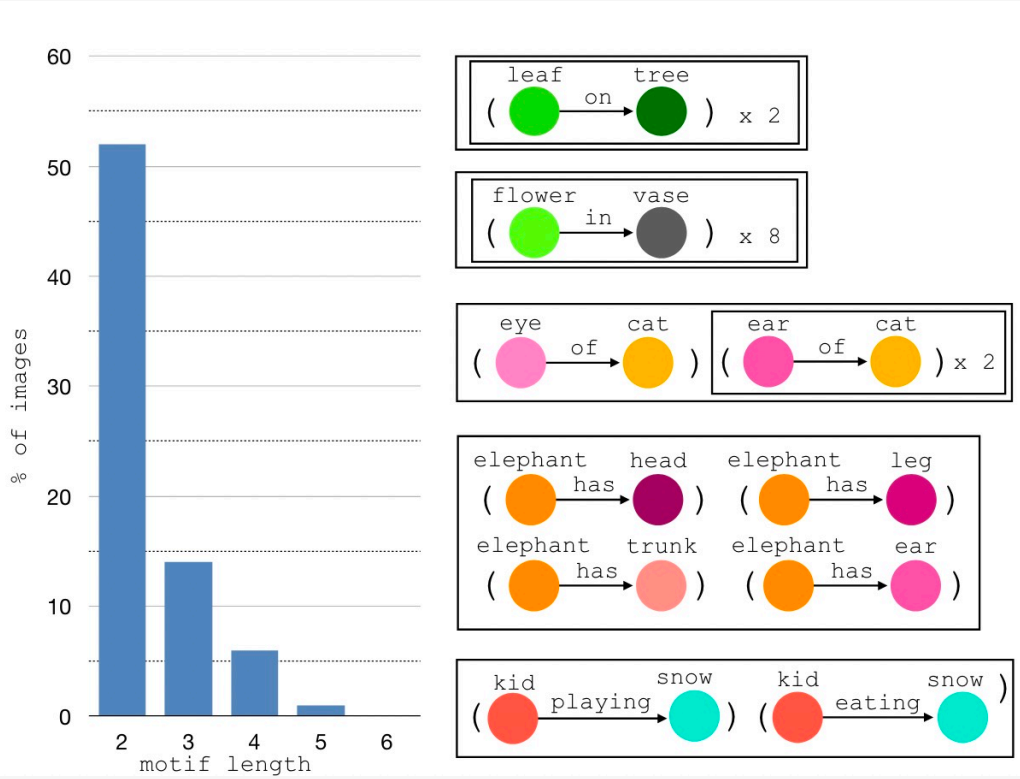
CVPR 2018

Neural Motifs: Scene graph parsing with global context - Motivation

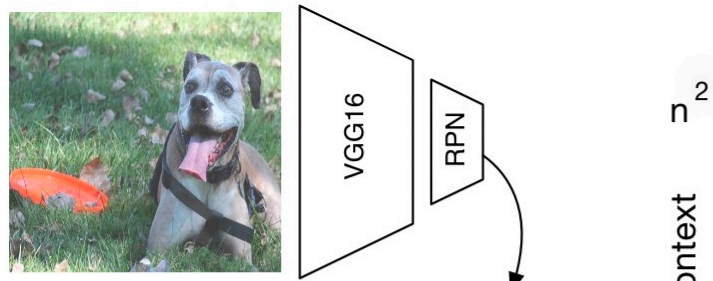
- 物体标签对关系推理的帮助
- Motifs的提出

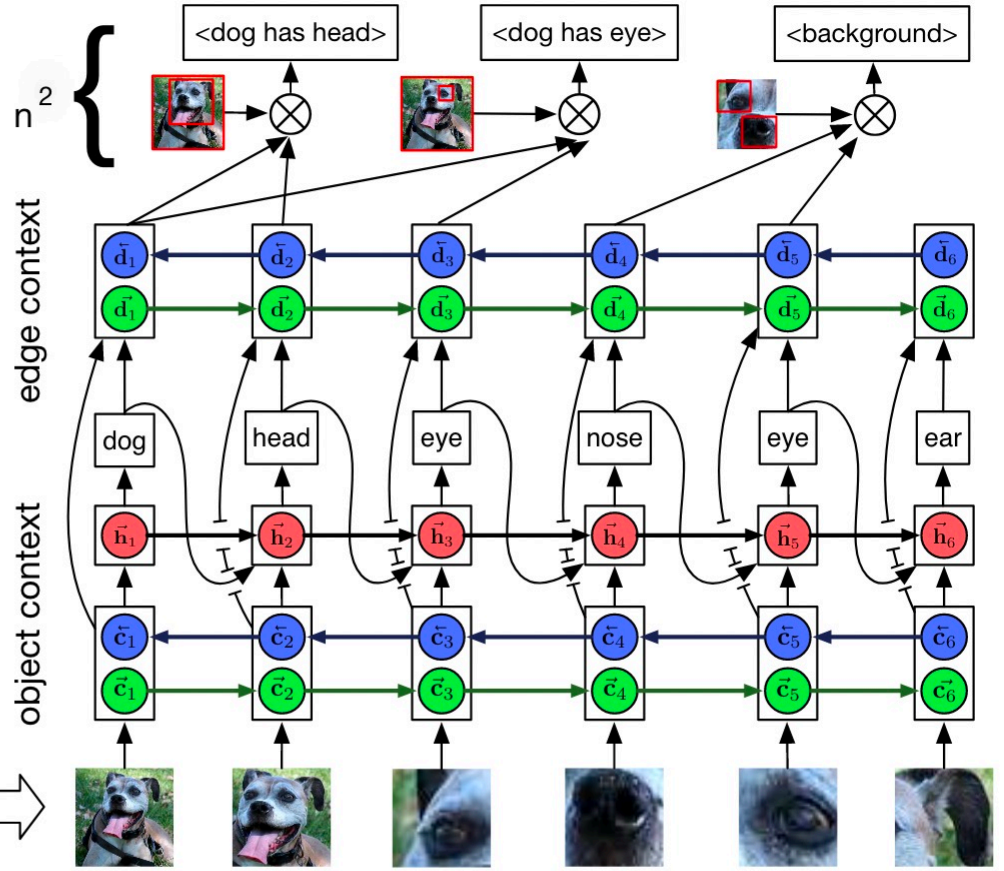






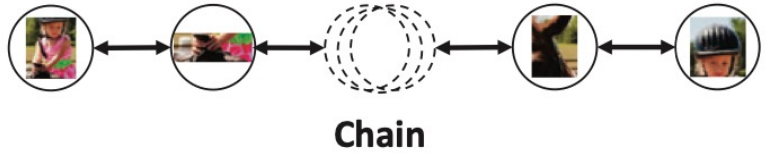
Neural Motifs: Scene graph parsing with global context - Problem



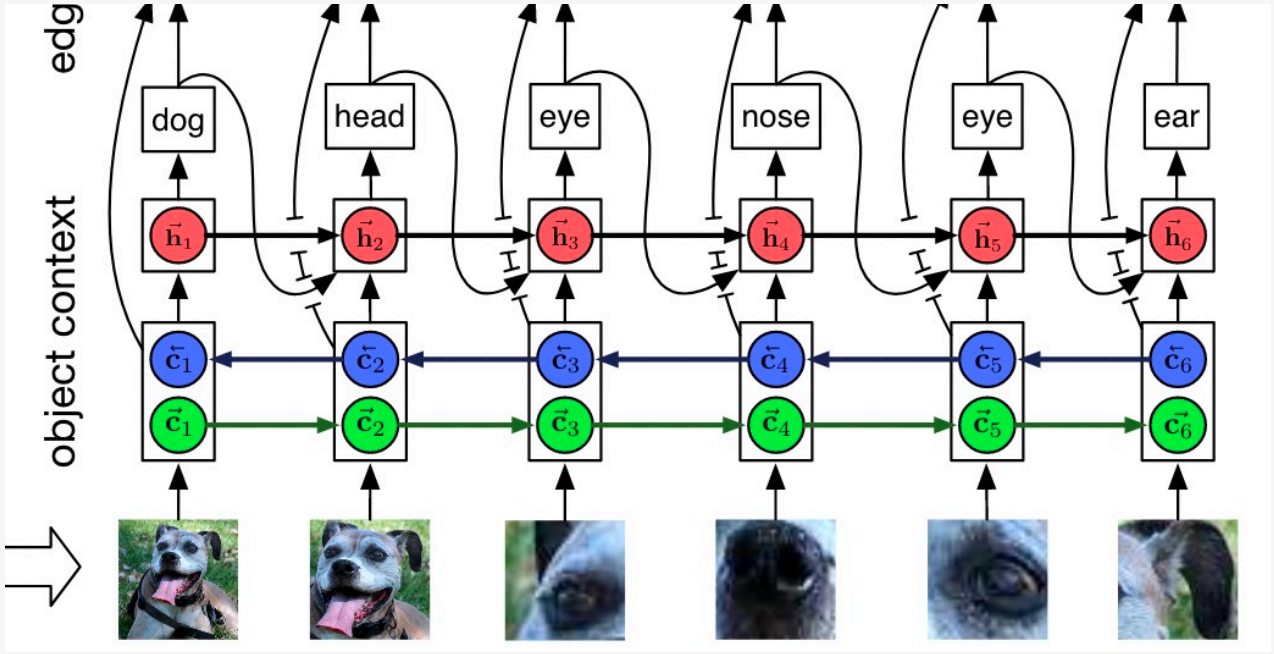


- 物体区域提取
  - Faster R-CNN
- 物体标签信息
  - bi-LSTM
  - LSTM
- 物体关系
  - bi-LSTM
  - softmax

$$\Pr(G \mid I) = \Pr(B \mid I) \Pr(O \mid B, I) \Pr(R \mid B, O, I). \quad (1)$$



Neural Motifs: Scene graph parsing with global context - object context

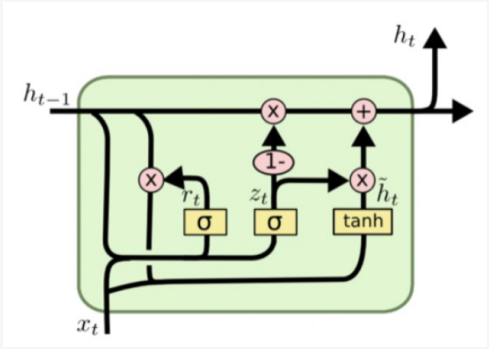


$$\Pr(O \mid B, I)$$

Encoding:

$$\mathbf{C} = \text{biLSTM}([\mathbf{f}_i; \mathbf{W}_1 \mathbf{l}_i]_{i=1,\ldots,n}),$$

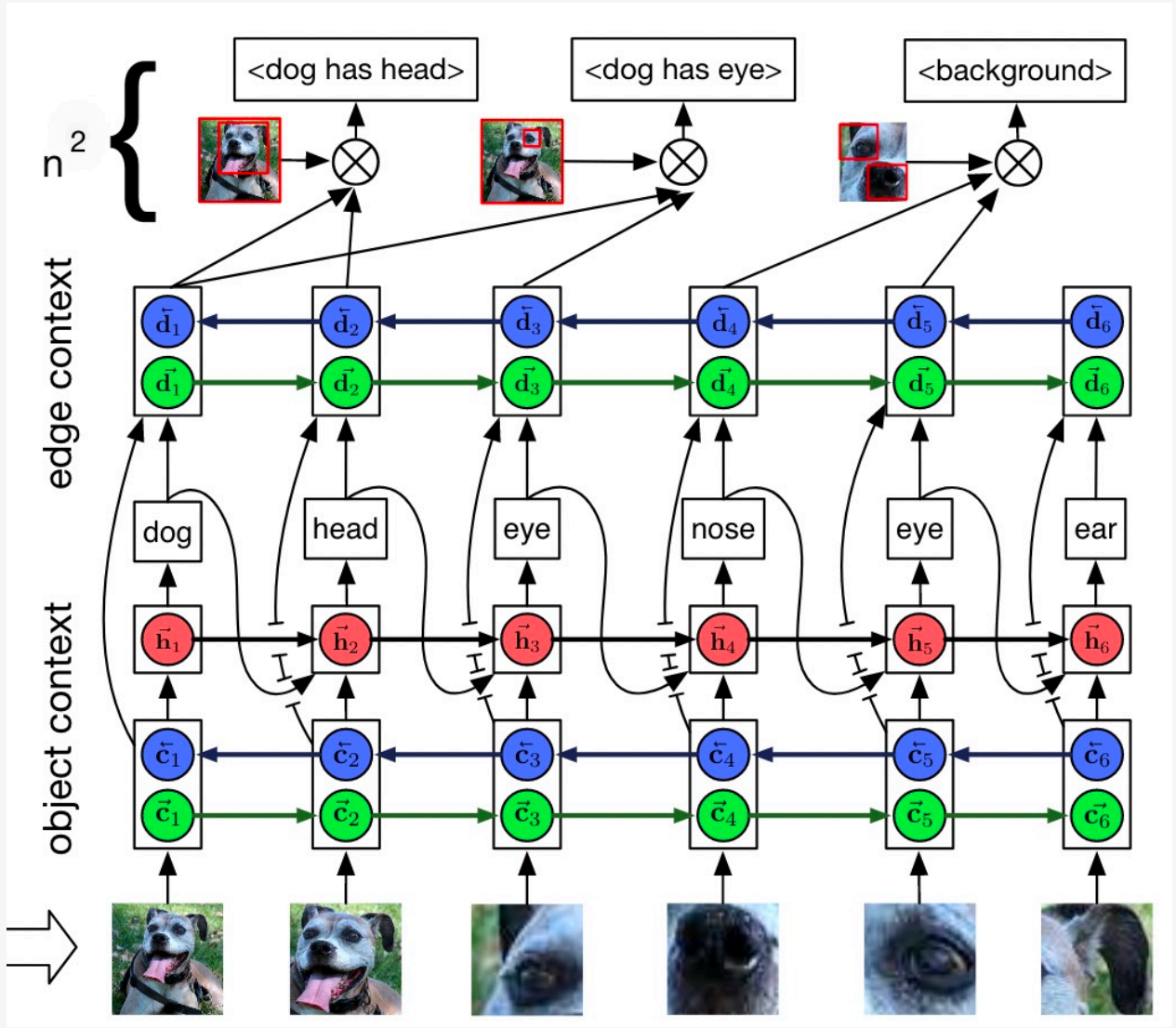
Decoding:



$$\mathbf{h}_i = \text{LSTM}_i\left([\mathbf{c}_i; \hat{\mathbf{o}}_{i-1}]\right)$$

$$\hat{\mathbf{o}}_i = \text{argmax}\left(\mathbf{W}_o\, \mathbf{h}_i\right) \in \mathbb{R}^{|\mathcal{C}|} \text{ (one-hot)}$$

Neural Motifs: Scene graph parsing with global context - edge context



$$\Pr(R \mid B, O, I).$$

Encoding:

$$\mathbf{D} = \text{biLSTM}([\mathbf{c}_i; \mathbf{W}_2\hat{\mathbf{o}}_i]_{i=1,\ldots,n}), \tag{5}$$

where the *edge context* $\mathbf{D} = [\mathbf{d}_1, \ldots, \mathbf{d}_n]$ contains the states for each bounding region at the final layer, and $\mathbf{W}_2$ is a parameter matrix mapping $\hat{\mathbf{o}}_i$ into $\mathbb{R}^{100}$.

Decoding:

$$\mathbf{g}_{i,j} = (\mathbf{W}_h\mathbf{d}_i) \circ (\mathbf{W}_t\mathbf{d}_j) \circ \mathbf{f}_{i,j} \tag{6}$$

$$\Pr(x_{i \to j} \mid B, O) = \text{softmax}\left(\mathbf{W}_r\mathbf{g}_{i,j} + \mathbf{w}_{o_i,o_j}\right). \tag{7}$$

$\mathbf{W}_h$ and $\mathbf{W}_t$ project the head and tail context into $\mathbb{R}^{4096}$. $\mathbf{w}_{o_i,o_j}$ is a bias vector specific to the head and tail labels.

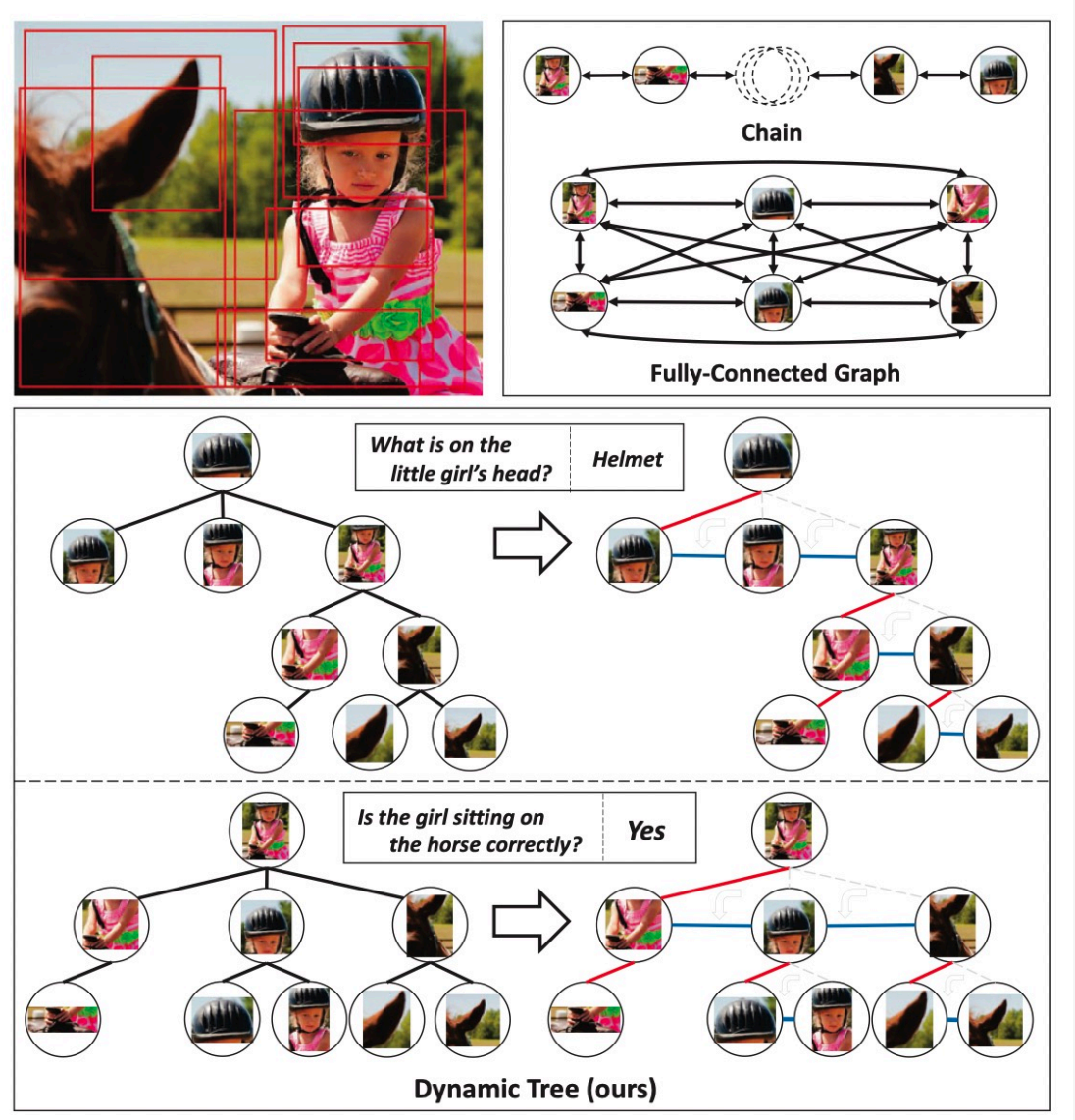
Learning to Compose Dynamic Tree Structures for Visual Contexts

Kaihua Tang, Hanwang Zhang, Baoyuan Wu, Wenhan Luo, Wei Liu

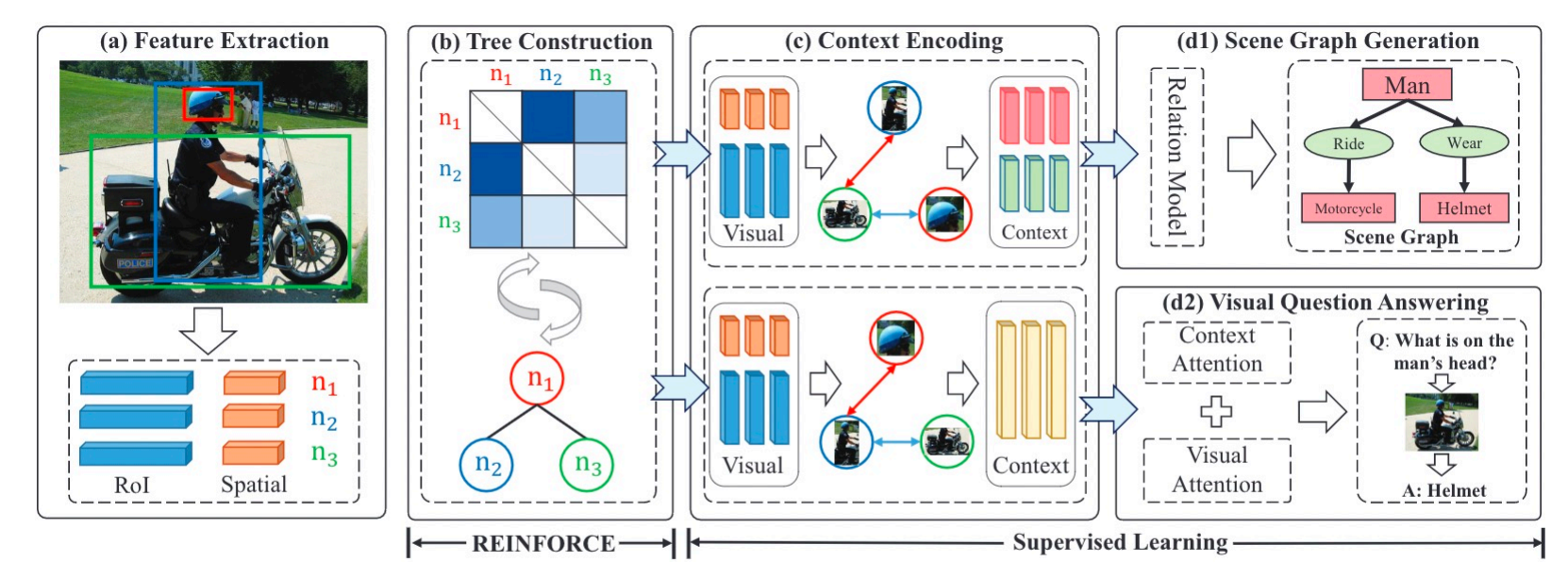
Nanyang Technological University, Tencent AI Lab

CVPR 2019

Learning to Compose Dynamic Tree Structures for Visual Contexts — Problem

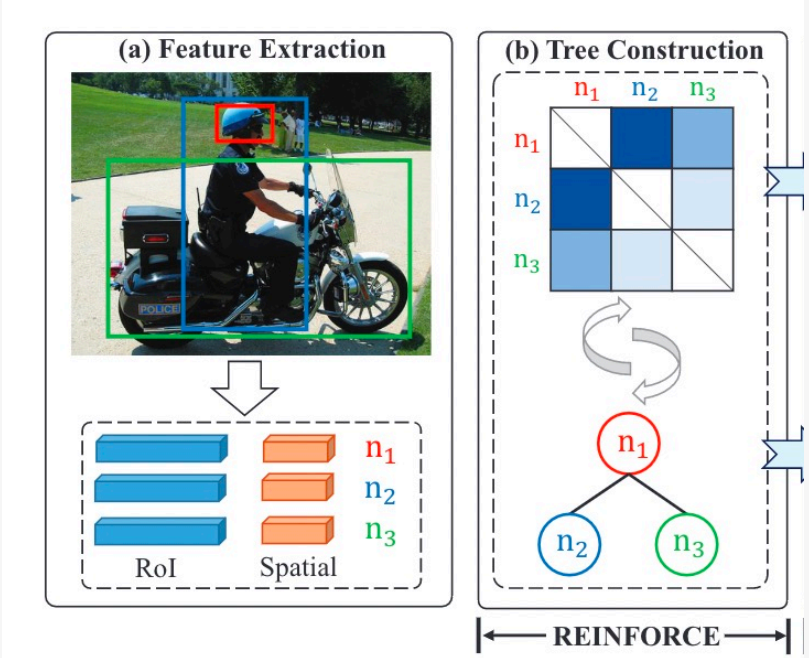


Learning to Compose Dynamic Tree Structures for Visual Contexts — Approach



- Faster-RCNN得到object proposals
- 学习参数矩阵用以构建VCTREE，树的构建采用hybrid learning
- 使用构建好的vctree，用 Bidirectional Tree LSTM encode contexual cues
- encoded context 根据不同的task进行解码。

Learning to Compose Dynamic Tree Structures for Visual Contexts — Approach



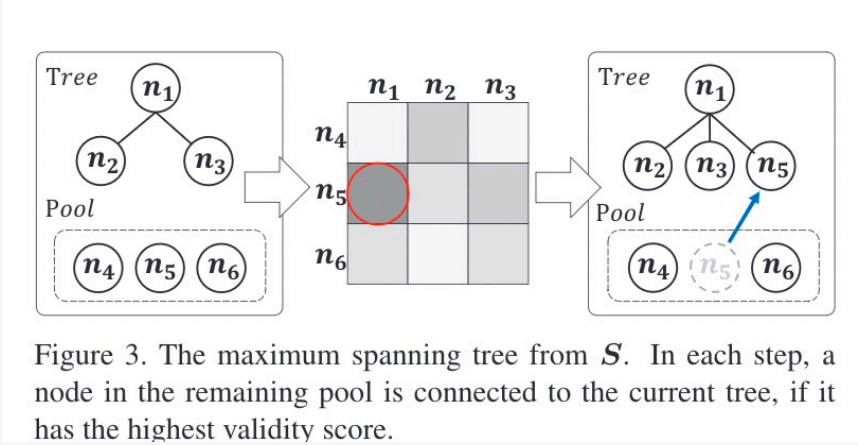


Figure 3. The maximum spanning tree from $\boldsymbol{S}$. In each step, a node in the remaining pool is connected to the current tree, if it has the highest validity score.

Feature Extraction

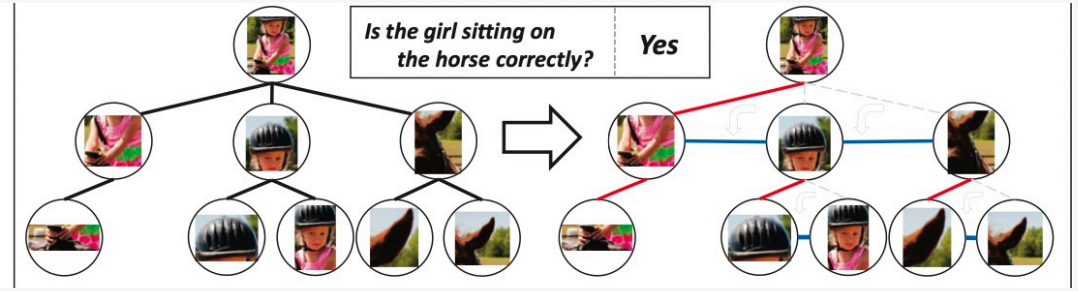
Spatial feature: $b_i \in R^8$

$$(x_1, y_1, x_2, y_2), (\frac{x_1 + x_2}{2}, \frac{y_1 + y_2}{2}), (x_2 - x_1, y_2 - y_1)$$

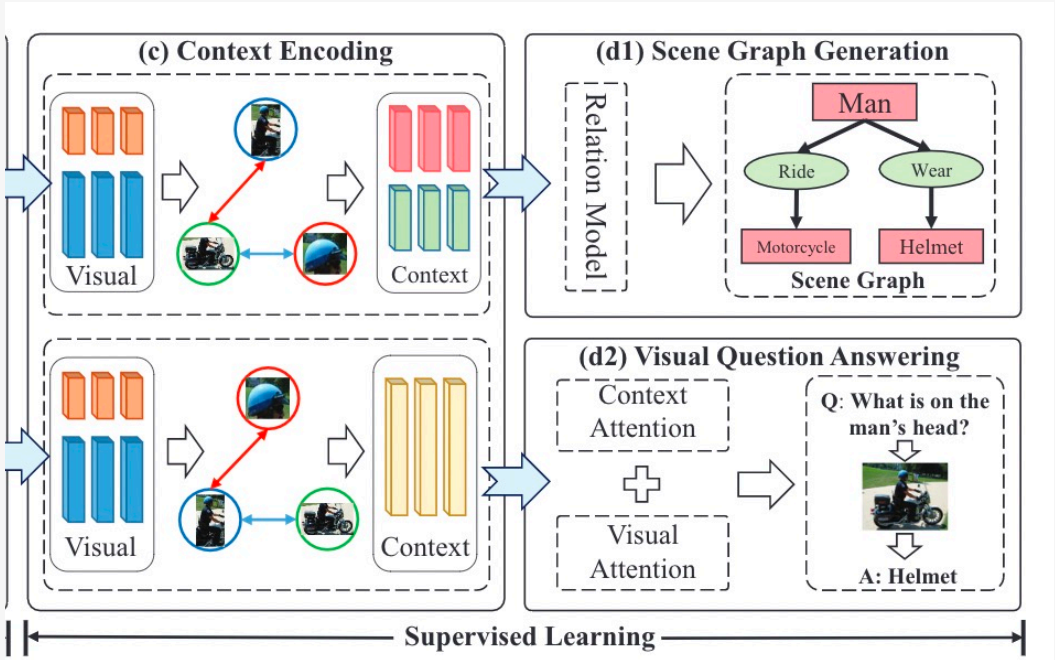
Tree Construction

VCTREE construction aims to learn a score matrix S

$$\begin{cases} \boldsymbol{S}_{ij} = f(\boldsymbol{x}_i, \boldsymbol{x}_j) \cdot g(\boldsymbol{x}_i, \boldsymbol{x}_j, \boldsymbol{q}), \\ f(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sigma\left(\mathrm{MLP}(\boldsymbol{x}_i, \boldsymbol{x}_j)\right), \\ g(\boldsymbol{x}_i, \boldsymbol{x}_j, \boldsymbol{q}) = \sigma(h(\boldsymbol{x}_i, \boldsymbol{q})) \cdot \sigma(h(\boldsymbol{x}_j, \boldsymbol{q})), \end{cases}$$



Learning to Compose Dynamic Tree Structures for Visual Contexts — Approach



Context Encoding

Context Encoder: $D = \text{BiTreeLSTM}(\{\boldsymbol{z}_i\}_{i=1,2,\ldots,n}),$

message passing：

$$\vec{\boldsymbol{h}}_i = \text{TreeLSTM}(\boldsymbol{z}_i, \vec{\boldsymbol{h}}_p),$$
$$\overleftarrow{\boldsymbol{h}}_i = \text{TreeLSTM}(\boldsymbol{z}_i, [\overleftarrow{\boldsymbol{h}}_l; \overleftarrow{\boldsymbol{h}}_r]),$$

object context input： $[\boldsymbol{x}_i; \boldsymbol{W}_1\hat{\boldsymbol{c}}_i]$

edge context input： onject contextoutput

Context Decoding

object decoding $[\boldsymbol{d}_i^o; \boldsymbol{W}_2\boldsymbol{c}_p],$

relation decoding

- $d_{ij} = MLP([d_i^r; d_j^r])$ as context feature
- $b_{ij} = MLP([b_i; b_j; b_{i\cup j}; b_{i\cap j}])$ as the bounding box pair feature
- v_{ij} as the RoI Align feature from the union bounding box of the object pair

Learning to Compose Dynamic Tree Structures for Visual Contexts — Result

	Scene Graph Generation			Scene Graph Classification			Predicate Classification		
Model	R@20	R@50	R@100	R@20	R@50	R@100	R@20	R@50	R@100
VRD [31]	-	0.3	0.5	-	11.8	14.1	-	27.9	35.0
AsscEmbed [34]	6.5	8.1	8.2	18.2	21.8	22.6	47.9	54.1	55.4
IMP$^\diamond$ [50]	14.6	20.7	24.5	31.7	34.6	35.4	52.7	59.3	61.3
TFR [21]	3.4	4.8	6.0	19.6	24.3	26.6	40.1	51.9	58.3
FREQ$^\diamond$ [57]	20.1	26.2	30.1	29.3	32.3	32.9	53.6	60.6	62.2
MOTIFS$^\diamond$ [57]	21.4	27.2	30.3	32.9	35.8	36.5	58.5	65.2	67.1
Graph-RCNN [51]	-	11.4	13.7	-	29.6	31.6	-	54.2	59.1
Chain	21.2	27.1	30.3	33.3	36.1	36.8	59.4	66.0	67.7
Overlap	21.4	27.3	30.4	33.7	36.5	37.1	59.5	66.0	67.8
Multi-Branch	21.5	27.3	30.6	34.3	37.1	37.8	59.5	66.1	67.8
VCTREE-SL	21.7	27.7	31.1	35.0	37.9	38.6	59.8	66.2	67.9
VCTREE-HL	**22.0**	**27.9**	**31.3**	**35.2**	**38.1**	**38.8**	**60.1**	**66.4**	**68.1**

Factorizable Net: An Efficient Subgraph-based Framework for Scene Graph Generation

Yikang LI,Wanli Ouyang,Bolei Zhou, Jianping Shi,Chao Zhang, and Xiaogang Wang

The Chinese University of Hong Kong, Hong Kong SAR, China

The University of Sydney, SenseTime Computer Vision Research Group,

MIT CSAIL USA,

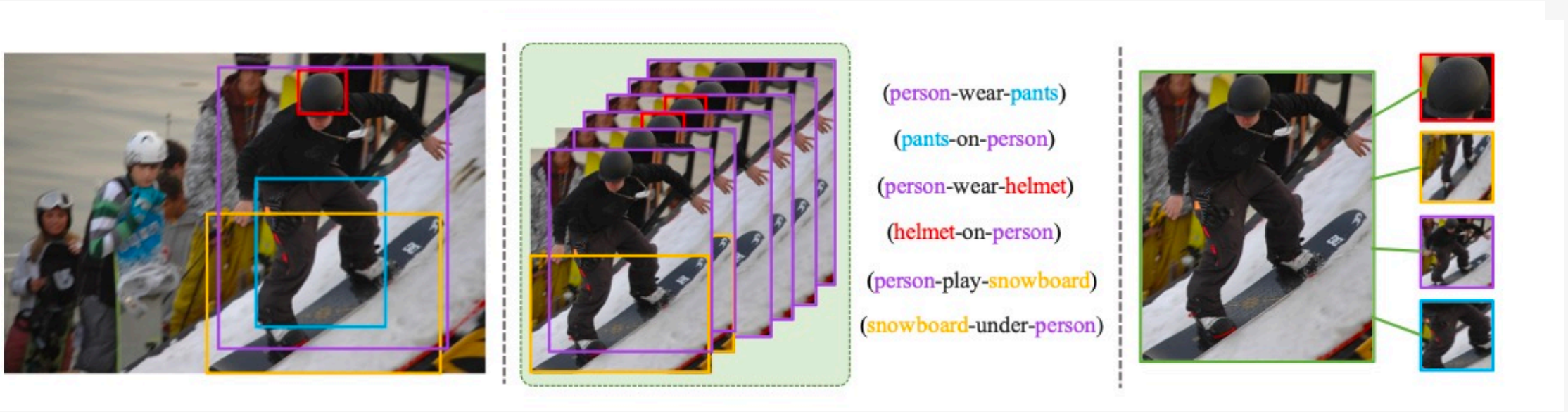
Sensetime Ltd, Beijing, China,

Samsung Telecommunication Research Institute, Beijing, China

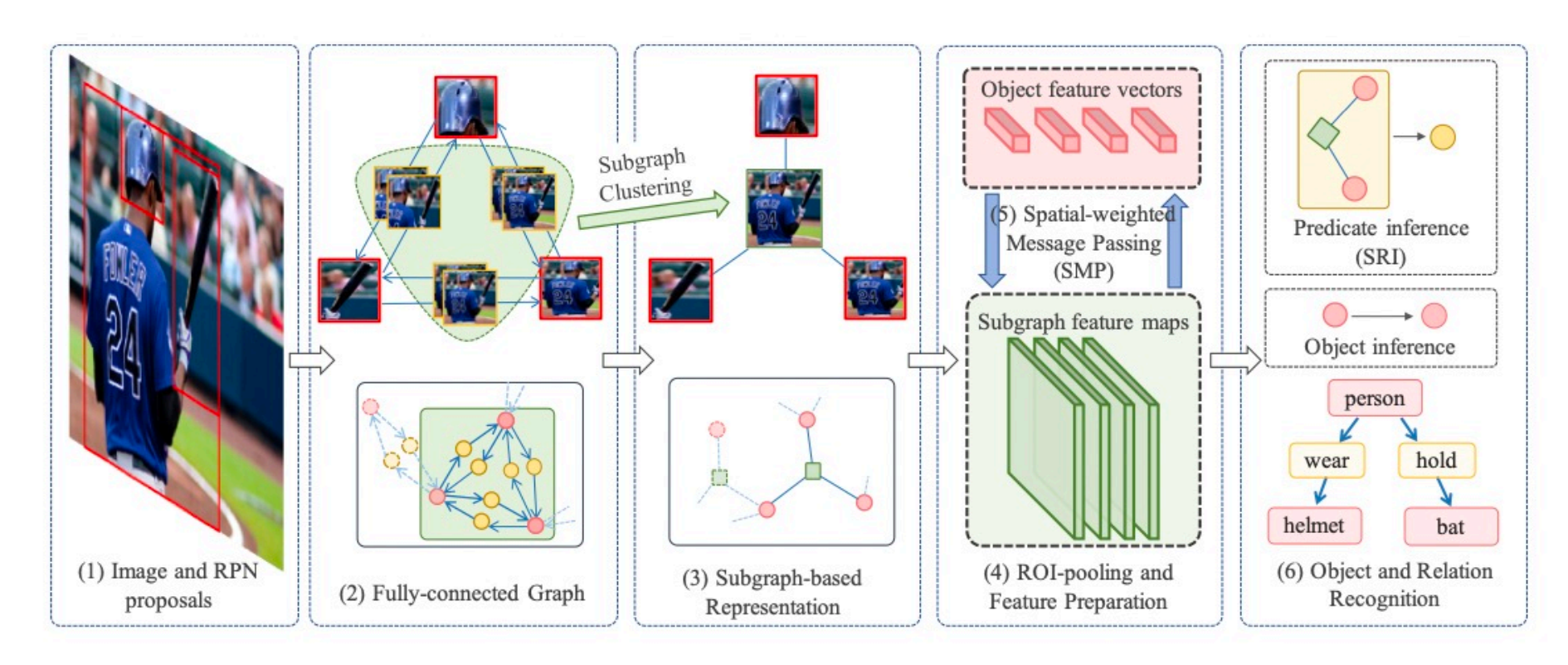
[ECCV 2018]

Factorizable Net - Problem

- phrase features 的数量决定模型的速度表现
- 最耗时的部分是phrase feature的处理
  - 寻找更简洁的场景图中间表示是解决问题的关键

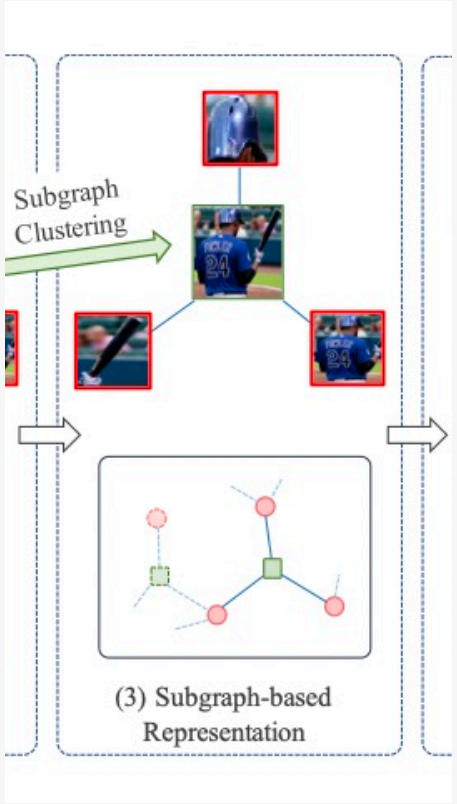


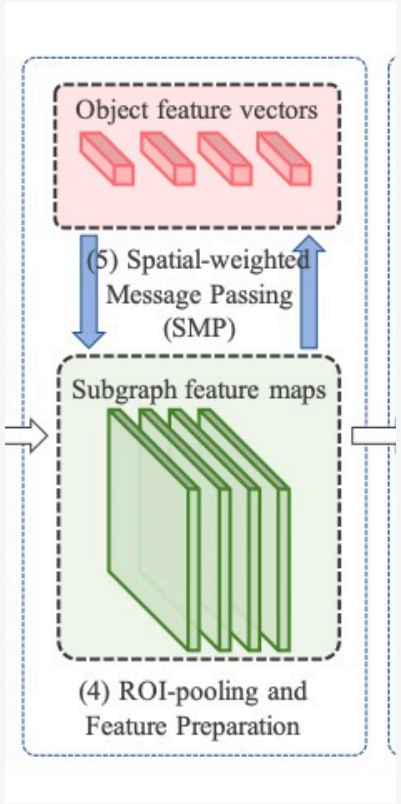
Factorizable Net - Approach



1. RPN被用来提取候选区
2. 候选框被组合成对，构成全连接图，两两有向连接
3. 将表示相似短语区域的边合并成子图，生成更简洁的连接图
4. 采用ROI池来获得相应的特征。
5. 消息在子图和对象特征之间传递，沿着分解的连接图进行特征细化
6. 根据对象特征预测对象，根据对象特征和子图特征推断谓词

Factorizable Net - Approach

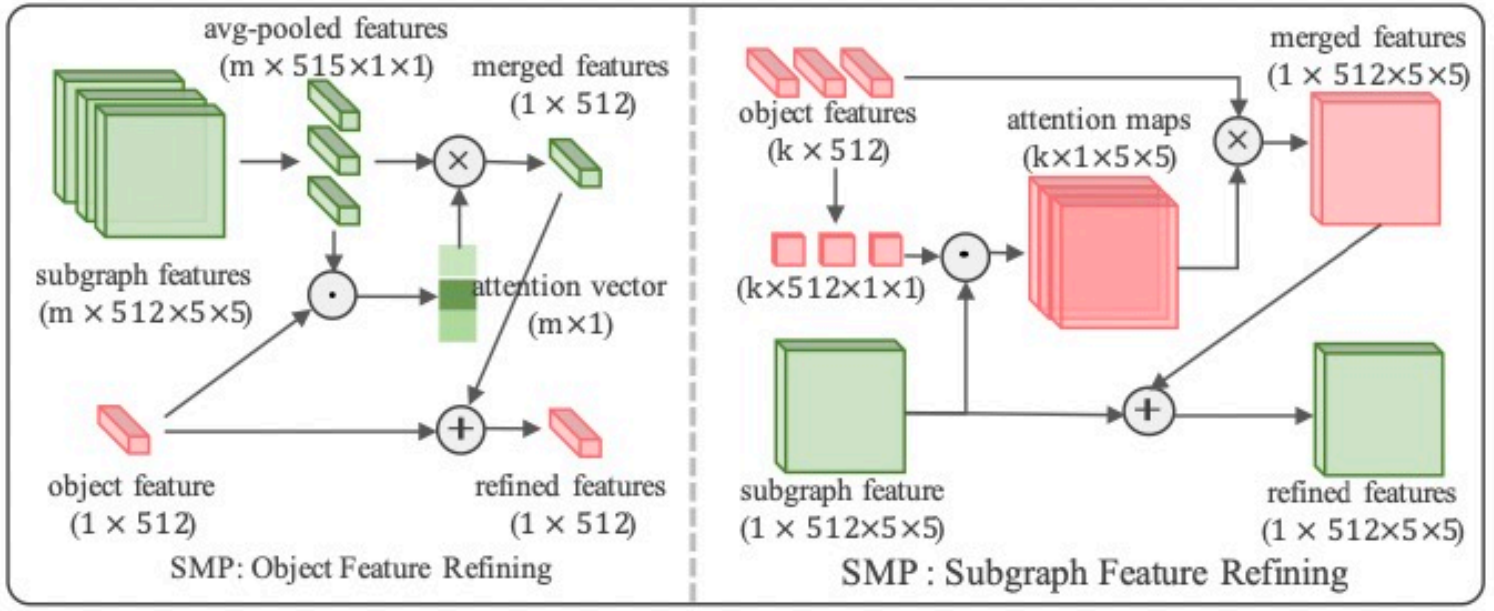




confidence score
bounding boxes location
non-maximum-suppression（NMS）

ROI-pooling生成subgraph features和object feature

Factorizable Net - Feature Refining with Spatial-weighted Message Passing（SMP）



Object to Subgraphs

$$\tilde{\mathbf{O}}_k(x,y) = \sum_{\mathbf{o_i} \in \mathbb{O}_k} \mathbf{P}_k(\mathbf{o}_i)(x,y) \cdot \mathbf{o}_i$$

$$\mathbf{P}_k(\mathbf{o}_i)(x,y) = \frac{\exp\left(\mathrm{FC}^{(att\_o)}\left(\mathrm{ReLU}\left(\mathbf{o}_i\right)\right) \cdot \mathbf{S}_k(x,y)\right)}{\sum_{\mathbf{S_k} \in \mathbb{C}_i} \exp\left(\mathrm{FC}^{(att\_o)}\left(\mathrm{ReLU}\left(\mathbf{o}_i\right)\right) \cdot \mathbf{S}_k(x,y)\right)}$$

$$\hat{\mathbf{S}}_k = \mathbf{S}_k + \mathrm{Conv}^{(o \to s)}\left(\mathrm{ReLU}\left(\tilde{\mathbf{O}}_k\right)\right)$$

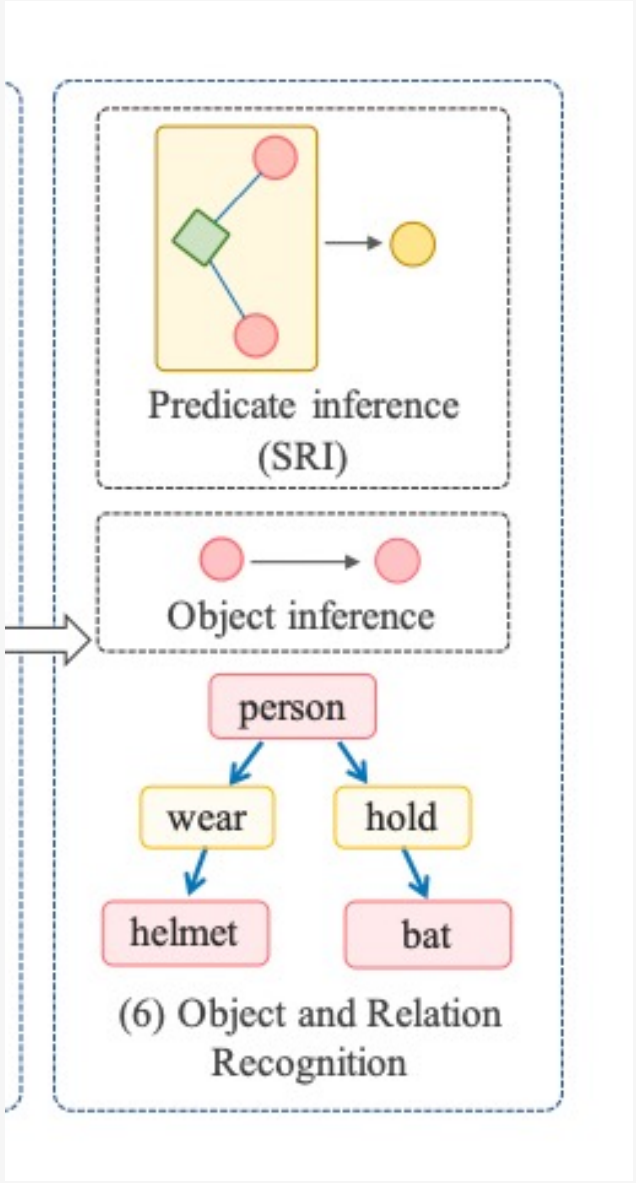
Subgraphs to Object

$$p_i(\mathbf{S}_k) = \frac{\exp\left(\mathbf{o}_i \cdot \mathrm{FC}^{(att\_s)}\left(\mathrm{ReLU}\left(\mathbf{s}_k\right)\right)\right)}{\sum_{\mathbf{S_k} \in \mathbb{C}_i} \exp\left(\mathbf{o}_i \cdot \mathrm{FC}^{(att\_s)}\left(\mathrm{ReLU}\left(\mathbf{s}_k\right)\right)\right)}$$

$$\tilde{\mathbf{s}}_i = \sum_{\mathbf{S_k} \in \mathbb{S}_i} p_i(\mathbf{S}_k) \cdot \mathbf{s}_k$$

$$\hat{\mathbf{o}}_i = \mathbf{o}_i + \mathrm{FC}^{(s \to o)}\left(\mathrm{ReLU}\left(\tilde{\mathbf{s}}_i\right)\right)$$

Factorizable Net - Spatial-sensitive Relation Inference



$$\mathbf{p}^{\langle i,k,j \rangle} = \mathbf{f}\left(\mathbf{o}_i, \mathbf{S}_k, \mathbf{o}_j\right)$$

$$\mathbf{S}_k^{(i)} = \mathrm{FC}\left(\mathrm{ReLU}\left(\mathbf{o}_i\right)\right) \otimes \mathrm{ReLU}\left(\mathbf{S}_k\right)$$

$$\mathbf{p}^{\langle i,k,j \rangle} = \mathrm{FC}^{(p)}\left(\mathrm{ReLU}\left(\left[\mathbf{S}_k^{(i)}; \mathbf{S}_k; \mathbf{S}_k^{(j)}\right]\right)\right)$$

$$\#\mathrm{FC}^{(p)} = C^{(p)} \times C \times W \times H \tag{10}$$

where $C^{(p)}$ denotes the number of predicate categories. C denotes the channel size. W and H denote the width and height of the feature map. Inspired by the bottleneck structure in [22], we introduce an additional 1×1 bottleneck convolution layer prior to $\mathrm{FC}^{(p)}$ to reduce the number of channels (omitted in Fig. 3). After adding an bottleneck layer with channel size equalling to C', the parameter size gets:

$$\#\mathrm{Conv}^{(bottleneck)} + \#\mathrm{FC}^{(p)} = C \times C' + C^{(p)} \times C' \times W \times H \tag{11}$$

If we take $C' = C/2$, as $\#\mathrm{Conv}^{(bottleneck)}$ is far less than $\#\mathrm{FC}^{(p)}$, we almost half the number of parameters.

Factorizable Net - Result

Speed shows the time spent for one inference forward pass (second/image).

ID	SubGraph	#SMP	2-D	SRI	#Boxes	PhrDet		SGGen		Speed
						R@50	R@100	R@50	R@100	
0	-	0	-	-	64	16.92	21.04	8.52	10.81	0.65
1	✓	0	-	-	64	16.50	20.79	8.49	10.33	0.18
2	✓	0	-	-	200	18.71	22.77	9.73	12.02	0.20
3	✓	0	✓	-	200	19.09	22.88	9.90	12.08	0.32
4	✓	1	✓	-	200	20.48	25.69	11.62	14.55	0.42
5	✓	1	✓	✓	200	22.54	28.31	12.83	16.12	0.44
6	✓	2	✓	✓	200	22.84	28.57	13.06	16.47	0.55

Dataset	Model	PhrDet		SGGen		Speed
		Rec@50	Rec@100	Rec@50	Rec@100	
VRD [37]	LP [37]	16.17	17.03	13.86	14.70	1.18*
	ViP-CNN [34]	22.78	27.91	17.32	20.01	0.78
	DR-Net [6]	19.93	23.45	17.73	20.88	2.83
	ILC [42]	16.89	20.70	15.08	18.37	2.70**
	Ours Full:1-SMP	25.90	30.52	18.16	21.04	**0.45**
	Ours Full:2-SMP	**26.03**	**30.77**	**18.32**	**21.20**	0.55
VG-MSDN [28, 35]	ISGG [58]	15.87	19.45	8.23	10.88	1.64
	MSDN [35]	19.95	24.93	10.72	14.22	3.56
	Ours-Full: 2-SMP	**22.84**	**28.57**	**13.06**	**16.47**	**0.55**
VG-DR-Net [6, 28]	DR-Net [6]	23.95	27.57	**20.79**	23.76	2.83
	Ours-Full: 2-SMP	**26.91**	**32.63**	19.88	**23.95**	**0.55**

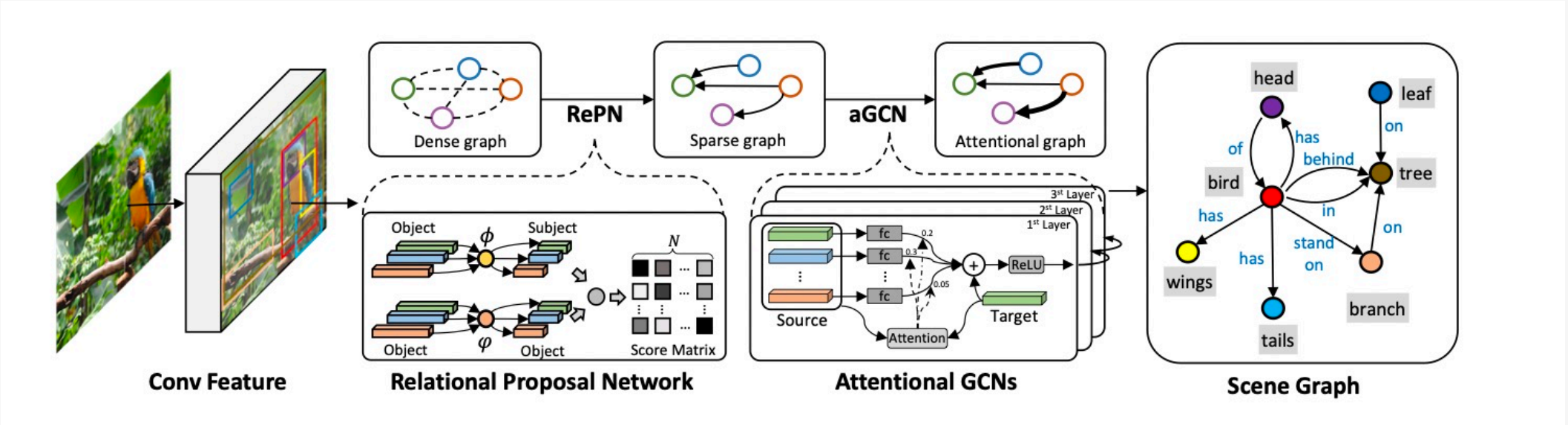
Graph R-CNN for Scene Graph Generation

Jianwei Yang, Jiasen Lu1, Stefan Lee, Dhruv Batra, and Devi Parikh,
Georgia Institute of Technology Facebook AI Research

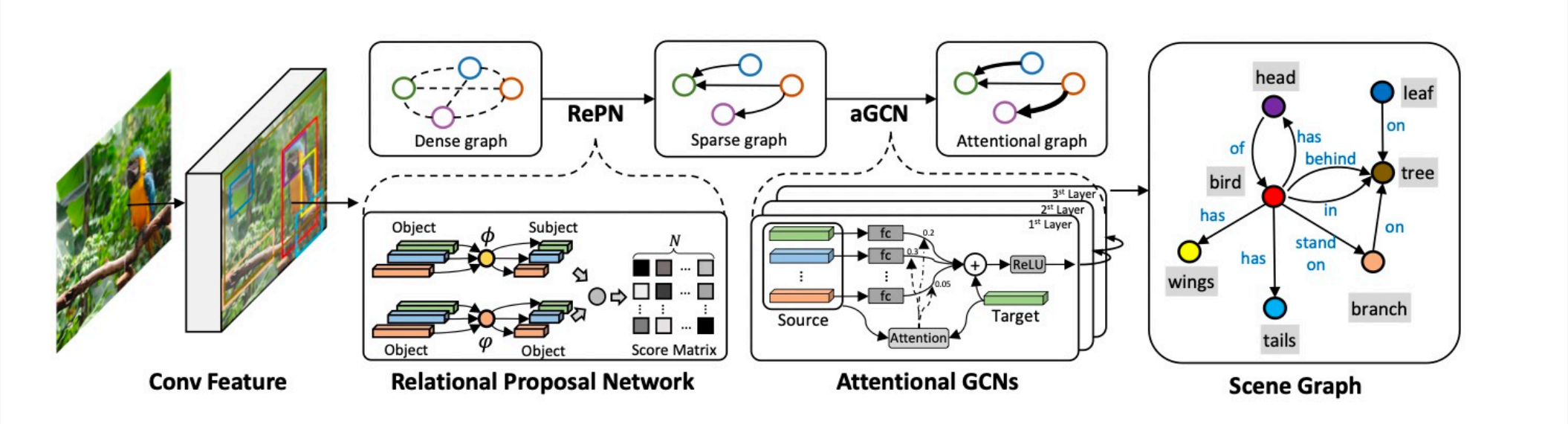
ECCV2018

Graph R-CNN - Contribution

- Relation Proposal Network（RePN）用来做关系过滤
- Attentional Graph Convolutional Network 用来做信息融合



Graph R-CNN - RePN



$$f(\boldsymbol{p}_i^o, \boldsymbol{p}_j^o) = \langle \Phi(\boldsymbol{p}_i^o), \Psi(\boldsymbol{p}_j^o) \rangle, i \neq j$$

We compute the overlap between two object pairs $\{u, v\}$ and $\{p, q\}$ as:

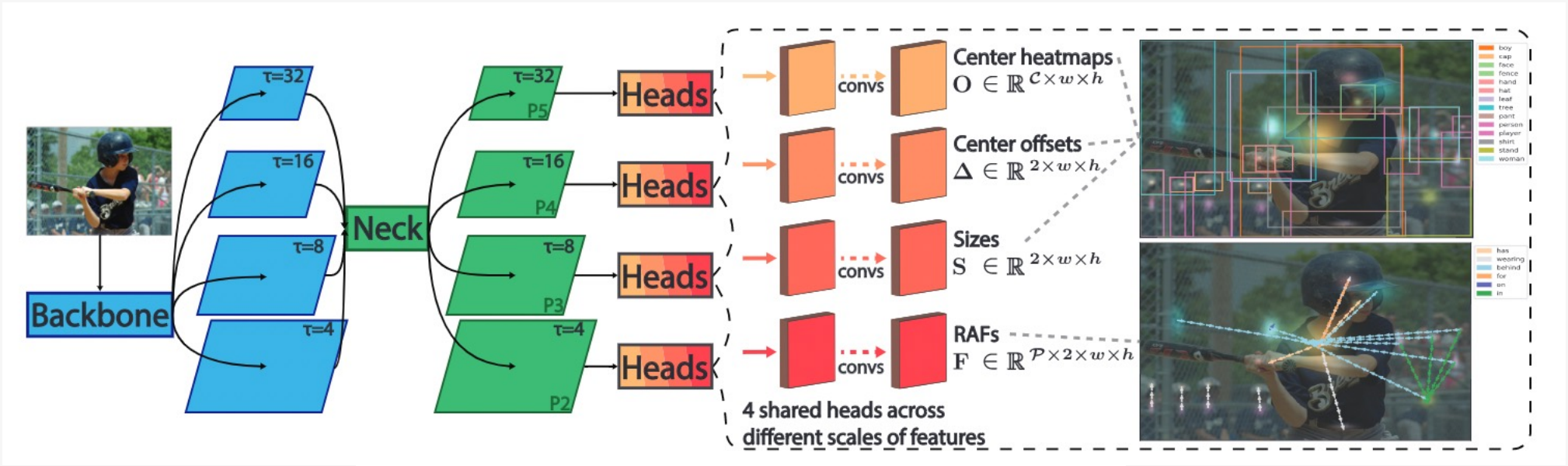
$$IoU(\{u, v\}, \{p, q\}) = \frac{I(r_u^o, r_p^o) + I(r_v^o, r_q^o)}{U(r_u^o, r_p^o) + U(r_v^o, r_q^o)}$$

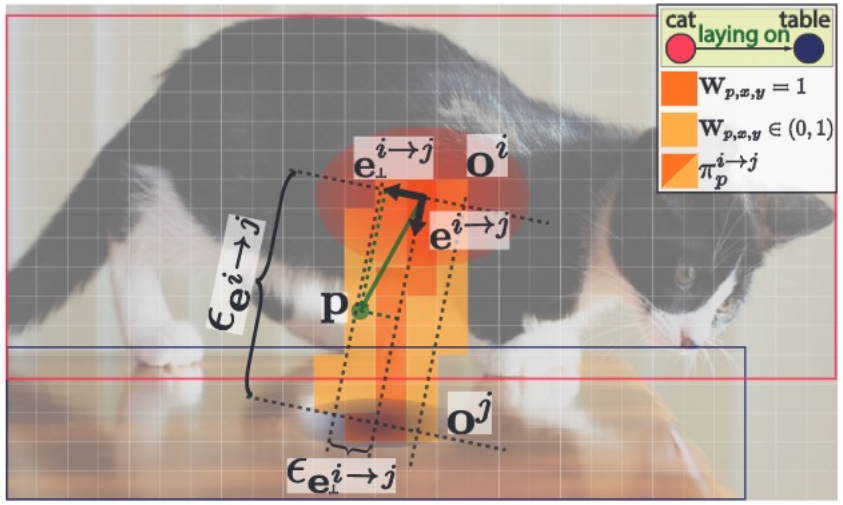
Fully Convolutional Scene Graph Generation

Hengyue Liu, Ning Yan, Masood Mortazavi, Bir Bhanu,
University of California, Riverside
Futurewei Technologies Inc.

CVPR 2021

Fully Convolutional Scene Graph Generation





Fully Convolutional Scene Graph Generation

Zero-shot Recall @K	PredCls		SGCls		SGDet	
Method	zsR@50/100		zsR@50/100		zsR@ 50/100	
MOTIFS-TDE [50]	14.4 / 18.2		3.4 / 4.5		2.3 / 2.9	
VTransE-TDE [50]	13.3 / 17.6		2.9 / 3.8		2.0 / 2.7	
VCTree-TDE [50]	14.3 / 17.6		3.2 / 4.0		2.6 / 3.2	
Knyazev *et al.* [28]	- / 21.5		- / 4.2		- / -	
FCSGG (Ours)	zsR @50/100	ng-zsR @50/100	zsR @50/100	ng-zsR @50/100	zsR @50/100	ng-zsR @50/100
HRNetW32-1S	8.3 / 10.7	12.9 / 19.2	1.0 / 1.2	2.3 / 3.5	0.6 / 1.0	1.2 / 1.6
HRNetW48-1S	8.6 / 10.9	12.8 / 19.6	1.7 / 2.1	2.9 / 4.4	1.0 / 1.4	1.8 / 2.7
ResNet50-4S-FPN$_{\times 2}$	8.2 / 10.6	11.7 / 18.1	1.3 / 1.7	2.4 / 3.8	0.8 / 1.1	1.0 / 1.7
HRNetW48-5S-FPN$_{\times 2}$	7.9 / 10.1	11.5 / 17.7	1.7 / 2.1	2.8 / 4.8	0.9 / 1.4	1.4 / 2.4
HRNetW48-5S-FPN$_{\times 2}$-f	7.8 / 10.0	11.4 / 17.6	1.6 / 2.0	2.8 / 4.8	0.8 / 1.4	1.4 / 2.3

Method	#Params (M)	Input Size	s / image
Pixels2Graphs [71]	94.8	512×512	3.55
VCTree-TDE [50]	360.8	600×1000	1.69
MOTIFS-TDE [50]	369.5	600×1000	0.87
KERN [6]	405.2	592×592	0.79
MOTIFS [50]	367.2	600×1000	0.66
FactorizableNet [33]	40.4	600×1000	0.59
VTransE-TDE [50]	311.6	600×1000	0.55
GB-NET-β [68]	444.6	592×592	0.52
Graph R-CNN [62]	80.2	800×1024	0.19
FCSGG (Ours)			
HRNetW32-1S	47.3	512×512	0.07
HRNetW48-1S	86.1	512×512	0.08
ResNet50-4S-FPN$_{\times 2}$	36.0	512×512	0.04
HRNetW48-5S-FPN$_{\times 2}$	87.1	640×1024	0.12
HRNetW48-5S-FPN$_{\times 2}$-f	87.1	640×1024	0.12