Referring Segmentation

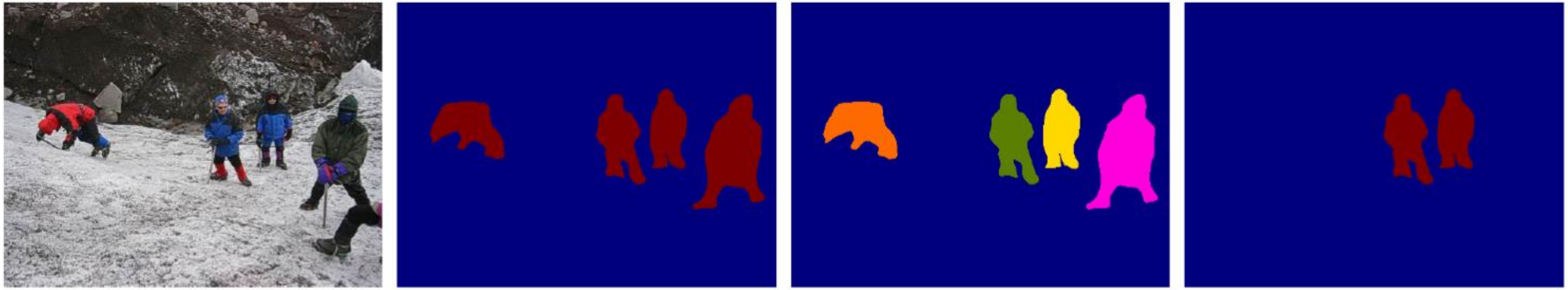
Mengxue

Segmentation from Natural Language Expressions

ECCV 2016

Ronghang Hu[1] Marcus Rohrbach[1,2] Trevor Darrell[1]
{ronghang, rohrbach, trevor}@eecs.berkeley.edu

[1]UC Berkeley EECS, CA, United States
[2]ICSI, Berkeley, CA, United States



(a) input image (b) object class segmentation of class ***people*** (c) object instance segmentation of class ***people*** (d) segmentation from expression *"people in blue coat"*

Contents

- Dataset
- Paper Sharing

Paper List :

Segmentation from Natural Language Expressions

Recurrent Multimodal Interaction for Referring Image Segmentation

MAttNet: Modular Attention Network for Referring Expression Comprehension

Referring Expression Object Segmentation with Caption-Aware Consistency

PhraseCut: Language-Based Image Segmentation in the Wild

Figure 2. **Example annotations from the VGPHRASECUT dataset.** Colors (blue, red, green) of the input phrases correspond to words that indicate attributes, categories, and relationships respectively.

Dataset	ReferIt [17]	Google RefExp [26]	RefCOCO [41]	VGPHRASECUT
# images	19,894	26,711	19,994	77,262
# instances	96,654	54,822	50,000	345,486
# categories	-	80	80	3103
multi-instance	No	No	No	Yes
segmentation	Yes	Yes	Yes	Yes
referring phrase	short phrases	long descriptions	short phrases	templated phrases

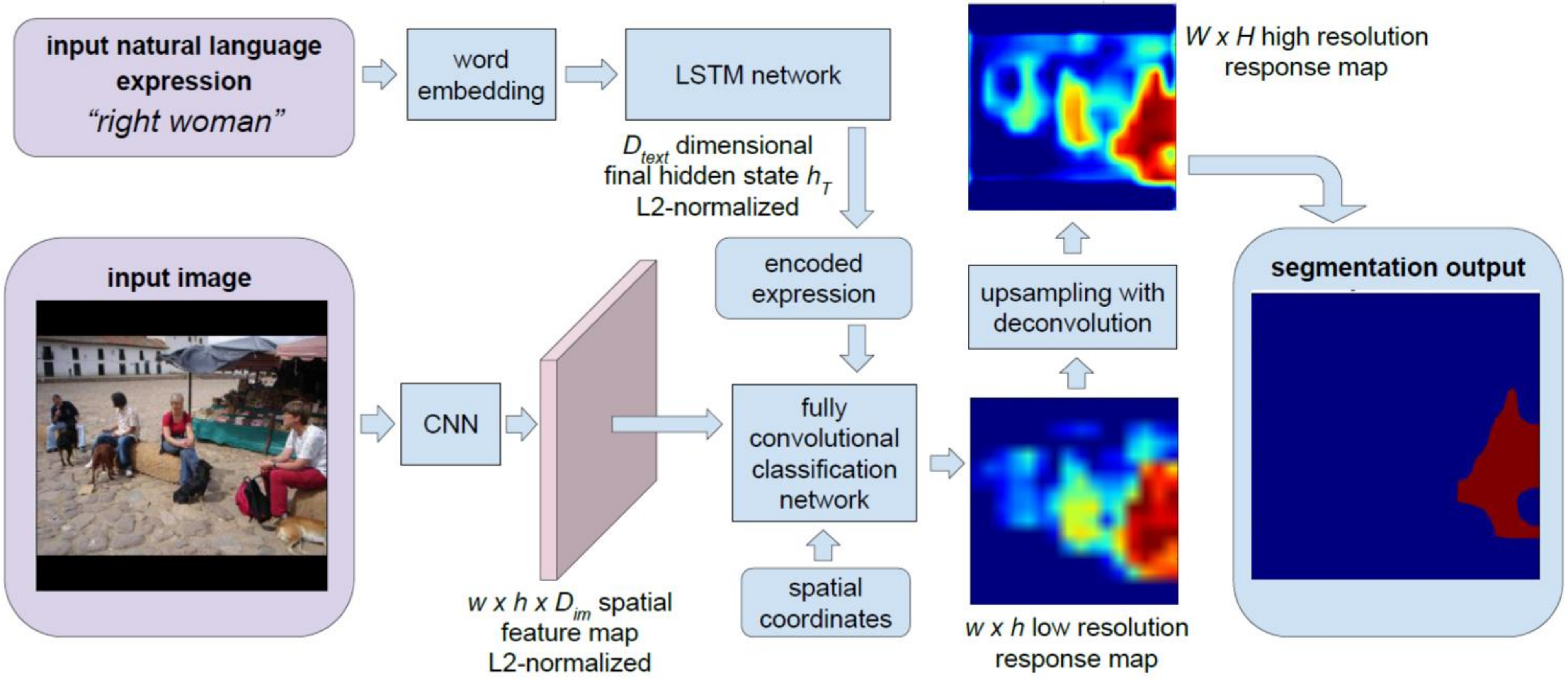
1

Mar 2016

Segmentation from Natural Language Expressions

ECCV 2016

Ronghang Hu[1] Marcus Rohrbach[1,2] Trevor Darrell[1]
{ronghang, rohrbach, trevor}@eecs.berkeley.edu



Word embedding

我们导入在维基百科上训练的GloVe向量

```
import gensim
import gensim.downloader as api
model = api.load('glove-wiki-gigaword-50')
```

单词“king”的词嵌入表示：

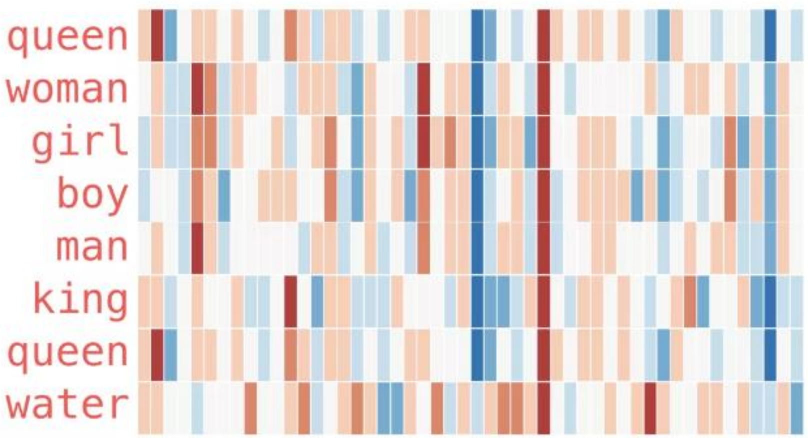
```
model["king"]
```

```
array([ 0.50451 ,  0.68607 , -0.59517 , -0.022801,  0.60046 , -0.13498 ,
       -0.08813 ,  0.47377 , -0.61798 , -0.31012 , -0.076666,  1.493   ,
       -0.034189, -0.98173 ,  0.68229 ,  0.81722 , -0.51874 , -0.31503 ,
       -0.55809 ,  0.66421 ,  0.1961  , -0.13495 , -0.11476 , -0.30344 ,
        0.41177 , -2.223   , -1.0756  , -1.0783  , -0.34354 ,  0.33505 ,
        1.9927  , -0.04234 , -0.64319 ,  0.71125 ,  0.49159 ,  0.16754 ,
        0.34344 , -0.25663 , -0.8523  ,  0.1661  ,  0.40102 ,  1.1685  ,
       -1.0137  , -0.21585 , -0.15155 ,  0.78321 , -0.91241 , -1.6106  ,
       -0.64426 , -0.51042 ], dtype=float32)
```

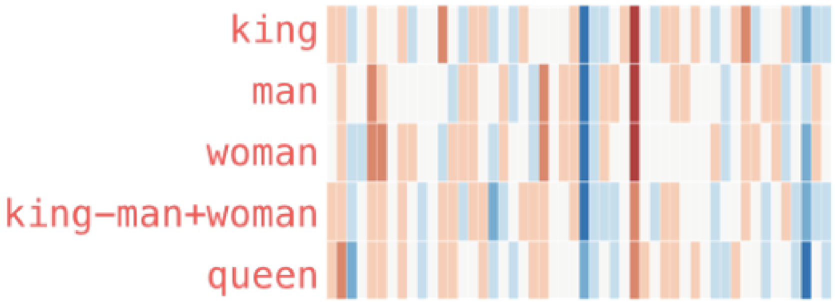
查看“king”最相似的单词

```
model.most_similar("king")
```

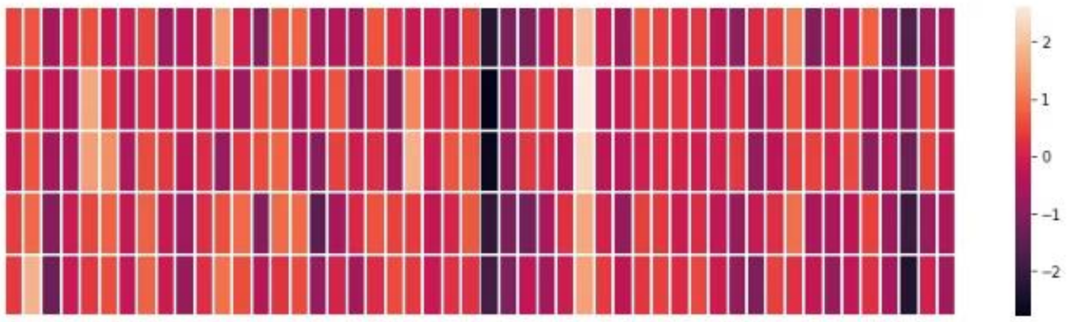
```
[('prince', 0.8236179351806641),
 ('queen', 0.7839042544364929),
 ('ii', 0.7746230363845825),
 ('emperor', 0.7736247181892395),
 ('son', 0.766719400882721),
 ('uncle', 0.7627150416374207),
 ('kingdom', 0.7542160749435425),
 ('throne', 0.7539913654327393),
 ('brother', 0.7492411136627197),
 ('ruler', 0.7434253096580505)]
```



king − man + woman ~= queen



→



Loss

$$Loss = \frac{1}{WH}\sum_{i=1}^{W}\sum_{j=1}^{H} L(v_{ij}, M_{ij})$$

$$L(v_{ij}, M_{ij}) = \begin{cases} \alpha_f \log(1 + \exp(-v_{ij})) & \text{if } M_{ij} = 1 \\ \alpha_b \log(1 + \exp(v_{ij})) & \text{if } M_{ij} = 0 \end{cases}$$

2

Recurrent Multimodal Interaction for Referring Image Segmentation

ICCV 2017

Chenxi Liu[1] Zhe Lin[2] Xiaohui Shen[2] Jimei Yang[2] Xin Lu[2] Alan Yuille[1]
Johns Hopkins University[1] Adobe Research[2]
{cxliu, alan.yuille}@jhu.edu {zlin, xshen, jimyang, xinl}@adobe.com

Segmentation from Natural Language Expressions:

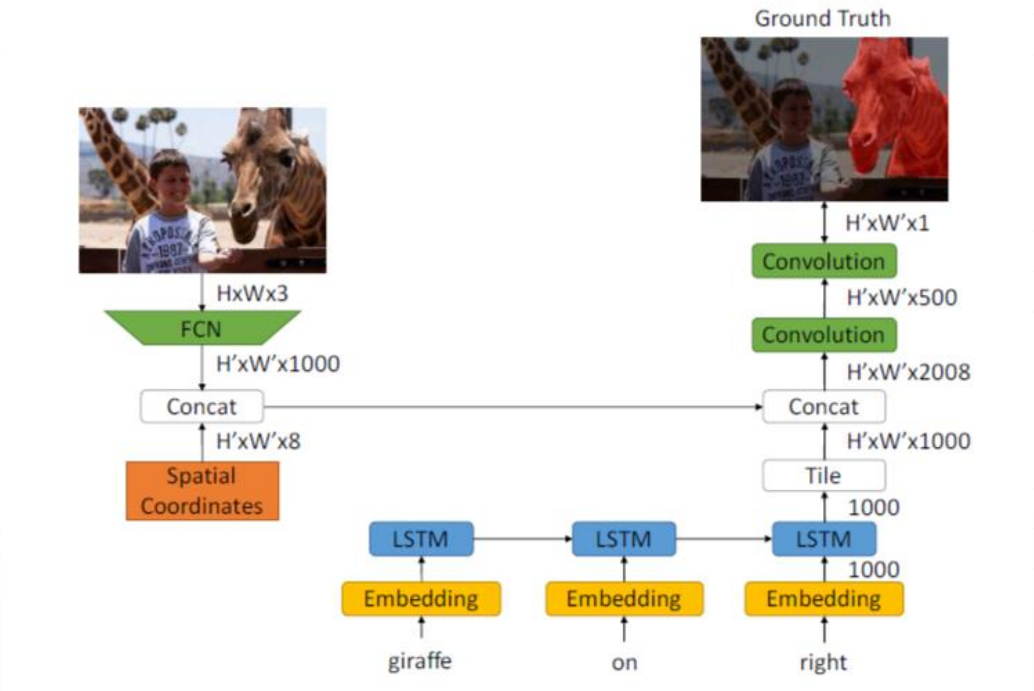
- Remember all information → Find the matching region
- **human**： image-sentence-image reading sequence 来回浏览确定目标区域

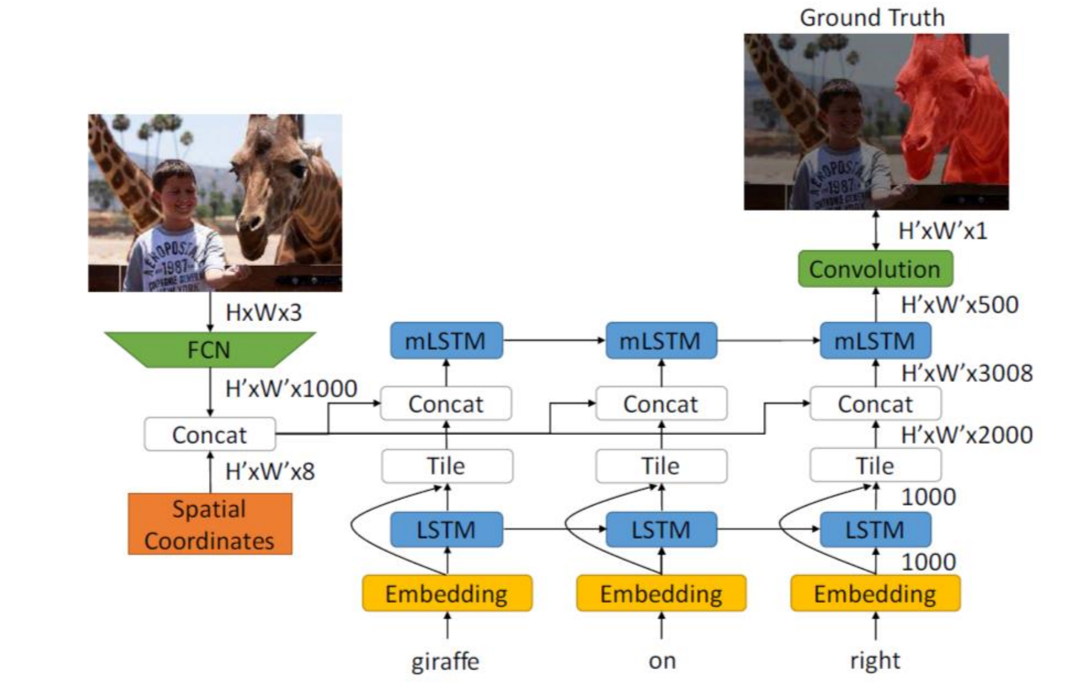
e.g. “The man on the right wearing blue.”

Step: 1.Find out pixels →“the man”
2.Delete pixels **not** “on the right”
3.Delete pixels **not** “wearing blue”

→ RMI

CNN+LSTM→RMI





$$\mathbf{l}_t \xrightarrow{LSTM} \mathbf{h}_T \xrightarrow{Concat} \begin{bmatrix} \mathbf{h}_T \\ \mathbf{v}^{ij} \end{bmatrix} \xrightarrow{Conv} \text{multimodal feature} \quad (8)$$

$$\mathbf{l}_t \xrightarrow{Concat} \begin{bmatrix} \mathbf{l}_t \\ \mathbf{v}^{ij} \end{bmatrix} \xrightarrow{mLSTM} \text{multimodal feature} \quad (9)$$

language-only LSTM

RMI

Experiments

	Google-Ref	UNC			UNC+			ReferItGame
	val	val	testA	testB	val	testA	testB	test
[12, 13]	28.14	-	-	-	-	-	-	48.03
R+LSTM	28.60	38.74	39.18	39.01	26.25	26.95	24.57	54.01
R+RMI	**32.06**	**39.74**	**39.99**	**40.44**	**27.85**	**28.69**	**26.65**	**54.55**
R+LSTM+DCRF	28.94	39.88	40.44	40.07	26.29	27.03	24.44	55.90
R+RMI+DCRF	**32.85**	**41.17**	**41.35**	**41.87**	**28.26**	**29.16**	**26.86**	**56.61**
D+LSTM	33.08	43.27	43.60	43.31	28.42	28.57	27.70	56.83
D+RMI	**34.40**	**44.33**	**44.74**	**44.63**	**29.91**	**30.37**	**29.43**	**57.34**
D+LSTM+DCRF	33.11	43.97	44.25	44.07	28.07	28.29	27.44	58.20
D+RMI+DCRF	**34.52**	**45.18**	**45.69**	**45.57**	**29.86**	**30.48**	**29.50**	**58.73**

3

MAttNet: Modular Attention Network for Referring Expression Comprehension

Licheng Yu[1], Zhe Lin[2], Xiaohui Shen[2], Jimei Yang[2], Xin Lu[2],
Mohit Bansal[1], Tamara L. Berg[1]

[1]University of North Carolina at Chapel Hill [2]Adobe Research
{licheng, tlberg, mbansal}@cs.unc.edu, {zlin, xshen, jimyang, xinl}@adobe.com

Target object: a red ball

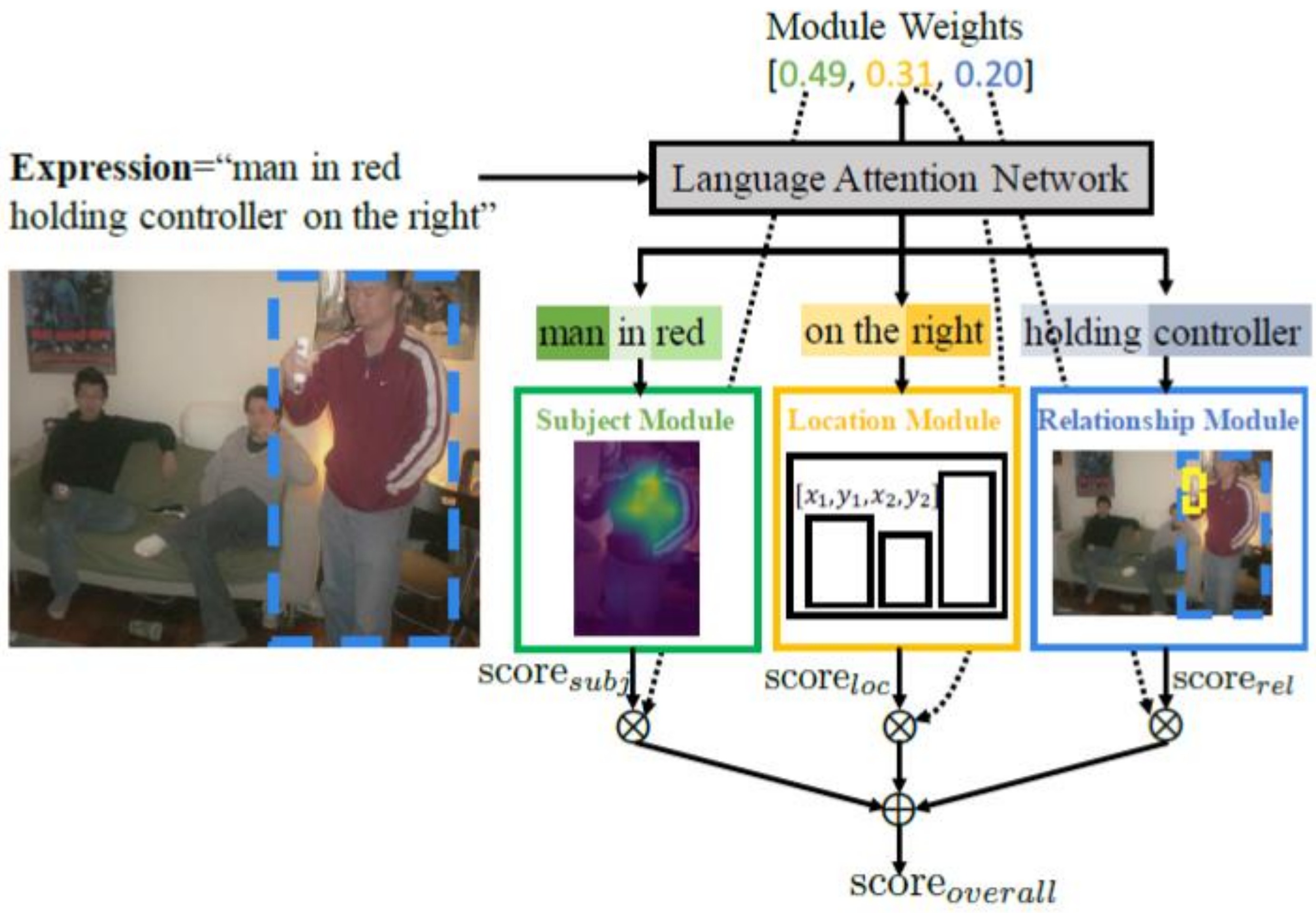
·a red ball among 10 black balls ←”the red ball” √

·a red ball is placed among 3 other red balls←“red ball on the right” ← location information

·a red ball is placed among 100 other red balls←“red ball next to the cat”← most distinguishing information

CVPR 2018

MAttNet



Language Attention Network

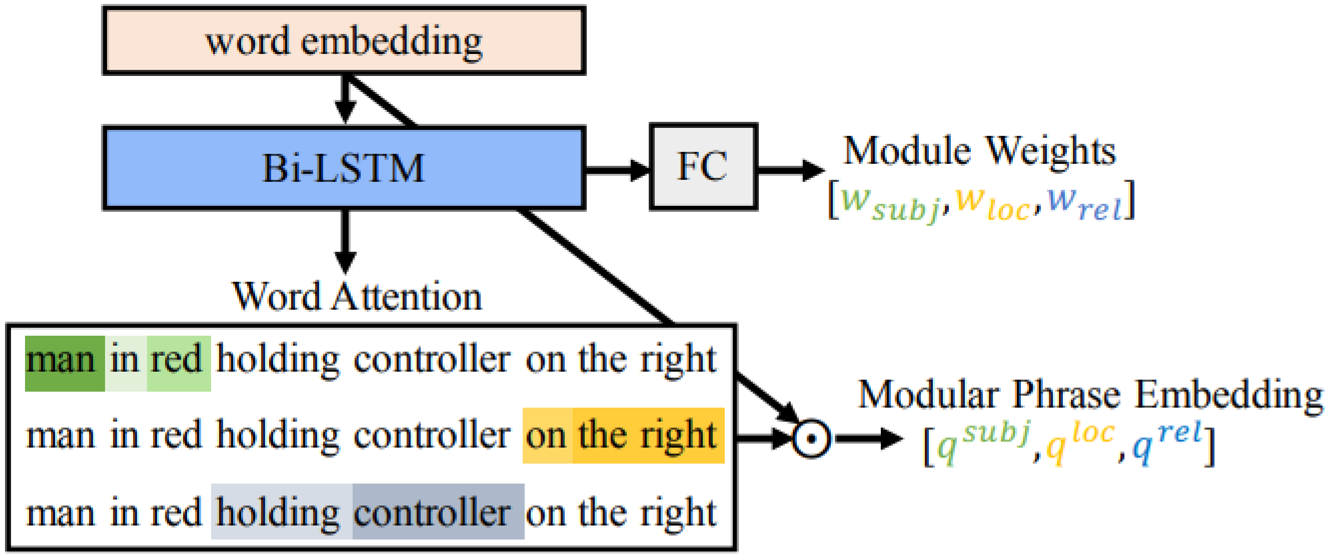


Figure 2: Language Attention Network

$$[w_{subj}, w_{loc}, w_{rel}] = \text{softmax}(W_m^T[h_0, h_T] + b_m)$$

$$r = \{u_t\}_{t=1}^T,$$

$$e_t = \text{embedding}(u_t)$$

$$\vec{h}_t = \text{LSTM}(e_t, \vec{h}_{t-1})$$

$$h_t = \text{LSTM}(e_t, h_{t+1})$$

$$h_t = [\vec{h}_t, h_t].$$

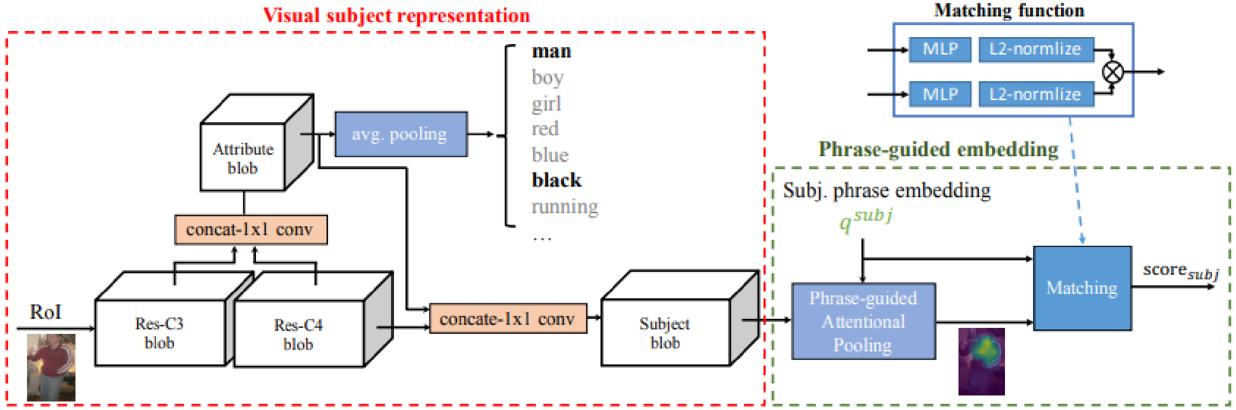
$$H = \{h_t\}_{t=1}^T$$

$$m \in \{\text{subj}, \text{loc}, \text{rel}\}$$

$$a_{m,t} = \frac{\exp(f_m^T h_t)}{\sum_{k=1}^T \exp(f_m^T h_k)}$$

$$q^m = \sum_{t=1}^T a_{m,t} e_t.$$

Visual Modules



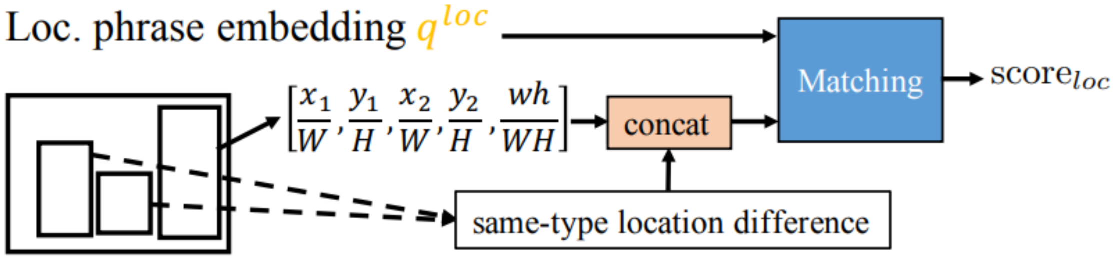


Figure 4: Location Module

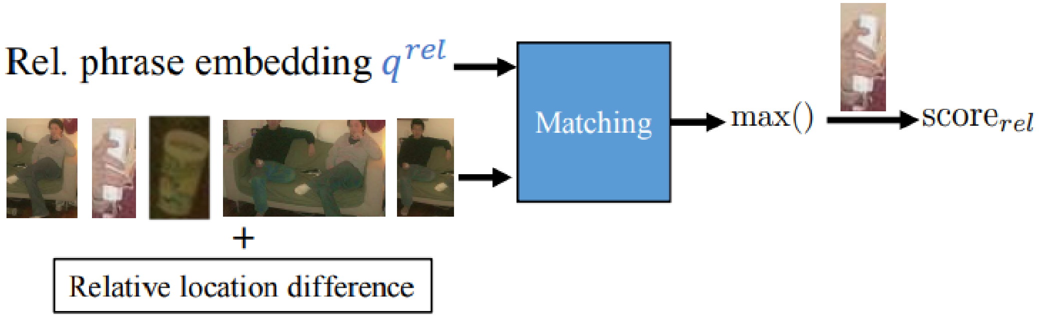
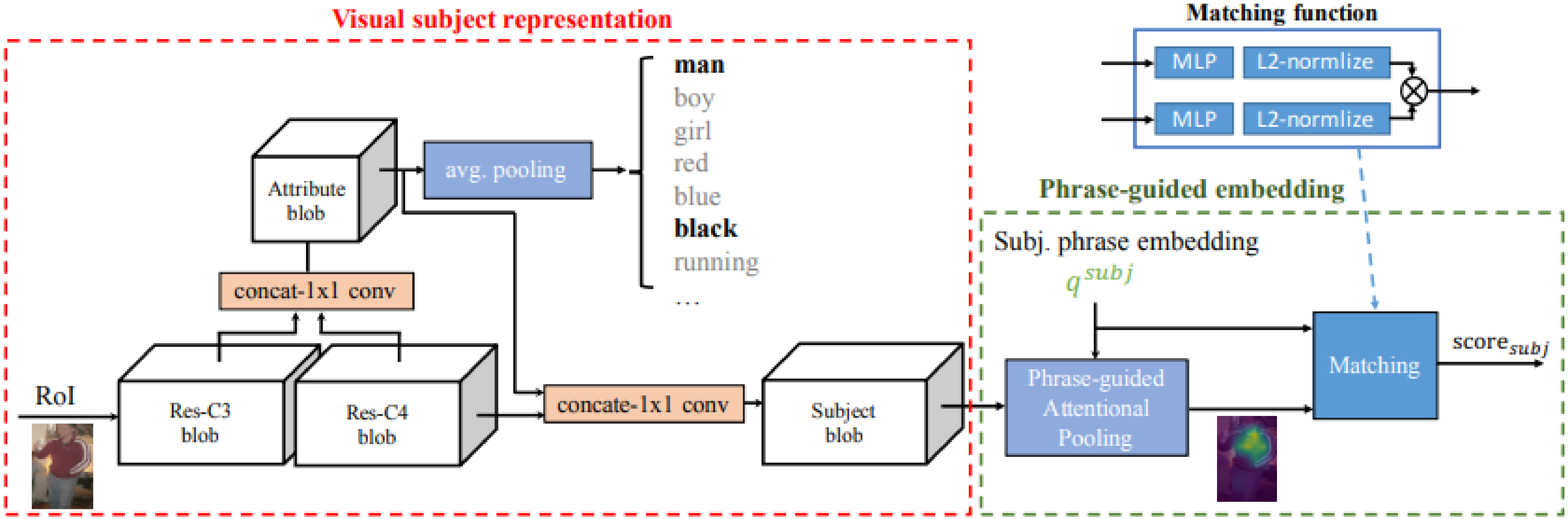


Figure 5: Relationship Module

$$S(o_i|q^{subj}), S(o_i|q^{loc}) \text{ and } S(o_i|q^{rel}).$$

Visual Modules--Subject Module

Attribute Prediction | Phrase-guided Attentional Pooling| Matching Function



Visual Modules--Location Module

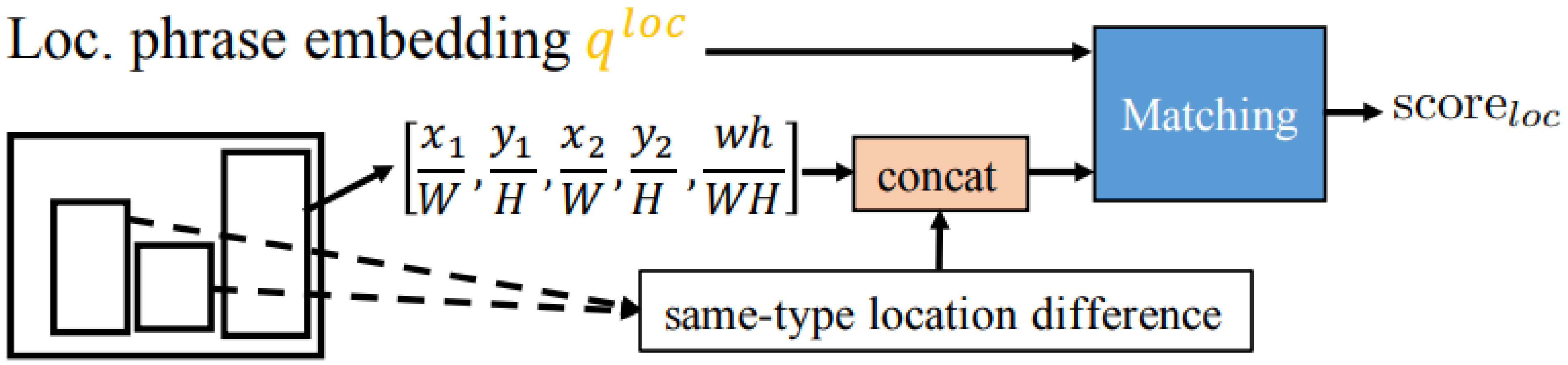


Figure 4: Location Module

Visual Modules--Relationship Module

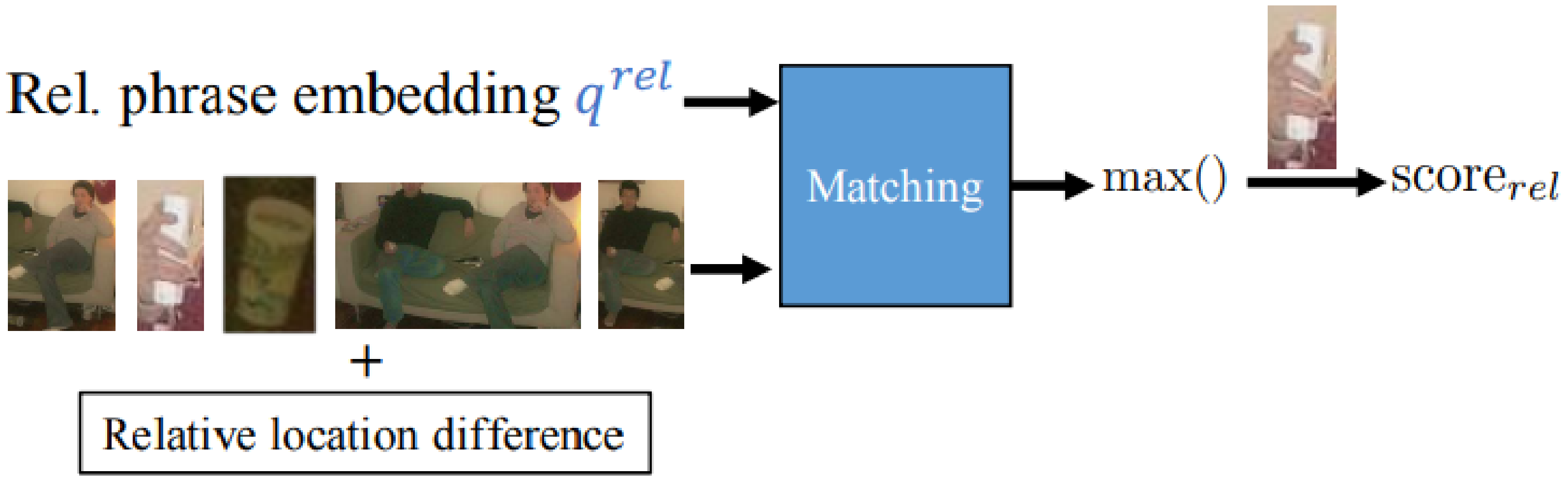


Figure 5: Relationship Module

$$\delta m_{ij} = \left[\frac{[\triangle x_{tl}]_{ij}}{w_i}, \frac{[\triangle y_{tl}]_{ij}}{h_i}, \frac{[\triangle x_{br}]_{ij}}{w_i}, \frac{[\triangle y_{br}]_{ij}}{h_i}, \frac{w_j h_j}{w_i h_i}\right].$$

$$\tilde{v}_{ij}^{rel} = W_r[v_{ij}; \delta m_{ij}] + b_r$$

Loss Function

$$S(o_i|r) = w_{subj}S(o_i|q^{subj}) + w_{loc}S(o_i|q^{loc}) + w_{rel}S(o_i|q^{rel})$$

$$\begin{aligned} L_{rank} = \sum_i [&\lambda_1 \max(0, \Delta + S(o_i|r_j) - S(o_i|r_i)) \\ &+ \lambda_2 \max(0, \Delta + S(o_k|r_i) - S(o_i|r_i))] \end{aligned}$$

$$L_{subj}^{attr} = \lambda_{attr} \sum_i \sum_j w_j^{attr} [\log(p_{ij}) + (1 - y_{ij}) \log(1 - p_{ij})]$$

$$L = L_{subj}^{attr} + L_{rank}.$$

Results: Referring Expression Comprehension

RefCOCO

Model	Backbone Net	Split	Pr@0.5	Pr@0.6	Pr@0.7	Pr@0.8	Pr@0.9	IoU
D+RMI+DCRF [14]	res101-DeepLab	val	42.99	33.24	22.75	12.11	2.23	45.18
MAttNet	res101-mrcn	val	**75.16**	**72.55**	**67.83**	**54.79**	**16.81**	**56.51**
D+RMI+DCRF [14]	res101-DeepLab	testA	42.99	33.59	23.69	12.94	2.44	45.69
MAttNet	res101-mrcn	testA	**79.55**	**77.60**	**72.53**	**59.01**	**13.79**	**62.37**
D+RMI+DCRF [14]	res101-DeepLab	testB	44.99	32.21	22.69	11.84	2.65	45.57
MAttNet	res101-mrcn	testB	**68.87**	**65.06**	**60.02**	**48.91**	**21.37**	**51.70**

RefCOCO+

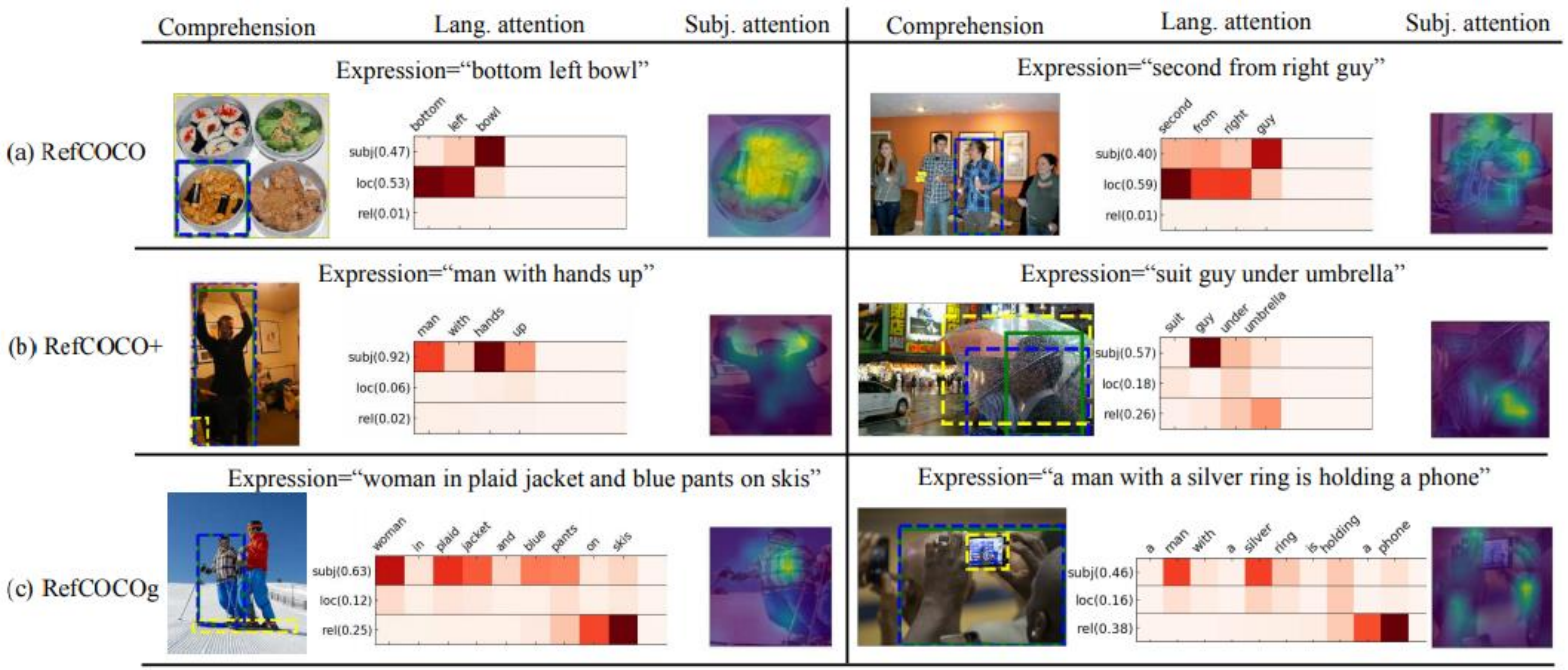
Model	Backbone Net	Split	Pr@0.5	Pr@0.6	Pr@0.7	Pr@0.8	Pr@0.9	IoU
D+RMI+DCRF [14]	res101-DeepLab	val	20.52	14.02	8.46	3.77	0.62	29.86
MAttNet	res101-mrcn	val	**64.11**	**61.87**	**58.06**	**47.42**	**14.16**	**46.67**
D+RMI+DCRF [14]	res101-DeepLab	testA	21.22	14.43	8.99	3.91	0.49	30.48
MAttNet	res101-mrcn	testA	**70.12**	**68.48**	**63.97**	**52.13**	**12.28**	**52.39**
D+RMI+DCRF [14]	res101-DeepLab	testB	20.78	14.56	8.80	4.58	0.80	29.50
MAttNet	res101-mrcn	testB	**54.82**	**51.73**	**47.27**	**38.58**	**17.00**	**40.08**

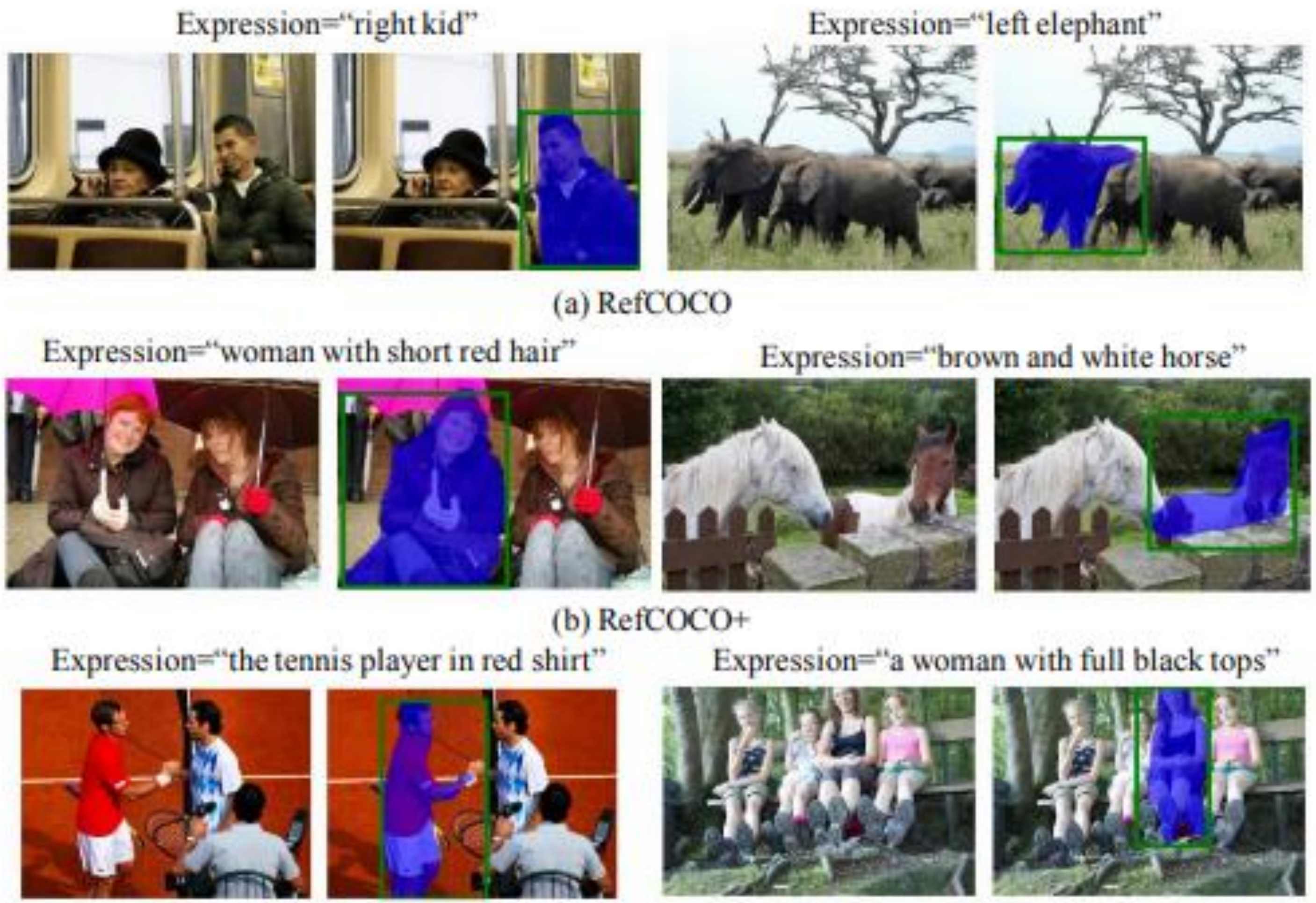
RefCOCOg

Model	Backbone Net	Split	Pr@0.5	Pr@0.6	Pr@0.7	Pr@0.8	Pr@0.9	IoU
MAttNet	res101-mrcn	val	64.48	61.52	56.50	43.97	14.67	47.64
MAttNet	res101-mrcn	test	65.60	62.92	57.31	44.44	12.55	48.61

Table 4: Comparison of segmentation performance on RefCOCO, RefCOCO+, and our results on RefCOCOg.

Precision@0.5 is the percentage of expressions where the IoU of the predicted segmentation and ground-truth is at least 0.5.


Comprehension
Lang. attention
Subj. attention
Comprehension
Lang. attention
Subj. attention
(a) RefCOCO
Expression="bottom left bowl"
bottom
left
bowl
subj(0.47)
loc(0.53)
rel(0.01)
Expression="second from right guy"
second
from
right
guy
subj(0.40)
loc(0.59)
rel(0.01)
(b) RefCOCO+
Expression="man with hands up"
man
with
hands
up
subj(0.92)
loc(0.06)
rel(0.02)
Expression="suit guy under umbrella"
suit
guy
under
umbrella
subj(0.57)
loc(0.18)
rel(0.26)
(c) RefCOCOg
Expression="woman in plaid jacket and blue pants on skis"
woman
in
plaid
jacket
and
blue
pants
on
skis
subj(0.63)
loc(0.12)
rel(0.25)
Expression="a man with a silver ring is holding a phone"
a
man
with
a
silver
ring
is
holding
a
phone
subj(0.46)
loc(0.16)
rel(0.38)



(a) RefCOCO

(b) RefCOCO+

(c) RefCOCOg

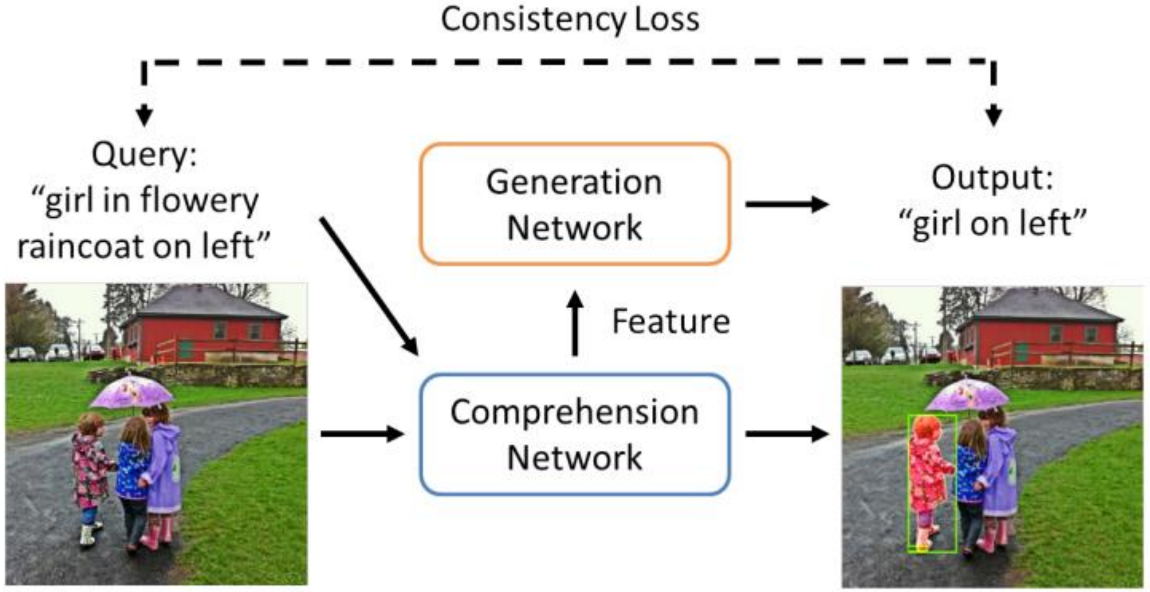
4

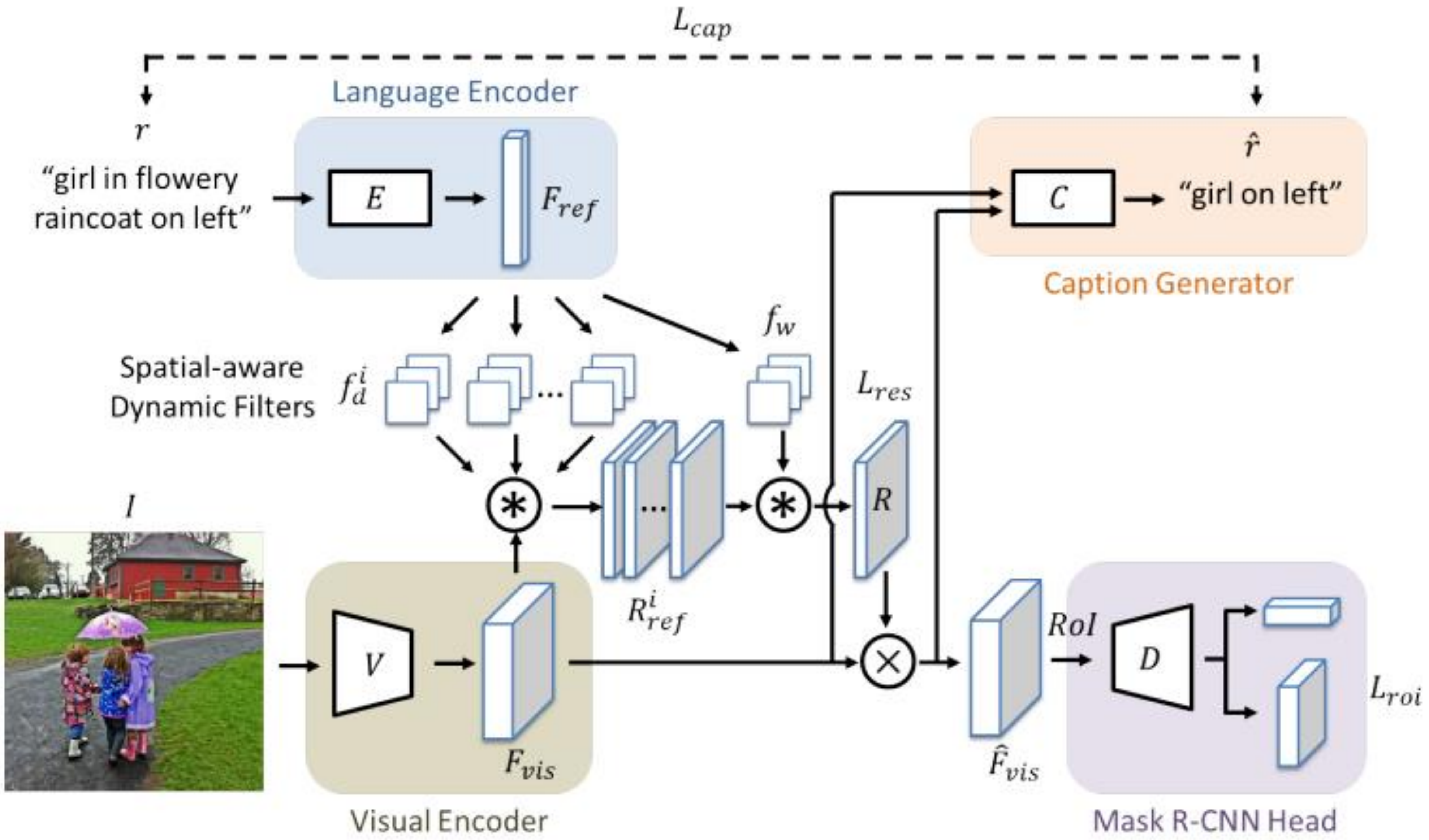
Referring Expression Object Segmentation with Caption-Aware Consistency

BMVC 2019

Yi-Wen Chen[1]
chenyiwena@gmail.com
Yi-Hsuan Tsai[2]
ytsai@nec-labs.com
Tiantian Wang[3]
tiantianwang.ice@gmail.com
Yen-Yu Lin[1]
yylin@citi.sinica.edu.tw
Ming-Hsuan Yang[34]
mhyang@ucmerced.edu

[1] Academia Sinica
[2] NEC Laboratories America
[3] University of California, Merced
[4] Google Cloud





L_{cap}
Language Encoder
r
"girl in flowery raincoat on left"
E
F_{ref}
$\hat{r}$
C
"girl on left"
Caption Generator
f_w
Spatial-aware Dynamic Filters
f_d^i
L_{res}
R
I
R_{ref}^i
V
F_{vis}
RoI
D
L_{roi}
$\hat{F}_{vis}$
Visual Encoder
Mask R-CNN Head

Segmentation from Referring Expression

- **Language Encoder**

  Matt bi-directional LSTM

$$\overrightarrow{h}_t = \overrightarrow{S}(e_t, \overrightarrow{h}_{t-1})$$
$$\overleftarrow{h}_t = \overleftarrow{S}(e_t, \overleftarrow{h}_{t+1})$$
$$F_{ref} = [\overrightarrow{h}_T, \overleftarrow{h}_1],$$

- **Visual Encoder**

  proposal-based Mask R-CNN、ResNet-101

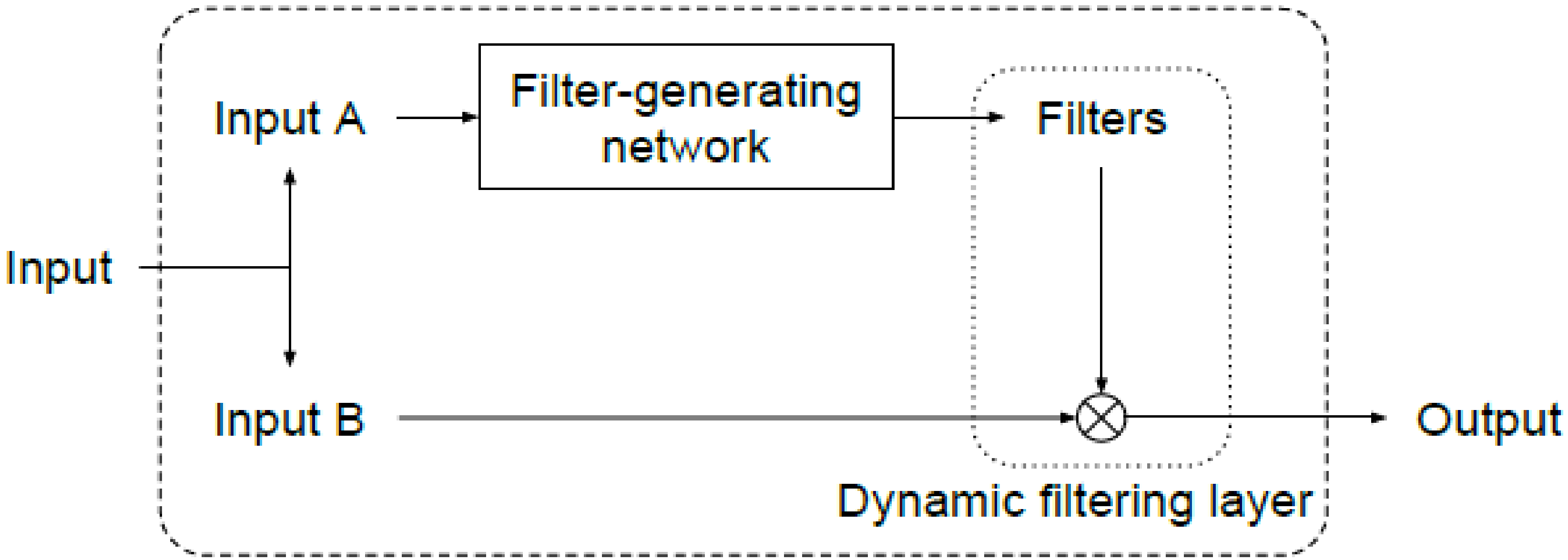
  $$F_{vis} = V(I),$$

- **Spatial-aware Dynamic Filters**
- **Baseline Objective**

$$L = L_{roi} + L_{res},$$

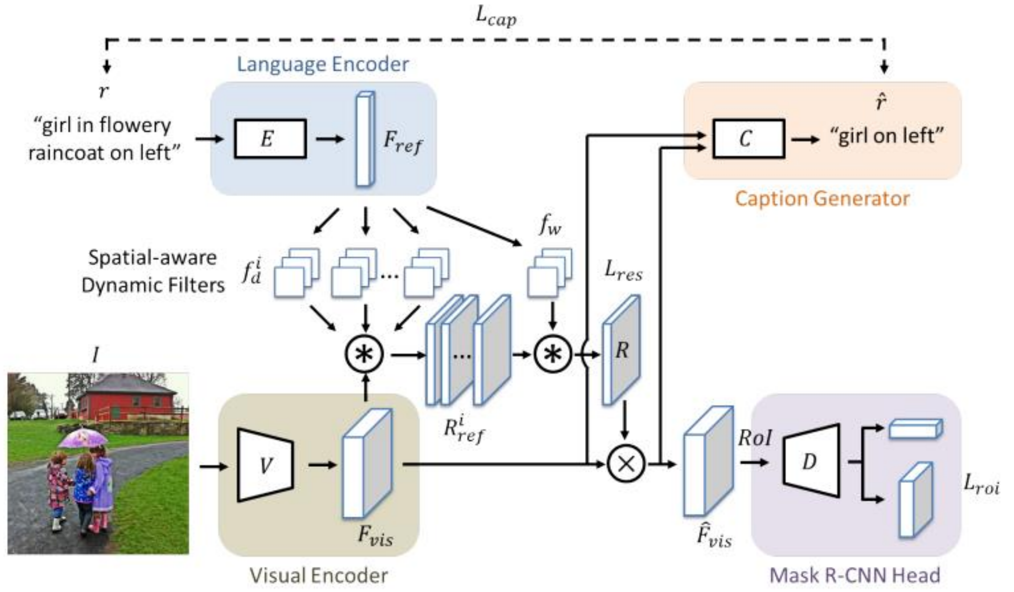
Dynamic Filter Network

Filter←→Kernel



- 1. **模型参数(model parameters):** 模型参数表示只在训练过程中被更新,在预先被初始化的层参数, 对于所有的样本都是相同的.
- 2. **动态产生参数(dynamically generated parameters)** :是由样本所决定的, 是动态生成的，不需要初始化. filter-generating 网络输出动态产生参数, 同时该网络还是有一部分的**模型参数**.

Spatial-aware Dynamic Filter



$$f_d^1 = \tanh(W_d^1 \cdot F_{ref} + b_d^1), \qquad R_{ref}^1 = f_d^1 * F_{vis}.$$

- Consider the entire image/only be able to catch the global structure but ignore spatially distributed objects

→spatial-aware dynamic filters: including up,down,left,right,horizontal and vertical middle regions←six additional fully connected layers

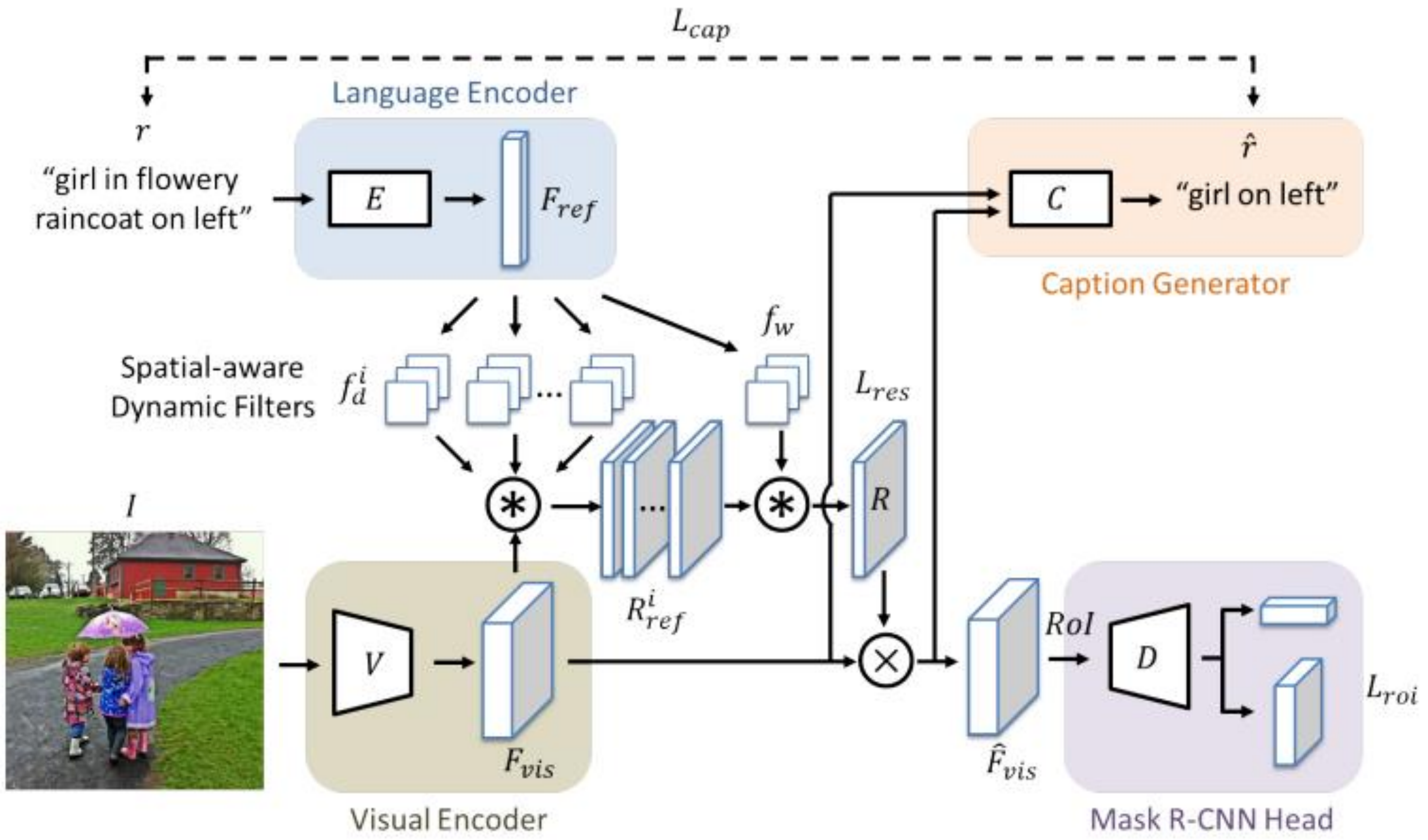
$$\{f_d^i\}_{i=2}^7 \qquad \{R_{ref}^i\}_{i=2}^7,$$

$$R_{con} = \text{concat}(R_{ref}^i) \qquad R = \sigma(f_w * R_{con}),$$

Loss:binary cross-entropy loss L_{res} → R & ground-truth object mask

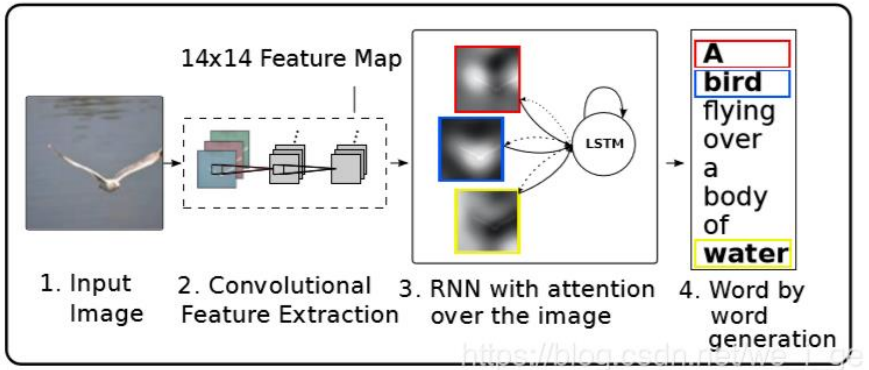
Loss

$$L = L_{roi} + L_{res},$$



- **Lroi** include: classification loss, bounding box loss, mask loss(defined in Mask R-CNN)
- **Lres** (in Spatial-aware Dynamic Filter)

A Joint Framework



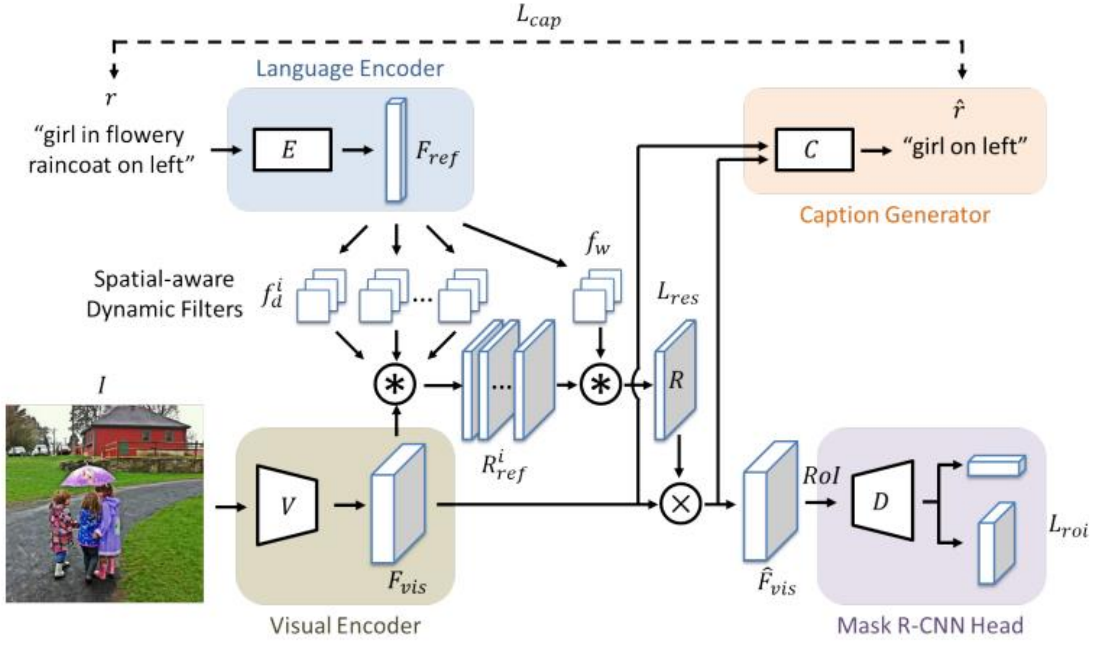
Show, Attend and Tell: Neural Image Caption Generation with Visual Attention[23]

- Caption-aware Consistency

$$L_{cap} = -\sum_{t=1}^{T} \log(p_{\theta_c}(w_t|w_1, \ldots, w_{t-1})),$$

- Overall Objective

$$L = L_{roi} + L_{res} + \alpha L_{cap},$$



location

Model	Info.	RefCOCO			RefCOCOg		
		val	testA	testB	val*	val	test
Nagaraja *et al.* [17]	C	57.30	58.60	56.40	-	-	49.50
Luo *et al.* [14]	J	-	67.94	55.18	49.07	-	-
Liu *et al.* [13]	Attr, J	-	72.08	57.29	52.35	-	-
Yu *et al.* [25]	J	-	73.78	63.83	59.84	-	-
MAttNet [26]	Attr, Attn, L, R	76.65	81.14	69.99	-	66.58	67.27
VC [27]	C	-	73.33	67.44	62.30	-	-
baseline	-	72.65	76.65	65.75	54.18	58.09	58.32
+ spatial coords [5]	L	75.89	78.57	68.54	61.37	64.10	64.21
+ spatial-aware filters	L	76.98	79.30	69.75	61.65	65.18	65.28
+ caption-aware consistency	J	76.05	78.84	69.36	60.69	64.71	63.79
full model	L, J	**77.08**	**80.34**	**70.62**	**62.34**	**65.83**	**65.44**

segmentation

Table 2: Segmentation results of our method and the competing methods on two datasets.

Model	Backbone Net	RefCOCO			RefCOCOg		
		val	testA	testB	val*	val	test
D+RMI+DCRF [12]	Deeplab101	45.18	45.69	45.57	-	-	-
RRN+LSTM+DCRF [9]	Deeplab101	55.33	57.26	53.95	36.45	-	-
MAttNet [26]	Res101	56.51	62.37	51.70	-	47.64	48.61
KWAN [20]	Deeplab101	-	-	-	36.92	-	-
DMN [16]	DPN92	49.78	54.83	45.13	36.76	-	-
Ours	Res101	58.90	61.77	53.81	44.32	46.37	46.95

Figure 3: Sample results of objects referred by various query expressions.

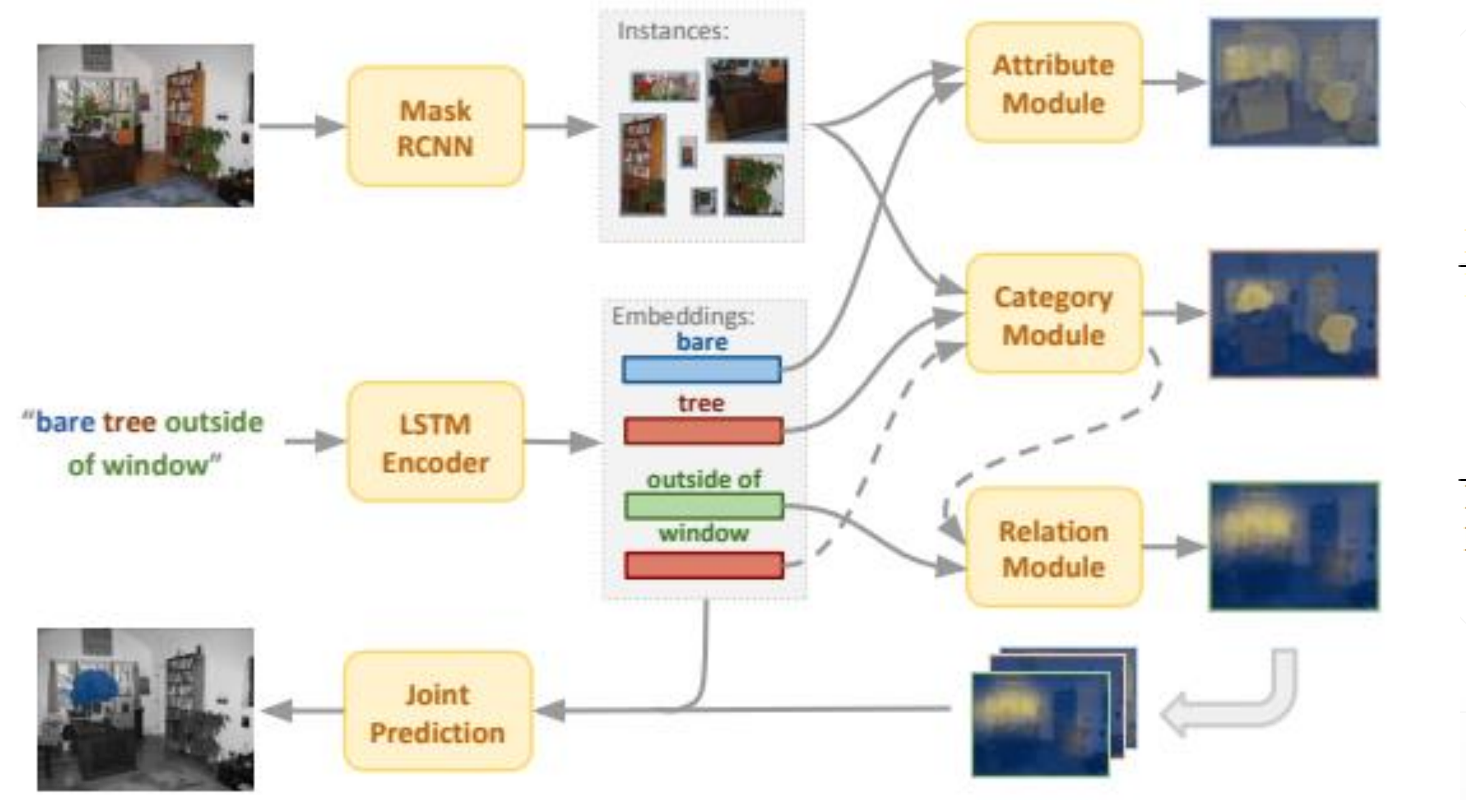
Figure 4: Sample results from different variants of the proposed model on RefCOCO.

PhraseCut: Language-based Image Segmentation in the Wild

CVPR 2020

Chenyun Wu[1] Zhe Lin[2] Scott Cohen[2] Trung Bui[2] Subhransu Maji[1]
[1]University of Massachusetts Amherst [2]Adobe Research
{chenyun,smaji}@cs.umass.edu, {zlin,scohen,bui}@adobe.com

Figure 2. **Example annotations from the VGPHRASECUT dataset**. Colors (blue, red, green) of the input phrases correspond to words that indicate attributes, categories, and relationships respectively.



Model	mean-IoU	cum-IoU	Pr@0.5	Pr@0.7	Pr@0.9
HULANet					
cat	39.9	48.8	40.8	25.9	5.5
cat+att	**41.3**	**50.8**	**42.9**	**27.8**	**5.9**
cat+rel	41.1	49.9	42.3	26.6	5.6
cat+att+rel	**41.3**	50.2	42.4	27.0	5.7
RMI	21.1	42.5	22.0	11.6	1.5
MattNet	20.2	22.7	19.7	13.5	3.0

- a modular approach for combining visual cues related to categories, attributes, and relationships
- a systematic approach to improving the performance on rare categories and attributes by leveraging predictions on more frequent ones

Figure 6. **Prediction results on VGPhraseCut dataset.** Rows from top to down are: (1) input image; (2) ground-truth segmentation and instance boxes; (3) MattNet baseline; (4) RMI baseline; (5) HULANet (cat + att + rel). See more results in the supplemental material.

Review

- Dataset
- Segmentation from Natural Language Expressions
- Recurrent Multimodal Interaction for Referring Image Segmentation
- MAttNet: Modular Attention Network for Referring Expression Comprehension
- Referring Expression Object Segmentation with Caption-Aware Consistency
- PhraseCut: Language-Based Image Segmentation in the Wild

Thank you