CRIPAC
智能感知与计算研究中心
Center for Research on Intelligent Perception and Computing

模式识别国家重点实验室
National Lab of Pattern Recognition

中国科学院自动化研究所
Institute of Automation Chinese Academy of Sciences

Improving Image and Sentence Matching with Multimodal Attention and Visual Attributes

Yan Huang

Center for Research on Intelligent Perception and Computing (CRIPAC)
National Laboratory of Pattern Recognition (NLPR)
Institute of Automation, Chinese Academy of Sciences (CASIA)

Mar. 26, 2018

CRIPAC

CRIPAC mainly focuses on the following research topic related
to national public security.

- Biometrics
- Image and Video Analysis
- Big Data and Multi-modal Computing
- Content Security and Authentication
- Sensing and Information Acquisition

CRAPIC receives regular fundings from various Government departments or agencies. It is also supported by funds of R&D projects from many other national and international sources.

CRIPAC members publish widely in leading national and international journals and conferences such as IEEE Transactions on PAMI, IEEE Transactions on Image Processing, International Journal of Computer Vision, Pattern Recognition, Pattern Recognition Letters, ICCV, ECCV, CVPR, ACCV, ICPR, ICIP, etc.

http://cripac.ia.ac.cn/en/EN/volumn/home.shtml

NVAIL

CRIPAC
智能感知与计算研究中心
Center for Research on Intelligent Perception and Computing

NVIDIA®

Artificial Intelligence Laboratory

Researches on artificial intelligence and deep learning

Outline

1 Image and Sentence Matching

2 Related Work

3 Improved Image and Sentence Matching

- 3.1 Context-modulated Multimodal Attention
- 3.2 Joint Semantic Concepts and Order Learning

4 Future Directions

Outline

1 Image and Sentence Matching

2 Related Work

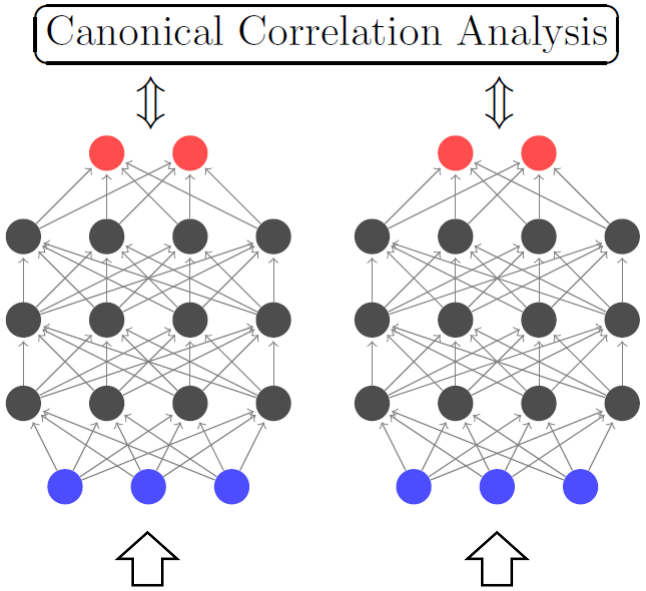
3 Improved Image and Sentence Matching

3.1 Context-modulated Multimodal Attention

3.2 Joint Semantic Concepts and Order Learning

4 Future Directions

Image and Sentence Matching



There are many kinds of vegetables

Image-sentence retrieval

Until April, the Polish forces had been slowly but steadily advancing eastward

In southeastern Washington, a stretch of the river passes through the Hanford Site, established in 1943 as part of the Manhattan Project

Tropical Storm Edouard was the first of eight named storms to form in September 2002, the most such storms for any month in the Atlantic

Image caption

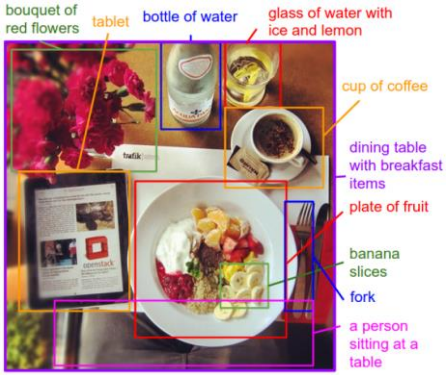


Image question answering

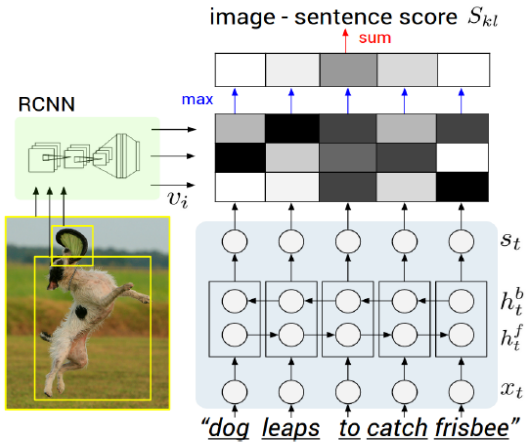
Image

GT Question 戴帽子的男孩在干什么？
What is the boy in green cap doing?

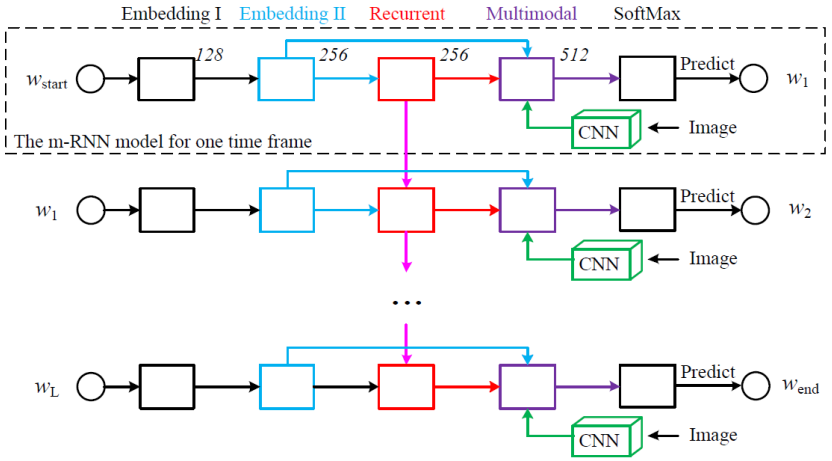
GT Answer 他在玩滑板。
He is playing skateboard.

The key challenge lies in how to well **measure the cross-modal similarity**

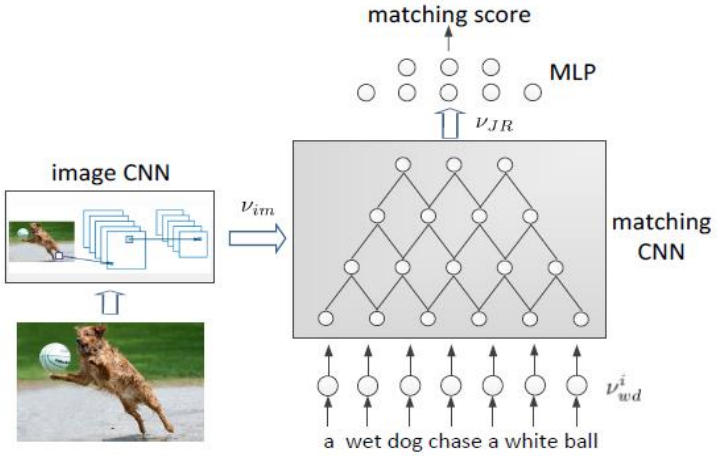
Related Work



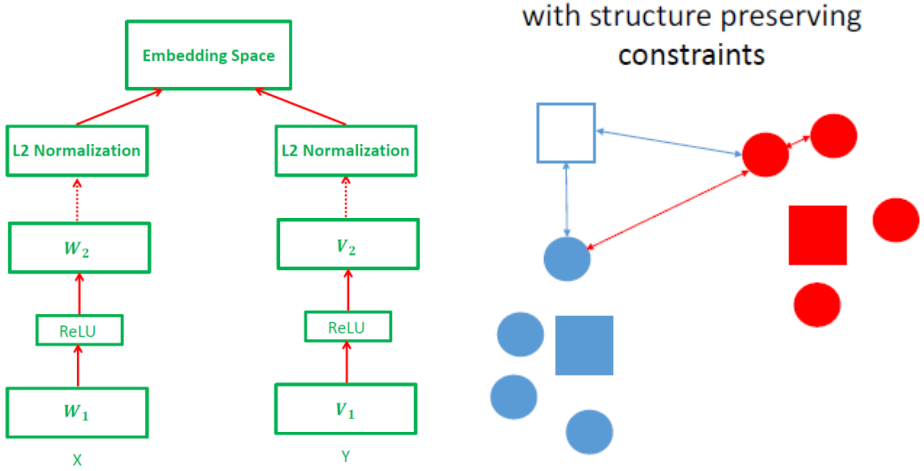
Karpathy et al., Deep Visual-Semantic Alignments for Generating Image Descriptions, CVPR, 2016.



Mao et al., Deep Captioning with Multimodal Recurrent Neural Networks, ICLR, 2015.



Ma et al., Multimodal Convolutional Neural Networks for Matching Image and Sentence, ICCV, 2015.



Wang et al., Learning Deep Structure-Preserving Image-Text Embeddings, CVPR, 2016.

Related Work

There are many kinds of vegetables

features

features

- Deep visual semantic embedding
  - Devise [1]
  - Order embedding [2]
  - Structure-preserving embedding [3]
- Deep canonical correlation analysis
  - Batch based learning [4]
  - Fisher vector on w2v [5]
  - Global + local correspondences [6]

- Sentence only describes **partial salient** image content
- Using **Global** image features might be inappropriate

[1] Frome et al., Devise: A deep visual-semantic embedding model. In NIPS, 2013.
[2] Vendrov et al., Order embeddings of images and language. In ICLR, 2016.
[3] Wang et al., Deep structure-preserving image-text embeddings. In CVPR, 2016..
[4] Yan and Mikolajczyk. Deep correlation for matching images and text. In CVPR, 2015.
[5] Klein et al., Associating neural word embeddings with deep image representations using fisher vectors. In CVPR, 2015.
[6] Plummer et al., Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In ICCV, 2015.

Outline

1 Image and Sentence Matching

2 Related Work

3 Improved Image and Sentence Matching

3.1 Context-modulated Multimodal Attention

3.2 Joint Semantic Concepts and Order Learning

4 Future Directions

Motivation

vegetables

~~people~~

~~fruit~~

~~bicycle~~

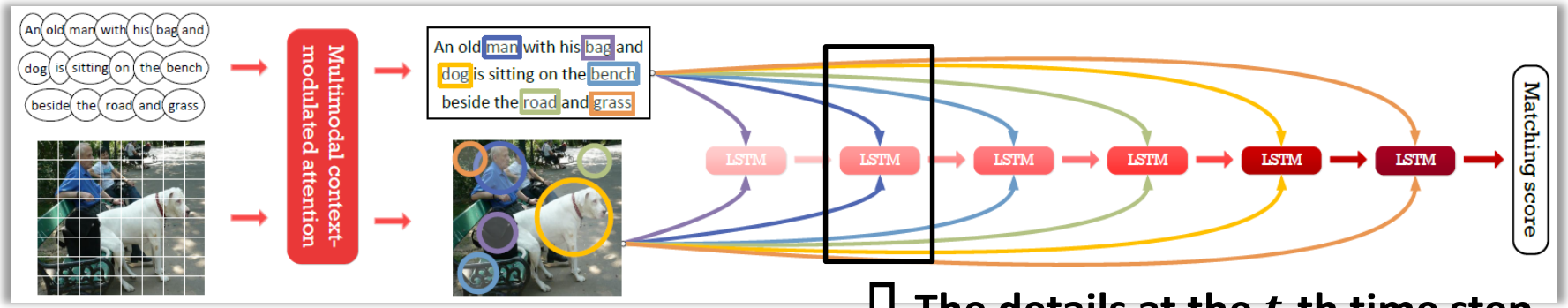
Association analysis

There are many kinds of vegetables

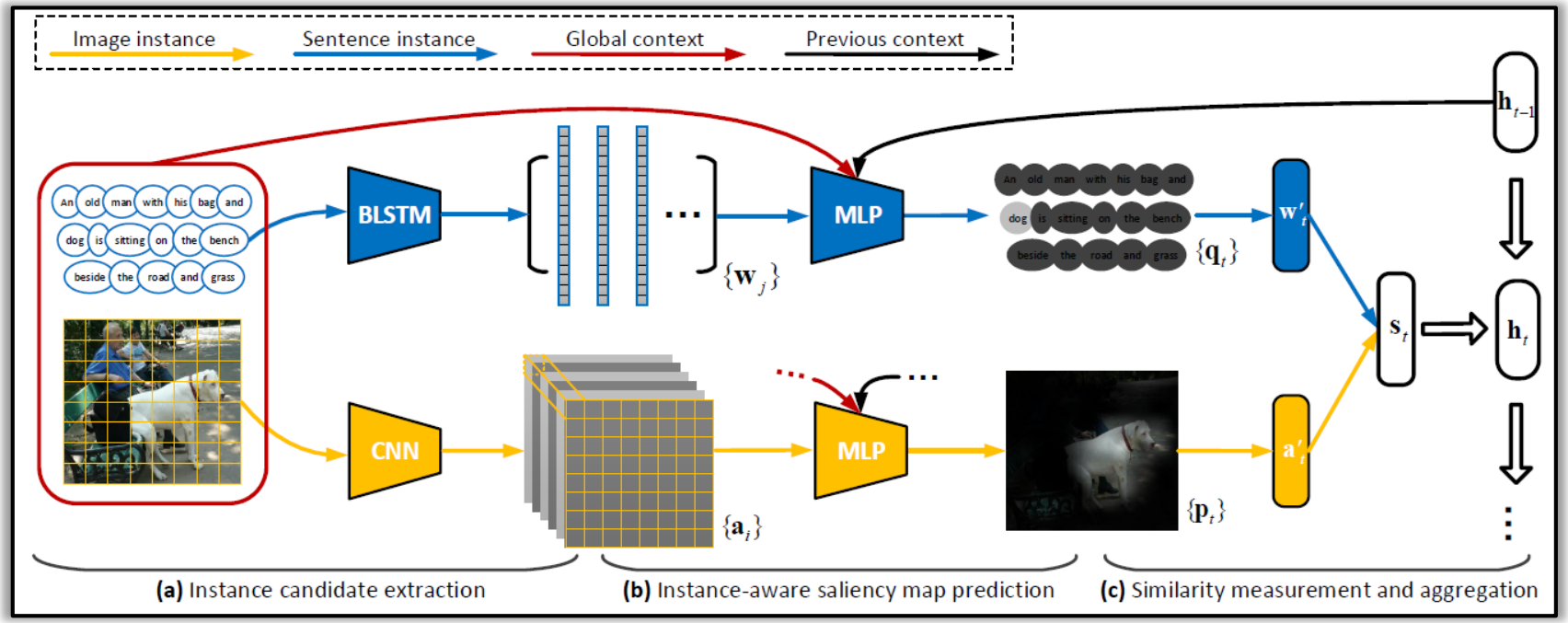
There of kinds vegetables are many

1. Image and sentence include much **redundant information**
2. Only **partial semantic instances** can be well associated

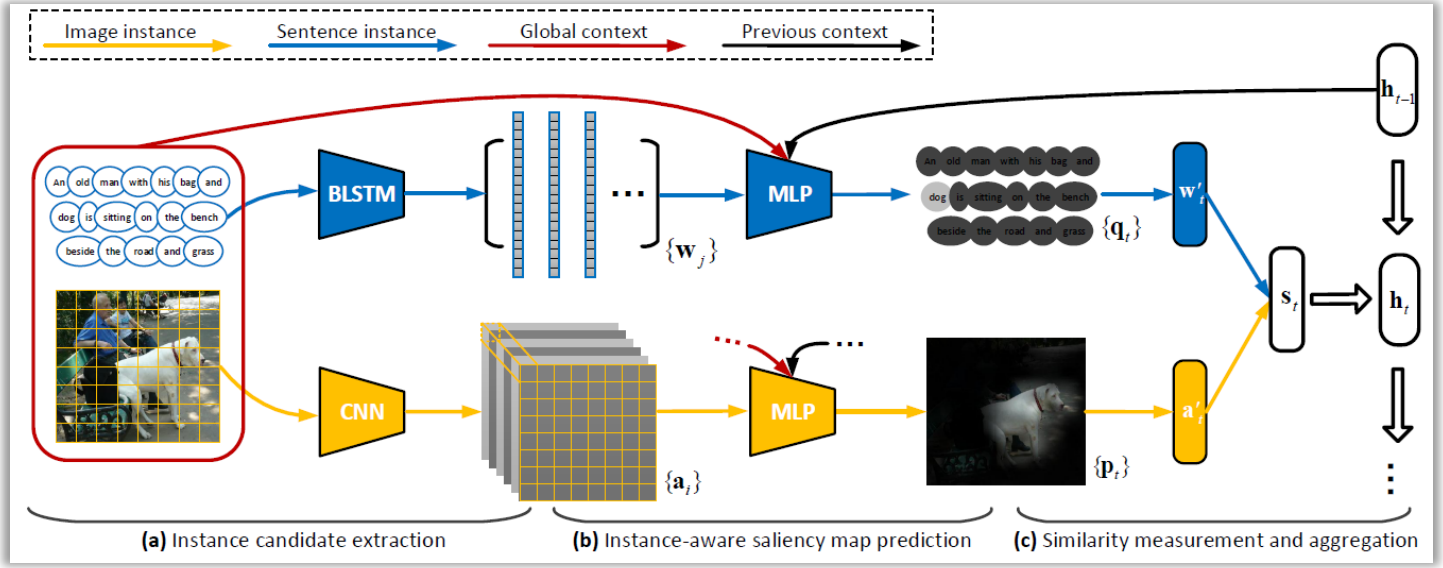
Instance-aware Image and Sentence Matching



The details at the *t*-th time step



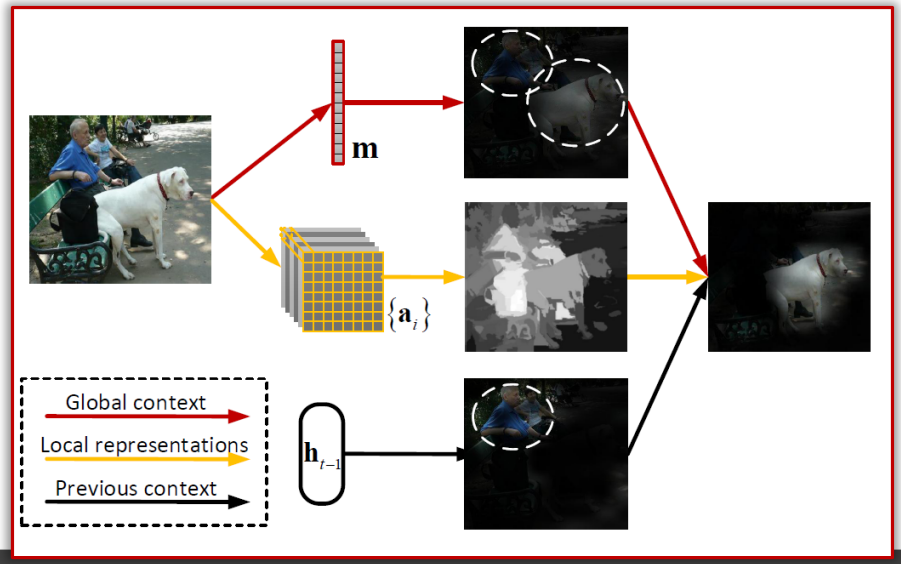
Details of LSTM at the t-th Timestep



- **Saliency probability of instance candidate:**

$$p_{t,i} = e^{\hat{p}_{t,i}} / \sum_{i=1}^{I} e^{\hat{p}_{t,i}}, \ \hat{p}_{t,i} = f_p(\mathbf{m}, \mathbf{a}_i, \mathbf{h}_{t-1}),$$
$$q_{t,j} = e^{\hat{q}_{t,j}} / \sum_{j=1}^{J} e^{\hat{q}_{t,j}}, \ \hat{q}_{t,j} = f_q(\mathbf{n}, \mathbf{w}_j, \mathbf{h}_{t-1})$$

$$f_p(\mathbf{m}, \mathbf{a}_i, \mathbf{h}_{t-1}) = \mathbf{w}_p(\sigma(\mathbf{m}W_\mathbf{m} + \mathbf{b}_\mathbf{m}) + \sigma(\mathbf{a}_i W_\mathbf{a} + \mathbf{b}_\mathbf{a}) + \sigma(\mathbf{h}_{t-1} W_\mathbf{h} + \mathbf{b}_\mathbf{h})) + b_p$$



- **Instance representation:**

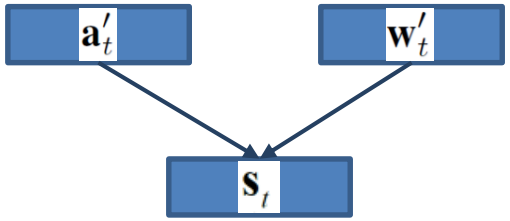
$$\mathbf{a}'_t = \sum_{i=1}^{I} p_{t,i} \mathbf{a}_i, \quad \mathbf{w}'_t = \sum_{j=1}^{J} q_{t,j} \mathbf{w}_j$$

Local Similarity Measurement and Aggregation

Image-sentence instance representation:

$$\mathbf{a}'_t = \sum_{i=1}^{I} p_{t,i}\mathbf{a}_i, \quad \mathbf{w}'_t = \sum_{j=1}^{J} q_{t,j}\mathbf{w}_j$$

Measure their local similarity with a two-way MLP

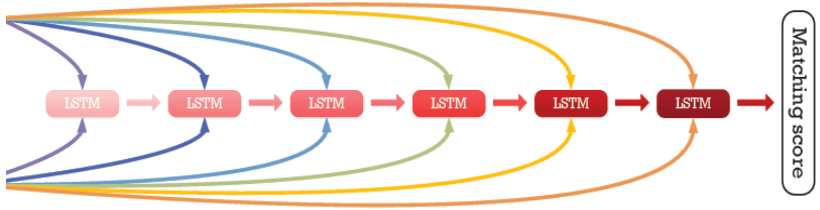


Feed into the current hidden state

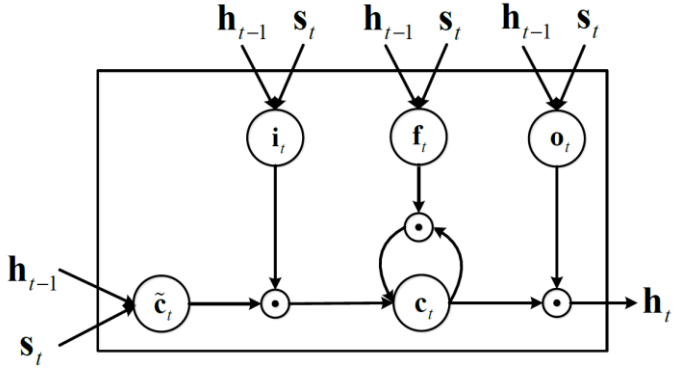
Global similarity:

$$s = \mathbf{w}_{hs}\left(\sigma\left(W_{hh}\mathbf{h}_t + \mathbf{b}_h\right)\right) + b_s$$

Aggregate all the similarities



Measure local similarities at all timesteps



$$\mathbf{i}_t = \sigma(W_{si}\mathbf{s}_t + W_{hi}\mathbf{h}_{t-1} + \mathbf{b}_i),$$
$$\mathbf{f}_t = \sigma(W_{sf}\mathbf{s}_t + W_{hf}\mathbf{h}_{t-1} + \mathbf{b}_f),$$
$$\mathbf{c}_t = \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \tanh(W_{sc}\mathbf{s}_t + W_{hc}\mathbf{h}_{t-1} + \mathbf{b}_c),$$
$$\mathbf{o}_t = \sigma(W_{so}\mathbf{s}_t + W_{ho}\mathbf{h}_{t-1} + \mathbf{b}_o), \mathbf{h}_t = \mathbf{o}_t \odot \tanh(\mathbf{c}_t)$$

Detailed formulation of LSTM at the *t*-th timestep

Model Learning

- **Structured objective function**
  - matched scores are larger than mismatched ones

$$\sum_{ik} \max\{0, m - s_{ii} + s_{ik}\} + \max\{0, m - s_{ii} + s_{ki}\}$$

- **Pairwise doubly stochastic regularization**
  - constrain the sum of saliency values of an instance candidate at all timesteps to be 1
  - encourages the model to **pay equal attention to every instance** rather than a certain one

$$\lambda\left(\sum_i (1 - \sum_t p_{t,i}) + \sum_j (1 - \sum_t q_{t,j})\right)$$

- **Optimize the objective using stochastic gradient descent**

Experimental Datasets

- Flickr 30k dataset
  - from the Flickr.com website
  - 31784 images
  - each image has 5 captions
  - use the public training, validation and testing splits, which contain 28000, 1000 and 1000 images, respectively
- Microsoft COCO dataset
  - 82783 images
  - each image has 5 captions
  - use the public training, validation and testing splits, with 82783, 4000 and 1000 images, respectively

1. A man in street racer armor is examining the tire of another racers motor bike.
2. The two racers drove the white bike down the road.
3. ……

1. A firefighter extinguishes a fire under the hood of a car.
2. a fireman spraying water into the hood of small white car on a jack
3. ……

Implementation Details

- Evaluation criterions
  - "**R@1**", "**R@5**" and "**R@10**", i.e., recall rates at the top 1, 5 and 10 results
  - "**Med r**" is the median rank of the first ground truth result
  - "**Sum**":

$$\text{Sum} = \underbrace{\text{R@1} + \text{R@5} + \text{R@10}}_{\text{Image annotation}} + \underbrace{\text{R@1} + \text{R@5} + \text{R@10}}_{\text{Image retrieval}}$$

- Feature extraction

	Image	Sentence
Global context	the feature vector in "fc7" layer of the 19-layer VGG network	the last hidden state of a visual-semantic embedding framework
Local representation	512 feature maps (size: 14x14) in "conv5-4" layer	multiple hidden states of a bidirectional LSTM

Implementation Details

- Five variants of the proposed sm-LSTMs
  - mean vector: use mean instead of weighted sum vector
  - attention: use conventional attention scheme
  - context: use global context modulation
  - ensemble: sum multiple cross-modal similarity matrices

	Mean vector	Attention	Context	Ensemble
sm-LSTM-mean	√			
sm-LSTM-att		√		
sm-LSTM-ctx			√	
sm-LSTM		√	√	
sm-LSTM*		√	√	√

Results on Flickr30K & Microsoft COCO

Table 1. Bidirectional image and sentence retrieval results on Flickr30k.

Method	Image Annotation				Image Retrieval				Sum
	R@1	R@5	R@10	Med r	R@1	R@5	R@10	Med r	
RVP (T+I) [4]	12.1	27.8	47.8	11	12.7	33.1	44.9	12.5	178.4
Deep Fragment [13]	14.2	37.7	51.3	10	10.2	30.8	44.2	14	188.4
DCCA [34]	16.7	39.3	52.9	8	12.6	31.0	43.0	15	195.5
NIC [31]	17.0	-	56.0	7	17.0	-	57.0	7	-
DVSA (BRNN) [14]	22.2	48.2	61.4	4.8	15.2	37.7	50.5	9.2	235.2
MNLM [15]	23.0	50.7	62.9	5	16.8	42.0	56.5	8	251.9
LRCN [7]	-	-	-	-	17.5	40.3	50.8	9	-
m-RNN [22]	35.4	63.8	73.7	3	22.8	50.7	63.1	5	309.5
FV$^{\dagger *}$ [17]	35.0	62.0	73.8	3	25.0	52.7	66.0	5	314.5
m-CNN* [21]	33.6	64.1	74.9	3	26.2	56.3	69.6	4	324.7
RTP+FV$^{\dagger *}$ [26]	37.4	63.1	74.3	-	26.0	56.0	69.3	-	326.1
RNN+FV† [19]	34.7	62.7	72.6	3	26.2	55.1	69.2	4	320.5
DSPE+FV† [32]	40.3	68.9	79.9	-	29.7	60.1	72.1	-	351.0
Ours:									
sm-LSTM-mean	25.9	53.1	65.4	5	18.1	43.3	55.7	8	261.5
sm-LSTM-att	27.0	53.6	65.6	5	20.4	46.4	58.1	7	271.1
sm-LSTM-ctx	33.5	60.6	70.8	3	23.6	50.4	61.3	5	300.1
sm-LSTM	42.4	67.5	79.9	**2**	28.2	57.0	68.4	4	343.4
sm-LSTM*	**42.5**	**71.9**	**81.5**	**2**	**30.2**	**60.4**	**72.3**	**3**	**358.7**

[4] Chen and Zitnick. Mind's eye: A recurrent visual representation for image caption generation. In CVPR, 2015.

[7] Donahue et al., Long-term recurrent convolutional networks for visual recognition and description. In CVPR, 2015.

[13] Karpathy et al., Deep fragment embeddings for bidirectional image sentence mapping. In NIPS, 2014.

[14] Karpathy and Li. Deep visual-semantic alignments for generating image descriptions. In CVPR, 2015.

[15] Kiros et al., Unifying visual-semantic embeddings with multimodal neural language models. TACL, 2015.

[17] Klein et al., Associating neural word embeddings with deep image representations using fisher vectors. In CVPR, 2015.

[19] Lev et al., Rnn fisher vectors for action recognition and image annotation. In ECCV, 2016.

Table 2. Bidirectional image and sentence retrieval results on COCO.

Method	Image Annotation				Image Retrieval				Sum
	R@1	R@5	R@10	Med r	R@1	R@5	R@10	Med r	
STD$^{\dagger *}$ [16]	33.8	67.7	82.1	3	25.9	60.0	74.6	4	344.1
m-RNN [22]	41.0	73.0	83.5	2	29.0	42.2	77.0	3	345.7
FV$^{\dagger *}$ [17]	39.4	67.9	80.9	2	25.1	59.8	76.6	4	349.7
DVSA [14]	38.4	69.9	80.5	**1**	27.4	60.2	74.8	3	351.2
MNLM [15]	43.4	75.7	85.8	2	31.0	66.7	79.9	3	382.5
m-CNN* [21]	42.8	73.1	84.1	2	32.6	68.6	82.8	3	384.0
RNN+FV† [19]	40.8	71.9	83.2	2	29.6	64.8	80.5	3	370.8
OEM [30]	46.7	-	88.9	2	37.9	-	85.9	**2**	-
DSPE+FV† [32]	50.1	79.7	89.2	-	39.6	75.2	86.9	-	420.7
Ours:									
sm-LSTM-mean	33.1	65.3	78.3	3	25.1	57.9	72.2	4	331.9
sm-LSTM-att	36.7	69.7	80.8	2	29.1	64.8	78.4	3	359.5
sm-LSTM-ctx	39.7	70.2	84.0	2	32.7	68.1	81.3	3	376.0
sm-LSTM	52.4	81.7	90.8	**1**	38.6	73.4	84.6	**2**	421.5
sm-LSTM*	**53.2**	**83.1**	**91.5**	**1**	**40.7**	**75.8**	**87.4**	**2**	**431.8**

[21] Ma et al., Multimodal convolutional neural networks for matching image and sentence. In ICCV, 2015.

[22] Mao et al., Explain images with multimodal recurrent neural networks. In ICLR, 2015.

[26] Plummer et al., Flickr30k entities: Collecting region to phrase correspondences for richer image to sentence models. In ICCV, 2015.

[34] Yan and Mikolajczyk. Deep correlation for matching images and text. In CVPR, 2015.

[30] Vendrov et al., Order embeddings of images and language. In ICLR, 2016.

[31] Vinyals et al., and D. Erhan. Show and tell: A neural image caption generator. In CVPR, 2015.

[32] Wang et al., Learning deep structure preserving image-text embeddings. In CVPR, 2016.

Analysis on Hyperparameters

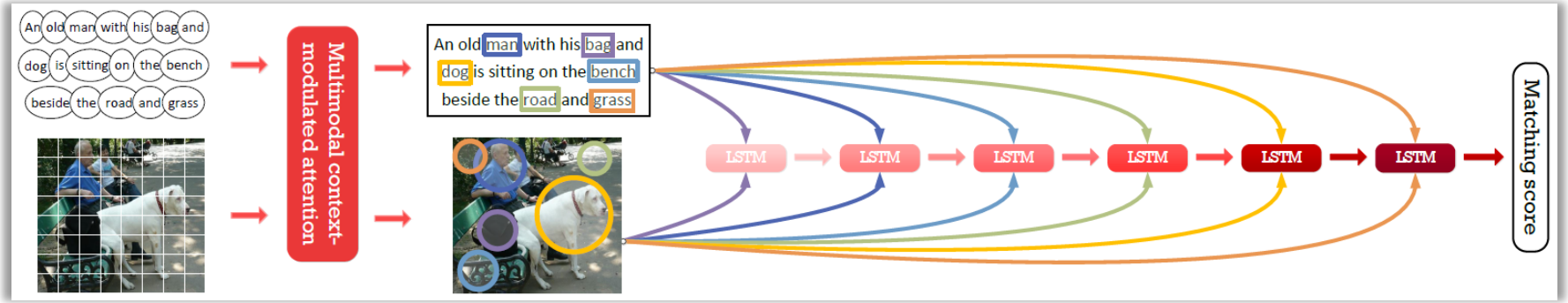


Table 3. The impact of **different numbers of timesteps** on the Flick30k dataset.

	Image Annotation			Image Retrieval		
	R@1	R@5	R@10	R@1	R@5	R@10
$T = 1$	38.8	65.7	76.8	28.0	56.6	68.2
$T = 2$	38.0	**68.9**	77.9	28.1	56.5	68.1
$T = 3$	**42.4**	67.5	**79.9**	**28.2**	**57.0**	**68.4**
$T = 4$	38.2	67.6	78.5	27.5	56.6	68.0
$T = 5$	38.1	68.2	78.4	28.1	56.0	67.9

$\boldsymbol{T}$: the number of timesteps in the sm-LSTM.

Table 4. The impact of **different values of the balancing parameter** on the Flick30k dataset.

	Image Annotation			Image Retrieval		
	R@1	R@5	R@10	R@1	R@5	R@10
$\lambda = 0$	37.9	65.8	77.7	27.2	55.4	67.6
$\lambda = 1$	38.0	66.2	77.8	27.4	55.6	67.7
$\lambda = 10$	38.4	67.4	77.7	27.5	56.1	67.6
$\lambda = 100$	**42.4**	**67.5**	**79.9**	**28.2**	**57.0**	**68.4**
$\lambda = 1000$	40.2	67.1	78.6	27.8	56.9	67.9

$\boldsymbol{\lambda}$: the balancing parameter between the structured objective and regularization.

Usefulness of Global Context

Table 5. **Attended image instances** at three different timesteps.

Input image	sm-LSTM			sm-LSTM-att		
	1^{st}step	2^{nd} step	3^{rd} step	1^{st}step	2^{nd} step	3^{rd} step

Instance-aware Saliency Maps

Figure 2. Visualization of **attended image and sentence instances** at three different timesteps.

Conclusion

- Conclusion
  - **selectively process redundant information** with context-modulated attention
  - **gradually accumulate salient information** with multi-modal LSTM-RNN

For more details, please refer to the following paper:

1. Yan Huang, Wei Wang, and Liang Wang**, Instance-aware Image and Sentence Matching with Selective Multimodal LSTM**. *CVPR*, pp. 2310-2318, 2017.

Outline

1 Image and Sentence Matching

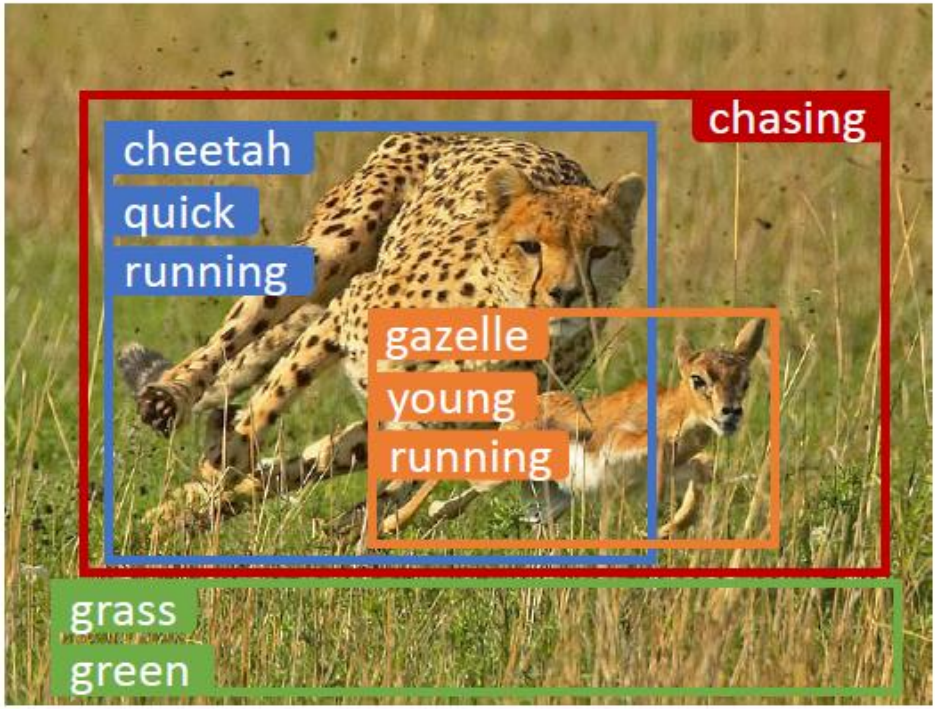
2 Related Work

3 Improved Image and Sentence Matching

- 3.1 Context-modulated Multimodal Attention
- 3.2 Joint Semantic Concepts and Order Learning

4 Future Directions

Only Instances Are Not Enough



- **Semantic concepts:**

Objects:	Properties:	Actions:
cheetah	quick	chasing
gazelle	young	running
grass	green	running

- **Semantic order:**

  cheetah chasing gazelle grass

  …

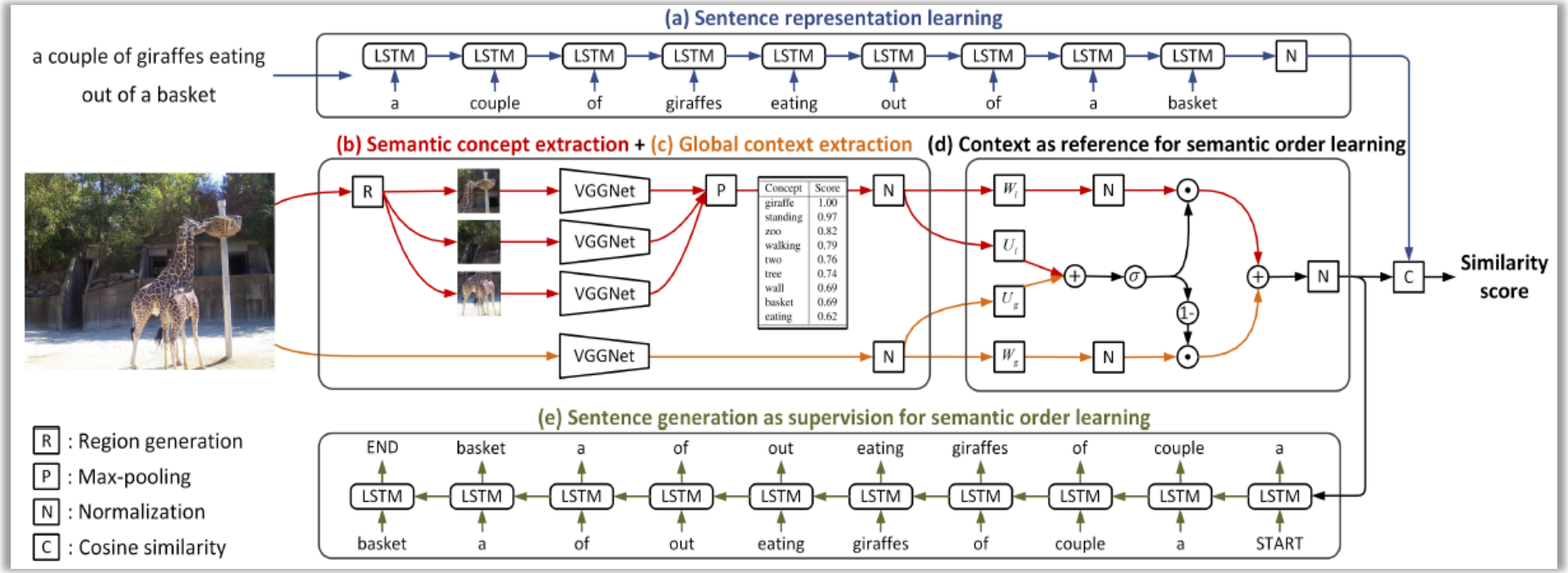
- **Matched sentence:**

  A quick cheetah is chasing a young gazelle on grass.

- "**Semantic concepts**" include separated instances and their descriptive properties
- Different "**semantic orders**" lead to various semantic meanings

Joint Semantic Concept and Order Learning

- **Improve image representations** by learning semantic concepts and then organizing them in the correct semantic order



- Multi-regional multilabel CNN for **concept prediction**
- Global context modulation & sentence generation for **order learning**

Semantic Concept Prediction

- **Process** the existing dataset, **select** the desired concept, and **reduce** the size of vocabulary
- **Learn** a multi-label CNN and perform **testing** in a multi-region way

A couple of giraffes eating out of basket

⇩

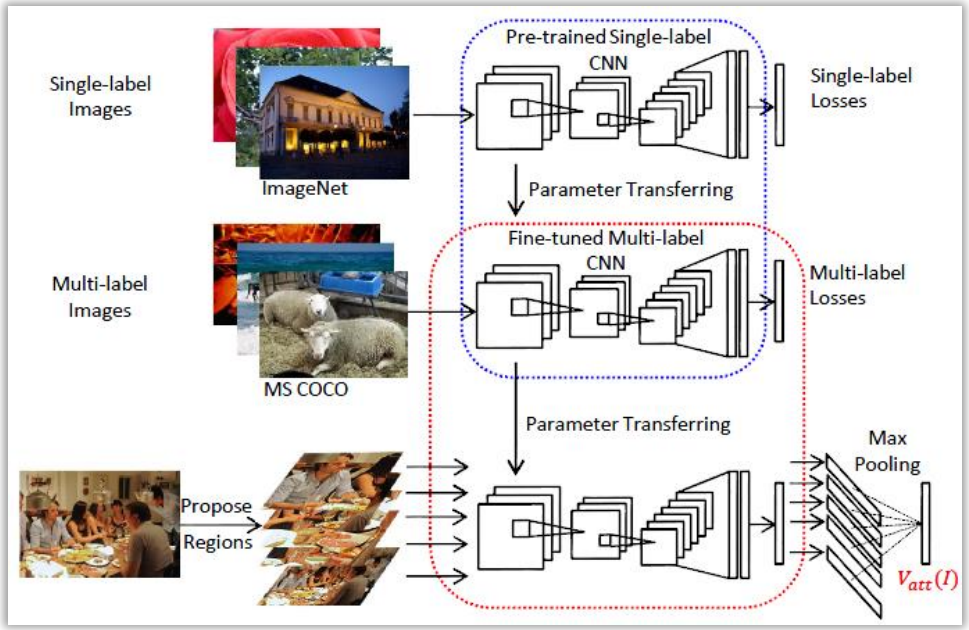
A, couple, of, giraffes, eating, out, of, basket

⇩

~~A~~, ~~couple~~, ~~of~~, giraffes, eating, ~~out~~, ~~of~~, basket

⇩

⇔ giraffe, eating, basket



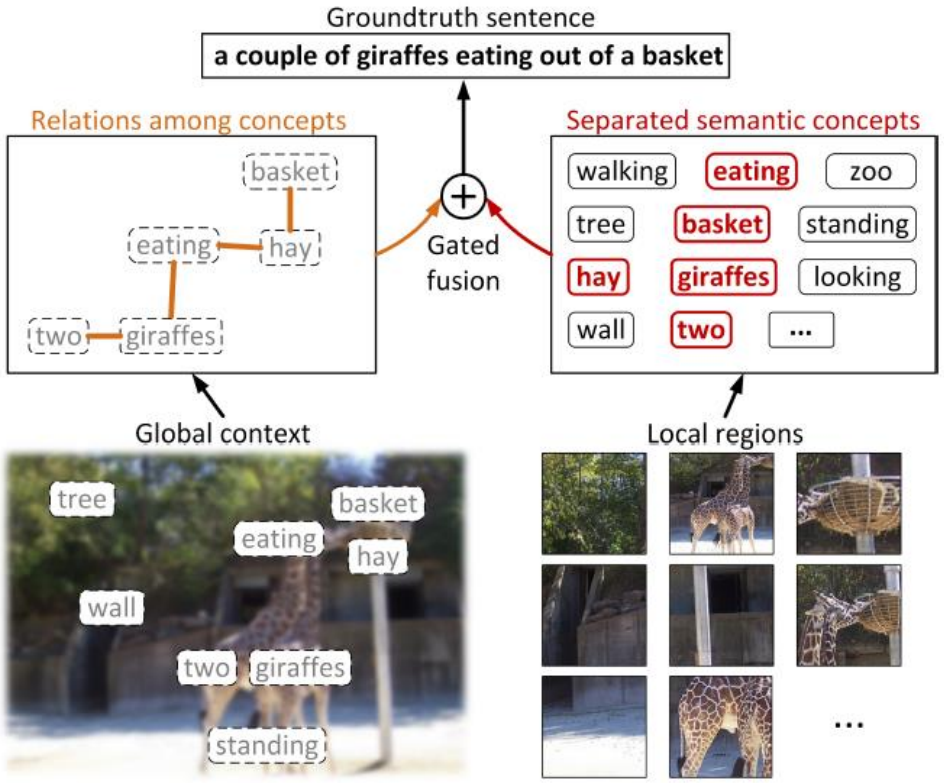
Multi-regional multi-label CNN [1]

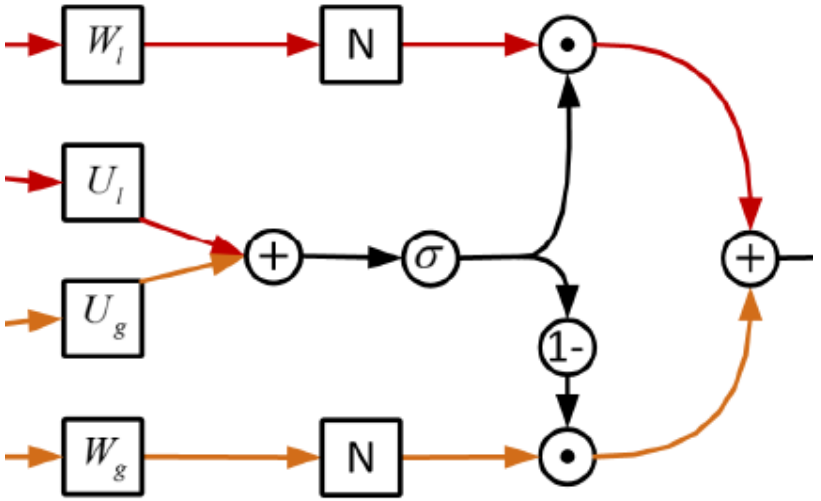
[1] Wu et al., What value do explicit high level concepts have in vision to language problems? In CVPR, 2016.
[2] Wei et al., Cnn: Single-label to multi-label. arXiv, 2014.
[3] Fang et al., From captions to visual concepts and back. In CVPR, 2015.
[4] Wu et al., Deep multiple instance learning for image classification and auto-annotation. In CVPR, 2015.
[5] Gong et al., Deep convolutional ranking for multilabel image annotation. arXiv, 2013.

Use Global Context as Reference

- **Directly learn the semantic order is very difficult !**

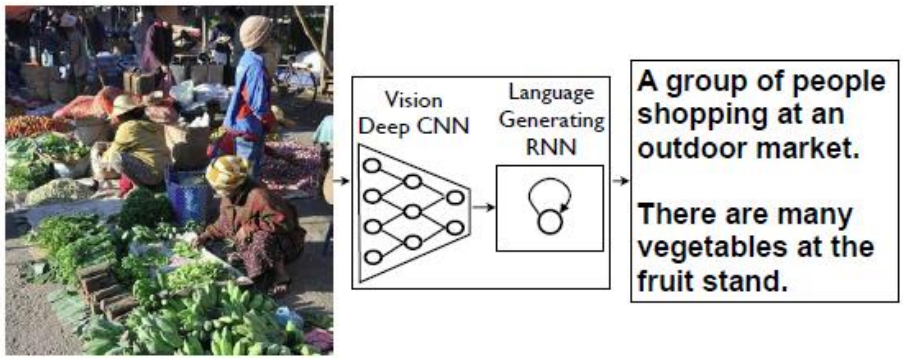
  - The global context indicates **the spatial relation** of concepts
  - Selectively **balance the importance** of concept and context

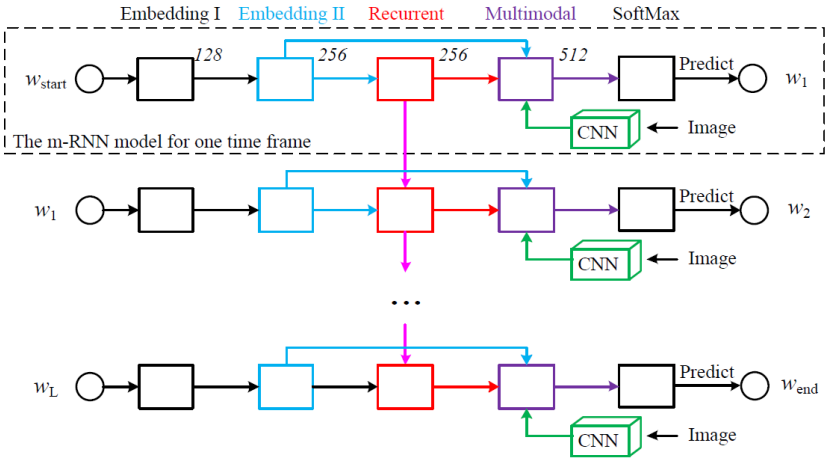




Use Sentence Generation as Supervision

- A straightforward approach is to **directly generate a sentence based on the predicted concepts**, but not working



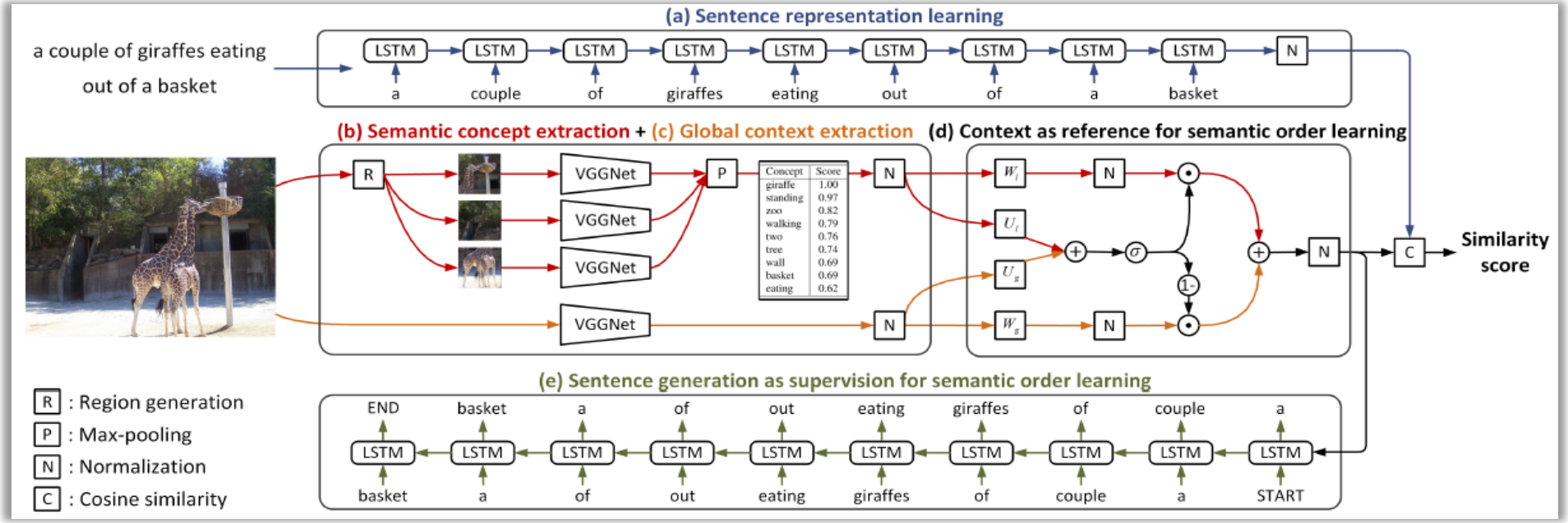


- Regard **the fused context and concepts** as image representation
- use **the groundtruth semantic order** to supervise it during sentence generation

$$L = L_{mat} + \lambda \times L_{gen}$$

$$\sum_{ik} \max\{0, m - s_{ii} + s_{ik}\} + \max\{0, m - s_{ii} + s_{ki}\} \qquad -\sum_t \log P(\mathbf{w}_t | \mathbf{w}_{t-1}, \mathbf{w}_{t-2}, \cdots, \mathbf{w}_0, \mathbf{x}, \mathbf{p})$$

Use Sentence Generation as Supervision



- Regard **the fused context and concepts** as image representation
- Use **the groundtruth semantic order** to supervise it during sentence generation

$$L = L_{mat} + \lambda \times L_{gen}$$

$$\sum_{ik} \max\{0, m - s_{ii} + s_{ik}\} + \max\{0, m - s_{ii} + s_{ki}\}$$

$$-\sum_t \log P(\mathbf{w}_t | \mathbf{w}_{t-1}, \mathbf{w}_{t-2}, \cdots, \mathbf{w}_0, \mathbf{x}, \mathbf{p})$$

Evaluation of Ablation Models

Table 6. The experimental settings of **ablation models**.

	1-crop	10-crop	context	concept	sum	gate	sentence	generation	sampling	shared	non-shared
ctx (1-crop)	√		√								
ctx		√	√								
ctx + sen		√	√				√				√
ctx + gen (S)		√	√					√	√		√
ctx + gen (E)		√	√					√		√	
ctx + gen		√	√					√			√
cnp				√							
cnp + gen				√				√			
cnp + ctx (C)		√	√	√	√						
cnp + ctx		√	√	√		√					
cnp + ctx + gen		√	√	√		√		√			√

- "**10-crop**" : crop 10 regions from images
- "**concept**" and "**context**": use semantic concepts and context
- "**sum**" and "**gate**": combine semantic concepts and context via feature concatenation and gated unit, respectively
- "**sentence**", "**generation**" and "**sampling**": image captionging, sentence generation and scheduled sampling
- "**share**" and "**non-shared**": two word embedding matrices

Evaluation of Ablation Models

Table 7. Comparison results by **ablation models** on the Flickr30k and MSCOCO datasets.

Method	Flickr30k dataset							MSCOCO dataset						
	Image Annotation			Image Retrieval			mR	Image Annotation			Image Retrieval			mR
	R@1	R@5	R@10	R@1	R@5	R@10		R@1	R@5	R@10	R@1	R@5	R@10	
ctx (1-crop)	29.8	58.4	70.5	22.0	47.9	59.3	48.0	43.3	75.7	85.8	31.0	66.7	79.9	63.8
ctx	33.8	63.7	75.9	26.3	55.4	67.6	53.8	44.7	78.2	88.3	37.0	73.2	85.7	67.9
ctx + sen	22.8	48.6	60.8	19.1	46.0	59.7	42.8	39.2	73.3	85.5	32.4	70.1	83.7	64.0
ctx + gen (S)	34.4	64.5	77.0	27.1	56.3	68.3	54.6	45.7	78.7	88.7	37.3	73.8	85.8	68.4
ctx + gen (E)	35.5	63.7	75.9	27.4	55.9	67.5	54.3	46.9	78.8	89.2	37.3	73.9	85.9	68.7
ctx + gen	35.6	66.3	76.9	27.9	56.8	68.2	55.3	46.9	79.2	89.3	37.9	74.0	85.9	68.9
cnp	30.9	60.9	72.4	23.1	52.5	64.8	50.8	59.5	86.9	93.6	48.5	81.4	90.9	76.8
cnp + gen	31.5	61.7	74.5	25.0	53.4	64.9	51.8	62.6	89.0	94.7	50.6	82.4	91.2	78.4
cnp + ctx (C)	39.9	71.2	81.3	31.4	61.7	72.8	59.7	62.8	89.2	95.5	53.2	85.1	93.0	79.8
cnp + ctx	42.4	72.9	81.5	32.4	63.5	73.9	61.1	65.3	90.0	96.0	54.2	85.9	93.5	80.8
cnp + ctx + gen	**44.2**	**74.1**	**83.6**	**32.8**	**64.3**	**74.9**	**62.3**	**66.4**	**91.3**	**96.6**	**55.5**	**86.5**	**93.7**	**81.8**

Table 8. Comparison results of **balancing parameter λ** on the Microsoft COCO dataset.

λ	Flickr30k dataset							MSCOCO dataset						
	Image Annotation			Image Retrieval			mR	Image Annotation			Image Retrieval			mR
	R@1	R@5	R@10	R@1	R@5	R@10		R@1	R@5	R@10	R@1	R@5	R@10	
0	42.4	72.9	81.5	32.4	63.5	73.9	61.1	65.3	90.0	96.0	54.2	85.9	93.5	80.8
0.01	43.1	72.8	83.5	32.8	63.2	73.6	61.5	66.3	91.2	96.5	55.4	86.5	93.7	81.6
1	**44.2**	**74.1**	**83.6**	**32.8**	**64.3**	**74.9**	**62.3**	**66.6**	**91.8**	**96.6**	**55.5**	**86.6**	**93.8**	**81.8**
100	42.3	73.8	83.1	32.5	63.3	74.0	61.5	65.0	90.5	96.1	54.9	86.3	93.7	81.1

Comparison with State-of-the-art Methods

Table 9. Bidirectional image and sentence retrieval results on two datasets.

	Method	Flickr30k dataset							MSCOCO dataset						
		Image Annotation			Image Retrieval			mR	Image Annotation			Image Retrieval			mR
		R@1	R@5	R@10	R@1	R@5	R@10		R@1	R@5	R@10	R@1	R@5	R@10	
VGGNet	m-RNN [21]	35.4	63.8	73.7	22.8	50.7	63.1	51.6	41.0	73.0	83.5	29.0	42.2	77.0	57.6
	FV [14]	35.0	62.0	73.8	25.0	52.7	66.0	52.4	39.4	67.9	80.9	25.1	59.8	76.6	58.3
	DVSA [12]	22.2	48.2	61.4	15.2	37.7	50.5	39.2	38.4	69.9	80.5	27.4	60.2	74.8	58.5
	MNLM [13]	23.0	50.7	62.9	16.8	42.0	56.5	42.0	43.4	75.7	85.8	31.0	66.7	79.9	63.8
	m-CNN [20]	33.6	64.1	74.9	26.2	56.3	69.6	54.1	42.8	73.1	84.1	32.6	68.6	82.8	64.0
	RNN+FV [16]	34.7	62.7	72.6	26.2	55.1	69.2	53.4	40.8	71.9	83.2	29.6	64.8	80.5	61.8
	OEM [29]	-	-	-	-	-	-	-	46.7	78.6	88.9	37.9	73.7	85.9	68.6
	VQA [18]	33.9	62.5	74.5	24.9	52.6	64.8	52.2	50.5	80.1	89.7	37.0	70.9	82.9	68.5
	RTP [26]	37.4	63.1	74.3	26.0	56.0	69.3	54.3	-	-	-	-	-	-	-
	DSPE [32]	40.3	68.9	79.9	29.7	60.1	72.1	58.5	50.1	79.7	89.2	39.6	75.2	86.9	70.1
	sm-LSTM [11]	42.5	71.9	81.5	30.2	60.4	72.3	59.8	53.2	83.1	91.5	40.7	75.8	87.4	72.0
	2WayNet [4]	**49.8**	67.5	-	**36.0**	55.6	-	-	55.8	75.2	-	39.7	63.3	-	-
	DAN [23]	41.4	73.5	82.5	31.8	61.7	72.5	60.6	-	-	-	-	-	-	-
	VSE++ [5]	41.3	69.0	77.9	31.4	59.7	71.2	58.4	57.2	85.1	93.3	45.9	78.9	89.1	74.6
	Ours	44.2	**74.1**	**83.6**	32.8	**64.3**	**74.9**	**62.3**	**66.6**	**91.8**	**96.6**	**55.5**	**86.6**	**93.8**	**81.8**
ResNet	RRF (Res) [19]	47.6	77.4	87.1	35.4	68.3	79.9	66.0	56.4	85.3	91.5	43.9	78.1	88.6	73.9
	DAN (Res) [23]	55.0	81.8	89.0	39.4	69.2	79.1	68.9	-	-	-	-	-	-	-
	VSE++ (Res) [5]	52.9	79.1	87.2	39.6	69.6	79.5	68.0	64.6	89.1	95.7	52.0	83.1	92.0	79.4
	Ours (Res)	**55.5**	**82.0**	**89.3**	**41.1**	**70.5**	**79.9**	**69.7**	**69.9**	**92.9**	**97.5**	**56.7**	**87.5**	**94.8**	**83.2**

Analysis of Image Annotation Results

Table 10. Results of image annotation by 3 ablation models.

Query	Retrieved top-4 relevant sentences		
	ctx	cnp + ctx	cnp + ctx + gen
①	1. a dinner table with various plates of food and a glass of water on the table 2. a table top with some plates of food on it 3. a table set for three with food and wine 4. a dinner table with three plates of gourmet hamburgers	1. a meal is being displayed on a table 2. a table with bowls of grains and fruit and a hand with a plate 3. a table top with some plates of food on it 4. a person holding a bowl of oats next to bowls of other condiments	1. a person holding a bowl of oats next to bowls of other condiments 2. a meal is being displayed on a table 3. a table with bowls of grains and fruit and a hand with a plate 4. a table that has some food on it
②	1. a man riding a skateboard up the side of a ramp 2. a man riding a skateboard up the side of a ramp 3. a man at a skate park with his foot on the side of the skateboard 4. a man on a skateboard performing a trick	1. its a cloudy night for a ride on the motorcycle 2. a motorcyclist surveys the sunlit road into the horizon 3. a close up of a person riding a motorcycle on a long empty road 4. a photo taken from a car looking at a skateboarder on the side of the road	1. a motorcyclist surveys the sunlit road into the horizon 2. a close up of a person riding a motorcycle on a long empty road 3. a photo taken from a car looking at a skateboarder on the side of the road 4. its a cloudy night for a ride on the motorcycle
③	1. a couple of giraffes look around the ground in the zoo 2. two giraffe standing near brick building 3. a pair of giraffes standing around in their enclosure 4. two giraffes roaming around an enclosed area on a sunny day	1. a pair of giraffes standing around in their enclosure 2. a couple of giraffes eating hay from a trough 3. two giraffes that are eating from a basket 4. two giraffes stand and eat food out of a basket	1. a couple of giraffes eating hay from a trough 2. a couple of giraffes eating out of a basket 3. two giraffes stand and eat food out of a basket 4. a couple of giraffes reach for a basket

Note: groundtruth matched sentences are marked as **red**, while some sentences sharing similar meanings as groundtruths are marked as underline.

Conclusion

- Conclusion
  - **multi-regional multi-label CNN** for semantic concept prediction
  - **gated fusion unit, and joint matching-generation learning** for semantic order learning

For more details, please refer to the following paper:

1. Yan Huang, Qi Wu, and Liang Wang**, Learning Semantic Concepts and Order for Image and Sentence Matching**. *CVPR*, 2018.

Outline

1 Image and Sentence Matching

2 Related Work

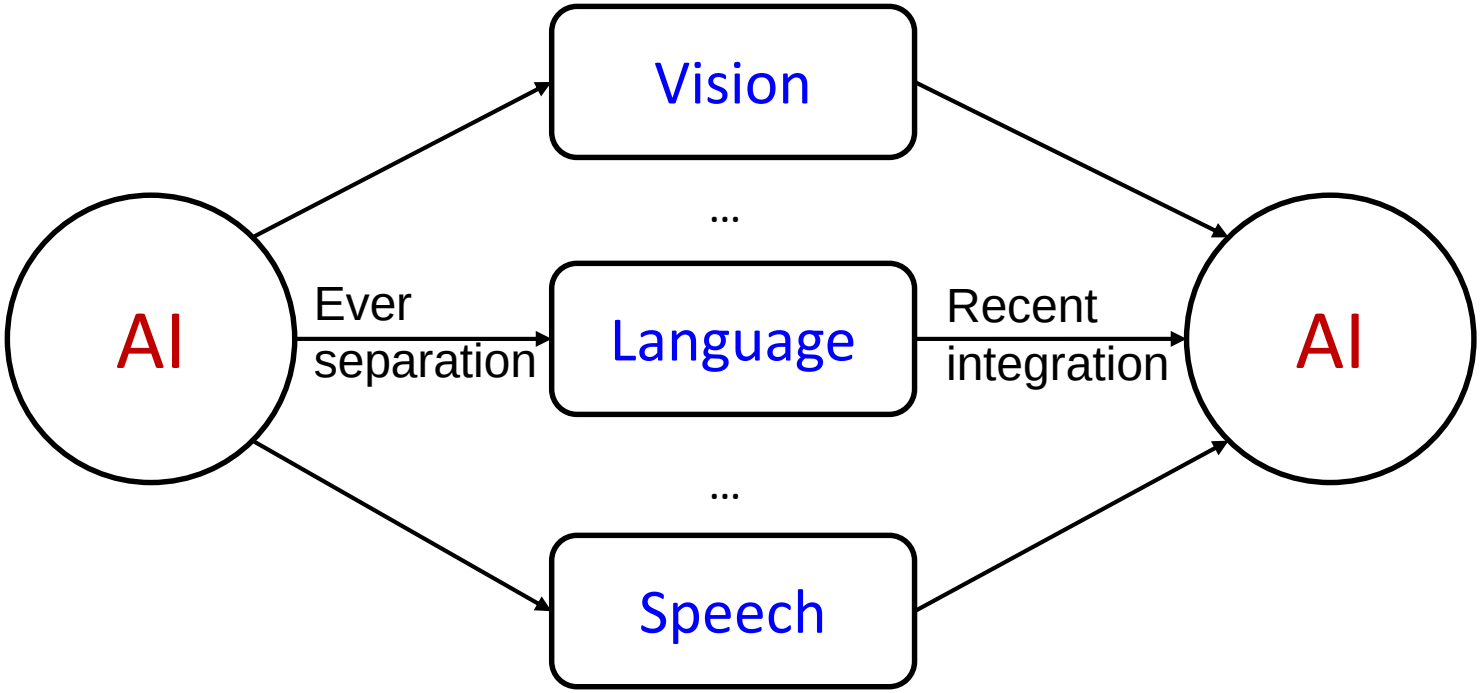
3 Improved Image and Sentence Matching

- 3.1 Context-modulated Multimodal Attention
- 3.2 Joint Semantic Concepts and Order Learning

4 Future Directions

Future Directions

- Understand vision jointly with language and speech in a unified framework



Acknowledgement

NVAIL

Artificial Intelligence Laboratory

Sponsor excellent hardware resources

中国科学院自动化研究所
INSTITUTE OF AUTOMATION
CHINESE ACADEMY OF SCIENCES

CRIPAC
智能感知与计算研究中心
Center for Research on Intelligent Perception and Computing

THANK YOU

yhuang@nlpr.ia.ac.cn