面向自然图像和文本的跨模态检索

汇报人：张一帆

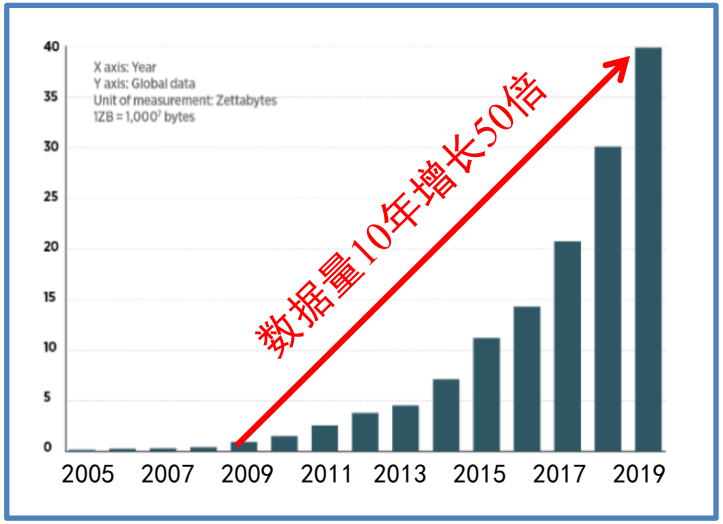
二〇一九年六月

内容提要

- 研究背景及问题定义
- 研究意义
- 研究内容
  - 内容一：
    嵌入空间表征学习法
  - 内容二：
    匹配得分学习法
- 总结

研究背景

- 数据量逐年增长



来源：联合国欧洲经济委员会，2013年

- 数据模态形式丰富

图像

视频

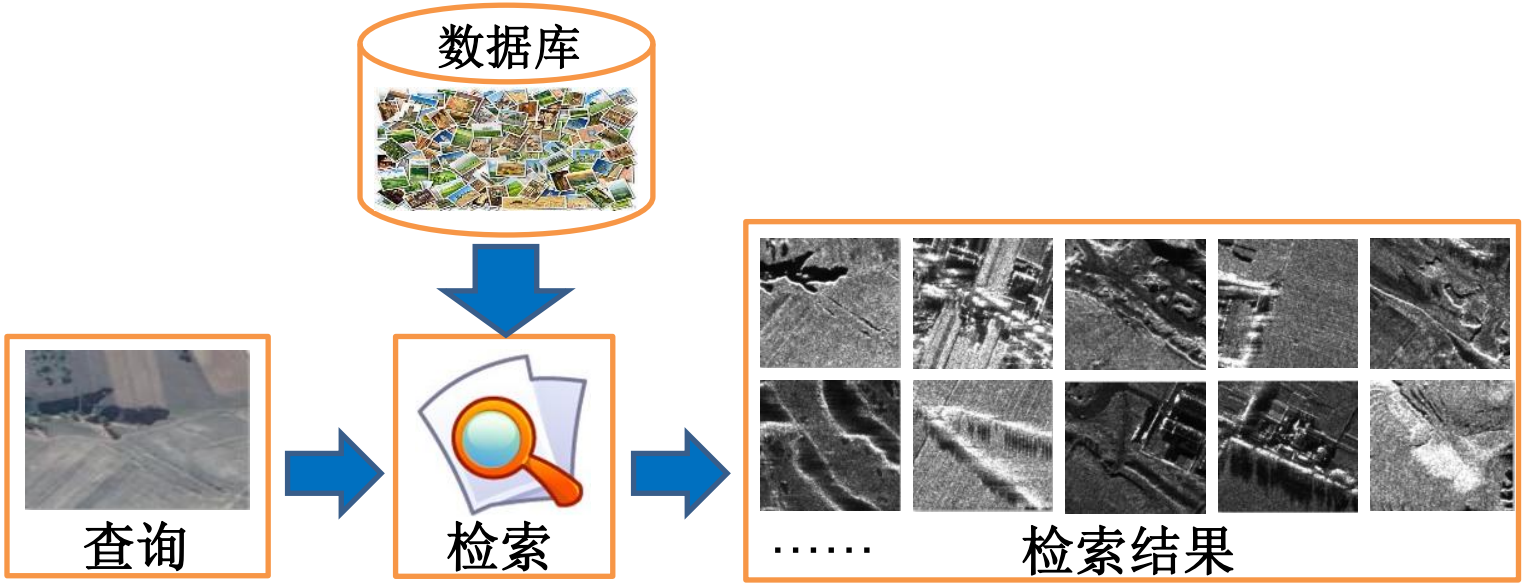
文本

音频

.....

问题定义

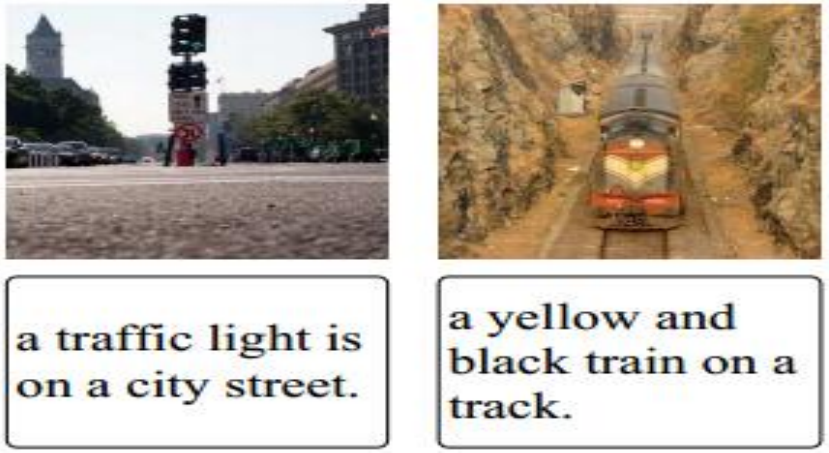
- 跨模态检索
  - 给定一种模态下的查询，从大规模数据库中快速找到相关的其它模态下的数据。
  - 相关的多模态数据：对于同一实体，通过不同方式或者角度采集到的数据。



研究意义

（1）搜索引擎管理

（2）智能视频监控



（3）辅助其它任务

（4）多模态信息交互

研究内容

- 如今，跨模态检索大致分为两类，即基于实例的跨模态检索和基于类别的跨模态检索。

- 面向基于实例的跨模态检索，讨论自然图像和文本间的跨模态检索

自然图像和文本的跨模态检索

- 问题定义：自然图像作为查询，检索有着相同或相似语义的文本句子（基于图像的文本检索），或者反之（基于文本的图像检索）。

检索

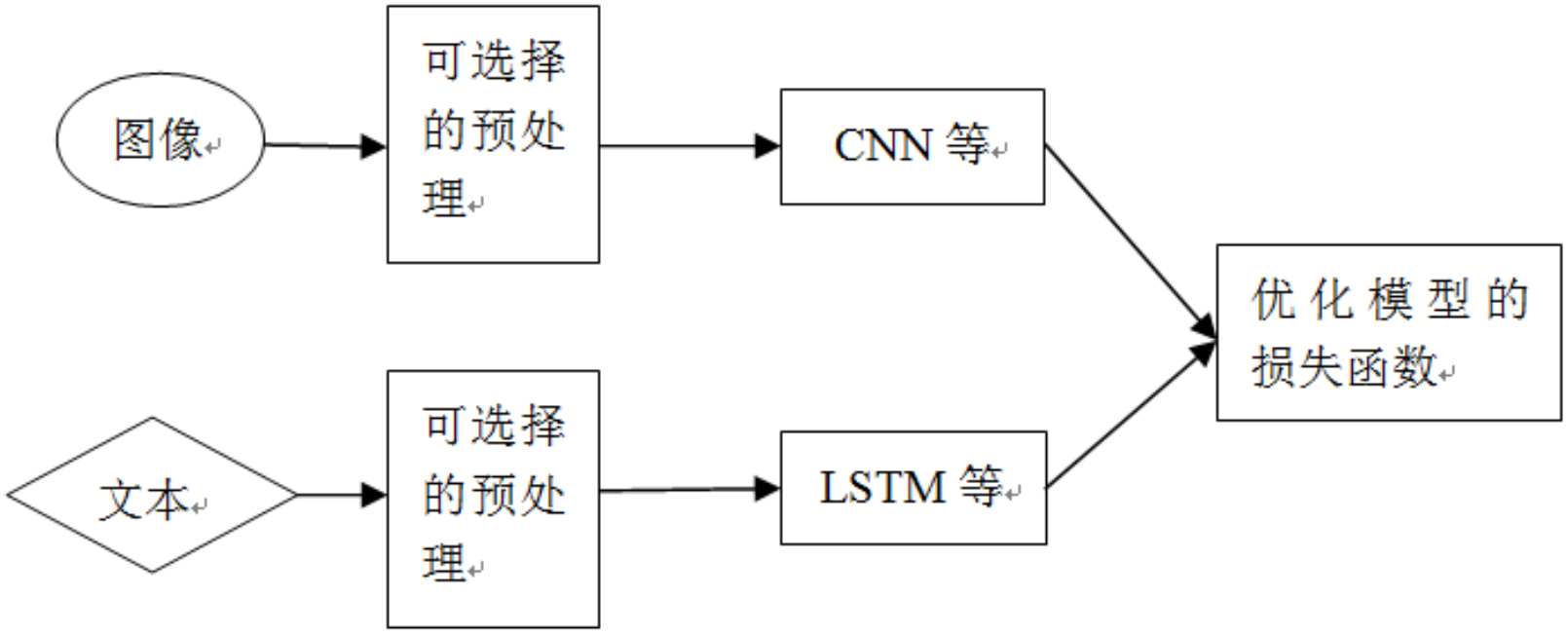
1. A black and white dog is running in a grassy garden surrounded by a white fence .
2. A boston terrier is running on lush green grass in front of a white fence .
3. A dog runs on the grass near a wooden fence .

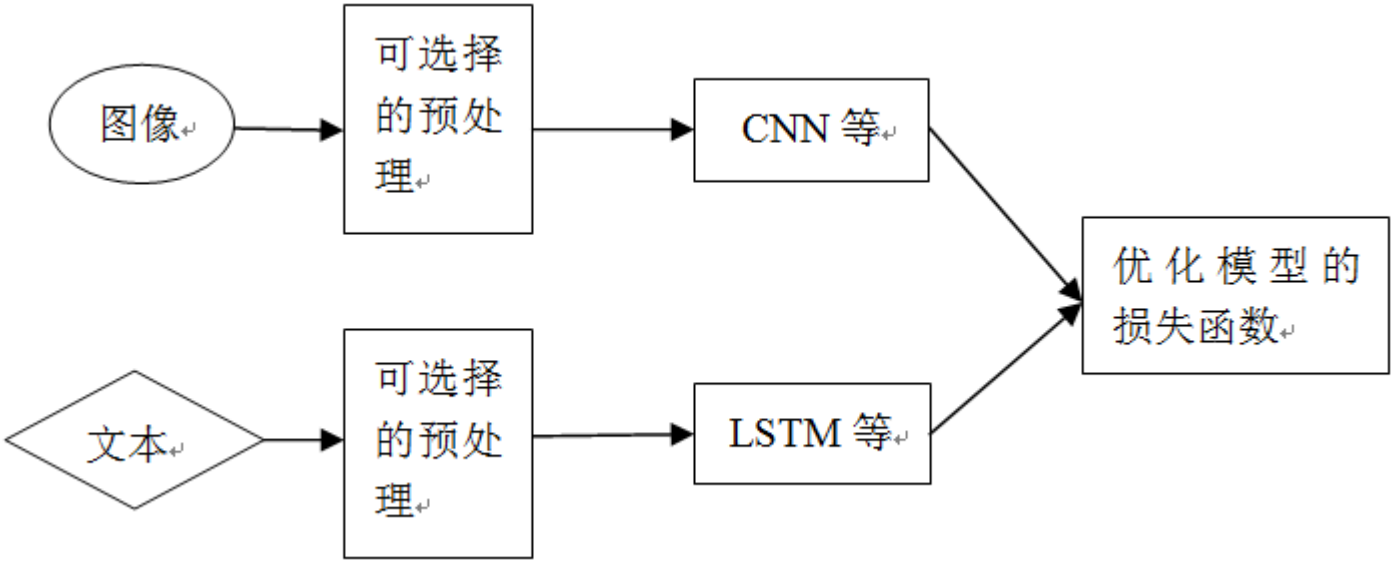
数据集

- 两个常用的数据集：MS-COCO和Flickr30K。
- 评价指标：Recall@K（R@K）和Med r
  - MS-COCO数据集
    - ✓ 123,287张图像，每个图像都有5个对应的文本句子
    - ✓ 5000张验证图像，5000张测试图像，其余图像用于训练
    - ✓ 记录在全部5000张测试图像上的结果和1000张测试图像上的5折平均结果
  - Flickr30K数据集
    - ✓ 31,784张图像， 每个图像都有5个对应的文本句子
    - ✓ 1000张验证图像，1000张测试图像，其余图像用于训练。
    - ✓ 记录在全部1000张测试图像上的结果

嵌入空间表征学习法

- 利用独立的深度模型将图像和文本映射到一个公共的特征空间，然后基于某种相似性准则完成排序

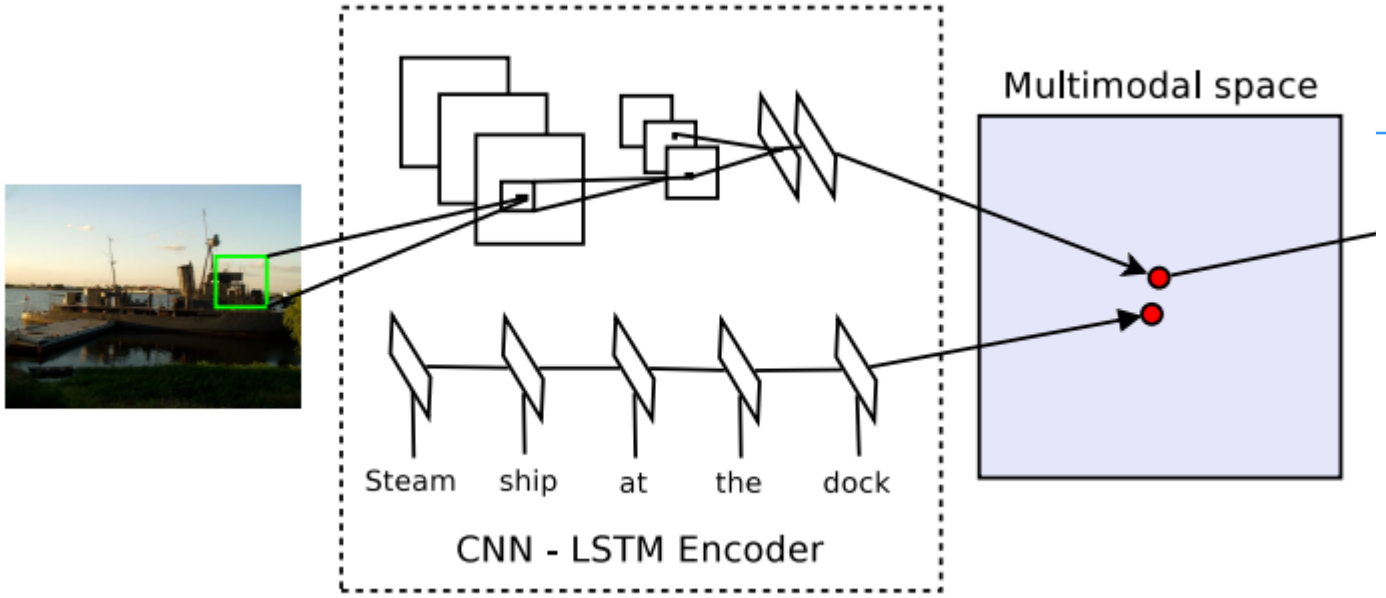




- 关于图像预处理：SIFT特征（BOW）、颜色直方图，边缘方向直方图、图像经过预训练的网络的特征等。
- 关于文本预处理：词频特征、潜在狄利克雷分布，word2vec等。

嵌入空间表征学习法

1.Unifying Visual-Semantic Embedding with Multimodal Neural Language Models
2. Learning Deep Structure-Preserving Image-Text Embeddings
3. Dual-Path Convolutional Image-Text Embeddings with Instance Loss
4. Dual Attention Networks for Multimodal Reasoning and Matching
5. VSE++: Improving Visual-Semantic Embeddings with Hard Negatives
6. Learning a Recurrent Residual Fusion Network for Multimodal Matching
7. Look, Imagine and Match: Improving Textual-Visual Cross-Modal Retrieval with Generative Models
8. Bidirectional Retrieval Made Simple
9. Learning Semantic Concepts and Order for Image and Sentence Matching



$$\min_{\boldsymbol{\theta}} \sum_{\mathbf{x}} \sum_{k} \max\{0, \alpha - s(\mathbf{x}, \mathbf{v}) + s(\mathbf{x}, \mathbf{v}_k)\} + \sum_{\mathbf{v}} \sum_{k} \max\{0, \alpha - s(\mathbf{v}, \mathbf{x}) + s(\mathbf{v}, \mathbf{x}_k)\}$$

Kiros R, Salakhutdinov R, Zemel R S. Unifying visual-semantic embeddings with multimodal neural language models[J]. arXiv preprint arXiv:1411.2539, 2014.

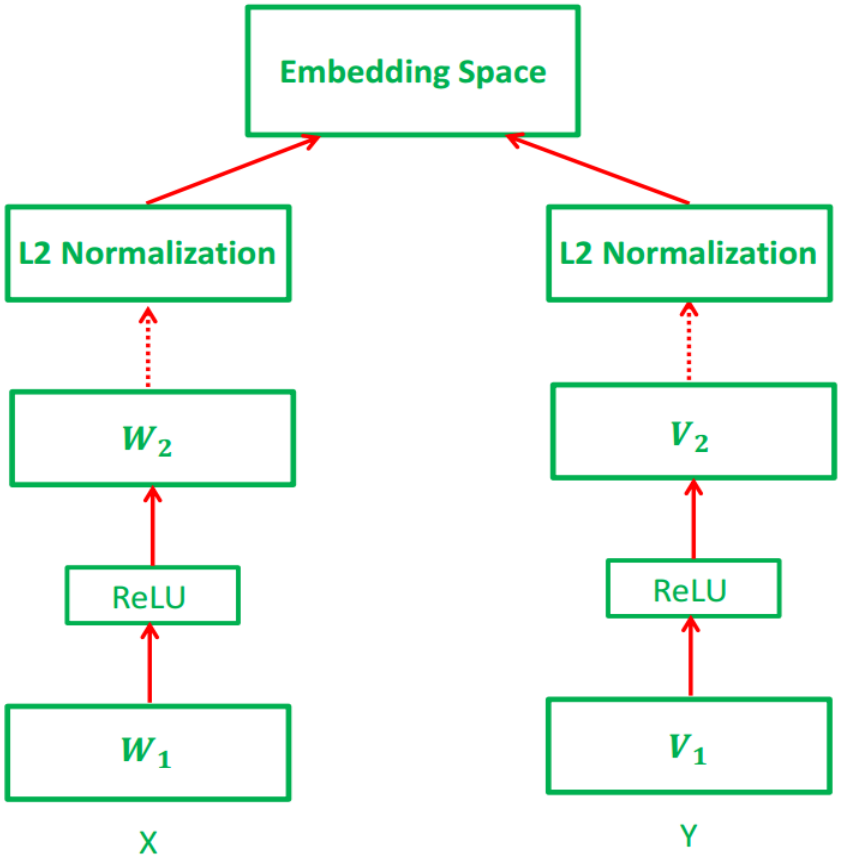
Flickr8K								
	Image Annotation				Image Search			
Model	**R@1**	**R@5**	**R@10**	**Med *r***	**R@1**	**R@5**	**R@10**	**Med *r***
Random Ranking	0.1	0.6	1.1	631	0.1	0.5	1.0	500
SDT-RNN [6]	4.5	18.0	28.6	32	6.1	18.5	29.0	29
† DeViSE [5]	4.8	16.5	27.3	28	5.9	20.1	29.6	29
† SDT-RNN [6]	6.0	22.7	34.0	23	6.6	21.6	31.7	25
DeFrag [15]	5.9	19.2	27.3	34	5.2	17.6	26.5	32
† DeFrag [15]	12.6	32.9	44.0	14	9.7	29.6	42.5	15
m-RNN [7]	14.5	37.2	48.5	11	11.5	31.0	42.4	15
Our model	13.5	36.2	45.7	13	10.4	31.0	43.7	14
Our model (OxfordNet)	**18.0**	**40.9**	**55.0**	**8**	**12.5**	**37.0**	**51.5**	**10**

Table 1: Flickr8K experiments. **R@K** is Recall@K (high is good). **Med *r*** is the median rank (low is good). Best results overall are **bold** while best results without OxfordNet features are underlined. A † infront of the method indicates that object detections were used along with single frame features.

Flickr30K								
	Image Annotation				Image Search			
Model	**R@1**	**R@5**	**R@10**	**Med *r***	**R@1**	**R@5**	**R@10**	**Med *r***
Random Ranking	0.1	0.6	1.1	631	0.1	0.5	1.0	500
† DeViSE [5]	4.5	18.1	29.2	26	6.7	21.9	32.7	25
† SDT-RNN [6]	9.6	29.8	41.1	16	8.9	29.8	41.1	16
† DeFrag [15]	14.2	37.7	51.3	10	10.2	30.8	44.2	14
† DeFrag + Finetune CNN [15]	16.4	40.2	54.7	8	10.3	31.4	44.5	13
m-RNN [7]	18.4	40.2	50.9	10	12.6	31.2	41.5	16
Our model	14.8	39.2	50.9	10	11.8	34.0	46.3	13
Our model (OxfordNet)	**23.0**	**50.7**	**62.9**	**5**	**16.8**	**42.0**	**56.5**	**8**

Table 2: Flickr30K experiments. **R@K** is Recall@K (high is good). **Med *r*** is the median rank (low is good). Best results overall are **bold** while best results without OxfordNet features are underlined. A † infront of the method indicates that object detections were used along with single frame features.

- 图像特征:VGG19 文本特征：Fisher vector



Wang L, Li Y, Lazebnik S. Learning deep structure-preserving image-text embeddings[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 5005-5013.

Loss函数

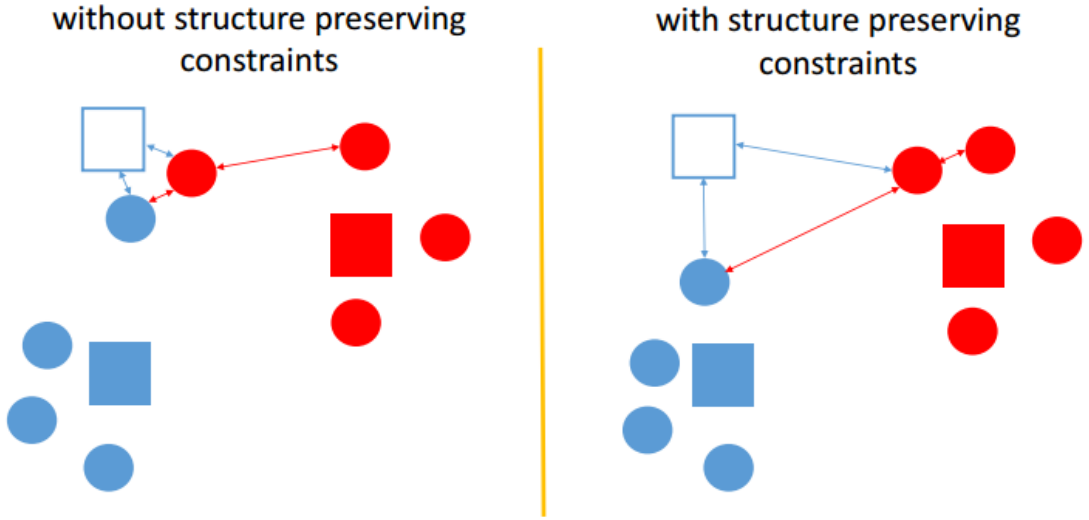


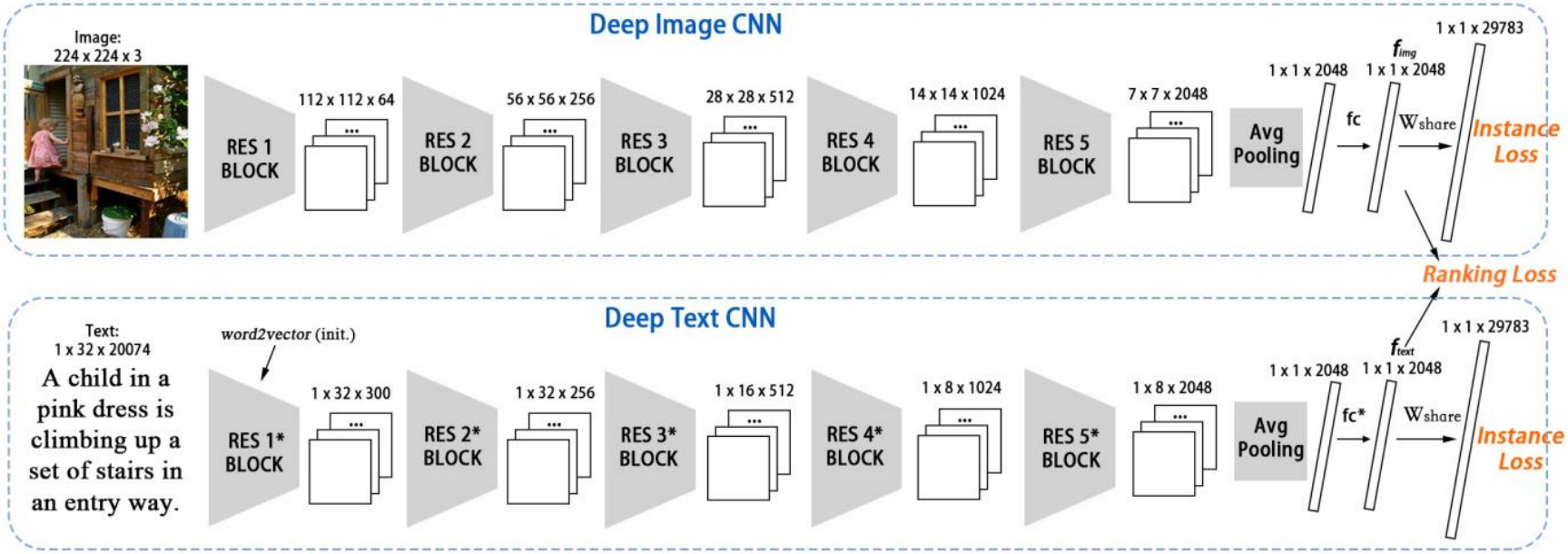
Figure 2. Illustration of the proposed structure-preserving constraints for joint embedding learning (see text). Rectangles represent images and circles represent sentences. Same color indicates matching images and sentences.

- Structure-preserving constraints

$$d(x_i, x_j) + m < d(x_i, x_k) \quad \forall x_j \in N(x_i), \forall x_k \notin N(x_i)$$

	Methods on Flickr30K	Image-to-sentence			Sentence-to-image		
		R@1	R@5	R@10	R@1	R@5	R@10
(a) State of the art	Deep CCA [33]	27.9	56.9	68.2	26.8	52.9	66.9
	mCNN(ensemble) [29]	33.6	64.1	74.9	26.2	56.3	69.6
	m-RNN-vgg [31]	35.4	63.8	73.7	22.8	50.7	63.1
	Mean vector [26]	24.8	52.5	64.3	20.5	46.3	59.3
	CCA (FV HGLMM) [26]	34.4	61.0	72.3	24.4	52.1	65.6
	CCA (FV GMM+HGLMM) [26]	35.0	62.0	73.8	25.0	52.7	66.0
	CCA (FV HGLMM) [37]	36.5	62.2	73.3	24.7	53.4	66.8
(b) Fisher vector	Linear + one-directional	33.5	61.7	73.6	21.0	47.4	60.5
	Linear + bi-directional	34.6	64.3	74.9	24.2	52.0	64.2
	Linear + bi-directional + structure	35.2	66.8	76.2	25.6	54.8	66.5
	Nonlinear + one-directional	37.5	65.6	76.9	22.4	50.9	63.3
	Nonlinear + bi-directional	39.3	68.0	78.3	28.1	59.2	71.2
	Nonlinear + bi-directional + structure	**40.3**	**68.9**	**79.9**	**29.7**	**60.1**	**72.1**
(c) Mean vector	Nonlinear + bi-directional	33.5	60.2	71.9	22.8	52.5	65.0
	Nonlinear + bi-directional + structure	35.7	62.9	74.4	25.1	53.9	66.5
(d) tf-idf	Nonlinear + bi-directional	38.7	66.6	76.9	27.6	57.0	69.0
	Nonlinear + bi-directional + structure	40.1	67.6	78.2	28.1	58.5	69.8

Table 1. Bidirectional retrieval results. The numbers in (a) come from published papers, and the numbers in (b-d) are results of our approach using different textual features. Note that the Deep CCA results in [33] were obtained with AlexNet [27]. The results of our method with AlexNet are still about 3% higher than those of [33] for image-to-sentence retrieval and 1% higher for sentence-to-image retrieval.



In the first training stage, we fixed the pre-trained image CNN, and train the text CNN only. The learning rate is 0.001. We stop training when instance loss converges. In the second stage, we combine the ranking loss as Eq. 9 (the margin $\alpha = 1$) and fine-tune the entire network.

Zheng Z, Zheng L, Garrett M, et al. Dual-Path Convolutional Image-Text Embedding with Instance Loss[J]. arXiv preprint arXiv:1711.05535, 2017.TCSVT

- 在Flickr30k, MSCOCO, CUHL-PEDES上进行了评测

Method	Visual	Textual	Image Query				Text Query			
			R@1	R@5	R@10	Med	R@1	R@5	R@10	Med r
DeVise [5]	ft AlexNet	ft skip-gram	4.5	18.1	29.2	26	6.7	21.9	32.7	25
Deep Fragment [6]	ft RCNN	fixed word vector from [58]	16.4	40.2	54.7	8	10.3	31.4	44.5	13
DCCA [59]	ft AlexNet	TF-IDF	16.7	39.3	52.9	8	12.6	31.0	43.0	15
DVSA [32]	ft RCNN (init. on Detection)	w2v + ft RNN	22.2	48.2	61.4	4.8	15.2	37.7	50.5	9.2
LRCN [60]	ft VGG-16	ft RNN	23.6	46.6	58.3	7	17.5	40.3	50.8	9
m-CNN [7]	ft VGG-19	4 × ft CNN	33.6	64.1	74.9	3	26.2	56.3	69.6	4
VQA-A [18]	fixed VGG-19	ft RNN	33.9	62.5	74.5	-	24.9	52.6	64.8	-
GMM-FV [17]	fixed VGG-16	w2v + GMM + HGLMM	35.0	62.0	73.8	3	25.0	52.7	66.0	5
m-RNN [16]	fixed VGG-16	ft RNN	35.4	63.8	73.7	3	22.8	50.7	63.1	5
RNN-FV [19]	fixed VGG-19	feature from [17]	35.6	62.5	74.2	3	27.4	55.9	70.0	4
HM-LSTM [21]	fixed RCNN from [32]	w2v + ft RNN	38.1	-	76.5	3	27.7	-	68.8	4
SPE [8]	fixed VGG-19	w2v + HGLMM	40.3	68.9	79.9	-	29.7	60.1	72.1	-
sm-LSTM [20]	fixed VGG-19	ft RNN	42.5	71.9	81.5	2	30.2	60.4	72.3	3
RRF-Net [61]	fixed ResNet-152	w2v + HGLMM	47.6	77.4	87.1	-	35.4	68.3	79.9	-
2WayNet [49]	fixed VGG-16	feature from [17]	49.8	67.5	-	-	36.0	55.6	-	-
DAN (VGG-19) [9]	fixed VGG-19	ft RNN	41.4	73.5	82.5	2	31.8	61.7	72.5	3
DAN (ResNet-152) [9]	fixed ResNet-152	ft RNN	55.0	81.8	89.0	1	**39.4**	69.2	79.1	2
Ours (VGG-19) Stage I	fixed VGG-19	ft ResNet-50† (w2v init.)	37.5	66.0	75.6	3	27.2	55.4	67.6	4
Ours (VGG-19) Stage II	ft VGG-19	ft ResNet-50† (w2v init.)	47.6	77.3	87.1	2	35.3	66.6	78.2	3
Ours (ResNet-50) Stage I	fixed ResNet-50	ft ResNet-50† (w2v init.)	41.2	69.7	78.9	2	28.6	56.2	67.8	4
Ours (ResNet-50) Stage II	ft ResNet-50	ft ResNet-50† (w2v init.)	53.9	80.9	**89.9**	**1**	39.2	**69.8**	80.8	**2**
Ours (ResNet-152) Stage I	fixed ResNet-152	ft ResNet-152† (w2v init.)	44.2	70.2	79.7	2	30.7	59.2	70.8	4
Ours (ResNet-152) Stage II	ft ResNet-152	ft ResNet-152† (w2v init.)	**55.6**	**81.9**	89.5	**1**	39.1	69.2	**80.9**	**2**

训练阶段loss

Method	Stage	Image Query		Text Query	
		R@1	R@10	R@1	R@10
Only Ranking Loss	I	6.1	27.3	4.9	27.8
Only Instance Loss	I	39.9	79.1	28.2	67.9
Only Instance Loss	II	50.5	86.0	34.9	75.7
Only Ranking Loss	II	47.5	85.4	29.0	68.7
Full model	II	55.4	89.3	39.7	80.8

类别数目

Methods	Image-Query R@1	Text-Query R@1
3000 categories (StageI)	38.0	26.1
10000 categories (StageI)	44.7	31.3
Our (StageI)	52.2	37.2

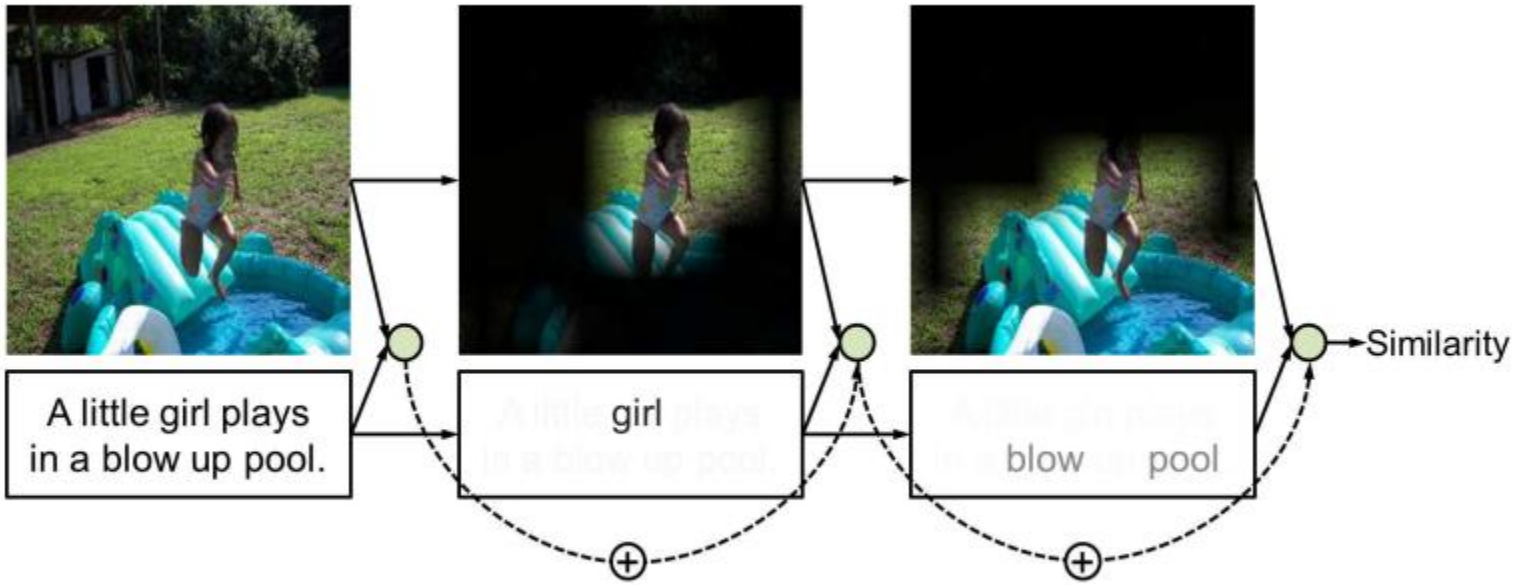
- Position shift

Method	Image Query		Text Query	
	R@1	R@10	R@1	R@10
Left alignment	34.1	73.1	23.6	61.4
Position shift	39.9	79.1	28.2	67.9

- Word2vec初始化

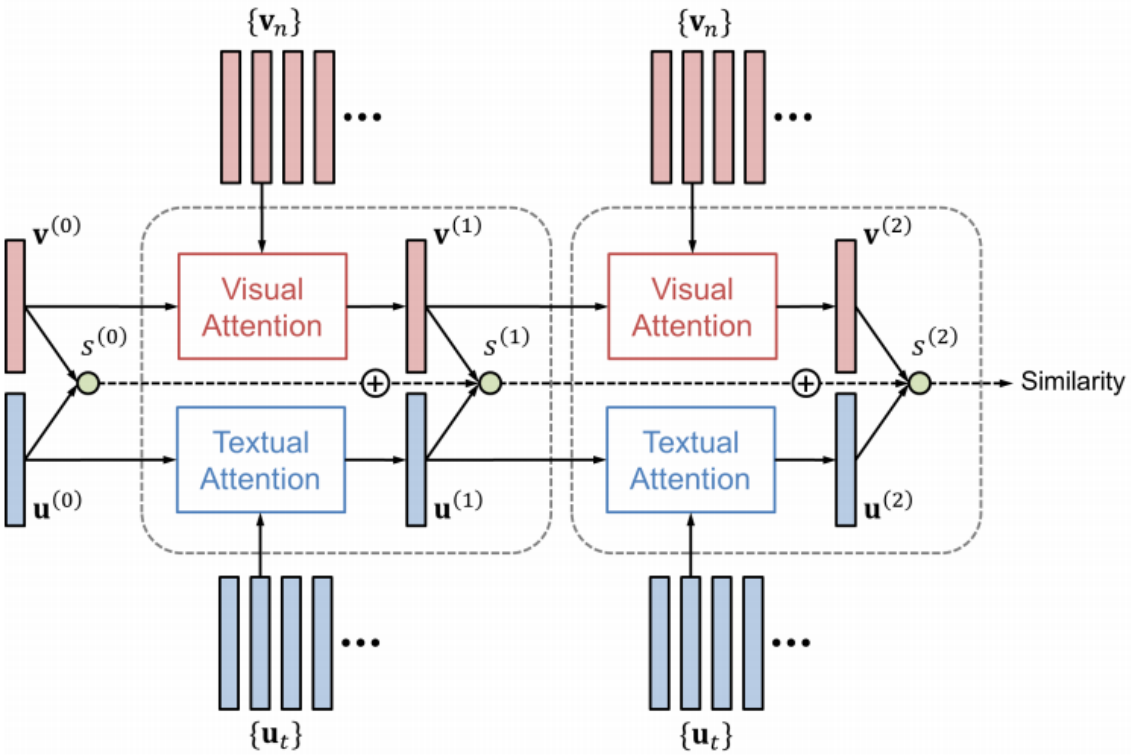
Method	Image Query		Text Query	
	R@1	R@10	R@1	R@10
Random initialization [52]	38.0	78.7	26.6	66.6
Word2vec initialization	39.9	79.1	28.2	67.9

Method	Visual	Textual	Image Query				Text Query			
			R@1	R@5	R@10	Med	R@1	R@5	R@10	Med r
1K test images										
DVSA [32]	ft RCNN	w2v + ft RNN	38.4	69.9	80.5	1	27.4	60.2	74.8	3
GMM-FV [17]	fixed VGG-16	w2v + GMM + HGLMM	39.4	67.9	80.9	2	25.1	59.8	76.6	4
m-RNN [16]	fixed VGG-16	ft RNN	41.0	73.0	83.5	2	29.0	42.2	77.0	3
RNN-FV [19]	fixed VGG-19	feature from [17]	41.5	72.0	82.9	2	29.2	64.7	80.4	3
m-CNN [7]	ft VGG-19	4 $\times$ ft CNN	42.8	73.1	84.1	2	32.6	68.6	82.8	3
HM-LSTM [21]	fixed CNN from [32]	ft RNN	43.9	-	87.8	2	36.1	-	86.7	3
SPE [8]	fixed VGG-19	w2v + HGLMM	50.1	79.7	89.2	-	39.6	75.2	86.9	-
VQA-A [18]	fixed VGG-19	ft RNN	50.5	80.1	89.7	-	37.0	70.9	82.9	-
sm-LSTM [20]	fixed VGG-19	ft RNN	53.2	83.1	91.5	1	40.7	75.8	87.4	2
2WayNet [49]	fixed VGG-16	feature from [17]	55.8	75.2	-	-	39.7	63.3	-	-
RRF-Net [61]	fixed ResNet-152	w2v + HGLMM	56.4	85.3	91.5	-	43.9	78.1	88.6	-
Ours (VGG-19) Stage I	fixed VGG-19	ft ResNet-50† (w2v init.)	46.0	75.6	85.3	2	34.4	66.6	78.7	3
Ours (VGG-19) Stage II	ft VGG-19	ft ResNet-50† (w2v init.)	59.4	86.2	92.9	1	41.6	76.3	87.5	2
Ours (ResNet-50) Stage I	fixed ResNet-50	ft ResNet-50† (w2v init.)	52.2	80.4	88.7	1	37.2	69.5	80.6	2
Ours (ResNet-50) Stage II	ft ResNet-50	ft ResNet-50† (w2v init.)	**65.6**	**89.8**	**95.5**	**1**	**47.1**	**79.9**	**90.0**	**2**
5K test images										
GMM-FV [17]	fixed VGG-16	w2v + GMM + HGLMM	17.3	39.0	50.2	10	10.8	28.3	40.1	17
DVSA [32]	ft RCNN	w2v + ft RNN	16.5	39.2	52.0	9	10.7	29.6	42.2	14
VQA-A [18]	fixed VGG-19	ft RNN	23.5	50.7	63.6	-	16.7	40.5	53.8	-
Ours (VGG-19) Stage I	fixed VGG-19	ft ResNet-50† (w2v init.)	24.5	50.1	62.1	5	16.5	39.1	51.8	10
Ours (VGG-19) Stage II	ft VGG-19	ft ResNet-50† (w2v init.)	35.5	63.2	75.6	3	21.0	47.5	60.9	6
Ours (ResNet-50) Stage I	fixed ResNet-50	ft ResNet-50† (w2v init.)	28.6	56.2	68.0	4	18.7	42.4	55.1	8
Ours (ResNet-50) Stage II	ft ResNet-50	ft ResNet-50† (w2v init.)	**41.2**	**70.5**	**81.1**	**2**	**25.3**	**53.4**	**66.4**	**5**



(b) DAN for multimodal matching. (m-DAN)

Nam H, Ha J W, Kim J. Dual attention networks for multimodal reasoning and matching[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 299-307.



$$\mathbf{h}_{\mathbf{v},n}^{(k)} = \tanh\left(\mathbf{W}_{\mathbf{v}}^{(k)}\,\mathbf{v}_n\right) \odot \tanh\left(\mathbf{W}_{\mathbf{v},\mathbf{m}}^{(k)}\,\mathbf{m}_{\mathbf{v}}^{(k-1)}\right)$$

$$\alpha_{\mathbf{v},n}^{(k)} = \mathrm{softmax}\left(\mathbf{W}_{\mathbf{v},\mathbf{h}}^{(k)}\,\mathbf{h}_{\mathbf{v},n}^{(k)}\right),$$

$$\mathbf{v}^{(k)} = \tanh\left(\mathbf{P}^{(k)} \sum_{n=1}^{N} \alpha_{\mathbf{v},n}^{(k)}\,\mathbf{v}_n\right),$$

$$\mathbf{v}^{(0)} = \tanh\left(\mathbf{P}^{(0)}\frac{1}{N}\sum_{n}\mathbf{v}_n\right)$$

$$\mathbf{u}^{(0)} = \frac{1}{T}\sum_{t}\mathbf{u}_t.$$

$$\mathbf{m}_{\mathbf{v}}^{(k)} = \mathbf{m}_{\mathbf{v}}^{(k-1)} + \mathbf{v}^{(k)}$$

$$\mathbf{m}_{\mathbf{u}}^{(k)} = \mathbf{m}_{\mathbf{u}}^{(k-1)} + \mathbf{u}^{(k)}$$

Table 2: Bidirectional retrieval results on the Flickr30K dataset compared with state-of-the-art methods.

Method	Image-to-Text				Text-to-Image			
	R@1	R@5	R@10	MR	R@1	R@5	R@10	MR
DCCA [34]	27.9	56.9	68.2	4	26.8	52.9	66.9	4
mCNN [19]	33.6	64.1	74.9	3	26.2	56.3	69.6	4
m-RNN-VGG [20]	35.4	63.8	73.7	3	22.8	50.7	63.1	5
GMM+HGLMM FV [14]	35.0	62.0	73.8	3	25.0	52.7	66.0	5
HGLMM FV [24]	36.5	62.2	73.3	-	24.7	53.4	66.8	-
SPE [30]	40.3	68.9	79.9	-	29.7	60.1	72.1	-
DAN (VGG)	41.4	73.5	82.5	2	31.8	61.7	72.5	3
DAN (ResNet)	**55.0**	**81.8**	**89.0**	**1**	**39.4**	**69.2**	**79.1**	**2**

(+) A woman in a brown vest is working on the computer.

(+) A man in a white shirt stands high up on scaffolding.

(+) A woman in a red vest working at a computer.

(+) Man works on top of scaffolding.

- Ranking loss

$$\ell_{SH}(i,c) = \sum_{\hat{c}}[\alpha - s(i,c) + s(i,\hat{c})]_+ + \sum_{\hat{i}}[\alpha - s(i,c) + s(\hat{i},c)]_+,$$

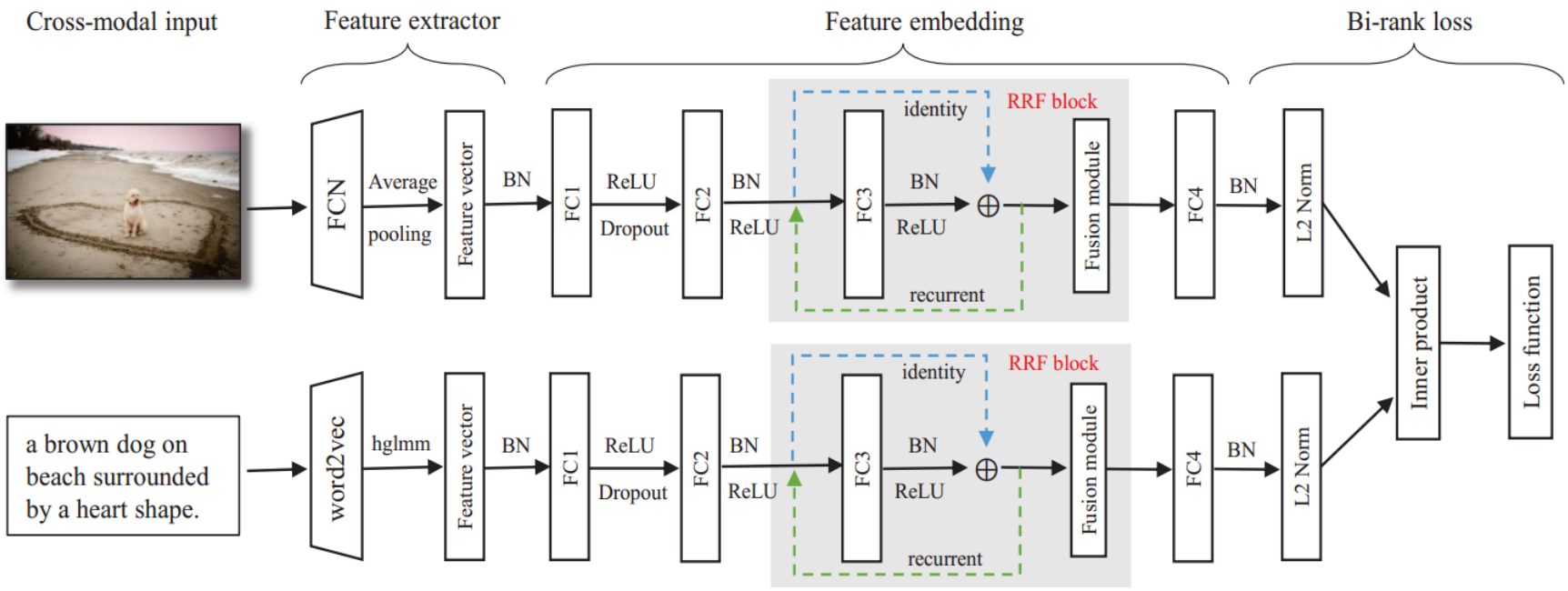
- Ranking loss with hard

$$\ell_{MH}(i,c) = \max_{c'}\left[\alpha + s(i,c') - s(i,c)\right]_+ + \max_{i'}\left[\alpha + s(i',c) - s(i,c)\right]_+$$

Faghri F, Fleet D J, Kiros J R, et al. VSE++: Improving Visual-Semantic Embeddings with Hard Negatives[J]. 2017.BMVC

#	Model	Trainset	Caption Retrieval				Image Retrieval			
			R@1	**R@5**	**R@10**	**Med r**	**R@1**	**R@5**	**R@10**	**Med r**
2.1	*VSE0*	*1C* (1 fold)	43.2	73.9	85.0	2.0	33.0	67.4	80.7	3.0
1.6	*VSE++*	*1C* (1 fold)	43.6	74.8	84.6	2.0	33.7	68.8	81.0	3.0
2.2	*VSE0*	*RC*	43.1	77.0	87.1	2.0	32.5	68.3	82.1	3.0
1.7	*VSE++*	*RC*	49.0	79.8	88.4	1.8	37.1	72.2	83.8	2.0
2.3	*VSE0*	*RC+rV*	46.8	78.8	89.0	1.8	34.2	70.4	83.6	2.6
1.8	*VSE++*	*RC+rV*	51.9	81.5	90.4	**1.0**	39.5	74.1	85.6	2.0
2.4	*VSE0 (FT)*	*RC+rV*	50.1	81.5	90.5	1.6	39.7	75.4	87.2	2.0
1.9	*VSE++ (FT)*	*RC+rV*	57.2	86.0	93.3	**1.0**	45.9	79.4	89.1	2.0
2.5	*VSE0 (ResNet)*	*RC+rV*	52.7	83.0	91.8	1.0	36.0	72.6	85.5	2.2
1.10	*VSE++ (ResNet)*	*RC+rV*	58.3	86.1	93.3	**1.0**	43.6	77.6	87.8	2.0
2.6	*VSE0 (ResNet, FT)*	*RC+rV*	56.0	85.8	93.5	1.0	43.7	79.4	89.7	2.0
1.11	*VSE++ (ResNet, FT)*	*RC+rV*	**64.6**	**90.0**	**95.7**	**1.0**	**52.0**	**84.3**	**92.0**	**1.0**

Table 2: The effect of data augmentation and fine-tuning. We copy the relevant results for *VSE++* from Table 1 to enable an easier comparison. Notice that after applying all the modifications, *VSE0* model reaches 56.0% for *R*@1, while *VSE++* achieves 64.6%.



Liu Y, Guo Y, Bakker E M, et al. Learning a recurrent residual fusion network for multimodal matching[C]//Proceedings of the IEEE International Conference on Computer Vision. 2017

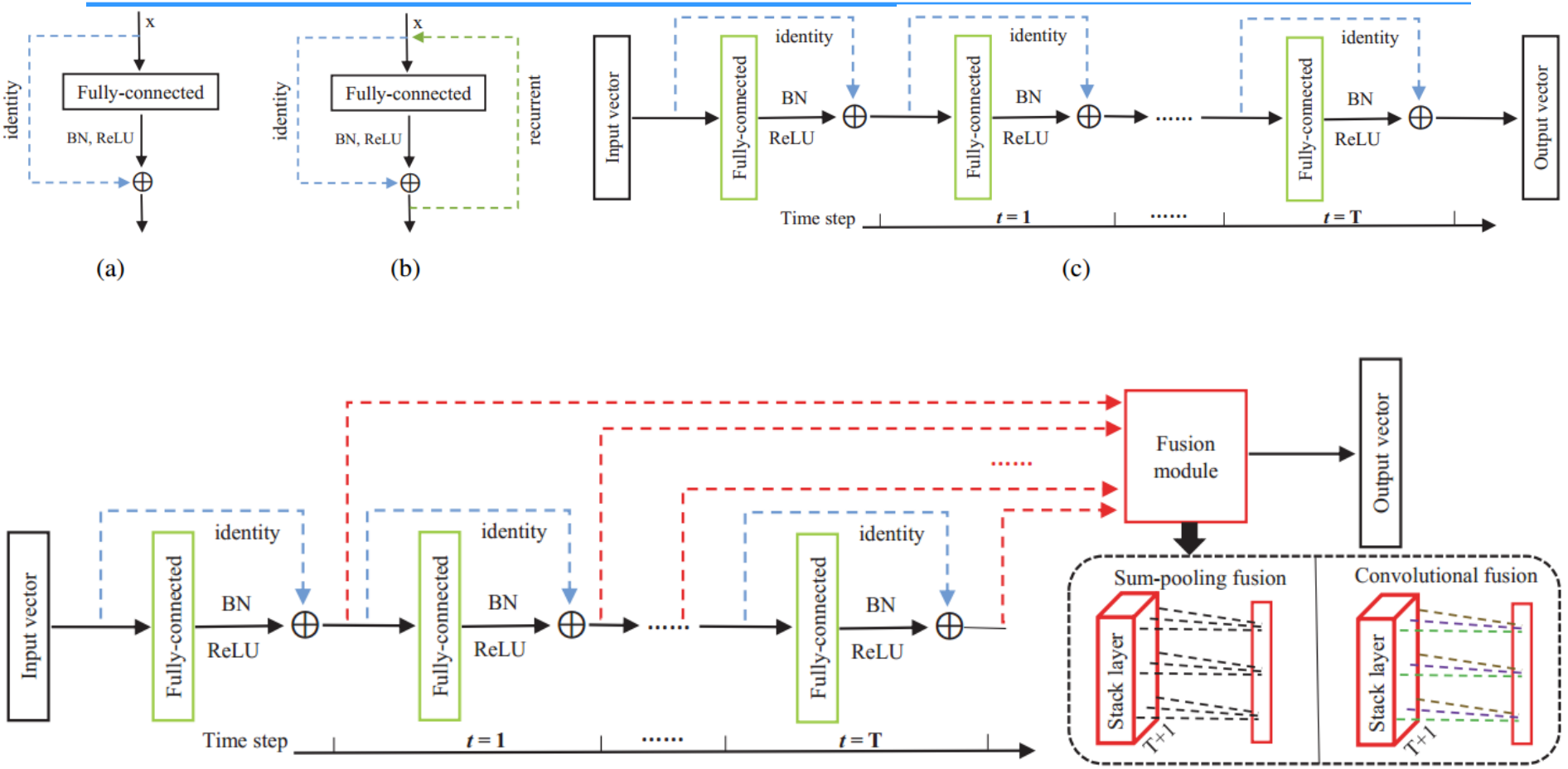

x
Fully-connected
BN, ReLU
identity
(a)
x
Fully-connected
BN, ReLU
identity
recurrent
(b)
Input vector
Fully-connected
identity
BN
ReLU
Output vector
Time step
t = 1
t = T
(c)
Fusion module
Sum-pooling fusion
Convolutional fusion
Stack layer
T+1

Table 1: Evaluation for the RRF-Net on the Flickr30K test set. Higher R@K is better. All of the four RRF-Net models outperform the baseline. When $T = 3$, it obtains better performance (in bold).

	Image to Text		Text to Image	
Method	R@1	R@5	R@1	R@5
Baseline	45.0	75.5	33.6	66.5
RRF-Net, T=1	46.4	76.1	34.3	67.3
RRF-Net, T=2	46.9	76.8	34.8	67.7
RRF-Net, T=3	**47.6**	**77.4**	**35.4**	**68.3**
RRF-Net, T=4	46.2	76.6	35.1	67.6

Table 2: Evaluation for fusion modules on the Flickr30K test set. The convolutional fusion shows better results by learning adaptive weights.

	Image to Text		Text to Image	
Method	R@1	R@5	R@1	R@5
RRF-Net w/o fusion module	45.8	75.9	34.2	67.1
RRF-Net with sum fusion	47.1	76.8	35.0	67.6
RRF-Net with conv fusion	47.6	77.4	35.4	68.3

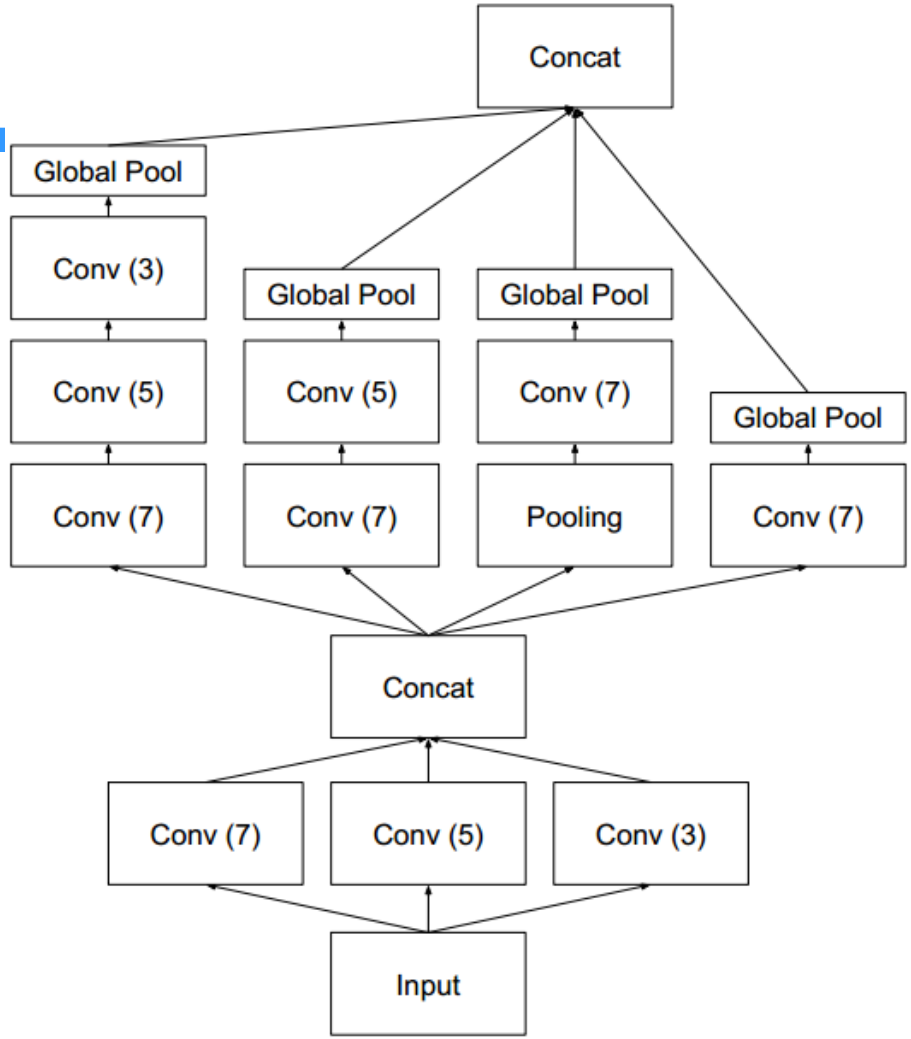


Figure 1. CHAIN-VSE-v1 module.

Wehrmann J, Barros R C. Bidirectional Retrieval Made Simple[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 7718-7726.

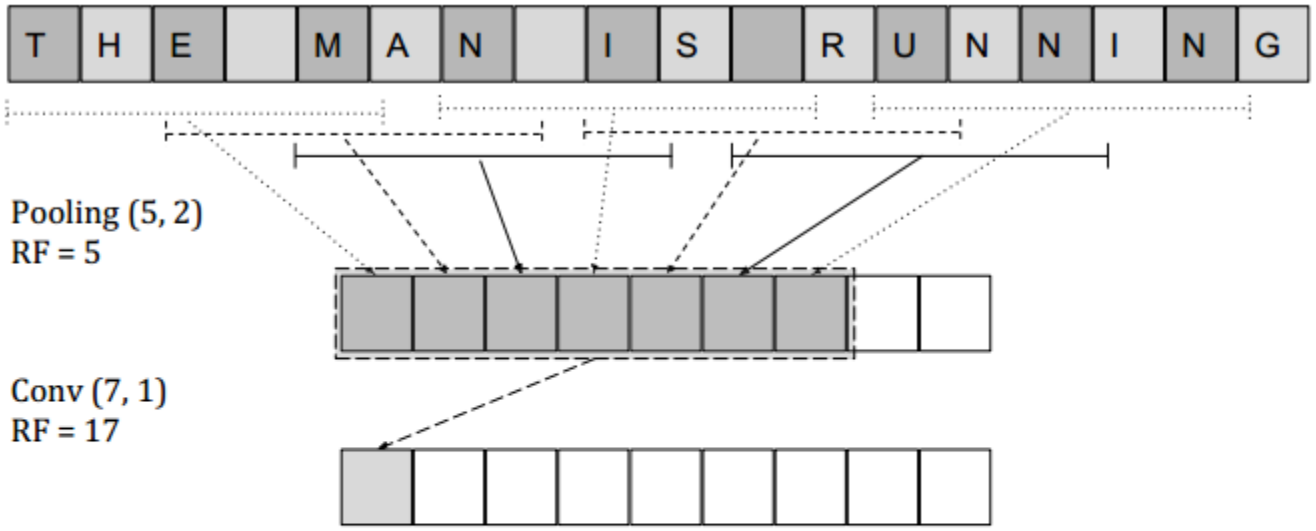


Figure 2. Impact of the pooling layer (filter size of 5 and stride of 2) in the receptive field of a character-based textual representation. For simplicity, we are ignoring the existence of the first inception module.

loss

$$s(x, y) = -|max\{0, y - x\}|^2$$

$$\mathcal{L} = \sum_{\mathbf{m}} \sum_{k} max\{0, \alpha - s(\mathbf{m}, \mathbf{v}) + s(\mathbf{m}, \mathbf{v}_k)\}$$
$$+ \sum_{\mathbf{v}} \sum_{k} max\{0, \alpha - s(\mathbf{v}, \mathbf{m}) + s(\mathbf{v}, \mathbf{m}_k)\}$$

- gooood

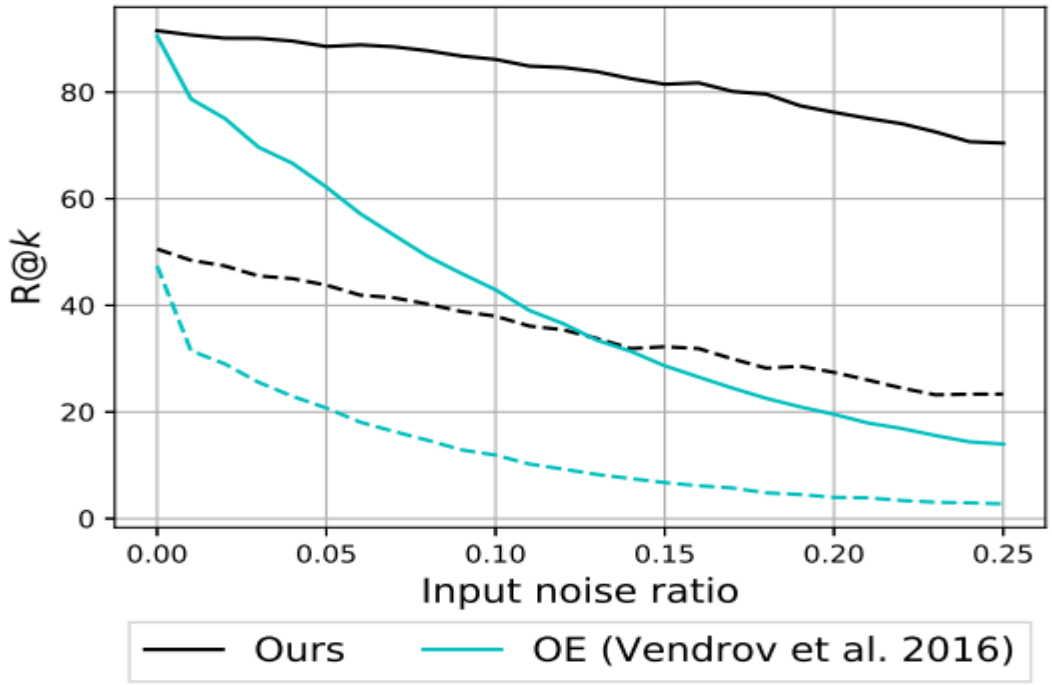
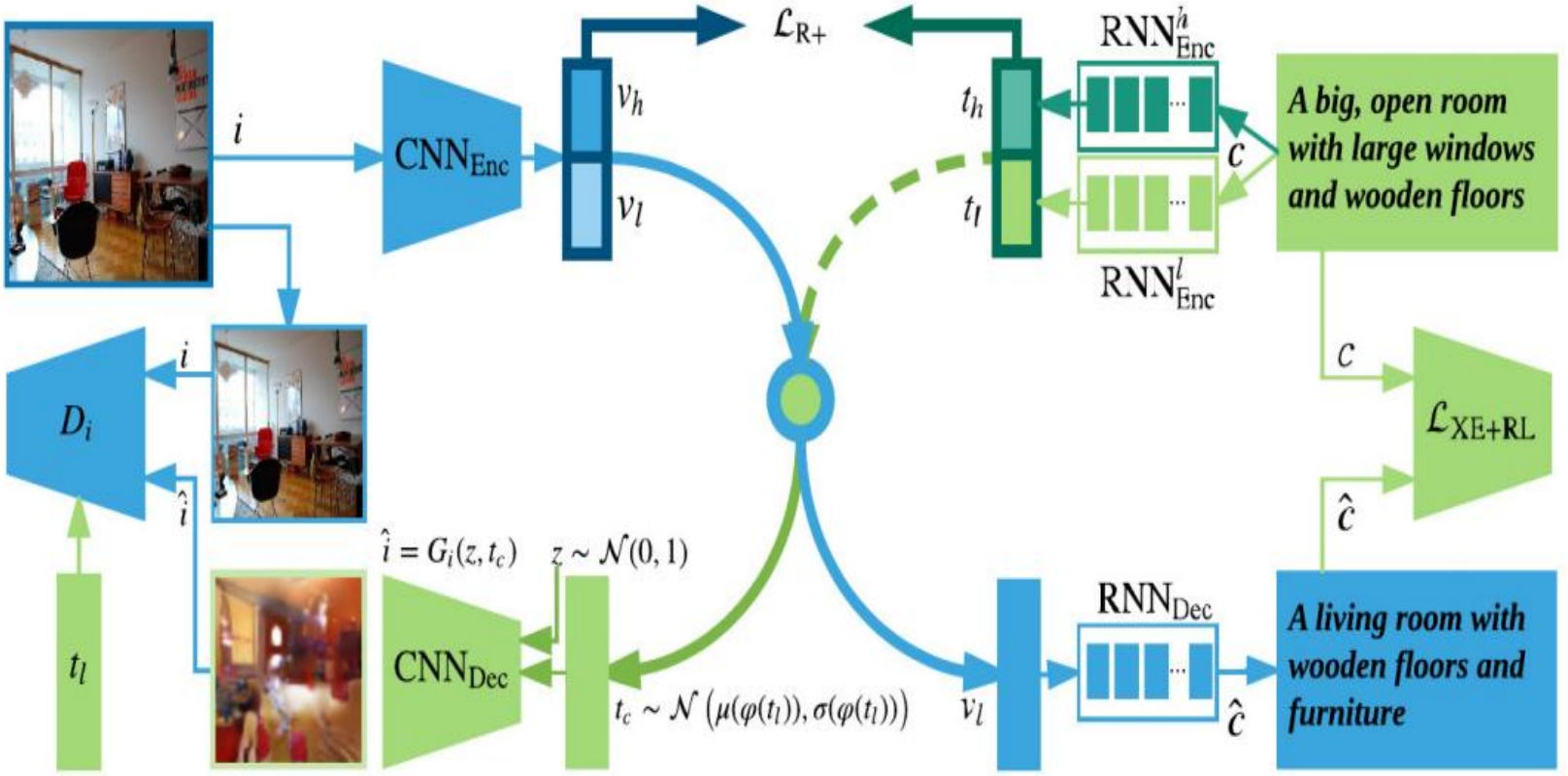


Figure 5. Analysis of performance given random input noise. Continuous lines depict $R@10$ values whereas dotted lines depict $R@1$ values.

Table 3. Bidirectional results on COCO-5cv test set. Bold values indicate the current state-of-the-art results.

Method	ConvNet	Image to text					Text to image				
		R@1	R@5	R@10	Med r	Mean r	R@1	R@5	R@10	Med r	Mean r
Order [29]	VGG-19	46.70	-	88.90	2.00	-	37.90	-	85.90	2.00	-
OECC [34]	VGG-19	47.20	78.60	88.90	2.00	5.60	37.50	74.60	87.00	2.00	7.30
SEAM-C [32]	VGG-19	50.70	81.40	90.90	1.40	4.90	40.30	75.70	87.40	2.00	7.40
Order (d=1024)	VGG-19	46.20	78.80	89.10	2.00	5.40	37.70	73.40	85.60	2.00	7.90
Order (d=4096)	VGG-19	48.80	79.20	89.70	1.60	5.20	38.70	74.10	86.40	2.00	7.60
Order (d=8192)	VGG-19	50.10	80.20	90.30	1.40	5.10	39.10	74.40	86.30	2.00	7.60
CHAIN-VSE-v1 (d=1024, p=1)	VGG-19	49.50	80.80	90.00	1.60	5.30	36.80	73.60	85.90	2.00	7.30
CHAIN-VSE-v1 (d=4096, p=1)	VGG-19	52.00	82.30	90.70	1.20	5.00	38.30	74.80	87.00	2.00	6.80
CHAIN-VSE-v1 (d=8192, p=1)	VGG-19	51.60	82.00	91.30	1.20	4.70	38.60	75.10	87.20	2.00	6.70
OECC [34]	IRv2	49.50	81.70	91.30	1.60	4.50	40.40	77.40	88.60	2.00	6.80
Order (d=1024)	IRv2	47.30	78.60	88.70	1.80	5.50	37.70	73.10	85.50	2.00	7.80
Order (d=4096)	IRv2	49.10	79.40	89.50	1.40	5.20	38.20	74.50	86.50	2.00	7.60
Order (d=8192)	IRv2	50.20	79.50	89.20	1.20	5.30	38.20	74.20	86.30	2.00	7.40
CHAIN-VSE-v1 (d=1024, p=1)	IRv2	50.50	83.60	92.20	1.60	4.30	39.00	76.20	88.10	2.00	6.80
CHAIN-VSE-v1 (d=4096, p=1)	IRv2	52.80	84.40	92.60	**1.00**	4.10	40.70	77.40	88.90	2.00	6.50
CHAIN-VSE-v1 (d=8192, p=1)	IRv2	53.70	85.10	93.10	**1.00**	3.90	40.70	77.60	89.00	2.00	6.30
CHAIN-VSE-v1 (d=1024, p=1)	ResNet-152	55.14	86.08	93.86	**1.00**	3.76	41.20	78.01	89.22	2.00	6.38
CHAIN-VSE-v1 (d=4096, p=1)	ResNet-152	57.76	87.88	**94.46**	**1.00**	3.42	42.96	79.20	90.01	2.00	6.05
CHAIN-VSE-v1 (d=8192, p=1)	ResNet-152	**59.40**	**87.98**	94.24	**1.00**	**3.37**	**43.47**	**79.78**	**90.22**	2.00	**5.90**



Gu J, Cai J, Joty S, et al. Look, Imagine and Match: Improving Textual-Visual Cross-Modal Retrieval with Generative Models[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 7181-7189.

- Image caption:当做一个预测分类问题

$$\mathcal{L}_{\text{xe}} = -\sum_{t=0}^{T-1} \log p_{\theta_t}(w_t | w_{0:t-1}, v_l; \theta_t)$$

- 文本生成图像 GAN, <real image, true caption> 为真，<fake image, true caption> and <real image, wrong caption>为假

Table 1: Cross-modal retrieval results on MSCOCO 1K-image test set (bold numbers are the best results).

Model	Image-to-Text			Text-to-Image		
	R@1	**R@10**	**Med**	**R@1**	**R@10**	**Med**
GRU(VGG19)	51.4	91.4	**1.0**	39.1	86.7	2.0
GRU_{Bi}(VGG19)	53.6	90.2	**1.0**	40.0	87.8	2.0
GXN(ResNet152)	59.4	94.7	**1.0**	47.0	92.6	2.0
GXN(fine-tune)	64.0	97.1	**1.0**	53.6	94.4	**1.0**
GXN(i2t,xe)	68.2	98.0	**1.0**	54.5	**94.8**	**1.0**
GXN(i2t,mix)	68.4	98.1	**1.0**	55.6	94.6	**1.0**
GXN(t2i)	67.1	**98.3**	**1.0**	56.5	**94.8**	**1.0**
GXN (i2t+t2i)	**68.5**	97.9	**1.0**	**56.6**	94.5	**1.0**

Huang Y, Wu Q, Wang L. Learning semantic concepts and order for image and sentence matching[C]//IEEE Conference on Computer Vision and Pattern Recognition. 2018, 2.

- 出发点：从图像学习一个concept,其实这些concept就是我们句子中的单词(动词，名词，形容词，数词）
- 得到这些概念了，关键是如何把他们正确的组织起来得到特征。两个办法：用全局上下文来进行组织（vgg最后的全连接层特征），然后有个image caption的loss

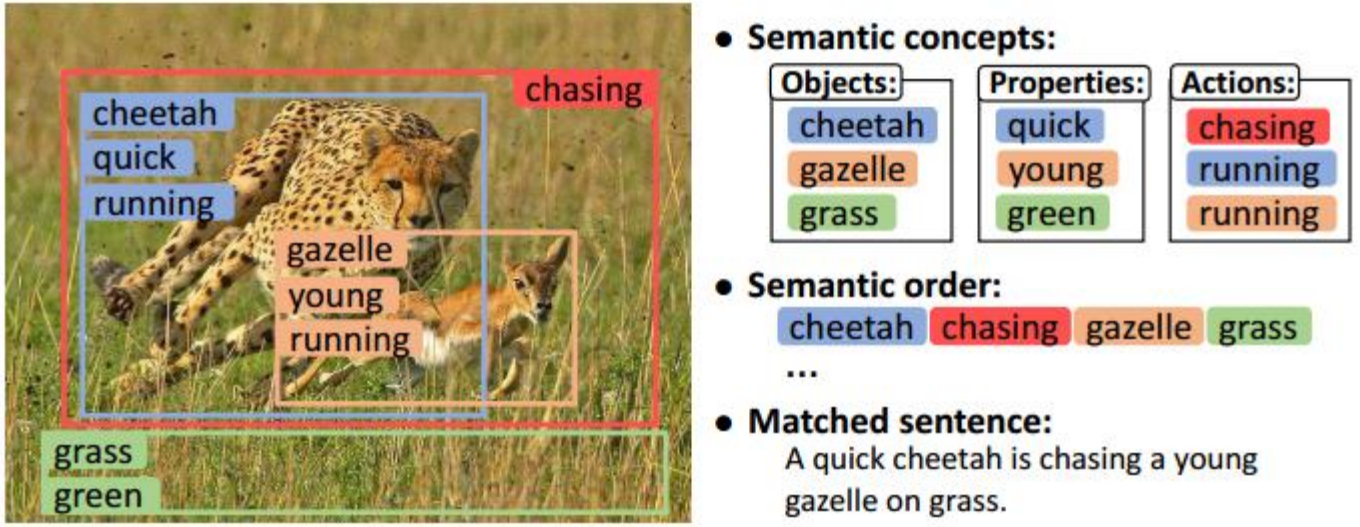


Figure 1. Illustration of the semantic concepts and order (best viewed in colors).

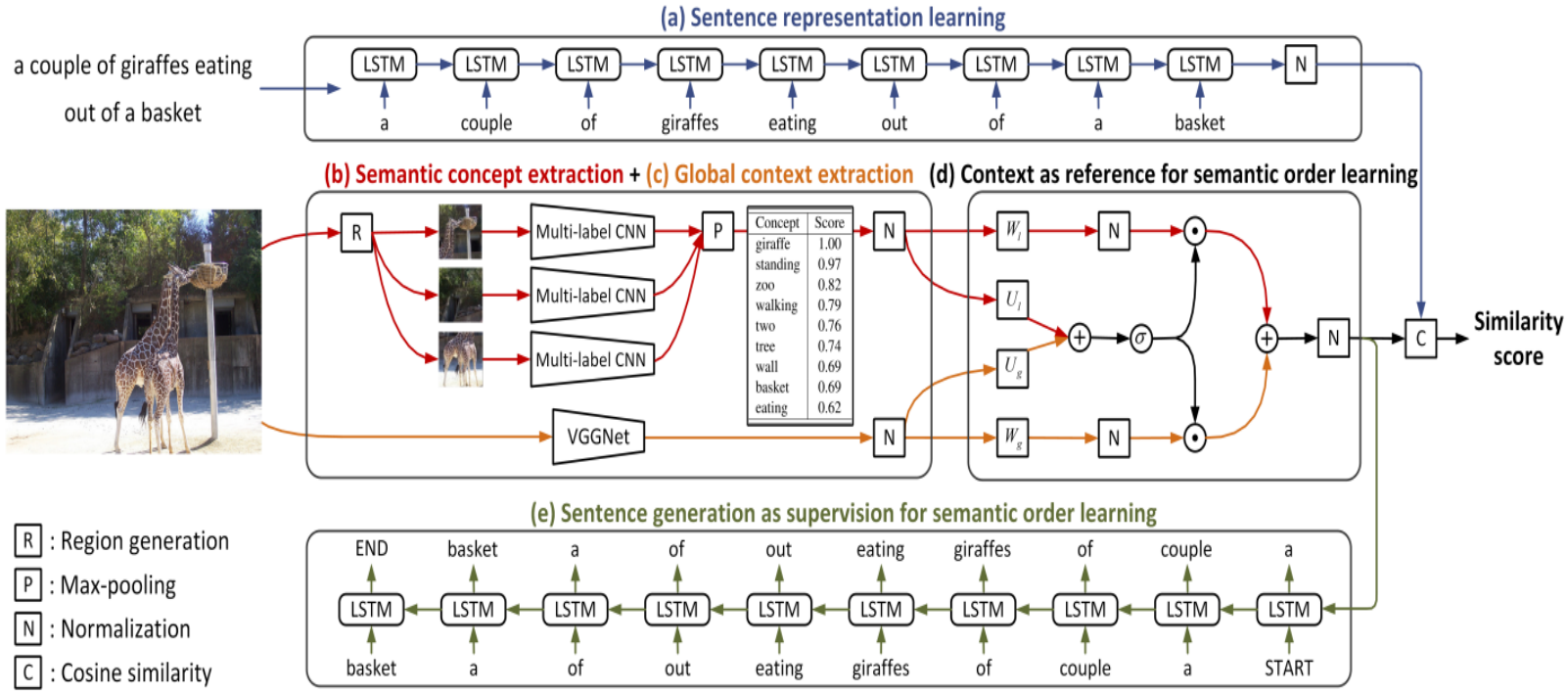


Figure 2. The proposed semantic-enhanced image and sentence matching model.

$$\hat{\mathbf{p}} = \|W_l \mathbf{p}\|_2, \ \hat{\mathbf{x}} = \|W_g \mathbf{x}\|_2, \ \mathbf{t} = \sigma(U_l \mathbf{p} + U_g \mathbf{x})$$
$$\mathbf{v} = \mathbf{t} \odot \hat{\mathbf{p}} + (\mathbf{1} - \mathbf{t}) \odot \hat{\mathbf{x}}$$

Huang Y, Wu Q, Wang L. Learning semantic concepts and order for image and sentence matching[C]//IEEE Conference on Computer Vision and Pattern Recognition. 2018, 2

Table 1. 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		√	√	√		√		√			√

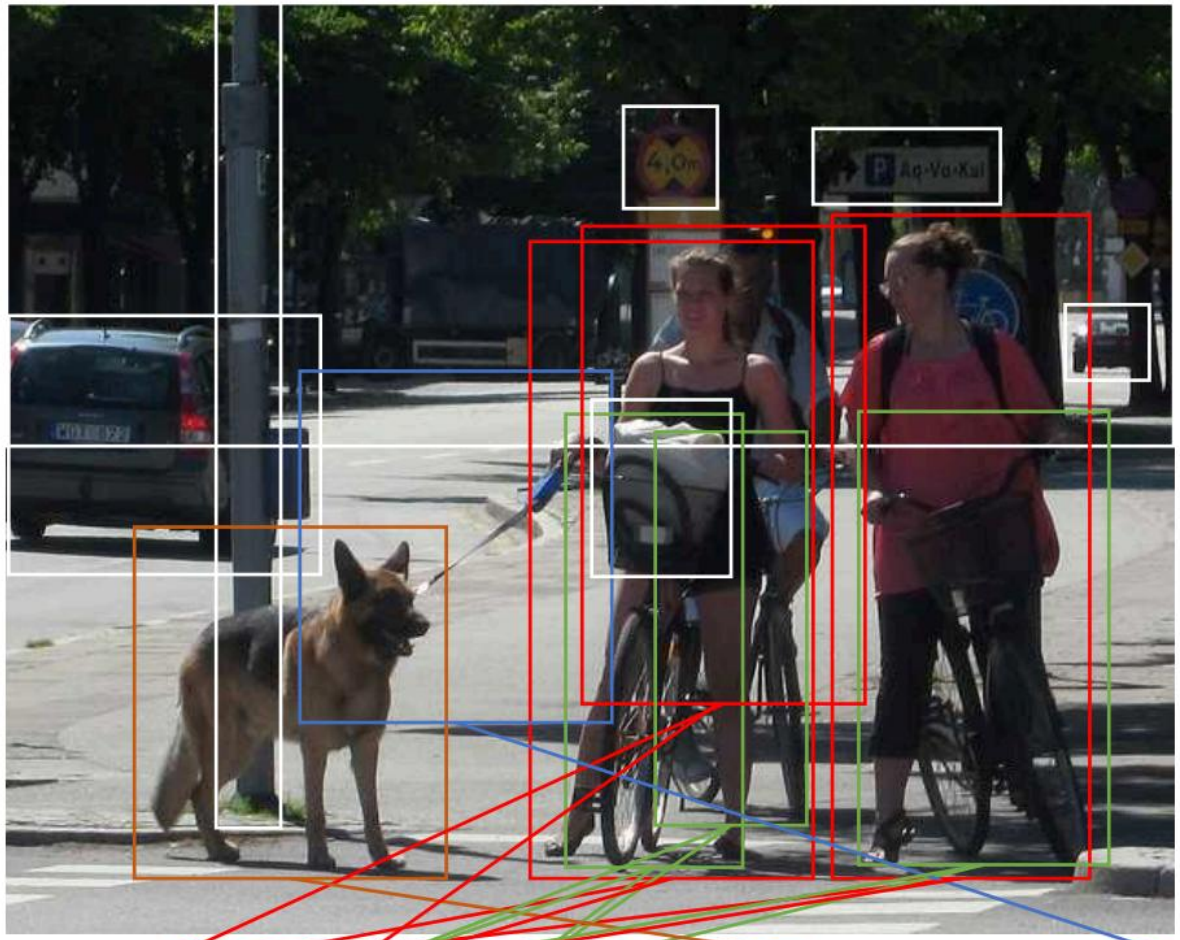
Table 2. Comparison results of image annotation and retrieval by ablation models on the Flickr30k and MSCOCO (1000 testing) datasets.

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	63.8	75.9	27.4	55.9	67.6	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 4. Comparison results of image annotation and retrieval on the Flickr30k and MSCOCO (1000 testing) 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	
m-RNN [24]	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 [16]	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 [14]	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 [15]	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 [23]	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 [18]	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 [31]	-	-	-	-	-	-	-	46.7	78.6	88.9	37.9	73.7	85.9	68.6
VQA [20]	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 [28]	37.4	63.1	74.3	26.0	56.0	69.3	54.3	-	-	-	-	-	-	-
DSPE [34]	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 [12]	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 [26]	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**
RRF (Res) [22]	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) [26]	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**	**80.1**	**69.7**	**69.9**	**92.9**	**97.5**	**56.7**	**87.5**	**94.8**	**83.2**

匹配得分学习法



匹配得分学习法

1. Deep Visual-Semantic Alignments for Generating Image Descriptions
2. Multimodal Convolutional Neural Networks for Matching Image and Sentence
3. Instance-aware Image and Sentence Matching with Selective Multimodal LSTM
4. Stacked Cross Attention for Image-Text Matching

Figure 1. Motivation/Concept Figure: Our model treats language as a rich label space and generates descriptions of image regions.

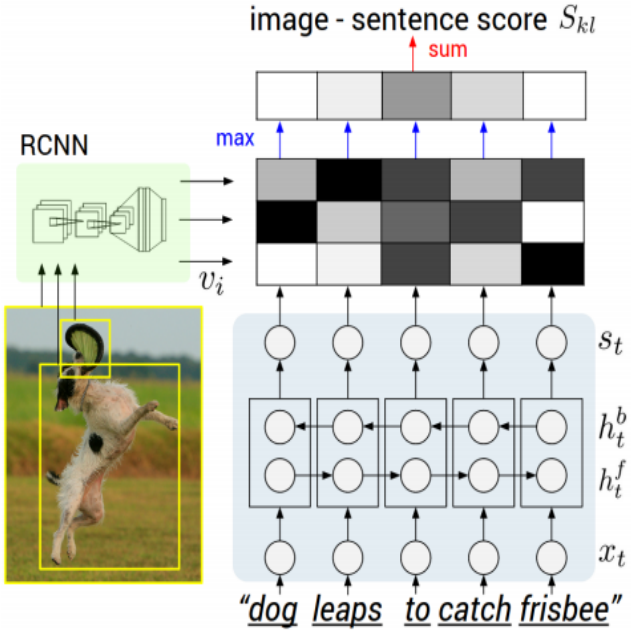


Figure 3. Diagram for evaluating the image-sentence score S_{kl}. Object regions are embedded with a CNN (left). Words (enriched by their context) are embedded in the same multimodal space with a BRNN (right). Pairwise similarities are computed with inner products (magnitudes shown in grayscale) and finally reduced to image-sentence score with Equation 8.

Karpathy A, Fei-Fei L. Deep visual-semantic alignments for generating image descriptions[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 3128-3137.

RCNN提取19个区域，用CNN提特征（4096）再转化为公共空间特征h（1000-1600）

双向RNN提每个单词的特征512，然后相加过全连接层变为h

$$S_{kl} = \sum_{t \in g_l} max_{i \in g_k} v_i^T s_t$$

$$\mathcal{C}(\theta) = \sum_k \Big[\underbrace{\sum_l max(0, S_{kl} - S_{kk} + 1)}_{\text{rank images}} + \underbrace{\sum_l max(0, S_{lk} - S_{kk} + 1)}_{\text{rank sentences}} \Big].$$

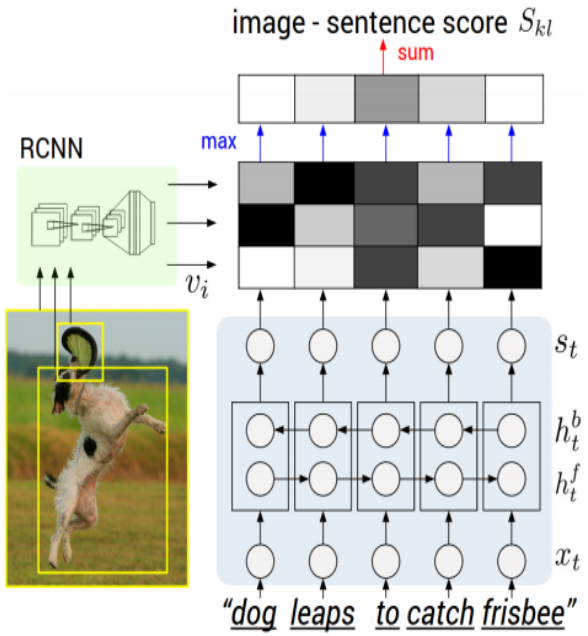


Figure 3. Diagram for evaluating the image-sentence score S_{kl}. Object regions are embedded with a CNN (left). Words (enriched by their context) are embedded in the same multimodal space with a BRNN (right). Pairwise similarities are computed with inner products (magnitudes shown in grayscale) and finally reduced to image-sentence score with Equation 8.

Model	Image Annotation				Image Search			
	R@1	R@5	R@10	Med *r*	R@1	R@5	R@10	Med *r*
Flickr30K								
SDT-RNN (Socher et al. [49])	9.6	29.8	41.1	16	8.9	29.8	41.1	16
Kiros et al. [25]	14.8	39.2	50.9	10	11.8	34.0	46.3	13
Mao et al. [38]	18.4	40.2	50.9	10	12.6	31.2	41.5	16
Donahue et al. [8]	17.5	40.3	50.8	9	-	-	-	-
DeFrag (Karpathy et al. [24])	14.2	37.7	51.3	10	10.2	30.8	44.2	14
Our implementation of DeFrag [24]	19.2	44.5	58.0	6.0	12.9	35.4	47.5	10.8
Our model: DepTree edges	20.0	46.6	59.4	5.4	15.0	36.5	48.2	10.4
Our model: BRNN	**22.2**	**48.2**	**61.4**	**4.8**	**15.2**	**37.7**	**50.5**	**9.2**
Vinyals et al. [54] (more powerful CNN)	23	-	63	5	17	-	57	8
MSCOCO								
Our model: 1K test images	38.4	69.9	80.5	1.0	27.4	60.2	74.8	3.0
Our model: 5K test images	16.5	39.2	52.0	9.0	10.7	29.6	42.2	14.0

Table 1. Image-Sentence ranking experiment results. **R@K** is Recall@K (high is good). **Med** *r* is the median rank (low is good). In the results for our models, we take the top 5 validation set models, evaluate each independently on the test set and then report the average performance. The standard deviations on the recall values range from approximately 0.5 to 1.0.

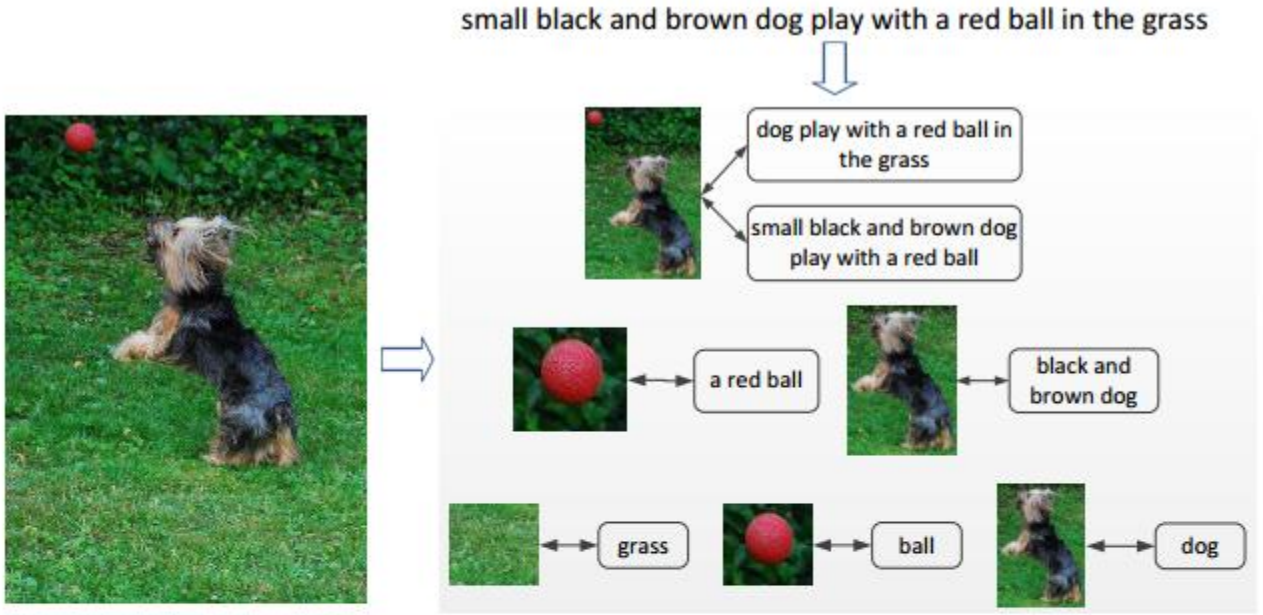


Figure 1. Multimodal matching relations between image and sentence. The words and phrases, such as "grass", "a red ball", and "small black and brown dog play with a red ball", correspond to the image areas of their grounding meanings. The sentence "small black and brown dog play with a red ball in the grass" expresses the meaning of the whole image.

Ma L, Lu Z, Shang L, et al. Multimodal convolutional neural networks for matching image and sentence[C]//Proceedings of the IEEE international conference on computer vision. 2015.

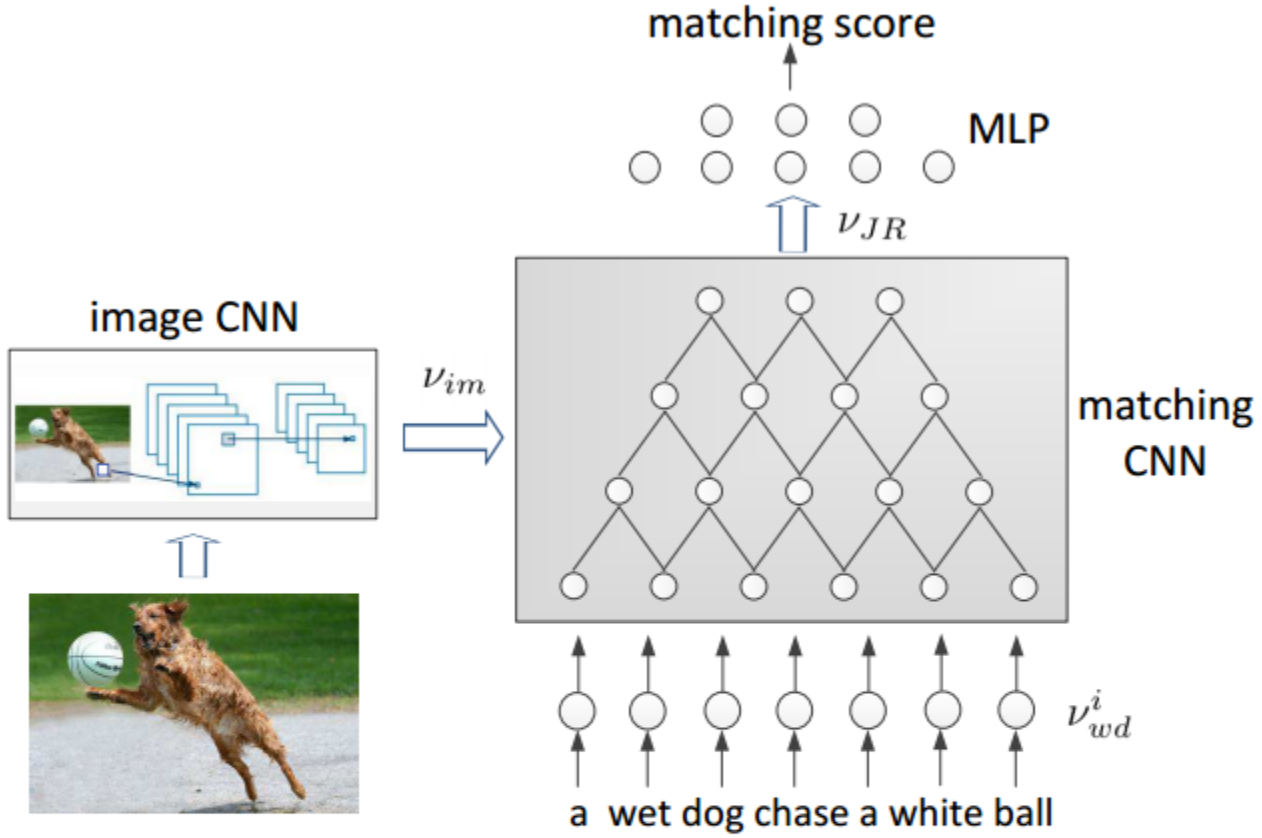
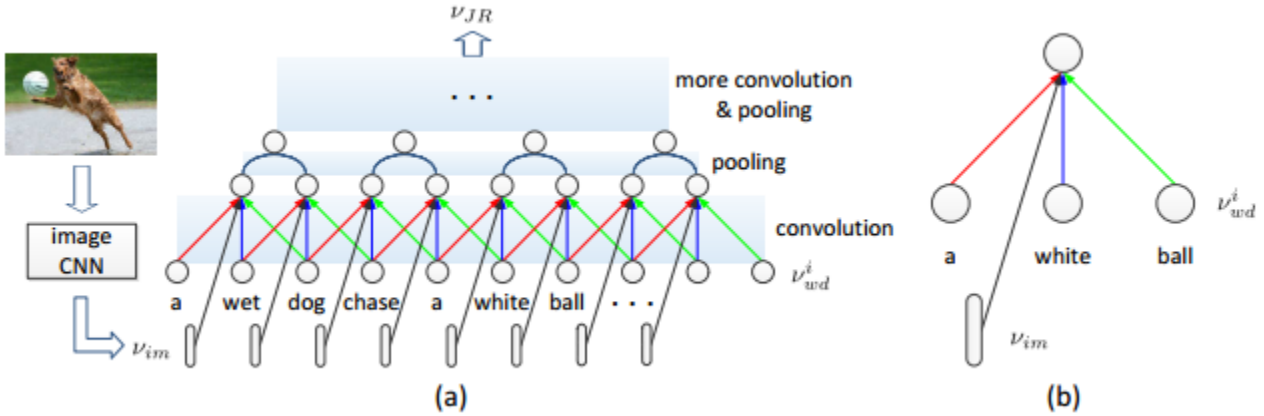
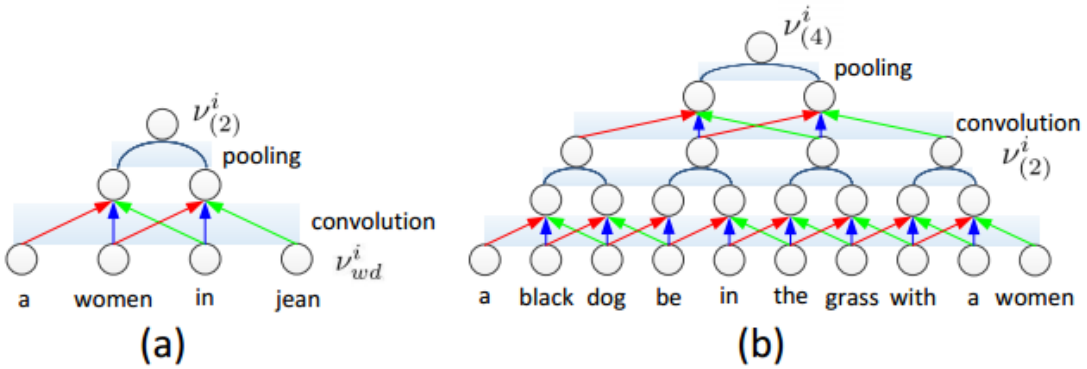


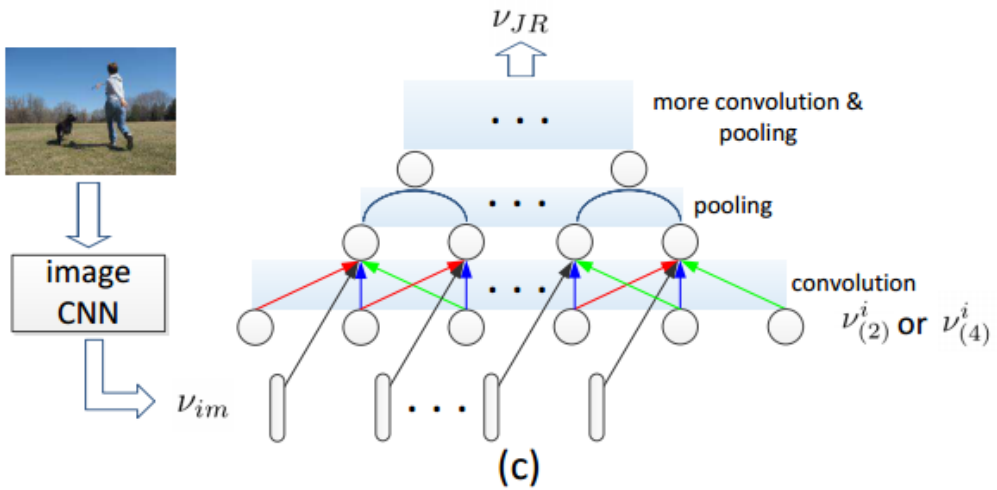
Figure 2. The m-CNN architecture for matching image and sentence. Image representation is generated by the image CNN. The matching CNN composes different fragments from the words of the sentence and learns the joint representation of image and sentence fragments. MLP summarizes the joint representation and produces the matching score.

word level





phrase level



Sentence-level matching

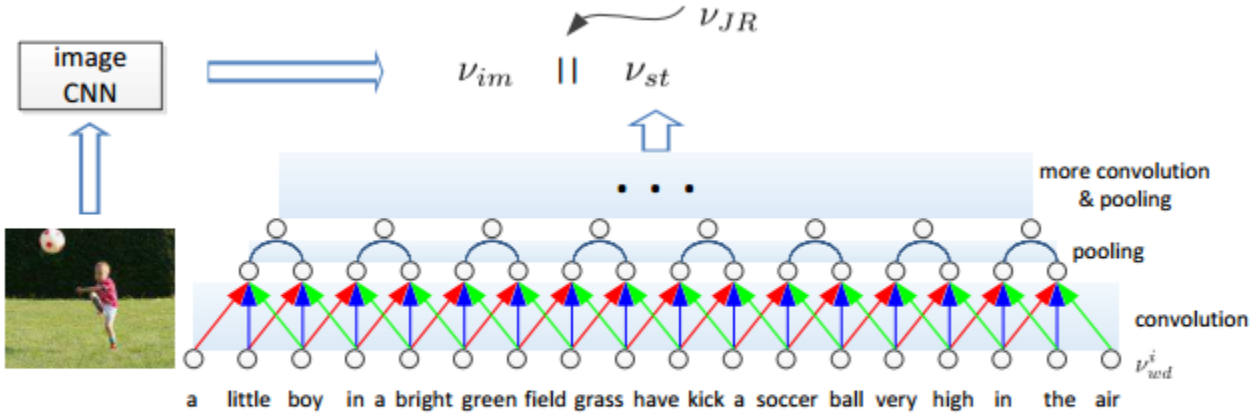


Figure 5. The sentence-level matching CNN. The joint representation is obtained by concatenating the image and sentence representations together.

Table 3. Bidirectional image and sentence retrieval results on Flickr30K.

	Sentence Retrieval				Image Retrieval			
	R@1	R@5	R@10	Med r	R@1	R@5	R@10	Med r
Random Ranking	0.1	0.6	1.1	631	0.1	0.5	1.0	500
DeViSE [6]	4.5	18.1	29.2	26	6.7	21.9	32.7	25
SDT-RNN [30]	9.6	29.8	41.1	16	8.9	29.8	41.1	16
MNLM [20]	14.8	39.2	50.9	10	11.8	34.0	46.3	13
MNLM-VGG [20]	23.0	50.7	62.9	5	16.8	42.0	56.5	8
m-RNN [24]	18.4	40.2	50.9	10	12.6	31.2	41.5	16
m-RNN-VGG [23]	**35.4**	63.8	73.7	**3**	22.8	50.7	63.1	5
Deep Fragment [16]	14.2	37.7	51.3	10	10.2	30.8	44.2	14
RVP (T) [3]	11.9	25.0	47.7	12	12.8	32.9	44.5	13
RVP (T+I) [3]	12.1	27.8	47.8	11	12.7	33.1	44.9	12.5
DVSA (DepTree) [17]	20.0	46.6	59.4	5.4	15.0	36.5	48.2	10.4
DVSA (BRNN) [17]	22.2	48.2	61.4	4.8	15.2	37.7	50.5	9.2
NIC [34]	17.0	*	56.0	7	17.0	*	57.0	7
LRCN [5]	*	*	*	*	17.5	40.3	50.8	9
OverFeat [28]:								
m-CNN$_{wd}$	12.7	30.2	44.5	14	11.6	32.1	44.2	14
m-CNN$_{phs}$	14.4	38.6	49.6	11	12.4	33.3	44.7	14
m-CNN$_{phl}$	13.8	38.1	48.5	11.5	11.6	32.7	44.1	14
m-CNN$_{st}$	14.8	37.9	49.8	11	12.5	32.8	44.2	14
m-CNN$_{ENS}$	20.1	44.2	56.3	8	15.9	40.3	51.9	9.5
VGG [29]:								
m-CNN$_{wd}$	21.3	53.2	66.1	5	18.2	47.2	60.9	6
m-CNN$_{phs}$	25.0	54.8	66.8	4.5	19.7	48.2	62.2	6
m-CNN$_{phl}$	23.9	54.2	66.0	5	19.4	49.3	62.4	6
m-CNN$_{st}$	27.0	56.4	70.1	4	19.7	48.4	62.3	6
m-CNN$_{ENS}$	33.6	**64.1**	**74.9**	**3**	**26.2**	**56.3**	**69.6**	**4**

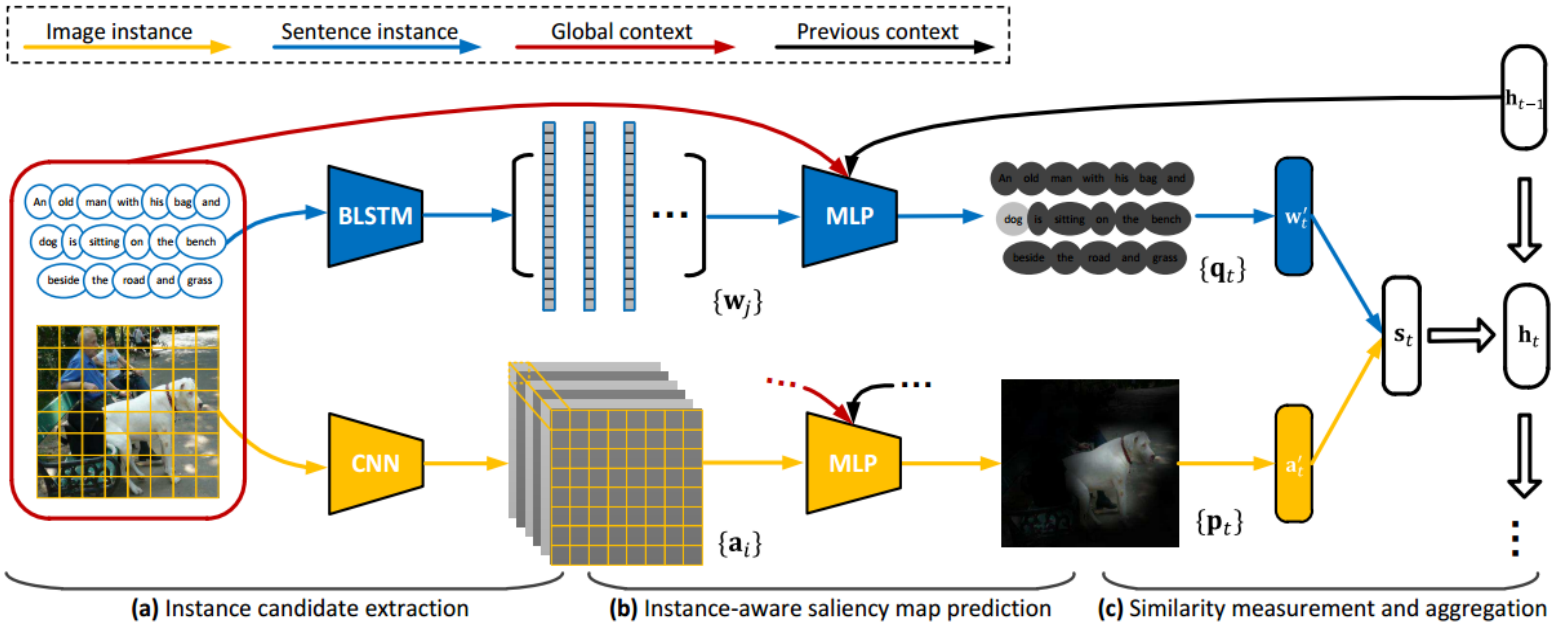


Figure 2. Details of the proposed sm-LSTM, including (a) instance candidate extraction, (b) instance-aware saliency map prediction, and (c) similarity measurement and aggregation (best viewed in colors).

A: 对于图像和句子都提候选区域

B: 用上下文注意力机制进行显著性图预测

C: 局部相似性计算和聚合

Huang Y, Wang W, Wang L. Instance-aware image and sentence matching with selective multimodal lstm[C]//The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017, 2(6): 7.

- A: 句子用双向LSTM,连接两个方向的单词特征；图像通过最后的卷积层来确定区域特征

- B: 不是所有的候选都是有效，所有要选择有效区域；另外并不是一对一的关系，CNN全连接层最后出来的特征作为图像全局上下文**m**，LSTM最后时间步长出来的特征作为句子全局上下文**n**。

$$p_{t,i} = e^{\hat{p}_{t,i}} / \sum_{i=1}^{I} e^{\hat{p}_{t,i}},\ \hat{p}_{t,i} = f_p(\mathbf{m}, \mathbf{a}_i, \mathbf{h}_{t-1}),$$

$$q_{t,j} = e^{\hat{q}_{t,j}} / \sum_{j=1}^{J} e^{\hat{q}_{t,j}},\ \hat{q}_{t,j} = f_q(\mathbf{n}, \mathbf{w}_j, \mathbf{h}_{t-1})$$

全局上下文求取显著性区域

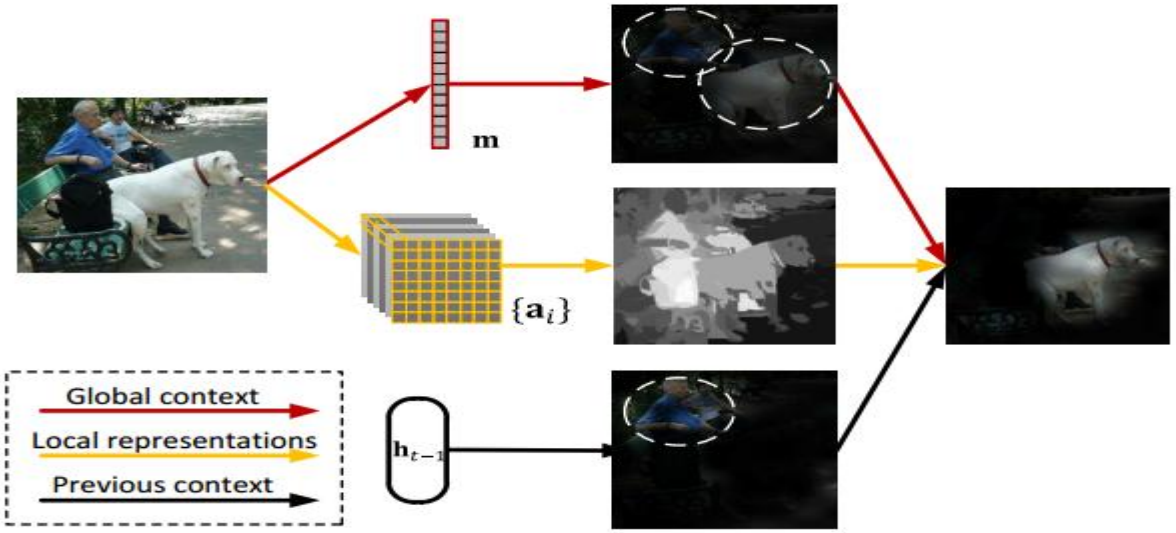


Figure 3. Illustration of context-modulated attention (the lighter areas indicate the attended instances, best viewed in colors).

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

相似性计算和聚合

- 得到每一个时间步的加权特征，然后过MLP

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

- 所有时间步长过LSTM

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

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

学习目标

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

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

where λ is a balancing parameter. By adding this constraint, the loss will be large when our model attends to the same instance for multiple times. Therefore, it encourages the model to pay equal attention to every instance rather than a certain one for information maximization. In our experiments, we find that using this regularization can further improve the performance.

Table 3. The impact of different numbers of timesteps on the Flick30k dataset. T: the number of timesteps in the sm-LSTM.

	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

Table 4. The impact of different values of the balancing parameter on the Flick30k dataset. λ: the balancing parameter between structured objective and regularization term.

	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

(a) Input image (b) Without global context (by sm-LSTM-att) (c) With global context (by sm-LSTM)

Figure 5. Attended image instances at three different timesteps, without or with global context, respectively (best viewed in colors).

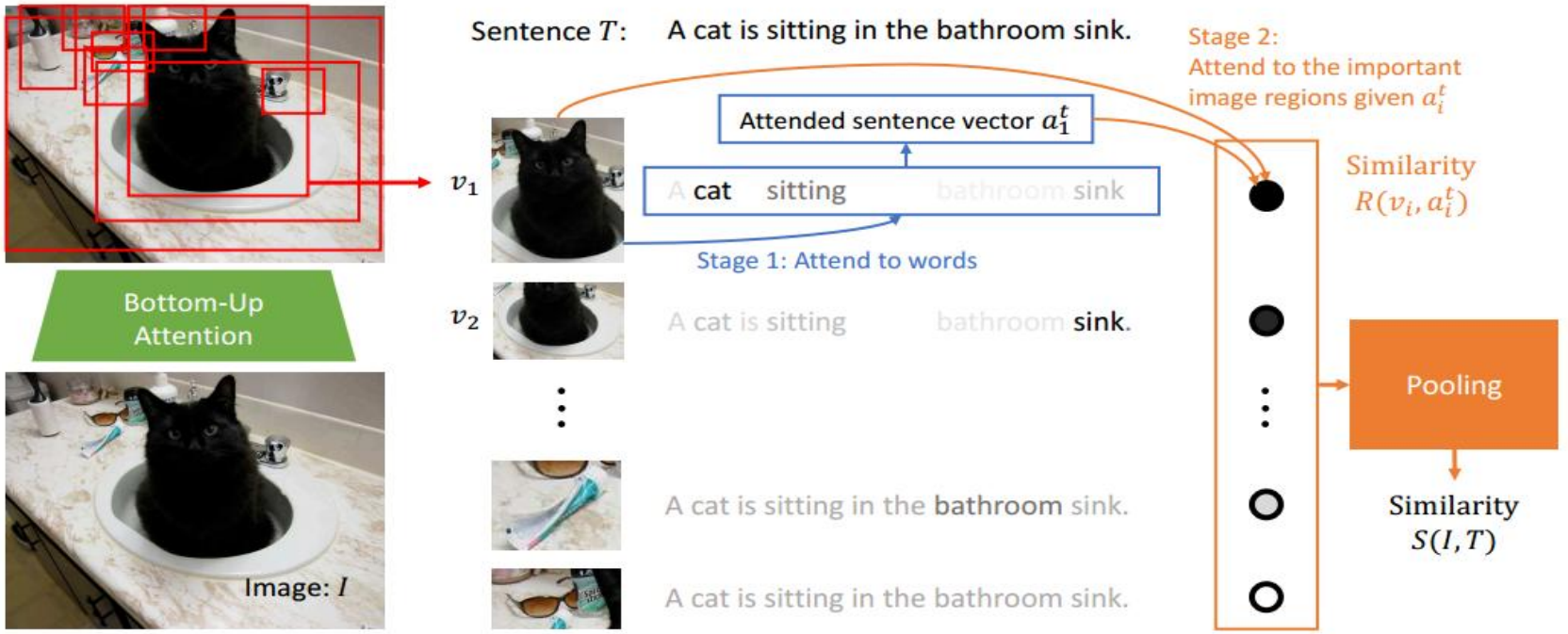


Fig. 2. Image-Text Stacked Cross Attention: At stage 1, we first attend to words in the sentence with respect to each image region feature v_i to generate an attended sentence vector a_i^t for i-th image region. At stage 2, we compare a_i^t and v_i to determine the importance of each image region, and then compute the similarity score.

Lee K H, Chen X, Hua G, et al. Stacked cross attention for image-text matching[C]//Proceedings of the European Conference on Computer Vision (ECCV). 2018.

有attention的图像区域和单词的对应，假设图像区域k个，单词n个，可利用如下计算相似性，加归一化处理：

$$s_{ij} = \frac{v_i^T e_j}{||v_i|| ||e_j||}, i \in [1,k], j \in [1,n].$$

$$\bar{s}_{ij} = [s_{ij}]_+ / \sqrt{\sum_{i=1}^{k} [s_{ij}]_+^2}, \text{ where } [x]_+ \equiv max(x,0)$$

对每一个图像区域，加权单词：

$$a_i^t = \sum_{j=1}^{n} \alpha_{ij} e_j,$$

$$\alpha_{ij} = \frac{exp(\lambda_1 \bar{s}_{ij})}{\sum_{j=1}^{n} exp(\lambda_1 \bar{s}_{ij})},$$

- 那么该区域与对应的加权单词的相似性可求得：

$$R(v_i, a_i^t) = \frac{v_i^T a_i^t}{||v_i|| ||a_i^t||}.$$

- 那么整个图像和句子的相似性：

$$S_{LSE}(I,T) = log(\sum_{i=1}^{k} exp(\lambda_2 R(v_i, a_i^t)))^{(1/\lambda_2)},$$

$$S_{AVG}(I,T) = \frac{\sum_{i=1}^{k} R(v_i, a_i^t)}{k}.$$

- 优化目标：

$$l_{hard}(I,T) = [\alpha - S(I,T) + S(I,\hat{T}_h)]_+ + [\alpha - S(I,T) + S(\hat{I}_h,T)]_+.$$

- Faster R-CNN提区域，ResNet101提图像区域特征2048（注意提特征的模型是在Visual Genomes训练，具有丰富的语义信息，有实例类别和属性类别，不光光是目标类别）
- 实例类别包括目标类别和其他显著的难以定位的stuff,如sky,glass,building. 属性就包括毛织的等
- 单词特征用双向GRU,然后取平均

Table 1. Comparison of the cross-modal retrieval restuls in terms of Recall@K(R@K) on Flickr30K. t-i denotes Text-Image. i-t denotes Image-Text. AVG and LSE denotes average and LogSumExp pooling respectively.

Method	Sentence Retrieval			Image Retrieval		
	R@1	R@5	R@10	R@1	R@5	R@10
DVSA (R-CNN, AlexNet) [19]	22.2	48.2	61.4	15.2	37.7	50.5
HM-LSTM (R-CNN, AlexNet) [32]	38.1	-	76.5	27.7	-	68.8
SM-LSTM (VGG) [16]	42.5	71.9	81.5	30.2	60.4	72.3
2WayNet (VGG) [9]	49.8	67.5	-	36.0	55.6	-
DAN (ResNet) [31]	55.0	81.8	89.0	39.4	69.2	79.1
VSE++ (ResNet) [10]	52.9	-	87.2	39.6	-	79.5
DPC (ResNet) [44]	55.6	81.9	89.5	39.1	69.2	80.9
SCO (ResNet) [17]	55.5	82.0	89.3	41.1	70.5	80.1
Ours (Faster R-CNN, ResNet):						
SCAN t-i LSE ($\lambda_1 = 9, \lambda_2 = 6$)	61.1	85.4	91.5	43.3	71.9	80.9
SCAN t-i AVG ($\lambda_1 = 9$)	61.8	87.5	93.7	45.8	74.4	83.0
SCAN i-t LSE ($\lambda_1 = 4, \lambda_2 = 5$)	67.7	88.9	94.0	44.0	74.2	82.6
SCAN i-t AVG ($\lambda_1 = 4$)	**67.9**	89.0	94.4	43.9	74.2	82.8
SCAN t-i AVG + i-t LSE	67.4	**90.3**	**95.8**	**48.6**	**77.7**	**85.2**

Method	Sentence Retrieval R@1	R@5	R@10	Image Retrieval R@1	R@5	R@10
			1K Test Images			
DVSA (R-CNN, AlexNet) [19]	38.4	69.9	80.5	27.4	60.2	74.8
HM-LSTM (R-CNN, AlexNet) [32]	43.9	-	87.8	36.1	-	86.7
Order-embeddings (VGG) [38]	46.7	-	88.9	37.9	-	85.9
SM-LSTM (VGG) [16]	53.2	83.1	91.5	40.7	75.8	87.4
2WayNet (VGG) [9]	55.8	75.2	-	39.7	63.3	-
VSE++ (ResNet) [10]	64.6	-	95.7	52.0	-	92.0
DPC (ResNet) [44]	65.6	89.8	95.5	47.1	79.9	90.0
GXN (ResNet) [13]	68.5	-	97.9	56.6	-	94.5
SCO (ResNet) [17]	69.9	92.9	97.5	56.7	87.5	**94.8**
Ours (Faster R-CNN, ResNet):						
SCAN t-i LSE ($\lambda_1 = 9, \lambda_2 = 6$)	67.5	92.9	97.6	53.0	85.4	92.9
SCAN t-i AVG ($\lambda_1 = 9$)	70.9	94.5	97.8	56.4	87.0	93.9
SCAN i-t LSE ($\lambda_1 = 4, \lambda_2 = 20$)	68.4	93.9	98.0	54.8	86.1	93.3
SCAN i-t AVG ($\lambda_1 = 4$)	69.2	93.2	97.5	54.4	86.0	93.6
SCAN t-i LSE + i-t AVG	**72.7**	**94.8**	**98.4**	**58.8**	**88.4**	**94.8**
			5K Test Images			
Order-embeddings (VGG) [38]	23.3	-	84.7	31.7	-	74.6
VSE++ (ResNet) [10]	41.3	-	81.2	30.3	-	72.4
DPC (ResNet) [44]	41.2	70.5	81.1	25.3	53.4	66.4
GXN (ResNet) [13]	42.0	-	84.7	31.7	-	74.6
SCO (ResNet) [17]	42.8	72.3	83.0	33.1	62.9	75.5
Ours (Faster R-CNN, ResNet):						
SCAN i-t LSE	46.4	77.4	87.2	34.4	63.7	75.7
SCAN t-i AVG + i-t LSE	**50.4**	**82.2**	**90.0**	**38.6**	**69.3**	**80.4**

Table 3. Effect of inferring the latent vision-language alignment at the level of regions and words. Results are reported in terms of Recall@K(R@K). Refer to Eqs. (9) (10) for the definition of Sum-Max. t-i denotes Text-Image. i-t denotes Image-Text.

Method	Sentence Retrieval			Image Retrieval		
	R@1	R@5	R@10	R@1	R@5	R@10
VSE++ (fixed ResNet, 1 crop) [10]	31.9	-	68.0	23.1	-	60.7
Sum-Max t-i	59.6	85.2	92.9	44.1	70.0	79.0
Sum-Max i-t	56.7	83.5	89.7	36.8	65.6	74.9
SCO [17] (current state-of-the-art)	55.5	82.0	89.3	41.1	70.5	80.1
SCAN t-i AVG ($\lambda_1 = 9$)	61.8	87.5	93.7	45.8	74.4	83.0
SCAN i-t AVG ($\lambda_1 = 10$)	67.9	89.0	94.4	43.9	74.2	82.8

总结

Thanks!