分类：计算机视觉 / 视频和图像处理

联系人 | 研究海报

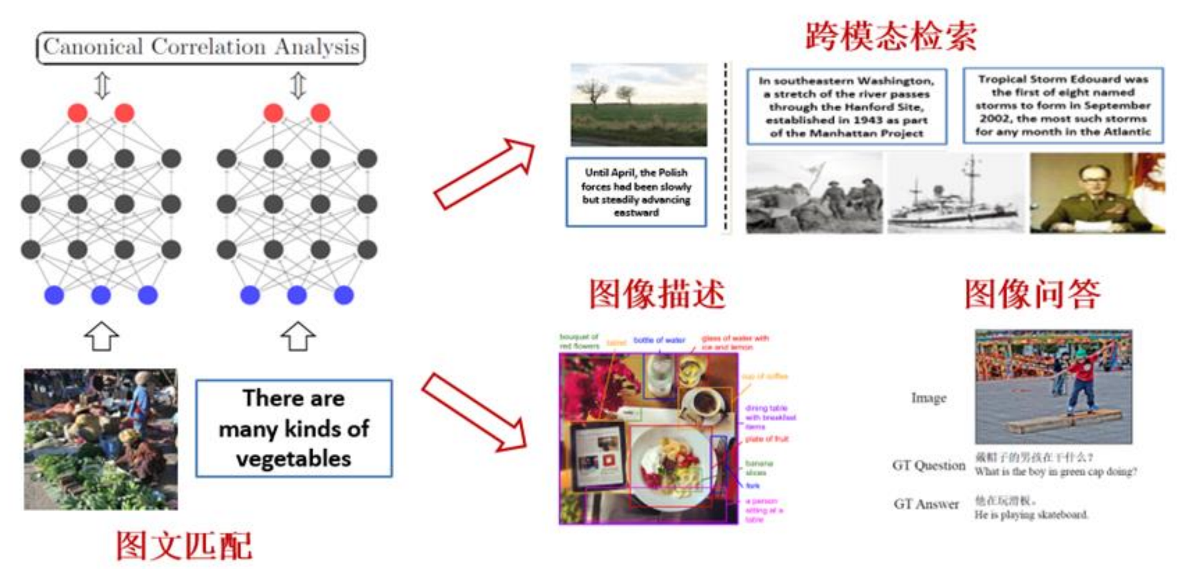
袁辉：HUI.YUAN@CRIPAC.IA.AC.CN | CNP20867

VSR++: *Improving Visual Semantic Reasoning for Fine-Grained Image-Text Matching*

Hui Yuan[1,2], Yan Huang[2], Dongbo Zhang[1,*], and Liang Wang[2,3,4]

1. The College of Automation and Electronic Information, Xiangtan University, Xiangtan, China
2. Center for Research on Intelligent Perception and Computing (CRIPAC), National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences (CASIA)
3. Center for Excellence in Brain Science and Intelligence Technology (CEBSIT)
4. Chinese Academy of Sciences, Artificial Intelligence Research (CAS-AIR)
5. hui.yuan@cripac.ia.ac.cn

Background



Motivation

Recently, the state-of-the-art performance of image-text matching is achieved by Visual Semantic Reasoning Network (VSRN)[5], which measures the global alignment between images and texts. Besides, existing other works [1,7] have demonstrated the effectiveness of mapping whole images and full texts to a common semantic vector space for imagetext matching. But they all suffer from the common drawback that they cannot well measure the local correspondence between image regions and text words.

As a result, they cannot well perform the image-text matching in a fine-grained manner. Especially in real-world applications, many gallery images and texts containing appearancesimilar regions and words that are very difficult to distinguish with each other. But how to well measure the local correspondence between regions and words is a challenging task, since the groundtruth annotation of local correspondence is unknown. Although some existing works [2,8] attempt to model the local correspondence in a weakly-supervised manner, how to combine them into the state-of-the-art framework of global alignment is investigated.

To this end, we propose an Improved Visual Semantic Reasoning model (VSR++), which can improve the global alignment by using additionally model the local correspondence, with the goal to improve the measurement of crossmodal similarity for fine-grained image-text matching. As an extension of VSRN, our VSR++ models the local correspondence between regions and words in the context of global alignment by bi-directional stacked cross-modal attention[2]. By incorporating the complementary advantages of global alignment and local correspondence, as well as balancing their relative importance, our VSR++ achieves the current state-of-the-art performance.

Model

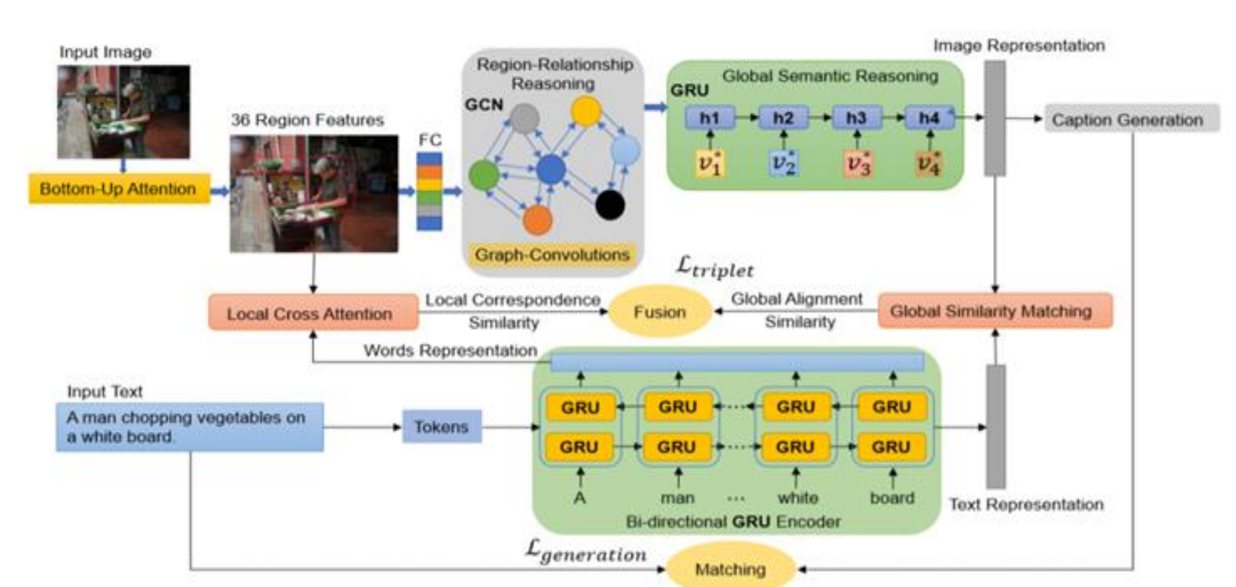


Fig. 1. Our proposed VSR++ model, which can incorporate the advantages of global image-text alignment and local region-word correspondence for fine-grained image-text matching.

A. Image Representation

We extract a set of features $V = \{v_1, \dots, v_k\}, v_i \in R^D$ from each image I by the bottom-up attention mechanism[3], such that each feature v_i encodes an object or a salient region in this image.

$$v_i = W_f f_i + b_f$$

B. Global Visual Semantic Similarity

- We first build up connections among image regions and perform region relationship reasoning with Graph Convolutional Networks (GCNs)[9] to generate features with semantic relationships.

$$R = (W_a \cdot v_i)^T (W_b \cdot v_j)$$

- After that, we also use the GRUs network to perform global semantic reasoning on these features with semantic relationships to generate the final global representation of the image.
- we use a bidirectional text-based GRU[4] encoder to map the whole text T to the same D-dimensional semantic vector space R^D as the text global representation T_{global}.
- Then we adopt the cosine similarity function to measure the similarity between the global image representation IG and the global text representation T_G.

$$S_G(I,T) = I_G \cdot T_G$$

C. Local Fine-Grained Correspondence

- Image-Text Local Cross-modal Attention

$$s_{ij} = v_i^T e_j, i \epsilon [1,k], j \epsilon [1,n]$$
$$w_{ij} = softmax(\lambda \hat{s}_{ij})$$
$$r_i^t = \sum_{j=1}^{n} w_{ij} e_j$$
$$R_i^t = v_i^T r_i^t$$
$$S_L(I,T) = \frac{\sum_{i=1}^{k} norm(R_i^t)}{k}$$

-Obtain the final image-text similarity S_L(I, T) in a locally fine-grained correspondence.

D. Model Learning Strategy

-we comprehensively fuse two similarity scores for global image-text alignment and local region-word correspondence, as well as balance their relative importance at a certain ratio.

$$S(I,T) = S_G(I,T) + \mu S_L(I,T)$$

-we adopt a hinge-based triplet ranking loss to learn the matching part.

$$\mathcal{L}_{triplet} = max[0, \alpha - S(I,T) + S(I,\hat{T})] + max[0, \alpha - S(I,T) + S(\hat{I},T)]$$

-The training loss[5,6] for text generation is represented as:

$$\mathcal{L}_{generation} = -\sum_{t=1}^{l} logp(y_t|y_{t-1}, V^*; \theta)$$

-In order to jointly match and generate for model learning, our final loss function is defined as follows:

$$\mathcal{L} = \mathcal{L}_{triplet} + \mathcal{L}_{generation}$$

Experimental Results

We implement our experiment in PyTorch framework with an NVIDIA GeForce GTX 2080Ti GPU. We use the Adam optimizer to train the model with 30 epochs. And It only takes about 10 hours to finish the training.

A. Evaluation of Ablation Models

1) “Global”, which only performs the global image-text alignment learning.
2) “Local”, which only performs the local fine-grained correspondence learning.
3) ”Fusion-loss“, which only considers the mutual influence of the training loss of the two modules (global image-text alignment and local fine-grained correspondence) during the training process, but does not fuse their similarity.4) ”Fusion-similarity”, which associates the similarity of the two modules and learns the model in a unified framework.
5) ”VSR++(GRU)”, a network that only uses GRU instead of Bi-GRU as the text encoder in our full VSR++ model.
6) ”VSR++(full)”, which denotes the full VSR++ model.

-μ represents the association parameter.

Methods	Flickr30k dataset					
	Text Retrieval			Image Retrieval		
	R@1	R@5	R@10	R@1	R@5	R@10
Global	71.3	90.6	96.0	54.7	81.8	88.2
Local	67.4	90.3	95.8	48.6	77.7	85.2
Fusion-loss	71.5	90.6	95.8	55.1	82.0	88.2
Fusion-similarity	72.2	92.5	97.0	56.1	82.3	89.0
VSR++ (GRU)	72.0	92.1	96.5	55.6	82.0	88.5
VSR++ (full)	**72.6**	**92.7**	**97.2**	**56.3**	**82.7**	**89.0**

Methods	Flickr30k dataset					
	Text Retrieval			Image Retrieval		
	R@1	R@5	R@10	R@1	R@5	R@10
VSR++ (μ=0.5)	66.4	89.1	94.6	53.9	81.7	88.2
VSR++ (μ=1.0)	69.3	91.5	96.1	56.0	82.6	89.0
VSR++ (μ=1.5)	72.0	92.2	97.1	56.1	82.7	89.0
VSR++ (μ=2.5)	72.4	93.0	96.8	54.9	81.8	88.8
VSR++ (μ=3.0)	72.1	93.3	96.7	54.5	81.4	88.4
VSR++ (μ=2.0)	**72.6**	**92.7**	**97.2**	**56.3**	**82.7**	**89.0**

B. Comparisons With The State-of-the-art

Results on MS-COCO

ethods	MS-COCO 1K dataset						
	Text Retrieval			Image Retrieval			mR
	R@1	R@5	R@10	R@1	R@5	R@10	
VSE++ [1]	64.6	89.1	95.7	52.0	83.1	92.0	79.4
SCO [3]	69.9	92.9	97.5	56.7	87.5	94.8	83.2
SCAN [2]	72.7	94.8	98.4	58.8	88.4	94.8	84.7
GVSE [12]	72.2	94.1	98.1	60.5	89.4	95.8	85.0
SAEM [6]	71.2	94.1	97.7	57.8	88.6	94.9	84.0
VSRN [4]	76.2	94.8	98.2	62.8	89.7	95.1	86.1
VSR++	**76.6**	**95.2**	**98.2**	**63.4**	**90.6**	**95.7**	**86.6**

Results on Flickr30K

Methods	Flickr30k dataset						
	Text Retrieval			Image Retrieval			mR
	R@1	R@5	R@10	R@1	R@5	R@10	
VSE++ [1]	52.9	79.1	87.2	39.6	69.6	79.5	68.0
SCO [3]	55.5	82.0	89.3	41.1	70.5	80.1	69.7
SCAN [2]	67.4	90.3	95.8	48.6	77.7	85.2	77.5
GVSE [12]	68.5	90.9	95.5	50.6	79.8	87.6	78.8
SAEM [6]	69.1	91.0	95.1	52.4	81.1	88.1	79.4
VSRN [4]	71.3	90.6	96.0	54.7	81.8	88.2	80.4
VSR++	**72.6**	**92.7**	**97.2**	**56.3**	**82.7**	**89.0**	**81.8**

C. Visualization and Analysis

-Qualitative results of two different methods in the image-to-text retrieval.

Fig. 2. Qualitative results of image-to-text retrieval for two different methods (VSRN and VSR++) on Flickr30K dataset. For each image query, we show separately the top 3 ranked text caption in two different methods. Ground-truth matched texts are with check marks, while some mismatched texts are with x marks. Besides, locally mismatched phrases are with red font marks.

- Qualitative results of two different methods in the textto-image retrieval.

Fig. 3. Qualitative results of the text-to-image retrieval for two different methods (VSRN and VSR++) on Flickr30K dataset. The first row of images represents the retrieval result of the VSRN model, and the second row of images represents the retrieval result of our VSR++ model. For each text query, we show top 3 ranked images from left to right, where matched images are with green boxes and mismatched images are with red boxes.

Conclusion

(1) We improve the VSRN by additionally modeling the local correspondences between regions and words for finegrained image-text matching.

(2) We propose an effective learning strategy to balance the relative importance of global alignment and local correspondences, which can well exploit their complementary properties.

(3) Our model achieves the state-of-the-art performance on the task of the image-text matching on MS-COCO and Flickr30K datasets.

This paper has been accepted by ICPR2020.

REFERENCE:

[1] J. R. K. Fartash Faghri, David J Fleet and S. Fidler, “Vse++: Improving visual-semantic embeddings with hard negatives,” in BMVC, 2018.
[2] G. H. H. Kuang-Huei Lee, Xi Chen and X. He, “Stacked cross attention for image-text matching,” in ECCV, 2018.
[3] e. a. Anderson, Peter, “Bottom-up and top-down attention for image captioning and visual question answering,” in CVPR, 2018.
[4] M. Schuster and K. K. Paliwal, “Bidirectional recurrent neural networks,” in IEEE Transactions on Signal Processing 45.11 (1997): 2673- 2681.
[5] K. L. Y. L. Kunpeng Li, Yulun Zhang and Y. Fu, “Visual semantic reasoning for image-text matching,” in ICCV, 2019.
[6] J. D. R. M. T. D. Subhashini Venugopalan, Marcus Rohrbach and K. Saenko, “Sequence to sequence-video to text,” in ICCV, 2015.
[7] e. a. Frome, Andrea, “Devise: A deep visual-semantic embedding model,” in NIPS, 2013.
[8] J. A. F.-F. L. Karpathy, A., “Deep fragment embeddings for bidirectional image sentence mapping,” in NIPS, 2014.
[9] T. N. Kipf and M. Welling, “Semi-supervised classification with graph convolutional networks,” in ICLR, 2017.