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

Hui Yuan, Yan Huang, Dongbo Zhang*,
Zerui Chen, Wenlong Cheng, and Liang Wang

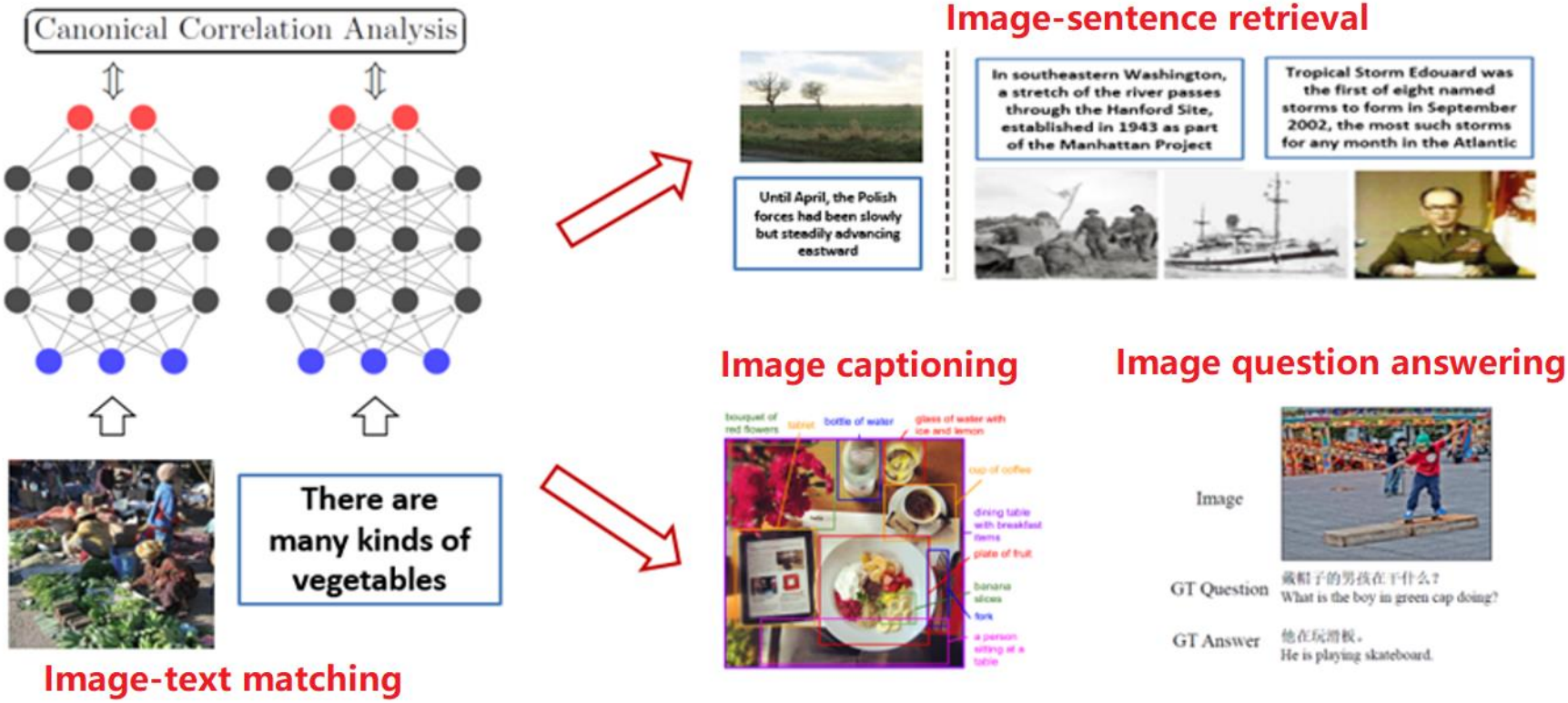
The College of Automation and Electronic Information, Xiangtan University, Xiangtan, China
Center for Research on Intelligent Perception and Computing (CRIPAC)
National Laboratory of Pattern Recognition (NLPR)
Institute of Automation, Chinese Academy of Sciences (CASIA)

Outline

- **Introduction**
- **Method**
- **Experiments**

➢Introduction

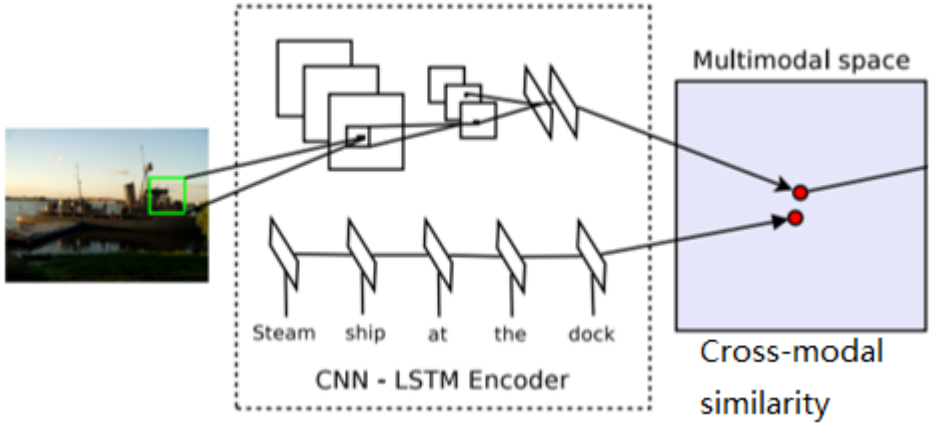
Image-Text Matching



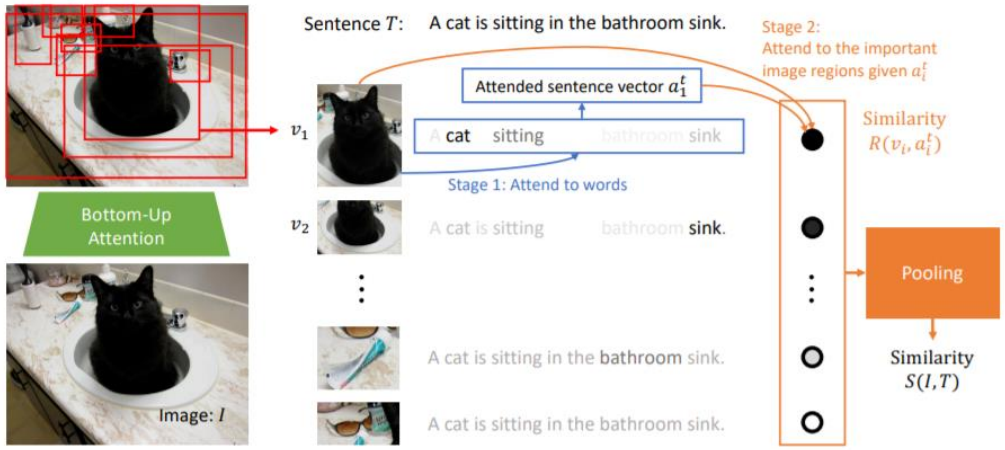
two retrieval sub-tasks:

- ✓ image annotation: given an image query to find mached texts
- ✓ image search :given a text query to retrieve matched images

Motivation



Coarse-grained image-text matching

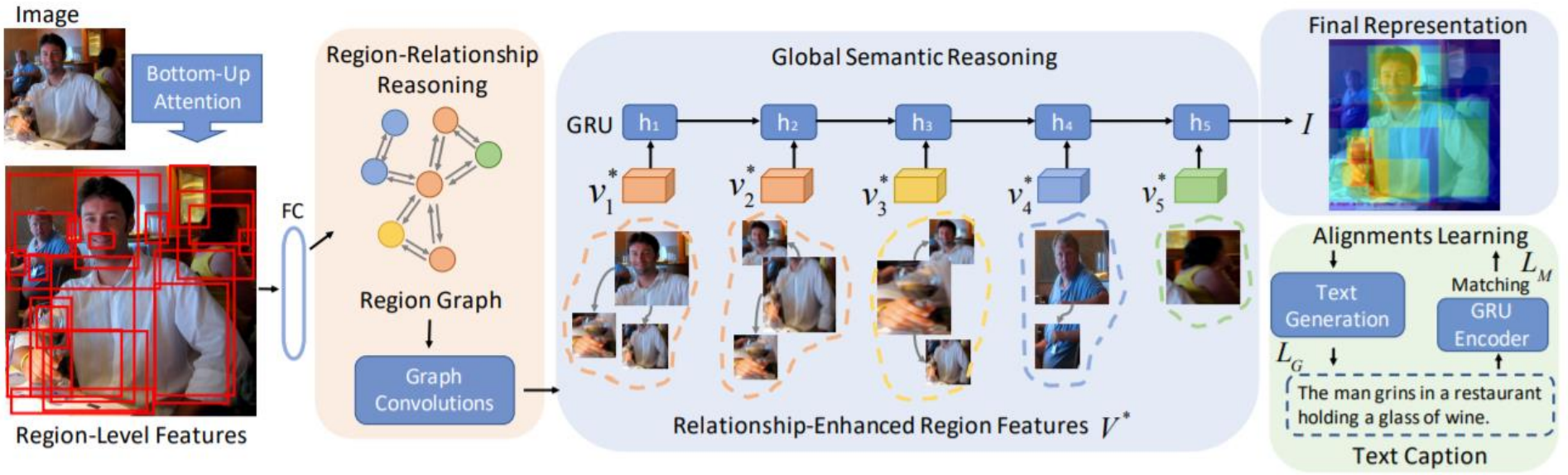


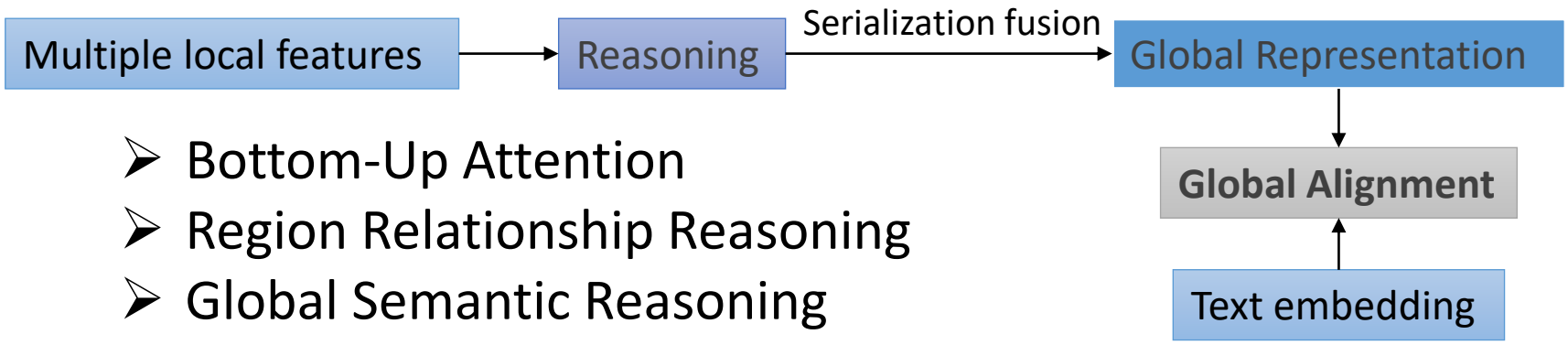
Fine-grained image-text matching

Incorporating the complementary advantages of global alignment and local correspondence, as well as balancing their relative importance.

➢Method

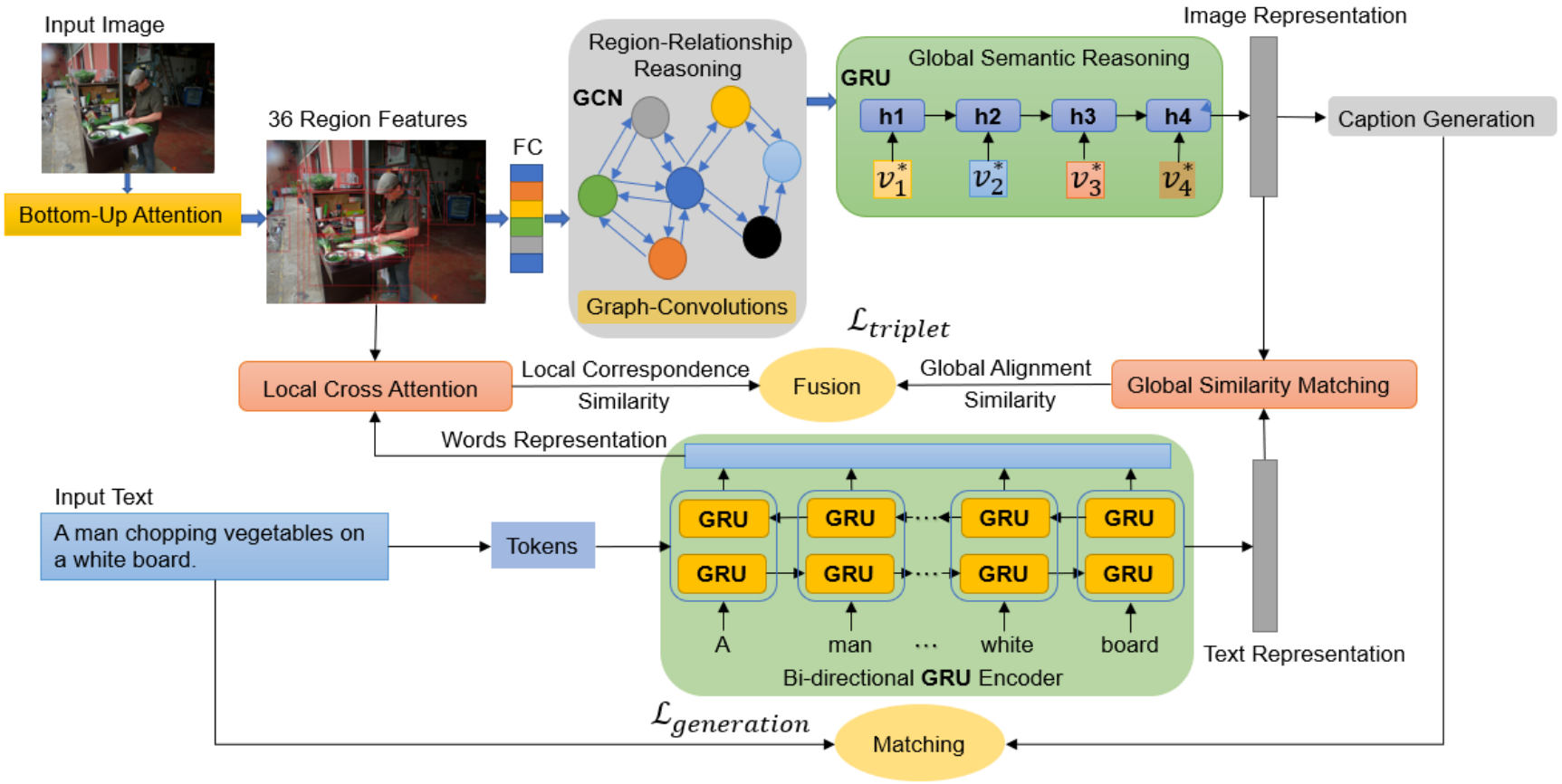
Visual Semantic Reasoning Network(VSRN)





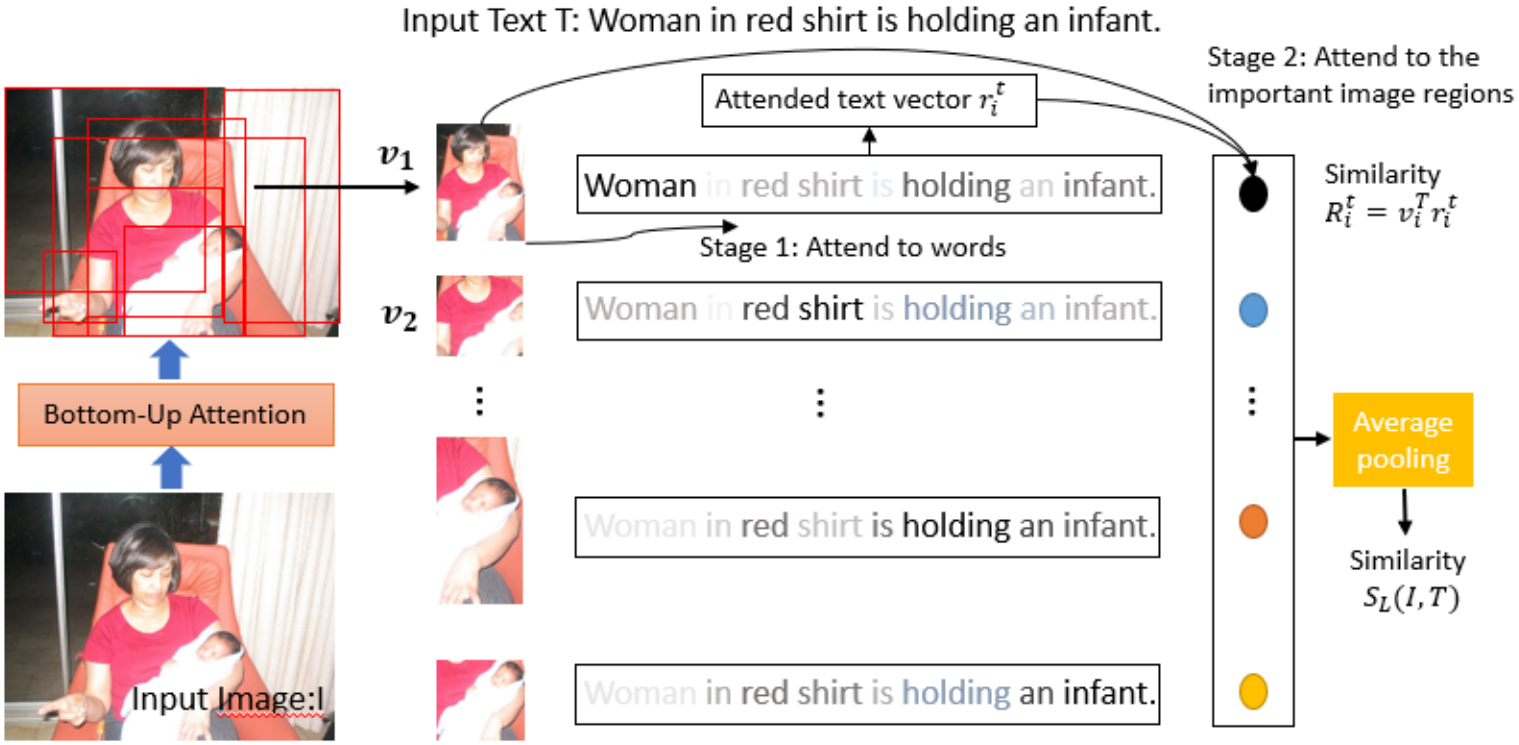
- Bottom-Up Attention
- Region Relationship Reasoning
- Global Semantic Reasoning
- GRU Encoder Text:
- Text Generation Modual

Ours(VSR++) Framework



- Incorporating the complementary advantages of global alignment and local correspondence in our VSR++ method.
- develop a suitable learning strategy to balance their relative importance.

Local Cross-Modal Attention



$$s_{ij} = v_i^T e_j, i\epsilon[1,k], j\epsilon[1,n], \quad (1)$$

$$w_{ij} = softmax(\lambda \hat{s}_{ij}) \quad (2)$$

$$r_i^t = \sum_{j=1}^{n} w_{i,j}\, e_j \quad (3)$$

$$R_i^t = v_i^T\, r_i^t, \quad (4)$$

$$S_L(I,T) = \frac{\sum_{i=1}^{k} norm(R_i^t)}{k} \quad (5)$$

Model Learning Strategy

- Global-local similarity fusion

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

- triplet ranking loss cross-modal learning

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

Caption generation:

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

maximize the log-likelihood of the predicted output caption.

- Our final loss function: $\mathcal{L} = \mathcal{L}_{triplet} + \mathcal{L}_{generation}$

➢ Experiments

Ablation Study

Table 1: Ablation studies on Flickr30k to investigate the effect of different network structures and different association ways. Results are reported in terms of recall@k(R@K).

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**

- Significantly improve the cross-modal similarity.
- Bi-GRU is better than GRU encoding.
- Similarity fusion is significantly better than other fusion ways.

Table 2: Ablation studies on Flickr30k to analyze the impact of different values of the association parameter μ between the global image-text alignment and local region-word correspondence.

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**

- The optimal fusion coefficient μ=2.0.

Comparisons with the SOTA

Table 3: The result of VSR++ on MS-COCO (1K test) dataset.

Methods	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**

R@1 **+0.4** **+0.6**

Table 4: The result of VSR++ on Flickr30K dataset.

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**

R@1 **+1.3** **+1.6**

Visualization Analysis

Image -> Text Retrieval

Query Image (a)

Method(a):VSRN(Global image-text alignment)
1: A very young child in a denim baseball cap eats a green apple. ✔
2: An infant wearing a hat, holding an apple in his right hand. ✔
3: A young child has a popsicle stick in his mouth. ✘

Method(b):VSR++(Global and Local Association)
1: A very young child in a denim baseball cap eats a green apple. ✔
2: A young boy in a blue hat and bib eating a green apple. ✔
3: An infant wearing a hat, holding an apple in his right hand. ✔

Query Image (b)

Method(a):VSRN(global image-text alignment)
1: Two young men, one in a white jersey and one in a red jersey, playing basketball. ✔
2: Two young men playing basketball , one defending the other attempting to make a basket. ✔
3: Two young boys playing in a dirt top playground. ✘

Method(b):VSR++(global and local association learning)
1: 3 basketball players vying for the ball and one in red jersey trying to take ball from guy in white jersey. ✔
2: A player in a white and red uniform goes for a layup , while the player with a red and black uniform blocks the attempted shot. ✔
3: Two young men, one in a white jersey and one in a red jersey, playing basketball. ✔

Text -> Image Retrieval

Query (a): A couple is sitting on the sand with their feet in the water , and they are shaking hands.

Query (b): Six people ride mountain bikes through a jungle environment.

Thank you!

Suggestion Questions

湘潭大学 XIANGTAN UNIVERSITY

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