Multilateral Semantic Relations Modeling for Image Text Retrieval

ZhengWang
School
of Computer Science and
Engineering
UESTC, Chengdu, SiChuan,
China

Zhenwei Gao
School
of Computer Science and
Engineering
UESTC, Chengdu,
SiChuan, China

Kangshuai Guo
School
of Computer Science and
Engineering
UESTC, Chengdu,
SiChuan, China

Yang Yang
School
of Computer Science and
Engineering
UESTC, Chengdu,
SiChuan, China

Yang Yang
School
of Computer Science and
Engineering
UESTC, Chengdu,
SiChuan, China

Heng Tao Shen
School
of Computer Science and
Engineering
UESTC, Chengdu,
SiChuan, China

Introduction

A man that is standing in the dirt with a bat.

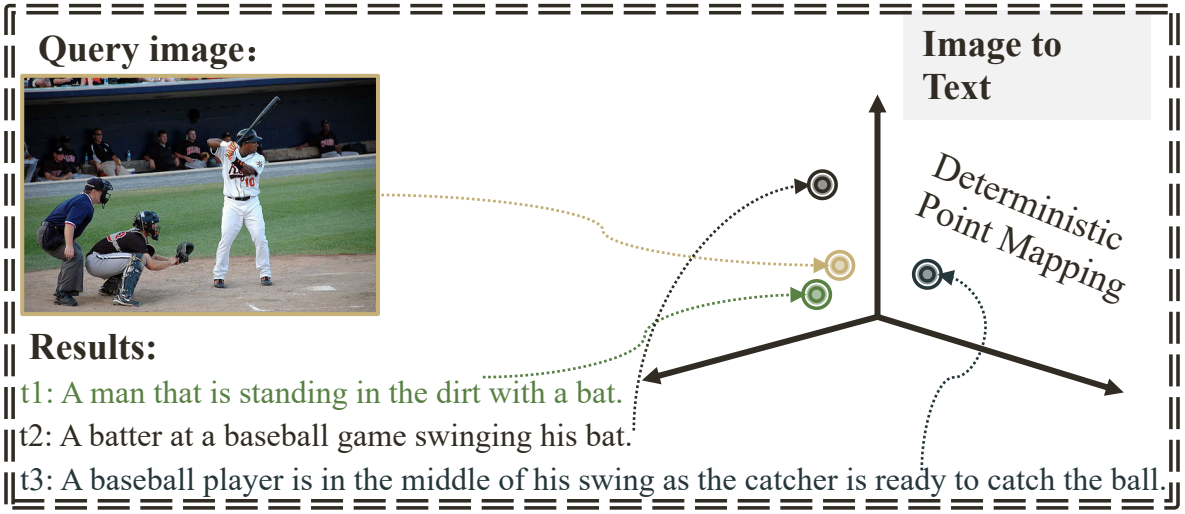
A batter at a baseball game swinging his bat.

A baseball player is in the middle of his swing as the catcher is ready to catch the ball.

Goal of cross-modal retrieval:

Learning embedding functions from **image / text** to a **shared embedding space**, where matching image-caption pairs are closer than non matching pairs in that space.

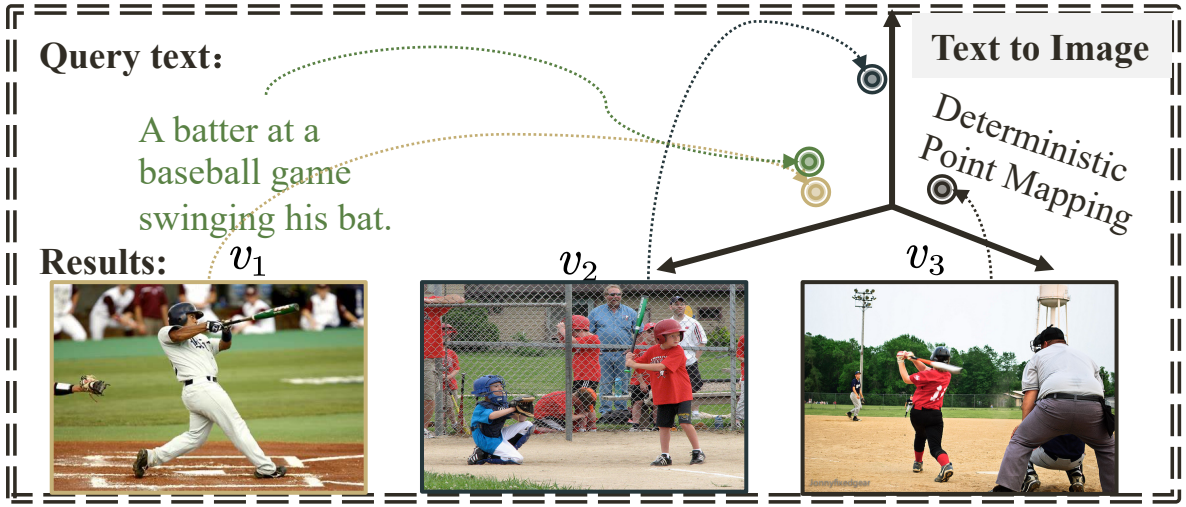
Existing Challenges



“One-to-many mapping” challenges in cross-modal retrieval tasks:

- An image can potentially be matched with a number of different captions.

Existing Challenges

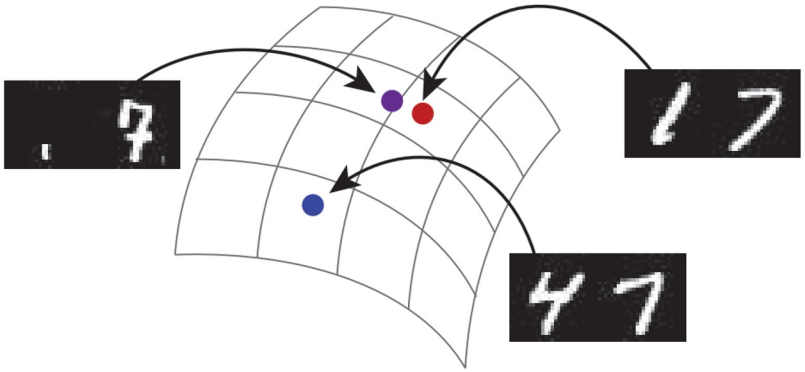


“One-to-many mapping” challenges in cross-modal retrieval tasks:

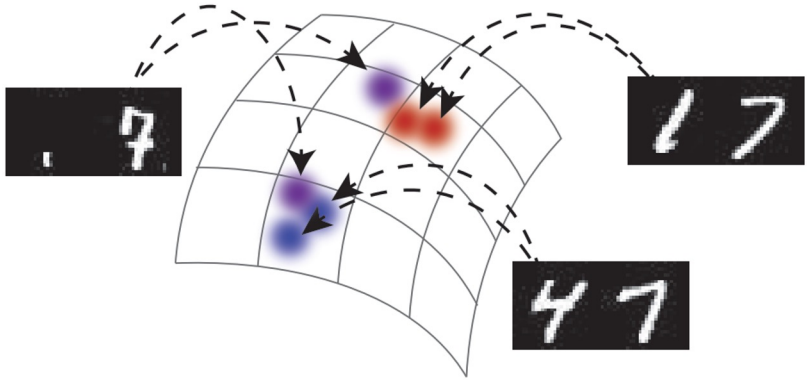
- A caption also semantically match more than one picture.

Probabilistic embedding

- Each embedding is a Gaussian distribution, instead of a point vector.
- Can handle “ambiguous” inputs.



(a) Point embedding.



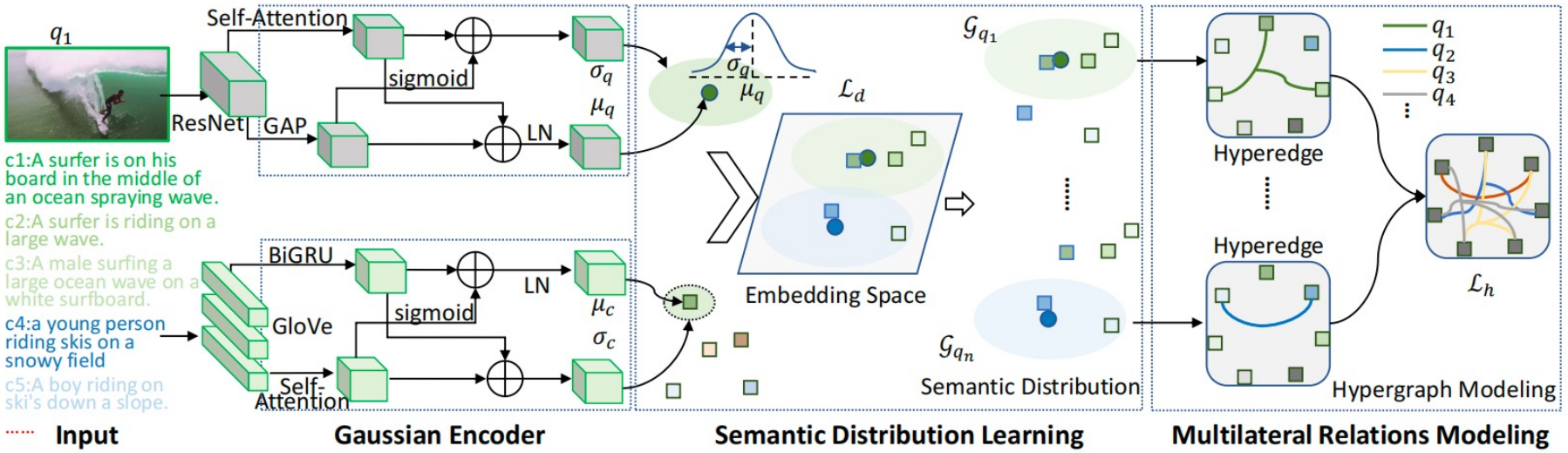
(b) Stochastic embedding.

ICLR 19, Modeling Uncertainty with Hedged Instance Embedding

Motivation

- A hyperedge can connect more than three nodes

Our Method


q_1
Self-Attention
sigmoid
ResNet
GAP
LN
σ_q
μ_q
c1:A surfer is on his board in the middle of an ocean spraying wave.
c2:A surfer is riding on a large wave.
c3:A male surfing a large ocean wave on a white surfboard.
c4:a young person riding skis on a snowy field
c5:A boy riding on ski's down a slope.
BiGRU
GloVe
Self-Attention
sigmoid
LN
μ_c
σ_c
$\mathcal{L}_d$
Embedding Space
$\mathcal{G}_{q_1}$
$\mathcal{G}_{q_n}$
Semantic Distribution
Hyperedge
Hyperedge
q_1
q_2
q_3
q_4
$\mathcal{L}_h$
Hypergraph Modeling
Input
Gaussian Encoder
Semantic Distribution Learning
Multilateral Relations Modeling

Evaluation Metrics:
Plausible Match R-Precision (PMRP[2]):

Query image

“Plausible” image

A group of planes sitting on a runway, in the day.

An outside view of airplanes and buildings at an airport.

The various airplanes are waiting for repairs at the terminals.

The view of runway from behind the windows of airport.

“Plausible” captions

CVPR 21, Probabilistic Embeddings for Cross-Modal Retrieval

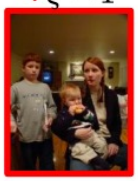
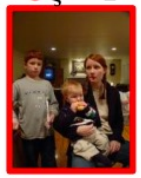
Experimental results

COCO Caption results

Methods	Dimension	1K Test				5K Test			
		Image-to-text		Text-to-image		Image-to-text		Text-to-image	
		PMRP	R@1	PMRP	R@1	PMRP	R@1	PMRP	R@1
VSE++ (BMVC'18) [8]	1024	-	64.60	-	52.00	-	41.30	-	30.30
PVSE M=1 (CVPR'19) [34]	1024	40.30	66.70	41.90	53.50	29.30	41.70	30.10	30.60
PVSE M=2 (CVPR'19) [34]	1024 × 2	42.80	69.20	43.70	55.20	31.80	45.20	32.00	32.40
VSRN (ICCV'19) [16]	2048	41.20	76.20	42.40	62.80	29.70	53.00	29.90	40.50
VSRN +AOQ (ECCV'20) [4]	2048 × 2	44.70	**77.50**	45.60	**63.50**	33.00	**55.10**	33.50	**41.10**
PCME μ only (CVPR'21) [6]	1024	45.00	68.00	45.90	54.60	34.00	43.50	34.30	31.70
PCME (CVPR'21) [6]	1024 × 2	45.10	68.80	46.00	54.60	34.10	44.20	34.40	31.90
PCME (CVPR'21)†	1024 × 2	45.10	65.90	46.00	53.30	34.10	41.70	34.40	31.20
P2RM (ACM MM'22) [41]	1024 × 2	45.90	66.60	46.42	54.22	35.52	42.12	35.11	31.50
MSRM (Ours)	1024 × 2	**46.43**	68.85	**47.35**	56.12	**35.62**	44.32	**35.81**	33.40

- Our methods shows the best PMRP scores among recent state-of-the-art COCO retrieval methods
- Although recent methods achieved impressive R@1 scores, their PMRP scores are much lower than us.

Retrieval Examples with our method and PCME

Query	Our MSRM	PCME
	1. A couple of men are loading a truck with glass. $\zeta=0$ 2. Many men work together to put objects in a truck. $\zeta=0$ 3. A man bending into the back of a truck on a street. $\zeta=1$ 4. A couple are approaching a man sitting down outside of a small shop. $\zeta=3$ 5. A man reaches in the back of a truck . $\zeta=0$ 6. A truck with a bunch of people in back of it. $\zeta=1$	1. A couple of men are loading a truck with glass. $\zeta=0$ 2. A man bending into the back of a truck on a street. $\zeta=1$ 3. A man reaches in the back of a truck. $\zeta=0$ 4. A couple are approaching a man sitting down outside of a small shop $\zeta=3$ 5. A man leaning over the back of a truck in front of buildings. $\zeta=3$ 6. Some people trying to load an item onto a motorcycle. $\zeta=3$
Two children play while eating in a restaurant.	1 GT $\zeta=0$  2 $\zeta=0$  3 $\zeta=0$  4 $\zeta=1$  5 $\zeta=1$  6 $\zeta=1$ 	1 $\zeta=1$  2 $\zeta=2$  3 GT $\zeta=0$  4 $\zeta=0$  5 $\zeta=3$  6 $\zeta=1$ 

Contribution

- We introduce an interpretable method named Multilateral Semantic Relations Modeling to better resolve the one-to-many correspondence for image-text retrieval.
- We propose the Semantic Distribution Learning module to extract the true semantics of a query based on Mahalanobis distance, which can infer more accurate multiple matches.
- We leverage the hyperedge convolution to model the high-order correlations between a Gaussian query and candidates for further improving the accuracy.

Reference

[1] ICLR 19, Modeling Uncertainty with Hedged Instance Embedding

[2] CVPR 21, Probabilistic Embeddings for Cross-Modal Retrieval(PCME)

[3] BMVC 18, VSE++: Improving Visual-Semantic Embeddings with Hard Negatives(VSE0)

[4] CVPR 19, Polysemous Visual-Semantic Embedding for Cross-Modal Retrieval(PVSE)

[5] ICCV 19, Visual Semantic Reasoning for Image-Text Matching(VSRN)

[6] ECCV 20, Adaptive Offline Quintuplet Loss for Image-Text Matching(VSRN+AOQ)