A Distributed Representation Based Query Expansion Approach for Image Captioning

Semih Yagcioglu, Erkut Erdem, Aykut Erdem, Ruket Çakıcı

Hacettepe University
Computer Vision Lab

Middle East Technical University
Department of Computer Engineering

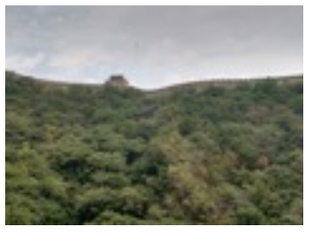










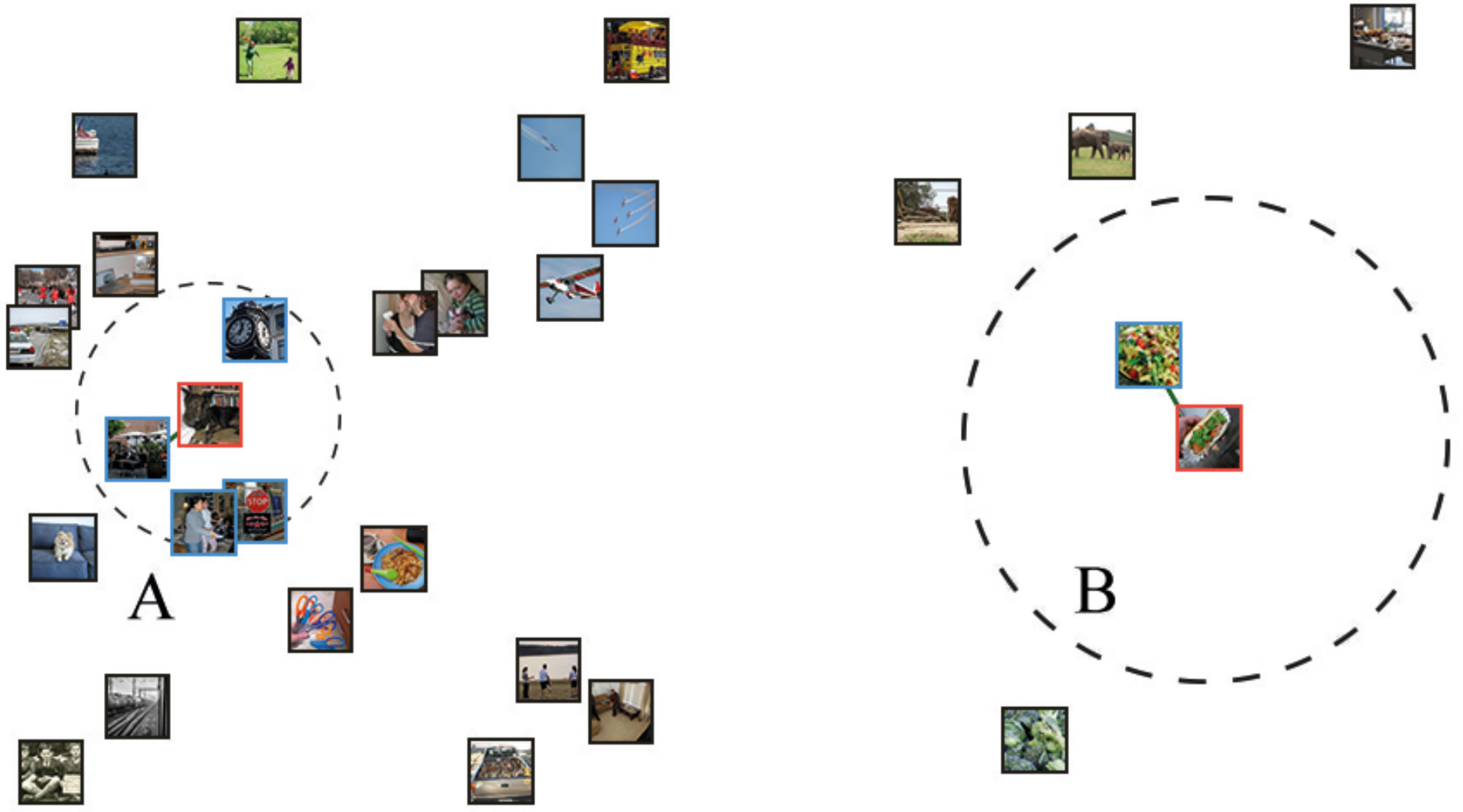




our approach

a simple data-driven **transfer based** approach
using **distributed representations**

image representation

- features from 16-layer VGG network (fc7)
- 4096 dimensions



$$\mathcal{N}(I_q) = \{(I_i, c_i) \mid dist(I_q, I_i) \leq (1+\epsilon) dist(I_q, I_{closest}),$$
$$I_{closest} = \arg\min\ dist(I_q, I_i), I_i \in \mathcal{T}\} \quad (1)$$

visual retrieval

and adaptive inlier selection

Query image I_q

Visually similar images

I_1

c_1: A man climbs up a snowy mountain.

I_2

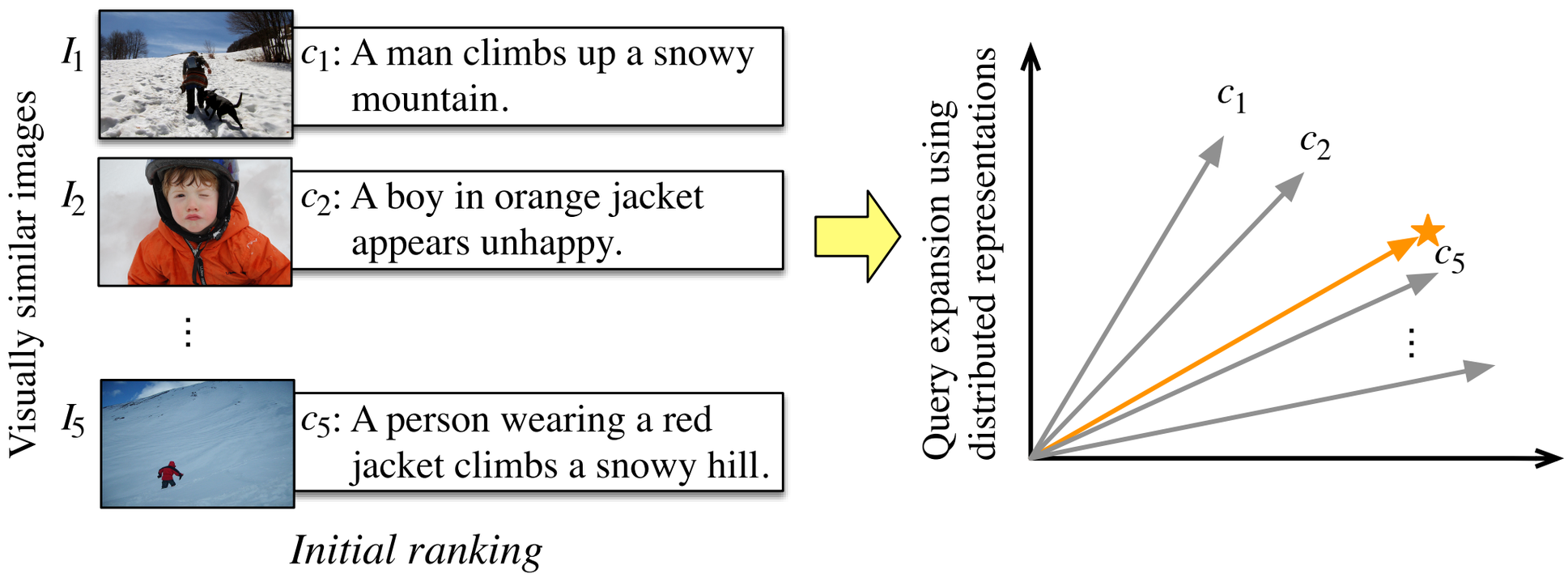
c_2: A boy in orange jacket appears unhappy.

⋮

I_5

c_5: A person wearing a red jacket climbs a snowy hill.

Initial ranking



our query expansion approach

swap modalities from the visual domain to a textual one

word representation

- *word2vec* model (Mikolov et al., 2013)
- *GloVe* model (Pennington et al., 2014)
  - word vectors, 500 dimensions
  - MS COCO captions as corpus (617K)

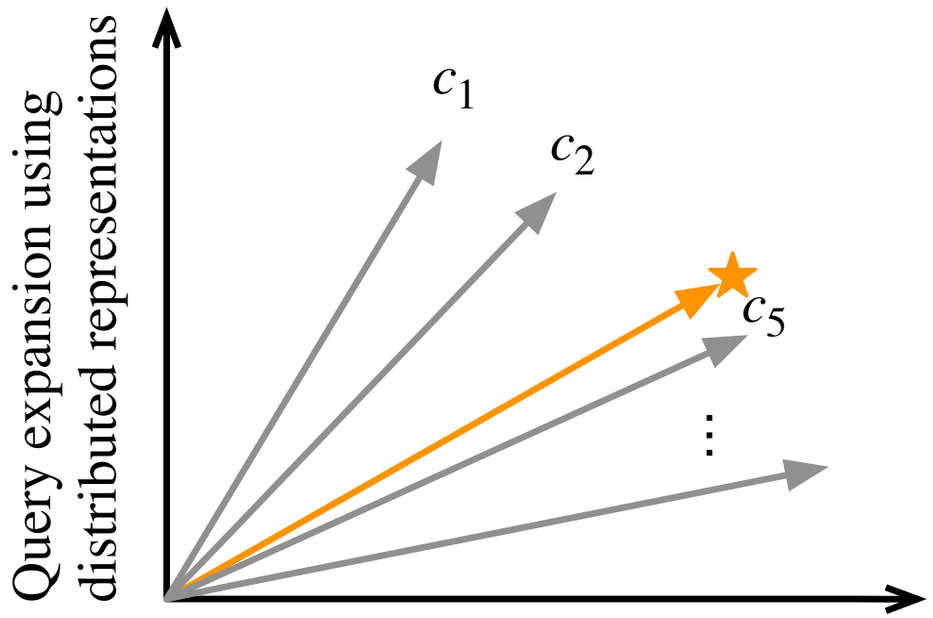
words to captions

- sum each word vector in a caption
- sentence vector c to represent captions

$$q = \frac{1}{NM} \sum_{i=1}^{N} \sum_{j=1}^{M} sim(I_q, I_i) \cdot c_i^j$$

calculating

the new textual query



transferred caption

c_5: A person wearing a red jacket climbs a snowy hill.

⋮

c_1: A man climbs up a snowy mountain.

c_2: A boy in orange jacket appears unhappy.

Final ranking

re-ranking

via cosine similarity

experimental setup

Dataset	# Images	# Captions
Flickr8K	8K	5
Flickr30K	30K	5
MS COCO	123K	5

the good, the bad and the ugly

results

a man in a black shirt and his little girl wearing orange are sharing a treat

a construction crew in orange vests
working near train tracks

a green bird perched on top of a tree filled with pink flowers

a white cat is sitting
in a bathroom sink

a boy is holding a dog
that is wearing a hat

a man wearing a santa hat holding a dog posing for a picture

a boy is holding a dog that is wearing a hat

quantitative evaluation

- VC (Ordonez et al. 2011)
- MC-KL, MC-SB (Mason and Charniak 2014)

- BLEU, METEOR, CIDEr

- Flickr8K, Flickr30K and MS COCO

quantitative evaluation

	Flickr8K			Flickr30K			MS COCO		
	BLEU	METEOR	CIDEr	BLEU	METEOR	CIDEr	BLEU	METEOR	CIDEr
OURS	**3.78**	**11.57**	**0.31**	**3.22**	**10.06**	0.20	**5.36**	**13.17**	**0.58**
MC-KL	2.71	10.95	0.15	2.02	9.92	0.07	4.04	12.56	0.37
MC-SB	3.08	9.06	0.27	2.76	8.06	**0.20**	5.02	11.78	0.56
VC	2.79	8.91	0.19	2.33	7.53	0.14	3.71	10.07	0.35
HUMAN	7.59	17.72	2.67	6.52	15.70	2.53	7.42	16.73	2.70

human evaluation

- rated for relevancy on a scale of 1 to 5
- Crowdflower with at least 5 annotators

	Rate	Variance
OURS	**2.73**	0.65
MC-SB	2.38	0.58
VC	2.27	0.66
MC-KL	2.03	0.62
HUMAN	4.84	0.26

concluding remarks

- a simple yet effective data-driven image captioning approach
- future work could focus on
  - other pooling approaches such as using Fisher vectors (Klein et al. 2015)
  - incorporating syntactic relations (Socher et al. 2015)
- source code will soon be available at
  - github.com/semihyagcioglu/image-captioning