A Distributed Representation Based Query Expansion Approach for Image Captioning

Semih Yagcioglu, Erkut Erdem, Aykut Erdem, Ruket Cakici

Hacettepe University Computer Vision Lab (HUCVL)

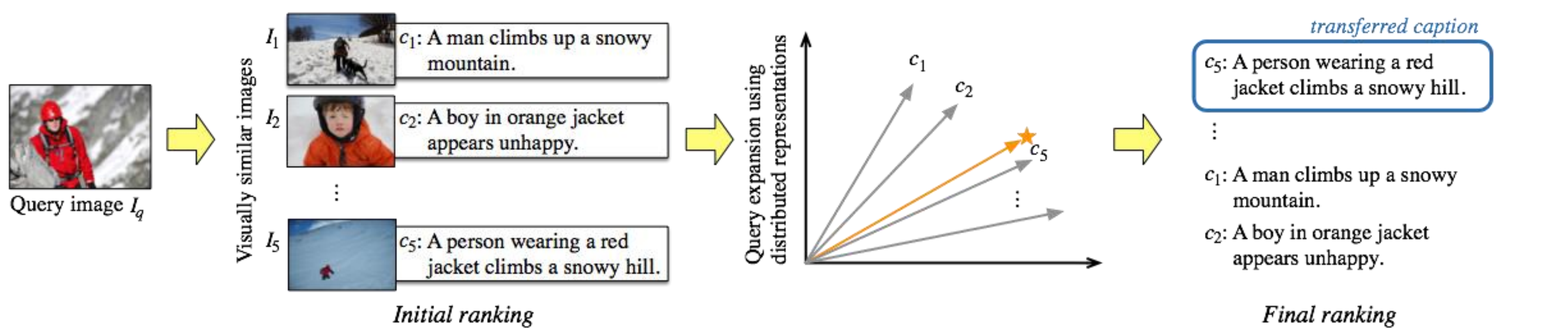
Dept. Of Computer Engineering, Hacettepe University, Ankara, TURKEY

Dept. of Computer Engineering, Middle East Technical University, Ankara, TURKEY

http://semihyagcioglu.com/projects/image-captioning

Introduction

In this study, we propose a **novel query expansion** approach for improving transfer based automatic image captioning. The core idea of our method is to translate the given visual query into a distributional semantics based form, which is generated by the average of the sentence vectors extracted from the captions of images visually similar to the input image. Using three image captioning benchmark datasets, we show that our approach provides more accurate results compared to the state-of-the art data-driven methods in terms of both automatic metrics and subjective evaluation.



A system overview of the proposed query expansion approach for image captioning.

Image Captioning

The aim of image captioning is to generate natural language descriptions for images.

A woman and a teenager is waiting near a busy street with their bicycles.

Contributions

- we take a new perspective to data-driven image captioning by proposing a novel query expansion step based on **compositional distributed semantics** to improve the results
- Our approach uses the weighted average of the distributed representations of retrieved captions to **expand the original query** in order to obtain captions that are **semantically more related** to the visual content of the input image.

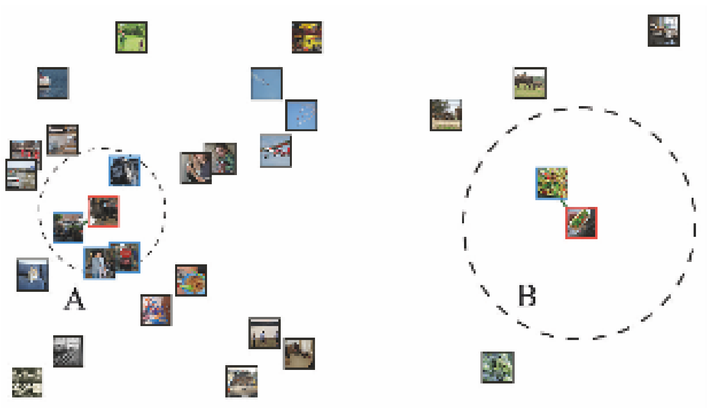
Method

We follow a simple approach as can be outlined follows:

- We first represent images with VGGNet architecture
- We retrieve visually similar images based on deep features
- Apply an adaptive neighborhood selection
- Compute distributed representations for captions
- We follow a query expansion from visual to textual domain
- We retrieve similar captions with distributed representations of captions
- By re-ranking we return the closest caption to describe the query image

Visual Retrieval

- We take the visual query and retrieve the visually similar images based on the L2 distance
- We further employ an **inlier selection approach** in which we select the neighbors adaptively based on the minimum distance of the closest neighbor



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

Query Expansion

- Our query expansion model on the distributional models of semantics where the meanings of words are represented with vectors that characterize the set of contexts they occur in a corpus.
- For a query image *Iq*, we first retrieve visually similar images from a large dataset of captioned images
- We swap modalities and **construct a new query** based on the distributed representations of captions. In particular, we expand the original visual query with a new textual query based on the weighted average of the vectors of the retrieved captions as follows:

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

- Here *N* and *M* respectively denote the total number of image-caption pairs in the candidate set *N* and the number of reference captions associated with each training image, and *sim(Iq , Ii)* refers to the visual similarity score of the image *Ii* to the query image *Iq* which is used to give more importance to the captions of images visually more close to the query image.

Results

Here are a few experimental results.

MC-KL	a black and white dog is playing or fighting with a brown dog in grass	a man is sitting on a blue bench with a blue blanket covering his face	a man in a white shirt and sunglasses is holding hands with a woman wearing a red shirt outside	one brown and black pigmented bird sitting on a tree branch
MC-SB	a dog looks behind itself	a girl looks at a woman s face	a woman and her two dogs are walking down the street	a tree with many leaves around it
VC	a brown and white dog jumping over a red yellow and white pole	a father feeding his child on the street	a girl is skipping across the road in front of a white truck	a black bear climbing a tree in forest area
OURS	a brown and white dog jumps over a dog hurdle	a man in a black shirt and his little girl wearing orange are sharing a treat	a girl jumps rope in a parking lot	a bird standing on a tree branch in a wooded area
HUMAN	a brown and white sheltie leaping over a rail	a man and a girl sit on the ground and eat	a girl is in a parking lot jumping rope	a painted sign of a blue bird in a tree in the woods

	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

Conclusion

In this study, we present a novel query expansion approach for image captioning, in which we utilize a distributional model of meaning for sentences. Extensive experimental results on three well-established benchmark datasets have demonstrated that our approach outperforms the state-of- the art data-driven approaches.

Contact:

Semih Yagcioglu

Dept. of Computer Engineering, Hacettepe University

Email: semih.yagcioglu@hacettepe.edu.tr

Acknowledgements

This study was supported in part by The Scientific and Technological Research Council of Turkey (TUBITAK), with award no 113E116.

Citation

- S. Yagcioglu, E. Erdem, A. Erdem, R. Çakıcı. A Distributed Representation Based Query Expansion Approach for Image Captioning. The 53rd Annual Meeting of the Association for Computational Linguistics and The 7th International Joint Conference of the Asian Federation of Natural Language Processing (ACL-IJCNLP 2015), Beijing, China, July 2015.