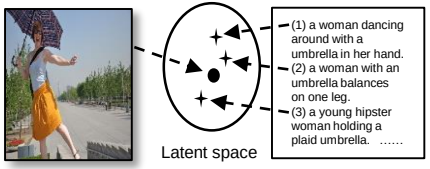
Learning a Recurrent Residual Fusion Network for Multimodal Matching

Yu Liu, Yanming Guo, Erwin M. Bakker, Michael S. Lew
LIACS Media Lab, Leiden University, The Netherlands

International Conference on Computer Vision 2017

Introduction

Background: Image and text matching is an important task in the computer vision field. It remains challenging due to the heterogeneous representations.



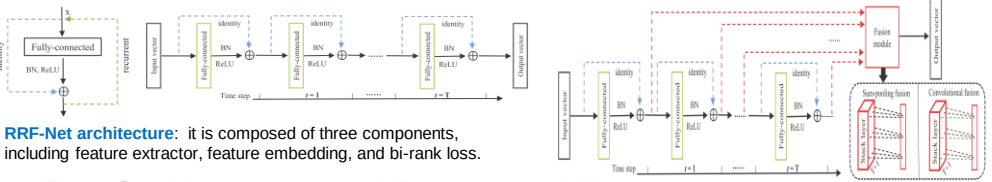
Goal: How to better match images and texts, while retaining the parameters?

Key ideas:

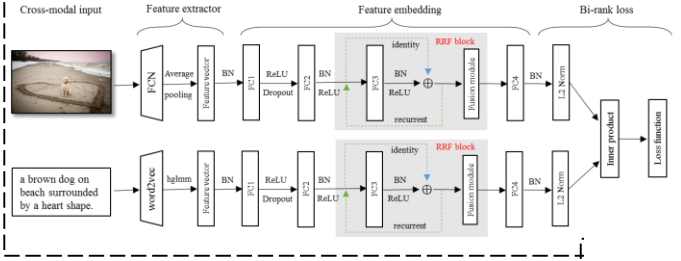
- We propose a deep matching network using recurrent residual fusion (RRF) as building blocks for improving feature embeddings between image and text.
- We employ a bi-rank loss function to enforce separability of two modalities in the embedding space.

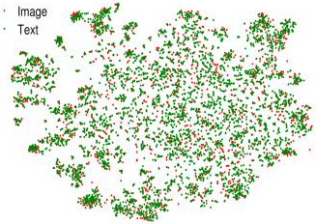
Method

RRF building block: identity connections, recurrent connections and a fusion module.



RRF-Net architecture: it is composed of three components, including feature extractor, feature embedding, and bi-rank loss.





Results

Ablation study: present comprehensive results on Flickr30K.

(1) Effect of recurrent steps.

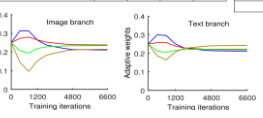
Method	Image to Text		Text to Image	
	R@1	R@5	R@1	R@5
Baseline	45.0	75.5	33.6	66.5
RRF-Net, T=1	46.4	76.1	34.3	67.3
RRF-Net, T=2	46.9	76.8	34.8	67.7
RRF-Net, T=3	**47.6**	**77.4**	**35.4**	**68.3**
RRF-Net, T=4	46.2	76.6	35.1	67.6

(2) Effect of fusion modules.

Method	Image to Text		Text to Image	
	R@1	R@5	R@1	R@5
RRF-Net w/o fusion module	45.8	75.9	34.2	67.1
RRF-Net with sum fusion	47.1	76.8	35.0	67.6
RRF-Net with conv fusion	47.6	77.4	35.4	68.3

(3) Effect of bi-rank loss.

Method	Image to Text		Text to Image	
	R@1	R@5	R@1	R@5
Baseline, bi-directional	43.4	73.8	32.5	65.4
Baseline, bi-rank	45.0	75.5	33.6	66.5
RRF-Net, bi-directional	46.4	76.5	34.1	67.4
RRF-Net, bi-rank	47.6	77.4	35.4	68.3



Comparison with the state-of-the-art approaches.

Method	Flickr30K dataset						MSCOCO dataset					
	Image to Text			Text to Image			Image to Text			Text to Image		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
DVSA [15]	22.2	48.2	61.4	15.2	37.7	50.5	38.4	69.9	80.5	27.4	60.2	74.8
SC-NLM [17]	23.0	50.7	62.9	16.8	42.0	56.5	NA	NA	NA	NA	NA	NA
Mean vector [18]	24.8	52.5	64.3	20.5	46.3	59.3	33.2	61.8	75.1	24.2	56.4	72.4
Deep CCA [42]	27.9	56.9	68.2	26.8	52.9	66.9	NA	NA	NA	NA	NA	NA
GMM+HGLMM [18]	35.0	62.0	73.8	25.0	52.7	66.0	39.4	67.9	80.9	25.1	59.8	76.6
m-RNN [27]	35.4	63.8	73.7	22.8	50.7	63.1	41.0	73.0	83.5	29.0	42.2	77.0
RNN-FV [20]	35.6	62.5	74.2	27.4	55.9	70.0	41.5	72.0	82.9	29.2	64.7	80.4
mCNN(ensemble) [25]	33.6	64.1	74.9	26.2	56.3	69.6	42.8	73.1	84.1	32.6	68.6	82.8
DSPE [39]	40.3	68.9	79.9	29.7	60.1	72.1	50.1	79.7	89.2	39.6	75.2	86.9
2WayNet [4]	**49.8**	67.5	NA	**36.0**	55.6	NA	55.8	75.2	NA	39.7	63.3	NA
RRF-Net	47.6	**77.4**	**87.1**	35.4	**68.3**	**79.9**	**56.4**	**85.3**	**91.5**	**43.9**	**78.1**	**88.6**

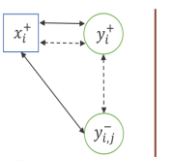
Model ensemble

Method	Image to Text		Text to Image	
	R@1	R@5	R@1	R@5
RRF-Net, $M = \{3\}$	47.6	77.4	35.4	68.3
RRF-Net, $M = \{1, 3\}$	49.1	78.4	36.8	69.8
RRF-Net, $M = \{1, 2, 3\}$	50.3	79.2	37.4	70.4
RRF-Net, $M = \{1, 2, 3, 4\}$	**50.8**	**79.5**	**37.6**	**70.9**

Cross-dataset generalization

Cross-dataset	Method	Image to Text			Text to Image		
		R@1	R@5	R@10	R@1	R@5	R@10
Train: Flickr30K	Baseline	23.7	52.0	64.0	17.9	42.8	57.4
Test: MSCOCO	RRF-Net, T=3	24.8	53.0	64.8	18.8	44.1	58.5
	RRF-Net, M={1,2,3,4}	**26.4**	**54.3**	**66.2**	**20.3**	**45.6**	**59.8**
Train: MSCOCO	Baseline	27.2	52.6	65.1	20.4	41.5	52.8
Test: Flickr30K	RRF-Net, T=3	28.8	53.8	66.4	21.3	42.7	53.7
	RRF-Net, M={1,2,3,4}	**31.5**	**56.0**	**68.6**	**23.4**	**45.0**	**56.8**

Bi-rank loss: it computes both image-to-text rank loss (Left) and text-to-image rank loss (Right).



$$l_{i2t} = \sum_{j=1}^{N} \Big(\alpha_1 \max[0, s(x_i^+, y_i^+) - s(x_i^+, y_{i,j}^-) + m] + \alpha_2 \max[0, s(x_i^+, y_i^+) - s(y_i^+, y_{i,j}^-) + m] \Big)$$

$$l_{t2i} = \sum_{j=1}^{N} \Big(\alpha_1 \max[0, s(y_i^+, x_i^+) - s(y_i^+, x_{i,j}^-) + m] + \alpha_2 \max[0, s(y_i^+, x_i^+) - s(x_i^+, x_{i,j}^-) + m] \Big)$$

$$l(x_i^+, y_i^+, X_i^-, Y_i^-) = \frac{\beta_1 l_{i2t} + \beta_2 l_{t2i}}{N}$$

References

[4] A. Eisenschtat et al. Linking image and text with 2-way nets. In CVPR, 2017.
[15] L. Ma et al. Multimodal convolutional neural networks for matching image and sentence. In ICCV, 2015..
[18] B. Klein et al. Associating neural word embeddings with deep image representations using fisher vectors. In CVPR, 2015.
[39] L. Wang et al. Learning deep structure-preserving image-text embeddings. In CVPR, 2016.

Codes are available