Multimodal Convolutional Neural Networks for Matching Image and Sentence

Lin Ma, Zhengdong Lu, Lifeng Shang, and Hang Li

Noah's Ark Lab, Huawei Technologies

http://mcnn.noahlab.com.hk/project.html

Contributions

- ★ ***m*-CNNs summarize the image, compose words of the sentence into different semantic fragments, and learn the matching relations and interactions between image and the composed fragments.**
- ★ ***m*-CNNs fully exploit complicated matching relations between image and sentence by letting image and the composed fragments of the sentence meet and interact at different levels.**
- ★ ***m*-CNNs outperform the state-of-the-art approaches for bidirectional image and sentence retrieval on the Flickr8K, Flickr30K, and Microsoft COCO datasets.**

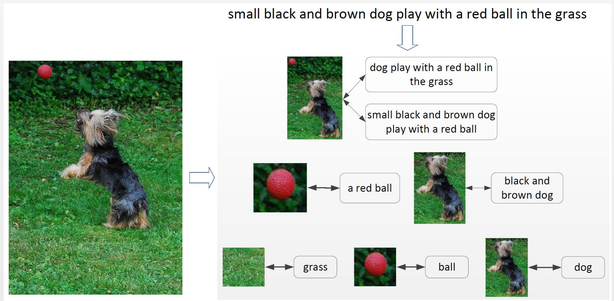
1: Matching between image and sentence

Playing an essential role in many image and sentence applications

- Image retrieval with sentence
- Image annotation with sentence

Occurring at different levels

- Word-level relations
- Phrase-level relations
- Sentence-level relations



2: Multimodal Convolutional Neural Network (*m*-CNN)

•**Image CNN**: Encoding image content

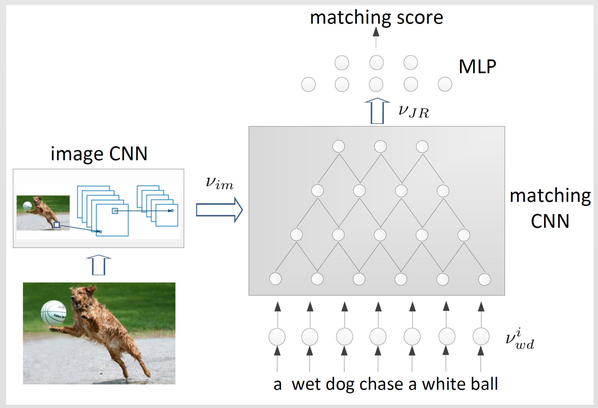
$$\nu_{im} = \sigma(\mathbf{w}_{im}(CNN_{im}(I)) + b_{im})$$

Matching CNN: learning the joint representation of image and sentence

- Composing words to different semantic fragments
- Learning interactions between image and word compositions

MLP: Summarizing the matching score

$$\mathbf{s}_{match} = \mathbf{w}_s\big(\sigma(\mathbf{w}_h(\nu_{JR}) + b_h)\big) + b_s$$



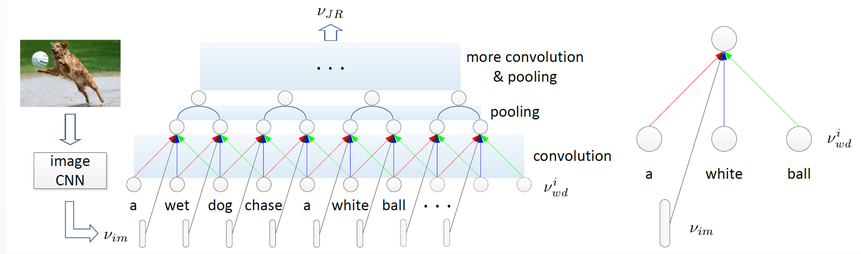
3: Matching CNN

Word-level matching CNN: image meets the word-level fragments of sentence:

- **Convolution**: composing higher semantic representations from words and learning interaction between image and word
- **Gating**: eliminating unexpected matching noises from convolution

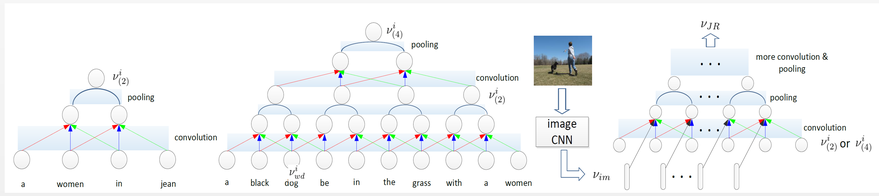
$$\nu^i_{(\ell,f)} = g(\vec{\nu}^i_{(\ell-1)}) \cdot \sigma(\mathbf{w}_{(\ell,f)}\vec{\nu}^i_{(\ell-1)} + b_{(\ell,f)}),\ g(x) = \begin{cases} 0, & x == \mathbf{0} \\ 1, & \text{otherwise} \end{cases}$$

- **Max-pooling**: filtering out unreliable compositions



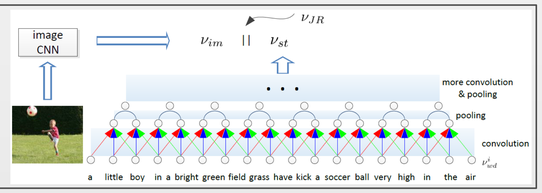
Phrase-level matching CNN: deferring the interactions until some higher semantic representations composed from words by convolution and max-pooling processes.

- Short phrase: objects and their relationships
- Long phrase: objects, their activities, and their relative positions



Sentence-level matching CNN: deferring the matching until the sentence is fully represented.

- Sentence CNN: three layers of convolution and max-pooling
- Multimodal layer: concatenating the image and sentence representations



Training: *m*-CNNs are trained with contrastive sampling using a ranking loss function.

$$e_\theta(x_n, y_n, y_m) = \max\big(0, \mu - \mathbf{s}_{match}(x_n, y_n) + \mathbf{s}_{match}(x_n, y_m)\big)$$

4: Experiment Results

Bidirectional image and sentence retrieval on **Flickr8K**

	Sentence Retrieval				Image Retrieval			
	R@1	R@5	R@10	Med r	R@1	R@5	R@10	Med r
Random Ranking	0.1	0.6	1.1	631	0.1	0.5	1.0	500
DeViSE	4.8	16.5	27.3	28.0	5.9	20.1	29.6	29
SDT-RNN	6.0	22.7	34.0	23.0	6.6	21.6	31.7	25
MNLM	13.5	36.2	45.7	13	10.4	31.0	43.7	14
MNLM-VGG	18.0	40.9	55.0	8	12.5	37.0	51.5	10
m-RNN	14.5	37.2	48.5	11	11.5	31.0	42.4	15
Deep Fragment	12.6	32.9	44.0	14	9.7	29.6	42.5	15
RVP (T)	11.6	33.8	47.3	11.5	11.4	31.8	45.8	12.5
RVP (T+I)	11.7	34.8	48.6	11.2	11.4	32.0	46.2	11
DVSA (DepTree)	14.8	37.9	50.0	9.4	11.6	31.4	43.8	13.2
DVSA (BRNN)	16.5	40.6	54.2	7.6	11.8	32.1	44.7	12.4
NIC	20.0	*	61.0	6	19.0	*	**64.0**	**5**
OverFeat :								
m-CNN$_{wd}$	8.6	26.8	38.8	18.5	8.1	24.7	36.1	20
m-CNN$_{phs}$	10.5	29.4	41.7	15	9.3	27.9	39.6	17
m-CNN$_{phl}$	10.7	26.5	38.7	18	8.1	26.6	37.8	18
m-CNN$_{st}$	10.6	32.5	43.6	14	8.5	27.0	39.1	18
m-CNN$_{ENS}$	14.9	35.9	49.0	11	11.8	34.5	48.0	11
VGG :								
m-CNN$_{wd}$	15.6	40.1	55.7	8	14.5	38.2	52.6	9
m-CNN$_{phs}$	18.0	43.5	57.2	8	14.6	39.5	53.8	9
m-CNN$_{phl}$	16.7	43.0	56.7	7	14.4	38.6	52.2	9
m-CNN$_{st}$	18.1	44.1	57.9	7	14.6	38.5	53.5	9
m-CNN$_{ENS}$	**24.8**	**53.7**	**67.1**	**5**	**20.3**	**47.6**	61.7	**5**

Bidirectional image and sentence retrieval on **Flickr30K**

	Sentence Retrieval				Image Retrieval			
	R@1	R@5	R@10	Med r	R@1	R@5	R@10	Med r
Random Ranking	0.1	0.6	1.1	631	0.1	0.5	1.0	500
DeViSE	4.5	18.1	29.2	26	6.7	21.9	32.7	25
SDT-RNN	9.6	29.8	41.1	16	8.9	29.8	41.1	16
MNLM	14.8	39.2	50.9	10	11.8	34.0	46.3	13
MNLM-VGG	23.0	50.7	62.9	5	16.8	42.0	56.5	8
m-RNN	18.4	40.2	50.9	10	12.6	31.2	41.5	16
m-RNN-VGG	**35.4**	63.8	73.7	**3**	22.8	50.7	63.1	5
Deep Fragment	14.2	37.7	51.3	10	10.2	30.8	44.2	14
RVP (T)	11.9	25.0	47.7	12	12.8	32.9	44.5	13
RVP (T+I)	12.1	27.8	47.8	11	12.7	33.1	44.9	12.5
DVSA (DepTree)	20.0	46.6	59.4	5.4	15.0	36.5	48.2	10.4
DVSA (BRNN)	22.2	48.2	61.4	4.8	15.2	37.7	50.5	9.2
NIC	17.0	*	56.0	7	17.0	*	57.0	7
LRCN	*	*	*	*	17.5	40.3	50.8	9
OverFeat :								
m-CNN$_{wd}$	12.7	30.2	44.5	14	11.6	32.1	44.2	14
m-CNN$_{phs}$	14.4	38.6	49.6	11	12.4	33.3	44.7	14
m-CNN$_{phl}$	13.8	38.1	48.5	11.5	11.6	32.7	44.1	14
m-CNN$_{st}$	14.8	37.9	49.8	11	12.5	32.8	44.2	14
m-CNN$_{ENS}$	20.1	44.2	56.3	8	15.9	40.3	51.9	9.5
VGG :								
m-CNN$_{wd}$	21.3	53.2	66.1	5	18.2	47.2	60.9	6
m-CNN$_{phs}$	25.0	54.8	66.8	4.5	19.7	48.2	62.2	6
m-CNN$_{phl}$	23.9	54.2	66.0	5	19.4	49.3	62.4	6
m-CNN$_{st}$	27.0	56.4	70.1	4	19.7	48.4	62.3	6
m-CNN$_{ENS}$	33.6	**64.1**	**74.9**	**3**	**26.2**	**56.3**	**69.6**	**4**

Bidirectional image and sentence retrieval on **Microsoft COCO**

	Sentence Retrieval				Image Retrieval			
	R@1	R@5	R@10	Med r	R@1	R@5	R@10	Med r
Random Ranking	0.1	0.6	1.1	631	0.1	0.5	1.0	500
m-RNN-vgg	41.0	73.0	83.5	2	29.0	42.2	77.0	3
DVSA	38.4	69.9	80.5	**1**	27.4	60.2	74.8	3
STV (uni-skip)	30.6	64.5	79.8	3	22.7	56.4	71.7	4
STV (bi-skip)	32.7	67.3	79.6	3	24.2	57.1	73.2	4
STV (combine-skip)	33.8	67.7	82.1	3	25.9	60.0	74.6	4
FV (Mean Vec)	33.2	61.8	75.1	3	24.2	56.4	72.4	4
FV (GMM)	39.0	67.0	80.3	3	24.2	59.2	76.0	4
FV (LMM)	38.6	67.8	79.8	3	25.0	59.5	76.1	4
FV (HGLMM)	37.7	66.6	79.1	3	24.9	58.8	76.5	4
FV (GMM+HGLMM)	39.4	67.9	80.9	2	25.1	59.8	76.6	4
VGG :								
m-CNN$_{wd}$	34.1	66.9	79.7	3	27.9	64.7	80.4	3
m-CNN$_{phs}$	34.6	67.5	81.4	3	27.6	64.4	79.5	3
m-CNN$_{phl}$	35.1	67.3	81.6	2	27.1	62.8	79.3	3
m-CNN$_{st}$	38.3	69.6	81.0	2	27.4	63.4	79.5	3
m-CNN$_{ENS}$	**42.8**	**73.1**	**84.1**	2	**32.6**	**68.6**	**82.8**	**3**

Contact: Dr. Lin Ma (forest.linma@gmail.com)