SCA-CNN: Spatial and Channel-wise Attention in Convolutional Networks for Image Captioning

Long Chen[1], Hanwang Zhang[2], Jun Xiao[1], Liqiang Nie[3], Jian Shao[1], Wei Liu[4], Tat-Seng Chua[5]

[1]Zhejiang University, [2]Columbia University, [3]Shandong University, [4]Tencent AI Lab, [5]National University of Singapore

IEEE 2017 Conference on Computer Vision and Pattern Recognition

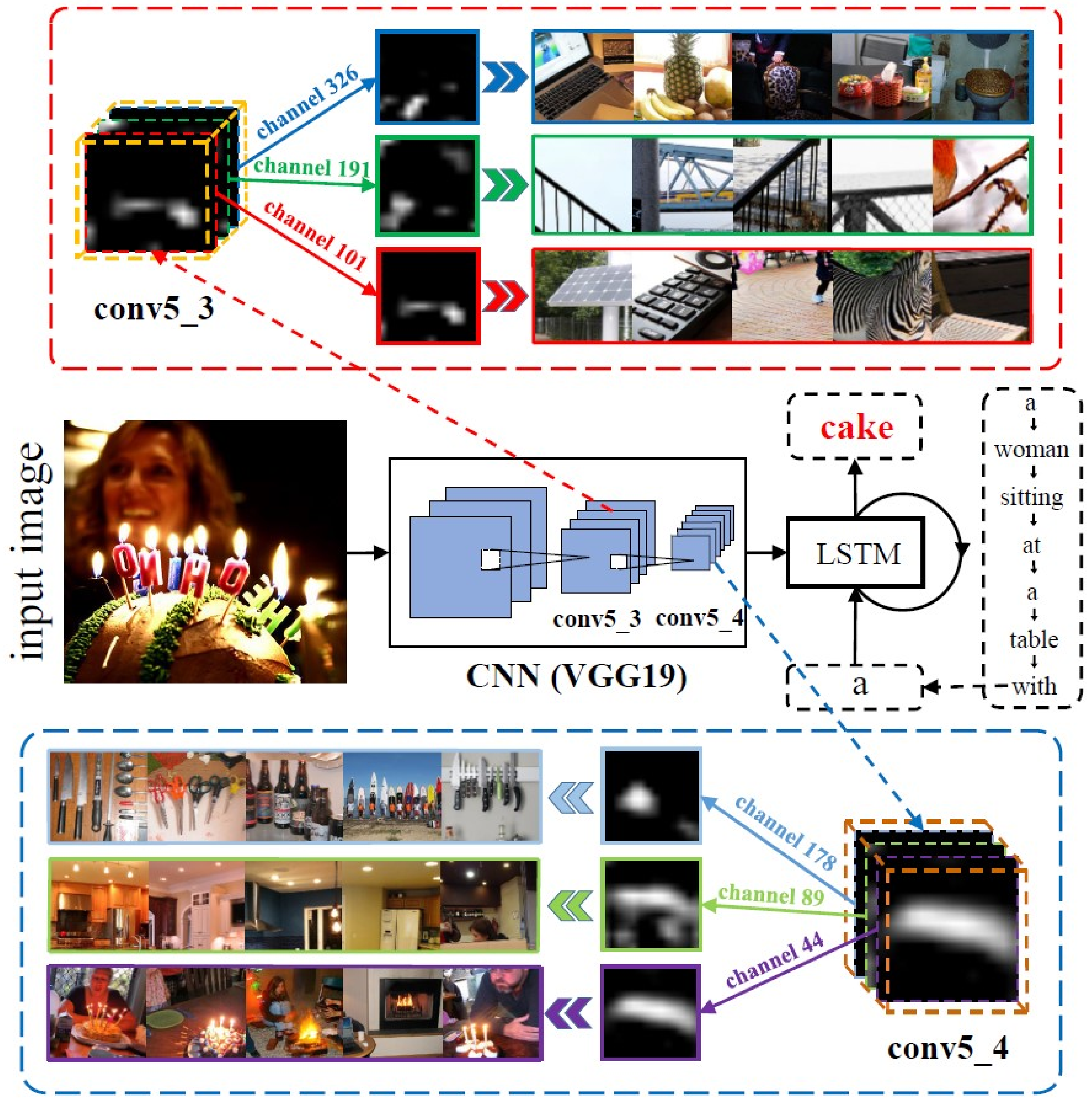
CVPR July 21-26 HONOLULU 2017

Introduction

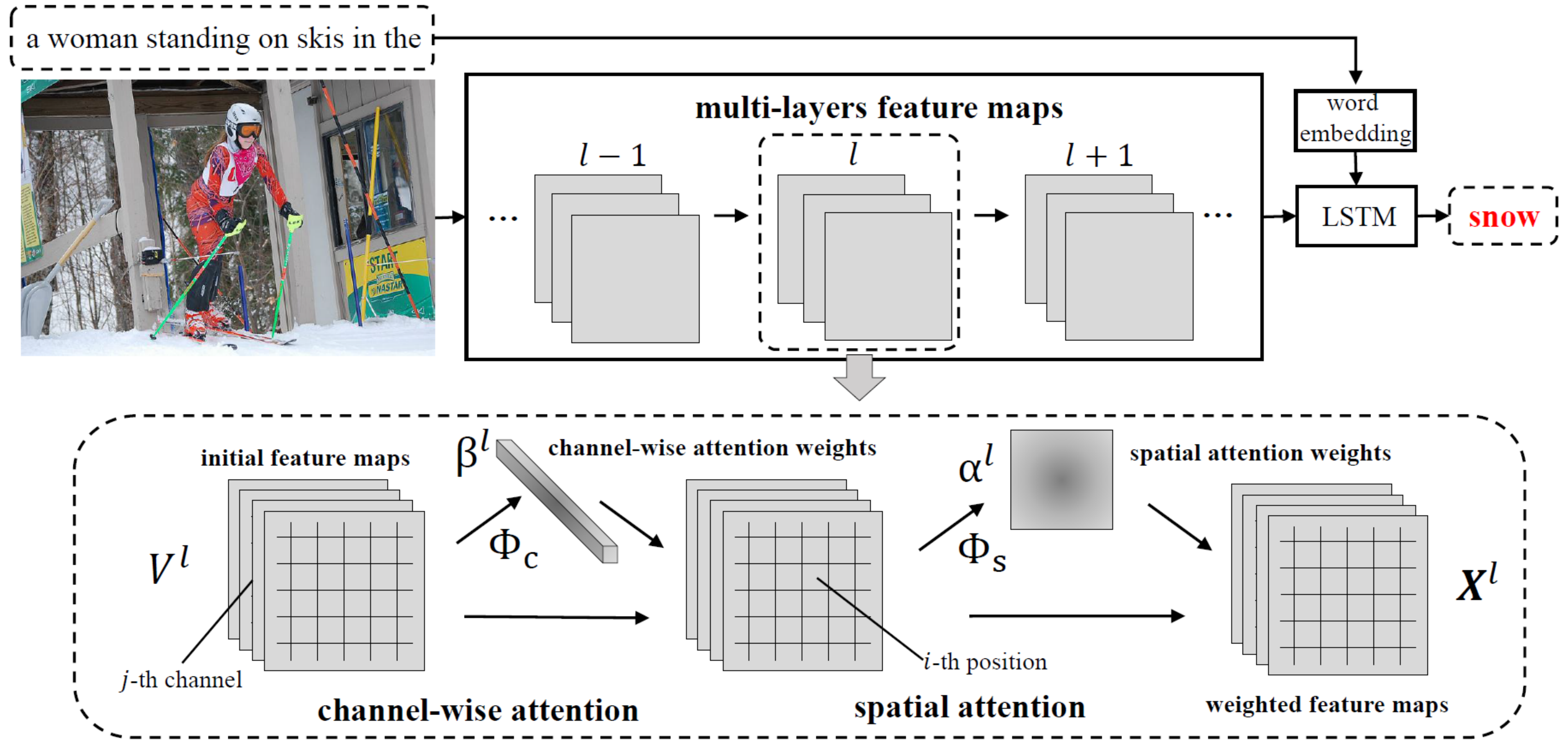
- Existed popular spatial attention mechanism only re-weight the last conv-layer feature map of a CNN.
- CNN features are naturally spatial, channel-wise and multi-layer. SCN-CNN exploits all of these features for image captioning.

Motivation

- Channel-wise: A channel-wise feature map is essentially a detector response map of the corresponding filter.
- Multi-layer: A feature map is dependent on its lower-layer ones.



Overview of SCA-CNN Architecture



SCA-CNN modulates V^l using the attention weights Υ^l in a recurrent and multi-layer fashion as:

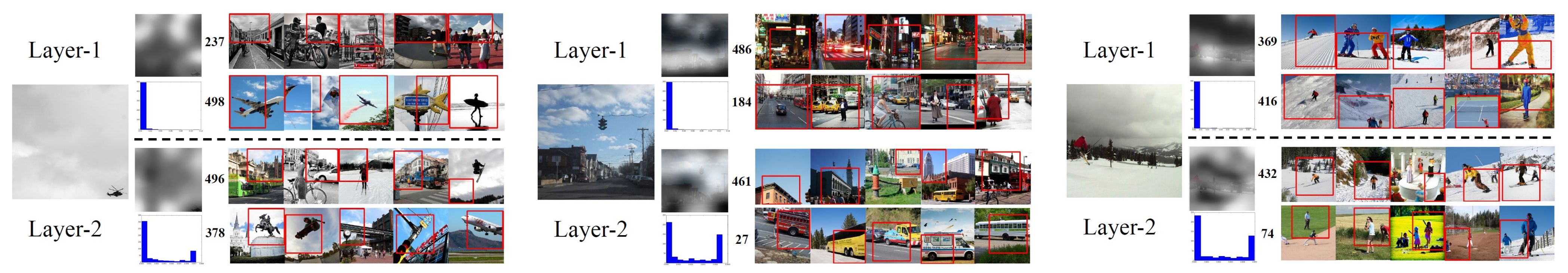
$$\boldsymbol{V}^l = \boldsymbol{CNN}(\boldsymbol{X}^l)$$
$$\Upsilon^l = \Phi(\boldsymbol{h}_{t-1}, \boldsymbol{V}^l)$$
$$\boldsymbol{X}^l = \boldsymbol{f}(\boldsymbol{V}^l, \Upsilon^l)$$

For the constrains of GPU memory, we decompose Υ^l into spatial attention weights α^l and channel-wise attention weights β^l .

Channel-Spatial variant:

$$\beta = \Phi_c(\boldsymbol{h}_{t-1}, \boldsymbol{V})$$
$$\alpha = \Phi_s(\boldsymbol{h}_{t-1}, \boldsymbol{f}_c(\boldsymbol{V}, \beta))$$
$$\boldsymbol{X} = \boldsymbol{f}(\boldsymbol{V}, \alpha, \beta)$$

Visualization of Spatial and Channel-wise Attention



a plane flying in the sky over a **cloudy** sky

a traffic light in the meddle of a **city** street

a man riding skis down a snow covered **slope**

Experimental Results

Q1: Evaluations of Channel-wise Attention

Dataset	Network	Method	B@4	MT	RG	CD
MS COCO	VGG	S	**28.2**	23.3	**51.0**	**85.7**
		C	27.3	22.7	50.1	83.4
		C-S	28.1	**23.5**	50.9	84.7
	ResNet	S	28.3	23.1	51.2	84.0
		C	29.5	23.7	51.8	91.0
		C-S	**30.4**	**24.5**	**52.5**	**91.7**

Q2: Evaluations of Multi-layer Attention

Dataset	Network	Method	B@4	MT	RG	CD
MS COCO	VGG	1-layer	28.1	23.5	50.9	48.7
		2-layers	**29.8**	**24.2**	**51.9**	**89.7**
		3-layers	29.4	24.0	51.7	88.4
	ResNet	1-layer	30.4	24.5	52.5	91.7
		2-layers	**31.1**	**25.0**	**53.1**	**95.2**
		3-layers	30.9	24.8	53.0	94.7

Q3:Comparision with State-of-The-Arts

Model	Flickr8k					Flickr30k					MS COCO				
	B@1	B@2	B@3	B@4	MT	B@1	B@2	B@3	B@4	MT	B@1	B@2	B@3	B@4	MT
Hard-Attention	67.0	45.7	31.4	21.3	20.3	**66.9**	43.9	29.6	19.9	18.5	71.8	50.4	35.7	25.0	23.0
emb-gLSTM	64.7	45.9	31.8	21.2	20.6	64.6	44.6	30.5	20.6	17.9	67.0	49.1	35.8	26.4	22.7
ATT	--	--	--	--	--	64.7	46.0	32.4	**23.0**	18.9	70.9	53.7	40.2	30.4	24.3
SCA-CNN-VGG	65.5	46.6	32.6	22.8	21.6	64.6	45.3	31.7	21.8	18.8	70.5	53.3	39.7	29.8	24.2
SCA-CNN-ResNet	**68.2**	**49.6**	**35.9**	**25.8**	**22.4**	66.2	**46.8**	**32.5**	22.3	**19.5**	**71.9**	**54.8**	**41.1**	**31.1**	**25.0**

References:

1. Show, attend and tell: Neural image caption generation with visual attention. In ICML, 2015
2. Guiding the long short term memory model for image caption generation. In ICCV, 2015
3. Image captioning with semantic attention. In CVPR, 2016

Conclusions

- SCA-CNN takes full advantage of characteristic of CNN to yield attentive image features: spatial, channel-wise, and multi-layer
- SCA-CNN achieves state-of-the-art performance on popular benchmarks for image captioning.
- SCA-CNN is not only a more powerful attention model, but also a better understanding of where(i.e., spatial) and what (i.e., channel-wise) the attention looks like in a CNN that evolves during. sentence generation.

WECHAT

FACEBOOK

CODE