MASK CAPTIONING NETWORK

Jian Han Lim and Chee Seng Chan

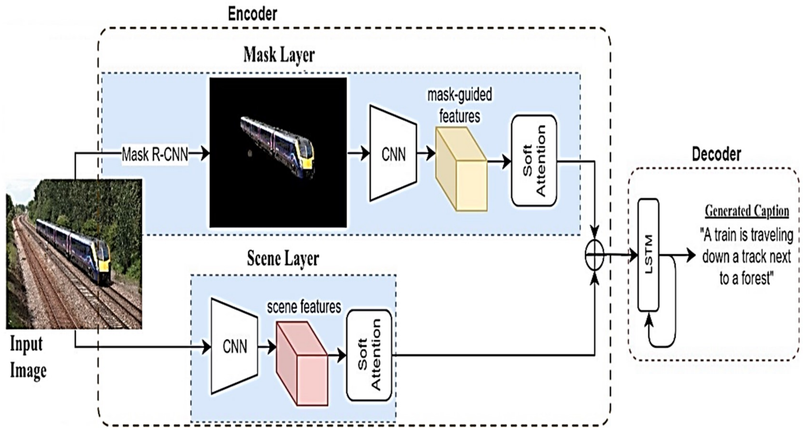
Centre of Image and Signal Processing, Faculty of Computer Science and Information Technology, University of Malaya, Malaysia

Introduction

- Image captioning: generate a sentence to describe an image
- Visual attention: de facto solution in image captioning task to detect and attend salient image regions for a better sentence generation.
- Will instance segmentation method improve the (encoder) performance in image captioning?

Main Contributions

- Propose a **Mask Captioning Network (MaC)** to detect salient regions in pixel level to eliminate the background information to focus on the image objects only
- Employ a much simpler solution to generate the scene features to ensure the overall meaning of the images are adapted into LSTM
- Our method outperforms baseline model and achieves comparable/better results with state-of-the-art methods



Methodology

- Leverage Mask R-CNN to produce a set of binary masks B and detection scores D
- Generate weighted mask B^W

$$B^w = \sum_{i=1}^{N} F_3(b_i) \odot d_i$$

- Attended mask features $\widehat{M}$:

$$M = f_m(B^w \odot I)$$

$$a^M = F_2((W_M M) \oplus (W_{M,h} h_{t-1}))$$

$$\alpha^m = softmax(W_{\alpha,m} a^M)$$

$$\widehat{M} = \sum_{i=1}^{k} \alpha_i^m M_i$$

- Attended scene features $\hat{S}$:

$$S = f_{CNN_s}(I)$$

$$a^S = F_2((W_S S) \oplus (W_{S,h} h_{t-1}))$$

$$\alpha^s = softmax(W_{\alpha,s} a^S)$$

$$\hat{S} = \sum_{i=1}^{k} \alpha_i^s S_i$$

- Concatenate $\widehat{M}$ and $\hat{S}$ before feed into LSTM at each time step t as:

$$x_t = \widehat{M} \oplus \hat{S}$$

$$h_t = LSTM(x_t, h_{t-1}, m_{t-1})$$

$$p(y_t | y_1, \cdots, y_{t-1}, I) = F_1(h_t)$$

Implementation Details

- Mask R-CNN pretrained on MSCOCO dataset
- ResNet-50 pretrained on ImageNet as image encoder
- Train all models under cross entropy loss using ADAM optimizer with mini-batch size of 32 and dropout rate 0.5
- Train LSTM using learning rate of 1e-4 for 8 epochs and finetune CNN with learning rate of 1e-5 up to 20 epochs

Results

Table 1: Performance of the proposed MaC model and state-of-the-art methods on MSCOCO dataset

Methods	MSCOCO						
	B-1	B-2	B-3	B-4	M	R	C
NIC [15]	66.6	45.1	30.4	20.3	-	-	-
ATT-FCN [16]	70.9	53.7	40.2	30.4	24.3	-	-
Hard-Attention [11]	71.8	50.4	35.7	25.0	23.04	-	-
Soft-Attention [11]	70.7	49.2	34.4	24.3	23.9	-	-
RA+SS [4]	**72.4**	55.5	41.8	31.3	24.8	53.2	95.5
Baseline	70.2	53.8	40.4	30.3	23.8	52.1	89.3
MaC_{mask}	69.8	53.1	39.7	30.0	23.6	51.7	89.6
MaC (D=0.5)	72.3	**56.0**	**42.6**	**32.4**	**25.0**	**53.7**	**96.8**
MaC (D=0.4)	72.3	55.9	42.3	32.0	24.8	53.6	95.9
MaC (D=0.6)	72.2	55.9	42.5	32.4	25.0	53.6	96.7

Table 2: Comparison on uniqueness of caption generated by proposed MaC model and baseline models.

Methods	MSCOCO		Flickr30k	
	Unique	Avg. length	Unique	Avg. length
Baseline	51.20%	8.92	89.50%	9.37
MaC_{mask}	48.36%	**9.07**	87.90%	**9.58**
MaC	**53.74%**	9.03	**89.70%**	9.37

(a) two men playing frisbee in a field.
(b) a couple of men playing a game of frisbee.

(a) a young man throwing a frisbee on a beach.
(b) a group of people playing a game of frisbee.

Comparison of captions generated by (a) MaC and (b) MaC_{mask}. Underline text indicates the scene in the sentence.