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, 50603 Kuala Lumpur, Malaysia
{jianhanl98@gmail.com, cs.chan@um.edu.my}

ABSTRACT

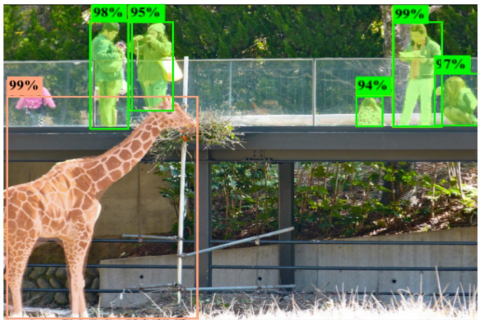
Nowadays, attention mechanisms have been widely adopted in image captioning task due to its outstanding performance. In this paper, we propose a Mask Captioning Network (MaC) that consists of an object layer and a background layer to capture the objects and scene of an image, independently to generate a much richer sentence. To this end, we leverage on Mask RCNN to detect the salient regions in pixel level in the object layer; while, in the background layer, a CNN model is used to encode the scene features. Experimental results show that our model significantly outperforms baseline models and achieves comparable results with the state-of-the-art methods on MSCOCO and Flickr30k datasets.

Index Terms— Image captioning, Deep learning, Scene understanding

1. INTRODUCTION

For the past few years, visual attention has been the de facto solution in image captioning task to detect and attend salient image regions for a much better sentence generation. For instance, in a very recent work, Anderson et al. [1] proposed a novel bottom-up and top-down attention mechanism to establish a closer link between the vision and language task by detecting a set of salient regions using Faster R-CNN [2]. Similarly, Yao et al. [3] also employed Faster R-CNN to detect objects within images and explore the visual relationship between the objects in the images by constructing semantic and spatial relation graphs for sentence generation. However, both of these works [1, 3] only used the salient regions for caption generation and the scene context of the images is excluded.

In [4], a new paradigm where a visual attention mechanism is employed with the used of selective search [5] to identify salient image regions. Then, scene vectors are predicted separately and altogether with the salient regions fed into LSTM so that the generated sentence is scene-specific. However, the work has a complicated process to obtain both the objects and scene vectors. That is to say, the visual regions are first extracted by selective search and then follow by a second process to train a classifier to select if the region is good or bad. For the scene context, text topics of images are first extracted using Latent Dirichlet Allocation, then follow by training a multilayer perceptron to predict the topic vector.



MaC: A giraffe standing next to a fence in a zoo.
MaC$_{mask}$: A giraffe standing in front of a crowd of people.
Baseline: A giraffe standing next to a wooden fence.

Fig. 1: It shows that our proposed model - MaC can generates a much richer captioning in comparison to other variants.

In this paper, we propose a new framework (namely as MaC - Mask Captioning network) that is similar to [4] but with refinements. For instance, our encoder design is as to [4], consisted of an object layer and background layer. However, we leverage on Mask R-CNN [6] to detect salient regions in pixel level as to Fig. 1, and our background layer only requires a CNN to encode the scene features. With the Mask RCNN, our work has few advantages. First, we do not require a second process (i.e to train a classifier) to select a good region since Mask RCNN produce binary mask with detection scores. So we can exploit the detection score to select a *good* region. Secondly, the binary mask of Mask RCNN is pixel level, eliminating all the background noises.

As a summary, the main contributions of this work are i) we propose a new image captioning model that leverage on Mask R-CNN to detect salient regions in pixel level to eliminate all the background information, merely focus on the image objects. ii) at the same, we also employ a much simpler solution (i.e. CNN only) to generate the scene features. With this, our proposed model can generate a much richer captioning as illustrate in Fig. 1; iii) our method outperforms baseline models and achieves comparable/better results with the state-of-the-art methods (Section 3, Table 1-2).

978-1-5386-6249-6/19/$31.00 ©2019 IEEE

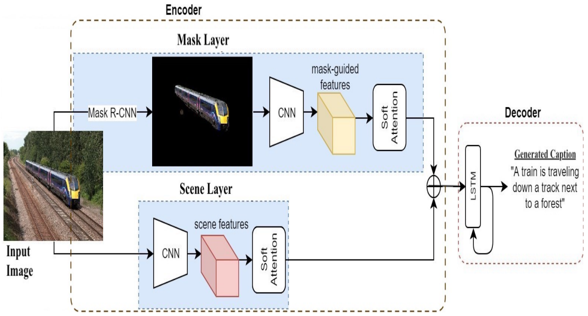


Fig. 2: Overview of the proposed framework. Our encoder consists two layers: (i) mask layer and (ii) scene layer, and the output is fed into the LSTM network for caption generation.

2. MASK CAPTIONING NETWORK

2.1. Overall Framework

Our proposed method follows the encoder-decoder framework, where an encoder is used to encode an image into image features and then feed into a decoder to generate captions as illustrated in Fig. 2. However in our encoder, as to [4], it consists of two layers: an object layer that we named as the mask layer and a background layer that we named as the scene layer. The idea is to employ the mask layer to focus on the image objects; and the scene layer to capture the background of the image. Then we apply soft attention and concatenate both features and feed them to LSTM model for sentence generation.

2.2. Mask Layer

Technically, in the mask layer, we leveraged on Mask R-CNN to produce a set of binary masks B and detection scores D, where $B = \{b_i\}_{i=1}^{N}$ with N salient regions in image I, $b_i \in \{0, 1\}^{m\times m}$ denotes the $m \times m$ binary mask for each salient region and $D = \{d_i\}_{i=1}^{N}$ with $d_i \in \mathbb{R} : 0 \leq d_i \leq 1$. Then, we generate the weighted mask B_w following:

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

where $\odot$ denotes the element-wise multiplication and $F_3(\cdot)$ represents the mask resize function. Our idea to generate the weighted masks is to exploit the confidence level of each mask in the image (i.e. the detection score D) to select a set of *good* mask features rather than training a separate classifier as to [4]. Finally, we encode B_w using a CNN model to generate mask features $M = f_m(B_w \odot I)$ where $f_m(\cdot)$ represents the CNN encoder.

2.3. Scene Layer

The mask features M generated in the mask layer only focus on the salient regions in the image and the background or the scene of the image is excluded. Intuitively, the scene or background of an image is an invaluable context that can affect the image scenario significantly. Imagine an image where a person runs in a park, the caption could be *A person is running in a park*. Otherwise, if the image is taken in a bank, the caption could be *A person is running in a bank*. It can be noticed that these two captions have very different meaning and perspective. The former can be interpret as a normal situation ("exercising"), while the latter could be a suspicious/dangerous situation as people seldom run in a bank. In [4], the author used a scene vector extractor to predict the scene vectors from the visual appearances. In our work, we propose a much simpler scene layer where we encoded using CNN only to generate the scene features S.

Technically, image I is resized as to the input of CNN and encoded to generate scene features $S = f_s(I)$ where $f_s(\cdot)$ represent the CNN encoder.

2.4. Sentence Generation

In the decoder, we concatenate the soft-attended mask features $\widehat{M}$ and soft-attended scene features $\widehat{S}$ before feed them into LSTM at each time step t as:

$$x_t = \widehat{M} \oplus \widehat{S} \tag{2}$$

$$h_t = LSTM(x_t, h_{t-1}, m_{t-1}) \tag{3}$$

$$p(y_t|y_1, \cdots, y_{t-1}, I) = F_1(h_t) \tag{4}$$

where $\oplus$ represents concatenation, h_{t-1} is the previous LSTM's hidden state, m_{t-1} is the previous memory cell, $F_1(\cdot)$ is a nonlinear function that outputs the probability of y_t, p is the probability of next word y_t with image I and previous words $y_1, \cdots, y_{t-1}$ to generate sentence.

Two different multi-layer perceptrons with softmax output are used to generate two attention distributions α^m, α^s over the mask features M and scene features S. The formula can be represented as:

$$a^M = F_2((W_M M) \oplus (W_{M,h} h_{t-1})) \tag{5}$$

$$\alpha^m = softmax(W_{\alpha,m} a^M) \tag{6}$$

$$\widehat{M} = \sum_{i=1}^{k} \alpha_i^m M_i \tag{7}$$

$$a^S = F_2((W_S S) \oplus (W_{S,h} h_{t-1})) \tag{8}$$

$$\alpha^s = softmax(W_{\alpha,s} a^S) \tag{9}$$

$$\widehat{S} = \sum_{i=1}^{k} \alpha_i^s S_i \tag{10}$$

where W denotes weights, $F_2(\cdot)$ represents multi-layer perceptron and k is the feature size. For simplicity, we do not explicitly represent bias term in this paper.

3. EXPERIMENTS

3.1. Dataset and Baselines

We conduct experiments on two popular image captioning datasets: MSCOCO [7] and Flickr30k [8], containing 123,287 and 31,783 images respectively. For a fair comparison, we followed the widely used split in [9] for both datasets: on MSCOCO, 113,287 images for training, 5,000 for validation and 5,000 for test; and for Flickr30k, 1,000 images for validation, 1,000 for test, and the rest for training. We converted all the captions into lower case and truncated captions longer than 20 words. For all experiments, we use a fixed vocabulary size of 10,000 for both datasets.

To compare the efficiency of the proposed method, MaC, we built two baseline models - the first one is implemented based on the soft attention model as to [10] and we refer this model as Baseline, and the second baseline is a mask layer only encoder and we refer this model as MaC$_{mask}$.

3.2. Implementation Details

In the mask layer: i) we use Mask R-CNN pretrained on MSCOCO dataset, generate top 100 binary masks B and remove those with detection scores D less than 0.5. ii) The image encoder in the mask layer is using ResNet-50 [11] pretrained on ImageNet [12] dataset. The mask features M are extracted using ResNet-50 without fully connected layers, resulting in $7 \times 7 \times 2048$ dimensional outputs. For the scene layer, we also used ResNet-50, pretrained on ImageNet without the fully connected layers to extract the scene features.

In the decoder, we used LSTM as our language generator and the dimension of the hidden layer and word embedding are both set to 512. We implement our model based on Tensorflow, and all experiments are trained by cross-entropy loss using Adam [13] optimizer with mini-batch size of 32 and dropout rate of 0.5. For the learning rate, we first train the LSTM decoder using learning rate of 1e-4 for 8 epochs and finetune the CNN with learning rate of 1e-5 up to 20 epochs. For inference stage, we set the beam size as 3.

3.3. Compared Approaches

To verify the MaC model, we compared with the following methods: (i) **NIC** [14] uses conventional CNN-LSTM based model which only injects image into LSTM at the initial time step. (ii) **ATT-FCN** [15] uses attributes as semantic attention to combine attributes and image in RNN for generating caption. (iii) **Hard-Attention & Soft-Attention** [10], "hard" stochastic attention and "soft" deterministic attention are used as spatial attention on convolutional features of an image. (iv) **RA+SS** [4] a work that similar to us utilizes visual attention to adapt visual features and scene features into LSTM.

Table 1: MSCOCO: Comparison between the MaC and the state-of-the-art methods, where B-N, M, R and C are short for BLEU-N, METEOR, ROUGE-L and CIDEr-D scores.

Methods	MSCOCO						
	B-1	B-2	B-3	B-4	M	R	C
NIC [14]	66.6	45.1	30.4	20.3	-	-	-
ATT-FCN [15]	70.9	53.7	40.2	30.4	24.3	-	-
Hard-Attention [10]	71.8	50.4	35.7	25.0	23.04	-	-
Soft-Attention [10]	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: Flickr30k: Comparison between MaC and the state-of-the-art methods, where B-N, M, R and C are short for BLEU-N, METEOR, ROUGE-L and CIDEr-D scores.

Methods	Flickr30k						
	B-1	B-2	B-3	B-4	M	R	C
NIC [14]	66.3	42.3	27.7	18.3	-	-	-
ATT-FCN [15]	64.7	46.0	32.4	**23.0**	18.9	-	-
Hard-Attention [10]	**66.9**	43.9	29.6	19.9	18.46	-	-
Soft-Attention [10]	66.7	43.4	28.8	19.1	18.49	-	-
RA+SS [4]	64.9	46.2	32.4	22.4	**19.4**	**45.1**	**47.2**
Baseline	63.2	44.9	31.5	21.9	17.8	44.1	42.3
MaC$_{mask}$	61.8	43.0	29.9	20.7	17.3	42.6	33.7
MaC (D=0.5)	64.7	**46.2**	**32.5**	22.7	18.5	45.0	43.4
MaC (D=0.4)	63.0	44.7	31.4	21.8	17.8	44.0	41.3
MaC (D=0.6)	63.6	45.3	32.0	22.4	18.2	44.3	42.4

3.4. Quantitative Results

Performance on MSCOCO and Flickr30k: Table 1-2 show the results on MSCOCO and Flickr30k datasets with a comparison with the state-of-the-art solutions mentioned in Section 3.3. First, we showed the performance of MaC with three different binary mask detection scores $D = 0.4$, $D = 0.5$ and 0.6 and found out $D = 0.5$ achieves the best results on both datasets. So it is selected for the rest of the experiments. On the MSCOCO dataset, it is noticed that MaC outperforms all the methods including semantic and visual attention-based approach. In particular, MaC achieves a relative improvement over similar work - RA+SS [4] on BLEU-4 and CIDEr-D scores from 31.3 to 32.4 and 95.5 to 96.8 respectively. Note that RA+SS [4] employed ResNet101 as their encoder. Comparing to the baseline models, MaC significantly outperforms the Baseline and MaC$_{mask}$ with a large margin. In particular, we can see that without the scene layer, MaC$_{mask}$ could not generate captions that are semantically correct as shown in Fig. 3. For instance, we notice that all the images in Fig. 3 involve frisbee and it will be rather confusing if the scene information is missing as shown in the second (i.e. field) and third (i.e. beach) images. In terms of quantitative analysis, we noticed that the CIDEr-D score improved by 8.4% from 89.3 to 96.8. On Flickr30k, MaC also achieves comparable

(a) a dog playing with a frisbee in the grass.
(b) a dog running with a frisbee in its mouth.

(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.

(a) A young boy holding a frisbee in his hand.
(b) A young boy in a blue shirt holding a frisbee.

Fig. 3: Comparison of captions generated by **(a)** MaC and **(b)** MaC$_{mask}$. Underline red text indicates the scene in the sentence.

(a) A group of kids playing soccer on a field.
(b) A group of kids playing soccer.
(c) A group of children playing soccer on a field.
(d) A group of five kids playing soccer together.

(a) A man holding an umbrella walking down a street.
(b) A man holding an umbrella in the rain.
(c) A woman holding an umbrella in the rain.
(d) A man with suit and tie holding an umbrella walking down the street in the rain.

(a) A herd of elephants walking across a lush green field.
(b) A herd of elephants walking across a field.
(c) A group of elephants walking through a field.
(d) A group of elephants walking by a tree in the jungle.

(a) A man and woman in a kitchen preparing food.
(b) A couple of people standing in a kitchen.
(c) two people in a kitchen preparing food.
(d) A man and woman in the kitchen preparing food.

Fig. 4: Comparison of captions generated from different baselines where **(a)** MaC, **(b)** MaC$_{mask}$, **(c)** Baseline, and **(d)** Groundtruth, respectively. Underline red text indicates the scene in the sentence.

results with the state-of-the-art methods and outperforms all the baseline models.

Evaluation on Uniqueness of Generated Caption: Table 3 shows the comparison on uniqueness of generated caption by MaC and the baseline models. We compare the generated captions with training captions, and a caption is considered as unique if the generated caption does not exist in the training captions. It shows that MaC is able to generate more unique captions when compared with Baseline and MaC$_{mask}$ on both datasets. Although, MaC$_{mask}$ has the longest average caption length but it also has the lowest uniqueness on both datasets. This is because it tends to generate caption that focuses on objects only, ignoring the scene.

Table 3: 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

3.5. Qualitative Results

Fig. 4[1] shows a few sample images, and the respective human-annotated ground truth caption and captions generated by different baselines in comparison to MaC. From these results, it shows that MaC can generate caption that captures both the objects and scene in an image, and MaC$_{mask}$ as expected generates caption without the scene context. For instance, in the first image of Fig. 4, MaC is able to generate “a group of kids playing soccer *on a field*”, but MaC$_{mask}$ only able to synthesize “a group of kids playing soccer” where the scene information is missing in the image. In the fourth image, MaC$_{mask}$ generates “A couple of people standing *in a kitchen*”, but missing of the main information which is “preparing food”. In contrast, MaC generates caption that captures the semantic relationship between the objects and scene. This is similar in the second image where MaC uses mask features (man, $umbrella$) and scene feature ($street$) to generate caption that is semantically correct by describing how the objects interact in the scene. Comparing to the Baseline, the scene (street) is missing in the generated caption.

4. CONCLUSION

This paper presented a new framework for image captioning, which explores the objects and scene in the image using mask layer and scene layer to generate captions. Experimental results on MSCOCO and Flickr30k datasets demonstrated the proposed model significantly outperforms baseline models and achieves comparable results with the state-of-the-art methods. Our future work is to explore other cues in the Mask R-CNN model such as size ratio to generate captions. We will also explore soft attention mechanism using object mask to decide whether and when to attend to the image.

[1] More results in supplementary material.

5. REFERENCES

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," in *CVPR*, 2018, vol. 3, p. 6.

[2] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *NIPS*, 2015, pp. 91–99.

[3] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei, "Exploring visual relationship for image captioning," in *ECCV*, 2018.

[4] Kun Fu, Junqi Jin, Runpeng Cui, Fei Sha, and Changshui Zhang, "Aligning where to see and what to tell: Image captioning with region-based attention and scene-specific contexts," *T-PAMI*, vol. 39, no. 12, pp. 2321–2334, 2017.

[5] Jasper RR Uijlings, Koen EA Van De Sande, Theo Gevers, and Arnold WM Smeulders, "Selective search for object recognition," *IJCV*, vol. 104, no. 2, pp. 154–171, 2013.

[6] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick, "Mask r-cnn," in *ICCV*. IEEE, 2017, pp. 2980–2988.

[7] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick, "Microsoft coco: Common objects in context," in *ECCV*. Springer, 2014, pp. 740–755.

[8] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier, "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions," *T-ACL*, vol. 2, pp. 67–78, 2014.

[9] Andrej Karpathy and Li Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3128–3137.

[10] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *ICML*, 2015, pp. 2048–2057.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.

[12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS*, 2012, pp. 1097–1105.

[13] D Kinga and J Ba Adam, "A method for stochastic optimization," in *ICLR*, 2015, vol. 5.

[14] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan, "Show and tell: A neural image caption generator," in *CVPR*, 2015, pp. 3156–3164.

[15] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo, "Image captioning with semantic attention," in *CVPR*, 2016, pp. 4651–4659.