

# Image Captioning with Integrated Bottom-Up and Multi-level Residual Top-Down Attention for Game Scene Understanding

Jian Zheng<sup>\*1</sup>, Sudha Krishnamurthy<sup>2</sup>, Ruxin Chen<sup>2</sup>, Min-Hung Chen<sup>3</sup>, Zhenhao Ge<sup>2</sup>, Xiaohua Li<sup>1</sup>

<sup>1</sup> State University of New York at Binghamton

<sup>2</sup> Sony Interactive Entertainment LLC

<sup>3</sup> Georgia Institute of Technology

{jzheng65, xli}@binghamton.edu, {sudha.krishnamurthy, ruxin.chen, zhenhao.ge}@sony.com, cmhungsteve@gatech.edu

## Abstract

*Image captioning has attracted considerable attention in recent years. However, little work has been done for game image captioning which has some unique characteristics and requirements. In this work we propose a novel game image captioning model which integrates bottom-up attention with a new multi-level residual top-down attention mechanism. Firstly, a lower-level residual top-down attention network is added to the Faster R-CNN based bottom-up attention network to address the problem that the latter may lose important spatial information when extracting regional features. Secondly, an upper-level residual top-down attention network is implemented in the caption generation network to better fuse the extracted regional features for subsequent caption prediction. We create two game datasets to evaluate the proposed model. Extensive experiments show that our proposed model outperforms existing baseline models.*

## 1. Introduction

Game image captioning is to enable machines to comprehend game scenes and output a decent description for a given image frame of game videos. Game image captioning can help players to have a better understanding of the game scenes and, especially, can provide special assistance to visually-impaired players or machine players. Nevertheless, according to our knowledge, little work has been done for image captioning with game images. There are no public datasets available for game image captioning. Game image captioning datasets are different from conventional im-

age captioning datasets because game images involve more action related contents. Unlike conventional action recognition, game image captioning needs to generate more delicate and semantic sentences to describe the complex game scenes instead of just a few words of action description. Although game images may be generated artificially, the source codes and image descriptions are usually not available to captioning machines.

Attention has been widely studied in image captioning [15]. Top-down attention alone is not enough to learn the attention distribution well, especially when the attention network is weakly-supervised. More recently, some efforts have been conducted to exploit both top-down attention and bottom-up attention [1], where bottom-up attention can be implemented based on Faster R-CNN [13]. However, existing bottom-up attention network may neglect some important spatial information in the detected salient regions.

In this paper, we propose a new image captioning model for game scene understanding by integrating bottom-up attention with a novel multi-level residual top-down attention mechanism. Specifically, a lower-level residual top-down attention network is added into the Faster R-CNN based bottom-up attention network to exploit better spatial information. An upper-level residual top-down attention network is implemented in the caption generation network to selectively attend to certain regions for better caption prediction. In addition, we create two game image datasets to validate the performance of the proposed model.

## 2. Related Work

Image captioning has been a hot research topic during the past several years [6][14]. Many top-down visual attention mechanisms have been developed and adopted in image

<sup>1</sup>This work is done during the first author's internship at SIE.

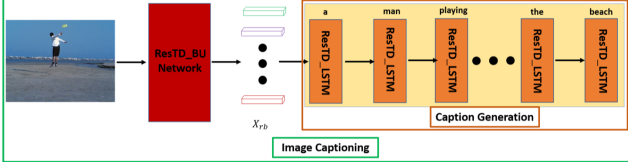


Figure 1. Block diagram of the proposed image captioning network.

captioning [16][9][10][1][8]. Recently, some approaches with both bottom-up attention and top-down attention are proposed [5, 12]. In particular, a combined bottom-up and top-down attention mechanism is proposed in [1], which is the current state-of-the-art in image captioning model. This paper enhances [1] with a novel multi-level residual top-down attention mechanism and addresses the specific problem of game image captioning.

### 3. Method

Fig. 1 shows our proposed image captioning model. The model consists of mainly two parts: a Residual Top-Down Bottom-Up (ResTD.BU) attention network and a Caption Generation network. ResTD.BU extracts regional feature vectors from an image, while the Caption Generation network takes the extracted feature vectors of different salient regions as input and predicts the natural language caption word by word.

#### 3.1. ResTD\_BU Network

The ResTD.BU network uses Faster R-CNN to realize bottom-up attention. Nevertheless, we replace the simple average pooling used in [1] with a novel residual network to introduce top-down attention, as illustrated with the red crossing and the ResTD\_1 block in Fig. 2.

Our ResTD.BU network takes a given image as input and feeds it to a deep ConvNet (we use ResNet-101 [4]) to learn the intermediate image feature representations. Then a Region Proposal Network (RPN) is applied over the intermediate feature maps to generate a large number of region proposals, among which only a small portion will be selected by applying non-maximum suppression (NMS) and intersection-over-union (IoU) threshold. The output of the RoI pooling layer is a set of  $N$  region feature maps, described as  $X_b = \{X_{b1}, X_{b1}, \dots, X_{bN}\}$ , where  $X_{bi} \in \mathcal{R}^{n_1 \times n_2 \times D}$ . We used  $n_1 = n_2 = 7, D = 2048$  in our experiments.

Next, instead of applying global average pooling, we propose to use the ResTD\_1 network to implement lower-level top-down attention on the region feature maps. Specifically, we take each  $X_{bi}$  as input and feed it to ResTD\_1 to obtain a feature vector  $x_{rbi}$  for each region. In the end, we have a set of regional feature vectors  $X_{rb} = \{x_{rb1}, x_{rb2}, \dots, x_{rbN}\}$  for each given image.

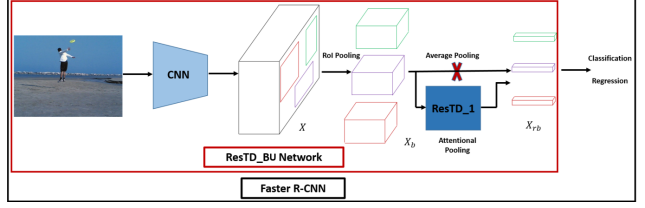


Figure 2. The ResTD\_BU Network: Faster R-CNN based bottom-up attention network with an integrated lower-level residual top-down attention network ResTD\_1.

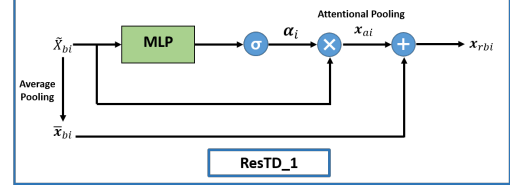


Figure 3. ResTD\_1: Lower-level residual top-down attention network.

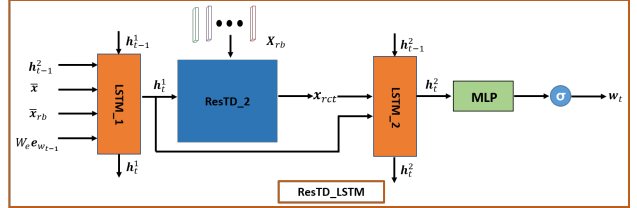


Figure 4. Block diagram of a ResTD\_LSTM unit.

The ResTD\_1 network is implemented as in Fig. 3, which is inspired by the residual network architecture [4]. The MLP is a two-layer fully-connected network,  $\sigma$  is the softmax function,  $\times$  and  $+$  are element-wise product and summation, respectively. The attention distribution among input data might not be learned well, thus important information could be lost during feature fusion. The averagely pooled feature vector  $\bar{x}_{bi}$  from  $\tilde{X}_{bi}$  can make up for the image information loss, where  $\tilde{X}_{bi} \in \mathcal{R}^{(n_1 \times n_2) \times D}$ .

#### 3.2. Caption Generation Network

The Caption Generation network consists of a set of residual top-down attention based LSTM units, which we call ResTD.LSTM. Top-down attention makes the network attend to selective regions when generating caption words. The network uses  $X_{rb}$  as input. As shown in Fig. 4, each ResTD.LSTM unit has three components: an upper-level residual top-down attention network (ResTD\_2), an LSTM for contextual information embedding (LSTM\_1), and an LSTM for caption prediction (LSTM\_2).

LSTM\_1 is used to combine image feature information with context information. Image feature information includes the global bottom-up feature vector  $\bar{x}_{rb}$ , which is obtained from  $X_{rb}$  through average pooling, and the global

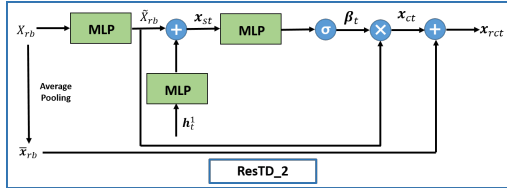


Figure 5. ResTD\_2: Upper-level residual top-down attention network.

feature vector of the entire image  $\bar{x}$ , which is obtained through global average pooling over  $X$ . Context information consists of two components, i.e.,  $W_e e_{w_{t-1}}$ , which is the word embedding predicted from the previous time step, and  $h_{t-1}^2$ , which is the hidden state of LSTM\_2. The LSTM\_1 mixes all these information to get a general contextual vector  $h_t^1$ .

The vector  $h_t^1$ , as well as the set of region feature vectors  $X_{rb}$  are fed into ResTD\_2 to learn the attention distribution over the  $N$  regions and to output the contextual vector  $x_{rct}$ . The implementation of ResTD\_2 is shown in Fig. 5. In addition to  $x_{rct}$ , the input to the caption prediction cell LSTM\_2 includes the hidden state  $h_{t-1}^2$ . To predict the next word  $w_t$ , we feed the output of LSTM\_2 to a single-layer perceptron MLP, followed by a softmax function.

## 4. Datasets & Experiments

To train the ResTD\_BU network, we use Visual Genome [7] and a game image dataset named GD\_Det, which is created by us for game object detection. To evaluate the overall image captioning model on game images, we use V-COCO [3] and our own game image captioning dataset GD\_ImgCap.

For game scene understanding, we are more interested in game objects and associated actions. Therefore, we only use the same 1600 object classes as [1] in Visual Genome. Our GD\_Det dataset contains 7 object classes: human, gun, axe, sword, monster, car, motorcycle. We split the dataset into 4,725 training images, 834 validation images, and 1,390 testing images. Since Visual Genome is much larger than GD\_Det, we train the ResTD\_BU network with these two datasets separately. We implement the ResTD\_BU network based on [2].

We created our image captioning dataset GD\_ImgCap with 7,558 game images collected from 131 different game videos, including 5,920 training images, 658 validation images, as well as 606 test images. The length of game image captions ranges from 2 to 19. On average, there are 2 captions for each game image. We combine V-COCO with GD\_ImgCap as the final game image captioning dataset. As a result, there are 15,331 training images, 1693 validation images, and 606 test images. We implement the image captioning model based on [11].

	GD_Det	Visual Genome
No_Att	58.22%	4.62%
Att	60.15%	6.36%
ResAtt	<b>61.88%</b>	<b>7.34%</b>

Table 1. Performance (mAP) comparison of Faster R-CNN object detection with/without top-down attention.

	avg_BLEU	CIDEr	METEOR	ROUGE	SPICE
SAT	16.53	99.37	13.32	33.74	26.97
TD	17.89	117.77	13.82	35.83	30.53
Att2in2	18.74	118.40	14.21	36.98	30.17
AdaAtt	19.18	127.84	14.43	36.66	30.09
Ours	<b>19.59</b>	<b>127.90</b>	<b>15.78</b>	<b>38.67</b>	<b>32.67</b>

Table 2. Performance comparison of our game captioning model with competing existing models on GD\_ImgCap.

	avg_BLEU	CIDEr	METEOR	ROUGE	SPICE
BU_Only	18.28	115.78	14.43	37.08	30.07
BU+Td	18.13	115.80	15.14	37.20	30.79
BU+ResTd	<b>19.59</b>	<b>127.90</b>	<b>15.78</b>	<b>38.67</b>	<b>32.67</b>

Table 3. The captioning performance under various model variants.

We compare the performance of our proposed image captioning model with four existing models, i.e., SAT [16], TD, Att2in2 [14], and AdaAtt [10]. TD is a baseline image captioning model with an architecture similar to our model. The difference is that in TD, we remove  $\bar{x}$  from LSTM\_1 and remove  $x_{rb}$  from ResTD\_2. The inputs to these four models are our baseline bottom-up attention regional features  $X_{rb}$  extracted without lower-level residual top-down attention introduced.

## 5. Experiment Results

**Detection Results:** Experiment results shown in Table 1 demonstrate the effectiveness of the proposed lower-level residual top-down attention network. Regular top-down attention (**Att**) improves the detection performance of Faster R-CNN, but our residual top-down attention network (**ResAtt**) has the best performance.

**Captioning Results:** We report the captioning results of our proposed model (**Ours**) with the other four models on GD\_ImgCap in Table 2. It can be seen that our model outperforms all the other four models. Especially, by comparing our model with the baseline model TD, we come to the conclusion that integrating bottom-up and multi-level residual top-down attention drastically improves captioning performance.

**Ablation Study:** To better understand our residual top-down attention, we compared our proposed model with its two variants: **BU\_only** where no top-down attention was used, and **BU+Td** where the residual top-down attention was replaced by conventional top-down attention. Results



<b>Human</b>	a woman standing inside of a wooden room and hitting wood with an axe.
<b>SAT</b>	a man running on wood and holding a gun.
<b>TD</b>	a woman standing on stone inside of a room and holding a gun.
<b>Att2in2</b>	a woman running on wood and holding a gun.
<b>AdaAtt</b>	a woman standing on stone inside of a room and holding a gun.
<b>Ours</b>	a woman standing inside of a room and hitting a wooden wall with an axe.

Figure 6. Captions generated from our model and some existing models for a game image.

shown in Table 3 indicates that residual top-down attention greatly improves game captioning performance.

**Caption Example:** Fig. 6 shows the captions generated from a game image. It can be seen that our proposed model outperforms the other four models in the sense that it generates more detailed caption that is closer to the human-annotated ground truth caption.

## 6. Conclusion

In this paper, we propose a novel game image captioning model with integrated bottom-up and multi-level residual top-down attention. To extract better regional feature representations, we introduce a lower-level residual top-down attention network into the Faster R-CNN based bottom-up network. In order to attend better to different regions during caption prediction, we employ an upper-level residual top-down attention network in the caption generation network. Two game image datasets are created. Extensive experiments are conducted to demonstrate the effectiveness of the proposed model on game images.

## References

[1] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, volume 3, page 6, 2018. 1, 2, 3

[2] X. Chen and A. Gupta. An implementation of faster rcnn with study for region sampling. *arXiv preprint arXiv:1702.02138*, 2017. 3

[3] S. Gupta and J. Malik. Visual semantic role labeling. *arXiv preprint arXiv:1505.04474*, 2015. 3

[4] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2

[5] J. Jin, K. Fu, R. Cui, F. Sha, and C. Zhang. Aligning where to see and what to tell: image caption with region-based attention and scene factorization. *arXiv preprint arXiv:1506.06272*, 2015. 2

[6] J. Johnson, A. Karpathy, and L. Fei-Fei. Denscap: Fully convolutional localization networks for dense captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4565–4574, 2016. 1

[7] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, et al. Visual genome: Connecting language and vision using crowd-sourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73, 2017. 3

[8] L. Li, S. Tang, L. Deng, Y. Zhang, and Q. Tian. Image caption with global-local attention. In *AAAI*, pages 4133–4139, 2017. 2

[9] C. Liu, J. Mao, F. Sha, and A. L. Yuille. Attention correctness in neural image captioning. In *AAAI*, pages 4176–4182, 2017. 2

[10] J. Lu, C. Xiong, D. Parikh, and R. Socher. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 6, page 2, 2017. 2, 3

[11] R. Luo. An image captioning codebase in pytorch. <https://github.com/ruotianluo/ImageCaptioning.pytorch>, 2017. 3

[12] M. Pedersoli, T. Lucas, C. Schmid, and J. Verbeek. Areas of attention for image captioning. *arXiv preprint arXiv:1612.01033*, 2016. 2

[13] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015. 1

[14] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel. Self-critical sequence training for image captioning. In *CVPR*, volume 1, page 3, 2017. 1, 3

[15] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2015. 1

[16] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057, 2015. 2, 3