The Summary of ”Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering”[1] Paper

Furkan Gül

I. Motivation

Image captioning and visual question answering (VQA) have been gathered considerable research interest at the edge of combining image and language understanding. To form high-quality outputs in both of these tasks, extracting features from an image must be performed at the fine grid level of image patches (like the images in Figure 1). That's why attention mechanisms for visual understanding have been widely preferred in image captioning and VQA problems. These mechanisms push the algorithm to focus on the salient regions/patches of images so that it improves the overall performance. In the human visual system, there are two types of signals: “top-down” signals for the current task (e.g. looking for something) and “bottom-up” signals for salient or novel stimuli. In this paper, the inspiration for the attention model architecture arises from the attention terminology in the human visual system. Thus, task-specific non-visual context attention mechanism is called “top-down” and attention for visual regions is called “bottom-up”. Most attention mechanisms used in image captioning and VQA are top-down types. This method does not give much attention to how each image regions are determined. In Figure 1, the resulting image regions in the left image are equally sized uniform grid. These regions do not consider the content of the image itself. To create captions like human-made, salient image patches and corresponding objects must be considered to form a basis for the attention mechanism (like the left image in Figure 1).

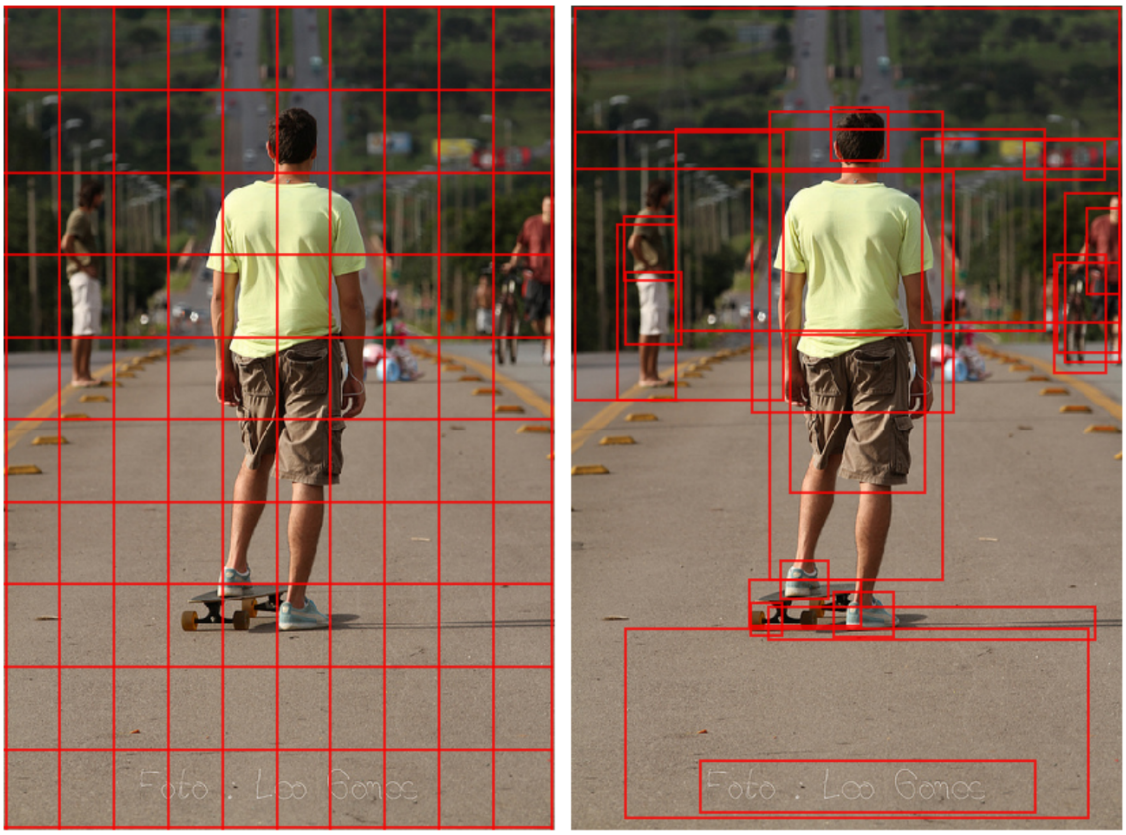

Fig. 1. Typically, attention models operate on CNN features corresponding to a uniform grid of equally-sized image regions (left). Their approach enables attention to be calculated at the level of objects and other salient image regions (right). [1]

II. Approach

A. Bottom-Up Attention Model

Image captioning and VQA models take the image features, *V*, as an input as follows:

$$V = \{\mathbf{v}_1, ..., \mathbf{v}_k\}, \mathbf{v}_i \in \mathbb{R}^D (1)$$

The spatial image feature, *V*, is the output of the bottom-up attention mechanism model coming from the last layer of a convolutional neural network (CNN). Each image feature, $\boldsymbol{v_i}$, encodes a salient region in the image. In this paper, spatial regions/patches are defined as the bounding boxes using Faster R-CCN [2] algorithm. In their work, Faster R-CNN together with the ResNET-101 [3] is implemented as the bottom-up attention model to select image regions that exceed a confidence threshold. For each selected region *i*, the size of the image feature vectors is 2048, denoted by D. Faster R-CNN operates like a "hard" attention model since it only selects a relatively small number of bounding box features for image regions/patches among a huge number of possible bounding boxes. The bottom-up model is pre-trained on ImageNet [4] classification. Finally, the bottom-up attention model outputs image feature vectors, V.

B. Captioning Model

The captioning model in their experiments consists of two LSTM [5] layers: top-down attention LSTM and language LSTM, using a general implementation [6]. Each layer is indicated with superscripts in the equations (like equations in Figure 2).

$$\boldsymbol{x}_t^1 = [\boldsymbol{h}_{t-1}^2, \bar{\boldsymbol{v}}, W_e \Pi_t]$$

$$a_{i,t} = \boldsymbol{w}_a^T \tanh\left(W_{va}\boldsymbol{v}_i + W_{ha}\boldsymbol{h}_t^1\right)$$

$$\boldsymbol{\alpha}_t = \text{softmax}\left(\boldsymbol{a}_t\right)$$

$$\hat{\boldsymbol{v}}_t = \sum_{i=1}^{K} \alpha_{i,t}\boldsymbol{v}_i$$

$$\boldsymbol{x}_t^2 = [\hat{\boldsymbol{v}}_t, \boldsymbol{h}_t^1]$$

$$p(y_t \mid y_{1:t-1}) = \text{softmax}\left(W_p \boldsymbol{h}_t^2 + \boldsymbol{b}_p\right)$$

$$p(y_{1:T}) = \prod_{t=1}^{T} p(y_t \mid y_{1:t-1})$$

$$L_{XE}(\theta) = -\sum_{t=1}^{T} \log(p_\theta(y_t^* \mid y_{1:t-1}^*))$$

$$L_R(\theta) = -\mathbf{E}_{y_{1:T} \sim p_\theta}[r(y_{1:T})]$$

Fig. 2. Equations for Top-Down Attention & Language LSTM

C. Top-Down Attention LSTM

It is the first LSTM layer in the captioning model. $\boldsymbol{x_t^1}$ is the input vector to the top-down attention LSTM at each time step. $\boldsymbol{h_{t-1}^2}$ is the previous output of the language LSTM. It represents the state of the language LSTM. $\overline{\boldsymbol{v}}$ is the mean pooled image feature vector. It is basically the overall content of the input image, and calculated as follows:

$$\overline{\boldsymbol{v}} = \frac{1}{k}\sum_i \boldsymbol{v_i} \tag{2}$$

$W_e \prod_t$ is an encoding of the previously generated word. W_e is word embedding matrix. The word embedding is randomly initialized without pretraining. $\prod_t$ is one-hot encoding of any word at time t. At each time step t, a normalized attention weight, $a_{i,t}$, is outputted for each k image features $\boldsymbol{v_i}$. $\hat{\boldsymbol{v_t}}$ is the attended image feature. The language LSTM takes this as an input.

D. Language LSTM

This is the second LSTM layer in the captioning model. $\boldsymbol{x_t^2}$ is the input vector to the language LSTM at each time step. $\boldsymbol{h_t^1}$ is the output of attention LSTM.

All equations for both LSTMs can be found in Figure 2. Last two equations in Figure 2 are the cross entropy loss and CIDEr [7] optimization. $y^*_{1:T}$ is the target ground truth sequence. θ represents parameters in the captioning model. By looking both equations in Figure 2 and the model architecture in Figure 3, the general structure of this two LSTM layered captioning model can be easily understood.

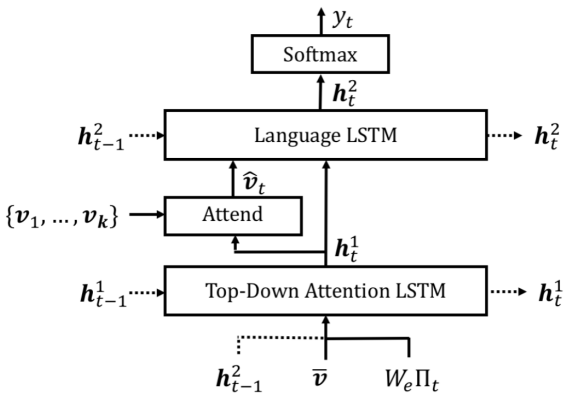

Fig. 3. Overview of the proposed captioning model. Two LSTM layers are used to selectively attend to spatial image features $\boldsymbol{v_1}, \ldots, \boldsymbol{v_k}$. These features can be defined as the spatial output of a CNN, or following their approach, generated using bottom-up attention. [1]

III. Dataset & Results

MSCOCO [8] dataset is used in their experiments to evaluate the suggested captioning model. There are at least 5 sentence descriptions for every image in this dataset. For both the validation and test phases with MSCOCO dataset, 5000 images are used. They convert all sentences to lower case, tokenize on white space, and filter words that have less than five occurences. Thus, the vocabulary size for their experiments is fixed and 10,010. For the evaluation of the caption model, following metrics are used: SPICE [9], CIDEr [7], METEOR [10], ROUGE-L [11], and BLEU [12]. They compare their Up-Down model against their baseline model and previous works done by other researchers. The baseline model uses a ResNet CNN architecture pretrained on ImageNet dataset to encode images instead of using the bottom up attention mechanisms.

In Figure 4, the performance of their baseline ResNet model and full Up-Down model are compared with the current SOTA Self-critical Sequence Training(SCST) [13] method on the Karpathy split [14]. To be a fair comparison, two optimization methods, Cross-Entropy and CIDEr Optimization, are used for reporting the results. As can be seen in Figure 4, Up-Down model has the best result at the time of 2017.

	Cross-Entropy Loss						CIDEr Optimization					
	BLEU-1	BLEU-4	METEOR	ROUGE-L	CIDEr	SPICE	BLEU-1	BLEU-4	METEOR	ROUGE-L	CIDEr	SPICE
SCST:Att2in [34]	-	31.3	26.0	54.3	101.3	-	-	33.3	26.3	55.3	111.4	-
SCST:Att2all [34]	-	30.0	25.9	53.4	99.4	-	-	34.2	26.7	55.7	114.0	-
Ours: ResNet	74.5	33.4	26.1	54.4	105.4	19.2	76.6	34.0	26.5	54.9	111.1	20.2
Ours: Up-Down	**77.2**	**36.2**	**27.0**	**56.4**	**113.5**	**20.3**	**79.8**	**36.3**	**27.7**	**56.9**	**120.1**	**21.4**
Relative Improvement	4%	8%	3%	4%	8%	6%	4%	7%	5%	4%	8%	6%

Fig. 4. Single-model image captioning performance on the MSCOCO Karpathy test split. Their baseline ResNet model obtains similar results to SCST, the existing state-of-the-art on this test set. Illustrating the contribution of bottom-up attention, their Up-Down model achieves significant (3–8 percentage) relative gains across all metrics regardless of whether cross-entropy loss or CIDEr optimization is used [1].

In the Figure 5, they report the performance of four ensembled models trained on the official MSCOCO evaluation server. At the time of submission, 18 July 2017, they take the lead in the leader board by outperforming all other submissions on the test server in terms of all metrics.

IV. A Couple of Words

I took a quick glance at the VQA model details and saw that it is not complicated as I expected. However, I did not go over the VQA model which uses VisualGenome [15] dataset in the summary since my project is about image captioning not

	BLEU-1		BLEU-2		BLEU-3		BLEU-4		METEOR		ROUGE-L		CIDEr		SPICE	
	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40
Review Net [48]	72.0	90.0	55.0	81.2	41.4	70.5	31.3	59.7	25.6	34.7	53.3	68.6	96.5	96.9	18.5	64.9
Adaptive [27]	74.8	92.0	58.4	84.5	44.4	74.4	33.6	63.7	26.4	35.9	55.0	70.5	104.2	105.9	19.7	67.3
PG-BCMR [24]	75.4	-	59.1	-	44.5	-	33.2	-	25.7	-	55	-	101.3	-	-	-
SCST:Att2all [34]	78.1	93.7	61.9	86.0	47.0	75.9	35.2	64.5	27.0	35.5	56.3	70.7	114.7	116.7	20.7	68.9
LSTM-A_3 [49]	78.7	93.7	62.7	86.7	47.6	76.5	35.6	65.2	27	35.4	56.4	70.5	116	118	-	-
Ours: Up-Down	**80.2**	**95.2**	**64.1**	**88.8**	**49.1**	**79.4**	**36.9**	**68.5**	**27.6**	**36.7**	**57.1**	**72.4**	**117.9**	**120.5**	**21.5**	**71.5**

Fig. 5. Highest ranking published image captioning results on the online MSCOCO test server. Their submission, an ensemble of 4 models optimized for CIDEr with different initializations, outperforms previously published work on all reported metrics. At the time of submission (18 July 2017), they also outperformed all unpublished test server submissions [1].

VQA. The next paper to be summarized will be "Multimodal Transformer with Multi-View Visual Representation for Image Captioning"[16] from 2019.

REFERENCES

[1] Peter Anderson, X. He, C. Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6077–6086, 2018.

[2] Shaoqing Ren, Kaiming He, Ross B. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39:1137–1149, 2015.

[3] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[4] Jia Deng, W. Dong, R. Socher, L. Li, K. Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR 2009*, 2009.

[5] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 9:1735–1780, 1997.

[6] J. Donahue, Lisa Anne Hendricks, Marcus Rohrbach, Subhashini Venugopalan, S. Guadarrama, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2625–2634, 2015.

[7] Ramakrishna Vedantam, C. L. Zitnick, and D. Parikh. Cider: Consensus-based image description evaluation. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4566–4575, 2015.

[8] Tsung-Yi Lin, M. Maire, Serge J. Belongie, J. Hays, P. Perona, D. Ramanan, Piotr Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. *ArXiv*, abs/1405.0312, 2014.

[9] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *ECCV*, 2016.

[10] Michael J. Denkowski and A. Lavie. Meteor universal: Language specific translation evaluation for any target language. In *WMT@ACL*, 2014.

[11] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *ACL 2004*, 2004.

[12] Kishore Papineni, S. Roukos, T. Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *ACL*, 2002.

[13] Steven J. Rennie, E. Marcheret, Youssef Mroueh, J. Ross, and V. Goel. Self-critical sequence training for image captioning. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1179–1195, 2017.

[14] A. Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39:664–676, 2017.

[15] R. Krishna, Yuke Zhu, O. Groth, J. Johnson, Kenji Hata, J. Kravitz, Stephanie Chen, Yannis Kalantidis, L. Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123:32–73, 2016.

[16] J. Yu, Jing Li, Zhou Yu, and Q. Huang. Multimodal transformer with multi-view visual representation for image captioning. *ArXiv*, abs/1905.07841, 2019.