The Summary of ”Hierarchy Parsing for Image Captioning” [1] Paper

Furkan Gül

I. Motivation

It is suggested to transform the problem from the viewpoint of parsing an image into a hierarchical structure of image patterns to represent the image in a better way. The main contribution is to create a top-down hierarchical tree from the image root to the middle layers of regions and the leaf layer of instances. Their solution also results in the elegant view of how to create and interpret the hierarchy of an image, and how to combine such hierarchy into general image captioning frameworks. The proposed design is considered as a feature refiner in general and can be seen in Figure 1.

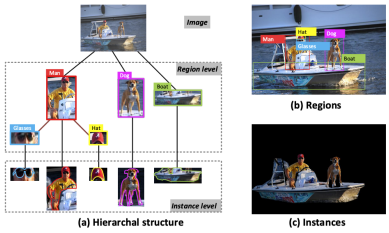

Fig. 1. Examples of (a) the hierarchal tree structure in a given image, (b) region levels and (c) instances in the image [1].

II. Visual Commonsense (VC) R-CNN

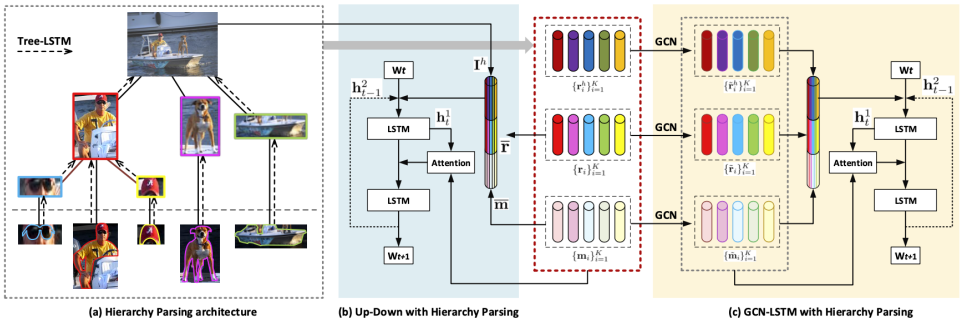

Fig. 2. HIerarchy Parsing (HIP) Architecture.

An overview of the proposed Hierarchy Parsing (HIP) architecture can be seen in Figure 2 (a). For HIP, both Mask R-CNN and Faster R-CNN [2] are implemented to detect and segment the set of object instances and their regions. Then, a three-level hierarchy is built where the entire image is initially decomposed into regions and every region is represented either at this level. Each region at middle layers is linked to the corresponding instance at leaf layer. Next, a Tree-LSTM structure is applied to the hierarchy from the bottom up with improved features of image regions. Generally, outputs represent image-level features. For sentence generation, a combination of features on three levels given by HIP could be easily given to a standard attention-based LSTM structure in Up-Down. Additionally, it is also easy to add HIP module into GCN-LSTM by improving the multi-level features with semantic relations by means of Graph Convolutional Networks (GCN).

III. Dataset & Results

All experiments in the paper are carried out on the most popular image captioning benchmark dataset, MSCOCO [3]. MSCOCO dataset which has more than 120000 images is used in their experiments to evaluate the suggested captioning model. There are at least 5 human-annotated sentence descriptions for every image in this dataset. They use Karparthy split [4] which is extensively preferred for reporting results in previous works. For both the validation and test phases with MSCOCO dataset, 5000 images are used. They convert all caption sentences to lower case and discard words that have less than five

	Cross-Entropy Loss					CIDEr-D Score Optimization				
	BLEU@4	METEOR	ROUGE-L	CIDEr-D	SPICE	BLEU@4	METEOR	ROUGE-L	CIDEr-D	SPICE
LSTM [29]	29.6	25.2	52.6	94.0	-	31.9	25.5	54.3	106.3	-
SCST [24]	30.0	25.9	53.4	99.4	-	34.2	26.7	55.7	114.0	-
ADP-ATT [21]	33.2	26.6	-	108.5	-	-	-	-	-	-
LSTM-A [35]	35.2	26.9	55.8	108.8	20.0	35.5	27.3	56.8	118.3	20.8
RFNet [13]	37.0	27.9	57.3	116.3	20.8	37.9	28.3	58.3	125.7	21.7
Up-Down [3]	36.2	27.0	56.4	113.5	20.3	36.3	27.7	56.9	120.1	21.4
Up-Down+HIP	37.0	28.1	57.1	116.6	21.2	38.2	28.4	58.3	127.2	21.9
GCN-LSTM [34]	37.1	28.1	57.2	117.1	21.1	38.3	28.6	58.5	128.7	22.1
GCN-LSTM+HIP	**38.0**	**28.6**	**57.8**	**120.3**	**21.4**	**39.1**	**28.9**	**59.2**	**130.6**	**22.3**

Fig. 3. Single image captioning model performance comparisons on the MSCOCO Karpathy test split. All values are reported as percentage (%).

occurrences. Thus, the vocabulary size for their experiments is fixed and 10201. For the evaluation of the caption model, following metrics are used: CIDEr [5], METEOR [6], SPICE [7], ROUGE [8], and BLEU [9].

Visual Genome [10] is used to train Faster R-CNN for detection. HIP is trained with a mini-batch size of 50, Adam [11] optimizer with 5e-4 learning rate, the beam size of 3. For the cross-entropy loss optimization, they set the maximum number of iterations to 30 epochs. For the optimization of the CIDEr-D score, SCST is applied at the fixed learning rate of 5e-5, and the maximum number of iterations set to 30 epochs. In Figure 3, they reports their implementation results with Up-Down [12] and GCN-LSTM [13] on the Karpathy test split. As it can be seen in Figure 3, HIP implemented versions outperform original Up-Down and GCN-LSTM models.

REFERENCES

[1] Ting Yao, Yingwei Pan, Yehao Li, and T. Mei. Hierarchy parsing for image captioning. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2621–2629, 2019.

[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] 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.

[4] 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.

[5] 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.

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

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

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

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

[10] 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.

[11] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2015.

[12] 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.

[13] Ting Yao, Yingwei Pan, Yehao Li, and T. Mei. Exploring visual relationship for image captioning. In *ECCV*, 2018.