The Summary of ”Visual Commonsense R-CNN” [1] Paper

Furkan Gül

I. Motivation

A novel unsupervised feature representation learning approach is proposed, Visual Commonsense Region-based Convolutional Neural Network (VC R-CNN). It is a kind of improved visual region encoder for high-level tasks like image captioning.

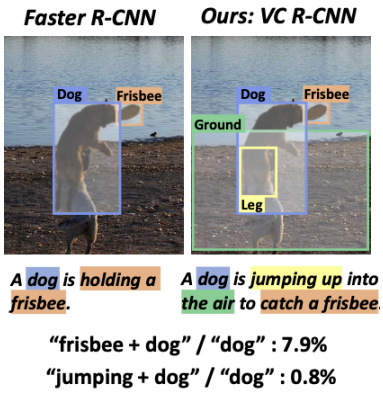

Fig. 1. Example of Cognitive Error in Image Captioning due to bias in the dataset [1].

The ratio in Figure 1 represents the co-occurrence percentage in ground-truth captions. By comparing with the Faster R-CNN [2] based features, suggested VC R-CNN features can fix the mistakes which results in more accurate visual relationships and visual attentions by increasing commonsense awareness. So, visual features must consist of more commonsense knowledge rather than having only visual appearances.

II. Visual Commonsense (VC) R-CNN

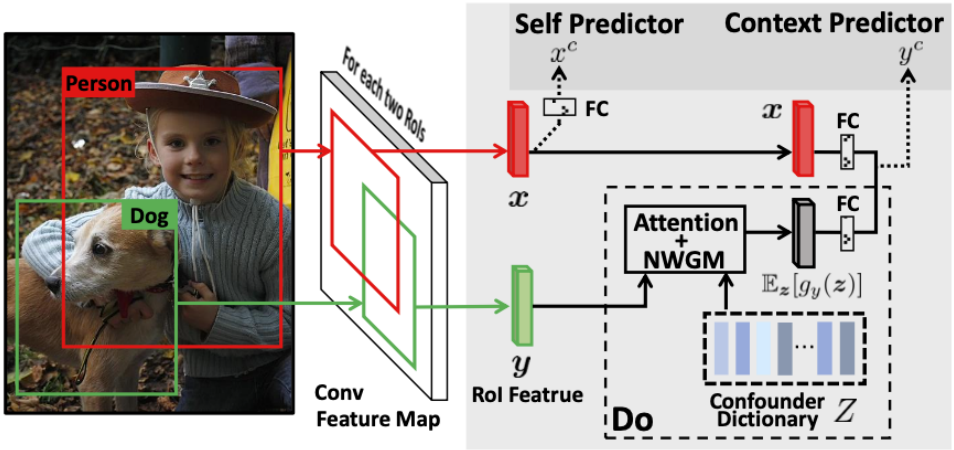

Fig. 2. VC R-CNN Model Framework [1].

The overall architecture of VC R-CNN can be seen in Figure 2. VC R-CNN takes an image as input and generates feature map from a CNN backbone. Any R-CNN visual backbone (like Faster R-CNN) can be implemented to exploit regions of interest (RoI) on the feature map. Then, each RoI is given to two sibling branches. The first branch is a Self Predictor to predict a class. The second one is a Context Predictor to predict context labels with our $\boldsymbol{Do}$ calculus. For tasks involving region proposal, VC R-CNN can be chosen as a ready-to-use region feature extractor for many high-level image tasks such as Image Captioning.

VC R-CNN is considered as a visual commonsense feature extractor for any region proposal. Then, extracted feature vectors are concatenated to the original image feature. Note that early concatenations for some models that contain a self-attention architecture (like AoANet) is not recommended.

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. For the evaluation of the caption model, following metrics are used: CIDEr [5], METEOR [6], SPICE [7], ROUGE [8], and BLEU [9].

Model	BLEU-4		METEOR		ROUGE-L		CIDEr-D	
Metric	c5	c40	c5	c40	c5	c40	c5	c40
Up-Down [2]	36.9	68.5	27.6	36.7	57.1	72.4	117.9	120.5
SGAE [70]	37.8	68.7	28.1	37.0	58.2	73.1	122.7	125.5
CNM [71]	37.9	68.4	28.1	36.9	58.3	72.9	123.0	125.3
AoANet [25]	37.3	68.1	28.3	37.2	57.9	72.8	124.0	126.2
Up-Down+VC	37.8	69.1	28.5	37.6	58.2	73.3	124.1	126.2
AoANet†+VC	**38.4**	**69.9**	**28.8**	**38.0**	**58.6**	**73.8**	**125.5**	**128.1**

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

VC R-CNN is trained with a mini-batch size of 8, SGD optimizer with weight decay of 0.0001 and momentum of 0.9. In Figure 3, they reports their implementation results with Up-Down [10] and AoANet [11] on the Karpathy test split. As it can be seen in Figure 3, VC implemented versions outperform original Up-Down and AoANet models.

References

[1] T. Wang, Jianqiang Huang, Hanwang Zhang, and Qianru Sun. Visual commonsense r-cnn. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10757–10767, 2020.

[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] 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.

[11] Lun Huang, Wenmin Wang, J. Chen, and Xiao-Yong Wei. Attention on attention for image captioning. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4633–4642, 2019.