Improving Descriptive Deficiencies with a Random Selection Loop for 3D Dense Captioning based on Point Clouds

Shinko Hayashi Zhiqiang Zhang Jinja Zhou

Hosei University, Tokyo, Japan

shinko.hayashi.3r@stu.hosei.ac.jp

Abstract

We propose adding a random selection loop to improve descriptive deficiencies in 3D dense captioning based on point clouds. Our method determines if descriptions are good or bad from the objectness score corresponding to the object, then if there are bad descriptions, use the next random points to evaluate descriptions and loop this process until no more bad descriptions or when these are no longer reduced, so that deficient descriptions (e.g., "Shelf" is described as "Lamp" as shown in Fig.2) decrease. This loop termination condition is configurable since it requires adjusting the execution time of evaluating descriptions to be not too long and its accuracy is not somewhat low. As a result, our work outperforms Scan2Cap [9] ***(14.87% CiDEr@0.5IoU improvement)****.*

1. Introduction

In computer vision research, there are many deep neural networks for visual scene understanding with natural language processing, such as dense captioning which generates sentences to describe how people or animals behave, or an event in the input image. However, images have a restricted field of vision, making it impossible to grasp the physical size and locations of objects. In contrast, 3D dense captioning uses 3D scenes, which allows the actual size and location relation to be determined and described more accurately. Scan2Cap [9] adapts PointNet++ [3] backbone and VoteNet [4] to detect objects in 3D scenes and introduces the Relational Graph module and Context-aware Captioning module to address the issue that object relations are often ignored when describing detected objects in 2D images.

However, when generating or evaluating descriptions, Chen et al. [9] randomly select 40,000 points from a point cloud, which happens to generate poor descriptions since some of those points may not be important in the point cloud. We improve this problem by adding a random selection loop. Since the points used in each loop are different, some descriptions of objects in the scene may also differ, so we store good descriptions based on the objectness score in each loop. By doing so, low objectness scores regardless of random points in the point cloud become less and corresponding objects are more correctly described, thus reducing the number of deficient descriptions.

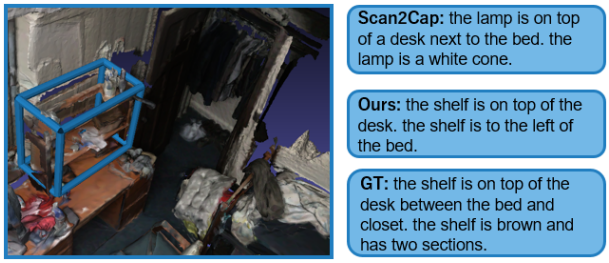

Figure 1. Comparing the results with Scan2Cap [9] in validation scene "scene0222". "Shelf" is sometimes misidentified as "Lamp" in Scan2Cap [9], but in this work, it is described the same as the ground truth.

2. Proposed Approach

To solve the problem of generating deficient descriptions due to using random points that may contain unimportant points in the point cloud as mentioned in section 1, we propose a random selection loop to improve the probability of obtaining high-quality point clouds. We use the objectness score corresponding to the object to determine if descriptions are good or bad when evaluating, then store the information of objects in the list. If the description is bad ($objectness\,score\ <\ 3$), store it and the corresponding object in the bad list; if good, store them in the good list. If there is no element in the bad list, the evaluation scores of descriptions are calculated, but if it exists, this process is done once more with other random 40,000 points, i.e., the next loop is started. By using other points, the objectness score of objects in the bad list may be better in the next loop, so descriptions and corresponding objects are stored on the good list, then previous ones are removed from the bad list. However, the objectness score of objects in the good list may be worse, in which case it is ignored. The goal is to repeat this process until all elements in the bad list are removed, and calculate the evaluation score of descriptions in the good list, however, since the element may not be reduced from a certain loop, the loop is interrupted midway (we interrupt it after 3 loops), in this case, calculate evaluation score of descriptions in the good list and bad list.

3. Experimental Results and Comparison

As in Chen et al. [9], we use the ScanRefer [5] dataset consisting of 51,583 descriptions for 11,046 objects in 800 ScanNet [1] scenes, and the standard metrics for im-

	scene0011				scene0222				scene0704			
	C@0.25IoU	B-4@0.25IoU	M@0.25IoU	R@0.25IoU	C@0.25IoU	B-4@0.25IoU	M@0.25IoU	R@0.25IoU	C@0.25IoU	B-4@0.25IoU	M@0.25IoU	R@0.25IoU
Scan2Cap [9]	49.78	30.34	27.63	51.85	54.51	36.08	24.49	52.73	70.94	30.14	28.16	54.85
Ours	**52.71**	**34.40**	**29.92**	**57.40**	**70.27**	**42.49**	**27.07**	**58.36**	**73.70**	**32.57**	**30.29**	**60.34**

Table 1. The example results for 3 validation scenes "scene0011", "scene0222", and "scene0704" with Scan2Cap [9] and ours. Metrics CiDEr [8], BLEU-4 [6], METEOR [2], and ROUGE [7] are abbreviated as C, B-4, M, R, respectively, with the percentage of the predicted bounding boxes whose IoU with the ground truth is greater than 0.25.

	scene0011				scene0222				scene0704			
	C@0.5IoU	B-4@0.5IoU	M@0.5IoU	R@0.5IoU	C@0.5IoU	B-4@0.5IoU	M@0.5IoU	R@0.5IoU	C@0.5IoU	B-4@0.5IoU	M@0.5IoU	R@0.5IoU
Scan2Cap [9]	31.01	24.99	24.65	45.21	37.60	23.54	20.45	42.81	50.40	22.37	23.27	46.28
Ours	**72.71**	**33.15**	**28.69**	**52.26**	**61.20**	**32.43**	**23.44**	**48.74**	**58.70**	**29.68**	**24.17**	**47.94**

Table 2. The example results for 3 validation scenes "scene0011", "scene0222", and "scene0704" with Scan2Cap [9] and ours. The percentage of the predicted bounding boxes whose IoU with the ground truth is greater than 0.5.

	C@0.25IoU	B-4@0.25IoU	M@0.25IoU	R@0.25IoU	C@0.5IoU	B-4@0.5IoU	M@0.5IoU	R@0.5IoU
Scan2Cap [9]	56.82	34.18	26.29	55.27	39.08	23.32	21.97	44.78
Ours	**63.15**	**35.16**	**28.25**	**58.28**	**44.89**	**26.38**	**23.82**	**47.16**

Table 3. The mean results for all validation scenes with Scan2Cap [9] and ours.

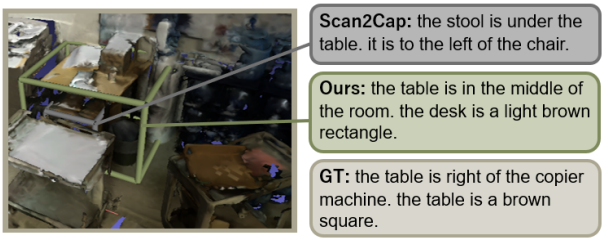

Figure 2. Comparing the results with Scan2Cap [9] in validation scene "scene0704". The object name "Table" in Scan2Cap [9] is sometimes misidentified as a stool under the table, but in this work, it is almost detected as "Table".

age captioning, CiDEr [8], BLEU-4 [6], METEOR [2], and ROUGE [7], combined with Intersection-over-Union (IoU) scores between predicted bounding boxes and the target bounding boxes. We compare our approach with Scan2Cap [9] on the validation split of ScanRefer [5]. At first, we adapt a random selection loop when evaluating one validation scene, and then compare the results on 3 validation scenes, "scene0011", "scene0222" and "scene0704". As shown in Tab. 1 and Tab. 2, our approach has better results, and some evaluation scores are considerably improved (72.71% compared to 31.01% on C@0.5IoU). Then we compare the results on all validation scenes. As shown in Tab. 3, our approach also outperforms Scan2Cap [9] in average evaluation scores. However, as shown in Tab. 4, some scenes don't take much execution time, but most take in our approach, so there are still areas that can be improved.

4. Conclusion and Future Work

In this work, we improve the results of Scan2Cap [9] which is 3D dense captioning based on point clouds. We propose adding a random selection loop, which results in lower objectness scores and the corresponding objects being correctly described regardless of random points in the point cloud, thus improving descriptive deficiencies. However, our approach takes time to evaluate most scenes, so we will improve this problem in future work.

	scene0011		scene0222		scene0704		all scenes	
	C@0.5IoU	Time(s)	C@0.5IoU	Time(s)	C@0.5IoU	Time(s)	C@0.5IoU	Time(s)
Scan2Cap [9]	31.01	**7.9**	37.60	**8.0**	50.40	**7.9**	39.08	**286.1**
Ours	**72.71**	22.9	**61.20**	23.5	**58.70**	15.0	**44.89**	858.2

Table 4. The CiDEr score and execution time for 3 validation scenes "scene0011", "scene0222", and "scene0704" and its mean results for all validation scenes with Scan2Cap [9] and ours. The percentage of the predicted bounding boxes whose IoU with the ground truth is greater than 0.5. Our method gives better results than Scan2Cap [9], but it is time-consuming.

References

[1] M. Savva M. Halber T. Funkhouser A. Dai, A. X. Chang and M. Nießner. ScanNet: Richly-annotated 3D reconstructions of in door scenes. In *CVPR*, 2017.

[2] S. Banerjee and A. Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *ACL*, 2005.

[3] H. Su C. R. Qi, L. Yi and L. J. Guibas. PointNet++: Deep hierarchical feature learning on point sets in a metric space. In *NeurIPS*, 2017.

[4] K. He C. R. Qi, O. Litany and L. J. Guibas. Deep hough voting for 3D object detection in point clouds. In *ICCV*, 2019.

[5] A. X. Chang D. Z. Chen and M. Nießner. ScanRefer: 3D object localization in RGB-D scans using natural language. In *ECCV*, 2020.

[6] T. Ward K. Papineni, S. Roukos and W. Zhu. Bleu: a method for automatic evaluation of machine translation. In *ACL workshop*, 2002.

[7] C. Lin. ROUGE: A package for sautomatic evaluation of summaries. In *Text summarization branches out*, 2004.

[8] C. L. Zitnick R. Vedantam and D. Parikh. CIDEr: Consensus-based image description evaluation. In *CVPR*, 2015.

[9] M. Niessner Z. Chen, A. Gholami and A. X. Chang. Scan2Cap: Context-Aware Dense Captioning in RGB-D Scans. In *CVPR*, 2021.