Improving Visual Relation Detection Using Depth Maps

Sahand Sharifzadeh[1], Sina Moayed Baharlou[2], Max Berrendorf[1], Rajat Koner[1], Volker Tresp[1,3]

[1]Ludwig Maximilian University of Munich, [2]Sapienza University of Rome, [3]Siemens AG

sharifzadeh@dbs.ifi.lmu.de

25th INTERNATIONAL CONFERENCE ON PATTERN RECOGNITION
Milan, Italy 10 | 15 January 2021

1. Introduction

- The goal of Visual Relation Detection is to predict the pairwise relations between objects in an image. We can summarize the context of an image, in form of a scene graph where nodes represent the objects and edges represent the relations or predicates between each pair of object (Figure 1).
- This is a fundamental task in scene understanding and can play an important role in many domains such as recommender systems, visual question answering and decision making. For example, detecting whether a man is *on* a bike or *next to* a bike is a crucial challenge in autonomous driving. Most works in this area rely on explicit image-based object information such as the class labels, bounding boxes and RGB features.

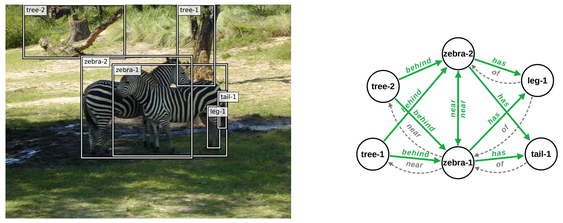

Figure 1: *An example of an image from VG and its corresponding scene graph.*

- We argue that depth maps can additionally provide valuable information about an object's relations as they provide the object's distance from the camera.

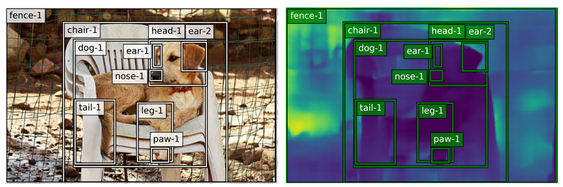

Figure 2: *An example of an image from VG and its generated depth map.*

- Unfortunately, most available image datasets do not provide depth maps, because the acquisition of depth maps is a cumbersome task requiring specialized hardware.
- We tackle this issue by synthetically generating the corresponding pseudo depth maps from 2D images of Visual Genome [1]. This is possible thanks to the large corpora of publicly available datasets of RGB-D pairs. Using a fully convolutional neural network and the set of RGB-D pairs from NYU-Depth-v2 [2], we can learn the mapping function of RGB images to their corresponding depth maps. We can then apply this network to the images from VG, generating their corresponding depth maps.

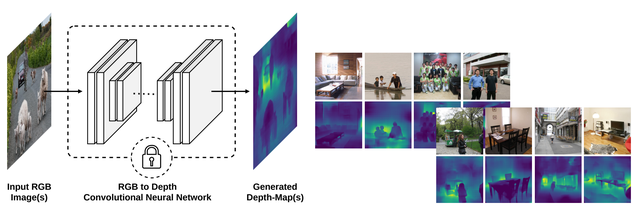

Figure 3: *We generate the corresponding depth maps VG images with a pre-trained fully convolutional neural network, and release it as VG-Depth.*

- We release the depth maps that are generated from the Visual Genome, as an extension to it, calling it **VG-Depth**.

2. Framework

- The object information extracted from depth maps and RGB images, i.e. **class labels** c, **location vectors** l, **image-based** v and **depth-based** d features, are the basis for relation detection in our simple yet effective framework.
- Note that unlike previous works in RGB-D object classification or image segmentation, the convolutional neural networks that we have employed for feature extraction are not shared between RGB and Depth modalities. While those works aim to get similar and complementary features from both modalities, we expect to extract different information from each. Therefore, unlike other works, we train a separate feature extractor CNN directly on the depth maps and specifically for the task of relation detection.

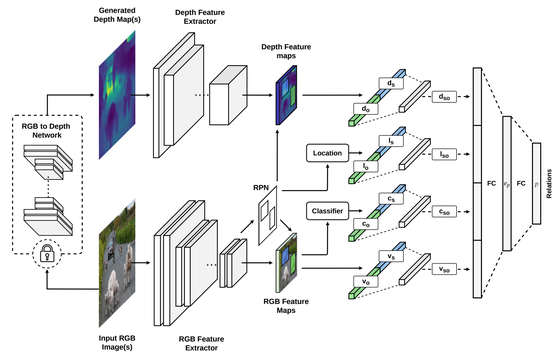

Figure 4: *Our full architecture fuses pairwise object information in order to classify the relation.*

3. Evaluation

- We test our approach on the Visual Genome and extensively study the effect of using different sources of object information in visual relation detection.
- We evaluated our approach using the standard [Micro] Recall@K Metric. This metric cannot properly reveal the improvements of under-represented relations in highly imbalanced datasets such as VG. For example the predicate *walking on* appears 648 times in the VG test set, while the predicate *wearing* appears 20,148 times. This means that the correct classification of *wearing* can highly affect the Micro Recall@K metric and prevent us from noticing the lack of accuracy in predicting *walking on*.
- We address this issue by proposing **Macro Recall@K**, where we compute the mean over Micro Recall@K per predicate, thereby eliminating the effect that over-represented classes have on Micro Recall@K:

$$\text{MACRO RECALL@K} = \sum_{(s,p,o)\in\mathcal{T}_p} \frac{\text{MICRO R@K}(p)}{|\mathcal{T}_p|} \tag{1}$$

- Table 5 presents our experimental results. We can see that our full model with depth maps, achieves the highest accuracy. It is interesting to note that when using only depth maps we can already achieve a significant accuracy, emphasizing the value of relational information that are stored within the depth maps alone. By comparing *Ours-v* to *Ours-v, d*, we can observe the improvements that depth maps bring. For an in-depth analysis of these results please refer to our paper.
- To get a better intuition of the improvements that we gain after including depth maps, we plotted the changes in prediction accuracy for each predicate in the Figure 6.
- Figure 7 shows some of our qualitative results.

	Strategy	Macro			Micro		
	Task	Predicate Pred.			Predicate Pred.		
	Metric	R@100	R@50	R@20	R@100	R@50	R@20
models	VTransE [Zhang *et al.*, 2017]	-	-	-	62.87	62.63	-
	Yu's-S [Yu *et al.*, 2017]	-	-	-	49.88	-	-
	Yu's-S+T [Yu *et al.*, 2017]	-	-	-	55.89	-	-
	IMP [Xu *et al.*, 2017]	-	-	-	53.00	44.80	-
	Graph R-CNN [Yang *et al.*, 2018b]	-	-	-	59.10	54.20	-
	NM [Zellers *et al.*, 2018]	14.39	13.20	10.25	67.10	65.20	58.50
ablations	Ours - d	9.51	8.46	6.35	54.72	51.90	43.86
	Ours - c	15.65	13.09	8.56	64.82	60.54	49.89
	Ours - v	13.88	12.24	8.99	61.72	58.50	50.41
	Ours - l	5.19	4.66	3.57	49.07	46.13	37.48
	Ours - v, d	15.47	14.04	10.83	62.88	60.52	53.07
	Ours - l, v, d	15.76	14.40	11.07	63.06	60.83	53.55
	Ours - l, c, d	21.67	19.56	15.12	67.97	66.09	59.13
	Ours - l, c, v	19.16	17.72	13.93	67.94	66.06	59.14
	Ours - l, c, v, d	**22.72**	**20.74**	**16.40**	**68.00**	**66.18**	**59.44**

Figure 5: *Quantitative ablations studies on Visual Genome, using different object information.*

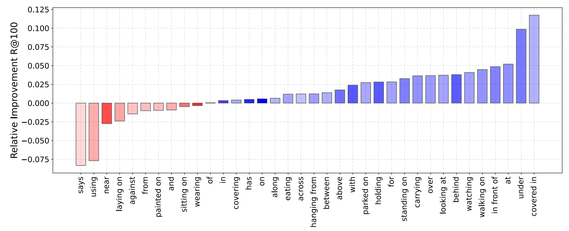

Figure 6: *Per predicate improvements after employing depth maps. We used darker shades for over-represented classes and lighter shades for under-represented ones. For example we can see that in general the accuracy of relations including the predicates such as under, in front of and behind has been improved.*

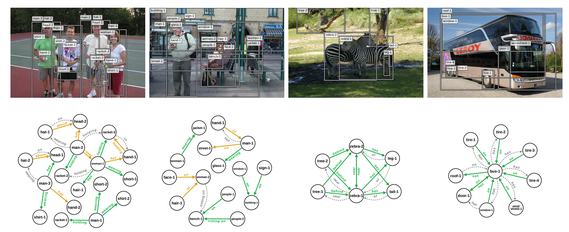

Figure 7: *Qualitative examples of predictions using our model.*

4. Conclusion

- We perform an extensive study on the effect of using different sources of object information in visual relation detection. We show in our empirical evaluations using the VG dataset, that our model can outperform competing methods by a margin of up to 8
- We release a new synthetic dataset VG-Depth, to compensate for the lack of depth maps in Visual Genome.
- We propose Macro Recall@K as a competitive metric for evaluating the visual relation detection performance in highly imbalanced datasets such as Visual Genome.

References

[1] Ranjay Krishna et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73, 2017.

[2] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *European conference on computer vision*, pages 746–760. Springer, 2012.