Grounding Scene Graphs on Natural Images via Visio-Lingual Message Passing

Aditay Tripathi[1] Anand Mishra[2] Anirban Chakraborty[1]
[1]Indian Institute of Science [2] Indian Institute of Technology Jodhpur
{aditayt,anirban}@iisc.ac.in mishra@iitj.ac.in

https://iiscaditaytripathi.github.io/sgl/

A. Dataset details

Dataset	#Samples	#Predicates	#Categories
VG-FO	93,323/40,124	40/40	150/150
VG-PO	69,917/39,699	40/40	125/150
VRD	39,699/6,869	70/70	100/100
COCO-stuff	74,121/3,074	6/6	183/183
SG	4,000/1,000	68/68	166/166

Table 1. Train/test statistics of the dataset used in our evaluation.

B. Scalability Analysis

Given N entities on the query graph and a set of object proposals $\mathcal{R}_u$ on the image containing M proposals, the augmented directed graph between proposal and query graph takes $O(MN)$-space. However, this complexity is nearly-linear with respect to the number of proposals (M) since even for complex queries $N \ll M$. Further, in theory, if the region proposals are fully connected, the proposal graph would take $O(M^2)$-space. However, in reality, the number of actual connections is restricted by the plausible set of relationships constrained by the visual semantic association between the objects in the scene. In fact, in our experiments, the actual number of edges in the proposal graph is just a tiny fraction of M^2 (refer Table 2).

Further, to illustrate the scalability of our proposed model with respect to the number of edges in the query graph, we compute the average inference time (in seconds) on NVIDIA Quadro-8000. The inference time does not increases significantly as the number of edges in the query graph increases from 1 to 8 (Refer Table 3), with a significant under-use of the GPU memory.

C. Limitations

Although VL-MPAG Net does not impose any theoretical limitations on grounding very-large-size scene graphs; our work is limited by availability of a balanced dataset with dense scene graph annotations. This limits us to show results of grounding scene graph with maximum of eight-edges only.

Proposals($\mathcal{R}_u$)	Average number of edges	$\|\mathcal{R}_u\|^2$
128	674.2	16,384
256	1,444.6	65,536
512	2,784.4	262,144
1024	5,051.9	1,048,576

Table 2. The number of edges in the proposal graph is averaged over all the queries in the test set of VG-FO dataset. The average number of edges in the proposal graph selected by VL-MPAG Net is significantly smaller than $|\mathcal{R}_u|^2$.

#Edges	1	2	3	4	5	6	7	8
Time (s)	0.12	0.12	0.12	0.12	0.13	0.13	0.13	0.14

Table 3. The Average inference time (shown in seconds) increases very slowly with the number of edges in the query graph. It shows that the proposed model is scalable with respect to size of the query graph.

D. Implementation Details

We implemented the model using the PyTorch v1.9.1 and PyTorch-geometric libraries with CUDA 10.2. The stochastic gradient descent with a momentum of 0.9 is used to train the models on one NVIDIA-Quadro RTX-8000. The model is end-to-end trained with the learning rate as 0.0001 and 0.02 for the fasterRCNN backbone and the remaining network, respectively with a decay by a factor of 0.1 after every 2 epoch with a maximum of 10 epochs. In our experiments, we use $m^+ = 0.3$ and $m^- = 0.7$ in Equation 10. The implementation of this work is available at **https://iiscaditaytripathi.github.io/sgl/**. We also provided splits for two representative datasets used in this paper.

E. Additional Visual Results

Please refer to the following two pages for additional visual results of the proposed model.

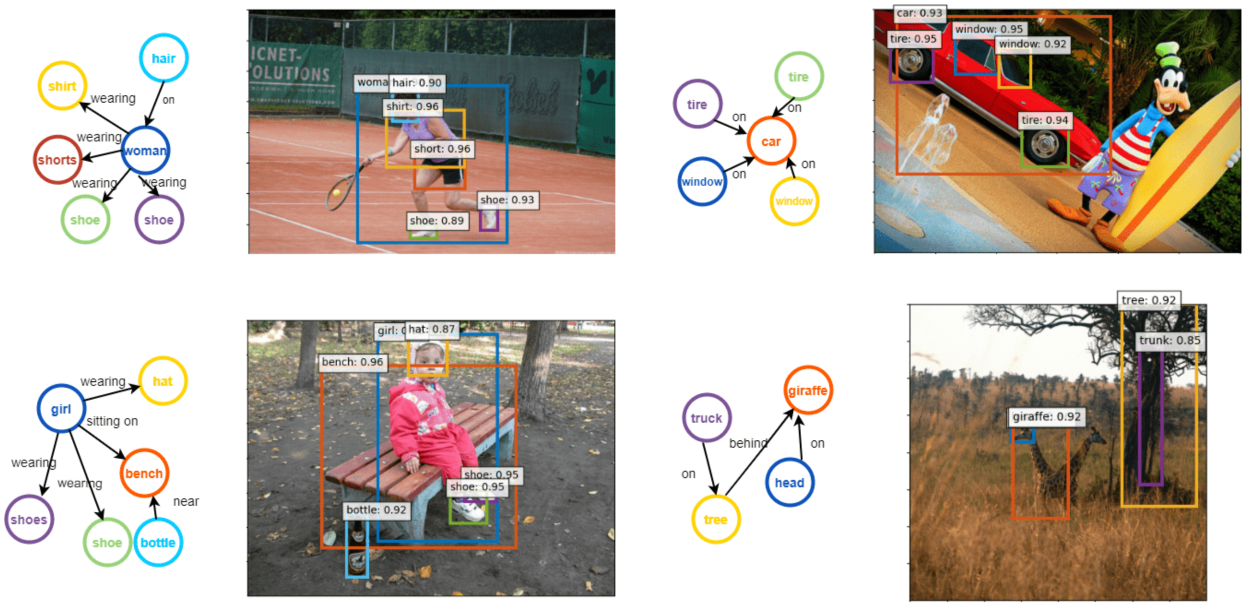

Figure 1. More results: Grounding Scene Graphs on Image.[**Best viewed in color**]

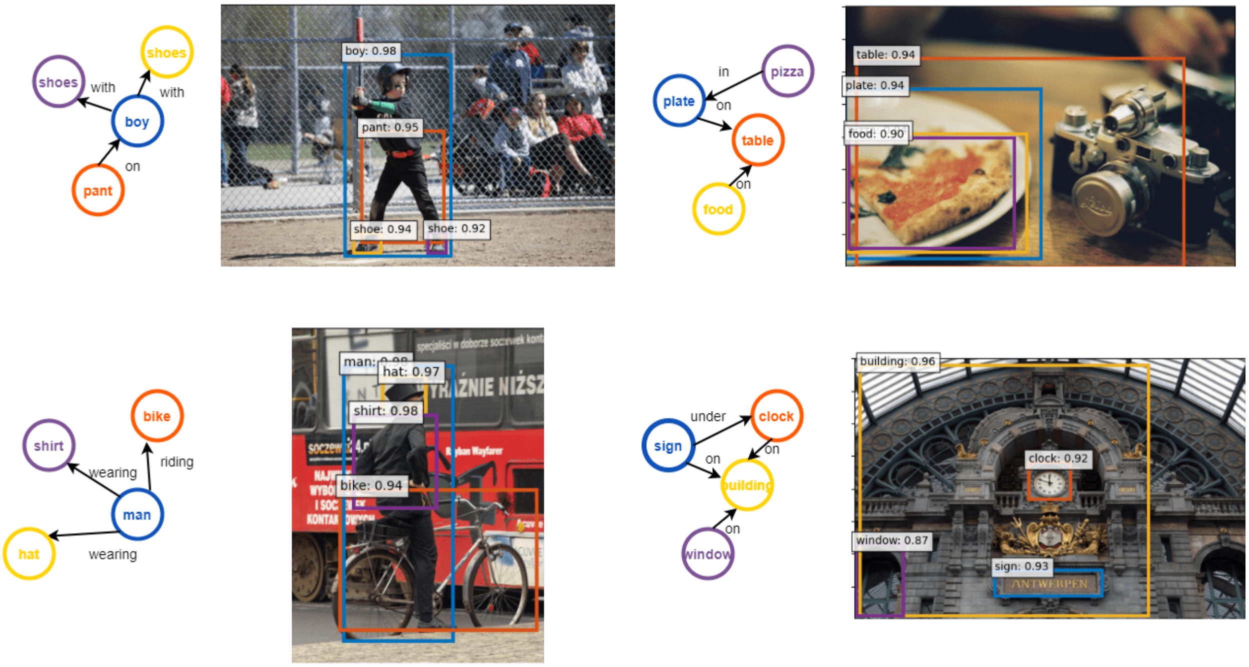

Figure 2. More results: Grounding Scene Graphs on Image.[**Best viewed in color**]

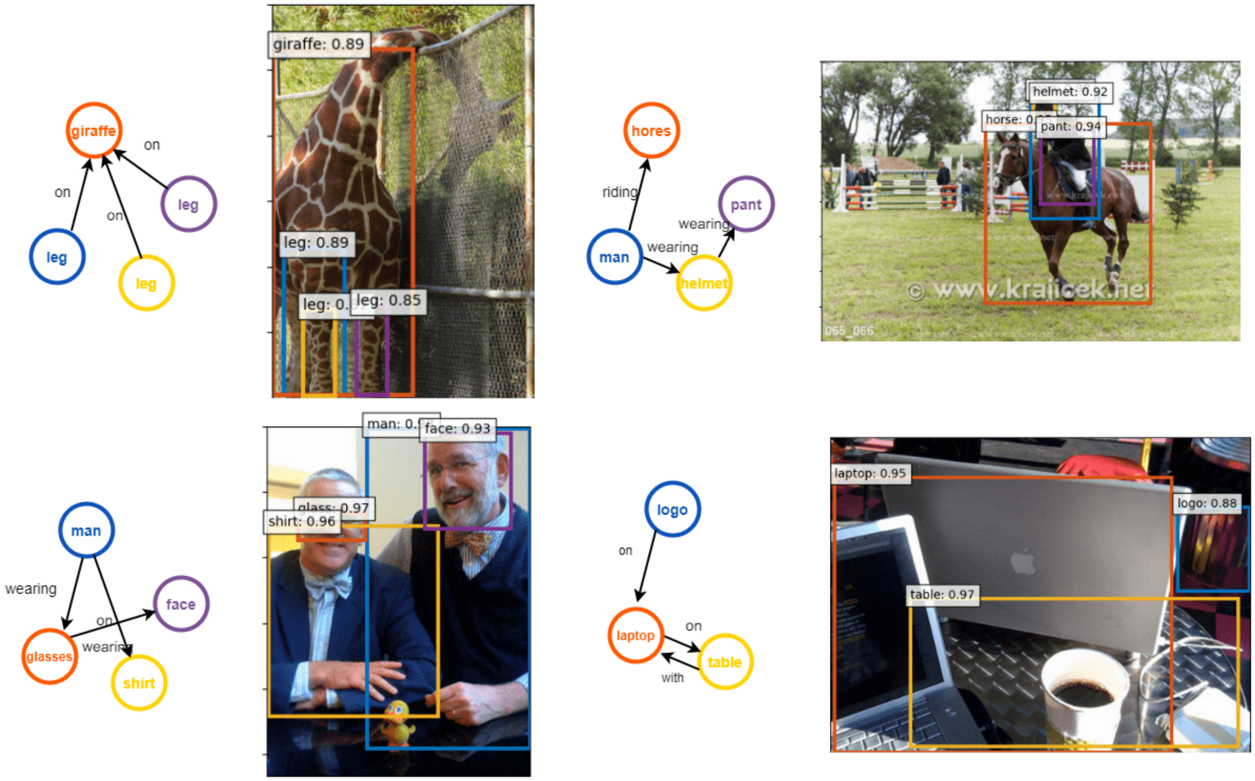

Figure 3. More results: Grounding Scene Graphs on Image. Last row correspond to some of the failure cases. In the first the model is confusing between two men and in the second image, the model is not able to detect the logo.[**Best viewed in color**]