Differentiable Scene Graphs Supplemental Materials

Moshiko Raboh[1*], Roei Herzig[1*], Jonathan Berant[1,4], Gal Chechik[2,3], Amir Globerson[1]
[1]Tel Aviv University, [2]Bar-Ilan University, [3]NVIDIA Research, [4]AI2

This supplementary material includes: (1) Model implementation details. (2) Details about the reasoning component in two steps ablation module.

1. Model Details

The model in Sec. 3.2 is implemented as follows.

Object Detector and Relation Feature Extractor. For object detection, we used Faster-RCNN with a 101-layers ResNet backbone. The RPN was trained with anchor scales of $\{4, 8, 16, 32\}$ and aspect ratios $\{0.5, 1, 2\}$. RPN proposals were filtered by non-maximum suppression with IOU-threshold of 0.5 and score higher than 0.8. We use at most 32 proposals per image. Both the entity features $\boldsymbol{f}_i$ and the relation features $\boldsymbol{f}_{i,j}$ are first extracted from the convolutional network feature map by the ROI-Align layer as $7\times7\times2048$ features. They are then reduced to a $7\times7\times512$ by convolution layer of size 1×1 and finally reduced to 1×512 by an average pooling layer.

Referring Relationship Classifier. The referring relationship classifier F_{RRC} is a fully-connected network with two layers of 512 hidden units each.

Bounding Box Refinement. The box refinement model applies a linear function to $\boldsymbol{z}_i$ to obtain four outputs $[dx, dy, dw, dh]$. Denote the RPN box by $[x, y, w, h]$. The refined box is then: $[dx \cdot w + x, dy \cdot h + y, e^{dw} \cdot w, e^{dh} \cdot h]$ (as in the correction used by Faster-RCNN).

Computational Estimation. Our model creates a graph with n nodes for objects and n^2 edges for relations. In the datasets we analyzed, using $n = 32$ objects within an image is sufficient. Adding the DSG component has a limited effect on complexity and run time. Specifically, as shown in Tab. 1, the DSG component adds 4M parameters and up to 1.5G operations (when $n = 32$) which is **only 10% of the parameters** and number of operations of the backbone network ($\sim$40M parameters and $\sim$15G operations). Adding DSG **increases training time by only 15%**. This is largely thanks to the fact that all n^2 relations can parallelized.

Differentiable Scene Graph Generator. We next describe the module that takes as input features $\boldsymbol{z}_i$ and $\boldsymbol{z}_{i,j}$ extracted by the RPN and outputs a set of vectors $\boldsymbol{z}'_i, \boldsymbol{z}'_{ij}$ corresponding to Differentiable Scene-Graph over entities and relationships in the image. For this model, we use the Graph Permutation Invariant (GPI) architecture introduced in [1]. A key property of this architecture is that it is invariant to permutations of the input that do not affect the labels.

	DSG Generator	Resnet101
Trainable parameters	$< 4M$	$> 40M$
Number of operations	$< 1.5G$	$> 15G$
	DSG	DSG -SG
Running time [sec]	0.054	0.045
Training time [sec]	0.19	0.165

Table 1. Analysis of running/training time and computational resources of DSGs.

The GPI transformation is defined as follows. First, the set of all input features is summarized via a permutation-invariant transformation into a single vector $\boldsymbol{g}$:

$$\boldsymbol{g} = \sum_{i=1}^{n} \boldsymbol{\alpha}(\boldsymbol{z}_i, \sum_{j \neq i} \boldsymbol{\phi}(\boldsymbol{z}_i, \boldsymbol{z}_{i,j}, \boldsymbol{z}_j)) \quad (1)$$

Here $\boldsymbol{\alpha}$ and $\boldsymbol{\phi}$ are fully connected networks. Then the new representations for entities and relations are computed via:

$$\boldsymbol{z}'_k = \boldsymbol{\rho}^{entity}(\boldsymbol{z}_k, \boldsymbol{g}) \ , \ \boldsymbol{z}'_{k,l} = \boldsymbol{\rho}^{relation}(\boldsymbol{z}_{k,l}, \boldsymbol{g}) \quad (2)$$

where $\boldsymbol{\rho}$ above are fully connected networks.

The three networks $\boldsymbol{\phi}$, $\boldsymbol{\alpha}$ and $\boldsymbol{\rho}$, described in GPI architecture are two fully-connected layers with 512 hidden units. The output size of ϕ and α is 512, and of $\boldsymbol{\rho}$ is 1024. We used the version with integrated attention mechanism replacing the sum operations in equation 1.

2. Model Ablations

Additional details about the TWO STEP model: Recall that in Two-step model a scene graph is first predicted, followed by a reasoning module. The reasoning module gets as an input a query $\langle$*subject*, *relation*, *object*$\rangle$ and a scene graph and outputs the nodes that represents the subject and the nodes that represents the object. In case the triplet $\langle$*subject*, *relation*, *object*$\rangle$ exists in the scene graph, the reasoning module simply returns the involved nodes. Otherwise, it selects the triplet in the scene graph that has the

highest probability to be the required triplet according to the probabilities provided by the scene graph.

References

[1] R. Herzig, M. Raboh, G. Chechik, J. Berant, and A. Globerson. Mapping images to scene graphs with permutation-invariant structured prediction. In *Advances in Neural Information Processing Systems (NIPS)*, 2018.