Layout Composition from *Attributed* Scene Graphs

Subarna Tripathi
Intel AI Lab,
San Diego, USA
{subarna.tripathi@intel.com

Anahita Bhiwandiwalla
Intel AI Lab,
Santa Clara, USA
{subarna.tripathi@intel.com

Scene graphs are a structured representation, with objects as nodes with *attributes*, and edges marking the semantic relationship between objects. Generating images from scene graphs, an emerging research direction, usually is a two-step process. First creating a scene layout using graph convolutional networks (GCN) and next generating a realistic RGB image from that layout. None of the existing methods performing scene graph to image generation [3, 5, 8] or layout generation [4, 7] use the attributes associated with the nodes. For example, for generating an image of a table, the system never gets *round* or *rectangular* as input. Here, we take one step forward to process *attributed* scene graphs while creating scene layout.

Most work on scene graphs uses the Visual Genome (VG) dataset [2] which provides human-annotated scene graphs. VG provides only bounding boxes for object instances but not their segmentation masks. To overcome the known issues of incomplete and incorrect annotations in VG, researchers often use synthetic scene graphs from COCO stuff [1], but those graphs are limited to simple geometric relationships (above, below, left, right, inside, surrounding). We also note that although COCO stuff has segmentation masks annotation, it lacks attribute annotations. In this work, we exploit supervision of segmentation [1] and attributes [6] for COCO instances from different sources in a layout composition training framework.

Training with Segmentation and Attributes: Since no scene graph datasets exist with both segmentation and attribute annotation, we combine COCO stuff and COCO attribute together. We divide the attributes set into two separate categories. Shape-altering attributes such as *round, rectangular, standing, running, square* and color-altering attributes such *red, white, smiling, shiny*. We create synthetic scene graph with pair-wise geometric relationships from COCO stuff. Additionally we curate shape-altering attributes for the objects belonging to the super categories of person, vehicle, animal and food by matching instance ids from both the datasets. COCO stuff and COCO attributes have annotations on train2017 and train2014 split of COCO respectively. If an instance does not have any attribute, we use a *sentinel*. Figure 1 shows how the GCN object embedding vector is used to predict the bounding box and each predicted bounding box along with category and attribute word vectors are used to predict the segmentation mask for each object. All of them are combined to form the scene layout.

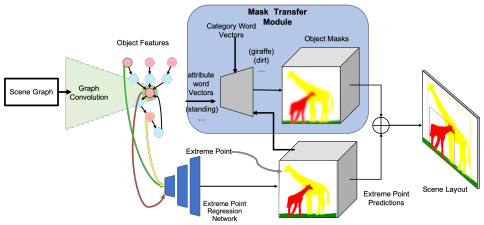

Figure 1: Mask prediction module uses the location, category and attribute word vectors.

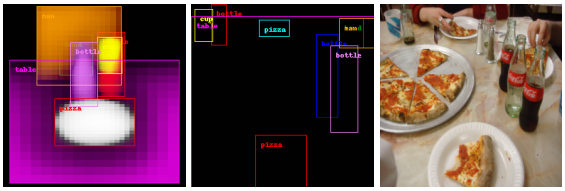

Figure 2: Generating object masks for VG scene graphs. Ground truth (GT) bounding boxes and image are shown for reference. GT segmentation masks not available.

Weakly Supervised Scene Layout Composition: Unlike the localization (bounding box or extreme points [7]) prediction net, the mask prediction network does not use the graph embedding vectors directly. Thus, we can generate precise scene layouts for VG scene graphs by leveraging weak supervision of bounding boxes, and keeping the weights of the mask generation network unchanged after training using COCO. Fig 2 shows an example of such generated scene layout from a VG scene graph where object category, attribute and relationship vocabulary are different from COCO.

Challenges and future work: Intersection-Over-Union (IOU) or even the recently proposed metric Relation Score [7, 8] can not measure compliance with an *attributed* scene graph. We need new metrics for this task. Proposed framework for handling *attributed* Scene Graph currently only

works for high frequency attributes. Long-tailed attribute distribution is a challenge, and we leave dealing with low-frequency attributes as future work.

References

[1] H. Caesar, J. Uijlings, and V. Ferrari. Coco-stuff: Thing and stuff classes in context. In *CVPR*. IEEE, 2018. 1

[2] R. K. et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *CoRR*, abs/1602.07332, 2016. 1

[3] J. Johnson, A. Gupta, and L. Fei-Fei. Image generation from scene graphs. *CVPR*, 2018. 1

[4] A. A. Jyothi, T. Durand, J. He, L. Sigal, and G. Mori. LayoutVAE: Stochastic scene layout generation from a label set. *ICCV*, 2019. 1

[5] O. Ashual and L. Wolf. https://www.youtube.com/watch?v=V2v0qEPsjr0tm, 2019. [ICCV 2019, Accessed: 2019-08-14]. 1

[6] G. Patterson and J. Hays. Coco attributes: Attributes for people, animals, and objects. *European Conference on Computer Vision*, 2016. 1

[7] S. Tripathi, S. N. Sridhar, S. Sundaresan, and H. Tang. Compact scene graphs for layout composition and patch retrieval. *CVPRW*, 2019. 1

[8] D. M. Vo and A. Sugimoto. Visual-relation conscious image generation from structured-text, 2019. 1