Hierarchical Relationships: A New Perspective to Enhance Scene Graph Generation

Bowen Jiang, Camillo J. Taylor
University of Pennsylvania
Philadelphia, PA, 19104, USA
{bwjiang, cjtaylor}@seas.upenn.edu

Abstract

This paper presents a finding that leveraging the hierarchical structures among labels for relationships and objects can substantially improve the performance of scene graph generation systems. The focus of this work is to create an informative hierarchical structure that can divide object and relationship categories into disjoint super-categories in a systematic way. Specifically, we introduce a Bayesian prediction head to jointly predict the super-category of relationships between a pair of object instances, as well as the detailed relationship within that super-category simultaneously, facilitating more informative predictions. The resulting model exhibits the capability to produce a more extensive set of predicates beyond the dataset annotations, and to tackle the prevalent issue of low annotation quality. While our paper presents preliminary findings, experiments on the Visual Genome dataset show its strong performance, particularly in predicate classifications and zero-shot settings, that demonstrates the promise of our approach.

1 Introduction

This work considers the scene graph generation problem [6, 12, 22, 24, 25, 8, 18] that deduces the objects in an image and their interconnected relationships. Unlike object detectors [15, 17, 1] which focus on individual object instances, scene graph models represent the entire image as a graph, where each object instance is a node and the relationship between a pair of nodes is a directed edge.

Existing literature has addressed the nuanced relationships among objects within visual scenes by designing complicated architectures [4, 11, 23, 3, 16, 2], but a gap remains due to the neglect of inherent hierarchical information in relationship categories, resulting in an incomplete understanding of object interactions. We adopt the definitions in Neural Motifs [25] to divide predominant relationships in scene graphs into *geometric*, *possessive*, and *semantic* super-categories and show how these categories can be explicitly utilized in a network.

In this work, we propose a novel classification scheme inspired by Bayes' rule, jointly predicting relationship super-category probabilities and conditional probabilities of relationships within each super-category. For each directed edge, the top one predicate under each super-category will participate in the ranking, although we still evaluate the recall scores within the top k most confident predicates in each image. Experimental results on the Visual Genome dataset [9] demonstrate that incorporating hierarchical relationship reasoning can enhance the performance of a baseline model by a large margin, indicating a promising and interesting preliminary finding.

2 Scene Graph Construction

Object detection backbone Our system adopts the widely-used two-stage design [22, 24, 25, 11, 4]. We leverage the Detection Transformer (DETR) [1] to predict object bounding boxes and labels. It has

NeurIPS 2023 New Frontiers in Graph Learning Workshop (NeurIPS GLFrontiers 2023).

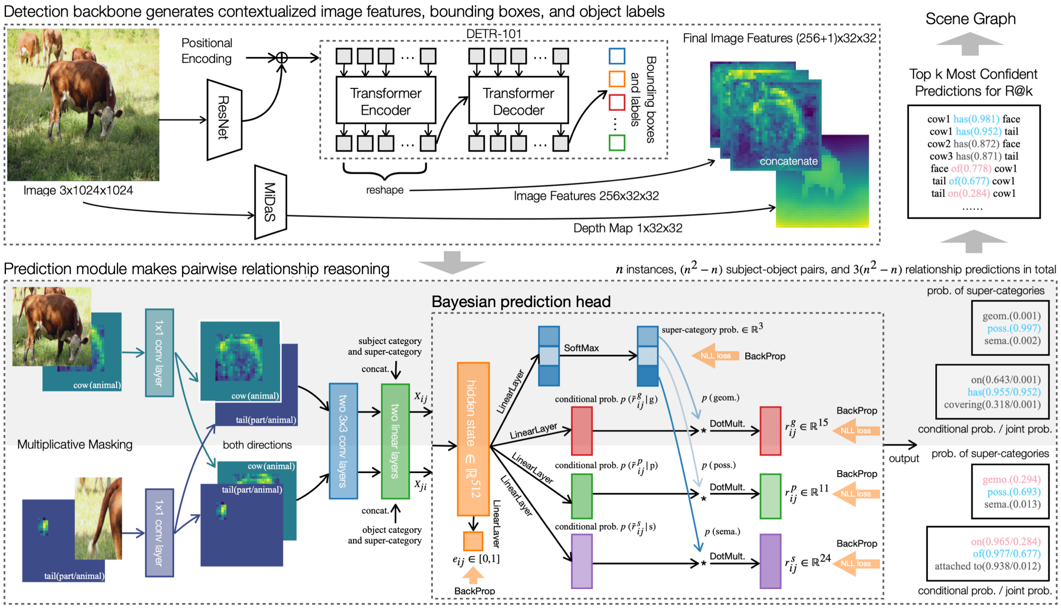

Figure 1: Illustration of the model architecture we construct to showcase the effectiveness of our Bayesian classification head that utilizes the relationship hierarchies. It performs pairwise relationship reasoning for each directed edge in the graph and predicts the probability distributions of both relationship super-categories and detailed categories within each super-category.

a ResNet-101 feature extraction backbone [5] and a transformer encoder [1, 21] that contextualizes the feature space with global information. Its output still preserves its spatial dimensions and serves as the image features $\boldsymbol{I} \in \mathbb{R}^{h \times s \times t}$, where h is the hidden dimensions, and s and t denote its spatial dimensions. We also employ the MiDaS [14] single-image depth estimation network to provide depth maps $\boldsymbol{D}$ for input images. The final image features $\boldsymbol{I}' = \text{concat}\{\boldsymbol{I}, \boldsymbol{D}\} \in \mathbb{R}^{(h+1) \times s \times t}$ serve as the inputs to the subsequent networks for relationship reasoning.

Direction-aware masking After extracting $\boldsymbol{I}'$, the model considers each pair of object instances. We construct a combined feature tensor by masking $\boldsymbol{I}'$ with the bounding boxes $\boldsymbol{M}_i, \boldsymbol{M}_j \in \mathbb{R}^{s \times t}$ of the subject and object, resulting in two feature tensors $\boldsymbol{I}'_i, \boldsymbol{I}'_j \in \mathbb{R}^{(h+1) \times s \times t}$. The order of these two tensors matters. For example, choosing <*bike, has, wheel*> or <*wheel, of, bike*> depends on which instance is considered as the subject and which one as the object. Therefore, we avoid using a union mask of the subject-object pair but instead, perform two separate passes through the model, depicted in Figure 1. One pass considers $\boldsymbol{I}'_i$ as the subject and concatenates them as $\boldsymbol{I}'_{ij}$, while the other pass swaps their roles to yield $\boldsymbol{I}'_{ji}$. Each one is fed to subsequent convolutional layers, flattened, concatenated with four one-hot vectors encoding the categories and super-categories associated with the subject and the object, and reduced to a 512-dimensional hidden vector denoted as $\boldsymbol{X}_{ij}$.

Bayesian prediction head Inspired by Baye's rule, the head predicts from $\boldsymbol{X}_{ij}$ a scalar connectivity score $0 \leq e_{ij} \leq 1$, the probability of three relationship super-categories $\boldsymbol{r}^{\text{sup}}_{ij} = [\mathbf{p}(\text{geo}),\ \mathbf{p}(\text{pos}),\ \mathbf{p}(\text{sem})] \in \mathbb{R}^3$, and the conditional probability distributions $\{\boldsymbol{r}^{\text{sub}}_{ij} = \mathbf{p}(\tilde{\boldsymbol{r}}^{\text{sub}}_{ij}|\text{sub}) \mid \text{sub} \in [\text{geo}, \text{pos}, \text{sem}]\}$ under each super-categories, respectively, each of which is computed by multiplying its conditional probability vector with the associated super-category probability.

$$e_{ij} = \text{Sigmoid}\left\{\boldsymbol{X}^{\top}_{ij}\boldsymbol{W}^{\text{conn}}\right\} \tag{1}$$

$$\boldsymbol{r}^{\text{sup}}_{ij} = [\mathbf{p}(\text{geo}),\ \mathbf{p}(\text{pos}),\ \mathbf{p}(\text{sem})] = \text{SoftMax}\left\{\boldsymbol{X}^{\top}_{ij}\boldsymbol{W}^{\text{sup}}\right\} \tag{2}$$

$$\boldsymbol{r}^{\text{sub}}_{ij} = \text{SoftMax}\left\{\boldsymbol{X}^{\top}_{ij}\boldsymbol{W}^{\text{sub}} * \mathbf{p}(\text{sub})\right\}, \text{for sub} \in [\text{geo}, \text{pos}, \text{sem}], \tag{3}$$

where $*$ is the scalar product and all $\boldsymbol{W}$s are the learnable parameter tensors of linear layers.

To train the Bayesian classification head, we apply one cross-entropy loss to the super-categories and another cross-entropy loss to the detailed relationships under the ground-truth super-category. Furthermore, we use a supervised contrastive loss [7] shown in Equation 4 to minimize the distances between hidden states corresponding to the same relation class (set $P(ij)$), while maximizing the distances between those from different relationship classes (set $N(ij)$).

$$\mathcal{L}_{\text{contrast}} \sim \sum_{p \in P(ij)} \log \frac{\exp\left(\boldsymbol{X}_{ij}^{\top} \boldsymbol{X}_p / \tau\right)}{\sum_{n \in N(ij)} \exp\left(\boldsymbol{X}_{ij}^{\top} \boldsymbol{X}_n / \tau\right)}, \text{ where } \tau \text{ is the temperature.} \tag{4}$$

The model yields three predicates on each edge, one from each disjoint super-category, which maintains the exclusivity among predicates within the same super-category. While keeping the same evaluation metrics that focus on recall scores within the top k most confident predicates in an image, all three predicates from each edge will participate in the ranking. Because there will be three times more candidates, we are not trivially relieving the graph constraints [13, 25] to make the task simple.

We find it common for one or two predicates from the same edge to appear within the top k of the ranking, providing potentially matched solutions at disjoint super-categories to enhance the performance, while pushing more valueless predicates out of the rank. This design leverages the super-category probabilities to guide the network's attention toward the appropriate conditional output heads, enhancing the interpretability and performance of the system. Furthermore, the use of separate but smaller linear layers does not exacerbate the number of parameters or scalability issues.

3 Experiments

Dataset and training techniques Our experiments are conducted on the Visual Genome dataset [9], following the same pre-processing procedures outlined in [22]. We filter out the top 150 object categories and 50 relationship predicates, resulting in nearly $75.7k$ training images and $32.4k$ testing images. We adopt the DETR pretrained by [10] and freeze its parameters throughout all our experiments. The subsequent model is trained using SGD with a learning rate of 1e-5, a step scheduler to reduce the learning rate by 10 at the third epoch, and a batch size of 16 for 3 epochs on four NVIDIA V100 GPUs. The average inference time is 0.16 seconds per image.

Evaluation metrics We assess performance using R@k and mR@k metrics [12, 20]. R@k measures the recall within the top k most confident predicates per image, while mR@k computes the average across all relationship classes. We conduct three tasks: (1) Predicate classification (PredCLS) predicts relationships with known bounding boxes and labels. (2) Scene graph classification (SGCLS) only assumes known bounding boxes. (3) Scene graph detection (SGDET) has no prior knowledge.

Results We compare our model with state-of-the-art methods and include preliminary ablation studies in Table 1 and 2. The most important comparison is between "ours" and "ours [a]" in Table 1, where we compare a model with a flat classification head and a model with the Bayesian classification head. The flat model without the hierarchical structure achieves inferior scores, as expected. However, despite the simplicity of our architectural design, adding the hierarchical relationships enhances the scores by a large margin particularly on the PredCLS task and allows our model to achieve a competitive performance. The approach also exhibits strong zero-shot performance in Table 3.

Figure 2 illustrates some predicted scene graphs. Each edge is annotated with the top predicate from three distinct super-categories. It is intriguing to discover that there are numerous reasonable predicates aligned with human intuitions but not annotated in the dataset, marked in blue. There are also scenarios where the top predicate under the second most likely super-category is picked by the dataset as the ground truth. Conventional algorithms would classify this edge as a false negative, but our approach acknowledges it as long as it still appears within the top k of the ranking. We strongly believe that creating an extensive set of predicates is beneficial for practical scene understanding.

4 Discussion and Future Work

We present a straightforward yet powerful scene graph generation algorithm that effectively exploits relationship hierarchies. The resulting system adopts a Bayesian prediction head, enabling simultaneous prediction of the super-categories and specific relationships within each super-category. This

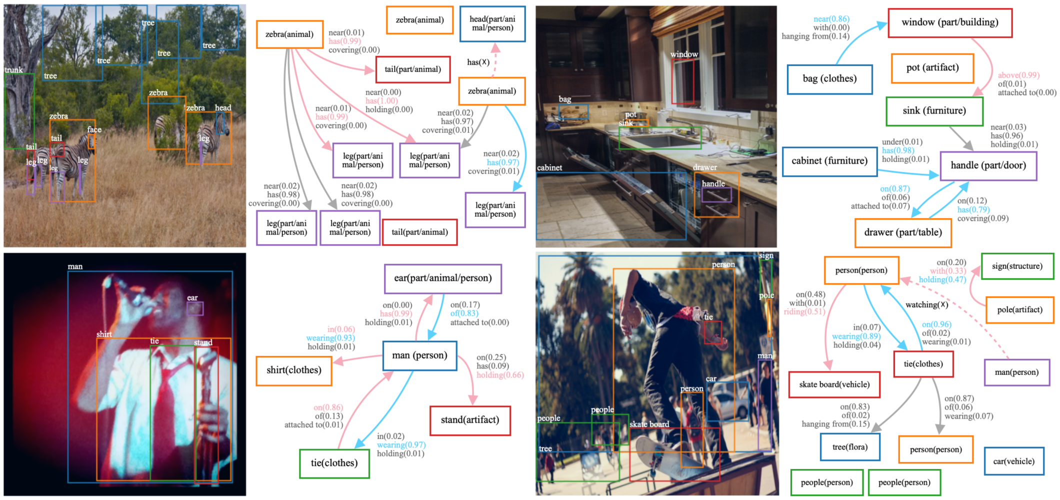

Figure 2: Examples of generated scene graphs. We only show the most confident edges with their top one predicates under all three super-categories and the super-category probabilities. We sketch (1) solid pink arrow: true positives. (2) dotted pink arrow: false negatives. (3) solid blue arrow: reasonably true positives not yet annotated in the dataset. (4) solid gray arrow: false positives.

Table 1: Test results and ablation studies. [a] means hierarchical relationship.

	PredCLS			SGCLS			SGDET		
Methods	R@20	R@50	R@100	R@20	R@50	R@100	R@20	R@50	R@100
NeuralMotifs [25]	58.5	65.2	67.1	32.9	35.8	36.5	21.4	27.2	30.3
HC-Net [16]	59.6	66.4	68.8	34.2	36.6	37.3	22.6	28.0	31.2
GPS-Net [11]	60.7	66.9	68.8	36.1	39.2	40.1	22.6	28.4	31.7
BGT-Net [4]	60.9	67.3	68.9	38.0	40.9	43.2	23.1	28.6	32.2
RelTR [3]	63.1	64.2	-	29.0	36.6	-	21.2	27.5	-
ours	61.1	73.6	78.1	30.6	36.0	37.6	22.8	29.3	32.1
ours w/o [a]	56.6	66.6	69.1	29.5	33.8	34.8	20.2	26.3	28.1
	mR@20	mR@50	mR@100	mR@20	mR@50	mR@100	mR@20	mR@50	mR@100
NeuralMotifs [25]	11.7	14.8	16.1	6.7	8.3	8.8	4.9	6.8	7.9
Motif+EB [18]	14.2	18.0	19.5	8.2	10.2	10.9	5.7	7.7	9.3
GPS-Net [11]	17.4	21.3	22.8	10.0	11.8	12.6	6.9	8.7	9.8
BGT-Net [4]	16.8	20.6	23.0	10.4	12.8	13.6	5.7	7.8	9.3
RelTR [3]	20.0	21.2	-	7.7	11.4	-	6.8	10.8	-
ours	14.4	20.6	23.7	7.7	10.4	11.9	4.1	6.8	8.7
ours w/o [a]	9.5	14.5	14.9	5.8	7.2	7.8	3.2	4.6	5.4

Table 2: More ablation studies (PredCLS). [b] ours w/o the depth maps. [c] ours w/o the supervised contrastive loss.

Methods	R@20	R@50	mR@20	mR@50
[b]	62.5	74.2	15.5	21.6
[c]	62.1	74.7	14.8	20.4

Table 3: Zero-shot recall [19] (PredCLS).

Methods	zsR@20	zsR@50
NeuralMotifs [25]	1.4	3.6
Motif+EB [18]	2.1	4.9
VC-TDE+EB [18]	9.6	15.1
ours	10.9	20.4
ours w/o [a]	9.4	14.7

preliminary study suggests that factorizing the final probability distribution over the relationship categories could enhance the scene graph generation performance, and produce a diverse set of predicates beyond dataset annotations. It also shows strong zero-shot performance. In the near future, we are going to perform more comprehensive ablation studies and experiments on different datasets, and most importantly, extend our hierarchical classification scheme, as a portable module, to other existing state-of-the-art scene graph generation algorithms and enhance their results.

5 Acknowledge

This work was supported by the National Science Foundation (NSF) grant CCF-2112665 (TILOS).

References

[1] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020.

[2] Tianshui Chen, Weihao Yu, Riquan Chen, and Liang Lin. Knowledge-embedded routing network for scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6163–6171, 2019.

[3] Yuren Cong, Michael Ying Yang, and Bodo Rosenhahn. Reltr: Relation transformer for scene graph generation. *arXiv preprint arXiv:2201.11460*, 2022.

[4] Naina Dhingra, Florian Ritter, and Andreas Kunz. Bgt-net: Bidirectional gru transformer network for scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2150–2159, 2021.

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[6] Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David Shamma, Michael Bernstein, and Li Fei-Fei. Image retrieval using scene graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3668–3678, 2015.

[7] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020.

[8] Rajat Koner, Suprosanna Shit, and Volker Tresp. Relation transformer network. *arXiv preprint arXiv:2004.06193*, 2020.

[9] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017.

[10] Rongjie Li, Songyang Zhang, and Xuming He. Sgtr: End-to-end scene graph generation with transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19486–19496, 2022.

[11] Xin Lin, Changxing Ding, Jinquan Zeng, and Dacheng Tao. Gps-net: Graph property sensing network for scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3746–3753, 2020.

[12] Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. Visual relationship detection with language priors. In *European conference on computer vision*, pages 852–869. Springer, 2016.

[13] Alejandro Newell and Jia Deng. Pixels to graphs by associative embedding. *Advances in neural information processing systems*, 30, 2017.

[14] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE transactions on pattern analysis and machine intelligence*, 2020.

[15] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

[16] Guanghui Ren, Lejian Ren, Yue Liao, Si Liu, Bo Li, Jizhong Han, and Shuicheng Yan. Scene graph generation with hierarchical context. *IEEE Transactions on Neural Networks and Learning Systems*, 32(2):909–915, 2020.

[17] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

[18] Mohammed Suhail, Abhay Mittal, Behjat Siddiquie, Chris Broaddus, Jayan Eledath, Gerard Medioni, and Leonid Sigal. Energy-based learning for scene graph generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13936–13945, 2021.

[19] Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang. Unbiased scene graph generation from biased training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3716–3725, 2020.

[20] Kaihua Tang, Hanwang Zhang, Baoyuan Wu, Wenhan Luo, and Wei Liu. Learning to compose dynamic tree structures for visual contexts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6619–6628, 2019.

[21] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[22] Danfei Xu, Yuke Zhu, Christopher B Choy, and Li Fei-Fei. Scene graph generation by iterative message passing. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5410–5419, 2017.

[23] Minghao Xu, Meng Qu, Bingbing Ni, and Jian Tang. Joint modeling of visual objects and relations for scene graph generation. *Advances in Neural Information Processing Systems*, 34:7689–7702, 2021.

[24] Jianwei Yang, Jiasen Lu, Stefan Lee, Dhruv Batra, and Devi Parikh. Graph r-cnn for scene graph generation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 670–685, 2018.

[25] Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. Neural motifs: Scene graph parsing with global context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5831–5840, 2018.