

CogTree: Cognition Tree Loss for Unbiased Scene Graph Generation

Jing Yu^{1*†}, Yuan Chai^{2*}, Yujing Wang³, Yue Hu¹ and Qi Wu⁴

¹Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China

²Intelligent Computing and Machine Learning Lab, School of ASEE, Beihang University, Beijing, China

³Key Laboratory of Machine Perception, MOE, School of EECS, Peking University, Beijing, China

⁴University of Adelaide, Australia

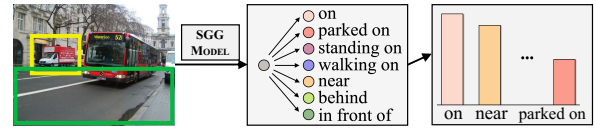
{yujing02, huyue}@ie.ac.cn, chaiyuan@buaa.edu.cn, yujwang@pku.edu.cn, qi.wu01@adelaide.edu.au

Abstract

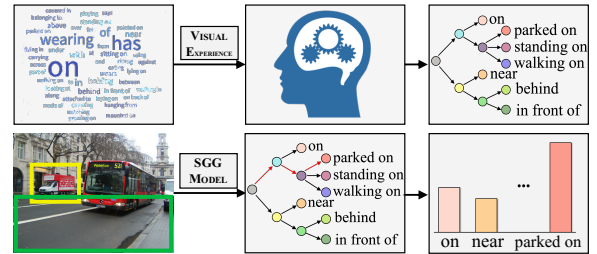
Scene graphs are semantic abstraction of images that encourage visual understanding and reasoning. However, the performance of Scene Graph Generation (SGG) is unsatisfactory when faced with biased data in real-world scenarios. Conventional debiasing research mainly studies from the view of balancing data distribution or learning unbiased models and representations, ignoring the correlations among the biased classes. In this work, we analyze this problem from a novel cognition perspective: automatically building a hierarchical cognitive structure from the biased predictions and navigating that hierarchy to locate the relationships, making the tail relationships receive more attention in a coarse-to-fine mode. To this end, we propose a novel debiasing Cognition Tree (CogTree) loss for unbiased SGG. We first build a cognitive structure CogTree to organize the relationships based on the prediction of a biased SGG model. The CogTree distinguishes remarkably different relationships at first and then focuses on a small portion of easily confused ones. Then, we propose a debiasing loss specially for this cognitive structure, which supports coarse-to-fine distinction for the correct relationships. The loss is model-agnostic and consistently boosting the performance of several state-of-the-art models. The code is available at: <https://github.com/CYVincent/Scene-Graph-Transformer-CogTree>.

1 Introduction

Making abstraction from an image into high-level semantics is one of the most remarkable capabilities of humans. Scene Graph Generation (SGG) [Xu *et al.*, 2017] — a task of extracting objects and their relationships in an image to form a graphical representation — aims to achieve the abstraction capability and bridge the gap between vision and language. SGG has greatly benefited the down-stream tasks of question answering [Norcliffe-Brown *et al.*, 2018;



(a) Flat thinking.



(b) Cognition-based hierarchical thinking.

Figure 1: (a) SGG model with conventional flat loss. (b) SGG model with our proposed cognition tree loss. The word size in the top-left box is proportional to the relationship frequency.

Zhu *et al.*, 2020] and visual understanding [Shi *et al.*, 2019; Jiang *et al.*, 2020]. Some works [Zhu *et al.*, 2020; Jiang *et al.*, 2020] feed the scene graphs into graph neural networks for relation-aware object representation. Some others [Hudson and Manning, 2019] perform sequential reasoning by traversing the relational graphs. Compared with independent objects, the rich relationships benefit explainable reasoning.

Although much effort has been made with high accuracy in object detection, the detected relationships are far from satisfaction due to the long-tailed data distribution. Only a small portion of the relationships have abundant samples (head) while most ones contain just a few (tail). This heavily biased training data causes biased relationship prediction. Tail relationships will be mostly predicted into head classes, which are not discriminative enough for high-level reasoning, such as falsely predicting *on* instead of *looking at*, and coarsely predicting *on* instead of *walking on*. Besides, tail relationships like *standing on*, *sitting on* and *lying on* can be even harder to distinguish from each other due to their visual similarity and scarce training data.

To tackle this problem, most research focuses on learning

*Equal contribution.

†Corresponding author.

unbiased models by re-weighting losses [Zareian *et al.*, 2020] or disentangling unbiased representations [Tang *et al.*, 2020]. However, humans can effectively infer the correct relationships even when some relationships appear more frequently than others. The essential difference between human and AI systems that has been ignored lies neither in the learning strategy nor feature representation, but in the way that the concepts are organized. To illustrate this discrepancy, we show an example `truck parked on street` in Figure 1. Existing models treat all the relationships independently for flat classification (Figure 1(a)). Because of the data bias, the head relationships, *e.g.* `on` and `near`, obtain high predicted probability. In contrast, the cognition theory [Sarafyazd and Jazayeri, 2019] supports that humans process information hierarchically in the prefrontal cortex (PFC). We intuitively start from making a rough distinction between remarkably different relationships. As shown in Figure 1(b), relationships belonging to the concept “on”, *e.g.* `on` and `parked on`, will firstly be distinguished from the ones about “near”, *e.g.* `near` and `behind`; then we move to distinguish the slight discrepancy among easily confused ones in one concept, *e.g.* `parked on`, `standing on` and `walking on`.

Inspired by the hierarchical reasoning mechanism in PFC, we propose a novel loss function, **Cognition Tree (CogTree)** loss, for unbiased scene graph generation. We first propose to build a hierarchy of the relationships, imitating the knowledge structure built from the independent relationships in human mind. The CogTree is derived from the prediction of a biased SGG model that satisfies the aforementioned thinking principles: distinguishing remarkably different relationships at first and then focusing on a small portion of easily confused ones. Then we design a CogTree-based loss to train the SGG network from scratch. This loss enables the network to surpass the noises from inter-concept relationships and then intra-concept relationships progressively. It frees the SGG models from the burden of distinguishing detailed discrepancy among all the relationships at one time, resulting in more accurate prediction via this coarse-to-fine strategy.

The main contributions are summarized as follows: (1) We propose a debiasing loss based on the inductive hierarchical structure, which supports coarse-to-fine distinction for the correct relationships while progressively eliminating the interference of irrelevant ones. (2) We automatically build the cognitive structure of the relationships from the biased SGG predictions, which reveals the inherent hierarchy that the relationships are organized after biased training; (3) The proposed loss is model-agnostic and can be applied to several representative SGG models. The results of extensive experiments indicate that the CogTree loss remarkably alleviates the bias in SGG and our loss achieves a new state-of-the-art performance compared with existing debiasing methods.

2 Related Work

Scene Graph Generation. SGG [Xu *et al.*, 2017] products graphical abstraction of an image and encourages visual relational reasoning and understanding in various down-stream tasks [Shi *et al.*, 2019; Zhu *et al.*, 2020]. Early works focus on object detection and relationship detection via independent

networks [Lu *et al.*, 2016; Zhang *et al.*, 2017], ignoring the rich contextual information. To incorporate the global visual context, recent works leverage message passing mechanism [Xu *et al.*, 2017; Chen *et al.*, 2019a], recurrent sequential architectures [Zellers *et al.*, 2018; Tang *et al.*, 2019] and contrastive learning [Zhang *et al.*, 2019] for more discriminative object and relationship representations. Although the accuracy is high in object detection, the relationship detection is far from satisfaction due to the heavily biased data. [Chen *et al.*, 2019b; Tang *et al.*, 2019] consider the biased SGG problem and propose mean Recall as the unbiased metric without corresponding debiasing solutions. The recent work [Liang *et al.*, 2019] prunes the predominant relationships and keeps the tail but informative ones in the dataset. [Tang *et al.*, 2020] proposes the first solution for unbiased SGG by counterfactual surgeries on causal graphs. We rethink SGG task from the cognition view and novelly solve the debiasing problem based on the coarse-to-fine structure of the relationships.

Biased Classification. Classification on highly-biased training data has been extensively studied in previous work, which can be divided into three categories: (1) balancing data distribution by data augmentation or re-sampling [Burnaev *et al.*, 2015; Li and Vasconcelos, 2019]; (2) debiasing learning strategies by re-weighting losses or training curriculums [Lin *et al.*, 2017b; Cui *et al.*, 2019]; (3) separating biased representations from the unbiased for prediction [Cadene *et al.*, 2019; Tang *et al.*, 2020]. Our CogTree loss belongs to the second category but differs from existing methods in that we are the first to leverage the hierarchical structure inherent in the relationships for re-weighting, which enables more discriminative representation learning by a coarse-to-fine mode.

3 Methodology

We first introduce the automatic CogTree building process. Then a novel debiasing loss based on the tree structure is proposed to train the SGG models. Since the CogTree loss is model-agnostic and applicable to various SGG models, we case-study two representative models, *i.e.* the widely compared MOTIFS [Zellers *et al.*, 2018] and the state-of-the-art VCTree [Tang *et al.*, 2019], and a newly proposed model, which first applies the transformer [Vaswani *et al.*, 2017] architecture to the SGG scenario and achieves superior performance. We name it as SG-Transformer and take it as a strong baseline. The framework is illustrated in Figure 2.

3.1 Automatic CogTree Building

Once the SGG model has been trained on the long-tailed data, the model tends to be biased on predicting “head” relationships. When different samples are predicted with the same relationship, they mostly share similar properties on either visual appearance (*e.g.* `walking on` and `standing on`) or high-level semantics (*e.g.* `has` and `with`). We regard these relationships sharing common properties as in one *concept*. We automatically induce concepts from the biased predictions and re-organize the relationships based on the concepts by a tree structure, denoted as Cognition Tree (CogTree). As shown on the top of Figure 2, this automatic process contains the following three steps:

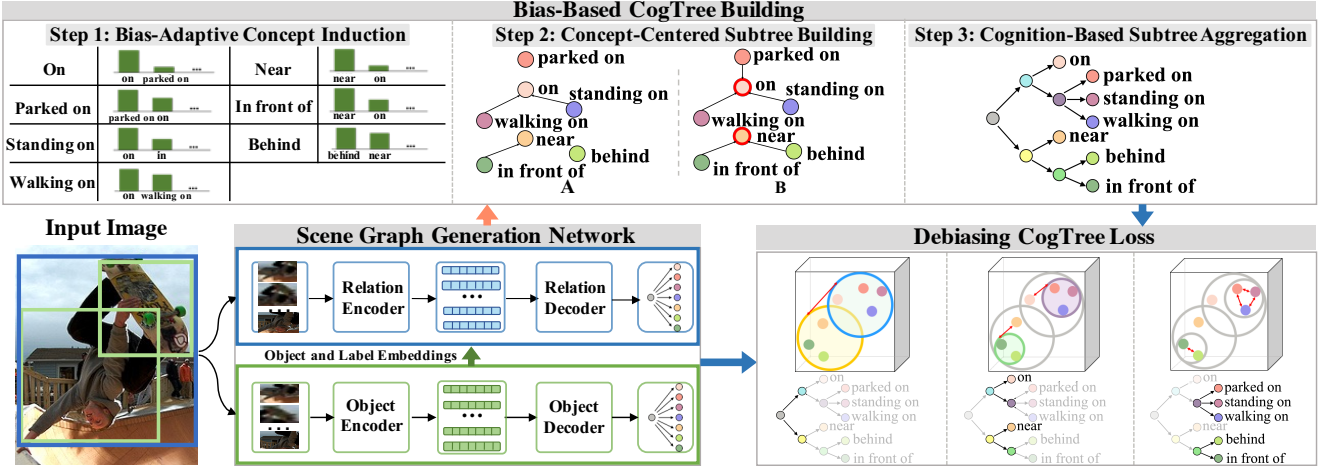


Figure 2: The overview of CogTree loss applied to SGG models. It contains three parts: Scene Graph Generation Network summarizes the framework of biased SGG models; Bias-Based CogTree Building organizes relationships by a coarse-to-fine tree structure based on biased predictions; Debiasing CogTree Loss supports network to distinguish relationships hierarchically.

Step 1: Bias-Based Concept Induction. For all the samples in the ground-truth class d_i , we predict their labels via a biased model and calculate the distribution of predicted label frequency, denoted as P_i . The most frequently predicted class d_j is regarded as the *concept relationship* of the ground-truth class d_i . As shown in Figure 2(Step 1), *on* is the concept relationship of itself, *standing on* and *walking on*. The above operation induces all the relationships into C concepts with corresponding concept relationships $\{c_i\}^C$.

Step 2: Concept-Centered Subtree Building. We reorganize all the relationships in each concept by a *Concept-Centered Subtree*. For the i^{th} subtree, the root is c_i while the leaves are the fine-grained relationships induced in c_i . Note that, if a subtree only contains a root c_i , e.g. *parked on* in Figure 2 (Step 2(A)), then c_i is considered not representative enough to summarize common properties. We further induce it to the most approximate concept relationship, which has the second highest frequency in P_i , e.g. induce *parked on* to *on* in Figure 2 (Step 2(B)). This process outputs T subtrees.

Step 3: Cognition-Based Subtree Aggregation. We aggregate the T subtrees into one hierarchical CogTree. As illustrated in Figure 3, CogTree contains four functional layers and each layer induces relationships into coarser groups than the layer that comes after it. Specifically, the *Root Layer* (y_0) contains a virtual node. The *Concept Layer* (y_1) distinguishes which concept the input belongs to. It contains T virtual nodes, each representing a subtree induced in Step 2. The following *Coarse-fine Layer* (y_2) tells whether the input can be described by a coarse-grained or fine-grained relationship in a certain concept. To this end, y_2 splits each node in y_1 into two nodes: one leaf node indicating the concept relationship while the other virtual node representing the cluster of fine-grained relationships in that concept. Each virtual node in y_2 links to its fine-grained relationships of the corresponding concept in the *Fine-grained Layer* (y_3). y_3 focuses on distinguishing the slight discrepancy among easily confused relationships, e.g. *standing on* and *walking on*.

3.2 Debiasing CogTree Loss

We propose a CogTree loss for training the SGG model for unbiased predictions. This loss encourages the model to classify relationships from coarse to fine according to the CogTree structure (see the bottom right of Figure 2). To better make the loss functions (3) and (4) understood, several notations are provided in advance as below:

Ground-Truth CogTree Path (GCP). It denotes the ground-truth classification path in CogTree for a training sample. Given the ground-truth label of a sample, we track the path from the root to the leaf with that label and denote this path as the ground-truth path $L_{path} = \{l_0, l_1, \dots, l_M\}$, where the k^{th} value l_m is the ground-truth node at layer y_m .

Predicted CogTree Probability (PCP). It means the predicted probability of each node in CogTree for a given sample by a biased SGG model. We denote the predicted probability set of all the classes as P_{pred} and denote the probability of class k as p_k . The PCP of node i is defined as below:

$$z_i = \begin{cases} p_k & \text{if } i \text{ is a leaf and } class(i) = k \\ \frac{1}{|C(i)|} \sum_{j \in C(i)} z_j & \text{if } i \text{ is not a leaf} \end{cases} \quad (1)$$

where $C(i)$ means the children of node i . $class(i)$ denotes the corresponding class of node i .

Class-Balanced Weight (CBW). It denotes the balance weight of each node in CogTree. Since the success of re-weighting strategy for debiasing, we adopt a well performed weighting factor [Cui *et al.*, 2019] to compute CBW of each leaf and uniformly define CBW of each node as below:

$$w_i = \begin{cases} (1 - \beta) / (1 - \beta^{n_i}) & \text{if } i \text{ is a leaf} \\ \frac{1}{|C(i)|} \sum_{j \in C(i)} w_j & \text{if } i \text{ is not a leaf} \end{cases} \quad (2)$$

where β is a hyper-parameter in the range of $[0, 1)$ and n_i is the sample number of the corresponding class of leaf i .

CogTree Loss. It contains two parts for complementary benefits: CogTree-based class-balanced (TCB) loss for debi-

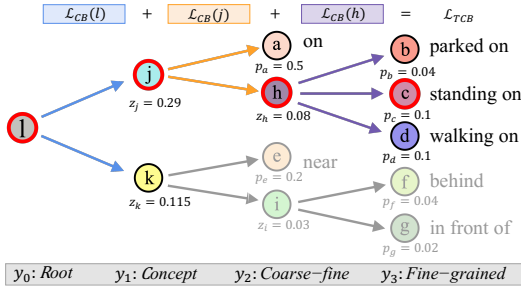


Figure 3: Illustration of calculating the TCB loss.

asing by coarse-to-fine relationship classification and class-balanced (CB) loss for debiasing by re-weighting on softmax cross-entropy loss. The two loss terms are defined below.

Given a sample with the ground-truth path L_{path} , as illustrated in Figure 3, we compute the class-balanced softmax cross-entropy loss on each layer of CogTree, and average the results to obtain the TCB loss as below:

$$\mathcal{L}_{TCB} = \frac{1}{|L_{path}|} \sum_{i \in L_{path}} -w_i \log\left(\frac{\exp(z_i)}{\sum_{z_j \in B(i)} \exp(z_j)}\right) \quad (3)$$

where $B(i)$ means the brothers of node i . TCB loss forces the network to surpass the noises from inter-concept relationships and learn concept-specific embeddings first, and then surpass the noises within one concept to refine relationship-specific embeddings, resulting in more fine-grained predictions.

Given a sample with the ground-truth class k , CB loss computes the class-balanced softmax cross-entropy based on the biased predicted probabilities P_{pred} :

$$\mathcal{L}_{CB} = -w_k \log\left(\frac{\exp(p_k)}{\sum_{p_j \in P_{pred}} \exp(p_j)}\right) \quad (4)$$

where w_k is the weighting factor [Cui *et al.*, 2019] for class k . Empirically, we found it the best to combine TCB with CB. In conclusion, our CogTree loss is:

$$\mathcal{L} = \mathcal{L}_{CB} + \lambda \mathcal{L}_{TCB} \quad (5)$$

where $\lambda = 1$ is the balancing weight. The SGG model is trained from scratch by the CogTree loss and predicting results via flat multi-class classifier without modification.

3.3 Scene Graph Generation Models

We case-study on three SGG models, *i.e.* MOTIFS, VCTree and our SG-Transformer. As shown in the bottom left of Figure 2, their architectures mainly contains two processes: object classification (bottom) for detecting the object positions and labels and relationship classification (top) for predicting the relationships between objects. We first review MOTIFS, and then introduce VCTree and SG-Transformer in a comparable fashion. All models are trained by the conventional cross-entropy loss [Zellers *et al.*, 2018].

MOTIFS. A pre-trained Faster R-CNN [Ren *et al.*, 2015] is applied to extract top K objects $O = \{o_i\}^K$ and describe object o_i by the RoIAlign visual feature v_i , the tentative object label l_i , and the spatial feature b_i . The above three kinds

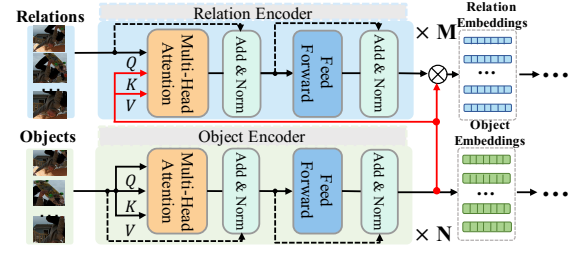


Figure 4: The encoder structure in SG-Transformer.

of features are concatenated and fed into $Encoder_o$ to obtain the refined object embedding m_i . m_i is then decoded by $Decoder_o$ to predict its final object label l_i^o . MOTIFS implemented $Encoder_o$ as a Bi-LSTM and adopts an LSTM as $Decoder_o$. The relationship encoder $Encoder_r$ first adopts another Bi-LSTM to capture the contextual information and then fuses the pairwise object features with the union region feature for decoding. The relationship decoder $Decoder_r$ is a fully connected layer followed by a softmax layer. **VC-Tree.** Different from MOTIFS, VCTree utilizes two independent TreeLSTMs as $Encoder_o$ and $Encoder_r$, respectively. It additionally takes the bounding box pair feature as another input of $Decoder_r$.

SG-Transformer. It adopts the transformer architecture as $Encoder_o$ and $Encoder_r$ by adjusting inputs for the SGG scenario. As shown in Figure 4, $Encoder_o$ contains N object-to-object (O2O) transformer blocks with the input of object embeddings $\{v_i\}^K$. $Encoder_r$ consists of M relation-to-object (R2O) transformer blocks with the input of relationship embeddings $\{r_{ij}\}^{K \times K}$ and the object embeddings $\{m_i\}^K$. r_{ij} is a combination of three kinds of features from objects o_i and o_j : spatial features, union region feature, word embeddings [Pennington *et al.*, 2014] of the predicted object labels. The output from $Encoder_r$ is concatenated with m_i and m_j for decoding. $Decoder_o$ and $Decoder_r$ contain a fully connected layer followed by a softmax layer.

4 Experiments

Dataset. We evaluate our models on the widely compared Visual Genome (VG) split [Xu *et al.*, 2017], with the 150 most frequent object classes and 50 most frequent relationship classes in VG [Krishna *et al.*, 2017]. The VG split only contains training set and test set and we follow [Zellers *et al.*, 2018] to sample a 5K validation set from the training set.

Tasks and Evaluation. The SGG task contains three sub-tasks [Zellers *et al.*, 2018]: (1) Predicate Classification (**PredCls**) takes the ground-truth object labels and bounding boxes for relationship prediction; (2) Scene Graph Classification (**SGCls**) takes ground-truth bounding boxes for object and relationship prediction; (3) Scene Graph Detection (**SGDet**) predict SGs from scratch. The conventional metric **Recall@K (R@K)** is reported but not considered as the main evaluation due to its bias. We adopt the recently proposed balanced metric **mean Recall@K (mR@K)** [Chen *et al.*, 2019b; Tang *et al.*, 2019]. mR@K calculates R@K for each class independently, and then average the results.

Model	Predicate Classification		Scene Graph Classification		Scene Graph Detection	
	mR@20 / 50 / 100	R@20 / 50 / 100	mR@20 / 50 / 100	R@20 / 50 / 100	mR@20 / 50 / 100	R@20 / 50 / 100
IMP+	- / 9.8 / 10.5	52.7 / 59.3 / 61.3	- / 5.8 / 6.0	31.7 / 34.6 / 35.4	- / 3.8 / 4.8	14.6 / 20.7 / 24.5
FREQ	8.3 / 13.0 / 16.0	53.6 / 60.6 / 62.2	5.1 / 7.2 / 8.5	29.3 / 32.3 / 32.9	4.5 / 6.1 / 7.1	20.1 / 26.2 / 30.1
KERN	- / 17.7 / 19.2	- / 65.8 / 67.6	- / 9.4 / 10.0	- / 36.7 / 37.4	- / 6.4 / 7.3	- / 27.1 / 29.8
MOTIFS	10.8 / 14.0 / 15.3	58.5 / 65.2 / 67.1	6.3 / 7.7 / 8.2	32.9 / 35.8 / 36.5	4.2 / 5.7 / 6.6	21.4 / 27.2 / 30.3
VCtree	14.0 / 17.9 / 19.4	60.1 / 66.4 / 68.1	8.2 / 10.1 / 10.8	35.2 / 38.1 / 38.8	5.2 / 6.9 / 8.0	22.0 / 27.9 / 31.3
MOTIFS*	11.5 / 14.6 / 15.8	59.5 / 66.0 / 67.9	6.5 / 8.0 / 8.5	35.8 / 39.1 / 39.9	4.1 / 5.5 / 6.8	25.1 / 32.1 / 36.9
MOTIFS + Focal	10.9 / 13.9 / 15.0	59.2 / 65.8 / 67.7	6.3 / 7.7 / 8.3	36.0 / 39.3 / 40.1	3.9 / 5.3 / 6.6	24.7 / 31.7 / 36.7
MOTIFS + Reweight	16.0 / 20.0 / 21.9	45.4 / 57.0 / 61.7	8.4 / 10.1 / 10.9	24.2 / 29.5 / 31.5	6.5 / 8.4 / 9.8	18.3 / 24.4 / 29.3
MOTIFS + Resample	14.7 / 18.5 / 20.0	57.6 / 64.6 / 66.7	9.1 / 11.0 / 11.8	34.5 / 37.9 / 38.8	5.9 / 8.2 / 9.7	23.2 / 30.5 / 35.4
MOTIFS + TDE	18.5 / 25.5 / 29.1	33.6 / 46.2 / 51.4	9.8 / 13.1 / 14.9	21.7 / 27.7 / 29.9	5.8 / 8.2 / 9.8	12.4 / 16.9 / 20.3
MOTIFS + CogTree	20.9 / 26.4 / 29.0	31.1 / 35.6 / 36.8	12.1 / 14.9 / 16.1	19.4 / 21.6 / 22.2	7.9 / 10.4 / 11.8	15.7 / 20.0 / 22.1
VCtree*	11.7 / 14.9 / 16.1	59.8 / 66.2 / 68.1	6.2 / 7.5 / 7.9	37.0 / 40.5 / 41.4	4.2 / 5.7 / 6.9	24.7 / 31.5 / 36.2
VCtree + TDE	18.4 / 25.4 / 28.7	36.2 / 47.2 / 51.6	8.9 / 12.2 / 14.0	19.9 / 25.4 / 27.9	6.9 / 9.3 / 11.1	14.0 / 19.4 / 23.2
VCtree + CogTree	22.0 / 27.6 / 29.7	39.0 / 44.0 / 45.4	15.4 / 18.8 / 19.9	27.8 / 30.9 / 31.7	7.8 / 10.4 / 12.1	14.0 / 18.2 / 20.4
SG-Transformer	14.8 / 19.2 / 20.5	58.5 / 65.0 / 66.7	8.9 / 11.6 / 12.6	35.6 / 38.9 / 39.8	5.6 / 7.7 / 9.0	24.0 / 30.3 / 33.3
SG-Transformer + CogTree	22.9 / 28.4 / 31.0	34.1 / 38.4 / 39.7	13.0 / 15.7 / 16.7	20.8 / 22.9 / 23.4	7.9 / 11.1 / 12.7	15.1 / 19.5 / 21.7

Table 1: State-of-the-art comparison on R@K and mR@K. Our re-implemented SGG models are denoted by the superscript *.

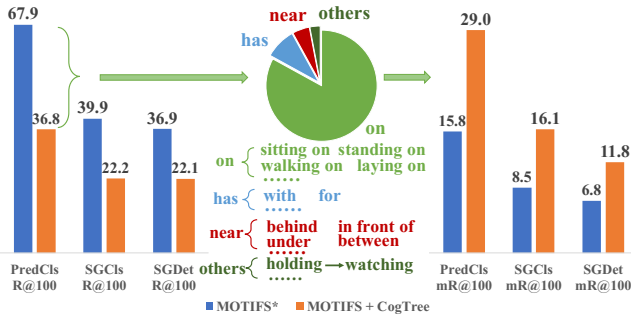


Figure 5: The pie chart shows the relationships that are correctly predicted by the baseline but regarded as incorrect by CogTree. The predictions by CogTree are shown under the pie chart. The increase of mR@K bringing a drop of R@K is commonly existed (see Table 1). Combining with the qualitative analysis, we believe that the contradiction of R@K and mR@K is caused by two reasons: (1) annotators prefer simple and vague labels due to bounded rationality [Simon, 1990]; (2) CogTree predicts more fine-grained relationships. Our analysis is consistent with the recent study [Tang *et al.*, 2020].

Implementation. The object detector is the pre-trained Faster R-CNN [Ren *et al.*, 2015] with ResNeXt-101-FPN [Lin *et al.*, 2017a]. λ is set to 1 and β is set to 0.999. SG-Transformer contains 3 O2O blocks and 2 R2O blocks with 12 attention heads. Models are trained by SGD optimizer with 5 epochs. The mini-batch size is 12 and the learning rate is 1.2×10^{-3} . Experiments are implemented with PyTorch and conducted with NVIDIA Tesla V100 GPUs.

4.1 State-of-the-Art Comparison

We evaluate the CogTree loss on three models: MOTIFS, VCtree and SG-Transformer, and compare the performance with the state-of-the-art debiasing approach TDE [Tang *et al.*, 2020]. We also compare with the biased models, including IMP+ [Xu *et al.*, 2017], FREQ [Zellers *et al.*, 2018], KERN [Chen *et al.*, 2019b]. The results are shown in Table 1. In the view of mR@K, CogTree remarkably improves all the baselines on all the metrics. The strongest baseline VCtree also achieves 3.6%~13.6% boost. CogTree con-

Method	MOTIFS	VCtree	SG-transformer
	R@20 / 50 / 100	R@20 / 50 / 100	R@20 / 50 / 100
baseline	35.0 / 36.7 / 37.0	35.0 / 36.5 / 36.8	38.0 / 39.6 / 39.9
baseline+CogTree	51.6 / 57.7 / 59.8	51.6 / 57.2 / 58.9	48.7 / 54.5 / 56.4

Table 2: R@K of the tail 45 classes on predicate classification.

Method	PredCls	SGCls	SGDet
	mR@20 / 50 / 100	mR@20 / 50 / 100	mR@20 / 50 / 100
Full loss	22.89 / 28.38 / 30.97	12.96 / 15.68 / 16.72	7.92 / 11.05 / 12.70
1 \mathcal{L}_{TCB}	21.08 / 27.08 / 29.41	12.15 / 15.07 / 16.15	7.70 / 10.39 / 12.07
2 \mathcal{L}_{CB}	18.02 / 23.40 / 25.25	10.76 / 13.13 / 13.88	6.74 / 9.56 / 11.29
3 \mathcal{L}_{TCE}	21.16 / 26.14 / 28.32	12.14 / 14.42 / 15.29	7.57 / 10.53 / 11.86
4 \mathcal{L}_{CE}	14.35 / 18.48 / 20.21	8.57 / 11.46 / 12.27	5.55 / 7.74 / 8.98
5 Fuse-subtree	16.20 / 20.17 / 22.12	8.71 / 10.66 / 11.61	5.36 / 7.19 / 8.28
6 Fuse-layer	13.77 / 18.87 / 20.77	8.17 / 10.39 / 11.32	5.86 / 8.02 / 9.05
7 $\mathcal{L}(\text{MAX})$	15.48 / 19.93 / 21.87	8.97 / 10.85 / 11.83	5.38 / 7.16 / 8.16
8 $\mathcal{L}(\text{SUM})$	11.31 / 15.67 / 17.98	6.58 / 8.82 / 9.86	1.86 / 3.09 / 3.68

Table 3: Ablation study of key components in CogTree loss.

sistently outperforms conventional debiasing methods: focal loss, reweight and resample, and the SOTA debiasing method TDE. Wherein, mR@20 achieves superior improvement compared with mR@50/100, which indicates that more accurate predictions have higher rank by CogTree. It is worth highlighting that CogTree outperforms TDE on both mR@K and most metrics of R@K on the strongest baseline VCtree.

Analysis of Labeling Bias. We observe a drop of R@K with the increase of mR@K from baselines to CogTree. Further analysis in Figure 5 reveals that annotators prefer simple and vague labels, which form the head classes. The baselines are biased on these head classes and result in high R@K. CogTree prefers fine-grained and semantic-rich tail labels, causing the low R@K by “incorrectly” classifying the head classes into reasonable tail ones. To fairly compare with baselines on the conventional metric R@K, we compute R@K of the tail 45 classes from the predictions of all the classes. We observe a significant improvement on R@K from baselines to CogTree in Table 2, which proves the strong debiasing ability of CogTree without sacrificing the performance of R@K.

4.2 Ablation Study

We investigate each loss terms, the weighting method, two tree building strategies and two methods for computing PCP

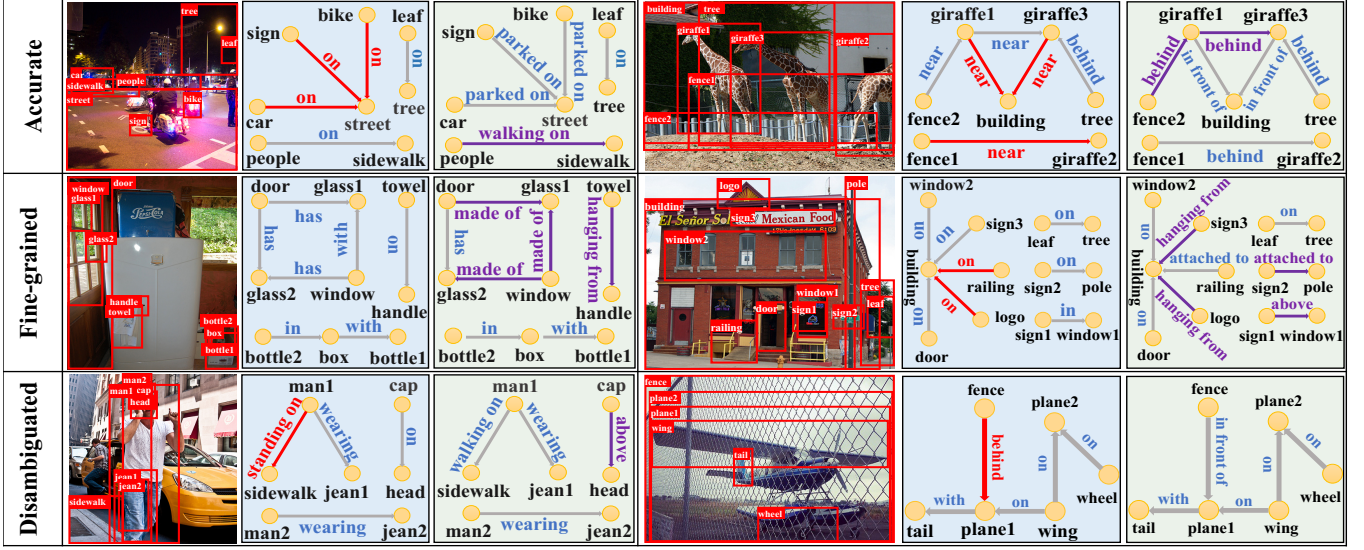


Figure 6: Visualization of scene graphs generated by SG-Transformer (blue) and SG-Transformer+CogTree (green). Compared with the ground-truth, the quality of predicted relationships are marked in three colors: red (false), blue (correct), purple (better).

and CBW: 1) \mathcal{L}_{TCB} : tree-based classification loss term in Eq. 3; 2) \mathcal{L}_{CB} : flat classification loss term in Eq. 4; 3) \mathcal{L}_{TCE} : we remove the weights $\{w_i\}$ from Eq. 3; 4) \mathcal{L}_{CE} : we remove the weight w_k from Eq. 4; 5) **Fuse-subtree**: we mix relationships belonging to different concepts in one subtree; 6) **Fuse-layer**: we mix the fine-grained and coarse relationships in the same layer of each subtree; 7) $\mathcal{L}(\text{MAX})$: we use MAX instead of AVERAGE to compute z_i and w_i for non-leaf nodes in Eq.1 and Eq.2; 8) $\mathcal{L}(\text{SUM})$: we use SUM instead of AVERAGE.

The results in Table 3 on the SG-Transformer baseline indicate that: 1) \mathcal{L}_{TCB} and \mathcal{L}_{CB} have complementary benefits for SGG, where \mathcal{L}_{TCB} has obvious greater influence. 2) Both of the loss terms benefit from the weighting strategy. 3) The CogTree is built based on two principles: a) Relationships belonging to the same concept are organized in one subtree and b) relationships in one subtree are organized in different layers from coarse to fine. A significant performance decrease is observed when the tree violates either of the principles. 4) AVERAGE performs the best because that both MAX and SUM increase the predicted probabilities and decrease the incorrect prediction penalty in Eq. 3.

4.3 Qualitative Analysis

We visualize several PredCls samples in Figure 6. SG-Transformer achieves obvious improvement when equipped with CogTree: (1) CogTree predicts more fine-grained relationships. As shown in the first two rows, the baseline mostly predicts the vague head classes `on` and `near` while CogTree predicts actions and locations like `parked on` and `in front of`. (2) CogTree accurately distinguishes visually and semantically similar relationships as shown in the last row. However, the baseline falsely predicts `walking on` as `standing on` and predicts `in front of` as `behind`, since flat classification cannot capture the detailed discrepancy. In summary, CogTree preferences towards fine-grained relationships instead of the vaguely biased ones.

Predicate Classification		Scene Graph Classification		Scene Graph Detection	
λ	mR@20 / 50 / 100	mR@20 / 50 / 100	mR@20 / 50 / 100	mR@20 / 50 / 100	mR@20 / 50 / 100
0.4	21.17 / 27.20 / 29.59	12.09 / 15.03 / 16.12	7.22 / 9.72 / 11.30		
0.7	21.28 / 26.39 / 28.69	12.22 / 14.91 / 15.88	8.62 / 11.30 / 12.70		
1	22.89 / 28.38 / 30.97	12.96 / 15.68 / 16.72	7.92 / 11.05 / 12.70		
1.3	22.06 / 27.11 / 29.21	12.35 / 15.04 / 15.87	7.84 / 10.48 / 12.19		
1.6	21.34 / 26.80 / 29.18	12.20 / 14.70 / 15.82	6.78 / 9.04 / 10.22		

Table 4: Performance with different balancing weights.

4.4 Parameter Analysis

In Table 4, we assess the effect of λ in Eq.5. $\lambda = 1$ achieves the best performance on most metrics. The performance drops slightly in the range of $[0.7, 1.3]$. We also vary the number of O2O and R2O blocks from 2 to 4 in SG-Transformer and obtain the highest mR@K with 3 O2O blocks and 2 R2O blocks. We use the above settings in our full model.

5 Conclusion

In this paper, we propose a CogTree loss to generate unbiased scene graphs with highly biased data. We novelly leverage the biased prediction from SGG models to organize the independent relationships by a tree structure, which contains multiple layers corresponding to the relationships from coarse to fine. We propose a CogTree loss specially for the above tree structure that supports hierarchical distinction for the correct relationships while progressively eliminating the irrelevant ones. The loss is model-agnostic and consistently boosting the performance of various SGG models with remarkable improvement. How to incorporate commonsense knowledge to optimize the CogTree structure will be our future work.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant No. 62006222).

References

- [Burnaev *et al.*, 2015] Evgeny Burnaev, Pavel Erofeev, and Artem Papanov. Influence of resampling on accuracy of imbalanced classification. In *ICMV*, pages 987521–987521, 2015.
- [Cadene *et al.*, 2019] Remi Cadene, Corentin Dancette, Hedi Ben younes, Matthieu Cord, and Devi Parikh. Rubi: reducing unimodal biases for visual question answering. In *NeurIPS*, pages 841–852, 2019.
- [Chen *et al.*, 2019a] Long Chen, Hanwang Zhang, Jun Xiao, Xiangnan He, Shiliang Pu, and Shih-Fu Chang. Counterfactual critic multi-agent training for scene graph generation. In *ICCV*, pages 4613–4623, 2019.
- [Chen *et al.*, 2019b] Tianshui Chen, Weihao Yu, Riquan Chen, and Liang Lin. Knowledge-embedded routing network for scene graph generation. In *CVPR*, pages 6163–6171, 2019.
- [Cui *et al.*, 2019] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *CVPR*, pages 9268–9277, 2019.
- [Hudson and Manning, 2019] Drew Hudson and Christopher D Manning. Learning by abstraction: the neural state machine. In *NeurIPS*, pages 5903–5916, 2019.
- [Jiang *et al.*, 2020] Xiaoze Jiang, Jing Yu, Zengchang Qin, Yingying Zhuang, Xingxing Zhang, Yue Hu, and Qi Wu. Dualvd: an adaptive dual encoding model for deep visual understanding in visual dialogue. In *AAAI*, pages 11125–11132, 2020.
- [Krishna *et al.*, 2017] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual genome: connecting language and vision using crowd-sourced dense image annotations. *IJCV*, 123:32–73, 2017.
- [Li and Vasconcelos, 2019] Yi Li and Nuno Vasconcelos. Repair: removing representation bias by dataset resampling. In *CVPR*, pages 9572–9581, 2019.
- [Liang *et al.*, 2019] Yuanzhi Liang, Yalong Bai, Wei Zhang, Xueming Qian, Li Zhu, and Tao Mei. Vrr-vg: refocusing visually-relevant relationships. In *ICCV*, pages 10403–10412, 2019.
- [Lin *et al.*, 2017a] Tsung-Yi Lin, Piotr Dollar, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, pages 2117–2125, July 2017.
- [Lin *et al.*, 2017b] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, pages 2980–2988, 2017.
- [Lu *et al.*, 2016] Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. Visual relationship detection with language priors. In *ECCV*, pages 852–869, 2016.
- [Norcliffe-Brown *et al.*, 2018] Will Norcliffe-Brown, Stathis Vafeias, and Sarah Parisot. Learning conditioned graph structures for interpretable visual question answering. In *NeurIPS*, pages 8334–8343, 2018.
- [Pennington *et al.*, 2014] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *EMNLP*, pages 1532–1543, 2014.
- [Ren *et al.*, 2015] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: towards real-time object detection with region proposal networks. In *NeurIPS*, pages 91–99, 2015.
- [Sarafyazd and Jazayeri, 2019] Morteza Sarafyazd and Mehrdad Jazayeri. Hierarchical reasoning by neural circuits in the frontal cortex. *Science*, 364(6441):eaav8911, 2019.
- [Shi *et al.*, 2019] Jiaxin Shi, Hanwang Zhang, and Juanzi Li. Explainable and explicit visual reasoning over scene graphs. In *CVPR*, pages 8376–8384, 2019.
- [Simon, 1990] Herbert A Simon. Bounded rationality. In *Utility and probability*, pages 15–18. Springer, 1990.
- [Tang *et al.*, 2019] Kaihua Tang, Hanwang Zhang, Baoyuan Wu, Wenhan Luo, and Wei Liu. Learning to compose dynamic tree structures for visual contexts. In *CVPR*, pages 6619–6628, 2019.
- [Tang *et al.*, 2020] Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang. Unbiased scene graph generation from biased training. In *CVPR*, pages 3716–3725, 2020.
- [Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, pages 5998–6008, 2017.
- [Xu *et al.*, 2017] Danfei Xu, Yuke Zhu, Christopher B Choy, and Li Fei-Fei. Scene graph generation by iterative message passing. In *CVPR*, pages 5410–5419, 2017.
- [Zareian *et al.*, 2020] Alireza Zareian, Svebor Karaman, and Shih-Fu Chang. Bridging knowledge graphs to generate scene graphs. In *ECCV*, pages 606–623, 2020.
- [Zellers *et al.*, 2018] Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. Neural motifs: scene graph parsing with global context. In *CVPR*, pages 5831–5840, 2018.
- [Zhang *et al.*, 2017] Ji Zhang, Mohamed Elhoseiny, Scott Cohen, Walter Chang, and Ahmed Elgammal. Relationship proposal networks. In *CVPR*, pages 5678–5686, 2017.
- [Zhang *et al.*, 2019] Ji Zhang, Kevin J. Shih, Ahmed Elgammal, Andrew Tao, and Bryan Catanzaro. Graphical contrastive losses for scene graph parsing. In *CVPR*, pages 11535–11543, 2019.
- [Zhu *et al.*, 2020] Zihao Zhu, Jing Yu, Yujing Wang, Yajing Sun, Yue Hu, and Qi Wu. Mucko: multi-layer cross-modal knowledge reasoning for fact-based visual question answering. In *IJCAI*, pages 1097–1103, 2020.