More Knowledge, Less Bias: Unbiasing Scene Graph Generation with Explicit Ontological Adjustment

Zhanwen Chen[1] Saed Rezayi[2] Sheng Li[1]
[1]School of Data Science, University of Virginia
{pct4et, shengli}@virginia.edu
[2]School of Computing, University of Georgia
saedr@uga.edu

Abstract

Scene graph generation (SGG) models seek to detect relationships between objects in a given image. One challenge in this area is the biased distribution of predicates in the dataset and the semantic space. Recent works incorporating knowledge graphs with scene graphs prove effective in improving recall for the tail predicate classes. Moreover, many recent SGG approaches with promising results explicitly redistribute the predicates in both the training process and in the prediction step. To incorporate external knowledge, we construct a commonsense knowledge graph by integrating ConceptNet and Wikidata. To explicitly unbias SGG with knowledge in the reasoning process, we propose a novel framework, Explicit Ontological Adjustment (EOA), to adjust the graph model predictions with knowledge priors. We use the edge matrix from the commonsense knowledge graph as a module in the graph neural network model to refine the relationship detection process. This module proves effective in alleviating the long-tail distribution of predicates. When combined, we show that these modules achieve state-of-the-art performance on the Visual Genome dataset in most cases. The source code is available at `https://github.com/zhanwenchen/eoa`.

1. Introduction

What is the relationship between the man and the ski as shown in Figure 1? Given such an image from the Visual Genome (VG) dataset [15], Scene Graph Generation (SGG) answers questions like this by combining symbolic and graphical representations extracted from this image. The output can be represented as a graph where scene objects are nodes or *entities* and their relationships as edges or *predicates*. Taking Figure 1 as an example, there are 12 such predicates such as ⟨man, ski, standing on⟩ and ⟨man, jacket, wearing⟩. Recent work has drawn attention to SGG thanks to its promising applications in downstream tasks such as image captioning [34, 7, 37], image retrieval [13, 23], image synthesis [12], and visual question answering (VQA) [6, 8, 31, 34, 42]. Additionally, SGG can offer interpretability for VQA [39, 24]. However, two challenges limit the performance of most existing SGG methods: (1). the long-tail distribution of predicate classes in practice, and (2). the limited volume and depth of commonsense knowledge in knowledge-based SGG models. As Guo *et al*. point out [9], common predicates like "on" crowd out rarer but more informative ones like "standing on." Common predicates have larger semantic spaces (*i.e*., more meanings), causing bias against less common but more informative relations. Moreover, despite the performance gains from introducing commonsense knowledge, only one source of structured ontology is used - ConceptNet [26], from which only 104 facts have been used in knowledge-based SGG [39]. The 104 facts are also limited to entities belonging to the VG ob-

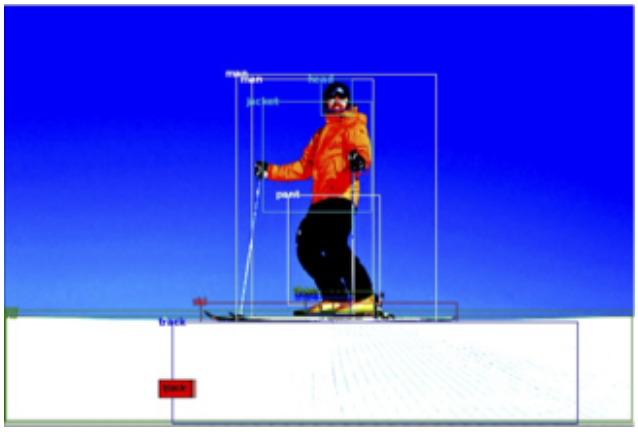

Figure 1: An example image from the Visual Genome dataset [15] with ground-truth relation labels including ⟨man, ski, standing on⟩ and ⟨man, jacket, wearing⟩

ject classes. This leaves room for further performance gains from not only using more knowledge but also from adding deeper knowledge outside of VG.

Recently, some works on SGG have successfully incorporated knowledge into the SGG task. KB-GAN [8] addresses the problem of object label imbalance by introducing external knowledge and a novel image reconstruction loss. KI-Net [41] incorporates off-scene higher-order knowledge ("suit worn over shirt") to solve the downstream VQA task. GB-Net [39] introduces a bipartite knowledge-and-scene graph and adapts the Gated Graph Sequence Neural Networks (GGNN) [17] to refine the scene graph (SG) edges and the SG-KG edges. However, as Zhang *et al*. point out [41], GB-Net only includes knowledge edges whose head and tail entities are both part of the VG visual classes[1], potentially limiting the depth of the reasoning process. To enrich the knowledge graph and address the limitation of reasoning with only on-scene concepts, we propose to incorporate additional knowledge by designing a commonsense knowledge graph extension procedure.

Many recent works have also attempted to address the biased distribution of predicates. In addition to knowledge-based approaches that improve the per-class performance of the tail classes [39], recent efforts have featured various data augmentation-based approaches such as Dynamic Label Frequency Estimation [3], sampling-based approaches such as Total Direct Effect [28], Balanced Predicate Learning (BPL) [9], and Semantic Adjustment (SA) [9]. Most relevant to our work are the BPL and the SA modules [9]. BPL conducts domain transfer by refining the trained model on an unbiased version of the original dataset with undersampled top-k predicates. The SA module uses the row-normalized confusion matrix to adjust the prediction logits in the model via matrix multiplication [9]. Although BPL and SA contribute to a significant improvement over all three SGG tasks on mean recall metrics, they do not take advantage of external commonsense knowledge.

Although KG-based SGG methods such as GB-Net [39] have reduced bias, they rely on a single knowledge source and limit the knowledge to on-scene entities. We improve KG-based SGG by taking advantage of the unbiasing effect of knowledge by enriching the commonsense KG. Specifically, we increase both the depth and the breadth of the KG by incorporating additional knowledge and off-scene knowledge with an external knowledge base (KB), Wikidata [30]. Additionally, we propose a novel framework called Explicit Ontological Adjustment (EOA) to rebalance predicates using knowledge as statistical priors. Our method outperforms state-of-the-art methods on the Visual Genome dataset on the Predicate Classification (PredCls) task and achieves a competitive performance gain on the Scene Graph Classification (SGCls) task, confirming a significant unbiasing effect of knowledge in relationship reasoning for SGG.

We summarize our contributions as follows:

- We build an enriched commonsense graph for SGG by incorporating higher-order and on/off-scene facts from ConceptNet and Wikidata [11].
- We propose the Explicit Ontological Adjustment (EOA) approach for the neural network architecture that includes BPL, SA, and a novel Ontological Adjustment (OA) module.
- Our method outperforms state-of-the-art methods on the Visual Genome dataset on the PredCls task and achieves competitive performance on the SGCls task, potentially suggesting a superior ability to unbias.

The rest of this paper is organized as follows. Section 2 reviews related work on SGG and discusses the various threads in solving the predicate bias issue in SGG, including resampling- and knowledge-based approaches. Section 3 describes our approaches to unbiasing SGG, including the commonsense KG extension process and the EOA neural network module. Section 4 shows extensive results and ablation studies on the Visual Genome dataset. Finally, Section 5 concludes the paper by discussing the results, potential next steps, and limitations.

2. Related Work

2.1. Scene Graph Generation (SGG)

SGG methods first extract region proposals using an object detector like Faster-RCNN [21] with feature extraction backbones such as VGG-16 [25] and ResNet [10]. SGG then classifies the individual regions into object classes before classifying region pairs into relation classes. The key to the main relationship reasoning process is joint reasoning with contextual information. For example, the Iterative Message Passing (IMP) technique [35] iteratively propagates contextual messages along the scene graph topology. MotifNet [40] encodes global context using LSTMs to assist with local predictions. Information propagation between region proposals is typically achieved with a graph neural network (GNN) [32]. Additional statistical metadata such as statistical correlation and semantic information also prove helpful. For example, MotifNet [40] achieves prediction improvement by reasoning with statistical correlations of object pairs and GloVe vectors.

2.2. Knowledge-Graph Based SGG Methods

Recently, some methods have introduced external knowledge to improve the reasoning process [2, 8, 39, 40]. MotifNet can also be seen as a knowledge-based approach

[1]We refer to entities that are part of the VG classes as on-scene entities and those not in VG as off-scene entities.

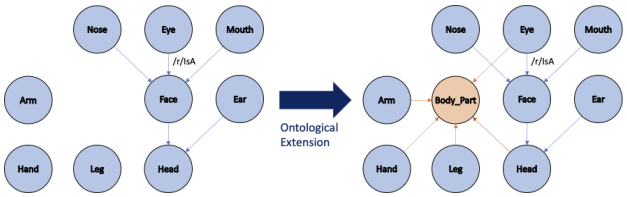

Figure 2: An example subgraph of the extended commonsense knowledge graph. Blue nodes from GB-Net [39] correspond to VG scene entities. The orange node "body_part" is an off-scene entity from Wikidata. Adding the off-scene "body_part" node makes the KG more densely connected and thus more informative.

for using prior triplet frequency to bias the relationship predictions [40]. KERN [2] jointly classifies object and relationship classes using prior knowledge of object and relationship co-occurrence. Incorporating such statistical priors, KERN modifies the task to be predicting the likeliest relation between objects given likelihood priors [2]. KB-GAN [8] uses commonsense knowledge from ConceptNet [26] to refine object and phrase features for SGG. KB-GAN also uses a Generative Adversarial Network (GAN) to regularize the whole SGG network. Most relevant to our work, GB-Net [39] extends the KERN approach by explicitly constructing the bipartite graph ontology to combine the scene graph and knowledge graph (KG). GB-Net also adapts a graph neural network, Gated Graph Sequence Neural Networks (GGNN) [17], to iteratively pass messages along this bipartite ontology. Our work is motivated in part by using knowledge to unbias SGG. Our first baseline, GB-Net, provides an innovative technique to propagate messages in both the knowledge graph and the scene graph in a dual message-passing network model. However, GB-Net only uses 104 facts [39] from ConceptNet [26], which leaves room for further performance gains from the additional breadth and depth of the KG. Furthermore, GB-Net [39] ignores off-scene knowledge that may be useful in reasoning about the scene. This limits the reasoning power of a KG where part of a dense knowledge subgraph is missing in the dataset. As Figure 2 illustrates, the original KG on the left side is missing an important off-scene node, "body_part," limiting the depth of reasoning. We address both issues by enriching the KG with a new data source, Wikidata [30].

2.3. Unbiased SGG Methods

Many recent works attempt to address the long-tailed distribution of relationship predicates in SGG. Instead of making predictions in isolation, the IMP method proposed by Xu *et al*. [35] incorporates the contextual information in the neural network model to address the uneven distribution of the relationship labels in the VG dataset. IMP achieves this by updating both entity and predicate representations in an interwoven fashion with Gated Recurrent Units (GRUs) [4]. Other more recent works focus on directly addressing the distribution bias. Tang *et al*. [28] propose the Total Direct Effect (TDE) loss function to measure the relative contribution of the visual features to relationship detection that is distinct from the context. Unlike traditional debiasing approaches, which cannot distinguish between desirable bias (*e.g*., "man wearing jacket" instead of "eating" it) and bad bias (*e.g*., "man on snow" instead of "man standing on snow"), TDE eliminates the harmful context bias with counterfactual evaluations. Another relevant recent work is the G2S framework involving the BPL and the SA modules proposed by Guo *et al*. [9]. Directly attacking the bad bias, *i.e*., the lack of informative predicates in both the semantic space and the VG sample space, SA and BPL are pipeline modules compatible with various models. SA [9] adjusts the model predictions with the confusion matrix, a proven strategy for reducing bias in long-tail classification tasks [19]. As for the sample-level predicate imbalance, BPL [9] offers a domain transfer-based solution by creating a data-augmented version of the original VG dataset. The "new" dataset is constructed by undersampling the top-K common predicates in the original dataset.

Inspired by the explicit adjustment technique in SA, we propose a novel technique, Explicit Ontological Adjustment (EOA), to explicitly adjust the distribution of predicates using knowledge-based priors. EOA uses the KG adjacency matrix to adjust the predicate logits in training to take advantage of the knowledge priors that are readily available in the KG. For example, knowing that "flying" is a "/r/MannerOf" "transport," the model is more likely to consider "flying," which is more informative than "transport".

3. Our Approach

In this section, we first introduce the preliminary SGG formalism. We explore the unbiasing potential of additional knowledge in two ways. The first is enriching the knowledge graph for SGG with off-scene entities. To achieve this, we adopt an entity linking algorithm to integrate multiple commonsense knowledge sources, including ConceptNet [26] and Wikidata [30]. We also use the knowledge graph to explicitly redistribute the long-tail distribution of predicates.

3.1. Preliminary

We first formally describe the Scene Graph Generation (SGG) problem. A scene graph (SG) consists of visual relations expressed as ⟨subject, predicate, object⟩ triplets [2]. This triplet can also be seen as an SG edge going from subject to object. An example is ⟨man, standing on, snow⟩ where "man" is the subject, "snow" is the object, and

[2]Similar to KG triplets ⟨head entity, relation, tail entity⟩.

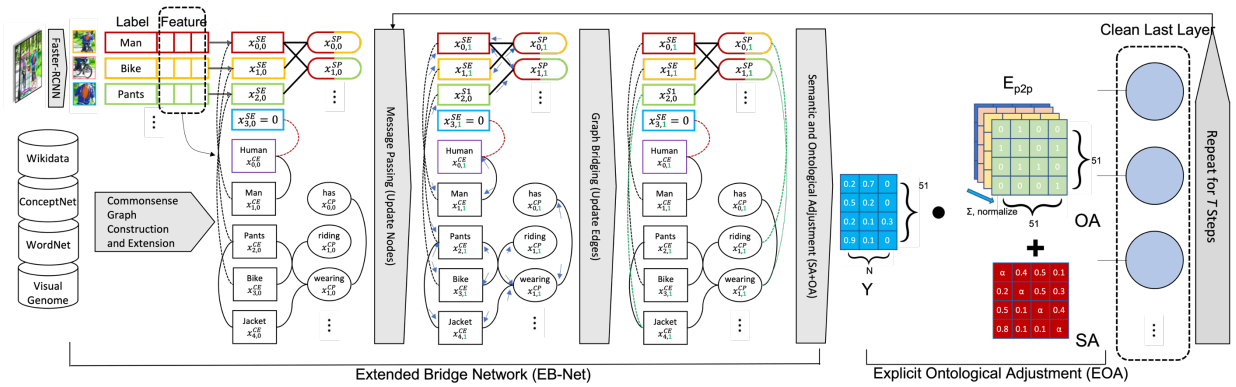

Figure 3: An overview of our entire pipeline with EB-Net and EOA. We first train the EB-Net after conducting the commonsense knowledge graph extension. In the last prediction step, we adjust the predicted logits by the sum of our ontological adjustment (OA) matrix and the confusion matrix C" similar to SA [9]. Lastly, we use the undersampled source dataset similar to BPL [9] to conduct domain transfer. The BPL process is not illustrated because it degrades figure readability.

"standing on" is the relation or predicate. Here, "man" and "snow" are visual instances belonging to the visual class set C for all scenes S in the dataset. The relation "standing on" describes the directed relation between this instance of "man" and this instance of "snow" that are localized in their bounding boxes. Each scene s_i, $i \in [1, |S|]$, consists of a variable number k_i of such relations $r_{s,m}$ where $m \in \{1, 2, ..., k_i\}$. For convenience, we do not consider attributes such as "red". The set of all relations in scene s, R_s, can be expressed as

$$R_s \subseteq C \times P \times C \tag{1}$$

where C is the set of all visual classes and can appear as subjects and/or objects; P is the set of all predicates in the dataset. Each visual class c_s in scene s has its bounding box b_{c_s} and its ground-truth class label l_{c_s}. The probability of a scene graph T_s can be represented as:

$$p(T_s|s) = p(B_s|s)\, p(C_s|b_s, s)\, p(R_s|C_s, B_s, s) \tag{2}$$

Namely, given an image, SGG algorithm detects the bounding boxes B_s ($p(B_s|s)$), classifies the bounding boxes into visual classes C_s ($p(C_s|B_s, s)$), and detects the relations R_s among the bounding boxes/classes ($p(R_s|C_s, B_s, s)$ [43].

Typically, the scene graph ontology can be expressed as a set of nodes N and edges E such that nodes consist of both visual class nodes N_C and predicate nodes N_P. Specifically,

$$G = \{N = N_C \cup N_P, E\} \tag{3}$$

The SG edges can be further formalized as two sets of directed edges [39]:

$$\{E^{P\to C}_{hasSubject}, E^{P\to C}_{hasObject}\} \tag{4}$$

When we introduce a knowledge graph (KG) to build a bipartite SG-KG, we have entities from both the scene and the KG. The existing entities and predicates from the image are now called scene entities and scene predicates. The entities and predicates from the KG are called concept entities and concept predicates. We denote scene entities as N_{SE} and scene predicates as $N_{CP}$with KG concepts corresponding to the visual classes or *concept entities* N_{CE}. We also proxy visual relations or *scene predicates* N_{SP} with ontological concepts corresponding to the visual relations, *i.e.*, concept predicates N_{CP}. We use two-directional edges, which can be expressed as

$$\begin{aligned}\{&E^{SE\to CE}_{classifiedTo}, E^{SP\to CP}_{classifiedTo},\\ &E^{CE\to SE}_{hasInstance}, E^{CP\to SP}_{hasInstance}\}\end{aligned} \tag{5}$$

. The SGG task becomes classifying the scene entities and scene predicates, given concept entities and concept predicates [39]:

$$\begin{aligned}p(N_{SE}, N_{SP}, E_S|I, N_{CE}, N_{CP}, E_C) = \\ p(N^?_{SE} N^?_{SP}, E_S|I)\times \\ p(E_B)|I, N_{CE}, N_{CP}, E_C, N^?_{SE}, N^?_{SP}, E_S),\end{aligned} \tag{6}$$

where E_B are the bridge edges. The unknown nodes in the graph to be learned are denoted with question marks.

3.2. Framework Overview

The overall pipeline of our framework is shown in Figure 3, which contains two major components. To fully exploit the unbiasing potential of knowledge, our first component aims to enrich the KG with additional knowledge by increasing its volume and depth. Since GB-Net [39] exhaustively adds ConceptNet [26] entities that are related to scene entities, we can benefit from using an additional

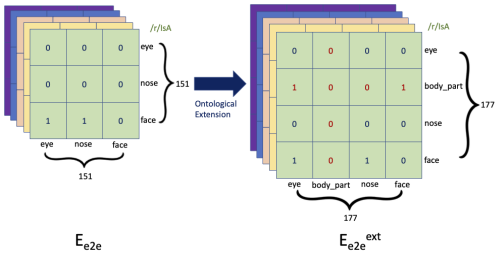

Figure 4: The effect of ontological extension on the knowledge edge matrix. The red values correspond to the addition of the off-scene "body_part" entity and its edges. We include 26 off-scene entities from Wikidata.

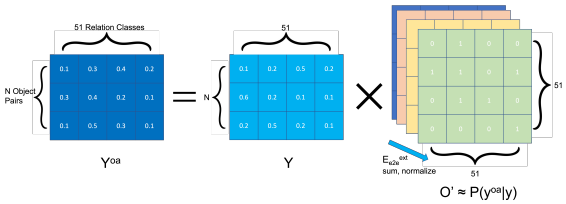

Figure 5: The Ontological Adjustment (OA) module rebalances the distribution of predicates with knowledge. The Ontological Matrix O' from Equation 9 on the right side approximates the conditional probability of knowledge-adjusted relations y^{oa} given the predicated relations y in the neural network model using commonsense knowledge.

data source, Wikidata [11], and extend the knowledge graph topology to add off-scene concept entities. We call this extended bipartite graph Extended Bridge Network (EB-Net).

Our second component explicitly resamples the biased distribution of relations based on the KG. Inspired by the SA approach [9] that adjusts the model predictions with the confusion matrix, we use the adjacency matrix from the knowledge graph to adjust the distribution of the prediction logits in the model. We name this method Explicit Ontological Adjustment (EOA).

3.3. Commonsense Knowledge Graph Extension

To enrich the commonsense knowledge graph (KG), we combine the ConceptNet KG [26] used by GB-Net [39] with a subset of Wikidata that includes commonsense knowledge, Wikidata-CS [11]. We choose Wikidata-CS because it shares the same ontology KG edge types (*e.g.*, "/r/PartOf" and "/r/RelatedTo") with ConceptNet, ensuring their conceptual coherence. To link the two KBs, we use the BLINK entity linking algorithm [33] to associate each ConceptNet entity class to Wikidata. To increase the depth of our KG, we relax GB-Net's [39] restriction to on-scene entities and include off-scene ones such as "body_part." For each VG object label, we randomly select one neighbor to prevent the KG from becoming unwieldy. The result is a total of 305 Wikidata KG edges or facts. We keep the 51 most relevant ones such as ⟨arm, /r/IsA, body_part⟩ by manually removing confusing ones such as ⟨bed, /r/PartOf, member⟩ and irrelevant ones such as ⟨boot, /r/UsedFor, torture⟩. Together with the 104 edges from GB-Net [39], we have a combined total of 152 unique edges (with three duplicates between the two KBs). We introduce a total of 26 off-scene entities.

In addition to extending the ontology, we also need to modify the other components of GB-Net [39], including word similarity, entity covariance, predicate covariance, triplet conditional probability, and predicate counts statistics that are previously used by MotifNet [40] and IMP [35]. Since the existing statistical matrices are based on the original 150 VG concept entities, we remap them to include our new off-scene entities with trivial values. After this statistics remapping step, we have constructed our final knowledge graph, Extended Bridge Network (EB-Net). Figure 4 illustrates the effect of this extension.

3.4. Explicit Ontological Adjustment

Inspired by the logit resampling technique of Semantic Adjustment (SA)[9], we use commonsense knowledge to redistribute relations by adjusting the model predictions. We propose a novel method, Ontological Adjustment (OA), to model the underlying probability distribution of relations given *a priori* commonsense knowledge. Intuitively, the model should weigh conceptually related predicates more favorably than unrelated ones. For example, "standing on" should have a higher probability when given a prediction of "on" than "has" because "standing on" and "on" are conceptually related while "standing on" and "has" are not. Notably, unlike SA which derives correlation from a sample-based confusion matrix, OA uses external knowledge which does not suffer from the sample-level relation imbalance.

The adjustment task for the ontological can be expressed by Equation 7:

$$P(y^{oa}|o_{subj}, o_{obj}) = P(y^{oa}|y) \times P(y|o_{subj}, o_{obj}) \quad (7)$$

, where $P(y|o_{subj}, o_{obj}) \in \mathbb{R}^5 1$ is the original predication of an SGG model for 51 predicate categories between subject o_i and object o_j. $P(y^{oa}|y)$ is the predicate conceptual relatedness given by the OA matrix, and $P(y^{oa}|y)$ is the conceptually adjusted prediction of the OA model. Figure 5 illustrates this adjustment process.

To conduct this redistribution, we take advantage of the knowledge-based predicate priors readily available in the EB-Net knowledge graph (KG) edge matrix and multiply it with the original logits. Specifically, we use the predicate-to-predicate subgraph E_{p2p}^{ext} in the overall KG. We then sum E_{p2p}^{ext} along the KG edge type dimension which consists of four edge types. The first three are Wikidata ontological edge types describing how two predicates are related: "Re-

latedTo," "MannerOf," and "MannerOf" in the reverse direction. The last KG edge type is the predicate covariance matrix from [2]. We refer to this summed E^{ext}_{p2p} as the Ontological Matrix O. To avoid dramatically decreasing the probability of predicates without facts, we add a diagonal identity matrix to O. Lastly, we row-normalize O so that each element $o_{k,l}$ represents the conditional probability of $P(r_k|r_l)$:

$$o_{k,l} = \frac{o_{k,l}}{\sum_{m=1}^{51} o_{k,m}}, \tag{8}$$

The final ontological matrix is given by Equation 9

$$O' = Row\_Normalize(O + I_{51}). \tag{9}$$

, which is matrix-multiplied with the original predicted logits Y to get the adjusted logits L^{oa}:

$$Y^{oa} = O' \cdot Y. \tag{10}$$

Finally, we incorporate our baseline SA[9] method which adjusts the logits with a row-normalized confusion matrix C' to measure semantic distance. The final adjusted logits Y^{oa+sa} becomes:

$$\begin{aligned} Y^{oa+sa} &= (C' + O') \cdot Y \\ &= C' \cdot Y + O' \cdot Y. \end{aligned} \tag{11}$$

Lastly, we include the baseline Balanced Predicate Transfer (BPL) [9] pipeline which conducts a domain transfer from the original Visual Genome dataset [15] with a long-tail predicate distribution to its undersampled version with fewer common predicates. BPL undersamples the top-k (k=15) predicates ($n \leq 2000$), leaving other predicates as-is. This domain transfer pipeline requires training the model on VG before retraining the model with additional fully-connected clean classifier layers[9].

3.5. Discussions

In our commonsense knowledge graph extension process, we choose Wikidata [11] as our new data source based on the availability of the BLINK [33] entity linking algorithm and the shared formats and edge types between Wikidata-CS and ConceptNet. However, there exist many alternative commonsense knowledge graphs including Cyc and OpenCyc [16], NELL [20], WebChild 2.0 [27], Atomic [22], and COMET [1]. By extending with more than one source, the knowledge graph may be enriched further with more diverse knowledge. For example, ATOMIC features inferential rules; ATOMIC and COMET also contain social knowledge [22, 1]. In addition, the ontological structure of the commonsense knowledge graph itself can be studied. For example, we can potentially benefit from having even deeper knowledge. One way to study this is by using a KG hop greater than 1. Furthermore, we only extend the concept entity ontology but not the concept predicate ontology, potentially limiting the unbiasing potential of our model. Future work can include the latter effort as well. Lastly, the random neighbor selection strategy in the entity linking step does not guarantee relevant facts. However, Wikidata-CS [11] does not offer scores for each edge. Thus, the only alternative would be the manual exclusion of many more edges.

For the EOA module, there may be alternative network designs of the EOA matrix. Summing across the edge type dimension can potentially destroy any effect associated with each edge type. Therefore, it may be beneficial to preserve the edge type dimension and compress the logit representation using another fully-connected neural network layer. Additionally, we could remove the edge directions by treating both directions as one for all edge types. Section 4.3 discusses these variations of the ontological adjustment process and their impacts.

4. Experiments

In this section, we first discuss the dataset we used, the SGG tasks, our baselines, and evaluation metrics. We also describe our implementation details, such as hyperparameters and training. Furthermore, we analyze our results in the context of our baselines. Lastly, we examine our model components in ablation studies.

4.1. Datasets and Settings

SGG Tasks. To evaluate the performance of the proposed approach and baselines on scene graph generation, we focus on the following two tasks, predicate classification (PredCls) and scene graph classification (SGCls), which have been commonly used in the SGG literature. In particular, the PredCls task requires the algorithm to classify predicates given both ground-truth bounding boxes and ground-truth object labels, and the SGCls task removes the object labels given in PredCls and thus must also classify the objects in the given bounding boxes.

Dataset. In our experiments, we use the Visual Genome [15] dataset that consists of 108,077 images, each with bounding boxes, object labels, and relation labels.

Baselines. We first include the experiment results from classic methods such as IMP [35], MotifNet [40], and VCTree [29]. Additionally, we reprint the reported performance of our most relevant baseline, GB-Net [39]. Moreover, we include another relevant baseline that balances the biased distribution, G2S (including BPL and SA modules) [9]. Finally, we report recent methods that have demonstrated competitive unbiasing performance, including PCPL [36], DT2-ACBS [5], DLFE [3], and CogTree [38].

Evaluation Metrics. The Graph Constraint (GC) limits a model to only one guess when predicting a relation. For the same model, its unconstrained (UC) performance would

Table 1: Evaluation in terms of mean triplet recall at top 20, 50, and top 100 (mR@K), with and without Graph Constraint (GC), for Predicate Classification (PredCls) and Scene Graph Classification (SGCls) tasks. Numbers are in percentage. The highest-performing method for each metric is shown in bold, and the second best one is in blue.

	PredCls						SGCls					
	mR@20		mR@50		mR@100		mR@20		mR@50		mR@100	
Model	UC	C	UC	C	UC	C	UC	C	UC	C	UC	C
IMP+ [35]	-	-	20.3	9.8	28.9	10.5	-	-	12.1	9.8	16.9	10.5
Neural Motifs [40]	-	10.8	-	14.0	-	15.3	-	6.3	-	7.7	-	8.2
VCTree [29]	-	14.0	-	17.9	-	19.4	-	8.2	-	10.1	-	10.8
PCPL [36]	-	-	50.6	35.2	62.6	37.8	-	-	**26.8**	18.6	**32.8**	19.6
GB-Net [39]	-	-	41.1	19.3	55.4	20.9	-	-	21.4	9.6	29.1	10.2
DT2-ACBS [5]	-	27.4	-	35.9	-	**39.7**	-	**18.7**	-	**24.8**	-	**27.5**
G2S: Transformer + BPL + SA [9]	-	26.7	-	31.9	-	34.2	-	15.7	-	18.5	-	19.4
G2S: MotifNet + BPL + SA [9]	-	24.8	-	29.7	-	31.7	-	14.0	-	16.5	-	17.5
G2S: VCTree + BPL + SA [9]	-	26.2	-	30.6	-	32.6	-	17.2	-	20.1	-	21.2
MotifNet+DLFE [3]	-	22.1	-	26.9	-	28.8	-	12.8	-	15.2	-	15.9
SG-Transformer + CogTree [38]	-	22.9	-	28.4	-	31.0	-	13.0	-	15.7	-	16.7
EB-Net + EOA (Ours)	**39.8**	**30.8**	**54.9**	**36.7**	**66.3**	39.2	**19.6**	14.9	26.7	17.3	32.5	18.3

be higher than constrained (C) because it is allowed multiple guesses for the same task. We list both C and UC results in our study.

Given our motivation to unbias SGG, we report the top-k mean triplet recall (mR@K) results. As Chen *et al*. [2] and Tang *et al*. [29] point out, the traditional top-k recall (R@K) metric is biased by the long-tail distribution of relation labels. A model can perform well on the R@K metric by guessing the most frequent relations. Both papers adopt the mR@K metric that takes the average across all R@Ks for each predicate, where a low R in a tail class would greatly reduce the mR. Conversely, a perfect unbiasing method would have R@Ks equal to mR@Ks because it would achieve the same R@K for all relations. Many works have since only reported mR@K [28, 9] results. In our experiments, we report the mR@K results for PredCls and SGCls with K=20, 50, and 100, with and without the GC (i.e., UC and C).

Implementation Details. We use the Adam [14] optimizer with a learning rate of 1×10^{-3} and an epsilon of 1×10^{-3}. We use single-GPU training with 30 epochs each on NVIDIA A5000 (24GB) GPUs. We optimize the code such that each task takes around 24 hours to train. We train a PredCls model first and use the best validation epoch to initialize an SGCls model.

4.2. Results and Analysis

As shown in Table 1, our model achieves a significant improvement over all other models for the PredCls task. Specifically, compared with the most relevant baselines, GB-Net and BPL+SA, our model achieves a significant increase of 17.4% in mR@50 over GB-Net [39] and an increase of 4.8% over BPL+SA [9]. These results confirm the effectiveness of commonsense knowledge in unbiasing.

For the SGCls task, our model does not outperform most models though it achieves comparable performance to some strong baselines such as the BPL/SA models. Our model actually obtains the second-best results in some cases. Given the additional bounding box classification task, the cause could be a difference in the object detection backends. For example, GB-Net [39] uses the VGG-16 backbone [25] while the BPL/SA [9] methods use ResNet101-FPN [10, 18]. Desai *et al*. show that switching the detector backbone from VGG to ResNet101-FPN improves VCTree by 2.3% in mR@50 for PredCls and 1.1% in mR@50 for SGCls [5]. As a result, the performance of our model could be further improved by using the ResNet101-FPN backbone. Another strategy is incorporating the Total Direct Effect (TDE) loss by Tang *et al*. [28] to separate the visual reasoning from the biased relation label distribution.

4.3. Ablation Studies

We further study the respective contributions of the enriched knowledge graph and the EOA used in our framework. To verify that additional knowledge translates to performance gains, we deconstruct the final KG and train the model with different combinations of the components. To study the mechanism of EOA, we study the impact of the variations of the OA matrix on performance.

4.3.1 KG Components and Additional Knowledge

As discussed in Section 3.3, GB-Net exploits not only the ConceptNet knowledge graph but also other information

Table 2: Ablation study with regard to EOA on Visual Genome. Numbers are in percentage.

	PredCls						SGCls					
	mR@20		mR@50		mR@100		mR@20		mR@50		mR@100	
Model	UC	C	UC	C	UC	C	UC	C	UC	C	UC	C
EB-Net + BPL + EOA Naïve	34.4	26.7	47.6	31.9	58.7	33.9	19.3	14.3	26.1	16.5	31.7	17.4
EB-Net + BPL + EOA Plus	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
EB-Net + BPL + EOA Folded	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
EB-Net + BPL + EOA Merge	**39.8**	**30.8**	**54.9**	**36.7**	**66.3**	**39.2**	**19.6**	**14.9**	**26.7**	**17.3**	**32.5**	**18.3**
EB-Net + BPL + EOA Compress	32.6	24.7	45.8	28.9	57.4	13.9	13.5	9.6	22.9	12.4	26.6	13.2

Table 3: Ablation study with regard to the role of knowledge in reasoning. Numbers are in percentage.

Model	PredCls (mR@50)
CN	18.1
CN+Wiki	18.6
CN+Stat+Emb	19.3
CN+Wiki+Stat+Emb	**19.5**

such as conditional probabilities and word embedding. To investigate the role of additional knowledge and highlight the contribution of our EB-Net enrichment process, we dissect GB-Net's graph components and examine each one's contribution. GB-Net consists of ConceptNet edges, a statistical matrix, and a GloVe embedding matrix [39]. As Table 3 shows, the statistical and embedding matrices together give a 0.9-1.2 boost to mR@50 on PredCls, while additional Wikidata edges yield a non-trivial 0.2-0.5 performance gain.

4.3.2 Alternative EOA Formulations

We also explore alternative ways to adjust the predicates with knowledge-based correlation. We have designed five distinct EOA formulations described below, whose results are shown in Table 2

1. **Naïve.** We construct a naïve predicate-to-predicate (p2p) matrix by summing the KG p2p edge matrix of size $51\times51\times4$ in the last (edge type) dimension. There are three edge types - two undirected "r/RelatedTo" types, one directed "r/MannerOf" type, and a predicate covariance matrix from GB-Net [39]. The naïve model performed worse than EB-Net+BPL, possibly because the knowledge matrix is relatively sparse, and this may adjust the logit values to zero.

2. **Folded.** To examine the effect of knowledge edge direction, we build an undirected knowledge matrix where we add the naïve EOA matrix to its transpose. This diagonally folded model consistently outputs non-numerical results. This suggests that direction does matter.

3. **Plus.** Concerned about the potentially harmful effect of sparsity in the knowledge matrix, we try to shift the value of the edge matrix by one, but this model also outputs non-numerical results.

4. **Merge.** We combine the EOA matrix with the SA confusion matrix by summing the two before adjusting the predicate prediction logits. This is equivalent to the products of two adjustments summed together in the iterative message propagation process. This formulation proves most effective for PredCls and is the final model that we report in Table 1.

5. **Compress.** Lastly, we use a fully-connected compression neural network to preserve all the channel dimensions of the edge matrix consisting of three edge types and one covariance matrix. The training process for this method is slow to converge, and it underperforms compared with Merge.

5. Conclusion

In this paper, we demonstrate that commonsense knowledge has a significant contribution to reducing bias in relation reasoning in SGG. We illustrate this effect by designing a new framework for SGG. First, our model increases the volume and depth of the knowledge graph by incorporating additional facts, consisting of both on-scene and off-scene class entities. Second, our model adjusts the long-tail prediction logits with the knowledge-based statistical priors. Through extensive experiments, we show that our model with the above two improvements can achieve competitive performance compared to the state-of-the-art SGG methods. Additionally, unbiasing SGG methods like ours necessarily sacrifice recall for mean recall until we reach perfect unbiasing. This trade-off may not be acceptable in all situations and requires further investigation.

Acknowledgement: This work is supported by the U.S. Army Research Office Award under Grant Number W911NF-21-1-0109.

References

[1] Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Çelikyilmaz, and Yejin Choi. Comet: Commonsense transformers for automatic knowledge graph construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2019.

[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 (CVPR)*, 2019.

[3] Meng-Jiun Chiou, Henghui Ding, Hanshu Yan, Changhu Wang, Roger Zimmermann, and Jiashi Feng. Recovering the unbiased scene graphs from the biased ones. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1581—-1590, 2021.

[4] Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder–decoder approaches. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 103–111, Doha, Qatar, Oct. 2014. Association for Computational Linguistics.

[5] A. Desai, T. Wu, S. Tripathi, and N. Vasconcelos. Learning of visual relations: The devil is in the tails. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15384–15393, 2021.

[6] Anthony Fader, Luke Zettlemoyer, and Oren Etzioni. Open question answering over curated and extracted knowledge bases. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, page 1156–1165, 2014.

[7] Jiuxiang Gu, Shafiq Joty, Jianfei Cai, and Gang Wang. Unpaired image captioning by language pivoting. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 503–519, 2018.

[8] Jiuxiang Gu, Handong Zhao, Zhe Lin, Sheng Li, Jianfei Cai, and Mingyang Ling. Scene graph generation with external knowledge and image reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[9] Yuyu Guo, Lianli Gao, Xuanhan Wang, Yuxuan Hu, Xing Xu, Xu Lu, Heng Tao Shen, and Jingkuan Song. From general to specific: Informative scene graph generation via balance adjustment. In *ICCV*, 2021.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[11] Filip Ilievski, Pedro A. Szekely, and Daniel Schwabe. Commonsense knowledge in wikidata. In Lucie-Aimée Kaffee, Oana Tifrea-Marciuska, Elena Simperl, and Denny Vrandecic, editors, *Proceedings of the 1st Wikidata Workshop (Wikidata 2020) co-located with 19th International Semantic Web Conference(OPub 2020), Virtual Conference, November 2-6, 2020*, volume 2773, 2020.

[12] Justin Johnson, Agrim Gupta, and Li Fei-Fei. Image generation from scene graphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1219–1228, 2018.

[13] 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 (CVPR)*, 2015.

[14] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *The Third International Conference on Learning Representations*, 2015.

[15] 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.

[16] Douglas B. Lenat. CYC: A large-scale investment in knowledge infrastructure. *Commun. ACM*, 38(11):32–38, 1995.

[17] Yujia Li, Daniel Tarlow, Marc Brockschmidt, and Richard S. Zemel. Gated graph sequence neural networks. In Yoshua Bengio and Yann LeCun, editors, *The 4th International Conference on Learning Representations*, 2016.

[18] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 936–944, 2017.

[19] Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. In *The 9th International Conference on Learning Representations*, 2021.

[20] T. Mitchell, W. Cohen, E. Hruschka, P. Talukdar, J. Betteridge, A. Carlson, B. Dalvi, M. Gardner, B. Kisiel, J. Krishnamurthy, N. Lao, K. Mazaitis, T. Mohamed, N. Nakashole, E. Platanios, A. Ritter, M. Samadi, B. Settles, R. Wang, D. Wijaya, A. Gupta, X. Chen, A. Saparov, M. Greaves, and J. Welling. Never-ending learning. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, page 2302–2310. AAAI Press, 2015.

[21] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1137–1149, 2017.

[22] Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A. Smith, and Yejin Choi. ATOMIC: an atlas of machine commonsense for if-then reasoning. In *The Thirty-Third AAAI Conference on Artificial Intelligence*, pages 3027–3035, 2019.

[23] Sebastian Schuster, Ranjay Krishna, Angel Chang, Li Fei-Fei, and Christopher D Manning. Generating semantically precise scene graphs from textual descriptions for improved image retrieval. In *Proceedings of the fourth workshop on vision and language*, pages 70–80, 2015.

[24] Jiaxin Shi, Hanwang Zhang, and Juanzi Li. Explainable and explicit visual reasoning over scene graphs. In *Proceedings*

of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019.

[25] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In Yoshua Bengio and Yann LeCun, editors, *The 3rd International Conference on Learning Representations*, 2015.

[26] Robyn Speer, Joshua Chin, and Catherine Havasi. Conceptnet 5.5: An open multilingual graph of general knowledge, 2017.

[27] Niket Tandon, Gerard de Melo, and Gerhard Weikum. WebChild 2.0 : Fine-grained commonsense knowledge distillation. In *Proceedings of ACL 2017, System Demonstrations*, pages 115–120, Vancouver, Canada, July 2017. Association for Computational Linguistics.

[28] Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang. Unbiased scene graph generation from biased training. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3713–3722, 2020.

[29] 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 (CVPR)*, 2019.

[30] Denny Vrandečić and Markus Krötzsch. Wikidata: A free collaborative knowledgebase. *Commun. ACM*, 57(10):78–85, sep 2014.

[31] Peng Wang, Qi Wu, Chunhua Shen, Anthony Dick, and Anton Van Den Hengel. Fvqa: Fact-based visual question answering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(10):2413–2427, 2017.

[32] Xiaolong Wang, Yufei Ye, and Abhinav Gupta. Zero-shot recognition via semantic embeddings and knowledge graphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[33] Ledell Wu, Fabio Petroni, Martin Josifoski, Sebastian Riedel, and Luke Zettlemoyer. Zero-shot entity linking with dense entity retrieval. In *EMNLP*, 2020.

[34] Qi Wu, Chunhua Shen, Peng Wang, Anthony Dick, and Anton Van Den Hengel. Image captioning and visual question answering based on attributes and external knowledge. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6):1367–1381, 2017.

[35] 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 (CVPR)*, 2017.

[36] Shaotian Yan, Chen Shen, Zhongming Jin, Jianqiang Huang, Rongxin Jiang, Yaowu Chen, and Xian-Sheng Hua. *PCPL: Predicate-Correlation Perception Learning for Unbiased Scene Graph Generation*, page 265–273. 2020.

[37] Xu Yang, Kaihua Tang, Hanwang Zhang, and Jianfei Cai. Auto-encoding scene graphs for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[38] Jing Yu, Yuan Chai, Yujing Wang, Yue Hu, and Qi Wu. Cogtree: Cognition tree loss for unbiased scene graph generation. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, pages 1274–1280, 2021.

[39] Alireza Zareian, Svebor Karaman, and Shih-Fu Chang. Bridging knowledge graphs to generate scene graphs. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[40] Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. Neural motifs: Scene graph parsing with global context. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5831–5840, 2018.

[41] Yifeng Zhang, Ming Jiang, and Qi Zhao. Explicit knowledge incorporation for visual reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1356–1365, 2021.

[42] Handong Zhao, Quanfu Fan, Dan Gutfreund, and Yun Fu. Semantically guided visual question answering. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1852–1860, 2018.

[43] Guangming Zhu, Liang Zhang, Youliang Jiang, Yixuan Dang, Haoran Hou, Peiyi Shen, Mingtao Feng, Xia Zhao, Qiguang Miao, Syed Afaq Ali Shah, et al. Scene graph generation: A comprehensive survey. *arXiv preprint arXiv:2201.00443*, 2022.