Scene Graph Generation via Conditional Random Fields

Weilin Cong   William Wang   Wang-Chien Lee

Department of Computer Science Engineering, Pennsylvania State University

{wxc272, wqw5158, wlee}@cse.psu.edu

Abstract

Despite the great success object detection and segmentation models have achieved in recognizing individual objects in images, performance on cognitive tasks such as image caption, semantic image retrieval, and visual QA is far from satisfactory. To achieve better performance on these cognitive tasks, merely recognizing individual object instances is insufficient. Instead, the interactions between object instances need to be captured in order to facilitate reasoning and understanding of the visual scenes in an image. Scene graph, a graph representation of images that captures object instances and their relationships, offers a comprehensive understanding of an image. Failing to distinguish subjects and objects in the visual scenes and address the "semantic compatibility" issue, existing techniques on scene graph generation do not perform well with ambiguous object instances in the real-world datasets. In this work, we propose Scene Graph Generation via Conditional Random Fields (SG-CRF), a novel scene graph generation model for predicting object instances and its corresponding relationships in an image. SG-CRF learns the sequential order of subject and object in a relationship triplet and the semantic compatibility of object instance nodes and relationship nodes in a scene graph efficiently. Experiments empirically show that SG-CRF outperforms the state-of-the-art methods on three different datasets, i.e., CLEVR, VRD, and Visual Genome, raising the Recall@100 from 24.99% to 49.95%, from 41.92% to 50.47%, and from 54.69% to 54.77%, respectively.

1. Introduction

In the past few years, research on perceptual tasks such as object detection [3, 26, 20, 25] and segmentation [4, 33, 10] have achieved great success. However, cognitive tasks such as image caption [30, 32], visual QA [28, 1], and semantic image retrieval [8] still face major challenges, as a deeper understanding of the visual scenes in images is required for computers to succeed in these tasks. Towards a better comprehension of visual scenes, it is essential to
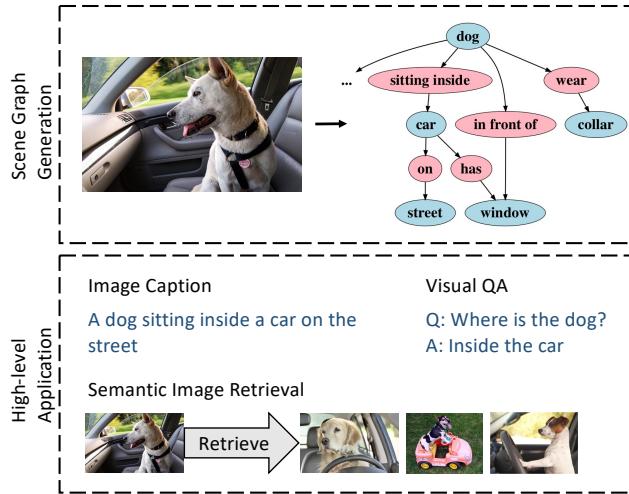

Figure 1: An image and its scene graph generated using all object instances and their relationships in the image.

identify object instances as well as capture the rich semantics embedded in various relationships between object instances in an image. For example, to generate a caption *"a dog is sitting inside a car on the street"*, to answer the question *"Where is the dog?"*, or to retrieve images with similar semantic information (as illustrated in Figure 1), the cognitive models need to not only capture the object instances but also the relationships in an image. Owing to its ability to represent object instances as well as the semantics between them, *scene graphs* have been leveraged in the aforementioned tasks with growing research interests.

Scene graph [8] is a graph representation of visual scenes in images. By capturing object instances and their relationships in an image, it enables a comprehensive understanding of the image. The nodes in a scene graph represent object instances and relationships, where two object instances and a relationship form a relationship triplet $\langle subject - relationship - object \rangle$. Subject and object in the triplet indicates the roles of the object instances, and the edges are pointing from the subject to the relationship and

1

from the relationship to the object.[1] For example, in Figure 1, the scene graph contains object instances, e.g., *"dog"*, *"car"*, *"street"*, *etc*., which are connected with relationships, e.g., *"sitting inside"*, *"on"*, *etc*., to form relationship triplets. *"car"* is a subject in the relationship triplet $\langle car - on - street \rangle$ and an object in relationship triplet $\langle dog - sitting\,inside - car \rangle$. The problem of generating scene graphs from images is an active research topic in computer vision [8, 6, 29].

A major challenge in scene graph generation is reasoning about relationships. The occurrence of relationships varies depending on the object instances involved, which makes the learning of relationships challenging. By fine-tuning the visual features with complex neural networks, previous works achieve good performance by focussing only on a small set of common relationships [27, 22, 31]. Consequently, the performance dramatically decreases when they face real-world datasets with a limited amount of examples per relationship [22] and datasets with a lot of ambiguous entities [11]. Without considering *semantic compatibility* between object instances and relationships, those models could mistakenly assign a high likelihood score to $\langle dog - driving - car \rangle$ for an image with a dog *sitting inside* the car[2]. Furthermore, these models ignore the sequential orders of subjects and objects involved in relationships, result in confusing subjects as objects and vise versa, which may generate prediction such as $\langle car - sitting\,inside - dog \rangle$.

To circumvent these issues, we introduce an end-to-end model , namely ***Scene Graph via Conditional Random Fields (SG-CRF)***, for scene graph generation. Taking an image as input, *SG-CRF* outputs a scene graph that consists of i) object instances localized in the image by bounding boxes, and ii) relationships between each pair of object instances. To distinguish subjects from objects in relationships, we propose an efficient *Relation Sequence Layer (RSL)* that captures the sequential order of subject and object involved. To match ambiguous entities with semantically compatible relationships, we propose a novel *Semantic Compatibility Network (SCN)* that learns the *semantic compatibility* (*i.e.*, the likelihood distribution of a node given all its 1-hop neighbors) of nodes in a scene graph via Conditional Random Fields. For example, in Figure 1, the semantic compatibility of $\langle dog - wear - collar \rangle$ captures the probability of predicting a relationship node as *"wear"*, given its 1-hop neighbors as *"dog"* and *"collar"*. As such, *SG-CRF* can reason the relationship according to the object instances involved. Additionally, *SG-CRF* performs zero-shot relationships detection by leveraging similar relationships, which is crucial for real-world images with complex

---

[1]Note that an object in one relationship triplet could be the subject in another and vice versa.

[2]Notice that $\langle dog - sitting\,inside - car \rangle$ is more semantically compatible than $\langle dog - driving - car \rangle$

relationships and many ambiguous entities.

The major contributions of this work are as follows:

- We reveal via experiments the pitfalls in existing works on scene graph generation, *i.e.*, failing to distinguish subjects from objects and ignoring the semantic compatibility, and propose *SG-CRF* to address these issues.

- We propose a novel RSL to capture the sequential order of subject and object involved in each relationship.

- We propose a novel SCN to learn semantic compatibility of relationships. We show how SCN iteratively optimizes primitive scene graphs by visualizing the internal states of our model.

- We empirically show that SG-CRF outperforms the start-of-the-art methods [22, 31] on three datasets, *i.e.*, CLEVR, VRD, and Visual Genome, raising the Recall@100 from 24.99% to 49.95%, from 41.92% to 50.47%, and from 54.69% to 54.77%, respectively.

The paper is organized as follows. We first review the related work in Section 2 and formulate our research problem in Section 3. We introduce the proposed *SG-CRF* model with implementation details in Section 4, and show experiment results in Section 5. Finally, we conclude the paper in Section 6.

2. Related Work

In this section, we review the literature on relationship reasoning and conditional random fields.

2.1. Relationship Reasoning

One of the major challenges in scene graph generation is reasoning about relationships between subjects and objects.

Lu *et al*. [22] attempt to independently predict object and relationship categories using a visual module, and fine-tune the likelihood of relationship prediction by leveraging language priors from *word2vec*[24] word embeddings. However, [22] ignores the surrounding context to infer individual components of a scene graph in isolation. Nevertheless, individual predictions of object instances and relationships can largely benefit from their surrounding context.

Instead of independently predicting object instances and relationships categories, Xu *et al*. [31] investigate the problem of relationship reasoning by jointly inference relationship with its surrounding context according to the topological structures of scene graphs, *i.e.*, fine-tuning the visual features for each node in the scene graph by leveraging information from its surrounding context. Although the performance of scene graph generation is improved compared with [22], our observation suggests that [31] is likely to confuse subjects from objects, and its performance decreases dramatically when facing real-world images with complex relationships and a lot of ambiguous entities.

More existing works are developed upon [31]. Li *et al.* [16] jointly train the scene graph generation model [31] with an image caption model to capture the semantic levels mutual connections between scene graph generation task and image caption. Li *et al.* [15] further propose to enable message passing within convolutional layers [31], to capture the lower-level visual features for relationship prediction.

However, all existing works focus on enhancing visual features without realizing the deficiencies of their approaches, *i.e.*, unable to distinguish subjects from objects and ignore the semantic compatibility of components in relationships. Their performance decreases greatly when images with complex relationships between object instances are given. In this work, we show that the performance of scene graph generation is improved by addressing these issues, instead of fine-tuning the visual features using a complex neural network structure.

2.2. Conditional Random Fields

Conditional Random Fields (CRFs), a classical tool for modeling complex structures consisting of a large number of interrelated parts, has been used extensively in graph inference. The key idea of using CRFs for graph inference is to incorporate dependencies between vertices in a graph. Much effort has been expended on image segmentation [10, 33, 17], named-entity recognition [23, 13] and image retrieval [8] using CRFs.

Krähenbühl *et al.* [10] propose an efficient CRFs meanfield approximate inference algorithm for image segmentation. They model each image as a fully connected grid graph and use CRFs to refine segmentation results obtained from Fully Convolutional Network [21]. Zheng *et al.* [33, 17] combines the strengths of CNNs with CRFs , and formulate mean-field inference as Recurrent Neural Networks. In the mean time, CRFs reasoning is widely used to classify named object instances [23, 13] in text into predefined categories. Inspired by the great success of CRFs in image segmentation and named-entity recognition, Johnson *et al.* [8] design a CRFs model that reasons about the connections between an image and its ground-truth scene graph, and use these scene graphs as queries to retrieve images with similar semantic meanings.

Our work is closely related to the image segmentation model [33] in that we also coarse-to-fine optimize the initial prediction using CRFs. The critical difference is how we incorporate the dependencies between nodes in a graph. [33] achieves this by assigning each pixel a predefined weight, which is calculated based on each input image by assuming the closer (spacial and color) the pixels, the more likely the same category. Instead, we achieve this by measuring the semantic compatibility of nodes in the scene graph, *i.e.*, assign each ground-truth label a trainable word embedding, and encode the semantic compatibility of nodes in terms of

the likelihood distribution of one node in the scene graph given the word embeddings of all its 1-hop neighbors. Furthermore, scene graphs consist of two independent category sets, object instance category **O** and relationship category **R**. It is impractical to directly incorporate the dependencies between two independent sets. As image segmentation only consists of one category set, *i.e.*, pixel category, word embeddings **E** are introduce in *SG-CRF* to make **O** conditionally dependent on **R** given **E**.

Our work is related to [31] in that we also employ message passing to generate scene graphs. The critical difference is how we use message passing. [31] use message passing to iteratively fine-tune the features of each node in the scene graph in visual features via the Recurrent Neural Network. The performance decrease greatly after two iterations because noises are aggregated in visual features as the number of iterations increases. Instead, we use a message passing to capture the semantic compatibility in the word semantic level. The performance of *SG-CRF* is monotonically improved and converge to the optimal in an average of 1.9 iterations on real-world datasets [11].

3. Problem Formulation

In this work, we aim to generate a scene graph from an image by detecting object instances and predicting relationships simultaneously. Formally, given an image I as input, *SG-CRF* outputs a scene graph $SG = (V_o, V_r, E)$ that consists of object instances localized in the image by bounding boxes, and relationships between each pair of object instances. Here V_o stands for object instance nodes, V_r stands for relationship nodes, and E stands for edges between object instance nodes and relationship nodes.

We denote the label of i-th object instance as $o_i \in V_o$ and its bounding box coordinate as $o_i^{bbox} \in \mathbb{R}^4$. Moreover, we denote the relationship between i-th and j-th object instance as $r_{i \to j} \in V_r$. Edge $(o_i, r_{i \to j}) \in E$ is automatically removed if o_i is classified as *"Background"* or $r_{i \to j}$ is classified as *"No-Relation"*. Let I denote the given input image and SG denote the output scene graph. We formulate the objective for *SG-CRF* as maximizing the following probability function by finding the optimal predictions of $o_i, o_i^{bbox}, r_{i \to j}$.

$$ \mathrm{P}(\mathrm{SG}|\mathrm{I}) = \prod_{o_i \in V_o} \mathrm{P}(o_i, o_i^{bbox}|\mathrm{I}) \prod_{r_{i \to j} \in V_r} \mathrm{P}(r_{i \to j}|\mathrm{I}) \quad (1) $$

To achieve this goal, we aim to adopt CRFs for SG-CRF and face the following challenges: (1) *Relationship reasoning*. Reasoning about relationships is critical to scene graph generation, but the appearance characteristics of relationships vary significantly, making relational reasoning challenging; (2) *CRFs modeling*. CRFs has been used extensively in graph inference, however leveraging CRFs in
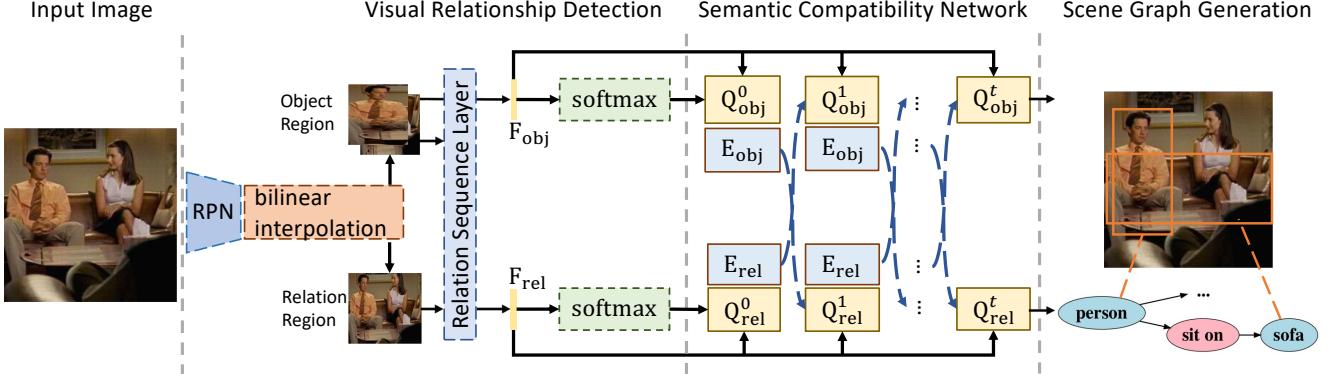
Figure 2: An overview of *SG-CRF* pipeline. *SG-CRF* first generates proposal regions with extracted features. Features are then fed into the *Visual Relationship Detection* component which outputs an initial grounding of the scene graph. To fined-tune the accuracy, our proposed *Semantic Compatibility Network* component extracts word embedding for each object instance and relationship to determine semantic compatibility utilizing marginal probability estimation.

scene graph is not straight forward. An efficient way to incorporate the dependencies between nodes in scene graph is important; (3) *Evaluation*. A comprehensive experiment for evaluating scene graph generation methods is important to showcase the effectiveness of relationship reasoning. Different evaluation setups can help us better understand the performance of *SG-CRF*.

4. The *SG-CRF* Model

Figure 2 presents an overview of our proposed model *SG-CRF*. Our model *SG-CRF* consists of two components: *Visual Relationship Detection (VRD)* and *Semantic Compatibility Network (SCN)*. In the following, we first present our approach of CRFs for scene graph inference in Section 4.1, then introduce the *VRD* component in Section 4.2 and the *SCN* component in Section 4.3. Implementation details are specified in Section 4.4.

4.1. CRFs for Scene Graph

CRFs for scene graph can be formulated as finding the optimal $x^* = \arg\max_x P(X)$ in the form of Gibbs distribution [14]:

$$P(X) = \frac{1}{Z(X)} \exp\big(-\underbrace{\sum_i \psi_u(x_i) - \sum_{j \neq i} \psi_p(x_i, x_j)}_{-E(x)}\big) \quad (2)$$

where the Gibbs energy $E(x)$ is composed of unary and pairwise potentials. The unary potential $\psi_u(x_i)$ measures the cost of assigning i-th node x_i, and pairwise potential $\psi_p(x_i, x_j)$ measures the cost of assigning x_i to i-th node given label assignment x_j of j-th node. For example, in Figure 2, $\psi_u(x_i)$ measures the cost of assigning object instances as *"man"* and *"sofa"* respectively, and $\psi_p(x_i, x_j)$ measures the cost of assigning one of the object instances

as *"man"* after acknowledging another object instance is a *"sofa"*. $Z(X)$ is the partition function [12]. Maximizing the above probability distribution yields the best label assignment for the given image.

In our model, unary potentials ψ_u are computed independently by a *VRD* component, detailed in Section 4.2. The pairwise potentials ψ_p are calculated by the proposed *SCN* component, detailed in Section 4.3.

4.2. Visual Relationship Detection

Given an image as input, *VRD* first generates a set of object proposals using Region Proposal Network (RPN) [26]. *VRD* then extracts the visual features inside the object proposal of the i-th object instance as an object feature embedding F_{obj}^i, and extracts visual features of the union-box over i-th and j-th proposal boxes as a relationship feature embedding $F_{rel}^{i \rightarrow j}$. Previous works tried to fine-tune the relationship feature embeddings $F_{rel}^{i \rightarrow j}$ by combining visual inputs with language priors [22] and passing visual features between object instances and relationships [31]. However, they are likely to confuse the subjects as objects. Instead, we introduce a *RSL* to capture the sequential order of subject and object involved in each relationship.

In practice, we propose two methods to achieve this goal: (1) TransE-concat: inspired by Translation Embedding (TransE) in representing large-scale knowledge bases [2, 19], we interpret the sequential order of subject and object as a *vector translation* and concatenate with the original relationship feature embedding $\hat{F}_{rel}^{i \rightarrow j} = [F_{obj}^i - F_{obj}^j, F_{rel}^{i \rightarrow j}]$; (2) Triple-concat: without using translation embeddings, we intuitively concatenate the relationship feature embedding with its corresponding subject and object feature embeddings $\hat{F}_{rel}^{i \rightarrow j} = [F_{obj}^i, F_{rel}^{i \rightarrow j}, F_{obj}^j]$. Both object instance feature embeddings F_{obj} and relationship feature embedding $\hat{F}_{rel}$ are used to generate the unary poten-

tials $\psi_u = \{\psi_u(\mathrm{x}_i), \psi_u(\mathrm{x}_{i \to j})\}$ as follows.

$$\psi_u(\mathrm{x}_i) = f(W_{obj} \otimes \hat{F}_{obj}^i + b_{obj}) \qquad (3)$$

$$\psi_u(\mathrm{x}_{i \to j}) = f(W_{rel} \otimes \hat{F}_{rel}^{i \to j} + b_{rel}) \qquad (4)$$

Experiments in Section 5.3 demonstrate that these novel ideas in *VRD* not only help to extract meaningful visual relation features but also give *SG-CRF* the ability to distinguish subjects from objects.

4.3. Semantic Compatibility Network

As shown in Figure 2, we propose the *SCN* to learn the semantic compatibility of nodes in the scene graph, and improve the accuracy of scene graph generation. We model this by assigning each ground-truth object instance and each relationship category a trainable word embedding. Pairwise potential $\psi_p(\mathrm{x}_i, \mathrm{x}_j)$ is formulated as predicting the marginal probability estimation of i-th node given the label word embeddings of the j-th node, where the j-th node is one of the 1-hop neighbors of the i-th node. Due to the densely connected structure of the scene graph, the exact maximization is NP-hard. For example, knowing one of the object instances in Figure 2 as *"man"* can affect the confidence of assigning another object instance as *"sofa"*, which in turn can affects the confidence of predicting the first object instance as *"man"*. The *SCN* approximates scene graph inference by mean-field approximation algorithm, detailed in Algorithm 1.

Given the unary potentials ψ_u as input, *SCN* first generates the initial grounding $Q^0 = \{Q_{obj}^0, Q_{rel}^0\}$ using *Softmax* normalization. *SCN* then gathers label word embeddings $E = \{E_{obj}, E_{rel}\}$ for each object instance and relationship according to the Q^0, and iteratively coarse-to-fine updates marginal probability estimation Q^t. An mean-field iteration can be expressed as $Q^t = \mathrm{Mean\,Field}(\psi_u, L_e, Q^{t-1})$, where ψ_u denotes all unary potentials learnt by *VRD* component, L_e denotes the word embeddings for ground-truth categories, W_o, b_o and W_r, b_r denote the parameters shared among all iteration. During each mean-field iteration, $\mathrm{Mean\,Field}(\psi_u, L_e, Q^{t-1})$ first lookups the word embeddings E of predicted label of each node according to the previous estimation of marginal probability Q^{t-1}. Then the pairwise potential ψ_p of each node is calculated based on the label word embeddings of its 1-hop neighbors. We use W_o, b_o if it is an object instance node and use W_r, b_r if it is a relationship node. After that, we update unary potential $\psi_u(\mathrm{x}_i)$ of each node x_i using its corresponding pairwise potential $\psi_p(\mathrm{x}_i, \mathrm{x}_j)$. Finally, we take *Softmax* to normalize the updated unary potential $\hat{\psi}_u$ as the current estimation of marginal probability, which is used for next mean-field iteration. The output Q^T of last mean-field iteration is the likelihood distribution of each node in a scene graph.

Algorithm 1 Individual steps of *Semantic Compatibility Network*.

---

procedure Mean Field(ψ_u, L_e, Q^{t-1})
$\quad \hat{Q}^{t-1} \leftarrow \mathrm{argmax}(Q^{t-1})$
$\quad E \leftarrow L_e(\hat{Q}^{t-1})$ $\qquad \triangleright$ Embeddings Lookup
$\quad$**for** x_i in all nodes **do**
$\qquad$**if** $\mathrm{x}_i \in V_o$ **then**
$\qquad\quad W, b = W_o, b_o$
$\qquad$**if** $\mathrm{x}_i \in V_r$ **then**
$\qquad\quad W, b = W_r, b_r$
$\qquad \psi_p(\mathrm{x}_i, \mathrm{x}_j) \leftarrow f(W \otimes \sum_{\mathrm{x}_j \in N(\mathrm{x}_i)} E_j + b)$
$\qquad\qquad\qquad\qquad\qquad \triangleright$ Message Passing
$\qquad \hat{\psi}_u(\mathrm{x}_i) \leftarrow \psi_u(\mathrm{x}_i) + \psi_p(\mathrm{x}_i, \mathrm{x}_j)$
$\qquad\qquad\qquad\qquad\qquad \triangleright$ Unary Update
$\quad$**return** $\hat{\psi}_u$

Semantic Compatibility Network
for $t : 0 \to T$ **do**
$\quad$**if** $t = 0$ **then**
$\qquad Q^0 \leftarrow \mathrm{Softmax}(\psi_u)$
$\quad$**else**
$\qquad \hat{\psi}_u \leftarrow \mathrm{Mean\,Field}(\psi_u, L_e, Q^{t-1})$
$\qquad Q^t \leftarrow \mathrm{Softmax}(\hat{\psi}_u)$

4.4. Implementation Details

Our scene graph generation network adopts the Imagenet pretrained Faster R-CNN [26] network with the ResNet-50 [5] architecture as the base to incorporate *VRD* and *SCN* for relationship prediction.

During training, RPN [26] generates 256 region proposals. Proposal is positive if it has IoU > 0.7 with some ground-truth regions and is negative if IoU < 0.3. The classification layer takes all positive proposals as input to output class probabilities and bounding box coordinates. Then, we apply Non-Maximum Suppression (NMS) for each class with IoU > 0.5 to de-duplicate bounding boxes with high overlap. During testing, NMS with IoU > 0.5 is applied on all proposals generated by RPN. We use Softmax to produce the final scores for the object instance and relationship prediction, and use the cross-entropy loss for object instance and focal loss [18] for the relationship due to the sparsity of the annotation. We replace the RoI pooling layer with bilinear interpolation operation and train the network including ResNet-50 by stochastic gradient descent with momentum [9]. We set 512 as the size of label word embeddings and 5 as the number of mean-field iterations to perform.

We train our model in a two-stage process: first pre-train the *VRD* component of the network, and then append the *SCN* on the network for end-to-end training. We choose not to train from the beginning with *SCN* as the unary potentials produced by the *VRD* are so poor that performing inference

on them produces meaningless results while increasing the computational time.

5. Experiments

In this section, we evaluate the effectiveness of *SG-CRF* against the state-of-the-art methods[22, 31] on three different visual relationship datasets [7, 22, 11]. We describe the datasets in Section 5.1, the performance metric and baselines in Section 5.2, and analyze results on scene graph generation across three different datasets in Section 5.3.

5.1. Datasets

We evaluate *SG-CRF* on CLEVR [7], VRD [22] and Visual Genome [11].

CLEVR [7] is a synthetic dataset generated from scene graphs, where the relationships are limited to 4 spatial relationships (left, right, front, behind) and 48 object categories. The dataset contains 70,000 training images and 15,000 testing images, each image is accompanied with scene-graph annotations as the ground truth for object locations and relationships. [7] allows us to test the effects of *RSL* without confounding noise in real-world datasets.

VRD [22] is a widely used real-world relationship detection dataset. The dataset has 4,000 training images and 1,000 testing images. There are 100 object categories, 70 relationships categories, and 37,993 relationship instances in the dataset, 4.9% of the relationships are for zero-shot evaluations. With only a few examples per object instance and relationships category, this dataset allows us to evaluate *SCN* when facing relationships infrequent in real-world.

Visual Genome [11] is the most extensive dataset for relationship detection. The dataset contains 108,077 images with over 2.3 million relationship instances. We select 100 most common object instances and 70 most common relationships categories. We manually filter out images with ambiguous annotations, reserve 10,000 images for validation, and treat the remaining images as training data. Experiments on Visual Genome represent a large-scale evaluation of our method, showing that our model can significantly improve relationship detection in scene graphs.

5.2. Metrics, Setups and Baselines

We adopt the Recall@K (R@K) metric to evaluate the performance of *SG-CRF*. The R@K metric measures the total instances where the ground-truth relationship is predicted in the top K most confident relationship predictions over total predictions. This choice is due to the sparsity of the relationship annotations in VRD [22] and Visual Genome [11]. Metrics like Average Precision (AP) would falsely penalize the detection if we do not have that particular ground-truth. Furthermore, missing proposal boxes are inevitable as the number of object instances and relationships increase, which makes fair evaluation difficult.

For example, after non-maximal suppression (NMS), highly overlapping regions such as *"photo frame"* and *"photo"* are suppressed, and relationship $\langle photo - in - photo\,frame \rangle$ cannot be predicted correctly if missing any of these object instances. Nevertheless, we can neither tell if a prediction is wrong nor the proposal boxes itself is missing.

Following previous works [22, 31], we evaluate our model under the following setups:

Scene Graph Generation (SGGEN): given an image, the task is to predict relationships as well as the bounding box location and object labels. We consider an object instance as correctly detected if its IoU >0.5 with the ground-truth box.

Scene Graph Classification (SGCLS): given an image and a set of ground-truth object bounding boxes, the task is to predict a set of possible relationships between pairs of object instances along with object labels. This setup allows us to evaluate the performance of relation triplet classification, without affected by inaccurate object proposal boxes.

Relationship Classification (RELCLS): given an image and a set of ground-truth object bounding boxes, the task is to predict a set of possible relationships between pairs of object instances. This setup aims to evaluate the performance of relationship classification isolated from other factors.

We create two different baseline models to showcase the effectiveness of the model proposed. The first baseline **VRD** is created inspired by [22]. Although [22] consists of a visual module and a language module, we only compare with its visual module because the language module can be added independently as a performance boost to all visual-based models, including ours. Note that baseline *VRD* is similar to our *Visual Relationship Detection* component without *RSL*. The second baseline **SG-Dual** is inspired by [31]. This baseline adds a message passing scheme to baseline *VRD*.

In the following sections, we analyze our experimental results on different datasets respectively.

5.3. Results

Evaluation on CLEVR. We examine the effects of *RSL* without confounding noise in the real-world dataset on CLEVR dataset [7]. Besides the baseline model mentioned above, we create two additional baselines, *i.e.*, **VRD$_{\text{TransE}}$** and **VRD$_{\text{Triple}}$**, which denote two different *RSL* implementations built upon the *VRD* baseline model, as proposed in Section 4.2. From the empirical results in Table 1 and the results shown in Figure 3, we observe that (1) the baseline models ignores the sequential order of object instances evolved in a relationship, thus are likely to confuse subjects as objects. The R@K performance of baseline models are close to random-guessing on a dataset with only four mutually-exclusive relationship categories, *i.e.*, left, right, front, behind; (2) both TransE-concat and Triple-

Table 1: Performances compared against baseline methods on the CLEVR [7] dataset.

Task	SGGEN		SGCLS		RELCLS	
Metric	R@50	R@100	R@50	R@100	R@50	R@100
VRD	22.32	24.99	22.32	24.99	22.32	25.00
VRD$_{TransE}$	43.08	47.91	43.08	47.91	43.08	47.92
VRD$_{Triple}$	**43.88**	**48.61**	**43.88**	**48.62**	**43.89**	**48.63**
SG-Dual	22.30	24.98	22.31	24.98	22.31	24.99
SG-CRF	**44.61**	**49.93**	**44.61**	**49.93**	**44.62**	**49.95**

Table 2: Performances compared against baseline methods on the VRD [22] and Visual Genome [11] dataset.

	Method	SGGEN		SGCLS		RELCLS	
		R@50	R@100	R@50	R@100	R@50	R@100
VRD	VRD	13.39	14.29	16.51	17.74	26.36	28.75
	SG-Dual	21.27	21.99	26.31	27.30	40.06	41.92
	SG-CRF	**24.98**	**25.48**	**31.46**	**32.13**	**49.16**	**50.47**
VG	VRD	13.76	14.66	17.39	18.64	32.89	35.53
	SG-Dual	22.89	23.37	29.27	29.98	**53.16**	54.69
	SG-CRF	**22.95**	**23.54**	**29.29**	**30.14**	53.11	**54.77**

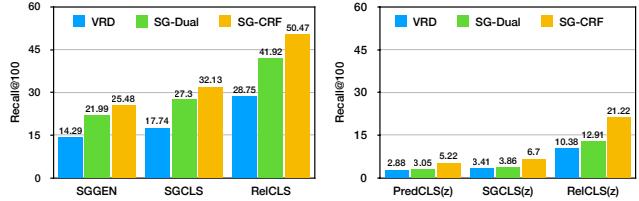Figure 3: Scene graph generation result of baselines on CLEVR [7] dataset.

Figure 4: Performances compared against baselines on the VRD [22] dataset. Baselines with suffix *(z)* perform zero-shot learning.

concat are effective in distinguishing subjects from objects. Triple-concat method slightly outperforms TransE-concat method, because the subtract operation in translation embeddings compress information, which are helpful for relationship prediction, whereas Triplet-concat directly concatenates without compression maintaining as many useful information as possible; (3) spatial information is compressed when feeding images which consist of 3D spatial relationships into models with a 2D receptive field, even though *SG-CRF* has difficulty distinguishing left-right and front-back in several circumstances. In the following sections we use Triple-concat by default.

Evaluation on VRD and Visual Genome. We compare *SG-CRF* against baselines on the real-world dataset VRD [22] and Visual Genome [11]. According to the results shown in Table 2 and the comparative results shown in Figure 5, we observe that (1) *VRD* is likely to confuse subjects from objects. Taking advantage of the contextual information of object instances and its relationships, *SG-Dual* learns the relation triplet orders of frequent relationships in the dataset, and achieves good performance on predicting common relationships such as $\langle zebra - has - tail \rangle$; (2) However, *SG-Dual*'s performance decreases with am-

biguous and infrequent relationships in the dataset, resulting in cyclic relationships such as $\langle fence - behind - tree - behind - fence \rangle$; (3) leveraging the proposed *RSL* and semantic compatibility learnt by *SCN*, *SG-CRF* demonstrates its robustness and effectiveness on VRD [22] and Visual Genome [11].

Zero-shot Evaluation. To extend the evaluation, we further perform zero-shot learning, *i.e.*, inferring unseen relationships in the test set using a similar relationships from the training set. This is practical since it is impossible to build a model trained with every possible relationship. According to the R@100 result shown in Figure 4, we observe that: (1) while the performances of *SG-CRF* and baseline models [22, 31] dramatically decrease, *SG-CRF* still outperforms the others. For example, during zero-shot scene graph generation, the R@100 of *SG-CRF* outperforms the baseline *SG-Dual* by 64.36% and outperforms the baseline *VRD* by 104.43%; (2) visual features are not discriminative enough for baseline models [22, 31] to predict unseen relationships, whereas *SG-CRF* utilizes the semantic compatibility of labels to generalize similar relationships to enable zero-shot predictions. For example, we infer unseen relationship $\langle building - behind - zebra \rangle$ by using similar relationships $\langle building - behind - horse \rangle$ seen before.

Demonstrate Effectiveness of Mean-field Iterations. Figure 6 shows all internal scene graph generation states Q^i of *SG-CRF* until the prediction result become stable (stable at i-th iteration if $Q^i = Q^{i-1}$), where Q^0 denotes the initial state before mean-field algorithm. To keep the visualization interpretable, we only show the relationship predictions for the pairs of object instances that have ground-truth relationship annotations. Taking advantage of the mean-field approximation in CRFs, *Semantic Relation Network* iter-
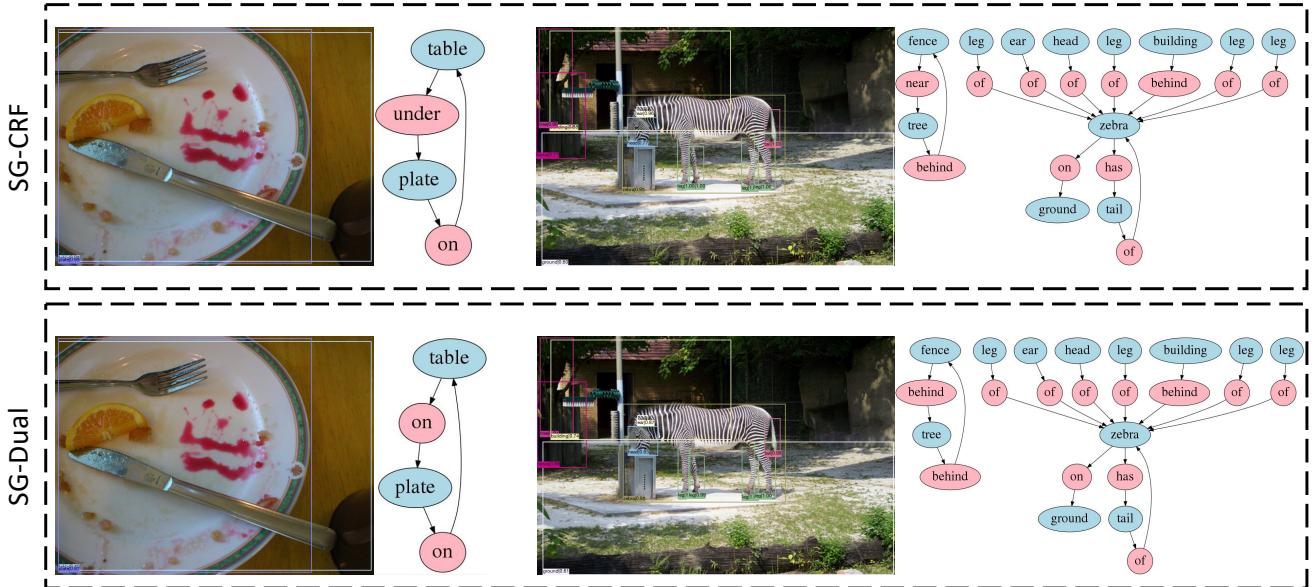
Figure 5: Scene graph generation result of baselines on VRD [22] dataset.
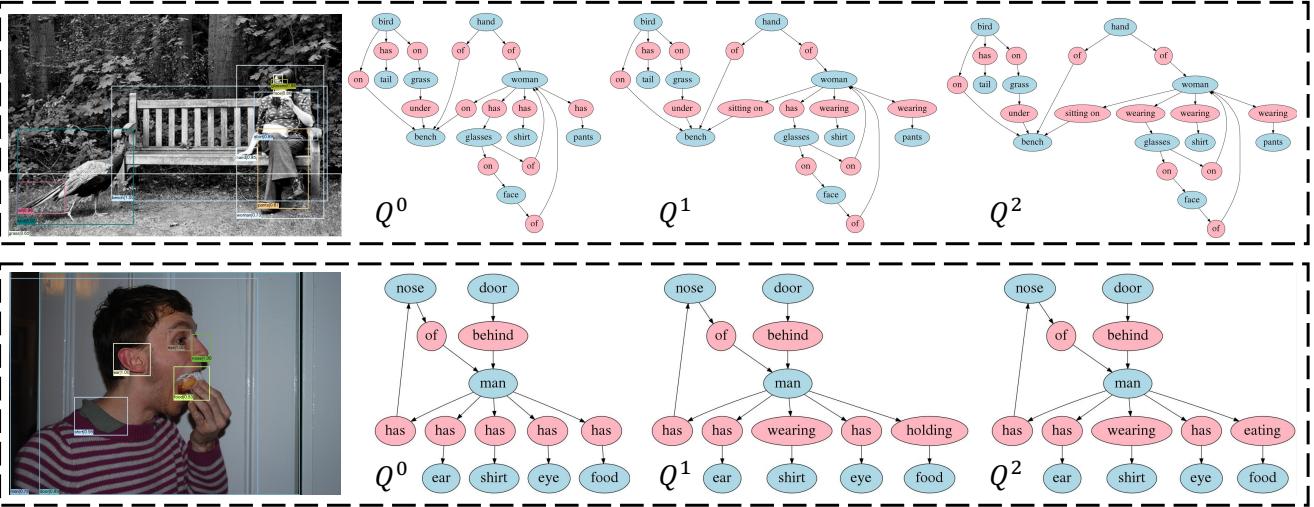

Figure 6: Scene graph generation internal states and result of *SG-CRF* on Visual Genome [11] dataset. Q^i stands for i-th internal state.

atively optimizes the primitive scene graph generated directly from an image. We observe that the result of scene graph classification becomes stable with an average of 1.7 iterations on Visual Genome dataset and 1.9 iterations on VRD dataset.

6. Conclusion

In this work, we study the problem of generating precise scene graph from an image. Our model, *SG-CRF*, learns the sequential order of subject and object via a proposed *Relation Sequence Layer*, and learns the semantic compatibility

in addition to visual features via a novel *Semantic Compatibility Network*. Experiments empirically demonstrate its effectiveness in predicting scene graph on real-world datasets.

References

[1] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang. Bottom-up and top-down attention for image captioning and visual question answering. *arXiv preprint*. 1

[2] A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, and O. Yakhnenko. Translating embeddings for modeling multi-

relational data. In *Advances in neural information processing systems*, pages 2787–2795, 2013. 4

[3] R. Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015. 1

[4] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, pages 2980–2988. IEEE, 2017. 1

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

[6] J. Johnson, A. Gupta, and L. Fei-Fei. Image generation from scene graphs. *arXiv preprint*, 2018. 2

[7] J. Johnson, B. Hariharan, L. van der Maaten, L. Fei-Fei, C. L. Zitnick, and R. Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pages 1988–1997. IEEE, 2017. 6, 7

[8] J. Johnson, R. Krishna, M. Stark, L.-J. Li, D. Shamma, M. Bernstein, and L. Fei-Fei. Image retrieval using scene graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3668–3678, 2015. 1, 2, 3

[9] D. Kinga and J. B. Adam. A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015. 5

[10] P. Krähenbühl and V. Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In *Advances in neural information processing systems*, pages 109–117, 2011. 1, 3

[11] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. 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. 2, 3, 6, 7, 8

[12] J. Lafferty, A. McCallum, and F. C. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. 2001. 4

[13] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer. Neural architectures for named entity recognition. *arXiv preprint arXiv:1603.01360*, 2016. 3

[14] L. Landau and E. Lifshitz. *Statistical Physics: Course of Theoretical Physics/Translated from the Russian by E. Peierls and RF Peierls*. Pergamon Press, 1958. 4

[15] Y. Li, W. Ouyang, X. Wang, and X. Tang. Vip-cnn: Visual phrase guided convolutional neural network. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pages 7244–7253. IEEE, 2017. 3

[16] Y. Li, W. Ouyang, B. Zhou, K. Wang, and X. Wang. Scene graph generation from objects, phrases and region captions. In *ICCV*, 2017. 3

[17] X. Liang, X. Shen, J. Feng, L. Lin, and S. Yan. Semantic object parsing with graph lstm. In *European Conference on Computer Vision*, pages 125–143. Springer, 2016. 3

[18] T. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar. Focal loss for dense object detection. pages 2999–3007, 2017. 5

[19] Y. Lin, Z. Liu, M. Sun, Y. Liu, and X. Zhu. Learning entity and relation embeddings for knowledge graph completion. In *AAAI*, volume 15, pages 2181–2187, 2015. 4

[20] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg. Ssd: Single shot multibox detector. *arXiv preprint arXiv:1512.02325*, 2015. 1

[21] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 3

[22] C. Lu, R. Krishna, M. Bernstein, and L. Fei-Fei. Visual relationship detection with language priors. In *European Conference on Computer Vision*, pages 852–869. Springer, 2016. 2, 4, 6, 7, 8

[23] A. McCallum and W. Li. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, pages 188–191. Association for Computational Linguistics, 2003. 3

[24] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc., 2013. 2

[25] J. Redmon, S. Divvala, R. Girshick, and A. 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. 1

[26] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015. 1, 4, 5

[27] M. A. Sadeghi and A. Farhadi. Recognition using visual phrases. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1745–1752. IEEE, 2011. 2

[28] D. Teney, L. Liu, and A. van den Hengel. Graph-structured representations for visual question answering. *CoRR, abs/1609.05600*, 3, 2016. 1

[29] D. Teney, L. Liu, and A. van den Hengel. Graph-structured representations for visual question answering. *arXiv preprint*, 2017. 2

[30] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. In *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*, pages 3156–3164. IEEE, 2015. 1

[31] D. Xu, Y. Zhu, C. B. Choy, and L. Fei-Fei. Scene graph generation by iterative message passing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, 2017. 2, 3, 4, 6, 7

[32] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning*, pages 2048–2057, 2015. 1

[33] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. H. Torr. Conditional random fields as recurrent neural networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1529–1537, 2015. 1, 3