Weakly Supervised Visual Semantic Parsing

Alireza Zareian, Svebor Karaman, and Shih-Fu Chang
Columbia University, New York, NY, USA
{az2407,sk4089,sc250}@columbia.edu

Abstract

Scene Graph Generation (SGG) aims to extract entities, predicates and their semantic structure from images, enabling deep understanding of visual content, with many applications such as visual reasoning and image retrieval. Nevertheless, existing SGG methods require millions of manually annotated bounding boxes for training, and are computationally inefficient, as they exhaustively process all pairs of object proposals to detect predicates. In this paper, we address those two limitations by first proposing a generalized formulation of SGG, namely Visual Semantic Parsing, which disentangles entity and predicate recognition, and enables sub-quadratic performance. Then we propose the Visual Semantic Parsing Network, VSPNET*, based on a dynamic, attention-based, bipartite message passing framework that jointly infers graph nodes and edges through an iterative process. Additionally, we propose the first graph-based weakly supervised learning framework, based on a novel graph alignment algorithm, which enables training without bounding box annotations. Through extensive experiments, we show that* VSPNET *outperforms weakly supervised baselines significantly and approaches fully supervised performance, while being several times faster. We publicly release the source code of our method*[1].

1. Introduction

Deep learning has excelled in various tasks such as object detection [33] and speech recognition [1], but it falls short of tasks that require deeper semantic understanding and reasoning, such as Visual Question Answering (VQA) [47]. Motivated by the success of structured representations in natural language processing [2, 34, 37], computer vision has started to adopt *scene graphs* to improve performance and explainability, in various tasks such as VQA [35, 12], image captioning [42], and image retrieval [14]. The task of Scene Graph Generation (SGG) [40] aims to represent an image with a set of entities (nodes) and predicates (directed edges), as illustrated in Figure 1 (bottom). Several methods have been proposed to address this problem [40, 20, 41, 48], but despite their success, important challenges remain unaddressed.

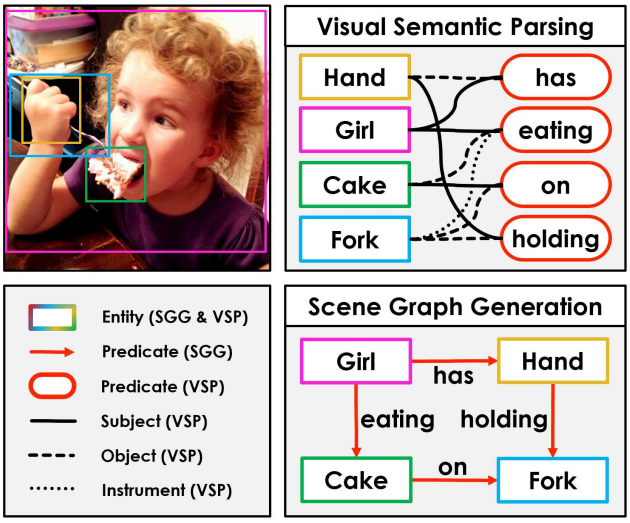

Figure 1. An example of structured scene understanding formulated as Scene Graph Generation, where predicates are edges, compared to the proposed Visual Semantic Parsing, where predicates are nodes and edges represent semantic roles.

Most existing methods are computationally inefficient, as they exhaustively process every pair of object proposals, in order to detect predicates. This results in a quadratic order with respect to the number of proposals. Extending to higher-order interactions has not been studied, and would make this problem even more complex. Furthermore, existing SGG methods require bounding box annotation for each object (node) in ground truth graphs, over the entire training data, which is an expensive constraint. We argue that SGG should ideally be disentangled from bounding box localization, so it can focus on high-level semantic and relational reasoning rather than low-level boundary analysis. However, weakly supervised SGG has barely been studied, and the performance is far from supervised methods [50].

To advance structured scene understanding, we propose the Visual Semantic Parsing Network (VSPNET), which

[1] https://github.com/alirezazareian/vspnet

aims to address the two mentioned limitations, *i.e*., computation and supervision costs. To this end, we generalize the formulation of SGG to represent predicates as nodes in the same semantic space as entity nodes, and instead, represent *semantic roles* (*e.g*. subject and object) as edges. Figure 1 (top) illustrates the proposed *Visual Semantic Parsing* (VSP) formalism. This not only allows us to break the quadratic complexity, but also can support higher-order interactions that cannot be expressed using the existing SGG formulation. For instance, the semantic structure of a `girl eating cake` *with* `fork` can be represented as a predicate node, `eating`, connected to three entity nodes `girl`, `cake` and `fork`, via three types of edges that are labeled with *subject*, *object* and *instrument* roles respectively.

Based on this new VSP formulation, we propose a dynamic, attention-based, bipartite message passing framework, which jointly infers node labels and edge labels through an iterative process, resulting in a VSP graph, and in turn a scene graph. VSPNET consists of a *role-driven* attention mechanism to dynamically estimate graph edges, along with a novel three-stage message aggregation network to route messages efficiently throughout the graph. These two modules successively refine nodes and edges of the graph, enabling a joint inference through global reasoning. The proposed architecture does not need to process all pairs of object proposals and hence is computationally efficient. Finally and most importantly, we propose a novel framework to train VSPNET in weakly supervised settings, by defining a two-stage optimization problem and devising a novel graph alignment algorithm to solve it.

Through extensive experiments on the Visual Genome dataset, we show that our method achieves significantly higher accuracy compared to weakly supervised counterparts, approaching fully supervised baselines. We also show that VSPNET is easily extendable to the fully supervised setting, where it can utilize bounding box annotations to further improve performance, and outperform the state of the art. Moreover, we show that our method is several times faster than all baselines, and qualitatively demonstrate its ability to extract higher-order interactions, which are beyond the capability of any existing method.

2. Related work

Structured scene understanding: Deep learning often simplifies computer vision into classification or detection tasks that aim to extract visual concepts such as objects or actions in isolation. Lu *et al*. [23] took a key step forward by defining Visual Relationship Detection (VRD) [49, 50, 21, 7, 31, 45, 51, 13], which aims to classify relationships between pairs of objects detected in a scene. Their definition of "relationship", also known as *predicate*, includes verbs (*e.g*. `eating`), spatial positions (*e.g*. `above`), and comparative adjectives (*e.g*. `taller than`). Human-Object Interaction (HOI) detection [9, 4, 15, 32] is a specialized version of VRD that focuses on verbs with a human subject. More recently, Xu *et al*. [40] redefined VRD as Scene Graph Generation (SGG) [20, 27, 48, 41, 19, 39], which aims to jointly detect all objects and predicates in a scene, and represent it as a graph that captures the holistic scene content. SGG assumes exactly two entities (subject and object) involved in each predicate, which is not always the case in the real world. Situation Recognition (SR) [44, 43, 24] resolves that limitation by detecting a verb and all of its arguments in a scene, but does not localize the objects, and is limited to one verb per image. Our proposed VSP can be seen as a generalization of both SGG and SR, representing images with semantic graphs that could contain any number of predicates, localized entities, and semantic roles.

Scene graph generation: The majority of SGG methods start by extracting object proposals from the input image, perform some kind of information propagation (*e.g*. Bi-LSTMs in [48] or Graph Convolutional Nets in [41]) to incorporate context, and then classify each proposal to an entity class, as well as each pair of proposals to a predicate class [40, 20, 48, 19, 39]. This process has a quadratic order and is thus inefficient. Recent methods have tried to reduce the computation by pruning the fully connected graph using a light-weight model [41], or by factorizing the graph into smaller sub-graphs [19]. However, they still suffer from quadratic order. Newell and Deng [27] proposed a method that does not rely on proposals at all, and directly extracts entities and predicates from a pair of feature maps. Our method is similar in that we allocate a constant, sub-quadratic number of predicates and infer their connection to entities, rather than processing all pairs of entities. In contrast with [27] though, we base our graph on object proposals and exploit message passing to incorporate context.

Neural message passing: Recent deep learning methods have increasingly utilized Message Passing (MP) in various computer vision tasks [22, 5, 15]. Most SGG methods use MP to propagate information among object proposals [40, 20, 19, 41]. Instead of relying on a static, often fully-connected graph, we propose a dynamic, bipartite graph that is refined using attention to route messages between relevant entity-predicate pairs. In contrast with other dynamic MP methods that refine graph edges in each step, which have been used in other tasks such as HOI [32] and video object detection [46], we define edges between entities and predicates rather than pairs of entities, leading to computational efficiency, while incorporating the rich semantic role structure through three-stage aggregation.

Weakly supervised learning: Weak Supervision (WS) has been advocated in several areas, such as object, action, and relation detection [3, 36, 50], and is motivated by the fact that manual annotation of boundaries is time consuming. Most WS object detection methods are based upon multiple

instance learning [8], which assumes each ground truth object corresponds to one out of many proposals, but the correspondence is unknown. WSDDN [3] dedicates a network branch to select a proposal for each ground truth. Zhang *et al*. [50] adopted WSDDN for VRD, selecting a pair of proposals for each ground truth relation. In contrast, we define a global optimization problem where the entire output graph has to be aligned with the ground truth graph, rather than considering each predicate independently. Peyre *et al*. [30] defined a global optimization for WS VRD too, but it is limited to a linear regression model for relationship recognition. Our novel WS formulation allows learning with gradient descent, which enables us to train a deep network with a complex message passing architecture.

3. Method

In this section, we first formalize our problem in Section 3.1, then detail our method and its two-fold contributions: the VSPNET architecture for constructing a semantic graph from an image (Section 3.2), and a graph alignment algorithm for weakly supervised training of the proposed network (Section 3.3). Figure 2 illustrates the general pipeline of our method.

3.1. Problem formulation

Given an image I, the goal of SGG is to produce a graph $G_{\text{SGG}} = (\mathcal{N}, \mathcal{E})$ where each node in $\mathcal{N}$ is represented by an entity class $c_i \in \mathcal{C}_e$ and a bounding box b_i, and each edge assigns a predicate class to an ordered pair of nodes, *i.e*., $\mathcal{E} : \mathcal{N} \times \mathcal{N} \mapsto \mathcal{C}_p$. The direction of predicate edges usually follow the order they would appear in an English phrase. For instance, a `person sitting on chair` would be represented as an edge labeled `sitting on`, going from the node `person` to the node `chair`, not the other way.

Nevertheless, this notation is inherently limiting, as it restricts predicates to have exactly two arguments present in the scene. This constraint may be acceptable for relational predicates such as prepositions, but certainly not for verbs, which constitute an important group of predicates. To relax this constraint, we follow [44] to adopt the formulation of Semantic Role Labeling [28], where predicates are represented as nodes, and edges represent semantic roles that entities play in each predicate. Accordingly, we define Visual Semantic Parsing (VSP) as predicting a bipartite graph $G_{\text{VSP}} = (\mathcal{N}_e, \mathcal{N}_p, \mathcal{E})$, where

$$\begin{aligned} \mathcal{N}_e &= \left\{ \left(c_i \in \mathcal{C}_e, b_i \in \mathbb{R}^4\right) \right\}_{i=1}^{n_e}, \\ \mathcal{N}_p &= \left\{ c_k \in \mathcal{C}_p \right\}_{k=1}^{n_p}, \text{ and} \\ \mathcal{E} &: \mathcal{N}_p \times \mathcal{N}_e \mapsto \mathcal{C}_r. \end{aligned} \tag{1}$$

Every scene graph G_{SGG} has an equivalent VSP graph G_{VSP} where each predicate has exactly two roles, subject and object, meaning $\mathcal{C}_r = \{s, o\}$. However, an arbitrary VSP graph does not necessarily map to a scene graph, as a predicate may connect to less or more than two entities, potentially involving other semantic roles such as instrument. Hence, VSP is a generalization of SGG.

In this paper we employ the VSP formalism, not only because it covers a wider range of semantics, but also because it naturally leads to a more efficient model architecture. In order to consider all possible relationships, most existing methods process a fully connected graph with n_e^2 edges, where n_e is usually the number of proposals which is typically 300. This is while more than 99% of graphs in Visual Genome have less than 20 predicates, and the largest one has 53. VSP allows us to replace the n_e^2 edges with a constant number of predicate nodes n_p, far less than n_e^2.

3.2. Visual semantic parsing network

We propose VSPNET, which takes an image as input and generates a VSP graph. To this end, we utilize an object proposal network to initialize a set of *entity nodes*, and devise another module to initialize a set of *predicate nodes*. The goal of VSPNET is to classify each entity and predicate node into entity and predicate classes including background, and classify each entity-predicate pair into predefined edge types (semantic roles) including no-edge. These are two co-dependent tasks as incorporating nodes would be helpful for edge classification and vice versa. But since both of them are unknown and to be determined, our model successively infers each given the other.

More specifically, VSPNET is based on a novel bipartite message passing framework that propagates information from entities to predicates and vice versa, through a *role-driven* attention mechanism that estimates edges. After nodes are updated using the estimated edges, we update edges by recomputing the attention using the new node representations, and repeat this process for u iterations. To incorporate each semantic role separately, we designate an attention head for each role. This leads to a complex routing problem where messages from a potentially large number of nodes have to be propagated through multiple types of edges. Accordingly, we propose a three-stage message aggregation network to efficiently route and collect relevant messages for updating each node.

Formally, we define $H_e^{(0)} \in \mathbb{R}^{n_e \times d_e}$ to be the initial hidden state of n_e entity nodes, and initialize each row using the appearance (RoI [33]) features of the corresponding object proposal, as well as its bounding box coordinates, by feeding them into two fully connected networks $e_a(.)$ and $e_b(.)$, and adding the two outputs. We also define $H_p^{(0)} \in \mathbb{R}^{n_p \times d_p}$ to be the initial hidden state of n_p predicate nodes. $H_p^{(0)}$ is a trainable matrix, randomly initialized before training but fixed during test. Given $H_e^{(t)}$ and $H_p^{(t)}$, we compute a set of attention matrices $\tilde{A}_r^{(t)} \in \mathbb{R}^{n_p \times n_e}$, each

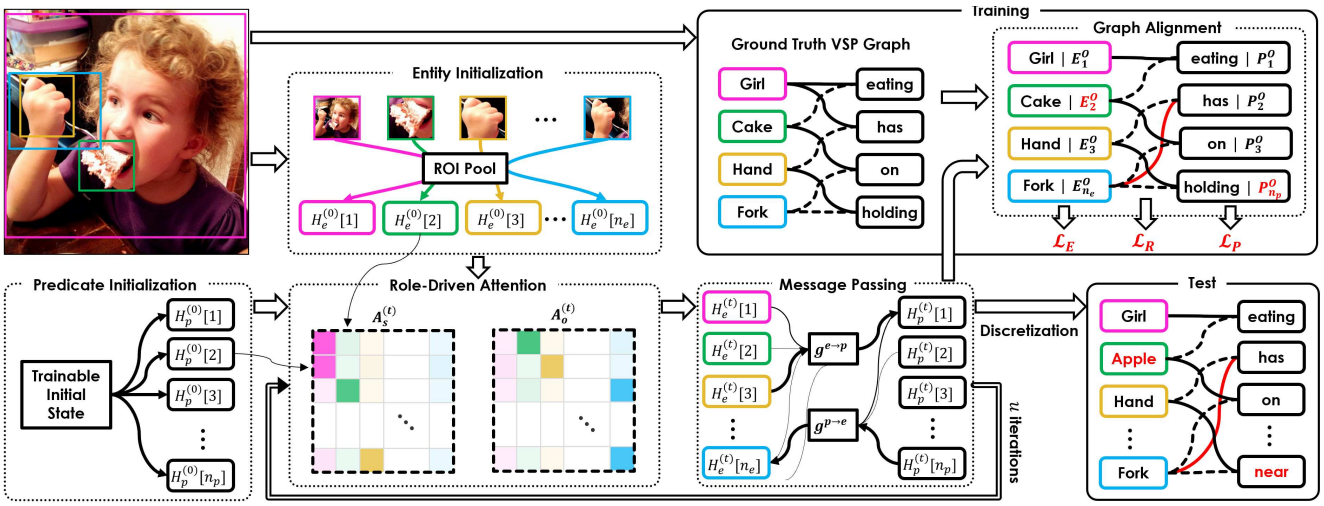

Figure 2. Overview of our proposed framework: Given an input image and object proposals, a scene graph is produced by an iterative process involving a multi-headed attention module that infers edges between entities and predicates, and a novel message passing module to propagate information between nodes and update their states. To define a classification loss for each node and edge, the ground truth graph is aligned to our output graph through a novel weakly supervised algorithm. Red represents mistake. Best viewed in color.

representing a semantic role class r in $\mathcal{C}_r$:

$$\tilde{A}_r^{(t)}[k,i] = \left\langle f_r^p\left(H_p^{(t)}[k]\right),\ f_r^e\left(H_e^{(t)}[i]\right)\right\rangle, \tag{2}$$

where $\langle .,. \rangle$ represents dot product, $H[k]$ represents the kth row of H, and f_r^p and f_r^e are trainable fully connected networks to compute the query and key vectors of the attention. We further stack $\tilde{A}_r^{(t)}$ to build the 3-dimensional tensor $\tilde{A}^{(t)}$ that represents the entire role-driven attention. In our experiments, no predicate can take more than one entity for each role, and no entity-predicate pair can have more than one semantic role. Hence, we normalize $\tilde{A}^{(t)}$ such that:

$$A_r^{(t)}[k,i] = \frac{\exp\left(\tilde{A}_r^{(t)}[k,i]\right)}{p_\varnothing + \sum_{r'=1}^{n_r}\exp\left(\tilde{A}_{r'}^{(t)}[k,i]\right)} \times \frac{\exp\left(\tilde{A}_r^{(t)}[k,i]\right)}{p_\varnothing + \sum_{i'=1}^{n_e}\exp\left(\tilde{A}_r^{(t)}[k,i']\right)}. \tag{3}$$

This can be interpreted as applying two softmax functions in paralell on $\tilde{A}^{(t)}$, once normalizing along the axis of roles, and once along the axis of entities, and then multiplying the two normalized matrices, element-wise. The constant $p_\varnothing$ is added to each denominator to allow the sum to be less than one, *e.g.* no role between an entity-predicate pair.

After computing attention matrices, we use them to propagate information from each entity to its relevant predicates and vice versa. To this end, we propose a three-stage message aggregation framework, that computes the incoming message to update each node, by aggregating outgoing messages from all other nodes, and separately processing them in the context of each semantic role. More specifically:

$$\begin{aligned} M_p^{(t)}[k] &= g^{e\to p}\left(A^{(t)}, H_e^{(t)}\right) \\ &= g^{p\leftarrow}\left(\sum_{r=1}^{n_r} g_r^e\left(\sum_{i=1}^{n_e} A_r^{(t)}[k,i]\, g^{e\to}\left(H_e^{(t)}[i]\right)\right)\right), \end{aligned} \tag{4}$$

where $g_r^{e\to}$, g_r^e, and $g_r^{p\leftarrow}$ are independent, trainable fully connected networks, respectively called *send head*, *pool head*, and *receive head*. Note that the pool head consists of n_r separate networks applied on the pooled messages for each role. Similarly, the incoming message to update each entity is computed as:

$$\begin{aligned} M_e^{(t)}[i] &= g^{p\to e}\left(A^{(t)}, H_p^{(t)}\right) \\ &= g^{e\leftarrow}\left(\sum_{r=1}^{n_r} g_r^p\left(\sum_{k=1}^{n_p} A_r^{(t)}[k,i]\, g^{p\to}\left(H_p^{(t)}[k]\right)\right)\right). \end{aligned} \tag{5}$$

After collecting messages for each node, we update their state using two Gated Recurrent Units (GRU) [6].

$$\begin{aligned} H_e^{(t+1)}[i] &= \text{GRU}_e\left(H_e^{(t)}[i], M_e^{(t)}[i]\right), \text{and} \\ H_p^{(t+1)}[k] &= \text{GRU}_p\left(H_p^{(t)}[k], M_p^{(t)}[k]\right). \end{aligned} \tag{6}$$

This process is repeated for a constant number of times u, and the final states $H_e^{(u)}$ and $H_p^{(u)}$ are passed through another pair of fully connected networks (h_e, h_p) to produce semantic embeddings E^O and P^O for entity and predicate nodes. The final state of the adjacency matrices $A_r^{(u)}$ are stacked together and named A^O.

After the message passing process, we have a continuous and fully differentiable output graph $G^O_{\text{soft}} = (E^O, P^O, A^O)$. In order to produce a valid, discrete graph as defined in Eq. (1), we apply a two-step *discretization* process. First, we convert E^O and P^O to discrete labels by picking the nearest neighbor of each of their rows among a dictionary of entity and predicate class embeddings. Next, we threshold the attention matrix A^O and suppress non-maximum roles for each entity-predicate pair. This leads to a discrete graph $G^O = (\mathcal{N}^O_e, \mathcal{N}^O_p, \mathcal{E}^O)$. In the next subsection, we define our cost function, where we also need the opposite process: converting a ground truth graph $G^T = (\mathcal{N}^T_e, \mathcal{N}^T_p, \mathcal{E}^T)$ to a soft representation $G^T_{\text{soft}} = (E^T, P^T, A^T)$. To this end, we stack the class embedding of entity and predicate nodes to get matrices E^T and P^T, and encode the edges into a binary adjacency matrix A^T.

3.3. Weakly supervised training

We train our model using pairs of image and unlocalized ground truth graph. Specifically, we need to compare the soft output graph G^O_{soft} (*i.e*. before discretization) to the target G^T_{soft} to calculate a differentiable cost to be minimized. To this end, we find an alignment (*i.e*., node correspondence) between the two graphs, and then define the overall cost as a summation of loss terms over aligned nodes and edges. Formally, we define an alignment $\mathcal{I}$ as:

$$\begin{aligned} &\mathcal{I} = (\mathcal{I}_e, \mathcal{I}_p), \text{ where} \\ &\mathcal{I}_e = \Big\{(i,j) | i \in \{1...n^O_e\}, j \in \{1...n^T_e\}\Big\}, \text{ and} \\ &\mathcal{I}_p = \Big\{(k,l) | k \in \{1...n^O_p\}, l \in \{1...n^T_p\}\Big\}, \end{aligned} \tag{7}$$

where $n^O_e = n_e$ and $n^O_p = n_p$ are the number of output entity and predicate nodes, while n^T_e and n^T_p are the number of ground truth entity and predicate nodes. $\mathcal{I}_e$ is a valid entity alignment if for any output node i there is at most one target node j, and for each j there is at most one i, where $(i,j) \in \mathcal{I}_e$. A similar constraint holds for $\mathcal{I}_p$. Moreover, $\mathcal{I}_e$ is a *maximal alignment* if all output entities **or** all target entities are aligned, whichever is fewer, *i.e*.

$$\begin{aligned} &|\mathcal{I}_e| = \min(n^O_e, n^T_e), \text{ and similarly,} \\ &|\mathcal{I}_p| = \min(n^O_p, n^T_p), \end{aligned} \tag{8}$$

where $|.|$ denotes set cardinality. Given an alignment $\mathcal{I}$ between output and target graphs, our objective function is:

$$\mathcal{L}(G^O, G^T, \mathcal{I}) = \mathcal{L}_E + \mathcal{L}_P + \lambda \mathcal{L}_R, \tag{9}$$

which is a combination of costs for entity recognition, predicate recognition, and semantic role labeling.

Our weakly supervised training framework is independent of how we define each loss term, as long as they are a summation of costs over aligned nodes. For instance, if we define the entity loss $\mathcal{L}_E$ and predicate loss $\mathcal{L}_P$ as mean square errors of entity and predicate embeddings, and if we define the role loss $\mathcal{L}_R$ to be a binary cross entropy on all attention scores, we can write:

$$\mathcal{L}_E(G^O, G^T, \mathcal{I}) = \frac{1}{|\mathcal{I}_e|} \sum_{(i,j)\in\mathcal{I}_e} \left\| E^O_i - E^T_j \right\|^2_2, \tag{10}$$

$$\mathcal{L}_P(G^O, G^T, \mathcal{I}) = \frac{1}{|\mathcal{I}_p|} \sum_{(k,l)\in\mathcal{I}_p} \left\| P^O_k - P^T_l \right\|^2_2, \tag{11}$$

$$\mathcal{L}_R(G^O, G^T, \mathcal{I}) = \frac{1}{n_r} \sum_{r=1}^{n_r} \mathcal{L}_r, \tag{12}$$

where for role r,

$$\mathcal{L}_r = \frac{1}{|\mathcal{I}|} \sum_{(i,j)\in\mathcal{I}_e} \sum_{(k,l)\in\mathcal{I}_p} \mathcal{X}\big(A^O_r[k,i], A^T_r[l,j]\big), \tag{13}$$

where $|\mathcal{I}| = |\mathcal{I}_e||\mathcal{I}_p|$, and

$$\mathcal{X}(p,q) = -q \log p - (1-q)\log(1-p). \tag{14}$$

Since $\mathcal{L}_R$ is in a different scale than $\mathcal{L}_E$ and $\mathcal{L}_P$, we use a hyperparameter λ to balance its significance in Eq. (9).

The main challenge of weakly supervised learning is that the alignment $\mathcal{I}$ is not known, and thus our training involves the following nested optimization:

$$\phi^* = \underset{\phi}{\operatorname{argmin}}\, \mathbb{E}\big[\min_{\mathcal{I}} \mathcal{L}(G^O, G^T, \mathcal{I})\big], \tag{15}$$

where ϕ is the collection of model parameters that lead to G^O, and the expectation is estimated by averaging over minibatches sampled from training data. Note that the inner optimization is subject to the constraints in Eq. (8). Inspired by the EM algorithm [25], we device an alternating optimization approach: We use the Adam Optimizer [16] for the outer optimization, and propose an iterative alignment algorithm to solve the inner optimization in the following.

There are no efficient exact algorithms for solving the inner optimization in Eq. (15). Hence, we propose an iterative algorithm to approximate the optimal alignment. We show that given an entity alignment $\mathcal{I}_e$, it is possible to find the optimal predicate alignment $\mathcal{I}_p$ in polynomial time, and similarly from $\mathcal{I}_p$ to $\mathcal{I}_e$. Accordingly, we perform those two steps iteratively in a coordinate-descent fashion, which is guaranteed to converge to a local optima.

Supposing $\mathcal{I}_e$ is given, we intend to find $\mathcal{I}_p$ that minimizes $\mathcal{L}$. Since $\mathcal{L}_E$ is constant with respect of $\mathcal{I}_p$, the problem reduces to minimizing $\mathcal{L}_P + \lambda\mathcal{L}_R$, which can be written:

$$\mathcal{L}_P + \lambda \mathcal{L}_R = \frac{1}{|\mathcal{I}_p|} \sum_{(k,l)\in\mathcal{I}_p} W^P_{kl}, \tag{16}$$

Method	Supervision	SGGEN		PHRDET	
		R@50	R@100	R@50	R@100
VtransE-MIL [50]	Weak	0.7	0.9	1.5	2.0
PPR-FCN [50]		1.5	1.9	2.4	3.2
VSPNET w/o iterative alignment	Weak	1.3	1.6	8.0	10.2
VSPNET w/ fewer alignment steps		1.8	2.0	9.9	11.9
VSPNET w/o three-stage MP		2.4	2.8	16.7	19.8
VSPNET w/o role-driven MP		2.5	2.9	15.7	18.7
VSPNET w/ fewer MP steps		2.5	2.8	15.5	18.3
VSPNET (Ours)		**3.1**	**3.5**	**17.6**	**20.4**
VtransE [50]	Full	5.5	6.0	9.5	10.4
S-PPR-FCN [50]		6.0	6.9	10.6	11.1
VSPNET (Ours)		**8.9**	**9.9**	**24.0**	**27.8**

Table 1. Results on VG preprocessed by [50]. All numbers are in percentage and baselines were borrowed from [50]

where W^P is a pairwise cost function between output and target predicate nodes, measuring not only their semantic embedding distance, but also the discrepancy of their connectivity in graph. More specifically:

$$W^P_{kl} \triangleq \left\|P^O_k - P^T_l\right\|^2_2 + \frac{\lambda}{n_r|\mathcal{I}_e|}\sum_{(i,j)\in\mathcal{I}_e}\sum_{r=1}^{n_r}\mathcal{X}\big(A^O_r[k,i], A^T_r[l,j]\big). \tag{17}$$

Note that the optimization of Eq. (16) is subject to Eq. (8), which makes $|\mathcal{I}_p|$ a constant. Hence, this problem is equivalent to maximum bipartite matching with fully connected cost function W^P, which can be solved in polynomial time using the Kuhn-Munkres algorithm [26].

Similarly, given $\mathcal{I}_p$, we can solve for $\mathcal{I}_e$, and repeat alternation. Every step leads to a lower or equal loss since either $\mathcal{L}_P + \mathcal{L}_R$ is minimized while $\mathcal{L}_E$ is fixed, or $\mathcal{L}_E + \mathcal{L}_R$ is minimized while $\mathcal{L}_P$ is fixed. Since $\mathcal{L}$ cannot become negative, these iterations must converge. We have observed that the convergence value of $\mathcal{L}$ is not sensitive to whether we start by initializing $\mathcal{I}_e$ or $\mathcal{I}_p$, nor does it depend on the initialization value. In our experiments we initialize $\mathcal{I}_p$ to an empty set and proceed with updating $\mathcal{I}_e$. We denote by v the number of iterations used for this alignment procedure.

Our method can be naturally extended to the fully supervised setting by adding a term in Eq. 10, to maximize the overlap between the aligned pairs of bounding boxes. Specifically, we redefine $\mathcal{L}_E$ as:

$$\mathcal{L}^{\text{sup}}_E(G^O, G^T, \mathcal{I}) = \frac{1}{|\mathcal{I}_e|}\sum_{(i,j)\in\mathcal{I}_e}\Big(\left\|E^O_i - E^T_j\right\|^2_2 - \lambda_B \log\big(\text{IoU}[B^O_i - B^T_j] + \epsilon\big)\Big), \tag{18}$$

where B^O and B^T are the set of output and ground truth bounding boxes respectively, and λ_B and ϵ are hyperparameters selected by cross-validation. Note that the gradient of the added term with respect to model parameters is zero, and hence this only affects alignment.

4. Experiments

We apply our framework on the Visual Genome (VG) dataset [17] for the task of scene graph generation, and compare to both weakly and fully supervised baselines. Through quantitative analysis, we show that VSPNET significantly outperforms the weakly and fully supervised state of the art, while being several times faster than existing methods. Furthermore, ablation experiments show the contribution of each proposed module, namely iterative alignment, role-driven attention, and three-stage message aggregation. We finally provide qualitative evidence that our method is able to produce VSP graphs, which are beyond the expressive capacity of conventional scene graphs.

4.1. Implementation details

We use an off-the-shelve Faster R-CNN [33] pretrained on the Open Images dataset [18] to extract object proposals that are needed as inputs to VSPNET. We extract proposal coordinates and features once for all images, and keep them fixed while training and evaluating our model. We do not stack VSPNET on top of Faster R-CNN and do not fine-tune Faster R-CNN during training. We use the original implementation of GRU [6] with 1024-dimensional states (d_e and d_p). The initialization heads e_a and e_b, the attention heads f^e_r and f^p_r, and the message passing heads, $g^{e\rightarrow}$, g^e_r, $g^{p\leftarrow}$, $g^{p\rightarrow}$, g^p_r, and $g^{e\leftarrow}$, are all fully connected networks with two 1024-dimensional layers. The embedding prediction heads h_e and h_p are each single-layer networks that map 1024-D GRU states to the 300-D embedding space. All fully connected networks use leaky ReLU activation functions [11]. Through cross-validation, we set $\lambda = 10$, $u = 3$, and $v = 3$. We use GloVe embeddings [29] to represent each class, and we fine-tune it during training.

Method	Supervision	SGGEN			SGCLS		PREDCLS	
		Time	R@50	R@100	R@50	R@100	R@50	R@100
IMP [40]	Full	1.64	3.4	4.2	21.7	24.4	44.7	53.1
MSDN [20]		3.56	7.7	10.5	19.3	21.8	63.1	66.4
MotifNet [48]		2.07	6.9	9.1	23.8	27.2	41.8	48.8
Assoc. Emb. [27]		1.19	9.7	11.3	26.5	30.0	**68.0**	**76.2**
Graph R-CNN [41]		0.83	11.4	13.7	29.6	31.6	54.2	59.1
VSPNET (Ours)		**0.11**	**12.6**	**14.2**	**31.5**	**34.1**	67.4	73.7
VSPNET (Ours)	Weak	**0.11**	**4.7**	**5.4**	**30.5**	**32.7**	**57.7**	**62.4**

Table 2. Results on VG [40]. Recall numbers (%) are from [41]. Inference time is in seconds per image, partially borrowed from [19].

The number of predicate nodes n_p is an important choice. Having more predicate nodes will increase recall but also inference time. Since SGG methods are conventionally evaluated at 100 and 50 predicates, we set $n_p = 100$. To output only 50 predicates, we rank the predicate nodes with respect to their confidence, which is defined as the product of three classification confidence scores, for subject, object and predicate. To report inference time in Table 2, we compute the average inference time per image on the test set, using identical settings for all methods (NVIDIA TITAN X, 200 proposals, VGG backbone). The time includes the extraction of proposals and their features.

4.2. Task definition

The Visual Genome dataset consists of 108,077 images with manual annotation of objects and relationships, with open-vocabulary classes. [40] and [50] preprocess the annotated objects and relationships to produce scene graphs with a fixed vocabulary. [40] keeps 150 most frequent entity and 50 most frequent predicate classes, while [50] cuts at 200 and 100 respectively. We perform two sets of experiments, based on both [40] and [50], to be able to compare to the performances reported by each paper separately. We follow their preprocessing, data splits, and evaluation protocol, but we assume bounding boxes are not available during weakly supervised training.

The main evaluation metric dubbed SGGEN, measures the accuracy of `subject-predicate-object` triplets. A detected triplet is considered correct if the predicted class for subject, object, and predicate are all correct, and the subject and object bounding boxes have an Intersection over Union (IoU) of at least 0.5 with ground truth. To evaluate, the top K triplets predicted by the model are matched to ground truth triplets. The number of correctly matched triplets is divided by the total number of triplets in the ground truth to compute recall at K. This value is averaged over all images leading to R@50 and R@100. Since SGGEN is highly affected by the quality of object proposals. we also report SGCLS, which assumes ground truth bounding boxes are given at test time, instead of proposals. Another metric, PREDCLS assumes ground truth bounding are given, and true object classes are given too. [50] also evaluates using PHRDET, which stands for Phrase Detection. This metric is similar to SGGEN, with the difference that instead of evaluating the bounding box of subject and object separately, the goal is to predict a union bounding box enclosing both the object and subject. To this end, for each detected triplet, we get the union box of its subject and object, and match with that of ground truth triplets at IoU ≥ 0.5.

4.3. Results

Table 1 shows our quantitative results on VG compared to VtransE [49] and PPR-FCN [50], in both Weakly Supervised (WS) and Fully Supervised (FS) settings, following the evaluation settings of [50]. Our VSPNET achieves the best WS performance, with SGGEN performance more than two times higher and PHRDET more than six times higher than the state of the art. Moreover, the FS extension of our method outperforms the FS variants of those baselines significantly. On the PHRDET measure, even our WS method outperforms all FS baselines. Furthermore, we provide ablative variants of our method as extra rows in Table 1, to study the effect of each proposed component in isolation.

In VSPNET **w/o iterative alignment**, we replace the proposed alignment algorithm with a heuristic baseline, where we align entities by minimizing $\mathcal{L}_E$ and independently align predicates to minimize $\mathcal{L}_P$, in a one-step process. Our alignment algorithm leads to more than twice the performance of this ablation. We make a similar observation by reducing the number of alignment steps v from 3 to 1, denoted as VSPNET **w/ fewer alignment steps**. Furthermore, in VSPNET **w/o three-stage MP**, we replace the proposed three-stage message aggregation framework with a conventional average pooling, that computes the sum of all messages after multiplying by the attention weights. In VSPNET **w/o role-driven MP**, we keep the three-stage message aggregation, but remove the role-driven attention, and replace $A_r(t)$ with a constant, uniformly distributed attention. Finally, in VSPNET **w/ fewer MP steps**, we only reduce the number of MP steps, u, from 3 to 1. All these three ablations lead to inferior performance, proving the ef-

fectiveness of our proposed message passing framework.

To compare to more recent methods, we also perform experiments on the original version of VG that was used by [40], and follow the evaluation protocol of [41]. Table 2 compares VSPNET to all the numbers reported by [41]. The FS version of our method outperforms all state-of-the-art methods in all metrics, except slightly outperformed by Assoc. Emb. [27] in PREDCLS only. In addition to superior accuracy, our method is several times faster than all methods. It is also 5 times faster than Factorizable Net [19], which is the fastest SGG method (0.55 seconds per image), although not shown in Table 2, because their reported recall is computed differently than ours.

Furthermore, our WS method shows competitive performance and even outperforms some FS methods. Although there is a performance drop from FS to WS, that is mainly due to the difficulty of object localization in the WS setting. In SGCLS, it achieves a performance very close to FS VSPNET, and outperforms all other FS baselines. This suggests that if some day we have access to very accurate proposals, our WS model would perform as accurately as FS methods. Note that although SGCLS provides ground truth bounding boxes, the WS model only treats them as input proposals, and is still trained with unlocalized ground truth and unknown alignment. Also note that all baselines in Table 2 train their Faster R-CNN on VG directly, using annotated bounding boxes that we assume not available in WS settings. Hence, we use an off-the-shelve Faster R-CNN that is pretrained on another dataset in all our experiments. This makes the comparison in Table 2 somewhat unfair, to our disadvantage. Adopting the backbone used by the baselines would improve our results, but violates WS constraints.

To illustrate the expressive power of our novel VSP formulation, we train our model on the V-COCO dataset [10], which annotates human actions in images, as well as objects and instruments of those actions. While this dataset has been primarily used for HOI in the literature [32, 38], we adopt it for VSP, by aggregating all action annotations of each image into a single semantic graph, and connecting them to the related objects through 3 types of semantic role: subject, object, and instrument. The resulting VSP graphs have unique properties that are not seen in scene graphs, as shown in Figure 3, such as verbs with more than two entities (*e.g.* `person cutting cake` *with* `knife`), and verbs with only one entity (*e.g.* `person smiling`). After training our model on the training set of V-COCO, we apply it on the test set and visualize output graphs in Figure 3. Our method successfully generates VSP graphs containing interactions that are not possible with any SGG method.

5. Conclusion

We proposed a method to parse an image into a semantic graph that includes entities, predicates, and semantic roles.

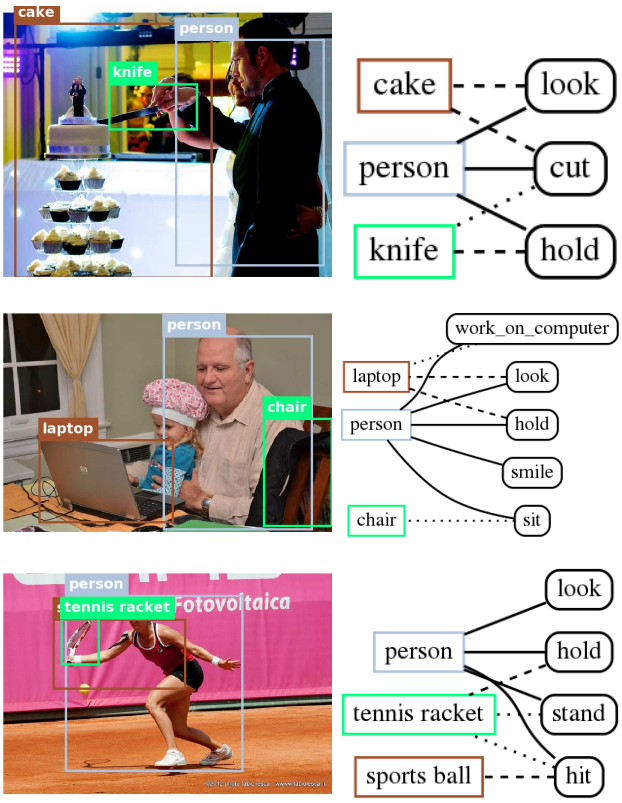

Figure 3. Example VSP graphs generated by our method. Solid, dashed, and dotted lines represent subject, object, and instrument.

Unlike prior works, our method does not require bounding box annotations for training, and does not rely on exhaustive processing of all object proposal pairs. Moreover, it is able to extract more flexible graphs where any number of entities are involved in each predicate. To this end, we proposed a generalized formulation of Scene Graph Generation (SGG) that disentangles predicates from entities, and enables sub-quadratic performance. Based on that, we proposed VSPNET, based on a dynamic, attention-based, bipartite message passing framework. We also introduced the first graph-based weakly supervised learning framework based on a novel graph alignment algorithm. We compared our method to the state of the art through extensive experiments, and achieved significant performance improvements in both weakly supervised and fully supervised settings, while several times faster than every existing method.

Acknowledgements: This work was supported by the U.S. DARPA AIDA Program No. FA8750-18-2-0014. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

References

[1] Dzmitry Bahdanau, Jan Chorowski, Dmitriy Serdyuk, Philemon Brakel, and Yoshua Bengio. End-to-end attention-based large vocabulary speech recognition. In *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 4945–4949. IEEE, 2016.

[2] Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. Abstract meaning representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186, 2013.

[3] Hakan Bilen and Andrea Vedaldi. Weakly supervised deep detection networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2846–2854, 2016.

[4] Yu-Wei Chao, Yunfan Liu, Xieyang Liu, Huayi Zeng, and Jia Deng. Learning to detect human-object interactions. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 381–389. IEEE, 2018.

[5] Xinlei Chen, Li-Jia Li, Li Fei-Fei, and Abhinav Gupta. Iterative visual reasoning beyond convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7239–7248, 2018.

[6] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

[7] Bo Dai, Yuqi Zhang, and Dahua Lin. Detecting visual relationships with deep relational networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3076–3086, 2017.

[8] Thomas G Dietterich, Richard H Lathrop, and Tomás Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artificial intelligence*, 89(1-2):31–71, 1997.

[9] Georgia Gkioxari, Ross Girshick, Piotr Dollár, and Kaiming He. Detecting and recognizing human-object interactions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8359–8367, 2018.

[10] Saurabh Gupta and Jitendra Malik. Visual semantic role labeling. *arXiv preprint arXiv:1505.04474*, 2015.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.

[12] Drew Hudson and Christopher D Manning. Learning by abstraction: The neural state machine. In *Advances in Neural Information Processing Systems*, pages 5901–5914, 2019.

[13] Seong Jae Hwang, Sathya N Ravi, Zirui Tao, Hyunwoo J Kim, Maxwell D Collins, and Vikas Singh. Tensorize, factorize and regularize: Robust visual relationship learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1014–1023, 2018.

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

[15] Keizo Kato, Yin Li, and Abhinav Gupta. Compositional learning for human object interaction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 234–251, 2018.

[16] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[17] 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.

[18] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Tom Duerig, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *arXiv preprint arXiv:1811.00982*, 2018.

[19] Yikang Li, Wanli Ouyang, Bolei Zhou, Jianping Shi, Chao Zhang, and Xiaogang Wang. Factorizable net: an efficient subgraph-based framework for scene graph generation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 335–351, 2018.

[20] Yikang Li, Wanli Ouyang, Bolei Zhou, Kun Wang, and Xiaogang Wang. Scene graph generation from objects, phrases and region captions. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1261–1270, 2017.

[21] Xiaodan Liang, Lisa Lee, and Eric P Xing. Deep variation-structured reinforcement learning for visual relationship and attribute detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 848–857, 2017.

[22] Yong Liu, Ruiping Wang, Shiguang Shan, and Xilin Chen. Structure inference net: object detection using scene-level context and instance-level relationships. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6985–6994, 2018.

[23] Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. Visual relationship detection with language priors. In *European Conference on Computer Vision*, pages 852–869. Springer, 2016.

[24] Arun Mallya and Svetlana Lazebnik. Recurrent models for situation recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 455–463, 2017.

[25] Todd K Moon. The expectation-maximization algorithm. *IEEE Signal processing magazine*, 13(6):47–60, 1996.

[26] James Munkres. Algorithms for the assignment and transportation problems. *Journal of the society for industrial and applied mathematics*, 5(1):32–38, 1957.

[27] Alejandro Newell and Jia Deng. Pixels to graphs by associative embedding. In *Advances in neural information processing systems*, pages 2171–2180, 2017.

[28] Martha Palmer, Daniel Gildea, and Nianwen Xue. Semantic role labeling. *Synthesis Lectures on Human Language Technologies*, 3(1):1–103, 2010.

[29] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[30] Julia Peyre, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Weakly-supervised learning of visual relations. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5179–5188, 2017.

[31] Bryan A Plummer, Arun Mallya, Christopher M Cervantes, Julia Hockenmaier, and Svetlana Lazebnik. Phrase localization and visual relationship detection with comprehensive image-language cues. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1928–1937, 2017.

[32] Siyuan Qi, Wenguan Wang, Baoxiong Jia, Jianbing Shen, and Song-Chun Zhu. Learning human-object interactions by graph parsing neural networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 401–417, 2018.

[33] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[34] Arpit Sharma, Nguyen H Vo, Somak Aditya, and Chitta Baral. Towards addressing the winograd schema challenge—building and using a semantic parser and a knowledge hunting module. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015.

[35] Jiaxin Shi, Hanwang Zhang, and Juanzi Li. Explainable and explicit visual reasoning over scene graphs. *arXiv preprint arXiv:1812.01855*, 2018.

[36] Zheng Shou, Hang Gao, Lei Zhang, Kazuyuki Miyazawa, and Shih-Fu Chang. Autoloc: Weakly-supervised temporal action localization in untrimmed videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 154–171, 2018.

[37] David Wadden, Ulme Wennberg, Yi Luan, and Hannaneh Hajishirzi. Entity, relation, and event extraction with contextualized span representations. *arXiv preprint arXiv:1909.03546*, 2019.

[38] Bo Wan, Desen Zhou, Yongfei Liu, Rongjie Li, and Xuming He. Pose-aware multi-level feature network for human object interaction detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9469–9478, 2019.

[39] Sanghyun Woo, Dahun Kim, Donghyeon Cho, and In So Kweon. Linknet: Relational embedding for scene graph. In *Advances in Neural Information Processing Systems*, pages 558–568, 2018.

[40] Danfei Xu, Yuke Zhu, Christopher B Choy, and Li Fei-Fei. Scene graph generation by iterative message passing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5410–5419, 2017.

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

[42] Xu Yang, Kaihua Tang, Hanwang Zhang, and Jianfei Cai. Auto-encoding scene graphs for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10685–10694, 2019.

[43] Mark Yatskar, Vicente Ordonez, Luke Zettlemoyer, and Ali Farhadi. Commonly uncommon: Semantic sparsity in situation recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7196–7205, 2017.

[44] Mark Yatskar, Luke Zettlemoyer, and Ali Farhadi. Situation recognition: Visual semantic role labeling for image understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5534–5542, 2016.

[45] Ruichi Yu, Ang Li, Vlad I Morariu, and Larry S Davis. Visual relationship detection with internal and external linguistic knowledge distillation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1974–1982, 2017.

[46] Yuan Yuan, Xiaodan Liang, Xiaolong Wang, Dit-Yan Yeung, and Abhinav Gupta. Temporal dynamic graph lstm for action-driven video object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1801–1810, 2017.

[47] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. *arXiv preprint arXiv:1811.10830*, 2018.

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

[49] Hanwang Zhang, Zawlin Kyaw, Shih-Fu Chang, and Tat-Seng Chua. Visual translation embedding network for visual relation detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5532–5540, 2017.

[50] Hanwang Zhang, Zawlin Kyaw, Jinyang Yu, and Shih-Fu Chang. Ppr-fcn: weakly supervised visual relation detection via parallel pairwise r-fcn. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4233–4241, 2017.

[51] Bohan Zhuang, Lingqiao Liu, Chunhua Shen, and Ian Reid. Towards context-aware interaction recognition for visual relationship detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 589–598, 2017.