Neural Motifs: Scene Graph Parsing with Global Context

Rowan Zellers[1] Mark Yatskar[1,2] Sam Thomson[3] Yejin Choi[1,2]
[1]Paul G. Allen School of Computer Science & Engineering, University of Washington
[2]Allen Institute for Artificial Intelligence
[3]School of Computer Science, Carnegie Mellon University
{rowanz, my89, yejin}@cs.washington.edu, sthomson@cs.cmu.edu
https://rowanzellers.com/neuralmotifs

Abstract

We investigate the problem of producing structured graph representations of visual scenes. Our work analyzes the role of motifs: regularly appearing substructures in scene graphs. We present new quantitative insights on such repeated structures in the Visual Genome dataset. Our analysis shows that object labels are highly predictive of relation labels but not vice-versa. We also find that there are recurring patterns even in larger subgraphs: more than 50% of graphs contain motifs involving at least two relations. Our analysis motivates a new baseline: given object detections, predict the most frequent relation between object pairs with the given labels, as seen in the training set. This baseline improves on the previous state-of-the-art by an average of 3.6% relative improvement across evaluation settings. We then introduce Stacked Motif Networks, a new architecture designed to capture higher order motifs in scene graphs that further improves over our strong baseline by an average 7.1% relative gain. Our code is available at github.com/rowanz/neural-motifs.

1. Introduction

We investigate scene graph parsing: the task of producing graph representations of real-world images that provide semantic summaries of objects and their relationships. For example, the graph in Figure 1 encodes the existence of key objects such as people ("man" and "woman"), their possessions ("helmet" and "motorcycle", both possessed by the woman), and their activities (the woman is "riding" the "motorcycle"). Predicting such graph representations has been shown to improve natural language based image tasks [17, 43, 51] and has the potential to significantly expand the scope of applications for computer vision systems. Compared to object detection [36, 34] , object interactions [48, 3] and activity recognition [13], scene graph parsing poses unique challenges since it requires reasoning about the complex dependencies across all of these components.

Elements of visual scenes have strong structural regularities. For instance, people tend to wear clothes, as can be seen in Figure 1. We examine these structural repetitions, or *motifs*, using the Visual Genome [22] dataset, which provides annotated scene graphs for 100k images from COCO [28], consisting of over 1M instances of objects and 600k relations. Our analysis leads to two key findings. First, there are strong regularities in the local graph structure such that the distribution of the relations is highly skewed once the corresponding object categories are given, but not vice versa. Second, structural patterns exist even in larger subgraphs; we find that over half of images contain previously occurring graph motifs.

Based on our analysis, we introduce a simple yet powerful baseline: given object detections, predict the most frequent relation between object pairs with the given labels, as seen in the training set. The baseline improves over prior state-of-the-art by 1.4 mean recall points (3.6% relative), suggesting that an effective scene graph model must capture both the asymmetric dependence between objects and

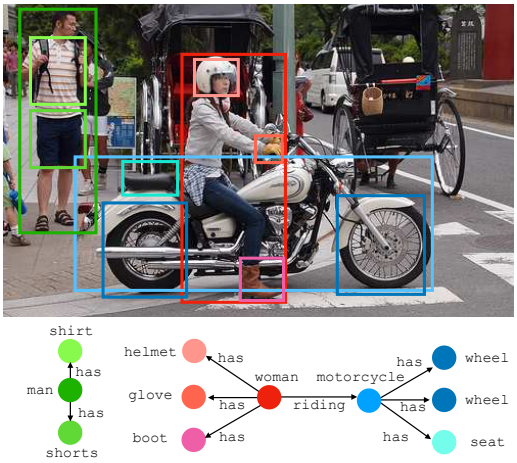

Figure 1. A ground truth scene graph containing entities, such as woman, bike or helmet, that are localized in the image with bounding boxes, color coded above, and the relationships between those entities, such as riding, the relation between woman and motorcycle or has the relation between man and shirt.

their relations, along with larger contextual patterns.

We introduce the *Stacked Motif Network (*MOTIFNET*)*, a new neural network architecture that complements existing approaches to scene graph parsing. We posit that the key challenge in modeling scene graphs lies in devising an efficient mechanism to encode the global context that can directly inform the local predictors (i.e., objects and relations). While previous work has used graph-based inference to propagate information in both directions between objects and relations [47, 25, 24], our analysis suggests strong independence assumptions in local predictors limit the quality of global predictions. Instead, our model predicts graph elements by staging bounding box predictions, object classifications, and relationships such that the global context encoding of all previous stages establishes rich context for predicting subsequent stages, as illustrated in Figure 5. We represent the global context via recurrent sequential architectures such as Long Short-term Memory Networks (LSTMs) [15].

Our model builds on Faster-RCNN [36] for predicting bounding regions, fine tuned and adapted for Visual Genome. Global context across bounding regions is computed and propagated through bidirectional LSTMs, which is then used by another LSTM that labels each bounding region conditioned on the overall context and all previous labels. Another specialized layer of bidirectional LSTMs then computes and propagates information for predicting edges given bounding regions, their labels, and all other computed context. Finally, we classify all n^2 edges in the graph, combining globally contextualized representations of head, tail, and image representations using using low-rank outer products [19]. The method can be trained end-to-end.

Experiments on Visual Genome demonstrate the efficacy of our approach. First, we update existing work by pretraining the detector on Visual Genome, setting a new state-of-the-art (improving on average across evaluation settings 14.0 absolute points). Our new simple baseline improves over previous work, using our updated detector, by a mean improvement of 1.4 points. Finally, experiments show Stacked Motif Networks is effective at modeling global context, with a mean improvement of 2.9 points (7.1% relative improvement) over our new strong baseline.

2. Formal definition

A *scene graph*, G, is a structured representation of the semantic content of an image [17]. It consists of:

- a set $B = \{b_1, \ldots, b_n\}$ of *bounding boxes*, $b_i \in \mathbb{R}^4$,
- a corresponding set $O = \{o_1, \ldots, o_n\}$ of *objects*, assigning a class label $o_i \in \mathcal{C}$ to each b_i, and
- a set $R = \{r_1, \ldots, r_m\}$ of binary relationships between those objects.

Each relationship $r_k \in \mathcal{R}$ is a triplet of a start node $(b_i, o_i) \in B \times O$, an end node $(b_j, o_j) \in B \times O$, and a relationship label $x_{i \to j} \in \mathcal{R}$, where $\mathcal{R}$ is the set of all predicate types, including the "background" predicate, BG, which indicates that there is no edge between the specified objects. See Figure 1 for an example scene graph.

Type	Examples	Classes	Instances
	Entities		
Part	arm, tail, wheel	32	200k (25.2%)
Artifact	basket, fork, towel	34	126k (16.0%)
Person	boy, kid, woman	13	113k (14.3%)
Clothes	cap, jean, sneaker	16	91k (11.5%)
Vehicle	airplane, bike, truck,	12	44k (5.6%)
Flora	flower, plant, tree	3	44k (5.5%)
Location	beach, room, sidewalk	11	39k (4.9%)
Furniture	bed, desk, table	9	37k (4.7%)
Animal	bear, giraffe, zebra	11	30k (3.8%)
Structure	fence, post, sign	3	30k (3.8%)
Building	building, house	2	24k (3.1%)
Food	banana, orange, pizza	6	13k (1.6%)
	Relations		
Geometric	above, behind, under	15	228k (50.0%)
Possessive	has, part of, wearing	8	186k (40.9%)
Semantic	carrying, eating, using	24	39k (8.7%)
Misc	for, from, made of	3	2k (0.3%)

Table 1. Object and relation types in Visual Genome, organized by super-type. Most, 25.2% of entities are parts and 90.9% of relations are geometric or possessive.

3. Scene graph analysis

In this section, we seek quantitative insights on the structural regularities of scene graphs. In particular, (a) how different types of relations correlate with different objects, and (b) how higher order graph structures recur over different scenes. These insights motivate both the new baselines we introduce in this work and our model that better integrates the global context, described in Section 4.

3.1. Prevalent Relations in Visual Genome

To gain insight into the Visual Genome scene graphs, we first categorize objects and relations into high-level types. As shown in Table 1, the predominant relations are *geometric* and *possessive*, with clothing and parts making up over one third of entity instances. Such relations are often obvious, e.g., houses tend to have windows. In contrast, *semantic* relations, which correspond to activities, are less frequent and less obvious. Although nearly half of relation types are semantic in nature, they comprise only 8.7% of relation instances. The relations "using" and "holding" account for 32.2% of all semantic relation instances.

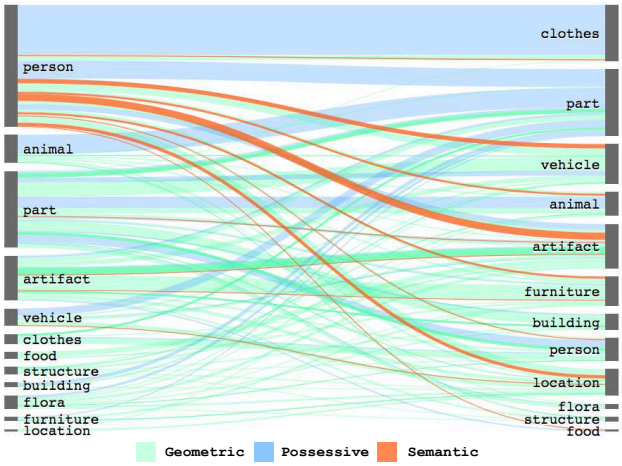

Figure 2. Types of edges between high-level categories in Visual Genome. Geometric, possessive and semantic edges cover 50.9%, 40.9%, and 8.7%, respectively, of edge instances in scene graphs. The majority of semantic edges occur between people and vehicles, artifacts and locations. Less than 2% of edges between clothes and people are semantic.

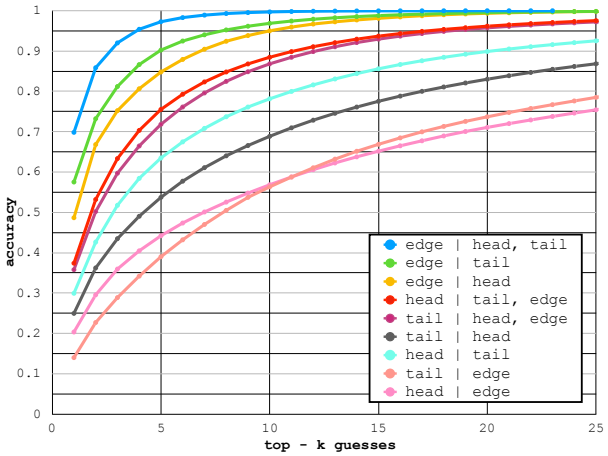

Figure 3. The likelihood of guessing, in the top-k, head, tail, or edge labels in a scene graph, given other graph components (i.e. without image features). Neither head nor tail labels are strongly determined by other labels, but given the identity of head and tail, edges (`edge | head, tail`) can be determined with 97% accuracy in under 5 guesses. Such strong biases make it critical to condition on objects when predicting edges.

Using our high-level types, we visualize the distribution of relation types between object types in Figure 2. Clothing and part entities are almost exclusively linked through possessive relations while furniture and building entities are almost exclusively linked through geometric relations. Geometric and spatial relationships between certain entities are interchangeable, for example, when a "part" is the head object, it tends to connect to other entities through a geometric relation (e.g. wheel on bike); when a "part" is the tail object, it tends to be connected with possessive relations (e.g. bike has wheel). Nearly all semantic relationship are headed by people, with the majority of edges relating to artifacts, vehicles, and locations. Such structural predictability and the prevalence of geometric and part-object relations suggest that common sense priors play an important role in generating accurate scene graphs.

In Figure 3, we examine how much information is gained by knowing the identity of different parts in a scene graphs. In particular, we consider how many guesses are required to determine the labels of head (h), edge (e) or tail (t) given labels of the other elements, only using label statistics computed on scene graphs. Higher curves imply that the element is highly determined given the other values. The graph shows that the local distribution of relationships has significant structure. In general, the identity of edges involved in a relationship is not highly informative of other elements of the structure while the identities of head or tail provide significant information, both to each other and to edge labels. Adding edge information to already given head or tail information provides minimal gain. Finally, the graph shows edge labels are highly determined given the identity of object pairs: the most frequent relation is correct 70% of the time, and the five most frequent relations for the pair contain the correct label 97% of the time.

3.2. Larger Motifs

Scene graphs not only have local structure but have higher order structure as well. We conducted an analysis of repeated motifs in scene graphs by mining combinations of object-relation-object labels that have high pointwise mutual information with each other. Motifs were extracted iteratively: first we extracted motifs of two combinations, replaced all instances of that motif with an atomic symbol and mined new motifs given previously identified motifs. Combinations of graph elements were selected as motifs if both elements involved occurred at least 50 times in the Visual Genome training set and were at least 10 times more likely to occur together than apart. Motifs were mined until no new motifs were extracted. Figure 4 contains example motifs we extracted on the right, and the prevalence of motifs of different lengths in images on the left. Many motifs correspond to either combinations of parts, or objects that are commonly grouped together. Over 50% of images in Visual Genome contain a motif involving at least two combinations of object-relation-object, and some images contain motifs involving as many as 16 elements.

4. Model

Here we present our novel model, *Stacked Motif Network* (MOTIFNET). MOTIFNET decomposes the probability of a graph G (made up of a set of bounding regions B,

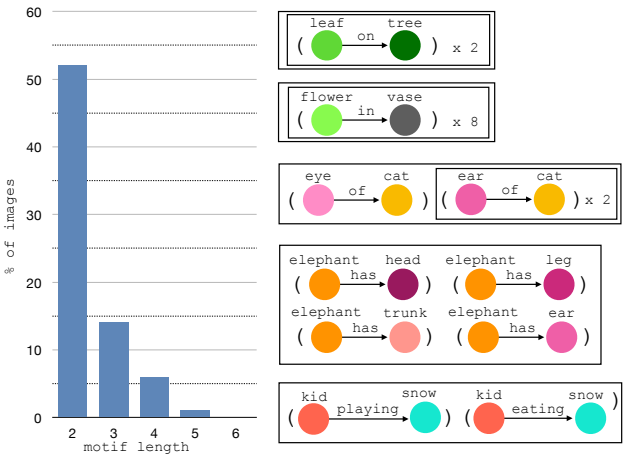

Figure 4. On the left, the percent of images that have a graph motif found in Visual Genome using pointwise mutual information, composed of at least a certain length (the number of edges it contains). Over 50% of images have at least one motif involving two relationships. On the right, example motifs, where structures repeating many times is indicated with plate notation. For example, the second motif is length 8 and consists of 8 flower-in-vase relationships. Graph motifs commonly result from groups (e.g., several instances of "leaf on tree"), and correlation between parts (e.g., "elephant has head," "leg," "trunk," and "ear.").

object labels O, and labeled relations R) into three factors:

$$\Pr(G \mid I) = \Pr(B \mid I)\, \Pr(O \mid B, I)\, \Pr(R \mid B, O, I). \quad (1)$$

Note that this factorization makes no independence assumptions. Importantly, predicted object labels may depend on one another, and predicted relation labels may depend on predicted object labels. The analyses in Section 3 make it clear that capturing these dependencies is crucial.

The *bounding box* model ($\Pr(B \mid I)$) is a fairly standard object detection model, which we describe in Section 4.1. The *object* model ($\Pr(O \mid B, I)$; Section 4.2) conditions on a potentially large set of predicted bounding boxes, B. To do so, we linearize B into a sequence that an LSTM then processes to create a contextualized representation of each box. Likewise, when modeling *relations* ($\Pr(R \mid B, O, I)$; Section 4.3), we linearize the set of predicted labeled objects, O, and process them with another LSTM to create a representation of each object in context. Figure 5 contains a visual summary of the entire model architecture.

4.1. Bounding Boxes

We use Faster R-CNN as an underlying detector [36]. For each image I, the detector predicts a set of region proposals $B = \{b_1, \ldots, b_n\}$. For each proposal $b_i \in B$ it also outputs a feature vector $\mathbf{f}_i$ and a vector $\mathbf{l}_i \in \mathbb{R}^{|\mathcal{C}|}$ of (non-contextualized) object label probabilities. Note that because BG is a possible label, our model has not yet committed to any bounding boxes. See Section 5.1 for details.

4.2. Objects

Context We construct a contextualized representation for object prediction based on the set of proposal regions B. Elements of B are first organized into a linear sequence, $[(b_1, \mathbf{f}_1, \mathbf{l}_1), \ldots, (b_n, \mathbf{f}_n, \mathbf{l}_n)]$.[1] The *object context*, $\mathbf{C}$, is then computed using a bidirectional LSTM [15]:

$$\mathbf{C} = \text{biLSTM}([\mathbf{f}_i; \mathbf{W}_1 \mathbf{l}_i]_{i=1,\ldots,n}), \quad (2)$$

$\mathbf{C} = [\mathbf{c}_1, \ldots, \mathbf{c}_n]$ contains the final LSTM layer's hidden states for each element in the linearization of B, and $\mathbf{W}_1$ is a parameter matrix that maps the distribution of predicted classes, $\mathbf{l}_1$, to $\mathbb{R}^{100}$. The biLSTM allows all elements of B to contribute information about potential object identities.

Decoding The context $\mathbf{C}$ is used to sequentially decode labels for each proposal bounding region, conditioning on previously decoded labels. We use an LSTM to decode a category label for each contextualized representation in $\mathbf{C}$:

$$\mathbf{h}_i = \text{LSTM}_i\left([\mathbf{c}_i; \hat{\mathbf{o}}_{i-1}]\right) \quad (3)$$

$$\hat{\mathbf{o}}_i = \text{argmax}\left(\mathbf{W}_o\, \mathbf{h}_i\right) \in \mathbb{R}^{|\mathcal{C}|} \text{ (one-hot)} \quad (4)$$

We then discard the hidden states $\mathbf{h}_i$ and use the object class commitments $\hat{\mathbf{o}}_i$ in the relation model (Section 4.3).

4.3. Relations

Context We construct a contextualized representation of bounding regions, B, and objects, O, using additional bidirectional LSTM layers:

$$\mathbf{D} = \text{biLSTM}([\mathbf{c}_i; \mathbf{W}_2 \hat{\mathbf{o}}_i]_{i=1,\ldots,n}), \quad (5)$$

where the *edge context* $\mathbf{D} = [\mathbf{d}_1, \ldots, \mathbf{d}_n]$ contains the states for each bounding region at the final layer, and $\mathbf{W}_2$ is a parameter matrix mapping $\hat{\mathbf{o}}_i$ into $\mathbb{R}^{100}$.

Decoding There are a quadratic number of possible relations in a scene graph. For each possible edge, say between b_i and b_j, we compute the probability the edge will have label $x_{i \to j}$ (including BG). The distribution uses global context, $\mathbf{D}$, and a feature vector for the union of boxes [2], $\mathbf{f}_{i,j}$:

$$\mathbf{g}_{i,j} = (\mathbf{W}_h \mathbf{d}_i) \circ (\mathbf{W}_t \mathbf{d}_j) \circ \mathbf{f}_{i,j} \quad (6)$$

$$\Pr(x_{i \to j} \mid B, O) = \text{softmax}\left(\mathbf{W}_r \mathbf{g}_{i,j} + \mathbf{w}_{o_i, o_j}\right). \quad (7)$$

$\mathbf{W}_h$ and $\mathbf{W}_t$ project the head and tail context into $\mathbb{R}^{4096}$. $\mathbf{w}_{o_i, o_j}$ is a bias vector specific to the head and tail labels.

[1] We consider several strategies to order the regions, see Section 5.1.

[2] A union box is the convex hull of the union of two bounding boxes.

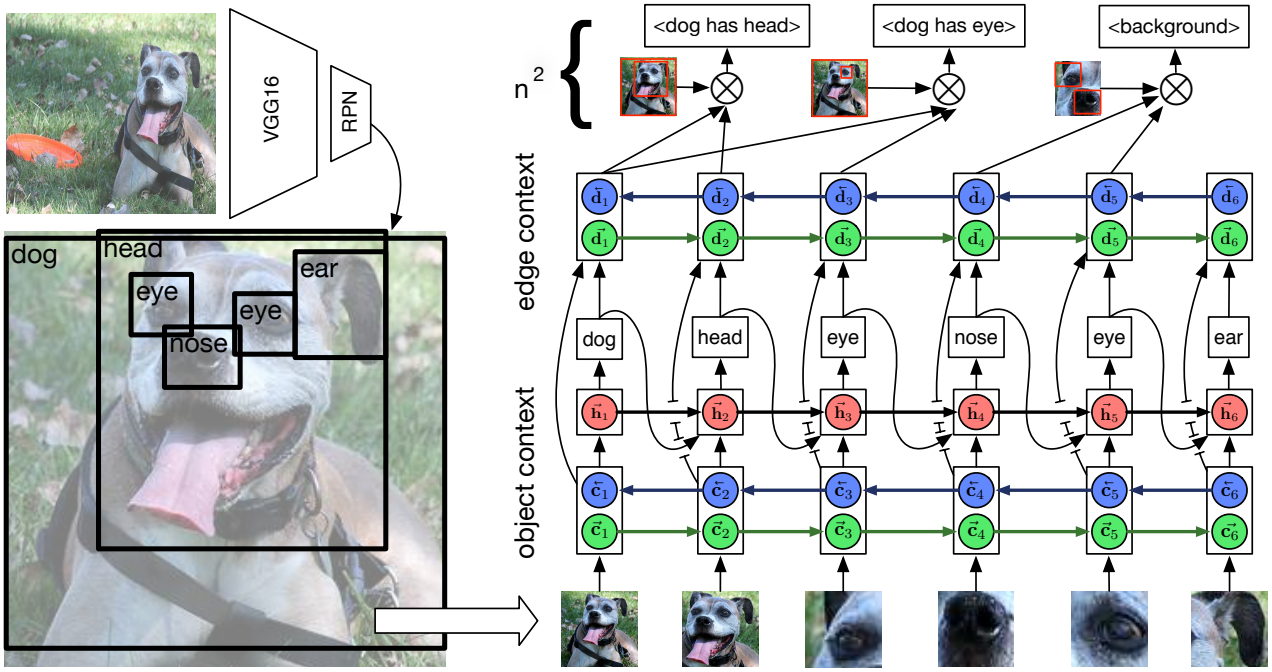

Figure 5. A diagram of a Stacked Motif Network (MOTIFNET). The model breaks scene graph parsing into stages predicting bounding regions, labels for regions, and then relationships. Between each stage, global context is computed using bidirectional LSTMs and is then used for subsequent stages. In the first stage, a detector proposes bounding regions and then contextual information among bounding regions is computed and propagated (object context). The global context is used to predict labels for bounding boxes. Given bounding boxes and labels, the model constructs a new representation (edge context) that gives global context for edge predictions. Finally, edges are assigned labels by combining contextualized head, tail, and union bounding region information with an outer product.

5. Experimental Setup

In the following sections we explain (1) details of how we construct the detector, order bounding regions, and implement the final edge classifier (Section 5.1), (2) details of training (Section 5.2), and (3) evaluation (Section 5.3).

5.1. Model Details

Detectors Similar to prior work in scene graph parsing [47, 25], we use Faster RCNN with a VGG backbone as our underling object detector [36, 40]. Our detector is given images that are scaled and then zero-padded to be 592x592. We adjust the bounding box proposal scales and dimension ratios to account for different box shapes in Visual Genome, similar to YOLO-9000 [34]. To control for detector performance in evaluating different scene graph models, we first pretrain the detector on Visual Genome objects. We optimize the detector using SGD with momentum on 3 Titan Xs, with a batch size of $b = 18$, and a learning rate of $lr = 1.8 \cdot 10^{-2}$ that is divided by 10 after validation mAP plateaus. For each batch we sample 256 RoIs per image, of which 75% are background. The detector gets 20.0 mAP (at 50% IoU) on Visual Genome; the same model, but trained and evaluated on COCO, gets 47.7 mAP at 50% IoU. Following [47], we integrate the use the detector freezing the convolution layers and duplicating the fully connected layers, resulting in separate branches for object/edge features.

Alternating Highway LSTMs To mitigate vanishing gradient problems as information flows upward, we add highway connections to all LSTMs [14, 41, 58]. To additionally reduce the number of parameters, we follow [14] and alternate the LSTM directions. Each alternating highway LSTM step can be written as the following wrapper around the conventional LSTM equations [15]:

$$\mathbf{r}_i = \sigma(\mathbf{W}_g[\mathbf{h}_{i-\delta}, \mathbf{x}_i] + \mathbf{b}_g) \quad (8)$$

$$\mathbf{h}_i = \mathbf{r_i} \circ \text{LSTM}(\mathbf{x}_i, \mathbf{h}_{i-\delta}) + (1 - \mathbf{r_i}) \circ \mathbf{W}_i \mathbf{x}_i, \quad (9)$$

where $\mathbf{x}_i$ is the input, $\mathbf{h}_i$ represents the hidden state, and δ is the direction: $\delta = 1$ if the current layer is even, and -1 otherwise. For MOTIFNET, we use 2 alternating highway LSTM layers for object context, and 4 for edge context.

RoI Ordering for LSTMs We consider several ways of ordering the bounding regions:

(1) LEFTRIGHT (default): Our default option is to sort the regions left-to-right by the central x-coordinate: we expect this to encourage the model to predict edges between nearby objects, which is beneficial as objects appearing in relationships tend to be close together.

(2) CONFIDENCE: Another option is to order bounding regions based on the confidence of the maximum non-background prediction from the detector: $\max_{j \neq \text{BG}} \mathbf{l}_i^{(j)}$, as this lets the detector commit to "easy" regions, obtaining context for more difficult regions.[3]

(3) SIZE: Here, we sort the boxes in descending order by size, possibly predicting global scene information first.

(4) RANDOM: Here, we randomly order the regions.

Predicate Visual Features To extract visual features for a predicate between boxes b_i, b_j, we resize the detector's features corresponding to the union box of b_i, b_j to 7x7x256. We model geometric relations using a 14x14x2 binary input with one channel per box. We apply two convolution layers to this and add the resulting 7x7x256 representation to the detector features. Last, we apply finetuned VGG fully connected layers to obtain a 4096 dimensional representation.[4]

5.2. Training

We train MOTIFNET on ground truth boxes, with the objective to predict object labels and to predict edge labels given ground truth object labels. For an image, we include all annotated relationships (sampling if more than 64) and sample 3 negative relationships per positive. In cases with multiple edge labels per directed edge (5% of edges), we sample the predicates. Our loss is the sum of the cross entropy for predicates and cross entropy for objects predicted by the object context layer. We optimize using SGD with momentum on a single GPU, with $lr = 6 \cdot 10^{-3}$ and $b = 6$.

Adapting to Detection Using the above protocol gets good results when evaluated on scene graph classification, but models that incorporate context underperform when suddenly introduced to non-gold proposal boxes at test time.

To alleviate this, we fine-tune using noisy box proposals from the detector. We use per-class non-maximal suppression (NMS) [38] at 0.3 IoU to pass 64 proposals to the object context branch of our model. We also enforce NMS constraints during decoding given object context. We then sample relationships between proposals that intersect with ground truth boxes and use relationships involving these boxes to finetune the model until detection convergence.

We also observe that in detection our model gets swamped with many low-quality RoI pairs as possible relationships, which slows the model and makes training less stable. To alleviate this, we observe that nearly all annotated relationships are between overlapping boxes,[5] and classify all relationships with non-overlapping boxes as BG.

[3] When sorting by confidence, the edge layer's regions are ordered by the maximum non-background object prediction as given by Equation 4.

[4] We remove the final ReLU to allow more interaction in Equation 6.

[5] A hypothetical model that perfectly classifies relationships, but only between boxes with nonzero IoU, gets 91% recall.

5.3. Evaluation

We train and evaluate our models on Visual Genome, using the publicly released preprocessed data and splits from [47], containing 150 object classes and 50 relation classes, but sample a development set from the training set of 5000 images. We follow three standard evaluation modes: (1) **predicate classification** (PREDCLS): given a ground truth set of boxes and labels, predict edge labels, (2) **scene graph classification** (SGCLS): given ground truth boxes, predict box labels and edge label and (3) **scene graph detection** (SGDET): predict boxes, box labels, and edge labels. The annotated graphs are known to be incomplete, thus systems are evaluated using recall@K metrics.[6]

In all three modes, recall is calculated for relations; a ground truth edge (b_h, o_h, x, b_t, o_t) is counted as a "match" if there exist predicted boxes i, j such that b_i and b_j respectively have sufficient overlap with b_h and b_t,[7] and the objects and relation labels agree. We follow previous work in enforcing that for a given head and tail bounding box, the system must not output multiple edge labels [47, 29].

5.4. Frequency Baselines

To support our finding that object labels are highly predictive of edge labels, we additionally introduce several frequency baselines built off training set statistics. The first, FREQ, uses our pretrained detector to predict object labels for each RoI. To obtain predicate probabilities between boxes i and j, we look up the empirical distribution over relationships between objects o_i and o_j as computed in the training set.[8] Intuitively, while this baseline does not look at the image to compute $\Pr(x_{i \to j} | o_i, o_j)$, it displays the value of *conditioning* on object label predictions o. The second, FREQ-OVERLAP, requires that the two boxes intersect in order for the pair to count as a valid relation.

6. Results

We present our results in Table 6. We compare MOTIFNET to previous models not directly incorporating context (VRD [29] and ASSOC EMBED [31]), a state-of-the-art approach for incorporating graph context via message passing (MESSAGE PASSING) [47], and its reimplemenation using our detector, edge model, and NMS settings (MESSAGE PASSING+). Unfortunately, many scene graph models are evaluated on different versions of Visual Genome; see the supp for more analysis.

Our best frequency baseline, FREQ+OVERLAP, improves over prior state-of-the-art by 1.4 mean recall, pri-

[6] Past work has considered these evaluation modes at recall thresholds R@50 and R@100, but we also report results on R@20.

[7] As in prior work, we compute the intersection-over-union (IoU) between the boxes and use a threshold of 0.5.

[8] Since we consider an edge $x_{i \to j}$ to have label BG if o has no edge to j, this gives us a valid probability distribution.

		Scene Graph Detection			Scene Graph Classification			Predicate Classification			Mean
	Model	R@20	R@50	R@100	R@20	R@50	R@100	R@20	R@50	R@100	
models	VRD [29]		0.3	0.5		11.8	14.1		27.9	35.0	14.9
	MESSAGE PASSING [47]		3.4	4.2		21.7	24.4		44.8	53.0	25.3
	MESSAGE PASSING+	14.6	20.7	24.5	31.7	34.6	35.4	52.7	59.3	61.3	39.3
	ASSOC EMBED [31]⋆	6.5	8.1	8.2	18.2	21.8	22.6	47.9	54.1	55.4	28.3
	FREQ	17.7	23.5	27.6	27.7	32.4	34.0	49.4	59.9	64.1	40.2
	FREQ+OVERLAP	20.1	26.2	30.1	29.3	32.3	32.9	53.6	60.6	62.2	40.7
	MOTIFNET-LEFTRIGHT	21.4	27.2	30.3	**32.9**	**35.8**	**36.5**	**58.5**	**65.2**	**67.1**	**43.6**
ablations	MOTIFNET-NOCONTEXT	21.0	26.2	29.0	31.9	34.8	35.5	57.0	63.7	65.6	42.4
	MOTIFNET-CONFIDENCE	**21.7**	**27.3**	**30.5**	32.6	35.4	36.1	58.2	65.1	67.0	43.5
	MOTIFNET-SIZE	21.6	**27.3**	30.4	32.2	35.0	35.7	58.0	64.9	66.8	43.3
	MOTIFNET-RANDOM	21.6	**27.3**	30.4	32.5	35.5	36.2	58.1	65.1	66.9	43.5

Table 2. Results table, adapted from [47] which ran VRD [29] without language priors. All numbers in %. Since past work doesn't evaluate on R@20, we compute the mean by averaging performance on the 3 evaluation modes over R@50 and R@100. ⋆: results in [31] are without scene graph constraints; we evaluated performance with constraints using saved predictions given to us by the authors (see Table in supp).

marily due to improvements in detection and predicate classification, where it outperforms MESSAGE PASSING+ by 5.5 and 6.5 mean points respectively. MOTIFNET improves even further, by 2.9 additional mean points over the baseline (a 7.1% relative gain).

Ablations To evaluate the effectiveness of our main model, MOTIFNET, we consider several ablations in Table 6. In MOTIFNET-NOCONTEXT, we predict objects based on the fixed detector, and feed non-contexualized embeddings of the head and tail label into Equation 6. Our results suggest that there is signal in the vision features for edge predictions, as MOTIFNET-NOCONTEXT improves over FREQ-OVERLAP. Incorporating context is also important: our full model MOTIFNET improves by 1.2 mean points, with largest gains at the lowest recall threshold of R@20. [9] We additionally validate the impact of the ordering method used, as discussed in Section 5.1; the results vary less than 0.3 recall points, suggesting that MOTIFNET is robust to the RoI ordering scheme used.

7. Qualitative Results

Qualitative examples of our approach, shown in Figure 6, suggest that MOTIFNET is able to induce graph motifs from detection context. Visual inspection of the results suggests that the method works even better than the quantitative results would imply, since many seemingly correct edges are predicted that do not exist in the ground truth.

There are two common failure cases of our model. The first, as exhibited by the middle left image in Figure 6 of a skateboarder carrying a surfboard, stems from predicate ambiguity ("wearing" vs "wears"). The second common failure case occurs when the detector fails, resulting in cascading failure to predict any edges to that object. For example, the failure to predict "house" in the lower left image resulted in five false negative relations.

[9] The larger improvement at the lower thresholds suggests that our models mostly improve on relationship ordering rather than classification. Indeed, it is often unnecessary to order relationships at the higher thresholds: 51% of images have fewer than 50 candidates and 78% have less than 100.

8. Related Work

Context Many methods have been proposed for modeling semantic context in object recognition [7]. Our approach is most closely related to work that models object co-occurrence using graphical models to combine many sources of contextual information [33, 11, 26, 10]. While our approach is a type of graphical model, it is unique in that it stages incorporation of context allowing for meaningful global context from large conditioning sets.

Actions and relations have been a particularly fruitful source of context [30, 50], especially when combined with pose to create human-object interactions [48, 3]. Recent work has shown that object layouts can provide sufficient context for captioning COCO images [52, 28]; our work suggests the same for parsing Visual Genome scene graphs. Much of the context we derive could be interpreted as commonsense priors, which have commonly been extracted using auxiliary means [59, 39, 5, 49, 55]. Yet for scene graphs, we are able to directly extract such knowledge.

Structured Models Structured models in visual understanding have been explored for language grounding, where language determines the graph structures involved in prediction [32, 20, 42, 16]. Our problem is different as we must reason over all possible graph structures. Deep sequential models have demonstrated strong performance for tasks such as captioning [4, 9, 45, 18] and visual question answering [1, 37, 53, 12, 8], including for problems not traditionally not thought of as sequential, such as multilabel classifi-

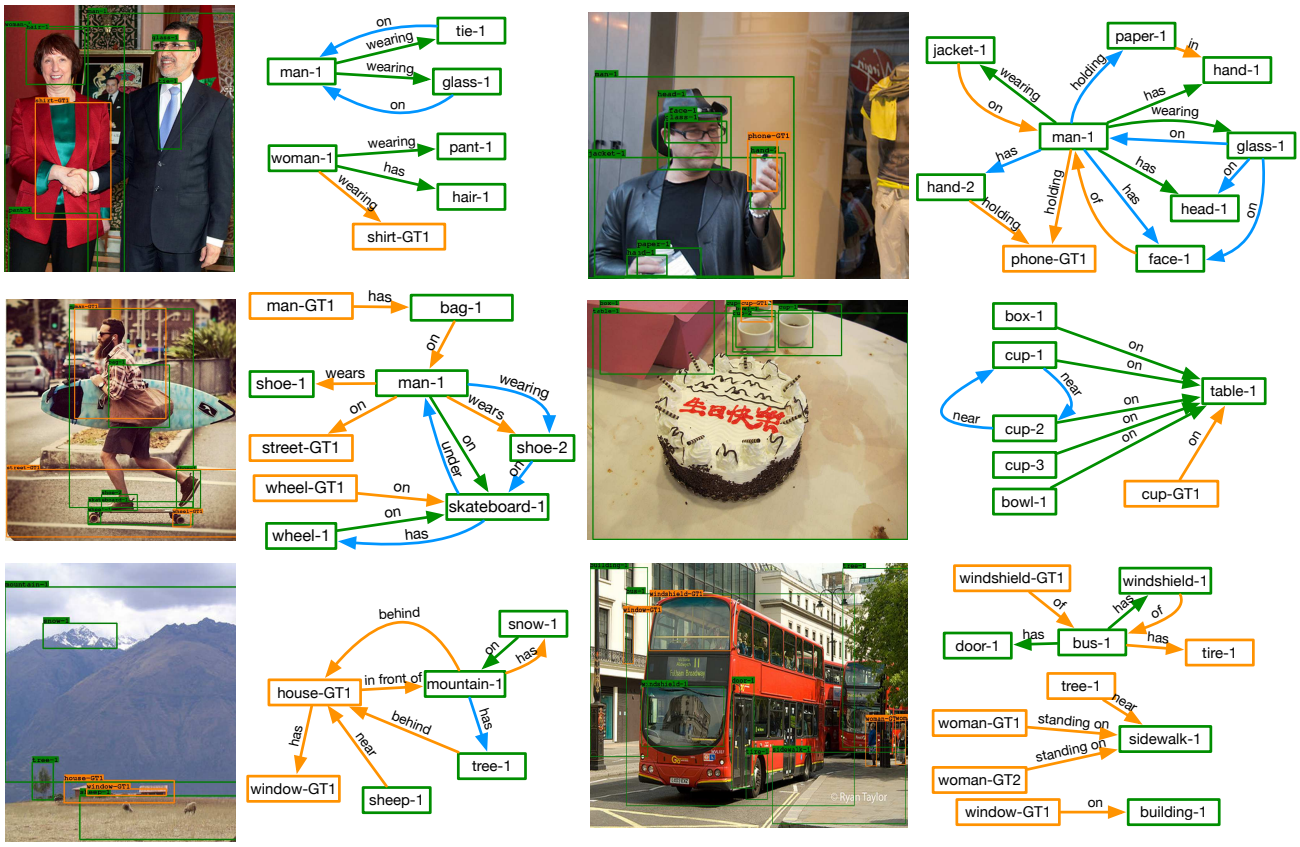

Figure 6. Qualitative examples from our model in the Scene Graph Detection setting. Green boxes are predicted and overlap with the ground truth, orange boxes are ground truth with no match. Green edges are true positives predicted by our model at the R@20 setting, orange edges are false negatives, and blue edges are false positives. Only predicted boxes that overlap with the ground truth are shown.

cation [46]. Indeed, graph linearization has worked surprisingly well for many problems in vision and language, such as generating image captions from object detections [52], language parsing [44], generating text from abstract meaning graphs [21]. Our work leverages the ability of RNNs to memorize long sequences in order to capture graph motifs in Visual Genome. Finally, recent works incorporate recurrent models into detection and segmentation [2, 35] and our methods contribute evidence that RNNs provide effective context for consecutive detection predictions.

Scene Graph Methods Several works have explored the role of priors by incorporating background language statistics [29, 54] or by attempting to preprocess scene graphs [56]. Instead, we allow our model to directly learn to use scene graph priors effectively. Furthermore, recent graph-propagation methods were applied but converge quickly and bottle neck through edges, significantly limiting information exchange [47, 25, 6, 23]. On the other hand, our method allows global exchange of information about context through conditioning and avoids uninformative edge predictions until the end. Others have explored creating richer models between image regions, introducing new convolutional features and new objectives [31, 57, 25, 27]. Our work is complementary and instead focuses on the role of context. See the supplemental section for a comprehensive comparison to prior work.

9. Conclusion

We presented an analysis of the Visual Genome dataset showing that motifs are prevalent, and hence important to model. Motivated by this analysis, we introduced strong baselines that improve over prior state-of-the-art models by modeling these intra-graph interactions, while mostly ignoring visual cues. We also introduced our model MOTIFNET for capturing higher order structure and global interactions in scene graphs that achieves additional significant gains over our already strong baselines.

Acknowledgements

We thank the anonymous reviewers along with Ali Farhadi and Roozbeh Mottaghi for their helpful feedback. This work is supported by the National Science Foundation Graduate Research Fellowship (DGE-1256082), the NSF grant (IIS-1524371, 1703166), DARPA CwC program through ARO (W911NF-15-1-0543), IARPA's DIVA grant, and gifts by Google and Facebook.

References

[1] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh. VQA: visual question answering. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 2425–2433, 2015. 7

[2] L. Castrejon, K. Kundu, R. Urtasun, and S. Fidler. Annotating object instances with a polygon-rnn. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 4485–4493, 2017. 8

[3] Y.-W. Chao, Z. Wang, Y. He, J. Wang, and J. Deng. Hico: A benchmark for recognizing human-object interactions in images. In *Proceedings of the IEEE International Conference on Computer Vision*, 2015. 1, 7

[4] X. Chen and C. Lawrence Zitnick. Mind's eye: A recurrent visual representation for image caption generation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2422–2431, 2015. 7

[5] X. Chen, A. Shrivastava, and A. Gupta. Neil: Extracting visual knowledge from web data. In *International Conference on Computer Vision (ICCV)*, pages 1409–1416. IEEE, 2013. 7

[6] B. Dai, Y. Zhang, and D. Lin. Detecting visual relationships with deep relational networks. 2017. 8

[7] S. K. Divvala, D. Hoiem, J. H. Hays, A. A. Efros, and M. Hebert. An empirical study of context in object detection. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1271–1278. IEEE, 2009. 7

[8] Z. et al. Building a large-scale multimodal knowledge base for visual question answering. *arXiv preprint arXiv:1507.05670*, 2015. 7

[9] H. Fang, S. Gupta, F. Iandola, R. K. Srivastava, L. Deng, P. Dollar, J. Gao, X. He, M. Mitchell, J. C. Platt, C. Lawrence Zitnick, and G. Zweig. From captions to visual concepts and back. June 2015. 7

[10] A. Farhadi, M. Hejrati, M. A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth. Every picture tells a story: Generating sentences from images. In *European Conference on Computer Vision*, pages 15–29. Springer, 2010. 7

[11] C. Galleguillos and S. Belongie. Context based object categorization: A critical survey. *Computer vision and image understanding*, 114(6):712–722, 2010. 7

[12] H. e. a. Gao. Are you talking to a machine? dataset and methods for multilingual image question answering. *arXiv preprint arXiv:1505.05612*, 2015. 7

[13] G. Guo et al. A survey on still image based human action recognition. *Pattern Recognition*, 2014. 1

[14] L. He, K. Lee, M. Lewis, and L. Zettlemoyer. Deep semantic role labeling: What works and whats next. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2017. 5

[15] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, Nov. 1997. 2, 4, 5

[16] R. Hu, J. Andreas, M. Rohrbach, T. Darrell, and K. Saenko. Learning to reason: End-to-end module networks for visual question answering. *arXiv preprint arXiv:1704.05526*, 2017. 7

[17] 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

[18] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137, 2015. 7

[19] J.-H. Kim, K. W. On, W. Lim, J. Kim, J.-W. Ha, and B.-T. Zhang. Hadamard Product for Low-rank Bilinear Pooling. In *The 5th International Conference on Learning Representations*, 2017. 2

[20] C. Kong, D. Lin, M. Bansal, R. Urtasun, and S. Fidler. What are you talking about? text-to-image coreference. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3558–3565, 2014. 7

[21] I. Konstas, S. Iyer, M. Yatskar, Y. Choi, and L. Zettlemoyer. Neural amr: Sequence-to-sequence models for parsing and generation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 146–157, 2017. 8

[22] 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. 1

[23] Y. Li, W. Ouyang, X. Wang, et al. Vip-cnn: Visual phrase guided convolutional neural network. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7244–7253. IEEE, 2017. 8

[24] Y. Li, W. Ouyang, X. Wang, and X. Tang. ViP-CNN: Visual Phrase Guided Convolutional Neural Network. *arXiv:1702.07191 [cs]*, Feb. 2017. arXiv: 1702.07191. 2

[25] Y. Li, W. Ouyang, B. Zhou, K. Wang, and X. Wang. Scene graph generation from objects, phrases and region captions. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017. 2, 5, 8

[26] L.-J. Li et al. What, where and who? classifying events by scene and object recognition. In *CVPR*, 2007. 7

[27] X. Liang, L. Lee, and E. P. Xing. Deep Variation-structured Reinforcement Learning for Visual Relationship and Attribute Detection. *arXiv:1703.03054 [cs]*, Mar. 2017. arXiv: 1703.03054. 8

[28] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollr, and C. L. Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision (ECCV)*, Zrich, 2014. 1, 7

[29] C. Lu, R. Krishna, M. Bernstein, and L. Fei-Fei. Visual relationship detection with language priors. In *European Conference on Computer Vision*, 2016. 6, 7, 8

[30] M. Marszalek, I. Laptev, and C. Schmid. Actions in context. In *Computer Vision and Pattern Recognition, 2009. CVPR*

2009. IEEE Conference on, pages 2929–2936. IEEE, 2009. 7

[31] A. Newell and J. Deng. Pixels to graphs by associative embedding. In *Advances in neural information processing systems*, 2017. 6, 7, 8

[32] B. Plummer, L. Wang, C. Cervantes, J. Caicedo, J. Hockenmaier, and S. Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. *arXiv preprint arXiv:1505.04870*, 2015. 7

[33] A. Rabinovich, A. Vedaldi, C. Galleguillos, E. Wiewiora, and S. Belongie. Objects in context. In *Computer vision, 2007. ICCV 2007. IEEE 11th international conference on*, pages 1–8. IEEE, 2007. 7

[34] J. Redmon and A. Farhadi. YOLO9000: Better, Faster, Stronger. *arXiv:1612.08242 [cs]*, Dec. 2016. arXiv: 1612.08242. 1, 5

[35] M. Ren and R. S. Zemel. End-to-end instance segmentation and counting with recurrent attention. *arXiv preprint arXiv:1605.09410*, 2016. 8

[36] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *arXiv:1506.01497 [cs]*, June 2015. arXiv: 1506.01497. 1, 2, 4, 5

[37] M. Ren et al. Image question answering: A visual semantic embedding model and a new dataset. *arXiv preprint arXiv:1505.02074*, 2015. 7

[38] A. Rosenfeld and M. Thurston. Edge and curve detection for visual scene analysis. *IEEE Trans. Comput.*, 20(5):562–569, May 1971. 6

[39] F. Sadeghi, S. K. Divvala, and A. Farhadi. Viske: Visual knowledge extraction and question answering by visual verification of relation phrases. In *Conference on Computer Vision and Pattern Recognition*, pages 1456–1464, 2015. 7

[40] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014. 5

[41] R. K. Srivastava, K. Greff, and J. Schmidhuber. Training very deep networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'15, pages 2377–2385, Cambridge, MA, USA, 2015. MIT Press. 5

[42] S. Tellex, T. Kollar, S. Dickerson, M. R. Walter, A. G. Banerjee, S. Teller, and N. Roy. Approaching the symbol grounding problem with probabilistic graphical models. *AI magazine*, 32(4):64–76, 2011. 7

[43] D. Teney, L. Liu, and A. van den Hengel. Graph-structured representations for visual question answering. *CVPR*, 2017. 1

[44] O. Vinyals, Ł. Kaiser, T. Koo, S. Petrov, I. Sutskever, and G. Hinton. Grammar as a foreign language. In *Advances in Neural Information Processing Systems*, pages 2773–2781, 2015. 8

[45] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3156–3164, 2015. 7

[46] J. Wang, Y. Yang, J. Mao, Z. Huang, C. Huang, and W. Xu. Cnn-rnn: A unified framework for multi-label image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2285–2294, 2016. 8

[47] D. Xu, Y. Zhu, C. B. Choy, and L. Fei-Fei. Scene Graph Generation by Iterative Message Passing. *arXiv:1701.02426 [cs]*, Jan. 2017. arXiv: 1701.02426. 2, 5, 6, 7, 8

[48] B. Yao et al. Modeling mutual context of object and human pose in human-object interaction activities. In *CVPR*, 2010. 1, 7

[49] M. Yatskar, V. Ordonez, and A. Farhadi. Stating the obvious: Extracting visual common sense knowledge. In *Proceedings of NAACL*, 2016. 7

[50] M. Yatskar, L. Zettlemoyer, and A. Farhadi. Situation Recognition: Visual Semantic Role Labeling for Image Understanding. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 7

[51] X. Yin and V. Ordonez. Obj2text: Generating visually descriptive language from object layouts. In *EMNLP*, 2017. 1

[52] X. Yin and V. Ordonez. Obj2text: Generating visually descriptive language from object layouts. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2017. 7, 8

[53] L. e. a. Yu. Visual madlibs: Fill in the blank image generation and question answering. *arXiv preprint arXiv:1506.00278*, 2015. 7

[54] R. Yu, A. Li, V. I. Morariu, and L. S. Davis. Visual relationship detection with internal and external linguistic knowledge distillation. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 8

[55] R. Zellers and Y. Choi. Zero-shot activity recognition with verb attribute induction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2017. 7

[56] H. Zhang, Z. Hu, Y. Deng, M. Sachan, Z. Yan, and E. P. Xing. Learning Concept Taxonomies from Multimodal Data. *arXiv:1606.09239 [cs]*, June 2016. arXiv: 1606.09239. 8

[57] H. Zhang, Z. Kyaw, S.-F. Chang, and T.-S. Chua. Visual translation embedding network for visual relation detection. *CVPR*, 2017. 8

[58] Y. Zhang, G. Chen, D. Yu, K. Yaco, S. Khudanpur, and J. Glass. Highway long short-term memory rnns for distant speech recognition. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5755–5759, March 2016. 5

[59] Y. Zhu, A. Fathi, and L. Fei-Fei. Reasoning about object affordances in a knowledge base representation. In *European conference on computer vision*, pages 408–424. Springer, 2014. 7