HHS Public Access

Author manuscript

Proc IEEE Int Conf Comput Vis. Author manuscript; available in PMC 2020 March 26.

Published in final edited form as:

Proc IEEE Int Conf Comput Vis. 2019 ; 2019: 2580–2590. doi:10.1109/iccv.2019.00267.

Scene Graph Prediction with Limited Labels

Vincent S. Chen, **Paroma Varma**, **Ranjay Krishna**, **Michael Bernstein**, **Christopher Ré**, **Li Fei-Fei**

Stanford University

Abstract

Visual knowledge bases such as Visual Genome power numerous applications in computer vision, including visual question answering and captioning, but suffer from sparse, incomplete relationships. All scene graph models to date are limited to training on a small set of visual relationships that have thousands of training labels each. Hiring human annotators is expensive, and using textual knowledge base completion methods are incompatible with visual data. In this paper, we introduce a semi-supervised method that assigns probabilistic relationship labels to a large number of unlabeled images using few' labeled examples. We analyze visual relationships to suggest two types of image-agnostic features that are used to generate noisy heuristics, whose outputs are aggregated using a factor graph-based generative model. With as few as 10 labeled examples per relationship, the generative model creates enough training data to train any existing state-of-the-art scene graph model. We demonstrate that our method outperforms all baseline approaches on scene graph prediction by 5.16 recall@ 100 for PREDCLS. In our limited label setting, we define a complexity metric for relationships that serves as an indicator ($R^2 = 0.778$) for conditions under which our method succeeds over transfer learning, the de-facto approach for training with limited labels.

1. Introduction

In an effort to formalize a structured representation for images, Visual Genome [27] defined **scene graphs,** a formalization similar to those widely used to represent knowledge bases [13, 18, 56]. Scene graphs encode objects (e.g. person, bike) as nodes connected via pairwise relationships (e.g., riding) as edges. This formalization has led to state-of-the-art models in image captioning [3], image retrieval [25,42], visual question answering [24], relationship modeling [26] and image generation [23]. However, all existing scene graph models ignore more than 98% of relationship categories that do not have sufficient labeled instances (see Figure 2) and instead focus on modeling the few relationships that have thousands of labels [31, 49, 54].

Hiring more human workers is an ineffective solution to labeling relationships because image annotation is so tedious that seemingly obvious labels are left unannotated. To complement human annotators, traditional text-based knowledge completion tasks have leveraged numerous semi-supervised or distant supervision approaches [6, 7, 17, 34], These

vincentsc@cs.stanford.edu.

methods find syntactical or lexical patterns from a small labeled set to extract missing relationships from a large unlabeled set. In text, pattern-based methods are successful, as relationships in text are usually **document-agnostic** (e.g. <Tokyo - is capital of – Japan>). Visual relationships are often incidental: they depend on the contents of the particular image they appear in. Therefore, methods that rely on external knowledge or on patterns over concepts (e.g. most instances of dog next to frisbee are playing with it) do not generalize well. The inability to utilize the progress in text-based methods necessitates specialized methods for visual knowledge.

In this paper, we automatically generate missing relationships labels using a small, labeled dataset and use these generated labels to train downstream scene graph models (see Figure 1). We begin by exploring how to define **image-agnostic** features for relationships so they follow patterns across images. For example, eat usually consists of one object consuming another object smaller than itself, whereas look often consists of common objects: phone, laptop, or window (see Figure 3). These rules are not dependent on raw pixel values; they can be derived from image-agnostic features like object categories and relative spatial positions between objects in a relationship. While such rules are simple, their capacity to provide supervision for unannotated relationships has been unexplored. While image-agnostic features can characterize *some* visual relationships very well, they might fail to capture *complex relationships* with high variance. To quantify the efficacy of our image-agnostic features, we define “subtypes” that measure spatial and categorical complexity (Section 3).

Based on our analysis, we propose a semi-supervised approach that leverages image-agnostic features to label missing relationships using as few as 10 labeled instances of each relationship. We learn simple heuristics over these features and assign probabilistic labels to the unlabeled images using a generative model [39,46]. We evaluate our method’s labeling efficacy using the completely-labeled VRD dataset [31] and find that it achieves an FI score of 57.66, which is 11.84 points higher than other standard semi-supervised methods like label propagation [57]. To demonstrate the utility of our generated labels, we train a state-of-the-art scene graph model [54] (see Figure 6) and modify its loss function to support probabilistic labels. Our approach achieves 47.53 recall@ 100[1] for predicate classification on Visual Genome, improving over the same model trained using only labeled instances by 40.97 points. For scene graph detection, our approach achieves within 8.65 recall@ 100 of the same model trained on the original Visual Genome dataset with 108 × more labeled data. We end by comparing our approach to transfer learning, the de-facto choice for learning from limited labels. We find that our approach improves by 5.16 recall@ 100 for predicate classification, especially for relationships with high complexity, as it generalizes well to unlabeled subtypes.

Our contributions are three-fold. (1) We introduce the first method to complete visual knowledge bases by finding missing visual relationships (Section 5.1). (2) We show the utility of our generated labels in training existing scene graph prediction models (Section 5.2). (3) We introduce a metric to characterize the complexity of visual relationships and

[1]Recall@K is a standard measure for scene graph prediction [31].

show it is a strong indicator ($R^2 = 0.778$) for our semi-supervised method's improvements over transfer learning (Section 5.3).

2. Related work

Textual knowledge bases were originally hand-curated by experts to structure facts [4,5,44] (e.g. <Tokyo - capital of – Japan>). To scale dataset curation efforts, recent approaches mine knowledge from the web [9] or hire non-expert annotators to manually curate knowledge [5, 47]. In semi-supervised solutions, a small amount of labeled text is used to extract and exploit patterns in unlabeled sentences [2, 21, 33–35, 37]. Unfortunately, such approaches cannot be directly applied to visual relationships; textual relations can often be captured by external knowledge or patterns, while visual relationships are often local to an image.

Visual relationships have been studied as spatial priors [14, 16], co-occurrences [51], language statistics [28, 31, 53], and within entity contexts [29]. Scene graph prediction models have dealt with the difficulty of learning from incomplete knowledge, as recent methods utilize statistical motifs [54] or object-relationship dependencies [30, 49, 50, 55]. All these methods limit their inference to the top 50 most frequently occurring predicate categories and ignore those without enough labeled examples (Figure 2).

The de-facto solution for limited label problems is **transfer learning** [15, 52], which requires that the source domain used for pre-training follows a similar distribution as the target domain. In our setting, the source domain is a dataset of frequently-labeled relationships with thousands of examples [30, 49, 50, 55], and the target domain is a set of limited label relationships. Despite similar objects in source and target domains, we find that transfer learning has difficulty generalizing to new relationships. Our method does not rely on availability of a larger, labeled set of relationships; instead, we use a small labeled set to annotate the unlabeled set of images.

To address the issue of gathering enough training labels for machine learning models, **data programming** has emerged as a popular paradigm. This approach learns to model imperfect labeling sources in order to assign training labels to unlabeled data. Imperfect labeling sources can come from crowdsourcing [10], user-defined heuristics [8, 43], multi-instance learning [22, 40], and distant supervision [12, 32]. Often, these imperfect labeling sources take advantage of domain expertise from the user. In our case, imperfect labeling sources are automatically generated heuristics, which we aggregate to assign a final probabilistic label to every pair of object proposals.

3. Analyzing visual relationships

We define the formal terminology used in the rest of the paper and introduce the image-agnostic features that our semi-supervised method relies on. Then, we seek quantitative insights into how visual relationships can be described by the properties between its objects. We ask (1) what image-agnostic features can characterize visual relationships? and (2) given limited labels, how well do our chosen features characterize the complexity of relationships? With these in mind, we motivate our model design to generate heuristics that do not overfit to the small amount of labeled data and assign accurate labels to the larger, unlabeled set.

3.1 Terminology

A scene graph is a multi-graph $\mathbb{G}$ that consists of objects o as nodes and relationships r as edges. Each object $o_i = \{b_i, c_i\}$ consists of a bounding box b_i and its category $c_i \in \mathbb{C}$ where $\mathbb{C}$ is the set of all possible object categories (e.g. dog, frisbee). Relationships are denoted <subject predicate - object> or $<o - p - o'>$. $p \in \mathbb{P}$ is a predicate, such as ride and eat. We assume that we have a small labeled set $\{(o, p, o') \in D_p\}$ of annotated relationships for each predicate p. Usually, these datasets are on the order of a 10 examples or fewer. For our semi-supervised approach, we also assume that there exists a large set of images D_U without any labeled relationships.

3.2 Defining image-agnostic features

It has become common in computer vision to utilize pretrained convolutional neural networks to extract features that represent objects and visual relationships [31, 49, 50]. Models trained with these features have proven robust in the presence of enough training labels but tend to overfit when presented with limited data (Section 5). Consequently, an open question arises: what other features can we utilize to label relationships with limited data? Previous literature has combined deep learning features with extra information extracted from categorical object labels and relative spatial object locations [25, 31]. We define categorical features, $< o, –, o' >$, as a concatenation of one-hot vectors of the subject o and object o'. We define spatial features as:

$$\frac{x - x'}{w}, \frac{y - y'}{h}, \frac{(y + h) - (y' + h')}{h},$$
$$\frac{(x + w) - (x' + w')}{w}, \frac{h'}{h}, \frac{w'}{w}, \frac{w'h'}{wh}, \frac{w' + h'}{w + h}$$

where $b = [y, x, h, w]$ and $b' = [y', x', h', w']$ are the top-left bounding box coordinates and their widths and heights.

To explore how well spatial and categorical features can describe different visual relationships, we train a simple decision tree model for each relationship. We plot the importances for the top 4 spatial and categorical features in Figure 3. Relationships like fly place high importance on the difference in y-coordinate between the subject and object, capturing a characteristic spatial pattern, look, on the other hand, depends on the category of the objects (e.g. phone, laptop, window) and not on any spatial orientations.

3.3 Complexity of relationships

To understand the efficacy of image-agnostic features, we'd like to measure how well they can characterize the complexity of particular visual relationships. As seen in Figure 4, a visual relationship can be defined by a number of image-agnostic features (e.g. a person can ride a bike, or a dog can ride a surfboard). To systematically define this notion of complexity, we identify **subtypes** for each visual relationship. Each subtype captures one way that a relationship manifests in the dataset. For example, in Figure 4, ride contains one categorical subtype with <person - ride - bike> and another with <dog - ride - surfboard>. Similarly, a person might carry an object in different relative spatial orientations (e.g. on her

head, to her side). As shown in Figure 5, visual relationships might have significantly different degrees of spatial and categorical complexity, and therefore a different number of subtypes for each. To compute spatial subtypes, we perform mean shift clustering [11] over the spatial features extracted from all the relationships in Visual Genome. To compute the categorical subtypes, we count the number of unique object categories associated with a relationship.

With access to 10 or fewer labeled instances for these visual relationships, it is impossible to capture *all* the subtypes for given relationship and therefore difficult to learn a good representation for the relationship as a whole. Consequently, we turn to the rules extracted from image-agnostic features and use them to assign labels to the unlabeled data in order to capture a larger proportion of subtypes in each visual relationship. We posit that this will be advantageous over methods that only use the small labeled set to train a scene graph prediction model, especially for relationships with high complexity, or a large number of subtypes. In Section 5.3, we find a correlation between our definition of complexity and the performance of our method.

4. Approach

We aim to automatically generate labels for missing visual relationships that can be then used to train any downstream scene graph prediction model. We assume that in the long-tail of infrequent relationships, we have a small labeled set $\{(o, p, o') \in D_p\}$ of annotated relationships for each predicate p (often, on the order of a 10 examples or less). As discussed in Section 3, we want to leverage image-agnostic features to learn rules that annotate unlabeled relationships.

Our approach assigns probabilistic labels to a set D_U of un-annotated images in three steps: (1) we extract image-agnostic features from the objects in the labeled D_p and from the object proposals extracted using an existing object detector [19] on unlabeled D_U (2) we generate heuristics over the image-agnostic features, and finally (3) we use a factor-graph based generative model to aggregate and assign probabilistic labels to the unlabeled object pairs in D_U. These probabilistic labels, along with D_p, are used to train any scene graph prediction model. We describe our approach in Algorithm 1 and show the end-to-end pipeline in Figure 6.

Algorithm 1 Semi-supervised Alg. to Label Relationships

1: **INPUT**: $\{(o, p, o') \in D_p\} \forall p \in \mathbb{P} -$ A small dataset of object pairs (*o, o'*) with multi-class labels for predicates.

2: **INPUT**: $\{(o, o')\} \in D_U\} -$ A A large unlabeled dataset of images with objects but no relationship labels.

3: **INPUT**: $f(\cdot,\cdot)$ — A function that extracts features from a pair of objects.

4: **INPUT**: $DT(\cdot)$ — A decision tree.

5: **INPUT**: G (·) — A generative model that assigns probabilistic labels given multiple labels for each datapoint

6: **INPUT**: train(·) — Function used to train a scene graph detection model.

7: Extract features and labels,
$X_p, Y_p := \{f(o,o'), p \text{ for } (o,p,o') \in D_p\}, X_U := \{(f(o,o') \text{ for } (o,o') \in D_U\}$

8: Generate heuristics by fitting J decision trees $DT_{fit}(X_p)$

9: Assign labels to $(o,o') \in D_U$, $\Lambda = DT_{predict}(X_U)$ for J decision trees.

10: Learn generative model $G(\Lambda)$ and assign probabilistic labels $\tilde{Y}_U := G(\Lambda)$

11: Train scene graph model, SGM := train$(D_p + D_U, Y_p + \tilde{Y}_U)$

12: **OUTPUT**: SGM(·)

Feature extraction:

Our approach uses the image-agnostic features defined in Section 3, which rely on object bounding box and category labels. The features are extracted from ground truth objects in D_p or from object detection outputs in D_U by running existing object detection models [19].

Heuristic generation:

We fit decision trees over the labeled relationships' spatial and categorical features to capture image-agnostic rules that define a relationship. These image-agnostic rules are threshold-based conditions that are automatically defined by the decision tree. To limit the complexity of these heuristics and thereby prevent overfitting, we use shallow decision trees [38] with different restrictions on depth over each feature set to produce J different decision trees. We then predict labels for the unlabeled set using these heuristics, producing a $\Lambda \in \mathbb{R}^{J \times |D_U|}$ matrix of predictions for the unlabeled relationships.

Moreover, we only use these heuristics when they have high confidence about their label; we modify Λ by converting any predicted label with confidence less than a threshold (empirically chosen to be 2 × random) to an *abstain,* or no label assignment. An example of a heuristic is shown in Figure 6: if the subject is above the object, it assigns a positive label for the predicate carry.

Generative model:

These heuristics, individually, are noisy and may not assign labels to all object pairs in D_U. As a result, we aggregate the labels from all J heuristics. To do so, we leverage a factor graph-based generative model popular in text-based weak supervision techniques [1, 39, 41, 45, 48]. This model learns the accuracies of each heuristic to combine their individual labels; the model's output is a probabilistic label for each object pair.

The generative model G uses the following distribution family to relate the latent variable $Y \in \mathbb{R}^{|D_U|}$, the true class, and the labels from the heuristics, Λ:

$$\pi_\phi(\Lambda, Y) = \frac{1}{Z_\phi} \exp\left(\phi^T \Lambda Y\right)$$

where Z_ϕ is a partition function to ensure $\boldsymbol{\pi}$ is normalized. The parameter $\phi \in \mathbb{R}^J$ encodes the average accuracy of each heuristic and is estimated by maximizing the marginal likelihood of the observed heuristic Λ. The generative model assigns probabilistic labels by computing $\pi_\phi(Y \mid \Lambda(o, o'))$ for each object pair (o, o') in D_U

Training scene graph model:

Finally, these probabilistic labels are used to train any scene graph prediction model. While scene graph models are usually trained using a cross-entropy loss [31, 49, 54], we modify this loss function to take into account errors in the training annotations. We adopt a noise-aware empirical risk minimizer that is often seen in logistic regression as our loss function:

$$L_\theta = \mathbb{E}_{Y \sim \pi}\left[\log\left(1 + \exp\left(-\theta^T V^T Y\right)\right)\right]$$

where θ is the learned parameters, $\boldsymbol{\pi}$ is the distribution learned by the generative model, Y is the true label, and V are features extracted by any scene graph prediction model.

5. Experiments

To test our semi-supervised approach for completing visual knowledge bases by annotating missing relationships, we perform a series of experiments and evaluate our framework in several stages. We start by discussing the datasets, baselines, and evaluation metrics used. (1) Our first experiment tests our generative model's ability to find missing relationships in the completely-annotated VRD dataset [31]. (2) Our second experiment demonstrates the utility of our generated labels by using them to train a state-of-the-art scene graph model [54]. We compare our labels to those from the large Visual Genome dataset [27]. (3) Finally, to show that our semi-supervised method's performance compared to strong baselines in limited label settings, we compare extensively to transfer learning; we focus on a subset of relationships with limited labels, allow the transfer learning model to pretrain on frequent relationships, and demonstrate that our semi-supervised method outperforms transfer learning, which has seen more data. Furthermore, we quantify when our method outperforms transfer learning using our metric for measuring relationship complexity (Section 3.3).

Eliminating synonyms and supersets.

Typically, past scene graph approaches have used 50 predicates from Visual Genome to study visual relationships. Unfortunately, these 50 treat synonyms like laying on and lying on as separate classes. To make matters worse, some predicates can be considered a superset of others (i.e. above is a superset of riding). Our method, as well as the baselines, is unable to differentiate between synonyms and supersets. For the experiments in this section, we eliminate all supersets and merge all synonyms, resulting in 20 unique predicates. In the Supplementary Material we include a list of these predicates and report our method's performance on all 50 predicates.

Dataset.

We use two standard datasets, VRD [31] and Visual Genome [27], to evaluate on tasks related to visual relationships or scene graphs. Each scene graph contains objects localized as bounding boxes in the image along with pairwise relationships connecting them, categorized as action (e.g., carry), possessive (e.g., wear), spatial (e.g., above), or comparative (e.g., taller than) descriptors. Visual Genome is a large visual knowledge base containing $108K$ images. Due to its scale, each scene graph is left with incomplete labels, making it difficult to measure the precision of our semi-supervised algorithm. VRD is a smaller but completely annotated dataset. To show the performance of our semi-supervised method, we measure our method's generated labels on the VRD dataset (Section 5.1). Later, we show that the training labels produced can be used to train a large scale scene graph prediction model, evaluated on Visual Genome (Section 5.2).

Evaluation metrics.

We measure precision and recall of our generated labels on the VRD dataset's test set (Section 5.1). To evaluate a scene graph model trained on our labels, we use three standard evaluation modes for scene graph prediction [31]: (i) scene graph detection (SGDET) which expects input images and predicts bounding box locations, object categories, and predicate labels, (ii) scene graph classification (SGCLS) which expects ground truth boxes and predicts object categories and predicate labels, and (iii) predicate classification (PREDCLS), which expects ground truth bounding boxes and object categories to predict predicate labels. We refer the reader to the paper that introduced these tasks for more details [31]. Finally, we explore how relationship complexity, measured using our definition of subtypes, is correlated with our model's performance relative to transfer learning (Section 5.3).

Baselines.

We compare to alternative methods for generating training labels that can then be used to train downstream scene graph models, ORACLE is trained on all of Visual Genome, which amounts to 108× the quantity of labeled relationships in D_p; this serves as the upper bound for how well we expect to perform. DECISION TREE [38] fits a single decision tree over the image-agnostic features, learns from labeled examples in D_p, and assigns labels to D_U LABEL PROPAGATION [57] employs a widely-used semi-supervised method and considers the distribution of image-agnostic features in D_U before propagating labels from D_p to D_U

We compare to a strong frequency baselines: (FREQ) uses the object counts as priors to make relationship predictions, and FREQ+OVERLAP increments such counts only if the bounding boxes of objects overlap. We include a TRANSFER LEARNING baseline, which is the de-facto choice for training models with limited data [15, 52], However, unlike all other methods, transfer learning requires a source dataset to pretrain. We treat the source domain as the remaining relationships from the top 50 in Visual Genome that do not overlap with our chosen relationships. We then fine tune with the limited labeled examples for the predicates in D_p. We note that TRANSFER LEARNING has an unfair advantage because there is overlap in objects between its source and target relationship sets. Our experiments will show that even with this advantage, our method performs better.

Ablations.

We perform several ablation studies for the image-agnostic features and heuristic aggregation components of our model. (CATEG.) uses only categorical features, (SPAT.) uses only spatial features, (DEEP) uses only deep learning features extracted using ResNet50 [20] from the union of the object pair's bounding boxes, (CATEG. + SPAT.) uses both categorical concatenated with spatial features, (CATEG. + SPAT. + DEEP) combines combines all three, and OURS (CATEG. + SPAT. + WORDVEC) includes word vectors as richer representations of the categorical features. (MAJORITY VOTE) uses the categorical and spatial features but replaces our generative model with a simple majority voting scheme to aggregate heuristic function outputs.

5.1 Labeling missing relationships

We evaluate our performance in annotating missing relationships in D_U Before we use these labels to train scene graph prediction models, we report results comparing our method to baselines in Table 1. On the fully annotated VRD dataset [31], OURS (CATEG. + SPAT.) achieves 57.66 F1 given only 10 labeled examples, which is 17.41, 13.88, and 1.55 points better than Label Propagation, Decision Tree and Majority Vote, respectively.

Qualitative error analysis.—We visualize labels assigned by Ours in Figure 7 and find that they correspond to image-agnostic rules explored in Figure 3. In Figure 7(a), OURS predicts fly because it learns that fly typically involves objects that have a large difference in y-coordinate. In Figure 7(b), we correctly label look because phone is an important categorical feature. In some difficult cases, our semi-supervised model fails to generalize beyond the image-agnostic features. In Figure 7(c), we mislabel hang as sit by incorrectly relying on the categorical feature chair, which is one of sit's important features. In Figure 7(d), ride typically occurs directly above another object that is slightly larger and assumes <book - ride - shelf> instead of <book - sitting on - shelf>. In Figure 7(e), our model reasonably classifies <glasses cover - face>. However, sit exhibits the same semantic meaning as cover in this context, and our model incorrectly classifies the example.

5.2 Training Scene graph prediction models

We compare our method's labels to those generated by the baselines described earlier by using them to train three scene graph specific tasks and report results in Table 2. We improve over all baselines, including our primary baseline, TRANSFER LEARNING, by 5.16 recall@100 for PREDCLS.We also achieve within 8.65 recall@100 of ORACLE for SGDET. We generate higher quality training labels than DECISION TREE and LABEL PROPAGATION, leading to an 13.83 and 22.12 recall@100 increase for PREDCLS.

Effect of labeled and unlabeled data.—In Figure 8 (left two graphs), we visualize how SGCLS and PREDCLS performance varies as we reduce the number of labeled examples from $n = 250$ to $n = 100, 50, 25, 10$. We observe greater advantages over TRANSFER LEARNING as n decreases, with an increase of 5.16 recall@100 PREDCLS when $n = 10$. This result matches our observations from Section 3 because a larger set of labeled examples gives TRANSFER LEARNING information about a larger proportion of subtypes for each relationship. in Figure 8

(right two graphs), we visualize our performance as the number of unlabeled data points increase, finding that we approach ORACLE performance with more unlabeled examples.

Ablations.—OURS (CATEG. + SPAT. + DEEP.) hurts performance by up to 7.51 recall@100 for PREDCLS because it overfits to image features while OURS (CATEG. + SPAT.) performs the best. We show improvements of 0.71 recall@100 for SGDET over OURS (MAJORITYVOTE), indicating that the generated heuristics indeed have different accuracies and should be weighted differently.

5.3 Transfer learning vs. semi-supervised learning

Inspired by the recent work comparing transfer learning and semi-supervised learning [36], we characterize when our method is preferred over transfer learning. Using the relationship complexity metric based on spatial and categorical subtypes of each predicate (Section 3), we show this trend in Figure 9. When the predicate has a high complexity (as measured by a high number of subtypes), OURS (CATEG. + SPAT.) outperforms TRANSFER LEARNING (Figure 9, left), with correlation co-efficient $R^2 = 0.778$. We also evaluate how the number of subtypes in the unlabeled set (D_U) affects the performance of our model (Figure 9, center). We find a strong correlation ($R^2 = 0.745$); our method can effectively assign labels to unlabeled relationships with a large number of subtypes. We also compare the difference in performance to the proportion of subtypes captured in the labeled set (Figure 9, right). As we hypothesized earlier, TRANSFER LEARNING suffers in cases when the labeled set only captures a small portion of the relationship's subtypes. This trend ($R^2 = 0.701$) explains how OURS (CATEG. + SPAT.) performs better when given a small portion of labeled subtypes.

6. Conclusion

We introduce the first method that completes visual knowledge bases like Visual Genome by finding missing visual relationships. We define categorical and spatial features as image-agnostic features and introduce a factor-graph based generative model that uses these features to assign probabilistic labels to unlabeled images. Our method outperforms baselines in F1 score when finding missing relationships in the complete VRD dataset. Our labels can also be used to train scene graph prediction models with minor modifications to their loss function to accept probabilistic labels. We outperform transfer learning and other baselines and come close to oracle performance of the same model trained on a fraction of labeled data. Finally, we introduce a metric to characterize the complexity of visual relationships and show it is a strong indicator of how our semi-supervised method performs compared to such baselines.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements.

This work was partially funded by the Brown Institute of Media Innovation, the Toyota Research Institute ("TRI"), DARPA under Nos. FA87501720095 and FA86501827865, NIH under No. U54EB020405, NSF under Nos. CCF1763315 and CCF1563078, ONR under No. N000141712266, the Moore Foundation, NXP, Xilinx, LETI-CEA, Intel, Google, NEC, Toshiba, TSMC, ARM, Hitachi, BASF, Accenture, Ericsson, Qualcomm, Analog Devices, the Okawa Foundation, and American Family Insurance, Google Cloud, Swiss Re, NSF Graduate

Research Fellowship under No. DGE-114747, Joseph W. and Hon Mai Goodman Stanford Graduate Fellowship, and members of Stanford DAWN: Intel, Microsoft, Teradata, Facebook, Google, Ant Financial, NEC, SAP, VMWare, and Infosys. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views, policies, or endorsements, either expressed or implied, of DARPA, NIH, ONR, or the U.S. Government.

References

[1]. Alfonseca Enrique, Filippova Katja, Delort Jean-Yves, and Garrido Guillermo. Pattern learning for relation extraction with a hierarchical topic model. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2, pages 54–59. Association for Computational Linguistics, 2012 5

[2]. Anderson Carolyn J, Wasserman Stanley, and Faust Katherine. Building stochastic blockmodels. Social networks, 14(1– 2):137–161, 1992 2

[3]. Anderson Peter, Fernando Basura, Johnson Mark, and Gould Stephen. Spice: Semantic propositional image caption evaluation. In European Conference on Computer Vision, pages 382–398. Springer, 2016 1

[4]. Auer Sören, Bizer Christian, Kobilarov Georgi, Lehmann Jens, Cyganiak Richard, and Ives Zachary. Dbpedia: A nucleus for a web of open data In The semantic web, pages 722–735. Springer, 2007 2

[5]. Bollacker Kurt, Evans Colin, Paritosh Praveen, Sturge Tim, and Taylor Jamie. Freebase: a collaboratively created graph database for structuring human knowledge. In Proceedings of the 2008 ACM SIGMOD international conference on Management of data, pages 1247–1250. AcM, 2008. 2

[6]. Bordes Antoine, Glorot Xavier, Weston Jason, and Bengio Yoshua. A semantic matching energy function for learning with multi-relational data. Machine Learning, 94(2):233–259, 2014 1

[7]. Bordes Antoine, Usunier Nicolas, Alberto Garcia-Duran Jason Weston, and Yakhnenko Oksana. Translating embeddings for modeling multi-relational data. In Advances in neural information processing systems, pages 2787–2795, 2013 1

[8]. Bunescu Razvan and Mooney Raymond. Learning to extract relations from the web using minimal supervision. In Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, pages 576–583, 2007 2

[9]. Carlson Andrew, Betteridge Justin, Kisiel Bryan, Settles Burr, Hruschka Estevam R Jr, and Mitchell Tom M. Toward an architecture for never-ending language learning. In AAAI, volume 5, page 3 Atlanta, 2010 2

[10]. Cheng Justin and Bernstein Michael S. Flock: Hybrid crowd-machine learning classifiers. In Proceedings of the 18th ACM conference on computer supported cooperative work & social computing, pages 600–611. ACM, 2015 2

[11]. Cheng Yizong. Mean shift, mode seeking, and clustering. IEEE transactions on pattern analysis and machine intelligence, 17(8):790–799, 1995 3

[12]. Craven Mark, Kumlien Johan, et al. Constructing biological knowledge bases by extracting information from text sources. In ISMB, volume 1999, pages 77–86, 1999 2

[13]. Culotta Aron and Sorensen Jeffrey. Dependency tree kernels for relation extraction. In Proceedings of the 42nd annual meeting on association for computational linguistics, page 423 Association for Computational Linguistics, 2004 1

[14]. Dai Bo, Zhang Yuqi, and Lin Dahua. Detecting visual relationships with deep relational networks. In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 3298–3308. IEEE, 2017 2

[15]. Donahue Jeff, Jia Yangqing, Vinyals Oriol, Hoffman Judy, Zhang Ning, Tzeng Eric, and Darrell Trevor. Decaf: A deep convolutional activation feature for generic visual recognition. In International conference on machine learning, pages 647–655, 2014 2, 6

[16]. Galleguillos Carolina, Rabinovich Andrew, and Belongie Serge. Object categorization using co-occurrence, location and appearance. In Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on, pages 1–8. IEEE, 2008 2

[17]. Gardner Matt, Talukdar Partha, Krishnamurthy Jayant, and Mitchell Tom. Incorporating vector space similarity in random walk inference over knowledge bases. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 397–406, 2014 1

[18]. Zhou GuoDong Su Jian, Jie Zhang, and Min Zhang. Exploring various knowledge in relation extraction. In Proceedings of the 43rd annual meeting on association for computational linguistics, pages 427–434. Association for Computational Linguistics, 2005 1

[19]. He Kaiming, Gkioxari Georgia, Piotr Dollár, and Ross Gir-shick Mask r-cnn. In Computer Vision (ICCV), 2017 IEEE International Conference on, pages 2980–2988. IEEE, 2017 4

[20]. He Kaiming, Zhang Xiangyu, Ren Shaoqing, and Sun Jian. Deep residual learning for image recognition. arXiv preprint arXiv:1512.03385, 2015 6

[21]. Hoff Peter. Modeling homophily and stochastic equivalence in symmetric relational data. In Advances in neural information processing systems, pages 657–664, 2008 2

[22]. Hoffmann Raphael, Zhang Congle, Ling Xiao, Zettle-moyer Luke, and Weld Daniel S. Knowledge-based weak supervision for information extraction of overlapping relations. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1, pages 541–550. Association for Computational Linguistics, 2011 2

[23]. Johnson Justin, Gupta Agrim, and Fei-Fei Li. Image generation from scene graphs. arXiv preprint arXiv:1804.01622, 2018 1

[24]. Johnson Justin, Hariharan Bharath, Maaten Laurens van der, Hoffman Judy, Fei-Fei Li, Zitnick C Lawrence, and Gir-shick Ross. Inferring and executing programs for visual reasoning. arXiv preprint arXiv:1705.03633, 2017 1

[25]. Johnson Justin, Krishna Ranjay, Stark Michael, Li Li-Jia, Shamma David, Bernstein Michael, and Fei-Fei Li. Image retrieval using scene graphs. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 3668–3678, 2015 1, 3

[26]. Krishna Ranjay, Chami Ines, Bernstein Michael, and Fei- Fei Li. Referring relationships. In IEEE Conference on Computer Vision and Pattern Recognition, 2018 1

[27]. Krishna Ranjay, Zhu Yuke, Groth Oliver, Johnson Justin, Hata Kenji, Kravitz Joshua, Chen Stephanie, Kalantidis Yannis, Li Li-Jia, Shamma David A, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. International Journal of Computer Vision, 123(1):32–73, 2017 1, 5, 6

[28]. Li Yikang, Ouyang Wanli, Wang Xiaogang, and Tang Xiao'Ou. Vip-cnn: Visual phrase guided convolutional neural network. In Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on, pages 7244–7253. IEEE, 2017 2

[29]. Li Yikang, Ouyang Wanli, Zhou Bolei, Wang Kun, and Wang Xiaogang. Scene graph generation from objects, phrases and region captions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1261–1270, 2017 2

[30]. Liang Xiaodan, Lee Lisa, and Xing Eric P. Deep variation-structured reinforcement learning for visual relationship and attribute detection. In Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on, pages 4408–4417. IEEE, 2017 2

[31]. Lu Cewu, Krishna Ranjay, Bernstein Michael, and Fei-Fei Li. Visual relationship detection with language priors. In European Conference on Computer Vision, pages 852–869. Springer, 2016 1,2, 3, 5, 6, 7

[32]. Mintz Mike, Bills Steven, Snow Rion, and Jurafsky Dan. Distant supervision for relation extraction without labeled data. In Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2, pages 1003–1011. Association for Computational Linguistics, 2009 2

[33]. Nickel Maximilian. Tensor factorization for relational learning. PhD thesis, lmu, 2013 2

[34]. Nickel Maximilian, Tresp Volker, and Kriegel Hans-Peter. A three-way model for collective learning on multi-relational data. In ICML, volume 11, pages 809–816, 2011 1, 2

[35]. Nickel Maximilian, Tresp Volker, and Kriegel Hans-Peter. Factorizing yago: scalable machine learning for linked data. In Proceedings of the 21st international conference on World Wide Web, pages 271–280. ACM, 2012 2

[36]. Oliver Avital, Odena Augustus, Raffel Colin, Cubuk Ekin D, and Goodfellow Ian J. Realistic evaluation of deep semi-supervised learning algorithms. arXiv preprint arXiv:1804.09170, 2018 8

[37]. Orbanz Peter and Roy Daniel M. Bayesian models of graphs, arrays and other exchangeable random structures. IEEE transactions on pattern analysis and machine intelligence, 37(2):437–461, 2015 2 [PubMed: 26353253]

[38]. Ross Quinlan J. Induction of decision trees. Machine learning, 1(1):81–106, 1986 5, 6, 7

[39]. Ratner Alexander J, De Sa Christopher M, Wu Sen, Selsam Daniel, and Re Christopher. Data programming: Creating large training sets, quickly In Lee DD, Sugiyama M, Luxburg UV, Guyon I, and Garnett R, editors, Advances in Neural Information Processing Systems 29, pages 3567–3575. Curran Associates, Inc, 2016 2, 5 [PubMed: 29872252]

[40]. Riedel Sebastian, Yao Limin, and McCallum Andrew. Modeling relations and their mentions without labeled text. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases, pages 148–163. Springer, 2010 2

[41]. Roth Benjamin and Klakow Dietrich. Combining generative and discriminative model scores for distant supervision. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pages 24–29, 2013 5

[42]. Schuster Sebastian, Krishna Ranjay, Chang Angel, Fei-Fei Li, and Manning Christopher D. Generating semantically precise scene graphs from textual descriptions for improved image retrieval. In Proceedings of the fourth workshop on vision and language, pages 70–80, 2015 1

[43]. Shin Jaeho, Wu Sen, Wang Feiran, De Sa Christopher, Zhang Ce, and Ré Christopher. Incremental knowledge base construction using deepdive. Proceedings of the VLDB Endowment, 8(11):1310–1321, 2015 2 [PubMed: 27144081]

[44]. Fabian M Suchanek Gjergji Kasneci, and Weikum Gerhard. Yago: a core of semantic knowledge. In Proceedings of the 16th international conference on World Wide Web, pages 697–706. ACM, 2007 2

[45]. Takamatsu Shingo, Sato Issei, and Nakagawa Hiroshi. Reducing wrong labels in distant supervision for relation extraction. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1, pages 721–729. Association for Computational Linguistics, 2012 5

[46]. Varma Paroma, Bryan D He Payal Bajaj, Khandwala Nishith, Banerjee Imon, Rubin Daniel, and Ré Christopher. Inferring generative model structure with static analysis. In Advances in Neural Information Processing Systems, pages 239–249, 2017 2 [PubMed: 29391769]

[47]. Vrandečić Denny and Krötzsch Markus. Wikidata: a free collaborative knowledgebase. Communications of the ACM, 57(10):78–85, 2014 2

[48]. Xiao Tong, Xia Tian, Yang Yi, Huang Chang, and Wang Xiaogang. Learning from massive noisy labeled data for image classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 2691–2699, 2015 5

[49]. Xu Danfei, Zhu Yuke, 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, volume 2, 2017 1, 2, 3, 5

[50]. Yang Jianwei, Lu Jiasen, Lee Stefan, Batra Dhruv, and Parikh Devi. Graph r-cnn for scene graph generation. arXiv preprint arXiv:1808.00191, 2018 2, 3

[51]. Yao Bangpeng and Fei-Fei Li. Modeling mutual context of object and human pose in human-object interaction activities. In Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on, pages 17–24. IEEE, 2010 2

[52]. Yosinski Jason, Clune Jeff, Bengio Yoshua, and Lipson Hod. How transferable are features in deep neural networks? In Advances in neural information processing systems, pages 3320–3328, 2014 2, 6

[53]. Yu Ruichi, Li Ang, Morariu Vlad I, and Davis Larry S. Visual relationship detection with internal and external linguistic knowledge distillation. arXiv preprint arXiv:1707.09423, 2017 2

[54]. Zellers Rowan, Yatskar Mark, Thomson Sam, and Choi Yejin. Neural motifs: Scene graph parsing with global context. arXiv preprint arXiv:1711.06640, 2017 1, 2, 5, 7

[55]. Zhang Ji, Kalantidis Yannis, Rohrbach Marcus, Paluri Manohar, Elgammal Ahmed, and Elhoseiny Mohamed. Large-scale visual relationship understanding. arXiv preprint arXiv:1804.10660, 2018 2

[56]. Zhou Guodong, Zhang Min, Ji DongHong, and Zhu Qiaoming. Tree kernel-based relation extraction with context-sensitive structured parse tree information. In Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), 2007 1

[57]. Zhu Xiaojin and Ghahramani Zoubin. Learning from labeled and unlabeled data with label propagation. Technical Report, 2002 2, 6, 7

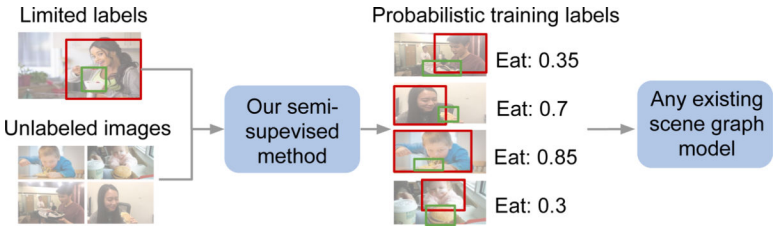

Figure 1.

Our semi-supervised method automatically generates probabilistic relationship labels to train any scene graph model.

NUM. LABELED ($\leq n$)	200	175	150	125	100	75	50	25	10	5
% RELATIONSHIPS	99.09	99.00	98.87	98.74	98.52	98.15	97.57	96.09	92.26	87.28

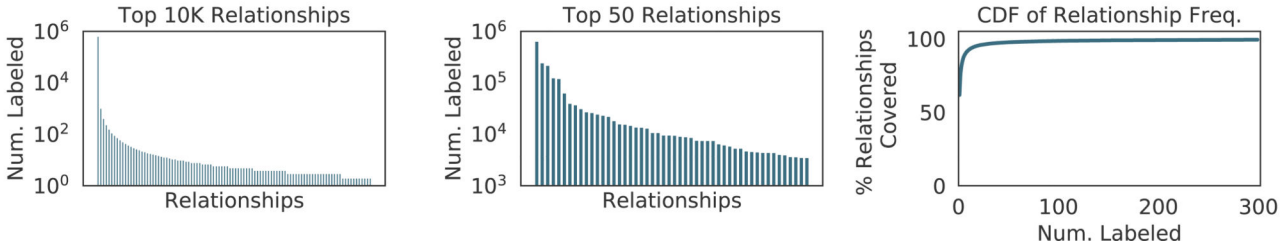

Figure 2.

Visual relationships have a long tail (left) of infrequent relationships. Current models [49,54] only focus on the top 50 relationships (middle) in the Visual Genome dataset, which all have thousands of labeled instances. This ignores more than 98% of the relationships with few labeled instances (right, top/table).

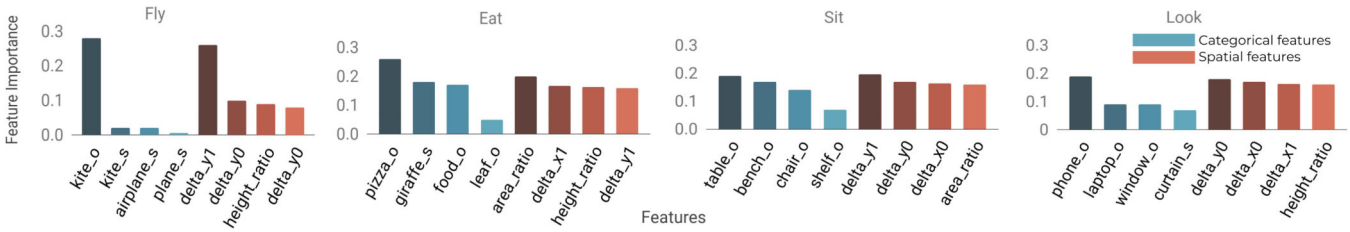

Figure 3.

Relationships, such as fly, eat, and sit can be characterized effectively by their categorical (s and o refer to subject and object, respectively) or spatial features. Some relationships like fly rely heavily only on a few features — kites are often seen high up in the sky.

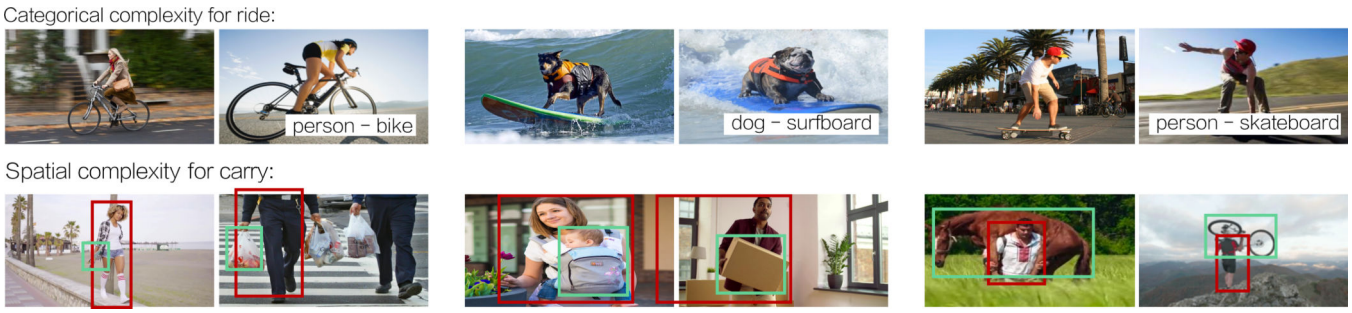

Figure 4.

We define the number of subtypes of a relationship as a measure of its complexity. Subtypes can be categorical — one subtype of ride can be expressed as <person - ride - bike> while another is <dog - ride - surfboard>. Subtypes can also be spatial—carry has a subtype with a small object carried to the side and another with a large object carried overhead.

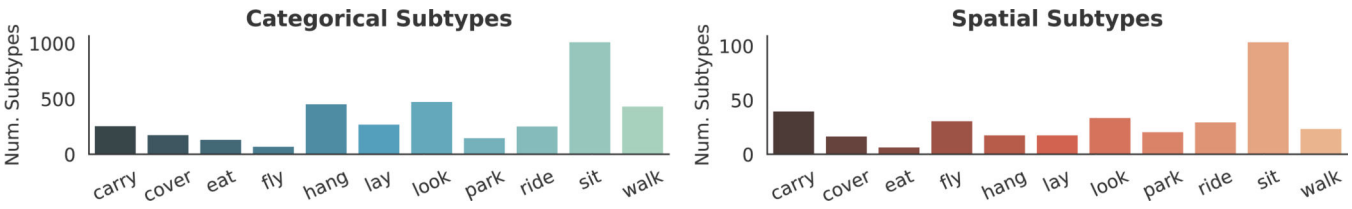

Figure 5.

A subset of visual relationships with different levels of complexity as defined by spatial and categorical subtypes. In Section 5.3, we show how this measure is a good indicator of our semi-supervised method's effectiveness compared to baselines like transfer learning.

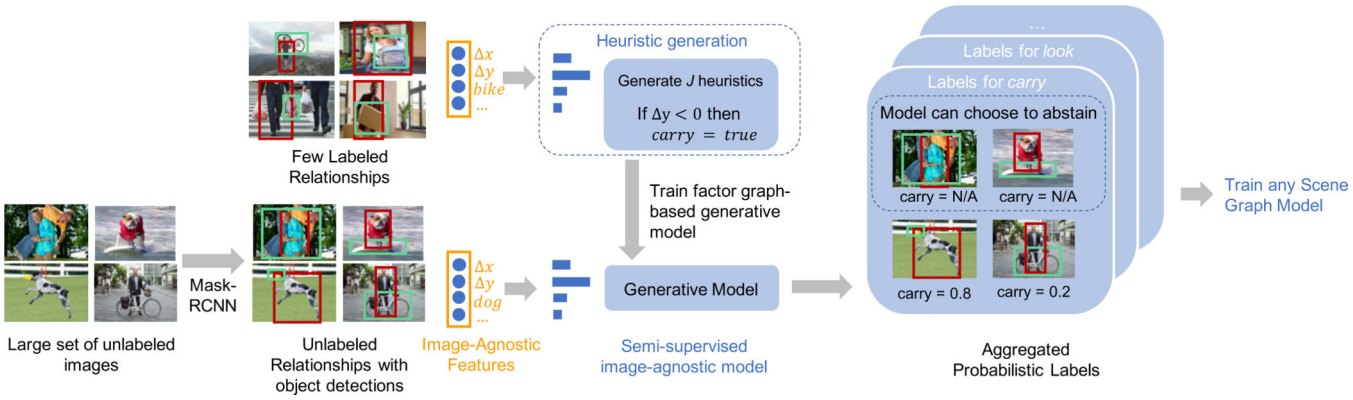

Figure 6.

For a relationship (e.g., carry), we use image-agnostic features to automatically create heuristics and then use a generative model to assign probabilistic labels to a large unlabeled set of images. These labels can then be used to train any scene graph prediction model.

Figure 7.

(a) Heuristics based on spatial features help predict <man - fly - kite>. (b) Our model learns that look is highly correlated with phone, (c) We overfit to the importance of chair as a categorical feature for sit, and fail to identify hang as the correct relationship, (d) We overfit to the spatial positioning associated with ride, where objects are typically longer and directly underneath the subject, (e) Given our image-agnostic features, we produce a reasonable label for <glass - cover - f ace>. However, our model is incorrect, as two typically different predicates (sit and cover) share a semantic meaning in the context of <glasses - ? - f ace>.

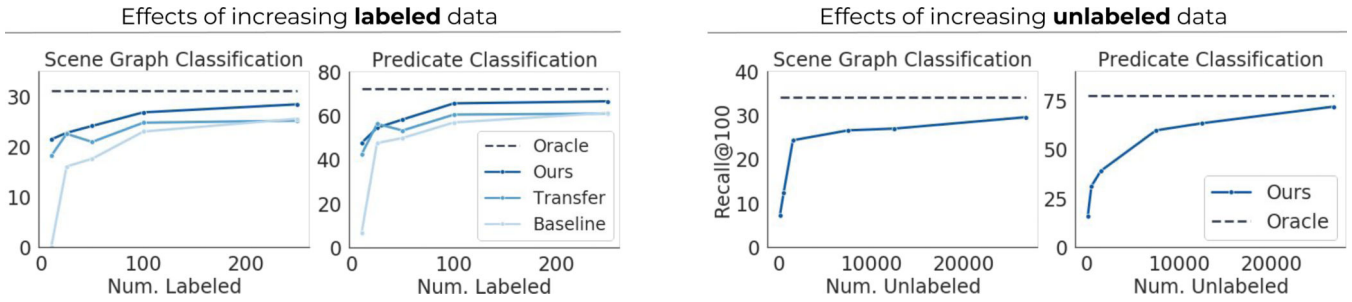

Figure 8.

A scene graph model [54] trained using our labels outperforms both using TRANSFER LEARNING labels and using only the BASELINE labeled examples consistently across scene graph classification and predicate classification for different amounts of available labeled relationship instances. We also compare to ORACLE, which is trained with 108× more labeled data.

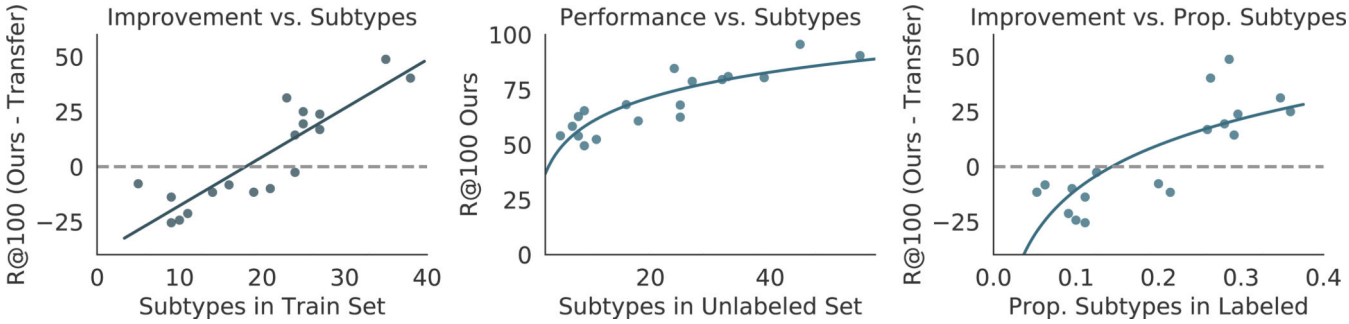

Figure 9.

Our method's improvement over transfer learning (in terms of R@100 for predicate classification) is correlated to the number of subtypes in the train set (left), the number of subtypes in the unlabeled set (middle), and the proportion of subtypes in the labeled set (right).

Table 1.

We validate our approach for labeling missing relationships using only $n = 10$ labeled examples by evaluating our probabilistic labels from our semi-supervised approach over the fully-annotated VRD using macro metrics dataset [31],

Model ($n = 10$)	**Prec.**	**Recall**	**F1**	**Acc.**
RANDOM	5.00	5.00	5.00	5.00
DECISION TREE	46.79	35.32	40.25	36.92
LABEL PROPAGATION	**76.48**	32.71	45.82	12.85
OURS (MAJORITY VOTE)	55.01	57.26	56.11	40.04
OURS (CATEG. + SPAT.)	54.83	**60.79**	**57.66**	**50.31**

Table 2.

Results for scene graph prediction tasks with $n = 10$ labeled examples per predicate, reported as recall @K. A state-of-the-art scene graph model trained on labels from our method outperforms those trained with labels generated by other baselines, like transfer learning.

		Scene Graph Detection			Scene Graph Classification			Predicate Classification		
	Model	**R@20**	**R@50**	**R@100**	**R@20**	**R@50**	**R@100**	**R@20**	**R@50**	**R@100**
Baselines	BASELINE [$n = 10$]	0.00	0.00	0.00	0.04	0.04	0.04	3.17	5.30	6.61
	FREQ	9.01	11.01	11.64	11.10	11.08	10.92	20.98	20.98	20.80
	FREQ+OVERLAP	10.16	10.84	10.86	9.90	9.91	9.91	20.39	20.90	22.21
	TRANSFER LEARNING	11.99	14.40	16.48	17.10	17.91	18.16	39.69	41.65	42.37
	DECISION TREE [38]	11.11	12.58	13.23	14.02	14.51	14.57	31.75	33.02	33.35
	LABEL PROPAGATION [57]	6.48	6.74	6.83	9.67	9.91	9.97	24.28	25.17	25.41
Ablations	OURS (DEEP)	2.97	3.20	3.33	10.44	10.77	10.84	23.16	23.93	24.17
	OURS (SPAT.)	3.26	3.20	2.91	10.98	11.28	11.37	26.23	27.10	27.26
	OURS (CATEG.)	7.57	7.92	8.04	20.83	**21.44**	**21.57**	43.49	44.93	45.50
	OURS (CATEG. + SPAT. + DEEP)	7.33	7.70	7.79	17.03	17.35	17.39	38.90	39.87	40.02
	OURS (CATEG. + SPAT. + WORDVEC)	8.43	9.04	9.27	20.39	20.90	21.21	45.15	46.82	47.32
	OURS (MAJORITY VOTE)	16.86	18.31	18.57	18.96	19.57	19.66	44.18	45.99	46.63
	OURS (CATEG. + SPAT.)	**17.67**	**18.69**	**19.28**	**20.91**	21.34	21.44	**45.49**	**47.04**	**47.53**
	ORACLE [$n_{\text{ORACLE}} = 108n$]	24.42	29.67	30.15	30.15	30.89	31.09	69.23	71.40	72.15

Proc IEEE Int Conf Comput Vis. Author manuscript; available in PMC 2020 March 26.