Referring Relationships

Ranjay Krishna†, Ines Chami†, Michael Bernstein, Li Fei-Fei
Stanford University
{ranjaykrishna, chami, msb, feifeili}@cs.stanford.edu

Abstract

Images are not simply sets of objects: each image represents a web of interconnected relationships. These relationships between entities carry semantic meaning and help a viewer differentiate between instances of an entity. For example, in an image of a soccer match, there may be multiple `persons` *present, but each participates in different relationships: one is* `kicking` *the* `ball`*, and the other is* `guarding` *the* `goal`*. In this paper, we formulate the task of utilizing these "referring relationships" to disambiguate between entities of the same category. We introduce an iterative model that localizes the two entities in the referring relationship, conditioned on one another. We formulate the cyclic condition between the entities in a relationship by modelling predicates that connect the entities as shifts in attention from one entity to another. We demonstrate that our model can not only outperform existing approaches on three datasets — CLEVR, VRD and Visual Genome — but also that it produces visually meaningful predicate shifts, as an instance of interpretable neural networks. Finally, we show that by modelling predicates as attention shifts, we can even localize entities in the absence of their category, allowing our model to find completely unseen categories.*

1. Introduction

Referring expressions in everyday discourse help identify and locate entities[1] in our surroundings. For instance, we might point to the "person kicking the ball" to differentiate from the "person guarding the goal" (Figure 1). In both these examples, we disambiguate between the two `persons` by their respective relationships with other entities [23]. While one `person` is `kicking` the `ball`, the other is `guarding` the `goal`. The eventual goal is to build computational models that can identify which entities others are referring to [34].

† = equal contribution

[1]We use the term "entities" for what is commonly referred to as "objects" to differentiate from the term `object` in <subject-predicate-object> relationships.

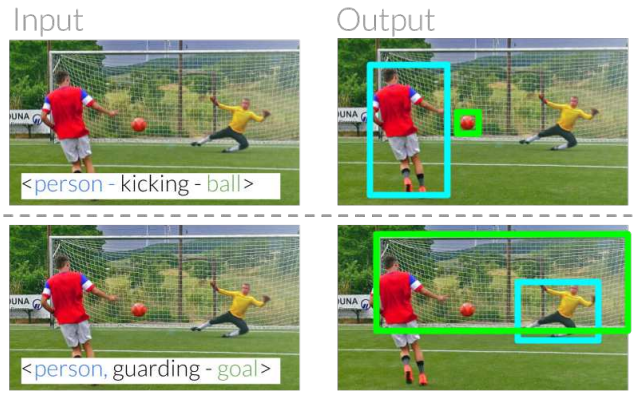

Figure 1: Referring relationships disambiguate between instances of the same category by using their relative relationships with other entities. Given the relationship <`person - kicking - ball`>, the task requires our model to correctly identify which `person` in the image is kicking the `ball` by understanding the predicate `kicking`.

To enable such interactions, we introduce referring relationships — a task where, given a relationship, models should know which entities in a scene are being referred to by the relationship. Formally, the task expects an input image along with a relationship, which is of the form <`subject - predicate - object`>, and outputs localizations of both the `subject` and `object`. For example, we can express the above examples as <`person - kicking - ball`> and <`person - guarding - goal`> (Figure 1). Previous work has attempted to disambiguate entities of the same category in the context of referring expression comprehension [28, 24, 41, 42, 11]. Their task expects a natural language input, such as "a person guarding the goal", resulting in evaluations that require both natural language as well as computer vision components. It can be challenging to pinpoint whether errors made by these models occur from either the language or the visual components. By interfacing with a structured relationship input, our task is a special case of referring expressions that alleviates the need to model language.

Referring relationships retain and refine the algorithmic challenges at the core of prior tasks. In the object localization literature, some entities such as `zebra` and `person` are highly discriminative and can be easily detected, while others such as `glass` and `ball` tend to be harder to localize [29]. These difficulties arise due to, for example, small size and non-discriminative composition. This difference in difficulty translates over to the referring relationships task. To tackle this challenge, we use the intuition that detecting one entity becomes easier if we know where the other one is. In other words, we can find the `ball` conditioned on the `person` who is `kicking` it and vice versa. We train this cyclic dependency by rolling out our model and iteratively passing messages between the subject and the object through an operator defined by the `predicate`. We describe this operator in more detail in Section 3.

However, modelling this `predicate` operator is not straightforward, which leads us to our second challenge. Traditionally, previous visual relationship papers have learned an appearance-based model for each predicate [20, 23, 26]. Unfortunately, the drastic appearance variance of predicates, depending on the entities involved, makes learning predicate appearance models challenging. For example, the appearance for the predicate `carrying` can vary significantly between the following two relationships: <`person - carrying - phone`> and <`truck - carrying - hay`>. Instead, inspired by the moving spotlight theory in psychology [18, 35], we bypass this challenge by using predicates as a visual attention shift operation from one entity to the other. While one shift operation learns to move attention from the `subject` to the `object`, an inverse predicate shift similarly moves attention from the `object` back to the `subject`. Over multiple iterations, we operationalize these asymmetric attention shifts between the `subject` and the `object` as different types of message operations for each `predicate` [37, 9].

In summary, we introduce the task of referring relationships, whose structured relationship input allows us to evaluate how well we can unambiguously identify entities of the same category in an image. We evaluate our model[2] on three vision datasets that contain visual relationships: CLEVR [12], VRD [23] and Visual Genome [17]. 33%, 60.3%, and 61% of relationships in these datasets refer to ambiguous entities, i.e. entities that have multiple instances of the same category. We extend our model to perform attention saccades [36] using relationships belonging to a scene graph [14]. Finally, we demonstrate that in the absence of a `subject` or the `object`, our model can still disambiguate between entities while also localizing entities from new categories that it has never seen before.

[2]Our model was coded using Keras with a Tensorflow backend and is available at https://github.com/StanfordVL/ReferringRelationships.

2. Related Work

To properly situate the task of referring relationships, we explore the evolution of visual relationships as a representation. Next, we survey the inception of referring expression comprehension as a similar task, summarize how attention has been used in the deep learning literature, and survey other technical approaches that are similar to our approach.

There is a long history of vision papers moving beyond simple object detection and **modelling the context** around the entities [27, 31] or even studying object co-occurrences [8, 19, 25] to improve classification and detection itself. Our task on referring relationships was motivated by such papers. Unlike these models, we utilize a formal definition for context in the form of a **visual relationship**.

Pushing along this thread, visual relationships were initially limited to spatial relationships: `above`, `below`, `inside` and `around` [8]. Relationships were then extended to include human interactions, such as `holding` and `carrying` [40]. Extending the definition further, the task of visual relationship detection was introduced along with a dataset of spatial, comparative, action and verb predicates [23]. More recently, relationships were formalized as part of an explicit formal representation for images called scene graphs [14, 17], along with a dataset of scene graphs called Visual Genome [17]. These scene graphs encode the entities in a scene as nodes in a graph that are connected together with directed edges representing their relative relationships. Scene graphs have shown to improve a number of computer vision tasks, including semantic image retrieval [33], image captioning [1] and object detection [30]. Newer work has extended models for relationship detection to use co-occurrence statistics [26, 32, 37] and have even formulated the problem in a reinforcement learning framework [21]. These papers focused primarily on detecting visual relationships categorically — they output relationships given an input image. In contrast, we focus on the inverse problem of localizing the entities that take part in an input relationship. We disambiguate entities in a query relationship from other entities of the same category in the image. Moreover, while all previous work has attempted to learn visual features of predicates, we propose that the visual appearances of predicates are too varied and can be more effectively learnt as an attention shift, conditioned on the entities in the relationship.

Such an inverse task of disambiguating between different regions in an image has been studied under the task of **referring expression comprehension** [24]. This task uses an input language description to find the referred entities. This work has been motivated by human-robot interaction, where the robot would have to disambiguate which entities the human user is referring to [34]. Models for their task have been extended to include global image contrasts [41], visual relationships [11] and reward-based reinforcement systems

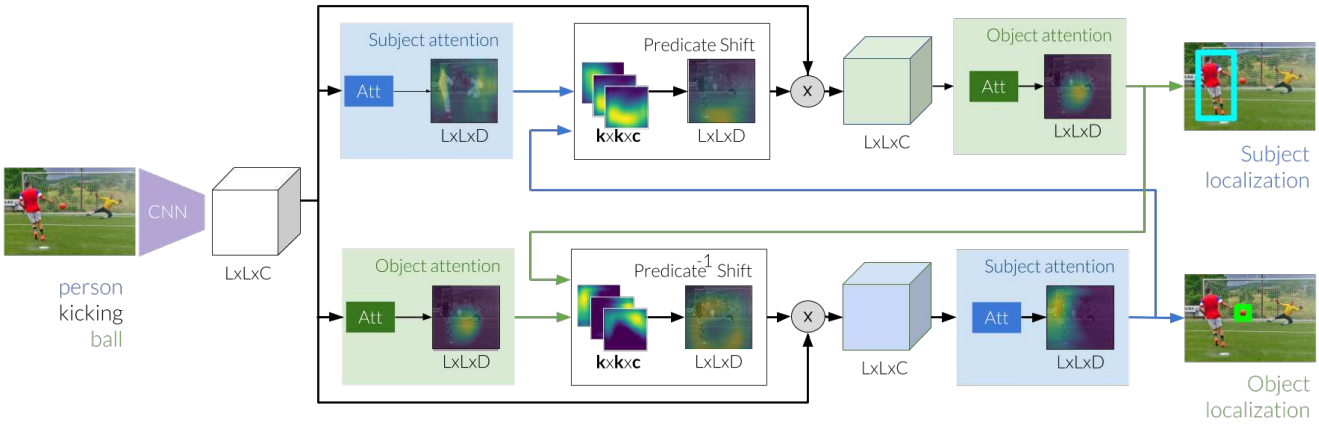

Figure 2: Referring relationships' inference pipeline begins by extracting image features, which are then used to generate an initial grounding of the `subject` and `object` independently. Next, these estimates are used to shift the attention using the `predicate` from the `subject` to where we expect the `object` to be. We modify the image features by focusing our attention to the shifted area when refining our new estimate of the `object`. Simultaneously, we learn an inverse shift from the initial `object` to the `subject`. We iteratively pass messages between the subject and object through the two predicate shift modules to finally localize the two entities.

that encourage the generation of unique expressions for different image regions [41]. Unfortunately, all these models require the ability to process both natural language as well as visual constructs. This requirement makes it difficult to disentangle the mistakes as a result of poor language modelling or visual understanding. In an effort to ameliorate these limitations, we propose the referring relationships task — simplifying referring expressions by replacing the language inputs with a structured relationship. We focus solely on the visual component of the model, avoiding confounding errors from language processing.

One key observations about predicates is their large variance in visual appearance [23]. For example, consider these two relationships: <person - carrying - phone> and <truck - carrying - hay>. We use an insight from psychology [18, 35], specifically the moving spotlight theory, which suggests that visual attention can be modelled as a spotlight that can be conditioned on and directed towards specific targets. The use of attention has been explored to improve image captioning [38, 2] and even stacked to improve question answering [13, 39]. In comparison, we model two discriminative **attention shifting** operations for each unique predicate, one conditioned on the `subject` to localize the `object` and an inverse predicate shift conditioned on the `object` to find the `subject`. Each predicate utilizes both the current estimate of the entities as well as image features to learn how to shift, allowing it to utilize both spatial and semantic features.

Our work also has similarities to **knowledge bases**, where predicates are often projections in a defined semantic space [3, 6, 22]. Such a method was recently used for visual relationship detection [43]. While these methods have seen success in knowledge base completion tasks, they have only led to a marginal gain for modelling visual relationships. However, unlike these methods, we do not model predicates as a projection in semantic space but as a shift in attention conditioned on an entity in a relationship. Our method can be thought of as a special case of deformable parts model [7] with two deformable parts, one for each entity. Finally, our messaging passing algorithm can be thought of as a domain-specific specialized version to the message passing in graph convolution approximation methods [9, 15].

3. Referring relationships model

Recall that our aim is to use the input referring relationship to disambiguate entities in an image by localizing the entities involved in the relationship. Formally, the input is an image $\mathbf{I}$ with a referring relationship, R =<S - P - O>, which are the `subject`, `predicate` and `object` categories, respectively. The model is expected to localize both the `subject` and the `object`.

3.1. Problem formulation

We begin by using a pre-trained convolutional neural network (CNN) to extract a $L \times L \times C$ dimensional feature map from the image $\boldsymbol{\mu} = \text{CNN}(I)$. That is, for each image, we extract a 3-dimensional tensor of shape $L \times L \times C$, where L is the spatial size of the feature map while C is the number of feature channels. Our goal is to decide if each $L \times L$ image region belongs to the subject or object or neither. We can model this problem by representing the image by two binary random variables $\mathbf{X}, \mathbf{Y}$. For $i = 1 \ldots L \times L$,

$X_i > \tau$ implies that the subject occupies the region i and $Y_i > \tau$ implies that the object occupies that region, for some hyperparameter threshold τ. We now define a graph $G = (\mathcal{V}_X \cup \mathcal{V}_Y, \mathcal{E})$, where $\mathcal{V}_X = \{x_i\}$, $\mathcal{V}_Y = \{y_i\}$ are the nodes of the graph represented by the image regions and $\mathcal{E} = (x_i, y_j)$ represents an edge from every x_i to y_j. Given the image and relationship, we want to assign $\mathbf{x}^*$ and $\mathbf{y}^*$ with $\mathbf{x}^*, \mathbf{y}^* = \arg\max_{\mathbf{x},\mathbf{y}} \Pr(\mathbf{X} = \mathbf{x}, \mathbf{Y} = \mathbf{y} | \boldsymbol{\mu}, R)$.

This optimization problem can be reduced to inference on a densely connected graph which can be very expensive. As shown in previous work [44, 16], dense graph inference can be approximated by mean field in Conditional Random Fields (CRF). Such papers allow fully differential inference assuming weighted gaussians as pairwise potentials [44]. To achieve greater flexibility in a more principled training framework, we design a general model where the messaging passing during inference is a series of learnt convolutions. More specifically, we design our model with two types of modules: attention and predicate shift modules. While attention models attempt to locate a specific category in an image, the predicate shift modules learn to move attention from one entity to another.

3.2. Symmetric stacked attention shifting (SSAS) model

Before we specify our attention and shift operators, let's revisit the challenges in referring relationships to motivate our design decisions. The two challenges are (1) the difference in difficulty in object detection and (2) the drastic appearance variance of predicates. First, the difference in difficulty arises because some objects like zebra and person are highly discriminative and can be easily detected while others like glass and ball tend to be harder to localize. We can overcome this problem by conditioning the localization of one entity on the other. If we know where the person is, we should be able to estimate the location of the ball that they are kicking.

Second, predicates tend to vary in appearance depending on the objects involved in the relationship. To deal with the wide appearance variance of predicates, we move away from how previous work [23] attempted to learn appearance features of predicates and instead treat predicates as a mechanism for shifting the attention from one object to another. Relationships like above should learn to focus attention down from the subject when locating the object, and the predicate left of should focus the attention to the right of the subject. Inversely, once we locate the object, the model should use left of to focus attention to the left to confirm its initial estimate of the subject. Note that not all predicates are spatial, so we also ensure that we can model their visual appearances by conditioning the shifts on the image features as well.

Attention modules. With these design goals in mind, we formulate the attention module as an initial estimate of the subject and object localizations by approximating the maximizers $\mathbf{x}^*$, $\mathbf{y}^*$ with the soft attention $\mathrm{Att}(\cdot)$:

$$\hat{\mathbf{x}}^0 = \mathrm{Att}(\boldsymbol{\mu}, S) = \mathrm{ReLU}(\boldsymbol{\mu} \cdot \mathrm{Emb}(S)) \quad (1)$$

$$\hat{\mathbf{y}}^0 = \mathrm{Att}(\boldsymbol{\mu}, O) = \mathrm{ReLU}(\boldsymbol{\mu} \cdot \mathrm{Emb}(O)) \quad (2)$$

where $Emb(\cdot)$ embeds the entity into a C dimensional semantic space. Note that $\mathrm{ReLU}(\cdot)$ is the Rectified Linear Unit operator. $\hat{\mathbf{x}}^0$, $\hat{\mathbf{y}}^0$ denote the initial attention over the subject and object, which are not conditioned on the predicate at all and only use the entities.

Predicate shift modules. Inspired by the message passing protocol in CRF's [44], we design a more general message passing function to transfer information between the two entities. Each message is passed from the subject's estimate to localize the object and vice versa. In practice, we want the message passed from the subject to the object to be different from the one passed from the object back to the subject. So, we learn two asymmetric attention shifts, one that shifts the location from the subject to its estimate of where it thinks the object is and another one that does the inverse from the object to the subject. We denote these shift operations as $\mathrm{Sh}(\cdot)$ and $\mathrm{Sh}^{-1}(\cdot)$, respectively and define them as n convolutions applied in series to the initial estimated assignments:

$$\hat{\mathbf{x}}^0_{shift} = \mathrm{Sh}^{-1}(\hat{\mathbf{y}}^0, P) = \bigcirc_l^n \mathrm{ReLU}(\hat{\mathbf{y}}^0 * F_l^{-1}(P)) \quad (3)$$

$$\hat{\mathbf{y}}^0_{shift} = \mathrm{Sh}(\hat{\mathbf{x}}^0, P) = \bigcirc_l^n \mathrm{ReLU}(\hat{\mathbf{x}}^0 * F_l(P)). \quad (4)$$

where the $\bigcirc_l^n$ implies that we perform the operation n times, each parametrized by $F_l^{-1}(P)$ and $F_l(P)$ which correspond to learned convolution filters for the inverse predicate and the predicate operations respectively. The $*$ operator indicates a convolution with kernels $F_l^{-1}(P)$ and $F_l(P)$ of size $k_l = k$ with c_l channels. We set $c_n = 1$ for the last convolution to ensure that $\hat{\mathbf{x}}^0_{shift}$ and $\hat{\mathbf{y}}^0_{shift}$ have dimension LxLx1. While we do not enforce the two shift operators to be inverses of one another, for most predicates, we empirically find that $\mathrm{Sh}^{-1}(\cdot)$ in fact learns the inverse attention shift of $\mathrm{Sh}(\cdot)$. Note that we do not provide any supervision to our shifts and the model is tasked to learn these shifts to improve its entity localizations. The outputs of these two predicate shift operators is a new estimate attention mask over where the our model expects to find the object, $\hat{\mathbf{y}}^0_{shift}$, conditioned on its initial estimate of the subject, $\hat{\mathbf{x}}^0$ and vice versa from $\hat{\mathbf{y}}^0$ to $\hat{\mathbf{x}}^0_{shift}$.

Each predicate learns its own set of shift and inverse shift functions. And by allowing multiple channels c_l for each set of kernels, our model can formulate shifts as a mixture. For example, carrying might want to focus on the top of the object when the relationship is <person - carrying - phone> while focusing towards the bottom when the relationship is <person - carrying - bag>.

Since we want every image region X_i to pass a message to all other regions Y_j, we enforce that $n > L/k$, i.e. we need a minimum of L/k number of convolutions in series. We arrive at this restriction because the maximum spatial distance that a message needs to travel is $\sqrt{2}L$ and the furtherest image region it can send a message to in each iteration is $\sqrt{2}k$, where L is the image feature size and k is the kernel size of each predicate shift convolution.

Running iterative inference. Once we have these estimates, we can modify our image features with using a element-wise multiplication across the C channels in the feature map. We can then pass it back to the `subject` and `object` attention modules to update their locations:

$$\hat{\mathbf{x}}^1 = \text{Att}(\hat{\mathbf{x}}^0_{shift} \times \boldsymbol{\mu}, S) \quad (5)$$

$$\hat{\mathbf{y}}^1 = \text{Att}(\hat{\mathbf{y}}^0_{shift} \times \boldsymbol{\mu}, O) \quad (6)$$

We can continuously update these locations, conditioned on one another. This amounts to running a maximum a posteriori inference on one entity while using the other entity's previous location. We finally output $\hat{\mathbf{x}}^t$ and $\hat{\mathbf{y}}^t$ where t is a hyper-parameter that determines the number of iterations for which we run inference.

Image Encoding. We extract image features using an ImageNet pre-trained [29] ResNet50's [10] last activation layer of conv4 which outputs a $14 \times 14 \times 512$ dimensional representation and finetune the features. We find that our model performs best with predicate convolution filters with kernel size 5×5 and 10 channels.

Training details. We use RMSProp as our optimization function with an initial learning rate of 0.0001 decaying by 30% when the validation loss does not decrease for 3 consecutive epochs. We train for a total of 30 epochs and embed all of our objects and predicates in a 512 dimensional space.

4. Experiments

We start our experiments by evaluating our model's performance on referring relationships across three datasets, where each dataset provides a unique set of characteristics that complement our experiments. Next, we evaluate how to improve our model in the absence of one of the entities in the input referring relationship. Finally, we conclude by demonstrating how our model can be modularized and used to perform attention saccades through a scene graph.

4.1. Datasets and Baselines

CLEVR. CLEVR is a synthetic dataset generated from scene graphs [12], where the relationships between objects are limited to 4 spatial predicates (`left`, `right`, `front`, `behind`) and 48 distinct entity categories. With over $5M$ relationships where 30% are ambiguous, along with the ease of localizing object categories, this dataset also allows us to explicitly test the effects of our predicate attention shifts without confounding errors from poor image features or noise in real world datasets.

VRD. Visual relationship detection (VRD) is the most widely benchmarked dataset for relationship detection in real world images [23]. It consists of 100 object and 70 predicate categories in $5k$ images, with 60% ambiguous relationships out of a total of $38k$. With a few examples per object and predicate category, this dataset allows us to evaluate how our model performs when starved for data.

Visual Genome. Visual Genome is the largest dataset for visual relationships in real images that is publicly available [17]. It contains $100k$ images with over $2.3M$ relationship instances. We use version 1.4, which focuses on the 100 most common objects with the 70 most common predicate categories. Our experiments on Visual Genome represent a large scale evaluation of our method where 61% of relationships refer to ambiguous entities.

Evaluation Metrics. Recall that the output of our model is localizing the subject and the object of the referring relationship. To evaluate how our model performs, we report the Mean Intersection over Union (IoU), a common metric used in localizing salient parts of an image [4, 5]. This metric measures the average intersection over union between the predicted image regions to those in the ground truth bounding boxes. Next, we report the KL-divergence, which measures the dissimilarity between the two saliency maps and heavily penalizes false positives.

Baseline models. We create three competitive baseline models inspired by related work in entity co-occurrence [8], spatial attention shifts [18] and visual relationship detection [23]. The first model tests how much we can leverage only the entities' **co-occurrence**, without using the predicate. This model simply embeds the `subject` and the `object` and combines them to collectively attend over the image features. The next baseline embeds the entities along with the predicate using a series of dense layers, similar to the vision component in relationship embeddings used in visual relationship detection (**VRD**) [23, 11]. This model has access to the entire relationship when finding the two entities. Finally, the third baseline replaces our learnt predicate shifts with a **spatial shift** that we statistically learn for each predicate in the dataset (see supplementary for details). This final model tests whether our model utilizes both semantic information from images and not just the spatial information from the entities to make predictions.

4.2. Results

Quantitative results. Across all the datasets, we find that the **co-occurrence** model is unable to disambiguate between instances of the same category and only performs well when there is only one instance of that category in an image. The **spatial shift** model does better than the other baselines on CLEVR, where the predicates are spatial and

	Mean IoU ↑						KL divergence ↓					
	CLEVR		VRD		Visual Genome		CLEVR		VRD		Visual Genome	
	S	O	S	O	S	O	S	O	S	O	S	O
Co-occurence [8]	0.691	0.691	0.347	0.389	0.414	0.490	0.839	0.839	2.598	2.307	1.501	1.271
Spatial shift [18]	0.740	0.740	0.320	0.371	0.399	0.469	0.643	0.643	2.612	2.318	1.512	1.293
VRD [23, 11]	0.734	0.732	0.345	0.387	0.417	0.480	1.024	1.014	2.492	2.171	1.483	1.255
SSAS(iter1)	0.742	0.748	0.358	0.398	**0.426**	**0.491**	0.623	0.640	1.936	1.710	1.483	1.235
SSAS(iter2)	0.777	**0.779**	0.365	0.404	0.422	0.487	0.597	**0.595**	1.783	**1.549**	1.458	1.212
SSAS(iter3)	**0.778**	0.778	**0.369**	**0.410**	0.421	0.482	**0.595**	0.596	**1.741**	1.576	**1.457**	**1.205**

Table 1: Results for referring relationships on CLEVR [12], VRD [23] and Visual Genome [17]. We report Mean IoU and KL divergence for the subject and object localizations individually.

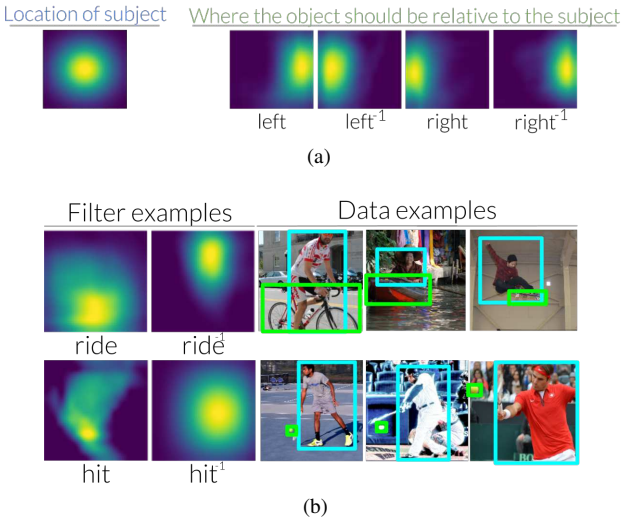

Figure 3: (a) Relative to a subject in the middle of an image, the predicate `left` will shift the attention to the right when using the relationship <`subject-left of-object`> to find the object. Inversely, when using the `object` to find the `subject`, the inverse predicate `left` will shift the attention to the left. We visualize all 70 VRD, 6 CLEVR and 70 Visual Genome predicate and inverse predicate shifts in our supplementary material. (b) We also show that these shifts are intuitive when looking at the dataset that was used to learn them. For example, we find that `ride` usually corresponds to an `object` below the `subject`.

worse on the real world datasets, implying that it is insufficient to model predicates only as spatial shifts. Surprisingly, when evaluating on the CLEVR dataset, we find that **VRD** model does not properly utilize the predicate and leads to marginal gains over the **co-occurrence** models. In comparison, we find that our **SSAS** variants perform better across all metrics. We gain over a 0.32 Mean IoU on CLEVR. This gain however, is smaller on Visual Genome and VRD as these datasets are noisy and incomplete, penalizing our model for making predictions that are not annotated in the datasets. KL, which only penalizes false predictions highlights that our models are more precise than our baselines. Across the different ablations of SSAS, we notice that having more iterations is better; but the performance saturates after 3 iterations because the predicate shifts and the inverse predicate shifts learn near inverse operations of one another.

Interpreting our results. We can interpret the predicate shifts by synthetically initializing the subject to be at the center of an image, as shown in Figure 3(a). When applying the `left` predicate shift, we see that the model has learnt to focus its attention to the right, expecting to find the `object` to the right of the `subject`. Similarly, the inverse predicate shift learns to do nearly the opposite by focusing attention in the other direction. When visualizing these shifts next to the dataset examples in Visual Genome, we see that the shifts represent the biases that exist in the dataset (Figure 3(b)). For example, since most entities that can be `ridden` are below the `subject`, the shifts learn to focus attention down to find the `object` and up to find the `subject`. We also find that that our model learns to encode dataset bias in these shifts. Since the perspective of most images in the training set for `hit` are of people playing tennis or baseball facing left, our model also captures this bias by learning that `hit` should focus attention to the bottom left to find the entity being hit.

Figure 4 shows numerous examples of how our model shifts attention over multiple iterations. We see that generally across all our test cases the subject and object attention modules learn to use the image features to localize all instances initially on iteration 0. For example, in Figure 4(a), all the regions that contain `person` are initially activated. But after the predicate and the inverse predicate shifts, we see that the model learns to move the attention in opposite directions for the predicate `left`. In the second iteration, both the people are uniquely localized in the image. Figure 4(b) clearly shows that we can easily locate all instances of `purple metal cylinders` in the image since it is

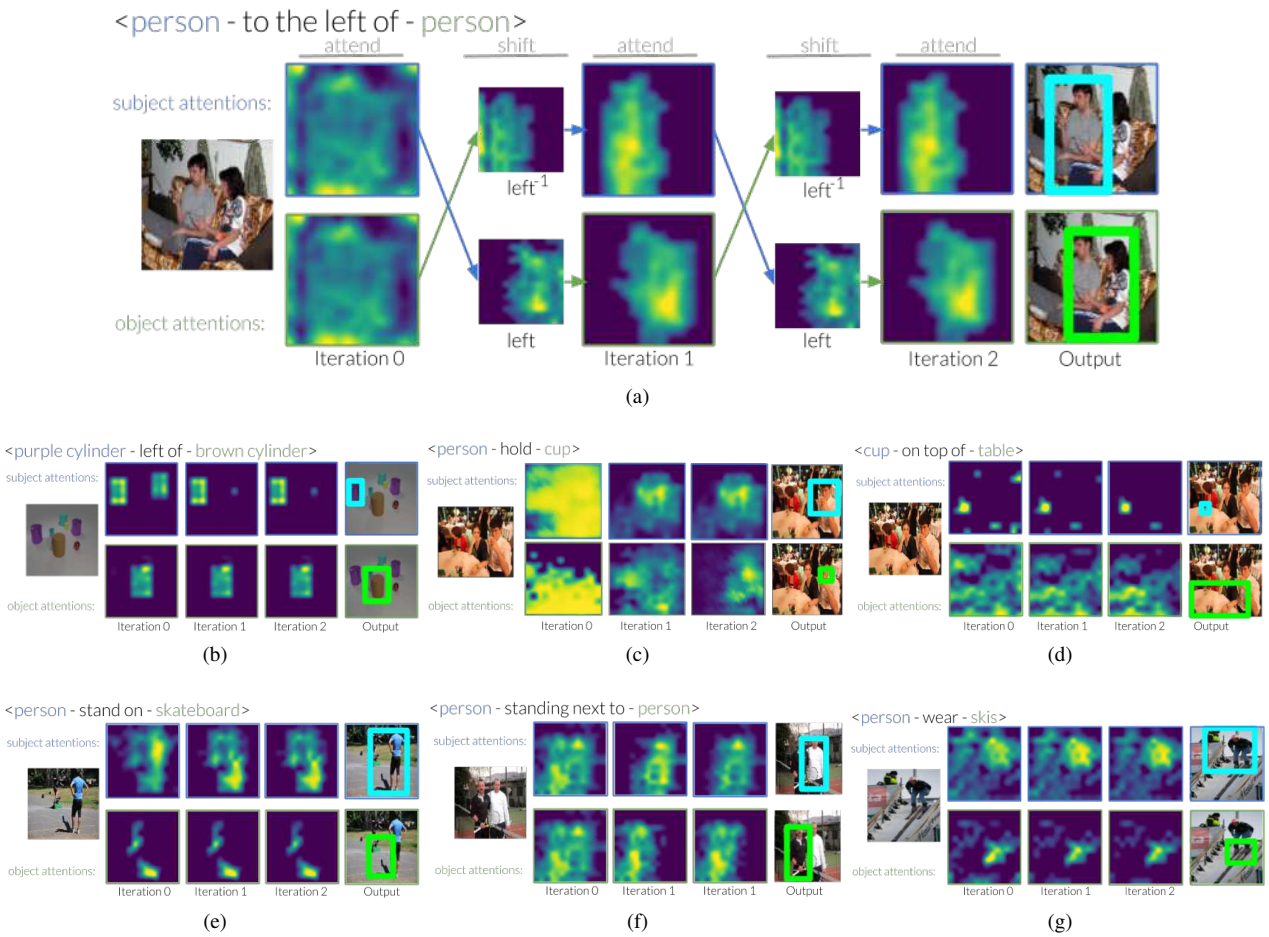

Figure 4: Example visualizations of how attention shifts across multiple iterations from the CLEVR and Visual Genome datasets. On the first iteration the model receives information only about the entities that it is trying to find and hence attempts to localize all instances of those categories. In later iterations, we see that the predicate shifts the attention, allowing our model to disambiguate between different instances of the same category.

easy to detect entities in CLEVR. Our model learns to identify which `purple metal cylinders` we are actually referring to on successive iterations while suppressing the other instance.

In Figure 4(c), even though both the subject and object have multiple instances of `person` and `cup`, we can disambiguate which `person` is actually `holding` the `cup`. For the same image in Figure 4(d), our model is able to distinguish the `cup` being `held` in the previous referring relationship from the one that is `on top of` the `table`. In cases where a referring relationship is not unique, like the example in Figure 4(e), we manage to find all instances that satisfy the relationship we care about. Here, we return both `persons riding` the `skateboards`. Having learnt from the dataset, that most relationships with `stand next to` annotate the `subject` to the left of the `object`, our model emulates this behaviour in Figure 4(f). However, our model does make a fair share of mistakes - for example, in Figure 4(g), it finds both the `persons` and isn't able to distinguish which one is wearing the `skis`.

4.3. Localizing unseen categories

Now that we have evaluated our model, one natural question to ask is how important is it for the model to receive both the entities of the relationship as input? Can it localize the `person` from Figure 1 if we only use `<___ - kicking - ball>` as input? Or can we localize both the subject and the object with only `<___ - kicking - ___>`? We are also interested in taking this task a step further and studying whether we can localize categories that we have never seen before. Previous work has shown that we can localize **seen categories** in novel relationship combinations [23] but we want to know if it is possible to localize **unseen categories**.

We remove all instances of categories like `pants`,

Figure 5: We can decompose our model into its attention and shift modules and stack them to attend over the nodes of a scene graph. Here we demonstrate how our model can be used to start at one node (phone) and traverse a scene graph using the relationships to connect the nodes and localize all the entities in the phrase <phone on the person next to another person wearing a jacket>. A second examples attends over the entities in <hat worn by person to the right of another person above the table>.

	No subject	No object	Only predicate	
	S-IoU	O-IoU	S-IoU	O–IoU
VRD [23]	0.208	0.008	0.024	0.026
SSAS (iter 1)	0.331	0.359	0.332	0.361
SSAS (iter 2)	0.333	0.360	**0.334**	0.361
SSAS (iter 3)	**0.335**	**0.363**	**0.334**	**0.365**

Table 2: Referring relationships results in the absence of the entities under three test conditions: **no subject** where the input is <___ - predicate - object>, **no object** where the input is <subject - predicate - ___> and **only predicate** where the input is <___ - predicate - ___>

hydrant, etc. that are not in ImageNet (CNN$(\cdot)$ was pretrained on ImageNet) from our training set and attempt to localize these novel categories using their relationships. We do not make any changes to our model but alter the training script to randomly (we use a drop rate of 0.3) mask out the subject or object or both in the referring relationships during each iteration. The model learns to attend over general object categories when the entities are masked out. We find that we can in fact localize these missing entities, even if they are from unseen categories. We report results for this experiment on the VRD dataset in Table 2.

4.4. Attention saccades through a scene graph

A ramification of our model design results in its modularity — the attention and shift modules expect inputs and produce outputs that are image features of shape $L \times L \times C$. We can decompose these modules and stack them like Lego blocks, allowing us to perform more complicated tasks. One particularly interesting extension to referring relationships is attention saccades [36]. Instead of using a single relationship as input, we can extend our model to take an entire scene graph as input. Figure 5 demonstrates how we can iterate between the attention and shift modules to traverse a scene graph. We can start from the phone and can localize the jacket worn by the "woman on the right of the man using the phone". A scene graph traversal can be evaluated by decomposing the graph into a series of relationships. We do not quantitatively evaluate these saccades here, as its evaluations are already captured by the referring relationships in the graph.

5. Conclusion

We introduced the task of referring relationships, where our model utilizes visual relationships to disambiguate between instances of the same category. Our model learns to iteratively use predicates as an attention shift between the two entities in a relationship. It updates its belief of where the subject and object are by conditioning its predictions on the previous location estimate of the object and subject, respectively. We show improvements on CLEVR, VRD and Visual Genome datasets. We also demonstrate that our model produces interpretable predicate shifts, allowing us to verify that the model is in fact learning to shift attention. We even showcase how our model can be used to localize completely unseen categories by relying on partial referring relationships and how it can be extended to perform attention saccades on scene graphs. Improvements in referring relationships could pave the way for vision algorithms to detect unseen entities and learn to grow its understanding of the visual world.

Acknowledgements. Toyota Research Center (TRI) provided funds to assist the authors with their research but this article solely reflects the opinions and conclusions of its authors and not TRI or any other Toyota entity. We thank John Emmons, Justin Johnson and Yuke Zhu for their helpful comments.

References

[1] P. Anderson, B. Fernando, M. Johnson, and S. Gould. Spice: Semantic propositional image caption evaluation. In *European Conference on Computer Vision*, pages 382–398. Springer, 2016. 2

[2] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014. 3

[3] A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, and O. Yakhnenko. Translating embeddings for modeling multi-relational data. In *Advances in neural information processing systems*, pages 2787–2795, 2013. 3

[4] Z. Bylinskii, T. Judd, A. Borji, L. Itti, F. Durand, A. Oliva, and A. Torralba. Mit saliency benchmark, 2015. 5

[5] Z. Bylinskii, A. Recasens, A. Borji, A. Oliva, A. Torralba, and F. Durand. Where should saliency models look next? In *European Conference on Computer Vision*, pages 809–824. Springer, 2016. 5

[6] T. Dettmers, P. Minervini, P. Stenetorp, and S. Riedel. Convolutional 2d knowledge graph embeddings. *arXiv preprint arXiv:1707.01476*, 2017. 3

[7] P. Felzenszwalb, D. McAllester, and D. Ramanan. A discriminatively trained, multiscale, deformable part model. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008. 3

[8] C. Galleguillos, A. Rabinovich, and S. Belongie. 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, 5, 6

[9] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, and G. E. Dahl. Neural message passing for quantum chemistry. *arXiv preprint arXiv:1704.01212*, 2017. 2, 3

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

[11] R. Hu, M. Rohrbach, J. Andreas, T. Darrell, and K. Saenko. Modeling relationships in referential expressions with compositional modular networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4418–4427. IEEE, 2017. 1, 2, 5, 6

[12] J. Johnson, B. Hariharan, L. van der Maaten, L. Fei-Fei, C. L. Zitnick, and R. Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. *arXiv preprint arXiv:1612.06890*, 2016. 2, 5, 6

[13] J. Johnson, B. Hariharan, L. van der Maaten, J. Hoffman, L. Fei-Fei, C. L. Zitnick, and R. Girshick. Inferring and executing programs for visual reasoning. *arXiv preprint arXiv:1705.03633*, 2017. 3

[14] 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. 2

[15] T. N. Kipf and M. Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016. 3

[16] P. Krähenbühl and V. Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In *Advances in neural information processing systems*, pages 109–117, 2011. 4

[17] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73, 2017. 2, 5, 6

[18] D. LaBerge, R. L. Carlson, J. K. Williams, and B. G. Bunney. Shifting attention in visual space: tests of moving-spotlight models versus an activity-distribution model. *Journal of Experimental Psychology: Human Perception and Performance*, 23(5):1380, 1997. 2, 3, 5, 6

[19] L. Ladicky, C. Russell, P. Kohli, and P. H. Torr. Graph cut based inference with co-occurrence statistics. In *European Conference on Computer Vision*, pages 239–253. Springer, 2010. 2

[20] Y. Li, W. Ouyang, and X. Wang. Vip-cnn: A visual phrase reasoning convolutional neural network for visual relationship detection. *arXiv preprint arXiv:1702.07191*, 2017. 2

[21] X. Liang, L. Lee, and E. P. Xing. Deep variation-structured reinforcement learning for visual relationship and attribute detection. *arXiv preprint arXiv:1703.03054*, 2017. 2

[22] Y. Lin, Z. Liu, M. Sun, Y. Liu, and X. Zhu. Learning entity and relation embeddings for knowledge graph completion. In *AAAI*, pages 2181–2187, 2015. 3

[23] C. Lu, R. Krishna, M. Bernstein, and L. Fei-Fei. Visual relationship detection with language priors. In *European Conference on Computer Vision*, pages 852–869. Springer, 2016. 1, 2, 3, 4, 5, 6, 7, 8

[24] J. Mao, J. Huang, A. Toshev, O. Camburu, A. L. Yuille, and K. Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20, 2016. 1, 2

[25] T. Mensink, E. Gavves, and C. G. Snoek. Costa: Co-occurrence statistics for zero-shot classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2441–2448, 2014. 2

[26] B. A. Plummer, A. Mallya, C. M. Cervantes, J. Hockenmaier, and S. Lazebnik. Phrase localization and visual relationship detection with comprehensive linguistic cues. *arXiv preprint arXiv:1611.06641*, 2016. 2

[27] 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. 2

[28] A. Rohrbach, M. Rohrbach, R. Hu, T. Darrell, and B. Schiele. Grounding of textual phrases in images by reconstruction. In *European Conference on Computer Vision*, pages 817–834. Springer, 2016. 1

[29] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015. 2, 5

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

[31] R. Salakhutdinov, A. Torralba, and J. Tenenbaum. Learning to share visual appearance for multiclass object detection. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1481–1488. IEEE, 2011. 2

[32] A. Santoro, D. Raposo, D. G. Barrett, M. Malinowski, R. Pascanu, P. Battaglia, and T. Lillicrap. A simple neural network module for relational reasoning. *arXiv preprint arXiv:1706.01427*, 2017. 2

[33] S. Schuster, R. Krishna, A. Chang, L. Fei-Fei, and C. D. Manning. Generating semantically precise scene graphs from textual descriptions for improved image retrieval. In *Proceedings of the fourth workshop on vision and language*, volume 2, 2015. 2

[34] M. Shridhar and D. Hsu. Grounding spatio-semantic referring expressions for human-robot interaction. *arXiv preprint arXiv:1707.05720*, 2017. 1, 2

[35] G. Sperling and E. Weichselgartner. Episodic theory of the dynamics of spatial attention. *Psychological review*, 102(3):503, 1995. 2, 3

[36] A. Torralba, A. Oliva, M. S. Castelhano, and J. M. Henderson. Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search. *Psychological review*, 113(4):766, 2006. 2, 8

[37] D. Xu, Y. Zhu, C. B. Choy, and L. Fei-Fei. Scene graph generation by iterative message passing. *arXiv preprint arXiv:1701.02426*, 2017. 2

[38] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning*, pages 2048–2057, 2015. 3

[39] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola. Stacked attention networks for image question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 21–29, 2016. 3

[40] B. Yao and L. Fei-Fei. 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

[41] L. Yu, P. Poirson, S. Yang, A. C. Berg, and T. L. Berg. Modeling context in referring expressions. In *European Conference on Computer Vision*, pages 69–85. Springer, 2016. 1, 2, 3

[42] L. Yu, H. Tan, M. Bansal, and T. L. Berg. A joint speaker-listener-reinforcer model for referring expressions. *arXiv preprint arXiv:1612.09542*, 2016. 1

[43] H. Zhang, Z. Kyaw, S.-F. Chang, and T.-S. Chua. Visual translation embedding network for visual relation detection. *arXiv preprint arXiv:1702.08319*, 2017. 3

[44] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. H. Torr. Conditional random fields as recurrent neural networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1529–1537, 2015. 4