

# Situation Recognition: Visual Semantic Role Labeling for Image Understanding

Mark Yatskar<sup>1</sup>, Luke Zettlemoyer<sup>1</sup>, Ali Farhadi<sup>1,2</sup>

<sup>1</sup>Computer Science & Engineering, University of Washington, Seattle, WA

<sup>2</sup>Allen Institute for Artificial Intelligence (AI2), Seattle, WA

[my89, lsz, ali]@cs.washington.edu







					
CLIPPING		JUMPING		SPRAYING	
ROLE	VALUE	ROLE	VALUE	ROLE	VALUE
AGENT	MAN	AGENT	BOY	AGENT	MAN
SOURCE	SHEEP	SOURCE	CLIFF	SOURCE	SPRAY CAN
TOOL	SHEARS	OBSTACLE	-	OBSTACLE	WATER
ITEM	WOOL	DESTINATION	WATER	DESTINATION	ICEBERG
PLACE	FIELD	PLACE	LAKE	PLACE	OUTDOOR
ROLE	VALUE	ROLE	VALUE	ROLE	VALUE
AGENT	VET	AGENT	BEAR	AGENT	FIREMAN
SOURCE	DOG	SOURCE	ICEBERG	SOURCE	HOSE
TOOL	CLIPPER	OBSTACLE	WATER	SUBSTANCE	PAINT
ITEM	CLAW	DESTINATION	ICEBERG	DESTINATION	WALL
PLACE	ROOM	PLACE	OUTDOOR	PLACE	ALLEYWAY
ROLE	VALUE	ROLE	VALUE	ROLE	VALUE
AGENT	FIREMAN	AGENT	BOY	AGENT	MAN
SOURCE	HOSE	SOURCE	CLIFF	SOURCE	SPRAY CAN
SUBSTANCE	WATER	OBSTACLE	-	OBSTACLE	WATER
DESTINATION	FIRE	DESTINATION	WATER	DESTINATION	ICEBERG
PLACE	OUTSIDE	PLACE	LAKE	PLACE	OUTDOOR

Figure 1. Six images that depict situations where actors, objects, substances, and locations play roles in an activity. Below each image is a *realized frame* that summarizes the situation: the left columns (blue) list activity-specific roles (derived from FrameNet, a broad coverage verb lexicon) while the right columns (green) list values (from ImageNet) for each role. Three different activities are shown, highlighting that visual properties can vary widely between role values (e.g., clipping a sheep’s wool looks very different from clipping a dog’s nails).

## Abstract

*This paper introduces situation recognition, the problem of producing a concise summary of the situation an image depicts including: (1) the main activity (e.g., clipping), (2) the participating actors, objects, substances, and locations (e.g., man, shears, sheep, wool, and field) and most importantly (3) the roles these participants play in the activity (e.g., the man is clipping, the shears are his tool, the wool is being clipped from the sheep, and the clipping is in a field). We use FrameNet, a verb and role lexicon developed by linguists, to define a large space of possible situations and collect a large-scale dataset containing over 500 activities, 1,700 roles, 11,000 objects, 125,000 images, and 200,000 unique situations. We also introduce structured prediction baselines and show that, in activity-centric images, situation-driven prediction of objects and activities outperforms independent object and activity recognition.*

## 1. Introduction

When we look at an image, we instantly and effortlessly recognize not only what is happening (e.g., clipping) but who and what is involved (e.g., a person, shears, a sheep,

wool) and how these entities relate to each other, *i.e.* the *roles* that they play (e.g., the person does the clipping, the shears are the clipping tool, and the wool is being clipped from the sheep). In this paper, we argue for explicitly encoding such semantic roles, a key missing ingredient in current paradigms of recognition, in image understanding. We introduce *situation recognition*, a problem that involves predicting activities along with actors, objects, substances, and locations and how these pieces fit together (semantic roles). For example, the leftmost table in Figure 1 shows one such representation: a situation where a man (agent) is clipping (activity) wool (item) from a sheep (source) using shears (tool) in a field (place).

Situation recognition generalizes activity recognition and human-object interaction, using the assignment of roles to define how actors, objects, substances, and locations participate in activities. For example, Figure 1 has image pairs that depict the same overall activity but look very different when the participating entities change for the different roles. Previous work has presented models for some aspects of a complete situation, including activity scene models [35] and models of very specific activities paired with a few prototypical objects, such as playing a musical in-

strument [48]. However, our formulation provides a more complete representation of the different roles that each of the participants can play, and allows us to scale to hundreds of different activities. In essence, we are building representations that support the understanding not just of “What is happening?” but also “Who is doing it?” (the *agent* role), “What are they doing it to?” (*patient*), “What are they doing it with?” (*tool*), “Where did it start?” (*source*), and so on, as appropriate for each activity.

It is difficult to know a priori what roles entities can play in each activity. However, we can draw inspiration from the way verbs are used in the English language by building on FrameNet [14], a linguist-authored verb lexicon. FrameNet pairs every verb with a *frame*, which specifies a set of *semantic roles*. Semantic roles categorize how objects can participate in the activity described by a verb. For example, the two rightmost images of Figure 1 show frames for spraying, which includes semantic roles such as *agent* and *destination*. Such frames have been used to build semantic parsers that match verbs to their arguments in English sentences, for example see [3]. However, here we instead use them to define the space of possible situations, much like how WordNet [13] was used to define ImageNet [41] object classes. For each frame, the verb defines an activity label, and the semantic roles specify how WordNet entities participate in the activity. For example, Figure 1 shows situations where the FrameNet verb *spraying* has a semantic role *tool* that is filled with WordNet synsets such as *spray can* or *hose*.

To demonstrate the generality of the situation recognition task, we introduce *imSitu*, a collection of over 125,000 images depicting 200,000 distinct situations. Each situation includes one of 500 possible activities and values for up to 6 activity-specific roles (3.5 on average and 1,700 unique roles in total with 190 types). The images were gathered from Google image search with query expansion techniques and labeled with complete situations on Amazon Mechanical Turk. The annotators specified one of 80,000 possible WordNet synsets for each role, providing over 11,000 unique values for this image collection. In addition to being large scale, this data is also high quality. For example, even though the space of possible values is very large, 2 out of 3 annotators provided the same synset for over 75% of roles. Sections 4 and 5 provide the full details of the data collection and statistics.

To support future work on the *imSitu* data, we provide results for a baseline model — a Conditional Random Field (CRF) which includes CNN [43] features (fine tuned by backpropagating the CRF error). This approach significantly outperforms a 5000-way classifier that predicts one of the 10 most frequent situations per verb. The CRF achieves 32.3% top-1 and 58.9% top-5 accuracy for activity prediction and predicts entire situations correctly 14.2% of

the time. When compared to independent models trained on the same activity-centric data, the approach improves top-1 accuracy for object recognition by 8.6% and top-1 activity recognition by 1.2%, demonstrating that the model benefits significantly from the context that is provided by jointly predicting the full situation. These results suggest that situation recognition with the *imSitu* dataset has the potential to become a strong benchmark for the study of objects, activities, and their interactions through semantic roles.

## 2. Formal Task Definition

In situation recognition, we assume discrete sets of verbs  $V$ , nouns  $N$ , and frames  $F$ . Each frame  $f \in F$  is paired with a discrete set of semantic roles  $E_f$ . For example, Figure 1 shows six different situations, representing the verbs *clipping*, *jumping*, and *spraying*. While some semantic roles, e.g. *agent*, are shared across all three frames, others (e.g., *tool*) only appear for some. Additionally, each semantic role  $e \in E_f$  is paired with a noun value  $n_e \in N \cup \{\emptyset\}$ , where  $\emptyset$  indicates the value is either not known or does not apply. For example, in the first image in Figure 1, the semantic role *item* takes the value *wool*. In this paper, the verb set  $V$  and frame set  $F$  are derived from FrameNet, while the noun set  $N$  is drawn from WordNet. We refer to the set of pairs of semantic roles and their values as a realized frame,  $R_f = \{(e, n_e) : e \in E_f\}$ . In the third image of Figure 1,  $R_f = \{(\text{agent}, \text{boy}), (\text{source}, \text{cliff}), (\text{obstacle}, \emptyset), (\text{destination}, \text{water}), (\text{place}, \text{lake})\}$ . Finally, a realized frame is valid if and only if each value  $e \in E_f$  is assigned exactly one noun  $n_e$ .

Now, given an image, our task is to predict a situation,  $S = (v, R_f)$ , specified by a verb  $v \in V$  and a valid realized frame  $R_f$ . For example, in the last image of Figure 1, the predicted situation is  $S = (\text{spraying}, \{(\text{agent}, \text{fireman}), (\text{source}, \text{hose}), (\text{substance}, \text{water}), (\text{destination}, \text{fire}), (\text{place}, \text{outside})\})$ .

## 3. Related Work

Activity recognition in still images has been widely studied [21], and it is generally accepted that objects and scenes are important for recognition [31]. These intuitions are often built directly into datasets by framing activity recognition as a discrete classification problem, with a small set of multi-word category labels that combine a verb with a scene or object [4, 10, 22, 44, 48, 49]. Although recent work has scaled the number of classes [30], they are still hand selected and it can be difficult to know what should be included in the set. For example, while “cutting-vegetables” is a category in Stanford-40, many others possibilities, like “cutting-grass” or the more generic “cutting,” are missing (similar examples can be found in all current activity recognition datasets). In contrast, our task formulation uses lin-

guistic resources to define a very large and significantly more comprehensive space of possible situations.

Many methods have been proposed for modeling semantic context in activity recognition [6]. Our approach is most closely related to work that models object co-occurrence [38] and uses graphical models to combine many sources of contextual information [17, 12]. Actions have been a particularly fruitful source of context [35], especially when combined with pose to create human-object interactions [33, 47]. However, we present the first approach to define how multiple objects participate in a single activity, allowing us to systematically recover activity-specific facts such as “Who is doing it?” (the *agent* role), “What are they doing it to?” (*patient*), etc.

There is also significant related work in the intersection of language and vision. WordNet [13] is used to define ImageNet [5] classes, much like how we use FrameNet [14] to define our situation space. Recent work has also explored other areas of cross pollination, including video recognition [20], cross modal mappings [42, 29, 16], coreference [8, 28], and affordances [52]. In particular, sentence generation is closely related and has received significant attention [50, 26, 2, 11, 46, 34, 24, 36, 32]. Our situations are inspired by semantic role labeling models [3, 27], which are designed to provide a type of shallow semantics for verbs; in essence, our frames correspond to simple declarative sentences. However, we sidestep the evaluation challenges that come with generating sentences [45, 7], while also providing visual evidence for verbs that should aid captioning. At least partially motivated by the same concerns, there are recent efforts to formulate Visual Question Answering (VQA) tasks [1, 39, 51, 18, 9], where the system must answer questions like “What is the person using to cut the grass?” In a pilot study on a VQA dataset [1], we found that up to 20% of questions ask about a semantic role, suggesting that situation recognition could be beneficial.

Finally, situation recognition is related to two parallel efforts to define visual semantic role labeling tasks. Both provide instance-level information with bounding regions for objects [23, 40]. We instead focus on classification, annotate an order of magnitude more images and are the first to consider more than two semantic roles.

## 4. Dataset Collection

We introduce imSitu, a dataset of images labeled with situations. Our annotation approach is scalable, the image labeling is done on Mechanical Turk and covers over 500 verbs with 125,000 images, and is relatively affordable, annotation cost approximately \$80 per verb.

### 4.1. Filtering and Labeling FrameNet

FrameNet is a rich resource that pairs verbs with frames and semantic roles. It is designed to cover, as much as pos-

sible, all English verbs and all roles they can take, not just those that can be visually recognized in an image. For example, it would include verbs such as *attempt* with roles such as *goal* that take other verbs as arguments. To define our recognition task, we manually filtered FrameNet to find verbs and roles that could be reliably recognized in images, and provide English labels for use in the crowdsourcing interface. This was done by a small set of trusted annotators.

**Finding Visual Verbs and Roles** We gathered 9683 candidate verbs and asked annotators to determine if they could be reliably recognized in images, and, if so, to provide a support image.<sup>1</sup> Verbs that were not recognizable generally fell into one of a few classes, including: (a) abstract, such as “presuming,” (b) representational, such as “thinking,” where we could find a supporting image evocative of the verb but did not depict it literally happening (c) technical, including “blanching,” where crowd workers were unlikely to know the word’s meaning, or otherwise just (d) hard, including “insufflating,” where the annotator does not know the word or what it would look like. Annotators were first calibrated to confirm they understood these categories and confusing cases were publicly discussed. In total, 1053 verbs (10.9%) were marked as visually recognizable. To find visual roles, annotators were shown visual verbs and their example images and asked to select the subset of visually recognizable semantic roles, a generally easier task.

**Labeling Verbs and Roles** To support later crowd sourcing, the annotators also provided simple English descriptions of the visual verbs and roles. They wrote a single sentence that summarizes all of the roles for each verb. For example, for the verb *clipping* in Figure 1, the sentence would be “An AGENT clips an ITEM from a SOURCE using a TOOL in a PLACE.” This sentence was shown to crowd workers to define the roles that each verb supports.

**Example Creation** Finally, to help crowd workers understand how to produce situation annotations, a few example image labels were produced for each verb. Five computer science undergraduates read definitions for all 1053 candidate verbs and retrieved three images that correspond to each verb from Google Image Search. If the annotators were unable to find such images, the verb was removed. Overall, 580 verbs passed this filtering stage.

### 4.2. Image Annotation

The final image annotations were gathered on Amazon Mechanical Turk in a two-stage process, that involved first filtering automatically collected images and then filling in the role values for target frames.

<sup>1</sup>We extended the nearly 5,000 verbs in FrameNet to include additional verbs from PropBank [27], a closely related verb lexicon, that were mapped to FrameNet as part of the SemLink project [37].





Figure 2. A word cloud of verbs in imSitu where larger words have a larger rate of unseen value-role combinations. Verbs with a low rate, e.g. “flossing” (0.7%) have specific meaning as compared to verbs such as “putting” (15.5%) or “biting” (7.7%).

**Candidate Image Filtering** Candidate images were retrieved by searching for phrases related to a target activity in Google Image Search. Phrases were mined from a subset of Google Syntactic N-Grams [19] that focuses on verb-argument structure. The phrases we extracted contain the target verb and include all descendants of the verb in a syntactic parse. We selected 450 such phrases, picking the most frequent 150 that contain “n-subj,” “d-obj,” or “p-obj” dependencies. For example, “cutting” would have the p-obj “scissors.” Using dependencies guarantees that the queried words occur in different syntactic positions relative to the target verb. We retrieved 200 full-color medium-sized images that pass safe search and consider all returned images as candidates. Workers were instructed to select images that contain the desired activity and (1) are not modified or computer generated and (2) contain at least some part of the main entity doing the action in the image.

**Value Filling** Selected images were next presented for value filling. Workers were shown a definition of the target verb, a sentence summarizing the semantic roles associated with verb and example images of realized frames for that verb. They were asked to chose a category from an auto-complete drop-down menu, that also presents synset definitions, to fill slots; to select the most specific WordNet synset, and if more than one could apply, select the most relevant. For groups, they were asked to either find a word that refers to the group (for example, “people,” “couple”) or simply use the singular (“person”). They were required to annotate at least one value per image and not to fill in values that could not be reasonably inferred from the image.

### 4.3. Diversity and Coverage

The goal of imSitu is to include as many verbs as possible and have samples for all unique combinations of semantic roles and values. This is challenging because situations are structured and there can be a combinatorial number of possible realized frames. We adopted a dynamic strategy to



Figure 3. A word cloud of verbs in imSitu where larger words have a larger true positive rate for images retrieved from Google Image Search. Verbs with low rates, *i.e.* “fanning” (1%), were cost prohibitive to annotate. For all verbs, the average rate was 6.6%.

increase diversity while not wasting money on verbs where we have already seen most combinations. First, candidate images from Google Image Search were presented for filtering by uniformly drawing images from query phrases, thus maximizing the diversity of types of images. 200 images were annotated in this way with full structures, providing a lower bound on the number of images per verb in imSitu. Then, we dynamically decided whether to continue to collect more annotations.

The rate at which unseen combinations occur can be approximated by splitting the data into a train and test set and computing how often a value appears in a semantic role in the test set but never appeared in train set. We refer to this as the out of vocabulary (OOV) rate of a verb, and compute it by averaging 1000 random splits of the data. Figure 2 visualizes the current OOV rate for a sample of verbs currently in imSitu. If during the collection process the OOV rate of verb was greater than 5%, we continued to collect images, up to a maximum of 400 images. While for some verbs this significantly improved the OOV rate, other verbs will always have a high rate. For example, despite collecting 400 images of the verb “making” and “putting,” both have an OOV rate of 15%. This is a fundamental challenge in situation recognition. On the other hand, “baptizing” has an OOV of zero with just 200 image samples. The final global OOV rate in imSitu is 3.5%.

### 4.4. Cost

During the collection process, every verb had a hard constraint of costing no more than \$120 and was discontinued when it exceeded this amount. The largest contributor to the cost of collecting imSitu was the true positive rate of candidates retrieved from Google Image Search. Figure 3 shows the true positive rates for a sample of verbs currently in imSitu. Over 25% of verbs were cost prohibitive to collect directly from Google Image Search results. In cases when we were able to collect at least 50 images but exceeded a cost

verbs	504
images	126,102
realized frames / image	3
total annotations	1,481,851
unique entities ( $\geq 3$ )	11,538 (6794)
semantic roles / verb (range)	3.55 (1 - 6)
semantic roles (types)	1788 (190)
images / verb (range)	250.2 (200 - 400)
unique realized frames ( $\geq 3$ )	205,095 (21,505)
out of vocabulary rate (range)	3.5% ( 0% - 15.8% )
train / dev / test	75,702 / 25,200 / 25,200

Table 1. Summary statistics of imSitu.

	Majority	1-link	2-link	3-link
all Roles	76.8	81.5	84.8	86.5
w/o Place	81.5	84.6	88.2	89.9

Table 2. Agreement statistics for situation role annotations in imSitu, with and without the Place role. Majority means that at least 2 of 3 Turker annotations agree. N-link means that a majority agree under the relaxed criteria of two synsets matching if they are within N links of each other in the WordNet hierarchy.

threshold before collecting 200 images, we made a second effort. A new set of queries for Google Image Search was constructed by pairing the verb with a noun that occurred in an annotated frame. The returned images were used to reseed the filter phase of our annotation. This second round allowed us to reduce the percentage of failed verbs to 13%. Overall, failed verbs contributed \$7 to the cost of annotating each verb.

## 5. Dataset Statistics

Table 1 provides summary statistics about imSitu, collected as described in the last section. In this section, we summarize the overall annotator agreement and highlight several interesting aspects of the data.

**Agreement** Quality control at scale is challenging. We used an automatic algorithm that discards annotations from workers that it estimates to be unreliable. The details are described in the supplementary material.

All images were annotated by three crowd workers. We measure agreement by comparing the values that workers annotated for semantic roles. We say that two semantic role annotations on a single image agree when they indicate the same WordNet synset (or  $\emptyset$ ). Furthermore, we compute a relaxed version of this criterion, allowing two annotations to match if the synsets are within 1, 2 or 3 links in the WordNet hierarchy. As a point of reference, the following synsets are all 3 links away from each other: “musical instrument” and “trumpet,” “child” and “little girl,” and “girl” and “person.” Table 2 summarizes agreement in imSitu.

While the agreement numbers are very high, especially considering Turkers can select one of 80,000 values for each semantic role, there are systematic sources of ambiguity. *Place*, a semantic role present in all frames, is highly am-

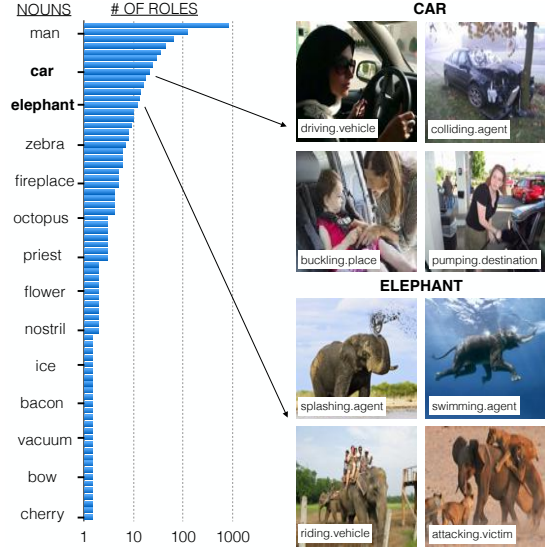


Figure 4. The number of semantic roles a noun can participate in, on a log-scale. 62% of nouns in imSitu appear with more than one semantic role. The most frequent noun, “man” appears in 44.6% of the roles. On the right are the different roles the nouns “car” and “elephant” participate in. Some roles can define particular viewpoints, such the role “place” being assigned “car” commonly indicates the interior view of the car.

biguous because it can be identified in three ways: a close interacting object (e.g., reading at a “desk”), an overall location type (e.g., reading in an “office”) or a coarse identifier (e.g., reading “inside”). Table 2 demonstrates that *place* is indeed a major contributor to disagreement, accounting for over 25% cases where workers failed to produce a majority. This type of disagreement provides a number of alternative correct answers. Other sources of disagreement are described in the supplementary material.

**Entity-Role Relations** Figure 4 shows a uniform sample of nouns and the number of semantic roles they participate in. As expected there is a large variance; for example, “man” can take up to 798 roles while “basin” only takes 1 role. We also compute the inverse of these statistics: the number of nouns that a role can take, as shown in Figure 5.

**Entity-Verb Relations** Figure 6 shows the number of entities a sample of verbs can take. As expected, less structured verbs like “putting” have 653 entities and heavily structured verbs like “flossing” only take 42 nouns.

## 6. Structured Prediction of Frames

Our CRF for predicting a situation,  $S = (v, R_f)$ , given an image  $i$ , decomposes over the verb  $v$  and semantic role-value pairs  $(e, n_e)$  in the realized frame  $R_f = \{(e, n_e) : e \in E_f\}$ . The CRF parameters  $\theta$  can be trained directly from our situation-labeled data. The full distribution, with poten-

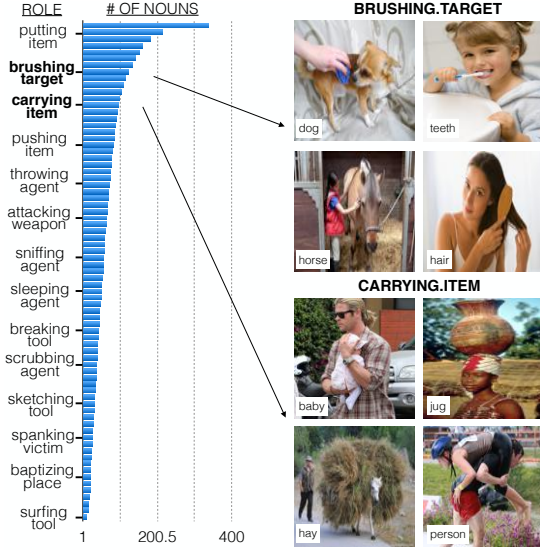


Figure 5. On the left, the number of nouns that can participate in a sample of semantic roles (not all labeled). On average 64.7 nouns appear per role. Some roles, such as the “tool” of “surfing” take very few values, indicating the majority of the information about the situation is indicated by the verb. On the right are examples of nouns that fill the “target” of “brushing” (the thing being brushed) and the “item” of carrying (the thing being carried), showing significant visual variation when the values are changed.

tials for verbs  $\psi_v$  and semantic roles  $\psi_e$  takes the form:

$$p(S|i; \theta) \propto \psi_v(v, i; \theta) \prod_{(e, n_e) \in R_f} \psi_e(v, e, n_e, i; \theta) \quad (1)$$

Computing the normalization is efficient: we can enumerate all valid verb-semantic role pairs and then for all pairs sum all possible semantic role values.

Each potential in the CRF is log linear:

$$\psi_v(v, i; \theta) = e^{\phi_v(v, i) \theta} \quad (2)$$

$$\psi_e(v, e, n_e, i; \theta) = e^{\phi_e(v, e, n_e, i) \theta} \quad (3)$$

where  $\phi_e$  and  $\phi_v$  encode scores from the output of a CNN. To learn this model, we assume that for an image  $i$  in dataset  $D$  there can, in general, be a set  $A_i$  of possible ground truth situations. We optimize the log-likelihood of observing at least one situation  $S \in A_i$ :

$$\sum_{i \in D} \log \left( 1 - \prod_{S \in A_i} (1 - p(S|i; \theta)) \right) \quad (4)$$

**CRF Features** In Equation 2 and 3 we introduce two feature functions that are implemented by adapting a neural network pretrained on the ImageNet Challenge [41]. We use VGG Large Network [43] in Caffe [25] with the final layers reduced to dimensionality 1024. The output of VGG

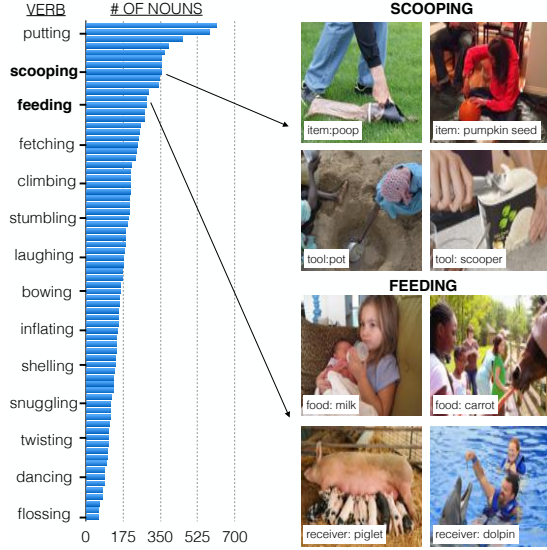


Figure 6. On the left, the number of nouns that appear with a sample of verbs (not all labeled). Some verbs (e.g. putting or scooping) require the ability to predict hundreds of noun values, while others (e.g. flossing) can only happen in a few canonical ways. On average, 199 nouns occur with a verb. On the right are example nouns for “scooping” and “feeding” and the roles they play.

is used as the input to a fully connected layer which predicts potential values in our CRF, similar to neural networks used for semantic role labeling in sentences [15]. At training time, we optimize Equation 4 with stochastic gradient ascent using a batch size of 192. We fine tune all layers of VGG for 30 epochs and reduce the initial learning rate of 1e-5 by a factor of ten for every ten epochs.

## 7. Experiments

We present the first results for situation prediction in im-Situ and also compare performance to baselines that independently recognize activities and objects.

### 7.1. Situation Recognition

**Metrics** We measure accuracy for different components of predicted situations. Because the evaluation data has situations provided by multiple annotators, we consider verb predictions (verb) and semantic role-value pair predictions (value) correct if they match any of the annotations. A realized frame is correct if it strictly matches all semantic role-value pairs provided by a single annotation (value-full) or if each pair matches at least one annotation (value-any). We also report accuracy with ground truth verbs.

**Systems** In addition to the CRF model described in Section 6, we also present a simple discrete classification baseline. The classifier selects one of the 10 most frequent realized frames for each verb seen in the training data, producing a 5040-class problem. For training, each realized



		top-1 predicted verb				top-5 predicted verbs				ground truth verbs		
		verb	value	value-any	value-full	verb	value	value-all	value-full	value	value-all	value-full
dev	Discrete Classifier	26.4	4.0	0.4	0.2	51.1	7.8	0.6	0.4	14.4	0.9	0.6
	CRF	32.2	24.6	14.3	11.2	58.6	42.7	22.7	17.5	65.9	29.5	22.3
test	Discrete Classifier	26.8	4.1	0.3	0.2	51.2	7.8	0.5	0.4	14.4	0.8	0.6
	CRF	32.3	24.6	14.2	11.2	58.9	42.8	22.5	17.5	65.7	29.0	22.0

Table 3. Situation prediction results in imSitu. Structured prediction outperforms classification of ten most common situations per activity.

frame is assigned as a positive example to the classifier output with the fewest number of differences. The classifier uses the same VGG features and fine tuning procedure as the CRF but with an initial learning rate of  $1e-3$ .

**Quantitative Results** Table 3 summarizes our experiments on the imSitu development set. We also ran these experiments once on the imSitu test which confirms our development results. Overall, the CRF outperforms the discrete classifier by large margins. Verb accuracy is 32.5% and rises to 59% in the top-5. We can isolate the performance of assigning values to semantic roles by considering prediction accuracy given ground truth verbs. The discrete classifier is significantly worse in this context at value and full prediction because it cannot assign new combinations of entities to roles at test time.

**Qualitative Results** Figure 7 shows a random selection of predictions from the CRF model on the development images where it predicted the correct verb. Over two thirds of the cases are correct or have only one incorrect role assignment. Furthermore, many of the errors are actually somewhat plausible. For example, the pole vaulter in the bottom right image is going over a horizontal pole. Other cases show similar reasonable errors, including confusing a cow with a horse, in the image second from the top and right.

## 7.2. Activity and Object Recognition

**Metrics** We evaluate activity and object recognition using top-1 and top-5 accuracy. For activity recognition, we treat the situation activity label as the gold standard. For object recognition, we assume any synset value annotated in a labeled frame is a gold standard object in the image.

**Systems** For activity recognition, we adapt our situation CRF by maximizing the potential in Equation 1 and predicting the corresponding verb. As a baseline, we train a discrete classifier for all verbs in imSitu, using VGG features and an identical fine tuning setup as the CRF but with an initial learning rate of  $1e-3$ . For object recognition, we use our CRF to compute probability of observing any synset in the dataset by marginalizing Equation 1 over verbs and predicting the synset with the maximum marginal probability. As a baseline, we train a discrete classifier for all noun synsets in imSitu. We create pseudo-examples for every unique synset associated with an image and train the classifier on this expanded dataset, using identical training setup

		activity		object	
		top-1	top-5	top-1	top-5
dev	Activity	30.6	57.4	-	-
	Object	-	-	64.9	94.1
	Situation	32.25	58.6	72.9	95.0
test	Activity	31.1	57.7	-	-
	Object	-	-	64.1	94.2
	Situation	32.3	58.9	72.7	94.8

Table 4. Object and activity recognition results in imSitu. Joint prediction of object and activity through situation recognition improves over independently predicting either object or activity.

as the CRF but with an initial learning rate of  $1e-3$ .

**Quantitative Results** Table 4 summarizes our experiments on the imSitu development set. We also ran these experiments once on the imSitu test data, which confirms our development results. Our situation CRF significantly outperforms predicting either activities or objects in isolation, by 1.2% and by 8.6% at top-1, respectively. Overall, the results are encouraging; the context provided by situations is helping significantly, and improved models that more accurately reason about how objects interact with activities have significant potential to improve all three recognition tasks.

## 8. Conclusion

We introduced the problem of situation recognition and described the construction of imSitu, a large new situation recognition data set. Key to the formulation was the use of semantic roles to represent how objects, actors, and other entities participate in different activities. The situation recognition task is challenging but provides strong context for recognizing activities and objects. Future work involves developing more accurate models and using them in applications, including image captioning and visual QA.

**Acknowledgements** This research was supported in part by the NSF (IIS-1252835, IIS-1338054), DARPA under the CwC program through the ARO (W911NF-15-1-0543), ONR (N00014-13-1-0720), two Allen Distinguished Investigator Awards, the Allen Institute for AI, and an AWS in Education Grant award. We thank Vicente Ordonez for critical feedback on the draft and Jayant Krishnamurthy for help with the quality control algorithms. We also appreciate the undergraduate and turk annotators, and in particular Diana Wang for her tireless effort. Also, Eunsol Choi, Ricardo Martin, Nicholas Fitzgerald, Chloe Kiddon, Yannis Konstantas, Sam Thomson and the reviewers provided feedback that greatly improved the work.

					
<b>SKIDDING</b>	<b>RAFTING</b>	<b>WEeping</b>	<b>WILTING</b>	<b>HITCHHIKING</b>	<b>REARING</b>
AGENT CAR	AGENT PEOPLE	AGENT WOMAN	AGENT PLANT	AGENT WOMAN	AGENT HORSE
PLACE ROAD	PLACE WATERFALL	PLACE OUTDOORS	PLACE FLOWERBD	PLACE ROAD	PLACE GRASS
					
<b>SHELLING</b>	<b>APPREHENDING</b>	<b>STRETCHING</b>	<b>SPEARING</b>	<b>NUZZLING</b>	<b>CLINGING</b>
AGENT PEOPLE	AGENT SOLDIER	AGENT WOMAN	AGENT MAN	AGENT HORSE	AGENT SLOTH
ITEM PEANUT	VICTIM MAN	ITEM ARM	VICTIM FISH	ITEM HORSE	ITEM TREE
PLACE KITCHEN	PLACE FIELD	PLACE ROOM	PLACE WATER	PLACE OUTDOORS	PLACE OUTDOORS
					
<b>ATTACKING</b>	<b>SIGNING</b>	<b>THROWING</b>	<b>STROKING</b>	<b>CARRYING</b>	<b>WHIPPING</b>
AGENT MAN	AGENT MAN	AGENT MAN	AGENT PERSON	AGENT WOMAN	AGENT JOCKEY
VICTIM PLAYER	SIGNEDITEM BOOK	ITEM BASEBALL	OBJECT CAT	ITEM JAR	ITEM HORSE
TOOL CHAIR	TOOL PEN	DESTINATION CATCHER	PART NECK	AGENTPART HEAD	TOOL CROP
PLACE STADIUM	PLACE SHOP	PLACE BALLPARK	PLACE -	PLACE OUTDOORS	PLACE RACETRACK
					
<b>BATHING</b>	<b>WIPING</b>	<b>EATING</b>	<b>BRUSHING</b>	<b>STAPLING</b>	<b>PILOTING</b>
AGENT MAN	AGENT BOY	AGENT MAN	AGENT WOMAN	AGENT PERSON	AGENT MAN
COAGENT CAT	SUBSTANCE DIRT	FOOD SOUP	TOOL BRUSH	ITEM FABRIC	VEHICLE AIRPLANE
TOOL HAND	SOURCE HAND	CONTAINER CAN	TARGET TEETH	SURFACE WOOD	START -
SUBSTANCE SOAP	TOOL SHIRT	TOOL SPOON	TOOL -	TOOL STAPLEGUN	END -
PLACE BUCKET	PLACE OUTDOORS	PLACE ROOM	PLACE -	PLACE INSIDE	PLACE -
					
<b>SHAVING</b>	<b>COOKING</b>	<b>EMPTYING</b>	<b>SHAVING</b>	<b>ATTACHING</b>	<b>VAULTING</b>
AGENT BARBER	AGENT WOMAN	AGENT MAN	AGENT MAN	AGENT MAN	AGENT WOMAN
COAGENT MAN	FOOD VEGETABLE	ITEM WATER	COAGENT -	ITEM WOOD	START GROUND
BODYPART FACE	CONTAINER PAN	CONTAINER BUCKET	BODYPART HEAD	DESTINATION -	OBSTACLE POLE
TOOL BLADE	HEATSOURCE OVEN	DESINATION GROUND	SUBSTANCE -	TOOL HAND	END GROUND
SUBSTANCE CREAM	TOOL SPOON	TOOL HAND	TOOL RAZOR	GLUE SCREW	TOOL POLE
PLACE SALON	PLACE KITCHEN	PLACE OUTDOORS	PLACE -	PLACE OUTDOORS	PLACE OUTDOORS

Figure 7. Example realized situations from imSitu. Below each image is a table where the first row is the activity, the left column is semantic roles, and the right column is values for those roles. On the left outlined in gold are examples of gold standard annotated data. On the right is random output from our CRF model when it correctly predicted the activity. Incorrect semantic role values are highlighted in red, whereas correct ones are green.



## References

- [1] S. Antol et al. Vqa: Visual question answering. *arXiv preprint arXiv:1505.00468*, 2015. 3
- [2] X. Chen et al. Learning a recurrent visual representation for image caption generation. *arXiv:1411.5654*, 2014. 3
- [3] D. Das. *Semi-Supervised and Latent-Variable Models of Natural Language Semantics*. PhD thesis, CMU, 2012. 2, 3
- [4] V. Delaitre et al. Recognizing human actions in still images: a study of bag-of-features and part-based representations. In *BMVC*, 2010. 2
- [5] J. Deng et al. Construction and Analysis of a Large Scale Image Ontology. Vision Sciences Society, 2009. 3
- [6] S. Divvala et al. An empirical study of context in object detection. In *CVPR*, 2009. 3
- [7] D. Elliott et al. Comparing automatic evaluation measures for image description. In *ACL*, 2014. 3
- [8] V. R. et al. Linking people with “their” names using coreference resolution. In *ECCV*, 2014. 3
- [9] Z. et al. Building a large-scale multimodal knowledge base for visual question answering. *arXiv preprint arXiv:1507.05670*, 2015. 3
- [10] M. Everingham et al. The pascal visual object classes challenge 2009. In *2th PASCAL Challenge Workshop*, 2009. 2
- [11] H. Fang et al. From captions to visual concepts and back. *arXiv:1411.4952*, 2014. 3
- [12] A. Farhadi et al. Every picture tells a story: Generating sentences from images. In *ECCV 2010*, pages 15–29. 2010. 3
- [13] C. Fellbaum. *WordNet*. Wiley Online Library, 1998. 2, 3
- [14] C. J. Fillmore et al. Background to framenet. *International Journal of lexicography*, 2003. 2, 3
- [15] N. FitzGerald et al. Semantic role labelling with neural network factors. In *EMNLP*, 2015. 6
- [16] A. Frome et al. Devise: A deep visual-semantic embedding model. In *NIPS*, 2013. 3
- [17] C. Galleguillos et al. Context based object categorization: A critical survey. *CVIU*, 2010. 3
- [18] H. e. a. Gao. Are you talking to a machine? dataset and methods for multilingual image question answering. *arXiv preprint arXiv:1505.05612*, 2015. 3
- [19] Y. Goldberg et al. A dataset of syntactic-ngrams over time from a very large corpus of english books. In *\*SEM*, 2013. 4
- [20] S. Guadarrama et al. Youtube2text: Recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition. In *ICCV*, 2013. 3
- [21] G. Guo et al. A survey on still image based human action recognition. *Pattern Recognition*, 2014. 2
- [22] A. Gupta et al. Beyond nouns: Exploiting prepositions and comparative adjectives for learning visual classifiers. In *ECCV*. 2008. 2
- [23] S. Gupta et al. Visual semantic role labeling. *arXiv preprint arXiv:1505.04474*, 2015. 3
- [24] M. Hodosh et al. Framing image description as a ranking task: Data, models and evaluation metrics. *JAIR*, 2013. 3
- [25] Y. Jia et al. Caffe: Convolutional architecture for fast feature embedding. *arXiv:1408.5093*, 2014. 6
- [26] A. Karpathy et al. Deep visual-semantic alignments for generating image descriptions. *arXiv:1412.2306*, 2014. 3
- [27] P. Kingsbury and M. Palmer. From treebank to propbank. In *LREC*. Citeseer, 2002. 3
- [28] C. Kong et al. What are you talking about? text-to-image coreference. In *CVPR*, 2014. 3
- [29] A. Lazaridou et al. Is this a wampimuk? In *ACL*, 2014. 3
- [30] D.-T. Le et al. Tuhoi: Trento universal human object interaction dataset. *V&L Net 2014*, 2014. 2
- [31] L.-J. Li et al. What, where and who? classifying events by scene and object recognition. In *CVPR*, 2007. 2
- [32] T.-Y. Lin et al. Microsoft coco: Common objects in context. In *ECCV*. 2014. 3
- [33] S. Maji et al. Action recognition from a distributed representation of pose and appearance. In *CVPR*, 2011. 3
- [34] J. Mao et al. Explain images with multimodal recurrent neural networks. *arXiv:1410.1090*, 2014. 3
- [35] M. Marszałek et al. Actions in context. In *CVPR*, 2009. 1, 3
- [36] V. Ordonez et al. Im2text: Describing images using 1 million captioned photographs. In *NIPS*, 2011. 3
- [37] M. Palmer. Semlink: Linking propbank, verbnet and framenet. In *GLC*, pages 9–15, 2009. 3
- [38] A. Rabinovich et al. Objects in context. In *ICCV*, 2007. 3
- [39] M. Ren et al. Image question answering: A visual semantic embedding model and a new dataset. *arXiv preprint arXiv:1505.02074*, 2015. 3
- [40] M. Ronchi et al. Describing common human visual actions in images. In *BMVC*, 2015. 3
- [41] O. Russakovsky et al. ImageNet Large Scale Visual Recognition Challenge. *CoRR*, 2014. 2, 6
- [42] C. Silberger et al. Grounded models of semantic representation. In *EMNLP*, 2012. 3
- [43] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014. 2, 6
- [44] K. Soomro et al. Ucf101: A dataset of 101 human actions classes from videos in the wild. 2012. 2
- [45] R. Vedantam et al. Cider: Consensus-based image description evaluation. *arXiv:1411.5726*, 2014. 3
- [46] O. Vinyals et al. Show and tell: A neural image caption generator. *arXiv:1411.4555*, 2014. 3
- [47] B. Yao et al. Modeling mutual context of object and human pose in human-object interaction activities. In *CVPR*. 3
- [48] B. Yao et al. Grouplet: A structured image representation for recognizing human and object interactions. In *CVPR*, 2010. 2
- [49] B. Yao et al. Human action recognition by learning bases of action attributes and parts. In *ICCV*, 2011. 2
- [50] M. Yatskar et al. See no evil, say no evil: Description generation from densely labeled images. *\*SEM*, 2014. 3
- [51] L. e. a. Yu. Visual madlibs: Fill in the blank image generation and question answering. *arXiv preprint arXiv:1506.00278*, 2015. 3
- [52] Y. Zhu et al. Reasoning about object affordances in a knowledge base representation. In *ECCV*. 2014. 3