Commonly Uncommon:
Semantic Sparsity in Situation Recognition

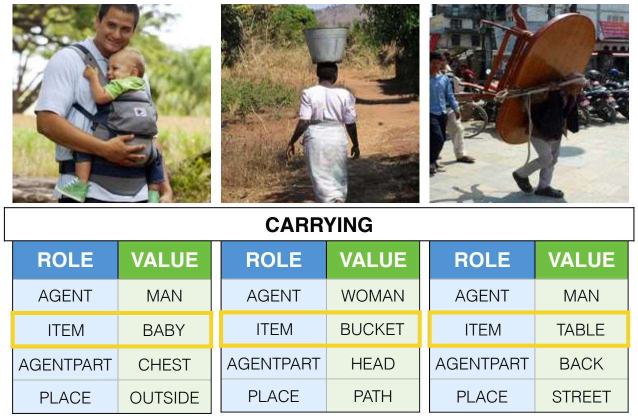
Mark Yatskar[1], Vicente Ordonez[2,3], Luke Zettlemoyer[1], Ali Farhadi[1,2]
[1]Computer Science & Engineering, University of Washington, Seattle, WA
[2]Allen Institute for Artificial Intelligence (AI2), Seattle, WA
[3]Department of Computer Science, University of Virginia, Charlottesville, VA.
[my89, lsz, ali]@cs.washington.edu, vicente@cs.virginia.edu

Abstract

Semantic sparsity is a common challenge in structured visual classification problems; when the output space is complex, the vast majority of the possible predictions are rarely, if ever, seen in the training set. This paper studies semantic sparsity in situation recognition, the task of producing structured summaries of what is happening in images, including activities, objects and the roles objects play within the activity. For this problem, we find empirically that most substructures required for prediction are rare, and current state-of-the-art model performance dramatically decreases if even one such rare substructure exists in the target output.We avoid many such errors by (1) introducing a novel tensor composition function that learns to share examples across substructures more effectively and (2) semantically augmenting our training data with automatically gathered examples of rarely observed outputs using web data. When integrated within a complete CRF-based structured prediction model, the tensor-based approach outperforms existing state of the art by a relative improvement of 2.11% and 4.40% on top-5 verb and noun-role accuracy, respectively. Adding 5 million images with our semantic augmentation techniques gives further relative improvements of 6.23% and 9.57% on top-5 verb and noun-role accuracy.

1. Introduction

Many visual classification problems, such as image captioning [29], visual question answering [2], referring expressions [23], and situation recognition [44] have structured, semantically interpretable output spaces. In contrast to classification tasks such as ImageNet [37], these problems typically suffer from *semantic sparsity*; there is a combinatorial number of possible outputs, no dataset can cover them all, and performance of existing models degrades significantly when evaluated on rare or unseen inputs [3, 46, 9, 44]. In this paper, we consider situation recognition, a prototypical structured classification problem with significant semantic sparsity, and develop new models and semantic data augmentation techniques that significantly improve performance by better modeling the underlying semantic structure of the task.



CARRYING					
ROLE	VALUE	ROLE	VALUE	ROLE	VALUE
AGENT	MAN	AGENT	WOMAN	AGENT	MAN
ITEM	BABY	ITEM	BUCKET	ITEM	TABLE
AGENTPART	CHEST	AGENTPART	HEAD	AGENTPART	BACK
PLACE	OUTSIDE	PLACE	PATH	PLACE	STREET

Figure 1: Three situations involving `carrying`, with semantic roles `agent`, the carrier, `item`, the carried, `agentpart`, the part of the agent carrying, and `place`, where the situation is happening. For carrying, there are many possible carry-able objects (nouns that can fill the `item` role), which is an example of semantic sparsity. Such rarely occurring substructures are challenging and cause significant errors, affecting not only performance on role-values but also verbs.

Situation recognition [44] is the task of producing structured summaries of what is happening in images, including activities, objects and the roles those objects play within the activity. This problem can be challenging because many activities, such as `carrying`, have very open ended semantic roles, such as `item`, the thing being carried (see Figure 1); nearly any object can be carried and the training data will never contain all possibilities. This is a prototypical instance of semantic sparsity: rare outputs constitute

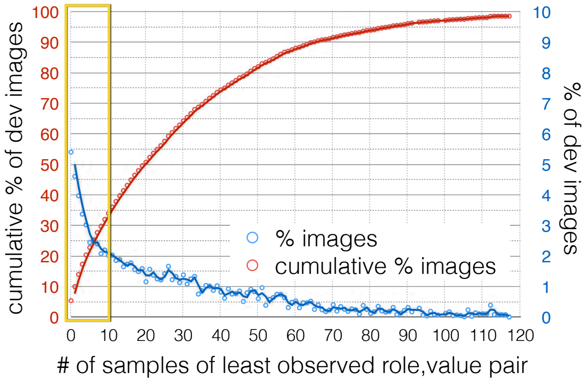


Figure 2: The percentage of images in the imSitu development set as a function of the total number of training examples for the least frequent role-noun pair in each situation. Uncommon target outputs, those observed fewer than 10 times in training (yellow box), are common, constituting 35% of all required predictions. Such semantic sparsity is a central challenge for situation recognition.

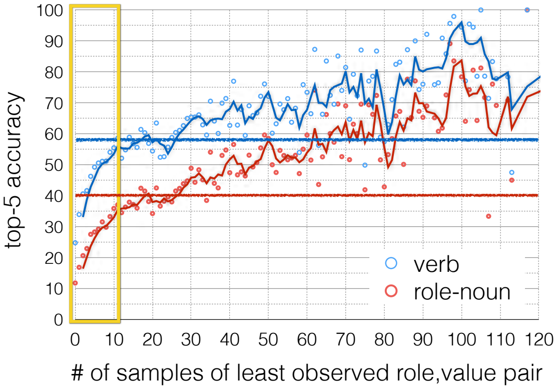


Figure 3: Verb and role-noun prediction accuracy of a baseline CRF [44] on the imSitu dev set as a function of the frequency of the least observed role-noun pair in the training set. Solid horizontal lines represent average performance across the whole imSitu dev set, irrespective of frequency. As even one target output becomes uncommon (highlighted in yellow box), accuracy decreases.

a large portion of required predictions (35% in the imSitu dataset [44], see Figure 2), and current state-of-the-art performance for situation recognition drops significantly when even one participating object has few samples for it's role (see Figure 3). We propose to address this challenge in two ways by (1) building models that more effectively share examples of objects between different roles and (2) semantically augmenting our training set to fill in rarely represented noun-role combinations.

We introduce a new compositional Conditional Random Field formulation (CRF) to reduce the effects of semantic sparsity by encouraging sharing between nouns in different roles. Like previous work [44], we use a deep neural network to directly predict factors in the CRF. In such models, required factors for the CRF are predicted using a global image representation through a linear regression unique to each factor. In contrast, we propose a novel tensor composition function that uses low dimensional representations of nouns and roles, and shares weights across all roles and nouns to score combinations. Our model is compositional, independent representations of nouns and roles are combined to predict factors, and allows for a globally shared representation of nouns across the entire CRF.

This model is trained with a new form of semantic data augmentation, to provide extra training samples for rarely observed noun-role combinations. We show that it is possible to generate short search queries that correspond to partial situations (i.e. "man carrying baby" or "carrying on back" for the situations in Figure 1) which can be used for web image retrieval. Such noisy data can then be incorporated in pre-training by optimizing marginal likelihood, effectively performing a soft clustering of values for unlabeled aspects of situations. This data also supports, as we will show, self training where model predictions are used to prune the set of images before training the final predictor.

Experiments on the imSitu dataset [44] demonstrate that our new compositional CRF and semantic augmentation techniques reduce the effects of semantic sparsity, with strong gains for relatively rare configurations. We show that each contribution helps significantly, and that the combined approach improves performance relative to a strong CRF baseline by 6.23% and 9.57% on top-5 verb and noun-role accuracy, respectively. On uncommon predictions, our methods provide a relative improvement of 8.76% on average across all measures. Together, these experiments demonstrate the benefits of effectively targeting semantic sparsity in structured classification tasks.

2. Background

Situation Recognition Situation recognition has been recently proposed to model events within images [19, 36, 43, 44], in order to answer questions beyond just "What activity is happening?" such as "Who is doing it?", "What are they doing it to?", "What are they doing it with?". In general, formulations build on semantic role labelling [17], a problem in natural language processing where verbs are automatically paired with their arguments in a sentence (for example, see [8]). Each semantic role corresponds to a question about an event, (for example, in the first image of Figure 1, the semantic role agent corresponds to "who is doing the carrying?" and agentpart corresponds to "how is the item being carried?").

We study situation recognition in imSitu [44], a large-scale dataset of human annotated situations containing over 500 activities, 1,700 roles, 11,000 nouns, 125,000 images. imSitu images are collected to cover a diverse set of situations. For example, as seen in Figure 2, 35% of situations annotated in the imSitu development set contain at least one rare role-noun pair. Situation recognition in imSitu is a strong test bed for evaluating methods addressing semantic sparsity: it is large scale, structured, easy to evaluate, and has a clearly measurable range of semantic sparsity across different verbs and roles. Furthermore, as seen in Figure 3, semantic sparsity is a significant challenge for current situation recognition models.

Formal Definition In situation recognition, we assume a discrete sets of verbs V, nouns N, and frames F. Each frame $f \in F$ is paired with a set of semantic roles E_f. Every element in V is mapped to exactly one f. The verb set V and frame set F are derived from FrameNet [13], a lexicon for semantic role labeling, while the noun set N is drawn from WordNet [34]. Each semantic role $e \in E_f$ is paired with a noun value $n_e \in N \cup \{\varnothing\}$, where $\varnothing$ indicates the value is either not known or does not apply. The set of pairs of semantic roles and their values is called a realized frame, $R_f = \{(e, n_e) : e \in E_f\}$. Realized frames are valid only if each $e \in E_f$ is assigned exactly one noun n_e.

Given an image, the task is to predict a situation, $S = (v, R_f)$, specified by a verb $v \in V$ and a valid realized frame R_f, where f refers to a frame mapped by v . For example, in the first image of Figure 1, the predicted situations is S = (carrying, {(agent,man), (item,baby), (agentpart,chest), (place,outside)}).

3. Methods

This section presents our compositional CRFs and semantic data augmentation techniques.

3.1. Compositional Conditional Random Field

Figure 4 shows an overview of our compositional conditional random field model, which is described below.

Conditional Random Field 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 $R_f = \{(e, n_e) : e \in E_f\}$, similarly to previous work [44]. The full distribution, with potentials for verbs ψ_v and semantic roles ψ_e takes the form:

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

The CRF admits efficient inference: we can enumerate all verb-semantic roles that occur and then sum all possible semantic role values that occurred in a dataset.

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 ϕ_e and ϕ_v encode scores computed by a neural network. To learn this model, we assume that for an image i in dataset Q there can, in general, be a set A_i of possible ground truth situations [1]. We optimize the log-likelihood of observing at least one situation $S \in A_i$:

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

Compositional Tensor Potential In previous work, the CRF potentials (Equation 2 and 3) are computed using a global image representation, a p-dimensional image vector $g_i \in \mathcal{R}^p$, derived by the VGG convolutional neural network [40]. Each potential value is computed by a linear regression with parameters, θ, unique for each possible decision of verb and verb-role-noun (we refer to this as image regression in Figure 4), for example for the verb-role-noun potential in Equation 3:

$$\phi_e(v, e, n_e, i, \theta) = g_i^T \theta_{v,e,n_e} \quad (5)$$

Such a model does not directly represent the fact that nouns are reused between different roles, although the underlying neural network could hypothetically learn to encode such reuse during fine tuning. Instead, we introduce compositional potentials that make such reuse explicit.

To formulate our compositional potential, we introduce a set of m-dimensional vectors $D = \{d_n \in \mathcal{R}^m | n \in N\}$, one vector for each noun in N, the set of nouns. We create a set matrices $T = \{H_{(v,e)} \in \mathcal{R}^{p\times o} | (v, e) \in E_f\}$, one matrix for each verb, semantic role pair occurring in all frames E_f, that map image representations to o-dimensional verb-role representations. Finally, we introduce a tensor of global composition weights, $C \in \mathcal{R}^{m\times o\times p}$. We define a tensor weighting function, T, which takes as input a verb, v, semantic role, e, noun, n, and image representation, g_i as:

$$T(v, e, n, g_i) = C \odot (d_n \otimes g_i^T H_{(v,e)} \otimes g_i) \quad (6)$$

The tensor weighting function constructs an image specific verb-role representation by multiplying the global image vector and the verb-role matrix $g_i^T H_{(v,e)}$. Then, it combines a global noun representation, the image specific role representation, and the global image representation with outer products. Finally, it weights each dimension of the outer product with a weight from C. The weights in C indicate which features of the 3-way outer product are important. The final potential is produced by summing up all of the elements of the tensor produced by T:

$$\phi_e(v, e, n_e, i) = \sum_{x=0}^{M}\sum_{y=0}^{O}\sum_{z=0}^{P} T(v, e, n_e, g_i)[x, y, z] \quad (7)$$

[1] imSitu provides three realized frames per example image.

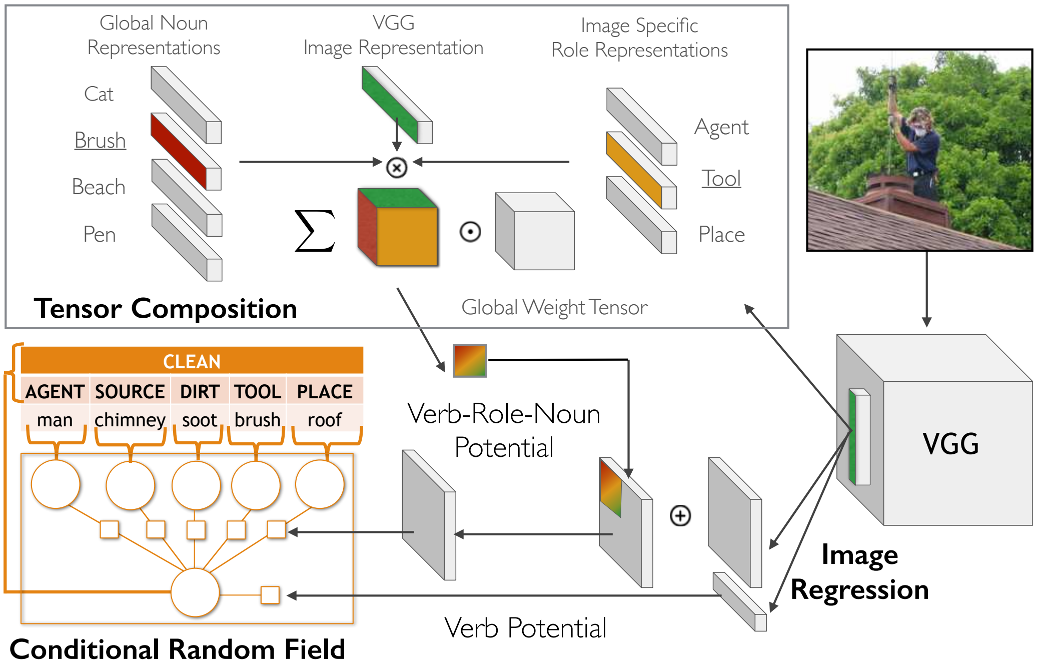


Figure 4: An overview of our compositional Conditional Random Field (CRF) for predicting situations. A deep neural network is used to compute potentials in a CRF. The verb-role-noun potential is built from a global bank of noun representations, image specific role representations and a global image representation that are combined with a weighted tensor product. The model allows for sharing among the same nouns in different roles, leading to significant gains, as seen in Section 5.

The tensor produced by T in general will be high dimensional and very expressive. This allows use of small dimensionality representations, making the function more robust to small numbers of samples for each noun.

The potential defined in Equation 7 can be equivalently formulated as :

$$\phi_e(v, e, n_e, i) = g_i^T A(d_{n_e} \otimes g_i^T H_{(v,e)}) \qquad (8)$$

Where A is a matrix with the same parameters as C but flattened to layout the noun and role dimensions together. By aligning terms with Equation 5, one can see that tensor potential offers an alternative parametrized to the linear regression that uses many more general purpose parameters, those of C. Furthermore, it eliminates any one parameter from ever being uniquely associated with one regression, instead compositionally using noun and verb-role representations to build up the parameters of the regression.

3.2. Semantic Data Augmentation

Situation recognition is strongly connected to language. Each situation can be thought of as simple declarative sentence about an activity happening in an image. For example, the first situation in Figure 1 could be expressed as "man carrying baby on chest outside" by knowing the prototypical ordering of semantic roles around verbs and inserting prepositions. This relationship can be used to reduce semantic sparsity by using image search to find images that could contain the elements of a situations.

We convert annotated situations to phrases for semantic augmentation by exhaustively enumerating all possible sub-pieces of realized situations that occur in the imSitu training set (see Section 4 for implementation details). For example, in first situation of Figure 1, we get the pieces: $(\texttt{carrying}, \{(\texttt{agent}, \texttt{man})\})$, $(\texttt{carrying}, \{(\texttt{agent}, \texttt{man}), (\texttt{item}, \texttt{baby})\})$, ect. Each of these substructures is converted deterministically to a phrase using a template specific for every verb. For example, the template for carrying is "{agent} carrying {item} {with agentpart} {in place}." Partial situations are realized into phrases by taking the first gloss in Wordnet of the synset associated with every noun in the substructure, inserting them into the corresponding slots of the template, and discarding unused slots. For example, the phrases for the sub-pieces above are realized as "man carrying" and "man carrying baby." These phrases are used to retrieve images from Google image search and construct a set, $W = \{(i, v, R_f)\}$, of images annotated with a verb and partially complete realized frames, by assigning retrieved images to the sub-piece that generated the retrieval query.[2]

[2] While these templates do not generate completely fluent phrases, preliminary experiments found them sufficiently accurate for image search

Pre-training Images retrieved from the web can be incorporated in a pre-training phase. The images retrieved only have partially specified realized situations as labels. To account for this, we instead compute the marginal likelihood, $\hat{p}$, of the partially observed situations in W:

$$\hat{p}(S|i;\theta) \propto \psi_v(v,i;\theta) \prod_{(e,n_e)\in R_f} \psi_e(v,e,n_e,i;\theta) \times \prod_{e\notin R_f \wedge e\in E_f} \sum_n \psi_e(v,e,n,i;\theta) \tag{9}$$

During pretraining, we optimize the marginal log-likelihood of W. This objective provides a partial clustering over the unobserved roles left unlabeled during the retrieval process.

Self Training Images retrieved from the web contain significant noise. This is especially true for role-noun combinations that occur infrequently, limiting their utility for pretraining. Therefore, we also consider filtering images in W after a model has already been trained on fully supervised data from imSitu. We rank images in W according to $\hat{p}$ as computed by the trained model and filter all those not in the top-k for every unique R_f in W. We then pretrain on this subset of W, train again on imSitu, and then increase k. We repeat this process until the model no longer improves.

4. Experimental Setup

Models All models were implemented in Caffe [21] and use a pretrained VGG network [40] for the base image representation with the final two fully connected layers replaced with two fully connected layers of dimensionality 1024. We finetune all layers of VGG for all models. For our tensor potential we use noun embedding size, $m = 32$, and role embedding size $o = 32$, and the final layer of our VGG network as the global image representation where $p = 1024$. Larger values of m and o did seem to improve results but were too slow to pretrain so we omit them. In experiments where we use the image regression in conjunction with a compositional potential, we remove regression parameters associated with combinations seen fewer than 10 times on the imSitu training set to reduce overfitting.

Baseline We compare our models to two alternative methods for introducing effective sharing between nouns. The first baseline (Noun potential in Table 1 and 2) adds a potential into the baseline CRF for nouns independent of roles. We modify the probability, from Equation 9 of a situation, S, given an image i, to not only decompose by pairs of roles, e and nouns n_e in a realized frame R_f, but also nouns n_e:

$$p(S|i;\theta) \propto \psi_v(v,i;\theta) \prod_{(e,n_e)\in R_f} \psi_e(v,e,n_e,i;\theta)\psi_{n_e}(n_e,i) \tag{10}$$

The added potential, ψ_{n_e}, is computed using a regression from a global image representation for each unique n_e.

The second baseline we consider is compositional but does not use a tensor based composition method. The model instead constructs many verb-role representations and combines them with noun representations using inner-products (Inner product composition in Table 1 and 2). In this model, as in the tensor model in Section 3, we use a global image representation $g_i \in \mathcal{R}^p$ and a set noun vectors, $d_n \in \mathcal{R}^m$ for every noun n. We also assume t verb-role matrices $H_{t,v,e} \in \mathcal{R}^{o\times p}$ for every verb-role in E_f. We compute the corresponding potential as in Equation 11:

$$\phi_e(v,e,n_e,i) = \sum_k d_{n_e}^T H_{(k,v,e)} q_i \tag{11}$$

The model is motivated by compositional models used for semantic role labeling [14] and allows us to trade-off the need to reduce parameters associated with nouns and expressivity. We grid search values of t such that $t \cdot o$ was at most 256, the largest size network we could afford to run and $o = m$, a requirement on the inner product. We found the best setting at $t = 16$, $o = m = 16$.

Decoding We experimented with two decoding methods for finding the best scoring situation under the CRF models. Systems which used the compositional potentials performed better when first predicting a verb v^m using the max-marginal over semantic roles: $v^m = \arg\max_v \sum_{(e,n_e)} p(v, R_f|i)$ and then predict a realized frame, R_f^m, with max score for v^m: $R_f^m = \arg\max_{R_f} p(v^m, R_f|i)$. All other systems performed better maximizing jointly for both verb and realized frame.

Optimization All models were trained with stochastic gradient descent with momentum 0.9 and weight decay 5e-4. Pretraining in semantic augmentation was conducted with initial learning rate of 1e-3, gradient clipping at 100, and batch size 360. When training on imSitu data, we use an initial learning rate of 1e-5. For all models, the learning rate was reduced by a factor of 10 when the model did not improve on the imSitu dev set.

Semantic Augmentation In experiments with semantic augmentation, images were retrieved using Google image search. We retrieved 200 medium sized, full-color, safe search filtered images per query phrase. We produced over 1.5 million possible query phrases from the imSitu training set, the majority extremely rare. We limited the phrases to any that occur between 10 and 100 times in imSitu and for phrases that occur between 3 and 10 times we accepted only those containing at most one noun. Roughly 40k phrases were used to retrieve 5 million images from the web. All duplicate images occurring in imSitu were removed. For pretraining, we ran all experiments up to 50k updates (roughly

because often no phrase could retrieve correct images. Longer phrases tended to have much lower precision.

			top-1 predicted verb			top-5 predicted verbs			ground truth verbs		
			verb	value	value-all	verb	value	value-all	value	value-all	mean
imSitu	1	Baseline: Image Regression [44]	32.25	24.56	14.28	58.64	42.68	22.75	65.90	29.50	36.32
	2	Noun Potential + reg	27.64	21.21	12.21	53.95	39.95	21.45	68.87	32.31	34.70
	3	Inner product composition + reg	32.13	24.77	14.71	58.33	42.93	23.14	66.79	30.2	36.62
	4	Tensor composition	31.73	24.04	13.73	58.06	42.64	22.7	68.73	32.14	36.72
	5	Tensor composition + reg	32.91	25.39	14.87	59.92	44.5	24.04	69.39	33.17	38.02
+ SA	6	Baseline : Image Regression	32.40	24.14	15.17	59.10	44.04	24.40	68.03	31.93	37.53
	7	Tensor composition + reg	34.04	26.47	**15.73**	61.75	46.48	**25.77**	**70.89**	**35.08**	39.53
	8	Tensor composition + reg + self train	**34.20**	**26.56**	15.61	**62.21**	**46.72**	25.66	70.80	34.82	**39.57**

Table 1: Situation recognition results on the full imSitu development set. The results are divided by models which were only trained on imSitu data, rows 1-5, and models which use web data through semantic data augmentation, marked as +SA in rows 6-8. Models marked with +reg also include image regression potentials used in the baseline. Our tensor composition model, row 5, significantly outperforms the existing state of the art, row 1, addition of a noun potential, row 2, and a compositional baseline, row 3. The tensor composition model is able to make better use of semantic data augmentation (row 8) than the baseline (row 6).

			top-1 predicted verb			top-5 predicted verbs			ground truth verbs		
			verb	value	value-all	verb	value	value-all	value	value-all	mean
imSitu	1	Baseline: image regression [44]	19.89	11.68	**2.85**	44.00	24.93	**6.16**	50.80	9.97	19.92
	2	Noun potential + reg	15.88	9.13	1.86	38.22	22.28	5.46	54.65	11.91	19.92
	3	Inner product composition + reg	18.96	10.69	1.89	42.53	23.28	3.69	49.54	6.46	19.63
	4	Tensor composition	19.78	11.28	2.26	42.66	24.42	5.57	54.06	11.47	21.43
	5	Tensor composition + reg	**21.12**	11.89	2.20	45.14	25.51	5.36	53.58	10.62	21.93
+ SA	6	Baseline : image regression	19.95	11.44	2.13	43.08	24.56	4.95	51.55	8.41	20.76
	7	Tensor composition + reg	20.08	11.58	2.22	44.82	26.02	5.55	55.45	11.53	22.16
	8	Tensor composition + reg + self train	20.52	**11.91**	2.34	**45.94**	**26.99**	6.06	**55.90**	**12.04**	**22.71**

Table 2: Situation prediction results on the rare portion imSitu development set. The results are divided by models which were only trained on imSitu data, rows 1-5, and models which use web data through semantic data augmentation, marked as +SA in rows 6-8. Models marked with +reg also include image regression potentials used in the baseline. Semantic data augmentation with the baseline hurts for rare cases. Semantic augmentation yields larger relative improvement on rare cases and a composition-based model is required to realize these gains.

4 epochs). For self training, we only self train on rare realized frames (those 10 or fewer times in imSitu train set). Self training yielded diminishing gains after two iterations and we ran the first iteration at k=10 and the second at k=20.

Evaluation We use the standard data split for imSitu[44] with 75k train, 25k development, and 25k test images. We follow the evaluation setup defined for imSitu, evaluating verb predictions (verb) and semantic role-value pair predictions (value) and full structure correctness (value-all). We report accuracy at top-1, top-5 and given the ground truth verb and the average across all measures (mean). We also report performance for examples requiring rare (10 or fewer examples in the imSitu training set) predictions.

5. Results

Compositional Tensor Potential Our results on the full imSitu dev set are presented in Table 1 in rows 1-5. Overall results demonstrate that adding a noun potential (row 2) and our baseline composition model (row 3) are ineffective and perform worse than the baseline CRF (row 1). We hypothesize that systematic variation in object appearance between roles is challenging for these models. Our tensor composition model (row 4) is able to better capture such variation and effectively share information among nouns, reflected by improvements in value and value-all accuracy given ground truth verbs while maintaining high top-1 and top-5 verb accuracy. However, as expected, many situations cannot be predicted only compositionally based on nouns (consider that a horse sleeping looks very different than a horse swimming and nothing like a person sleeping). Combination of the image regression potential and our tensor composition potential (row 5) yields the best performance, indicating they are modeling complementary aspects of the problem. Our final model (row 5) only trained on imSitu data outperforms the baseline on every measure, improving over 1.70 points overall.

Results on the rare portion of the imSitu dataset are presented in Table 2 in rows 1-5. Our final model (row 5) provides the best overall performance (mean column) on rare cases among models trained only on imSitu data, improving by 0.64 points on average. All models struggle to get correctly entire structures (value-all columns), indicating rare predictions are extremely hard to get completely correct while the baseline model which only uses image regression potentials performs the best. We hypothesize that image regression potentials may allow the model to more easily coordinate predictions across roles simultaneously because role-noun combinations that always co-occur will always have the same set of regression weights.

		top-1 predicted verb			top-5 predicted verbs			ground truth verbs		
		verb	value	value-all	verb	value	value-all	value	value-all	mean
imSitu	Baseline: Image Regression [44]	32.34	24.64	14.19	58.88	42.76	22.55	65.66	28.96	36.25
	Tensor composition + reg	32.96	25.32	14.57	60.12	44.64	24.00	69.2	32.97	37.97
+ SA	Baseline : Image Regression	32.3	24.95	14.77	59.52	44.08	23.99	67.82	31.46	37.36
	Tensor composition + reg + self train	**34.12**	**26.45**	**15.51**	**62.59**	**46.88**	**25.46**	**70.44**	**34.38**	**39.48**

Table 3: Situation prediction results on the full imSitu test set. Models were run exactly once on the test set. General trends are identical to experiments run on development set.

		top-1 predicted verb			top-5 predicted verbs			ground truth verbs		
		verb	value	value-all	verb	value	value-all	value	value-all	mean
imSitu	Baseline: Image Regression [44]	**20.61**	11.79	**3.07**	44.75	24.85	5.98	50.37	9.31	21.34
	Tensor composition + reg	19.96	11.57	2.30	44.89	25.26	4.87	53.39	10.15	21.55
+ SA	Baseline : Image Regression	19.46	11.15	2.13	43.52	24.14	4.65	51.21	8.26	20.57
	Tensor composition + reg + self train	20.32	**11.87**	2.52	**47.07**	**27.50**	**6.35**	**55.72**	**12.28**	**22.95**

Table 4: Situation prediction results on the rare portion of imSitu test set. Models were run exactly once on the test set. General trends established on the development set are supported.

Semantic Data Augmentation Our results on the full imSitu development set are presented in Table 1 in rows 6-8. Overall results indicate that semantic data augmentation helps all models, while our tensor model (row 7) benefits more than the baseline (row 6). Self training improves the tensor model slightly (row 8), making it perform better on top-1 and top-5 predictions but hurting performance given gold verbs. On average, our final model outperforms the baseline CRF trained on identical data by 2.04 points.

Results on the rare portion of the imSitu dataset are presented in Table 2 in rows 6-8. Surprisingly, on rare cases semantic augmentation hurts the baseline CRF (line 6). Rare instance image search results are extremely noisy. On close inspection, many of the returned results do not contain the target activity at all but instead contain target nouns. We hypothesize that without an effective global noun representation, the baseline CRF cannot extract meaningful information from such extra data. On the other hand, our tensor model (line 7) improves on these rare cases overall and with self training improves further (line 8).

Overall Results Experiments show that (a) our tensor model is able perform better in comparable data settings, (b) our semantic augmentation techniques largely benefit all models, and (c) our tensor model benefits more from semantic augmentation. We also present our full performance on top-5 verb across all numbers of samples in Figure 5. While our compositional CRF with semantic augmentation outperforms the baseline CRF, both models continue to struggle on uncommon cases. Our techniques seem to give most benefit for examples requiring predictions of structures seen between 5 and 35 times, while providing somewhat less benefit to even rarer ones. It is challenging future work to make further improvements for extremely rare outputs.

We also evaluated our models on the imSitu test set exactly once. The results are summarized in Table 3 for the full imSitu test set and in Table 4 for the rare portion. General trends established on the imSitu dev set are supported. We provide examples in Figure 6 of predictions our final system made on rare examples from the development set.

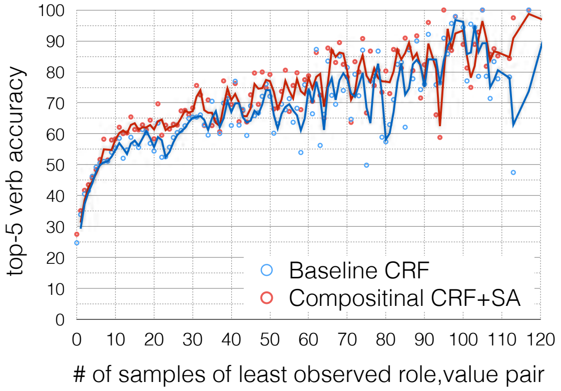


Figure 5: Top-5 verb accuracy on the imSitu development set. Our final compositional CRF with semantic data augmentation outperforms the baseline CRF on rare cases (fewer than 10 training examples), but both models continue to struggle with semantic sparsity. For our final model, the largest improvement relative to the baseline are for cases with 5-35 examples on the training set.

6. Related Work

Learning to cope with semantic sparsity is closely related to zero-shot or k-shot learning. Attribute-based learning [24, 25, 12], cross-modal transfer [39, 28, 15, 26] and using text priors [32, 18] have all been proposed but they study classification or other simplified settings. For the structured case, image captioning models [45, 22, 7, 11, 33, 20, 35, 31] have been observed to suffer from a lack of diversity and generalization [42]. Recent efforts to gain insight on such issues extract subject-verb-object (SVO) triplets from captions and count prediction failures on rare tuples [3]. Our use of imSitu to study semantic sparsity circumvents the need for intermediate processing of captions and general-

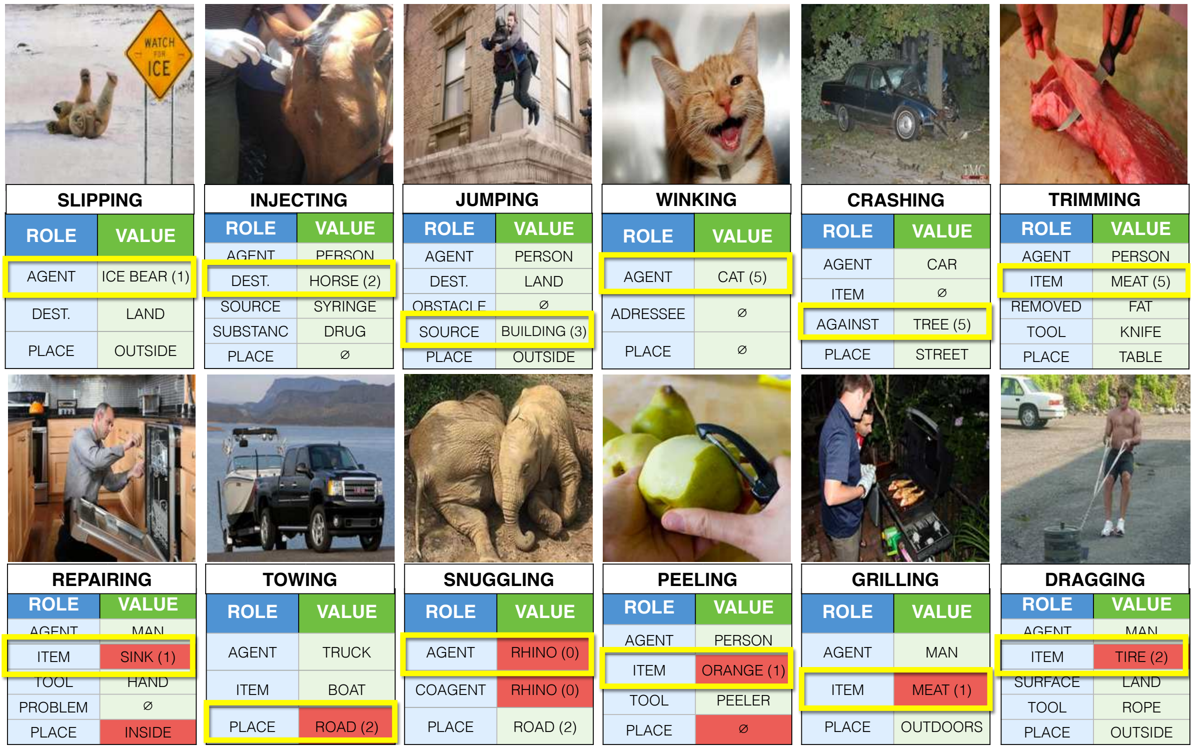


Figure 6: Output from our final model on development examples containing rare role-noun pairs. The first row contains examples where the model correctly predicts the entire structures in the top-5 (top-5, value-all). We highlight the particular role-noun pairs that make the examples rare with a yellow box and put the number occurrences of it in the imSitu training set. The second row contains examples where the verb was correctly predicted in the top-5 but not all the values were predicted correctly. We highlight incorrect predictions in red. Many such predictions occur zero times in the training set (ex. the third image on the second row). All systems struggle with such cases.

izes to verbs with more than two arguments.

Compositional models have been explored in a number of applications in natural language processing, such as sentiment analysis [41], dependency parsing [27], text similarity [4], and visual question answering [1] as effective tools for combining natural language elements for prediction. Recently, bilinear pooling [30] and compact bilinear pooling [16] have been proposed as second-order feature representations for tasks such as fine grained recognition and visual question answer. We build on such methods, using low dimensional embeddings of semantic units and expressive outer product computations.

Using the web as a resource for image understanding has been studied through NEIL [6], a system which continuously queries for concepts discovered in text, and Levan [10], which can create detectors from user specified queries. Web supervision has also been explored for pretraining convolutional neural networks [5] or for fine-grained bird classification [5] and common sense reasoning [38]. Yet we are the first to explore the connection between semantic sparsity and language for automatically generating queries for semantic web augmentation and we are able to show improvement on a large scale, fully supervised structured prediction task.

7. Conclusion

We studied situation recognition, a prototypical instance of a structured classification problem with significant semantic sparsity. Despite the fact that the vast majority of the possible output configurations are rarely observed in the training data, we showed it was possible in introduce new compositional models that effectively share examples among required outputs and semantic data augmentation techniques that significantly improved performance. In the future, it will be important to introduce similar techniques for related problems with semantic sparsity and generalize these ideas to the zero-shot learning.

Acknowledgements This work is in part supported by ONR N00014-13-1-0720, NSF IIS-1338054, NSF-1652052, NRI-1637479, NSF RI-1525251, DARPA under the CwC program through the ARO (W911NF-15-1-0543), two Allen Distinguished Investigator Award, and the Allen Institute for Artificial Intelligence. We thank Jayant Krishnamurthy for helpful insight into what the tensor methods were doing. Also Eunsol Choi, Ricardo Martin, Nicholas Fitzgerald, Yannis Konstas, Kenton Lee, and the reviewers who all feedback greatly improved the work.

References

[1] J. Andreas, M. Rohrbach, T. Darrell, and D. Klein. Neural module networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 39–48, 2016.

[2] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh. Vqa: Visual question answering. In *International Conference on Computer Vision*, 2015.

[3] Y. Atzmon, J. Berant, V. Kezami, A. Globerson, and G. Chechik. Learning to generalize to new compositions in image understanding. *arXiv preprint arXiv:1608.07639*, 2016.

[4] M. Baroni and A. Lenci. Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*, 36(4):673–721, 2010.

[5] X. Chen and A. Gupta. Webly supervised learning of convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1431–1439, 2015.

[6] X. Chen, A. Shrivastava, and A. Gupta. Neil: Extracting visual knowledge from web data. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1409–1416, 2013.

[7] X. Chen et al. Learning a recurrent visual representation for image caption generation. *arXiv:1411.5654*, 2014.

[8] D. Das. *Semi-Supervised and Latent-Variable Models of Natural Language Semantics*. PhD thesis, CMU, 2012.

[9] J. Devlin, S. Gupta, R. Girshick, M. Mitchell, and C. L. Zitnick. Exploring nearest neighbor approaches for image captioning. *arXiv preprint arXiv:1505.04467*, 2015.

[10] S. Divvala, A. Farhadi, and C. Guestrin. Learning everything about anything: Webly-supervised visual concept learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3270–3277, 2014.

[11] H. Fang et al. From captions to visual concepts and back. *arXiv:1411.4952*, 2014.

[12] A. Farhadi et al. Describing objects by their attributes. In *CVPR*, 2009.

[13] C. J. Fillmore et al. Background to framenet. *International Journal of lexicography*, 2003.

[14] N. FitzGerald et al. Semantic role labelling with neural network factors. In *EMNLP*, 2015.

[15] A. Frome et al. Devise: A deep visual-semantic embedding model. In *NIPS*, 2013.

[16] Y. Gao, O. Beijbom, N. Zhang, and T. Darrell. Compact bilinear pooling. *arXiv preprint arXiv:1511.06062*, 2015.

[17] D. Gildea and D. Jurafsky. Automatic labeling of semantic roles. *Computational linguistics*, 28(3):245–288, 2002.

[18] S. Guadarrama et al. Youtube2text: Recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition. In *ICCV*, 2013.

[19] S. Gupta and J. Malik. Visual semantic role labeling. *arXiv preprint arXiv:1505.04474*, 2015.

[20] M. Hodosh et al. Framing image description as a ranking task: Data, models and evaluation metrics. *JAIR*, 2013.

[21] Y. Jia et al. Caffe: Convolutional architecture for fast feature embedding. *arXiv:1408.5093*, 2014.

[22] A. Karpathy et al. Deep visual-semantic alignments for generating image descriptions. *arXiv:1412.2306*, 2014.

[23] S. Kazemzadeh, V. Ordonez, M. Matten, and T. L. Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 787–798, 2014.

[24] C. H. Lampert, H. Nickisch, and S. Harmeling. Attribute-based classification for zero-shot visual object categorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(3):453–465, 2014.

[25] C. H. Lampert et al. Learning to detect unseen object classes by between-class attribute transfer. In *CVPR*, 2009.

[26] A. Lazaridou et al. Is this a wampimuk? In *ACL*, 2014.

[27] T. Lei, Y. Zhang, R. Barzilay, and T. Jaakkola. Low-rank tensors for scoring dependency structures. Association for Computational Linguistics, 2014.

[28] J. Lei Ba, K. Swersky, S. Fidler, et al. Predicting deep zero-shot convolutional neural networks using textual descriptions. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4247–4255, 2015.

[29] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*. 2014.

[30] T.-Y. Lin, A. RoyChowdhury, and S. Maji. Bilinear cnn models for fine-grained visual recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1449–1457, 2015.

[31] T.-Y. Lin et al. Microsoft coco: Common objects in context. In *ECCV*. 2014.

[32] 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.

[33] J. Mao et al. Explain images with multimodal recurrent neural networks. *arXiv:1410.1090*, 2014.

[34] G. A. Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

[35] V. Ordonez et al. Im2text: Describing images using 1 million captioned photographs. In *NIPS*, 2011.

[36] M. Ronchi and P. Perona. Describing common human visual actions in images. In *British Machine Vision Conference (BMVC)*, 2015.

[37] 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*, 2014.

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

[39] C. Silberer et al. Grounded models of semantic representation. In *EMNLP*, 2012.

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

[41] R. Socher, A. Perelygin, J. Y. Wu, J. Chuang, C. D. Manning, A. Y. Ng, and C. Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the conference on empirical methods in natural language processing (EMNLP)*, volume 1631, page 1642. Citeseer, 2013.

[42] O. Vinyals et al. Show and tell: A neural image caption generator. *arXiv:1411.4555*, 2014.

[43] S. Yang, Q. Gao, C. Liu, C. Xiong, S.-C. Zhu, and Y. J. Chai. Grounded semantic role labeling. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 149–159. Association for Computational Linguistics, 2016.

[44] M. Yatskar, L. Zettlemoyer, and A. Farhadi. Situation recognition: Visual semantic role labeling for image understanding. In *Conference on Computer Vision and Pattern Recognition*, 2016.

[45] M. Yatskar et al. See no evil, say no evil: Description generation from densely labeled images. **SEM*, 2014.

[46] B. Zhou, Y. Tian, S. Sukhbaatar, A. Szlam, and R. Fergus. Simple baseline for visual question answering. *arXiv preprint arXiv:1512.02167*, 2015.