Grounding Semantic Roles in Images

Carina Silberer[†♣]
carina.silberer@upf.edu

Manfred Pinkal[†]
pinkal@coli.uni-saarland.de

[†]Department of Computational Linguistics
Saarland University, Saarbrücken, Germany
[♣]Universitat Pompeu Fabra
Barcelona, Spain

Abstract

We address the task of visual semantic role labeling (vSRL), the identification of the participants of a situation or event in a visual scene, and their labeling with their semantic relations to the event or situation. We render candidate participants as image regions of objects, and train a model which learns to ground roles in the regions which depict the corresponding participant. Experimental results demonstrate that we can train a vSRL model without reliance on prohibitive image-based role annotations, by utilizing noisy data which we extract automatically from image captions using a linguistic SRL system. Furthermore, our model induces frame–semantic visual representations, and their comparison to previous work on supervised visual verb sense disambiguation yields overall better results.

1 Introduction

Images of everyday scenes can be interpreted and described in many ways, depending on the perceiver and the context in which the image is presented. The latter may be natural language data or a visual sequence. As an example, consider the two scenes in Figure 1 and the question *What is the man doing?* The interpretation of the first target image (left) in isolation would allow many answers. Taking into account the visual context, however, may disprove many of those answers (e.g., *He is questioning the women.*). For the target image on the right, the reason for *Why there is so much food on the table?* can be inferred from its textual context.

As the examples illustrate, the interpretation of a (visual) scene is related to the determination of its events, their participants and the roles they play therein (i.e., distill *who did what to whom, where, why and how*), and this may require a joint processing or reasoning with possibly multiple (extra-)linguistic information sources (e.g., text, images). In NLP, the well-established and studied task of semantic role labeling (SRL) aims to extract such knowledge in the form of shallow semantic structures from natural language texts (e.g., *questioning(Agent:man, Theme:women)*); see, e.g., Gildea and Jurafsky (2002); Palmer et al. (2010), for an overview). It is considered an essential task towards text understanding, and was shown to be beneficial for applications such as information extraction (see Roth and Lapata (2016) and the references therein) and question answering (Shen and Lapata, 2007). In computer vision research, recent efforts have been made on *visual* SRL or *situation recognition*, a task coined by transferring the use of semantic roles to produce similar structured meaning descriptions for visual scenes (e.g., Yang et al. (2016); Yatskar et al. (2016)). To facilitate the endeavor of joint processing over multiple sources, it is desirable to induce representations of texts and visual scenes which do encode this kind of information, and in, essentially, a congruent and generic way. The latter would furthermore support the induction of a desired level of abstraction as needed.

In this paper we propose an approach towards this goal: We address the task of visual SRL (vSRL) and learn frame–semantic representations of images. Specifically, we present a model that learns to ground the semantic roles of a *semantic frame* in image *regions*, which may be crucial for, e.g., human-robot interaction or surveillance (e.g., *Who/Where* is the robber?). For example, the image shown in Figure 2 evokes the ARREST frame, and its semantic roles `Authorities`, `Suspect`, and `Place` are grounded in the image regions (delineated by bounding boxes) which depict their corresponding fillers. While being trained on this task, our model learns distributed *situation representations* (for images and frames), and *participant representations* (for image regions and roles) which

Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 2616–2626
Brussels, Belgium, October 31 - November 4, 2018.

Figure 1: Example images along with their visual (left) or textual (right) contexts.

capture the visual–frame-semantic features of situations and participants, respectively.

We train our model on data that we automatically extract by running a linguistic SRL system on image captions—human produced data that is abundant and requires less time and expertise than frame-semantic annotations. Supervised SRL has suffered from data sparsity since it relies on labor-intensive human annotations. Analogous issues on manually annotated images have been addressed by Yatskar et al. (2017). By leveraging existing efforts made in NLP, we explore whether we can alleviate the supervision bottleneck in visual SRL

Our experiments yield promising results, and our models are even able to make correct predictions for erroneous data points. Furthermore, we evaluate the induced situation representations on the task of supervised visual verb sense disambiguation, where it outperforms or is comparable to previous work (on motion or non-motion verbs, respectively).

2 Related Work

Yatskar et al. (2016) introduced the *ImSitu* dataset for the task of *situation recognition*, i.e., the problem of, given an image, predicting a structured output which specifies the depicted activity (e.g., *jumping*) and its associated semantic roles paired with their nominal fillers (e.g., {*(agent, bear), (obstacle, water)*}. To address the task, Yatskar et al. (2016, 2017) train conditional random field (CRF) models on ImSitu (Yatskar et al., 2016) and on additional training data for rarely occurring noun-role combinations which they source from the web (Yatskar et al., 2017). Mallya and Lazebnik (2017) assume that the roles associated with each activity are in a fixed order, and treat the above task as one of recognizing activities and generating a sequence of nouns, for which they use a recurrent neural network. They show how hereby learned features can be transferred to tackle image caption generation. Li et al. (2017) explicitly model role dependencies through a gated graph neural network. Given an image, they instantiate a fully connected graph with a verb and its roles as nodes. Each node's hidden state vector is initialized with image features from two CNNs, which were pre-trained for the prediction of verbs and nouns, respectively. Using a softmax layer augmented with hidden state vectors, they predict the verb and the nominal fillers of its roles.

In contrast to above works on ImSitu, we do not link the roles of a verb to their lexical fillers. We address the related task of explicitly grounding roles in the corresponding image regions, since our focus is on the relation between semantic roles and the typical visual features of their fillers (e.g., a `Body_part` is typically not a *bike* but *arms*). Gupta and Malik (2015) introduced this task as *visual semantic role labeling*. Similarly, Yang et al. (2016) formulate a CRF that jointly processes a cooking video and its natural language descriptions in order to ground the semantic roles associated with the verbs in corresponding object tracks. Both of these studies are limited to a small number of activities performed by people and a few semantic roles (26 and 11 verbs, 3 and 6 roles, respectively).

Unlike related work, our approach does not rely on manual role annotations of images, but exploits a linguistic SRL system for data creation. With more than 1k frame-specific roles, our data is of a larger scope than Gupta and Malik (2015) and Yang et al. (2016). Further, unlike the CRF-based approaches, our model induces frame-semantic representations during training.

ARREST		PLACING	
r_1, r_2	Authorities	r_1	Agent
r_5	Suspect	r_5	Theme
r_3	Place	r_3	Place
		r_4	Goal

target outputs

Figure 2: Example image (from *Flickr30k Entities*) augmented with frame-semantic annotations. Top: Image with objects rendered by bounding boxes. Bottom: Annotations which show the frames which the image evokes, and their roles, linked to their filler objects.

3 Grounding Semantic Roles in Images

We first define the task of vSRL and then present our model and our approach for data creation.

3.1 Task Definition: vSRL

Our approach is based on the linguistic theory of frame-semantics (Fillmore, 1982), which underlies the idea that words evoke semantic frames. Frames describe prototypical situations or events and contain semantic roles. For example, in the sentence *They arrested him for assault*, the argument *they* fills the `Authorities` role, *him* is the `Suspect`, and *assault* the `Charges` of the ARREST frame, which was evoked by the verb *arrest*.

Let F be a set of frames, E be the set of all semantic role labels, and E_f be the inventory of roles associated with the frame f (e.g., $E_{\text{ARREST}} = \{$`Authorities`, `Suspect`, `Charges`, `Offense`, `Place`$\}$)[1]. Assume we are given an image i, which evokes a frame f, and a set of image regions R_i, which render one or several objects in i. The task of vSRL is to link each role $e \in E_f$ to the object $r \in R_i$ that fills role e in the situation or event which f describes. We call a role e to be *realized* in an image, if it can be grounded in an image (region). The object r shown in the image region is called the *filler* or *realization* of e. The structure $A_f = \{(r, e) | r \in R_i, e \in E_f\}$ overall represents the frame f in the image i.

In SRL, the task of identifying the frame which a *predicate* evokes is a prerequisite, but it is usually treated as a subtask of SRL. We follow this approach and consider the identification of the frames evoked by an image as a subtask of vSRL. We formulate two further subtasks for vSRL, namely *role prediction*—determining the correct role for a relevant image region, and *role grounding*—linking a realized role to its filler.

Note that not all roles of a frame may be realized in an image, and not all objects may play a role in an evoked frame. Figure 2, for instance, shows an image with some of its objects delineated by six bounding boxes $R_i = \{r_1, r_2, r_3, r_4, r_5, r_6\}$. The target outputs (bottom, Fig. 2) are the frames ARREST and PLACING, as well as their realized roles which are aligned with their fillers (marked by colors). The FrameNet roles `Charges` and `Offense` are not realized in the image, i.e., they cannot be grounded. The vehicle, box r_4, in turn, does not participate in the ARREST frame.

3.2 Model: Visual-Frame–Semantic Embedder

Our model, illustrated in Figure 3, is formulated as a neural network architecture. Its input is a tuple $q = (i, r, f, e) \in Q$ of an image i, an object which is delineated by bounding box r, a frame $f \in F$, and a role label $e \in E_f$ (e.g., $q =$ (img$_1$, r_5, ARREST, `Suspect`); cf. Fig. 2). The model output is a score $s(q) \in [-1, 1]$ which quantifies the visual–frame-semantic correspondence between the box r and the role e of f (Fig. 3, right).

More specifically, the model maps visual encodings of i and r (e.g., vectors of a pre-trained CNN), and frame-semantic representations of f and e (randomly initialized embeddings) to common visual–frame-semantic spaces (*cross-modal layers* in Fig. 3).

We assume that images capture different frame-semantic features than image regions—an image encodes the whole scene and its participants and thus evokes a frame, while individual image regions of participants capture the participant-specific features of the semantic roles they fill. We therefore distinguish between two different cross-modal spaces: a situation space for images and frames, and a participant space for regions and roles. Using the respective representations in these spaces, the model then estimates the situation similar-

[1] We use FrameNet 1.5 (Ruppenhofer et al., 2006).

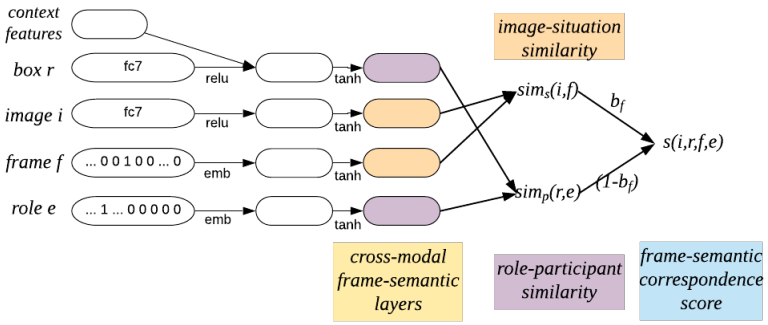


Figure 3: The **ImgObjLoc** model which scores the correspondence between a semantic role and its frame, respectively, and a candidate role filler (an image region) and the whole image, respectively.

ity, $\text{sim}_s(i, f)$, between the image and the frame, and the participant-role similarity, $\text{sim}_p(r, e)$, between the box and the role. Finally, the overall frame-semantic score $s(q)$ is the aggregation of sim_s and sim_p:

$$\text{s}(q) = b_f \, \text{sim}_s(i, f) + (1 - b_f) \, \text{sim}_p(r, e), \quad (1)$$

where parameter $b_f \in \theta$ weights the contribution of the situation and participant scores to the overall score and is learned along all model parameters θ.

By definition of the output function s (Equ. 1), each role-object pair is scored independently of the decisions made for the other roles and regions of the same frame and image, respectively. Technically, this allows for the use of partially labeled training data, where not every realized role of a frame has been linked to its filler, as we will explain in Section 3.3.

Below we describe how we use our model to address the subtasks of role prediction and grounding (Section 3.1), respectively, for which we will report experimental results in Section 5.[2] In any case, the method is based on the visual-frame–semantic correspondence $\text{s}(q)$ (Equ. 1), where we discard all candidates of role-filler pairings with a score less than zero.

Role Prediction Given an image i, we formulate the role prediction problem as a mapping L:

$$\begin{aligned} L : &\{i\} \times R_i \rightarrow F \times E \\ L(i, r) = &\underset{(f,e), f \in F, e \in E_f}{\arg\max} \; s(i, r, f, e) \end{aligned} \quad (2)$$

That is, the predicted role (and the frame it is associated with) which an image region $r \in R_i$ of i fills is that $e \in E$ to which r is most similar in the visual–frame-semantic space.

[2]We refer to Appendix A.2 in the supplemental material for the production of the structure $A_{i,f}$ of all role–filler pairs for a frame f evoked by image i.

Role Grounding is the equivalent to linguistic semantic role labeling.[3] Given a frame f realized in i, we ground each role $e \in E_f$ in the region $r \in R_i$ with the highest visual–frame-semantic similarity to e:

$$\begin{aligned} G : &\{i\} \times \{f\} \times E_f \rightarrow R_i \\ G(i, f, e) = &\underset{r \in R_i}{\arg\max} \, s(i, r, f, e) \end{aligned} \quad (3)$$

Training We train the model by using a ranking criterion designed to give higher scores to true cross-modal frame-semantic combinations (i, r, f, e) than to mismatches, by a margin M. To this end, for each positive example $q = (i, r, f, e)$ of a training set Q, we sample K negative examples $q'_k = (i, r, f', e')$ of a frame f' and role $e' \in E_{f'}$ not true for image i and box r,[4] and learn model parameters θ by minimizing the maximum margin hinge loss function on the tuples (q, q') (Equ. 4) . Ideally, using this loss function would guide the parameter learning towards mapping images and the frames they evoke, and regions and the roles they fill, respectively, nearby each other in the cross-modal spaces.

$$\theta = \arg\min_{\theta} \sum_{q \in Q} \frac{1}{K} \sum_{k=1}^{K} \max(0, M - \text{s}(q) + \text{s}(q'_k)) \quad (4)$$

Margin M is found during hyperparameter optimization on a validation set.

[3]More formally, the task of SRL is the determination of the arguments and their semantic roles of a predicate in a sentence.

[4]We could extend the model to also sample a negative image and box for f and e, and a negative role r' for f that is filled by another box in the image. We refrain from this since we create our training data from automatically labeled data, which hence could contain erroneous role-filler pairs.

(1a) [r5 A man] is being placed in [r4 a police car] by [r1 a uniformed officer].	(2a) PLACING (Theme:r5/A man, Goal:r4/a police car, Agent:r1/a uniformed officer)
(1b) [r1,r2 The police] arresting [r5 someone] on [r3 a busy city street].	(2b) ARREST (Authorities:r1,r2/The police, Suspect:r5/someone, Place:r3/on a busy city street)
(1c) [r5 A young guy] is getting arrested.	(2c) ARREST (Suspect:r5/A young guy)

Figure 4: Flickr30k captions for the image in Fig. 2. Left: Flickr30k Entities annotations of the mentioned objects with unique entity ids. Right: Frame-semantic annotations of the sentences, output by PathLSTM (Roth, 2016; Roth and Lapata, 2016).

3.3 Using Linguistic Knowledge for Data Creation

SRL systems in NLP research use training data which have been carefully created by linguistic experts (e.g., Ruppenhofer et al. (2006); Palmer et al. (2005)) for many years. To train our model on the *visual* SRL task, we build upon the annotation efforts made in NLP. The exploitation of existing resources which were developed for the analogous goal means to get around the time-consuming and costly annotation effort involved in the creation of training data. Moreover, adopting an established framework in NLP for shallow semantic representations (FrameNet, Ruppenhofer et al. (2006), in our case), including the therein defined frame and role labels, could facilitate cross-modal interactions—advances in vSRL can help to improve SRL and vice versa, or jointly draw inferences from both modalities (e.g., a text and its illustration).

Our data creation approach is to use a (linguistic) SRL system to extract frame-semantic annotations from a corpus of images paired with captions. We use the Flickr30k Entities dataset (Plummer et al., 2015)[5] which contains 30k images and five captions per image. We chose this dataset since its captions are augmented with entity mention annotations, associating them with the 276k manually annotated bounding boxes (i.e., entities are grounded in the image). To create the set $Q = \{(i^{(j)}, r^{(k_j)}, f^{(l_j)}, e^{(l_j,k_j)}) | j \in \{1, \ldots, 30k\}\}$ of training instances, we run PathLSTM (Roth, 2016; Roth and Lapata, 2016) on all captions, and extract all semantic frame annotations whose roles are filled by a grounded entity. As a result, our training corpus comprises images, the frames they evoke, and the associated semantic roles paired with their grounded fillers (i.e., bounding boxes).

Sentences (1a)–(1c) in Figure 4 (left), for example, are three human produced captions for the image in Figure 2, in which entity mentions are linked to their image regions (indicated by colors). Using PathLSTM, we extract the grounded frame-semantic annotations (2a)–(2c) (Fig. 4, right), which results in the following six instances of our corpus Q:

(img$_1$, r_5, PLACING, Theme)
(img$_1$, r_1, PLACING, Agent)
(img$_1$, r_4, PLACING, Goal)

(img$_1$, r_1_r_2, ARREST, Authorities)
(img$_1$, r_5, ARREST, Suspect)
(img$_1$, r_3, ARREST, Place)

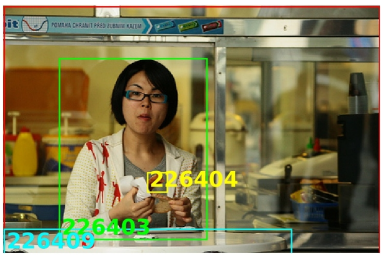
4 Data

Training Data We adopt the training, validation and test splits provided in the Flickr30k Entities dataset (Plummer et al., 2015) and create our dataset Q with the method described above. Some verbs and the frame types which they evoke occur very frequently in the set of annotations (e.g., BEING_LOCATED) and therefore allow the induction of a finer-grained frame inventory. Specifically, we transform each frame which is evoked by an individual *verb* (e.g., *stand* or *sit*) for at least 100 images (as obtained from the captions) in the Flickr30k Entities training split to a finer-grained frame type by concatenating it with the verb (e.g., BEING_LOCATED-*sit*). Finally, we keep all *frame types* (fine-grained or coarse) which had been assigned to at least 100 different images. This amounts to an inventory of 252 frame types (102 coarse types, e.g., STATEMENT), 1,409 frame-specific role types (e.g., STATEMENT.speaker), 169 role labels (e.g., Speaker) and 76,939 training instances. We derive our validation and test splits from the original splits on the basis of above modifications. See Table 1 for the quantitative details on the dataset, which we henceforth call *Flickr30k Roles*.

[5] See web.engr.illinois.edu/~bplumme2/Flickr30kEntities

	# inst.	# types				
	frame.role	frame.role		frames		roles
		fine	coarse	fine	coarse	
train	76,939	1,409	426	252	102	169
val	7,171	755	421	239	102	143
test	7,229	756	426	242	102	146

Table 1: Overview of our *Flickr30K Roles* dataset.



INGESTION-*eat*			1–1
Role	Entity ID	(obj. name)	
Ingestor	226403	woman	1
Ingestibles	226404	lunch/hotdog	1
Place	226408	at_diner	1
Ingestor	226409	table/at_table	0
instances			*label*

Figure 5: Automatically derived instances in *Flickr30k Roles* (colored, left columns) and the human correctness judgments of the frame, verb, and role fillers (right-most column; 1 is *correct*, 0 *wrong*). The object names were presented to facilitate the annotation, but are not part of the instance.

Reference Data *Flickr30k Roles* may contain false instances due to its creation on the basis of automatic frame–semantic annotations. Im-Situ (Yatskar et al., 2016) is, to the best of our knowledge, the only existing benchmark dataset for vSRL. As explained in Section 2, however, it is image-based, and does not provide explicit links between roles and the regions which depict their fillers. It cannot be used for the evaluation of role prediction and grounding without additional annotations.

We therefore created a set of reference instances by presenting a subset of the *Flickr30k Roles* test data to two human subjects (both students of computational linguistics) for annotation. We chose all instances which agree in their frame label with instances extracted from at least two other captions of the underlying image. This amounts to 201 images and 715 instances. The annotators were presented with an image with relevant objects rendered by bounding boxes, along with the automatically grounded semantic frame annotations. Figure 5 gives an example image along with the 4 automatically obtained instances. They were asked to judge the correctness of the frame (e.g., INGESTION, Fig. 5), the verb (in the case of a fine-grained frame type; e.g., *eat*) and each of the role–filler links (e.g., `Ingestor`–226403). They further linked wrong role assignments to their correct fillers when possible. We created the reference set as the intersection of all correct instances of the two annotators (frame and role–filler linkings), which amounts to 554 instances.

Visual Representations We use high-dimensional distributed vectors to represent images and regions (bounding boxes), and represent the latter by additional contextual features. These encode a region's relative location and size with respect to the whole image (cf. (Mao et al., 2016)):

$$\left[\frac{x_{tl}}{W}, \frac{y_{tl}}{H}, \frac{x_{br}}{W}, \frac{y_{br}}{H}, \frac{w \cdot h}{W \cdot H}\right], \quad (5)$$

where (x_{tl}, y_{tl}) and (x_{br}, y_{br}) are the coordinates of the top left and bottom right corners of the bounding box, H and W are the height and width of the image, and h and w the height and width of the box. These features have been found useful for referring expression generation/interpretation for objects in images (Mao et al., 2016). We hypothesize that the relative position and size of an object can be likewise informative for the roles it can(not) realize. For example, an object that is located at the bottom of an image is probably rather the `Patient` of a KICKing event than the `Agent`.

5 Experiments

We first evaluate our model in terms of different aspects related to visual SRL on the two subtasks role prediction and grounding (see Section 3).

Our second experiment assesses the usefulness of the learned frame-semantic image representations on the task of visual verb disambiguation: given an image and a verb, assign the correct sense of the verb, i.e., the one that describes the action depicted in the image (e.g., *play* an instrument; *play* sport). This task is different from visual SRL, but forms a prerequisite for it, since in frame semantics, roles are defined on the basis of frames evoked by verb *senses*.

Model Details For each bounding box and image, we use the VGG16 network (Simonyan and Zisserman, 2014), trained on ImageNet (Deng et al., 2009), to extract a $4,096$-dimensional feature vec-

		Fine-grained frame types							Coarse frame types			
		top-1-pred.			top-5 preds.			gt fr.	top-1-pred.		top-5 preds.	
		frame	fr.role	role	frame	fr.role	role	role	frame	fr.role	frame	fr.role
test set	Image-only	19.0	9.4	16.7	44.1	28.6	52.3	47.9	23.7	12.0	55.8	36.3
	ImgObject	18.7	12.8	24.1	44.9	33.8	61.2	64.3	22.6	15.5	55.5	41.4
	ImgObjLoc	18.6	13.5	25.9	46.8	35.7	62.2	65.7	23.0	16.7	56.5	43.2
reference	Image-only	**27.8**	13.2	17.2	55.2	39.3	57.3	50.2	**30.8**	14.6	67.8	46.6
	ImgObject	22.6	15.7	22.4	59.6	44.3	66.9	69.0	25.1	16.7	**68.8**	51.0
	ImgObjLoc	24.9	**17.4**	**23.6**	**60.2**	**47.3**	**68.6**	**70.3**	28.4	**19.7**	67.4	**53.3**

Table 2: Role prediction accuracy on the *Flickr30k Roles* test data and on its human corrected subset.

tor from the fully connected fc7 layer. To transform the feature vectors into the visual–frame-semantic embedding space, we use two two-layer networks which are composed of a layer with rectified linear activation units (relu) followed by a layer with tanh activations (see Fig. 3, top left). We furthermore concatenate the first hidden layer (relu layer) of each image region (i.e., box) with a vector of contextual features (relative box size and location, Equ. 5).

Frames and roles, in turn, are encoded as one-hot vectors and mapped to randomly initialized embedding layers, which are then transformed into the visual–frame-semantic representations using tanh activation layers (Fig. 3, bottom). We use the cosine similarity to quantify visual-frame–semantic correspondences in the cross-modal space (Equ. 1).

Throughout our experiments we compare our model (**ImgObjLoc**), which takes into account the contextual features (Equ. 5), to a model that does not use contextual box features (**ImgObject**), and one that only uses the image as visual input (**Image-only**). **Image-only** derives its cross-modal *role* representation by augmenting both, the image *and* the box input layers with the image's fc7 feature vector.

The network parameters were optimized using AdaGrad (Duchi et al., 2011) with a learning rate of 0.003. We monitored the role prediction performance on the validation set of Flickr30k Roles and kept the best performing model. See Appendix A.1 for further details on the model hyperparameters.

5.1 Exp.1: Semantic Role Prediction and Local Grounding

In the role prediction evaluation, the model is given an image and a bounding box, which represents a candidate role filler, and needs to predict the frame and role which the entity (or entities) in the box fills.

In the grounding experiment, the model is given an image, a frame and an associated role which is realized in the image, and needs to determine the correct role filler from a list of boxes. We report results on using ground truth boxes as well as box proposals, extracted with selective search (Uijlings et al., 2013). Regarding the latter, we apply the intersection over union (IoU) metric (e.g., Everingham et al. (2010)), and consider a role to be grounded in the correct box proposal $\tilde{r}$ if the area of overlap between $\tilde{r}$ and the reference box, divided by the area of their union, exceeds 50%.

Results We report top-1 and top-k accuracy (i.e., the frame and role is among the top-k scored predictions) on the *Flickr30k Roles* test and reference sets for both subtasks (recall that Flickr30k Entities provides ground truth alignments between entity mentions and objects).

Table 2 gives the results on role *prediction* with ground truth bounding boxes (i.e., for all entities which fill at least one semantic role). We report the accuracy for predicting the correct frame and role (columns fr.role), for predicting the correct frame (columns frame), and the correct role regardless of its frame (columns role; e.g., a prediction of STATEMENT.Speaker would be considered correct even if the reference was SPEAK_ON_TOPIC.Speaker). We further give results for the coarse frame types, where verbs are stripped off the frame labels (i.e., STATEMENT-*speak* is STATEMENT). Since the role prediction performance is equal for both frame types, we report the results for the fine-grained frames only.

As Table 2 shows, the models which use participant representations extracted from the relevant image regions (**ImgObject** and **ImgObjLoc**) perform better than **Image-only** which considers the

			Fine-grained frame types							Fine-grained frame types					
			top-1 pred. filler			top-3 pred. fillers				top-1 pred. filler			top-3 pred. fillers		
			frame	fr.role	role	frame	fr.role	role		frame	fr.role	role	frame	fr.role	role
test set	Random	gt	37.7	23.6	25.3	70.8	56.5	59.4	props	5.5	3.7	4.1	15.7	10.6	11.6
	ImgObject	gt	55.9	55.1	58.0	83.2	84.0	78.7	props	10.5	11.3	11.7	21.8	21.4	21.2
	ImgObjLoc	gt	56.6	56.6	59.4	83.1	85.1	79.7	props	11.5	12.8	13.3	22.3	22.6	22.5
reference	Random	gt	54.7	25.7	25.7	91.7	65.5	65.5	props	8.1	3.8	3.8	22.9	11.8	11.8
	ImgObject	gt	78.9	62.1	62.1	95.8	88.2	83.6	props	13.7	12.8	12.8	39.6	30.9	28.2
	ImgObjLoc	gt	**80.8**	**63.9**	**63.9**	**97.9**	**91.8**	**86.4**	props	**18.6**	**16.9**	**16.9**	**43.8**	**35.5**	**34.6**

Table 3: Role grounding accuracy on the *Flickr30k Roles* test data and on its human corrected subset. Instances with less than 2 (for top-1) or 3 (for top-3) gt filler candidates were discarded.

PathLSTM	Roles		
high prec.	`Ingestor`	`Source`	
	`Carrier`	`Speaker`	
low prec.	`Body_part`	`Activity`	`Seller`
	`Buyer`	`Manner`	`Purpose`

Table 4: Roles which were most difficult to predict by **ImgObjLoc**, in the order of their total frequency in the reference set (top left to bottom right), distinguished by the prediction precision of PathLSTM.

global image only, except for the top-1 frame prediction. This indicates that the two models are able to learn useful role-specific visual representations. Contextual features in the form of the relative size and location of a region (cf. Equ. 5) seems to be also beneficial, due to **ImgObjLoc** yielding the overall best results.

These features are furthermore beneficial for role *grounding* in automatically selected bounding boxes: When using automatically selected boxes, **ImgObjLoc** is significantly more effective than **ImgObject** in all settings (rows props, right block in Table 3). The Random baseline, which assigns each role randomly to a box in the image, performs unsurprisingly worst.

Interestingly, the models perform substantially better on the reference set than on the noisy test set (top and bottom blocks in Tables 2,3).[6] This indicates that they were able to generalize over wrong role-filler pairs in the training data, and are able to make correct predictions even for erroneous instances (see the qualitative analysis below). When assuming that the correct frame has been identified (columns gt fr.), the best role prediction accuracy reaches 70.3% on the reference set, and grounding accuracy with box proposals is at 35.5% (**ImgObjLoc**, Tables 2,3, respectively).

Finally, frame prediction proves to be a difficult task, especially for fine-grained frame types (e.g., BEING_LOCATED-*sit*; left block in Table 2).

Qualitative Analysis Notably, our analysis revealed that **ImgObjLoc** could correctly predict roles for cases in which PathLSTM failed, especially for highly visual entities (e.g., `performance` vs. `location`, `goal` vs. `path`). Overall, **ImgObjLoc** was often able to identify location roles which PathLSTM had missed, but may confuse the specific labels (e.g., `area` vs. `path` or `location`) for reasons discussed below. See Figure 6 for the recall of **ImgObjLoc** on the reference set for individual roles (top-20).

In an error analysis of the predictions of **ImgObjLoc** we identified several classes of errors. Typical errors in *role* prediction were in cases in which an image region contained multiple objects, and the system predicted a label for an object which was occluded by the target or vice versa (e.g., `ingestibles` vs. `source`; `clothing` vs. `wearer` or `body_part`; `path` vs. `area`). We found that this error was propagated from noise in the training data. Table 4 shows the roles which were most difficult to predict by **ImgObjLoc**, and which the textual SRL system (PathLSTM) could predict with a high precision (top; as calculated from the human annotations, cf. Section 4), or with a low precision (bottom), respectively. As may be expected, among these are also highly non-visual roles, such as `manner` and `purpose`.

Other noise propagated from the training data was caused by wrong frame predictions of PathLSTM (e.g., TRAVERSE-*pass* instead of BRINGING-*carry*; CONTAINING-*hold*.`contents` vs. IN-

[6] The accuracy scores on the uncorrected instances in the reference set yield comparable or worse accuracy scores than those on the test set, except for the top-5 predicted frames.

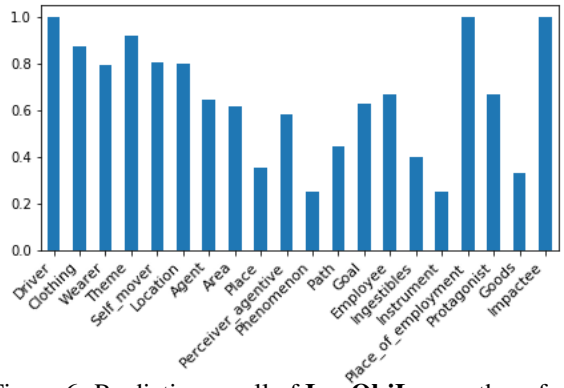


Figure 6: Prediction recall of **ImgObjLoc** on the reference set for the top-20 roles, ordered by their frequency.

Features	Motion	Non-motion
Random	76.7 ± 0.86	78.5 ± 0.39
MFS$^+$	76.1	80.0
CNN$^+$	82.3	80.0
Gella–CNN+O$^+$	83.0	80.0
Gella–CNN+C$^+$	82.3	80.3
CNN (reproduced)	83.1	79.8 ± 0.53
ImgObjLoc	**84.8** ± **0.69**	**80.4** ± **0.57**

Table 5: Sense prediction accuracy for motion (left) and non-motion verbs (right) using different image representations. $^+$ marks results taken from Gella et al. (2018). MFS is the most frequent sense heuristic.

GESTION.ingestibles). Frequent patterns of incorrect frame predictions were furthermore a failure of the system to distinguish between fine-grained frames (e.g., BEING_LOCATED-*sit* vs. -*lie* or SELF_MOTION-*walk* vs. -*run*), or between motion and non-motion actions (e.g., POSTURE vs. SELF_MOTION).

Finally, we observed that often the reference did not contain an actually valid frame which had been predicted by the system for an image, due to different levels of frame specificity, i.e., the output of **ImgObjLoc** was more specific (e.g., ASSISTANCE-*help*.helper vs. WEARING.wearer; OPERATE_VEHICLE-*ride*.vehicle vs. PERCEPTION_ACTIVE-*look*.location_of_perceiver) or it was more general (e.g., WEARING.wearer vs. IMPACT-*hit*.impactor).

5.2 Exp.2: Visual Verb Sense Disambiguation

We evaluate the effectiveness of the frame-semantic *image* representations that can be extracted with our **ImgObjLoc** model on the VerSe (visual Verb Sense disambiguation) dataset (Gella et al., 2018). It covers 90 verbs and 163 senses used to annotate 3,510 images. We follow the supervised method applied in (Gella et al., 2018), divide VerSe into training and test data, and train logistic regression classifiers for sense prediction on 19 motion verbs and 19 non-motion verbs (those which have at least 20 images and at least 2 senses). Input to the sense classifiers are the frame-semantic image representations (second top cross-modal layer in Fig. 3) of the VerSe images, which we extract with the **ImgObjLoc** model, trained on Flickr30k Roles.

Table 5 gives the mean accuracy obtained on the test data (of 100 runs). Our **ImgObjLoc** vectors outperform all comparison models on motion verbs, including CNN-based image features and the best-performing models of (Gella et al., 2018), namely Gella–CNN+O and Gella–CNN+C (CNN features concatenated with predicted object labels and image captions, respectively). On non-motion verbs, the best models, including our own, perform only comparably to the most frequent sense heuristic. Note that we examine the simplest representation **ImgObjLoc** can yield, i.e., frame-semantic representations for individual images. More complex representations are left for future work. See Appendix A.3 for examples.

6 Conclusions

We addressed the task of grounding semantic roles of frames which an image evokes in the corresponding image regions of its fillers. We found that our model can be trained without the need of manual role annotations of image data, and that the frame-semantic image representations it learns can be used for related tasks. Encouraged by our findings, future work includes the exploration of the model and its learned frame-semantic representations for tasks such as the interpretation of multimodal scenes and stories and referring expressions.

Acknowledgments

We thank the anonymous reviewers, Leonie Harter, Christine Schäfer, Michael Roth, Gemma Boleda, Anna Rohrbach and Bernt Schiele. This research was supported by the German Research Foundation (DFG EXC 285), by the European Research Council (ERC Horizon 2020 grant agreement No 715154), and the Spanish Ramon y Cajal programme (grant RYC-2015-18907). This paper reflects the authors' view only, and the EU is not responsible for any use that may be made of the information it contains.

References

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A Large-scale Hierarchical Image Database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. IEEE.

John Duchi, Elad Hazan, and Yoram Singer. 2011. Adaptive Subgradient Methods for Online Learning and Stochastic Optimization. *The Journal of Machine Learning Research*, 12:2121–2159.

Mark Everingham, Luc Gool, Christopher K. Williams, John Winn, and Andrew Zisserman. 2010. The Pascal Visual Object Classes (VOC) Challenge. *International Journal of Computer Vision*, 88(2):303–338.

Charles J. Fillmore. 1982. *Frame Semantics*. Hanshin Publishing Co., Seoul, South Korea.

S. Gella, F. Keller, and M. Lapata. 2018. Disambiguating Visual Verbs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. In press.

Daniel Gildea and Daniel Jurafsky. 2002. Automatic Labeling of Semantic Roles. *Computational Linguistics*, 28(3):245–288.

Saurabh Gupta and Jitendra Malik. 2015. Visual Semantic Role Labeling. *CoRR*, abs/1505.04474.

Ruiyu Li, Makarand Tapaswi, Renjie Liao, Jiaya Jia, Raquel Urtasun, and Sanja Fidler. 2017. Situation Recognition With Graph Neural Networks. In *The IEEE International Conference on Computer Vision (ICCV)*.

Arun Mallya and Svetlana Lazebnik. 2017. Recurrent Models for Situation Recognition. *arXiv preprint arXiv:1703.06233*.

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L. Yuille, and Kevin Murphy. 2016. Generation and Comprehension of Unambiguous Object Descriptions. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational Linguistics*, 31(1):71–106.

Martha Palmer, Daniel Gildea, and Nianwen Xue. 2010. Semantic Role Labeling. volume 3, pages 1–103. Morgan & Claypool Publishers.

Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. *CoRR*, abs/1505.04870.

Michael Roth. 2016. Improving Frame Semantic Parsing via Dependency Path Embeddings. In *Book of Abstracts of the 9th International Conference on Construction Grammar*, pages 165–167, Juiz de Fora, Brazil.

Michael Roth and Mirella Lapata. 2016. Neural Semantic Role Labeling with Dependency Path Embeddings. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1192–1202, Berlin, Germany.

Josef Ruppenhofer, Michael Ellsworth, Miriam R.L. Petruck, Christopher R. Johnson, and Jan Scheffczyk. 2006. *FrameNet II: Extended Theory and Practice*. International Computer Science Institute, Berkeley, California. Distributed with the FrameNet data.

Dan Shen and Mirella Lapata. 2007. Using Semantic Roles to Improve Question Answering. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*.

Karen Simonyan and Andrew Zisserman. 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition. *CoRR*, abs/1409.1556.

Jasper RR Uijlings, Koen EA Van De Sande, Theo Gevers, and Arnold WM Smeulders. 2013. Selective Search for Object Recognition. *International Journal of Computer Vision*, 104(2):154–171.

Shaohua Yang, Qiaozi Gao, Changsong Liu, Caiming Xiong, Song-Chun Zhu, and Joyce Y. Chai. 2016. 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, San Diego, California. Association for Computational Linguistics.

Mark Yatskar, Vicente Ordonez, Luke Zettlemoyer, and Ali Farhadi. 2017. Commonly Uncommon: Semantic Sparsity in Situation Recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Mark Yatskar, Luke Zettlemoyer, and Ali Farhadi. 2016. Situation Recognition: Visual Semantic Role Labeling for Image Understanding. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

A Supplemental Material

A.1 Model Parameters

We optimized the network parameters using AdaGrad (Duchi et al., 2011) with a learning rate of 0.003. Dropout with rate 0.25 was added on top of the visual–frame-semantic layers. We monitored the role prediction performance on the validation

Algorithm 1 vSRL algorithm which grounds each semantic role e of a frame f^* in at most one region $r \in R_i$ of image i. $s(.)$ denotes the visual-frame–semantic correspondence score (Equation (1)).

Require: a frame f^* for image i, R_i // Equ. (6)
  $S_{i,f^*} \leftarrow ((s(i,r,f^*,e),r,e) : r \in R_i, e \in E_{f^*})$, sorted in descending order of $s(.)$
  $A_{i,f^*} \leftarrow \emptyset$, grounded realization of f^*
  for all $(s,r,e) \in S_{i,f^*}$ **do**
    if $(r,.) \notin A_{i,f^*} \wedge (.,e) \notin A_{i,f^*}$ **then**
      $A_{i,f^*} \leftarrow A_{i,f^*} \cup \{(r,e)\}$
    end if
  end for
  return A_{i,f^*}

set of Flickr30k Roles and kept the best performing model, which was obtained after about 20 epochs for each model. For all models, the first visual hidden layer has 1000 dimensions, and all other layers have 250 units. The margin M (Equation 4 in the main paper) was set to 0.3, and $K = 10$ negative frame-role examples were sampled for each training instance.

A.2 Model: vSRL

The full vSRL task requires, given an image i, the computation of the set $A_{i,f}$ of role-object pairs which comprises the semantic roles of a frame f grounded to their fillers, i.e., $A_{i,f} = \{(r,e)|r \in R_i, e \in E_f\}$. Using our model, we first determine the frame f^* which image i evokes on the basis of all role-filler predictions, i.e.,

$$f^* = \arg\max_{f \in F} \sum_{\{r \in R_i, e \in E_f\}} s(i,r,f,e) \quad (6)$$

We then apply a simple algorithm (Algorithm 1) which chooses the filler–role pairs with maximum similarity from the set S_{i,f^*} of all scored frame-specific filler-role pairings for image i given frame f^* (cf. Equation 1 in the main paper), such that every role is grounded in at most one region, and every region fills at most one role (line 6, Algorithm 1).

A.3 Examples for VSD

Figure 7 shows example images of non-motion verbs for which ImgObjLoc achieved a high (*serve*, 95%) and a low accuracy (*reach*, 50%), their ground truth senses (gt) and the predictions of **ImgObjLoc** (sys).


serve (95%):
dish out, hand out something, often food
sys: ✓


serve (95%):
put a ball into play
sys: ✓


reach (50%)
extend physically/by influence
sys: ✓


reach (50%)
extend physically/by influence
sys: ✓


reach (50%):
gt: *pass or transfer something*
✗sys: *attain/arrive at a state, real or abstract*


reach (50%):
gt: *pass or transfer something*
✗sys: *extend physically/by influence*

Figure 7: VSD: Example images of non-motion verbs, their verb senses (gt) and our system's predictions (sys).