CRAPES:Cross-modal Annotation Projection for Visual Semantic Role Labeling

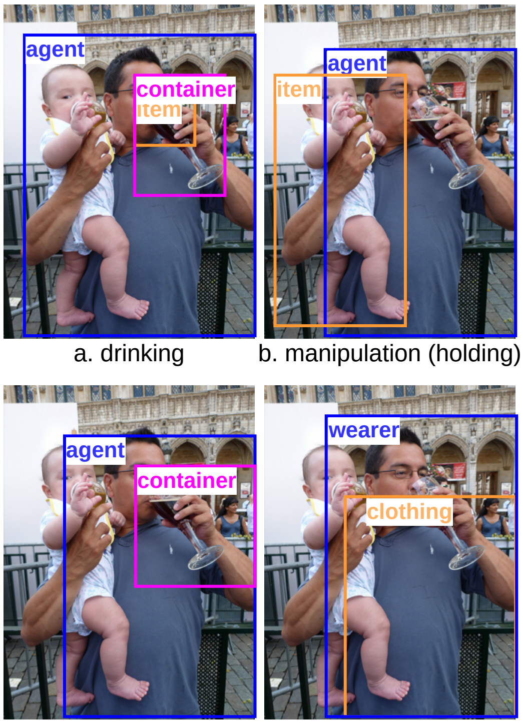
Abhidip Bhattacharyya and **Martha Palmer** and **Christoffer Heckman**
University of Colorado Boulder
firstname.lastname@colorado.edu

Abstract

Automatic image comprehension is an important yet challenging task that includes identifying actions in an image and corresponding action participants. Most current approaches to this task, now termed **G**rounded **S**ituation **R**ecognition (GSR), start by predicting a verb that describes the action and then predict the nouns that can participate in the action as arguments to the verb. This problem formulation limits each image to a single action even though several actions could be depicted. In contrast, text-based **S**emantic **R**ole **L**abeling (SRL) aims to label all actions in a sentence, typically resulting in at least two or three predicate argument structures per sentence. We hypothesize that expanding GSR to follow the more liberal SRL text-based approach to action and participant identification could improve image comprehension results. To test this hypothesis and to preserve generalization capabilities, we use general-purpose vision and language components as a front-end. This paper presents our results, a substantial 28.6 point jump in performance on the SWiG dataset, which confirm our hypothesis. We also discuss the benefits of loosely coupled broad-coverage off-the-shelf components which generalized well to out of domain images, and can decrease the need for manual image semantic role annotation.

1 Introduction

Automatic image comprehension can positively contribute to many modern applications, such as description generation, cross-modal retrieval, and human-robot interaction. To comprehend an image it is important to identify the action(s) and participants in the action such asan agent (who is performing the action), a patient (who is being affected by the action), and an instrument. To address this problem (Yatskar et al., 2016b; Pratt et al., 2020) proposed the task of grounded situation recognition (GSR). Many approaches (Pratt et al., 2020; Cooray et al., 2020; Cho et al., 2021) have been proposed to perform the task of GSR. Most of these frameworks have two steps: in the first step verbs are predicted, and in the second step nouns and roles are predicted in an auto-regressive manner. Some other methods deployed include another layer to refine the quality of detection (Cho et al., 2021; Wei et al., 2021; Cheng et al., 2022).



a. drinking b. manipulation (holding)

c. drinking d. wearing

Figure 1: a. depicts a GT example from SWiG where the man is the agent of drinking. b., c., and d. show frames extracted by our method. Bounding boxes depict grounding and role annotation for each frame.

One fundamental limitation of these models derives from the problem formulation. In the current formulation, verb frames would compete for an image, limiting the expressiveness of the image's semantic representation. In reality, various actions can co-exist in an image, even sharing participants. This limitation of one frame per image is imposed by the predominant dataset of GSR: the SWiG dataset (Pratt et al., 2020). For example, Fig-

*Proceedings of the The 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)*, pages 61–70
July 13-14, 2023 ©2023 Association for Computational Linguistics

ure 1a depicts a ground-truth (GT) annotation of an image from SWiG and has a GT annotation only with respect to a drinking frame. In fact, there are other frames, such as holding, wearing.

Semantic role labeling (SRL) of natural text, on the other hand, is a well researched problem in the domain of computational linguistics. Semantic role annotation, based on paradigms such as PropBank or Framenet (Palmer et al., 2005; Fillmore et al., 2003), is used to train semantic parsers that then convey knowledge about *who is doing what to whom, when* as predicate-argument structure labeling. In other words, given an action in a sentence, it identifies who is performing the action (the agent), who is affected by the action (the patient), what instrument is being used, etc. to comprehend the meaning of the sentence. Semantic roles of a sentence have the capability representing more than one predicate-argument structure for that sentence. Current text-based SRL systems have gained remarkable accuracy. However, SRL of images has yet to enjoy similar success.

We hypothesize that expanding GSR to follow the more liberal text-based SRL approach to action, participant identification could improve image comprehension results. Here, we propose a framework (CRAPEs) with cross-modal *annotation projection* (AP) for visual semantic role labeling. AP is a well-known paradigm in text-based cross-lingual semantic role labeling (Kozhevnikov and Titov, 2013; Padó and Lapata, 2009; Akbik et al., 2015; Jindal et al., 2022) that has not been previously extended to cross-modal applications. Moreover, to preserve generalization capabilities, we focus on reusing general-purpose vision and language (V+L) components and text-based SRL components. This framework offers the following advantages over traditional GSR approaches:

- With our updated formulation of GSR, this framework can be trained to accommodate co-existing verb frames in an image. It can also be specialised to one verb frame per image.
- Additionally, image representations can be learned separately from the SRL task; in doing so, CRAPEs can leverage advantages of large-scale multi-modal image representations.
- Success of text-based SRL systems trained on large, broad-coverage corpora of frames and roles, is helpful in widening its ability for detecting out-of-domain frames.
- Moreover the two modules can be trained separately, thereby decreasing the need for manual image semantic role annotation.
- As image representation and SRL are not tightly coupled, CRAPEs can be extended to alternative semantic role labeling paradigms, such as FrameNet or PropBank.

2 Related Work

(Yatskar et al., 2016b) proposed the task of situation recognition (SR) together with an image situation recognition dataset (imSitu). Based on the architecture, methods for SR can be stratified into the following categories: 1) Conditional random field (*CRF*) (Yatskar et al., 2016b), 2) CRF-based model with data augmentation (*CRF+dataAug*) (Yatskar et al., 2016a), 3) RNN model with a VGG backbone for vision features (*VGG+RNN*) (Mallya and Lazebnik, 2017), 4) *graph based models* (Li et al., 2017; Suhail and Sigal, 2019), and 5) query based models such as *CAQ* (Cooray et al., 2020).

The idea of grounding nouns in the image was coined by (Pratt et al., 2020), thereby proposing the task of GSR and the SWiG dataset. A recurrent framework with ResNet-50 embedding was used to detect the verb and then the noun for each role. A RetinaNet backbone was used for object grounding. (Cooray et al., 2020; Cho et al., 2021) model visual SRL as query based vision reasoning. (Cooray et al., 2020) adopt a top-down attention model (Anderson et al., 2018) and deploy inter-dependent queries to model relations among semantic roles. (Cho et al., 2021) use a transformer encoder to classify verbs and to create image representations. Then the image representation was queried with the concatenation of roles and verbs. However, most of these aforementioned approaches use two-stage frameworks where in the first step the verb is predicted independently and then nouns and roles are predicted in an autoregressive manner depending on the verb. However, subsequent work (Cho et al., 2022; Wei et al., 2021) identified that this emphasis on the detection of the verb may confuse the prediction. Furthermore, verb miss-classification may result in miss-recognition of semantic roles.

Therefore, they adopted a three-stage framework. In the first two stages candidate verbs and nouns were detected. The third stage mostly refined the prediction. During the detection of the candidate, information flows either from verb to noun (Wei et al., 2021) or from noun to verb (Cho et al., 2022). This ignores the semantic dependency in the other

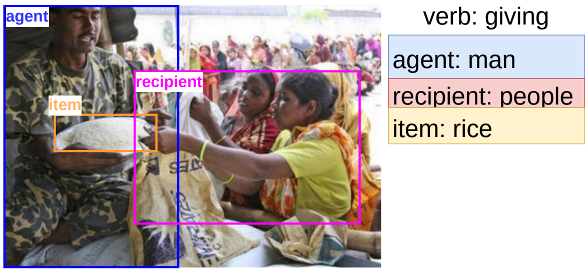


Figure 2: An example of GSR from the SWiG dataset.

direction. Moreover, this refinement can be done in only one iteration. (Cheng et al., 2022) solved these issues by designing an iterative method through message passing between verb and noun prediction modules. Recently, (Li et al., 2022) addressed the task of GSR, even though their main goal was to propose a pre-training schema using event based cross-modal alignment. All of these methods are limited to predicting one verb per image. None of these models acknowledge the existence of multiple actions and therefore multiple verb frames.

3 Approach

To detect semantic roles in images we adopted the idea of AP, as discussed above, from cross-lingual semantic role labeling in the text domain. In AP, auto-predicted semantic roles from source language is transferred to a target language using soft word alignments. Alignment is learned using large-scale parallel corpus. In the case of GSR we consider the image as our target domain.

3.1 Problem Formulation

Given an image $\mathcal{I}$ the task of GSR is to detect structured verb frame(s) $\mathcal{G} = \{v, \mathcal{R}_v\}$ where $v \in V$ is the action (verb) in the image. $\mathcal{R}_v = \{(r_v, n^r, b_v^r) | r_v \in \mathscr{R}_v, n^r \in \mathcal{N}, b_v^r \in \mathbb{R}^4\}$ where $\mathscr{R}_v = \{r_v^1, .., r_v^m\}$ set of semantic role types associated with the verb v. Therefore, each role is a triplet of a role type r_v, a noun label n^r and a bounding box (bbox), b_v^r that is grounded with respect to the v and the role of the noun n^r . For example in Figure 2 the given image is annotated with the verb "giving". The verb has role types *agent*, *recipient* and *item*. The nouns for these roles are `man`, `people` and `rice`, respectively.

Issues with current approaches. As discussed above in section 2, current methods (Pratt et al., 2020; Cho et al., 2021; Li et al., 2017) modeled this problem as:

$$\mathcal{P}(\mathcal{G}|\mathcal{I}) = \mathcal{P}(v|\mathcal{I})\mathcal{P}(\mathcal{R}_v|v, \mathcal{I}). \tag{1}$$

There are two complications with this kind of formulation: first, action prediction without knowledge of participants results in inaccurate verb prediction. Second, errors in verb prediction can adversely affect accuracy of noun and role prediction. To address this issue, recent methods (Wei et al., 2021; Cho et al., 2021) adopt a three stage framework. (Wei et al., 2021) formulated the problem as given in Equation 2:

$$\begin{gathered}\mathcal{P}(\mathcal{G}|\mathcal{I}) = \mathcal{P}(V_c|\mathcal{I})\mathcal{P}(\mathcal{R}_{\mathcal{V}_c}|V_C, \mathcal{I}) \\ \mathcal{P}(v, \mathcal{R}_v|V_C, \mathcal{R}_{\mathcal{V}_c}\mathcal{I}).\end{gathered} \tag{2}$$

In this formulation candidate verbs are detected first, then candidate nouns. In the final stage these candidates are used to refine the final result. (Cho et al., 2021) on the other hand, used candidate nouns to detect the verb and ultimately refined the frame predictions (Equation 3):

$$\begin{aligned}&\mathcal{P}(\mathcal{G}|\mathcal{I}) = \\ &\quad \mathcal{P}(\mathcal{N}_{\mathcal{V}_c}|\mathcal{I})\mathcal{P}(v|\mathcal{N}_{\mathcal{V}_c}, \mathcal{I})\mathcal{P}(v, \mathcal{R}_v|\mathcal{N}_{\mathcal{V}_c}, \mathcal{I}).\end{aligned} \tag{3}$$

Both the approaches used nouns to determine the verb at some point, ignoring the restrictions applied in the other direction. Moreover, even with these revised formulations, verbs compete with each other for a given image. On contrast, in a scene image more than one verb can coexist.

3.2 Methodology

To overcome the limitation imposed by the traditional formulation, we propose an alternative formulation given as:

$$\begin{aligned}\mathcal{P}(\mathcal{G}|\mathcal{I}) &= \sum_i \mathcal{P}(\mathcal{G}_i|\mathcal{I}) \\ &= \underbrace{\mathcal{P}(\mathcal{T}|\mathcal{I})}_{\text{V+L}} \underbrace{\sum_i \mathcal{P}(\mathcal{G}_i|\mathcal{T}, \mathcal{I})}_{\text{SRL}}\end{aligned}$$

To capture the complete essence of the intertwined relations of a verb and its roles, we use a V+L model which creates a text-based holistic representation $\mathcal{T}$ using self-attention. Text-based SRL then extracts all possible predicate-argument structures. The soft alignments from the V+L model is used to project the SRL back to the image (Figure 3). To

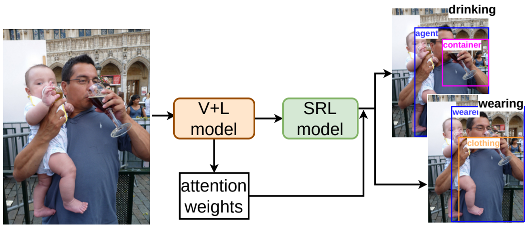


Figure 3: Our overall framework. 1. The V+L model projects the image into the text domain. The SRL annotator detects the semantic roles and the action. Attention from the V+L model is used to align semantic roles

preserve generalization capabilities, we used off-the-shelf general-purpose V+L components and a text based SRL system. Being trained on data outside the SWiG dataset, this framework has more potential to detect out-of-domain frames.

3.3 Pipeline

Our framework has two modules: 1) V+L model, and 2) text-based SRL system. (refer to Figure 3)

V+L. We chose Oscar (Li et al., 2020; Zhang et al., 2021) to this end. Oscar is a transformer based architecture that learns generic image-text representations for V+L understanding and generation tasks. Typically Oscar model would take three inputs- word tokens, object labels and object features. One of the novelties of Oscar lies in the notion of the 'view' of the data during pre-training. In a dictionary view elements from similar semantic spaces are considered together (words and object labels). On the other hand, in the modality view elements from the same modality are considered together. We trained Oscar with image region features $\mathcal{I} = \{(\varsigma_i, l_i) | \varsigma_i \in \mathbb{R}^d \; l_i \in \Sigma \; d = 2056\}$. We used (Zhang et al., 2021) to extract 2048 dimensional image region features and then concatenated with 6 positional features for the region (normalised coordinates of bounding boxes, height, width). Σ denotes the vocabulary for the language model. For the purpose of CRAPES, two separate models of Oscar are trained on the Flickr30k and the SWiG datasets, see Table 1. During inference the captions generated by Oscar are passed to the SRL module.

SRL. We experimented with two text based FrameNet SRL systems. For a given sentence T consisting of tokens $< t_1, t_2, .., t_k >$ a typical SRL system produces collections of verbs and their roles. Briefly $T_{srl} = \{(v, \mathcal{R}_v^T)\}$ where $\mathcal{R}_v^T$ is set of semantic roles given the verb v. It is a collection of tuples of the form $\{(r_v^i, (s_v^i, e_v^i))\}$ where $r_v^i \in \mathscr{R}_v$ is the semantic role and (s_v^i, e_v^i) marks the start and end indices of the phrase spanned by the role. For our experimentation we used an off the shelf annotator span-finder (Xia et al., 2021) for FrameNet annotation. We trained a second SRL consisting of BERT-base model with CRF at the top layer, on SWiG frames (see Table 1).

Cross-modal Annotation Projection. Our SRL system detects the semantic roles and the nouns from the text given by the V+L model. For grounding the roles to image bboxes we used attention weights from the V+L model. For each role span, corresponding cross-modal attention is retrieved from the V+L model. Attention is aggregated over all the tokens in the span:

$$\text{role}(\text{bbox}_j) = r_v^t, \text{ where}$$
$$j = \arg\max(\alpha_i) \text{ and } \alpha_i = \sum_{l,h} \alpha_{l,h}(i, s_v^t, e_v^t),$$

where l and h are spans over number of attention layers and head accordingly.

4 Experiments

4.1 Experimental Set up

Data Preparation. We experimented with SWiG (Pratt et al., 2020). SwiG provides FrameNet semantic role labeling of images. The SwiG dataset provides grounding for all visible semantic roles in terms of image bboxes. SWiG contains 126102 images with 504 verbs and 190 semantic role types, and each verb is accompanied by 1 to 6 semantic roles. The official splits are $75K/25K/25K$ images for training, dev, and test set, respectively. Unlike Flickr30k, this dataset does not have any textual image descriptions.

Data augmentation. Figure 4 presents an overview of data flow during training. To train CRAPES with SWiG, we created templates for each verb frame using roles. For each image, the corresponding verb frame and template are retrieved. Roles in the template were replaced with the corresponding noun values from the annotation of the image to generate the sentence. This sentence along with the image is used to train the V+L model, and the sentence with the roles is used to train the BERT+CRF SRL model.

Evaluation Metric. We used the following metric (Pratt et al., 2020) to report our results. 1)

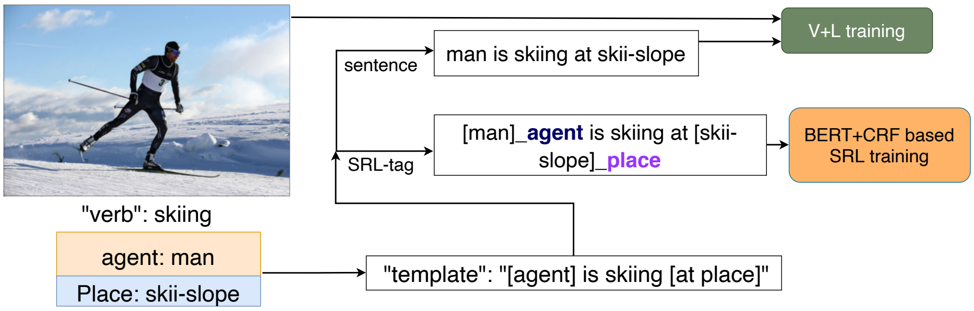

V+L training
sentence
man is skiing at skii-slope
SRL-tag
[man]_agent is skiing at [skii-slope]_place
BERT+CRF based SRL training
"verb": skiing
agent: man
Place: skii-slope
"template": "[agent] is skiing [at place]"

Figure 4: Training pipeline of CRAPES for the SWiG dataset. SWiG images are not accompanied by sentences. Using the ground truth (GT) frames, template sentences are created. The image and sentence pair is used to train the V+L model. Sentence and frames are used to train the BERT+CRF srl model

Model	**Description**	Annotation
CRAPES$_1$	Oscar with flickr, LOME framenet	FN
CRAPES$_2$	Oscar with SWiG, BERT+CRF on SWiG	FN

Table 1: Different versions of CRAPES based on training data of V+L and different SRL models. In last column FN stands for Framenet.

verb: the accuracy of verb prediction; 2) value: accuracy of noun prediction for individual roles; 3) value-all: accuracy of the prediction of nouns for the whole role set; 4) grounded-value (grnd): accuracy of noun prediction with correct grounding (bboxes) for individual semantic roles; 5) grounded-value-all (grnd-all): accuracy of noun prediction with correct grounding (bboxes) for the whole role set.

Implementation Details. We used the pre-trained Oscar base model ($H = 768$) fine-tuned for caption generation. This model was trained on the MSCOCO dataset (Lin et al., 2014). We trained two separate versions of Oscar with the Flickr30k train (Young et al., 2014) and SWiG dev datasets with an AdamW Optimizer (Loshchilov and Hutter, 2019) for 20 epochs with learning rate 3×10^{-5}. We trained the text-based BERT+CRF SRL system on the template generated sentences of the train split of the SWiG dataset.

4.2 Quantitative Results

A quantitative comparison with recent approaches on the SWiG benchmark based on both SR and GSR is presented in Table 2, using the categorization from section 2. We report our results on SWiG with the top-1 set up. CRAPES leads in the value, value-all, and grnd metrics. CRAPES has a dramatic absolute gain of 28.6 points and relative gain of 76% in value with respect to GSRFormer, the previous SOTA. Similarly, in val-all and grnd it has a relative gain of 31% and 15% accordingly. Oscar pretraining tasks (Li et al., 2020) have a major role in these improvements. As discussed in subsection 3.3 Oscar pretraining tasks were designed around two major views on how to use object labels. The first view considered object labels as members of text modality where as the second one considered them as part of the image modality. This form of training enables OSCAR to include object labels in the generated description. These object-labels contribute toward the noun prediction task in GSR. Moreover, OSCAR fine-tuned with template generated sentences is able to replicate similar structures during inference. Similarly, our BERT+CRF based SRL parser, trained on a similar domain of sentences, is able to annotate them with semantic roles. So Table 2 *firmly supports our hypotheses about the benefits of reusing general-purpose V+L components*. However, there are still certain image-verb frame combinations that confuse our system. We discuss this in our qualitative analysis.

4.3 Discussion

Table 1 lists different versions of CRAPES. Table 3 presents performance of CRAPES on FrameNet annotation. From Table 3 apparently the performance of CRAPES$_1$ is poor. However, this version of CRAPES actually gave atomic frames and parallel frames for a given image. Because of Oscar being trained on human generated sentences and the LOME parser being trained on text corpora for FrameNet, CRAPES$_1$ is able to predict out-of-domain verbs and frames. The current metrics can not reflect this capability adequately. Fig-

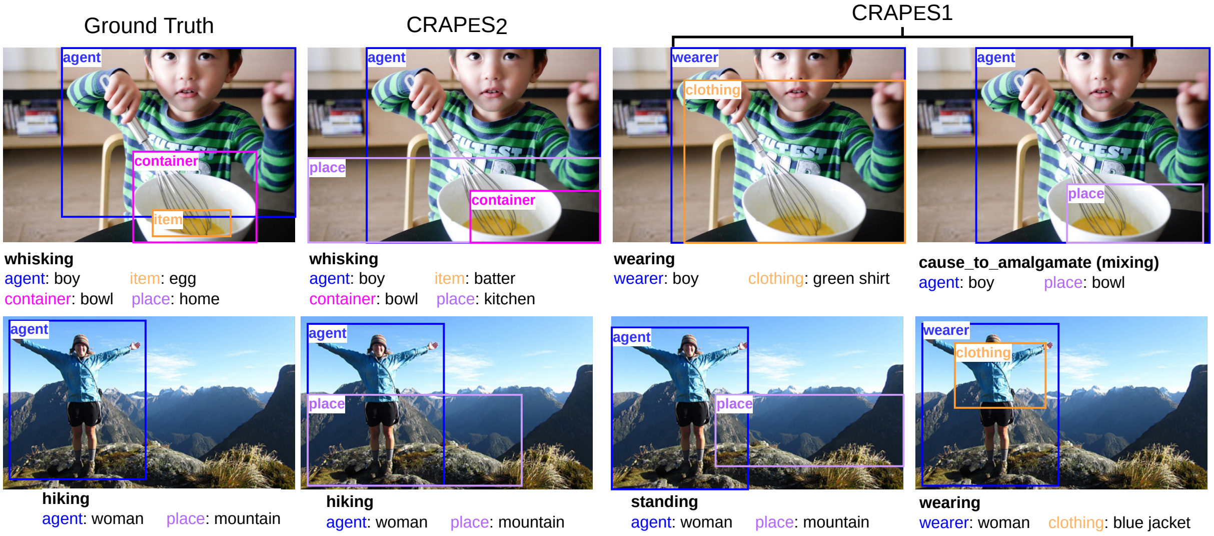


Figure 5: Examples of predictions made by CRAPES. The first column lists the GT image and frame from the SWiG test set. The second column lists the prediction from CRAPES$_2$ (V+L and SRL parser trained on SWiG). Last two columns depicts parallel frames detected by CRAPES$_1$ (V+L trained on Flickr30k and LOME parser)

Model	**value**	**val-all**	**verb**	**grnd**	**grnd-all**
situation recognition					
CRF (Yatskar et al., 2016b)	24.6	14.2	32.3	–	–
CRF+dataAug (Yatskar et al., 2016a)	26.45	15.51	34.12	–	–
VGG+RNN (Mallya and Lazebnik, 2017)	27.45	16.36	35.90	–	–
FC-Graph (Li et al., 2017)	27.52	19.25	36.72	–	–
CAQ (Cooray et al., 2020)	30.23	18.47	38.19	–	–
Kernel-Graph (Suhail and Sigal, 2019)	35.41	19.38	43.27	–	–
grounded situation recognition					
ISL (Pratt et al., 2020)	30.09	18.62	39.36	22.73	7.72
JSL (Pratt et al., 2020)	31.44	18.87	39.94	24.86	9.66
GSRTR (Cho et al., 2021)	32.52	19.63	41.06	26.04	10.44
SituFormer (Wei et al., 2021)	35.24	21.86	44.20	29.22	13.41
CoFormer (Cho et al., 2022)	35.98	22.22	44.66	29.05	12.21
CLIP Event (Li et al., 2022)	33.1	20.1	45.6	26.1	10.6
GSRFormer (Cheng et al., 2022)	37.48	23.32	46.53	31.53	14.23
CRAPES$_2$	**66.08**	**30.64**	41.86	**36.73**	6.47

Table 2: Performance (%) of state-of-the-art GSR methods on the SWiG dataset test set based on top-1 verb.

Model	**value**	**val-all**	**verb**	**grnd**	**grnd-all**
CRAPES$_1$	18.12	0.357	5.72	14.33	0.63
CRAPES$_2$	65.98	30.53	41.86	35.13	5.78
+union of BBoxes	65.98	30.53	41.86	35.13	6.1
attention from lower4 layer	66.08	30.64	41.86	36.31	6.47

Table 3: Performance (%) of SWiG test set with different combinations of V+L and Framenet parsers

Model	**grnd**	**grnd-all**
attention from top 3 layer	35.13	5.78
+ include union of boxes	35.87	6.10
attention from 5 – 8 layer	36.31	6.26
attention from all layer	36.35	6.31
attention from layer 1 – 4	36.73	6.47

Table 4: Affect of attention layers on bbox grounding reported on SWiG test set

ure 5 demonstrates examples of the frames predicted by CRAPES. Frames like `wearing` and `cause-to-amalgamate` (first row of Figure 5), will be considered as misclassifications by the current metrics with respect to GT.

However, CRAPES lags in terms of grounded-value-all. Note that this metric required that all bboxes be annotated correctly with nouns from

GT verb	Competing verbs in CRAPES$_2$
retraining	arresting, detaining, subduing, handcuffing
hunting	pouncing, shooting, chasing, attacking
teaching	lecturing, educating, helping, preaching
cooking	frying, baking, chopping, stirring, scooping
filming	videotaping, photographing, recording, carrying
raking	hoeing, shoveling, clearing, sweeping
tying	lacing, stitching, adjusting, stapling
watering	sprinkling, moistening, gardening, spraying, wetting

(a) Examples of verb confusions by CRAPES$_2$

GT verb	Co-existing verbs in CRAPES$_1$
cooking	wearing, cause-to-amalgamate, cutting, standing
baking	wearing, cause-to-amalgamate, cutting, measure_volume
teaching	wearing, standing, sitting, reading, writing, speaking
lecturing	wearing, standing, sitting, reading, talking
arresting	walking, arresting, striking, law_enforcement_agency, hostile_encounter
detaining	walking, arresting, striking, law_enforcement_agency, attacking

(b) Examples of verb co-existence detection by CRAPES$_1$

Table 5: Comparison between frame competitions and frame co-existance

the GT annotation. Therefore missing one bbox annotation can affect the metric for an image significantly. One possible reason for the poor performance could be the distribution shift between the V + L model and the SRL model. Another source of error is a limitation of the interpretability of the attention weights. To align bounding boxes with SRL we used attention between bboxes and words from Oscar attention layers. In our experiment we noticed that the 5th head from layers 5 and 6 mostly attended to bboxes. However, to our surprise, it did not provide much improvement. Attention from the lower 4 layers gave us the best result, meriting further investigation. Table 4 shows experimental results of using alignment from different attention layers.

4.4 Qualitative Results

One of the main advantages of CRAPES is that it can predict out-of-domain frames that are otherwise not present in the SWiG dataset. Figure 1 depicts one such example from SWiG where the GT annotation contains only the frame for 'drinking'. CRAPES$_1$ detects the action 'drinking' along with two other frames 'holding' and 'wearing'. These frames are not only missing in the GT image, they were not listed in the vocabulary of the SWiG dataset. The LOME FrameNet parser, trained on the FrameNet v1.7 corpus, a huge text base corpus for SRL, enables CRAPES$_1$ to detect those frames. Moreover, CRAPES can accommodate coexisting verb frames. This is because Oscar, being trained on Flickr30k sentences, learned to create holistic representations of the image. Similar examples can be found in the last two columns of Figure 5 where CRAPES$_1$ provides parallel frames, not present in the GT annotation. *This shows the efficacy of our reformulation of the GSR and the advantage of reusing general-purpose SRL systems.*

For the sake of bench-marking we trained CRAPES$_2$ with template generated sentences from the SWiG dataset. Predictions made by CRAPES$_2$ contained one frame per image as desired by the SWiG dataset. This demonstrates the flexibility of the overall framework. The second column of Figure 5 depicts some example predictions by CRAPES$_2$.

CRAPES does commit mistakes which can be categorized mainly into three types: 1) *the predicted verb is different than GT*. Figures6a,b depict two examples from the SWiG dataset where CRAPES detected a different frame. These are indeed very plausible mistakes. Table 5a shows examples of some GT verbs along with a list of verbs that CRAPES$_2$ confused with the GT verb. This fact is supported by CRAPES$_1$ as well. Table 5b lists examples of parallel verb frames detected by CRAPES$_1$ for GT images with a given verb. For example *cooking* is often confused with baking(Table 5a). From Table 5b it can be observed that both of these verbs have similar co-existing frames like cutting, cause-to-amalgamate. Similar phenomena can be noticed for *arresting* and *detaining*; 2) *predicted noun for a role is different than GT*. In the first image of column CRAPES$_2$ from Figure 5,the noun for role *item* is predicted as *batter*. 3) *grounded bbox for a noun is different than GT*. In Figure 6c the action jogging is attributed to a different bbox in the image. Mistakes made by CRAPES are reasonable, relevant and plausible. For these examples, predictions are different than the GT but still

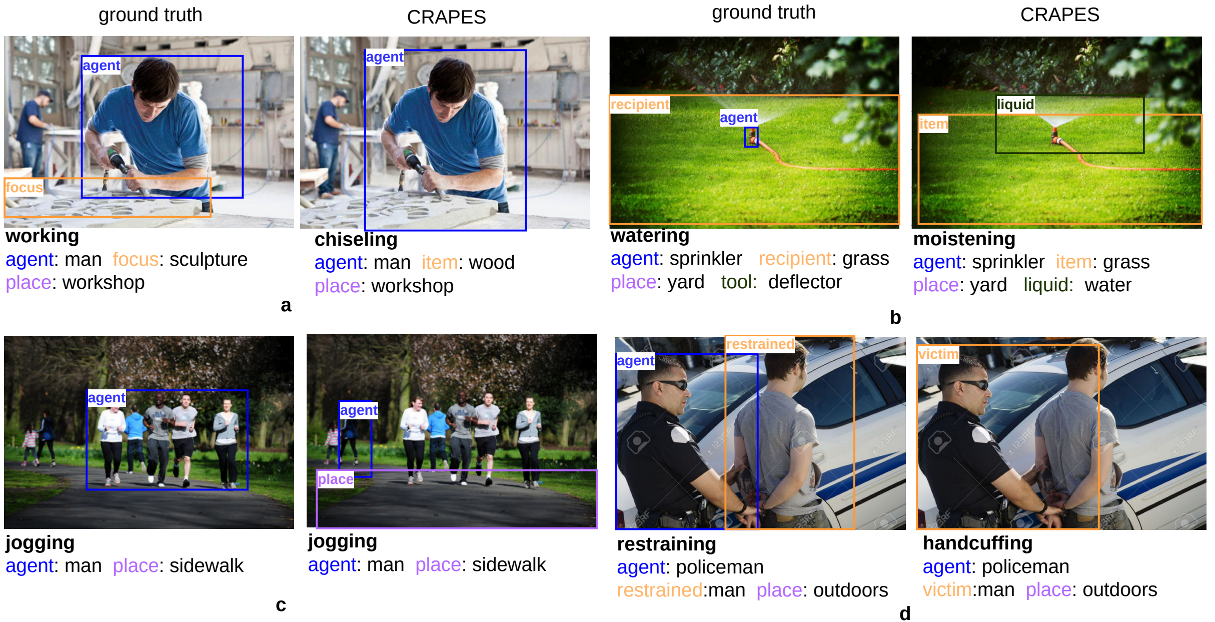


Figure 6: Reasonable mistakes made by CRAPES. For each image left column shows GT annotations and right column depicts mistakes made by CRAPES$_2$. For a,b,c prediction of CRAPES$_2$ can not be classified as wrong. For d CRAPES$_2$ struggled to detect correct bbox.

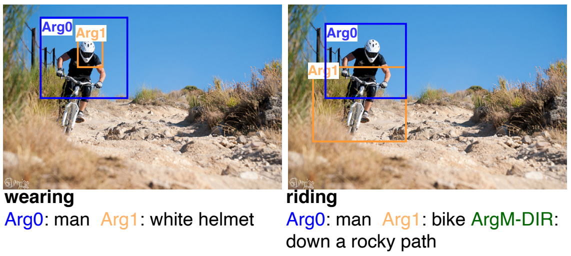


Figure 7: Parallel frames detected by CRAPES in Flickr30k images using PropBank style role labeling.

relevant to the given image. However, sometimes CRAPES struggles to ground the roles (Figure 6d).

5 Future work

Current GSR models cannot go beyond the SWiG dataset. Moreover predicted semantic roles are restricted to follow a particular paradigm of SRL. On the contrary, having independent V+L enables CRAPES to work on other image datasets. In addition, having a separate SRL module enables extension to other SRL paradigms. We performed preliminary experiments on the Flickr30k dataset with PropBank (Palmer et al., 2005) annotation. Figure 7 depicts one such example. We would like to extend our experiments to the version of Flickr30k used in (Bhattacharyya et al., 2022). However, our preliminary experiments suggest that experiments with Flickr30k are more challenging for several reasons.

- Flickr30k does not provide semantic roles for images. Therefore, we need to follow a similar approach to (Bhattacharyya et al., 2022) in creating silver standard data.
- The silver standard data will have multiple frames for an image. Current metrics of GSR presuppose one GT frame per image.
- Flickr30k images are general scene images with many agents,objects and actions, whereas images in SWiG focus mostly on one salient action and a small number of participants.
- As pointed out by (Bhattacharyya et al., 2022), PropBank annotation of Flickr30k has abstract conceptual roles such as temporal, direction, manner, purpose, etc. denoted with ArgM-. It is hard to learn concrete representations for these roles, let alone ground them in an image.

Our formulation of CRAPES can accommodate PropBank SRL experiments on Flickr30k. However, a more rigorous study with human evaluation is required to correctly measure the potential of CRAPES. Therefore, this a critical future direction for us. It requires a new dataset with images annotated with more than one frame. One choice is to extend the SWiG dataset to accommodate more than one frame per image. Another choice is to enhance the current Flickr30k annotation. Ideally we would do both. However, the current proposed evaluation metrics for GSR are incompatible wih a multi-frame scenario. More robust and appropriate evaluation metrics also need to be developed.

6 Conclusion

In this paper we identified a fundamental issue in the problem formulation of the GSR task. The current formulation limits an image to a single verb frame. We propose an alternate formulation allowing for multiple actions as implemented in **Cr**oss-modal **A**nnotation **P**rojection for Visual **Se**mantic Role Labeling (CRAPES). A V+L model trained on image-text parallel corpora and an SRL module trained independently on text corpora allow the model to integrate domain-specific knowledge with out-of-domain knowledge, which dramatically improves over the SOTA by 28.6 points. In addition, CRAPES can accommodate co-existing verb frames for an image (CRAPES$_1$) yet can also be trained to select only one verb frame for a given image (CRAPES$_2$). Moreover, inter module independence allows CRAPES to extend its labeling to alternative paradigms of SRL (such as FrameNet or PropBank). However one major area for improvement is `grnd-all`, that requires better semantic comprehension and guidance of attention weights produced by the V+L module. Therefore, improving on `grnd-all` along with Flickr30k and PropBank will be our next endeavour. We will also explore extending datasets to have multiple ground truth frames per image and more appropriate evaluation metrics for reporting results on those datasets.

7 Acknowledgements

We gratefully acknowledge the support of DARPA KAIROS Program No. FA8750-19-2-1004. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of DARPA or the U.S. government.

References

Alan Akbik, Laura Chiticariu, Marina Danilevsky, Yunyao Li, Shivakumar Vaithyanathan, and Huaiyu Zhu. 2015. Generating high quality proposition Banks for multilingual semantic role labeling. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 397–407, Beijing, China. Association for Computational Linguistics.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering.

Abhidip Bhattacharyya, Cecilia Mauceri, Martha Palmer, and Christoffer Heckman. 2022. Aligning images and text with semantic role labels for fine-grained cross-modal understanding. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4944–4954, Marseille, France. European Language Resources Association.

Zhi-Qi Cheng, Qi Dai, Siyao Li, Teruko Mitamura, and Alexander Hauptmann. 2022. Gsrformer: Grounded situation recognition transformer with alternate semantic attention refinement. In *Proceedings of the 30th ACM International Conference on Multimedia*, MM '22, page 3272–3281, New York, NY, USA. Association for Computing Machinery.

Junhyeong Cho, Youngseok Yoon, and Suha Kwak. 2022. Collaborative transformers for grounded situation recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Junhyeong Cho, Youngseok Yoon, Hyeonjun Lee, and Suha Kwak. 2021. Grounded situation recognition with transformers. In *British Machine Vision Conference (BMVC)*.

Thilini Cooray, Ngai-Man Cheung, and Wei Lu. 2020. Attention-based context aware reasoning for situation recognition. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4735–4744.

Charles Fillmore, Christopher Johnson, and Miriam Petruck. 2003. Background to framenet. *International Journal of Lexicography*, 16.

Ishan Jindal, Alexandre Rademaker, Michał Ulewicz, Ha Linh, Huyen Nguyen, Khoi-Nguyen Tran, Huaiyu Zhu, and Yunyao Li. 2022. Universal proposition bank 2.0. In *Proceedings of the Language Resources and Evaluation Conference*, pages 1700–1711, Marseille, France. European Language Resources Association.

Mikhail Kozhevnikov and Ivan Titov. 2013. Cross-lingual transfer of semantic role labeling models. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1190–1200, Sofia, Bulgaria. Association for Computational Linguistics.

Manling Li, Ruochen Xu, Shuohang Wang, Luowei Zhou, Xudong Lin, Chenguang Zhu, Michael Zeng, Heng Ji, and Shih-Fu Chang. 2022. Clip-event: Connecting text and images with event structures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16420–16429.

R. Li, M. Tapaswi, R. Liao, J. Jia, R. Urtasun, and S. Fidler. 2017. Situation recognition with graph neural networks. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 4183–4192, Los Alamitos, CA, USA. IEEE Computer Society.

Xiujun Li, Xi Yin, Chunyuan Li, Xiaowei Hu, Pengchuan Zhang, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. *ECCV 2020*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common objects in context.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Arun Mallya and Svetlana Lazebnik. 2017. Recurrent models for situation recognition. In *Proceedings - 2017 IEEE International Conference on Computer Vision, ICCV 2017*, Proceedings of the IEEE International Conference on Computer Vision, pages 455–463, United States. Institute of Electrical and Electronics Engineers Inc. Funding Information: This work was partially supported by the National Science Foundation under Grants CIF-1302438 and IIS-1563727, Xerox UAC, the Sloan Foundation, and a Google Research Award Publisher Copyright: © 2017 IEEE.; 16th IEEE International Conference on Computer Vision, ICCV 2017 ; Conference date: 22-10-2017 Through 29-10-2017.

Sebastian Padó and Mirella Lapata. 2009. Cross-lingual annotation projection of semantic roles. *J. Artif. Int. Res.*, 36(1):307–340.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The proposition bank: An annotated corpus of semantic roles. 31(1):71–106.

Sarah Pratt, Mark Yatskar, Luca Weihs, Ali Farhadi, and Aniruddha Kembhavi. 2020. Grounded situation recognition. *ArXiv*, abs/2003.12058.

Mohammed Suhail and Leonid Sigal. 2019. Mixture-kernel graph attention network for situation recognition. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10362–10371.

Meng Wei, Long Chen, Wei Ji, Xiaoyu Yue, and Tat-Seng Chua. 2021. Rethinking the two-stage framework for grounded situation recognition. *arXiv preprint arXiv:2112.05375*.

Patrick Xia, Guanghui Qin, Siddharth Vashishtha, Yunmo Chen, Tongfei Chen, Chandler May, Craig Harman, Kyle Rawlins, Aaron Steven White, and Benjamin Van Durme. 2021. LOME: Large ontology multilingual extraction. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 149–159.

Mark Yatskar, Vicente Ordonez, Luke Zettlemoyer, and Ali Farhadi. 2016a. Commonly uncommon: Semantic sparsity in situation recognition. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6335–6344.

Mark Yatskar, Luke Zettlemoyer, and Ali Farhadi. 2016b. Situation recognition: Visual semantic role labeling for image understanding. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5534–5542.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. 2:67–78.

Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. 2021. Vinvl: Making visual representations matter in vision-language models. *CVPR 2021*.