Towards Context-aware Interaction Recognition for Visual Relationship Detection

Bohan Zhuang* Lingqiao Liu* Chunhua Shen† Ian Reid
The University of Adelaide, Australia; and Australian Centre for Robotic Vision

Abstract

Recognizing how objects interact with each other is a crucial task in visual recognition. If we define the context of the interaction to be the objects involved, then most current methods can be categorized as either: (i) training a single classifier on the combination of the interaction and its context; or (ii) aiming to recognize the interaction independently of its explicit context. Both methods suffer limitations: the former scales poorly with the number of combinations and fails to generalize to unseen combinations, while the latter often leads to poor interaction recognition performance due to the difficulty of designing a context-independent interaction classifier.

To mitigate those drawbacks, this paper proposes an alternative, context-aware interaction recognition framework. The key to our method is to explicitly construct an interaction classifier which combines the context, and the interaction. The context is encoded via word2vec into a semantic space, and is used to derive a classification result for the interaction. The proposed method still builds one classifier for one interaction (as per type (ii) above), but the classifier built is adaptive to context via weights which are context dependent. The benefit of using the semantic space is that it naturally leads to zero-shot generalizations in which semantically similar contexts (subject-object pairs) can be recognized as suitable contexts for an interaction, even if they were not observed in the training set. Our method also scales with the number of interaction-context pairs since our model parameters do not increase with the number of interactions. Thus our method avoids the limitation of both approaches. We demonstrate experimentally that the proposed framework leads to improved performance for all investigated interaction representations and datasets.

1. Introduction

Object interaction recognition is a fundamental problem in computer vision and it can serve as a critical component for solving many visual recognition problems such as action recognition [2, 22, 26, 35, 39], visual phrase recognition [11, 14, 28], sentence to image retrieval [12, 20] and visual question answering [19, 36, 37]. Unlike object recognition in which the object appearance and its class label have a clear association, the interaction patterns, e.g., "eating", "playing", "stand on", usually have a vague connection to visual appearance. This phenomenon is largely caused by the same interaction being involved with different objects as its context, i.e. the subject and object of an interaction type. For example, "cow eating grass" and "people eating bread" can be visually dissimilar although both of them have the same interaction type "eating". Thus the subject and object associated with the interaction – also known as the *context* of the interaction – could play an important role in interaction recognition.

In existing literature, there are two ways to model the interaction and its context. The first one treats the combination of interaction and its context as a single class. For example, in this approach, two classifiers will be built to classify "cow eating grass" and "people eating bread." To recognize the interaction "eating", images that are classified as either "cow eating grass" or "people eating bread" will be considered as having interaction "eating". This treatment has been widely used in defining action (interaction) classes in many action (interaction) recognition benchmarks [2, 22, 26, 35, 39]. This approach, however, suffers from poor scalability and generalization ability. The number of possible combinations of the interaction and its context can be huge, and thus it is very inefficient to collect training images for each combination. Also, this method fails to generalize to an unseen combination even if both its interaction type and context are seen in the training set.

To handle these drawbacks, another way is to model the interaction and the context separately [4, 10, 18, 30]. In this case, the interaction is classified independently of its context, which can lead to poor recognition performance due

*First two authors contributed equally to this work.
†C. Shen is the corresponding author.

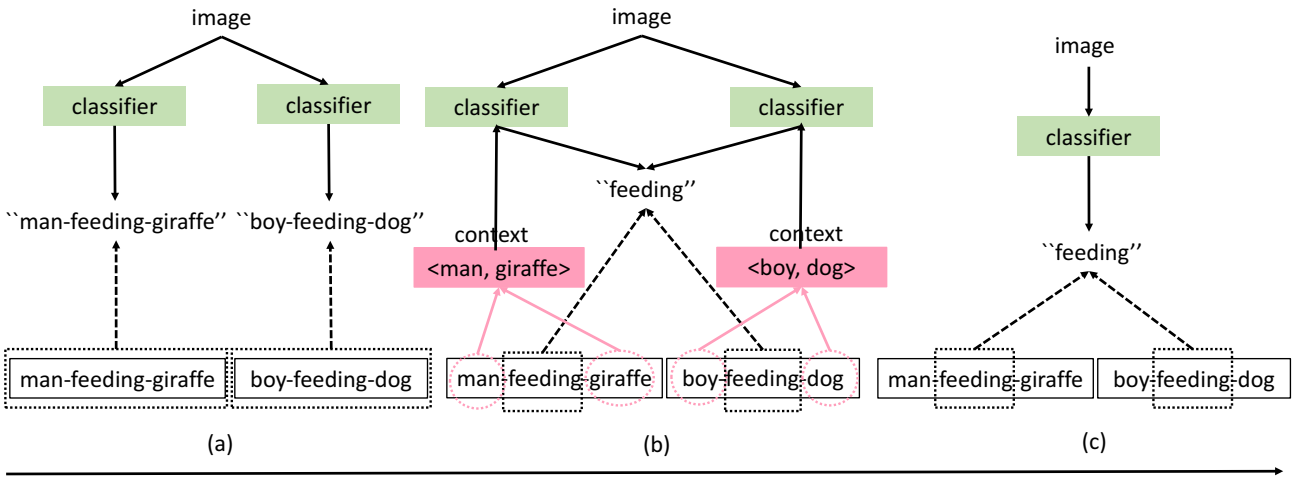

Figure 1: Comparison of two baseline interaction recognition methods and the proposed approach. The two baseline methods take two extremes. For one extreme, (a) treats the combination of the interaction and its context as a single class. For another extreme, (c) classifies the interaction separately from its context. Our method (b) lies somewhere between (a) and (c). We still build one classifier for each interaction but the classifier parameter is also adaptive to the context of the interaction, as shown in the example in (b).

to the difficulty of associating the interaction with certain visual appearance in the absence of context information. To overcome the imperfection of interaction classification, some recent works employ techniques such as language priors [18] or structural learning [14–16] to avoid generating an unreasonable combination of interaction and context. However, the context-independent interaction classifier is still used as a building block, and this prevents the system from gaining more accurate recognition from visual cues.

The solution proposed in this paper aims to overcome the drawbacks of both methods. To avoid the explosion of the number of classes, we still separate the classification of the interaction and the context into two stages. However, different to the second method, the interaction classifier in our method is designed to be adaptive to its context. In other words, for the same interaction, different contexts will result in different classifiers and our method will encourage interactions with similar contexts to have similar classifiers. By doing so, we can achieve context-aware interaction classification while avoiding treating each combination of context and interaction as a single class. Based on this framework, we investigate various feature representations to characterize the interaction pattern. We show that our framework can lead to performance improvements for all the investigated feature representations. Moreover, we augment the proposed framework with an attention mechanism, which leads to further improvements and yields our best performing recognition model. Through extensive experiments, we demonstrate that the proposed methods achieve superior performance over competing methods. Code is available at https://bitbucket.org/jingruixiaozhuang/iccv2017_vrd.

2. Related work

Action recognition: Action is one of the most important interaction patterns and action recognition in images/videos has been widely studied [2, 22, 26, 35, 39]. Various action recognition datasets such as Stanford 40 actions [38], UCF-101 [33] and HICO [3] have been proposed, but most of them focus on actions (interactions) with limited number of context. For example, in the relatively large HICO [3] dataset, there are only 600 categories of human-object interactions. Thus the interplay of the interaction and its context has not been explored in the works of this direction.

Visual relationships: Some recent works focus on the detection of visual relationships. A visual relationship is composed of an interaction and its context, i.e. subject and object. Thus this direction is most relevant to this paper. In fact, the interaction recognition can be viewed as the most challenging part of the visual relationship detection. Some recent works in visual relationship detection have made progress in improving the detection performance and the detection scalability. The work in [18] leveraged language priors to produce relationship detections that make sense to human beings. The latest approaches [14, 16, 40] attempt to learn the visual relationship detector in an end-to-end manner and explicitly reason the interdependency among relationship components at the visual feature level.

Language-guided visual recognition: Our method uses language information to guide the visual recognition. This corresponds to the recent trend in utilizing language information for benefiting visual recognition. For example, language information has also been incorporated in phrase grounding [11, 25, 28] tasks. In [11, 28], attention model is

employed to extract linguistic cues from phrases. Language guided attention has also been widely used in visual question answering [6, 12, 21, 27] and has recently been applied to one-shot learning [34].

3. Methods

3.1. Context-aware interaction classification framework

In general, an interaction and its context can be expressed as a triplet $\langle O1\text{-}P\text{-}O2 \rangle$, where P denotes the interaction, and $O1$ and $O2$ denote its subject and object respectively. In our study, we assume the interaction context ($O1$,$O2$) has been detected by a detector (i.e. we are given bounding boxes and lables for both subject $O1$ and object $O2$) and the task we are addressing is to classify their interaction type P. To recognize the interaction, existing works take two extremes in designing the classifier. One is to directly build a classifier for each P and assume that the same classifier applies to P with different context. Another takes the combination of $\langle O1\text{-}P\text{-}O2 \rangle$ as a single class and build a classifier for each combination. As discussed in the introduction section, the former does not fully leverage the contextual information for interaction recognition while the latter suffers from the scalability and generalization issues. Our proposed method lies between those two extremes. Specifically, we still allocate one classifier for each interaction type, however we make the classifier parameters adaptive to the context of the interaction. In other words, the classifier is a function of the context. The schematic illustration of this idea is shown in Figure 1.

Formally, we assume that the interaction classifier takes a linear classifier form $y_p = \mathbf{w}_p^\top \phi(I), \ \ \mathbf{w}_p \in \mathbb{R}^d$, where y_p is the classification score for the p-th interaction and $\phi(I)$ is the feature representation extracted from the input image. The classifier parameters for the p-th interaction $\mathbf{w}_p$ are a function of $(O1, O2)$, that is, the context of the p-th interaction. It is designed as the summation of the following two terms:

$$\mathbf{w}_p(O1, O2) = \bar{\mathbf{w}}_p + r_p(O1, O2), \tag{1}$$

where the first term $\bar{\mathbf{w}}_p$ is independent of the context; it plays a role which is similar to the traditional context-independent interaction classifier. The second term $r_p(O1, O2)$ can be viewed as an auxiliary classifier generated from the information of context $(O1, O2)$. Note that the summation of two classifiers has been widely used in transfer learning [1, 5, 24] and multi-task learning [7, 23], e.g., one term corresponds to the classifier learned in the target domain and another corresponds to the classifier learned in the source domain.

Intuitively, for two interaction-context combinations, if both of them share the same interaction and their contexts are similar, the interaction in those combinations tends to be associated with similar visual appearance. For example, $\langle$*boy, playing, football*$\rangle$ and $\langle$*man, playing, soccer*$\rangle$ share similar context, so the interaction "playing" should suggest similar visual appearance for these two combinations. This inspires us to design $\mathbf{w}_p(O1, O2)$ to allow semantically similar contexts to generate similar interaction classifiers, as demonstrated in Figure 2. To realize this idea, we first represent the object and subject through their word2vec embedding which maps semantically similar words into similar vectors and then generate the auxiliary classifier r_p by concatenating their embeddings. Formally, r_p is designed as:

$$r_p(O1, O2) = \mathbf{V}_p f(\mathbf{Q} E(O1, O2)), \tag{2}$$

where $E(O1, O2) \in \mathbb{R}^{2e}$ is the concatenation of the e-dimensional word2vec embeddings of $(O1, O2)$, and $\mathbf{Q} \in \mathbb{R}^{m \times 2e}$ is a projection matrix to project $E(O1, O2)$ to a low-dimensional (e.g. 20) semantic embedding space. $f(\cdot)$ is the RELU function and $\mathbf{V_p}$ transforms the context embedding to the auxiliary classifier. Note that $\mathbf{V}_p$ and $\bar{\mathbf{w}}_p$ in Eq. (1) are distinct per interaction type p while the projection matrix $\mathbf{Q}$ is shared across all interactions. All of these parameters are learnt at training time.

Remark: Many recent works [14, 16, 25, 40] on visual relationship detection takes a structural learning alike formulation to simultaneously predict $O1, O2$ and P. The unary term used in their framework is still a context-independent classifier and such choice may lead to poor recognition accuracy in identifying interaction from the visual cues. To improve these techniques, one could replace their unary terms with our context-aware interaction recognition module. On the other hand, their simultaneous prediction framework could also benefit our method in achieving better visual relationship performance. Since our focus is to study the interaction part, we do not pursue this direction in this paper and leave it for future work.

3.2. Feature representations for interactions recognition

One remaining issue in implementing the framework in Eq. (1) is the design of $\phi(I)$, that is, the feature representation of the interaction. It is clear that the choice of the feature representation can have significant impact on the interaction prediction performance. In this section, we investigate two types of feature representations to characterize the interaction. We evaluate these feature representations in Sec. 4.1.1.

3.2.1 Spatial feature representation

Our method assumes that the context has been detected and therefore the interaction between the subject and the object

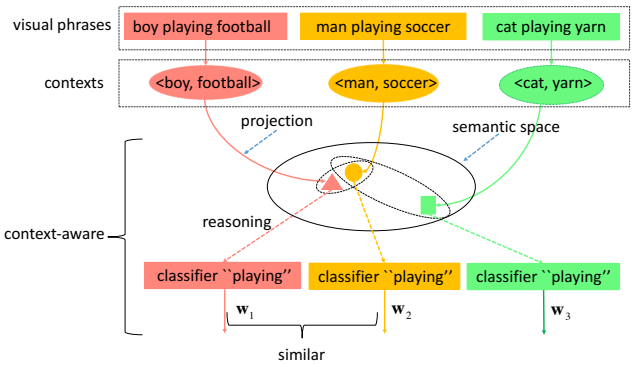

Figure 2: An example of the proposed context-aware model. The same interaction "playing" is associated with various contexts. The contexts of the first two phrases are semantically similar, resulting in two similar context-aware classifiers. Since the last two contexts are far away from each other in the semantic space, their corresponding context-aware classifiers may not similar despite sharing the same label. In this way, we explicitly consider the visual appearance variations introduced by changing context, thus more accurate and generalizable interaction classifiers can be learned.

could be characterized by the spatial features of the detection bounding boxes. These kind of features have been previously employed [11, 25, 40] to recognize the visual relationship of objects. In our study, we use both the spatial features from each bounding box and the spatial features from their mutual relationship. Formally, let (x, y, w, h) and (x', y', w', h') be the bounding box coordinates of the *subject* and *object*, respectively. Given the bounding boxes, the spatial feature for a single box is a 5-dimentional vector represented as $[\frac{x}{W_I}, \frac{y}{H_I}, \frac{x+w}{W_I}, \frac{y+h}{H_I}, \frac{S_b}{S_I}]$, where S_b and S_I are the areas of region b and image I, W_I and H_I are the width and height of the image I. And the pairwise spatial vector is denoted as $[\frac{x-x'}{w'}, \frac{y-y'}{h'}, \log\frac{w}{w'}, \log\frac{h}{h'}]$. We concatenate them together to get a 14-dimentional feature representation (using both subject and object bounding boxes). Then the spatial feature directly passes through the context-aware classifier defined in Eq. (1) for the interaction classification.

3.2.2 Appearance feature representation

Besides spatial features, we can also use appearance features, e.g. the activations of a deep neural network to depict the interaction. In our study, we first crop the union region of the subject and object bounding boxes, and rescale the region to $224 \times 224 \times 3$ as the input of a VGG-16 [32] CNN. We then apply the mean-pooling to the activations of the $conv5\_3$ layer as our feature representation $\phi(I)$. This feature is then fed into our context-aware interaction classifier in Eq. (1). To improve the performance, we treat the context-aware interaction classifier as a newly added layer and fine-tune this layer with the VGG-16 net in an end-to-end fashion.

3.3. Improving appearance representation with attention and context-aware attention

The discriminative visual cues for interaction recognition may only appear in a small region of the input image or the image region. For example, to see if "man riding bike" occurs, one may need to focus on the region near human feet and bike pedal. This consideration motivates us to use attention module to encourage the network "focus on" discriminative regions. Specially, we can replace the mean-pooling layer in Sec. 3.2.2 with an attention-pooling layer.

Formally, let $\mathbf{h}_{ij} \in R^c$ denote the last convolutional layer activations at the spatial location (i, j), where $i = 1, 2, ..., M$ and $j = 1, 2, ..., N$ are the coordinates of the feature map and M, N are the height and width of the feature map respectively, c is the number of channels. The attention pooling layer pools the convolutional layer activations into a c-dimensional vector through:

$$\begin{aligned} \bar{a}(\mathbf{h}_{ij}) &= \frac{a(\mathbf{h}_{ij})+\varepsilon}{\sum_i \sum_j (a(\mathbf{h}_{ij})+\varepsilon)}, \\ \widetilde{\mathbf{h}} &= \frac{1}{MN} \sum_{ij} \bar{a}(\mathbf{h}_{ij}) \mathbf{h}_{ij}, \end{aligned} \tag{3}$$

where $a(\mathbf{h}_{ij})$ is the attention generation function which produces an attention value for each location (i, j). The attention value is then normalized (ε is a small constant) and used as a weighting factor to pool the convolutional activations $\mathbf{h}_{ij}$. We consider two designs of $a(\mathbf{h}_{ij})$.

Direct attention: The first attention generation function is simply designed as $a(\mathbf{h}_{ij}) = f(\mathbf{w}_{att}^\top \mathbf{h}_{ij} + b)$, where $\mathbf{w}_{att}$ and b are the weight and bias of the attention model.

Context-aware attention In the above attention generation function, the attention value is solely determined by $\mathbf{h}_{ij}$. Intuitively, however, it makes sense that different attention is required for different classification tasks. For example, to examine "man riding bike" and examine "man playing football", different regions-of-interest should be focused on. We therefore propose to use a context-aware attention generator; i.e. we design $\mathbf{w}_{att}$ as a function of $(P, O1, O2)$. We can follow the framework in Eq. (1) to calculate:

$$\mathbf{w}_{att}(P, O1, O2) = \bar{\mathbf{w}}_p^{\mathbf{a}} + \mathbf{V}_p^a f(\mathbf{Q}E(O1, O2)), \tag{4}$$

where $\bar{\mathbf{w}}_p^{\mathbf{a}}$ is the attention weight for the p-th interaction independent of its context and $\mathbf{V}_p^a$ transforms the semantic embedding of the context to the auxiliary attention weight for the p-th interaction. Note that in this case $\mathbf{w}_{att}$ depends on the interaction class P and therefore different attention-pooling vectors $\widetilde{\mathbf{h}}_p$ will be generated for different P. $\widetilde{\mathbf{h}}_p$ will be then sent to the context-aware classifier for interaction P to obtain the decision value for P and the class that

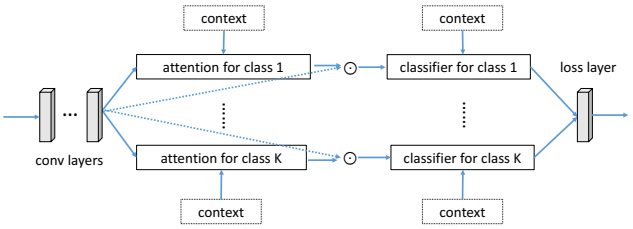

Figure 3: Detailed illustration of the context-aware attention model. For each interaction class, there is a corresponding attention model imposed on the feature map to select the interaction-specific discriminative feature regions. Different attention-pooling vectors will be generated for different interaction classes. The generated pooling vector will be then sent to the corresponding context-aware classifier to obtain the decision value.

produces the maximal decision value will be considered as the recognized interaction. This structure is illustrated in Figure 3.

3.4. Implementation details

For all the above methods, we use the standard multi-class cross-entropy loss to train the models. The Adam algorithm [13] is applied as the optimization method. The methods that use appearance features involve convolutional layers from the standard VGG-16 network together with some newly added layers. For the former we initialize those layers with the parameters pretrained on ImageNet [29] and for the latter we randomly initialize the parameters. We set the learning rate to 0.001 and 0.0001 for the new layers and VGG-16 layers respectively.

4. Experiments

To investigate the performance of the proposed methods, we analyse the effects of the context-aware interaction classifier, the attention models and various feature representations. Eight methods are implemented and compared:

1. "**Baseline1-app**": We directly fine-tune the VGG-16 model to classify the interaction categories. Inputs are the union of subject and object boxes. This baseline models the interaction and its context separately, which corresponds to the approach described in Figure 1 (c).

2. "**Baseline1-spatial**": We directly train a linear classifier to classify the spatial features described in Sec. 3.2.1 into multiple interaction categories.

3. "**Baseline2-app**": We treat the combination of the interaction and its context as a single class and fine-tune the VGG-16 model for classification. This corresponds to using appearance feature to implement the method in Figure 1 (a).

4. "**Baseline2-spatial**": Similar to "Baseline2-app". We train a linear classifier to classify the spatial features into the classes derived from the combination of the interaction and its context.

5. "**AP+C**": We apply the context-aware classifier to the appearance representation described in Sec. 3.2.2.

6. "**AP+C+AT**": The basic attention-pooling representation described in Sec. 3.3 with the classifier in **AP+C**.

7. "**AP+C+CAT**": The context-aware attention-pooling representation described in Sec. 3.3 with the classifier in **AP+C**.

8. "**Spatial+C**": We apply the context-aware classifier to the spatial features described in Sec. 3.2.1.

Besides those methods, we also compare the performance of our methods against those reported in the related literature. However, it should noted that these methods may use different feature representation, detectors or pre-training strategies.

4.1. Evaluation on the visual relationship dataset

We first conduct experiments on the Visual Relationship Detection (VRD) dataset [18]. This dataset is designed for evaluating the visual relationship ($\langle$*subject, predicate, object*$\rangle$) detection, where the "predicate" in those datasets is equivalent to the "interaction" in our paper and we will use them interchangeably thereafter. It contains 4000 training and 1000 test images including 100 object classes and 70 predicates. In total, there are 37993 relationship instances with 6672 relationship types, out of which 1877 relationships occur only in the test set but not in the training set.

Following [18], we evaluate on three tasks: (1) For **predicate detection**, the input is an image and a set of ground-truth object bounding boxes. The task is to predict the possible interactions between pairs of objects. Since the interaction recognition is the main focus of this paper, the performance of this task provides the most relevant indication of the quality of the proposed method. (2) In **phrase detection**, we aim to predict $\langle$*subject-predicate-object*$\rangle$ and localize the entire relationship in one bounding boxes. (3) For **relationship detection**, the task is to recognize $\langle$*subject-predicate-object*$\rangle$ and localize both subject and object bounding boxes. Both boxes should have at least 0.5 overlap with the ground truth bounding boxes in order to be regarded as a correct prediction. For the second and third tasks, we use the object detection results (both bounding boxes and corresponding detection scores) provided in [18]. This allows us to fairly compare the performance of the proposed interaction recognition framework without the influence of detection.

We use the Recall@100 and Recall@50 as our evaluation metric following [18]. Recall@x computes the fraction of times the correct relationship is calculated in the top x predictions, which are ranked by the product of the objectness confidence scores and the classification probabilities of the interactions. As discussed in [18], we do not use the mean average precision (mAP), which is a pessimistic evaluation metric because it cannot exhaustively annotate all possible relationships in an image.

4.1.1 Detection results comparison

In this section, we evaluate the performance of three detection tasks on the Visual Relationship Detection (VRD) benchmark dataset and provide the comprehensive analysis. We compare all the eight methods and the results in [18,31]. The results are shown in Table 1. From it we can make the following observations:

The effect of context-aware modeling: To validate the main point in this paper, we compare the proposed method against two context-interaction modeling baselines, i.e. baseline1-app, baseline2-app, baseline1-spatial and baseline2-spatial). By analysing the results, we can see that the proposed context-aware modeling methods (methods with "AP") achieves much better performance than the four baselines. The improvement achieved by use context-aware modeling is consistently observed for both spatial features and appearance features. This justifies that the context information is crucial for interaction prediction.

Various feature representations: We also quantitatively investigate the performance of the proposed context-aware framework under various feature types. As can be seen in Table 1, the appearance feature representation performs consistently better than the spatial feature representation, especially for the baseline2 setting. This may be because the visual feature representation has richer discriminative power than the 14-dimensional spatial feature. Also, with our context-aware recognition framework, we can significantly boost the performance of both features and interestly in this case the gap between two types of features is largely diminished, e.g. AP+C+CAT vs. Spatial+C.

The effect of attention models: We also investigate the impacts of the attention scheme employed in our model by comparing AP+C, AP+C+AT and AP+C+CAT. The best results are obtained by utilizing the context-aware attention model. This justifies our postulate that it is better to make the network attend on the discriminative regions of feature maps.

Comparison with [31] and [18]: Finally, we compare our methods with the methods in [31] and [18]. As seen, our methods achieve better performance than these two competing methods. Since our methods use the same object detection in [18], our result is most comparable to it. Note that our model does not employ explicit language priors modeling as in [18] and our improvement purely comes from the visual cue. This again demonstrates the power of context-aware interaction recognition.

Method	**Predicate Det.**		Phrase Det.		Relationship Det.	
	R@100	R@50	R@100	R@50	R@100	R@50
Visual Phrase [31]	1.91	0.97	0.07	0.04	-	-
Language Priors [18]	47.87	47.87	17.03	16.17	14.70	13.86
Baseline1-app	18.13	18.13	6.02	5.42	5.54	5.01
Baseline1-spatial	17.77	17.77	5.24	4.77	4.54	4.19
Baseline2-app	27.23	27.23	9.30	7.91	8.34	7.03
Baseline2-spatial	13.85	13.85	4.15	3.06	3.63	2.63
Spatial+C	51.17	51.17	17.61	15.46	15.43	13.51
AP+C	52.36	52.36	18.69	16.91	16.46	14.88
AP+C+AT	53.12	53.12	19.08	17.30	16.89	15.40
AP+C+CAT	**53.59**	**53.59**	**19.24**	**17.60**	**17.39**	**15.63**

Table 1: Evaluation of different methods on the visual relationship benchmark dataset. The results reported include visual phase detection (Phrase Det.), visual relationship detection (Relationship Det.) and predicate detection (Predicate Det.) measured by Top-100 recall (R@100) and Top-50 recall (R@50).

To better evaluate our approach, we further visualize some test examples of AP+C+CAT in Figure 4. We can see that our predictions are reasonable in most cases.

4.1.2 Zero-shot learning performance evaluation

An important motivation of our method is to make the interaction classifier generalizable to unseen combinations of the interaction and context. In this section, we report the performance of our method on a zero-shot learning setting. Specifically, we train our models on the training set and evaluate their interaction classification performance on the 1877 unseen visual relationships in the test set. The results are reported in Table 3. From the table, we can see that the proposed methods work especially well in the zero-shot learning. For example, our best performed method (AP+C+CAT) almost doubled the performance on predicate detection in comparison with the Language Priors [18] method. This big improvement can be largely attributed to the advantage of using the context-aware scheme to model the interaction. In the Language Priors [18] method, the visual term for recognizing interaction is context-independent. Without context information to constrain the appearance variations, the learned interaction classifier tends to overfit the training set and fails to generalize to images with unseen interaction-context combinations. In comparison, with context-aware modeling, we explicitly consider the visual appearance variations introduced by changing context, thus more accurate and generalizable interaction classifier can be learned.

One interesting observation made in Table 3 is that the spatial feature representation produces better performance than the appearance based representation, as is evident from the superior performance of Spatial+C over AP methods.

Method	Phrase Det.		Relationship Det.		Zero-Shot Phrase Det.		Zero-Shot Relationship Det.	
	R@100	R@50	R@100	R@50	R@100	R@50	R@100	R@50
CLC (CCA+Size+Position) [25]	20.70	16.89	18.37	15.08	**15.23**	**10.86**	**13.43**	**9.67**
VTransE [40]	22.42	19.42	15.20	14.07	3.51	2.65	2.14	1.71
Vip-CNN [14]	**27.91**	22.78	20.01	17.32	-	-	-	-
VRL [16]	22.60	21.37	20.79	18.19	10.31	9.17	8.52	7.94
Faster-RCNN + (AP+C+CAT)	25.26	23.88	23.39	20.14	11.28	10.73	10.17	9.57
Faster-RCNN + (AP+C+CAT) + Language Priors	25.56	**24.04**	**23.52**	**20.35**	11.30	10.78	10.26	9.54

Table 2: Results for visual relationship detection on the visual relationship benchmark dataset. Notice that we simply replace the detector with Faster-RCNN to extract a set of candidate object proposals without end-to-end jointly training the detector [14, 16, 40] with the proposed method. And in CLC [25], they use features and detection results from a Faster RCNN trained on external MSCOCO [17] dataset and additional cues (e.g. size and position) are incorporated.

Method	**Predicate Det.**		Phrase Det.		Relationship Det.	
	R@100	R@50	R@100	R@50	R@100	R@50
Language Priors [18]	8.45	8.45	3.75	3.36	3.52	3.13
Baseline1-app	7.44	7.44	3.08	2.82	2.91	2.74
Baseline1-spatial	7.27	7.27	2.14	2.14	2.14	2.14
Baseline2-app	7.36	7.36	2.22	1.71	2.05	1.54
Baseline2-spatial	0.43	0.43	0.09	0.09	0.09	0.09
Spatial+C	**16.42**	**16.42**	6.24	5.82	5.65	5.30
AP+C	15.06	15.06	5.82	5.05	5.22	4.62
AP+C+AT	15.00	15.00	5.62	5.02	5.36	4.76
AP+C+CAT	16.37	16.37	**6.59**	**5.99**	**5.99**	**5.47**

Table 3: Results for zero-shot visual relationship detection on the visual relationship benchmark dataset.

We speculate this is because spatial relationship features are more object independent and are less prone to overfiting the training set.

To intuitively evaluate zero-shot performance, we add some test examples of AP+C+CAT in Figure 5. We can make reasonable predictions on unseen interaction-context combinations in most cases.

4.1.3 Extensions and comparison with state-of-the-art methods

Since the main focus of above experiments is to validate the advantage of the proposed methods over four competing baselines, we did not explore some techniques which could potentially further improve the visual relationship detection performance on the VRD dataset. To make our method achieve more comparable performance on the visual relationship and visual phrase detection tasks, we may consider two straightfoward extensions for our method: (1) use a better detector and (2) incorporate the language term trained in [18]. In the following part, we will examine the performance attained by applying these extensions and compare the resultant performance against the very latest state-of-the-art approaches [14, 16, 25, 40] on the VRD dataset.

Improved detector: We first examine the effect of using a better detector by replacing the detection results obtained in [18] with that obtained by a Faster-RCNN detector [8]. Note that the Faster-RCNN detector has also been used in [14, 16, 25, 40] and using it will make our method comparable with the current state-of-the-arts. In our implementation, only the top 50 candidate object proposals, ranked by objectness confidence scores are extracted for mining relationships in per test image. The result of this modification is reported in Table 2 with our method annotated as Faster-RCNN + (AP+C+CAT). As seen, our method achieves best performance on phrase detection R@50, relationship detection, zero-shot phrase and relationship detection. Note that our method can be further incorporated into the end-to-end relationship detection framework such as [14] to achieve even better performance.

Language priors: Language priors make significant contribution to [18] and in this section we apply the language priors released by [18] to investigate its impact. Following [18], we multiply our best performed model Faster-RCNN + (AP+C+CAT) with the language priors for interactions to obtain the final detection scores and the result is shown in Table 2 with the annotation Faster-RCNN + (AP+C+CAT) + Language Priors. Interestingly, the introduction of the language priors only introduces a marginal performance improvement. We suspect that is due to that our method builds a classifier with the information of both the interaction and context, and the correlation of interaction and context has been implicitly encoded. Therefore adding the language priors does not bring further benefit.

4.2. Evaluation on the visual phrase dataset

Following [18], we also run additional experiments on the Visual Phrase [31] dataset. It has 17 phrases, out of which 12 of these phrases can be represented as triplet relationships as in the VRD dataset. We use the setting of [18] to conduct the experiment and report the R@50 and R@100 results in Table 4. Since the Visual Phrase dataset does not provide detection results, we apply the RCNN [9] model to produce a set of candidate object regions and corresponding detection scores. As seen from Table 4, AP+C+CAT again achieves the best performance. In comparison with the performance of [18], our method improves most in the zero-shot learning setting. This is consistent with the observation made in Sec. 4.1.2.

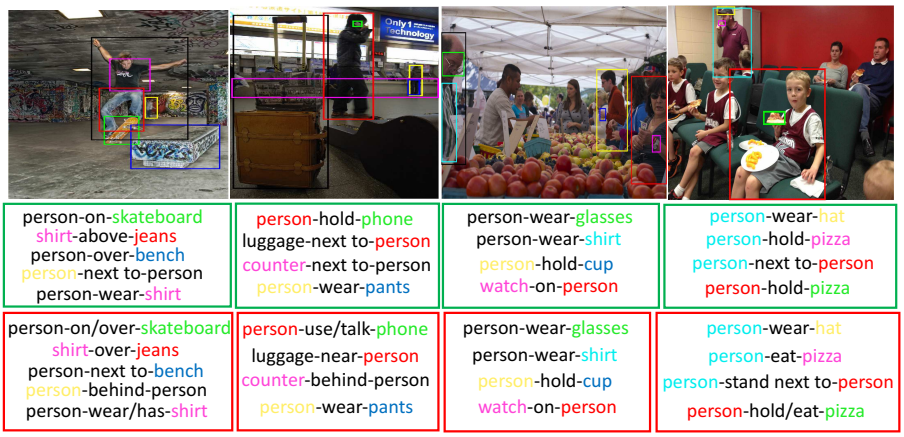

Figure 4: Qualitative examples of interaction recognition. We only predict the interaction between the ground-truth context bounding boxes. The phrases in the green bounding boxes are predicted while the phrases shown in the red bounding boxes are ground-truth.

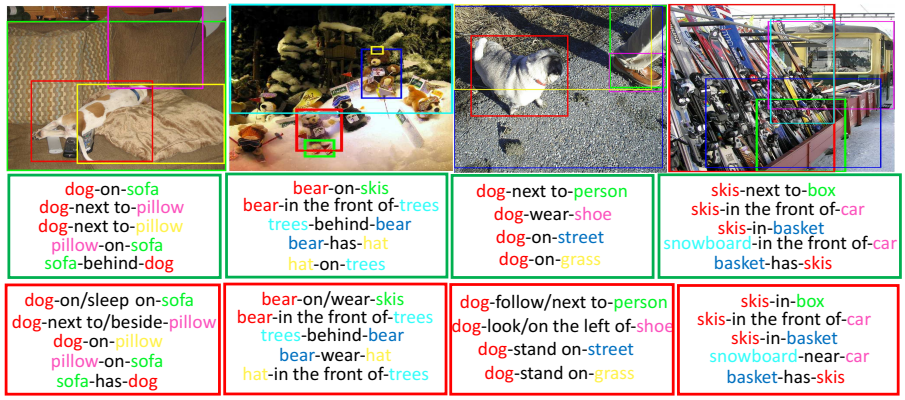

Figure 5: Qualitative examples of zero-shot interaction recognition. We only predict the interaction between the ground-truth context bounding boxes. The phrases in the green bounding boxes are predicted while the phrases shown in the red bounding boxes are ground-truth.

Method	Phrase Detection		Zero-Shot Phrase Detection	
	R@100	R@50	R@100	R@50
Visual Phrase [31]	52.7	49.3	-	-
Language Priors [18]	82.7	78.1	23.9	11.4
Baseline1-app	70.1	65.6	12.4	10.5
Baseline1-spatial	68.3	63.6	10.3	8.9
Baseline2-app	77.5	72.3	11.0	9.2
Baseline2-spatial	15.7	10.4	1.1	0.5
Spatial+C	84.9	80.8	27.6	15.7
AP+C	85.9	81.6	28.5	16.4
AP+C+AT	86.2	82.1	28.8	17.9
AP+C+CAT	**86.8**	**82.9**	**30.2**	**18.7**

Table 4: Comparison of performance on the Visual Phrase dataset.

5. Conclusion

In this paper, we study the role of context in recognizing the object interaction pattern. After identifying the importance of using context information, we propose a context-aware interaction classification framework which is accurate, scalable and enjoys good generalization ability to recognize unseen context-interaction combinations. Further, we investigate various ways to derive the visual representation for interaction patterns and extend the context-aware framework to design a new attention-pooling layer. With extensive experiments, we validate the advantage of the proposed methods and produce the state-of-the-art performance on two visual relationship detection datasets.

Acknowledgements LL was in part supported by ARC DECRA Fellowship DE170101259. CS was in part supported by ARC Future Fellowship FT120100969. IDR was in part supported by ARC Laureate Fellowship FL130100102.

References

[1] A. Arnold, R. Nallapati, and W. W. Cohen. A comparative study of methods for transductive transfer learning. In *International Conference on Data Mining Workshops*, pages 77–82. IEEE, 2007. 3

[2] H. Bilen, B. Fernando, E. Gavves, A. Vedaldi, and S. Gould. Dynamic image networks for action recognition. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 3034–3042, 2016. 1, 2

[3] Y.-W. Chao, Z. Wang, Y. He, J. Wang, and J. Deng. Hico: A benchmark for recognizing human-object interactions in images. In *Proc. IEEE Int. Conf. Comp. Vis.*, pages 1017–1025, 2015. 2

[4] C. Desai, D. Ramanan, and C. C. Fowlkes. Discriminative models for multi-class object layout. *Int. J. Comp. Vis.*, 95(1):1–12, 2011. 1

[5] C. Do and A. Y. Ng. Transfer learning for text classification. In *Proc. Adv. Neural Inf. Process. Syst.*, pages 299–306, 2005. 3

[6] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 2625–2634, 2015. 3

[7] T. Evgeniou and M. Pontil. Regularized multi–task learning. In *International Conference on Data Mining*, pages 109–117. ACM, 2004. 3

[8] R. Girshick. Fast r-cnn. In *Proc. IEEE Int. Conf. Comp. Vis.*, pages 1440–1448, 2015. 7

[9] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 580–587, 2014. 7

[10] A. Gupta and L. S. Davis. Beyond nouns: Exploiting prepositions and comparative adjectives for learning visual classifiers. In *Proc. Eur. Conf. Comp. Vis.*, pages 16–29. Springer, 2008. 1

[11] R. Hu, M. Rohrbach, J. Andreas, T. Darrell, and K. Saenko. Modeling relationships in referential expressions with compositional modular networks. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, July 2017. 1, 2

[12] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 3128–3137, 2015. 1, 3

[13] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[14] Y. Li, W. Ouyang, X. Wang, and X. Tang. Vip-cnn: Visual phrase guided convolutional neural network. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2017. 1, 2, 3, 7

[15] Y. Li, W. Ouyang, B. Zhou, K. Wang, and X. Wang. Scene graph generation from objects, phrases and region captions. In *Proc. IEEE Int. Conf. Comp. Vis.*, 2017. 2

[16] X. Liang, L. Lee, and E. P. Xing. Deep variation-structured reinforcement learning for visual relationship and attribute detection. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2017. 2, 3, 7

[17] 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 *Proc. Eur. Conf. Comp. Vis.*, pages 740–755. Springer, 2014. 7

[18] C. Lu, R. Krishna, M. Bernstein, and L. Fei-Fei. Visual relationship detection with language priors. In *Proc. Eur. Conf. Comp. Vis.*, pages 852–869. Springer, 2016. 1, 2, 5, 6, 7, 8

[19] J. Lu, C. Xiong, D. Parikh, and R. Socher. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, July 2017. 1

[20] L. Ma, Z. Lu, L. Shang, and H. Li. Multimodal convolutional neural networks for matching image and sentence. In *Proc. IEEE Int. Conf. Comp. Vis.*, pages 2623–2631, 2015. 1

[21] M. Malinowski, M. Rohrbach, and M. Fritz. Ask your neurons: A neural-based approach to answering questions about images. In *Proc. IEEE Int. Conf. Comp. Vis.*, pages 1–9, 2015. 3

[22] A. Mallya and S. Lazebnik. Learning models for actions and person-object interactions with transfer to question answering. In *Proc. Eur. Conf. Comp. Vis.*, pages 414–428. Springer, 2016. 1, 2

[23] S. Parameswaran and K. Q. Weinberger. Large margin multitask metric learning. In *Proc. Adv. Neural Inf. Process. Syst.*, pages 1867–1875, 2010. 3

[24] N. Patricia and B. Caputo. Learning to learn, from transfer learning to domain adaptation: A unifying perspective. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 1442–1449, 2014. 3

[25] B. A. Plummer, A. Mallya, C. M. Cervantes, J. Hockenmaier, and S. Lazebnik. Phrase localization and visual relationship detection with comprehensive linguistic cues. *arXiv preprint arXiv:1611.06641*, 2016. 2, 3, 4, 7

[26] V. Ramanathan, C. Li, J. Deng, W. Han, Z. Li, K. Gu, Y. Song, S. Bengio, C. Rosenberg, and L. Fei-Fei. Learning semantic relationships for better action retrieval in images. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 1100–1109, 2015. 1, 2

[27] M. Ren, R. Kiros, and R. Zemel. Image question answering: A visual semantic embedding model and a new dataset. *Proc. Adv. Neural Inf. Process. Syst.*, 1(2):5, 2015. 3

[28] A. Rohrbach, M. Rohrbach, R. Hu, T. Darrell, and B. Schiele. Grounding of textual phrases in images by reconstruction. In *Proc. Eur. Conf. Comp. Vis.*, pages 817–834. Springer, 2016. 1, 2

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

[30] F. Sadeghi, S. K. Kumar Divvala, and A. Farhadi. Viske: Visual knowledge extraction and question answering by visual verification of relation phrases. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 1456–1464, 2015. 1

[31] M. A. Sadeghi and A. Farhadi. Recognition using visual phrases. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 1745–1752, 2011. 6, 7, 8

[32] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 4

[33] K. Soomro, A. R. Zamir, and M. Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 2

[34] O. Vinyals, C. Blundell, T. Lillicrap, D. Wierstra, et al. Matching networks for one shot learning. In *Proc. Adv. Neural Inf. Process. Syst.*, pages 3630–3638, 2016. 3

[35] L. Wang, Y. Qiao, and X. Tang. Action recognition with trajectory-pooled deep-convolutional descriptors. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 4305–4314, 2015. 1, 2

[36] Q. Wu, C. Shen, L. Liu, A. Dick, and A. van den Hengel. What value do explicit high level concepts have in vision to language problems? In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 203–212, 2016. 1

[37] Q. Wu, P. Wang, C. Shen, A. Dick, and A. van den Hengel. Ask me anything: Free-form visual question answering based on knowledge from external sources. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 4622–4630, 2016. 1

[38] B. Yao, X. Jiang, A. Khosla, A. L. Lin, L. Guibas, and L. Fei-Fei. Human action recognition by learning bases of action attributes and parts. In *Proc. IEEE Int. Conf. Comp. Vis.*, pages 1331–1338, 2011. 2

[39] B. Zhang, L. Wang, Z. Wang, Y. Qiao, and H. Wang. Real-time action recognition with enhanced motion vector cnns. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, June 2016. 1, 2

[40] H. Zhang, Z. Kyaw, S.-F. Chang, and T.-S. Chua. Visual translation embedding network for visual relation detection. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2017. 2, 3, 4, 7