Weakly-supervised learning of visual relations

Julia Peyre[1,2] Ivan Laptev[1,2] Cordelia Schmid[2,4] Josef Sivic[1,2,3]

Abstract

This paper introduces a novel approach for modeling visual relations between pairs of objects. We call relation a triplet of the form $(subject, predicate, object)$ where the predicate is typically a preposition (eg. 'under', 'in front of') or a verb ('hold', 'ride') that links a pair of objects $(subject, object)$. Learning such relations is challenging as the objects have different spatial configurations and appearances depending on the relation in which they occur. Another major challenge comes from the difficulty to get annotations, especially at box-level, for all possible triplets, which makes both learning and evaluation difficult. The contributions of this paper are threefold. First, we design strong yet flexible visual features that encode the appearance and spatial configuration for pairs of objects. Second, we propose a weakly-supervised discriminative clustering model to learn relations from image-level labels only. Third we introduce a new challenging dataset of unusual relations (UnRel) together with an exhaustive annotation, that enables accurate evaluation of visual relation retrieval. We show experimentally that our model results in state-of-the-art results on the visual relationship dataset [32] significantly improving performance on previously unseen relations (zero-shot learning), and confirm this observation on our newly introduced UnRel dataset.

1. Introduction

While a great progress has been made on the detection and localization of individual objects [41, 53], it is now time to move one step forward towards understanding complete scenes. For example, if we want to localize "a person sitting on a chair under an umbrella", we not only need to detect the objects involved : "person", "chair", "umbrella", but also need to find the correspondence of the semantic relations "sitting on" and "under" with the correct pairs of objects in the image. Thus, an important challenge is automatic understanding of how entities in an image interact with each other.

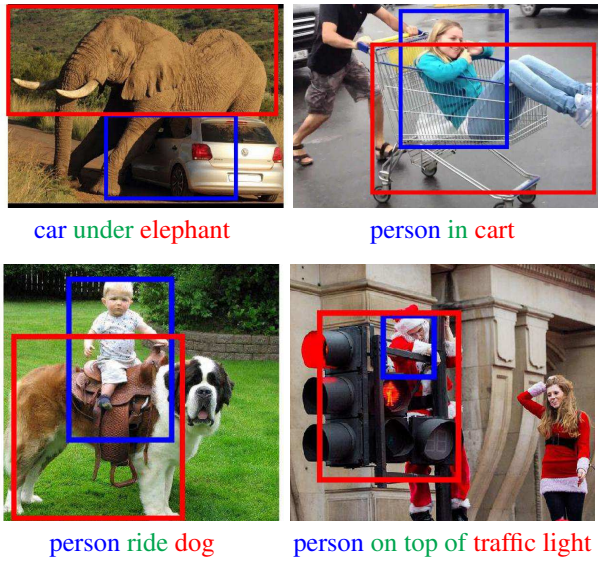

Figure 1: Examples of top retrieved pairs of boxes in UnRel dataset for unusual queries (indicated below each image) with our weakly-supervised model described in 3.2.

This task presents two main challenges. First, the appearance of objects can change significantly due to interactions with other objects (person cycling, person driving). This visual complexity can be tackled by learning "visual phrases" [44] capturing the pair of objects in a relation as one entity, as opposed to first detecting individual entities in an image and then modeling their relations. This approach, however, does not scale to the large number of relations as the number of such visual phrases grows combinatorially, requiring large amounts of training data. To address this challenge, we need a method that can share knowledge among similar relations. Intuitively, it seems possible to generalize frequent relations to unseen triplets like those depicted in Figure 1 : for example having seen "person ride horse" at training could help recognizing "person ride dog" at test time.

The second main challenge comes from the difficulty to provide exhaustive annotations on the object level for relations that are by their nature non mutually-exclusive (i.e. "on the left of" is also "next to"). A complete labeling of R relations for all pairs of N objects in an image would indeed

[1]Département d'informatique de l'ENS, Ecole normale supérieure, CNRS, PSL Research University, 75005 Paris, France.
[2]INRIA
[3]Czech Institute of Informatics, Robotics and Cybernetics at the Czech Technical University in Prague.
[4]Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, 38000 Grenoble, France.

require $\mathcal{O}(N^2R)$ annotations *for each image*. Such difficulty makes both learning and evaluation very challenging. For learning, it would be desirable to learn relations from image-level annotations only. For evaluation, current large-scale datasets [28, 32] do not allow retrieval evaluation due to large amount of missing annotations.

Contributions. The contributions of this work are threefold. First, to address the combinatorial challenge, we develop a method that can handle a large number of relations by sharing parameters among them. For example, we learn a single "on" classifier that can recognize both "person on bike" and "dog on bike", even when "dog on bike" has not been seen in training. The main innovation is a new model of an object relation that represents a pair of boxes by explicitly incorporating their spatial configuration as well as the appearance of individual objects. Our model relies on a multimodal representation of object configurations for each relation to handle the variability of relations. Second, to address the challenge of missing training data, we develop a model that, given pre-trained object detectors, is able to learn classifiers for object relations from image-level supervision only. It is, thus, sufficient to provide an image-level annotation, such as "person on bike", without annotating the objects involved in the relation. Finally, to address the issue of missing annotations in test data, we introduce a new dataset of unusual relations (UnRel), with exhaustive annotation for a set of unusual triplet queries, that enables to evaluate retrieval on rare triplets and validate the generalization capabilities the learned model.

2. Related Work

Alignment of images with language. Learning correspondences between fragments of sentences and image regions has been addressed by the visual-semantic alignment which has been used for applications in image retrieval and caption generation [6, 25, 26]. With the appearance of new datasets providing box-level natural language annotations [27, 28, 33, 38], recent works have also investigated caption generation at the level of image regions for the tasks of natural language object retrieval [20, 33, 42] or dense captioning [22]. Our approach is similar in the sense that we aim at aligning a language triplet with a pair of boxes in the image. Typically, existing approaches do not explicitly represent relations between noun phrases in a sentence to improve visual-semantic alignment. We believe that understanding these relations is the next step towards image understanding with potential applications in tasks such as Visual Question Answering [2].

Learning triplets. Triplet learning has been addressed in various tasks such as mining typical relations (knowledge extraction) [7, 43, 52, 54], reasoning [21, 35, 45], object detection [17, 44], image retrieval [23] or fact retrieval [11]. In this work, we address the task of relationship detection in images. This task was studied for the special case of human-object interactions [9, 10, 18, 39, 40, 49, 50, 51], where the triplet is in the form $(person, action, object)$. Contrary to these approaches, we do not restrict the $subject$ to be a person and we cover a broader class of predicates that includes prepositions and comparatives. Moreover, most of the previous work in human-object interaction was tested on small datasets only and does not explicitly address the combinatorial challenge in modeling relations [44]. Recently, [32] tried to generalize this setup to non-human subjects and scale to a larger vocabulary of objects and relations by developing a language model sharing knowledge among relations for visual relation detection. In our work we address this combinatorial challenge by developing a new visual representation that generalizes better to unseen triplets without the need for a strong language model. This visual representation shares the spirit of [14, 23, 30] and we show it can handle multimodal relations and generalizes well to unseen triplets. Our model also handles a weakly-supervised set-up when only image-level annotations for object relations are available. It can thus learn from complex scenes with many objects participating in different relations, whereas previous work either uses full supervision or typically assumes only one object relation per image, for example, in images returned by a web search engine. Finally, we also address the problem to evaluate accurately due to missing annotations also pointed out in [11, 32]. We introduce a new dataset of unusual relations exhaustively labeled for a set of triplet queries, the UnRel dataset. This dataset enables the evaluation of relation retrieval and localization. Our dataset is related to the "Out of context" dataset of [8] which also presents objects in unusual configurations. However, the dataset of [8] is not annotated with relations and does not match the vocabulary of objects in [32], which prevents direct comparisons with existing methods that use data from [32] for training.

Weak supervision. Most of the work on weakly-supervised learning for visual recognition has focused on learning objects [4, 12, 36]. Here, we want to tackle the task of weakly-supervised detection of relations. This task is more complex as we need to detect the individual objects that satisfy the specific relation. We assume that pre-trained detectors for individual objects are available and learn relations among objects with image-level labels. Our work uses a discriminative clustering objective [3], that has been successful in several computer vision tasks [5, 24], but has not been so far, to the best of our knowledge, used for modeling relations.

Zero-shot learning. Zero-shot learning has been mostly explored for object classification [13, 29, 46, 48] and recently for the task of describing images with novel objects [19, 47]. In our work, we address zero-shot learning of relations in the form of triplets $(subject, predicate, object)$, where each term has already

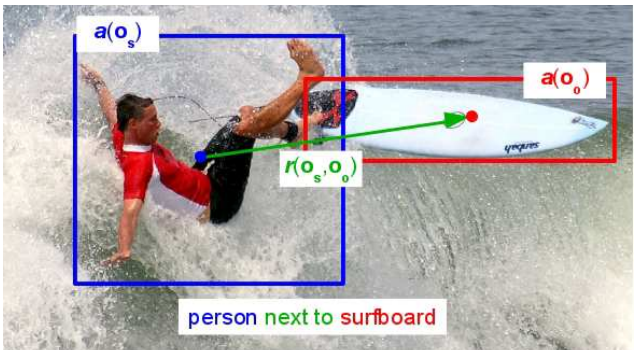

Figure 2: Our visual representation is the composition of appearance features for each object $[\boldsymbol{a}(\boldsymbol{o_s}), \boldsymbol{a}(\boldsymbol{o_o})]$ and their spatial configuration $\boldsymbol{r}(\boldsymbol{o_s}, \boldsymbol{o_o})$ represented by the green arrow.

been seen independently during training, but not in that specific combination. We develop a model to detect and localize such zero-shot relations.

3. Representing and learning visual relations

We want to represent triplets $t = (s, r, o)$ where s is the subject, o the object and r is the predicate. s and o are nouns and can be objects like "person", "horse", "car" or regions such as "sky", "street", "mountain". The predicate r is a term that links the subject and the object in a sentence and can be a preposition ("in front of", "under"), a verb ("ride", "hold") or a comparative adjective ("taller than"). To detect and localize such triplets in test images, we assume that the candidate object detections for s and o are given by a detector trained with full supervision. Here we use the object detector [15] trained on the Visual Relationship Detection training set [32]. In 3.1, we will explain our representation of a triplet $t = (s, r, o)$ and show in 3.2 how we can learn to detect triplets in images given weak image-level supervision for relations.

3.1. Visual representation of relations

A triplet $t = (s, r, o)$ such as "person next to surfboard" in Figure 2 visually corresponds to a pair of objects (s, o) in a certain configuration. We represent such pairs by the spatial configuration between object bounding boxes $(\boldsymbol{o_s}, \boldsymbol{o_o})$ and the individual appearance of each object.

Representing spatial configurations of objects. Given two boxes $\boldsymbol{o_s} = [x_s, y_s, w_s, h_s]$, $\boldsymbol{o_o} = [x_o, y_o, w_o, h_o]$, where (x, y) are the coordinates of the center of the box, and (w, h) are the width and height of the box, we encode the spatial configuration with a 6-dimensional vector:

$$\boldsymbol{r}(\boldsymbol{o_s}, \boldsymbol{o_o}) = [\underbrace{\frac{x_o - x_s}{\sqrt{w_s h_s}}}_{r_1}, \underbrace{\frac{y_o - y_s}{\sqrt{w_s h_s}}}_{r_2}, \underbrace{\sqrt{\frac{w_o h_o}{w_s h_s}}}_{r_3}, \underbrace{\frac{\boldsymbol{o_s} \cap \boldsymbol{o_o}}{\boldsymbol{o_s} \cup \boldsymbol{o_o}}}_{r_4}, \underbrace{\frac{w_s}{h_s}}_{r_5}, \underbrace{\frac{w_o}{h_o}}_{r_6}] \quad (1)$$

where r_1 and r_2 represent the renormalized translation between the two boxes, r_3 is the ratio of box sizes, r_4 is the overlap between boxes, and r_5, r_6 encode the aspect ratio of each box respectively. Directly adopting this feature as our representation might not be well suited for some spatial relations like "next to" which are multimodal. Indeed, "s next to o" can either correspond to the spatial configuration "s left of o" or "s right of o". Instead, we propose to discretize the feature vector (1) into k bins. For this, we assume that the spatial configurations $\boldsymbol{r}(\boldsymbol{o_s}, \boldsymbol{o_o})$ are generated by a mixture of k Gaussians and we fit the parameters of the Gaussian Mixture Model to the training pairs of boxes. We take the scores representing the probability of assignment to each of the k clusters as our spatial representation. In our experiments, we use $k = 400$, thus the spatial representation is a 400-dimensional vector. In Figure 3, we show examples of pairs of boxes for the most populated components of the trained GMM. We can observe that our spatial representation can capture subtle differences between configurations of boxes, see row 1 and row 2 of Figure 3, where "person on board" and "person carry board" are in different clusters.

Representing appearance of objects. Our appearance features are given by the fc7 output of a Fast-RCNN [15] trained to detect individual objects. In our experiments, we use Fast-RCNN with VGG16 pre-trained on ImageNet. As the extracted features have high dimensionality, we perform PCA on the L2-normalized features to reduce the number of dimensions from 4096 to 300. We concatenate the appearance features of the subject and object and apply L2-normalization again.

Our final visual feature is a concatenation of the spatial configuration $\boldsymbol{r}(\boldsymbol{o_s}, \boldsymbol{o_o})$ and the appearance of objects $[\boldsymbol{a}(\boldsymbol{o_s}), \boldsymbol{a}(\boldsymbol{o_o})]$. In our experiments, it has a dimensionality of $d = 1000$. In the fully supervised setup, where each relation annotation is associated with a pair of object boxes in the image, we use ridge regression to train a multi-way relation classifier to assign a relation to a given visual feature. Training is performed jointly on all relation examples of the training set.

In the next section, we describe how we learn relation classifiers with only weak, image-level, annotations.

3.2. Weakly-supervised learning of relations

Equipped with pre-trained detectors for individual objects, our goal here is to learn to detect and localize relations between objects, given image-level supervision only. For example, for a relation "person falling off horse" we are given (multiple) object detections for "person" and "horse", but do not know which objects participate in the relation, as illustrated in Figure 4. Our model is based on a weakly-supervised discriminative clustering objective [3], where we introduce latent variables to model which pairs of objects

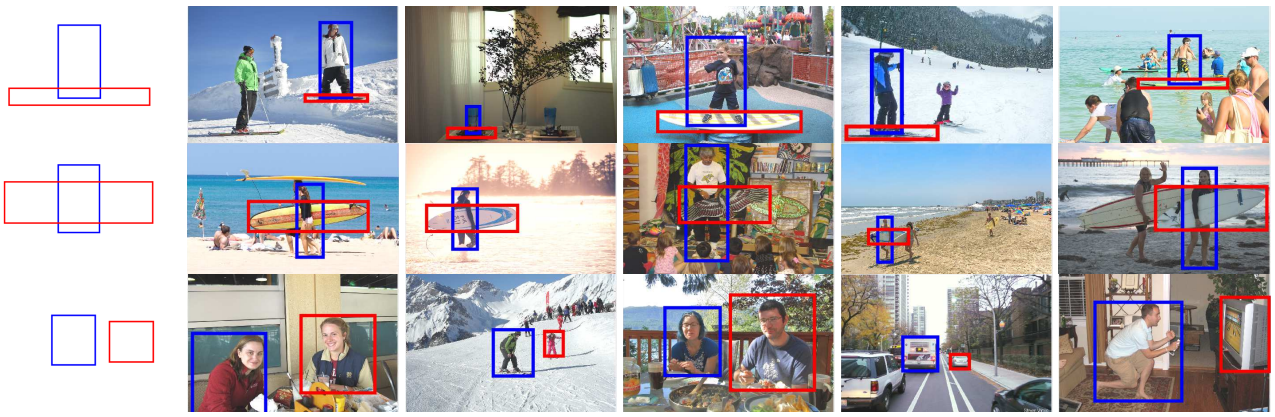

Figure 3: Examples for different GMM components of our spatial configuration model (one per row). In the first column we show the spatial configuration corresponding to the mean of the pairs of boxes per component. Note that our representation can capture subtle differences between spatial configurations, see e.g., row 1 and 2.

participate in the relation. We train a classifier for each predicate r and incorporate weak annotations in the form of constraints on latent variables. Note that the relation classifiers are shared across object categories (eg. the relations "person on horse" and "cat on table" share the same classifier "on") and can thus be used to predict unseen triplets.

Discriminative clustering of relations. Our goal is to learn a set of classifiers $W = [\mathbf{w_1}, ..., \mathbf{w_R}] \in \mathbb{R}^{d \times R}$ where each classifier $\mathbf{w_r}$ predicts the likelihood of a pair of objects (s, o) to belong to the r^{th} predicate in a vocabulary of R predicates. If strong supervision was provided for each pair of objects, we could learn W by solving a ridge regression :

$$\min_{W \in \mathbb{R}^{d \times R}} \frac{1}{N} \|Z - XW\|_F^2 + \lambda \|W\|_F^2 \tag{2}$$

where $Z \in \{0, 1\}^{N \times R}$ are the ground truth labels for each of the N pairs of objects across all training images, and $X = [\mathbf{x_1}, ..., \mathbf{x_N}]^T$ is a $N \times d$ matrix where each row $\mathbf{x_k}$ is the visual feature corresponding to the k^{th} pair of objects. However, in a weakly-supervised setup the entire matrix Z is unknown. Building on DIFFRAC [3], our approach is to optimize the cost :

$$\min_{Z \in \mathcal{Z}} \min_{W \in \mathbb{R}^{d \times R}} \frac{1}{N} \|Z - XW\|_F^2 + \lambda \|W\|_F^2 \tag{3}$$

which treats Z as a latent assignment matrix to be learnt together with the classifiers $W \in \mathbb{R}^{d \times R}$. Minimizing the first term encourages the predictions made by W to match the latent assignments Z, while the second term is a L2-regularization on the classifiers W. This framework enables to incorporate weak annotations by constraining the space of valid assignment matrices $Z \in \mathcal{Z}$. The valid matrices $Z \in \{0, 1\}^{N \times R}$ satisfy the multiclass constraint $Z\mathbf{1}_R = \mathbf{1}_N$ which assigns a pair of objects to one and only one predicate r. We describe in the next paragraph how to incorporate the weak annotations as constraints.

Weak annotations as constraints. For an image, we are given weak annotations in the form of triplets $t = (s, r, o) \in \mathcal{T}$. Having such triplet (s, r, o) indicates that at least one of the pairs of objects with object categories (s, o) is in relation r. Let us call $\mathcal{N}_t$ the subset of pairs of objects in the image that correspond to object categories (s, o). The rows of Z that are in subset $\mathcal{N}_t$ should satisfy the constraint :

$$\sum_{n \in \mathcal{N}_t} Z_{nr} \geq 1 \tag{4}$$

This constraint ensures that at least one of the pair of objects in the subset $\mathcal{N}_t$ is assigned to predicate r. For instance, in case of the image in Figure 4 that contains 12 candidate pairs of objects that potentially match the triplet $t = (person, falling\ off, horse)$, the constraint (4) imposes that at least one of them is in relation $falling\ off$.

Triplet score. At test time, we can compute a score for a pair of boxes $(\boldsymbol{o}_s, \boldsymbol{o}_o)$ relative to a triplet $t = (s, r, o)$ as

$$\begin{aligned} S((\boldsymbol{o}_s, \boldsymbol{o}_o) \mid t) = v_{rel}((\boldsymbol{o}_s, \boldsymbol{o}_o) \mid r) + \alpha_{sub} v_{sub}(\boldsymbol{o}_s \mid s) \\ + \alpha_{obj} v_{obj}(\boldsymbol{o}_o \mid o) + \alpha_{lang} \ell((s, o) \mid r), \end{aligned} \tag{5}$$

where $v_{rel}((\boldsymbol{o}_s, \boldsymbol{o}_o)|r) = \mathbf{x}_{(\boldsymbol{o}_s, \boldsymbol{o}_o)} \mathbf{w_r}$ is the score returned by the classifier $\mathbf{w}_r$ for predicate r (learnt by optimizing (3)) for the visual representation $\mathbf{x}_{(\boldsymbol{o}_s, \boldsymbol{o}_o)}$ of the pair of boxes. $v_{sub}(\boldsymbol{o}_s|s)$ and $v_{obj}(\boldsymbol{o}_o|o)$ are the object class likelihoods returned by the object detector. $\ell((s, o)|r)$ is a score of a language model that we can optionally combine with our visual model.

Optimization. We optimize the cost in (3) on pairs of objects in the training set using a variant of the Frank-Wolfe algorithm [34, 37]. The hyperparameters $(\alpha_{sub}, \alpha_{obj}, \alpha_{lang})$ are optimized on an held-out fully-annotated validation set which has no overlap with our training and test sets. In our experiments we use the validation split of [22] of the Visual

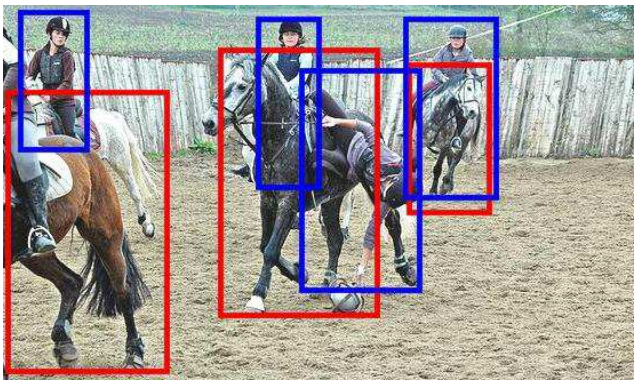

Figure 4: Image from the COCO dataset [31] associated with caption : *"A person falling off the side of a horse as it rides"*. The boxes correspond to the possible candidates for object category $person$ (blue) and $horse$ (red). Our model has to disambiguate the right pair for the relation "falling off" among 12 candidate pairs.

Genome dataset [28]. Unless otherwise specified, the candidate pairs, both at training and test time, are the outputs of the object detector [15] that we fine-tuned on the Visual Relationship Detection dataset [32]. For each image, we keep the object candidates whose confidence scores is above 0.3 among the top 100 detections. Non-maximum suppression with threshold 0.3 is applied to handle multiple detections. This results in an average of 18 object detections per image, i.e. around 300 pairs of boxes.

4. Experiments

In this section, we evaluate the performance of our model on two datasets for different evaluation setups. First, we evaluate our new visual representation for relations on the Visual Relationship Detection dataset [32]. We show results with our weakly-supervised model learned from image-level supervision and present large improvements over the state of the art for detecting unseen triplets (zero-shot detection). Second, we evaluate our model for the task of unusual triplets retrieval and localization on our new UnRel dataset.

4.1. Recall on Visual Relationship Detection dataset

Dataset. We evaluate our method on the Visual Relationship Detection dataset [32] following the original experimental setup. This dataset contains 4000 training and 1000 test images with ground truth annotations for relations between pairs of objects. Due to the specific train/test split provided by [32], 10% of test triplets are not seen at training and allow for evaluation of zero-shot learning. Some of these triplets are rare in the linguistic and visual world (e.g. "laptop on stove"), but most of them are only infrequent in the training set or have not been annotated (e.g. "van on the left of car"). Around 30K triplets are annotated in the training set, with an average of 7.5 relations per image. The dataset contains 100 objects and 70 predicates, i.e. $100 \times 100 \times 70$ possible triplets. However there are only 6672 different annotated triplets.

Evaluation set-up. Following [32], we compute `recall@x` which corresponds to the proportion of ground truth pairs among the `x` top scored candidate pairs in each image. We use the scores returned by (5) to sort the candidate pairs of boxes. The evaluation is reported for three setups. In **predicate detection**, candidate pairs of boxes are ground truth boxes, and the evaluation only focuses on the quality of the predicate classifier. In the other two more realistic setups, the subject and object confidence scores are provided by an object detector and we also check whether the candidate boxes intersect the ground truth boxes : either both subject and object boxes should match (**relationship detection**), or the union of them should match (**phrase detection**). For a fair comparison with [32], we report results using the same set of R-CNN [16] object detections as them. The localization is evaluated with IoU = 0.5.

Benefits of our visual representation. We first evaluate the quality of our visual representation in a fully supervised setup where the ground truth spatial localization for each relation is known, i.e. we know which objects in the image are involved in each relation at training time. For this, we solve the multi-label ridge regression in (2). Training with one-vs-rest SVMs gives similar results. We compare three types of features described in Section 3.1 in Table 1: [S] the spatial representation (f.), [A] the appearance representation (g.) and [S+A] the concatenation of the two (h.). We compare with the Visual Phrases model [44] and several variants of [32] [1] : Visual model alone (b.), Language (likelihood of a relationship) (c.), combined Visual+Language model (d.). In row (e.) we also report the performance of the full language model of [32], that scores the candidate pairs of boxes based on their predicted object categories, that we computed using the model and word embeddings provided by the authors. Because their language model is orthogonal to our visual model, we can combine them together (i.). The results are presented on the complete test set (column All) and on the zero-shot learning split (column Unseen). Table 1 shows that our combined visual features [S+A] improve over the visual features of [32] by 40% on the task of predicate detection and more than 10% on the hardest task of relationship detection. Furthermore, our purely visual features without any use of language (h.) reach comparable performance to the combined Visual+Language features of [32] and reach state-of-the-art performance (i.) when combined with the language scores of [32]. The good performance of our spatial features [S] alone (f.) confirms the observation we made in Figure 3 that our spatial clusters group pairs of objects in similar relations. That could partly explain why the visual model of [32] has low performance.

[1] When running the evaluation code of [32], we found slighlty better performance than what is reported in their paper. See appendix [1] for more details.

		Predicate Det.		Phrase Det.		Relationship Det.	
		All	Unseen	All	Unseen	All	Unseen
	Full sup.						
a.	Visual Phrases [44]	0.9	-	0.04	-	-	-
b.	Visual [32]	7.1	3.5	2.2	1.0	1.6	0.7
c.	Language (likelihood) [32]	18.2	5.1	0.08	0.00	0.08	0.00
d.	Visual + Language [32]	47.9	8.5	16.2	3.4	13.9	3.1
e.	Language (full) [32]	48.4	12.9	15.8	4.6	13.9	4.3
f.	Ours [S]	42.2	22.2	13.8	7.4	12.4	7.0
g.	Ours [A]	46.3	16.1	14.9	5.6	12.9	5.0
h.	Ours [S+A]	50.4	**23.6**	16.7	**7.4**	14.9	**7.1**
i.	Ours [S+A] + Language [32]	**52.6**	21.6	**17.9**	6.8	**15.8**	6.4
	Weak sup.						
j.	Ours [S+A]	46.8	19.0	16.0	6.9	14.1	6.7
k.	Ours [S+A] - Noisy	46.4	17.6	15.1	6.0	13.4	5.6

Table 1: Results on Visual Relationship Detection dataset [32] for R@50. See appendix [1] for results with R@100.

Their model learns a classifier only based on the appearance of the union of the two object boxes and lacks information about their spatial configuration.

Weak supervision. We evaluate our weakly-supervised classifiers W learned on image-level labels as described in Section 3.2. We use the ground truth annotations of the Visual Relationship Detection dataset as image-level labels. We report the results using our combined spatial and appearance features (j.) in Table 1. We see that when switching to weak supervision the recall@50 only drops from 50.4% to 46.8% for predicate detection and has limited influence on the other tasks. This is an interesting result as it suggests that, given pre-trained object detectors, weak image-level annotations are enough to learn good classifiers for relations. To investigate this further we have also tried to learn relation classifiers directly from noisy image-level labels without inferring at training time which objects participate in which relation. For each triplet $t = (s, r, o)$ in an image containing candidate pairs of boxes $(\boldsymbol{o_s}, \boldsymbol{o_o})$ we randomly select one of the pairs as being in relation r and discard the other object pairs. This is equivalent to training in a fully-supervised setup but with noisy labels. The performance obtained by this classifier (k.) is below our weakly-supervised learning set-up but is surprisingly high. We believe that this is related to a particular bias present in the Visual Relationship Detection dataset [32], which contains many images with only two prominent objects involved in a specific relation (more than half of the triplets fall into this category). To underline the ability of the weakly-supervised model to disambiguate the correct bounding boxes, we evaluate in a more difficult setup where we replace the candidate test pairs of [32] by all candidate pairs formed by objects of confidence scores above 0.3. This multiplies by 5 the number of candidate pairs, resulting in an increased level of ambiguity. In this more challenging setup, our approach obtains a recall@50 for Phrase Detection (resp. Relationship Detection) of 17.9% (resp. 12.0%) compared to the "Ours [S+A] Noisy" baseline which drops to 15.3% (resp. 10.1%).

Unseen triplets. Following [32] we report results on the "zero-shot split" of the test set containing only the test triplets not seen in training. Results for both of our fully-supervised and weakly-supervised methods are shown in Table 1 (column Unseen). Interestingly, our fully supervised model almost triples the performance on the unseen triplets compared to the Visual+Language model of [32]. Even using weak supervision, our recall of 19.0% is significantly better than their fully supervised method. We believe that this improvement is due to the strength of our visual features that generalize well to unseen triplets.

Figure 5 shows examples of predictions of both seen and unseen triplets (last row) by our model [S+A] trained with weak-supervision. We note that many of the misclassified relations are in fact due to missing annotations in the dataset (yellow column). First, not all pairs of objects in the image are labeled; second, the pairs that are labeled are not labelled exhaustively, i.e. "person riding horse" can be labelled as "person on horse" and predicting "riding" for this pair of objects is considered as an error. Not having exhaustive annotation per object pair is therefore an issue as predicates are not necessary mutually exclusive. We tackle this problem in the next section by introducing a new exhaustively labeled dataset that enables retrieval evaluation. Our real errors (red column) are mostly due to two reasons: either the spatial configuration is challenging (e.g."person on table"), or the spatial configuration is roughly correct but the output predicate is incorrect (e.g. "van has car" has similar configuration to "person has bag").

4.2. Retrieval of rare relations on UnRel Dataset

Dataset. To address the problem of missing annotations, we introduce a new challenging dataset of unusual relations, UnRel, that contains images collected from the web with

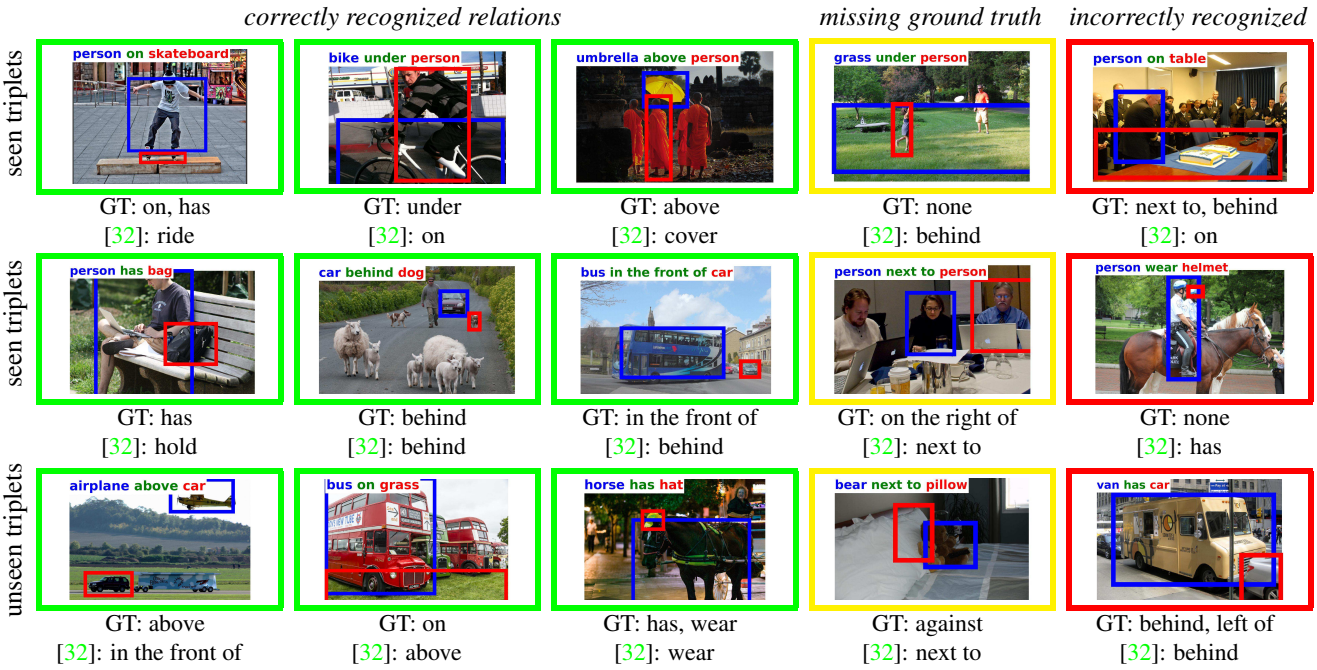

Figure 5: Relationship detections on the test set of [32]. We show examples among the top scored triplets detected for each relation by our weakly-supervised model described in 3.2. The triplet is correctly recognized if both the object detections and the relation match ground truth (in green), else the triplet is incorrect (in red). We also show examples of correctly predicted relations where the ground truth is erroneous : either missing or incomplete (in yellow). The last row shows zero-shot triplets that are not in the training set. See the appendix [1] for additional qualitative results.

unusual language triplet queries such as "person ride giraffe". We exhaustively annotate these images at box-level for the given triplet queries. UnRel dataset has three main advantages. First, it is now possible to evaluate retrieval and localization of triplet queries in a clean setup without problems posed by missing annotations. Second, as the triplet queries of UnRel are rare (and thus likely not seen at training), it enables evaluating the generalization performance of the algorithm. Third, other datasets can be easily added to act as confusers to further increase the difficulty of the retrieval set-up. Currently, UnRel dataset contains more than 1000 images queried with 76 triplet queries.

Setup. We use our UnRel dataset as a set of positive pairs to be retrieved among all the test pairs of the Visual Relationship Dataset. We evaluate retrieval and localization with mean average precision (mAP) over triplet queries $t = (s, r, o)$ of UnRel in two different setups. In the first setup (with GT) we rank the manually provided ground truth pairs of boxes $(\boldsymbol{o}_s, \boldsymbol{o}_o)$ according to their predicate scores $v_{rel}((\boldsymbol{o}_s, \boldsymbol{o}_o) \mid r)$ to evaluate relation prediction without the difficulty of object detection. In the second setup (with candidates) we rank candidate pairs of boxes $(\boldsymbol{o}_s, \boldsymbol{o}_o)$ provided by the object detector according to predicate scores $v_{rel}((\boldsymbol{o}_s, \boldsymbol{o}_o) \mid r)$. For this second setup we also evaluate the accuracy of localization : a candidate pair of boxes is positive if its IoU with one ground truth pair is above 0.3. We compute different localization metrics : $mAP–subj$ computes the overlap of the predicted subject box with the ground truth subject box, $mAP–union$ computes the overlap of the predicted union of subject and object box with the union of ground truth boxes and $mAP–subj/obj$ computes the overlap of both the subject and object boxes with their respective ground truth. Like in the previous section, we form candidate pairs of boxes by taking the top-scored object detections given by [15]. We keep at most 100 candidate objects per image, and retain at most 500 candidate pairs per image. For this retrieval task where it is important to discriminate the positive from negative pairs, we found it is important to learn an additional "no relation" class by adding an extra column to W in (3). The negative pairs are sampled at random among the candidates that do not match the image-level annotations.

Results. Retrieval results are shown in Table 2. Our classifiers are trained on the training subset of the Visual Relationship Dataset. We compare with two strong baselines. The first baseline is our implementation of [32] (their trained models are not available online). For this, we trained a classifier [41] to output predicates given visual features extracted from the union of subject and object bounding boxes. We do not use the language model as its score does not affect the retrieval results (only adding a constant offset to all retrieved images). We verified our implementation on the Visual Relationship Dataset where results of [32] are available. As the second baseline, we use the DenseCap

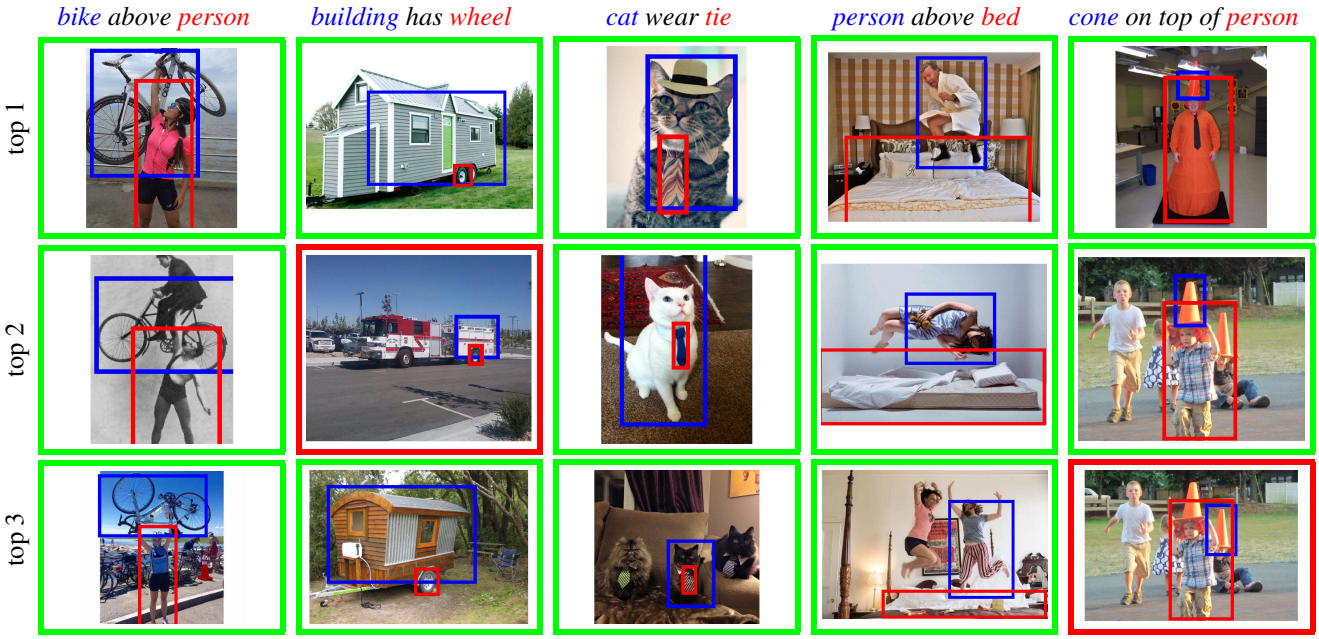

Figure 6: Top 3 retrieved pairs of boxes for a set of UnRel triplet queries (first line is best) with our weakly-supervised model. The pair is marked as positive (green) if the candidate subject and object boxes coincide with a ground truth subject and object boxes with $IoU \geq 0.3$. We provide more qualitative results in appendix [1].

[22] model to generate region candidates for each image and sort them according to the score of the given triplet query. Note that this is a strong baseline as we use the pre-trained model released by the authors which has been trained on 77K images of [28] in a fully supervised manner using localized language descriptions, compared to our model trained on only 4K training images of [32]. DenseCap outputs only a single bounding box (not a pair of boxes) but we interpret its output as either a subject box or a union of boxes. We cannot compare with the Visual Phrases [44] approach as it requires training data for each triplet, which is not available for these rare queries. We report the chance as the performance given by random ordering of the proposals. Results in Table 2 show significant improvements of our method over the baselines. Note that our weakly-supervised method outperforms these strong baselines that are fully supervised. This confirms our results from the previous section that (i) our visual features are well suited to model relations, (ii) they generalize well to unseen triplets, and (iii) training from weak image-level supervision is possible and results only in a small loss of accuracy compared to training from fully supervised data. Examples of retrieved unusual relations are shown in Figure 6.

	With GT	With candidates		
	-	union	subj	subj/obj
Chance	38.4	8.6	6.6	4.2
Full sup.				
DenseCap [22]	-	6.2	6.8	-
Reproduce [32]	50.6	12.0	10.0	7.2
Ours [S+A]	**62.6**	**14.1**	**12.1**	**9.9**
Weak sup.				
Ours [S+A]	58.5	13.4	11.0	8.7
Ours [S+A] - Noisy	55.0	13.0	10.6	8.5

Table 2: Retrieval on UnRel (mAP) with IoU=0.3

5. Conclusion

We have developed a new powerful visual descriptor for representing object relations in images achieving state-of-the-art performance on the Visual Relationship Detection dataset [32], and in particular significantly improving the current results on unseen object relations. We have also developed a weakly-supervised model for learning object relations and have demonstrated that, given pre-trained object detectors, object relations can be learnt from weak image-level annotations without a significant loss of recognition performance. Finally, we introduced, UnRel, a new evaluation dataset for visual relation detection that enables to evaluate retrieval without missing annotations and assess generalization to unseen triplets. Our work opens-up the possibility of learning a large vocabulary of visual relations directly from large-scale Internet collections annotated with image-level natural language captions.

Acknowledgements. This work was partly supported by ERC grants Activia (no. 307574), LEAP (no. 336845) and Allegro (no. 320559), CIFAR Learning in Machines & Brains program and ESIF, OP Research, development and education Project IMPACT No. CZ.02.1.01/0.0/0.0/15_003/0000468.

References

[1] Supplementary material (appendix) for the paper. http://arxiv.org/abs/1707.09472. 5, 6, 7, 8

[2] J. Andreas, M. Rohrbach, T. Darrell, and D. Klein. Neural module networks. In *CVPR*, 2016. 2

[3] F. R. Bach and Z. Harchaoui. Diffrac: a discriminative and flexible framework for clustering. In *NIPS*, 2008. 2, 3, 4

[4] H. Bilen and A. Vedaldi. Weakly supervised deep detection networks. In *CVPR*, 2016. 2

[5] P. Bojanowski, R. Lajugie, F. Bach, I. Laptev, J. Ponce, C. Schmid, and J. Sivic. Weakly supervised action labeling in videos under ordering constraints. In *ECCV*, 2014. 2

[6] A. Chang, W. Monroe, M. Savva, C. Potts, and C. D. Manning. Text to 3d scene generation with rich lexical grounding. *ACL*, 2015. 2

[7] X. Chen, A. Shrivastava, and A. Gupta. Neil: Extracting visual knowledge from web data. In *ICCV*, 2013. 2

[8] M. J. Choi, A. Torralba, and A. S. Willsky. Context models and out-of-context objects. *Pattern Recognition Letters*, 2012. 2

[9] V. Delaitre, J. Sivic, and I. Laptev. Learning person-object interactions for action recognition in still images. In *NIPS*, 2011. 2

[10] C. Desai, D. Ramanan, and C. Fowlkes. Discriminative models for static human-object interactions. In *CVPR Workshops*, 2010. 2

[11] M. Elhoseiny, S. Cohen, W. Chang, B. Price, and A. Elgammal. Sherlock: Scalable fact learning in images. *AAAI*, 2016. 2

[12] H. Fang, S. Gupta, F. N. Iandola, R. Srivastava, L. Deng, P. Dollár, J. Gao, X. He, M. Mitchell, J. C. Platt, C. L. Zitnick, and G. Zweig. From captions to visual concepts and back. In *CVPR*, 2015. 2

[13] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, M. A. Ranzato, and T. Mikolov. Devise: A deep visual-semantic embedding model. In *NIPS*. 2013. 2

[14] C. Galleguillos, A. Rabinovich, and S. Belongie. Object categorization using co-occurrence, location and appearance. In *CVPR*, 2008. 2

[15] R. Girshick. Fast R-CNN. In *ICCV*, 2015. 3, 5, 7

[16] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014. 5

[17] A. Gupta and L. S. Davis. Beyond nouns: Exploiting prepositions and comparative adjectives for learning visual classifiers. In *ECCV*, 2008. 2

[18] A. Gupta, A. Kembhavi, and L. S. Davis. Observing human-object interactions: Using spatial and functional compatibility for recognition. *PAMI*, 2009. 2

[19] L. A. Hendricks, S. Venugopalan, M. Rohrbach, R. Mooney, K. Saenko, and T. Darrell. Deep compositional captioning: Describing novel object categories without paired training data. In *CVPR*, 2016. 2

[20] R. Hu, H. Xu, M. Rohrbach, J. Feng, K. Saenko, and T. Darrell. Natural language object retrieval. *CVPR*, 2016. 2

[21] R. Jenatton, N. L. Roux, A. Bordes, and G. R. Obozinski. A latent factor model for highly multi-relational data. In *NIPS*, 2012. 2

[22] J. Johnson, A. Karpathy, and L. Fei-Fei. Densecap: Fully convolutional localization networks for dense captioning. In *CVPR*, 2016. 2, 4, 8

[23] J. Johnson, R. Krishna, M. Stark, L.-J. Li, D. A. Shamma, M. S. Bernstein, and L. Fei-Fei. Image retrieval using scene graphs. In *CVPR*, 2015. 2

[24] A. Joulin, K. Tang, and L. Fei-Fei. Efficient image and video co-localization with frank-wolfe algorithm. In *ECCV*, 2014. 2

[25] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, 2015. 2

[26] A. Karpathy, A. Joulin, and L. Fei-Fei. Deep fragment embeddings for bidirectional image sentence mapping. In *NIPS*, 2014. 2

[27] S. Kazemzadeh, V. Ordonez, M. Matten, and T. L. Berg. Referitgame: Referring to objects in photographs of natural scenes. In *EMNLP*, 2014. 2

[28] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, M. Bernstein, and L. Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. In *IJCV*, 2016. 2, 5, 8

[29] A. Lazaridou, E. Bruni, and M. Baroni. Is this a wampimuk? cross-modal mapping between distributional semantics and the visual world. In *ACL*, 2014. 2

[30] C. Li, D. Parikh, and T. Chen. Automatic discovery of groups of objects for scene understanding. In *CVPR*, 2012. 2

[31] 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 *ECCV*, 2014. 5

[32] C. Lu, R. Krishna, M. Bernstein, and L. Fei-Fei. Visual relationship detection with language priors. In *ECCV*, 2016. 1, 2, 3, 5, 6, 7, 8

[33] J. Mao, J. Huang, A. Toshev, O. Camburu, A. Yuille, and K. Murphy. Generation and comprehension of unambiguous object descriptions. *CVPR*, 2016. 2

[34] A. Miech, J.-B. Alayrac, P. Bojanowski, I. Laptev, and J. Sivic. Learning from video and text via large-scale discriminative clustering. *ICCV*, 2017. 4

[35] D. Movshovitz-Attias and W. W. Cohen. Kb-lda: Jointly learning a knowledge base of hierarchy, relations, and facts. In *ACL*, 2015. 2

[36] M. Oquab, L. Bottou, I. Laptev, and J. Sivic. Is object localization for free? weakly-supervised learning with convolutional neural networks. In *CVPR*, 2015. 2

[37] A. Osokin, J.-B. Alayrac, I. Lukasewitz, P. K. Dokania, and S. Lacoste-Julien. Minding the gaps for block Frank-Wolfe optimization of structured SVMs. In *ICML*, 2016. 4

[38] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *ICCV*, 2015. 2

[39] A. Prest, C. Schmid, and V. Ferrari. Weakly supervised learning of interactions between humans and objects. *PAMI*, 2011. 2

[40] V. Ramanathan, C. Li, J. Deng, W. Han, Z. Li, K. Gu, Y. Song, S. Bengio, C. Rossenberg, and L. Fei-Fei. Learning

semantic relationships for better action retrieval in images. In *CVPR*, 2015. 2

[41] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *NIPS*. 2015. 1, 7

[42] A. Rohrbach, M. Rohrbach, R. Hu, T. Darrell, and B. Schiele. Grounding of textual phrases in images by reconstruction. *ECCV*, 2016. 2

[43] F. Sadeghi, S. K. Divvala, and A. Farhadi. Viske: Visual knowledge extraction and question answering by visual verification of relation phrases. In *CVPR*, 2015. 2

[44] M. A. Sadeghi and A. Farhadi. Recognition using visual phrases. In *CVPR*, 2011. 1, 2, 5, 6, 8

[45] R. Socher, D. Chen, C. D. Manning, and A. Ng. Reasoning with neural tensor networks for knowledge base completion. In *NIPS*, 2013. 2

[46] R. Socher, M. Ganjoo, C. D. Manning, and A. Ng. Zero-shot learning through cross-modal transfer. In *NIPS*, 2013. 2

[47] S. Venugopalan, L. A. Hendricks, M. Rohrbach, R. Mooney, T. Darrell, and K. Saenko. Captioning images with diverse objects. *arXiv preprint arXiv:1606.07770*, 2016. 2

[48] Y. Xian, Z. Akata, G. Sharma, Q. Nguyen, M. Hein, and B. Schiele. Latent embeddings for zero-shot classification. *CVPR*, 2016. 2

[49] B. Yao and L. Fei-Fei. Grouplet: A structured image representation for recognizing human and object interactions. In *CVPR*, 2010. 2

[50] B. Yao and L. Fei-Fei. Modeling mutual context of object and human pose in human-object interaction activities. In *CVPR*, 2010. 2

[51] 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 *ICCV*, 2011. 2

[52] M. Yatskar, V. Ordonez, and A. Farhadi. Stating the obvious: Extracting visual common sense knowledge. In *NAACL*, 2016. 2

[53] S. Zagoruyko, A. Lerer, T.-Y. Lin, P. O. Pinheiro, S. Gross, S. Chintala, and P. Dollár. A multipath network for object detection. *BMVC*, 2016. 1

[54] Y. Zhu, A. Fathi, and L. Fei-Fei. Reasoning about object affordances in a knowledge base representation. In *ECCV*, 2014. 2