The Thirty-Second AAAI Conference
on Artificial Intelligence (AAAI-18)

HCVRD: A Benchmark for Large-Scale Human-Centered Visual Relationship Detection

Bohan Zhuang,[*] Qi Wu,[*†] Chunhua Shen, Ian Reid, Anton van den Hengel
Australian Centre for Robotic Vision,
The University of Adelaide, Australia

Abstract

Visual relationship detection aims to capture interactions between pairs of objects in images. Relationships between objects and humans represent a particularly important subset of this problem, with implications for challenges such as understanding human behavior, and identifying affordances, amongst others. In addressing this problem we first construct a large-scale human-centric visual relationship detection dataset (HCVRD), which provides many more types of relationship annotations (nearly 10K categories) than the previous released datasets. This large label space better reflects the reality of human-object interactions, but gives rise to a long-tail distribution problem, which in turn demands a zero-shot approach to labels appearing only in the test set. This is the first time this issue has been addressed. We propose a webly-supervised approach to these problems and demonstrate that the proposed model provides a strong baseline on our HCVRD dataset.

Introduction

The challenge in visual relationship detection (Li, Ouyang, and Wang 2017; Liang, Lee, and P. Xing 2017; Lu et al. 2016) is to capture interactions between pairs of objects in an image. In this paper, rather than detect interactions between arbitrary objects, we focus on capturing the relationships between a human and an object. Recognising human-object relationships is a problem of significant practical import, and a subtly different challenge to the more general case. Humans have a far wider variety of modes of interaction than general objects, and they have agency, meaning that more can be drawn from human-object interactions than from other interactions. For example, a human can interact with a bicycle in multiple ways (such as carry, hold, ride, park, push *etc.*), but the relationships between bicycles and other objects are far simpler. The human interactions also imply intent, and possibly provide information about the past or future that is typically lacking from object-object relationships. Previous work (Chao et al. 2017; 2015) has similarly recognised that human-object interactions of particular interest, and have proposed several datasets.

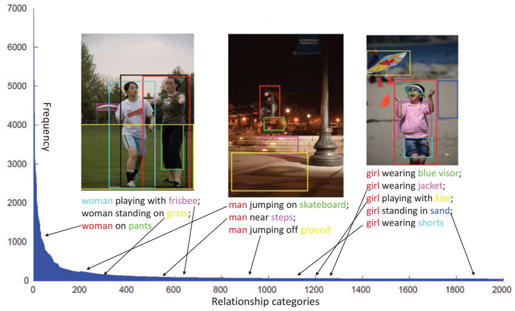


Figure 1: The long-tail label distribution of our HCVRD dataset. We only show the top-2000 relationships because the tail is too long. Three example images are also shown, with our webly-supervised model detected results. The color of human and objects in the phrases correspond to the color of the bounding boxes. The arrows indicate the 'location' of the relationship in the label distribution. As we can see, most of the relationships lie in the tail. Some of them such as 'girl wearing blue visor' is not even in the top-2000.

As in so many problems of practical interest, the label space of realistic human-centric visual relationship detection (HCVRD) exhibits a long tail distribution, meaning that there are very few, to zero, training examples for the vast majority of labels. This is a fundamental problem for the standard deep learning approach, which relies on large numbers of examples for each class. If deep learning is to progress from easy, and often artificially simplified problems for which copious training data is available, datasets will need to better reflect the practical reality of the majority of problems. The main contribution of this paper is a large-scale human-centric visual relationship detection (HCVRD) dataset that accurately depicts the long-tail label distribution of the problem, thus necessitating zero-shot recognition.

We formulate the human-centric visual relationship detection problem as that of detecting relationship triplets $\langle$*human, predicate, object*$\rangle$ in the image, with bounding boxes on the human subject and object. The HCVRD dataset is constructed based on the Visual Genome (Krishna et al. 2017). Compared to previous *human-object interaction* works (Chao et al. 2017; 2015), there are several differences. First,

[*] First two authors contributed equally to this work.

[†] Qi Wu is the corresponding author.

we have more fine-grained labels. For the 'human' item in the triplet, we are not satisfied only detecting a 'human' subject, instead, we have four sub-categories which are man(adult), woman(adult), boy and girl. This is valuable because the gender and age can affect the way that a humans interact with objects. For example, we are unlikely to find 'a man holding a Barbie' but this relationship is more commonly seen for 'a girl'. Except for the 'human' type, our 'predicate' covers a much wider range of 'relationships' than the 'interactions' in the previous setting. The dataset contains 9852 different relationships, nearly 20 times more than the HICO dataset (Chao et al. 2015). Such a large label space leads to a long-tail label distribution, *i.e.* some labels appear less than 10 times. Additionally, we provide 18,471 zero-shot relationships, *i.e.* relationships that never appear in the training split. To the best of our knowledge, this is the biggest dataset with these two forms of labels provided and that is labeled with both human-centric visual relationships and corresponding 'human' and 'object' bounding boxes.

Motivated by above challenges, our second contribution is developing methods for (i) automatically augmenting the training set using weakly labeled data crawled from the web; and (ii) performing zero-shot recognition by comparing the query data to web-retrieved data. While not radically novel in approach, our methods address the issues raised in long-tail datasets and provide, we believe, a strong baseline for further works based on our HCVRD dataset and similar data.

Related work

Visual relationship detection (Li, Ouyang, and Wang 2017; Liang, Lee, and P. Xing 2017; Lu et al. 2016) has attracted a lot of attention, thanks to the fast development of visual object detection (Deng et al. 2009), action recognition (Yao et al. 2011; Wang et al. 2013) and related problems. Recently, Lu *et al.* (Lu et al. 2016) propose a model that uses language priors from semantic word embeddings to finetune the likelihood of a predicted relationship. The recently released Visual Genome dataset (Krishna et al. 2017) provides a large-scale annotation of images containing objects, attributes and relationships. All these works are interested in the visual relationships between arbitrary two objects in the image. Although this direction is quite interesting and challenging, our focus is different. We are more concerned specifically with human-object interactions because of their particular importance in understanding factors such as intention and affordances, as well as the fact that, in practical terms, human-centric photos account for a large portion of images on the Internet.

Our work is closely related to the studies of human-object interactions (HOI), which mainly focus on learning the human actions on an object. Earlier methods, such as (Gupta, Kembhavi, and Davis 2009; Yao and Fei-Fei 2010b; 2012; Paisitkriangkrai, Shen, and van den Hengel 2014) develop joint models of body pose configuration and object location within the image. Yao and Fei-Fei (Yao and Fei-Fei 2010a) learn spatial groupings of low-level (SIFT) features for recognizing human-object interactions. Delaitre *et al.* (Delaitre, Sivic, and Laptev 2011) introduce a person-object interaction feature representation based on spatial co-occurrences of individual body parts and objects while (Desai, Ramanan, and Fowlkes 2010; Hu et al. 2013) learn a discriminative model.

Chao *et al.* (Chao et al. 2015) introduce a Humans Interacting with Common Objects (HICO) dataset which contains 117 actions and 80 objects. Each image in the dataset contains only one label of $\langle$*action, object*$\rangle$ and there are 520 such pair categories in total. Most recently, the HICO-DET (Chao et al. 2017), an incremental version of the HICO dataset is proposed. In the HICO-DET, the bounding boxes of the human and the objects in an image are annotated. Gupta *et al.* (Gupta and Malik 2015) provide a Verbs-COCO dataset which has similar settings with HICO-DET, but there are only 26 action categories. They allow for multiple persons in a single image, but restrict that each person has only one type of action on one object. Our HCVRD dataset has no such restrictions, thus, we have at least one human in the image, and each human can have multiple relationships with multiple objects. Moreover, we do not restrict that the relationship between the human and an object must be a verb or action; we provide a rich set of predicates comprising more than 900 categories. Consider we have 1,824 object categories, we finally have more than 9000 relationship triplets $\langle$*human, predicate, object*$\rangle$ (it is not necessary that a human must have relationships with every object category). A such big label space leads to a long-tailed distribution problem (see Figure 1) of the data, *i.e.* some classes may have thousands of training examples while some only have very few (less than 10).

Learning from web data (Xiao et al. 2015; Sukhbaatar and Fergus 2015) is also a related research area. Chen *et al.* (Chen and Gupta 2015) propose to pretrain a CNN on simple examples and adapt it to harder images by leveraging the structure of data and categories in a two-step manner. To better dealing with label noise in Web images, Zhuang (Zhuang et al. 2017) propose a random grouping and attention unified strategy to effectively reduce the noise of web image annotations.

Our work is also related to the few and zero-shot learning. The few-shot learning (Ravi and Larochelle 2017) problem focuses on when training sets only contain few labeled examples while the zero-shot (Lampert, Nickisch, and Harmeling 2014) aims to recognise instances for which no examples have ever been seen in training. A more comprehensive review about the zero-shot learning can be found in (Xian, Schiele, and Akata 2017). The learning from imbalanced data (He and Garcia 2009) is also related to our work.

The HCVRD Dataset

Our dataset comprises two parts, publicly available separately or together from Hiddenforblindreview. The main part comprises a carefully curated set harvested from the large Visual Genome dataset (Krishna et al. 2017). In addition we have created a supplementary component of 788,160 images drawn from the top 100 image-search results for each relationship triple.

Constructing HCVRD dataset

Our proposed human-centric visual relationship detection (HCVRD) dataset is constructed based on the Visual Genome dataset (Krishna et al. 2017), which provides detailed scene annotations, such as objects, attributes and relationships (defined as $\{$sub, predicate, obj$\}$). Since we are only interested

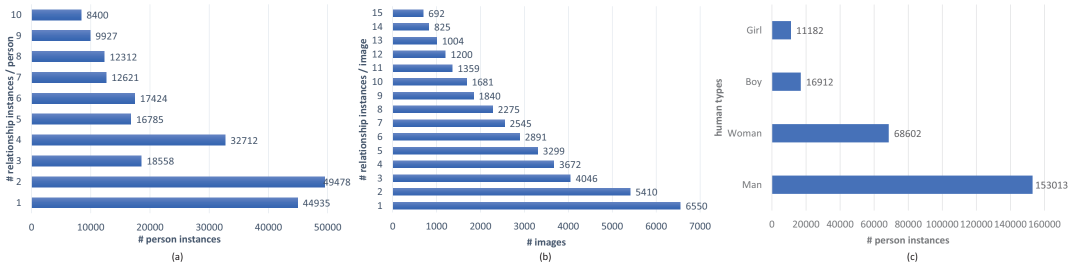


Figure 2: Statistics of the HCVRD dataset, the distribution of the (a): number of different relationships for each identified person. (b): number of relationships in each image. (c): human types.

Datasets	#relationships (no zero-shot)	#predicates	#objects	#images	#zero-shot relationships
Verbs-COCO (Gupta and Malik 2015)	-	26	80	10346	-
Stanford 40 actions (Yao et al. 2011)	40	35	28	9532	-
MPII Human Pose (Andriluka et al. 2014)	410	-	66	40522	-
HICO-DET (Chao et al. 2017)	520	117	80	47774	-
Ours	9852	927	1824	52855	18471

Table 1: Comparison of the existing human-object interaction detection datasets.

in the relationships involving human subjects, the first step is to extract all the human-related relationships from the 2.3 million relationships pool in the Visual Genome (Krishna et al. 2017). This is done automatically by searching all the relationships that their 'subject' include a 'human' concept (we use the WordNet (Leacock and Chodorow 1998) to define a 'human' concept vocabulary including human, person, people, man, male, woman, boy, girl *etc.* .)

It is worth noting that there are some relationships that only appear once in the dataset. We annotate a 'zero-shot' tag on those labels so that they can test under the zero-shot setting. This is one of the significant differences with previous human-object interaction dataset, such as the HICO (Chao et al. 2017). The zero-shot setting can verify the generalization ability of an algorithm, *i.e.* the ability to detect unseen relationships in the training set.

The collected relationships are still noisy and should be carefully processed. We first manually correct the annotations that contain misspellings and noisy characters (e.g. comma). We then eliminate the attribute predicates (such as "has", "is", "are") because these predicates are too abstract and may lead to a weak discriminative model. We further normalize the predicates by eliminating the tense using a lexical analysis toolkit (Bird, Klein, and Loper 2009) and finally have 927 predicate categories, which cover a wide range of types, such as action, spatial, preposition, comparative and verb and so on. We then merge some semantically similar objects by using the GloVe (Pennington, Socher, and Manning 2014) (*i.e.* two words are merged if their similarity calculated based on the Global Vector words representation is bigger than a threshold) and normalize (singularization and eliminate the article) the remaining object names while keeping their fine-grained attributes (e.g. black shirt, yellow shirt). Furthermore, we manually divide the 'human' subject into four more fine-grained classes according to the image content, which are man(adult), woman(adult), boy and girl. This is a valuable setting because the gender and age can affect the way that the human interacts with objects.

Dataset Statistics

Table 1 provides summary statistics about our proposed HCVRD dataset, compared with some human-object interactions dataset. In the following part, we highlight several interesting aspects of the data.

We finally have 52,855 images with 1,824 object categories and 927 predicates. In total, the dataset contains 256,550 relationships instances with 9,852 non zero-shot relationship types and 18,471 zero-shot relationships types. There are on average 10.63 predicates per object category. We use 31,586 images for training and construct two test splits. The first test split contains 10,000 images where all the relationships occur in the training set. Another test split includes all the zero-shot relationships, *i.e.* relationships in this split are never occurred in the training split. The distribution of human-object relationships in our dataset (see Figure 1) highlight the long-tail effect of infrequent relationships. Specifically, there are 370 relationships that appear more than 100 times and 7,474 relationships appear fewer than 10 times.

Figure 2 (a) shows a distribution of the number of different relationship instances that occurred on a person. Unlike past datasets where each person only can have one relationship, each people in the HCVRD dataset has on average 2.62 relationships with other objects. Figure 2 (b) shows the distribution of number of relationship instances in each image. Our HCVRD dataset has a large number of images with more than one relationship instance. On average there are 6.13 relationship instances annotated per image. Figure 2 (c) shows the distribution of human types (such as man, woman, boy and girl) in our dataset.

Supplementary web data

In addition to the curated main dataset described above we have collected a supplementary set of 788,160 images which are also available for download, and which we use in our

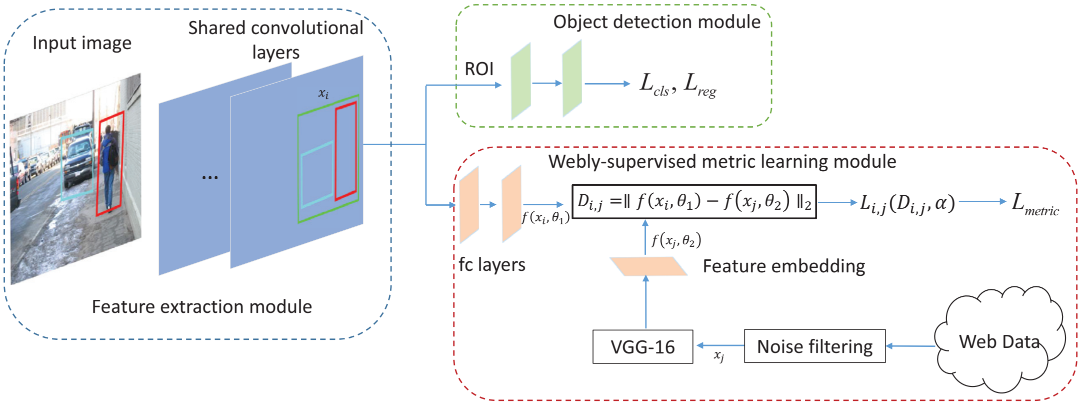

Input image
Shared convolutional layers
x_i
...
Feature extraction module
Object detection module
ROI
L_{cls}, L_{reg}
Webly-supervised metric learning module
$f(x_i, \theta_1)$
$D_{i,j} = \| f(x_i, \theta_1) - f(x_j, \theta_2) \|_2$
$L_{i,j}(D_{i,j}, \alpha)$
L_{metric}
fc layers
$f(x_j, \theta_2)$
Feature embedding
VGG-16
x_j
Noise filtering
Web Data

Figure 3: The framework of the proposed model. The model consists of (a):a feature extraction module, (b): an object detection module, (c) a webly-supervised metric learning module. The three modules can be jointly trained in an end-to-end manner.

model for metric learning, to provide a baseline for recognition in long-tailed data. To collect these images we automatically crawl using Google Images as the source of candidates. We treat all the 9,852 relationships as the query list and process each category independently, taking the top 100 images returned as representing that relationship class.

For most basic categories commonly appearing in the visual world, the top results returned by Google image search are quite clean so that we can directly learn useful visual representations from them. However some returned images may have wildly different content from the query triple, and this can adversely affect training of the model. To mitigate this issue, we employ the weakly-supervised noise robust approach of (Zhuang et al. 2017) to filter the noisy images fully automatically.

More specifically, (Zhuang et al. 2017) relies on a random group training process that randomly groups multiple web data (images) into a single training instance as the input of a classification neural network (we use a separate network for this purpose, performing 1-of-9,852 classification). As the size of the group increases, the chances diminish exponentially that a training instance (i.e. a group) does not contain imagery of the true relationship. (Zhuang et al. 2017) shows that this simple "trick" can lead to sizeable gains in accuracy when training with weakly labelled data. To determine which image or images from a group contain true positive imagery, an attentional pooling layer is employed on the last convolutional layer to determine which neuron activations have contributed to the classification. More specifically, we use the attention weights to decide a confidence score for each individual image in the random group. We then sort all images of a given relationship category according to their confidence scores, and retain the top 80% (discarding the remaining 20%). This process yields a relatively clean (though still weakly labelled) set of supplementary data that covers the entire set of 9,852 relationship categories with 80 images per category (hence 788,160).

A webly-supervised model

One of the biggest challenges in our proposed dataset is the long-tail distribution of the labels. Nearly 80% of the relationship labels in our dataset have fewer than 10 training examples. This issue creates a big challenge to the conventional supervised learning models, especially for those deep convolutional neural network based models, which normally require a large number of examples to train. Part of our purpose in creation of the dataset is to stimulate research in this important direction. To this end, we propose a strong baseline model for recognition in long-tailed data based on a so-called "webly"-supervised learning approach. Such an approach aims to leverage (practically) unlimited weakly labelled web data to overcome the restriction of limited training examples and the long-tail distribution.

An overview of our proposed webly-supervised relationship detection (WSRD) model is shown in Figure 3. Our model is divided into three sub-modules: the feature extraction module, the detection module and the distance metric learning module. The feature extraction module is a stack of convolutional layers and max-pooling layers which have the same configuration as the VGG-16 (Simonyan and Zisserman 2015) or the ResNet (He et al. 2016). The detection module is in the style of Faster-RCNN (Ren et al. 2015), which is used to detect the object and human subject (in its sub-category). A bounding box that encompasses the detected human-object pair (i.e, contains both human and object) is sent to a deep metric learning module, which performs inference by finding the nearest-neighbour match in the web-crawled data amongst all triples sharing the same human and object labels. This determines the predicate category. The neighbourhood distances are computed using the learned distance metric (i.e. in the feature space).

The three sub-modules can be learned in an end-to-end manner. For the efficiency, the feature map generated by the feature extraction module is shared as input to following two modules. We use the VGG-16 (Simonyan and Zisserman 2015) network as a basic building block for our model. We discuss the detection module and the distance metric learning module in more detail in the following sections.

Method	Predicate Det.				Phrase Det.				Relationship Det.			
	R@50		R@100		R@50		R@100		R@50		R@100	
	top-1	top-3	top-1	top-3	top-1	top-3	top-1	top-3	top-1	top-3	top-1	top-3
Multilabel	0.87	2.78	0.87	2.78	0.44	0.92	0.50	0.95	0.03	0.07	0.04	0.09
JointCNN	2.68	7.36	2.68	7.36	2.35	5.63	2.39	6.14	0.21	0.44	0.22	0.53
SeparateCNN	29.00	44.37	29.00	45.87	8.24	10.53	8.92	13.81	0.48	0.60	0.50	0.66
Ours	31.08	47.66	31.08	48.98	10.03	13.05	10.75	16.94	0.53	0.68	0.59	0.72

Table 2: Evaluation of different methods on the proposed dataset. The results reported include visual relationship detection (Relationship Det.) and predicate detection (Predicate Det.) measured by Top-100 recall (R@100) and Top-50 recall (R@50).

Method	Predicate Det.				Phrase Det.				Relationship Det.			
	R@50		R@100		R@50		R@100		R@50		R@100	
	top-1	top-3	top-1	top-3	top-1	top-3	top-1	top-3	top-1	top-3	top-1	top-3
Multilabel	0.45	1.09	0.45	1.09	0.22	0.58	0.24	0.62	0.01	0.01	0.01	0.01
JointCNN	0.02	0.03	0.02	0.03	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
SeparateCNN	15.94	26.73	15.94	26.73	0.49	1.55	0.58	1.96	0.04	0.08	0.05	0.10
Ours-without web data	18.01	29.35	18.01	29.35	0.73	2.15	0.80	2.43	0.06	0.10	0.07	0.13
Ours	24.55	36.59	24.55	36.59	1.76	3.62	1.91	4.56	0.12	0.16	0.14	0.21

Table 3: Results for human-object relationship detection on the long-tail benchmark subset.

Detection module

The object (and human subject) detection module structure is identical to that of the Faster-RCNN (Ren et al. 2015). Taking the output of the feature extraction module (`Conv5_3` feature map) as the input, the Region Proposal Network (RPN) is used to generate object proposals. During training, we extract features with RoIPool for each object proposal, followed by the bounding box regression loss L_{reg} and a classification loss L_{cls} to learn the detector/classifier in a manner identical to (Ren et al. 2015). During inference, we use this module to detect all human subjects and objects in the images. We apply non-maximum suppression (NMS) to reduce the number of proposals with the IoU (Intersection of Union) threshold 0.3 and objectiveness scores higher than 0.2. These filtered boxes are further grouped to all possible ⟨*human, object*⟩ pairs and a bounding box that fully contains the human and object boxes is associated to each pair. These "union" bounding boxes are (separately and individually) the input to the distance metric learning module.

Distance metric learning module

As noted above, this module accepts a union region of the detected human and object, and computes the feature-space distance between the proposed region and all of the web-crawled visual relationship data. The nearest class label of the web data is assigned to the proposed region. The distance metric function is learned via deep metric learning on the web-crawled (supplementary) data.

More specifically, the deep metric learning process aims to learn a semantic feature embedding (a feature space) for which similar examples are mapped close to one another while dissimilar examples are mapped further apart. To this end, we construct a set of positive pairs and a set of negative pairs by drawing from the main dataset and the web data. Each positive pair $(\mathbf{x}_i, \mathbf{x}_j)$ contains a sample from the main HCVRD dataset and a sample from the web data with the same label, while each negative pair is similarly drawn one from each, but with non-matching labels. We follow (Oh Song et al. 2016) to incrementally add the positive and negative pairs. Specifically, we first sample a few anchor pairs and then active mining hard negative images to the batch, more details can be found in (Oh Song et al. 2016).

During the training, the ground truth predicate region $\mathbf{x}_i$'s corresponding `Conv5_3` feature map is used as part of the input for the metric learning module. In the inference, we first detect the human and objects and get all the possible union bounding boxes' corresponding `Conv5_3` feature map as the input, separately and individually. Then the convolutional feature map is sent to two fully connected layers and the output $f(x_i, \theta_1)$ serves as part of the input for the metric learning functions (see equation (1)), where f is the feed-forward function and θ_1 is the learnable parameters of the feature extraction module with the fully connected layers. Another input $f(x_j, \theta_2)$ of the metric learning functions is from the collected web data, which is passed through a pre-trained VGG-16 model and a learnable feature embedding layer with parameter θ_2. Following (Oh Song et al. 2016), the metric is then learned using a structured loss function based on the sampled positive and negative pairs of training samples:

$$\begin{aligned} &L_{mec} = \frac{1}{2|\mathbb{P}|} \sum_{(i,j)\in\mathbb{P}} \max(0,\ L_{i,j})^2, \\ &L_{i,j} = \log\Big(\sum_{(i,k)\in\mathbb{N}} \exp(\alpha - D_{i,k}) + \sum_{(j,l)\in\mathbb{N}} \exp(\alpha - D_{j,l})\Big) + D_{i,j} \end{aligned} \tag{1}$$

where $\mathbb{P}$ is the set of positive pairs and $\mathbb{N}$ is the set of negative pairs, $D_{i,j} = \|f(x_i, \theta_1) - f(x_j, \theta_2)\|_2$ is distance between two embedding feature vectors. The α is the learnable margin parameter.

The two modules can be jointly trained in an end-to-end manner. The model employs multi-task loss for human-object relationship detection:

$$L = L_{reg} + L_{cls} + L_{mec} \tag{2}$$

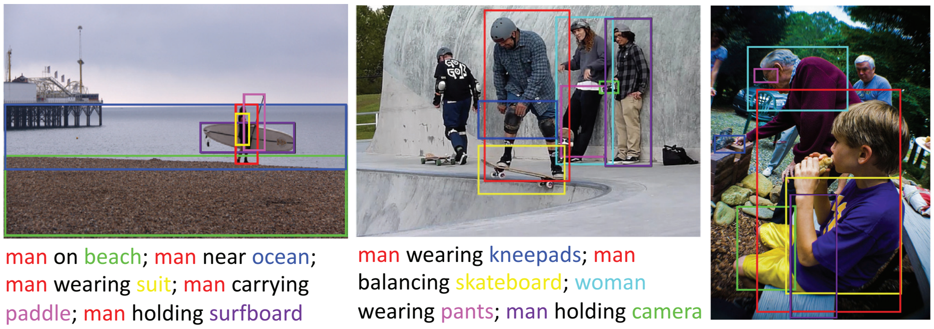
where L_{reg} and L_{cls} are the regression loss and cross-entropy loss in the detection module.

Method	Predicate Det.				Phrase Det.				Relationship Det.			
	R@50		R@100		R@50		R@100		R@50		R@100	
	top-1	top-3	top-1	top-3	top-1	top-3	top-1	top-3	top-1	top-3	top-1	top-3
(a) Without metric learning module	22.55	33.12	22.55	33.87	5.87	7.33	6.04	9.44	0.29	0.43	0.34	0.49
(b) Without noise filtering	30.36	46.12	30.37	46.68	9.92	12.96	10.67	16.36	0.49	0.64	0.57	0.70
(c) Ours (full model)	31.08	47.66	31.08	48.98	10.03	13.05	10.75	16.94	0.53	0.68	0.59	0.72

Table 4: Ablation studies on the HCVRD benchmark non-zeroshot test set.

Method	Predicate Det.				Phrase Det.				Relationship Det.			
	R@50		R@100		R@50		R@100		R@50		R@100	
	top-1	top-3	top-1	top-3	top-1	top-3	top-1	top-3	top-1	top-3	top-1	top-3
Multilabel	-	-	-	-	-	-	-	-	-	-	-	-
JointCNN	-	-	-	-	-	-	-	-	-	-	-	-
SeparateCNN	2.75	4.98	2.99	5.93	0.06	0.11	0.07	0.16	0.01	0.05	0.03	0.08
Ours	8.15	12.34	8.57	13.42	0.88	1.43	0.92	1.84	0.03	0.09	0.05	0.12

Table 5: Results for human-object relationship detection on the zero-shot benchmark test set.



man on beach; man near ocean; man wearing suit; man carrying paddle; man holding surfboard

man wearing kneepads; man balancing skateboard; woman wearing pants; man holding camera

man with glasses; man using forks; boy wearing yellow pants; boy wearing purple shirt ; boy on bench

Figure 4: Qualitative examples of the predicate detection. The color of human and objects in the phrases correspond to the color of the bounding boxes. We only predict the interactions between the ground-truth bounding box pairs.

Experiments

Implementation details

We set the feature embedding size in the metric learning module as 256. For training efficiency, we initialize the feature extraction module with the pre-trained VGG-16. We then pretrained the detection module and fix it while training the metric learning module. The learning rate is initialized to 0.0001 and decreased by a factor of 10 after every 5 epochs. During the inference, we first retrieve the top 20 nearest neighbor relationships and select those including both detected human and object categories. Then we use the top-ranked selected candidates for evaluation.

Evaluation Setup

We evaluate our human-object interactions task using Recall@100 and Recall@50, following the setting of Visual Relationship Detection (VRD) task (Liang, Lee, and P. Xing 2017; Lu et al. 2016). Recall@x computes the fraction of times the correct relationship is calculated in the top x predictions, which are ranked by the final distances. We evaluate on three tasks: (1) For **predicate detection**, the goal is to predict the accuracy of *predicate* recognition, where the groundtruth labels and bounding boxes for both the *object* and *human* are given. (2) In **phrase detection**, we aim to predict $\langle$*human-predicate-object*$\rangle$ and localize the entire relationship in one bounding box. (3) For **relationship detection**, the task is to recognize $\langle$*human-predicate-object*$\rangle$ and localize both human and object bounding boxes, where both boxes should have at least 0.5 overlap (IoU) with the ground-truth in order to be regarded as correct prediction. In the real world applications, different relationships may share very similar semantic meanings (e.g. "man holding phone", "man talking on phone", "man using phone") and it's difficult to differentiate them. Hence, in many cases, one "appropriate" prediction may be judged "incorrect" due to the limitation of the test annotations, which is a common problem of the current VRD evaluation metric. One possible solution is to employ the human evaluation, which is cost however. In this paper, we instead report both top-1 and top-3 results under different Recalls to evaluate the model.

Baselines

We benchmark the following approaches on our new dataset and results are reported in Table 2.

Multilabel classification A person can concurrently perform different interactions with different target objects, e.g. a person can "ride bicycle" and "drink water" at the same time. Thus we treat the human-object relationship detection task as a multilabel classification problem where we apply a sigmoid cross entropy loss on top of the classification layer. Specifically, we treat the union of a human and its correlated objects as the input during training. During the testing, we use our object detection module to return the regions. We use VGG-16 model as the basis building block.

JointCNN This implements the Visual phrases (Sadeghi and Farhadi 2011). We train a VGG-16 model to jointly predict the three components of a relationship. Specifically, we treat each relationship category separately and train a 9,852 way classification model.

SeparateCNN Following the visual model of (Lu et al. 2016), we first train a VGG-16 model to classify the 1,824 objects. Similarly, we train a second model to classify each of the 927 predicates using the union of the bounding boxes of the participating human and the object in that relationship.

For *JointCNN* and *Multilabel* baselines, we empirically find that due to the long-tail property of the dataset, the learned models are seriously biased. It causes the predictions only fall into those labels with large numbers of training examples. To solve the problem of extreme classification with enormous number of categories, we instead propose to employ the metric learning approach with web data to perform efficient nearest neighbor inference on the learned metric space. By comparing *ours* with the two baselines, we find significant performance increase on all evaluation metrics.

For the *SeperateCNN* baseline, since the training data for human, objects and predicates are relatively adequate respectively, its performance is competitive with our proposed method. In other words, the human, objects and predicates are predicted separately, hence, the label prediction space is much smaller than above two baseline approaches. However, compared to predicate detection results, the performance of phrase and relationship detection decreases a lot. It shows that detecting such wide range of objects is a major challenge for visual relationship detection.

Long-tail evaluation

Due to the long-tail distribution of the categories in the dataset, the infrequent relationships will contribute not much to the final testing performance. But in real world applications, the relationships in long-tail should not be ignored. So we select those relationships that appear less than 10 times as a subset (i.e. there are totally 7,474 relationships) and report the performance in Table 3. From the table, we can see that our approach performs steadily better than the baseline methods. For the baseline methods, the lack of training data is a main challenge for obtaining accurate predictions. The main motivation of the proposed method is to utilize web data to tackle this limitation. With the always available web data, we can learn the distance metrics and efficiently infer nearest neighbor relationships on the learned metric space.

Ablation study

With vs. without metric learning module Metric learning module is the key component of our system. To evaluate its impact, we implement a variant without the metric learning module. For the detected union bounding boxes of relationships and web data, we directly extract the 4096-dimentional feature vector for each sample using the pretrained VGG-16 model. We then compute the cosine similarity between the test sample and all mean vectors of the relationship categories that contain both detected human and object types. We then retrieve the nearest neighbor relationship categories as our predictions. Table 4 (a) vs. (c) shows that learning the semantic feature embeddings via distance metric contributes a lot to the final performance.

With vs. without web data We also evaluate the influence of the web data by only using the training data of the dataset. Since one motivation of introducing web data is to solve the scarceness of training data, we report this variant under the long-tail setting in Table 3 as *Ours-without web data*. By comparing it with *Ours* in Table 3, we find that removing web data causes an obvious performance degradation, which proves the effectiveness of introducing the web data. We find that the web data can help on some relationships that rarely happened in the dataset, such as 'man cooking on street' and 'man peddling rickshaw'.

With vs. without noise filtering We further remove the noise filtering step to investigate the affect of noisy labels. The results are shown in Table 4 (b). Table 4 (b) vs. (c) shows that removing noise filtering have less affect to the performance compared to removing metric learning module. This is because for relationships that commonly used in the visual content, top results returned by Google images search are pretty clean. Noise filtering provides an auxiliary to further improve the quality of web data.

Zero-shot evaluation

It is quite important to make the model generalizable to unseen human-object relationships. 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 relationship detection performance on the $18,471$ unseen visual relationships in the zero-shot test split. Given the detected human and objects in a relationship, we first get all their possible interactions to form a search space. We then collect web data and extract feature embeddings to get the nearest neighbors relationships for the test sample. The results are reported in Table 5. We can see that the proposed method works more robust. This can be attributed to the introduction of the external web data for efficient nearest neighbor search. For the "separateCNN" baseline, by predicting the predicates separately from its objects, it is difficult to capture the appearance variations due to the weak and even ambiguous visual features.

Conclusion

We have proposed a large-scale human-centric visual relationship detection (HCVRD) dataset, which is significantly larger and broader than previous datasets. Human-centric relationships represent an important subclass of all relationships, not only because the human has agency, but also due to their practical importance for other challenges. Increasing the scale of data available better captures the reality of the task, but rises two important practical problems, the long-tail distribution issue and the zero-shot problem, which are both reflected in our proposed HCVRD dataset. Motivated by the practical importance of the task, our webly-supervised method addresses the issues and provides a strong baseline for further works based on our HCVRD dataset and similar data.

References

Andriluka, M.; Pishchulin, L.; Gehler, P.; and Schiele, B. 2014. 2d human pose estimation: New benchmark and state of the art analysis. In *CVPR*.

Bird, S.; Klein, E.; and Loper, E. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Chao, Y.-W.; Wang, Z.; He, Y.; Wang, J.; and Deng, J. 2015. Hico: A benchmark for recognizing human-object interactions in images. In *ICCV*.

Chao, Y.-W.; Liu, Y.; Liu, X.; Zeng, H.; and Deng, J. 2017. Learning to Detect Human-Object Interactions. *arXiv preprint arXiv:1702.05448*.

Chen, X., and Gupta, A. 2015. Webly supervised learning of convolutional networks. In *ICCV*.

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

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *CVPR*.

Desai, C.; Ramanan, D.; and Fowlkes, C. 2010. Discriminative models for static human-object interactions. In *CVPR workshops*.

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

Gupta, A.; Kembhavi, A.; and Davis, L. S. 2009. Observing human-object interactions: Using spatial and functional compatibility for recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

He, H., and Garcia, E. A. 2009. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*.

Hu, J.-F.; Zheng, W.-S.; Lai, J.; Gong, S.; and Xiang, T. 2013. Recognising human-object interaction via exemplar based modelling. In *ICCV*.

Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.-J.; Shamma, D. A.; et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*.

Lampert, C. H.; Nickisch, H.; and Harmeling, S. 2014. Attribute-based classification for zero-shot visual object categorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Leacock, C., and Chodorow, M. 1998. Combining local context and WordNet similarity for word sense identification. *WordNet: An electronic lexical database*.

Li, Y.; Ouyang, W.; and Wang, X. 2017. ViP-CNN: Visual Phrase Guided Convolutional Neural Network. In *CVPR*.

Liang, X.; Lee, L.; and P. Xing, E. 2017. Deep Variation-structured Reinforcement Learning for Visual Relationship and Attribute Detection. In *CVPR*.

Lu, C.; Krishna, R.; Bernstein, M.; and Fei-Fei, L. 2016. Visual relationship detection with language priors. In *ECCV*.

Oh Song, H.; Xiang, Y.; Jegelka, S.; and Savarese, S. 2016. Deep metric learning via lifted structured feature embedding. In *CVPR*.

Paisitkriangkrai, S.; Shen, C.; and van den Hengel, A. 2014. Strengthening the effectiveness of pedestrian detection with spatially pooled features. In *ECCV*.

Pennington, J.; Socher, R.; and Manning, C. D. 2014. Glove: Global Vectors for Word Representation. In *EMNLP*.

Ravi, S., and Larochelle, H. 2017. Optimization as a model for few-shot learning. In *ICLR*.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*.

Sadeghi, M. A., and Farhadi, A. 2011. Recognition using visual phrases. In *CVPR*.

Simonyan, K., and Zisserman, A. 2015. Very deep convolutional networks for large-scale image recognition. In *ICLR*.

Sukhbaatar, S., and Fergus, R. 2015. Learning from noisy labels with deep neural networks. In *ICLR workshops*.

Wang, Z.; Shi, Q.; Shen, C.; and Van Den Hengel, A. 2013. Bilinear programming for human activity recognition with unknown mrf graphs. In *CVPR*.

Xian, Y.; Schiele, B.; and Akata, Z. 2017. Zero-Shot Learning-The Good, the Bad and the Ugly. *arXiv preprint arXiv:1703.04394*.

Xiao, T.; Xia, T.; Yang, Y.; Huang, C.; and Wang, X. 2015. Learning from massive noisy labeled data for image classification. In *CVPR*.

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

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

Yao, B., and Fei-Fei, L. 2012. Recognizing human-object interactions in still images by modeling the mutual context of objects and human poses. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Yao, B.; Jiang, X.; Khosla, A.; Lin, A. L.; Guibas, L.; and Fei-Fei, L. 2011. Human action recognition by learning bases of action attributes and parts. In *ICCV*.

Zhuang, B.; Liu, L.; Li, Y.; Shen, C.; and Reid, I. 2017. Attend in Groups: A Weakly-Supervised Deep Learning Framework for Learning From Web Data. In *CVPR*.