Recognizing Open-Vocabulary Relations between Objects in Images

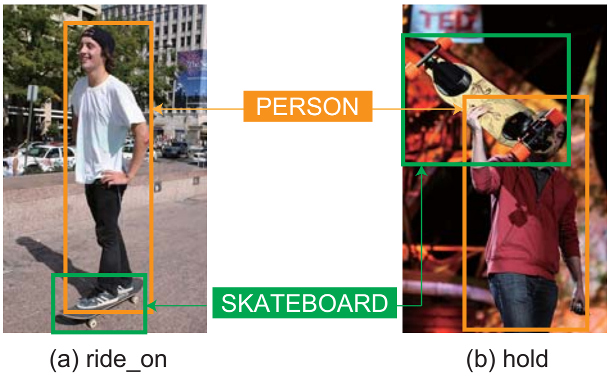
Masayasu Muraoka[*] **Sumit Maharjan**[†] **Masaki Saito**[‡]
Kota Yamaguchi[‡] **Naoaki Okazaki**[†] **Takayuki Okatani**[‡] **Kentaro Inui**[†]
IBM Research – Tokyo[*] Tohoku University[†‡]
mmuraoka@jp.ibm.com[*]
{sumit,okazaki,inui}@ecei.tohoku.ac.jp[†]
{msaito,kyamagu,okatani}@vision.is.tohoku.ac.jp[‡]

Abstract

How can we describe the relations between objects in a picture? As recent deep neural networks have exhibited impressive performance in identifying individual entities in a picture, in this study we turn our attention to recognize inter-object relations. To recognize open-domain relations, (a) we propose collecting relational concepts automatically from an image-text corpus. In addition, using collected relational instances, (b) we train a classifier to recognize inter-object relations. A relation recognition experiment conducted in our study suggests that relative information calculated from objects improves relation recognition effectively.

1 Introduction

Generating image descriptions draws considerable attention in the natural language processing and computer vision communities. Recent studies have addressed this task by using a Deep Neural Network (DNN) (Kiros et al., 2014; Karpathy and Fei-Fei, 2015; Vinyals et al., 2015; Donahue et al., 2015; Johnson et al., 2015). Even though these studies provide elegant end-to-end solutions, they essentially extract visual features trained for an object recognition task, and plug them into a (variant of) neural language model. In other words, these studies essentially utilize the language model to put the 'pieces' of recognized objects into a sentence.



(a) ride_on (b) hold

Figure 1: Different relations between a man and skateboard.

One possible drawback of this approach is that these studies do not necessarily recognize the structure of objects in an image, whereas a sentence typically exhibits a syntactic/semantic structure. More specifically, they do not consider the spatial (positional, magnitude, tangent, etc.) or action relations between objects in an image. Therefore, it may be relatively easy to generate the description *a man rides on a skateboard* for Figure 1 (a) because the major relation between the person and skateboard is *ride_on*. In contrast, we need to focus on the positional relationship between the man and skateboard in Figure 1 (b), and verbalize the relationship as *hold* for generating the description *a man holds a skateboard*.

Now that DNN models have reached the level of a human's ability for recognizing objects, as shown in the ImageNet Large Scale Visual Recognition Challenge 2015 (ILSVRC2015) (He et al., 2015), we believe that the primary and important next step toward image understanding is to recognize relations

[*] This work was conducted while the author was in Tohoku University.

between objects in an image. Recognizing relations between objects also opens up new applications such as image retrieval using a subject-verb-object (SVO) triplet (Farhadi et al., 2010), reasoning with relational knowledge grounded with both the image and text (Sadeghi et al., 2015), and enrichment of common-sense knowledge with visual information.

However, only a few studies have addressed relation recognition between objects. Elliott and de Vries (2015), Kong et al. (2014), and Lin et al. (2015) classified a pair of objects into a relation from a small number of manually-defined relations, but their types are restricted to positional ones (e.g., *close_to*, *on_top_of*, and *in_front_of*), not including other types of relations such as actions (e.g., *ride*, *throw*, and *eat*).

In this paper, we present the first approach for open-vocabulary relation recognition between objects in images. The contributions of this paper are two fold.

(a) We propose to automatically extract relation instances between objects in images, e.g., *ride_on*(PERSON, SKATEBOARD), using the IBM Model and the dependency information of descriptions.

(b) We train a classifier that recognizes relations between objects with novel features (e.g., positional or regional feature and more), and demonstrate the effectiveness through the experiments.

2 Related work

2.1 Relation recognition between objects

Elliott and de Vries (2015) proposed Visual Dependency Representation (VDR) to represent dependency relations between objects in images. VDR categorizes a relation of a pair of objects in five positional relations: *beside*, *above*, *below*, *on*, and *surrounds*. They reported that the VDR-based method could achieve a comparable performance to that using DNN. Although they did not evaluate the effectiveness of VDR in relation recognition, the results indicated the importance of identifying inter-object relations for description generation.

Kong et al. (2014) proposed the use of a Markov Random Field (MRF) for building a relational graph representing inter-object relations. A node in the graph denotes either an object in an image or a noun in a caption describing the image. MRF trains the mapping of objects to nouns. Their approach considers two types of relations (*close–to*, *on–top–of*) as the edge potential functions of MRF to capture a spatial relation between the objects. Extending the work of Kong et al. (2014), Lin et al. (2015) addressed a task for generating multiple sentences for indoor scenes, and built a scene graph from an image. In addition, they incorporated attribute expressions (e.g., the color and size of an object) to vertices of the graph in order to generate detailed descriptions of the scene. In their work, a relation is defined by eight labels (*next–to*, *near*, *top–of*, *above*, *in–front–of*, *behind*, *to–left–of*, *to–right–of*).

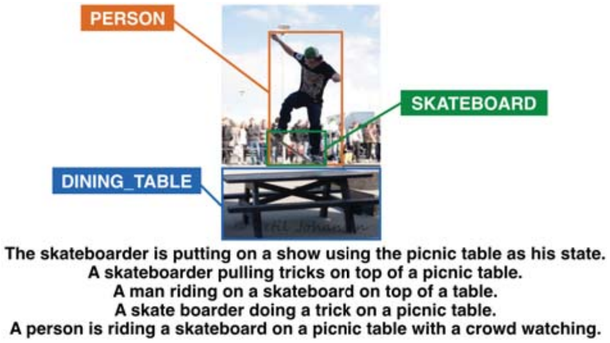
Unlike the previous work (Elliott and de Vries, 2015; Kong et al., 2014; Lin et al., 2015), we do not define relations in advance. Instead, we extract the vocabulary of relations between objects that are used frequently to describe the scenes of images in a dataset. This approach can naturally include action relations such as *look_at*, *throw*, and *eat*, which have never previously been explored.

The closest work to this paper is that by Aditya et al. (2015), in that they do not pre-define a relation vocabulary but instead extract relations from image descriptions. Their approach associates object categories (e.g., PERSON) with words (e.g., man) by using the WordNet hierarchy[1]. They extracted inter-object relation instances by applying a semantic parser named the XYZ Parser (formally named the K-parser)[2] to image descriptions.

Our approach is different from that of Aditya et al. (2015) in two aspects. Firstly, we bridge object categories and textual expressions by using an alignment model for statistical machine translation. As we will show in Section 5.3, our alignment model outperforms the method using the WordNet hierarchy. Secondly, they did not develop a relation recognizer between two image objects, but only extracted relation instances for constructing a knowledge base. In con-

[1] It is trivial to map an object category to textual expressions because category labels were originally defined in ImageNet dataset and taken from the WordNet hierarchy.

[2] `http://kparser.org`



The skateboarder is putting on a show using the picnic table as his state.
A skateboarder pulling tricks on top of a picnic table.
A man riding on a skateboard on top of a table.
A skate boarder doing a trick on a picnic table.
A person is riding a skateboard on a picnic table with a crowd watching.

Figure 2: An instance in the MS COCO dataset.

trast, we build a classifier that predicts a relation between a pair of objects.

2.2 Caption generation from images

Description generation is a fascinating application of image understanding. A number of studies applied DNNs for generating image descriptions with the availability of a large amount of training data (Karpathy and Fei-Fei, 2015; Vinyals et al., 2015; Donahue et al., 2015; Johnson et al., 2015; Kiros et al., 2014). A typical approach combines a Convolutional Neural Network (CNN) with a variant of a Recurrent Neural Network (RNN). We can view this approach as an instance of an *encoder-decoder model*, where an encoder (CNN) represents an input image with abstract features, and a decoder (RNN) realizes a sentence from the feature representation.

This architecture seemingly has the ability to recognize object categories as well as relations between objects in an image. However, the end-to-end models adopted in these studies make an analysis of the internal mechanism for generating image descriptions intractable. Furthermore, these models do not encode spatial relationships between image objects. Thus, no one has demonstrated that these studies really recognize relations between objects.

3 Dataset for image and description

We explore relations between objects using the MS COCO dataset (Lin et al., 2014)[3]. MS COCO is a large-scale collection of images depicting various objects in the scene, with an emphasis on the contextual relationship between multiple objects. The dataset was originally designed for various tasks including language generation, object segmentation, and context understanding between multiple objects. The dataset contains 328k images, distributed under the Creative Commons Attribution 4.0 License[4] and Flickr Terms of Use[5].

The dataset annotates objects with a single category (out of 80 categories) and a bounding box (e.g., the blue, green and yellow rectangles in Figure 2). A bounding box is represented by four values (x, y, w, h), where x and y represent the top-left coordinates of the bounding box, and w and h are the width and height, respectively, of the box. Throughout this work, we use the object categories and bounding boxes annotated in the dataset as the ground truth.

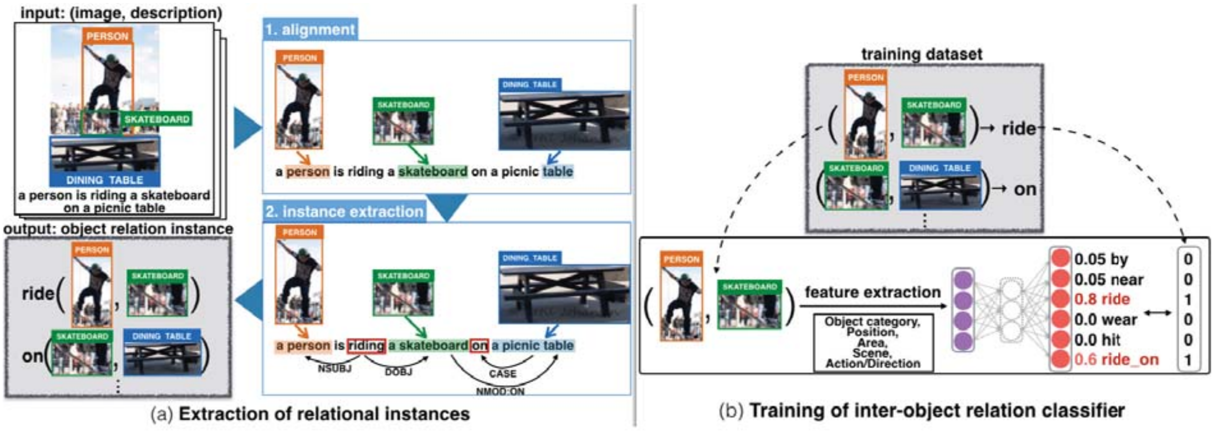
In addition, MS COCO includes five manually written descriptions (sentences) per image (see Figure 2). We utilize these image descriptions to discover relations between objects. For example, the third sentence in Figure 2 expresses the *ride_on* relation between *a man* and *a skateboard*. If we could ground the man with a yellow bounding box (PERSON) and *a skateboard* with a green box (SKATEBOARD), we could understand the meaning of $\textit{ride\_on}(o_1, o_2)$ relation via the image: the object o_1 has a contact with o_2, and o_1 is usually located above o_2. Unfortunately, because the MS COCO dataset does not have alignments between images and words in its descriptions, we estimate the alignments, as will be explained in Section 5.1.

There are other publicly available datasets, such as the VLT2K (Elliott and Keller, 2013), PASCAL VOC (Everingham et al., 2014), Stanford 40 Actions (Yao et al., 2011), and HICO (Chao et al., 2015). Those datasets contain relation information, although the information restricts only positional or action ones: VLT2K has only positional relations, PASCAL VOC and Stanford 40 Actions contain action relations (e.g., *walking* and *running*), and HICO has human-object relations (e.g., *riding a bike*). We aim at a generic natural image understanding that might involve object-object relationships other than people, and we consider the MS

[3] http://mscoco.org/

[4] https://creativecommons.org/licenses/by/4.0/legalcode

[5] https://info.yahoo.com/legal/us/yahoo/utos/utos-173.html


input: (image, description)
PERSON
SKATEBOARD
DINING TABLE
a person is riding a skateboard on a picnic table
output: object relation instance
ride(
on(
1. alignment
a person is riding a skateboard on a picnic table
2. instance extraction
a person is riding a skateboard on a picnic table
NSUBJ
DOBJ
CASE
NMOD:ON
training dataset
→ ride
→ on
feature extraction
Object category, Position, Area, Scene, Action/Direction
0.05 by
0.05 near
0.8 ride
0.0 wear
0.0 hit
0.6 ride_on
0
0
1
0
0
1

(a) Extraction of relational instances

(b) Training of inter-object relation classifier

Figure 3: Our approach to open-vocabulary relation recognition between image objects: (a) automatic acquisition of relation instances and (b) training of a classifier to recognize inter-object relations.

COCO dataset to be more appropriate for our purpose of open-vocabulary recognition.

4 Our approach

Figure 3 illustrates our approach. We first associate objects in an image with their corresponding expressions in the description, adapting an alignment model for statistical machine translation (Section 5.1). Using the alignments and dependency parses of image descriptions, we extract relation instances whose arguments are grounded to image objects, and whose relations include various expressions that are commonly perceived and described for two objects (Section 5.2). Unlike the previous rule-based approaches (Elliott and Keller, 2013; Elliott and de Vries, 2015; Kong et al., 2014; Lin et al., 2015), our approach does not require hand-crafted relation labels or manual annotations between objects.

Using the relation instances, we train a relation recognizer that predicts a relation for a given pair of unseen image objects (Section 6). The relation recognizer is modeled by a three-layer neural network, whose input provides various features for two given objects: object categories, relative coordinates and intersection areas, etc. The relation instances include multiple relations between the same pair of objects in the image because MS COCO involves five independent descriptions. For example, the relation between the PERSON and SKATEBOARD is described by *ride_on* and *ride* in Figure 2. Thus, we design the recognizer such that it can also handle multiple relations between a pair of objects rather than force a single relation as the ground truth.

5 Extracting relation instances

5.1 Aligning image objects and text

Although the MS COCO dataset contains only 80 object categories (e.g., PERSON or CAR), each object category is referred to by a number of expressions. For example, the object category PERSON can be described by *man*, *person*, *skateboarder*, *skate boarder*, etc., as shown in Figure 2. Thus, we need to identify the correspondences between objects in an image and their referring expressions in the dataset.

In this study, we cast the problem of object-word alignment as a translation task, where the input language is a set of object categories and the output language is a description. Here, we use the IBM Model (Brown et al., 1993) to obtain the translation probability $P(w|c)$, where c denotes an object category in an image and w denotes a word in its description. For instance, the IBM Model gives a higher probability for $P(w = \mathit{man}|c = \text{PERSON})$ after seeing the training instances:

PERSON, SKATEBOARD
a man is riding a skateboard

PERSON, DONUT
a man who is eating a donut

We use the GIZA++ (Och and Ney, 2003) implementation[6] to estimate the alignments.

[6] https://github.com/moses-smt/giza-pp

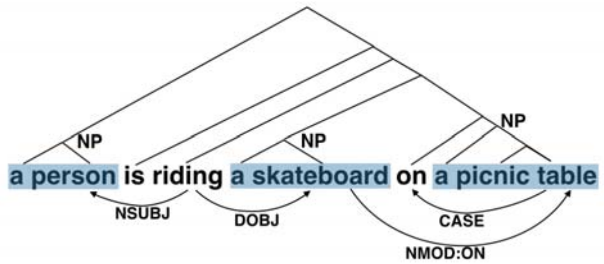


Figure 4: An output from the Stanford CoreNLP. The upper part depicts a phrase-structure tree, and the lower part shows a dependency tree. Phrases in blue represent noun phrases.

5.2 Extracting relation instances

We extract object-relation instances from a description with the image objects aligned. Suppose that we have a description that is aligned with the image objects.

*a man/*PERSON *riding*
*a skateboard/*SKATEBOARD
*on a picnic table/*DINING_TABLE

Here, we denote an object category in uppercase letters followed by a word and a slash. We extract two relation instances from the example:

ride(*a man*/PERSON,
a skateboard/SKATEBOARD),
on(*a skateboard*/SKATEBOARD,
a picnic table/DINING_TABLE)

Because PERSON, SKATEBOARD, and DINING_TABLE are associated with the objects in the image, the two relation instances describe the relations between PERSON and SKATEBOARD objects as *ride* and between SKATEBOARD and DINING_TABLE objects as *on*.

We design a method for extracting relation instances from dependency trees of image descriptions, inspired by the methods for Open Information Extraction (Schmitz et al., 2012; Nakashole et al., 2012; Xu et al., 2013; Moro and Navigli, 2013). We first parse a description using the Stanford CoreNLP (Manning et al., 2014)[7]. We find a set of the longest noun phrases (NPs) whose phrase structures are located at nodes of height no greater than three from their leaves (in blue in Figure 4)[8].

Table 1: Result of object-word alignment.

	Precision	Recall	F1
IBM Model	**.880**	**.743**	**.806**
WordNet	.738	.565	.638

We extract inter-object relation instances using the following templates.

1. $v(o_1, o_2)$: $o_1 \xleftarrow{\text{nsubj}} v \xrightarrow{\text{dobj}} o_2$
   e.g., *ride*(*a man*, *a skateboard*)
2. $v\_p(o_1, o_2)$: $o_1 \xleftarrow{\text{nsubj}} v \xrightarrow{\text{nmod}} o_2 \xrightarrow{\text{case}} p$
   e.g., *ride_on*(*a man*, *a skateboard*)
3. $p(o_1, o_2)$: $o_1 \xrightarrow{\text{nmod}} o_2 \xrightarrow{\text{case}} p$
   e.g., *on*(*a skateboard*, *a picnic table*)

Templates 1 and 3 extract their example instances from Figure 4. Template 2 is used to extract the example from the sentence, "A man is riding on a skateboard." In Template 1, we attach a particle (compound:prt) if any to the verb for extracting *take_off*(*a man*, *the hat*) from the sentence, "A man is taking off the hat." In this way, we extracted 156,293 instances with 5,153 distinct relations from the MS COCO dataset.

5.3 Experiments

Table 1 reports the quality of the object-word alignments in terms of precision, recall, and F1. The performances were measured on a test set with 50 images that were sampled randomly from the MS COCO dataset; we annotated the gold-standard alignments for the 250 descriptions corresponding to the 50 images.

The alignment method presented in this paper achieved a reasonably high precision (0.880) despite its simplicity. Because we use only aligned descriptions as the source for relation extraction in Section 5.2, the precision is more important than the recall.

In contrast, the method using the WordNet hierarchy, which has been commonly used in previous work (Elliott and de Vries, 2015; Aditya et al., 2015), underperformed the presented alignment method. The recall of the WordNet method was relatively low because WordNet is prone to suffer from textual variations. For example, WordNet includes *skateboarder* as a descendant of the synset *person*, but does not include *skate border* nor *border*. The precision of the WordNet method was also lower than the IBM Model because some object categories

[7] We used Stanford CoreNLP 3.5.2. http://stanfordnlp.github.io/CoreNLP/

[8] This finds noun phrases with four words at most.

Table 2: The 10 most frequent relations extracted from the MS COCO dataset.

Relation	# of instances	Relation	# of instances
on	19,666 (12.58%)	of	4,096 (2.62%)
in	14,300 (9.15%)	next_to	3,974 (2.54%)
with	13,047 (8.35%)	ride	3,711 (2.37%)
hold	5,136 (3.29%)	sit_on	3,265 (2.09%)
at	4,345 (2.78%)	on_top_of	2,393 (1.53%)

(e.g., PERSON, FOOD, VEHICLE) are general concepts in WordNet and are mapped to general words (e.g., *building* and *group*) inappropriately.

5.4 Collected relation instances

Table 2 lists the 10 most frequent relations extracted from the MS COCO dataset. We can see from the table that our approach extracts not only spacial relations consisting of prepositions (e.g., *on* and *next_to*) but also predicative relations representing actions (e.g., *hold*, *ride* and *sit_on*).

Figure 5 visualizes some interesting examples of relations. A relation instance $r(o_1, o_2)$ consists of a relation expression r and objects o_1 and o_2 in the image. Each object o has the bounding box $(o.x, o.y, o.w, o.h)$. Therefore, we can compute the means and standard deviations of objects o_1 and o_2 that appear as the arguments of the relation r. In this way, we can visualize a rough interpretation of spatial relationships between objects referred to by the relation r.

In Figure 5, we normalize the image coordinates of all bounding boxes to the range of $[0, 1]$, and transform the position of o_2 to a relative coordinate with respect to o_1. The center of the ellipse in each visualization indicates the mean of the center of the objects. A bright ellipse represents the mean size of bounding boxes for the object, and a dark ellipse indicates the standard deviation of the center coordinates. For example, the visualization of *above*(o_1, o_2) reflects the meaning of *above* that the y-coordinate of o_1 is greater than that of o_2.

The previous work (Elliott and de Vries, 2015; Kong et al., 2014; Lin et al., 2015) pre-defined rules to represent relations. For example, the *above* relation holds if an object o_1 has a greater y-coordinate than that of o_2 and if no overlap exists between the two objects. We would like to stress here that we could acquire similar rules automatically from the statistics of a large-scale dataset with image annotations. In addition, it is non-trivial to define rules manually for action verbs such as *wear* or *eat*. These spatial relationships will be encoded as features for the relation recognizer described in the next section.

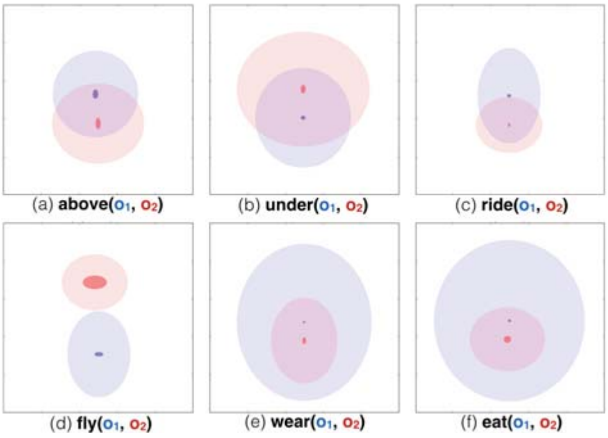


Figure 5: Visualization of positional relations between objects.

6 Recognizing inter-object relations

6.1 Relation recognizer

Using the relation instances acquired in the previous section, we train a classifier to recognize relations between two objects in an unseen image. Let R denote a set of relations extracted in Section 5.2. We model $P(r|o_1, o_2)$, the probability that two objects o_1 and o_2 have the relation $r \in R$ in the image. Note that multiple relations may hold true at the same time (e.g., *ride* and *ride_on*). Thus, we formalize the recognition task as a multi-label classification problem. We design a three-layer neural network[9] whose top layer uses the sigmoid activation function σ,

$$P(r|o_1, o_2) = \sigma(\boldsymbol{w}_r \cdot \boldsymbol{h}_{o_1,o_2} + b_r), \qquad (1)$$

$$\boldsymbol{h}_{o_1,o_2} = \mathrm{ReLU}(H\boldsymbol{x}_{o_1,o_2} + \boldsymbol{b}_h). \qquad (2)$$

Here, $\boldsymbol{x}_{o_1,o_2} \in \mathbb{R}^d$ is a feature vector for the two object o_1 and o_2. The matrix $H \in \mathbb{R}^{d \times h}$, vector $\boldsymbol{w}_r \in \mathbb{R}^d$, and bias terms $\boldsymbol{b}_h \in \mathbb{R}^h$, $b_r \in \mathbb{R}$ are the model parameters. $\mathrm{ReLU}(.)$ represents the leaky rectified linear unit function (Xu et al., 2015), $\mathrm{ReLU}(x) = \max(x, ax)$. We use the default slope

[9] We used Chainer (Tokui et al., 2015) to implement this network. `http://chainer.org/`

coefficient value $a = 0.2$. When predicting relations, we identify all relations $r \in R$ satisfying $P(r|o_1, o_2) \geq 0.5$. We found empirically that a hidden layer helps mitigate the difficulty of learning specific relations.

We compute the input vector $\boldsymbol{x}_{o_1,o_2}$ for the objects o_1 and o_2 by using the following features.

Category (160 dims) We encode a one-hot vector representing 80 categories of an object. We concatenate two one-hot vectors corresponding to the two objects ($2 \times 80 = 160$ dimensions in total).

Position (8 dims) Scaling the coordinates of every image in the range of $[0, 1]$, we encode the position of the center of the bounding box for o_1 and the relative position of o_2 with respect to o_1. In addition, we encode the sizes of the two objects.

Area (5 dims) We encode the following values as features: the areas of o_1 and o_2; the ratio of the area of o_1 to that of o_2; the area of the union of o_1 and o_2; and the ratio of the area of the intersection of o_1 and o_2 to that of the union of o_1 and o_2.

Scene (205 dims) We take the `fc8` layer of Place CNN (Zhou et al., 2014) to incorporate the scene of the image. We expect that this feature can capture a scene-specific relation, e.g., the *look_at* relation between PERSON and GIRAFFE objects when the scene of the image is in a zoo.

Action/Direction (20 dims) It may be difficult to identify a relation for a PERSON object with only the above features because there are high ambiguities among relations between a PERSON and the other object. For this reason, we made an attempt to manually annotate action states (*standing*, *walking*, *running*, *sitting/lying*, *unknown*) and directions (*left*, *right*, *frontal*, *back*, *unknown*) of a person. We asked nine experts to annotate, and assigned one annotator per image. The experts discussed the criteria of the annotation every time the need arose. We encode a one-hot vector representing the truths of 10 attributes per person (i.e., 20 features for two people). Because these features currently require manual annotations, we will explore the usefulness of these features in the experiment.

Concatenating the above features, we form a 398-dimensional vector $\boldsymbol{x}_{o_1,o_2}$ as the input for the neural network. A training instance consists of a tuple $(\boldsymbol{x}_{o_1,o_2}, \boldsymbol{y})$, where $\boldsymbol{y}$ represents a n-hot vector for the gold relations $\{r_1, r_2, \cdots r_n\}$ for the objects. In order to remove infrequent relations, we employ only the top 80% of the frequent relations in the extracted relation instances. In this way, we obtained 43,290 relation instances with 133 distinct relations ($|R| = 133$) for the experiments.

Table 3: Performance of object relation prediction.

	Precision	Recall	F1
Category only	.385	.149	.205
All features	.304	**.255**	**.250**
w/o Category	.241	.217	.199
w/o Scene	**.393**	.195	.241
w/o Action/Direction	.336	.218	.239
w/o Area	.302	.250	.243
w/o Position	.296	.246	.245

We initialized the model parameters randomly according to $\mathcal{N}(0, \sqrt{1/d})$ or $\mathcal{N}(0, \sqrt{1/h})$ (depending on the layer of the parameters). We determined the dimension of the hidden layer $h \in \{100, 200, 300, 400, 500\}$ such that it yielded the best performance on 10-fold cross validation. We used the cross-entropy loss function and RMSProp to train the model parameters.

6.2 Results

To the best of our knowledge, this is the first research to evaluate open-vocabulary relation recognition between image objects. Therefore, we built a test set by sampling 1,000 images randomly from those in the MS COCO dataset that were left unused in the training data. We annotate gold relations manually to the object pairs mentioned in the descriptions. The test set consists of 454 instances.

Table 3 shows the performance of relation recognition and the results of ablation tests that remove one of the five features types. The classifier that was trained with all features in Section 6 (*All features*) achieved a 0.250 F1 score whereas the one trained with only the object category feature (*Category only*) achieved a 0.205 F1 score. We can consider *Category only* as a language model since it uses only text information (i.e., object categories). Although the ablation test for *Category* also reveals the importance of the category information (a 0.051 reduction of F1 score), showing the largest contribution among five features, the difference of F1

Table 4: Top-1 result of object relation prediction.

	Precision	Recall	F1
Majority baseline	.282	.281	.281
Category only	.467	.495	.480
All features	.452	.474	.463
w/o Category	.342	.352	.347
w/o Scene	**.479**	**.520**	**.499**
w/o Action/Direction	.449	.469	.459
w/o Area	.441	.465	.453
w/o Position	.420	.436	.428

scores (0.045) between *All features* and *Category only* indicates the importance of spatial and visual features for recognizing relations between objects. We speculate the reason of the largest contribution of the *Category* feature is that, compared to the spatial or visual features, it can reduce relation candidates given two objects. We might be able to specify the relations if *Category* information (e.g., PERSON and SKATEBOARD) rather than the spacial or visual information (e.g., ”o_1 is upper of o_2”) is given when looking at a picture. The ablation test for *Action/Direction* shows that understanding the state of a person is also useful for recognizing relations.

We also evaluate the performance of relation recognition in terms of top-1 predictions, as it is important practically for the application of description generation to predict at least one true-positive relation. Defining a top-1 prediction as the relation r to which the classifier yielded the highest probability $P(r|o_1, o_2)$ of all relations, we regard a prediction as correct if the predicted relation r is included in the set of gold relations.

Table 4 reports the performance of the top-1 evaluation. We added a *Majority baseline* that always predicts the most frequent relation *on* in the training set. The full-feature model achieved a 0.463 F1 score. In contrast to our expectation, the best result of a 0.499 F1 score was obtained without the scene features. This is probably because the relation recognizer overfitted to the training data with the abstract features of the `fc8` layer of Place CNN, which may have the potential to discriminate against individual images. Removing the scene features, the relation recognizer could outperform the *Category only* baseline.

However, we also encounter pairs of image objects for which the proposed method cannot predict relations in principle. Figure 6 shows a typical example of these instances. In this image, the dog is located around the center of the couch, but the spatial relationship is insufficient to describe the scene. Furthermore, we need to recognize the state of the dog in order to differentiate *sleep_on*, *lie_on*, and *sit_on*. It may be necessary to recognize the fine-grained properties about objects, e.g., whether or not the animal has its eyes closed.

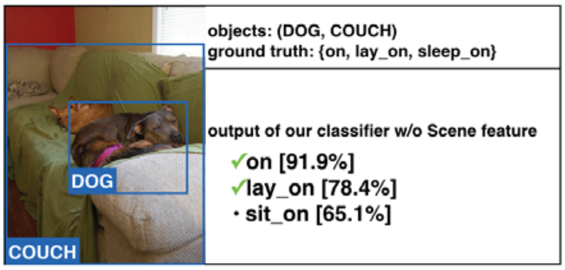

objects: (DOG, COUCH)
ground truth: {on, lay_on, sleep_on}
output of our classifier w/o Scene feature
✓on [91.9%]
✓lay_on [78.4%]
• sit_on [65.1%]
DOG
COUCH

Figure 6: An example of our inter-object relation classifier. A green tick indicates that the relation is true (included in the gold labels).

7 Conclusion

In this paper, we presented the first approach for open-vocabulary relation recognition between objects in images. In order to extract expressions that refer specifically to relations between objects, we successfully adopted a word alignment model developed for statistical machine translation. Using the relation instances whose arguments are grounded to image objects, we could train a relation recognizer that predicts a relation for a given pair of objects in an image. The experimental results demonstrated that the spatial features contributed to the task of relation recognition.

An immediate future work would be further analysis to explore important features/attributes for relation recognition, e.g., features/attributes expressing an object, two objects, or the whole scene of an image. We plan to demonstrate the usefulness of relation recognition between image objects in applications including image description generation, image retrieval, and even image recognition with the commonsense knowledge extracted from the image descriptions.

Acknowledgments

We gratefully acknowledge anonymous annotators for annotating a large amount of gold labels within a limited time period. We thank anonymous reviewers for their thoughtful comments and suggestions. This research was supported by JSPS KAKENHI Grant Numbers JP15H01702, JP15H05318 and JP15H05919.

References

Somak Aditya, Yezhou Yang, Chitta Baral, Cornelia Fermüller, and Yiannis Aloimonos. 2015. From images to sentences through scene description graphs using commonsense reasoning and knowledge. *CoRR*, abs/1511.03292.

Peter F Brown, Vincent J Della Pietra, Stephen A Della Pietra, and Robert L Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational linguistics*, 19(2):263–311.

Yu-Wei Chao, Zhan Wang, Yugeng He, Jiaxuan Wang, and Jia Deng. 2015. HICO: A benchmark for recognizing human-object interactions in images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1017–1025.

Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. 2015. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2625–2634.

Desmond Elliott and Arjen de Vries. 2015. Describing images using inferred visual dependency representations. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 42–52.

Desmond Elliott and Frank Keller. 2013. Image description using visual dependency representations. In *Proceedings of the 2014 Conference on Empirical Methods In Natural Language Processing (EMNLP)*, pages 1292–1302.

Mark Everingham, S. M. Ali Eslami, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. 2014. The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision*, 111(1):98–136.

Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. 2010. Every picture tells a story: Generating sentences from images. In *Computer Vision–ECCV 2010*, pages 15–29.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep residual learning for image recognition. *CoRR*, abs/1512.03385.

Justin Johnson, Andrej Karpathy, and Fei-Fei Li. 2015. Densecap: Fully convolutional localization networks for dense captioning. *CoRR*, abs/1511.07571.

Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3128–3137.

Ryan Kiros, Ruslan Salakhutdinov, and Richard S. Zemel. 2014. Unifying visual-semantic embeddings with multimodal neural language models. *CoRR*, abs/1411.2539.

Chen Kong, Dahua Lin, Mayank Bansal, Raquel Urtasun, and Sanja Fidler. 2014. What are you talking about? text-to-image coreference. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3558–3565.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft COCO: Common objects in context. In *Computer Vision–ECCV 2014*, pages 740–755.

Dahua Lin, Sanja Fidler, Chen Kong, and Raquel Urtasun. 2015. Generating multi-sentence natural language descriptions of indoor scenes. In *Proceedings of the British Machine Vision Conference (BMVC)*, pages 93.1–93.13.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60.

Andrea Moro and Roberto Navigli. 2013. Integrating syntactic and semantic analysis into the open information extraction paradigm. In *Proceedings of the 23th International Joint Conference on Artificial Intelligence*, IJCAI'13, pages 2148–2154.

Ndapandula Nakashole, Gerhard Weikum, and Fabian Suchanek. 2012. Patty: a taxonomy of relational patterns with semantic types. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1135–1145.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Fereshteh Sadeghi, Santosh K Divvala, and Ali Farhadi. 2015. Viske: Visual knowledge extraction and question answering by visual verification of relation phrases. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1456–1464.

Michael Schmitz, Robert Bart, Stephen Soderland, Oren Etzioni, et al. 2012. Open language learning for information extraction. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 523–534.

Seiya Tokui, Kenta Oono, Shohei Hido, and Justin Clayton. 2015. Chainer: a next-generation open source framework for deep learning. In *Proceedings of Workshop on Machine Learning Systems (LearningSys) in The Twenty-ninth Annual Conference on Neural Information Processing Systems (NIPS)*.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3156–3164.

Ying Xu, Mi-Young Kim, Kevin Quinn, Randy Goebel, and Denilson Barbosa. 2013. Open information extraction with tree kernels. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 868–877.

Bing Xu, Naiyan Wang, Tianqi Chen, and Mu Li. 2015. Empirical evaluation of rectified activations in convolutional network. *CoRR*, abs/1505.00853.

Bangpeng Yao, Xiaoye Jiang, Aditya Khosla, Andy Lai Lin, Leonidas Guibas, and Li Fei-Fei. 2011. Human action recognition by learning bases of action attributes and parts. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1331–1338.

Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. 2014. Learning deep features for scene recognition using places database. In *Advances in neural information processing systems*, pages 487–495.