Detecting Unseen Visual Relations Using Analogies

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

Abstract

We seek to detect visual relations in images of the form of triplets $t = (subject,\ predicate,\ object)$, such as "person riding dog", where training examples of the individual entities are available but their combinations are unseen at training. This is an important set-up due to the combinatorial nature of visual relations : collecting sufficient training data for all possible triplets would be very hard. The contributions of this work are three-fold. First, we learn a representation of visual relations that combines (i) individual embeddings for subject, object and predicate together with (ii) a visual phrase embedding that represents the relation triplet. Second, we learn how to transfer visual phrase embeddings from existing training triplets to unseen test triplets using analogies between relations that involve similar objects. Third, we demonstrate the benefits of our approach on three challenging datasets : on HICO-DET, our model achieves significant improvement over a strong baseline for both frequent and unseen triplets, and we observe similar improvement for the retrieval of unseen triplets with out-of-vocabulary predicates on the COCO-a dataset as well as the challenging unusual triplets in the UnRel dataset.

1. Introduction

Understanding interactions between objects is one of the fundamental problems in visual recognition. To retrieve images given a complex language query such as "a woman sitting on top of a pile of books" we need to recognize individual entities "woman" and "a pile of books" in the scene, as well as understand what it means to "sit on top of something". In this work we aim to recognize and localize unseen interactions in images, as shown in Figure 1, where the individual entities ("person", "dog", "ride") are available at training, but not in this specific combination. Such ability is important in practice given the combinatorial nature of visual relations where we are unlikely to obtain sufficient training data for all possible relation triplets.

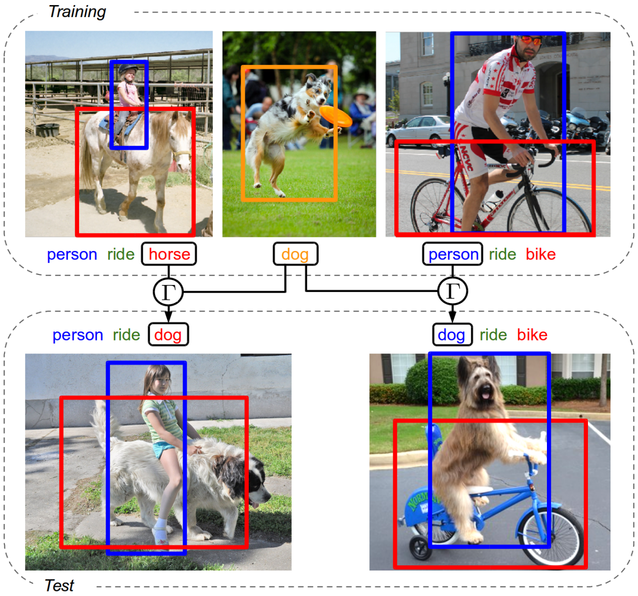


Figure 1: Illustration of transfer by analogy with our model described in 3.2. We transfer visual representations of relations seen in the training set such as "person ride horse" to represent new unseen relations in the test set such as "person ride dog".

Existing methods [7, 24, 27] to detect visual relations in the form of triplets $t = (subject, predicate, object)$ typically learn generic detectors for each of the entities, i.e. a separate detector is learnt for subject (e.g. "person"), object (e.g. "horse") and predicate (e.g. "ride"). The outputs of the individual detectors are then aggregated at test time. This *compositional approach* can detect unseen triplets, where subject, predicate and object are observed separately but not in the specific combination. However, it often fails in practice [30, 46], due to the large variability in appearance of the visual interaction that often heavily depends on the objects involved; it is indeed difficult for a single "ride" detector to capture visually different relations such as "person ride horse" and "person ride bus".

An alternative approach [40] is to treat the whole triplet as a single entity, called a visual phrase, and learn a separate detector for each of the visual phrases. For instance, separate detectors would be learnt for relations "person ride horse" and "person ride surfboard". While this approach better

[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.

handles the large variability of visual relations, it requires training data for each triplet, which is hard to obtain as visual relations are combinatorial in their nature and many relations are unseen in the real world.

In this work we address these two key limitations. First, what is the right representation of visual relations to handle the large variability in their appearance, which depends on the entities involved? Second, how can we handle the scarcity of training data for unseen visual relation triplets? To address the first challenge, we develop a hybrid model that combines compositional and visual phrase representations. More precisely, we learn a compositional representation for subject, object and predicate by learning separate visual-language embedding spaces where each of these entities is mapped close to the language embedding of its associated annotation. In addition, we also learn a relation triplet embedding space where visual phrase representations are mapped close to the language embedding of their corresponding triplet annotations. At test time, we aggregate outputs of both compositional and visual phrase models. To address the second challenge, we learn how to transfer visual phrase embeddings from existing training triplets to unseen test triplets using analogies between relations that involve similar objects. For instance, as shown in Figure 1, we recognize the unseen triplet "person ride dog" by using the visual phrase embedding for triplet "person ride horse" after a transformation that depends on the object embedding for "dog" and "horse". Because we transfer training data only from triplets that are visually similar, we expect transferred visual phrase detectors to better represent the target relations compared to a generic detector for a relation "ride" that may involve also examples of "person ride train" and "person ride surfboard".

Contributions. Our contributions are three fold. First, we take advantage of both the compositional and visual phrase representations by learning complementary visual-language embeddings for subject, object, predicate and the visual phrase. Second, we develop a model for transfer by analogy to obtain visual-phrase embeddings of never seen before relations. Third, we perform experimental evaluation on three challenging datasets where we demonstrate the benefits of our approach on both frequent and unseen relations.

2. Related work

Visual relation detection. Learning visual relations belongs to a general class of problems on relational reasoning [3, 4, 15, 22, 41] that aim to understand how entities interact. In the more specific set-up of visual relation detection, the approaches can be divided into two main groups: (i) compositional models, which learn detectors for subject, object and predicates separately and aggregate their outputs; (ii) and visual phrase models, which learn a separate detector for each visual relation. Visual phrase models such as [40] have demonstrated better robustness to the visual diversity of relations than compositional models. However, with the introduction of datasets with a larger vocabulary of objects and predicates [6, 23], visual phrase approaches have been facing severe difficulties as most relations have very few training examples. Compositional methods [9, 11, 17, 27, 30, 33, 42], which allow sharing knowledge across triplets, have scaled better but do not cope well with unseen relations. To increase the expressiveness of the generic compositional detectors, recent works have developed models of statistical dependencies between the subject, object and predicate, using, for example, graphical models [7, 24], language distillation [45], or semantic context [48]. Others [1, 8, 31, 38] have proposed to combine unigram detectors with higher-order composites such as bigrams (subject-predicate, predicate-object). In contrast to the above methods that model a discrete vocabulary of labels, we learn visual-semantic (language) embeddings able to scale to out-of-vocabulary relations and to benefit from powerful pre-learnt language models.

Visual-semantic embeddings. Visual-semantic embeddings have been successfully used for image captioning and retrieval [18, 19]. With the introduction of datasets annotated at the region level [23, 32], similar models have been applied to align image regions to fragments of sentences [14, 44]. In contrast, learning embeddings for visual relations still remains largely an open research problem with recent work exploring, for example, relation representations using deformations between subject and object embeddings [46]. Our work is, in particular, related to models [47] learning separate visual-semantic spaces for subject, object and predicate. However, in contrast to [47], we additionally learn a visual phrase embedding space to better deal with appearance variation of visual relations, and develop a model for analogy reasoning to infer embeddings of unseen triplets.

Unseen relations and transfer learning. Learning visual phrase embeddings suffers from the problem of lack of training data for unseen relations. This has been addressed by learning factorized object and predicate representations [13] or by composing classifiers for relations from simpler concepts [20, 29]. In contrast, our approach transfers visual relation representations from seen examples to unseen ones in a similar spirit to how previous work dealt with inferring classifiers for rare objects [2]. The idea of sharing knowledge from seen to unseen triplets to compensate for the scarcity of training data has been also addressed in [34] by imposing constraints on embeddings of actions. Different from this work, we formulate the transfer as an analogy between relation triplets. To achieve that, we build on the computational model of analogies developed in [35] but extend it to representations of visual relations. This is related to [39] who also learn visual analogies as vector operations in an embedding space, but only consider visual inputs while we learn analogy models for joint image-language embeddings.

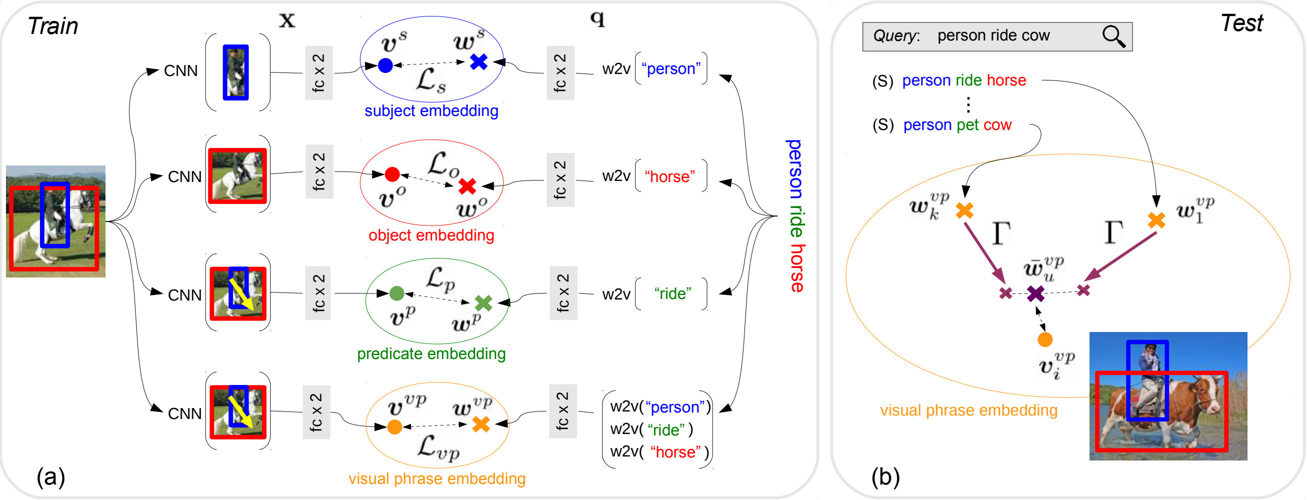


Figure 2: **Model overview.** Our model consists of two parts : (a) learning embedding spaces for subject, object, predicate and visual phrase by optimizing the joint loss $\mathcal{L}_{joint} = \mathcal{L}_s + \mathcal{L}_o + \mathcal{L}_p + \mathcal{L}_{vp}$ combining the input visual $\mathbf{x}$ and language $\mathbf{q}$ representations; (b) at test time, we are given a new unseen triplet ("person ride cow"). We find similar but seen triplets ("person ride horse" and "person pet cow"), transform their embeddings $\boldsymbol{w}_k^{vp}$ with analogy transformation Γ to compute an estimate of the embedding $\bar{\boldsymbol{w}}_u^{vp}$ of the triplet "person ride cow" and use this estimated embedding to retrieve relevant images by nearest neighbour search on the (embedded) visual descriptors v_i^{vp}.

3. Model

In this section we describe our model for recognizing and localizing visual relations in images. As illustrated in Figure 2, our model consists of two parts. First, we learn different visual-language embedding spaces for the subject (s), the object (o), the predicate (p) and the visual phrase (vp), as shown in Figure 2(a). We explain how to train these embeddings in Section 3.1. Second, we transfer visual phrase embeddings of seen triplets to unseen ones with analogy transformations, as shown in Figure 2(b). In Section 3.2 we explain how to train the analogy transformations and form visual phrase embeddings of new unseen triplets at test time.

Notation for relation triplets. The training dataset consists of N candidate pairs of bounding boxes, each formed by a subject candidate bounding box proposal and object candidate bounding box proposal. Let $\mathcal{V}_s$, $\mathcal{V}_o$ and $\mathcal{V}_p$ be the vocabulary of subjects, objects and predicates, respectively. We call $\mathcal{V}_{vp} = \mathcal{V}_s \times \mathcal{V}_p \times \mathcal{V}_o$ the vocabulary of triplets. A triplet t is of the form $t = (s, p, o)$, e.g. $t = (person, ride, horse)$. Each pair of candidate subject and object bounding boxes, $i \in \{1, ..., N\}$, is labeled by a vector $(y_t^i)_{t \in \mathcal{V}_{vp}}$ where $y_t^i = 1$ if the i^{th} pair of boxes could be described by relation triplet t, otherwise $y_t^i = 0$. The labels for subject, object and predicate naturally derive from the triplet label.

3.1. Learning representations of visual relations

We represent visual relations in joint visual-semantic embedding spaces at different levels of granularity : (i) at the unigram level, where we use separate subject, object and predicate embeddings, and (ii) at the trigram level using an a visual phrase embedding of the whole triplet. Combining the different types of embeddings results in a more powerful representation of visual relations as will be shown in section 4. In detail, as shown in Figure 2(a), the input to visual embedding functions (left) is a candidate pair of objects i encoded by its visual representation $\mathbf{x}_i \in \mathbb{R}^{d_v}$. This representation is built from (i) pre-computed appearance features obtained from a CNN trained for object detection and (ii) a representation of the relative spatial configuration of the object candidates. The language embeddings (right in Figure 2(a)) take as input a triplet t encoded by its language representation $\mathbf{q}_t \in \mathbb{R}^{d_q}$ obtained from pre-trained word embeddings. We provide more details about these representations in 4.2. Next we give details of the embedding functions.

Embedding functions. Our network projects the visual features $\mathbf{x}_i$ and language features $\mathbf{q}_t$ into separate spaces for the subject (s), the object (o), the predicate (p) and the visual phrase (vp). For each input type $b \in \{s, o, p, vp\}$, we embed the visual features and language features into a common space of dimensionality d using projection functions

$$\boldsymbol{v}_i^b = f_v^b(\mathbf{x}_i), \tag{1}$$

$$\boldsymbol{w}_t^b = f_w^b(\mathbf{q}_t), \tag{2}$$

where $\boldsymbol{v}_i^b$ and $\boldsymbol{w}_t^b$ are the output visual and language representations, and the projection functions $f_v^b : \mathbb{R}^{d_v} \rightarrow \mathbb{R}^d$ and $f_w^b : \mathbb{R}^{d_q} \rightarrow \mathbb{R}^d$ are 2-layer perceptrons, with ReLU non linearities and Dropout, inspired by [44]. Additionally, we L2 normalize the output language features while the output visual features are not normalized, which we found to work well in practice.

Training loss. We train parameters of the embedding functions (f_v^b, f_w^b) for each type of input b (i.e subject, object, predicate and visual phrase) by maximizing log-likelihood

$$\mathcal{L}_b = \sum_{i=1}^{N} \sum_{t \in \mathcal{V}_b} \mathbb{1}_{y_t^i=1} \log \left(\frac{1}{1+e^{-\boldsymbol{w}_t^{b\,T} \boldsymbol{v}_i^b}} \right) + \sum_{i=1}^{N} \sum_{t \in \mathcal{V}_b} \mathbb{1}_{y_t^i=0} \log \left(\frac{1}{1+e^{\boldsymbol{w}_t^{b\,T} \boldsymbol{v}_i^b}} \right), \quad (3)$$

where the first attraction term pushes closer visual representation $\boldsymbol{v}_i^b$ to its correct language representation $\boldsymbol{w}_t^b$ and the second repulsive term pushes apart visual-language pairs that do not match. As illustrated in Figure 2, we have one such loss for each input type and optimize the joint loss that sums the individual loss functions $\mathcal{L}_{joint} = \mathcal{L}_s + \mathcal{L}_o + \mathcal{L}_p + \mathcal{L}_{vp}$. A similar loss function has been used in [28] to learn word representations, while visual-semantic embedding models [19, 44] typically use triplet ranking losses. Both loss functions work well, but we found embeddings trained with log-loss (3) easier to combine across different input types as their outputs are better calibrated.

Inference. At test time, we have a language query in the form of triplet t that we embed as $(\boldsymbol{w}_t^b)_b$ using Eq. (2). Similarly, pairs i of candidate object boxes in the test images are embedded as $(\boldsymbol{v}_i^b)_b$ with Eq. (1). Then we compute a similarity score $S_{t,i}$ between the triplet query t and the candidate object pair i by aggregating predictions over the different embedding types $b \in \{s, p, o, vp\}$ as

$$S_{t,i} = \prod_{b \in \{s,p,o,vp\}} \frac{1}{1+e^{-\boldsymbol{w}_t^{b\,T} \boldsymbol{v}_i^b}}. \quad (4)$$

Interpretation of embedding spaces. The choice of learning different embedding spaces for subject, object, predicate and visual phrase is motivated by the observation that each type of embedding captures different information about the observed visual entity. In Figure 3 we illustrate the advantage of learning separate predicate (p) and visual-phrase (vp) embedding spaces. In the p space, visual entities corresponding to "person ride horse" and "person ride car" are mapped to the same point, as they share the same predicate "ride". In contrast, in the vp space, the same visual entities are mapped to two distinct points. This property of the vp space is desirable to handle both language polysemy (i.e., "ride" has different visual appearance depending on the objects involved and thus should not be mapped into a single point) and synonyms (i.e., "person jump horse" and "person ride horse" projections should be close even if they do not share the same predicate).

3.2. Transferring embeddings to unseen triplets by analogy transformations

We propose to explicitly transfer knowledge from seen triplets at training to new unseen triplets at test time by analogy reasoning. The underlying intuition is that if we have seen examples of "person ride horse", it might be possible to use this knowledge to recognize the relation "person ride cow", as "horse" and "cow" have similar visual appearance. As illustrated in Figure 2(b), this is implemented as an *analogy transformation* in the visual phrase embedding space, where a representation of the source triplet (e.g. "person ride horse") is transformed to form a representation of target triplet (e.g. "person ride cow"). There are two main steps in this process. First, we need to learn how to perform the analogy transformation of one visual phrase embedding (e.g. "person ride horse") to another (e.g. "person ride cow"). Second, we need to identify which visual phrases are suitable for such transfer by analogy. For example, to form a representation of a new relation "person ride cow" we want to transform the representation of "person ride horse" but not "person ride bus". We describe the two steps next.

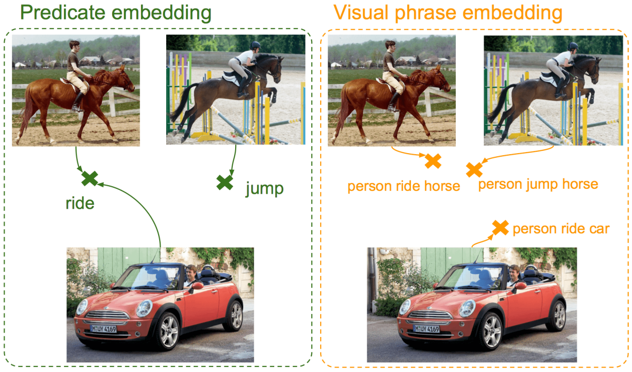


Figure 3: Illustration of the differences between predicate (p) (left) and visual phrase (vp) (right) embeddings. In the p space, visually different relations such as "person ride horse" and "person ride car" map to the same location defined by the predicate "ride". In contrast, they are mapped to distinct locations in the visual phrase space that considers the entire relation triplet.

Transfer by analogy. To transform the visual phrase embedding $\boldsymbol{w}_t^{vp}$ of a source triplet $t = (s, p, o)$ to the visual phrase embedding $\boldsymbol{w}_{t'}^{vp}$ of a target triplet $t' = (s', p', o')$ we learn a transformation Γ such that

$$\boldsymbol{w}_{t'}^{vp} = \boldsymbol{w}_t^{vp} + \Gamma(t, t'). \quad (5)$$

Here, Γ could be interpreted as a correction term that indicates how to transform $\boldsymbol{w}_t^{vp}$ to $\boldsymbol{w}_{t'}^{vp}$ in the joint visual-semantic space vp to compute a target relation triplet t' that is analogous to source triplet t. This relates to neural word representations such as [28] where word embeddings of similar concepts can be linked by arithmetic operations such as "$king$" $-$ "man" $+$ "$woman$" $=$ "$queen$". Here, we would like to perform operations such as "$person\ ride\ horse$" $-$ "$horse$" $+$ "cow" $=$ "$person\ ride\ cow$".

Form of Γ. To relate the visual phrase embeddings of t and t' through Γ we take advantage of the decomposition of the triplet into subject, predicate and object. In detail, we use the visual phrase embeddings of individual subject, predicate and object to learn how to relate the visual phrase embeddings of triplets. Using this structure, we redefine the analogy transformation given by Eq. (5) as

$$\boldsymbol{w}_{t'}^{vp} = \boldsymbol{w}_t^{vp} + \Gamma \begin{bmatrix} \boldsymbol{w}_{s'}^{vp} - \boldsymbol{w}_s^{vp} \\ \boldsymbol{w}_{p'}^{vp} - \boldsymbol{w}_p^{vp} \\ \boldsymbol{w}_{o'}^{vp} - \boldsymbol{w}_o^{vp} \end{bmatrix}, \tag{6}$$

where $t = (s, p, o)$ and $t' = (s', p', o')$ denote the source and target triplet, and $\boldsymbol{w}_s^{vp}$, $\boldsymbol{w}_p^{vp}$, $\boldsymbol{w}_o^{vp}$ are visual phrase embeddings of subject, predicate and object, respectively, constructed using Eq. (2) as $\boldsymbol{w}_s^{vp} = f_w^{vp}(\mathbf{q}_{[s,0,0]})$, $\boldsymbol{w}_p^{vp} = f_w^{vp}(\mathbf{q}_{[0,p,0]})$, $\boldsymbol{w}_o^{vp} = f_w^{vp}(\mathbf{q}_{[0,0,o]})$. Here $[s, 0, 0]$ denotes the concatenation of word2vec embeddings of subject s with two vectors of zeros of size d. For example, the analogy transformation of $t = (person, ride, horse)$ to $t' = (person, ride, camel)$ using Eq. (6) would result in

$$\boldsymbol{w}_{t'}^{vp} = \boldsymbol{w}_t^{vp} + \Gamma \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \\ \boldsymbol{w}_{camel}^{vp} - \boldsymbol{w}_{horse}^{vp} \end{bmatrix}. \tag{7}$$

Intuitively, we would like Γ to encode how the change of objects, observable through the embeddings of source and target objects, $\boldsymbol{w}_o^{vp}$, $\boldsymbol{w}_{o'}^{vp}$, influences the source and target triplet embeddings $\boldsymbol{w}_t^{vp}$, $\boldsymbol{w}_{t'}^{vp}$. Please note that here we have shown an example of a transformation resulting from a change of object, but our formulation, given by Eq. (6), allows for changes of subject or predicate in a similar manner. While different choices for Γ are certainly possible, we opt for

$$\Gamma(t, t') = MLP \begin{bmatrix} \boldsymbol{w}_{s'}^{vp} - \boldsymbol{w}_s^{vp} \\ \boldsymbol{w}_{p'}^{vp} - \boldsymbol{w}_p^{vp} \\ \boldsymbol{w}_{o'}^{vp} - \boldsymbol{w}_o^{vp} \end{bmatrix}, \tag{8}$$

where MLP is a 2-layer perceptron without bias. We also compare different forms of Γ in Section 4.

Which triplets to transfer from? We wish to apply the transformation by analogy Γ only between triplets that are similar. The intuition is that to obtain representation of an unseen target triplet $t' = (person, ride, camel)$, we wish to use only similar triplets such as $t = (person, ride, horse)$ but not triplets such as $t = (person, ride, skateboard)$. For this, we propose to decompose the similarity between triplets t and t' by looking at the similarities between their subjects, predicates and objects measured by the dot-product of their representations in the corresponding individual embedding spaces. The motivation is that the subject, object and predicate spaces do not suffer as much from the limited training data compared to the visual phrase space. In detail, we define a weighting function G as :

$$G(t, t') = \sum_{b \in \{s,p,o\}} \alpha_b {\boldsymbol{w}_t^b}^T \boldsymbol{w}_{t'}^b, \tag{9}$$

where ${\boldsymbol{w}_t^b}^T \boldsymbol{w}_{t'}^b$ measures similarity between embedded representations $\boldsymbol{w}^b$ and scalars α_b are hyperparameters that reweight the relative contribution of subject, object and predicate similarities. As we constrain $\sum_b \alpha_b = 1$ the output of $G(t, t') \in [0, 1]$. For a target triplet t', we define as $\mathcal{N}_{t'}$ the set of k most similar source triplets according to G.

Learning Γ. We fit parameters of Γ by learning analogy transformations between triplets in the training data. In particular, we generate training data pairs of source t and target t' triplets. Given the generated data, we optimize log-likelihood similar to Eq. (3) but using visual features of the real target triplet and language features of the source triplet transformed with the analogy transformation Γ. The optimization is performed w.r.t. to both the parameters of Γ and parameters of the embedding functions. Details are given in the Section A.1 of the Appendix.

Aggregating embeddings. At test time, we compute the visual phrase embedding of an unseen triplet u by aggregating embeddings of similar seen triplets $t \in \mathcal{N}_u$ transformed using the analogy transformation:

$$\bar{\boldsymbol{w}}_u^{vp} = \sum_{t \in \mathcal{N}_u} G(t, u)\ (\boldsymbol{w}_t^{vp} + \Gamma(t, u)), \tag{10}$$

where $\boldsymbol{w}_t^{vp}$ is the visual phrase embedding of source triplet t obtained with Eq. (2), $\Gamma(t, u)$ is the analogy transformation between source triplet t and unseen triplet u computed by Eq. (8) and $G(t, u)$ is a scalar weight given by Eq. (9) that re-weights the contribution of the different source triplets. This process is illustrated in Figure 2(b).

4. Experiments

In this section we evaluate the performance of our model for visual relation retrieval on three challenging datasets : HICO-DET [5], UnRel [30] and COCO-a [37]. Specifically, we numerically assess the two components of our model : (i) learning the visual phrase embedding together with the unigram embeddings and (ii) transferring embeddings to unseen triplets by analogy transformations.

4.1. Datasets and evaluation set-ups

HICO-DET. The HICO-DET [6, 5] dataset contains images of human-object interactions with box-level annotations. The interactions are varied : the vocabulary of objects matches the 80 COCO [26] categories and there are 117 different predicates. The number of all possible triplets is $1 \times 117 \times 80$ but the dataset contains positive examples for only 600 triplets. All triplets are seen at least once in training. The authors separate a set of 138 rare triplets, which are the triplets that appear fewer than 10 times at training. To conduct further analysis of our model, we also select a set of 25 triplets that we treat as unseen, exclude them completely

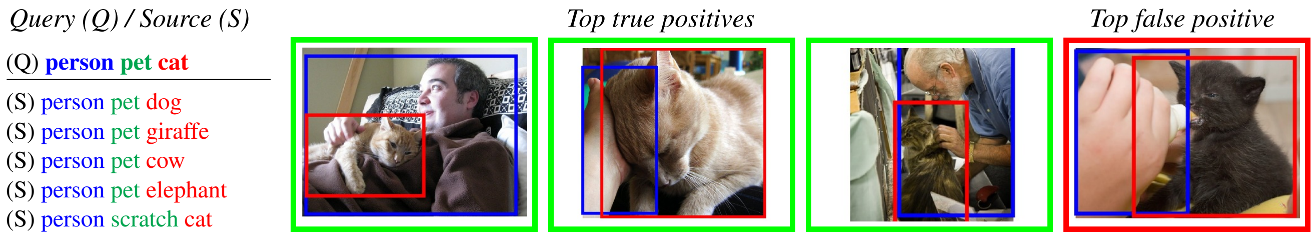


Figure 4: Top retrieved positive (green) and negative (red) detections with our model (s+o+vp+transfer) on unseen triplets excluded from HICO-DET. For a target triplet (Q) (e.g. "person pet cat"), our model automatically learns to select meaningful source triplets (S) involving visually similar objects or predicates ("person pet dog", "person scratch cat") and transforms their visual phrase embeddings by analogy transformation Γ. The top false positive corresponds to a visually related action ("feed"). Additional examples are in Section C of the Appendix.

from the training data in certain experiments, and try to retrieve them at test time using our model. These triplets are randomly selected among the set of non-rare triplets in order to have enough test instances on which to reliably evaluate.

UnRel. UnRel [30] is an evaluation dataset containing visual relations for 76 unusual triplet queries. In contrast to HICO-DET and COCO-a, the interactions do not necessarily involve a human, and the predicate is not necessarily an action (it can be a spatial relation, or comparative). The vocabulary of objects and predicates matches those of Visual Relation Detection Dataset [27]. UnRel is only an evaluation dataset, so similar to [30] we use the training set of Visual Relationship Dataset as training data.

COCO-a. The COCO-a dataset [37] is based on a subset of COCO dataset [26] augmented with annotations of human-object interactions. Similar to HICO-DET, the vocabulary of objects matches the 80 COCO categories. In addition, COCO-a defines 140 predicates resulting in a total of 1681 different triplets. The released version of COCO-a contains 4413 images with no pre-defined train/test splits. Given this relatively small number of images, we use COCO-a as an evaluation dataset for models trained on HICO-DET. This results in an extremely challenging set-up with 1474 unseen triplets among which 1048 involve an out-of-vocabulary predicate that has not been seen at training in HICO-DET.

Evaluation measure. On all datasets, we evaluate our model in a retrieval setup. For each triplet query in the vocabulary, we rank the candidate test pairs of object bounding boxes using our model and compute the performance in terms of Average Precision. Overall, we report mean Average Precision (mAP) over the set of triplet queries computed with the evaluation code released by [5] on HICO-DET and [30] on UnRel. On COCO-a, we use our own implementation as no evaluation code is released.

4.2. Implementation details

Candidate pairs. We use pre-extracted candidate pairs of objects from an object detector trained for the vocabulary of objects specific to the dataset. On HICO-DET, we train the object detector on the COCO training data using Detectron [10]. To be comparable to [11], we use a Faster-R-CNN [36] with ResNet-50 Feature Pyramid Network [25]. We post-process the candidate detections by removing candidates whose confidence scores are below 0.05 and apply an additional per-class score thresholding to maintain a fixed precision of 0.3 for each object category. At test time, we use non-maximum suppression of 0.3. For COCO-a, we re-train the object detector excluding images from COCO that intersect with COCO-a. On UnRel, we use the same candidate pairs as [30] to have directly comparable results.

Visual representation. Following [30], we first encode a candidate pair of boxes $(\boldsymbol{o_s}, \boldsymbol{o_o})$ by the appearance of the subject $\boldsymbol{a}(\boldsymbol{o_s})$, the appearance of the object $\boldsymbol{a}(\boldsymbol{o_o})$, and their mutual spatial configuration $\boldsymbol{r}(\boldsymbol{o_s}, \boldsymbol{o_o})$. The appearance features of the subject and object boxes are extracted from the last fully-connected layer of the object detector. The spatial configuration $\boldsymbol{r}(\boldsymbol{o_s}, \boldsymbol{o_o})$ is a 8-dimensional feature that concatenates the subject and object box coordinates renormalized with respect to the union box. The visual representation of a candidate pair is a 1000-dimensional vector, aggregating the spatial and appearance features of the objects (more details in Section A.2 of the Appendix). For the subject (resp. object) embeddings, we only consider the appearance of the subject (resp. object) without the spatial configuration.

Language representation. For a triplet $t = (s, p, o)$, we compute the word embeddings e_s (resp. e_p, e_o) for subject (resp. predicate, object) with a Word2vec [28] model trained on GoogleNews. The representation of a triplet is taken as the concatenation of the word embeddings $\mathbf{q}_t = [\boldsymbol{e}_s; \boldsymbol{e}_p; \boldsymbol{e}_o] \in \mathbb{R}^{900}$.

Embedding functions. The embedding projection functions are composed of two fully connected layers, with a ReLU non-linearity. For the visual projection functions, we use Dropout. The dimensionality of the joint visual-language spaces is set to $d = 1024$ for HICO-DET and COCO-a. We use $d = 256$ for UnRel as the training set is much smaller.

Training details. We train our model with Adam optimizer [21] using a learning rate 0.001. We first learn the parameters of the projection functions by optimizing $\mathcal{L}_{joint}$, then activate the analogy loss $\mathcal{L}_\Gamma$ to learn the parameters of

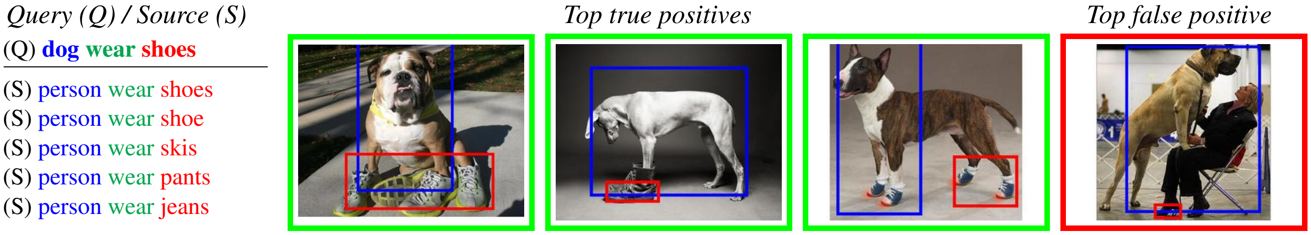


Figure 5: Top retrieved positive (green) and negative (red) detections with our model (p+vp+transfer) on UnRel triplets. The embedding of the unseen query triplet (Q) is formed from the embedding of seen source triplets (S) via analogy transformation. While transfer with analogy on HICO-DET is often done through change of object, here, for retrieving the unseen triplet “dog wear shoes”, our model samples source triplets involving a different subject, “person”, in interaction with similar objects (e.g. “person wear shoes”, ”person wear skis”). Additional examples are in Section D of the Appendix.

transfer and finetune the visual phrase embeddings. The hyperparameters α_s, α_o, α_p and k are optimized by grid-search on the validation set. More details on optimization and batch sampling are provided in Section A.2 of the Appendix.

4.3. Evaluating visual phrases on seen triplets

We first validate the capacity of our model to detect triplets seen at training and compare with recent state-of-the-art methods. In Table 1, we report mAP results on HICO-DET in the Default setting defined by [5] on the different subsets of triplets (full), (rare), (non rare) as described in 4.1. First, we compute three variants of our model : (i) the compositional part using all unigram terms (s+o+p), which can be viewed as a strong fully compositional baseline, (ii) the visual phrase part combined with object scores (s+o+vp), and (iii) our full model (s+o+p+vp) that corresponds to the addition of the visual phrase representation on top of the compositional baseline (section 3.1). The results show that our visual phrase embedding is beneficial, leading to a consistent improvement over the strong compositional baseline on all sets of triplets, improving the current state-of-the art [9] by more than 30% in terms of relative gain. We provide ablation studies in Section B of the Appendix as well as experiments incorporating bigrams modules (sr+ro) leading to improved results.

4.4. Transfer by analogy on unseen triplets

Next, we evaluate the benefits of transfer by analogy focusing on the challenging set-up of triplets never seen at training time. While the HICO-DET dataset contains both seen (evaluated in previous section) and manually constructred unseen triplets (evaluated here), in this section we consider additional two datasets that contain only unseen triplets. In particular, we use UnRel to evaluate retrieval of unusual (and unseen) triplets and COCO-a to evaluate retrieval of unseen triplets with out-of-vocabulary predicates.

	full	rare	non-rare
Chao [5]	7.8	5.4	8.5
Gupta [12]	9.1	7.0	9.7
Gkioxari [11]	9.9	7.2	10.8
GPNN [33]	13.1	9.3	14.2
iCAN [9]	14.8	10.5	16.1
s+o+p	18.7	13.8	20.1
s+o+vp	17.7	11.6	19.5
s+o+p+vp	**19.4**	**14.6**	**20.9**

Table 1: Retrieval results on HICO-DET dataset (mAP).

	Base	With aggregation G			
	-	Γ=$\emptyset$	Γ=0	Γ=$linear$	Γ=$deep$
s+o+p	23.2	-	-	-	-
s+o+vp+transfer	24.1	9.6	24.8	27.6	**28.6**
s+o+p+vp+transfer	23.6	12.5	24.5	25.4	**25.7**
supervised	33.7	-	-	-	-

Table 2: mAP on the 25 zero-shot test triplets of HICO-DET with variants of our model trained on the *trainval* set excluding the positives for the zero-shot triplets. The first column shows the results without analogy transfer (Section 3.1) while the other columns display results with transfer using different forms of analogy transformation Γ (Section 3.2). Last line (supervised) is the performance of (s+o+p+vp) trained will all training instances.

Evaluating unseen triplets on HICO-DET. First, we evaluate our model of transfer by analogy on the 25 zero-shot triplets of HICO-DET. In Table 2, we show results for different types of analogy transformations applied to the visual phrase embeddings to be compared with the base model not using analogy (first column). First, Γ=$\emptyset$ corresponds to aggregation of visual phrase embeddings of source triplets without analogy transformation. Then, we report three variants of an analogy transformation, where visual phrase embeddings are trained with analogy loss and the embedding of source triplet is either (i) aggregated without transformation (Γ=0), or transformed with (ii) a linear transformation (Γ=$linear$) or (iii) a 2-layer perceptron (Γ=$deep$). The results indicate that forming visual phrase embeddings of unseen test triplets by analogy transformations of similar seen triplets, as described in 3.2, is beneficial, with the best model (s+o+vp+transfer using Γ=$deep$) providing a significant improvement over the compositional baseline (from mAP of 23.2 to 28.6), thus partly filling the gap to the fully supervised setting (mAP of 33.7). It is also interesting to note that, when aggregating visual phrase embeddings of differ-

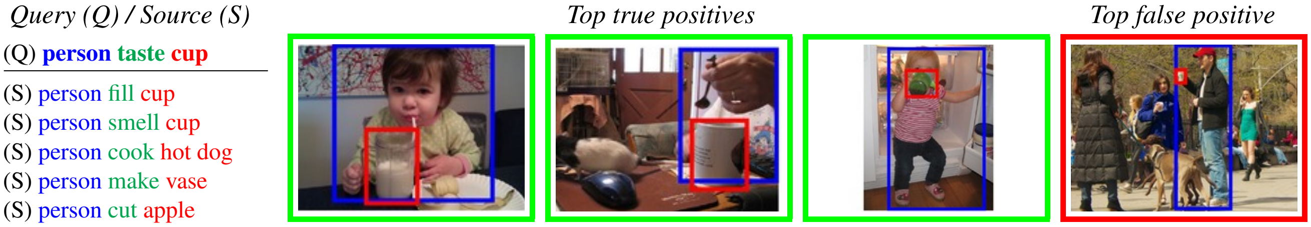


Figure 6: Top retrieved positives (green) and negatives (red) detections with our model (s+o+vp+transfer) of COCO-a triplets. The embedding of the query triplet (Q) to retrieve is formed with the embedding of source triplets (S) by analogy. For retrieving out-of-vocabulary triplets such as "person taste cup", our model of transfer by analogy automatically samples relevant source triplets involving similar predicates and objects (e.g. "person smell cup", "person make vase"). Additional examples are in Section E of the Appendix.

ent source triplets as described in Eq. (10), transforming the visual phrase embedding via analogy prior to the aggregation is necessary, as indicated by the significant drop of performance when $\Gamma = \emptyset$. In Figure 4 we show qualitative results for retrieval of unseen triplets with the (s+o+vp+transfer) model. For a query triplet (Q) such as "person pet cat" we show the top 3 retrieved candidate pairs (green), and the top 1 false positive (red). Also, for each target triplet, we show the source triplets (S) used in the transfer by analogy (Eq. (10)). We note that the source triplets appear relevant to the query.

Evaluating unseen (unusual) triplets on UnRel. Table 3 shows numerical results for retrieval on the UnRel dataset. Similar to [30], we also do not use subject and object scores as we found them uninformative on this dataset containing hard to detect objects. For transfer by analogy we use $\Gamma = deep$. First, we observe that our (p+vp+transfer) method improves over all other methods, significantly improving the current state-of-the-art [30] on this data, as well as outperforming the image captioning model of [16] trained on a larger corpus. Note that we use the same detections and features as [30], making our results directly comparable. Second, the results confirm the benefits of transfer by analogy (p+vp+transfer) over the fully compositional baseline (p) with a consistent improvement in all evaluation metrics. Interestingly, contrary to HICO-DET, using visual phrase embeddings without transfer (p+vp) does not bring significant improvements over (p). This is possibly due to the large mismatch between training and test data as the UnRel dataset used for testing contains unusual relations, as shown in the qualitative examples in Figure 5. This underlines the importance of the transfer by analogy model.

	With GT	With candidates		
	-	union	subj	subj/obj
DenseCap [16]	-	6.2	6.8	-
Lu [27]	50.6	12.0	10.0	7.2
Peyre [30] full	62.6	14.1	12.1	9.9
p	62.2	16.8	15.2	12.6
vp	53.4	13.2	11.7	9.4
p+vp	61.7	16.4	14.9	12.6
vp+transfer	53.7	13.7	12.0	9.7
p+vp+transfer	**63.9**	**17.5**	**15.9**	**13.4**

Table 3: Retrieval on UnRel (mAP) with IoU=0.3.

	all	out of vocabulary
s+o+p	4.3	4.2
s+o+vp	6.0	6.2
s+o+p+vp	5.1	5.1
s+o+vp+transfer	**6.9**	**7.3**
s+o+p+vp+transfer	5.2	5.1

Table 4: Retrieval on unseen triplets of COCO-a (mAP). We show the performance on all unseen triplets (first column) and on unseen triplets involving out-of-vocabulary predicates (second column).

Evaluating unseen (out-of-vocabulary) triplets on COCO-a. Finally, we evaluate our model trained on HICO-DET dataset for retrieval on the unseen triplets of COCO-a dataset. This is an extremely challenging setup as the unseen triplets of COCO-a involve predicates that are out of the vocabulary of the training data. The results shown in Table 4 demonstrate the benefits of the visual phrase representation as previously observed on HICO-DET and UnRel datasets. Furthermore, the results also demonstrate the benefits of analogy transfer : compared to the fully compositional baseline (s+o+p) our best analogy model (s+o+vp+transfer) obtains a relative improvement of 60% on all, and more than 70% on the out of vocabulary triplets. Qualitative results are shown in Figure 6.

5. Conclusion

We have developed a new approach for visual relation detection that combines compositional and visual phrase representations. Furthermore, we have proposed a model for transfer by analogy able to compute visual phrase embeddings of never seen before relations. We have demonstrated benefits of our approach on three challenging datasets involving unseen triplets.

Acknowledgements. This work was partly supported by ERC grants Activia (no.307574), LEAP (no.336845), Allegro (no.320559), CIFAR Learning in Machines & Brains program, the MSR-Inria joint lab, Louis Vuitton ENS Chair on Artificial Intelligence, DGA project DRAAF, and European Regional Development Fund under the project IMPACT (reg. no. CZ.02.1.01/0.0/0.0/15_003/0000468)

References

[1] Yuval Atzmon, Jonathan Berant, Vahid Kezami, Amir Globerson, and Gal Chechik. Learning to generalize to new compositions in image understanding. *arXiv:1608.07639*, 2016. 2

[2] Yusuf Aytar and Andrew Zisserman. Tabula rasa: Model transfer for object category detection. In *ICCV*, 2011. 2

[3] Trapit Bansal, Arvind Neelakantan, and Andrew McCallum. Relnet: End-to-end modeling of entities & relations. *arXiv:1706.07179*, 2017. 2

[4] Peter W. Battaglia, Razvan Pascanu, Matthew Lai, Danilo Jimeneze Rezende, and Koray Kavukcuoglu. Interaction networks for learning about objects, relations and physics. In *NIPS*, 2016. 2

[5] Yu-Wei Chao, Yunfan Liu, Xieyang Liu, Huayi Zeng, and Jia Deng. Learning to detect human-object interactions. In *WACV*, 2018. 5, 6, 7, 11

[6] Yu-Wei Chao, Zhan Wang, Yugeng He, Jiaxuan Wang, and Jia Deng. Hico: A benchmark for recognizing human-object interactions in images. In *ICCV*, 2015. 2, 5

[7] Bo Dai, Yuqi Zhang, and Dahua Lin. Detecting visual relationships with deep relational networks. In *CVPR*, 2017. 1, 2

[8] Santosh Kumar Divvala, Ali Farhadi, and Carlos Guestrin. Learning everything about anything: Webly-supervised visual concept learning. In *CVPR*, 2014. 2

[9] Chen Gao, Yuliang Zou, and Jia-Bin Huang. Ican: Instance-centric attention network for human-object interaction detection. In *BMVC*, 2018. 2, 7, 12

[10] Ross Girshick, Ilija Radosavovic, Georgia Gkioxari, Piotr Dollár, and Kaiming He. Detectron. https://github.com/facebookresearch/detectron, 2018. 6

[11] Georgia Gkioxari, Ross Girshick, and Kaiming He. Detecting and recognizing human-object interactions. In *CVPR*, 2018. 2, 6, 7, 12

[12] Saurabh Gupta and Jitendra Malik. Visual role semantic labeling. *arXiv:1505.04474*, 2015. 7

[13] Seong Jae Hwang, Sathya N. Ravi, Zirui Tao, Hyunwoo J. Kim, Maxwell D. Collins, and Vikas Singh. Tensorize, factorize and regularize: Robust visual relationship learning. In *CVPR*, 2018. 2

[14] Hamid Izadinia, Fereshteh Sadeghi, Santosh Kumar Divvala, Yejin Choi, and Ali Farhadi. Segment-phrase table for semantic segmentation, visual entailment and paraphrasing. In *ICCV*, 2015. 2

[15] Rodolphe Jenatton, Nicolas L Roux, Antoine Bordes, and Guillaume R Obozinski. A latent factor model for highly multi-relational data. In *NIPS*, 2012. 2

[16] Justin Johnson, Andrej Karpathy, and Li Fei-Fei. Densecap: Fully convolutional localization networks for dense captioning. In *CVPR*, 2016. 8

[17] Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David A Shamma, Michael S Bernstein, and Li Fei-Fei. Image retrieval using scene graphs. In *CVPR*, 2015. 2

[18] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, 2015. 2

[19] Andrej Karpathy, Armand Joulin, and Li Fei-Fei. Deep fragment embeddings for bidirectional image sentence mapping. In *NIPS*, 2014. 2, 4

[20] Keizo Kato, Yi Li, and Abhinav Gupta. Compositional learning for human object interaction. In *ECCV*, 2018. 2

[21] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 6, 12

[22] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *ICLR*, 2016. 2

[23] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, Michael Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. In *IJCV*, 2016. 2

[24] Yikang Li, Wanli Ouyang, and Xiaogang Wang. Vip-cnn: A visual phrase reasoning convolutional neural network for visual relationship detection. In *CVPR*, 2017. 1, 2

[25] Tsung-Yi Lin, Piotr Dollár, Ross B. Girshick, Kaiming He, Bharath Hariharan, and Serge J. Belongie. Feature pyramid networks for object detection. In *CVPR*, 2017. 6

[26] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, 2014. 5, 6

[27] Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. Visual relationship detection with language priors. In *ECCV*, 2016. 1, 2, 6, 8

[28] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *NIPS*, 2013. 4, 6

[29] Ishan Misra, Abhinav Gupta, and Martial Hebert. From red wine to red tomato: Composition with context. In *CVPR*, 2017. 2

[30] Julia Peyre, Ivan Laptev, Cordelia Schmid, and Josef Sivic. Weakly-supervised learning of visual relations. In *ICCV*, 2017. 1, 2, 5, 6, 8

[31] Bryan A. Plummer, Arun Mallya, Christopher M. Cervantes, Julia Hockenmaier, and Svetlana Lazebnik. Phrase localization and visual relationship detection with comprehensive linguistic cues. In *ICCV*, 2017. 2

[32] Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *ICCV*, 2015. 2

[33] Siyuan Qi, Wenguan Wang, Baoxiong Jia, Jianbing Shen, and Song-Chun Zhu. Learning human-object interactions by graph parsing neural networks. In *ECCV*, 2018. 2, 7

[34] Vignesh Ramanathan, Congcong Li, Jia Deng, Wei Han, Zhen Li, Kunlong Gu, Yang Song, Samy Bengio, Chuck Rossenberg, and Li Fei-Fei. Learning semantic relationships for better action retrieval in images. In *CVPR*, 2015. 2

[35] Scott E. Reed, Yi Zhang, Yuting Zhang, and Honglak Lee. Deep visual analogy-making. In *NIPS*, 2015. 2

[36] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *NIPS*, 2015. 6

[37] Matteo Ruggero Ronchi and Pietro Perona. Describing common human visual actions in images. In *BMVC*, 2015. 5, 6

[38] Fereshteh Sadeghi, Santosh K Divvala, and Ali Farhadi. Viske: Visual knowledge extraction and question answering by visual verification of relation phrases. In *CVPR*, 2015. 2

[39] Fereshteh Sadeghi, C. Lawrence Zitnick, and Ali Farhadi. Visalogy: Answering visual analogy questions. In *NIPS*, 2015. 2

[40] Mohammad Amin Sadeghi and Ali Farhadi. Recognition using visual phrases. In *CVPR*, 2011. 1, 2

[41] Adam Santoro, David Raposo, David G.T. Barrett, Mateusz Malinowski, Razvan Pascanu, Peter Battaglia, and Timothy Lillicrap. A simple neural network module for relational reasoning. In *NIPS*, 2017. 2

[42] Liyue Shen, Serena Yeung, Judy Hoffman, Greg Mori, and Li Fei-Fei. Scaling human-object interaction recognition through zero-shot learning. In *WACV*, 2018. 2

[43] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *JMLR*, 2008. 11, 14, 18, 19, 20, 21

[44] Liwei Wang, Yin Li, and Svetlana Lazebnik. Learning deep structure-preserving image-text embeddings. In *CVPR*, 2016. 2, 3, 4

[45] Ruichi Yu, Ang Li, Vlad I. Morariu, and Larry S. Davis. Visual relationship detection with internal and external linguistic knowledge distillation. In *ICCV*, 2017. 2

[46] Hanwang Zhang, Zawlin Kyaw, Shih-Fu Chang, and Tat-Seng Chua. Visual translation embedding network for visual relation detection. In *CVPR*, 2017. 1, 2

[47] Ji Zhang, Yannis Kalantidis, Marcus Rohrbach, Manohar Paluri, Ahmed Elgammal, and Mohamed Elhoseiny. Large-scale visual relationship understanding. In *AAAI*, 2019. 2

[48] Bohan Zhuang, Lingqiao Liu, Chunhua Shen, and Ian Reid. Towards context-aware interaction recognition for visual relationship detection. In *ICCV*, 2017. 2

Overview

In this Appendix, we provide (i) technical details of our model described in Section 3 of the main paper (Section A), (ii) additional ablation studies to better understand the benefits of the different components of our model (Section B), (iii) additional qualitative results of our model for transfer by analogy on HICO-DET dataset (Section C), UnRel dataset (Section D) and COCO-a dataset (Section E), and (iv) a qualitative analysis of joint embedding spaces by the t-sne [43] visualization (Section F).

A. Additional details of our model

In this part, we provide additional details of our model that we could not include in the main paper due to space constraints. First, in Section A.1, we describe how we learn the analogy transformation including details on (i) sampling source triplets, (ii) training loss and (iii) optimization. Second, in Section A.2, we detail our visual representation and explain how we form mini-batches during training.

A.1. Learning analogy transformations

Sampling source triplets. Please recall (Section 3.2 of the main paper) that we fit parameters of Γ by learning analogy transformations between triplets available in the training data. To do this, we generate pairs of source t and target t' triplets as follows. For a target triplet t' in the training data, the source triplets for transfer by analogy are sampled in two steps : (i) for a given target triplet t', we first compute the similarity $G(t, t')$ given by Eq. (9) using all triplets t in the training data that occur at least 10 times (i.e. the non-rare triplets according to the definition of [5]), (ii) we sort this set of candidate source triplets, and retain the top k most similar triplets according to G. The outcome is a set of source triplets $\mathcal{N}_{t'}$, similar to the target triplet t' and hence suitable for learning the analogy transformation. Please note that we do not constrain the source triplets to share words with the target triplet, so all words may differ between source and target triplets. Also note that the procedure described above is similar at training and test time. In practice, we take $k = 5$, $\alpha_r = 0.8$, $\alpha_s = \alpha_o = 0.1$ for all datasets. These hyperparameters are optimized by grid-search on the validation set of HICO-DET.

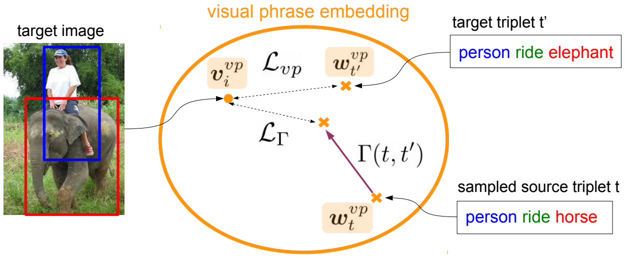


Figure A: Illustration of training the analogy transformation Γ. For each target triplet t' (e.g. "person ride elephant"), we randomly sample a source triplet t (e.g. "person ride horse"). The first part of the analogy loss $\mathcal{L}_\Gamma$ in Eq (11) encourages that the transformed visual phrase embedding $w_t^{vp} + \Gamma(t, t')$ is close to the corresponding visual representation v_i^{vp} of target triplet t.

Learning Γ. For each target triplet t' in the training batch, we randomly sample a relevant source triplet $t \in \mathcal{N}_{t'}$ as described above. We call $\mathcal{Q}$ the set of pairs of related triplets (t, t') formed like this. The parameters of Γ are learnt by maximizing the log-likelihood :

$$\mathcal{L}_\Gamma = \sum_{i=1}^{N} \sum_{(t,t')\in\mathcal{Q}} \mathbb{1}_{y^i_{t'}=1} \log\left(\frac{1}{1+e^{-(\boldsymbol{w}_t^{vp}+\Gamma(t,t'))^T \boldsymbol{v}_i^{vp}}}\right) + \sum_{i=1}^{N} \sum_{(t,t')\in\mathcal{Q}} \mathbb{1}_{y^i_{t'}=0} \log\left(\frac{1}{1+e^{(\boldsymbol{w}_t^{vp}+\Gamma(t,t'))^T \boldsymbol{v}_i^{vp}}}\right), \tag{11}$$

where $\boldsymbol{v}_i^{vp}$ are the visual features projected to the visual phrase space and $(\boldsymbol{w}_t^{vp} + \Gamma(t, t'))$ is the transformed visual phrase embedding of the source triplet t to target triplet t' following Eq. (6) in the main paper. Note that this loss is similar to the loss used for learning embeddings of visual relations, given by Eq. (3) in the main paper. The first attraction term pulls closer visual representation $\boldsymbol{v}_i^{vp}$ to its corresponding language representation $\boldsymbol{w}_t^{vp} + \Gamma(t, t')$ obtained via analogy transformation, i.e. where the visual representation matches the embedding of the target triplet t' obtained via analogy transformation. We illustrate this term in Figure A. The second repulsive term pushes apart visual-language pairs that do not match, i.e. where the visual representation does not match the target triplet t' obtained via the analogy transformation. The main idea behind Eq. (11) is to use the analogy transformation Γ to make the link between the language embedding of a source triplet t and the visual embedding of a target triplet t' in the joint vp space. For example, let us consider source-target pairs of triplets $\mathcal{Q} = \{(t_1, t'_1), (t_2, t'_2)\}$ in a mini-batch, where, $t_1 = (person, ride, horse)$, $t'_1 = (person, ride, elephant)$, $t_2 = (person, pet, cat)$, $t'_2 = (person, pet, sheep)$. The analogy loss in Eq. (11) learns Γ that transforms, in the joint vp space, the language embedding of the source triplet $(person, ride, horse)$ such that it is close to the visual embedding of the target triplet $(person, ride, elephant)$ (first term in the loss) but far from the visual embedding of the other target triplet $(person, pet, sheep)$ (second term in the loss).

Optimization details. First, we learn the parameters of embedding functions by optimizing $\mathcal{L}_{joint} = \mathcal{L}_s + \mathcal{L}_o +$

$\mathcal{L}_p + \mathcal{L}_{vp}$ (Eq. (3) in the main paper) for 10 epochs with Adam optimizer [21] using a learning rate 0.001. Then, we fix parameters of the embedding functions for s, o and p and only finetune parameters of the visual phrase embedding function vp while learning parameters of analogy transformation Γ. This is done by jointly optimizing $\mathcal{L}_{vp} + \lambda \mathcal{L}_\Gamma$ for 5 epochs with Adam optimizer [21] using a learning rate 0.001. In practice, we take $\lambda = 1$. In this joint optimization, we found it helpful to restrict back-propagation of gradients coming from $\mathcal{L}_\Gamma$ only to the parameters of analogy transformation Γ and parameters of the visual embedding functions f_v^b (Eq. (1)), i.e. we exclude back-propagation of gradients coming from $\mathcal{L}_\Gamma$ to parameters of language embedding functions f_w^b. These parameters are finetuned using gradients back-propagated from $\mathcal{L}_{vp}$.

A.2. Implementation details

Visual representation. As described in Section 4 of the main paper, a candidate pair of bounding boxes $(\boldsymbol{o_s}, \boldsymbol{o_o})$ is encoded by the appearance of the subject $\boldsymbol{a}(\boldsymbol{o_s})$, the appearance of the object $\boldsymbol{a}(\boldsymbol{o_o})$, and their mutual spatial configuration $\boldsymbol{r}(\boldsymbol{o_s}, \boldsymbol{o_o})$. The spatial configuration $\boldsymbol{r}(\boldsymbol{o_s}, \boldsymbol{o_o})$ is a 8-dimensional feature that concatenates the subject and object box coordinates renormalized with respect to the union box, i.e. we concatenate $[\frac{x_{min}-T}{A}, \frac{x_{max}-T}{A}, \frac{y_{min}-T}{A}, \frac{y_{max}-T}{A}]$ for subject and object boxes where T and A are the origin and the area of the union box, respectively. The visual representation of a candidate pair is then

$$\mathbf{x}_i = \begin{bmatrix} MLP_s(\boldsymbol{a}(\boldsymbol{o_s})) \\ MLP_o(\boldsymbol{a}(\boldsymbol{o_o})) \\ MLP_r(\boldsymbol{r}(\boldsymbol{o_s}, \boldsymbol{o_o})) \end{bmatrix}, \tag{12}$$

where MLP_s, MLP_o contain one layer that projects the appearance features into a vector of dimension 300 and MLP_r is a 2-layer perceptron projecting the spatial features into a vector of dimension 400, making the final dimension of $\mathbf{x}_i$ equal to 1000. Note that both p and vp use the same visual input (including spatial features) while s and o modules only use the appearance features.

Sampling batches. In practice, our model is trained with mini-batches containing 64 candidate object pairs. 25% of the candidate pairs are positive, i.e. the candidate subject and object are interacting. The rest are negative, randomly sampled among candidate pairs involving the same subject and object category (but not interacting). For training, we use candidates from both ground truth and object detector outputs. At test time, we only use candidate pairs from the object detector.

B. Ablation studies

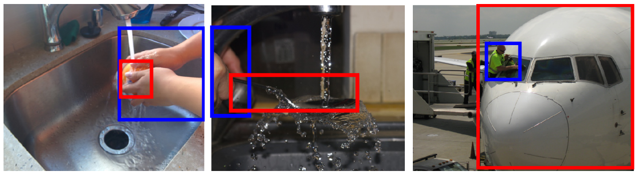
In this section, we perform ablation studies that complement the analysis in Section 4 of the main paper. We discuss the benefits of the different components of our model introduced in Section 3.1 of the main paper, and in particular the benefits of the visual phrase module. We also analyze the influence of pre-trained word embeddings and the effect of adding bigrams modules.

		full	rare	non-rare
(a)	s+o (obj.det.)	5.6	4.2	6.5
(b)	s+o	10.0	7.6	10.8
(c)	p	14.9	9.4	16.5
(d)	bigrams	14.9	9.6	16.5
(e)	vp	16.5	10.4	18.4
(f)	s+o+vp (*main paper*)	17.7	11.6	19.5
(g)	s+o+p (classifier)	18.0	13.4	19.4
(h)	s+o+p (random words)	18.4	13.7	19.8
(i)	s+o+p (*main paper*)	18.7	13.8	20.1
(j)	s+o+p (finetuned words)	18.8	14.5	20.1
(k)	s+o+p+vp (*main paper*)	19.4	14.6	20.9
(l)	s+o+p+bigrams	19.5	14.6	21.0
(m)	s+o+p+vp+bigrams	**20.0**	**15.0**	**21.5**

Table A: Ablation study on HICO-DET.

Benefits of different components of our model. Our contribution is a hybrid model which combines subject (*s*), object (*o*), predicate (*p*) and visual phrase (*vp*) modules. We show in Table A, which complements Table 1 of the main paper, that each of these modules is making a complementary contribution. The performance of our compositional model *s+o+p* builds on our strong unigram models *s+o* (row (b)) that already significantly improve over the baseline using only the object scores returned by pre-trained object detectors (row (a)) typically used by other methods [9, 11]. The strength of our modules for representing visual relations is clearly demonstrated by the good performance of our unigram predicate model *p* (row (c)) and the visual phrase model *vp* (row (e)) over using objects alone (cf. *s+o*, row (b)). In addition, *vp* alone performs better than *p* alone (row (e) > (c)). Importantly, these modules are complementary as clearly shown by the best performance of our combined model (row (k)) that can also easily incorporate bigrams (row (m)), see below.

Benefits of visual phrase (vp) model. The improvement thanks to the *vp* model is consistent over several datasets. We found qualitatively that the *vp* branch handles important unusual situations where the compositional model (*s+o+p*) fails, which happens when (at least) one of the *s*, *o* and *p* branches has a low score, e.g. due to object occlusion (Figure B(a)), unusual object appearance (Figure B(b)) or unusual spatial configuration (Figure B(c)). The visual phrase model (*vp*) can better handle these situations because it better models the specific appearance and spatial configuration of triplets seen in training.



(a) person wash orange (b) person wash spoon (c) person wash airplane

Figure B: Retrieval examples where *s+o+p+vp* is better than *s+o+p*.

Benefits of word vectors. In Table A we (1) show benefits of mapping input triplets to image-language embedding space instead of learning *s*, *o* and *p* classifiers (row (h) > (g)) and (2) confirm that using pre-trained word embeddings helps, but only slightly (row (i) > (h)). Because of the mismatch between word usage in the pre-training text corpus and our dataset, we also found that fine-tuning the pre-trained word embeddings is beneficial (row (j) > row (i)).

Incorporating bigrams. While our primary focus is to marry compositional (unigrams) and visual phrase (trigram) models, we can easily incorporate bigram branches (*sp+po*) in our model. As shown in Table A, bigrams provide an improvement over unigrams modelling subject and object independently *s+o* (row (d) > (b)) and combined with unigrams (row (l)) they reach comparable results to a combination of unigrams and trigram (row (k)). Interestingly, bigrams and trigram are complementary. Their combination leads to the overall best results (row (m)).

Alternative to weighting function G. We tested an alternative to the weighting function G (Eq. (9) of the main paper) taking as input word2vec embeddings instead of joint visual-semantic embeddings in s, o and p spaces. This lead to a slight performance drop (28.3 vs. 28.6 for our analogy transfer in Table 2 of the main paper). This result suggests that while pre-trained language embeddings are core ingredients to establish similarities between concepts, they can be further strengthen by using visual appearance.

C. Qualitative results on HICO-DET dataset

In Figure C we show additional examples of retrieved detections for unseen triplets that supplement Figure 4 of the main paper. These qualitative examples confirm that our model for transfer by analogy (s+o+vp+transfer) (Section 3.2 of the main paper) automatically selects relevant source triplets (S) given an unseen triplet query (Q). For instance, for the query triplet "person throw frisbee" (first row), our model selects (1) a source triplet that involves the same action, with a different, but similar, object "person throw sports ball", (2) two source triplets with the same object, and different, but related, actions "person catch frisbee", "person block frisbee" and (3) two other source triplets with different, but related, object and actions "person hit sports ball", "person serve sports ball". Similar conclusions hold for the other examples displayed. The top false positives indicate that the main failure mode is the confusion with another similar interaction (e.g. "lie on" is confused with "sit on" in row 3 or "inspect" is confused with "hold" in row 4. Some detections are also incorrectly classified as failure, as they are still some missing ground truth annotations (e.g. row 2, row 6).

D. Qualitative results on UnRel dataset

In Figure D we show additional qualitative results for our model (p+vp+transfer) for retrieval of unseen (unusual) triplets on the UnRel dataset supplementing results shown in Figure 5 of the main paper. We show the source triplets (S) automatically sampled by our analogy model that are used to form the visual phrase embedding of the target query (Q). The top true positive retrievals are shown in green, the top false positive retrieval is shown in red. The automatically sampled source triplets all appear relevant. Our method samples source triplets involving (1) a different subject ("dog ride bike" is transferred from "person ride bike", "building has wheel" is transferred from "truck has wheel"), (2) a different object ("person stand on horse" is transferred from "person stand on sand"), or (3) a different predicate ("cone on the top of person" is transferred from "sky over person"). The results confirm that our model works well not only for human-object interactions but also for more general interactions involving spatial relations (e.g. "in", "on the top of") or a subject different from a person (e.g. "cone", "car", "building", "dog"). There are two main failure modes illustrated by the top false positive detections. The first one is an incorrect object detection (e.g. "train" is confused with "building" in row 3, or "motorcycle" is confused with "bike" in row 2). The second failure mode is due to the confusion with another similar triplet, possibly due to the unusual character of UnRel queries which sometimes make it difficult to sample close enough source triplets for the transfer by analogy. For instance, it is hard to form a good embedding for "car in building" from source triplets "car in street", "bus in street", "person in street" as these source triplets have fairly different visual appearance (row 5).

E. Qualitative results on COCO-a dataset

In Figure E, we show additional qualitative results of our model for transfer by analogy (s+o+vp+transfer) on retrieval of unseen (out of vocabulary) triplets in the COCO-a dataset, complementing results in Figure 6 of the main paper. We display the source triplets (S) automatically sampled by our model for a target query (Q). Despite the fact that the target predicates are not seen in training, our model manages, most of the time, to sample relevant source triplets for transfer. For instance, our model would link the unseen triplet "person use laptop", involving the unseen predicate "use" (row 2) to source triplets such as “person type on laptop”, “person read laptop” or “person text on phone”, all involving a predicate that is relevant to the unseen target predicate “use”. The same holds for the unseen triplet “person touch horse” (row 3) for which our model samples source triplets involving contact interaction such as “person hug horse”, “person pet horse” or “person kiss horse”. The top false detections are informative : (i) either they correspond to interactions involving related triplets, which are likely to be sampled as source triplets (e.g. “person shear sheep” confused with “person caress sheep” in row 1), (ii) or they correspond to interactions with ambiguous semantics (e.g. “person get frisbee” or “person prepare kite” that involve ambiguous predicates that could correspond to a large variety of spatial configurations).

F. Visualization of joint embedding spaces

Here, we provide additional insights about the embedding spaces learnt on the HICO-DET dataset and UnRel dataset using the t-sne visualization [43] of the final learnt joint embedding. First, we show t-sne visualization [43] of joint embedding spaces learnt for objects and predicates on HICO-DET to better understand which concepts are close together in the learnt space. For the object embedding, as shown in Figure F, objects are grouped according to their visual and semantic similarity. The same holds for predicate embeddings shown in Figure G. We draw similar plots for UnRel dataset, showing the object embedding in Figure H and the predicate embedding in Figure I. The visualization of predicate embedding on UnRel dataset in Figure I is especially interesting as it involves spatial relations. We remark that our model is able to separate spatial relations such as “under” from “above” which are semantically very similar. Learning good embedding for unigrams is crucial in our model for transfer by analogy, as unigram embeddings directly influence the analogy transformation from the seen visual phrases to the unseen ones.

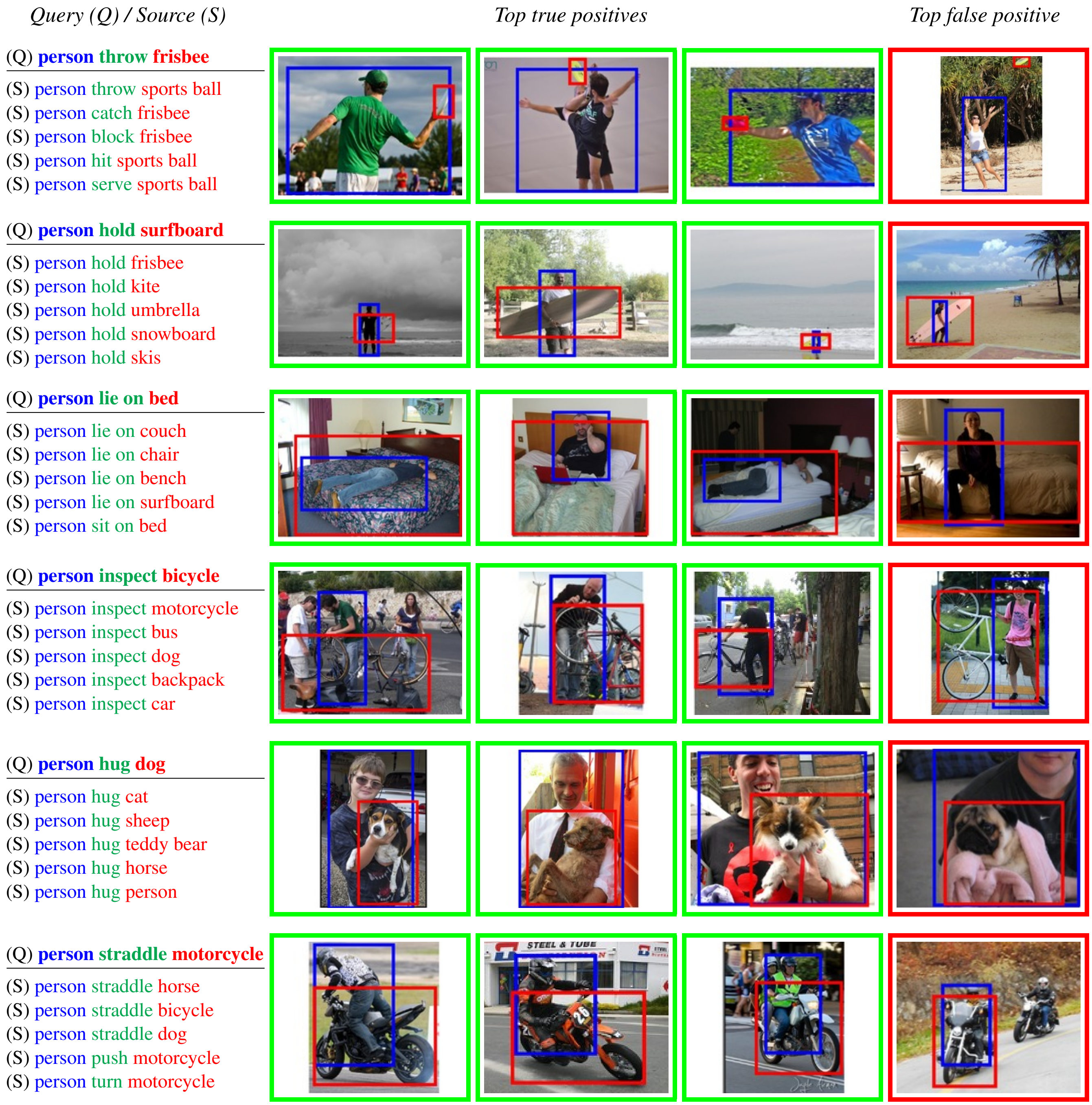


Figure C: **Retrieval examples on the HICO-DET dataset.** Top retrieved positives (green) and negatives (red) using our model (s+o+vp+transfer) for unseen triplet queries. The query is marked as (Q). The source triplets automatically selected by our model are marked as (S). For instance, for the query triplet "person throw frisbee" (first row), our model selects (1) a source triplet that involves the same action, with a different, but similar, object "person throw sports ball", (2) two source triplets with the same object, and different, but related, actions "person catch frisbee", "person block frisbee" and (3) two other source triplets with different, but related, object and actions "person hit sports ball", "person serve sports ball". The top false positives show the main failure mode: the interaction is confused with another similar interaction (e.g. "lie on" is confused with "sit on" in row 3 or "inspect" is confused with "hold" in row 4). Also, we note that some mistakes among the top false positives are due to missing ground truth annotations.

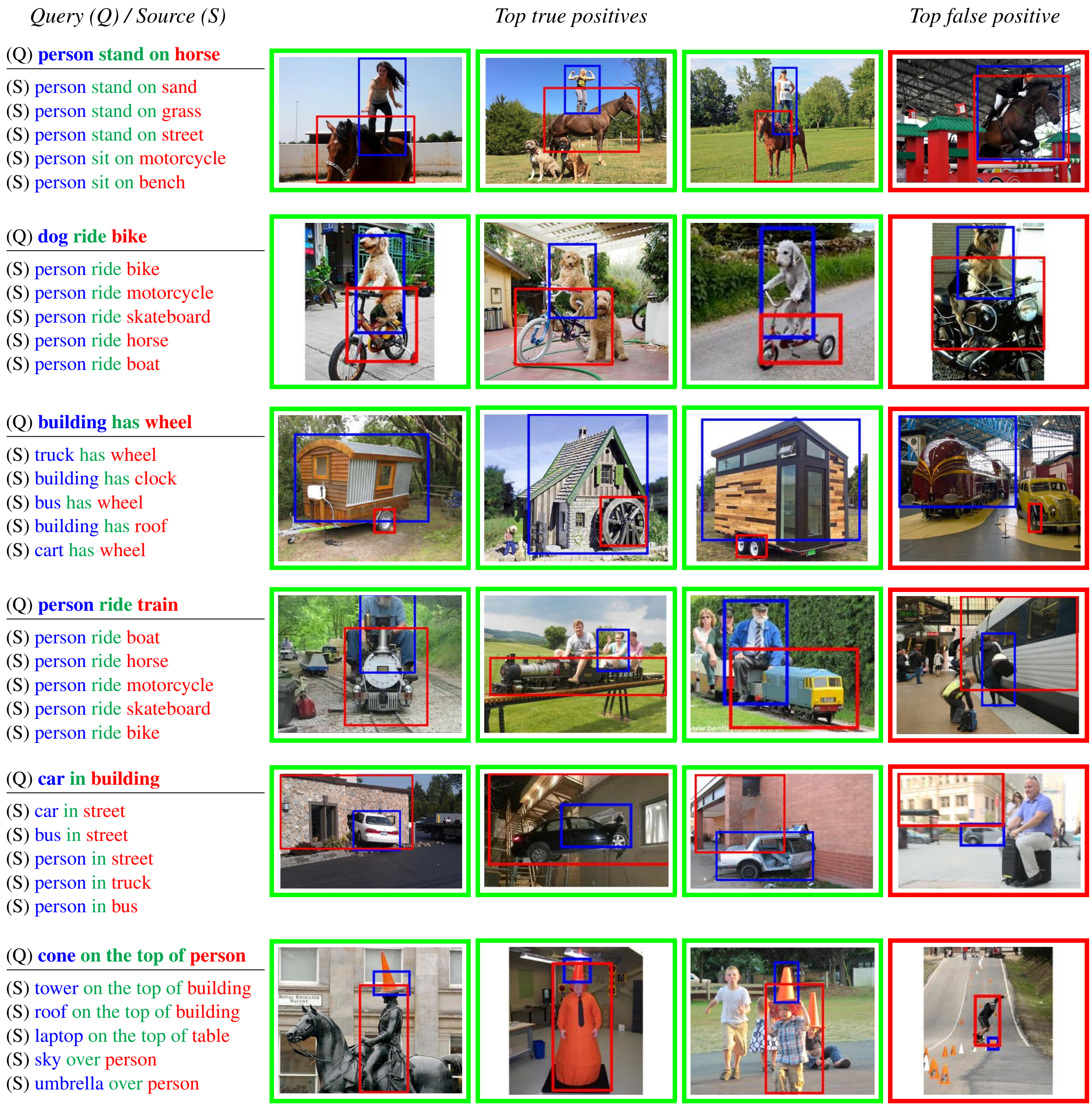


Figure D: **Querying for unseen (unusual) triplets on the UnRel dataset.** Examples of retrieval using our model (p+vp+transfer). The query triplet is marked as (Q). The source triplets (S) seen in training are automatically selected by our model described in 3.2 and used to transfer the visual phrase embedding using the analogy transformation. The automatically selected source triplets all appear relevant. Our method selects source triplets involving (1) a different subject ("dog ride bike" is transferred from "person ride bike", "building has wheel" is transferred from "truck has wheel"), (2) different object ("person stand on horse" is transferred from "person stand on sand"), or (3) different predicate ("cone on the top of person" is transferred from "sky over person").

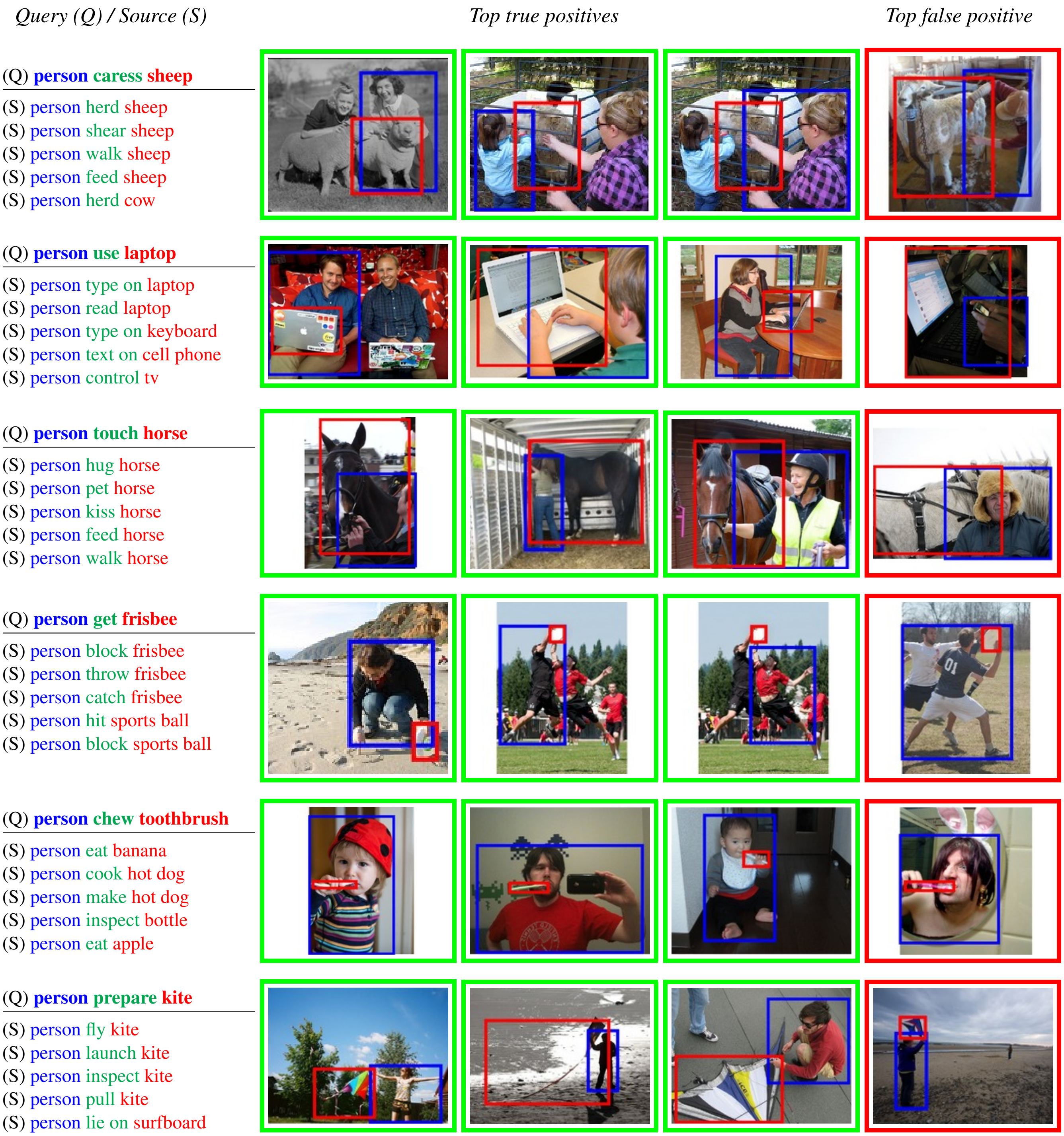


Figure E: **Querying for unseen (out of vocabulary) triplets on the COCO-a dataset.** Examples of retrieval using our model (s+o+vp+transfer). The query triplet is marked as (Q). The source triplets (S) seen in training are automatically selected by our model described in 3.2 and used to transfer the visual phrase embedding using the analogy transformation. The automatically selected source triplets all appear relevant despite the difficulty that all predicates involved in the shown triplet queries are unseen at training time. The transfer to unseen predicates is made possible by the use of pre-trained word2vec embeddings. Given out-of-vocabulary triplets such as "person use laptop" (row 2) or "person touch horse" (row 3), our model automatically samples source triplets involving a relevant predicate such as "person type on laptop" or "person hug horse". However, we also observe that sometimes the out-of-vocabulary predicate is ambiguous (e.g. "prepare" or "get"), which makes it challenging to identify relevant source triplets among the set of available training triplets (e.g. "person launch kite", "person catch frisbee").

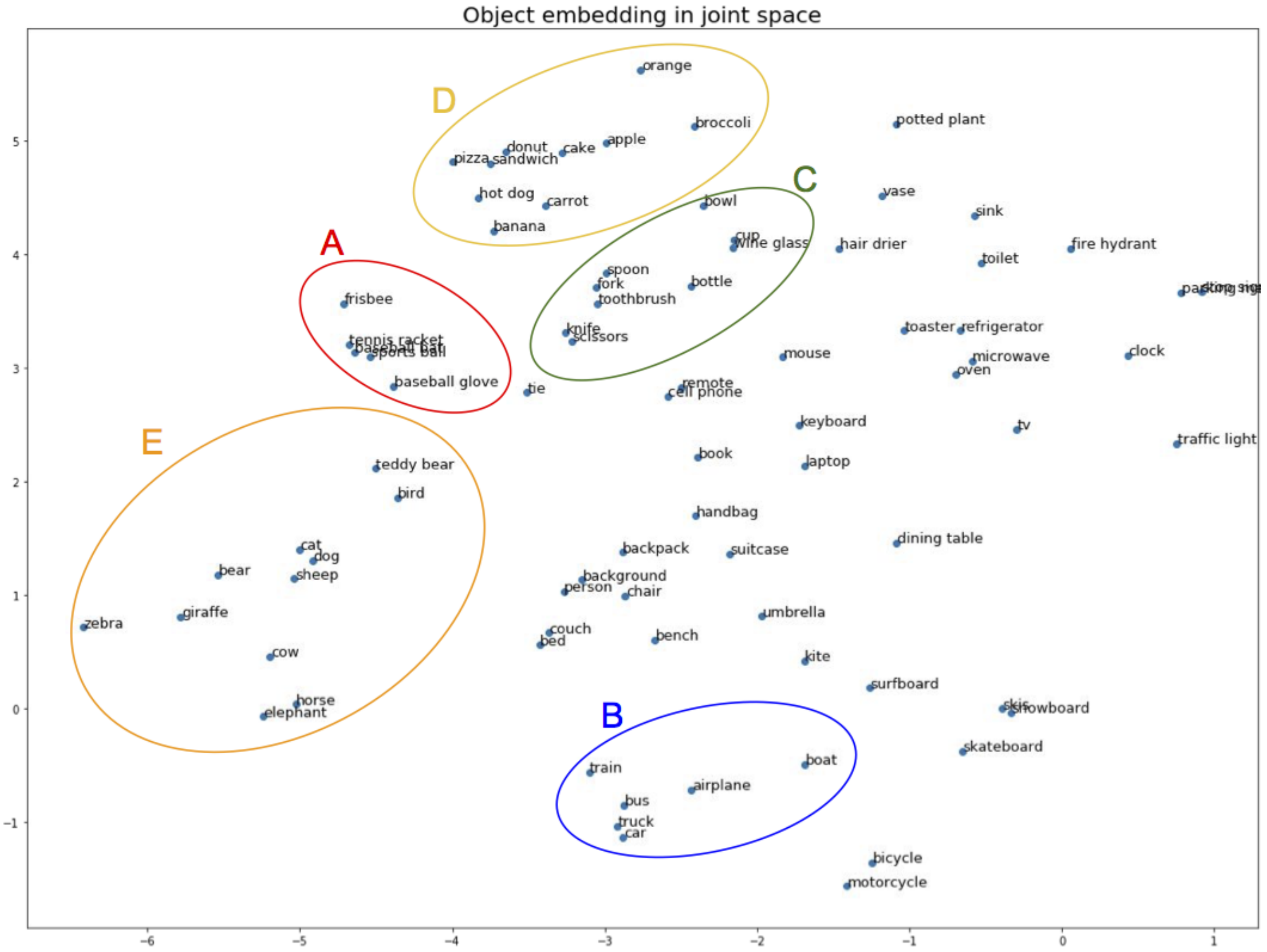


Figure F: Object embedding on HICO-DET visualized using T-sne [43]. Objects appear to be grouped according to their visual and semantic similarity. For example, we highlight regions corresponding to: (A) sports instruments (e.g. "tennis racket", "frisbee"), (B) big transportation (e.g. "bus", "train"), (C) eating utensils (e.g. "fork", "cup"), (D) food (e.g. "pizza", apple"), (E) animals (e.g. "giraffe", "bird"). Learning a good embedding for unigrams (here objects) is crucial in our model that uses the transfer by analogy, as unigram embeddings directly influence the analogy transformation from the seen visual phrases to the unseen ones.

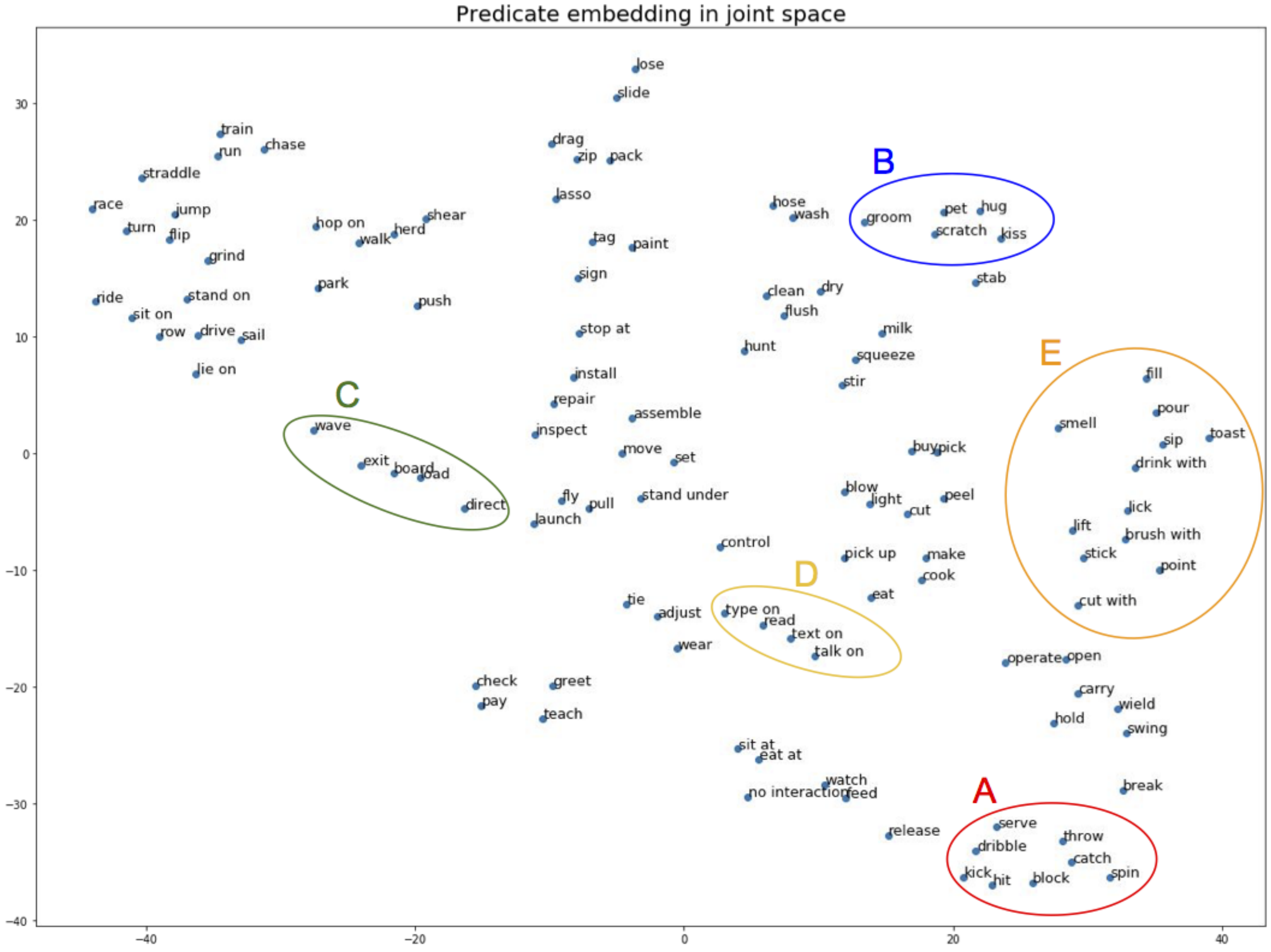


Figure G: Predicate embedding on HICO-DET visualized with T-sne [43]. The predicates are grouped according to their visual and semantic similarity. For example, we highlight regions corresponding to: (A) interactions related to sports (e.g. “throw”, “dribble”), (B) gentle interactions with an animal/person (e.g. “hug”, “kiss”), (C) interactions with transportation vehicles (e.g. “board”, “exit”), (D) interactions with (electronic) devices (e.g. “text on”, “read”), (E) interactions with food (e.g. “smell”, “lick”). Learning a good embedding for unigrams (here predicates) is crucial in our model that uses transfer by analogy, as unigram embeddings directly influence the analogy transformation from the seen visual phrases to the unseen ones.

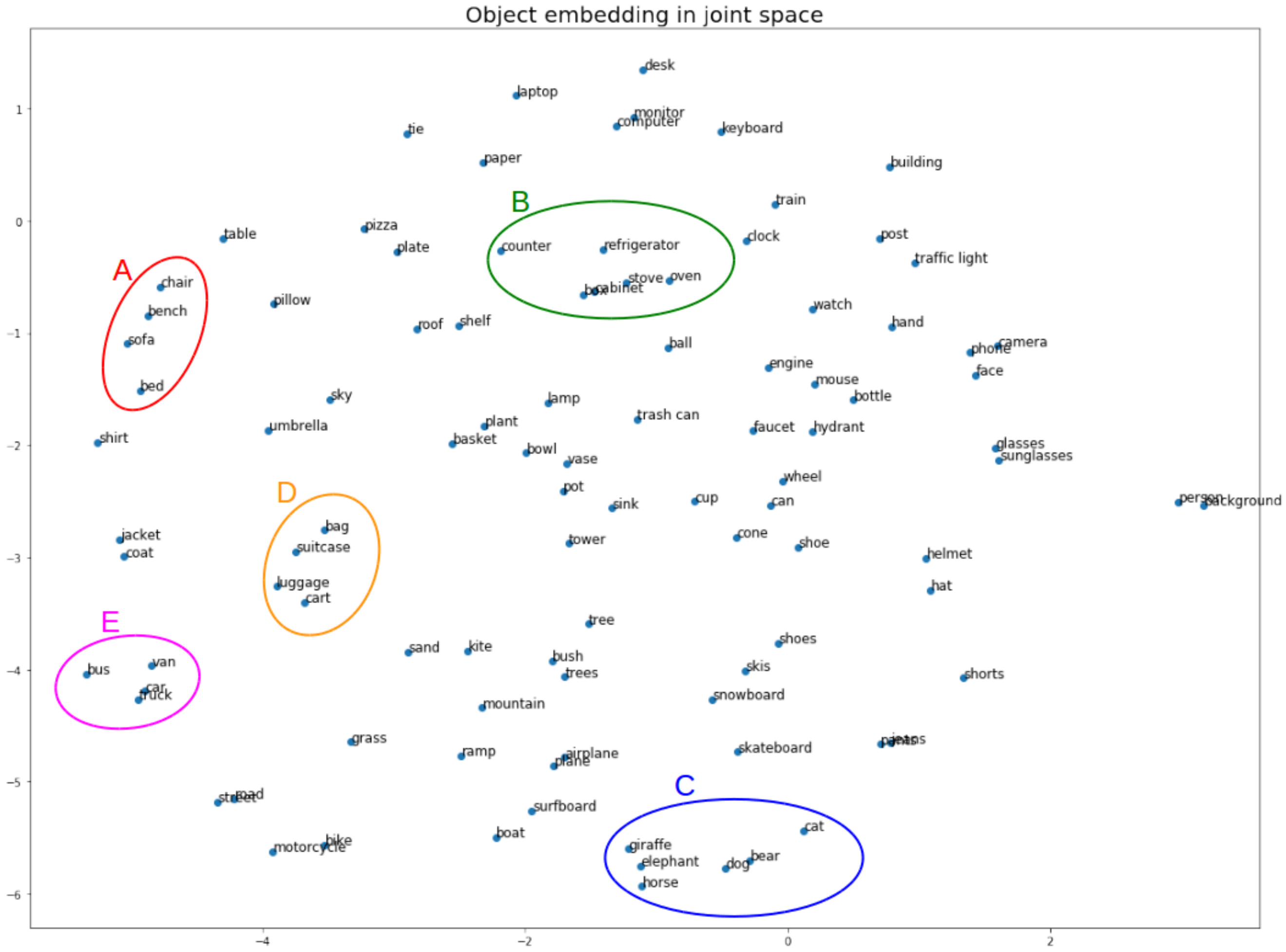


Figure H: Object embedding on the UnRel dataset visualized using T-sne [43]. Objects appear to be grouped according to their visual and semantic similarity. For example, we highlight regions corresponding to: (A) piece of furniture on which to sit (e.g. “chair”, “bench”), (B) kitchen furniture (e.g. “refrigerator”, “stove”), (C) animals (e.g. “giraffe”, “cat”), (D) bags and containers (e.g. “suitcase”, cart”), (E) motorized transportation (e.g. “bus”, “car”). Learning a good embedding for unigrams (here objects) is crucial in our model that uses the transfer by analogy, as unigram embeddings directly influence the analogy transformation from the seen visual phrases to the unseen ones.

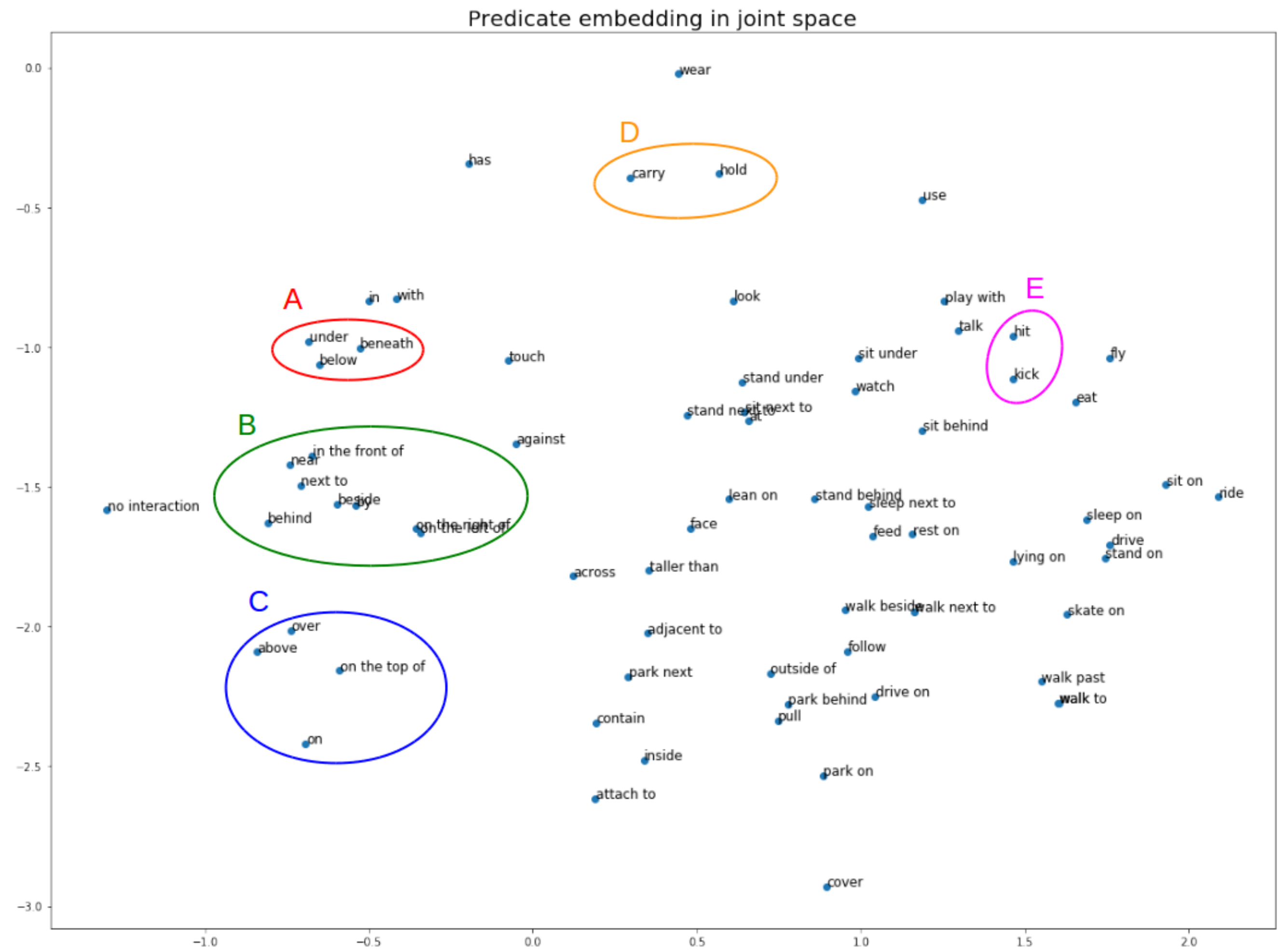


Figure I: Predicate embedding on the UnRel dataset visualized with T-sne [43]. The predicates are grouped according to their visual and semantic similarity. For example, we highlight regions corresponding to: (A) spatial relations related to “under” (e.g. “below”, “beneath”), (B) spatial relations related to “next to” (e.g. “near”, “beside”), (C) spatial relations related to “above” (e.g. “over”, “on the top of”), or similar actions (D) and (E). Note that it is remarkable that our visual-semantic embedding separates relations such as those in (A) from those in (C) while they are very similar from a strictly semantic point of view (in pre-trained word2vec embeddings). Learning a good embedding for unigrams (here predicates) is crucial in our model that uses transfer by analogy, as unigram embeddings directly influence the analogy transformation from the seen visual phrases to the unseen ones.