Zero-Shot Human-Object Interaction Recognition via Affordance Graphs

Alessio Sarullo,[1] **Tingting Mu**[1]
[1]University of Manchester
alessio.sarullo@manchester.ac.uk, tingting.mu@manchester.ac.uk

Abstract

We propose a new approach for Zero-Shot Human-Object Interaction Recognition in the challenging setting that involves interactions with unseen actions (as opposed to just unseen combinations of seen actions and objects). Our approach makes use of common-sense knowledge in the form of a graph that models affordance relations between actions and objects, i.e., whether an action can be performed on the given object or not. We propose a loss function with the aim of distilling the knowledge contained in the graph into the model, while also using the graph to regularise learnt representations by imposing a local structure on the latent space. We evaluate our approach on several datasets (including the popular HICO and HICO-DET) and show that it outperforms the current state of the art.

1 Introduction

Human-Object Interaction (HOI) Recognition is the task of identifying how people interact with the surrounding objects from the visual appearance of the scene and it is of paramount importance to understand the content of an image. It consists of producing a set of $\langle human, action, object \rangle$ triplets for the input image, providing a concise representation of the image semantics that can be used in higher-level tasks like Image Captioning (Anderson et al. 2018) or Human-Robot Interaction (Fang, Yuan, and Magnenat-Thalmann 2018).

One of the greatest difficulties when dealing with visual relations is that the number of possible triplets increases multiplicatively in the cardinality of the human, action and object spaces. Due to the practical challenges of building a dataset, it is common for only a subset of all possible interactions to be annotated, while a large number remains unlabelled; for instance, HICO (Chao et al. 2015) contains only 600 interactions out of the 9360 possible pairs (among the 9360-600=8760 unlabelled interactions, some are invalid like $\langle eating, bottle \rangle$, while some are valid but missing like $\langle carrying, knife \rangle$). This is why more and more approaches are focusing on Zero-Shot Learning (ZSL) for HOI Recognition (Shen et al. 2018; Kato, Li, and Gupta 2018; Peyre et al. 2019; Bansal et al. 2020). ZSL aims to alleviate the problems caused by the combinatorial growth of the number of possible interactions by allowing models to make predictions about previously unseen interactions.

Figure 1: Left to right: ⟨*eating*, *sandwich*⟩, ⟨*eating*, *pizza*⟩, ⟨*cooking*, *pizza*⟩. Both pairs of objects (*pizza* and *sandwich*) and actions (*eating* and *cooking*) are semantically similar, yet images that share an action look more similar than images that share an object.

We focus on actions, as they play a more significant role than objects in defining an interaction: several studies in Psychology (Norman 2013), Neurobiology (Chao and Martin 2000) and Computer Vision (Bansal et al. 2020) show that objects can be categorised and recognised based on their affordances, making the semantics of an object defined in term of actions. We illustrate this intuition via some visual examples provided in Figure 1. For this reason, we follow a challenging zero-shot setting that consists of predicting interactions containing unseen actions, instead of only new combinations of seen object and action classes.

Our model uses a Graph Convolutional Network (GCN) (Kipf and Welling 2017) to learn unseen classes in a semi-supervised manner (Wang, Ye, and Gupta 2018). The graph's connectivity determines how nodes are linked to each other and thus how information is aggregated in the learnt representations. We make use of a particular type of common-sense knowledge graph called an *affordance graph*, that is, a graph whose edges model *affordances* (Gibson 2014; Norman 2013): action-object pairs $\langle a, o \rangle$ where a can be performed on o (e.g., ⟨*hold*, *apple*⟩, because apples can be held). Such a graph enables the model to learn what interactions are affordable regardless of whether they appear in the training set, allowing it to perform ZSL.

The focus of this paper is to propose a new training objective function that aims to improve the representations learnt by the model. More specifically, the proposed objective function enhances the loss used by state-of-the-art approaches in two ways. First, it effectively distils action affordance in the unseen class representations by making use of

relations from the affordance graph to train unseen actions in a weakly-supervised way. As a result, the model learns to distinguish which unseen actions can be performed on a given object and which ones cannot. Second, it imposes a local structure on the latent space through a regulariser that clusters unseen class representations together with similar classes according to the affordance graph. Qualitative and quantitative results demonstrate that our model (shown in Figure 2) learns representations that are effective at differentiating actions based on affordances, and it outperforms the current state of the art on HICO, VG-HOI and COCO-a.

2 Related Work

Knowledge Usage in HOI Recognition Many works have been proposed to perform HOI Detection in recent years, the most similar to ours being the ones that make use of common-sense knowledge (Kato, Li, and Gupta 2018; Peyre et al. 2019; Bansal et al. 2020; Xu et al. 2019). In (Bansal et al. 2020) a language component is used to identify functionally similar objects, effectively augmenting the training data with new interaction instances. In the other works, the common-sense knowledge is used to obtain class representations, which are used for prediction. These representations come from word embeddings that are mapped through functions implemented as a Multi-Layer Perceptron (MLP) (Peyre et al. 2019) or a GCN (Kato, Li, and Gupta 2018; Xu et al. 2019). An important difference between these models lies in what representations are computed: while in (Peyre et al. 2019) action, object and interactions classes are all considered and the respective scores combined in a compositional way, in the other methods only representations for actions (Xu et al. 2019) or interactions (Kato, Li, and Gupta 2018) are used for prediction. Our approach is similar to (Peyre et al. 2019) regarding the compositional model and to (Kato, Li, and Gupta 2018; Xu et al. 2019) in the utilisation of external knowledge to build the graph used by the GCN, but differs from all of the above in the way we use the graph at training time to regularise action representation and to distil affordances into the model.

Zero-Shot Learning The growing field of ZSL primarily aims to overcome the difficulties of dealing with a non-exhaustively annotated dataset. A common framework to perform ZSL (Zhang, Xiang, and Gong 2017; Wang, Ye, and Gupta 2018) is to exploit some kind of common-sense knowledge to transfer to unseen classes what has been learnt about seen ones in a semi-supervised way. Representations are learnt for both classes and instances and compared through a similarity function to predict output probabilities. The model is trained by feeding the output for seen classes into a loss function such as least squares (Zhang, Xiang, and Gong 2017) or cross entropy (Wang, Ye, and Gupta 2018).

An interesting method to learn better representations is to add a regularisation loss (Mishra et al. 2018; Schönfeld et al. 2018). (Mishra et al. 2018) map label embeddings into the visual space, adding a reconstruction loss to make sure that the inverse transformation is also possible and thus the visual projection preserves semantics. A different technique is used in (Schönfeld et al. 2018), where a cross-reconstruction loss between images and labels is added in order to "pull together" representations of the same class from the two different sources (image and labels). Inspired by these works, we formulate a different regularisation loss that uses the affordance graph and is thus better suited to our goal of modelling action affordance.

A few recent approaches tackle ZSL in HOI Recognition/Detection (Shen et al. 2018; Kato, Li, and Gupta 2018; Peyre et al. 2019; Bansal et al. 2020). We compare our results to the works that considers unseen actions (Kato, Li, and Gupta 2018; Peyre et al. 2019), as they are the most closely related to ours.

3 Notation and Problem Statement

Let us denote the ordered set of objects and actions by $\mathcal{O}$ and $\mathcal{A}$, respectively. We will denote the elements of these sets by the corresponding lowercase letter or sometimes by the index only (for example we will write $a_k \in \mathcal{A}$ or $k \in \mathcal{A}$).

Our dataset is denoted by $\mathcal{D} = \{(I_i, \mathbf{T}_i)\}_{i=1}^{M}$. Here, I_i is the i-th image and $\mathbf{T}_i \in \{0,1\}^{|\mathcal{O}|\times|\mathcal{A}|}$ is its label matrix, with its jk-th element t_{ijk} being 1 if and only if example i is annotated with interaction $\langle a_k, o_j \rangle$ (note that an image can have multiple labels). Under the considered Zero-Shot Learning setting, we assume that there are no available visual examples for some objects and actions. This is equivalent to omit the corresponding labels from all images during training, although the affected images might still be annotated with other labels that have not been omitted. The omitted class set will be denoted with $\mathcal{U}$ (they are *unseen*), while $\mathcal{S}$ is the set of *seen* (i.e., trained-on) classes. Therefore, we have $\mathcal{O} = \mathcal{S}^O \cup \mathcal{U}^O$ and $\mathcal{A} = \mathcal{S}^A \cup \mathcal{U}^A$. Note that seen and unseen classes do not intersect, i.e., $\mathcal{S}^q \cap \mathcal{U}^q = \emptyset \; \forall q \in \{O, A\}$. The task is to learn a model that is able to predict any interaction $\langle a_k, o_j \rangle$, even when $o_j \in \mathcal{U}^O$ or $a_k \in \mathcal{U}^A$ (that is, when either or both object and action are unseen).

4 Proposed Method

4.1 Affordance Graph

The main motivation of this work is to improve zero-shot interaction recognition by using structured common-sense knowledge, which is expressed in the form of an *affordance graph*. We define it as a graph $\mathcal{G} = \langle \mathcal{V}, \mathcal{E} \rangle$ whose nodes $\mathcal{V}$ are objects and actions and edges $\mathcal{E}$ represent *affordances*: object node o_j is connected to action node a_k only if a_k can be performed on o_j, i.e., $\langle a_k, o_j \rangle$ constitutes a valid interaction. For example, *eat* and *apple* will be connected, but *eat* and *fork* will not because people cannot eat forks. This graph is undirected and bipartite: all links are symmetric and there are no connections between object nodes, nor between action nodes. We construct the affordance graph by mining interactions from external sources, to simulate a real-world scenario where no interaction information regarding unseen classes is available. Details about the construction process will be provided in Section 5.3.

4.2 Model Architecture

Overview Our model takes as input an image I, which is fed into a Convolutional Neural Network (CNN) to pro-

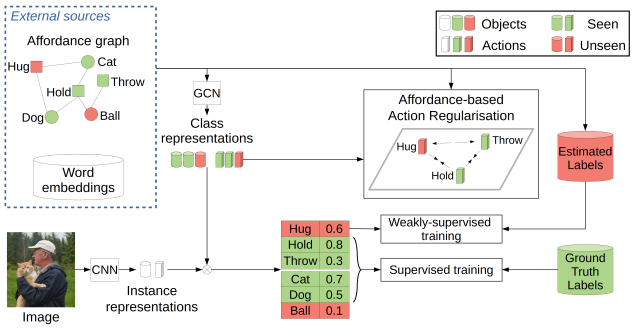


Figure 2: Overview of the proposed model for HOI Recognition. $\otimes$ indicates dot product. Best viewed in colour.

duce image-level visual features $\boldsymbol{v} = f_{CNN}(I)$. These features are fed into two identically structured modules indexed by variable q, one for objects ($q = O$) and one for actions ($q = A$). Specifically, for each module we compute a d-dimensional representation $\boldsymbol{x}^q = f_1^q(\boldsymbol{v})$ through a non-linear mapping f_1^q (e.g., an MLP). Vector $\boldsymbol{x}^q$ is compared to a set of d-dimensional class representations $\mathbf{Z}^q = [\boldsymbol{z}_1^q \mid \ldots \mid \boldsymbol{z}_{|\mathcal{S}^q \cup \mathcal{U}^q|}^q]$ via inner product, and the similarity scores are fed into the sigmoid function $\sigma(x) = 1/(1+e^{-x})$ to output probabilities $\boldsymbol{y}^q = \sigma(\mathbf{Z}^q \boldsymbol{x})$. We will now describe how to compute class representations $\mathbf{Z}^q$.

Class Representations We use a GCN to train unseen class representations in a semi-supervised way, effectively embedding the affordance relations contained in the graph into the learnt representations. We also incorporate additional semantic information computed from word embeddings, but, differently from previous work, we do not use them to initialise GCN's input embeddings. The reason is that the affordance graph and word embeddings provide different types of semantics: the former aims to capture affordances, while the latter co-occurrence statistics. As a result, for instance, "eat" and "drink" are distant according to affordances while close according to word embeddings, which results in a mismatch in action similarity that brings down the performance. However, co-occurrence semantics carried by word embeddings are useful for objects (e.g., "pizza" and "sandwich" have high similarity according to word embeddings, and indeed are both objects that can be eaten), so we use word embeddings to enrich the objects representations produced by the GCN. The final class representations $\mathbf{Z}^O \in \mathbb{R}^{|\mathcal{O}| \times d}$ and $\mathbf{Z}^A \in \mathbb{R}^{|\mathcal{A}| \times d}$ are

$$\mathbf{Z}^O = (\mathbf{Z}_{GCN})_{\mathcal{O},:} + f_2(\mathbf{W}^O) \quad (1)$$

$$\mathbf{Z}^A = (\mathbf{Z}_{GCN})_{\mathcal{A},:} \quad (2)$$

$$\mathbf{Z}_{GCN} = f_3\left(f_{GCN}\left(\mathbf{Z}_0\right)\right) , \quad (3)$$

where $(\mathbf{Z}_{GCN})_{\mathcal{O},:}$ and $(\mathbf{Z}_{GCN})_{\mathcal{A},:}$ denote the rows of $\mathbf{Z}_{GCN}$ corresponding to object and action classes (respectively), f_2 and f_3 are non-linear functions (e.g., MLPs), $\mathbf{W}^O \in \mathbb{R}^{|\mathcal{O}| \times d'}$ are d'-dimensional word embeddings and GCN's input embeddings $\mathbf{Z}_0 \in \mathbb{R}^{(|\mathcal{O}|+|\mathcal{A}|) \times d_0}$ are randomly initialised. We use $\mathbf{Z}^O$ and $\mathbf{Z}^A$ to predict class probabilities $\boldsymbol{y}^q = \sigma(\mathbf{Z}^q \boldsymbol{x})$ for $q \in \{O, A\}$. Note that these representations (and the corresponding probabilities) are computed for both seen and unseen classes. We use $\mathbf{Z}^{q-\mathcal{S}} = \mathbf{Z}^q_{\mathcal{S}^q,:}$ and $\mathbf{Z}^{q-\mathcal{U}} = \mathbf{Z}^q_{\mathcal{U}^q,:}$ to denote the sub-matrices of $\mathbf{Z}^q$ that only contain rows for seen or unseen classes, respectively.

Inference At inference time a score is assigned to every interaction by multiplying object and action scores, producing a matrix $\mathbf{Y} = \boldsymbol{y}^O \left(\boldsymbol{y}^A\right)^T \in [0,1]^{|\mathcal{O}| \times |\mathcal{A}|}$ whose element y_{jk} constitutes the probability for interaction $\langle a_k, o_j \rangle$.

4.3 Training

Our model is trained by minimising the following composite loss function, which is designed to optimise all network parameters $\boldsymbol{\Theta}$ (which include weights for MLPs and GCN, as well as GCN's initial representations $\mathbf{Z}_0$) through variables $\boldsymbol{y}^O$, $\boldsymbol{y}^A$ and $\mathbf{Z}^A$:

$$\min_{\boldsymbol{\Theta}} \sum_{i=1}^{M} \left[\ell\left(\boldsymbol{y}_i^{O-\mathcal{S}}, \boldsymbol{t}_i^O, \mathcal{S}^O\right) + \ell\left(\boldsymbol{y}_i^{A-\mathcal{S}}, \boldsymbol{t}_i^A, \mathcal{S}^A\right)\right] + \quad (4)$$
$$\sum_{i=1}^{M} \lambda\, \ell\left(\boldsymbol{y}_i^{A-\mathcal{U}}, \hat{\boldsymbol{t}}_i^A, \mathcal{U}^A\right) + \rho\, \mathcal{L}_{REG}\left(\mathbf{Z}^A\right) ,$$

where λ and ρ are hyperparameters that regulate the contribution of their respective terms, label vectors $\boldsymbol{t}_i^O \in \{0,1\}^{|\mathcal{O}|}$ and $\boldsymbol{t}_i^A \in \{0,1\}^{|\mathcal{A}|}$ are obtained from matrix $\mathbf{T}_i$ according to $t_{ij}^O = \max_{k \in \mathcal{A}} t_{ijk}$ and $t_{ik}^A = \max_{j \in \mathcal{O}} t_{ijk}$, and ℓ is the binary cross entropy loss:

$$\ell(\boldsymbol{y}, \boldsymbol{t}, \mathcal{J}) = \sum_{j \in \mathcal{J}} \left[t_j \log y_j + (1 - t_j) \log(1 - y_j)\right] , \quad (5)$$

where $\boldsymbol{y}$ are outputs, $\boldsymbol{t}$ target labels and $\mathcal{J}$ a set of indices. We also add L_2-regularisation to $\boldsymbol{\Theta}$ to prevent overfitting (not shown in Equation (4)).

The first term of Equation (4) implement a standard training loss, which uses ground truth labels to reward pairing instances with the corresponding seen classes and to penalise assigning the wrong class. The second term aims to train unseen actions in the same way. However, since ground truth labels are not available for unseen actions, we adopt a weakly-supervised approach and estimate labels $\hat{\boldsymbol{t}}^A$ as:

$$\hat{t}_k^A = \max_{j \in \mathcal{S}^O} m_{jk} s_{jk} \qquad \forall k \in \mathcal{U}^A \quad (6)$$

$$s_{jk} = \frac{1}{\sum_{h \in \mathcal{S}^A} t_{jh}} \sum_{h \in \mathcal{S}^A} t_{jh} \left[\boldsymbol{w}_h^T \boldsymbol{w}_k\right]_+ , \quad (7)$$

where $[x]_+ = \max(x, 0)$, $\mathbf{M} \in \{0,1\}^{|\mathcal{O}| \times |\mathcal{A}|}$ is the graph adjacency matrix[1] and $\boldsymbol{w}_k$ is the word embedding for the k-th action. Equation (7) computes a score that determines how likely unseen action k describes an image containing object j. This score is *not* binary, but rather a real value in $[0, 1]$. This is needed because binary estimated labels would incur the risk of introducing noise, since we cannot know which of

[1] $\mathbf{M}$ needs not be a square matrix because the graph is bipartite.

the affordable unseen actions are actually depicted in a particular image. Word embeddings are used to assign a score based on the similarity with labelled seen actions (which are compatible with object j, since they come from the ground truth) through the positive inner product $\left[\boldsymbol{w}_h^T \boldsymbol{w}_k\right]_+$, so that unseen actions similar to shown seen ones will be assigned a higher score: if $o_j =$ *person* and *hug* is a labelled seen action, *kiss* and *greet* are better unseen candidates than *teach*. Action affordance is distilled into the model according to Equation (6): score s_{jk} contributes to $\hat{t}_k^A$ only if $m_{jk} = 1$, that is, only if $\langle a_k, o_j \rangle$ is an affordable action. Since an image may contain multiple objects, the maximum score over objects is taken according to the Multiple Instance Learning framework (Mallya and Lazebnik 2016).

Additionally, we use the affordance graph as a regulariser for action classes, with the goal of learning better representations by inducing a structure onto the latent space based on affordances. Specifically, we want to group *functionally* similar actions, that is, actions that can be performed on the same objects. We use the following ranking margin loss:

$$\begin{aligned} \mathcal{L}_{REG}\left(\mathbf{Z}^A\right) &= \sum_{i \in \mathcal{U}^A} \sum_{j \in \mathcal{N}(i)} \sum_{k \notin \mathcal{N}(i)} \left[\gamma - c_{ij} + c_{ik}\right]_+ \\ c_{ij} &= \frac{\boldsymbol{z}_i^T \boldsymbol{z}_j}{||\boldsymbol{z}_i|| ||\boldsymbol{z}_j||} \quad \forall i, j \in \mathcal{A}\,, \end{aligned} \tag{8}$$

where $\gamma \in \mathbb{R}$ is the margin, c_{ij} is the cosine similarity between the i-th and j-th columns of $\mathbf{Z}^A$ ($\boldsymbol{z}_i$ and $\boldsymbol{z}_j$), and $\mathcal{N}(i)$ denotes the set of actions that are functionally similar to action node a_i, i.e., actions at distance 2 from a_i in the affordance graph (with nodes at distance 1 being objects).

5 Experiments

We compare our results to methods reported in (Kato, Li, and Gupta 2018) on HICO and VG-HOI, and in (Peyre et al. 2019) on COCO-a. The model has been implemented in Python 3.6 using PyTorch v0.4.1. Experiments have been run on a single NVIDIA GeForce GTX TITAN X GPU on a server with an Intel(R) Core(TM) i7-5930K CPU and 64GB of RAM running CentOS Linux 7.

5.1 Datasets

HICO and HICO-DET The HICO dataset (Chao et al. 2015) and its bounding-box-annotated variant HICO-DET (Chao et al. 2018) comprise 47k images, 80 object classes, and 117 action classes, including a null one. They are annotated with 600 interactions and each image may belong to more than one interaction class. We follow the predefined train/test split of 38,116/9,658 images, and we randomly sample 10% of the training set for validation in every run. In our Recognition experiment we follow (Kato, Li, and Gupta 2018), excluding the null action during training and testing and thus restricting the dataset to 116 actions and 520 HOIs.

VG-HOI VG-HOI (Kato, Li, and Gupta 2018) is a HOI dataset built out of Visual Genome (Krishna et al. 2017b). It comprises 10,799 train images and 4251 test images, for a total of 15,050. We use 10% of the training set for validation. There are 1392 objects, 495 actions and 6643 interactions, but for testing only the 532 that have at least 10 instances are used. The large number of classes and the low number of examples make this dataset extremely challenging.

COCO-a COCO-a (Ronchi and Perona 2015) contains 4413 images annotated with 145 action classes and 80 object classes (same as HICO), for a total of 1681 interactions. We use it as an evaluation dataset for our model trained on HICO-DET, following the challenging setting used in (Peyre et al. 2019): there are 1474 unseen interactions, 1048 of which involve an unseen action.

5.2 Evaluation

We use the standard mean Average Precision (mAP) as evaluation metric, reporting it as a percentage. We train every model multiple times (10 for HICO and COCO-a, 5 for VG-HOI and HICO-DET), reporting the average result on the test set. We run Student's t-tests against current state-of-the-art results and all reported improvements are statistically significant at the 99% confidence interval.

5.3 Affordance Graph Construction

To build the affordance graph, we mine interactions from external knowledge bases and add them to the ones that can be found in the training set. Specifically, we use four external sources: Visual Genome (Krishna et al. 2017b) (except for VG-HOI), ActivityNet Captions (Krishna et al. 2017a), imSitu (Yatskar, Zettlemoyer, and Farhadi 2016) and HCVRD (Zhuang et al. 2018). The former three contain image or video captions that we parse into action-object pairs using NLTK (Bird, Klein, and Loper 2009) and the dependency parser from AllenNLP (Gardner et al. 2018). On the other hand, HCVRD is annotated with triplets in the form $\langle subject, predicate, object \rangle$. In both cases, we select the ones where $subject$ is a person (for HCVRD), $predicate$ or $action \in \mathcal{A}$, and $object \in \mathcal{O}$. This results in graphs containing 80/111/1350 unique object/action/HOIs (respectively) for HICO, 1106/231/8713 for VG-HOI and 80/189/1755 for HICO-DET+COCO-A.

5.4 Experimental Setting for HOI Recognition

Compared Models We report the performance of four variants of our model: using none ($\lambda, \rho = 0$), either (λ or $\rho > 0$), or both (λ and $\rho > 0$) of the proposed losses.

The most similar method to ours is (Kato, Li, and Gupta 2018), which performs zero-shot learning on both action and objects. We compare our models to their best results, which are denoted by "GCNCL". We also report other competitive methods from (Kato, Li, and Gupta 2018), namely Semantic Embedding Space (SES) (Xu, Hospedales, and Gong 2015) and Triplet Siamese. We refer the reader to the corresponding papers for more details.

Zero-Shot Settings In order to make a fair comparison, we use the same seen/unseen splits as Task 2 from (Kato, Li, and Gupta 2018): the training set is made of 49 objects and 53 actions for HICO and 554 objects and 198 actions for VG-HOI. At test time all classes are included, following the Generalised Zero-Shot Learning setting.

Method	All	Unseen only
Triplet Siamese	10.38	7.76
SES	11.69	7.19
GCNCL+NV+A	11.94	7.50
Ours	13.02	6.64
Ours, $\lambda = 1$	14.95	9.78
Ours, $\rho = 10$	13.36	6.99
Ours, $\lambda = 1, \rho = 10$	**15.14**	**9.92**

Table 1: Results on HICO.

Method	All	Unseen only
Triplet Siamese	2.55	1.67
SES	2.07	0.96
GCNCL-I+A	4.00	2.63
GCNCL+A	4.07	2.44
Ours	4.83	3.52
Ours, $\lambda = 0.1$	4.99	3.85
Ours, $\rho = 100$	**5.25**	**4.13**
Ours, $\lambda = 0.1, \rho = 100$	4.94	3.64

Table 2: Results on VG-HOI.

Implementation Details We use a pre-trained ResNet-152 as image feature extractor, for a fair comparison to (Kato, Li, and Gupta 2018). Functions f_1, f_2 and f_3 are implemented by two fully-connected layers with output dimensions both equal to 1024, with ReLU non-linearity. After the non-linearity we add Dropout (Hinton et al. 2012) (at a 0.5 rate) for f_1 and f_3, but not for f_2, as suggested in (Peyre et al. 2019). We use Glorot initialisation (Glorot and Bengio 2010) to initialise the optimisation parameters $\mathbf{\Theta}$. Our GCN comprises two convolutional layers with output dimension 1024, the first of which is equipped with ReLU and Dropout (0.5 rate).

We keep the margin parameter γ in Equation (8) fixed at 0.3, whereas we experiment with different values of ρ and λ for the two datasets. The best ones (according to validation results) are the ones shown in the respective tables.

We use GloVe (Pennington, Socher, and Manning 2014) for our word embeddings. More specifically, we use the 300-dimensional embeddings trained on Gigaword and Wikipedia[2] and we normalise them. For compound words, we take the average of the components.

Finally, we train our model using minibatch Stochastic Gradient Descent (SGD) with momentum. We use a fixed learning rate of 0.001 and set the momentum and weight decay coefficients to 0.9 and $5 \cdot 10^{-4}$, respectively. We train our model for a maximum of 100 epochs on HICO and 150 on VG-HOI, with early stopping based on validation accuracy. We use a batch size of 64.

5.5 Results for HOI Recognition

Results on HICO Our results are summarised in Table 1. We see that our baseline model already compares very favourably to all the existing approaches, and adding either or both of the proposed losses upgrades our baseline's performance considerably. The best performing model, obtained with $\lambda = 1$ and $\rho = 10$, gains more than 3% over the current state of the art (GCNCL+NV+A) for the whole test set and around 2.4% for unseen classes only. This corresponds to sizeable ∼27% and ∼32% relative increases. It is worth mentioning that the graph building process results in 68 missing interactions out of HICO's 520, since they cannot be mined from our external sources. Despite this, no object is completely isolated in the affordance graph, whereas only 5 actions are (*hop_on*, *hunt*, *lose*, *stab*, *toast*). Most of these actions (namely *hunt*, *lose*, *stab* and *toast*) are too niche to be found in the other sources, and in fact even in HICO they only appear in one interaction each. On the other hand, *hop_on* can be found, but not with the meaning of "jumping on a ride" it has in HICO (and thus it is not paired with the same objects). Nonetheless, our model still performs very well, possibly due to the fact that additional interactions are added and they contribute to meaningful representations being learnt, even though they do not appear in HICO.

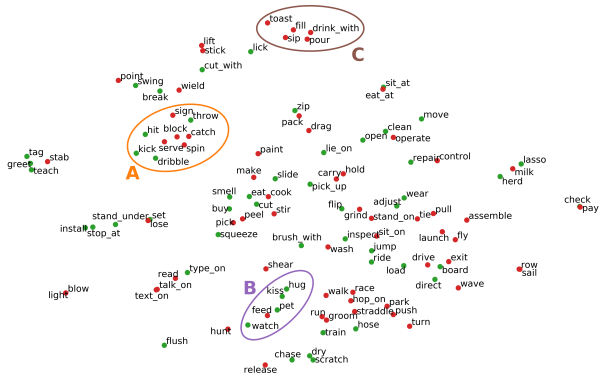


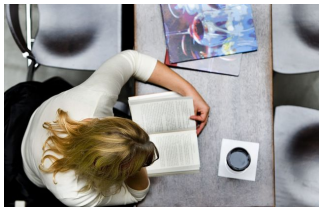
Figure 3: Visualisation of action class representations. Green dots represent seen actions and red dots unseen ones. Some clusters are highlighted: (*A*) sports actions (e.g., *catch*, *throw*), (*B*) actions regarding pets (e.g., *pet*, *feed*) and (*C*) action involving cups or glasses (e.g., *sip*, *pour*).

Results on VG-HOI Results are reported in Table 2. Our baseline is better than previous models, GCNCL+A in particular: ∼19% relative gain for all classes and almost 40% for unseen categories. Adding the proposed losses, especially the regularisation term, further improves performance.

Qualitative Results on HICO We show some predictions on HICO's test set examples in Figure 4, demonstrating that our model is able to correctly predict several previously unseen actions. We also show the representation space in Figure 3 using t-SNE (Maaten and Hinton 2008) on a model trained with both proposed losses ($\lambda, \rho > 0$). Some clusters are clearly identifiable, such as cluster A, which contains actions like *catch*, *throw* or *spin* that can be performed on small sport items like *sports_ball* or *frisbee*. This shows that the proposed approach is effective in grouping actions based on their affordance.

[2] Available at https://nlp.stanford.edu/projects/glove.

(a)
✓ *eat_at* dining_table
✓ sit_at dining_table



(b)
✓ *hold* book
✓ open book

(c)
✓ *carry* umbrella
✓ *hold* umbrella
✗ stand_under umbrella

(d)
✓ *pull* tie
✓ *tie* tie
✓ wear tie

Figure 4: Some predictions of our best model on HICO. Marks indicate whether the prediction matches the ground truth ✓ or not ✗ . Actions in *italic* are unseen.

Method	All
(Shen et al. 2018)	6.46
(Chao et al. 2018)	7.81
InteractNet (Gkioxari et al. 2018)	9.94
GPNN (Qi et al. 2018)	13.11
(Xu et al. 2019)	14.70
iCAN (Gao, Zou, and Huang 2018)	14.84
(Song et al. 2020)	15.27
(Wang et al. 2019)	16.24
No-frills (Gupta, Schwing, and Hoiem 2019)	17.18
(Li et al. 2019)	17.22
RPNN (Zhou and Chi 2019)	17.35
PMFNet (Wan et al. 2019)	17.46
(Peyre et al. 2019)	19.40
(Wang et al. 2020)	19.56
PPDM (Liao et al. 2020)	21.73
(Bansal et al. 2020)	**21.96**
Ours	18.74

Table 3: Results on HICO-DET in a fully supervised setting.

Method	All	Unseen
Ours	10.45	8.27
Ours, $\lambda = 0.1, \rho = 0.1$	**11.03**	**9.80**

Table 4: Baseline for ZS HOI Detection on HICO-DET.

5.6 Experimental Setup for HOI Detection

We describe the experimental setup for the HOI Detection experiments in the following. All unmentioned settings are identical to the ones described in Section 5.4 for the HOI Recognition experiments.

Zero-Shot Settings The focus of this experiments is Zero-Shot HOI Detection when there are unseen actions. On HICO-DET our training set contains the same unseen actions as the recognition experiment ($\sim$50% of the total, as described in Section 5.4), while on COCO-a there are 114 unseen actions, corresponding to 1048 unseen interactions. In both cases there are no unseen object classes, therefore we rely purely on the object scores provided by a pre-trained object detector (we use Mask R-CNN (He et al. 2017) with ResNet-50 as backbone).

Architectural Changes We adapted our model to deal with image regions instead of whole images. The object detector provides visual features $\boldsymbol{h}_i^{(h)}$, $\boldsymbol{h}_i^{(o)}$ and $\boldsymbol{h}_i^{(a)}$ for every person, object and region that represents a possible interaction (union of person and object bounding boxes), respectively. We compute the interaction representation as

$$\boldsymbol{x}_i = f_1([\boldsymbol{h}_i^{(h)}, \boldsymbol{s}_i^{(h)}]) + f_1([\boldsymbol{h}_i^{(o)}, \boldsymbol{s}_i^{(o)}]) + f_1(\boldsymbol{h}_i^{(a)}) \,, \quad (9)$$

where f_1 is defined as usual as an MLP, $[\cdot, \cdot]$ indicates concatenation and $\boldsymbol{s}^{(\cdot)}$ are object classification score vectors returned by the object detector.

Training Procedure During training, we keep all detected object bounding boxes and add the ground-truth ones that do not have any match, i.e., there is no detected box whose intersection-over-union (IoU) is greater than 0.5. We keep as positive interaction examples all human-object pairs whose subject and object are correctly classified and overlap with the subject/object (respectively) of a ground-truth interaction (again, the threshold for IoU is 0.5). Among the pairs that are not positive interactions, we sample negative ones, at a rate of 3 negatives per positive – this is a widely used ratio, see for example (Peyre et al. 2019). At inference time, we only keep human candidates with a confidence score greater than 0.7 and threshold object ones at 0.3. Every possible human-object pair in the image is considered as a candidate interaction and classified by the model.

When using the regularisation loss $\mathcal{L}_{REG}$ on HICO-DET, we found it beneficial to only enable it (that is, set $\rho > 0$) after the first 5 epochs. This allows the model to learn class representations first, and only later regularise them.

The model is trained with minibatch Stochastic Gradient Descent (SGD) with a learning rate of 0.001 and weight decay coefficient of $5 \cdot 10^{-4}$. We train our model for a maximum of 10 epochs when evaluating on HICO-DET and 20 on COCO-a, and use a batch size of 64.

5.7 Results for HOI Detection

HICO-DET While there are works on Zero-Shot Learning on HICO-DET for interactions (Shen et al. 2018; Peyre et al. 2019) and objects (Bansal et al. 2020), no previous approach has dealt with zero-shot actions (to the best of our knowledge). We provide in Table 4 a baseline for future reference. We also show in Table 3 how our approach compares against other methods in a fully supervised setting as a reference, where we can see that there is a noticeable increase

Method	Unseen HOIs All	Unseen HOIs With unseen actions
(Peyre et al. 2019) (best)	6.9	7.3
Ours	9.69	10.88
Ours, $\lambda = 0.1, \rho = 10$	**10.13**	**11.48**

Table 5: Results on COCO-a.

in mAP with respect to most methods in the literature. It is worth mentioning that some of the techniques that likely contribute to the outstanding results of the top three methods, such as fine-tuning the object detector on HICO-DET (Bansal *et al.*) or following a more intensive training regime while fine-tuning the feature extractor (50 epochs on 5 GPUs for (Wang et al. 2020), 110 epochs on 8 GPUs for PPDM), are applicable to our model as well – in fact, Bansal *et al.* report that their method only achieves 16.96% mAP without such fine-tuning. We leave this for future work.

COCO-a In Table 5 we show our results on COCO-a, reporting our baseline plus the best performing model against a state-of-the-art approach. Our approach performs much better, gaining around 2.7 points for all unseen interactions ($\sim$40% relative gain) and 3.7 points when dealing with interactions involving unseen actions (about 50% relative gain). Performance improve even further when setting λ and ρ to non-zero values, once again showing the effectiveness of the proposed losses.

6 Conclusion

We have proposed an effective approach that uses structured common-sense knowledge in the form of an affordance graph to improve Zero-Shot Human-Object Interaction Recognition. The proposed model learns regularised representations of unseen classes in a weakly supervised way using labels which are estimated through the affordance graph. Our method is able to predict unseen interactions in very challenging settings, in which the majority of actions are unseen during training. We evaluate our results on several datasets (including standard benchmarks like HICO and HICO-DET) and show that our approach performs significantly better than the current state of the art.

References

Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; and Zhang, L. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6077–6086.

Bansal, A.; Rambhatla, S. S.; Shrivastava, A.; and Chellappa, R. 2020. Detecting Human-Object Interactions via Functional Generalization. In *AAAI*, 10460–10469.

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

Chao, L. L.; and Martin, A. 2000. Representation of Manipulable Man-Made Objects in the Dorsal Stream. *NeuroImage* 12(4): 478–484. ISSN 10538119. doi:10.1006/nimg.2000.0635.

Chao, Y.-W.; Liu, Y.; Liu, X.; Zeng, H.; and Deng, J. 2018. Learning to Detect Human-Object Interactions. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 381–389. IEEE.

Chao, Y.-W.; Wang, Z.; He, Y.; Wang, J.; and Deng, J. 2015. Hico: A benchmark for recognizing human-object interactions in images. In *Proceedings of the IEEE International Conference on Computer Vision*, 1017–1025.

Fang, Z.; Yuan, J.; and Magnenat-Thalmann, N. 2018. Understanding Human-Object Interaction in RGB-D videos for Human Robot Interaction. In *Proceedings of Computer Graphics International 2018*, 163–167.

Gao, C.; Zou, Y.; and Huang, J.-B. 2018. iCAN: Instance-Centric Attention Network for Human-Object Interaction Detection. In *British Machine Vision Conference*.

Gardner, M.; Grus, J.; Neumann, M.; Tafjord, O.; Dasigi, P.; Liu, N.; Peters, M.; Schmitz, M.; and Zettlemoyer, L. 2018. Allennlp: A deep semantic natural language processing platform. *arXiv preprint arXiv:1803.07640* .

Gibson, J. J. 2014. *The ecological approach to visual perception: classic edition*. Psychology Press.

Gkioxari, G.; Girshick, R.; Dollár, P.; and He, K. 2018. Detecting and recognizing human-object interactions. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8359–8367. IEEE.

Glorot, X.; and Bengio, Y. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, 249–256.

Gupta, T.; Schwing, A.; and Hoiem, D. 2019. No-Frills Human-Object Interaction Detection: Factorization, Layout Encodings, and Training Techniques. In *Proceedings of the IEEE International Conference on Computer Vision*, 9677–9685.

He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. 2017. Mask r-cnn. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, 2980–2988. IEEE.

Hinton, G. E.; Srivastava, N.; Krizhevsky, A.; Sutskever, I.; and Salakhutdinov, R. R. 2012. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580* URL https://arxiv.org/abs/1207.0580.

Kato, K.; Li, Y.; and Gupta, A. 2018. Compositional learning for human object interaction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 234–251.

Kipf, T. N.; and Welling, M. 2017. Semi-supervised classification with graph convolutional networks. *International Conference on Learning Representations (ICLR)* .

Krishna, R.; Hata, K.; Ren, F.; Fei-Fei, L.; and Carlos Niebles, J. 2017a. Dense-captioning events in videos. In *Proceedings of the IEEE international conference on computer vision*, 706–715.

Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.-J.; Shamma, D. A.; Bernstein, M. S.; and Fei-Fei, L. 2017b. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. *International Journal of Computer Vision* 123(1): 32–73. ISSN 0920-5691, 1573-1405. doi:10.1007/s11263-016-0981-7. URL http://link.springer.com/10.1007/s11263-016-0981-7.

Li, Y.-L.; Zhou, S.; Huang, X.; Xu, L.; Ma, Z.; Fang, H.-S.; Wang, Y.; and Lu, C. 2019. Transferable Interactiveness Knowledge for Human-Object Interaction Detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3585–3594.

Liao, Y.; Liu, S.; Wang, F.; Chen, Y.; Qian, C.; and Feng, J. 2020. PPDM: Parallel Point Detection and Matching for Real-Time Human-Object Interaction Detection. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 479–487. Seattle, WA, USA: IEEE. ISBN 978-1-72817-168-5. doi:10.1109/CVPR42600.2020.00056. URL https://ieeexplore.ieee.org/document/9156683/.

Maaten, L. v. d.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9(Nov): 2579–2605.

Mallya, A.; and Lazebnik, S. 2016. Learning models for actions and person-object interactions with transfer to question answering. In *European Conference on Computer Vision*, 414–428. Springer.

Mishra, A.; Verma, V. K.; Reddy, M. S. K.; Arulkumar, S.; Rai, P.; and Mittal, A. 2018. A generative approach to zero-shot and few-shot action recognition. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 372–380. IEEE.

Norman, D. 2013. *The design of everyday things: Revised and expanded edition*. Basic books.

Pennington, J.; Socher, R.; and Manning, C. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 1532–1543.

Peyre, J.; Laptev, I.; Schmid, C.; and Sivic, J. 2019. Detecting Unseen Visual Relations Using Analogies. In *Proceedings of the IEEE International Conference on Computer Vision*, 1981–1990.

Qi, S.; Wang, W.; Jia, B.; Shen, J.; and Zhu, S.-C. 2018. Learning human-object interactions by graph parsing neural networks. In *European Conference on Computer Vision*, 407–423. Springer.

Ronchi, M. R.; and Perona, P. 2015. Describing Common Human Visual Actions in Images. *arXiv:1506.02203 [cs]* URL http://arxiv.org/abs/1506.02203. ArXiv: 1506.02203.

Schönfeld, E.; Ebrahimi, S.; Sinha, S.; Darrell, T.; and Akata, Z. 2018. Generalized Zero-and Few-Shot Learning via Aligned Variational Autoencoders. *arXiv preprint arXiv:1812.01784* .

Shen, L.; Yeung, S.; Hoffman, J.; Mori, G.; and Fei-Fei, L. 2018. Scaling Human-Object Interaction Recognition through Zero-Shot Learning. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 1568–1576. IEEE.

Song, Y.; Li, W.; Zhang, L.; Yang, J.; Kiciman, E.; Palangi, H.; Gao, J.; Kuo, C.-C. J.; and Zhang, P. 2020. Novel Human-Object Interaction Detection via Adversarial Domain Generalization. *arXiv:2005.11406 [cs]* URL http://arxiv.org/abs/2005.11406. ArXiv: 2005.11406.

Wan, B.; Zhou, D.; Liu, Y.; Li, R.; and He, X. 2019. Pose-aware Multi-level Feature Network for Human Object Interaction Detection. *arXiv:1909.08453 [cs]* URL http://arxiv.org/abs/1909.08453. ArXiv: 1909.08453.

Wang, T.; Anwer, R. M.; Khan, M. H.; Khan, F. S.; Pang, Y.; Shao, L.; and Laaksonen, J. 2019. Deep Contextual Attention for Human-Object Interaction Detection. *arXiv:1910.07721 [cs]* URL http://arxiv.org/abs/1910.07721. ArXiv: 1910.07721.

Wang, T.; Yang, T.; Danelljan, M.; Khan, F. S.; Zhang, X.; and Sun, J. 2020. Learning Human-Object Interaction Detection Using Interaction Points. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 4115–4124. Seattle, WA, USA: IEEE. ISBN 978-1-72817-168-5. doi:10.1109/CVPR42600.2020.00417. URL https://ieeexplore.ieee.org/document/9157334/.

Wang, X.; Ye, Y.; and Gupta, A. 2018. Zero-shot recognition via semantic embeddings and knowledge graphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6857–6866.

Xu, B.; Wong, Y.; Li, J.; Zhao, Q.; and Kankanhalli, M. S. 2019. Learning to Detect Human-Object Interactions With Knowledge. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Xu, X.; Hospedales, T.; and Gong, S. 2015. Semantic embedding space for zero-shot action recognition. In *2015 IEEE International Conference on Image Processing (ICIP)*, 63–67. IEEE.

Yatskar, M.; Zettlemoyer, L.; and Farhadi, A. 2016. Situation recognition: Visual semantic role labeling for image understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5534–5542.

Zhang, L.; Xiang, T.; and Gong, S. 2017. Learning a deep embedding model for zero-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021–2030.

Zhou, P.; and Chi, M. 2019. Relation Parsing Neural Network for Human-Object Interaction Detection. In *Proceedings of the IEEE International Conference on Computer Vision*, 843–851.

Zhuang, B.; Wu, Q.; Shen, C.; Reid, I. D.; and van den Hengel, A. 2018. HCVRD: A Benchmark for Large-Scale Human-Centered Visual Relationship Detection. In *AAAI*.