

Human Object Interaction Detection Primed with Context

Maya Antoun^a and Daniel Asmar^b

Vision and Robotics Lab, American University of Beirut, Bliss Street, Beirut, Lebanon

Keywords: Human Object Interaction, Scene Understanding, Deep Learning.

Abstract: Recognizing Human-Object Interaction (HOI) in images is a difficult yet fundamental requirement for scene understanding. Despite the significant advances deep learning has achieved so far in this field, the performance of state of the art HOI detection systems is still very low. Contextual information about the scene has shown improvement in the prediction. However, most works that use semantic features rely on general word embedding models to represent the objects or the actions rather than contextual embedding. Motivated by evidence from the field of human psychology, this paper suggests contextualizing actions by pairing their verbs with their relative objects at an early stage. The proposed system consists of two streams: a semantic memory stream on one hand, where verb-object pairs are represented via a graph network by their corresponding feature vector; and an episodic memory stream on the other hand in which human-objects interactions are represented by their corresponding visual features. Experimental results indicate that our proposed model achieves comparable results on the HICO-DET dataset with a pretrained object detector and superior results on HICO-DET with finetuned detector.


1 INTRODUCTION


Perception stands as one of the fundamental building blocks of a completely autonomous system. Living beings rely on perception for their survival; we perceive the environment around us, the objects we interact with, as well as other humans. Despite its apparent simplicity, visual perception is difficult to realize in autonomous agents primarily because of our shortcomings in understanding and replicating human solutions that involve higher levels of cognition. In the past decade, and with the notable developments in deep learning, significant steps forward have been achieved in scene perception and understanding. More specifically, and of particular interest to this paper is the problem of Human-Object Interaction (HOI). Given an input image, the aim of HOI is to localize and estimate the interactions between humans and the objects around them by predicting the triplet $\langle \text{human}, \text{predicate}, \text{object} \rangle$. Detecting these interactions requires both knowledge of human and object information as well as the possible interactions between them.

Researchers have solved the HOI problem using two types of methods: two-stage and single-stage methods. The first stage in a two-stage method is

the detection of the humans and objects using an off-the-shelf detector, then in the second stage the interaction between them is predicted using the extracted features. In single-stage systems, the object detection and interaction prediction are done in parallel or in an end-to-end manner. Most of the existing two-stage systems, (Gkioxari et al., 2018; Gao et al., 2018; Baldassarre et al., 2020; Hou et al., 2020; Li et al., 2020a), rely on interpreting the scene based on its appearance as well as the geometric layout of objects and people within the scene. In some of these works, contextual information is only incorporated through features from the union region of a human and object bounding box which may not always be shown in the features covering the union region. Other systems, (Li et al., 2019; Wan et al., 2019; Song et al., 2020) solve the HOI problem by estimating the pose of detected people as an addition to the spatial and visual features. However, the human performing the action on the object is not always visible in the image and can be occluded by different objects or other humans. Therefore, relying on the human visual and pose for action prediction is not sufficient.

Alternatively, other two-stage networks, (Liang et al., 2020; Bansal et al., 2020a; Kim et al., 2020b; Zhou et al., 2020; Li et al., 2020b; Sun et al., 2020; Wang et al., 2020a; Xu et al., 2019; Gao et al., 2020; Liu et al., 2020; Hou et al., 2021), predict the HOI prediction by integrating semantics into the network

^a  <https://orcid.org/0000-0001-6017-5804>

^b  <https://orcid.org/0000-0002-4932-9777>

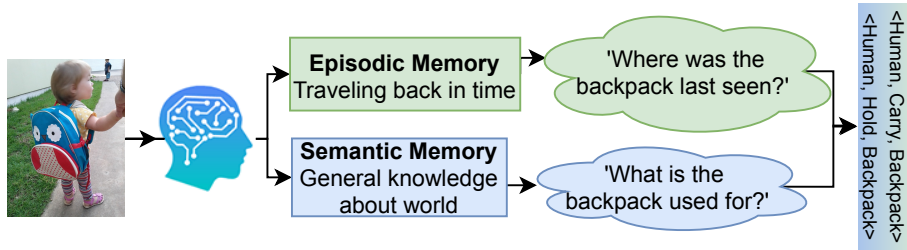


Figure 1: The perception behind HOI prediction. The episodic memory recalls contextual details from previous experiences and the semantic memory recalls facts, concepts, and ideas about the objects and the scene in question. The episodic memory and the semantic memory interplay and together help predict the correct interaction from the image.

architecture. Xu (Xu et al., 2019) construct a knowledge graph between object and action based on the semantic features of the ground-truth annotations of training dataset and external source. Bansal (Bansal et al., 2020a) integrate visual and spatial features with general word embedding of humans and objects. Gao (Gao et al., 2020) propose a dual relation graph by using spatial-semantic representation to describe each human-object pair. Liu (Liu et al., 2020) build a consistency graph that encodes the relations among objects, actions and interactions. Liang (Liang et al., 2020) build a dual-graph attention network that aggregates contextual visual, spatial, and semantic information. These works use general vector representation models to represent the actions and objects which does not take into consideration the semantic context of the object in the image expect for the established edges between the nodes in the graphs. Thus, the same action is represented by the same feature vector when mentioned with different objects.

To improve the HOI detection, recent works have developed one-stage pipelines to detect HOIs in a single shot. Single-stage methods, (Liao et al., 2020; Wang et al., 2020b; Kim et al., 2020a), localize the interaction with an interaction point or find the anchor box of a human-object pair. Contextual features are extracted around the detected point or box. The interacting triplets are predicted by matching the detected objects with the localized interaction and manually searching for the threshold. Later, single-stage methods were improved by using end-to-end transformer-based methods (Zou et al., 2021; Chen et al., 2021; Kim et al., 2021; Tamura et al., 2021). A transformer-based contextual self-attention mechanism is used to detect the interacting pairs and predict their interaction simultaneously. In these single-stage methods, contextual features are extracted visually from the image without any semantic representations. However, relying on visual context can be tricky in images where details are not well visible, such as in paintings and artwork. In our work we rely on improving the semantic contextual representation of the objects

and verbs. Our model does not only rely on visual features because context from semantic features helps the model become more robust to any type of images whether details, such as edges, are visible. Therefore, we only compare our work to two-stage methods only.

To solve the problem of extracting context from the semantic representation instead of the visual representation, we rely on human psychology for action perception. Nelissen (Nelissen et al., 2005) argue that action information without knowledge about the identity of the object acted upon, is not sufficient to provide a full understanding of the observed action. Also, Gallese (Gallese et al., 1996) state that the movement analysis in humans depends on the presence of objects. The cortical responses for goal directed actions are different from the responses evoked when the same action is executed but without the presence of the object. Moreover, Bub (Bub and Masson, 2006) show that observers build specific forms of gestural knowledge obtained from the conceptual representations of the objects. This suggests the importance of object priming in the representation of the action.

Two types of memories interplay to guide the visual search for the targets in a scene: (1) The episodic memory, located in the hippocampus part of the brain, answers questions about the position, colors, edges, and context in which the object was last seen in previous experiences; and (2) the semantic memory, located in the neocortex part of the brain, includes facts, concepts and ideas independent of personal experience. The semantic memory answers general questions, such as the affordance or the name and type or usual position of the object of interest. Semantic information are derived from the accumulation of the episodic memory. Therefore, they are interrelated to form, together, a complete picture of the scene (Figure1).

This paper is motivated by human psychology where the human brain, to infer an interaction, relies not only on the human performing the action but also on the object that they are interacting with. Similar to vision transformers, we use BERT, a transformer-

based word embedding model for contextual semantic representation of the action in context of the detected object in the GCN. By using the visual features of the object and the union box to find the similarity between the interaction visual and semantic features, we build the connection between both memories.

In this paper we apply the idea to HOI detection by priming context into the encoding of actions (*i.e.* verbs) at different levels of a deep network. We rely on the detected object’s visual-spatial features as well as its semantic relationship to actions. To benefit from the influence of the object on the interaction prediction, we change the semantic representation of the actions based on their presence with the object.

Our network consists of two streams. In the first episodic memory stream, features corresponding to the visual appearance, spatial features and the physical layout of people and objects are extracted as well as that of the action. The second stream is the semantic memory in which a graph convolutional (GCN) network is built between the objects and the actions. The objects and the actions are represented in the affordance-based graph by their personalized contextual vector representation extracted from a contextual word embedding model. The verb-object dependence is applied by representing the action features as their word embedding when presented with the detected object. The features from the episodic memory stream and the semantic memory stream will be used together to predict the human object interaction.

The main contributions of this paper include:

- We build an object related graph where the action nodes are represented by their contextual embedding when mentioned with the object which improves the model performance by contextualizing the graph.
- By feeding the visual features in the episodic memory to the semantic features from the semantic memory, we build the relationship between the two memories and enhance the graph output.
- Our approach outperforms two-stage state-of-the-arts on the challenging HICO-DET dataset with a finetuned object detector and shows comparable results with an object detector pretrained on COCO.

2 SYSTEM OVERVIEW

Figure 2 presents the flowchart of our proposed HOI detection system. Given an input image, the objective is to predict the triplet $\langle human, verb, object \rangle$ between each candidate human-object pair.

The model consists of two streams: an episodic memory stream, and a semantic memory stream. For the episodic memory stream, we use a pretrained object detection model to detect humans and objects inside query images. Then, using a feature extraction backbone, visual features are extracted from each human F_H^v and object F_O^v individually on one hand, and from the union of both human and object F_{Int}^v on the other hand. Moreover, a spatial attention feature map F_{sp} is created from the both human and object bounding boxes. For the human pose features F_H^p , 2D human body pose is extracted using a pretrained pose estimation model (RMPE (Fang et al., 2017)). (Section 2.1). These features are used together to predict the HOI.

For the semantic memory stream, semantic information is represented by the knowledge graph, which is built from the ground truth annotations of the training data. To render the training dataset more comprehensive, we augment it with the ConceptNet database (Speer et al., 2016), which builds additional nodes to the verb-object graph based on the affordance knowledge of the objects. First, semantic features from the detected object’s class F_O^s and its related verbs F_V^s are extracted using BERT(Devlin et al., 2018), a contextual word embedding model. The object’s semantic features F_O^s are concatenated with its visual features F_O^v to create the object node $F_O^{v,s}$. The graph network is updated through convolutions and the new interaction features F_{Int}^s are compared with its visual features F_{Int}^v for HOI prediction. (Section 2.2). Finally, the loss functions calculated from each module are added together, using a weighted sum, to get the final verb prediction loss function.

2.1 Episodic Memory Stream

Given an input image, a pretrained object detection model (Faster-RCNN) detects the candidate humans and objects and estimates the coordinates of the bounding boxes for humans BB_h and objects BB_o . A feature extraction backbone, ResNet-101, is used to extract visual features from the cropped human F_H^v and objects boxes F_O^v respectively. Moreover, visual features from the union human and object boxes are extracted and represent the interaction visual features F_{Int}^v . A spatial attention feature map F_{sp} is generated from the human and object bounding boxes following (Bansal et al., 2020b) and (Chao et al., 2018).

We use the two channel binary image representation to model the spatial relationship between a human and an object. The union of the two bounding boxes as a reference and re-scale it to a fixed size. Then, a binary image with two channels is created:

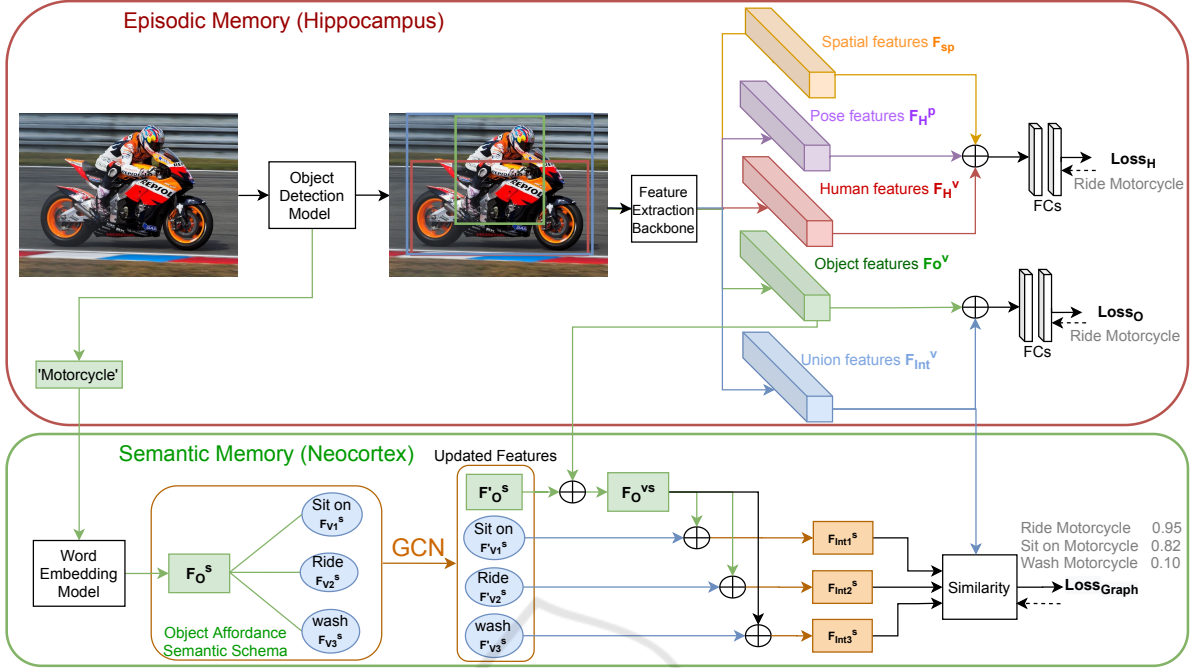


Figure 2: System flowchart: It consists of a visual-spatial module 2.1 where visual-spatial and pose features are extracted for the object, human, and the interaction between them. In the knowledge graph module 2.2, semantic features are extracted for the detected object and all the related candidate actions. The input features of the detected object-related actions are replaced with their contextual ones. A cross entropy loss is extracted from each spatial-visual branch and a cosine similarity is calculated between the candidate interaction semantic features and the interaction’s visual features.

value ones and zeros are filled in the human and object bounding boxes respectively in the first channel. In the second channel, value ones and zeros are filled in the object and human bounding boxes respectively. These two-channel binary images are fed into a two-layer convolutional network to extract the spatial attention feature map F_{sp} . To extract the human pose features F_H^P , we follow the work of (Li et al., 2019). We apply (Fang et al., 2017) to extract 17 keypoints from the union bounding box. Then, the keypoints are connected with lines of different gray value ranging from 0.15 to 0.95 representing the different body parts. Finally, the union box is reshaped to 64x64 to construct the pose map F_H^P .

For each detected human, the visual, spatial and pose features are concatenated together and fed to a fully connected layer followed by a Sigmoid activation function to find the action prediction score s_H based on human information. Similarly, an action prediction scores s_O is found from the concatenation of the object and union appearance features and feeding them to a fully connected layer followed by a Sigmoid activation function. These scores are used to extract the individual cross entropy losses \mathcal{L}_{cross}^H for the human and \mathcal{L}_{cross}^O for the object.

2.2 Semantic Memory Stream

To model the semantic representation of each object and action, we follow the work of (Xu et al., 2019). A graph convolutional network \mathcal{G} (GCN) (Kipf and Welling, 2016) is built, and whose aim is to model the relations between nodes N connected by edges E . The purpose of building a GCN, in our system, is to learn the features of the objects and candidate actions nodes by inspecting the relationship between them. An adjacency matrix is used to build the connection between the nodes. In our case, undirected edges are used where the connection between the nodes is the same in both direction.

We first use the training dataset ground truth annotation to extract the nodes and edges of the graph. To get a richer graph, we use an external dataset, ConceptNet, that includes all the affordance based relationships for all the objects in the database. To ensure that the affordance of the object is well represented in the graph, we extract all data with the *usedfor* relationships between them. This gives all possible triplets $\langle object, usedfor, action \rangle$. Thus the edges in the graph connects the objects with the actions that might occur with them and no connections are established between objects and actions that are never mentioned within the database. For example, if the detected ob-

ject is *motorcycle*, the actions that are connected to it in the graph include *sit on, ride, hold, wash, clean* and the actions that have no connection to it are *eat, cook, read*. Adding these affordance based nodes enriches the graph network with nodes that help in getting better action predictions.

As a first stage, the detected object semantic features are represented by a general word embedding features F_O^s . We rely on the pretrained BERT model (Devlin et al., 2018) to extract vector representation of each object word.

Based on the detected object, the interaction phrase composed by the verb followed by the object $\langle \text{verb}, \text{object} \rangle$ is fed to the word embedding model and the vector representation of the verb F_V^s (first word) is extracted in context of the object. These new representations are more specific to the detected object, and their features are tailored to that object. A context-based GCN is constructed where the object nodes are represented by the concatenation of the object semantic features F_O^v and the verb nodes are represented by their semantic word embedding features. The object-related verbs are represented by their contextual semantic features. whereas other non-related verbs are represented by their general semantic features. The adjacency matrix A of the GCN for all the networks is defined by the binary values of whether the nodes are connected or not.

Given the nodes features F_O^s and F_V^s and the adjacency matrix A , the semantic feature representation of the nodes at the $(i+1)^{th}$ layer are extracted using the forward pass of the GCN defined as:

$$F^{i+1} = \sigma(W^i F^i A') \quad (1)$$

where, A' is the normalized adjacency matrix, W^i is the learned weight at the i^{th} layer and F^i is the feature vector representation of the nodes at the i^{th} layer. σ is the non-linear activation function applied to the output of the convolution in order to represent the non-linear features in latent dimension. The output of the GCN is a feature vector representing the objects $F_O^{i,s}$ and the verbs $F_V^{i,s}$.

At the output of the GCN, we concatenate the updated object semantic features with its visual features with that of the candidate verbs. This concatenation represent the interaction semantic representation F_{Int}^s in the context of the object. We found that the addition of the object visual features yields a better representation of the object in the scene context. The cosine similarity between the visual F_{Int}^v and semantic interaction features is calculated, and the interaction with the highest similarity score s_g is considered to be the graph prediction. Inspired by (Salvador et al., 2017), the feature representations F_{Int}^s and F_{Int}^v are mapped

into the joint embedding space as: $\phi_v = W_v F_{Int}^v + b_v$ and $\phi_g = W_g F_{Int}^s + b_g$ respectively. W_v and W_g are the learned embedding weights. Thus the cosine similarity loss is defined as:

$$\mathcal{L}_{cos} = \begin{cases} 1 - \cos(\phi_v, \phi_g) & \text{if } y = 1 \\ \max(0, \cos(\phi_v, \phi_g) - \alpha) & \text{if } y = 0 \end{cases} \quad (2)$$

where, α is the margin and y is set to 1 if the candidate verb is the ground truth and zero if not.

The calculated losses from the episodic memory and the semantic memory modules are added, using a weighted sum, together to get the final loss function:

$$\mathcal{L}_{total} = \lambda_1 \mathcal{L}_{cos} + \lambda_2 \mathcal{L}_{cross}^H + \lambda_3 \mathcal{L}_{cross}^O \quad (3)$$

where, $\lambda_1, \lambda_2, \text{ and } \lambda_3$ are added weights to each individual loss function to control their contribution to the total loss. The final target is to minimize the total loss term in (3).

3 IMPLEMENTATION AND EXPERIMENTS

3.1 Experimental Setup

Experiments are performed on the HICO-DET (Chao et al., 2018) for Human Object Interaction prediction. HICO-DET is a large dataset for detecting HOIs. It includes 38,118 training and 9,658 testing images for 80 objects and 117 action verbs. HICO-DET annotates the images for full 600 human-object interactions. Interactions that occur less than 10 times in the training are categorized as Rare. We have 138 Rare and 462 Non-Rare interactions in the HICO-DET dataset. We also use ConceptNet (Speer et al., 2016) database to extract all the affordance based relationships and use the action verbs as additional nodes. ConceptNet includes data from other crowd sourced resources, expert-created resources, and games with a purpose such as Wiktionary which is a free multilingual dictionary and OpenCyc.

We follow the method of (Chao et al., 2018) to evaluate the performance of the proposed systems, using the metric of role mean average precision (role mAP). A prediction for a human-object interaction is considered correct if the human and object bounding boxes have an Intersection over Union (IoU) greater than 0.5 with the ground-truth boxes and if the verb class label of the interaction of the pair is correct.

We rely on the pretrained Faster-RCNN (Ren et al., 2015) for human and object detection for training. A threshold of 0.8 for human detection score and

Table 1: State-of-the-art comparison (mAP) on HICO-DET test set.

Method	Detector	Backbone	Default			Known Object		
			Full	Rare	Non Rare	Full	Rare	Non Rare
Bansal (Bansal et al., 2020a)	HICO-DET	ResNet-101	21.96	16.43	23.62	-	-	-
VCL(Hou et al., 2020)		ResNet-50	23.63	17.21	25.55	25.98	19.12	28.03
DRG (Gao et al., 2020)		ResNet-50-FPN	24.53	19.47	26.04	27.98	23.11	29.43
IDN (Li et al., 2020a)		ResNet-50	26.29	22.61	27.39	28.24	24.47	29.37
SCG(Zhang et al., 2020)		ResNet-50-FPN	31.33	24.72	33.31	-	-	-
OURS		ResNet-101	32.51	24.92	34.78	34.27	28.77	35.91
InteractNet(Gkioxari et al., 2018)	COCO	ResNet-50-FPN	9.94	7.16	10.77	-	-	-
GPNN (Qi et al., 2018)		ResNet-50	13.11	9.34	14.23	-	-	-
iCAN(Gao et al., 2018)		ResNet-50	14.84	10.45	16.15	16.26	11.33	17.73
Xu (Xu et al., 2019)		ResNet-50	14.7	13.26	15.13	-	-	-
Bansal(Bansal et al., 2020a)		ResNet-101	16.96	11.73	18.52	-	-	-
DRG (Gao et al., 2020)		ResNet-50-FPN	19.26	17.74	19.71	23.4	21.75	23.89
VCL (Hou et al., 2020)		ResNet-50	19.43	16.55	20.29	22.00	19.09	22.87
VSGNet (Ulutun et al., 2020)		ResNet-50 19.80	16.05	20.91	-	-	-	-
ConsNet(Liu et al., 2020)		ResNet-50-FPN	22.15	17.12	23.65	-	-	-
IDN(Li et al., 2020a)		ResNet-50	23.36	22.47	23.63	26.43	25.01	26.85
SCG(Zhang et al., 2020)		ResNet-50-FPN	21.85	18.11	22.97	-	-	-
OURS		ResNet-101	22.73	21.37	23.14	25.86	24.57	26.24

Table 2: State-of-the-art comparison (mAP) on V-COCO test set.

Method	Backbone	Scenario 1	Scenario 2
InteractNet(Gkioxari et al., 2018)	ResNet-50-FPN	40	-
GPNN(Qi et al., 2018)	ResNet-101	44	-
iCAN(Gao et al., 2018)	ResNet-50	45.3	52.4
TIN(Li et al., 2019)	ResNet-50	47.8	54.2
DRG(Gao et al., 2020)	ResNet-50-FPN	51	-
VSGNet(Ulutun et al., 2020)	ResNet-152	51.8	57
IDN(Li et al., 2020a)	ResNet-50	53.3	60.3
SCG(Zhang et al., 2020)	ResNet-50-FPN	54.2	60.9
OURS	ResNet-101	54.8	61.6

0.4 for object detection score is set. These values are chosen experimentally. ResNet-101 (He et al., 2016) is used as a feature extraction backbone. We finetune Faster-RCNN during testing only. The object nodes of the graph network are represented by their semantic features. We rely on the pretrained BERT model (Devlin et al., 2018) to extract vector representation of each word that has a size of 1x768. To get the candidate verbs semantic features, we feed the sentence composed of the verb and the object to BERT and extract the first word’s features as the verb features in the context of the object. We perform two convolutions on the input graph to get the final semantic vector representations of the object words and their connected verbs of the size 1x512. LeakyReLU with a negative slope of 0.2 (Wang et al., 2018) is used as the activation function after each layer of the graph. The total loss hyperparameters λ_1, λ_2 are set to 1 and λ_3 is set to 2. The margin for the cosine loss is set to 0.1. We use Stochastic Gradient Descent (SGD) to train the model for 10 epochs with a learning rate of 0.001, a weight decay of 0.0005, and a momentum of 0.9.

3.2 Experimental Testing and Results

We compare the mAP of our model with state-of-the-art methods in Table 1 on HICO-DET dataset. We report our results using a pretrained object detector on MS-COCO dataset (Lin et al., 2014) and using a fine-tuned object detector on Default and Known Object settings. We observe that our system shows comparable results when using a pretrained object detector compared to other state-of-the-art two-stage systems. When we used the fine-tuned detector, we were able to outperform state of the art two-stage methods by 0.97% and 1.03% on the Default and Known Object setting of the HICO-DET. The main reason why our model outperforms IDN with HICO-DET, but not with COCO is that our model was able to get higher cosine similarity between the interaction semantic and visual features which is due to more accurate detections. To analyze the contributions of each component of our model, we perform an ablation study and report the results in Table 3. We test our model without taking into consideration the HOI prediction from the human stream, including the visual and spatial and pose feature. Then, we test it without

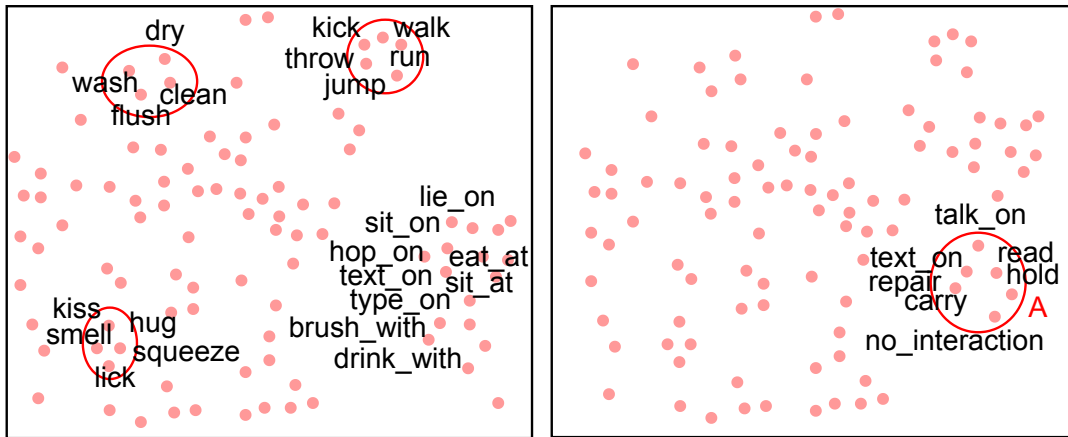


Figure 3: Visualisation of the input action class representations using tSNE (left) using BERT word embedding, and (right) after modifying the representation of cellphone-related actions.

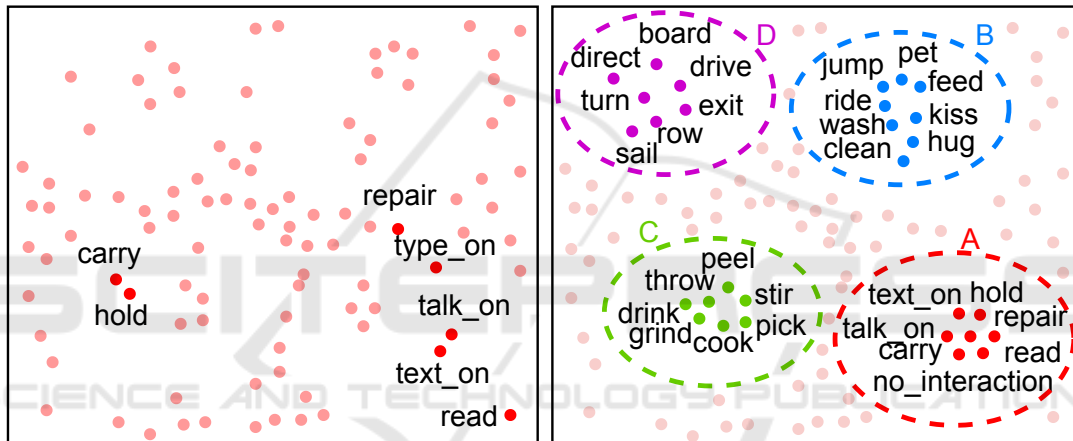


Figure 4: Visualisation of action class representations using tSNE at the output of the GCN (left) without object priming, and (right) with object ‘cellphone’ priming.

Table 3: Ablation study of the proposed system on HICO-DET in Default setting.

	Full	Rare	Non Rare
OURS	27.26	21.92	28.85
OURS- <i>w/o human</i>	23.21	19.45	24.33
OURS- <i>w/o object</i>	22.58	18.89	23.68
OURS- <i>w/o graph</i>	21.13	17.34	22.26

the object stream, including the object and the union visual features. At last, to show the importance of the semantic graph we test the model without the semantic memory stream. We can see from the results that the graph has the largest influence on the HOI prediction followed by the object and union features.

The results support our hypothesis about the importance of the presence of the object in the interaction prediction and the value of the semantic features for better HOI prediction.

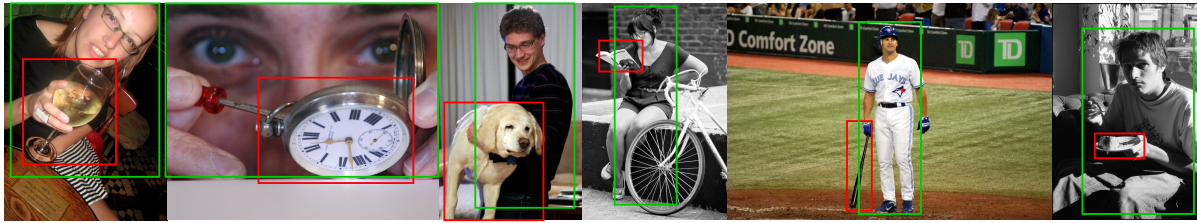
In Table 4, we test our model without object priming. The action verbs in the graph are represented by

their BERT word embedding vector without priming them with the detected object. The results support our main argument that object priming enhances the verb representation and thus it improves the HOI predictions. It improved the model mAP by 2.49 %, given that object priming adds context to the semantic representation of the related verbs, and thus enhances the HOI prediction accordingly.

In Figure 3, we compare the different input representations of the candidate actions using tSNE (Maaten and Hinton, 2008). Figure 3 (left) represents the general BERT word embedding on the candidate action verbs. We can see that BERT represents the

Table 4: Ablation study of the semantic memory module on HICO-DET in Default setting.

	Full	Rare	Non Rare
OURS	27.26	21.92	28.85
OURS- <i>w/o object priming</i>	24.77	18.61	26.61



(a) From left to right: hold wine glass, hold clock, hold dog, hold book, hold baseball bat, hold cake.



(b) From left to right: ride skateboard, ride boat, ride bicycle, ride snowboard.

Figure 5: HOI detections on the HICO-DET test images. Our model detects various forms of HOIs for same actions ‘hold’ in (a) and ‘ride’ in (b) with different set of objects.

actions in a general contextual manner. In Figure 3 (right), we modify the representation of detected object *cellphone* related actions by feeding the sentence composed by the verb and the object to BERT and extracting the feature vector of the verb. We can see that the actions related to cellphone are clustered together without the use of GCN. These activities include *text on*, *talk on*, *hold*, ..., and shows how much the addition of the object context in the semantic module helps the GCN by clustering the object-related actions at the input.

Figure 4 shows the representations of the candidate actions at the output of the graph convolutional network with and without the introduction of the detected object at the input GCN level. From Figure 4 (left), we see that the actions related (in red) to the detected object *cellphone* were not correctly clustered together. From Figure 4 (left), we can notice that Cluster A (in red), which refers to cellphone related activities, enclosed additional actions to the input ones such as *type on*, *pick up*. Moreover, we can see that the training of the model helped in clustering different activities related actions. For example, Cluster B (in blue) refers to pet related activities such as *pet*, *walk*, *feed*, *kiss*. Cluster C (in green) refers to food related activities including *peel*, *stir*, *pour*. Cluster D (in purple) refers to vehicle related activities such as *drive*, *board*, *load*, *sail*.

Figure 5 shows qualitative results of our method. We highlight the ability of our system to correctly predict interactions with objects that it was not trained on. This is due to the fact that the same verb is con-

nected to similar objects in the training and external datasets.

4 CONCLUSIONS

We present a novel model for Human Object Interaction detection which uses visual, spatial, pose, and graph semantic features from the input image to get the best output prediction. We showed that the presence of the object in the input semantic features plays a fundamental role in enhancing the action prediction by contextualizing the semantic representations in the scene. Visual-spatial features are extracted from the human, object, and interaction. A similarity is calculated between the visual-spatial features of the interaction and the semantic features from the graph output of the candidate interactions. The external dataset included all affordance based connection that can occur between a object and a verb. The constructed semantic graph helped in predicting interactions that the network was not trained on. We rely on prior work in the episodic stream. We contextualize the action semantic representation in the GCN. We connect both memories through the concatenation of the object visual and semantic features and the comparison of the interaction visual and semantic features. Our experiments demonstrated that our system improved the performance on HOI detection.

ACKNOWLEDGEMENTS

The authors would like to acknowledge the American University of Beirut (AUB) and the National Council for Scientific Research of Lebanon (CNRS-L) for granting a doctoral fellowship to Maya Antoun.

REFERENCES

- Baldassarre, F., Smith, K., Sullivan, J., and Azizpour, H. (2020). Explanation-based weakly-supervised learning of visual relations with graph networks. *arXiv preprint arXiv:2006.09562*.
- Bansal, A., Rambhatla, S. S., Shrivastava, A., and Chellappa, R. (2020a). Detecting human-object interactions via functional generalization. In *AAAI*, pages 10460–10469.
- Bansal, A., Rambhatla, S. S., Shrivastava, A., and Chellappa, R. (2020b). Spatial priming for detecting human-object interactions. *arXiv preprint arXiv:2004.04851*.
- Bub, D. and Masson, M. (2006). Gestural knowledge evoked by objects as part of conceptual representations. *Aphasiology*, 20(9):1112–1124.
- 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)*, pages 381–389. IEEE.
- Chen, M., Liao, Y., Liu, S., Chen, Z., Wang, F., and Qian, C. (2021). Reformulating hoi detection as adaptive set prediction. *arXiv preprint arXiv:2103.05983*.
- Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Fang, H.-S., Xie, S., Tai, Y.-W., and Lu, C. (2017). Rmpe: Regional multi-person pose estimation. In *Proceedings of the IEEE international conference on computer vision*, pages 2334–2343.
- Gallese, V., Fadiga, L., Fogassi, L., and Rizzolatti, G. (1996). Action recognition in the premotor cortex. *Brain*, 119(2):593–609.
- Gao, C., Xu, J., Zou, Y., and Huang, J.-B. (2020). Drg: Dual relation graph for human-object interaction detection. In *European Conference on Computer Vision*, pages 696–712. Springer.
- Gao, C., Zou, Y., and Huang, J.-B. (2018). ican: Instance-centric attention network for human-object interaction detection. *arXiv preprint arXiv:1808.10437*.
- Gkioxari, G., Girshick, R., Dollár, P., and He, K. (2018). Detecting and recognizing human-object interactions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8359–8367.
- He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.
- Hou, Z., Peng, X., Qiao, Y., and Tao, D. (2020). Visual compositional learning for human-object interaction detection. *arXiv preprint arXiv:2007.12407*.
- Hou, Z., Yu, B., Qiao, Y., Peng, X., and Tao, D. (2021). Affordance transfer learning for human-object interaction detection. *arXiv preprint arXiv:2104.02867*.
- Kim, B., Choi, T., Kang, J., and Kim, H. J. (2020a). Union-det: Union-level detector towards real-time human-object interaction detection. In *European Conference on Computer Vision*, pages 498–514. Springer.
- Kim, B., Lee, J., Kang, J., Kim, E.-S., and Kim, H. J. (2021). Hotr: End-to-end human-object interaction detection with transformers. *arXiv preprint arXiv:2104.13682*.
- Kim, D., Lee, G., Jeong, J., and Kwak, N. (2020b). Tell me what they’re holding: Weakly-supervised object detection with transferable knowledge from human-object interaction. In *AAAI*, pages 11246–11253.
- Kipf, T. N. and Welling, M. (2016). Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.
- Li, Y.-L., Liu, X., Wu, X., Li, Y., and Lu, C. (2020a). Hoi analysis: Integrating and decomposing human-object interaction. *Advances in Neural Information Processing Systems*, 33.
- Li, Y.-L., Xu, L., Liu, X., Huang, X., Xu, Y., Wang, S., Fang, H.-S., Ma, Z., Chen, M., and Lu, C. (2020b). Pastanet: Toward human activity knowledge engine. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 382–391.
- 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*, pages 3585–3594.
- Liang, Z., Guan, Y., and Rojas, J. (2020). Visual-semantic graph attention network for human-object interaction detection. *arXiv preprint arXiv:2001.02302*.
- 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 482–490.
- Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.
- Liu, Y., Yuan, J., and Chen, C. W. (2020). Consnet: Learning consistency graph for zero-shot human-object interaction detection. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 4235–4243.
- Maaten, L. v. d. and Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605.
- Nelissen, K., Luppino, G., Vanduffel, W., Rizzolatti, G., and Orban, G. A. (2005). Observing others: multi-

- ple action representation in the frontal lobe. *Science*, 310(5746):332–336.
- Qi, S., Wang, W., Jia, B., Shen, J., and Zhu, S.-C. (2018). Learning human-object interactions by graph parsing neural networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 401–417.
- Ren, S., He, K., Girshick, R., and Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99.
- Salvador, A., Hynes, N., Aytar, Y., Marin, J., Ofli, F., Weber, I., and Torralba, A. (2017). Learning cross-modal embeddings for cooking recipes and food images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3020–3028.
- 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 preprint arXiv:2005.11406*.
- Speer, R., Chin, J., and Havasi, C. (2016). Conceptnet 5.5: An open multilingual graph of general knowledge. *arXiv preprint arXiv:1612.03975*.
- Sun, X., Hu, X., Ren, T., and Wu, G. (2020). Human object interaction detection via multi-level conditioned network. In *Proceedings of the 2020 International Conference on Multimedia Retrieval*, pages 26–34.
- Tamura, M., Ohashi, H., and Yoshinaga, T. (2021). Qpic: Query-based pairwise human-object interaction detection with image-wide contextual information. *arXiv preprint arXiv:2103.05399*.
- Ulutun, O., Iftekhar, A., and Manjunath, B. S. (2020). Vs-gnet: Spatial attention network for detecting human object interactions using graph convolutions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13617–13626.
- Wan, B., Zhou, D., Liu, Y., Li, R., and He, X. (2019). Pose-aware multi-level feature network for human object interaction detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9469–9478.
- Wang, S., Yap, K.-H., Yuan, J., and Tan, Y.-P. (2020a). Discovering human interactions with novel objects via zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11652–11661.
- Wang, T., Yang, T., Danelljan, M., Khan, F. S., Zhang, X., and Sun, J. (2020b). Learning human-object interaction detection using interaction points. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4116–4125.
- 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*, pages 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*.
- Zhang, F. Z., Campbell, D., and Gould, S. (2020). Spatially conditioned graphs for detecting human-object interactions. *arXiv preprint arXiv:2012.06060*.
- Zhou, T., Wang, W., Qi, S., Ling, H., and Shen, J. (2020). Cascaded human-object interaction recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4263–4272.
- Zou, C., Wang, B., Hu, Y., Liu, J., Wu, Q., Zhao, Y., Li, B., Zhang, C., Zhang, C., Wei, Y., et al. (2021). End-to-end human object interaction detection with hoi transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11825–11834.