SCIENCE CHINA
Information Sciences

• LETTER •

Special Focus on Deep Learning for Computer Vision

June 2022, Vol. 65 160108:1–160108:2
https://doi.org/10.1007/s11432-021-3427-2

Human-object interaction detection via interactive visual-semantic graph learning

Tongtong WU[1], Fuqing DUAN[1*], Liang CHANG[1] & Ke LU[2]

[1] *College of Artificial Intelligence, Beijing Normal University, Beijing* 100875*, China;*
[2] *School of Computer and Control Engineering, University of Chinese Academy of Sciences, Beijing* 100049*, China*

Received 31 March 2021/Revised 10 August 2021/Accepted 15 October 2021/Published online 22 April 2022

Citation Wu T T, Duan F Q, Chang L, et al. Human-object interaction detection via interactive visual-semantic graph learning. Sci China Inf Sci, 2022, 65(6): 160108, https://doi.org/10.1007/s11432-021-3427-2

Dear editor,

Human-object interaction (HOI) detection is an important human-centric visual understanding task with several applications in visual monitoring, intelligent robot, etc. It aims at localizing and inferring interaction relationships between humans and objects in images. As a result of their success in object detection, many object detectors can be used to localize human and object instances. Therefore, the key to HOI detection mainly lies in the second part, namely interaction recognition. Many factors such as occlusion or self-occlusion, variation of the human gesture, and camera viewpoint, which bring ambiguity of the interaction recognition, cause it to be challenging.

Recently, owing to the success of deep learning and the availability of large-scale datasets, many deep learning-based HOI models have been proposed [1,2]. These methods usually use context information to improve model performance. However, they mainly adopt visual context information, and semantic context, which can provide effective prior knowledge, is still overlooked.

To address the issue mentioned above, we propose modeling the visual and semantic contexts using a visual graph and a semantic graph, and learning and refining the contextual information using a set of proposed graph update-modules, including graph inner update modules and graph cross update modules. This method reduces the model's burden to extract clues from raw images and provides key HOI prediction evidence.

Given an input image, the task is to detect all humans and objects' bounding boxes and predict their interactions. Like many existing methods, we use an off-the-shelf detector for detection and focus on the stage of interaction classification. We first use a detector to detect all human bounding boxes $\boldsymbol{B}_{\rm h}$ and object bounding boxes $\boldsymbol{B}_{\rm o}$. Our model's input is the image features from a backbone CNN and these bounding boxes. For each human bounding box $\boldsymbol{b}_{\rm h} \in \boldsymbol{B}_{\rm h}$ and object bounding box $\boldsymbol{b}_{\rm o} \in \boldsymbol{B}_{\rm o}$, we predict the HOI score $S_{\rm h,o,a}$ where "a" denotes an interaction. The score $S_{\rm h,o,a}$ depends on three parts, human score $S_{\rm h}$, object score $S_{\rm o}$ and interaction score $S_{\rm a}$. Specifically, the HOI score $S_{\rm h,o,a}$ has the following form:

$$S_{\rm h,o,a} = S_{\rm h} \times S_{\rm o} \times S_{\rm a}. \tag{1}$$

In this study, we only focus on the action score $S_{\rm a}$.

Overview. Figure 1 shows a high-level overview of the proposed network architecture. The network mainly consists of two branches, a visual branch which extracts visual features from the human-object pairs, and a graph branch which models contextual information using graphs and extracts contextual features from graphs. After these branches, the visual and contextual features are fused for classification.

Visual branch. This branch extracts visual characteristics from human-object pairs. Given a human/object bounding box, we crop the region from the image feature, and then send it through an RoI pooling layer, a residual block, and a global average pooling layer. After that, each human/object is defined as a visual embedding vector $\boldsymbol{f}$. Letting $\boldsymbol{b}_{\rm h} = \{x_{\rm h}^1, y_{\rm h}^1, x_{\rm h}^2, y_{\rm h}^2\}, \boldsymbol{b}_{\rm o} = \{x_{\rm o}^1, y_{\rm o}^1, x_{\rm o}^2, y_{\rm o}^2\}$, we embed the spatial relationship of the human-object pair as a vector by

$$\begin{aligned} \boldsymbol{f}_{\rm sp} = \Bigg\{ & \frac{x_{\rm h}^1}{W_{\rm u}}, \frac{y_{\rm h}^1}{H_{\rm u}}, \frac{x_{\rm h}^2}{W_{\rm u}}, \frac{y_{\rm h}^2}{H_{\rm u}}, \frac{x_{\rm o}^1}{W_{\rm u}}, \frac{y_{\rm o}^1}{H_{\rm u}}, \frac{x_{\rm o}^2}{W_{\rm u}}, \\ & \frac{y_{\rm o}^2}{H_{\rm u}}, \frac{A_{\rm h}}{A_{\rm u}}, \frac{A_{\rm o}}{A_{\rm u}}, \frac{A_{\rm h}}{A_{\rm i}}, \frac{A_{\rm o}}{A_{\rm i}}, \frac{x_{\rm h}^1 - x_{\rm o}^1}{x_{\rm o}^2 - x_{\rm o}^1}, \frac{y_{\rm h}^1 - y_{\rm o}^1}{y_{\rm o}^2 - y_{\rm o}^1} \Bigg\}, \end{aligned} \tag{2}$$

where $W_{\rm u}$ and $H_{\rm u}$ represent the width and height of the union bounding box, $A_{\rm h}$ and $A_{\rm o}$ represent the area of the two bounding boxes and $A_{\rm i}$ and $A_{\rm u}$ are the area of the intersection and union bounding boxes, respectively. The spatial relationship vector is applied to a fully connected layer, and outputs the spatial embedding vector $\boldsymbol{f}_{\rm sp}^{'}$. Finally, human embedding vector $\boldsymbol{f}_{\rm h}$, object embedding vector $\boldsymbol{f}_{\rm o}$, and spatial embedding vector $\boldsymbol{f}_{\rm sp}^{'}$ are concatenated and the concatenated vector is sent into a fully connected layer.

$$\boldsymbol{f}_{\rm vis} = {\rm relu}(\boldsymbol{W}_1 \times (\boldsymbol{f}_{\rm h} \oplus \boldsymbol{f}_{\rm sp}^{'} \oplus \boldsymbol{f}_{\rm o}) + \boldsymbol{b}_1), \tag{3}$$

where $\oplus$ denotes the concatenate operation.

[*] Corresponding author (email: fqduan@bnu.edu.cn)

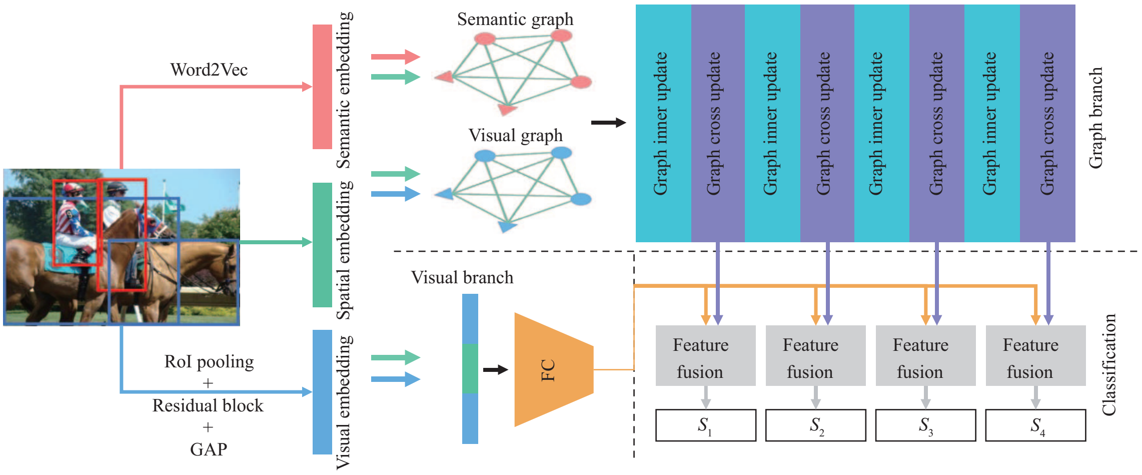

Figure 1 (Color online) Model architecture.

Graph branch. This branch uses graph convolution networks to capture useful contextual information. Graph convolution networks model the inputs through a graph composed of nodes and edges and extract features by traversing and updating the nodes. We propose two kinds of graphs, visual graph and semantic graph. In visual graph, human embedding vectors $\boldsymbol{f}_{\rm h}$ and object embedding vectors $\boldsymbol{f}_{\rm o}$ are defined as nodes. In semantic graph, we use Word2Vec [3] to generate word embedding vectors based on human and object categories. We add a virtual object node whose bounding box is the same size as the input image to retain information integrity. Thus, all contextual information, including background, is preserved. The virtual node is constructed based on the entire image feature map in visual graph and is a unit vector in semantic graph. The spatial embedding vectors $\boldsymbol{f}_{\rm sp}^{'}$ of node pairs are defined as edges. Finally, these visual nodes, semantic nodes, and edges are used to construct a fully connected visual graph and a semantic graph, respectively.

Graph update consists of two modules, a graph inner update module and a graph cross update module. Messages are transferred within each graph in graph inner update module and between the visual graph and semantic graph in graph cross update module. In graph inner update module, we use graph attention [4] method to update graph, while in graph cross update module, we use corresponding pairs of visual and semantic nodes to update each other's features through a basic attention method. The details of graph update modules are shown in Appendix A.

Classification. Given a human-object pair, we extract a visual feature vector from the output of the visual branch, and from the final updated graphs, we obtain corresponding human node feature vector $\boldsymbol{h}_{\rm h}$, object node feature vector $\boldsymbol{h}_{\rm o}$, and human-object edge feature vector $\boldsymbol{h}_{\rm e}$. The feature vectors $\boldsymbol{h}_{\rm h}$, $\boldsymbol{h}_{\rm o}$, and $\boldsymbol{h}_{\rm e}$ are concatenated, and the concatenated feature vector is sent into a fully connected layer to generate contextual features $\boldsymbol{h}_{\rm g}$. For visual graph and semantic graph, we calculate the contextual features $\boldsymbol{h}_{\rm g,vis}$ and $\boldsymbol{h}_{\rm g,sem}$, respectively, and concatenate $\boldsymbol{h}_{\rm g,vis}$ and $\boldsymbol{h}_{\rm g,sem}$ as $\boldsymbol{f}_{\rm g}$.

$$\boldsymbol{h}_{\rm g} = {\rm relu}(\boldsymbol{W}_2 \times (\boldsymbol{h}_{\rm h} \oplus \boldsymbol{h}_{\rm e} \oplus \boldsymbol{h}_{\rm o}) + \boldsymbol{b}_2), \tag{4}$$

$$\boldsymbol{f}_{\rm g} = \boldsymbol{h}_{\rm g,vis} \oplus \boldsymbol{h}_{\rm g,sem}. \tag{5}$$

The concatenated vector $\boldsymbol{f}_{\rm g}$ represents the final contextual features, and is fused with $\boldsymbol{f}_{\rm vis}$ by a basic attention method. The attention output $\boldsymbol{f}_{\rm out}$ is used directly through a fully connected layer for interaction classification by a multi-hot sigmoid cross entropy function.

Experiments. Experiments on prominent datasets HICO-DET and V-COCO show that our model is effective. The details of the experiment results and an ablation study are shown in Appendixes B and C, respectively.

Conclusion. In this study, we propose modeling the context visually and semantically by combining a visual graph and a semantic graph and learning a vital context in the HOI problem using a group of graph update-modules, including graph inner update modules and graph cross update modules. We fuse the contextual features from the visual graph and semantic graph with the visual characteristics of the human-object pairs in a network to detect HOIs. We evaluate our proposed model on two challenging datasets, HICO-DET and V-COCO, and demonstrate excellent performance. Our work can provide a reference for modeling contextual information in the HOI problem.

Acknowledgements This work was supported by National Key Research and Development Program of China (Grant No. 2018AAA0100802).

Supporting information Appendixes A–C. The supporting information is available online at info.scichina.com and link.springer.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.

References

1 Gao C, Zou Y L, Huang J-B. iCAN: instance-centric attention network for human-object interaction detection. 2018. ArXiv:1808.10437

2 Wang T, Anwer R M, Khan M H, et al. Deep contextual attention for human-object interaction detection. In: Proceedings of IEEE International Conference on Computer Vision, 2019

3 Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality. In: Proceedings of Advances in Neural Information Processing Systems, 2013. 3111–3119

4 Veličković P, Cucurull G, Casanova A, et al. Graph attention networks. 2017. ArXiv:1710.10903