TASHKENT UNIVERSITY OF
INFORMATION TECHNOLOGIES
NAMED AFTER MUHAMMAD AL-KHWARIZMI

MUHAMMAD AL-XORAZMIY NOMIDAGI
TOSHKENT AXBOROT TEXNOLOGIYALARI
UNIVERSITETI

BULLETIN OF TUIT:

MANAGEMENT AND COMMUNICATION TECHNOLOGIES



ISSN:2181-1083 DOI:
Abdurashidova Kamola Turgunbaevna, Abdukhakimov Fayzulla Kudratulla ugli, Chorshanbiyeva Sevinch Akramovna № 1(9)2024

Towards Identifying Interactions Between Human and Object Using Transformer-Based Models

Abdurashidova Kamola Turgunbaevna[1], Abdukhakimov Fayzulla Kudratulla ugli[2], Chorshanbiyeva Sevinch Akramovna[3]

[1] Associate professor at Tashkent University of Information Technologies named after Muhammad al-Khwarizmi, Department of Computer Systems
[2] Graduate student at Tashkent University of Information Technologies named after Muhammad al-Khwarizmi, Department of Computer Systems
[3] Undergraduate student at Tashkent University of Information Technologies named after Muhammad al-Khwarizmi, Department of Computer Systems

E-mail: kamolabdurashidova@gmail.com, abduhakimovfayzullo@gmail.com, chorshanbiyevasevinch1@gmail.com

ABSTRACT

Understanding the subtle interactions between individuals and objects is critical for a variety of applications, including robotics, computer vision, and human-computer interaction. This paper looks at the research landscape for using transformer-based models to discover and understand these relationships. Existing approaches based on their methodologies, datasets used, and application fields have been explored thoroughly. The paper identifies several existing SOTA (state-of-the-art) methodologies that use transformer-based models to detect interactions among people, and objects as well.

Keywords: *transformers, encoder-decoder architecture, attention mechanism, sequential data, computer vision.*

INTRODUCTION

Deep learning, a subset of artificial intelligence, processes data and generates decision-making patterns in the same way that a human brain does. Its strength is the ability to learn structures of data through various layers like in human's brain, allowing it to extract detailed properties and make predictions or classifications.

Deep learning has significantly contributed to diagnostic and treatment procedures in the healthcare field. In [13], for example, medical imaging data such as echocardiograms or MRI scans have been evaluated to discover small problems in the structure or function of the heart, allowing for earlier diagnosis to prevent heart illnesses. Furthermore, deep learning helps to understand human interactions, which is important

for fields such as human-computer interaction and sentiment analysis. As in [14], various methods of graph neural networks have been analyzed to identify interactions among people.

In a study outlined in [1], Human-Object Interaction (HOI) detection is formally described as the objective of forecasting a collection of human, object, interaction triplets within an image. Past methodologies have tackled this objective indirectly by initially conducting object detection and subsequently linking human-object pairs with separate post-processing stages. Specifically, earlier endeavors (referred to as sequential HOI detectors [2, 3, 4, 5]) have executed this linking through a subsequent neural network, resulting in significant time and computational resources consumption.

To address the inefficiencies of sequential HOI detectors, recent studies [6,7] have introduced parallel HOI detectors. These approaches explicitly pinpoint interactions using either interaction boxes (encompassing the center point of an object pair) [6] or union boxes (covering the regions of both objects) [8]. These localized interactions are then linked with object detection outcomes to form the human, object, interaction triplets. Rather than relying on time-intensive neural network inference, a straightforward matching process is employed, leveraging heuristics such as distance [6, 7] or IoU [8].

Nevertheless, prior studies on HOI detection have notable limitations in two regards: a) They necessitate supplementary post-processing procedures, such as suppressing nearly identical predictions and heuristic threshold setting. b) Despite evidence demonstrating that modeling relationships between objects enhances object detection, the efficacy of incorporating high-level dependencies for interactions in HOI detection remains largely unexplored.

The transformer architecture marks a significant paradigm shift in deep learning, particularly natural language processing (NLP). Transformers, as opposed to recurrent neural networks (RNNs) and convolutional neural networks (CNNs), rely exclusively on self-attention mechanisms to assess the importance of various input items. This attention technique enables the model to focus on all points in the input sequence at the same time, allowing it to effectively capture long-range dependencies. The transformer architecture is made up of two parts: an encoder and a decoder, each with numerous layers of self-attention mechanisms and feed-forward neural networks (Figure 1). During the encoding phase, the input sequence is embedded in high-dimensional representations that are then processed by numerous encoder layers, where self-attention mechanisms capture the relationships between the sequence's elements. During the decoding phase, the decoder creates the output sequence based on the encoded representations, using both self-attention and cross-attention techniques to maintain coherence and relevance between the input and output sequences.

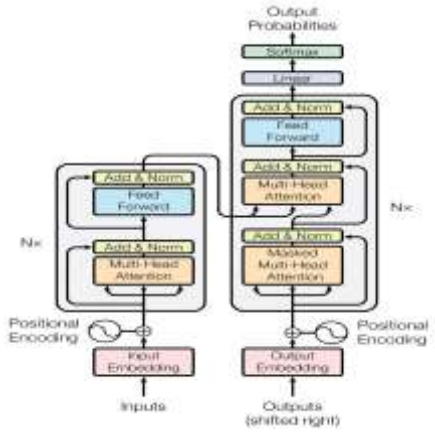


Figure 1. Overall architecture of transformers

METHODOLOGY

HOTR (Human-Object interaction TRansformer). This method introduces an efficient and precise HOI algorithm that forecasts a collection of human-object interactions within a scene simultaneously, employing a direct set prediction approach [9]. The general process of HOTR is depicted in Figure 2. The architecture comprises a transformer encoder-decoder setup with a common encoder and two separate decoders (instance decoder and interaction decoder). The outcomes of these decoders are linked using the proposed HO Pointers to produce the ultimate HOI triplets. Details regarding the HOTR architecture will be discussed, followed by an introduction to HO Pointers. Similar to DETR [10], the backbone CNN and a shared encoder gather the global context from the input image. Then, two sets of positional embeddings, namely the instance queries and the interaction queries, are introduced into the two parallel decoders (instance decoder and interaction decoder in Fig. 2).

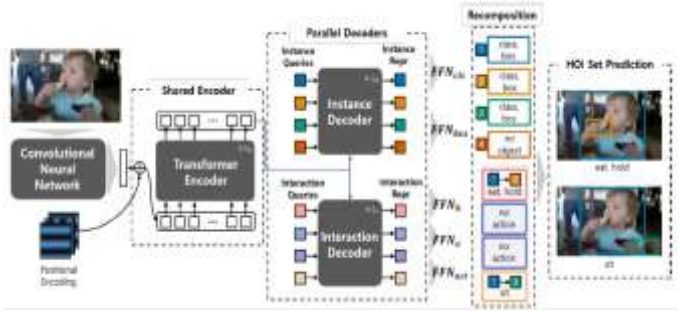


Figure 2. Overall architecture of HOTR model

The instance decoder transforms the instance queries into representations for object detection, while the interaction decoder transforms the interaction queries into representations for interaction detection. Feed-forward networks (FFNs) are applied to the interaction representation, resulting in the identification of a Human Pointer, an Object Pointer, and the interaction type. Essentially, the interaction representation assists in pinpointing human and object regions by guiding the relevant instance representations using the Human Pointer and Object Pointer (HO Pointers), rather than directly predicting the bounding box. This approach offers several advantages over direct regression. For example, it overcomes issues which encountered when an object is involved in multiple interactions, where direct regression may lead to inconsistent localization. The architecture addresses this challenge by employing separate instance and interaction representations, which are then linked using HO Pointers. Furthermore, the design enhances the efficiency of localization learning by eliminating the need to redundantly learn localization for each interaction. Notably, the experiments demonstrate that a shared encoder is more effective in learning HO Pointers compared to using two separate encoders.

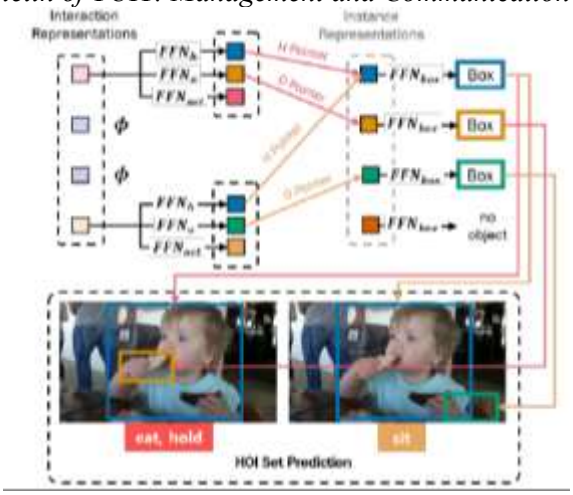


Figure 3. Conceptual illustration of how HO Pointers associates the interaction representations with instance representations

Figure 3 provides a general representation of how HO Pointers, namely the Human Pointer and Object Pointer, establish connections between the parallel predictions generated by the instance decoder and the interaction decoder. These pointers store the indices of the corresponding instance representations of the human and the object involved in the interaction. Once the interaction decoder transforms K interaction queries into K interaction representations, denoted as zi, they undergo processing through two feed-forward networks, $\mathrm{FFN_h} : \mathbb{R}^d \to \mathbb{R}^d$ and $\mathrm{FFN_o} : \mathbb{R}^d \to \mathbb{R}^d$, resulting in vectors $v_i^h$and v_i^o, respectively, such that $v_i^h = \mathrm{FFN_h}(z_i)$ and $v_i^o = \mathrm{FFN_o}(z_i)$. Subsequently, the Human, Object Pointers $\hat{c}_i^h$ and $\hat{c}_i^o$, representing the indices of the instance representations with the highest similarity scores, are derived:

$$\hat{c}_i^h = \mathrm{argmax}(\mathrm{sim}(v_i^h, \mu_j)), \quad (1)$$

$$\hat{c}_i^o = \mathrm{argmax}(\mathrm{sim}(v_i^o, \mu_j)),$$

where μ_j is the j-th instance representation and $sim(u,v) = u^T v/||u||||v||$.

QPIC. This method is a query-based HOI detector that detects individuals and objects in pairs using image-wide contextual data [12]. QPIC is based on DETR [10], an object detector, and incorporates crucial components such as the attention mechanism and query-based detection within a transformer framework [11]. The attention mechanism analyzes the entire visual region, selectively accumulating contextually relevant elements based on image content. Furthermore, QPIC's query design allows for the distinct extraction of features for numerous HOI instances, even if they are close together, by guaranteeing that each query focuses on a single human-object pair. This method permits attention calculations to be performed query-wise rather than location-wise, allowing for a clear identification of each query's target human-object combination using self-attention mechanisms.

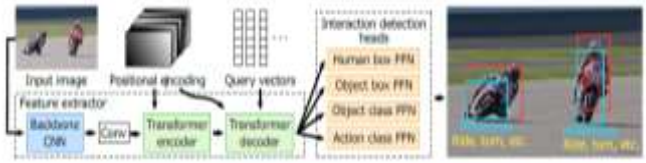


Figure 4. Overall architecture of the proposed QPIC

Figure 4 provides an overview of the QPIC architecture. Beginning with an input image $x \in R^{3\times H\times W}$, a feature map $z_b \in R^{D_b\times H'\times W'}$ is generated through a backbone network. Here, H and W denote the height and width of the input image, while H′ and W′ represent those of the output feature map, with D_b indicating the number of channels. Typically, H′ < H and W′ < W. Subsequently, z_b undergoes dimension reduction

from D_b to D_c through a projection convolution layer with a kernel size of 1 × 1. The transformer encoder then processes this feature map with reduced dimensions, $z_c \in R^{D_c \times H\prime \times W\prime}$, leveraging the self-attention mechanism to enrich it with contextual information. To supplement positional information, which the self-attention mechanism alone cannot inherently capture, a fixed positional encoding p $\in R^{D_c \times H\prime \times W\prime}$ [10, 12] is introduced alongside the feature map and fed into the encoder. The resulting encoded feature map, $z_e \in R^{D_c \times H\prime \times W\prime}$, is obtained as $z_e = f_{enc}$ (z_c, p), where f_{enc} (·, ·) represents a series of stacked transformer encoder layers. The transformer decoder takes a set of learnable query vectors Q = $\{q_i | q_i \in R^{D_c}\}_{i=1}^{N_q}$ and transforms them into embeddings D = $\{d_i | d_i \in R^{D_c}\}_{i=1}^{N_q}$ containing image-wide contextual information crucial for HOI detection. These embeddings reference the encoded feature map ze using the attention mechanism. Here, N_q represents the number of query vectors, each designed to capture at most one human-object pair and their interactions. N_q is set to be sufficiently large to accommodate the maximum number of human-object pairs in an image. The resulting decoded embeddings are obtained as D = f_{dec} (z_e, p, Q), where f_{dec} (·, ·, ·) denotes a series of stacked transformer decoder layers. Once again, a positional encoding p is employed to integrate spatial information.

The subsequent interaction detection heads further process the decoded embeddings to generate N_q prediction results. It's important to note that one or more HOIs corresponding to a human-object pair are mathematically defined by four vectors: a human-bounding-box vector normalized by the corresponding image size $b^{(h)} \in [0,1]^4$, a normalized object-bounding-box vector $b^{(o)} \in [0,1]^4$, an object-class one-hot vector c $\in \{0,1\}^{N_{obj}}$, where N_{obj} is the number of object classes, and an action-class vector a $\in \{0,1\}^{N_{act}}$, where $N_{act}$is the number of action classes. It's worth mentioning that a is not necessarily a one-hot vector because there may be multiple actions corresponding to a human-object pair. The interaction detection heads consist of four small feed-forward networks (FFNs): human-bounding-box FFN f_h, object-bounding-box FFN f_o, object-class FFN f_c, and action-class FFN f_a, each dedicated to predicting one of the aforementioned four vectors, respectively.

RESULTS AND DISCUSSIONS

HOTR achieves a new state-of-the-art performance on both V-COCO and HICO-DET datasets, while also being the fastest parallel detector. Table 1 displays the results on the V-COCO dataset for both Scenario1 and Scenario2. HOTR surpasses the state-of-the-art parallel HOI detector [6] in Scenario1 by a margin of 4.2 mAP.

Table 1. Comparison of performance on V-COCO test set.

Method	Backbone	$AP_{role}^{\#1}$	$AP_{role}^{\#2}$
VSRL	R50-FPN	31.8	-
InteractNet	R50-FPN	40.0	48.0
GPNN	R152	44.0	-
iCAN	R50	45.3	52.4
UnionDet	R50-FPN	47.5	56.2
QPIC	R50	**55.2**	**64.4**

Table 2 shows the HICO-DET results in the default configuration for each Full/Rare/Non-Rare class. Because of the inclusion of noisy object labeling in the HICO-DET dataset, improving the pre-trained object detector on the HICO-DET train set provides a useful prior that improves overall performance. As a result, the HICO-DET dataset is evaluated under two conditions: a) using pre-trained weights from MS-COCO that remain fixed during training (referred to as COCO in the Detector column), and b) assessing performance after fine-tuning the pre-trained detector on the HICO-DET train set. Under both situations, the model outperforms the current parallel HOI detector by 4.1 mAP and 4 mAP respectively.

Table 2. Performance comparison in HICO-DET

Method	Detector	Backbone	Feature	Full	Rare	Non-rare
DCA	COCO	R50	A+S	16.24	11.16	17.75
VSGNet	COCO	R152	A+S	19.80	16.05	20.91
VCL	HICO-DET	R50	A+S	23.63	17.21	25.55
PPDM	HICO-DET	HG104	A	21.10	14.46	23.09
QPIC	HICO-DET	R50	A	25.10	17.34	27.42

Table 3 compares the results for V-COCO. Among all baseline approaches, QPIC delivers cutting-edge performance. QPIC with the ResNet-50 backbone achieves a considerable gain of 5.6 mAP (10.5%) over ConsNet, the strongest baseline. Interestingly, contrary in the HICO-DET result, the ResNet-50 backbone outperforms the ResNet-101 backbone, probably due to the restricted number of training samples in V-COCO, which may not be enough to adequately train the larger network. Overall, these comparisons demonstrate QPIC's dataset-invariant effectiveness.

Table 3. Comparison against state-of-the-art methods on V-COCO

Method	Scenario 1	Scenario 2
VCL	48.3	-
DRG	51.0	-
VSGNet	51.8	57.0
ConsNet	53.2	-
QPIC(ResNet-50)	**58.8**	**61.0**
QPIC(ResNet-101)	58.3	60.7

CONCLUSION

To summarize, transformer-based models have considerable advantages in detecting human-object interactions (HOIs) because they efficiently capture long-range relationships and aggregate image-wide contextual information. Despite their computational cost and interpretability issues, transformer-based HOI detectors have exhibited cutting-edge performance and effectiveness in a variety of domains. Ongoing research initiatives aim to address these difficulties using approaches such as model distillation, efficient architectures, and interpretability methods, with the goal of improving scalability and trustworthiness. Transformer-based models have the potential to revolutionize HOI detection in a wide range of practical applications, including robotics and human-computer interaction. Future research could look into hybrid techniques that combine the strengths of transformers with other topologies in

order to improve the efficiency and interpretability of HOI detection systems. Finally, the continuous development of transformer-based models offers transformational results across a wide range of applications, influencing the future of human-centered technology.

REFERENCES

[1] Jitendra Gupta, Saurabh Malik. Visual semantic role labeling. *arXiv preprint arXiv:1505.04474*, 2015.

[2] Chen Gao, Yuliang Zou, and Jia-Bin Huang. ican: Instancecentric attention network for human-object interaction detection. arXiv preprint arXiv:1808.10437, 2018.

[3] Yong-Lu Li, Siyuan Zhou, Xijie Huang, Liang Xu, Ze Ma, Hao-Shu Fang, Yanfeng Wang, and Cewu Lu. Transferable interactiveness knowledge for human-object interaction detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 3585–3594, 2019.

[4] Yong-Lu Li, Liang Xu, Xinpeng Liu, Xijie Huang, Yue Xu, Shiyi Wang, Hao-Shu Fang, Ze Ma, Mingyang Chen, and Cewu Lu. Pastanet: Toward human activity knowledge engine. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 382–391, 2020.

[5] Oytun Ulutan, ASM Iftekhar, and Bangalore S Manjunath. Vsgnet: 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, 2020.

[6] TiancaiWang, Tong Yang, Martin Danelljan, Fahad Shahbaz Khan, Xiangyu Zhang, and Jian Sun. Learning human-object interaction detection using interaction points. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4116–4125, 2020.

[7] Yue Liao, Si Liu, FeiWang, Yanjie Chen, Chen Qian, and Jiashi Feng. 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, 2020.

[8] Bumsoo Kim, Taeho Choi, Jaewoo Kang, and Hyunwoo Kim. Uniondet: Union-level detection towards real-time human-object interaction detection. In *Proceedings of the European conference on computer vision (ECCV)*, 2020.

[9] Kim, B., Lee, J., Kang, J., Kim, E. S., & Kim, H. J. (2021). Hotr: End-to-end human-object interaction detection with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 74-83).

[10] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. *arXiv preprint arXiv:2005.12872*, 2020.

[11] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In NIPS, December 2017.

[12] Tamura, M., Ohashi, H., & Yoshinaga, T. (2021). Qpic: Query-based pairwise human-object interaction detection with image-wide contextual information. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 10410-10419).

[13] N Nasimova, B Muminov, R Nasimov, K Abdurashidova, M Abdullaev. Comparative Analysis of the Results of Algorithms for Dilated Cardiomyopathy and Hypertrophic Cardiomyopathy Using Deep Learning. 2021 International Conference on Information Science and Communications Technologies (ICISCT) (2021): pp. 1-5.

[14] Abdurashidova, K., Abdukhakimov, F., & Chorshanbiyeva, S. (2024). UNDERSTANDING HUMAN INTERACTION USING GRAPH NEURAL NETWORKS: AN OVERVIEW. *Talqin va tadqiqotlar*, (4 (41)).