Exploring Predicate Visual Context in Detecting of Human–Object Interactions

Frederic Z. Zhang[1†] Yuhui Yuan[2] Dylan Campbell[1] Zhuoyao Zhong[2] Stephen Gould[1]
[1]The Australian National University [2]Microsoft Research Asia
frederic.zhang@anu.edu.au yuhui.yuan@microsoft.com
https://github.com/fredzzhang/pvic

A. Bounding Box Pair Positional Embeddings

We provide more mathematical details on the positional embeddings for bounding box pairs used in cross-attention. Let us first revisit the notations from the main paper. We define sinusoidal embedding of a scalar as $\boldsymbol{\phi} : \mathbb{R} \rightarrow \mathbb{R}^d$,

$$\boldsymbol{\phi}(x)_{2i} = \sin\left(\frac{x}{\tau^{2i/d}}\right), \ \boldsymbol{\phi}(x)_{2i-1} = \cos\left(\frac{x}{\tau^{2i/d}}\right), \quad (1)$$

where $i = 1, \dots, d/2$ and τ is a temperature parameter we set as 20. With box width and height as modulation, the positional embeddings for one bounding box are as follows

$$\text{PE}(x, y, w, h) = \left[\boldsymbol{\phi}(y)\frac{h_{ref}}{h}, \boldsymbol{\phi}(x)\frac{w_{ref}}{w}\right] \in \mathbb{R}^{2d}, \quad (2)$$

where w_{ref}, h_{ref} are reference values learned from box appearance features $\mathbf{f}$ as follows

$$w_{\text{ref}}, h_{\text{ref}} = \sigma(\text{MLP}(\mathbf{f})), \quad (3)$$

where σ is the sigmoid function. As such, the positional embeddings for a human–object pair ($\mathbf{b}_h, \mathbf{b}_o \in \mathbb{R}^4$) are defined by concatenating the positional embeddings of the two boxes,

$$\mathbf{q}_p = \left[\mathbf{q}_p^h, \mathbf{q}_p^o\right] = [\text{PE}(\mathbf{b}_h), \text{PE}(\mathbf{b}_o)] \in \mathbb{R}^{4d}. \quad (4)$$

Denote the positional embeddings of an image patch with normalised spatial indices (i, j) by

$$\mathbf{k}_p = [\boldsymbol{\phi}(j), \boldsymbol{\phi}(i)] \in \mathbb{R}^{2d}. \quad (5)$$

Assuming the number of heads is one, the dot-product attention weights between positional embeddings are computed as

$$(W_k \mathbf{k}_p)^T (W_p \mathbf{q}_p) = \mathbf{k}_p^T W_k^T W_p \mathbf{q}_p, \quad (6)$$

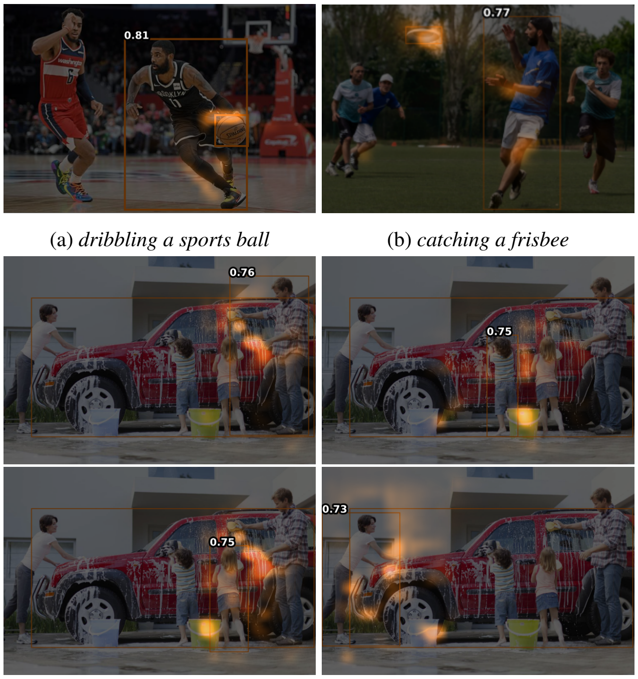
where $W_k \in \mathbb{R}^{2d \times 2d}, W_p \in \mathbb{R}^{2d \times 4d}$ are weight matrices associated with the linear transformations applied to the positional embeddings. In particular, matrix W_p can be partitioned into $\left[W_p^h, W_p^o\right]$, therefore decomposing the linear transformation on query (human–object pair) positional embeddings as follows

$$W_p \mathbf{q}_p = W_p^h \mathbf{q}_p^h + W_p^o \mathbf{q}_p^o. \quad (7)$$

For brevity of exposition, let us now assume that weight matrices $W_k, W_p^h, W_p^o \in \mathbb{R}^{2d \times 2d}$ are identity matrices. This simplifies the dot-product attention weights between positional embeddings in Eq. 6 as follows

$$\mathbf{k}_p^T \mathbf{q}_p^h + \mathbf{k}_p^T \mathbf{q}_p^o, \quad (8)$$

demonstrating that the concatenation of two modulated box positional embeddings results in a weighted sum of the pre-normalised attention weights.



(a) *dribbling a sports ball*

(b) *catching a frisbee*

(c) *washing a car*

Figure 1. Predicted human–object interactions and visualised attention weights on data in the wild.

[†] Work done at Microsoft Research Asia.

B. Demonstration on Data in The Wild

For more evidence on the two types of visual context exploited by our model, we provide additional qualitative results on data in the wild, with cross-attention weights overlaid. As shown in Figures 1a and 1b, the model extracts contextual features from image regions containing relevant human body parts, and successfully predicts the correct interactions with high scores. In Figure 1c, we highlight the other type of context, i.e. another involved object. Notably, three out of the four human–object pairs place high attention weights on the water buckets, which are indicators of the corresponding interaction.

C. Pipeline

For better clarity, we attach an illustration of the entire pipeline, as shown in Figure 2. Due to the popularity of the DETR framework, the first stage is depicted as a transformer-based object detector. But the method itself is detector-agnostic.

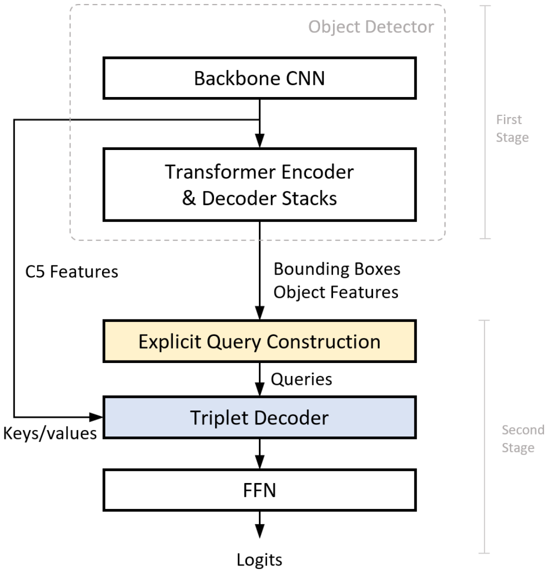


Figure 2. Illustration of the overall pipeline.