Video Interaction Recognition using an Attention Augmented Relational Network and Skeleton Data

Supplementary Material

1. Relative features (spatial object)

Following the new formulation of spatial objects above, we define appropriate distance and motion information as follows:

$$D_s(s_i) = (\| c_{i1}^{p1} - c_{i1}^{p2} \|, \| c_{i2}^{p_1} - c_{i2}^{p2} \|, ..., \| c_{iT}^{p1} - c_{iT}^{p2} \|) \tag{1}$$

$$M_s(s_i) = (\| c_{i1}^{p1} - c_{i2}^{p1} \|, \| c_{i2}^{p_1} - c_{i3}^{p1} \|, ..., \| c_{iT-1}^{p1} - c_{iT}^{p1} \| \frown \| c_{i1}^{p2} - c_{i2}^{p2} \|, \| c_{i2}^{p_2} - c_{i3}^{p2} \|, ..., \| c_{iT-1}^{p2} - c_{iT}^{p2} \|) \tag{2}$$

$$L_s(s_i) = (\| c_{i1}^{p1} - c_{i2}^{p2} \|, \| c_{i2}^{p_1} - c_{i3}^{p2} \|, ..., \| c_{iT-1}^{p1} - c_{iT}^{p2} \|) \tag{3}$$

in the formulation above, $D_s(s_i)$ represents the distance between the i^th joints of two actors over timesteps. $M_s(s_i)$ captures the motion each actor's i^th joint over the timesteps. For more details and notation see Sec. 3.1 of the main paper. While forming an object pair (i.e., relations) between s_i and s_j we calculate $D_s(s_i)$, $D_s(s_j)$, $M_s(s_i)$,$M_s(s_j)$, $L_s(s_i)$, and $L_s(s_j)$ and concatenate them with the relation.While forming an object pair (i.e., relations) between s_i and s_j we calculate $D_s(s_i)$, $D_s(s_j)$, $M_s(s_i)$, $M_s(s_j)$, $L_s(s_i)$, and $L_s(s_j)$; and concatenate them with the relation.

2. Relative features effect

Tab. 1 demonstrates the impact of individual relative features (i.e., D, M, L) and their combination on the classification accuracy for NTU RGB+D dataset. It is important to note that these experiments utilize our original temporal objects (rather than the baseline spatial objects). In the Tab. 1 L means the relative feature L is concatenated with the temporal object and $D \frown L$ means both relative features D and L are concatenated with the temporal object. For more details see Sec. 3.1 of the mian paper. As evident from the results, among the individual relative features, M (intra-motion) is the most effective. Furthermore, its combination with D (distance) forms the most effective pairing. Lastly, the concatenation of all three yields best performance.

Method	NTU RGB+D	
	X-Sub (%)	X-View (%)
L	87.98 ± 0.33	90.82 ± 0.14
D	88.61 ± 0.09	91.35 ± 0.08
M	89.07 ± 0.14	91.66 ± 0.11
$D \frown L$	88.79 ± 0.07	91.92 ± 0.11
$M \frown L$	89.75 ± 0.12	92.37 ± 0.08
$D \frown M$	89.97 ± 0.16	92.94 ± 0.49
AARN ($D \frown M \frown L$)	90.79 ± 0.65 (91.26)	93.42 ± 0.65 (93.88)

Table 1. Interaction classification accuracy (interaction classes only). The impact of relative features individually and in pair on classification accuracy.

3. Transmotion architecture

In this section we present the detailed architecture of the Transmotion attention module. Fig. 1 demonstrates the architecture.

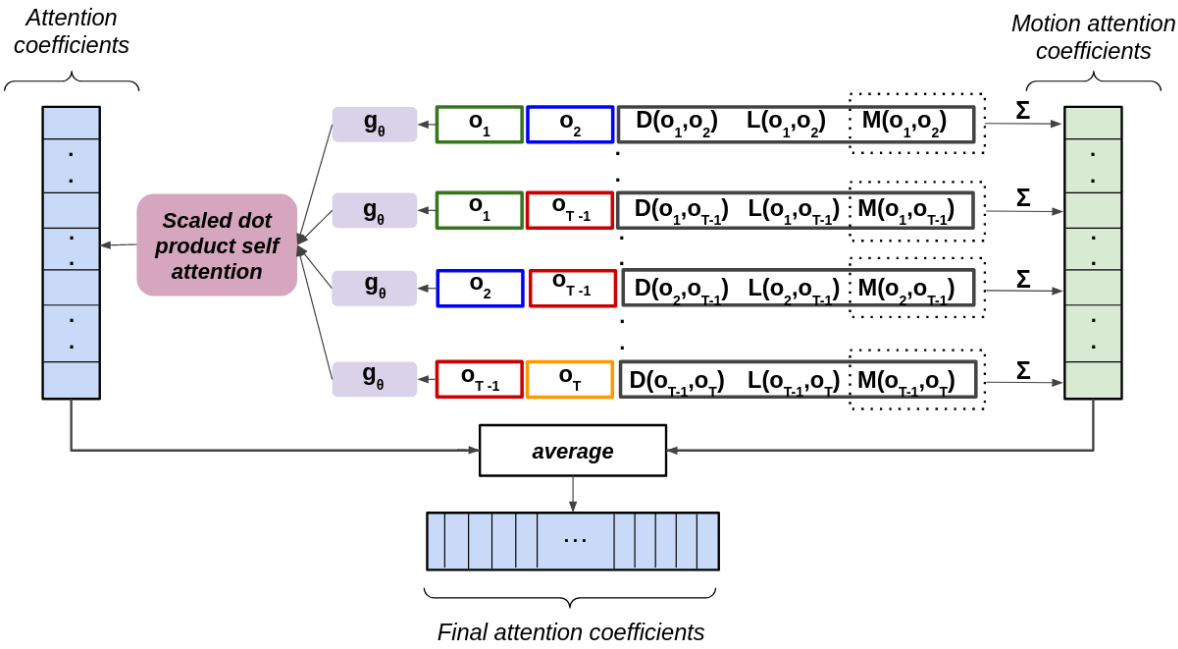

Figure 1. Transmotion module as a baseline. This is the integration function that combines scale dot product self attention with motion coefficients from Eq. (4) of the main paper. The final attention coefficients are produced by averaging the attention coefficients generated by each mechanism.

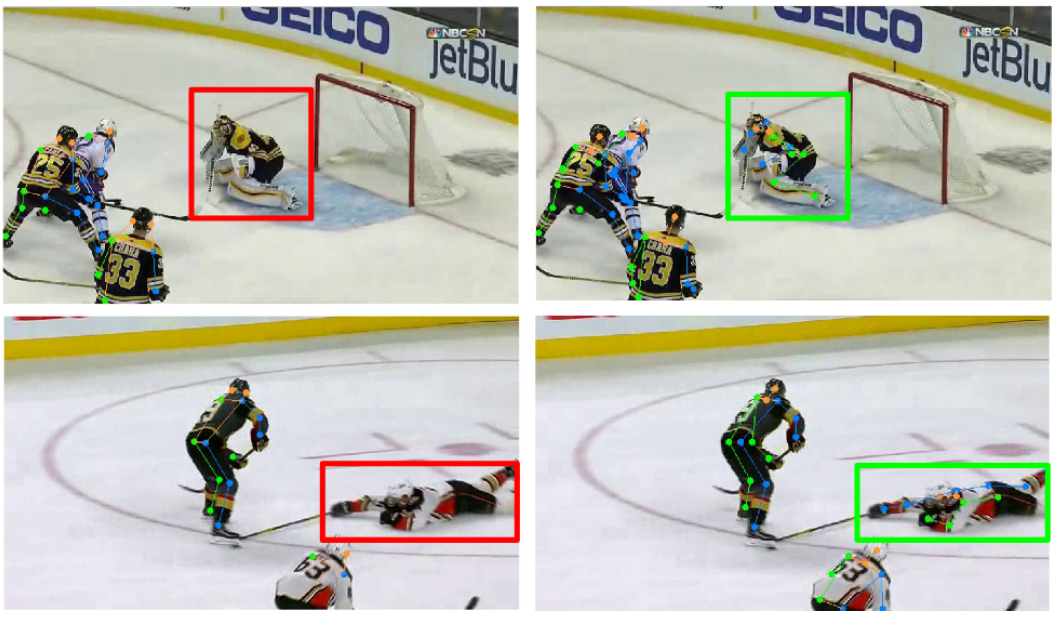

Figure 2. Fine-tuning pose estimators enables them to extract difficult poses such as a hockey goaltender with unusual posture and oversized jersey (top) and a fallen player wearing white jersey blending with the ice (bottom). Left: top-down pose estimator pretrained on COCO, right: the same model fine-tuned on HPD. Red and green bounding boxes indicate inaccurate and accurate poses, respectively