Supplementary Material for "Graph-based High-order Relation Modeling for Long-term Action Recognition"

Jiaming Zhou[1,5], Kun-Yu Lin[1], Haoxin Li[3], Wei-Shi Zheng[1,2,4*]
[1]School of Computer Science and Engineering, Sun Yat-sen University, China
[2]Peng Cheng Laboratory, Shenzhen 518005, China
[3]School of Electronics and Information Technology, Sun Yat-sen University, China
[4]Key Laboratory of Machine Intelligence and Advanced Computing, Ministry of Education, China
[5]Pazhou Lab, Guangzhou, China
jia_ming_zhou@outlook.com, kunyulin14@outlook.com, lihaoxin05@gmail.com, wszheng@ieee.org

1. Channel-shared GHRM

In order to exploit the high-order relations in the long-term actions, our proposed Graph-based High-order Relation Modeling (GHRM) module aims to incorporate the information from all the other graphs through the adjacent matrix $\mathcal{A}^i$ and the embedding layer ρ^i in the i-th graph $\mathcal{G}^i$ during graph reasoning for the i-th basic relation.

However, the number of parameters of the two trainable layers g^i (used to construct the adjacent matrix $\mathcal{A}^i$), ρ^i in the i-th graph are both $K \cdot C^2$, which means that if we want to model K basic relations in the long-term actions, we have to use K graphs, thus the number of parameters of our GHRM will increase to K times that of Vanilla-GCN. In order to make our model more lightweight and prevent overfitting, we adopt a channel sharing strategy on the layer g^i and the layer ρ^i during graph reasoning on each graph $\mathcal{G}^i$, which enables our GHRM to efficiently incorporate the information from all the other graphs in a channel-shared manner. In the following, we will show how to construct the adjacent matrix $\mathcal{A}^i$ in a channel-shared manner and the details of the channel-shared embedding layer ρ^i in each graph $\mathcal{G}^i$.

- Channel-shared Construction of Adjacent Matrix $\mathcal{A}^i$. To construct the adjacent matrix $\mathcal{A}^i$ in a channel-shared manner, for the u-th segment node $x_u^i \in \mathbb{R}^C$ and the v-th segment node $x_v^i \in \mathbb{R}^C$ in the i-th graph, if the edge $\mathcal{E}^i_{(u,v)}$ in $\mathcal{E}^i$ is 1 which means there is an edge between segment nodes x_u^i and x_v^i, GHRM will calculate the connection strength value $\mathcal{A}^i_{(u,v)}$ between them[1] by applying the channel-shared layer g^i as follows:

$$\begin{aligned}\mathcal{A}^i_{(u,v)} &= \frac{\exp((\tilde{x}_u)^T \tilde{x}_v)}{\sum_{w=1}^{T} \exp((\tilde{x}_u)^T \tilde{x}_w)},\\ \tilde{x}_u &= [g^i(x_u[0]), g^i(x_u[1]), \cdots, g^i(x_u[C-1])]^T,\\ x_u &= [x_u^1, \cdots, \beta^i(x_u^i), \cdots, x_u^K],\end{aligned} \tag{1}$$

where $x_u \in \mathbb{R}^{C\times K}$ is concatenated from the u-th segment node in all graphs. β^i is the embedding layer for the segment nodes in the i-th graph. $g^i : \mathbb{R}^K \to \mathbb{R}$ is the trainable layer in the i-th graph, which transforms feature x_u into feature $\tilde{x}_u \in \mathbb{R}^C$ in the i-th relation space in a channel-shared manner (*i.e.*, layer g^i is shared on the channel dimension). Therefore, by using the channel-shared layer g^i, our model is more lightweight, and the high-order relations will be exploited as the information from all the other graphs can still be incorporated when constrcuting the adjacent matrix $\mathcal{A}^i$ of the i-th graph.

- Channel-shared Embedding Layer ρ^i. Similarly, we can reformulate the embedding layer ρ^i in the i-th graph in a channel-shared manner as follows:

$$\begin{aligned}\rho^i(X_{agg}^1, \cdots, X_{agg}^K) &= \delta([X_{agg}^{(0)}\mathcal{W}^i, \cdots, X_{agg}^{(C-1)}\mathcal{W}^i]),\\ X_{agg}^{(c)} &= [X_{agg}^1[:,c], \cdots, \gamma^i(X_{agg}^i)[:,c], \cdots, X_{agg}^K[:,c]],\end{aligned} \tag{2}$$

where $X_{agg}^i \in \mathbb{R}^{T\times C}$ is the aggregated node feature in the i-th graph, $X_{agg}^{(c)} \in \mathbb{R}^{T\times K}$ is concatenated from the c-th channel of the aggregated node feature in all graphs. δ is a nonlinear function. $\mathcal{W}^i \in \mathbb{R}^{K\times 1}$ is the parameter of the channel-shared embedding layer ρ^i in the i-th graph. γ^i is the embedding layer for the aggregated node feature X_{agg}^i in the i-th graph. When embedding the i-th aggregated node feature X_{agg}^i in the i-th graph, the embedding layer ρ^i incorporates the information from all the other graphs in a

*Corresponding author

[1]If the edge $\mathcal{E}^i_{(u,v)}$ is 0, the connection strength value $\mathcal{A}^i_{(u,v)}$ between these two nodes is set to 0.

channel-shared manner, thus the high-order relations can be exploited.

According to the descriptions above, the number of parameters in the two channel-shared trainable layers g^i, ρ^i are both K, which is $\frac{1}{C^2}$ times that of original GHRM and $\frac{K}{C^2}$ times that of Vanilla-GCN. The channel sharing policy makes our GHRM more efficient while keeping the ability to incorporate the information from all the other graphs when graph reasoning on each graph, thus the high-order relations in the long-term actions can still be well exploited.

2. More Visualizations of Adjacent Matrices

In order to more intuitively understand the advantages of the proposed GHRM over Vanilla-GCN, we provide more visualizations of the adjacent matrices in Vanilla-GCN and GHRM. We randomly select two video samples from Charades [1], and use 16 graphs in both Vanilla-GCN and GHRM. The visualizations are shown in Figure 1 and Figure 2. We can see that the adjacent matrices in Vanilla-GCN can only model similar patterns, while the adjacent matrices in our GHRM can model diverse patterns. This phenomenon indicates that different graphs can interact with each other in GHRM, such that different basic relations will be figured out and the high-order relations in the long-term actions can be naturally exploited.

3. Analysis on the Window Size W in Temporal-GHRM

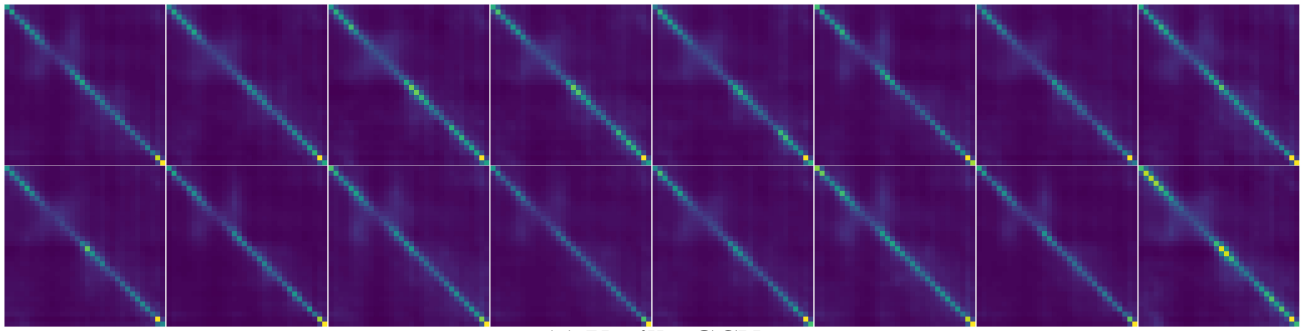
The Temporal-GHRM branch in the GHRM-layer aims to model the local temporal high-order relations in the long-term actions in a local manner, which restricts the number of temporal neighboring nodes of each node to window size W. Here we analyze the effects of using different window size W in the Temporal-GHRM branch on Charades. As shown in Table 1, using a larger window size generally brings better results, and our model achieves the best result when the window size is 7, which means that connecting each node with its six neighboring nodes is the most suitable for modeling the local temporal high-order relations.

Window Size W	mAP(%) on Charades
W=3	36.3
W=5	36.7
W=7	**37.3**
W=9	37.1

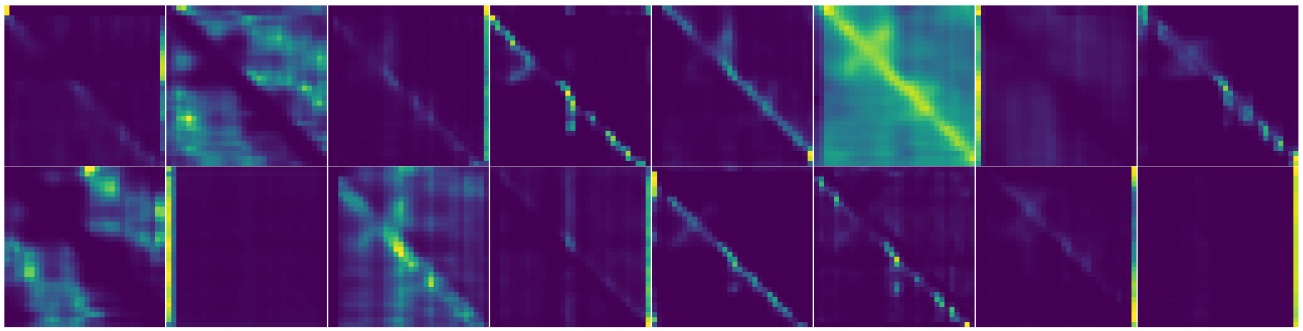
Table 1. **Results of using different window size W on Charades.** Our model achieves the best result when the window size is 7.

References

[1] Gunnar A. Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *European Conference on Computer Vision*, 2016. 2

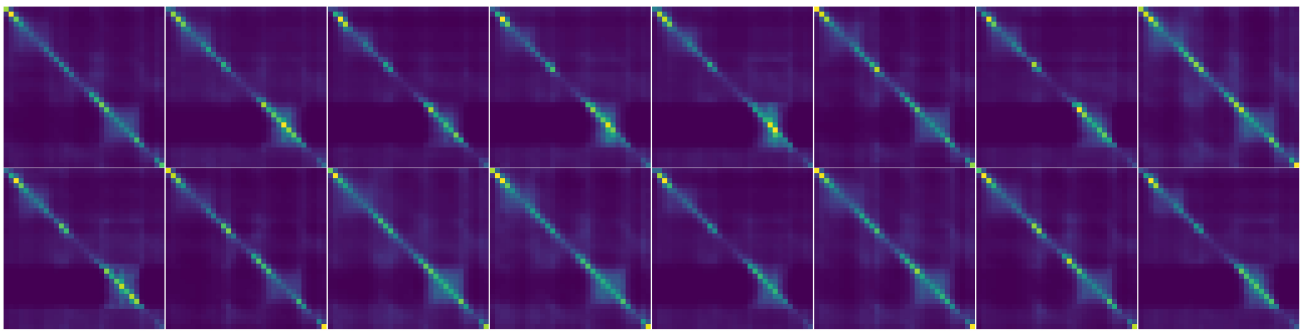


(a) Vanilla-GCN

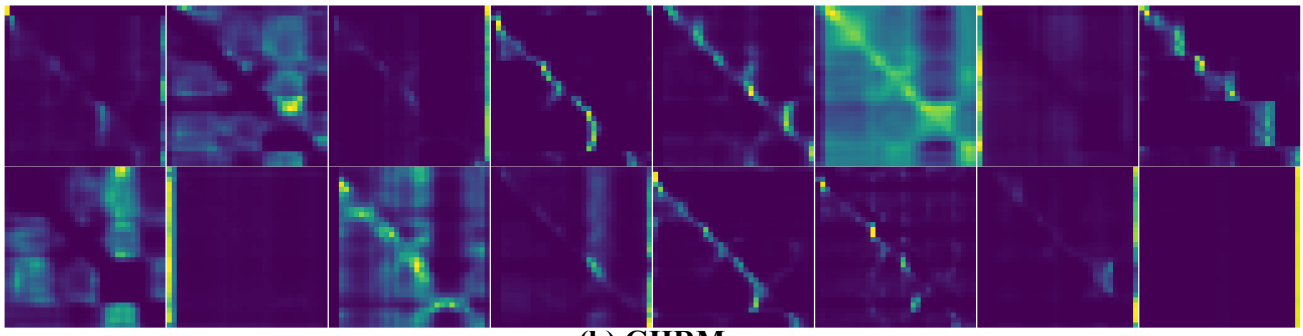


(b) GHRM

Figure 1. Visualization of the adjacent matrices for the first video sample. Figure (a) shows 16 adjacent matrices in Vanilla-GCN, and Figure (b) shows 16 adjacent matrices in GHRM.



(a) Vanilla-GCN



(b) GHRM

Figure 2. Visualization of the adjacent matrices for the second video sample. Figure (a) shows 16 adjacent matrices in Vanilla-GCN, and Figure (b) shows 16 adjacent matrices in GHRM.