Supplementary Material for “A Hybrid Egocentric Activity Anticipation Framework via Memory-Augmented Recurrent and One-shot Representation Forecasting”

Tianshan Liu and Kin-Man Lam
Department of Electronic and Information Engineering
The Hong Kong Polytechnic University
tianshan.liu@connect.polyu.hk, enkmlam@polyu.edu.hk

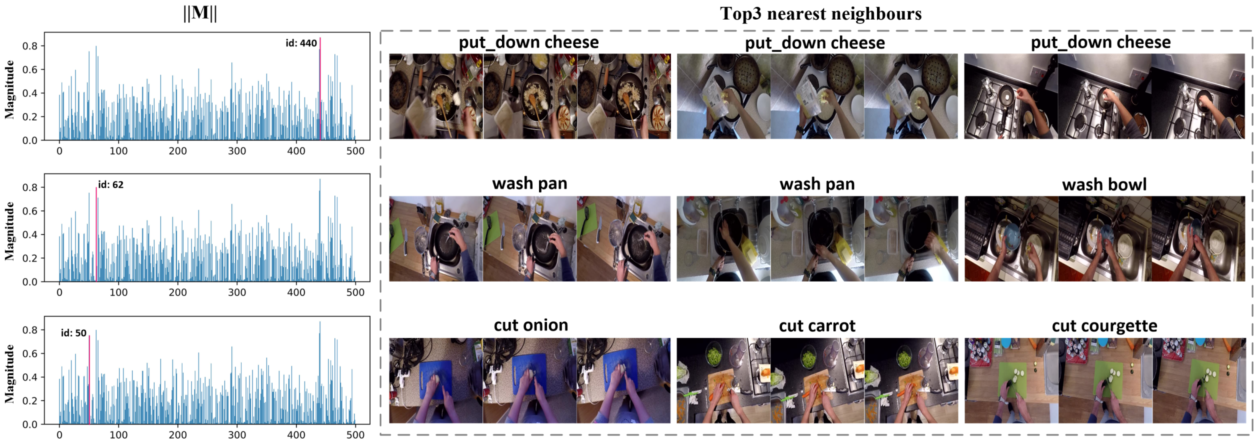

Figure 1. Visualizations of the samples that highly rely on the $\{440, 62, 50\}$-th memory items for forecasting the representations on the EK-55 validation set. The left illustrates the magnitude of the memory bank, $||\mathbf{M}|| \in \mathbb{R}^{500}$, and the selected memory items are highlighted in magenta. The right shows the Top-3 nearest neighbours retrieved by the corresponding memory item.

1. Visualization of Learned Memory Bank

To obtain a deeper understanding of the proposed memory bank, as shown in Fig. 1, we visualize several representative activated memory items when forecasting the representations of future activities during the inference phase. Generally, the memory items stored in the memory bank encode discriminative motion prototypes, e.g., “wash” (ID:62) and “cut” (ID:50). These encoded prototypes can regularize the forecasted (reconstructed) representations without deviating to uncontrollable motion patterns. Moreover, some memory items can even capture the specific motion-object (verb-noun) interactions, e.g., “put_down cheese” (ID:440), which provides more reliable prototypical activity semantics in recurrent feature anticipation.

2. Discussions on Negative Sampling Strategy

To guarantee the effectiveness of contrastive learning, the negative sampling strategy should ensure the diversity and the similarity of the negative samples compared to the positive sample. For the egocentric activity anticipation task, the original long untrimmed video is split into multiple clips according to the activity labels. Thus, these randomly sampled negatives for calculating the contrastive loss may come from videos that have the same or different video IDs as the positive sample. It implies that the positive sample and negative samples are recorded by the same actor and in the same scene, when they come from the same long video. These negative samples can be regarded as “hard negatives”, as they contain similar (but distinct) semantic information to the target positive sample. Minimizing the contrastive loss among these hard negatives forces the model to be aware of subtle yet crucial semantic differences between different egocentric activities.