Supplementary Material for Learning to Anticipate Future with Dynamic Context Removal

Xinyu Xu¹, Yong-Lu Li¹,², Cewu Lu¹ *
¹Shanghai Jiao Tong University ² Hong Kong University of Science and Technology
{xuxinyu2000, yonglu_li, lucewu}@sjtu.edu.cn

1. License

1.1. Dataset

We use four datasets in this work. They are EPIC-KITCHENS-100 [10], EPIC-KITCHENS-55 [11], EGTEA GAZE+ [19] and 50-Salads [23].
EPIC-KITCHENS-100 [10] and EPIC-KITCHENS-55 [11] are copyright by the same team and published under the Creative Commons Attribution-NonCommercial 4.0 International License [6]. We download data from its website [7].
EGTEA GAZE+ [19] is publicly available, and no license is specified. We download data from its website [5].
50-Salads [23] is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License [1]. We download data from its website [2].

1.2. Prior Work

We sincerely thank prior work RULSTM [14] and AVT [15] for their generous release of checkpoints and pre-extracted feature which greatly helps our experiments.
RULSTM [14] is publicly released in [8], and no license is specified.
AVT [15] is publicly released in [4], and licensed under the Apache License 2.0 [3].

2. Baseline

Deep Multimodal Regressor (DMR) [27] applies an unsupervised training scheme to minimize the representation gap between current observation and the multimodal future via deep regression network. Then the anticipative feature is sent to classifier to give results.
Anticipation TSN (ATSN) [11] is a variant of TSN [28] that has same model architecture but different input segment. The observed segment is sent to TSN architecture for a simple classification using future action label.

Verb-Noun Marginal Cross Entropy Loss (MCE) [13] is an effective loss function to boost the anticipation performance of ATSN. It focuses on predicting verb-noun composed action label, but still follows marginal constraints.
Forecasting HOI (FHOL) [20] adopts intentional hand movement and jointly predicts the egocentric hand motion, interaction hotspots and future action.
RULSTM [14] is the winner of EK55 2019 anticipation challenge [11]. It utilizes one rolling LSTM to summarize the past and another unrolling LSTM to anticipate future.
Modality attention mechanism (MATT) is proposed to make multi-modal prediction.
ImagineRNN [29] attempts to anticipate future by imagination. It narrows the gap of observation and action execution with an imagined intermediate and further improves performance by the residual anticipation.
ActionBanks [22] is the winner of EK55 2020 anticipation challenge [11]. It leverages different levels of past aggregation representation via attention mechanism to improve the anticipation performance.
Ego-OMG [12] annotates extra information about hand segmentation and (next) active objects to serve as intermediate knowledge in anticipation. It is encoded by graph network and LSTM then ensembled with additional CSN [25] branch to give the prediction.
Anticipative Video Transformer (AVT) [15] is a recent work as well as the winner of EK100 2021 anticipation challenge [10]. It proposes an end-to-end transformer [26] based architecture with causal attention on the head to anticipate future in the seq2seq manner.

3. Model Ensemble

We simply last fuse results from different models to give prediction on EPIC-KITCHENS [10, 11] series, and it also performs more complex fusion methods [14, 16]. Noticeably, We only use the transformer [26] version DCR in fusion. For EK100 validation set, we fuse TSM-DCR, TSN-DCR, FRCNN-DCR with weight 1:1:1. For EK55 validation set, we fuse TSM-DCR, irCSN152-DCR, TSN-DCR,
4. Additional Training Detail

We present additional training details as a supplement to Sec. 4.3, main text. The default transformer architecture starts with order-aware pre-training. In this phase, for all datasets, we set batch size 512 and optimize the network for 50 epoch with base learning rate is 1e-4. Then, the next stage is reconstructing future with dynamic context. We customize hyper-parameters for different datasets in Tab. 1. We can conclude some interesting empirical results in hyper-parameter tuning, e.g. small dataset suffers more from future uncertainty and increasing label smoothing [24] level is beneficial. We conduct LSTM [17] experiments on EK100 [10]. It’s not applied with pre-training but directly starts with the second stage. It is optimized with base learning rate 1e-2 for 100 epochs. We set batch size = 512, λcls=1, λrec=1, smoothing factor ϵ=0.2.

5. Additional Ablation Study

We give more ablation study of DCR-LSTM in Tab. 2. It has a little difference with the transformer reasoner since the order-aware pre-training is not used. Comparing with FRCNN-DCR with weight 1:1:1:1. The submissions to online test server is more challenging since the huge computation cost and additional training data of competitive baselines. Thus, we leverage the public model zoo from prior work AVT [15], which makes seven model ensemble and achieves state-of-the-art. On EK100 test set, we fuse results of TSM-DCR, TSN-DCR, AVT with weight 1:0.5:1. On EK55, we fuse TSM-DCR, irCSN152-DCR, AVT with weight 1:1:1 for test set S1, while 0.5:1.5:1.5 for test set S2. Results are listed in main text.

6. Easiness Schedule

Our instance-specific easiness schedule is visualize in Fig. 1. In each training epoch, we bound easiness range with minimal and maximal Te value among all instances and compute the average value as the red line in the figure. We can observe one fast easiness decreasing line as the bottom of the blue area while another slow easiness decreasing line as the upper bound. This indicates different difficulty in reasoning different video clips thus a finer-grained schedule is necessary with empirical supports.

7. Early Action Recognition

As a general training strategy, DCR has the potential to support a wider range of temporal predictive applications, not limited to action anticipation. We take 50% observation early action recognition as an example, i.e. recognizing video action based on 50% part video clips. Our method can have simple migration by modifying frame setting in Fig.3 and Eq.2 of main text. Detailed, we sample 40 frames from each action clip. First 50% frames are constant observation (blue in Fig.3 main text). Other 50% frames have results in Tab.7 main text, we can conclude some properties between transformer and LSTM. First, LSTM is more robust with context as it doesn’t collapse when we set Te = 1 in training but Te = 0 in testing. Second, for LSTM, global easiness schedules (e.g. exponential, linear) can achieve comparable performance with the local schedule in DCR. Third, transformer benefits more from feature level supervision but LSTM is less sensitive.
<table>
<thead>
<tr>
<th>Visible Context</th>
<th>Action Segment</th>
<th>Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>$T_e = 1$</td>
<td>![Cut Sandwich]</td>
<td>Cut Sandwich</td>
</tr>
<tr>
<td>$T_e = 0$</td>
<td>![Sandwich]</td>
<td></td>
</tr>
<tr>
<td>$T_e = 1$</td>
<td>![Throw Rubbish]</td>
<td>Throw Rubbish</td>
</tr>
<tr>
<td>$T_e = 0$</td>
<td>![Rubbish]</td>
<td></td>
</tr>
<tr>
<td>$T_e = 1$</td>
<td>![Wash Container]</td>
<td>Wash Container</td>
</tr>
<tr>
<td>$T_e = 0$</td>
<td>![Container]</td>
<td></td>
</tr>
<tr>
<td>$T_e = 1$</td>
<td>![Place Lid]</td>
<td>Place Lid</td>
</tr>
<tr>
<td>$T_e = 0$</td>
<td>![Lid]</td>
<td></td>
</tr>
</tbody>
</table>

Figure 2. Quantitative Cases. Transparent frames are not visible.

dynamic visibility (orange). One additional [CLS] is employed to predict label (yellow). Notably, we use the reconstruction quality of 1s future since last observation as easiness scheduling criterion in anticipation task in Eq.3 of main text. But it may not accessed in early action recognition scenario, we use the last frame reconstruction instead in a few exception cases. We conduct an experiment on EK55 [11], with the same setting following RULSTM [14]. We train top transformer on RGB-TSN, FLOW-TSN, OBJ-FRCNN three backbones then late-fuse by 1:1:1 to obtain the final prediction. Results are listed in Tab. 3. We outperform baselines by a clear margin, validating the generalization ability of our method.

8. Cases

We show cases of anticipation task with easiness $T_e = 0$ or $T_e = 1$ in Fig. 2. Some key frames are not visible in the difficult mode of $T_e = 0$.

References

[1] 50-salads license. [HTTP](http://creativecommons.org/licenses/by-nc-sa/4.0/).
[2] 50-salads website. [HTTP](http://cvip.coming.dundee.ac.uk/datasets/foodpreparation/50salads/).
[6] Epic-kitchens license. [HTTP](https://creativecommons.org/licenses/by-nc/4.0/).
[10] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Jian Ma, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Rescaling egocentric vision: Collection,
pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision (IJCV)*, 2021. 1, 2


[22] Fadime Sener, Dipika Singhania, and Angela Yao. Temporal aggregate representations for long-range video understanding. 2020. 1


