Supplementary Material for Learning to Anticipate Future with Dynamic Context Removal

Xinyu Xu[1], Yong-Lu Li[1,2], Cewu Lu[1] *
[1]Shanghai 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 (FHOI) [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,

*Cewu Lu is the corresponding author, member of Qing Yuan Research Institute and MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University, China and Shanghai Qi Zhi institute.

	ϵ	λ_{cls}	λ_{rec}	learning rate	batch size	epoch
EK100 [10]	0.2	0.5	1	1e-4	128	100
EK55 [11]	0.4	1	1	1e-4	128	100
EG+ [19]	0.4	0.5	1	5e-5	512	50
50S [23]	0.5	0.5	2	5e-5	64	50

Table 1. Hyper-parameters for DCR training details with transformer [26] head.

	EK100 [10] TSM	EK100 [10] TSN
DCR	**15.2**	**14.5**
classification	14.0	13.5
T_e = 1	14.1	13.1
T_e = 0	14.6	13.9
linear T_e	15.0	14.2
exponential T_e	**15.2**	14.4
w.o. L_{rec}	14.5	13.8
w.o. label smooth	14.0	13.3

Table 2. Ablation study of LSTM [17] version DCR.

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.

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, λ_{cls}=1, λ_{rec}=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 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 collaspe when we set $T_e = 1$ in training but $T_e = 0$ 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.

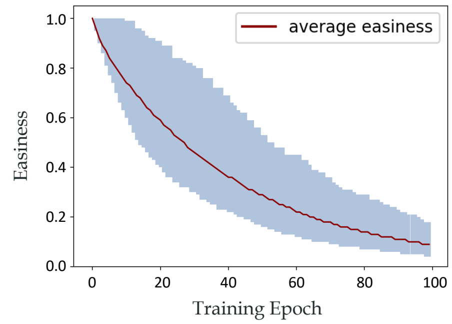

Figure 1. Easiness schedule for DCR training on EK100 [10].

	Top-1 Acc
TCN [9]	19.1
RL [21]	28.6
EL [18]	27.5
RULSTM [14]	30.9
DCR	**33.0**

Table 3. Results of early action recognition with 50% observation on EK55.

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 T_e 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

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://creativecommons.org/licenses/by-nc-sa/4.0/. 1

[2] 50-salads website. https://cvip.computing.dundee.ac.uk/datasets/foodpreparation/50salads/. 1

[3] Avt license. http://www.apache.org/licenses/. 1

[4] Avt website. https://github.com/facebookresearch/AVT. 1

[5] Egtea gaze+ website. http://cbs.ic.gatech.edu/fpv/#egtea_gaze_plus. 1

[6] Epic-kitchens license. https://creativecommons.org/licenses/by-nc/4.0/. 1

[7] Epic-kitchens website. https://epic-kitchens.github.io/. 1

[8] Rulstm website. https://github.com/fpv-iplab/rulstm. 1

[9] Shaojie Bai, J Zico Kolter, and Vladlen Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*, 2018. 2

[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

[11] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 720–736, 2018. 1, 2, 3

[12] Eadom Dessalene, Chinmaya Devaraj, Michael Maynord, Cornelia Fermuller, and Yiannis Aloimonos. Forecasting action through contact representations from first person video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, page 1–1, 2021. 1

[13] Antonino Furnari, Sebastiano Battiato, and Giovanni Maria Farinella. Leveraging uncertainty to rethink loss functions and evaluation measures for egocentric action anticipation. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018. 1

[14] Antonino Furnari and Giovanni Farinella. Rolling-unrolling lstms for action anticipation from first-person video. *IEEE transactions on pattern analysis and machine intelligence*, 2020. 1, 2, 3

[15] Rohit Girdhar and Kristen Grauman. Anticipative Video Transformer. In *ICCV*, 2021. 1, 2

[16] Xiao Gu, Jianing Qiu, Yao Guo, Benny Lo, and Guang-Zhong Yang. Transaction: Icl-sjtu submission to epic-kitchens action anticipation challenge 2021. *arXiv preprint arXiv:2107.13259*, 2021. 1

[17] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 2

[18] Ashesh Jain, Avi Singh, Hema S Koppula, Shane Soh, and Ashutosh Saxena. Recurrent neural networks for driver activity anticipation via sensory-fusion architecture. In *2016 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3118–3125. IEEE, 2016. 2

[19] Yin Li, Miao Liu, and James M. Rehg. In the eye of beholder: Joint learning of gaze and actions in first person video. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018. 1, 2

[20] Miao Liu, Siyu Tang, Yin Li, and James M Rehg. Forecasting human-object interaction: joint prediction of motor attention and actions in first person video. In *European Conference on Computer Vision*, pages 704–721. Springer, 2020. 1

[21] Shugao Ma, Leonid Sigal, and Stan Sclaroff. Learning activity progression in lstms for activity detection and early detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1942–1950, 2016. 2

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

[23] Sebastian Stein and Stephen J McKenna. Combining embedded accelerometers with computer vision for recognizing food preparation activities. In *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*, pages 729–738, 2013. 1, 2

[24] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016. 2

[25] Du Tran, Heng Wang, Matt Feiszli, and Lorenzo Torresani. Video classification with channel-separated convolutional networks. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5551–5560, 2019. 1

[26] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 1, 2

[27] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Anticipating visual representations from unlabeled video. In *CVPR*, 2016. 1

[28] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *European conference on computer vision*, pages 20–36. Springer, 2016. 1

[29] Yu Wu, Linchao Zhu, Xiaohan Wang, Yi Yang, and Fei Wu. Learning to anticipate egocentric actions by imagination. *IEEE Transactions on Image Processing*, 30:1143–1152, 2021. 1