Intentional Evolutionary Learning for Untrimmed Videos with Long Tail Distribution: Supplementary Material

Anonymous submission

1 Sliding window

Sliding window is a common data augmentation strategy. As shown in Figure 1, the window length can be denoted as W_{len}, the window sliding stride is W_{stride}, the start time of the window is w_s, and the end time is w_e. Therefore, one valid video clip can be determined by a sliding window with W_{len}, W_{stride}, and w_s or w_e. And then the label of the evolution action can be denoted as the last category contained in the specific window y' which starting from t_s^{pred} to w_e. Similarly, the label of the potential action can be denoted as the penultimate category of this window y which starting from t_s^{reco} to t_s^{pred}.

Given a random number t_{rand} between t_s^{reco} and t_s^{pred}, we take the video clip with length t_{rand} starting from w_s as the the observing video clip and use it as the model input. And then the action evolution time from category y to category y' can be denoted as $D = t_e^{reco} - t_{rand}$. Consequently, the ground truth of one observing video clip is a triplet of $\{y, y', D\}$.

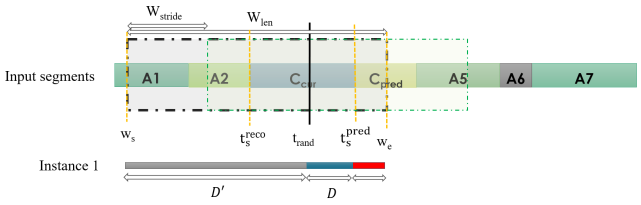

Figure 1: Sliding window diagram.

2 Experiment

Experimental Setup

Implementation details. We design the sliding window with length 200 and stride 20. If the length of a video is less than 200 segments, the entire video would be taken as one sliding window. The I3D (Carreira and Zisserman 2017) network pre-trained on Kinetics-400 (Carreira and Zisserman 2017) is employed for feature extraction and the Adam (Loshchilov and Hutter 2017) optimizer is used for end-to-end training. The learning rate, weight decay and dropout rate for THUMOS14 and ActivityNet v1.3 are both set to be 10^{-4}, 5×10^{-5}, and 0.7, respectively. For our ICCA loss, the update epoch is 10, and the mean precision from the previous round of the training set is employed as the error function. Moreover, all experiments are conducted on a single NVIDIA A100 GPU.

modal	top-1 acc		top-5 acc		time acc
	reco	pred	reco	pred	
rgb	**67.9**	64.8	**95.5**	**92.8**	38.11
flow	65.6	63.3	89.9	87.1	35.74
rgb+flow	66.4	**65.5**	92.5	88.7	**42.70**

Table 1: Multi-modal ablation experiments were conducted on the THUMOS14 dataset.

threshold ξ	top-1 acc		top-5 acc		mean precision	
	reco	pred	reco	pred	reco	pred
1e-6	**63.2**	59.2	**94.5**	**91.5**	27.66	37.10
1e-5	67.3	70.2	92.9	91.0	45.47	**54.61**
1e-4	66.3	**63.9**	95.7	90.7	**38.33**	47.30
1e-3	57.0	60.5	92.4	86.0	25.61	38.48

Table 2: Exploring the hyperparameter ξ of intention-oriented evolutionary learning on the THUMOS14 dataset.

Supplemental Experiments

Multi-modal fusion. We conducted multi-modal ablation experiments on the THUMOS14 dataset using RGB, flow, and the fusion of both. The fusion strategy used was early fusion, with simple concatenation of RGB and flow before the data was fed into the prediction backbone. As shown in Table 1, under the RGB modality, most optimal or sub-optimal results can be obtained, but with less computational complexity and memory usage. Therefore, we choose RGB as the single modality that runs through our experiments.

Threshold parameter ξ. We discussed the impact of different threshold parameters ξ on the results of intention-oriented evolutionary learning on the THUMOS14 dataset. As shown in Table 2, we can see that most good results for the top-1 accuracy and top-5 accuracy can be obtained when ξ is set to 1e-6 in the four orders of magnitude of ξ. Therefore, we chose this value for our experiments.

References

Carreira, J.; and Zisserman, A. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6299–6308.

Loshchilov, I.; and Hutter, F. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.