Prompt-Guided Zero-Shot Anomaly Action Recognition using Pretrained Deep Skeleton Features Supplementary Material

Fumiaki Sato*, Ryo Hachiuma*, Taiki Sekii
Konica Minolta, Inc.
{fumiaki.sato.jp,rhachiuma,taiki.sekii}@gmail.com

1. Implementation Details

In this section, we provide detailed information about the proposed network architecture, data augmentation, and hyperparameters employed during pretraining.

1.1. Architecture

In the DNN architecture, the dimension of the output feature vector at the first MLP layer is set to 64. We set the number of repeats in the residual block to $r = 5$ and the MLP bottleneck ratio to $\beta = 4$. The output dimension D_{out} at each residual MLP block is set to $64, 128, 128, 1024, 1024$ $(= S)$, respectively. We employ Batch Normalization [2] as the normalization layer and Leaky ReLU [4] as the activation function. Also, we employ MPNet [5] as the text encoder.

1.2. Data Augmentation

We apply two types of data augmentation during pretraining; augmentation onto the image space and along the temporal axis to the input joints. We randomly scale, shift, rotate, and flip the joint coordinates for the augmentation onto the image space. We randomly crop the input joints with a random size of the temporal window and a random start time. We also drop joints within a random interval range.

1.3. Hyperparameters

The hyperparameters employed in each experiment are listed in Tab. 1. We pretrain the model for 200 epochs on Kinetics-400 and NTU RGB+D datasets with $\alpha = 1$. We then set α to 0.2 and 0.3 to pretrain the model on each dataset for 40 and 10 epochs, respectively. Finally, the hyperparameters were simply found in a standard coarse-to-fine grid search or step-by-step tuning.

Table 1. Hyperparameters of each dataset during pretraining.

Pretraining dataset	Kinetics-400 [1]		NTU RGB+D 120 [3]	
Mixing ratio of the loss functions α	1	0.2	1	0.3
Optimizer	Stocastic Gradient Descent			
Number of epochs	200	40	200	10
Batch size	512	256	512	256
Learning rate	0.04	1.6	0.04	0.8
Weight decay	0.0001	0.000025	0.00005	0.000025
Momentum	0			
LR scheduler	linear			
Joint scaling	$[0.8, 1.2]$		$[0.6, 1.4]$	
Joint shift	0.2		0.4	
Joint rotate (°)	5		10	
Joint flip ratio	0.5			
Temporal crop window	100		150	
Temporal shift range	150		150	
Temporal FPS drop	5		3	

References

[1] Joao Carreira and Andrew Zisserman. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In *CVPR*, 2017. 1

[2] Sergey Ioffe and Christian Szegedy. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In *ICML*, 2015. 1

[3] Jun Liu, Amir Shahroudy, Mauricio Perez, Gang Wang, Ling-Yu Duan, and Alex C Kot. NTU RGB+D 120: A large-scale benchmark for 3D human activity understanding. *PAMI*, 42(10):2684–2701, 2020. 1

[4] Andrew L Maas, Awni Y Hannun, Andrew Y Ng, et al. Rectifier Nonlinearities Improve Neural Network Acoustic Models. In *ICML*, 2013. 1

[5] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. MPNet: Masked and Permuted Pre-training for Language Understanding. In *NeurIPS*, 2020. 1

* Equal contribution.