Efficient Few-Shot Action Recognition via Multi-Level Post-Reasoning: Supplementary Material

Cong Wu[1], Xiao-Jun Wu[1]*, Linze Li[1], Tianyang Xu[1], Zhenhua Feng[1,2,3], and Josef Kittler[2,3]

[1] School of Artificial Intelligence and Computer Science, Jiangnan University, China
[2] Centre for Vision, Speech and Signal Processing, University of Surrey, UK
[3] School of Computer Science and Electronic Engineering, University of Surrey, UK
{congwu,linze.li}@stu.jiangnan.edu.cn;
{wu_xiaojun,tianyang.xu,fengzhenhua}@jiangnan.edu.cn;
j.kittler@surrey.ac.uk

1 Datasets

SSV2 (Something-Something V2) [2] and K400 (Kinetics-400) [1] are currently the most commonly used large-scale datasets in action recognition. The SSV2 dataset encompasses 174 categories with 220,847 video sequences, comprising 168,913 samples in the training set, 24,777 in the validation set, and 27,157 in the test set. The K400 dataset encompasses over 306k video sequences, covering 400 categories, with each category containing at least 400 samples. HMDB-51 [3] and UCF-101 [4] are two relatively smaller datasets in action recognition. HMDB-51 consists of 51 categories with 6,766 videos, primarily focusing on facial and body actions. UCF-101 includes 101 categories with 13,320 videos, which were recorded in unconstrained real-world settings.

The aforementioned datasets can generally be categorised into two groups: temporal-related datasets, such as SSV2, and spatial-related datasets, represented by K400, HMDB-51 and UCF-101. The characteristics of these datasets necessitate different design considerations for models. For instance, SSV2 emphasises fine-grained actions, particularly highlighting interactions between individuals and objects. On the other hand, HMDB-51, UCF-101 and K400 places a greater emphasis on comprehending spatial information. Therefore, large pre-trained models can often directly bring huge accuracy improvements in spatial-related datasets, while more refined image-to-video adaptation is required in temporal-related datasets.

2 Implementation Details

Following the previous setting [5], for SSV2 and K400 datasets, samples from 64/12/24 classes are selected as training/validation/test sets; and we choose

* Corresponding Author.

Table 1: The implementation details of our proposed EMP-Net.

Dataset	SSV2 Small&Full	HMDB-51	UCF-101	K400
Optimizer	Adam, Momentum=0.9, Nesterov=True			
Epoch	10			
Batchsize	2			
Warm up epoch	1			
Learning rate	1e-3		1e-4	
Warm up learning rate	1e-5		1e-6	
Learning rate policy	reduced to 1/10 of the previous epoch at [4,6] epoch			
Train tasks	10000		5000	
Train tasks	10000			
Data augmentation	Color augmentation, Random erase			
Frame	8			

Table 2: Text prompt for action label.

{}
A photo of action {}
A picture of action {}
Human action of {}
{}, an action
{}, this is an action
{}, a video of action
Playing action of {}
Playing a kind of action, {}
Doing a kind of action, {}
Look, the human is {}
Can you recognize the action of {}?
Video classification of {}
A video of {}
The man is {}
The woman is {}

samples from 31/10/10 classes for training/validation/test for HMDB-51, and samples from 70/10/21 classes for training/validation/test for UCF-101. Additionally, two versions of the SSV2 dataset are used: SSV2 Small and Full. The former contains only 100 samples for each class, while the latter retains all samples in each class. The training and evaluation details are summarised in Table 1. We also include multiple text prompts for each category, as illustrated in Table 2. During training, one prompt strategy is randomly selected from the candidates; for evaluation, we compute the average of all candidates as the final textual feature.

The CLIP-ViT-B/16 is used as the backbone network. As for the newly introduced transformer architectures, we set the head as 4, the channel dimension of the head as 64, the channel dimension of the feed-forward operation as 128, and the dropout rate for the attention and feed-forward operation as 0.2 and 0.05 respectively.

3 Visualisation Analysis

To further analyse the effectiveness of the proposed components in EMP-Net, we visualise the classification effects of different models, as depicted in Fig. 1. Specifically, for a given query video, we present the corresponding most similar

Fig. 1: Visualisation of classification effects of different models. We show the support video that is most similar to the query video under different paradigms.

support videos after incorporating post-reasoning and multi-level step by step. Theoretically, as discussed in the main paper, the similarity between the query video and the support video serves as the basis for label inference. Firstly, from this figure, we observe that due to significant similarities among certain actions, it is challenging to endow the network with sufficient discrimination solely by relying on the prior knowledge of the CLIP model. For example, 'putting sth on a surface' and 'failing to put sth' exhibit notable resemblance, as do 'dropping sth into sth' and 'sth falling'. Consequently, a baseline lacking effective reasoning ability struggles to correctly classify these instances. Furthermore, while the introduction of post-reasoning alleviates this issue, providing reliable clues for categories such as 'failing to put something into something' and 'putting something on the edge of something' remains challenging, prompting the introduction of the multi-level mechanism. Ultimately, the integration of post-reasoning and multi-level mechanisms provides more comprehensive clues, resulting in optimal performance.

References

1. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: proceedings of the IEEE Conference on Computer Vision and

Pattern Recognition. pp. 6299–6308 (2017). https://doi.org/10.1109/CVPR.2017.502

2. Goyal, R., Ebrahimi Kahou, S., Michalski, V., Materzynska, J., Westphal, S., Kim, H., Haenel, V., Fruend, I., Yianilos, P., Mueller-Freitag, M., et al.: The" something something" video database for learning and evaluating visual common sense. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 5842–5850 (2017). https://doi.org/10.1109/ICCV.2017.622
3. Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., Serre, T.: Hmdb: a large video database for human motion recognition. In: 2011 International conference on computer vision. pp. 2556–2563. IEEE (2011). https://doi.org/10.1109/ICCV.2011.6126543
4. Soomro, K., Zamir, A.R., Shah, M.: Ucf101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402 (2012)
5. Wang, X., Zhang, S., Cen, J., Gao, C., Zhang, Y., Zhao, D., Sang, N.: Clip-guided prototype modulating for few-shot action recognition. International Journal of Computer Vision **132**(6), 1899–1912 (2024). https://doi.org/10.1007/s11263-023-01917-4