Supplementary Material: Few-shot Action Recognition with Prototype-centered Attentive Learning

Xiatian Zhu
xiatian.zhu@samsung.com
Antoine Toisoul
antoine.toisoul@gmail.com
Juan-Manuel Pérez-Rúa
jmpr@fb.com
Brais Martinez
brais.mart@gmail.com
Li Zhang
lizhangfd@fudan.edu.cn
Tao Xiang
t.xiang@surrey.ac.uk

Samsung AI Centre Cambridge, UK

In this supplementary document, we conducted ablation study of our method on the most challenging fine-grained video action dataset Sth-Sth-100 [3] with more subtle inter-class differences.

1 Importance of pretraining

In our model, the feature embedding TSN model is pretrained on the whole training set before the episodic training stage. Instead, the current state-of-the-art OTAM [1] skipped this pretraining step. Table 1 shows surprisingly that this pretraining step is vital. Our PAL model only with pretraining is already more effective than OTAM. This suggests for the first time that a strong feature embedding is important for action classification in a FSL context, confirming the similar conclusion drawn in the image counterparts [2, 4, 5]. To evaluate how our pretraining can benefit existing few-shot action models, we integrated it in our own OTAM implementation due to the absence of officially released code. The results show that once the feature embedding is well pretrained, OTAM's time warping becomes insignificant, even introducing a negative impact in 5-shot case.

2 Contributions of model components

To understand the benefits of the two components in our PAL, namely Hybrid Attentive Learning (HAL) and Prototype-centered Contrastive Learning (PCL), we conducted a de-

Method	Sth-Sth-100 1-shot	5-shot
OTAM [1]	42.8	52.3
Pretrained PAL	42.7	**59.6**
Pretraining PAL + OTAM*	**43.1**	58.4

Table 1: Importance of pretraining. *: Our implementation.

tailed component analysis by comparing the performances with and without them. From Table 2 the following observations can be made. (1) Our model pretraining sets out with strong performance. (2) Each component is useful and the two components are complementary to each other in improving the classification accuracy. Concretely, the two components jointly improve the accuracy by 3.7% and 3.0% over pretraining in the 1-shot and 5-shot settings, respectively.

Method	Sth-Sth-100 1-shot	5-shot
Pretrained PAL	42.7	59.6
HAL	45.3	62.1
PCL	44.5	61.4
HAL+PCL	**46.4**	**62.6**

Table 2: Model component analysis. Two main components of our PAL were evaluated. **HAL**: Hybrid Attentive Learning; **PCL**: Prototype-centered Contrastive Learning.

References

[1] Kaidi Cao, Jingwei Ji, Zhangjie Cao, Chien-Yi Chang, and Juan Carlos Niebles. Few-shot video classification via temporal alignment. In *CVPR*, 2020.

[2] Wei-Yu Chen, Yen-Cheng Liu, Zsolt Kira, Yu-Chiang Frank Wang, and Jia-Bin Huang. A closer look at few-shot classification. In *ICLR*, 2019.

[3] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fründ, Peter Yianilos, Moritz Mueller-Freitag, Florian Hoppe, Christian Thurau, Ingo Bax, and Roland Memisevic. The "something something" video database for learning and evaluating visual common sense. In *ICCV*, 2017.

[4] Bin Liu, Yue Cao, Yutong Lin, Qi Li, Zheng Zhang, Mingsheng Long, and Han Hu. Negative margin matters: Understanding margin in few-shot classification. In *ECCV*, 2020.

[5] Yonglong Tian, Yue Wang, Dilip Krishnan, Joshua B Tenenbaum, and Phillip Isola. Rethinking few-shot image classification: a good embedding is all you need? In *ECCV*, 2020.