Supplemental Materials: Self-supervised Motion Learning from Static Images

Ziyuan Huang[1,2], Shiwei Zhang[2], Jianwen Jiang[2], Mingqian Tang[2], Rong Jin[2], Marcelo H. Ang Jr[1]
[1] National University of Singapore, Singapore
[2] Alibaba Group, China
ziyuan.huang@u.nus.edu, mpeangh@nus.edu.sg
{zhangjin.zsw, jianwen.jjw, mingqian.tmq,jinrong.jr}@alibaba-inc.com

In this supplemental material, we provide additional details on the pre-training and fine-tuning strategies.

Pre-training. During the self-supervised pre-training, we use Adam [4] as our optimizer and use the learning rate of 0.001. The batch size is 10 source images per GPU for 16 GPUs. We use warm-up [1] of 10 epochs starting with a learning rate of 0.0001 and the total number of epochs is 100 except for kinetics, where 2 warm-up epochs and in total 20 epochs are used to train the model. A half-period cosine schedule [5] is adopted. A dropout of 0.5 is used during the pre-training of the video models. Besides the MoSI-specific augmentations, including random sizes and locations of the static mask, we also apply a frame-wise random color jittering following [2, 3], for the model to learn not only low-level pixel-correspondence, but also semantic correspondences. The weight decay is set to 1e-4.

Fine-tuning. During fine-tuning on the downstream action classification task, warm-up is applied with a starting learning rate of 1/10 the base learning rate. The weight decay is set to 0.001. Ten warmup epochs is used and the model is trained in total for 300 epochs. Data augmentation include spatial random crop, random horizontal flip, and clip-wise random color jittering with the same parameter as in self-supervised training.

References

[1] Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017. 1

[2] Tengda Han, Weidi Xie, and Andrew Zisserman. Video representation learning by dense predictive coding. In *ICCV Workshops*, pages 0–0, 2019. 1

[3] Tengda Han, Weidi Xie, and Andrew Zisserman. Memory-augmented dense predictive coding for video representation learning. *arXiv preprint arXiv:2008.01065*, 2020. 1

[4] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 1

[5] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 1