Improving Video Model Transfer with Dynamic Representation Learning Supplemental Material

Yi Li Nuno Vasconcelos
Department of Electrical and Computer Engineering
University of California, San Diego

A. Implementation Details

Optimization hyperparameters. Table 1 shows pretraining and fine-tuning hyperparameters used on each dataset. Stochastic gradient descent (SGD) of mini-batch size $N = 64$ is used to optimize all models except TSM ResNet-50, where we used SGD with $N = 32$ to save GPU resources and scaled the learning rate accordingly. Due to a substantial difference in training set size and data statistics, we determine these individually on each dataset on a left-out validation set. Default hyperparameters for the proposed DRL iterations, as described in Algorithm 1, 2 and 3 of main text, are provided in table 2. These are used in all experiments unless otherwise noted.

Dataset	FT	Epochs	Initial LR	LR Step (Epochs)	WD	Freeze BN
miniKinetics	✗	50	0.1	20	10^{-4}	✗
	✓	25	0.01	10		✓
Kinetics	✗	100	0.1	30	10^{-4}	✗
	✓	25	0.01	10		✓
UCF-101	✗	100	0.1	30	10^{-3}	✗
	✓	30	0.001	20		✓
HMDB-51	✗	100	0.1	30	10^{-3}	✗
	✓	30	0.001	20		✓
Diving-48	✗	100	0.1	30	10^{-3}	✗
	✓	50	0.01	20		✗

Table 1. Optimization hyperparameters by training dataset. **FT**—fine-tuning, **LR**—learning rate, **WD**—weight decay, **BN**—batch normalization [6]. At multiples of LR step, learning rate is reduced by $10\times$.

Distillation weight α	0.5
Adversarial input weight β	0.5
(Alg. 1) Perturbation strength ϵ	8/255
(Alg. 2 & 3) Dynamic loss weight λ	0.5

Table 2. Default DRL hyperparameters.

Preprocessing. During training, videos are resized and randomly cropped to the desired spatio-temporal dimension, after which color jittering and random horizontal flipping are applied. To ensure the fairness of the dynamic score metric across model architectures, we use an adaptive sampling frame rate to ensure that the *duration* of input clips is fixed at 1 second. At test time, model outputs are aggregated over center crops of 10 1-second clips sampled uniformly from each input video.

Training resources. Experiments are performed on NVIDIA GeForce GTX 1080 Ti GPUs on an internal cluster. Data parallelism is used to distribute batches to multiple GPUs when training larger networks (3D ResNet-50 [5], TSM ResNet-50 [10]). Total training time per episode on Kinetics-400 [7] varies from 2 to 5 days depending on model architecture.

B. Extended Results

Feature visualizations. Figure 1 shows the t-SNE [13] visualization of feature representations extracted from UCF and HMDB videos, using TSM ResNet-50 [10] with standard and DRL pretraining on Kinetics. It can be observed that DRL improves representation quality, with video features forming more pronounced clusters in the t-SNE plots. This translates to superior linear classification accuracy on both datasets, as reported in Table 2 of main text.

Model predictions. Figure 2, 3 and 4 contain frames from sample test videos and their corresponding predictions by baseline and DRL-trained models. We notice that DRL frequently corrects mistakes from the baseline model in a few scenarios:

- Actions with a long temporal span—Fig. 2a, 3d, 4d;
- Actions in uncommon scene—Fig. 2c, 2d, 3b, 4b;
- Actions without co-occurring objects—Fig. 3a, 3c, 3d.

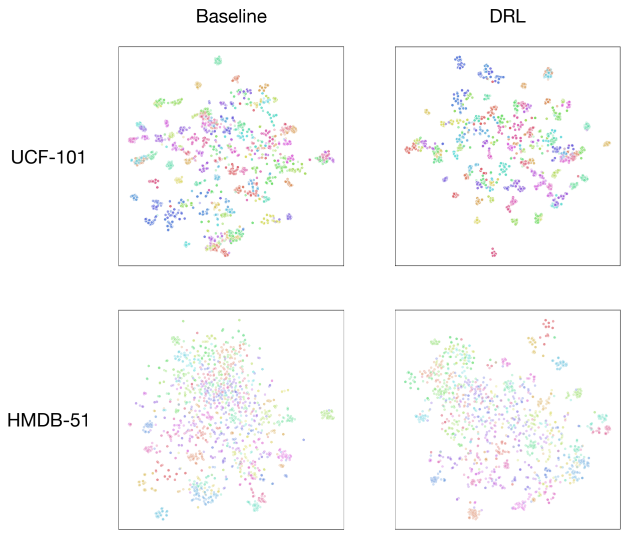


Figure 1. t-SNE [13] visualization of UCF [12] and HMDB [8] video features, extracted from TSM ResNet-50 [10] models with standard (**left**) and DRL (**right**) pretraining.

C. Limitations and Future Work

Spatial appearance vs. temporal dynamics. While experiments have confirmed the benefit of dynamic video representations, we note that an inherent trade-off exists between spatial and temporal modeling within a given video network. It is possible that spatial modeling is beneficial and should be exploited for recognizing certain actions, such as those that involve human-object interactions. By introducing an objective and interpretable measure of spatial-temporal bias of models, we expect that this work stimulates more research in the vision community to study this trade-off, as a guidance to building robust video action recognition systems.

Inductive biases of video networks. Towards the goal of building video representations with more dynamics, a parallel direction to this work is to design model architectures with stronger inductive bias for long-range temporal modeling. Recent progresses on video transformer models [1, 2, 3] present a promising direction thanks to the ability of self-attention layers to aggregate global information. However, without careful design and proper regularizations, even transformer models have been found to ignore temporal orders of input video sequence [3]. The findings of this work show that unless bias is explicitly penalized, the networks will leverage it. Through the formulation of dynamic score and DRL, we also anticipate more follow-up research on analyzing the inductive biases of video recognition models, using both 2D/3D convolutional and transformer-based architectures.

D. Assets & Licenses

All datasets used in this work are publicly available [4, 7, 8, 9, 11, 12, 14]. Table 3 lists the download page URL and license (if provided) of each individual dataset.

References

[1] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. *arXiv preprint arXiv:2103.15691*, 2021. 2

[2] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? *arXiv preprint arXiv:2102.05095*, 2021. 2

[3] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. *arXiv preprint arXiv:2104.11227*, 2021. 2

[4] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The" something something" video database for learning and evaluating visual common sense. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5842–5850, 2017. 2, 3

[5] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6546–6555, 2018. 1

[6] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. PMLR, 2015. 1

[7] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 1, 2, 3

[8] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *2011 International conference on computer vision*, pages 2556–2563. IEEE, 2011. 2, 3, 5

[9] Yingwei Li, Yi Li, and Nuno Vasconcelos. Resound: Towards action recognition without representation bias. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 513–528, 2018. 2, 3

[10] Ji Lin, Chuang Gan, and Song Han. Tsm: Temporal shift module for efficient video understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7083–7093, 2019. 1, 2

[11] Joanna Materzynska, Guillaume Berger, Ingo Bax, and Roland Memisevic. The jester dataset: A large-scale video dataset of human gestures. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019. 2, 3

[12] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 2, 3

(a) Triple jump

Baseline: Long jump$^{.588}$
DRL: Triple jump$^{.563}$

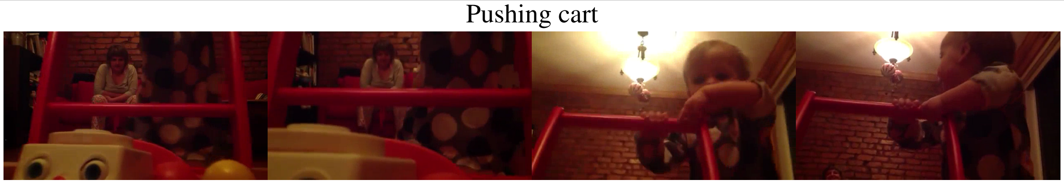
(b) Changing wheel

Baseline: Checking tires$^{.620}$
DRL: Changing wheel$^{.570}$

(c) Feeding birds

Baseline: Feeding fish$^{.661}$
DRL: Feeding birds$^{.299}$

(d) Pushing cart



Baseline: Reading book$^{.103}$
DRL: Pushing cart$^{.326}$

Figure 2. Sample model predictions on test videos from Kinetics [7].

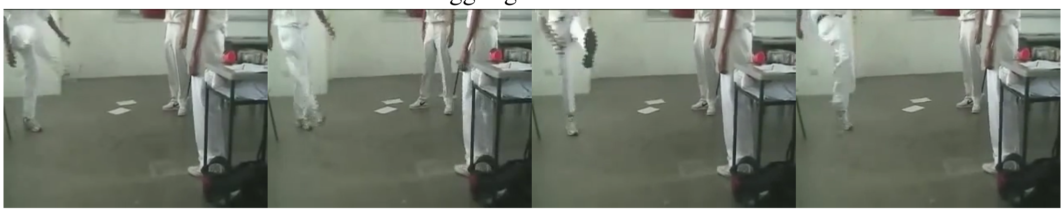
Dataset	Source URL	License
Kinetics [7]	https://deepmind.com/research/open-source/kinetics	CC BY 4.0
UCF-101 [12]	https://www.crcv.ucf.edu/data/UCF101.php	–
HMDB-51 [8]	https://serre-lab.clps.brown.edu/resource/hmdb-a-large-human-motion-database/	CC BY 4.0
Diving-48 [9]	http://www.svcl.ucsd.edu/projects/resound/dataset.html	–
Something V2 [4]	https://developer.qualcomm.com/software/ai-datasets/something-something	Research Use
Jester [11]	https://developer.qualcomm.com/software/ai-datasets/jester	Research Use
Mimetics [14]	https://europe.naverlabs.com/research/computer-vision/mimetics/	–

Table 3. Download URL and license (if applicable) of datasets used in this work.

[13] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 1, 2

[14] Philippe Weinzaepfel and Grégory Rogez. Mimetics: Towards understanding human actions out of context. *International Journal of Computer Vision*, 129(5):1675–1690, 2021. 2, 3, 4

(a) Juggling soccer ball



Baseline: Skipping rope$^{.174}$

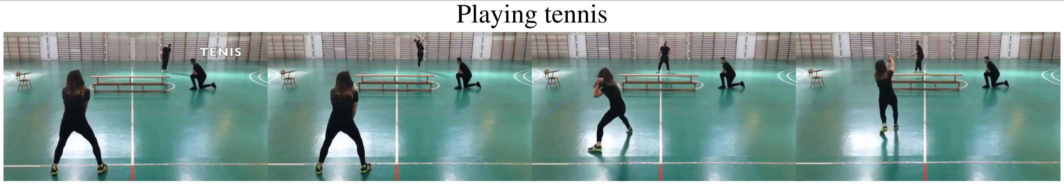
DRL: Juggling soccer ball$^{.562}$

(b) Juggling balls

Baseline: Punching person (boxing)$^{.175}$

DRL: Juggling balls$^{.961}$

(c) Playing tennis



Baseline: Catching or throwing baseball$^{.266}$

DRL: Playing tennis$^{.405}$

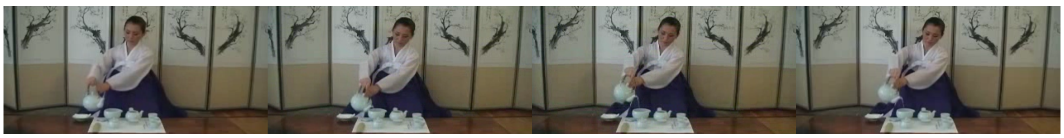
(d) Archery

Baseline: Punching person (boxing)$^{.380}$

DRL: Archery$^{.222}$

Figure 3. Sample model predictions on test videos from Mimetics [14].

(a) Pour



Baseline: Sword exercise$^{.559}$
DRL: Pour$^{.656}$

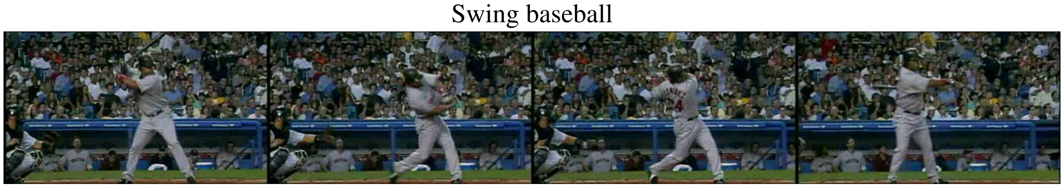
(b) Fall on the floor

Baseline: Punch$^{.802}$
DRL: Fall on the floor$^{.661}$

(c) Dribble

Baseline: Shoot ball$^{.511}$
DRL: Dribble$^{.892}$

(d) Swing baseball



Baseline: Throw$^{.838}$
DRL: Swing baseball$^{.601}$

Figure 4. Sample model predictions on test videos from HMDB-51 [8].