Rethinking Training Data for Mitigating Representation Biases in Action Recognition

Kensho Hara, Yuchi Ishikawa, Hirokatsu Kataoka
National Institute of Advanced Industrial Science and Technology (AIST)
Tsukuba, Ibaraki, Japan
{kensho.hara, yuchi.ishikawa, hirokatsu.kataoka}@aist.go.jp

Abstract

The purpose of this study is to train spatiotemporal 3D convolutional neural networks (3D CNNs) that properly leverage temporal information to recognize actions. Though 3D CNNs are an effective framework in action recognition, some studies showed the biases of video datasets for generic action recognition lead 3D CNNs to recognize not dynamic motions but static cues, such as objects, scenes, and people. On the other hand, video datasets for fine-grained action recognition, which classifies various actions in a specific domain, are expected to have small biases compared with the datasets for generic action recognition. In this study, we examine the biases of various video datasets, which include both generic and fine-grained action recognition tasks, for training 3D CNNs. Based on the results of experiments, the following conclusions could be obtained: (i) The representation biases learned from fine-grained action recognition datasets are smaller than those of generic action recognition datasets. (ii) The models pretrained on fine-grained action recognition datasets, of which the biases are small, leverage temporal information to recognize actions rather than static information. (iii) The models that leverage temporal information achieve better performance on fine-grained action recognition whereas the performance of the models pretrained on biased datasets is better on generic action recognition. We should evaluate models on both generic and fine-grained recognition datasets to properly evaluate their performance.

1. Introduction

Spatiotemporal 3D convolutional neural networks (3D CNNs) are an effective framework in action recognition. The 3D convolutions can theoretically extract spatiotemporal feature representations, which are important for action recognition, from raw videos. Various network architectures of 3D CNNs have been proposed, such as C3D [19], I3D [3], 3D ResNet [8], and SlowFast network [6].

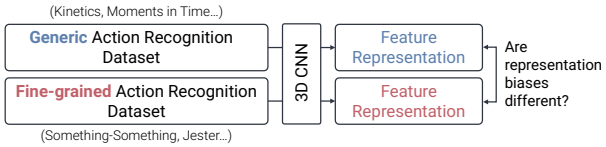


Figure 1: Overview of this study.

Some studies showed video datasets for action recognition have biases toward recognizing not dynamic motions but static cues, such as objects, scenes, and people [4, 9, 12, 13]. Such datasets, called generic action recognition datasets in this study, include various action categories in various domains, and the domain difference enables recognition of actions indirectly. For instance, a dataset includes videos of tennis swing action only in a tennis court, models trained on the dataset may recognize the action based on the scene (tennis court). Using such biased models leads to degenerate transferability and make incorrect recognition of novel actions in the same static cues.

Video datasets for fine-grained action recognition have also been released [7, 12, 15, 17] besides those for generic action recognition, such as Kinetics [10], UCF-101 [18], and HMDB-51 [11]. Most fine-grained action recognition datasets consist of videos in a specific domain. Because the static cues, such as objects and scenes, in the same domain are similar, it is difficult to recognize actions based on only the static cues. Training models on fine-grained recognition datasets likely avoids the effect of biases. Though the methods that mitigate the biases in action recognition are proposed [4, 12, 13], these methods require additional training costs or decrease the number of training data. Therefore, exploring video datasets with small biases would be a better solution.

In this study, we examine training 3D CNNs on various video datasets, which include both generic and fine-grained action recognition tasks, in order to build the models that properly leverage temporal information to recognize actions. We conducted various experiments to show (i) the differences of the biases toward the static representations and learned feature representations between generic and fine-grained action recognition datasets, (ii) the differences of transferability of models trained on generic and fine-grained action recognition datasets. As shown in Figure 1, our experimental results indicate that the CNN trained on a fine-grained action recognition datset leverages the temporal information in videos for action recognition. See Section 4 in details of our experimental results.

2. Related Work

2.1. Representation Biases in Action Recognition

Studies have been conducted to analyse and to mitigate representation biases. Huang *et al*. showed 3D CNNs did not significantly focus on motion information in classification of the videos of Kinetics by confirming performance drop with and without motion information in videos. Li *et al*. mitigated scene, object and people biases by resampling the original video datasets [12, 13]. The resampling approach reduces not only representation biases but also the number of training data, which is important for training 3D CNNs. Choi *et al*. introduced the adversarial loss for scene types to mitigate scene biases [4]. Their method needs additional training costs for detecting humans and classifying scenes. Unlike the abovementioned studies, we examine various existing video datasets for training 3D CNNs without representation biases.

2.2. Video Datasets

Video datasets for action recognition are mainly divided into generic and fine-grained action recognition datasets. Most of video datasets for action recognition provide the generic action recognition task, which tries classifying various action categories in various domains. The granularity of categories of generic action recognition is larger than that of fine-grained action recogition, described later. Such datasets include videos in a wide variety of domains, such as daily activities, sports, and entertainments. Popular datasets for generic action recognition are Kinetics-400, 600, 700 [1, 2, 10], Moments in Time [16], ActivityNet [5], UCF-101 [18], and HMDB-51 [11].

Fine-grained action recognition datasets have been recently released. Such datasets provide the videos in a specific domain. Something-Something includes the videos of fine-grained actions of human-object interactions [7]. Something-Something v2, which is the extended version of Something-Something, includes a larger number of videos [14]. Jester is a dataset for hand gesture recognition [15]. Diving48 [12] and FineGym [17] are video datasets in sports (gymnastics and diving).

We use the datasets for training 3D CNNs and compare the biases of trained models.

3. Experimental configuration

3.1. Summary

In order to examine how we should use video datasets to train action recognition models without the biases toward static representations, we conducted the two experiments described below. In Section 3.2, We explain the definition of representation biases discussed in this study. The details of the datasets are described in Section 3.3. Implementation details are described in Section 3.4.

We first examined the biases of video datasets for both generic and fine-grained action recognition tasks. According to previous works [4, 9, 12, 13], training 3D CNNs on some generic action recognition datasets, such as Kinetics, leads to biases toward the static representations. Here, we examined the differences of the biases toward the static representations between generic and fine-grained action recognition datasets. We trained the same model on different datasets in both generic and fine-grained recognition tasks, and compared biases of them. See Section 4.1 for details.

We then conducted an experiment to explore transferability of the models trained on each dataset. We assume that the models with the large biases toward the static representations can work well on the datasets of generic action recognition, which static cues are important to, but perform poorly on the datasets for fine-grained action recognition, which focuses on classifying dynamic motions. Therefore, we finetuned the models on both generic and fine-grained action recognition tasks, and compared performance of them. See Section 4.2 for details.

3.2. Definition of representation bias

To evaluate the representation biases, we confirmed the accuracies when we shuffled the input video frames during testing. If the representation biases of trained models are large, the models perform well even using the shuffled input frames because the models ignore temporal dynamics. On the other hand, the recognition performance of models without such biases significantly drop because the shuffle of input frames destroy temporal information of videos. Therefore, we define the representations biases of dataset D as the log ratio between the accuracies on original and shuffled inputs as

$$B_D = \log\left(A(D, \phi_{\text{original}})/A(D, \phi_{\text{shuffle}})\right), \qquad (1)$$

where A is accuracy using feature representation ϕ on dataset D and ϕ_{original} and ϕ_{shuffle} are feature representa-

tions of original and shuffled input frames, respectively.

3.3. Datasets

We use both generic and fine-grained action recognition datasets in this study.

Generic action recognition datasets. We use Kinetics-400 [10], Moments in Time [16], ActivityNet [5], UCF-101 [18], and HMDB-51 [11]. The Kinetics dataset is the most popular video dataset for training action recognition models. Different versions of Kinetics exist, such as Kinetics-400, -600, and -700, which are similar datasets with different numbers of action classes. We use the smallest Kinetics-400 in this study. Moments in Time is also a large-scale video dataset, which includes 1,000,000 videos from 339 action classes. ActivityNet (v1.3) provides samples from 200 human action classes with the total video length is 849 hours, and with the total number of action instances is 28,108. Unlike the other datasets, ActivityNet consists of untrimmed videos, which include non-action frames. We use ActivityNet as a trimmed video datset in this study. UCF-101 and HMDB-51 are popular video datasets as benchmark of action recognition. UCF-101 includes 13,320 action instances from 101 human action classes, and HMDB-51 includes 6,766 videos from 51 human action classes. We use train/test split 1 of UCF-101 and HMDB-51 in this study.

Fine-grained action recognition datasets. We use Something-Something v2 [14], Jester [15], Gym288 [17], and Diving48 [12]. The four fine-grained datasets consist of videos of different specific domains. Something-Something v2 includes 220,847 videos that capture 174 fine-grained action classes of human-object interaction scenarios. Jester is a video dataset of hand gestures that includes 148,092 videos with 27 hand gesture classes. Gym288 is a configuration of the FineGym dataset, which includes 288 gymnastic action classes with 38,980 videos. Diving48 has been built to produce a video datset with small biases, and provides samples from 48 diving actions with 18,404 videos.

3.4. Implementation

Training. We use 3D ResNet-50 [8] with 32 frames $\times 112 \times$ 112 pixels inputs. We use stochastic gradient descent with momentum to train the model. Training samples are randomly generated from videos in training data in order to perform data augmentation. First, we select a temporal position in a video by uniform sampling in order to generate a training sample. A 32-frame clip is then generated around the selected temporal position. If the video is shorter than 32 frames, then we loop it as many times as necessary. Next, we randomly select a spatial position and scale by uniform sampling. The sample is cropped at the selected spatial positions with the selected scale and square aspect ratio. We spatially resize the sample at 112×112 pixels so that the size of training sample is the 3 channels $\times$ 16 frames $\times$ 112 pixels $\times$ 112 pixels. Each sample is horizontally flipped with 50% probability. We also perform mean subtraction, which means that we subtract the mean values of Kinetics-400 from the sample for each color channel.

In our training, we use cross-entropy losses. The training parameters include a weight decay of 0.0001 and 0.9 for momentum. When training the networks, we start from learning rate 0.03, and divide it by 10 at every 100 epochs. The training continues until 250 epoch. We use 4 GPUs for the all training, and a mini-batch size is 32 clips per GPU.

Inference. We adopt the sliding window manner to generate input clips, (*i.e.*, each video is split into 32-frame clips with 16-frame strides), input each clip into the trained models, and estimate the clip class scores, which are averaged over all the clips of the video. The class that has the maximum score indicates the classified label.

Table 1: Training 3D ResNet-50 from scratch. *Something v2* indicates Something-Something v2. Larger *Log ratio*, which is the log ratio between the accuracies on original and shuffled inputs, indicates the smaller representation.

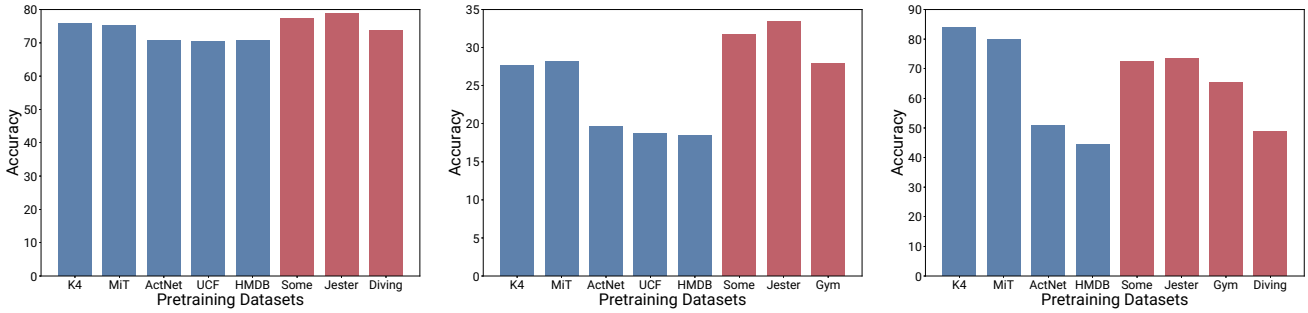
Dataset	Accuracy [%]		Log ratio
	Original inputs	Shuffled inputs	
Kinetics-400	64.3	32.7	0.68
Moments in Time	27.1	14.2	0.65
ActivityNet	33.4	29.6	0.12
UCF-101	45.3	32.0	0.35
HMDB-51	19.4	11.0	0.56
Something v2	44.3	2.3	2.95
Jester	92.0	20.1	1.52
Gym288	71.7	3.9	2.92
Diving48	15.1	4.2	1.29

4. Results and Discussion

4.1. Analyses of scratch training

We began by training 3D ResNet-50 on each dataset from scratch. We conducted this experiment to confirm the differences of representation biases learned from each dataset. We used Kinetics-400, Moments in Time, ActivityNet, UCF-101, HMDB-51, Something-Something v2, Jester, Diving48, and Gym288.

Table 1 shows accuracies on original and shuffled inputs as well as the log ratios between the accuracies. As can be seen in the table, the log ratios of generic action recognition



(a) Gym288. (b) Diving48. (c) UCF-101.

Figure 2: Finetuning 3D ResNet-50 on Gym288, Diving48, and UCF-101. Blue and red colors indicate generic and fine-grained action recognition datasets, respectively. K4, MiT, ActNet, UCF, HMDB, Some, Gym, and Diving means Kinetics-400, Moments in Time, ActivityNet, UCF-101, HMDB-51, Something-Something v2, Gym288, and Diving48, respectively.

datasets (top rows) are small compared with the ratios of fine-grained recognition datasets (bottom rows) though the accuracies vary due to the difficulty of recognition on each dataset. High accuracies on shuffled input frames mean that the trained models ignore the motion information, *i.e.*, the trained models capture the biases toward the static cues. These results indicate that the models trained on existing generic action recognition datasets leverage the static cues to recognize actions whereas the models trained on fine-grained recognition datasets focus on the dynamic motions.

4.2. Analyses of finetuning

Here, we conducted experiments of finetuning the pretrained models to explore transferability of the models. We assumed that the models with the large biases toward the static cues can work well on the datasets of generic action recognition, which the static cues are important to, but perform poorly on the datasets of fine-grained action recognition, which focuses on classifying dynamic motions.

Figure 2 shows the results of the models when finetuning on Gym288, Diving48, and UCF-101, respectively. Here, it can be seen that the accuracies of models pretrained on ActivityNet, UCF-101, HMDB-51, and Diving48 are relatively low due to the small number of pretraining data. We can also see that the the models pretrained on Something-Something v2 and Jester achieved better accuracies on Gym288 and Diving48, which are fine-grained recognition datasets, compared with Kinetics-400 and Moments in Time datasets, which are larger scale video datasets for the generic action recognition task. Something-Something v2 and Jester are also fine-grained action recognition datasets of different domains. These result indicate that representations with small biases are effective even though the representations are learned from different domains, and that the biases acquired from generic action recognition datasets degenerate the transferability of feature representations to fine-grained action recognition even though the representations are learned using larger datasets. In addition, we can see that the Kinetics-400 and Moments in Time pretrained models achieved higher accuracies on UCF-101 compared with the accuracies of the models pretrained on Something-Something v2 and Jester. This result indicates that the biased feature representations works better on generic action recognition. If a model achieves good accuracy on generic action recognition, the representation of model may be biased and transferability is low especially on fine-grained action recognition datasets.

5. Conclusion

In this study, we examined training 3D CNNs on various video datasets in order to build the models that properly leverage temporal information to recognize actions. Based on the results of those experiments, the following conclusions could be obtained: (i) The representation biases learned from fine-grained action recognition datasets are smaller than those from generic action recognition datasets. (ii) The models pretrained on fine-grained action recognition datasets, of which the biases are small, leverage temporal information to recognize actions rather than static information. (iii) The models that leverage temporal information achieve better performance on fine-grained action recognition whereas the performance of the models pretrained on biased datasets is better on generic action recognition. We should evaluate recognition models on both generic and fine-grained recognition datasets to properly evaluate the performance of models.

Acknowledgment

This paper is based on results obtained from a project, JPNP20006, commissioned by the New Energy and Industrial Technology Development Organization (NEDO). Computational resource of AI Bridging Cloud Infrastructure (ABCI) provided by AIST was used.

References

[1] João Carreira, Eric Noland, Andras Banki-Horvath, Chloe Hillier, and Andrew Zisserman. A short note about Kinetics-600. *arXiv preprint, arXiv:1808.01340*, 2018. 2

[2] João Carreira, Eric Noland, Chloe Hillier, and Andrew Zisserman. A short note on the Kinetics-700 human action dataset. *arXiv preprint, arXiv:1907.06987*, 2019. 2

[3] João Carreira and Andrew Zisserman. Quo vadis, action recognition? A new model and the Kinetics dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4724–4733, 2017. 1

[4] Jinwoo Choi, Chen Gao, Joseph C. E. Messou, and Jia-Bin Huang. Why can't i dance in the mall? learning to mitigate scene bias in action recognition. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, pages 1–13, 2019. 1, 2

[5] Bernard Ghanem Fabian Caba Heilbron, Victor Escorcia and Juan Carlos Niebles. ActivityNet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 961–970, 2015. 2, 3

[6] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the International Conference on Computer Vision (ICCV)*, pages 6202–6211, 2019. 1

[7] R. Goyal, S. E. Kahou, V. Michalski, J. Materzynska, S. Westphal, H. Kim, V. Haenel, I. Fruend, P. Yianilos, M. Mueller-Freitag, F. Hoppe, C. Thurau, I. Bax, and R. Memisevic. The "something something" video database for learning and evaluating visual common sense. In *Proceedings of the International Conference on Computer Vision (ICCV)*, pages 5843–5851, 2017. 1, 2

[8] 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 (CVPR)*, pages 6546–6555, 2018. 1, 3

[9] De-An Huang, Vignesh Ramanathan, Dhruv Mahajan, Lorenzo Torresani, Manohar Paluri, Li Fei-Fei, and Juan Carlos Niebles. What makes a video a video: Analyzing temporal information in video understanding models and datasets. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7366–7375, 2018. 1, 2

[10] Will Kay, João Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The Kinetics human action video dataset. *arXiv preprint, arXiv:1705.06950*, 2017. 1, 2, 3

[11] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. HMDB: a large video database for human motion recognition. In *Proceedings of the International Conference on Computer Vision (ICCV)*, pages 2556–2563, 2011. 1, 2, 3

[12] 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 1–16, 2018. 1, 2, 3

[13] Yi Li and Nuno Vasconcelos. Repair: Removing representation bias by dataset resampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9572–9581, 2019. 1, 2

[14] Farzaneh Mahdisoltani, Guillaume Berger, Waseem Gharbieh, David Fleet, and Roland Memisevic. On the effectiveness of task granularity for transfer learning. *arXiv preprint, arXiv:1804.09235*, 2018. 2, 3

[15] J. Materzynska, G. Berger, I. Bax, and R. Memisevic. The jester dataset: A large-scale video dataset of human gestures. In *Proceedings of the International Conference on Computer Vision Workshop (ICCVW)*, pages 2874–2882, 2019. 1, 2, 3

[16] Mathew Monfort, Alex Andonian, Bolei Zhou, Kandan Ramakrishnan, Sarah Adel Bargal, Tom Yan, Lisa Brown, Quanfu Fan, Dan Gutfruend, Carl Vondrick, et al. Moments in time dataset: one million videos for event understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–8, 2019. 2, 3

[17] Dian Shao, Yue Zhao, Bo Dai, and Dahua Lin. Finegym: A hierarchical video dataset for fine-grained action understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2616–2625, 2020. 1, 2, 3

[18] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A dataset of 101 human action classes from videos in the wild. CRCV-TR-12-01, 2012. 1, 2, 3

[19] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3D convolutional networks. In *Proceedings of the International Conference on Computer Vision (ICCV)*, pages 4489–4497, 2015. 1