



Do Less and Achieve More:

Training CNNs for Action Recognition Utilizing Action Images from the Web

Shugao Ma¹, Sarah Adel Bargal¹, Jianming Zhang¹, Leonid Sigal², Stan Sclaroff¹
¹ Boston University, ² Disney Research



Main Question

“Can web action images be leveraged to train better CNN models and to reduce the burden of curating large amounts of training videos?”

Motivation

- Labeled web images tend to contain **discriminative** action **poses**, which highlight discriminative portions of a video’s temporal progression.
 - n images contain more unique content compared to n video frames, and images are easier to collect.
- Clearly, there exists a compromise between temporal information available in videos and discriminative poses and variety of unique content in images.

BU 101 Dataset



We collect action images that correspond with the 101 action classes in the **UCF101** video dataset. We manually filter for duplicate and irrelevant images *eg.* drawings or cartoons.

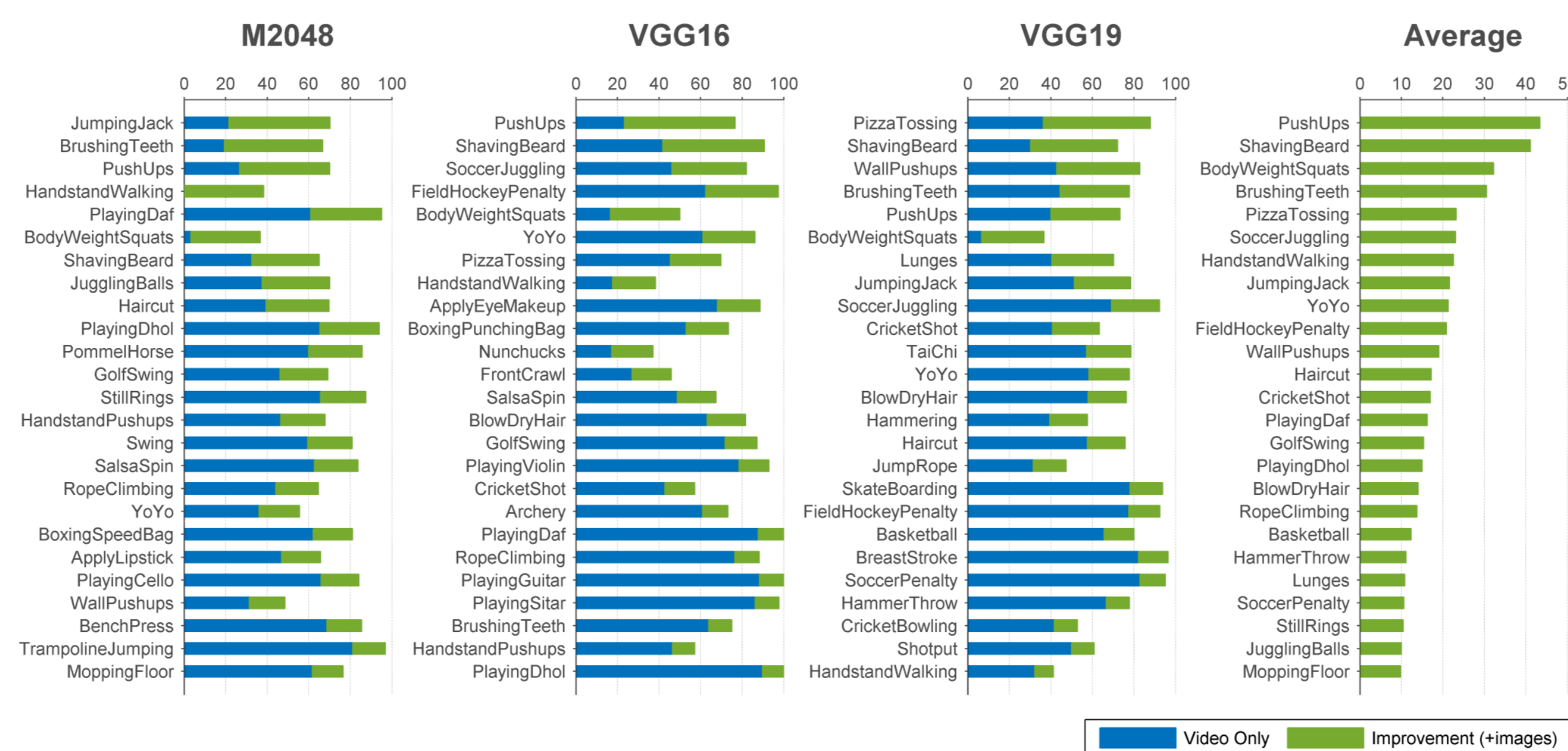
Images beneficial irrespective of CNN depth?

Three CNN models are used for action recognition on the dataset UCF 101 split 1. All architectures benefit.

Model	# layers	# param. (in Millions)	Accuracy video only	Accuracy video + images
M2048	7	91	66.1%	75.2%
VGG16	16	138	77.8%	83.5%
VGG19	19	144	78.8%	83.5%

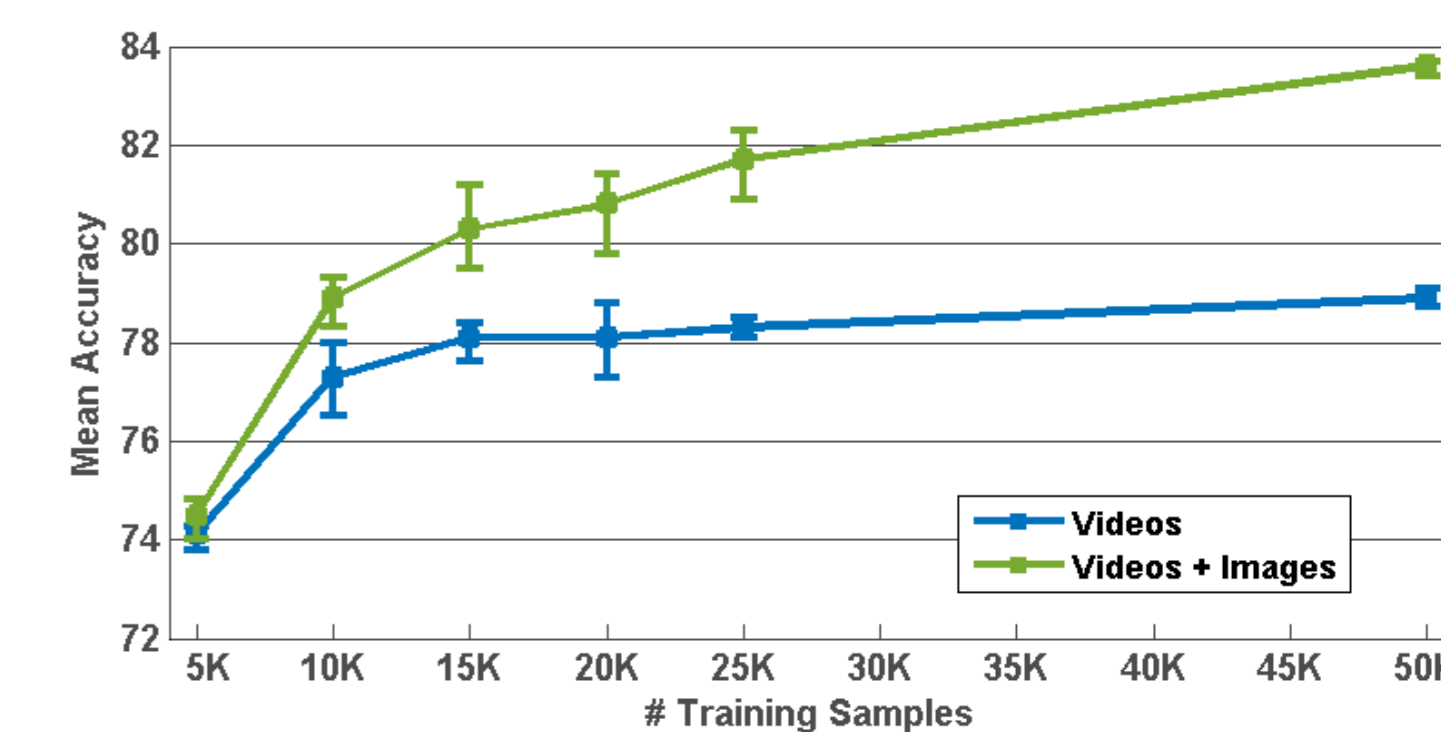
Which classes benefit most?

For UCF101 split 1, the top 25 classes benefiting from adding images are presented (absolute improvement).



Do images complement videos?

A consistent improvement in performance is achieved when half the video frames are replaced by web images on split 1 of UCF101 trained on VGG16.



State-of-the-art performance on UCF101

We obtain state-of-the-art performance when adding images and using motion features: Improved dense trajectories.

Model	Accuracy (%)
IDT-FV [Wang et al. ICCV'13]	85.9
Two-stream CNN [Simonyan et al. NIPS'14]	88.0
RCNN using LSTM [Ng et al. arXiv'15]	88.6
VGG16 + Images + IDT-FV	91.1

Scalability: ActivityNet

We test our approach by collecting a crawled unfiltered dataset for the larger scale dataset ActivityNet: ~800 hrs of video.

- State-of-the-art results on ActivityNet.
- Replacing 16.2M frames by 393K images obtains comparable accuracy.

Experiment	# Frames	# Images	mAP (%)
All vids	32.3M	none	47.7
1/2 vids	16.2M	none	40.9*
1/4 vids	8.1M	none	33.4*
1/2 vids + imgs	16.2M	393K	46.3*
1/4 vids + imgs	8.1M	393K	41.7*

Conclusion

We proposed a filtering technique for data of action videos, thereby reducing the amount of curated training videos needed.

