

Beyond Humanity: Leveraging Pre-trained Human Video Classification Models for Data-Efficient Cross-Species Wildlife Animal Action Recognition

Wenxin (Rose) Zhao
Dartmouth College
15 Thayer Dr, Hanover, NH 03755
wenxin.zhao.gr@dartmouth.edu

Abstract

This paper presents a transfer-learning approach for data-efficient video-based cross-species wildlife animal action recognition, using a pretrained model on human action dataset. It bridges the gap between the well-studied human-focused video classification and under-investigated animal action recognition, largely limited by insufficient structured, annotated data across animal species. By leveraging the SlowFast framework, a state-of-the-art architecture for video classification, and conducting on a small sample of the Animal Kingdom dataset, a benchmark on animal action recognition, the paper reveals a notable improvement in the mean Average Precision (mAP) score, with much fewer training data, when fine-tuned on a model pretrained with Kinetics-400 as compared to training from scratch or utilizing image-based model pretrained on ImageNet. This research demonstrated the promising nature of cross-domain transfer learning in video classification and has substantial inspiration for the advancement of animal behavior understanding and biodiversity conservation.

1. Introduction

Computer Vision has become invaluable in fostering global biodiversity conservation, through global-scale camera-trap biodiversity monitoring [1] [2]. The task of action recognition in videos has gained significant attention among the computer vision communities. While there have been substantial advancements in human action recognition, the same cannot be said for animal action recognition, primarily due to the limited availability of structured, annotated data for a wide range of species [3]. This poses a significant challenge in developing generalized models for animal action recognition across various species [4].

This project aims to tackle animal action recognition in videos, with a particular focus on developing a model capable of identifying actions among a wide range of ani-

mal species with limited data. Our primary focus will be to explore whether leveraging pre-trained models on human actions can be an effective transfer-learning technique and improve performance when applied to animal action recognition. Specifically, with Facebook’s Slowfast Framework [5], a SOTA architecture specializing in video classification, a pre-trained model using human action datasets will be fine-tuned on wildlife animal videos with labeled actions. By utilizing pre-trained models, we hope to take advantage of the knowledge acquired from the more extensive and diverse human action datasets, thereby mitigating the impact of limited data availability and advancing the state-of-the-art in cross-species action recognition.

2. Related Work

In the literature, numerous approaches have been developed for action recognition in videos, such as SlowFast [5], TimeSformer [6], and videoMAE [7]. However, these SOTAs are all trained on human datasets, such as Kinetics 400/600 [8], ActivityNet [9], HMDB [10], and UCF [11], largely because they are large-scale, structured, and accessible.

Current endeavors to animal action recognition, on the other hand, are limited, most of which focus on one specific species such as mammals [12]. Research such as [13] [14] [15] extracted skeletons of the animals and made predictions based on the relative location of the skeletons, a popular technique called pose estimation. However, such an approach can be limited when applied to wildlife camera traps, because different species would have drastically different anatomy and different movement patterns, and some actions can also be context-based [3]. There have not been notable attempts to create a generalized, foundational model across species using video inputs.

Furthermore, most animal datasets are only on very few types of animals such as cows [16], mice [14], monkeys [17], and fish [18], and usually in a controlled or lab environment. The Animal Kingdom dataset [3] stands out as the

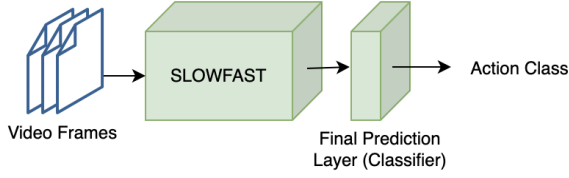


Figure 1. Training Pipeline: The Model takes in videos frames as input, and train/fine-tune on Slowfast architecture attached with the final classifier layer with custom labels, and outputs a predicted action class with a confidence score.

largest existing benchmark on cross-species action recognition for wildlife animals. The dataset contains 50h of video footage with annotations of 140 action classes across 850 species. On average, a video lasts 6 seconds, with a range between 1 to 117 seconds, and always contains at least one animal. This dataset stands out as a suitable candidate to build a generalized animal action recognition model.

This project seeks to bridge the gap between the advancement of human video classification models and animal behavior analysis, by leveraging an existing model trained on human actions to create a generalized model for wildlife animals.

3. Proposed Approach

In this project, we presented comparisons between training on the animal action dataset directly, fine-tuning a model pre-trained with human actions, and fine-tuning a model pre-trained with generic image-based object identification data. We also investigated model performance using fewer training data sizes, currently the bottleneck for biodiversity AI research [4].

We used the Animal Kingdom dataset as the training dataset. Wildlife conservation researchers are constantly in the fields around the world and often don't have a designated lab. It is common for them to encounter the limitation of computing power and data storage resources. Inspired by the circumstance, the dataset is sampled down to 9 action classes, each class with 100 randomly selected training videos, 10 validation videos, and 10 test videos. To limit the scope of the action recognition task, we use videos with only one action label and one kind of animal per clip.

The SlowFast framework [5] is implemented to train the data. Specifically, experiments were conducted where we first trained a model from scratch. Then we fine-tune a model pretrained on K400 (Kinetics-400, the human action video dataset) with the same training dataset and configurations, and compare their performances. Furthermore, to show how temporal human actions are more useful as pre-training dataset rather than generic image-based knowledge, we fine-tune another model pretrained with ImageNet (a

Model	10/class	100/class
From Scratch	0.27211	0.32320
Pre-trained K400	0.45641	0.53707
Pre-trained ImageNet	0.18941	0.33044

Table 1. mAP Score Results for models with different pretrained weights after training on 10 and 100 videos per class.

large image dataset for generic object detection) [19] and compare their performances. Lastly, to investigate the performance with limited training data size, the models were trained with only 10 training videos and 5 test videos per class, and then compared with ones utilizing all 900 training videos. Following [3] and [20], mean Average Precision (mAP) is used as the evaluation metric for each model. It is computed as the unweighted mean of all the per-class average precisions (AP) [21]. For each test video, the model predicts one or more action labels each associated with a confidence score. The evaluation then takes the predictions and the confidence scores to compute the Average Precision across all of the predictions and all the videos. Formally, AP is calculated as follows:

$$AP = \sum_{i=1}^N p(i) \Delta r(i) \quad (1)$$

where N is the number of predictions, p(i) is the precision, and r(i) is the recall [22].

4. Experimental Results

In the experiment, the videos are conformed with the required 30fps for the Slowfast framework and extracted into frames. For each input clip, Slowfast processes with a spatial crop size of 256, a video sampling rate of 2, and 8 frames per clip. Then we perform data augmentation on the sampled frames, specifically, random horizontal flip and adding PCA jittering with scales [256, 340]. The Slowfast architecture is set up where the inverse of the channel reduction ratio between the Slow and Fast pathways is 8, the frame rate reduction ratio between the Slow and Fast pathways is 4, the ratio of channel dimensions between the Slow and Fast pathways is 2, and Kernel dimension used for fusing information from Fast pathway to Slow pathway is 7. Then each model was trained with an SGD optimizer, 0.5 dropout rate, a cross-entropy loss function, a batch size of 8, and a sigmoid function on the activation layer for the output head. The learning rate starts out as 0.00085 and warms up linearly in each iteration until reaching 0.0375 on the fifth epoch, and keeps constant at 0.0375 for the remaining epochs. The total number of epochs to train is 20.



Video		
Ground Truth	Swimming	Eating
K400	Swimming	Keeping Still
From Scratch	Jumping	Eating

Table 2. Top-1 Prediction on two examples videos by K400 model and model from scratch. On the video of otters swimming, K400 correctly identifies while the one from scratch confuses the up/down wavy motion with jumping. In the Kangaroo video, the kangaroos displayed no motion and K400 misinterprets it as keeping still.

4.1. Quantitative Results

Tab. 1 shows the result of the experiments. The overall best-performing model is the one pretrained on K400 with 100 videos per action class. First, the mAP score is higher for K400 pretrained model than from scratch, demonstrating that transfer-learning from K400 is effective. Second, the K400 model has higher mAP than ImageNet model, which was trained on images so it had no temporal understanding prior to training. Lastly but notably, K400 model trained with merely 10 videos per class still yields a higher mAP than training from scratch with 100 videos per class, demonstrating its data-efficient learning nature.

Fig. 2 shows the confusion matrix produced by each model trained with 100 videos per class. In Fig. 2d, the K400 model confusion matrix exhibits a darker shade along the diagonal than the other two matrices, indicating a higher number of true positives and true negatives. This suggests the model’s ability in making accurate predictions across different classes.

4.2. Qualitative Analysis

While the K400 model outperforms quantitatively, its qualitative performance reveals areas where it excels and where it falls short. In Tab. 2, the Otter video is an example where the K400 pre-trained model predicted correctly but the one from scratch predicted wrong. The Kinetics-400 dataset contains 2588 footage labeled as swimming [8], and the pre-trained model may have learned to identify the water and waves in the video and associate them with the action swimming. Yet for the one from scratch, it would have a harder time identifying the otters’ movements, which are moving up and down in the water in the video, which could be disguised as jumping. Fig. 2b shows the model often confuses videos with “swimming” as the true label with

“jumping” and “flying”. This confusion also occurs in the K400 model, but with less frequency (0.1 compared to 0.2 for both classes) [Fig. 2d].

On the other hand, the Kangaroo video demonstrates the reverse, where the knowledge of human actions did not help. In the Kangaroo video, the animals barely moved in the video frames, and the kangaroos eating look nothing like humans eating. For this video, the K400 model was confused and concluded the result as “keeping still”. However, the model trained from scratch, which has more focus on animals and their actions, demonstrated a correct prediction.

5. Conclusion

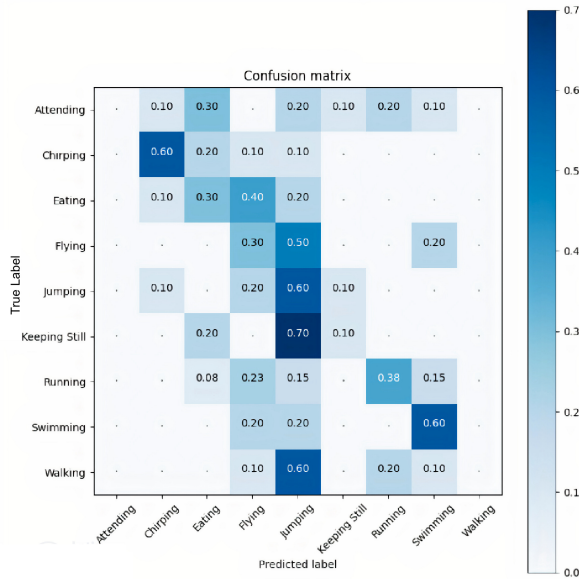
This paper demonstrates the effectiveness and data efficiency of transfer learning from the K400 human action videos to the cross-species animal action recognition task. Future work includes implementing more advanced video classification frameworks, including TimeSformer and videoMAE, as well as including a wider range of action classes, multi-action labels, and multiple animal species in the same frame.

6. Task Assignment

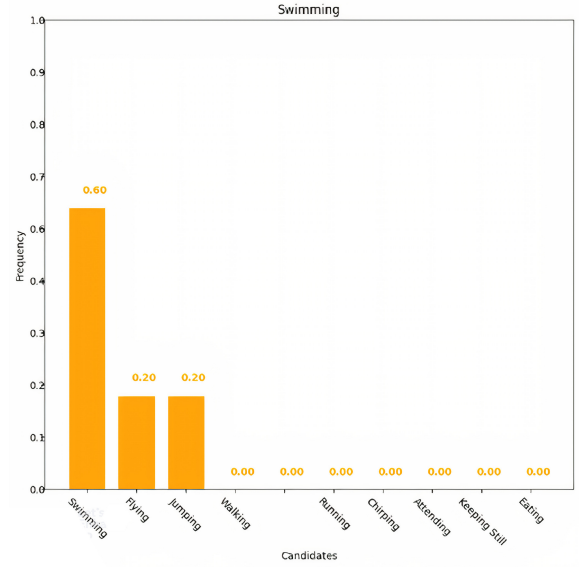
This project was conducted solely by the author, with help from Professor SouYoung Jin.

References

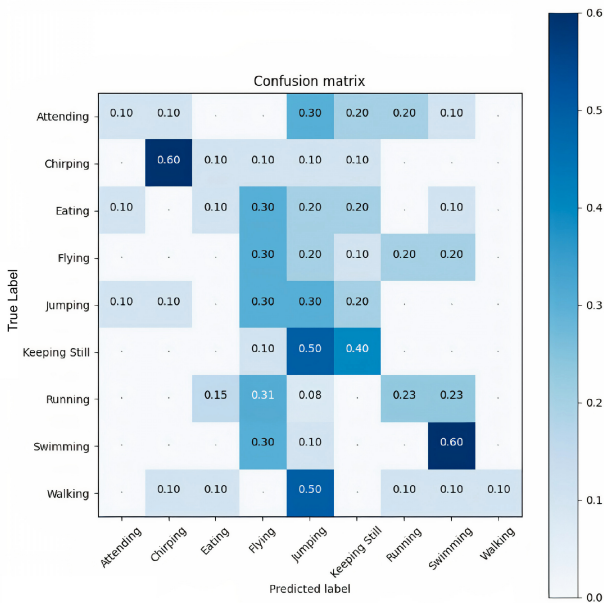
- [1] Fabiola Iannarilli, Ruth Oliver, Tanya Birch, Sara Beery, Eric Fegraus, Nicole Flores, Roland Kays, Jorge Ahumada, and Walter Jetz. Wildlife insights: How camera trap data can foster global biodiversity conservation. 2022. 1



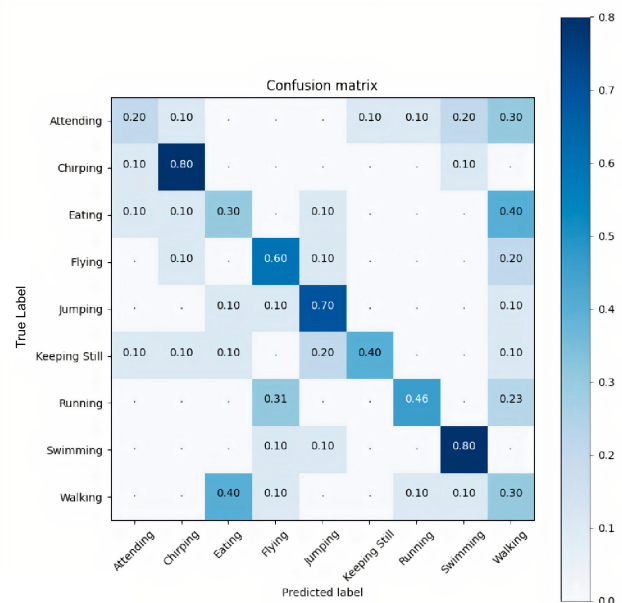
(a) Confusion Matrix for Model Trained from Scratch.



(b) Top Predictions from Model Trained from Scratch for videos with Swimming as Ground Truth. Swimming action is often confused with Flying and Jumping.



(c) Confusion Matrix for Model Pretrained with ImageNet.



(d) Confusion Matrix for Model Pretrained from with Kinetics-400.

Figure 2. (a,c,d) K400 model displays darker shades along the diagonal than other models, showing more True Positives and True Negatives. (b) Some action predictions may be confused with similar motions.

[2] Abhinav Singh, Marcin Pietrasik, Gabriell Natha, Nehla Ghouaie, Ken Brizel, and Nilanjan Ray. Animal detection in man-made environments. *CoRR*, abs/1910.11443, 2019. 1

[3] Xun Long Ng, Kian Eng Ong, Qichen Zheng, Yun Ni, Si Yong Yeo, and Jun Liu. Animal kingdom: A large and diverse dataset for animal behavior understanding. In *Pro-*

ceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 19023–19034, June 2022. 1, 2

[4] Lukas Ziegler, Oliver Sturman, and Johannes Bohacek. Big behavior: challenges and opportunities in a new era of deep behavior profiling. *Neuropsychopharmacology*, 46, 06 2020.

1, 2

- [5] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. *CoRR*, abs/1812.03982, 2018. 1, 2
- [6] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? *CoRR*, abs/2102.05095, 2021. 1
- [7] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training, 2022. 1
- [8] 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. *CoRR*, abs/1705.06950, 2017. 1, 3
- [9] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Nieves. Activitynet: A large-scale video benchmark for human activity understanding. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 961–970, 2015. 1
- [10] 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)*, 2011. 1
- [11] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. *CoRR*, abs/1212.0402, 2012. 1
- [12] Hang Yu, Yufei Xu, Jing Zhang, Wei Zhao, Ziyu Guan, and Dacheng Tao. AP-10K: A benchmark for animal pose estimation in the wild. *CoRR*, abs/2108.12617, 2021. 1
- [13] Liqi Feng, Yaqin Zhao, Yichao Sun, Wenxuan Zhao, and Jiaxi Tang. Action recognition using a spatial-temporal network for wild felines. *Animals*, 11(2), 2021. 1
- [14] Cristina Segalin, Jalani Williams, Tomomi Karigo, May Hui, Moriel Zelikowsky, Jennifer J Sun, Pietro Perona, David J Anderson, and Ann Kennedy. The mouse action recognition system (mars) software pipeline for automated analysis of social behaviors in mice. *eLife*, 10:e63720, nov 2021. 1
- [15] Jessy Lauer, Mu Zhou, Shaokai Ye, William Menegas, Tanmay Nath, Mohammed Mostafizur Rahman, Valentina Di Santo, Daniel Soberanes, Guoping Feng, Venkatesh N. Murthy, George Lauder, Catherine Dulac, Mackenzie W. Mathis, and Alexander Mathis. Multi-animal pose estimation and tracking with deeplabcut. *bioRxiv*, 2021. 1
- [16] Yun Liang, Fuyou Xue, Xiaoming Chen, Zexin Wu, and Xiangji Chen. A benchmark for action recognition of large animals. In *2018 7th International Conference on Digital Home (ICDH)*, pages 64–71, 2018. 1
- [17] Yuan Yao, Praneet Bala, Abhiraj Mohan, Eliza Bliss-Moreau, Kristine Coleman, Sienna M Freeman, Christopher J Machado, Jessica Raper, Jan Zimmermann, Benjamin Y Hayden, et al. Openmonkeychallenge: Dataset and benchmark challenges for pose estimation of non-human primates. *International Journal of Computer Vision*, 131(1):243–258, 2023. 1
- [18] Justin Kay, Peter Kulits, Suzanne Stathatos, Siqi Deng, Erik Young, Sara Beery, Grant Van Horn, and Pietro Perona. The caltech fish counting dataset: A benchmark for multiple-object tracking and counting, 2022. 1
- [19] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 2
- [20] Gunnar A. Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. *CoRR*, abs/1604.01753, 2016. 2
- [21] Shweta Bhardwaj, Mukundhan Srinivasan, and Mitesh M. Khapra. Efficient video classification using fewer frames. *CoRR*, abs/1902.10640, 2019. 2
- [22] Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. Youtube-8m: A large-scale video classification benchmark. *CoRR*, abs/1609.08675, 2016. 2