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

Wenxin Zhao[1]

[1]Dartmouth College, 15 Thayer Dr, Hanover, NH 03755, United States

Abstract

This paper presents a transfer learning approach for data-efficient video-based multi-species wildlife animal action recognition, using pre-trained models on human action datasets. 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 pre-trained with Kinetics-400 as compared to training from scratch or utilizing image-based model pre-trained on ImageNet. This research demonstrated the promising nature of cross-domain transfer learning for video classification and has substantial inspiration for advancing the understanding of animal behavior and biodiversity conservation.

Keywords

Transfer Learning, Video Classification, Wildlife Conservation, Action Recognition

1. Introduction

Computer Vision has become invaluable in fostering global biodiversity conservation, through global-scale camera-trap biodiversity monitoring [1][2] and through increasingly capable models and more computational power available. The task of video classification, especially human action classification, has gained significant attention among the computer vision communities [3][4][5]. While there have been substantial advancements in human action recognition [6][7], 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 [8]. This poses a significant challenge in developing generalized models for animal action recognition across various species [9].

This paper aims to tackle animal action recognition in videos, focusing on developing a model capable of identifying actions among a wide range of animal species with limited data. This can allow wildlife researchers to focus more on analysis than manual data collections[10], and inspire further studies for a deeper understanding of how and why animals behave [11]. 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, as opposed to training from scratch. Specifically, with Facebook's SlowFast Framework [12], a state-of-the-art architecture specializing in video classification, two pre-trained models on Kinetics-400 and ImageNet using human action datasets will be fine-tuned on wildlife animal videos with labeled actions. By utilizing pre-trained models, we hope to use 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 multi-species action recognition.

4th International Workshop on Camera Traps, AI, and Ecology, September 5 – 6, 2024, Hagenberg, Austria

wenxin.zhao.gr@dartmouth.edu (W. Zhao)

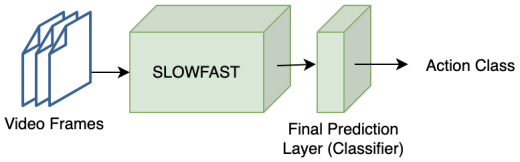


Figure 1: Training Pipeline: The Model takes video frames as input, trains and fine-tunes on SlowFast architecture attached to the final classifier layer with custom labels, and outputs a predicted action class with a confidence score.

2. Related Work

In the literature, numerous approaches have been developed for action recognition in videos, such as SlowFast [12], TimeSformer [13], and videoMAE [14]. However, these state-of-the-art models are all trained on human datasets, such as Kinetics 400/600 [15], ActivityNet [16], and UCF [3], largely because they are large-scale, structured, and accessible.

Current endeavors at animal action recognition, on the other hand, are limited. Research such as [17] [18] [19] [20] extracted skeletons of the animals and made predictions based on the relative motions of the joints, a popular technique called pose estimation. However, such an approach can be limited when applied to wildlife camera trap data, because different species would have drastically different anatomy and movement patterns, and some actions can also be context-based [8]. There have not been notable attempts to create a generalized, foundational model across species using video inputs.

Furthermore, most animal datasets contain only a few types of animals such as cows [21], mice [18], monkeys [22], apes [20] and fish [23], or a specific animal class such as mammals [24], and usually in a controlled or lab environment. The Animal Kingdom dataset [8] stands out as the largest existing benchmark on multi-species action recognition for wildlife animals. The dataset contains 50 hours 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 for building a generalized animal action recognition model.

This paper 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 paper, we presented comparisons between training on the animal action dataset from scratch, 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 [9].

We used the Animal Kingdom dataset as the training dataset. To limit the scope of the action recognition task, we use videos with only one action label and one animal species per clip, as opposed to multiple labels or species in one clip. Wildlife conservation researchers spend much of their time in the field worldwide with limited computing power and data storage resources. Inspired by the circumstance, we filter the training data to have only the 9 most labeled actions in the dataset defined in Table 3. Each class consists of 100 randomly selected training videos, 10 validation videos, and 10 test videos.

Figure 1 shows the training pipeline using the SlowFast framework. Videos are extracted into individual image frames to feed into the SlowFast architecture, where they go through two parallel convolution neural networks (the Slow pathway and the Fast pathway) [12]. At the end, we add a classifier layer (and discard the original classifier layer if using pre-trained models) that outputs the predictions of the nine action labels. We first trained a model from scratch (random initialization of weights) as our baseline result. Then we obtained weights of a model pre-trained on Kinetics-400 (K400,

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
Pre-trained K400 model shows the best mAP score in both cases training on 10 and 100 videos per class.

the human action video dataset) and used the same training dataset and configurations to fine-tune the weights and compare their performances. Furthermore, to show how temporal human actions can be more useful as a pre-training dataset than generic visual feature knowledge, we fine-tuned another model pre-trained with ImageNet (a large image dataset for generic object detection) [25] and compared 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 [8] and [26], 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 precision (AP), bounded between 0 and 1 [27]. 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 predictions and videos. Formally, AP is calculated as follows:

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

where N is the number of predictions, p(i) is the precision, and r(i) is the recall [5]. mAP is then calculated by taking the mean of these AP values as follows:

$$\text{mAP} = \frac{1}{N}\sum_{i=1}^{N} \text{AP}_i \tag{2}$$

where N is the number of classes.

3.1. Model Setup

Overall, the models underwent supervised learning with the labeled training data. In the experiment, the videos were conformed with the required 30 frames per second for the SlowFast framework and extracted into individual image 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 performed data augmentation on the sampled frames, specifically, random horizontal flip and adding Principal Components Analysis (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. Weights of both pre-trained models are obtained from SlowFast's official GitHub repository. Then each model was trained with a Stochastic Gradient Descent optimizer, a dropout rate of 0.5, 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 started as 0.00085 and warmed up linearly in each iteration until reaching 0.0375 on the fifth epoch, and kept constant at 0.0375 for the remaining epochs. The total number of epochs to train is 20.

Video		
Ground Truth K400 From Scratch	Swimming **Swimming** Jumping	Eating Keeping Still **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 misinterpreted it as keeping still.

4. Experimental Results

4.1. Quantitative Results

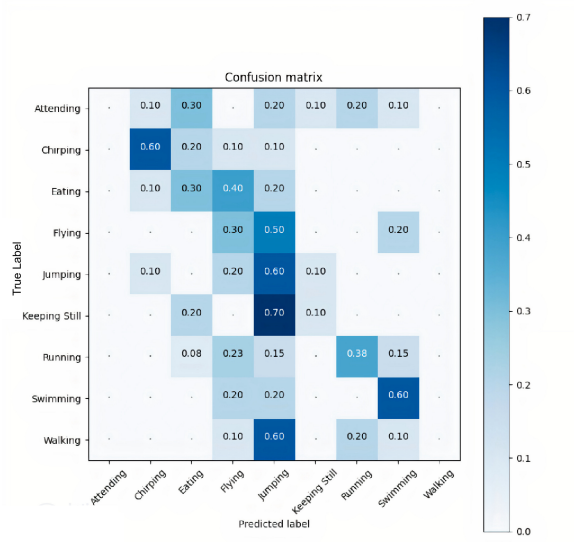
table 1 shows the result of the experiments, where the overall best-performing model is the one pre-trained on K400 with 100 videos per action class. First, the mAP score is higher for the K400 pre-trained model than from scratch, demonstrating that transfer learning from K400 is effective. On the other hand, the mAP of the ImageNet model shows an insignificant increase from the model from scratch, much less than that of the K400 model. It suggests that the action recognition model benefits more through transfer learning from a model with temporal understanding than a generic image classification model. Furthermore, the 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 to make accurate predictions across different classes.

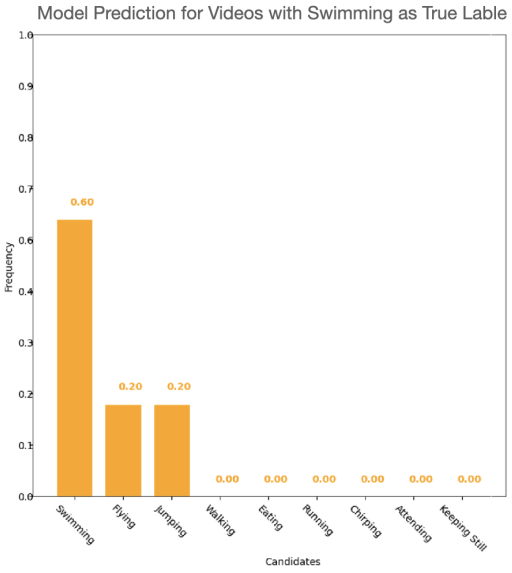
4.2. Qualitative Analysis

While the K400 model outperforms quantitatively, its qualitative performance reveals areas where it excels and where it falls short. To demonstrate, both the model from scratch and the K400 model were applied to unseen videos. In table 2, the Otter video is an example where the K400 pre-trained model predicted correctly but the one from scratch predicted wrong. The K400 dataset contains 2588 footage labeled as swimming [15], 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 the model from scratch had a harder time identifying the otters' movements (moving up and down in the water) in the video, which could be disguised as jumping. fig. 2b shows the model from scratch 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].

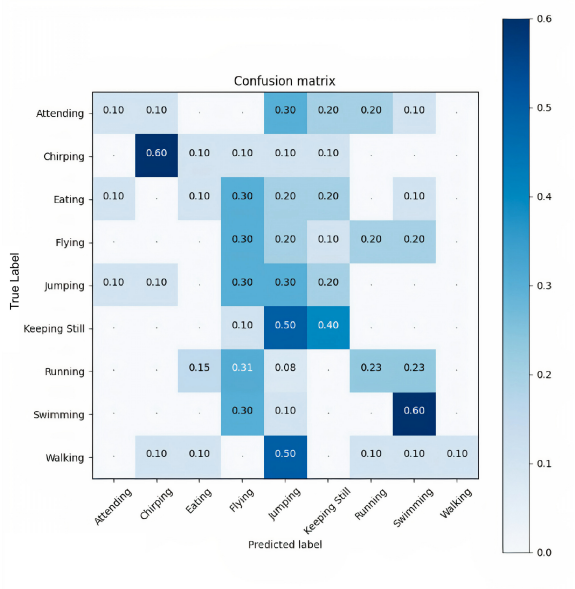
On the other hand, the Kangaroo video demonstrates the reverse, where knowledge of human actions did not help. In the Kangaroo video, the animals barely moved in the video frames, and the kangaroos eating looked nothing like humans eating. For this video, the K400 model was confused, and concluded the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



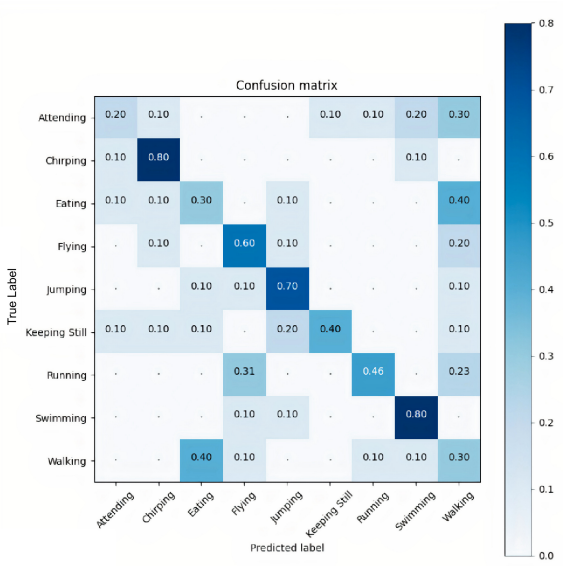
(a) Confusion Matrix for Model Trained from Scratch.



(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 pre-trained with ImageNet.



(d) Confusion Matrix for Model pre-trained with Kinetics-400.

Figure 2: (a,c,d) The 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.

5. Conclusion

This paper demonstrates the effectiveness and data efficiency of transfer learning from the K400 human action videos to the multi-species animal action recognition task, which outperforms ImageNet and models trained from scratch. Future work includes implementing more advanced video classification frameworks, including TimeSformer and videoMAE, incorporating a wider range of action classes, multi-action labels, and multiple animal species in the same frame, and evaluating the K400 model more

comprehensively with more test data to reveal actions it excels and confuses the most.

6. Acknowledgement

This work was advised by Dr. SouYoung Jin from Dartmouth College and sponsored by the Department of Computer Science at Dartmouth College.

Category	Action	Description
General	Keeping still	Animal makes no or minimal movement (i.e., animals staying still and alert)
Feeding	Eating	Include feeding, grazing, and gnawing
Sensing	Attending	Animal locates a stimulus of potential interest, and directs its attention (eyes, ears, face) towards it, and often keeping very still to observe the situation
Movement	Swimming	Animal swims in the water (e.g. fish), or on the surface of water (e.g. water birds)
Movement	Jumping	Animal makes large jumping movement from one spot to another (e.g. from lower to higher grounds), or on the same spot
Movement	Walking	Animal moves from one spot to another in a slow pace
Movement	Running	
Movement	Flying	
Communication	Chirping	

Table 3
Descriptions of the 9 most labeled actions used for training from the Animal Kingdom Dataset [8]

References

[1] F. Iannarilli, R. Oliver, T. Birch, S. Beery, E. Fegraus, N. Flores, R. Kays, J. Ahumada, W. Jetz, Wildlife insights: How camera trap data can foster global biodiversity conservation (2022).

[2] A. Singh, M. Pietrasik, G. Natha, N. Ghouaiel, K. Brizel, N. Ray, Animal detection in man-made environments, CoRR abs/1910.11443 (2019). URL: http://arxiv.org/abs/1910.11443. arXiv:1910.11443.

[3] K. Soomro, A. R. Zamir, M. Shah, UCF101: A dataset of 101 human actions classes from videos in the wild, CoRR abs/1212.0402 (2012). URL: http://arxiv.org/abs/1212.0402. arXiv:1212.0402.

[4] R. Goyal, S. Ebrahimi Kahou, V. Michalski, J. Materzynska, S. Westphal, H. Kim, V. Haenel, I. Fruend, P. Yianilos, M. Mueller-Freitag, F. Hoppe, C. Thurau, I. Bax, R. Memisevic, The "something something" video database for learning and evaluating visual common sense, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017.

[5] S. Abu-El-Haija, N. Kothari, J. Lee, P. Natsev, G. Toderici, B. Varadarajan, S. Vijayanarasimhan, Youtube-8m: A large-scale video classification benchmark, CoRR abs/1609.08675 (2016). URL: http://arxiv.org/abs/1609.08675. arXiv:1609.08675.

[6] B. Jiang, M. Wang, W. Gan, W. Wu, J. Yan, Stm: Spatiotemporal and motion encoding for action recognition, 2019. URL: https://arxiv.org/abs/1908.02486. arXiv:1908.02486.

[7] D. Lee, J. Lee, J. Choi, Cast: Cross-attention in space and time for video action recognition, 2023. URL: https://arxiv.org/abs/2311.18825. arXiv:2311.18825.

[8] X. L. Ng, K. E. Ong, Q. Zheng, Y. Ni, S. Y. Yeo, J. Liu, Animal kingdom: A large and diverse dataset for animal behavior understanding, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 19023–19034.

[9] L. Ziegler, O. Sturman, J. Bohacek, Big behavior: challenges and opportunities in a new era of deep behavior profiling, Neuropsychopharmacology 46 (2020). doi:10.1038/s41386-020-0751-7.

[10] E. Fazzari, D. Romano, F. Falchi, C. Stefanini, Animal behavior analysis methods using deep learning: A survey, 2024. URL: https://arxiv.org/abs/2405.14002. arXiv:2405.14002.

[11] A. E. Brown, B. de Bivort, Ethology as a physical science, bioRxiv (2018). URL: https://www.biorxiv.org/content/early/2018/02/02/220855. doi:10.1101/220855. arXiv:https://www.biorxiv.org/content/early/2018/02/02/220855.full.pdf.

[12] C. Feichtenhofer, H. Fan, J. Malik, K. He, Slowfast networks for video recognition, CoRR abs/1812.03982 (2018). URL: http://arxiv.org/abs/1812.03982. arXiv:1812.03982.

[13] G. Bertasius, H. Wang, L. Torresani, Is space-time attention all you need for video understanding?, CoRR abs/2102.05095 (2021). URL: https://arxiv.org/abs/2102.05095. arXiv:2102.05095.

[14] Z. Tong, Y. Song, J. Wang, L. Wang, Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training, 2022. arXiv:2203.12602.

[15] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, A. Zisserman, The kinetics human action video dataset, CoRR abs/1705.06950 (2017). URL: http://arxiv.org/abs/1705.06950. arXiv:1705.06950.

[16] F. C. Heilbron, V. Escorcia, B. Ghanem, J. C. Niebles, Activitynet: A large-scale video benchmark for human activity understanding, in: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 961–970. doi:10.1109/CVPR.2015.7298698.

[17] L. Feng, Y. Zhao, Y. Sun, W. Zhao, J. Tang, Action recognition using a spatial-temporal network for wild felines, Animals 11 (2021). URL: https://www.mdpi.com/2076-2615/11/2/485. doi:10.3390/ani11020485.

[18] C. Segalin, J. Williams, T. Karigo, M. Hui, M. Zelikowsky, J. J. Sun, P. Perona, D. J. Anderson, A. Kennedy, The mouse action recognition system (mars) software pipeline for automated analysis of social behaviors in mice, eLife 10 (2021) e63720. URL: https://doi.org/10.7554/eLife.63720. doi:10.7554/eLife.63720.

[19] J. Lauer, M. Zhou, S. Ye, W. Menegas, T. Nath, M. M. Rahman, V. D. Santo, D. Soberanes, G. Feng, V. N. Murthy, G. Lauder, C. Dulac, M. W. Mathis, A. Mathis, Multi-animal pose estimation and tracking with deeplabcut, bioRxiv (2021). URL: https://www.biorxiv.org/content/early/2021/04/30/2021.04.30.442096. doi:10.1101/2021.04.30.442096. arXiv:https://www.biorxiv.org/content/early/2021/04/30/2021.04.30.442096.full.pdf.

[20] M. Fuchs, E. Genty, K. Zuberbühler, P. Cotofrei, Asbar: an animal skeleton-based action recognition framework. recognizing great ape behaviors in the wild using pose estimation with domain adaptation, bioRxiv (2023). doi:10.1101/2023.09.24.559236. arXiv:https://www.biorxiv.org/content/early/2023/09/25/2023.09.24.559236.full.pdf.

[21] Y. Liang, F. Xue, X. Chen, Z. Wu, X. Chen, A benchmark for action recognition of large animals, in: 2018 7th International Conference on Digital Home (ICDH), 2018, pp. 64–71. doi:10.1109/ICDH.2018.00020.

[22] Y. Yao, P. Bala, A. Mohan, E. Bliss-Moreau, K. Coleman, S. M. Freeman, C. J. Machado, J. Raper, J. Zimmermann, B. Y. Hayden, et al., Openmonkeychallenge: Dataset and benchmark challenges for pose estimation of non-human primates, International Journal of Computer Vision 131 (2023) 243–258.

[23] J. Kay, P. Kulits, S. Stathatos, S. Deng, E. Young, S. Beery, G. V. Horn, P. Perona, The caltech fish counting dataset: A benchmark for multiple-object tracking and counting, 2022. arXiv:2207.09295.

[24] H. Yu, Y. Xu, J. Zhang, W. Zhao, Z. Guan, D. Tao, AP-10K: A benchmark for animal pose estimation in the wild, CoRR abs/2108.12617 (2021). URL: https://arxiv.org/abs/2108.12617. arXiv:2108.12617.

[25] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 248–255. doi:10.1109/CVPR.2009.5206848.

[26] G. A. Sigurdsson, G. Varol, X. Wang, A. Farhadi, I. Laptev, A. Gupta, Hollywood in homes: Crowdsourcing data collection for activity understanding, CoRR abs/1604.01753 (2016). URL: http://arxiv.org/abs/1604.01753. arXiv:1604.01753.

[27] S. Bhardwaj, M. Srinivasan, M. M. Khapra, Efficient video classification using fewer frames, CoRR abs/1902.10640 (2019). URL: http://arxiv.org/abs/1902.10640. arXiv:1902.10640.