Optimizing Inference in Behavior Classification of Fishes

Manu Tej Sharma Arrojwala
Vineeth Aljapur
Priyam Raut
Sachin Sarath Y Kothandaraman

ABSTRACT

Cichlid fish are amongst the most striking organisms to study behavioral diversity, and thus can serve as a powerful system for understanding the biological basis of natural behaviors. Since analysis of these behavioral trials is associated with video monitoring, automated methods will be necessary to fully analyze the data. 3D Convolutional Neural Networks have been highly successful in applications such as video classification, and action detection, which are essential facets in scientific communities studying live models. However, utilizing 3D CNN models is a time consuming and computationally intensive process. The popularity of Convolutional Neural Network (CNN) models imply that better performance of CNN model inference can deliver significant gain to a large number of users. Our project aims to optimize a 3D ResNet model used to detect fish behaviour, for faster inference with minimal loss of accuracy.

1 PROBLEM STATEMENT

One of the most characteristic behaviors displayed by cichlids is the construction of bowers which are sand structures made by male cichlids to attract females. They do this by building a sand castle, an elevated structure, or by digging a pit, a depression (Fig. 1a). Bowers are constructed by scooping sand from dispersed locations and spitting into a concentrated area to create a castle (Fig. 1b) or scooping sand from a concentrated area and spitting into dispersed locations to create a pit (Fig. 1c) [1]. Unlike most behaviors that are studied in the laboratory, this is repeated for days to weeks at a time during the mating season.

Typically, male cichlids build bowers by freely manipulating sand only in the presence of female cichlid(s). We study bower building in the lab by introducing male and female cichlids into a large sand tank and monitoring the fish using cameras connected to a raspberry pi computer which generates around 10 hours of video recordings daily for each tank. Traditionally, quantification of this social behavior is done manually where an experimenter has to look at entire recordings to find and label bower building events. This ensures that the accuracy of labeling remains high but it comes at the cost of looking at the entire video(s). Since each of our behavioral trials is associated with around 100 hours of video, automated methods will be necessary to fully analyze the data.

Convolutional neural networks with spatio-temporal 3D kernels (3D CNNs) have an ability to directly extract spatiotemporal features from videos for action recognition [2]. Residual networks (ResNets) are one of the most powerful architecture. ResNets[3] introduce shortcut connections that bypass a signal from one layer to the next. The connections pass through the gradient flows of networks from later layers to early layers, and ease the training of very deep networks. Applying the architecture of ResNets to 3D CNNs is expected to contribute further improvements of action recognition performance.

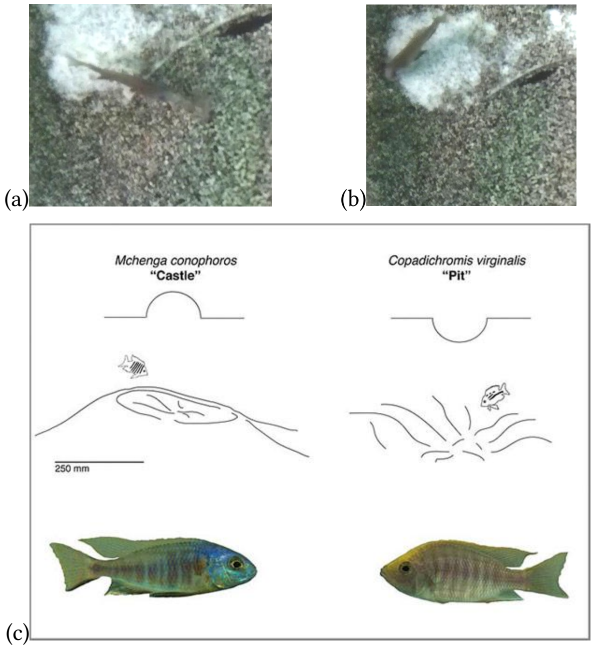


Figure 1: Bower building behavior of Cichlids: (a) spit by Mchenga conophoros (b) scoop by Mchenga conophorosa, (c) Castle building and pit digging in Mchenga conophoros and Copadichromis virginalis respectively

As a first step towards automation of this analysis, we used an out-of-box 3D CNN with ResNet-18 model[4]. The video data was labeled to indicated 10 behavioural classes. The model was trained with the resized data and later tested to determine the accuracy of the classification. However, we identified that using 3D ResNets were slow and the current model would not be optimal to analyze large scale video data. Action classification is slow and memory intensive because of 3D convolution (needs 6x Nvidia 980Ti). Another issue is since 50 % of the frames of our videos are empty, classifiying a set of continous set of empty frames is a waste of resources. Accuracy is not good as the model is not for our dataset. Hence, we intend to explore methods to optimize the speed of using the 3d ResNet model with minimal loss in accuracy of classification.

2 SIGNIFICANCE OF THE PROBLEM

Action classification plays a predominant role in biological research involving live models. Though existing 3D CNN models provide comparatively accurate classifications, they are computationally intensive and quite slow in their inference. Though it is possible to design new models that are optimal for classifying biologically significant behaviours in live models, We believe that identifying optimal ways to increase the speed of inference and reducing the computational load on the existing model with minimal loss of accuracy would be less intensive, and easier to utilize in biological research.

3 METHOD

(1) The first step is to increase the speed of the process. We want to implement this by removing the empty frames in the video which we estimate to be 50 % of the total frames using OpenCV and Shallow Neural Networks.
(2) We also want to add an additional layer of depth data to improve accuracy as well since we can track the extended phenotype from depth data, the Neural Network can use this context for better accuracy.

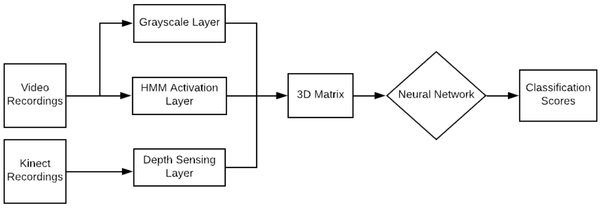
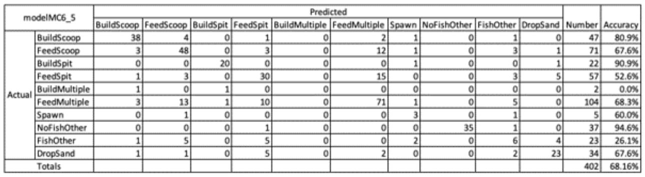


Figure 2: Flow Chart showing the implementation of depth sensing data to improve classification

(3) We would like to tweak 3D ResNet architecture and tune Hyperparameters to improve detection speed and accuracy by using different architectures[5].

4 VALIDATION

Since our current model works, we intend to compare the accuracy of classification from tweaked model with the present results, by checking if the model is capable of classifying the same behavioural classes accurately.



modelMC6_5		Predicted										Number	Accuracy
		BuildScoop	FeedScoop	BuildSpit	FeedSpit	BuildMultiple	FeedMultiple	Spawn	NoFishOther	FishOther	DropSand		
Actual	BuildScoop	38	4	0	1	0	2	1	0	1	0	47	80.9%
	FeedScoop	3	48	0	3	0	12	1	0	3	1	71	67.6%
	BuildSpit	0	0	20	0	0	0	1	0	0	1	22	90.9%
	FeedSpit	1	3	0	30	0	15	0	0	3	5	57	52.6%
	BuildMultiple	1	0	1	0	0	0	0	0	0	0	2	0.0%
	FeedMultiple	3	13	1	10	0	71	1	0	5	0	104	68.3%
	Spawn	0	1	0	0	0	0	3	0	1	0	5	60.0%
	NoFishOther	0	0	0	1	0	0	0	35	1	0	37	94.6%
	FishOther	1	5	0	5	0	0	2	0	6	4	23	26.1%
	DropSand	1	1	0	5	0	2	0	0	2	23	34	67.6%
	Totals											402	68.16%

Figure 3: Current validation for out of box model (3D CNN with ResNet-18) with only video recordings

5 PERFORMANCE EVALUATION

We will evaluate the performance of the model at each stage of the modification to the existing performance times as the baseline. The Inference rate at 75 % , 100 %, and 125% will be compared with the existing model, while ensuring the accuracy of the classification.

If the depth dimension parameter is integrated with the existing parameters, we can validate the accuracy of the tweaked model with the addition of depth dimension vs. without to verify if adding the depth information contributes to significant increase in accuracy.

6 RESOURCES

- Software
  - PyTorch
  - Tensorflow
  - Scikit-Learn,
  - Git
- Hardware
  - GPUs
- Data Sets
  - Custom Dataset of about 700 hr of video of Cichlids

7 GOALS

- 75 %: Adding low-level filter (OpenCV) to improve speed by removing empty frames
- 100 %: Implementing NoScope to remove empty frames by a cascade of shallow NNs
- 125 % To improve the accuracies for the 3D CNN by tuning the architecture and hyperparameters and adding an additional dimension of depth which is obtained by Kinect

REFERENCES

(1) Ryan A. York, Chinar Patil, et al., âĂIJEvolution of bower building in Lake Malawi cichlid fish: phylogeny, morphology, and behaviorâĂİ Front. Ecol. Evol., (2015)
(2) J. Carreira and A. Zisserman. Quo vadis, action recognition? A new model and the kinetics dataset. arXiv preprint, abs/1705.07750, 2017
(3) K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 770âĂŞ778, 2016.
(4) Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh, "Learning Spatio-Temporal Features with 3D Residual Networks for Action Recognition", Proceedings of the ICCV Workshop on Action, Gesture, and Emotion Recognition, 2017.
(5) Hara, Kensho, Hirokatsu Kataoka, and Yutaka Satoh. "Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet?." Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. 2018