AI System for Analyzing Footage from Animal-borne Cameras: a Sea Turtle Case Study

Brian Chen, Brian Allan, Nathan Robinson, Aiman S Soliman

Department of Computer Science, College of Engineering, University of Illinois at Urbana-Champaign

Introduction

- Advanced technologies provide huge amount of video data
- Rich and precise information in videos
- Unlikely to process it manually due to limited resources
- Applied AI/ML tools to analyze videos
- Understand sea turtle's behavior

Objectives

- Classify the behavior of sea turtle in the turtle footage
- Obtain *frame-wise* action labels for videos
- Behaviors include swimming, breathing, surface resting, and more

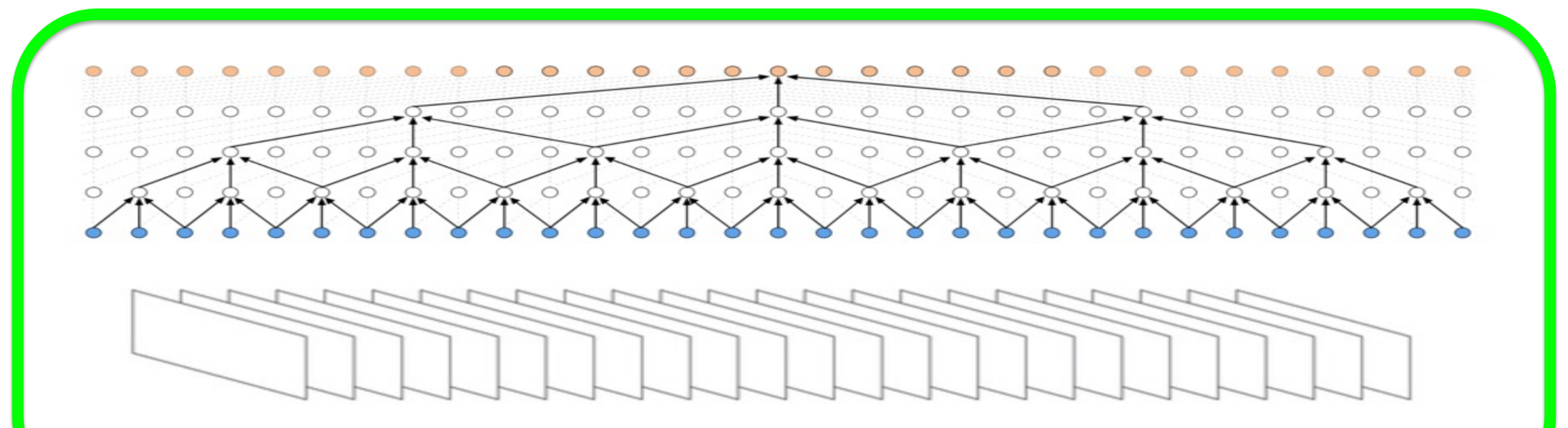
Swimming

Breathing

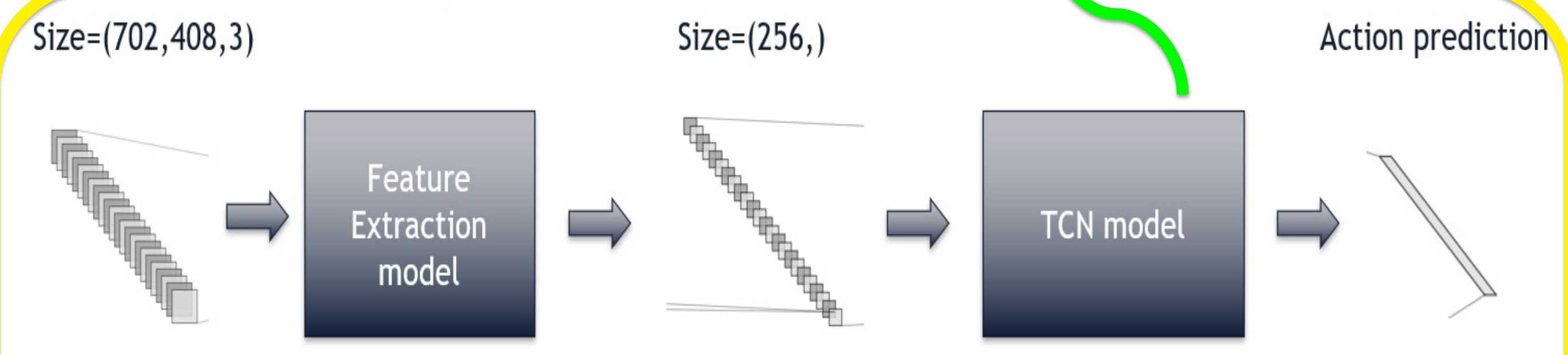
Method

Temporal Convolutional Network (TCN)

- Capture the relationship and dependencies between sequences of inputs
- Efficient and compact - less memory because of dilated convolutions over sequences of inputs



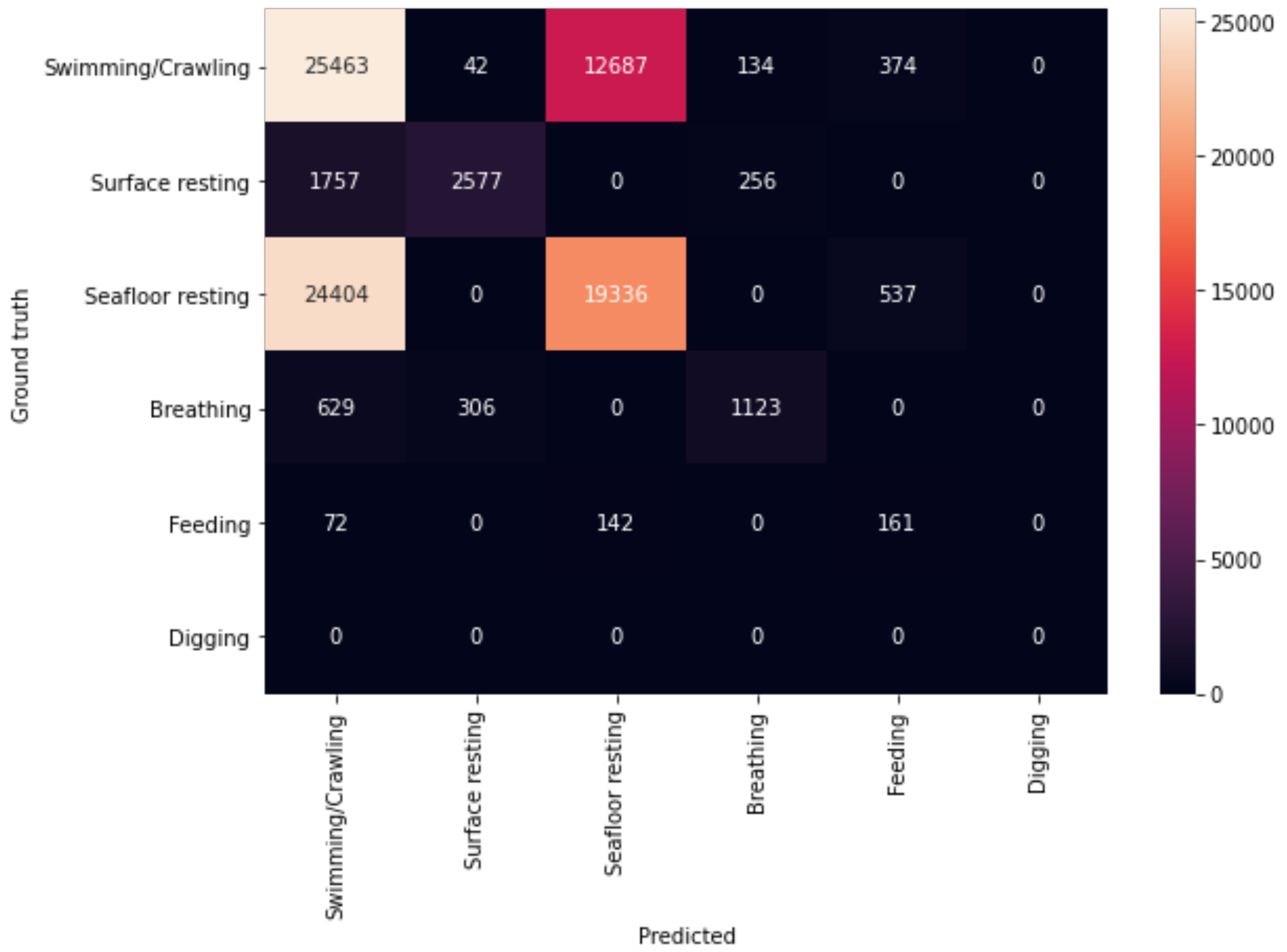
TCN Block



Framework overview; we use VGG16 as backbone in our experiments

Results & Analysis

- Unable to distinguish between swimming and seafloor resting
- We conjecture that the model **relies on the background information of frames** for prediction instead of *capturing motions across frames*
- Majority of feeding behavior being predicted as either seafloor resting or swimming suggests their similarity
- Unbalanced dataset is one of the main reasons of non-optimum performance



Confusion matrix for the model tested on six categories of turtle behavior

Activity	Swim	Surface rest	Seafloor rest	Breath	Feed	Dig
Frequency (%)	46.62	*5.45*	46.59	2.28	9.00	0.05

The frame-wise distribution of our dataset, which is **extremely unbalanced.** Prior works have shown that model tends to cheat by minimizing the loss in an easy way – predicting every frame as the majority category.

fps	1	30
Accuracy	98%	98%

Increasing frames per second (fps) does not improve accuracy. One of the experiments we did is to redo the annotation for our sea turtle dataset and experiment on inputs with different fps. As shown in above table, the extra labor invested and higher temporal frequency for annotation did not improve the performance.

Conclusion

The model has trouble distinguishing activities with similar backgrounds; swimming and resting both occur underwater. Our results suggest current SOTA models rely heavily on the spatial information and characteristics of frames to predict labels, instead of capturing the motions and correspondences across frames. Moreover, as our dataset is extremely unbalanced, the model tends to predict everything as the majority categories. Additionally, we show that using annotated dataset with higher temporal frequency does not improve the performance; one frame per second is sufficient for the purpose of video classification.

Future Work

- Better video classification frameworks such as Non-local network
- Synthetic dataset for fixing the unbalanced dataset
- Triplet loss for distinguishing similar activities

Reference

[1] Yazan Abu Farha and Juergen Gall. Ms-tcn: Multi-stage tem-poral convolutional network for action segmentation, 2019.

[2] lexandros Stergiou and Ronald Poppe. Spatio-temporal fast3d convolutions for human action recognition.2019 18thIEEE International Conference On Machine Learning And Applications (ICMLA), Dec 2019.

Acknowledgements

ILLINOIS
NCSA | National Center for Supercomputing Applications

SPIN

ILLINOIS