dongguk UNIVERSITY

Context R-CNN: Long Term Temporal Context for Per-Camera Object Detection

2021.05.31

Ershang Tian

Introduction

Context R-CNN

- leverages temporal context from the unlabeled frames of a novel camera to improve performance at that camera.
- The attention-based approach, ggregate contextual features from other frames to boost object detection performance on the current frame.

Introduction

Context R-CNN

- The model that can learn how to find and use other potentially easier examples from the same camera to help improve detection performance.

(a) Object moving out of frame.

(b) Object highly occluded.

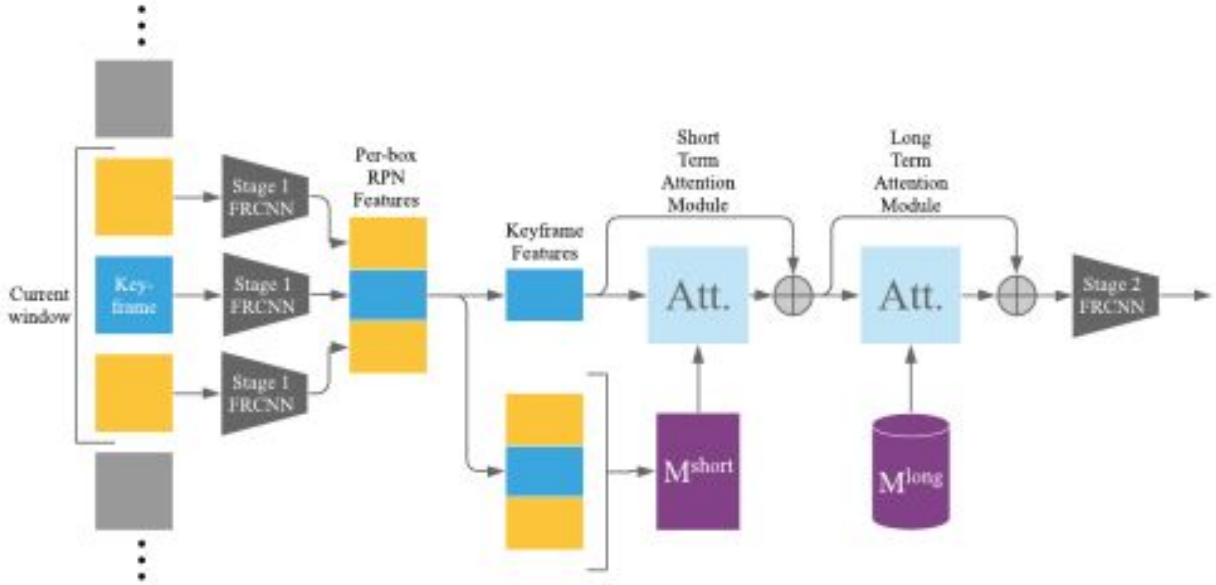
동국대학교
dongguk university

Introduction

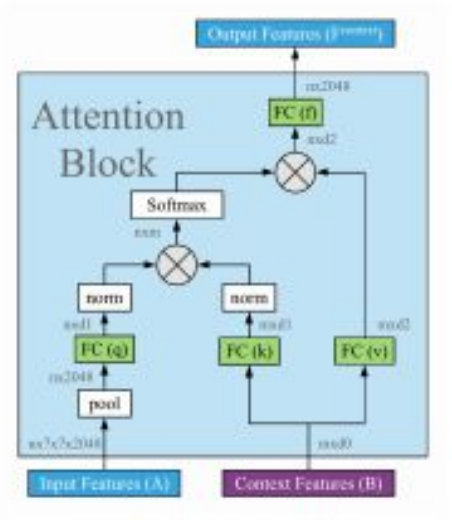
focus on two static-camera domains:

- (1) species detection using camera traps
- (2) vehicle detection in traffic cameras

Context R-CNN Architecture.



(a) High-level Context R-CNN architecture.



(b) Single attention block.

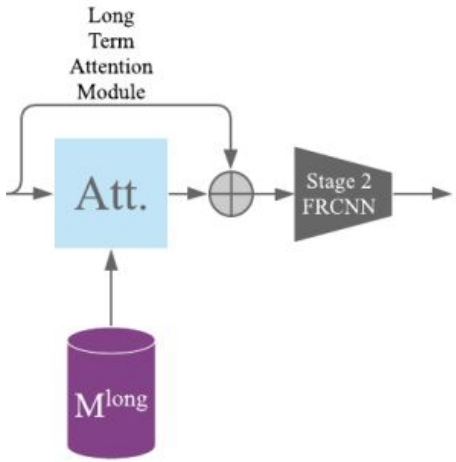
동국대학교
dongguk university

Building a memory bank

- **Long Term Memory Bank (M-long)**

- Given a keyframe i_t, for which want to detect objects, iterate over all frames from the same camera within a pre-defined time $i_{t-k} : i_{t+k}$, -k : it+k, running a frozen, pre-trained detector on each frame.

- Build our long-term memory bank (M-long) from feature vectors corresponding to resulting detections



동국대학교
dongguk university

Building a memory bank

- **Long Term Memory Bank (M-long)**

- Instance-level feature tensors after cropping proposals from the RPN and save only a spatially pooled representation of each such tensor concatenated with a spatiotemporal encoding of the DateTime and box position.

- Curate by limiting the number of proposals for which store features---consider multiple strategies for deciding which and how many features to save to memory banks

- By using these strategies able to construct memory banks holding up to 8500 contextual features — represent a month's worth of context from a camera.

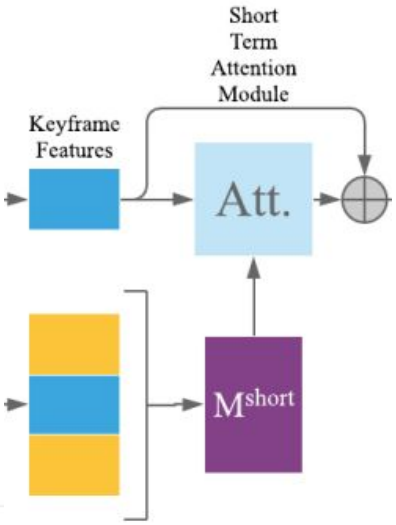
Building a memory bank

- Short Term Memory (M-short).

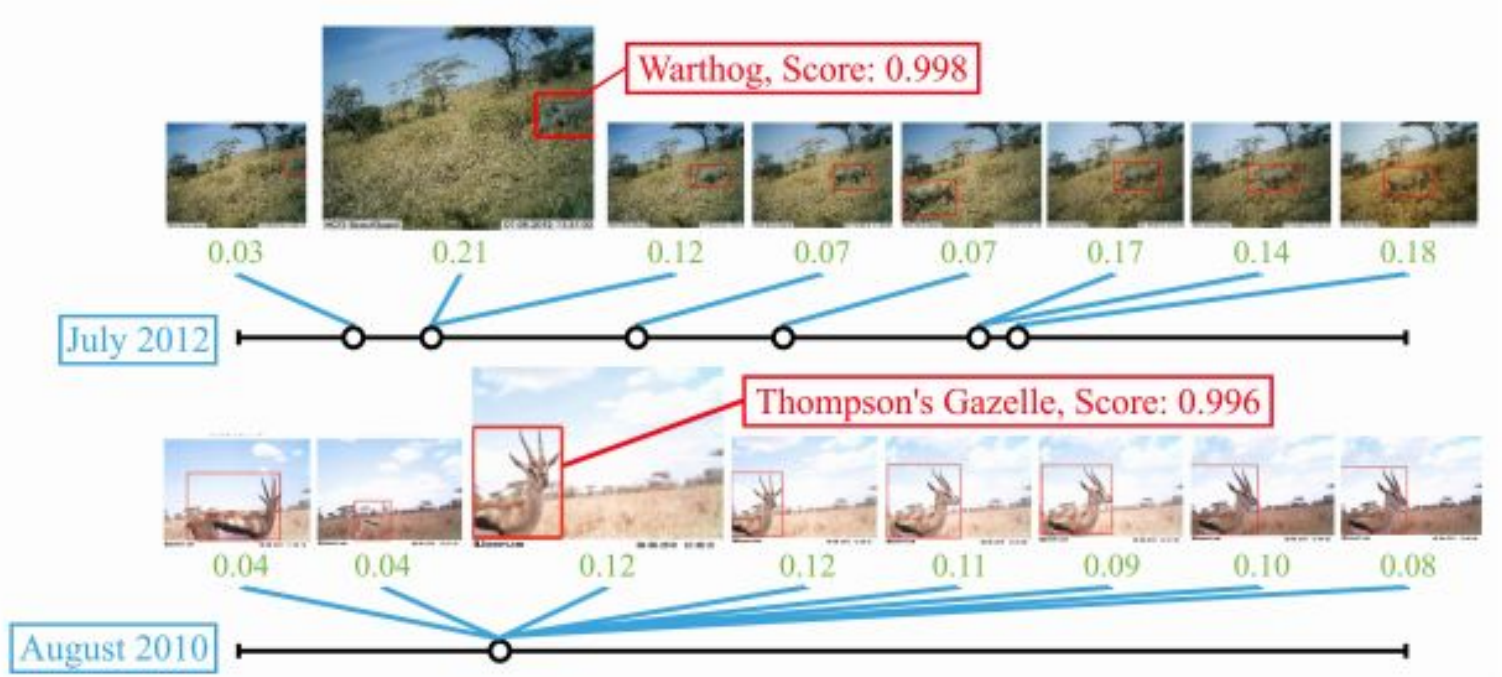
- hold features for all box proposals in memory.

- cropped instance-level features across a small window and globally pool across the spatial dimensions.

- A matrix of shape (# proposals per frame*# frames)×(feature depth) containing a single embedding vector per box proposal, that is then passed into the short term attention block.



Building a memory bank



동국대학교
dongguk university

Experiments

- Established object detection metrics: mAP at 0.5 IoU and Average Recall (AR)
- Results to a single-frame baseline for all three datasets
- In Snapshot Serengeti, investigating the effects of both short-term and long-term attention, the feature extractor, the long-term time horizon

동국대학교
dongguk university

Experiments

Main Results

Model	SS mAP	SS AR	CCT mAP	CCT AR	CC mAP	CC AR
Single Frame	37.9	46.5	56.8	53.8	38.1	28.2
Context R-CNN	**55.9**	**58.3**	**76.3**	**62.3**	**42.6**	**30.2**

(a) Results across datasets

SS	mAP	AR
Single Frame	37.9	46.5
Maj. Vote	37.8	46.4
ST Spatial	39.6	36.0
S3D	44.7	46.0
SF Attn	44.9	50.2
ST Attn	46.4	55.3
LT Attn	55.6	57.5
ST+LT Attn	**55.9**	**58.3**

(d) Comparison across models

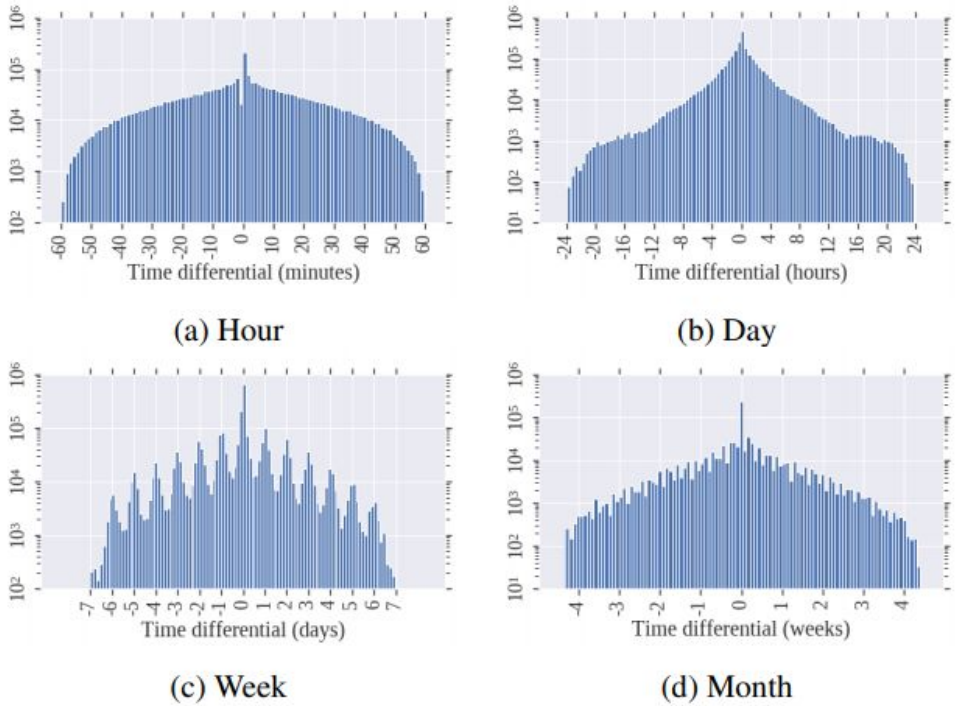
동국대학교
dongguk university

Experiments

Changing the Time Horizon

SS	mAP	AR
One minute	50.3	51.4
One hour	52.1	52.5
One day	52.5	52.9
One week	54.1	53.2
One month	**55.6**	**57.5**

(b) Time horizon



(a) Hour (b) Day

(c) Week (d) Month

동국대학교
dongguk university

Experiments

Contextual features for constructing M-long

SS	mAP	AR
One box per frame	**55.6**	**57.5**
COCO features	50.3	55.8
Only positive boxes	53.9	56.2
Subsample half	52.5	56.1
Subsample quarter	50.8	55.0

(c) Selecting memory

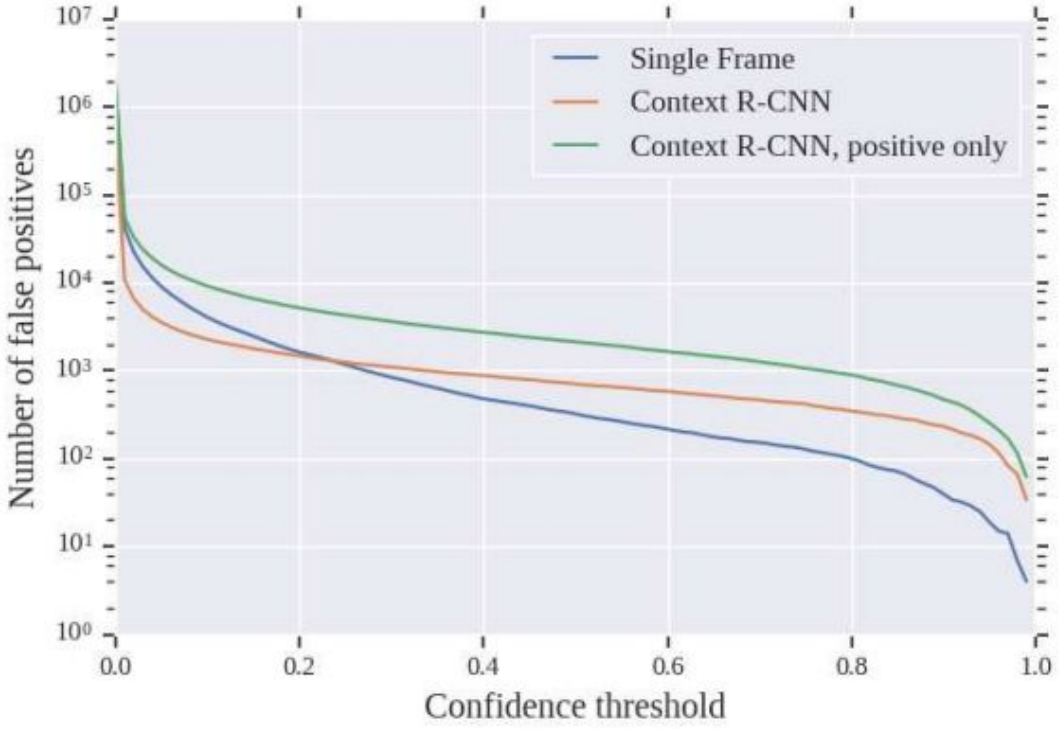
CC	mAP	AR
Single Frame	38.1	28.2
Top 1 Box	40.5	29.3
Top 8 Boxes	**42.6**	**30.2**

(e) Adding boxes to M^{long}

동국대학교 dongguk university

Experiments

Contextual features for constructing
M-long



Future Work

- A model that leverages percamera temporal context up to a month, and shows that in the static camera setting, attention-based temporal context is particularly beneficial.

- Context R-CNN, is general across static camera domains, improving detection performance over single-frame baselines on both camera trap and traffic camera data.

- Context R-CNN is adaptive and robust to passive-monitoring sampling strategies that provide data streams with low, irregular frame rates.

THANKS

Q&A

Thank you!