Into the Deep

An ecologist's journey into machine learning

Mathias Tobler
San Diego Zoo Global
Camera traps in ecology

- Inventories
- Activity patterns
- Density with capture-recapture
- Occupancy
- Resource use
From film to big data

2005
• 36 exposures per roll
• Film/batteries last 7-10 das
• 50 cameras, 60 days
• Area: 60 km²
• ~5000 images

2019
• >10,000 images/videos per SD card
• Battery life up to 9 months
• 250 cameras, 120 days
• Area: >2000 km²
• >250,000 images
Adoption of technology for research and conservation

New

Technology

incremental improvements

Data Analysis

Application / Study Design

Data Processing
Workflow pre AI

- **Raw data**
  - Download images from SD cards

- **Organize Data**
  - Organize image in folders by camera
  - Check and fix errors in EXIF data

- **Camera Base**
  - Import one folder at a time
  - Use EXIF data for Date/Time
  - Manually enter species information

- **Quality control**
  - Expert checks all images
  - Go through one species at a time

~ 1.5 Months
30,000 images,
10,000 subjects
Camera Base
Workflow with Zooniverse

Raw Images
- Download images from SD cards

Organize
- Organize in folders by camera
- Check and fix errors in EXIF data

Zooniverse
- Resize images
- Upload data to Zooniverse
- Wait for volunteers to process data
- Download and data

Camera Base
- Process Zooniverse results and import into Camera Base

Quality control
- Expert checks all images
- Go through one species at a time

~ 3 Months
30,000 images,
10,000 subjects
Data processing challenge

- Manual processing is slow (~2-300 images/hour for trained person)
- Availability of students/volunteers (often only 5-10 hour/week)
- Zooniverse – inconsistent processing rates
- Zooniverse – relatively high error rates ~10%
- Zooniverse – requires some level of engagement with volunteers

- Backlog of months to years
- Time from data collection to analysis is long
Recent publications

Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning
Mohammad Sadegh Norouzzadeh, Anh Nguyen, Margaret Kosmala, Alexandra Swanson, Meredith S. Palmer, Craig Packer, and Jeff Clune

Machine learning to classify animal species in camera trap images: Applications in ecology

Identifying animal species in camera trap images using deep learning and citizen science
Marco Willi, Ross T. Pitman, Anabelle W. Cardoso, Christina Locke, Alexandra Swanson, Amy Boyer, Marten Veldthuis, Lucy Fortson

FIGURE 1 Examples of camera trap images from the different projects used in this study. From left to right: Snapshot Serengeti, Camera Catalogue, Elephant Expedition, and Snapshot Wisconsin.
Workflow with AI

Raw Images
- Download images from SD cards

Organize
- Organize in folders by camera
- Check and fix errors in EXIF data

AI
- Read in organized images, use EXIF data
- Classify all images with machine learning

Camera Base
- Process classification results and import into Camera Base

Quality control
- Volunteers check images and fix misclassifications
- Expert checks problem
- Go through one species at a time

~ 5 hours
30,000 images, 10,000 subjects

~ 2 weeks
30,000 images, 10,000 subjects
Languages

- Computer science
- Data science
- Machine learning

- Quantitative Ecology
- Biology
- Statistics

- Field biology
- Resource management
- Government
Machine learning in R

- R package *keras*
- R package *tensorflow*
- R package *rTorch*

- Wrappers around python libraries
- All programming in native R
Ingredients

- Work
- NVIDIA
- LOTS
- Time
AI Training Workflow 1.0

1. Organized Images
2. Resize Images
3. Train species classifier in Keras (CNN)
AI Training Workflow 1.0

• Trained in Keras on Inception v3, ResNet 50, Xception, InceptionResNet, transfer learning from ImageNet
• 250,000 training images, including empty images
• Training accuracy on random validation subset ~95%
• Test accuracy on completely independent dataset ~75%
• Not much difference across models
• Classifier does not transfer well to new camera deployments
• Models seems to learn about static background in addition to species
MegaDetector

Speaking of models that might be useful to other people, we have trained a one-class animal detector trained on several hundred thousand bounding boxes from a variety of ecosystems. Lots more information – including download links – on the MegaDetector page.

Here’s a “teaser” image of what detector output looks like:

Image credit University of Washington.
MegaDetector
AI Training Workflow 2.0

- Organized Images
- Run MegaDetector
- Extract resized crops containing animals
- Train species classifier in Keras (CNN)
AI Training Workflow 2.0

- Trained in Keras on Inception v3, ResNet 50, Xception, InceptionResNet, transfer learning from ImageNet
- **124,000** training images with animals
- Training accuracy on independent validation set ~89%
- Test accuracy on independent dataset ~89%
- Not much difference across models
- Classifier does transfer much better to new camera trap survey
AI Classification Workflow 2.0

1. Organized Images
2. Extract EXIF data
3. Run MegaDetector
4. Extract crops containing animals
5. Run species classifier (ensemble model)
6. Process sequences
7. Rename images
8. Format results for import into Database
AI Classification Workflow 2.0

• Sequence classification improves accuracy by ~5%
• Ensemble models improve accuracy by ~3%
• Final accuracy Peru ~92%
• Final accuracy Snapshot Serengeti Season 1-3 ~ 94%
Quality control

- Camera Base
- Volunteers
- Check one species at a time
- Experts
- Still necessary for humans to review all images
Accuracy Zooniverse vs AI

- Small test dataset from Peru: 7500 images, 2500 subjects
- 39 species including Empty and Human
- Zooniverse: 10 classifications/subject
- AI Workflow 2.0
- Classification at subject level

Zooniverse: 91.9%
AI: 90.9%
## Accuracy Zooniverse vs AI

<table>
<thead>
<tr>
<th>Species</th>
<th>Subjects</th>
<th>Zooniverse</th>
<th>AI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brown agouti</td>
<td>342</td>
<td>0.97</td>
<td>0.99</td>
</tr>
<tr>
<td>Coati</td>
<td>20</td>
<td>0.60</td>
<td>0.40</td>
</tr>
<tr>
<td>Collared peccary</td>
<td>260</td>
<td>0.87</td>
<td>0.96</td>
</tr>
<tr>
<td>Empty</td>
<td>461</td>
<td>1.00</td>
<td>0.80</td>
</tr>
<tr>
<td>Giant anteater</td>
<td>98</td>
<td>0.97</td>
<td>0.97</td>
</tr>
<tr>
<td>Grey brocket deer</td>
<td>128</td>
<td>0.80</td>
<td>0.79</td>
</tr>
<tr>
<td>Long-nosed armadillo</td>
<td>66</td>
<td>0.94</td>
<td>0.76</td>
</tr>
<tr>
<td>Paca</td>
<td>262</td>
<td>0.89</td>
<td>0.97</td>
</tr>
<tr>
<td>Pale-winged trumpeter</td>
<td>80</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Razor-billed curassow</td>
<td>70</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Red brocket deer</td>
<td>109</td>
<td>0.72</td>
<td>0.94</td>
</tr>
<tr>
<td>Short-eared dog</td>
<td>41</td>
<td>0.98</td>
<td>1.00</td>
</tr>
<tr>
<td>Spix's guan</td>
<td>24</td>
<td>0.96</td>
<td>1.00</td>
</tr>
<tr>
<td>Tapir</td>
<td>239</td>
<td>0.95</td>
<td>0.98</td>
</tr>
<tr>
<td>White-lipped peccary</td>
<td>158</td>
<td>0.84</td>
<td>0.93</td>
</tr>
</tbody>
</table>

Species with >20 subjects
Accuracy Zooniverse vs AI
Accuracy thoughts

• Overall accuracy largely measures accuracy for few common species and ignores rare species.
• Accuracy vs training sample size, need >1000 images per species
• Poor image quality might prevent very high accuracy (for now)
• False positives much worse than false negatives (detection probability). Ecological models are good at dealing with false negatives (imperfect detections) but generally can’t handle false positives.
Lessons learned

• Dataset specific adaptations necessary
• No full black-box solutions for training
• Automated solutions for classifications possible
Next steps

- More training data
- Process images from paired cameras
- Finalize easy to use pipeline in R
- Include videos in pipeline
- Combine AI with humans
Open Questions

• Does class imbalance matter? How to best deal with it?
• Could we use images from the internet to augment poorly represented species?
• What is the minimum number of images needed to achieve a model with an accuracy of >95%?
• How bad are miss-classification in the training data?
• How to best deal with multiple species per image?
Thank you
Where’s the Bear (WTB)?
Using Advances in Cloud+Data Analytics to Help Conservation Science

Chandra Krintz
Dept. of Computer Science
UC Santa Barbara
ckrintz@ucsb.edu

Camera Trap Tech Symposium. Nov 8th, 2019
Cloud + Data Analytics: Revolutionizing Commerce

Math, Statistics, Machine Learning… (Code!)

What will you buy?
When will you buy it?
What will you pay?

Internet Activity ∨ Inference and Prediction

Cloud Computing

What Else Can We Revolutionize With It?
Wildlife Monitoring Using Camera Traps

- Alternative to labor-intensive observation & tracking
- Cost effective and scalable
- Safe and non-invasive
- Increasingly autonomous

Challenges:
- Vast amounts of data
- Inaccessible/remote areas
- Intermittent/no network connectivity
Connect Every Camera/Device to the Internet?

The “Internet of Things (IoT)”

Public Clouds

- Amazon Web Services
- Microsoft Azure
- Google Cloud Platform
Connect Every Camera/Device to the Internet?

Projecting the ‘Things’ Behind the Internet of Things

From 2014-2020, IoT grows at an annual compound rate of 23.1% CAGR

- Industrial and emerging ‘things’ increasingly captures a larger share of IoT market
- IoT primarily comprised of computers, mobile phones and tablet ‘things’

Sources: Gartner | Cisco | CompTIA
Data appliances
- Self managing, fault tolerant
- Hardened for hostile environments
- API-compatible with public clouds
- **Multiple** machine learning tools/models
**Data appliances**

- Self managing, fault tolerant
- Hardened for hostile environments
- API-compatible with public clouds
- **Multiple** machine learning tools/models
Deployment and Empirical Methodology
A New Kind of Computer Science Research

- Problem driven and empirical
- Societal and regional impact
- Multidisciplinary collaboration
- Repeatable, demonstrable, applied (tech-transfer ready)
- Engaging students & the community
Thanks!

- **Co-Lead:** Dr. Rich Wolski
- **Collaborators:** UCSB, LREC, CalPoly SLO, Fresno State, NCState, Powwow Energy, Sedgwick Reserve, Private Growers
- **Support:** Google, IBM Research, Microsoft Research, NSF, NIH, California Energy Commission

Students:

- Fatih Bakir
- Gareth George
- Nevena Golubovic
- Carly Larsson
- Wei-Tsung Lin
- Nazmus Saquib
- Michael Zhang

ckrintz@cs.ucsb.edu, rich@cs.ucsb.edu
http://www.cs.ucsb.edu/~ckrintz/racelab.html
DEEP LEARNING APPROACHES FOR ANIMAL RE-IDENTIFICATION

Stefan Schneider
University of Guelph
Ontario, Canada
About Me

- Undergrad at University of Waterloo
  - *Environmental Science & Business*

- Masters at University of Guelph
  - *Trophic Dynamics of Endangered Ecosystems*

- PhD at University of Guelph
  - *Deep Learning for Animal Re-Identification*
Animal Re-Identification

- Population Estimates
  - Diversity, Richness, Abundance
  - Larger overarching ecological interpretations of trophic interactions and population dynamics
  - Necessary for our understanding of sustainable practices and the protection of endangered species and ecosystems
  - Mark & Recapture
Methods for Animal Re-ID

- Tagging and Scarring

  - Advantages
    - Moderate-High reliability

  - Disadvantages
    - Expensive
    - Labourious
    - Invasive to the animal
Methods for Animal Re-ID

- Camera Trap / Video

  - Advantages
    - Lower Costs
    - Less Invasive

  - Disadvantages
    - Low-Moderate Reliability
    - Labourious to Analyze
    - Human Judgment Bias
Camera Trap Literature

- Karanth, 1995
  - First study using camera traps combine with a formal mark and recapture model

- 50% annual growth in publications using camera traps 1998-2015
  - Rowcliffe and Carbone, 2008; Burton et al., 2015

- Trolle, 2003 identified recognized a biased selection of study species
  - Spotted and striped felids

- Foster and Harmsen, 2011 criticized 47 studies for:
  - Leniencies for individual re-ID
  - Small sample sizes
Feature Engineering

- Robust History of all kinds of species
- Whale Sharks (Arzoumanian, 2005)
- Review (Schneider et al., 2018)
Deep Neural Networks

- Computational framework where a system's parameters are not hard-coded but optimized through training from large amounts of data.

- Multilayered function with modifiable weights capable of capturing logical relationships within data.

- With enough data and computation, the underlying relationship between the data and output can be mapped.
Convolutional Neural Network

- Introduce learnable filter maps which are capable of learning to capture meaningful spatial information within an image
  - Lines
  - Edges
  - Colours

- A simple CNN will have:
  - two or three convolution layers passed through non-linearity functions
  - two or three pooling layers
  - ending with fully-connected layers to return a classification output
Advances in Computer Vision

- Krizhevsky et al., 2012 demonstrated great success using CNNs for image classification
  - Conditioned returning only one classification for a given image

- Object Detection methods allow for objects within an image to be localized
  - Object Classification and position regression
    - YOLO – Redmond et al., 2015
    - SSD – Liu et al., 2016
    - Faster R-CNN – Ren et al., 2016
CRV 2018

- Schneider et al., 2018
  - Reconyx Camera Trap Dataset & Gold Standard Snapshot Serengeti
  - Faster RCNN vs VOLOv2
  - 4450 Bounding Box coordinates
    - Semi-Supervised Learning -> Learning from unlabelled data
Ecological Applications

- Object detection allows for the possible autonomous data collection of:
  - Numbers of individuals of *each species*
    - Sex
    - Age
    - Stance

- Which can be extrapolated to autonomously collect data related to:
  - Interspecies cohabitation
  - Species family dynamics
  - Travel direction
  - How these values change seasonally/annually
Object Detector - Octopus
Deep Learning for Animal Re-ID

- **Carter et al., 2014**
  - Ensemble of networks to re-ID green turtles in Australia

- **Freytag et al., 2016**
  - AlexNet on Chimpanzee faces

- **Brust et al., 2017**
  - YOLO to extract Gorilla faces followed by AlexNet for re-ID
Siamese Networks

- Limitation of traditional networks
  - Requires large amounts of data
  - Unrealistic for animal populations

- One-shot Learning
  - Siamese Network

- Bromley et al., 1993
Deep Learning for Animal Re-ID

- Deb et al., 2018
  - Siamese Network for Chimpanzee, Lemur and Golden Monkey re-ID
    - Verification
    - Closed set-identification
    - Open set-identification
Triplet Loss Networks

- Euclidean distance of embeddings
  - Minimize positive distance while maximizing negative

- FaceNet – Schroff et al., 2016  95% top-1 YTF 1,595 people
Data for Similarity Networks

- **Human Actors**
  - FaceScrub - 106,863 images of 530 male/females

- **Chimpanzee**
  - Chimpface - 5,599 images of 90 chimpanzees

- **Humpback Whale Fluke**
  - Humpback Whale Identification Challenge - 9,046 images of 4,251 humpback whales

- **Fruit Fly**
  - Jon Schneider & Graham Taylor - 244,760 images of 20 fruit flies

- **Siberian Tigers**
  - Amur Tiger Re-Identification in the Wild

- **Octopus**
  - Stefan Linquist - Octopolis - 5,192 images taken from video
Image Data
### Table 2: Summary of Performance Metrics for Siamese Similarity Learning Models by Species & Data Set

<table>
<thead>
<tr>
<th>Species</th>
<th>Model</th>
<th>mAP@1</th>
<th>mAP@5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td>AlexNet</td>
<td>0.699 ± 0.342</td>
<td>0.721 ± 0.332</td>
</tr>
<tr>
<td></td>
<td>VGG19</td>
<td>0.680 ± 0.288</td>
<td>0.703 ± 0.251</td>
</tr>
<tr>
<td></td>
<td>DenseNet201</td>
<td>0.734 ± 0.277</td>
<td>0.835 ± 0.875</td>
</tr>
<tr>
<td></td>
<td>ResNet152</td>
<td><strong>0.756 ± 0.282</strong></td>
<td><strong>0.856 ± 0.123</strong></td>
</tr>
<tr>
<td></td>
<td>InceptionV3</td>
<td>0.713 ± 0.227</td>
<td>0.854 ± 0.126</td>
</tr>
<tr>
<td>Chimpanzee</td>
<td>AlexNet</td>
<td>0.639 ± 0.221</td>
<td>0.863 ± 0.121</td>
</tr>
<tr>
<td></td>
<td>VGG19</td>
<td>0.645 ± 0.168</td>
<td>0.884 ± 0.094</td>
</tr>
<tr>
<td></td>
<td>DenseNet201</td>
<td>0.725 ± 0.134</td>
<td>0.871 ± 0.064</td>
</tr>
<tr>
<td></td>
<td>ResNet152</td>
<td><strong>0.775 ± 0.134</strong></td>
<td><strong>0.901 ± 0.097</strong></td>
</tr>
<tr>
<td></td>
<td>InceptionV3</td>
<td>0.743 ± 0.106</td>
<td>0.869 ± 0.125</td>
</tr>
<tr>
<td>Whale</td>
<td>AlexNet</td>
<td>0.509 ± 0.385</td>
<td>0.682 ± 0.334</td>
</tr>
<tr>
<td></td>
<td>VGG19</td>
<td>0.543 ± 0.397</td>
<td>0.689 ± 0.410</td>
</tr>
<tr>
<td></td>
<td>DenseNet201</td>
<td>0.521 ± 0.445</td>
<td>0.711 ± 0.312</td>
</tr>
<tr>
<td></td>
<td>ResNet152</td>
<td><strong>0.563 ± 0.202</strong></td>
<td><strong>0.757 ± 0.298</strong></td>
</tr>
<tr>
<td></td>
<td>InceptionV3</td>
<td><strong>0.576 ± 0.203</strong></td>
<td><strong>0.742 ± 0.390</strong></td>
</tr>
<tr>
<td>Fruit Fly</td>
<td>AlexNet</td>
<td>0.621 ± 0.078</td>
<td>0.875 ± 0.064</td>
</tr>
<tr>
<td></td>
<td>VGG19</td>
<td>0.590 ± 0.081</td>
<td>0.838 ± 0.090</td>
</tr>
<tr>
<td></td>
<td>DenseNet201</td>
<td>0.638 ± 0.153</td>
<td>0.843 ± 0.180</td>
</tr>
<tr>
<td></td>
<td>ResNet152</td>
<td><strong>0.693 ± 0.098</strong></td>
<td><strong>0.896 ± 0.109</strong></td>
</tr>
<tr>
<td></td>
<td>InceptionV3</td>
<td>0.522 ± 0.021</td>
<td>0.873 ± 0.143</td>
</tr>
<tr>
<td>Tiger</td>
<td>AlexNet</td>
<td>0.794 ± 0.396</td>
<td>0.858 ± 0.289</td>
</tr>
<tr>
<td></td>
<td>VGG19</td>
<td>0.735 ± 0.245</td>
<td>0.821 ± 0.243</td>
</tr>
<tr>
<td></td>
<td>DenseNet201</td>
<td><strong>0.803 ± 0.398</strong></td>
<td><strong>0.8756 ± 0.148</strong></td>
</tr>
<tr>
<td></td>
<td>ResNet152</td>
<td>0.789 ± 0.320</td>
<td><strong>0.877 ± 0.172</strong></td>
</tr>
<tr>
<td></td>
<td>InceptionV3</td>
<td>0.701 ± 0.307</td>
<td>0.843 ± 0.231</td>
</tr>
</tbody>
</table>

### Table 3: Summary of Performance Metrics for Triplet Loss Similarity Learning Models by Species & Data Set

<table>
<thead>
<tr>
<th>Species</th>
<th>Model</th>
<th>mAP@1</th>
<th>mAP@5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td>AlexNet</td>
<td>0.739 ± 0.284</td>
<td>0.804 ± 0.345</td>
</tr>
<tr>
<td></td>
<td>VGG19</td>
<td>0.811 ± 0.325</td>
<td>0.843 ± 0.251</td>
</tr>
<tr>
<td></td>
<td>DenseNet201</td>
<td><strong>0.914 ± 0.299</strong></td>
<td>0.947 ± 0.187</td>
</tr>
<tr>
<td></td>
<td>ResNet152</td>
<td>0.886 ± 0.301</td>
<td><strong>0.952 ± 0.093</strong></td>
</tr>
<tr>
<td></td>
<td>InceptionV3</td>
<td>0.903 ± 0.235</td>
<td>0.940 ± 0.124</td>
</tr>
<tr>
<td>Chimpanzee</td>
<td>AlexNet</td>
<td>0.739 ± 0.241</td>
<td>0.886 ± 0.166</td>
</tr>
<tr>
<td></td>
<td>VGG19</td>
<td>0.734 ± 0.188</td>
<td>0.890 ± 0.085</td>
</tr>
<tr>
<td></td>
<td>DenseNet201</td>
<td>0.792 ± 0.164</td>
<td>0.932 ± 0.049</td>
</tr>
<tr>
<td></td>
<td>ResNet152</td>
<td><strong>0.811 ± 0.155</strong></td>
<td><strong>0.961 ± 0.097</strong></td>
</tr>
<tr>
<td></td>
<td>InceptionV3</td>
<td>0.756 ± 0.136</td>
<td>0.940 ± 0.075</td>
</tr>
<tr>
<td>Whale</td>
<td>AlexNet</td>
<td>0.709 ± 0.374</td>
<td>0.782 ± 0.274</td>
</tr>
<tr>
<td></td>
<td>VGG19</td>
<td>0.743 ± 0.349</td>
<td>0.831 ± 0.287</td>
</tr>
<tr>
<td></td>
<td>DenseNet201</td>
<td>0.721 ± 0.304</td>
<td>0.801 ± 0.253</td>
</tr>
<tr>
<td></td>
<td>ResNet152</td>
<td>0.763 ± 0.252</td>
<td><strong>0.860 ± 0.275</strong></td>
</tr>
<tr>
<td></td>
<td>InceptionV3</td>
<td><strong>0.776 ± 0.243</strong></td>
<td>0.834 ± 0.290</td>
</tr>
<tr>
<td>Fruit Fly</td>
<td>AlexNet</td>
<td>0.671 ± 0.158</td>
<td>0.935 ± 0.041</td>
</tr>
<tr>
<td></td>
<td>VGG19</td>
<td>0.608 ± 0.161</td>
<td>0.954 ± 0.120</td>
</tr>
<tr>
<td></td>
<td>DenseNet201</td>
<td>0.660 ± 0.194</td>
<td>0.978 ± 0.084</td>
</tr>
<tr>
<td></td>
<td>ResNet152</td>
<td><strong>0.743 ± 0.163</strong></td>
<td><strong>0.986 ± 0.089</strong></td>
</tr>
<tr>
<td></td>
<td>InceptionV3</td>
<td>0.561 ± 0.125</td>
<td>0.967 ± 0.131</td>
</tr>
<tr>
<td>Tiger</td>
<td>AlexNet</td>
<td>0.830 ± 0.296</td>
<td>0.978 ± 0.217</td>
</tr>
<tr>
<td></td>
<td>VGG19</td>
<td>0.770 ± 0.205</td>
<td>0.940 ± 0.145</td>
</tr>
<tr>
<td></td>
<td>DenseNet201</td>
<td><strong>0.863 ± 0.193</strong></td>
<td>0.974 ± 0.148</td>
</tr>
<tr>
<td></td>
<td>ResNet152</td>
<td>0.811 ± 0.124</td>
<td><strong>0.996 ± 0.072</strong></td>
</tr>
<tr>
<td></td>
<td>InceptionV3</td>
<td>0.731 ± 0.117</td>
<td>0.933 ± 0.121</td>
</tr>
</tbody>
</table>
Results
Real World Applications - Parks Canada

- Parks Canada
  - University of Calgary - Saul Greenberg
    - Timelapse
      - AI systems prior to Microsoft AI
  
- 89,600 images of 55 Classifications from 36 locations

- Opportunity to practice training advanced networks
Parks Canada - Data
Classification Imbalance
Parks Canada - Results

- **Ratio Selection Technique**

- **Background Locations**
  - 500 Images
    - 80% Accuracy
  - 750 Images
    - 90% Accuracy

### TABLE I
**Summary of Performance Metrics for Parks Canada Species ID with Trained Locations**

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy</th>
<th>F1 Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>DenseNet201</td>
<td>0.956</td>
<td>0.794</td>
</tr>
<tr>
<td>Inception ResNetV2</td>
<td>0.929</td>
<td>0.724</td>
</tr>
<tr>
<td>InceptionV3</td>
<td>0.940</td>
<td>0.756</td>
</tr>
<tr>
<td>NASNetMobile</td>
<td>0.910</td>
<td>0.714</td>
</tr>
<tr>
<td>MobileNetV2</td>
<td>0.931</td>
<td>0.754</td>
</tr>
<tr>
<td>Xception</td>
<td>0.954</td>
<td>0.786</td>
</tr>
<tr>
<td>Ensemble</td>
<td>0.959</td>
<td>0.812</td>
</tr>
</tbody>
</table>

### TABLE II
**Summary of Performance Metrics for Parks Canada Species ID with Untrained Locations**

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy</th>
<th>F1 Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>DenseNet201</td>
<td>0.687 ± 0.057</td>
<td>0.698 ± 0.031</td>
</tr>
<tr>
<td>Inception ResNetV2</td>
<td>0.635 ± 0.049</td>
<td>0.654 ± 0.036</td>
</tr>
<tr>
<td>InceptionV3</td>
<td>0.651 ± 0.057</td>
<td>0.655 ± 0.029</td>
</tr>
<tr>
<td>NASNetMobile</td>
<td>0.637 ± 0.068</td>
<td>0.678 ± 0.034</td>
</tr>
<tr>
<td>MobileNetV2</td>
<td>0.643 ± 0.050</td>
<td>0.653 ± 0.030</td>
</tr>
<tr>
<td>Xception</td>
<td>0.685 ± 0.062</td>
<td>0.646 ± 0.034</td>
</tr>
<tr>
<td>Ensemble</td>
<td>0.712</td>
<td>0.708</td>
</tr>
</tbody>
</table>
Operation Wallacea

- "Operation Wallacea is a network of academics from European and North American universities, who design and implement biodiversity and conservation management research expeditions."

- Trained a Faster R-CNN network using TensorFlow API and ResNet-50

- Data
  - Hand Annotated
  - 87% Accuracy
  - IOU ~0.5
Amazon River

- Fish & Dolphin Abundance Relative to Areas of Deforestation
- Sonar Images from the back of a trolling boat
  - +/- 2.1 Fish and +/- 0.1 Dolphin
  - Dr. Bodmer
WACV2020

https://sites.google.com/view/wacv2020animalreid/WACV2020

March 1st, 2020

Snowmass Village, Colorado
Questions!

Thank You!
## Literature Review Results

<table>
<thead>
<tr>
<th>Animal</th>
<th>Year</th>
<th>Methodology</th>
<th>Test Size</th>
<th>Num. Classes</th>
<th>Top-1 Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sperm Whale</td>
<td>1990</td>
<td>Database similarity</td>
<td>56</td>
<td>1,015</td>
<td>59</td>
</tr>
<tr>
<td>Humpback Whale</td>
<td>1990</td>
<td>Database similarity</td>
<td>30</td>
<td>790</td>
<td>41.4</td>
</tr>
<tr>
<td>Grey Seal</td>
<td>1990</td>
<td>3-D Pattern Cell similarity</td>
<td>58</td>
<td>58</td>
<td>98.0</td>
</tr>
<tr>
<td>Sperm Whale</td>
<td>1998</td>
<td>Wavelet transformations</td>
<td>56</td>
<td>8</td>
<td>92.0</td>
</tr>
<tr>
<td>Cheetah</td>
<td>2001</td>
<td>3-D Pattern Cell similarity</td>
<td>1,000</td>
<td>NA</td>
<td>97.5</td>
</tr>
<tr>
<td>Whale/Dolphin</td>
<td>2003</td>
<td>XY Pair Euclidean Distance</td>
<td>52</td>
<td>36</td>
<td>50.0</td>
</tr>
<tr>
<td>Marbled Salamander</td>
<td>2004</td>
<td>Pixel histogram and local colours</td>
<td>69</td>
<td>NA</td>
<td>72.0</td>
</tr>
<tr>
<td>Whale Shark</td>
<td>2005</td>
<td>Star pattern recognition</td>
<td>27</td>
<td>NA</td>
<td>90.0</td>
</tr>
<tr>
<td>Elephant</td>
<td>2007</td>
<td>Polynomial multi-curve matching</td>
<td>332</td>
<td>268</td>
<td>75.0</td>
</tr>
<tr>
<td>African penguin</td>
<td>2009</td>
<td>Per feature AdaBoost classifier</td>
<td>N/A</td>
<td>NA</td>
<td>92-97.0</td>
</tr>
<tr>
<td>Tiger</td>
<td>2009</td>
<td>3-D Pattern Cell similarity</td>
<td>298</td>
<td>298</td>
<td>95.0</td>
</tr>
<tr>
<td>Manta Ray</td>
<td>2013</td>
<td>SIFT</td>
<td>720</td>
<td>265</td>
<td>51.0</td>
</tr>
<tr>
<td>Chimpanzee (C-Zoo)</td>
<td>2013</td>
<td>Support Vector Machine</td>
<td>478</td>
<td>120</td>
<td>84.0</td>
</tr>
<tr>
<td>Chimpanzee (C-Tai)</td>
<td>2013</td>
<td>Support Vector Machine</td>
<td>1,146</td>
<td>286</td>
<td>68.8</td>
</tr>
<tr>
<td>Green Turtle</td>
<td>2014</td>
<td>Feedforward Network</td>
<td>180</td>
<td>72</td>
<td>95.0</td>
</tr>
<tr>
<td>Chimpanzee (C-Zoo)</td>
<td>2016</td>
<td>Convolutional Network</td>
<td>478</td>
<td>120</td>
<td>92.0</td>
</tr>
<tr>
<td>Chimpanzee (C-Tai)</td>
<td>2016</td>
<td>Convolutional Network</td>
<td>1,146</td>
<td>286</td>
<td>75.7</td>
</tr>
<tr>
<td>Shark</td>
<td>2017</td>
<td>Naive Bayes Nearest Neighbour</td>
<td>2,456</td>
<td>85</td>
<td>82.0</td>
</tr>
<tr>
<td>Gorilla</td>
<td>2017</td>
<td>Convolutional Network</td>
<td>500</td>
<td>482</td>
<td>90.8</td>
</tr>
<tr>
<td>Elephant</td>
<td>2018</td>
<td>Support Vector Machine</td>
<td>2,078</td>
<td>276</td>
<td>59.0</td>
</tr>
<tr>
<td>Chimpanzee</td>
<td>2018</td>
<td>Siamese Network</td>
<td>5,599</td>
<td>90</td>
<td>59.9</td>
</tr>
<tr>
<td>Lemur</td>
<td>2018</td>
<td>Siamese Network</td>
<td>3,000</td>
<td>129</td>
<td>83.1</td>
</tr>
<tr>
<td>Golden Monkey</td>
<td>2018</td>
<td>Siamese Network</td>
<td>241 videos</td>
<td>49</td>
<td>78.7</td>
</tr>
</tbody>
</table>
Fine-grained species classification using unified data

Sayali Kulkarni
Research & Machine Intelligence
Google

Tanya Birch (Google)
Tomer Gadot (Google)
Ştefan Istrate (Google)
Chen Luo (Google)

Sara Beery (CalTech)
Eric Fegraus (CI)
Dave Thau (WWF)
Biodiversity is declining around the world

**Living Planet Index**
Population Abundance data shows 60% wildlife population decline by 2020

**Biodiversity Intactness Index**
Monitors changes in species composition

**IUCN Red List Index**
Monitors extinction risk across 105k species

**Species Habitat Index**
Measures change in species distribution

**Wildlife Picture Index**
Measures population abundance and more from camera trap data
Sensors to monitor biodiversity

SPACE

AERIAL

GROUND

SEA

http://dylanbrowndesigns.com/tutorials/creating-a-3d-cross-section-using-google-earth-and-photoshop
Camera trap data

- Data collected from camera traps setup in multiple remote sites
- Data is collected over decades
- $O(10M)$ images; $O(100)$ labels - labeled images from domain experts
https://www.wildlifeinsights.org

- Platform for analytics over wildlife data to better manage wildlife populations
- Experts can upload images from camera traps to automatically identify species using AI models
- Frees up domain experts time and enables them to focus actual conservation effort
Data
Combining multiple large datasets

- 20M Images
- 700-1000+ Classes (Species)

Licensing for incoming Wildlife Insights images CC0, CC-BY-4.0, CC-BY-NC

Data Partners:

- ZSL
- Conservation International
- Map of Life
- WCS
- Smithsonian Institution
- WWF
## Partner data ingested so far

<table>
<thead>
<tr>
<th>Dataset Name</th>
<th>Total images</th>
<th>Species</th>
<th>Geography</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>TEAM network</td>
<td>3.02M</td>
<td>490</td>
<td>Pantropical</td>
<td>Conservation International, TEAM partners</td>
</tr>
<tr>
<td>OneTam</td>
<td>741K</td>
<td>69</td>
<td>SF Bay area</td>
<td></td>
</tr>
<tr>
<td>WWF</td>
<td>118.3K</td>
<td>70</td>
<td>Peru, British Columbia, Spain, Pakistan</td>
<td>World Wildlife Fund</td>
</tr>
<tr>
<td>WCS</td>
<td>87.5K</td>
<td>26</td>
<td></td>
<td>Wildlife Conservation Society</td>
</tr>
<tr>
<td>Snapshot Serengeti</td>
<td>3.01M</td>
<td>46</td>
<td>Sub-saharan Africa</td>
<td>LiLa Science</td>
</tr>
<tr>
<td>North American Camera Trap Images</td>
<td>1.35M</td>
<td>50</td>
<td>five locations across the United States</td>
<td></td>
</tr>
<tr>
<td>CalTech Camera Traps</td>
<td>202.5K</td>
<td>20</td>
<td>Southwestern US</td>
<td></td>
</tr>
</tbody>
</table>
### WI Computer Vision Standard

Based on the Open Standards for Camera Trap Data (Forrester et al. 2016)

<table>
<thead>
<tr>
<th>Name</th>
<th>Required</th>
<th>Example value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class</td>
<td>Yes</td>
<td>Mammalia</td>
</tr>
<tr>
<td>Order</td>
<td>Yes</td>
<td>Cetartiodactyla</td>
</tr>
<tr>
<td>Family</td>
<td>Yes</td>
<td>Cervidae</td>
</tr>
<tr>
<td>Genus</td>
<td>Yes</td>
<td>Muntiacus</td>
</tr>
<tr>
<td>Species</td>
<td>Yes</td>
<td>muntjak</td>
</tr>
<tr>
<td>Date Time</td>
<td>Ideal</td>
<td>2018-10-09 10:28:14</td>
</tr>
<tr>
<td>Latitude</td>
<td>Ideal</td>
<td>-5.694427013</td>
</tr>
<tr>
<td>Longitude</td>
<td>Ideal</td>
<td>104.4210052</td>
</tr>
<tr>
<td>Burst_index</td>
<td>No</td>
<td>1</td>
</tr>
<tr>
<td>Burst_length</td>
<td>No</td>
<td>12</td>
</tr>
</tbody>
</table>

Hierarchical class labels standardized as Class > Order > Family > Genus > Species

- **Temporal information**
- **Spatial information**
- **Sequence information**

Model and evaluation
Per-frame Multi-class classifier

- Multiclass classification model over pretrained from Inception v4
- Data split for evaluation (90% Train, 10% Test):
  - Grids of 100sq.m. hashed either into train or test sets
    - reduces overfitting based on the background
Fine Grained Classification

Accuracy across all labels: 75.07%

<table>
<thead>
<tr>
<th>Class</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>mule deer</td>
<td>84.74%</td>
</tr>
<tr>
<td>southern pig-tailed macaque</td>
<td>88.93%</td>
</tr>
<tr>
<td>blue duiker</td>
<td>73.91%</td>
</tr>
<tr>
<td>central american agouti</td>
<td>78.37%</td>
</tr>
<tr>
<td>wild boar</td>
<td>84.28%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Class</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>bartlett's tinamou</td>
<td>74.51%</td>
</tr>
<tr>
<td>olive baboon</td>
<td>68.63%</td>
</tr>
<tr>
<td>banded civet</td>
<td>53.85%</td>
</tr>
<tr>
<td>north american porcupine</td>
<td>72.22%</td>
</tr>
<tr>
<td>handsome francolin</td>
<td>55.56%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Class</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>sun bear</td>
<td>41.07%</td>
</tr>
<tr>
<td>humboldt's white-fronted capuchin</td>
<td>42.11%</td>
</tr>
<tr>
<td>starred wood-quail</td>
<td>28.07%</td>
</tr>
<tr>
<td>red-fronted brown lemur</td>
<td>20.29%</td>
</tr>
<tr>
<td>marsh mongoose</td>
<td>15.28%</td>
</tr>
</tbody>
</table>

Common classes:

Rare classes (less than 100 training examples):
Precision - Recall tradeoff

Recall is important for blanks

Precision is important for species
Attention to input

Objects are correctly realized in making the predictions
Attention to input

Sometimes the model looks for clues all over the place, leading to errors or overfitting due to background features.
Work in progress
Fine-tuning with spatio-temporal slices

- **Bias on spotting the species by**
  - Location - example: presence in specific habitat
  - Time - example: migration patterns

- **Improving the current models that are purely based on image features using spatial and temporal information**

- **Quick hack:**
  - Adding hard filters based at inference time

- **Intermediate:**
  - Use a spatial/temporal prior at inference time, if available

- **Ideal**
  - Training the model with spatio-temporal features and expecting these to be available at inference time
Multi-task learning

- Classification at frame-level limits predictions of a single class

Predict and evaluate images with more than one animal

Images with multiple species

Training an object detector and classifier: jointly or chained
Intra-sequence attention

- Sequence bursts in camera traps where bursts are within few seconds of each other
- Significant redundancy among consecutive images
- Using the differentials across images for tracking and identifying the object of interest more accurately
Thank you!  
Looking forward to collaborations...

Pic credit: https://today.tamu.edu/
Long-term context for species detection

Sara Beery, Jonathan Huang, Vivek Rathod
Google Visual Dynamics Team
Wildlife Insights Team
The cameras are static, and animals are habitual.

We want to leverage long-term temporal context in detectors to learn to:
The cameras are static, and animals are habitual

We want to leverage long-term temporal context in detectors to learn to:
1. Ignore salient false positives

These rocks have not moved in a month, they’re probably not animals.
The cameras are static, and animals are habitual

We want to leverage long-term temporal context in detectors to learn to:
1. Ignore salient false positives
2. Improve per-location species classification, since each location has its own imbalanced species distribution

These are probably the same species, and if we’re confident about one, that should help us classify the other
Snapshot Serengeti: large-scale public camera trap dataset

6 seasons, 225 camera locations, 48 classes (long-tailed distribution), ~2M images with image-level labels, ~70% of the images are empty (false triggers), ~80K with ground-truth bboxes, all other bboxes come from the megadetector.
What to “remember”

- Use a static feature extractor
- Most salient object from each image
- Pool box classifier features from Faster-RCNN spatially
- Encode and store position in space and time
How to “remember”

- Build a memory bank for each camera location -> matrix of pooled box classifier features
- For each box proposal in current frame (keyframe), build a global context feature via non-local self-attention
- Add this context feature back to the box classifier feature and pass to second-stage predictor

\[
W = \text{softmax} \left( \frac{f(X)g(M)}{T\sqrt{d}} \right)
\]

\[
C = W^T h(M)
\]
Short-term attention

Add context to features for each proposal

Global Memory

Box predictions

Class predictions

Loss calculated from key frame GT
## Results

**16.9% mAP improvement from the single-frame baseline**

<table>
<thead>
<tr>
<th>Model</th>
<th>mAP</th>
<th>mAP@0.5</th>
<th>AR@1</th>
<th>AR@100</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single-frame Faster RCNN ResNet101</td>
<td>28.6%</td>
<td>39.4%</td>
<td>48.3%</td>
<td>54.8%</td>
</tr>
<tr>
<td>LSTD</td>
<td>23.3%</td>
<td>33.6%</td>
<td>46.0%</td>
<td>51.9%</td>
</tr>
<tr>
<td>S3D</td>
<td>30.9%</td>
<td>44.7%</td>
<td>49.5%</td>
<td>57.2%</td>
</tr>
<tr>
<td>Spatial average pooling</td>
<td>15.0%</td>
<td>31.6%</td>
<td>27.9%</td>
<td>40.6%</td>
</tr>
<tr>
<td>Temporally weighted spatial average pooling</td>
<td>26.9%</td>
<td>47.9%</td>
<td>34.2%</td>
<td>45.1%</td>
</tr>
<tr>
<td>Boxwise attention, T=1</td>
<td>38.6%</td>
<td>52.2%</td>
<td>49.5%</td>
<td>59.1%</td>
</tr>
<tr>
<td>Boxwise attention, T=.01</td>
<td>40.8%</td>
<td>53.7%</td>
<td>51.6%</td>
<td>60.9%</td>
</tr>
<tr>
<td>Framewise attention from keyframe memory features, T=.01</td>
<td>39.2%</td>
<td>53.2%</td>
<td>49.5%</td>
<td>61.7%</td>
</tr>
<tr>
<td>Boxwise attention from current box features, T=1</td>
<td>40.7%</td>
<td>53.6%</td>
<td>53.4%</td>
<td>61.5%</td>
</tr>
<tr>
<td>Boxwise attention from current box features, T=.01</td>
<td><strong>45.5%</strong></td>
<td><strong>60.4%</strong></td>
<td>51.3%</td>
<td>61.2%</td>
</tr>
</tbody>
</table>

**Single Frame Only**

- Long-term context
- Short-term context

Google Confidential + Proprietary
## Results

14.6% mAP improvement from S3D baseline

<table>
<thead>
<tr>
<th>Model</th>
<th>mAP</th>
<th>mAP@0.5</th>
<th>AR@1</th>
<th>AR@100</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single-frame Faster RCNN ResNet101</td>
<td>28.6%</td>
<td>39.4%</td>
<td>48.3%</td>
<td>54.8%</td>
</tr>
<tr>
<td>LSTD</td>
<td>23.3%</td>
<td>33.6%</td>
<td>46.0%</td>
<td>51.9%</td>
</tr>
<tr>
<td>S3D</td>
<td>30.9%</td>
<td>44.7%</td>
<td>49.5%</td>
<td>57.2%</td>
</tr>
<tr>
<td>Spatial average pooling</td>
<td>15.0%</td>
<td>31.6%</td>
<td>27.9%</td>
<td>40.6%</td>
</tr>
<tr>
<td>Temporally weighted spatial average pooling</td>
<td>26.9%</td>
<td>47.9%</td>
<td>34.2%</td>
<td>45.1%</td>
</tr>
<tr>
<td>Boxwise attention, T=1</td>
<td>38.6%</td>
<td>52.2%</td>
<td>49.5%</td>
<td>59.1%</td>
</tr>
<tr>
<td>Boxwise attention, T=.01</td>
<td>40.8%</td>
<td>53.7%</td>
<td>51.6%</td>
<td>60.9%</td>
</tr>
<tr>
<td>Framewise attention from keyframe memory features, T=.01</td>
<td>39.2%</td>
<td>53.2%</td>
<td>49.5%</td>
<td>61.7%</td>
</tr>
<tr>
<td>Boxwise attention from current box features, T=1</td>
<td>40.7%</td>
<td>53.6%</td>
<td>53.4%</td>
<td>61.5%</td>
</tr>
<tr>
<td>Boxwise attention from current box features, T=.01</td>
<td>45.5%</td>
<td>60.4%</td>
<td>51.3%</td>
<td>61.2%</td>
</tr>
</tbody>
</table>
## Results

9.9% improvement from S3D baseline using only short-term attention

<table>
<thead>
<tr>
<th>Model</th>
<th>mAP</th>
<th>mAP@0.5</th>
<th>AR@1</th>
<th>AR@100</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single-frame Faster RCNN ResNet101</td>
<td>28.6%</td>
<td>39.4%</td>
<td>48.3%</td>
<td>54.8%</td>
</tr>
<tr>
<td>LSTD</td>
<td>23.3%</td>
<td>33.6%</td>
<td>46.0%</td>
<td>51.9%</td>
</tr>
<tr>
<td>S3D</td>
<td>30.9%</td>
<td>44.7%</td>
<td>49.5%</td>
<td>57.2%</td>
</tr>
<tr>
<td>Spatial average pooling</td>
<td>15.0%</td>
<td>31.6%</td>
<td>27.9%</td>
<td>40.6%</td>
</tr>
<tr>
<td>Temporally weighted spatial average pooling</td>
<td>26.9%</td>
<td>47.9%</td>
<td>34.2%</td>
<td>45.1%</td>
</tr>
<tr>
<td>Boxwise attention, T=1</td>
<td>38.6%</td>
<td>52.2%</td>
<td>49.5%</td>
<td>59.1%</td>
</tr>
<tr>
<td>Boxwise attention, T=.01</td>
<td>40.8%</td>
<td>53.7%</td>
<td>51.6%</td>
<td>60.9%</td>
</tr>
<tr>
<td>Framewise attention from keyframe memory features, T=.01</td>
<td>39.2%</td>
<td>53.2%</td>
<td>49.5%</td>
<td>61.7%</td>
</tr>
<tr>
<td>Boxwise attention from current box features, T=1</td>
<td>40.7%</td>
<td>53.6%</td>
<td>53.4%</td>
<td>61.5%</td>
</tr>
<tr>
<td>Boxwise attention from current box features, T=.01</td>
<td>45.5%</td>
<td>60.4%</td>
<td>51.3%</td>
<td>61.2%</td>
</tr>
</tbody>
</table>
### Results

**4.7% improvement when adding long-term context**

<table>
<thead>
<tr>
<th>Model</th>
<th>mAP</th>
<th>mAP@0.5</th>
<th>AR@1</th>
<th>AR@100</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single-frame Faster RCNN ResNet101</td>
<td>28.6%</td>
<td>39.4%</td>
<td>48.3%</td>
<td>54.8%</td>
</tr>
<tr>
<td>LSTD</td>
<td>23.3%</td>
<td>33.6%</td>
<td>46.0%</td>
<td>51.9%</td>
</tr>
<tr>
<td>S3D</td>
<td>30.9%</td>
<td>44.7%</td>
<td>49.5%</td>
<td>57.2%</td>
</tr>
<tr>
<td>Spatial average pooling</td>
<td>15.0%</td>
<td>31.6%</td>
<td>27.9%</td>
<td>40.6%</td>
</tr>
<tr>
<td>Temporally weighted spatial average pooling</td>
<td>26.9%</td>
<td>47.9%</td>
<td>34.2%</td>
<td>45.1%</td>
</tr>
<tr>
<td>Boxwise attention, T=1</td>
<td>38.6%</td>
<td>52.2%</td>
<td>49.5%</td>
<td>59.1%</td>
</tr>
<tr>
<td>Boxwise attention, T=.01</td>
<td>40.8%</td>
<td>53.7%</td>
<td>51.6%</td>
<td>60.9%</td>
</tr>
<tr>
<td>Framewise attention from keyframe memory features, T=.01</td>
<td>39.2%</td>
<td>53.2%</td>
<td>49.5%</td>
<td>61.7%</td>
</tr>
<tr>
<td>Boxwise attention from current box features, T=1</td>
<td>40.7%</td>
<td>53.6%</td>
<td>53.4%</td>
<td>61.5%</td>
</tr>
<tr>
<td>Boxwise attention from current box features, T=.01</td>
<td>45.5%</td>
<td>60.4%</td>
<td>51.3%</td>
<td>61.2%</td>
</tr>
</tbody>
</table>
## Results

Decrease in softmax temperature improves mAP by 5%, but incurs 2% drop in AR

<table>
<thead>
<tr>
<th>Model</th>
<th>mAP</th>
<th>mAP@0.5</th>
<th>AR@1</th>
<th>AR@100</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single-frame Faster RCNN ResNet101</td>
<td>28.6%</td>
<td>39.4%</td>
<td>48.3%</td>
<td>54.8%</td>
</tr>
<tr>
<td>LSTD</td>
<td>23.3%</td>
<td>33.6%</td>
<td>46.0%</td>
<td>51.9%</td>
</tr>
<tr>
<td>S3D</td>
<td>30.9%</td>
<td>44.7%</td>
<td>49.5%</td>
<td>57.2%</td>
</tr>
<tr>
<td>Spatial average pooling</td>
<td>15.0%</td>
<td>31.6%</td>
<td>27.9%</td>
<td>40.6%</td>
</tr>
<tr>
<td>Temporally weighted spatial average pooling</td>
<td>26.9%</td>
<td>47.9%</td>
<td>34.2%</td>
<td>45.1%</td>
</tr>
<tr>
<td>Boxwise attention, T=1</td>
<td>38.6%</td>
<td>52.2%</td>
<td>49.5%</td>
<td>59.1%</td>
</tr>
<tr>
<td>Boxwise attention, T=.01</td>
<td>40.8%</td>
<td>53.7%</td>
<td>51.6%</td>
<td>60.9%</td>
</tr>
<tr>
<td>Framewise attention from keyframe memory features, T=.01</td>
<td>39.2%</td>
<td>53.2%</td>
<td>49.5%</td>
<td>61.7%</td>
</tr>
<tr>
<td>Boxwise attention from current box features, T=1</td>
<td>40.7%</td>
<td>53.6%</td>
<td>53.4%</td>
<td>61.5%</td>
</tr>
<tr>
<td>Boxwise attention from current box features, T=.01</td>
<td>45.5%</td>
<td>60.4%</td>
<td>51.3%</td>
<td>61.2%</td>
</tr>
</tbody>
</table>
When an animal hangs out in front of the camera for a long period of time, we leverage all of that context (not restricted to a short time window).
When there are only a few images of a class to compare to (porcupines are rare at this location), the model is able to combine information across the available frames.
In the extreme case where there is only a single relevant image, the model is able to learn to focus on the information that is relevant. It is adaptable to variable, sparse sampling.
In this case, our model was able to detect a wildebeest so difficult to see that our gold-standard annotators missed it.
In empty images, the model pays attention to the same salient position over time, and we see detection confidence decrease from the single-frame model.
Another example of the model attending to the same place in the frame.
“Rock” is not a labeled class, but our model has learned to pay attention to other examples of rocks, and learned that rocks are a salient background class.
Similarly, “bush” is not a labeled class, but our model has learned to pay attention to other examples of bushes and classify them as background.
Thanks!

Microsoft

AI for Earth

LILA BC
Labeled Information Library of Alexandria: Biology and Conservation

USGS
science for a changing world
If you would like to stay connected: Email
aiforconservation@gmail.com
to be invited to the
AI for Conservation Slack
(there’s a #cameratraps channel)
WILDBOOK
ANIMAL IDENTIFICATION FROM CAMERA TRAPS

WILD ME

Jason Parham
Senior Computer Vision Research Engineer, Ph.D. Candidate at RPI

November 8th, 2019  |  Camera Trap Tech Symposium
What is Wildbook?

• **Problem:** No OOTB Wildlife CMS or plugin CV system
• **Solution:** Wildbook - an web-based, open-source framework for wildlife data management and algorithms

  • rapidly share research techniques, tools, and investment across groups and species (encourage collaboration)
  • filter, search, export images and ecological metadata
  • decentralize conservation action by engaging citizen scientists

INTRODUCTION
Why is Wildbook Important?

• For biologists, data scientists, and statisticians, Wildbook provides:
  • a collaborative data management and analysis platform
  • A portal to engage the public in their research
  • a 20-year leap forward in IT technology
  • a way to quickly get access to new CV and AI research in a familiar, usable interface focused on their field

• For computer scientists, Wildbook provides:
  • a standard way to get wildlife data for CV research
  • a place to deploy successful research to biologists including standard APIs for integration

• For citizen scientists, Wildbook provides:
  • A way to directly contribute to wildlife research and conservation
The Power of a Name

- Individual identity of animals is the cornerstone of capture-mark-recapture (CMR)
- Related metadata: where, when, with whom?
- Capture-mark-recapture is a powerful technique supporting:
  - population dynamics
  - molecular ecology
  - animal biometrics
  - toxicology
  - social ecology studies
  - and more...
Motivation
Motivation

INTRODUCTION
Detection Pipeline Overview

**Raw Imagery from Citizen Scientists**

**Input Image**
- 8% - Masai Giraffe
- 4% - Reticulated Giraffe
- 1% - Sea Turtle
- 0% - Whale Fluke
- 45% - Grevy's Zebra
- 97% - Plains Zebra

**Aoi Classification**

**Image Classification**

**Annotation Localization**

**Annotation Classification**

**Background Segmentation**

**out**

Labeled Sightings Ready for ID Pipeline
Whole-Image Classification (WIC)

Trained to predict a multi-label, multi-class vector

High-pass filter to prevent irrelevant images from being processed

DETECTION PIPELINE
WIC for Camera Trap Data

Positive

Negative

Data provided by Dr. Megan McSherry, postdoc. at Princeton University
WIC ROC Curves on Camera Trap Data

DETECTION PIPELINE

Implicit Split (0%)
Detection Pipeline Example

Input Image

Grevy's: 70.60%  Plains: 83.31%

Localizations
Detection Pipeline Example
Identification Matches - HotSpotter

Work by Dr. Jonathan Crall
Identification Matches - CurvRank

Work by Hendrik Weideman
Camera Trap Detection - PNW Spotted Skunk
Camera Trap ID - Lynx

IDENTIFICATION PIPELINE
Camera Trap ID - Jaguar
Camera Trap ID - Jaguar

Identification Pipeline
“Reward the gift of data with the gift of knowledge”