Similarity Learning Networks for Animal Individual Re-Identification - Beyond the Capabilities of a Human Observer

Stefan Schneider
email sschne01@uoguelph.ca

Graham W. Taylor
email gwtaylor@uoguelph.ca

Stefan S. Linquist
email linquist@uoguelph.ca

Stefan C. Kremer
email skremer@uoguelph.ca

February 26, 2019

Key Terms - Animal Re-Identification, Camera Traps, Computer Vision, Convolutional Networks, Deep Learning, Density Estimation, Monitoring, Object Detection, Population Dynamics, Siamese Networks

Abstract

The ability of a researcher to re-identify (re-ID) an animal individual upon re-encounter is fundamental for addressing a broad range of questions in the study of ecosystem function, community and population dynamics, and behavioural ecology. Tagging animals during mark and recapture studies is the most common method for reliable animal re-ID however camera traps are a desirable alternative, requiring less labour, much less intrusion, and prolonged and continuous monitoring into an environment. Despite these advantages, the analyses of camera traps and video for re-ID by humans are criticized for their biases related to human judgment and inconsistencies between analyses. Recent years have witnessed the emergence of deep learning systems which re-ID humans based on image and video data with near perfect accuracy. Despite this success, there are limited examples of this approach for animal re-ID. Here, we demonstrate the viability of novel deep similarity learning methods on five species: humans, chimpanzees, humpback whales, octopus and fruit flies. Our implementation demonstrates the generality of this framework as the same process provides accurate results beyond the capabilities of a human observer. In combination with a species object detection model, this methodology will allow ecologists with camera/video trap data to re-identify individuals that exit and re-enter the camera frame. Our expectation is that this is just the beginning of a major trend that could stand to revolutionize the analysis of camera trap data and, ultimately, our approach to animal ecology.

Introduction

The ability to re-ID animals allows for population estimates which are used in a variety of ecological metrics including diversity, relative abundance distribution, and carrying capacity [1]. Ecologists have used a variety of techniques for re-ID including tagging, scarring, banding, and DNA analyses of hair follicles or feces [1]. While accurate, these techniques are laborious for the field research team, intrusive to the animal, and often expensive for the researcher.

Compared to traditional methods of field observations, camera traps are desirable due to their lower cost and reduced workload for field researchers. Camera traps also provide a unique advantage by recording the undisturbed behaviours of animals within their environment. These advantages have led to a 50% annual growth in publications using camera trap methods to assess population sizes between 1998 and 2008 and the trend has persisted until 2015 [2,3].

Despite their advantages, there are a number of practical and methodological challenges associated with the use of camera traps for animal re-ID. The discrimination of individual animals is often an expert skill requiring a considerable amount of training. Even among experienced researchers there remains an opportunity for human error and bias [4,5]. Historically, these limitations have restricted the use of camera traps to the re-ID of animals that bear conspicuous individual markings.

Recent decades have witnessed the emergence of deep learning systems that make use of large data volumes [6]. Modern deep learning systems no longer require 'hard-coded' feature extraction methods. Instead, these algorithms can learn, through their exposure to large amounts of data, the particular features that allow for the discrimination of individuals. [7]. These methods have been developed primarily outside the realm of ecology, first in the field of computer image recognition [8], and more recently in the security and social media industries [6]. Modern deep learning systems now consistently outperform feature engineered methods provided that they have access to large amounts of data [9,10].

When given ample data of all individuals of a population, traditional deep learning methodologies, such as Feedforward and Convolutional Neural Networks (CNNs), have demonstrated near perfect accuracy at re-ID. When considering animal re-ID however, capturing a library of photos of every individual within a population is infeasible. Monitoring animal populations is categorized as a 'one-shot' learning problem, where the system must accurately re-ID an individual based on seeing only one previous example. Here, we describe, test, and demonstrate the success of the Siamese similarity learning network for the animal re-ID one-shot learning problem for five species: humans, chimpanzees, humpback whales, octopus and fruit flies. Our results demonstrate how, with enough data, these systems can be used to re-ID animal individuals from camera trap data beyond the capabilities of a human observer.

Deep Learning Methods for Re-Identification

Prior to deep learning methods, for decades the approach to standardizing the statistical analysis of animal re-ID involved computer vision. 'Feature engineering' has been the most commonly used computational technique where algorithms are designed and implemented to focus exclusively on pre-determined traits, such as the detection of patterns of spots or stripes, to discriminate among individuals. The main limitations of this approach surround its impracticality [11]. Feature engineering requires programming experience, sufficient familiarity with the organisms to identify relevant features, and lacks in generality where once a feature detection algorithm has been designed for one species, it is unlikely to be useful for other taxa.

Modern deep learning systems have shown great success learning the necessary features for re-ID from data and removes the need for feature engineering. Despite its long history, there has been a rapid growth of interest in deep learning due to its success related to improved computational power and the availability of large data sets, both requirements for the model. In recent years, deep learning methods have dramatically improved performance levels in the fields of speech recognition, computer vision, drug discovery, genomics, artifical intelligence, and others becoming the standard computational approach for problems with large amounts of data [7]. For an intuitive description of the mechanisms and functionality of deep learning systems relevant to animal re-ID see Schneider et al. (2018) [12]

The success of deep learning methods for human re-identification is well documented when ample training images are available for each individual. In 2015, two research teams, Lisanti et al. and Martinel et al., demonstrated the successful capabilities of CNNs on human re-ID using the ETHZ data set, a data set composed of 8580 images of 148 unique individuals taken from mobile platforms, where CNNs were able to correctly classify individuals from the test set with 99.9% accuracy after seeing 5 images of an individual [9, 10]. In 2014, Taigman et al. introduced Deepface, a method of creating a 3-dimensional representation of the human face to provide more data to a neural network which returned an accuracy of 91.4% on the YouTube faces dataset containing videos of 1,595 individuals [13].

Despite the success of deep learning methods for human re-ID, few ecological studies have utilized its advantages. In 2014, Carter et al. published one of the first works using neural networks for animal re-ID, a tool for green turtle (*Chelonia mydas*) re-ID [14]. The authors collected 180 photos of 72 individuals from Lady Elliot Island in the southern Great Barrier Reef, both nesting and free swimming considering an undisclosed number of testing images. Their algorithm pre-processes the image by extracting a shell pattern, converting it to grey scale, unravelling the data into a raw input vector, and then training a simple feedforward network [14]. Each model produces an output accuracy of 80-85% accuracy, but the authors utilize an ensemble approach by training 50 different networks and having each vote for a correct classification. The ensemble approach returns an accuracy of 95%. Carter et al.'s work has been considered a large success and is currently used to monitor the southern Great Barrier Reef green turtle population [14].

In 2016, Freytag et al. trained the CNN architecture AlexNet on the isolated faces of chimpanzees considering two chimpanzee data sets: C-Zoo and C-Tai [15]. Freytag et al. (2016) report an improved accuracy of 92.0% and 75.7% in comparison to the original Support Vector Machine method of 84.0% and 68.8% [15, 16]. In 2017, Brust et al. trained the object detection method YOLO to extract cropped images of Gorilla (*Gorilla gorilla*) faces from 2,500 annotated images camera trap images of 482 individuals taken in the Western Lowlands of the Nouabalé -Nodki National Park in the Republic of Congo [17, 18]. Once the faces are extracted, Brust et al. (2017) followed the same procedure as Freytag et al. (2016) to train the CNN AlexNet achieving a 90.8% accuracy on a test size of 500 images [15, 17]. The authors close discussing how deep learning for ecological studies show promises for a whole realm of new applications if the fields of basic identify, spatio-temporal coverage and socio-ecological insights. [15, 17]

Traditional CNNs for re-ID requires a large number of labeled data for each individual and re-training the network for every new individual sighted, both of which are infeasible requirement for animal re-ID research efforts. In 1993, Bromley et al. introduced a suitable neural network architecture for this problem, titled a Siamese network, which learns to detect if two input images are similar or dissimilar [19]. Once trained, Siamese networks require only one labeled input image of an individual in order to accurately re-identify the second input image is of the same individual. In practice, one would train a Siamese network to learn a species' similarity and compare individuals using a known ground truth, such as a zoo or wildlife footage. In 2016, Schroff et al. introduced the Siamese-based network architecture FaceNet which currently holds the highest accuracy on the YouTube Faces data set with a 95.12% top-1 accuracy [20].

In 2018, Deb et al. (2018) addressed the one-shot learning probelm for animal re-ID considering three species: chimpanzees, lemurs, and golden monekys. She formulated the problem into three categories for testing successful re-ID: verification (determine if two images are the same individual), closed-set identification (identify an individual from a given set of images data), and open-set identification (identify an individual from a given set of images or conclude the individual is absent from the data) [21]. For chimpanzees, Deb et al. (2018) combined the C-Zoo and C-Tai data sets to create the *ChimpFace* data set containing 5,599 images of 90 chimpanzees. For lemurs, they consider a data set known as *LemurFace* from the Duke Lemer Center, North Carolina containing 3,000 face images of 129 lemur individuals from 12 different species. For golden monkeys, they extracted the faces of 241 short video clips (average 6 seconds) from Volcanoes National Park in Rwanda where 1,450 images of 49 golden monkey faces were cropped and extracted [21]. Deb et al., (2018) use a custom Siamese CNN containing four convolutional layers, followed by a 512 node fully connected layer [21]. Deb et al. (2018) report verification, closed-set, and open-set accuracies respectively for lemurs: 83.1%, 93.8%, 81.3%, golden monkeys: 78.7%, 90.4%, 66.1%, and chimpanzess: 59.9%, 75.8%, and 37.1% [21].

Similarity Learning Networks for Animal Re-Identification

When approaching the problem of animal re-ID, traditional softmax classifiers are not a viable option as they require a data set containing a large number of examples for every individual from the population. This is unrealistic to obtain when considering animal populations. In order to utilize deep learning for animal re-ID, I instead propose training a Siamese similarity network to successfully compare if two input images are of the same individual.

To test the capabilities of similarity networks on animal re-ID, we consider five species using the following data sets. Each species and data set provide a unique challenge allowing for a robust analysis of the performance of the methodology:

- FaceScrub: 106, 863 images of 530 male/female human individuals varying in pose [22].

- ChimpFace: 5, 599 images of 95 male/female chimpanzee individuals. This is a combination of two previous data sets: C-Tai and C-Zoo [15].
- HappyWhale: 9, 850 images of 4, 251 humpback whale (*Drosophila melanogaster* individuals offered as an expired Kaggle competition [23].
- FruitFly Data: 244, 760 images of 20 fruit flies in a variety of poses [24].
- Octopus: 5, 192 images of an unknown number of octopus (*Octopus vulgaris*). Images are captured from research footage and identifies individuals as being different only when on camera at the same time [25].

To train a similarity network for animal re-ID, we create an equal number of image pairs for each individual labeled as same and different. To measure the performance of a model, I consider verification accuracy on a created validation and test set for each species [21]. The validation data are created by randomly splitting the created pair data by a ratio of 1/10 split. This creates a data set of individuals seen during training, but in novel combinations. To test how well the model generalizes to individuals not seen during training, for each data set I also create a test set by randomly excluding 10% of the individuals to create a pairwise data set of unseen individuals. This second data set provides a better representation for how well the model generalizes to realistic scenarios of unseen individuals and should be considered as our primary metric for performance.

For the re-ID of Humans, we consider the publicly available FaceScrub data set [22] (Figure 1). This data set allows for a benchmark comparison of our methodology in comparison to other human similarity networks. After following the described data creation format, the training set contains 218,872 pairs of images, the validation set 21,231 pairs of images, and the test set 32,398 pairs of images created by considering 16 random individuals not included in the training data.

Figure 1: Example Images from the FaceScrub Data Set

For the re-ID of Chimpanzee, I will consider the publicly available C-Zoo and C-Tai data sets introduced by Loos and Ernst (2013) combined into the *Chimpface* dataset introduced by Deb et al (Figure 2). (2018) [16, 21]. This data set provides the unique opportunity of comparing the performance of similarity comparison models to the previously reported performance of feature engineering as well as deep learning methods. After following the described data creation format, the training set contains 148,516 pairs of images, the validation set 13,656 pairs of images, and the test set 16,398 pairs of images created by considering 8 random individuals not included in the training data.

To test re-identifying Humpback Whales, I will consider the Humpback Whale Identification Challenge data set offered as a Kaggle competition [23]. This data set provides a realistic representation of the real world application of animal re-ID as the 9,046 images only contain the fluke of the whale and are extremely sparse, having only an average of only 2 (+/- 8) individuals considering 4,251 individual classifications (Figure 8). After creating the pairwise data, the training set has 25,399 image pairs, validation set 2,655 image pairs, and test set 3,822 image pairs. There are much fewer images because there are far less pair combinations for each individuals.

To test the re-ID of Fruit Flies, we collaborate John Schneider to test our similarity network on a large library of high resolution images of fruit flies [24] (Figure 3). The large number of images available here provide an excellent opportunity to test the capabilities of similarity learning networks on an animal species where re-ID is beyond the capabilities of a human observer. After following the described data creation format, the training set contains 216, 421 pairs of images, the validation set 23, 423 pairs of images, and the test set 17,632 pairs of images created by considering 4 random individuals not included in the training data.

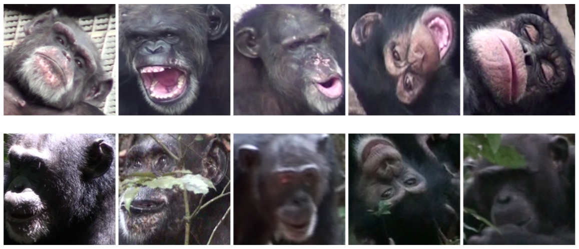

Figure 2: Example Images from the Chimpface Chimpanzee Data Set

Lastly, to test the re-ID capabilities considering octopus we consider a real world scenario where images are not readily curated into a data set, but instead extracted from video when multiple individuals are on screen [25]. This data set demonstrates the framework for real-world application as it requires an accurate object detector to identify octopus as they enter the video frame and a similarity comparison network to determine if these individuals has been previous seen before. Gathering the data required to train the object detector involves documenting the bounding box coordinates for each octopus in each frame of video to train the object detector as well as extract these isolated octopus images into labeled folders of unique individuals to train the similarity comparison network. An example of a trained a Faster R-CNN object detection model using the Inception architecture to localize octopus within an video sequence with 96.4% accuracy can be found at: https://www.youtube.com/watch?v=TXbv5pN4JRI [12].

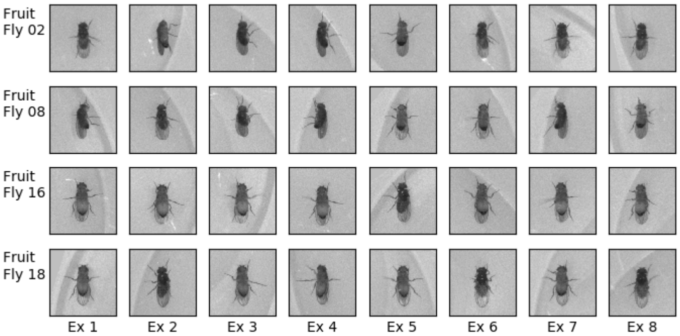

Figure 3: Example Images from the Fruit Fly Data Set

To collect bounding box and extract images of octopus individuals, I have created a universally applicable video data extraction tool to streamline the process. The user begins by selecting their video of interest. The software then asks the user to place their mouse cursor at the top left corner of an object of interest (eg. octopus) and follow along while the video plays and it records the X&Y position of the mouse cursor for every frame. The software then repeats the video for the user to follow the bottom right corner of the same object. The user can then select if there are additional objects and repeat/review the process until the video is fully labeled. The end result are the necessary files required for the two data formats: the bounding box coordinates per frame of each octopus individual used to train the object detector, and the extracted images of each octopus individual used to train the similarity network. A demo of this software is available at: https://youtu.be/YcTj0ayztA4 and is publicly available on GitHub at https://goo.gl/XomSao.

For our testing purposes, we labeled 16 short octopus videos with 3.5 average octopuses per video and an average length of 2:14 minutes. Three randomly selected videos are used to create the test set. We manually inspected each catalog of images and delete approximately 90% of images that appear redundant over time (ie. when an octopus remains stationary). We then follow the same Siamese data creation format as listed above for each video independently to create a current total of 80,916 training pairs, 9,080 validation pairs, and 17,968 unseen testing pairs. Pairwise interactions are only considered within each video segment where the octopus remains on screen as the identity of individuals across videos is uncertain.

Animal Re-Identification Results

Our network architecture follows the Siamese Network construct, composed of two sister networks where we implemented two VGG19 architectures, consisting of 13 2-dimensional convolution layers and 5 max pooling layers [26]. The features of the last convolution layers are then concatenated together and passed through two 2048 node fully connected layers, and lastly through a binary output representing similar/dissimilar pairs. Details of the training process can be found in Appendix A.

Considering the FaceScrub data set, the model produces a training set accuracy of 91.3%, a validation accuracy of 91.3%, and test set accuracy of individuals never seen during training of 89.7%. Considering state of the art models, such as FaceNet, our model does not match their level of performance, however it does perform well enough to indicate their is room for improvement in terms of architecture, and demonstrate the capabilities of this system for animal re-ID. This serves a platform of comparison for how similarity learning models perform on a variety of difference species and data set composition.

Considering the *Chimpface* data set, the model produces a training set accuracy of 88.2%, a validation accuracy of 87.5%, and test set accuracy of individuals never seen during training of 75.5%. This is a large improvement over the verification score of 59.9% using the same data reported by Deb et al. (2018) [21] (Figure 4 & 5).

Considering the humpback whale data set, the model returns a training set accuracy of 65.2%, validation accuracy of 62.3% and test accuracy of 61.4%. A low training accuracy suggests either the model is not complex enough to capture the representation of whale individuals, which is unlikely considering its success on *Chimpface*. As a result, we further data augmentation techniques

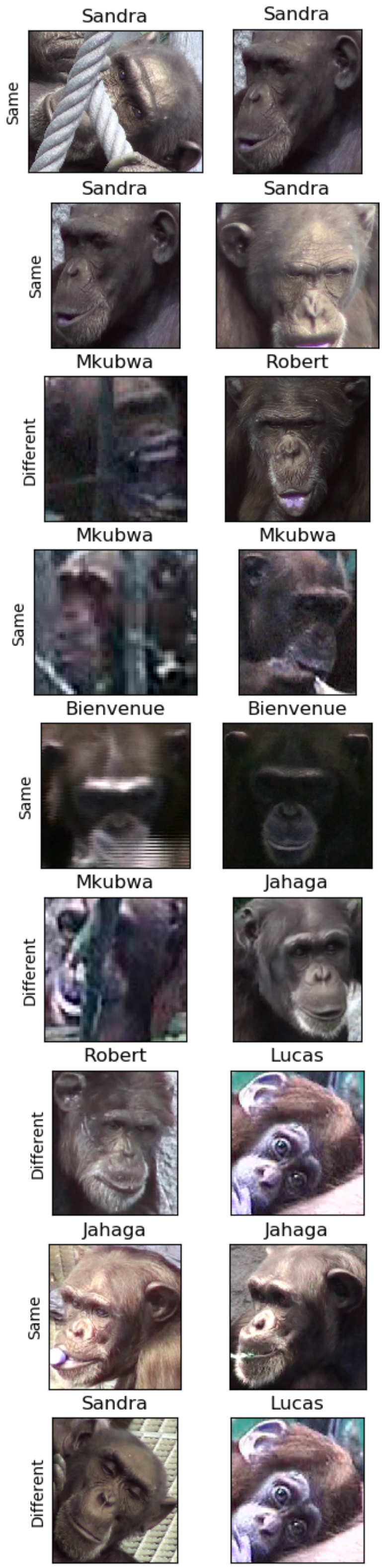

Figure 4: Model Output for Chimpanzee Individuals. Y label for each image is the model output and the X label is the name of each individual in the below photo

may help improve performance further (Figure 6). While a poor result in comparison to acceptable standards for monitoring, while the competition has since expired, our approach currently stands as the 4th best performing model considering 528 entries.

For the fruit fly data set, the model returns a training set accuracy of 82.4%, validation accuracy of 82.0% and test accuracy of 79.3%. While distinguishing between fruit fly individuals seems to be an impossible task for humans, our model was able to successfully learn features which accurately distinguish between individuals. Considering the data set, the success of this model is likely based on the very large number of training images available, as well as the standardized background of the images themselves.

Considering the created octopus data set, the model returns a training accuracy of 97.8%, validation accuracy of 95.3% and testing accuracy of 92.2% considering octopus it hasn't seen during training (Figure 7). These results show promise considering more data can still be collected. Pushing to achieve greater than 95% accuracy seems like a tangible objective with additional image augmentation techniques.

A summary of results can be found in Table 1.

Near Future Techniques for Animal Re-Identification

By considering modern deep learning approaches, ecologists can utilize improve accuracies without the requirement of hand-coded feature extraction methods by training a neural network from large amounts of data. The success of our models across multiple species and environments show that similarity learning networks are capable of solving the one-shot learning problem associated with animal population monitoring.

In practice, a wildlife re-ID system would work as follows. One would collect or find a data library of images of animal individuals for the species in consideration, ideally in the 1,000+. One would then organize the images into pairs and train a Siamese network to distinguish if two animal individuals are the same. In addition, one would train an object detector to localize the animal species from a camera trap image [12]. Once both models are trained, one can set up the object detector to extract animal species from images which are then fed into the Siamese network. One can be used to estimate population sizes by querying a database. Upon initialization, this database would be empty. As each individual enters into the camera, the network will query each existing animals within the database. If none are deemed to be similar, an image of the new individual will be added to the database and the process repeats for each individual that enters.

In order for such technqiues to become generally applicable, we foresee the greatest challenge for deep learning methods being the creation of large 1,000+ labeled datasets for animal individuals. Our proposed approach for data collection would be to utilize environments with known ground truths for individuals, such as national parks, zoos, or camera traps in combination with individuals being tracked by GPS, to build the datasets. We would recommend using video wherever possible to gather the greatest number of images for a given encounter with an individual. We encourage researchers with images of labeled animal individuals to make these datasets publicly available to further the research in this field. In addition to gathering the images, labeling the data is then also a labourious task, especially when training an object detection model where bounding boxes are required. One approach for solving this problem is known as weakly supervised learning, where one provides object labels to a network (ie. zebra) and the network returns the predicted coordinates of its location [27]. An alternative approach is to outsource the labeling task to online services, such as Zooniverse which can be time saving for researchers, but introduces inevitable error and variability [28].

While deep learning approaches are able to generalize to examples similar to those seen during training, we foresee various environmental, positional, and timing related challenges. Environmental difficulties may include challenging weather conditions, such as heavy rain, or extreme lighting/shadows, especially from video analysis that only make comparisons between similar weather conditions. A possible solution to limit these concerns may be to re-ID only during optimal weather conditions. A second is to include a robust amount of image augmentation. A positional challenge may occur if an individual were to enter the camera frame at extremely near or far distances. To solve this, one could limit animals to a certain range from the camera before considering it for re-ID. Lastly, a challenge may be if an individual's appearance were to change dramatically between sightings, such as being injured or the rapid growth of a youth. While a network would be robust to such changes given training examples, this would require examples be available as training data. To account for this issue we would consider having a 'human-in-the-loop' approach, where a human monitors results and relabels errorenous classifications for further training to improve performance [29].

While today fully autonomous re-ID is still in development, researchers can already use these systems to reduce manual labour for their studies. Examples include training networks to filter images by the presence/absence of animals, or species classifications [12, 30, 31]. Soon deep learning systems will accurately perform animal re-ID at which time one can create systems that autonomously extract from camera traps a variety of ecological metrics such as diversity, evenness, richness, relative abundance distribution, carrying capacity, and trophic function, contributing to overarching ecological interpretations of trophic interactions and population dynamics.

Conclusion

Population estimates are the underlying metric for many fundamental ecological questions and rely on the accurate re-identification of animals. Camera and video data have become increasingly common due to their relatively inexpensive method of data collection, however they are criticized for their unreliability and bias towards animals with obvious markings. Feature engineering methods for computer vision have shown success re-identifying animal individuals and removing biases from these analyses, however these methods require algorithms designed for feature extraction. Deep learning provides a promising alternative for ecologists as it learns these features from large amounts and has shown success for human and animal re-ID. By utilizing deep learning methods for object detection and similarity comparison, ecologists can utilize deep learning methods to autonomously re-identify animal individuals from camera trap data. Our results here demonstrate how such a tool would allow ecologists to automate population estimates.

Appendix A

In addition to the general architecture, we consider a variety of regularization techniques to improve training stability and prevent overfitting. Initializing the model with the Xavier (also known as Glorot normal) initialization helps improve the likelihood of successful training over random initialization which often stays within a plateu with no gradient information [32]. Also, before each fully connected layers I implement batch normalization, a technique shown to help improve generalization of a model [33]. To prevent overfitting, I also regularize the final two fully connected layers with a Ridge regression value of 0.01 which penalizes the complexity of the function by including a squared error term that the model must also minimize [34]. I also include a 50% dropout for both fully connected layers which for each mini-batch randomly prohibits the use of 50% of the nodes in the fully connected layers [35].

In addition to these techniques, to help improve generalization I also apply data augmentation methods during training. For every training example, there is a given probability that the image will be modified in some capacity while maintaining the correct labels. This technique helps improve generalization and reduce overfitting as it artificially creates additional training examples. For each training example there is a random chance of noise to be added for colour channel manipulation, per pixel manipulation, bluriness, and pixel dropbout. Image Augmentation are performed using the ImgAug library [36].

To train the weights of the neural network I select the Adaptive Momentum (Adam) optimizer with a learning rate of 0.001 [37]. Simply stated, Adam is regarded as the most capable optimizer, able to increase and decrease its step size relevant to the speed at which the error is decreasing and also adjusting its direction travel to the steepest descent similar to the optimizer RMS Prop [37]. To represent error, I consider the binary crossentropy loss considering the binary output of the similarity network. I train the model using mini-batch sizes of 128 training examples. As a metric for recording performance during training, I consider the accuracy on the test set of previously unseen individuals and only save the model if the current epoch is an improvement over the previous. In essence, this approach allows me to consider early stopping as a means of regularization and selects the model which best generalizes to unseen individuals. With the above considerations, I train the model for 100 epochs. Training and analyses of this model were performed using Python 3.6, Tensorflow 1.8, and Keras 2.2 using two GPUS: NVidia 1080 GTX and NVidia 680 GTX.

Table 1: Summary of Performance Metrics for Similarity Learning Models be Species & Data Set

Species	Total Images	Num. Individuals	Avg. Num. Images / Individual	Num. Pairs Used for Training	Val. Acc.	Test Acc.
Human	65,943	530	127.3	218,872	91.3	89.7
Chimpanzee	5, 599	90	71.0	148,516	87.5	75.5
Humpback Whale	9,046	4,251	2.1	25,399	62.3	61.4
Fruit Fly	244,760	20	12,238.0	218,872	82.0	79.3
Octopus	5,192	N/A	92.7	80,916	95.3	92.2

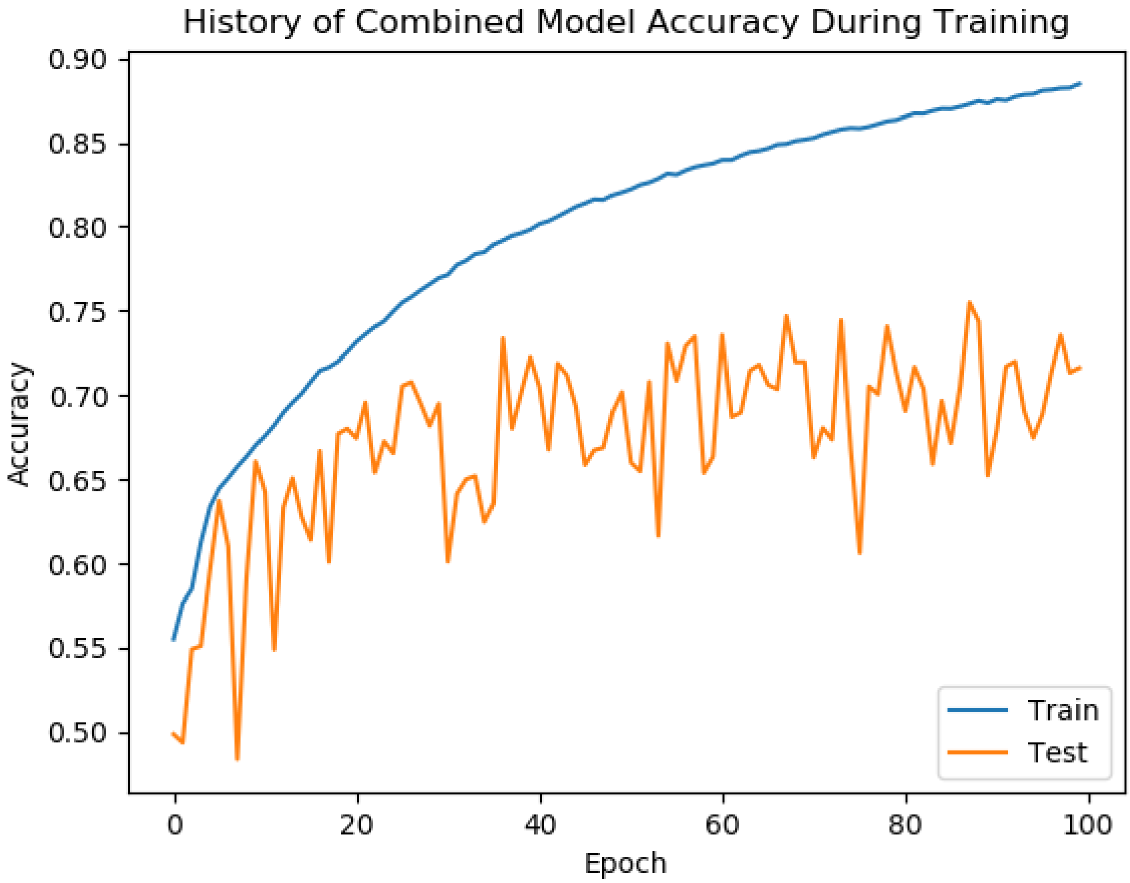

Figure 5: Chimpanzee Accuracy

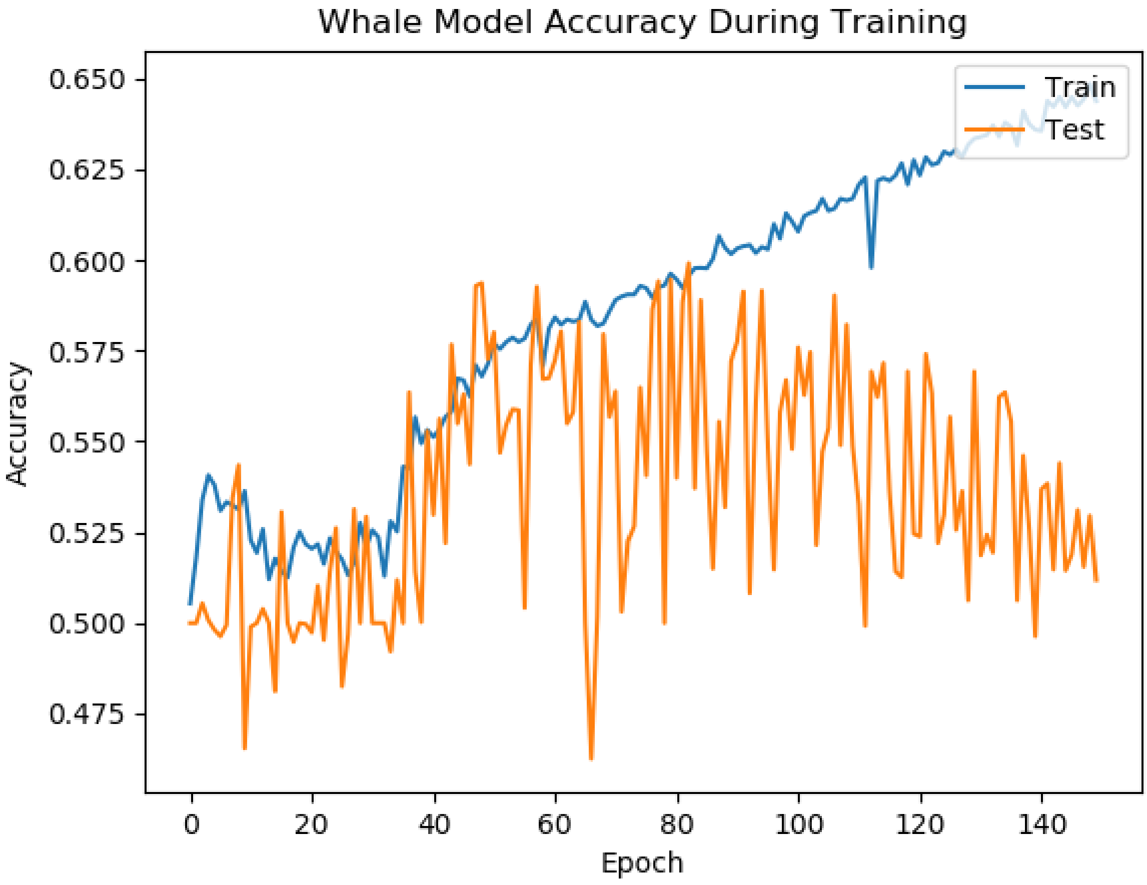

Figure 6: Humpback Whale Accuracy

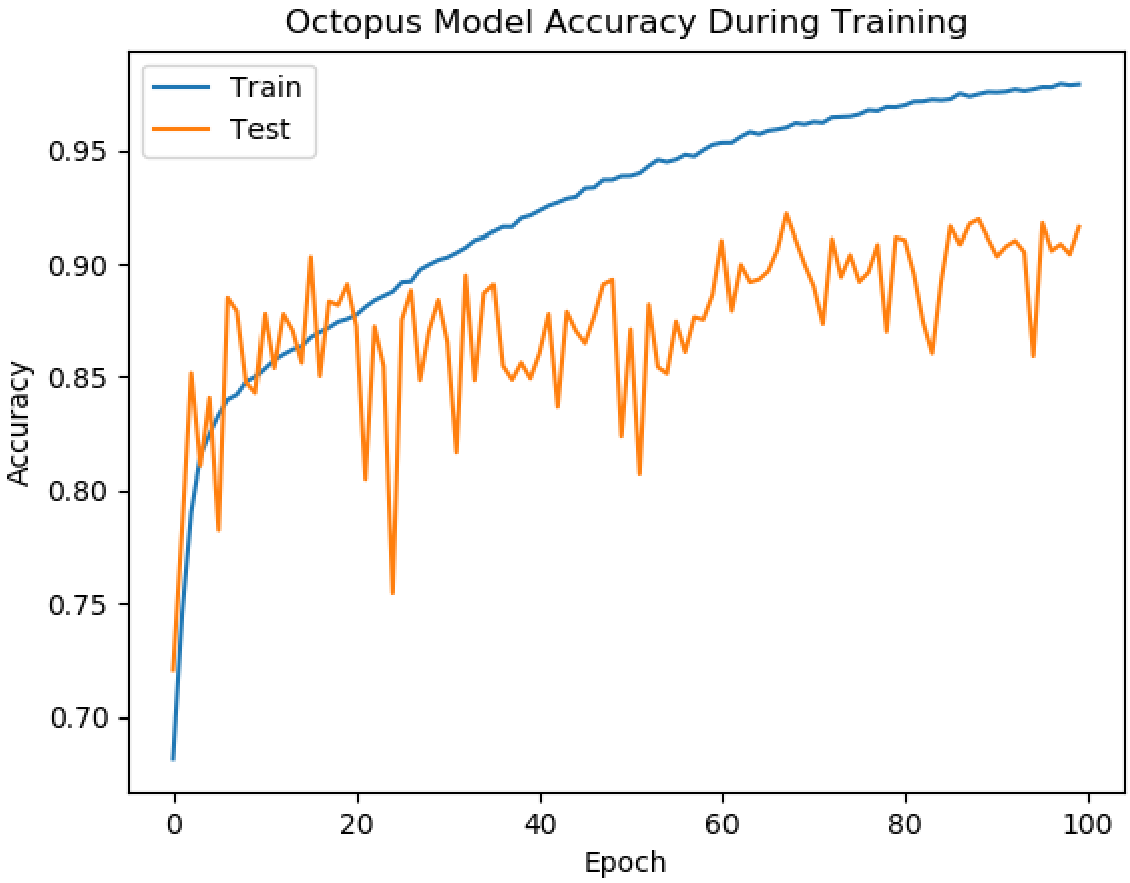

Figure 7: Octopus Accuracy

References

[1] C. J. Krebs *et al.*, “Ecological methodology,” Harper & Row New York, Tech. Rep., 1989.

[2] J. M. Rowcliffe, J. Field, S. T. Turvey, and C. Carbone, “Estimating animal density using camera traps without the need for individual recognition,” *Journal of Applied Ecology*, vol. 45, no. 4, pp. 1228–1236, 2008.

[3] A. C. Burton, E. Neilson, D. Moreira, A. Ladle, R. Steenweg, J. T. Fisher, E. Bayne, and S. Boutin, “Wildlife camera trapping: a review and recommendations for linking surveys to ecological processes,” *Journal of Applied Ecology*, vol. 52, no. 3, pp. 675–685, 2015.

[4] R. J. Foster and B. J. Harmsen, “A critique of density estimation from camera-trap data,” *The Journal of Wildlife Management*, vol. 76, no. 2, pp. 224–236, 2012.

[5] P. D. Meek, K. Vernes, and G. Falzon, “On the reliability of expert identification of small-medium sized mammals from camera trap photos,” *Wildlife Biology in Practice*, vol. 9, no. 2, pp. 1–19, 2013.

[6] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, “Scalable person re-identification: A benchmark,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1116–1124.

[7] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[8] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[9] G. Lisanti, I. Masi, A. D. Bagdanov, and A. Del Bimbo, “Person re-identification by iterative re-weighted sparse ranking,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 8, pp. 1629–1642, 2015.

[10] N. Martinel, A. Das, C. Micheloni, and A. K. Roy-Chowdhury, “Re-identification in the function space of feature warps,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 8, pp. 1656–1669, 2015.

[11] L. Hiby, P. Lovell, N. Patil, N. S. Kumar, A. M. Gopalaswamy, and K. U. Karanth, “A tiger cannot change its stripes: using a three-dimensional model to match images of living tigers and tiger skins,” *Biology letters*, pp. rsbl–2009, 2009.

[12] S. Schneider, G. Taylor, and S. Kremer, “Deep learning object detection methods for ecological camera trap data,” *Conference on Computer and Robot Vision*, to appear.

[13] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, “Deepface: Closing the gap to human-level performance in face verification,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1701–1708.

[14] S. J. Carter, I. P. Bell, J. J. Miller, and P. P. Gash, “Automated marine turtle photograph identification using artificial neural networks, with application to green turtles,” *Journal of experimental marine biology and ecology*, vol. 452, pp. 105–110, 2014.

[15] A. Freytag, E. Rodner, M. Simon, A. Loos, H. S. Kühl, and J. Denzler, “Chimpanzee faces in the wild: Log-euclidean cnns for predicting identities and attributes of primates,” in *German Conference on Pattern Recognition.* Springer, 2016, pp. 51–63.

[16] A. Loos and A. Ernst, “An automated chimpanzee identification system using face detection and recognition,” *EURASIP Journal on Image and Video Processing*, vol. 2013, no. 1, p. 49, 2013.

[17] C.-A. Brust, T. Burghardt, M. Groenenberg, C. Käding, H. S. Kühl, M. L. Manguette, and J. Denzler, “Towards automated visual monitoring of individual gorillas in the wild,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2820–2830.

[18] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You only look once: Unified, real-time object detection,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 779–788.

[19] J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, and R. Shah, “Signature verification using a ”siamese” time delay neural network,” in *Advances in Neural Information Processing Systems*, 1994, pp. 737–744.

[20] F. Schroff, D. Kalenichenko, and J. Philbin, “Facenet: A unified embedding for face recognition and clustering,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 815–823.

[21] D. Deb, S. Wiper, A. Russo, S. Gong, Y. Shi, C. Tymoszek, and A. Jain, “Face recognition: Primates in the wild,” *arXiv preprint arXiv:1804.08790*, 2018.

[22] H.-W. Ng and S. Winkler, “A data-driven approach to cleaning large face datasets,” in *2014 IEEE International Conference on Image Processing (ICIP).* IEEE, 2014, pp. 343–347.

[23] “Humpback whale identification challenge,” https://www.kaggle.com/c/whale-categorization-playground, accessed: 2018-05-15.

[24] J. Schneider, N. Murali, G. W. Taylor, and J. D. Levine, “Can drosophila melanogaster tell whos who?” *PloS one*, vol. 13, no. 10, p. e0205043, 2018.

[25] D. Scheel, S. Chancellor, M. Hing, M. Lawrence, S. Linquist, and P. Godfrey-Smith, “A second site occupied by octopus tetricus at high densities, with notes on their ecology and behavior,” *Marine and Freshwater Behaviour and Physiology*, vol. 50, no. 4, pp. 285–291, 2017.

[26] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *arXiv preprint arXiv:1409.1556*, 2014.

[27] Z.-H. Zhou, “A brief introduction to weakly supervised learning,” *National Science Review*, vol. 5, no. 1, pp. 44–53, 2017.

[28] R. Simpson, K. R. Page, and D. De Roure, “Zooniverse: observing the world’s largest citizen science platform,” in *Proceedings of the 23rd international conference on world wide web.* ACM, 2014, pp. 1049–1054.

[29] A. Holzinger, “Interactive machine learning for health informatics: when do we need the human-in-the-loop?” *Brain Informatics*, vol. 3, no. 2, pp. 119–131, 2016.

[30] “iWildcam 2018 camera trap challenge,” https://www.kaggle.com/c/iwildcam2018, accessed: 2018-07-11.

[31] M. S. Norouzzadeh, A. Nguyen, M. Kosmala, A. Swanson, M. Palmer, C. Packer, and J. Clune, “Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning,” *ArXiv:1703.05830v5*, 2017.

[32] X. Glorot and Y. Bengio, “Understanding the difficulty of training deep feedforward neural networks,” in *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, 2010, pp. 249–256.

[33] S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” *arXiv preprint arXiv:1502.03167*, 2015.

[34] A. E. Hoerl and R. W. Kennard, “Ridge regression: Biased estimation for nonorthogonal problems,” *Technometrics*, vol. 12, no. 1, pp. 55–67, 1970.

[35] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Dropout: A simple way to prevent neural networks from overfitting,” *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[36] “imgaug,” https://imgaug.readthedocs.io/en/latest/index.html, accessed: 2010-20-02.

[37] D. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.