Animal Identification with Convolutional Neural Networks

Arya Subramanyam
arya.subramanyam@gmail.com

Abstract

Animal identification is an important practice with numerous applications in the realms of wildlife conservation, poaching, animal behavior research, and biodiversity monitoring. It is essential for animals to be identified accurately, especially when the amounts of useful data collected may be limited. We train 3 Convolutional Neural Networks (CNNs), including a basic CNN designed from scratch, ResNet-18 trained from scratch, and a version ResNet-18 pre-trained on ImageNet, to determine which CNN yields the most accurate and precise animal identification. The ANIMAL-10N dataset serves as our benchmark for evaluating the performance of the CNN architectures. After training, various experiments were performed on the models to learn about their abilities and the impacts of transfer learning on such tasks. We calculated each neural network's top-1 classification accuracy, precision and recall. ResNet-18 with transfer learning emerged as the most accurate model for animal classification with a peak training accuracy of 87.6% over 20 epochs. Its classification patterns were quantified used the Confusion Matrix and its features were visualized using Saliency Maps. The contribution this paper makes is exploring the learning behaviour and features of CNNs for animal identification and analyzing the effects of transfer learning, as well as shallow versus deep networks on their classification effectiveness. The results showcase that transfer learning makes the biggest contribution to achieving highly-accurate animal classification.

1 Introduction

Our planet has witnessed a 68% decline in bird, reptile, fish, mammal, and amphibian populations on average, since 1970. (1) These catastrophic findings reveal the urgency in protecting and conserving our wildlife and taking measures to decrease poaching. Daily, vasts amount of image and video data is collected from jungles, forests and other habitats. This data is used to identify and track animals in the wild to help researchers track population dynamics, monitor migration patterns, analyze behaviours and assess animal health conditions. Additionally, animal recognition and tracking data equips governments and other bodies to initiate targetted efforts against illegal wildlife tracking and reveals trends that are used for decision-making to protect endangered species.

A variety of methods and devices are used for wild animal monitoring. A commonly used recognition-based tool is motion-sensitive camera traps to capture wildlife on film. These devices are typically placed in remote locations and use infrared sensors. Camera traps are growing in popularity as they are easily available and convenient to deploy and operate. They capture high-resolution images at any time of the day, and are effective devices to track wildlife covertly and consistently. The images they capture are often in a variety of complex scenes among diverse natural settings. Since animals can be pictured in both day or night, different illumination may occur, along with varying angles and scales. (2) Furthermore, some camera traps utilise flash which often inform animals of their presence, due to which animals tend to avoid those locations. (3) In addition, cameras may malfunctions due to climate conditions which results in a much lower than expected amount of actual, valuable images of animals. With these limited pictures in varying conditions, the importance of accurate animal identification is immense.

There is a body of related works which attempts to achieve animal identification using deep neural networks, however, some have trained their models on small datasets consisting of only a couple of thousand of images. (4) Others use manually-crafted features for recognition, specifying qualities and characteristics such as textures and colours to detect animals. (5) Some research has focused on camera trap pictures to detect wildlife. (6) Additionally, there has even been previous work focusing on differences within species such as dogs using the Stanford dogs (SD) and Oxford IIIT-Pet dataset (OX) for animal breed classification. (7) In recent times, CNNs have shown great abilities in image recognition with 98%+ with a single convolutional layer, surpassing the 97% accuracy of simple neural networks. (8) This paper will explore the abilities of 3 difference CNNs to learn about their accuracy for this context.

This paper takes inspiration from a similar feat to analyse CNNs and their capabilities for image classification. The paper takes a focus on artist identification from images of fine art. (9) We apply their concepts to the domain of animal classification instead. The following notable contributions are made through this effort:

- Examine and visualise the learning behaviours of CNNs for animal identification.
- Experiment on networks that do and do not use transfer learning, as well as shallow and deep networks to determine their effectiveness for animal identification.

2 Method

2.1 Dataset

2.1.1 Overview

We have used the ANIMAL-10N dataset to train, as well as, test our CNNs to identify animals. This dataset was obtained from Deep Lake, a data lake for deep learning applications. It was created by crawling various search engines (Google, Bing) by searching for the predecided classes as keywords. These pictures were then labelled by a team of 15 participants. It has $50,000$ pictures for training, and $5,000$ pictures for testing. There are noisy labels within the dataset that have been introduced spontaneously through human error. The noisy labels exist at an approximate rate of 4%. (10) Each image within this dataset is of shape 64x64 (RGB).

The 10 classes of animals within the dataset include 5 pairs of confusing aninals. They are as follows: (cat, lynx), (jaguar, cheetah), (wolf, coyote), (chimpanzee, orangutan), (hamster, guinea pig). The distribution of testing images across all labels is within a close range, with at most 5466 images for a the label of cat and at least 4608 for the label of lynx. (11)

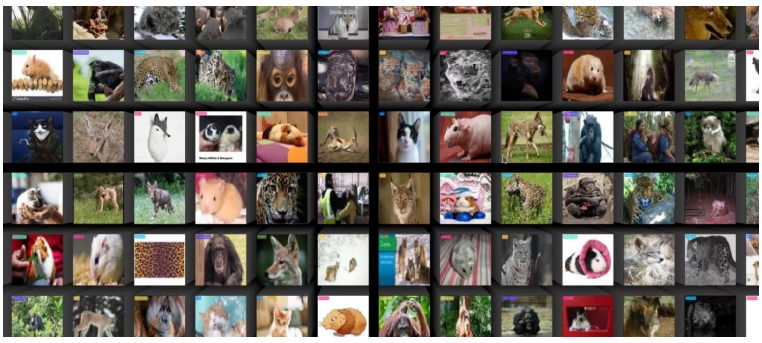

Fig. 1: Animal-10N Visualization (10)

2.1.2 Preprocessing

We conduct some data augmentation and transformations to the images before inputting them into our CNNs. The same preprocessing is applied to both, training and testing sets. We first rescale all the images to 224x224. We hypothesis that this rescaling is justified as detail at the pixel-level would not be necessary for animal identification. Next, the images are flipped horizontally with a probability of 50% to increase the randomness of our data. This will add variety to our dataset and prevent

overfit from occuring. Lastly, the images are normalized using the mean and standard deviation of ImageNet as these values have been proven to work on 'natural' images and could be treated as a reliable baseline. (12) Their values are as follows:

$$mean = [0.485, 0.456, 0.406]$$
$$std = [0.229, 0.224, 0.225]$$

2.2 Models

For this experiment, we train three distinct CNN-based models.

2.2.1 Baseline CNN

The architecture of the Baseline CNN was motivated by the Baseline CNN utilised in the Artist Identification Experiment we take inspiration from. (9) This model was built and trained from scratch on the ANIMAL-10N dataset. It essentially exists as a baseline comparison for the other CNNs being experimented on. The architecture of this network can be seen in Table 1. Layer by layer, the image is down-sampled by a factor of two. This has been done in order to lessen the complexity of computation due to the limited resources that are available for training. However, there may be some disadvatanges to this approach. It is possible that the model would not be able to easily learn details of the images and other fine-grained features which may assist with the animal identification.

Input Size	Layer
3x22x224	2D Conv, kernel = 3, stride = 2, padding = 1
32x112x112	2D Maxpool, kernel = 2, stride = 2
32x56x56	2D Conv, kernel = 3, stride = 2, padding = 1
32x28x28	2D Maxpool, kernel = 2, stride = 2
1x6272	Fully-connected
1x228	Fully-connected

Tab. 1: Architecture of Baseline CNN model. Batch Normalization and ReLU Layers have not been included in this table.

2.2.2 ResNet-18 Trained from Scratch

We utlise the ResNet-18 architecture but with a modified fully-connected layer to output scores for the 10 labels in the dataset. We train this network from scratch to ensure that it learns animal classification. The architecture of this model allows for gradient flow directly to deep layers from shallow ones, providing the possibility of improved training. The choice of ResNet-18 was made to ensure faster training and lower memory usage than versions with more layers. Its architecture is shown in Figure 2.

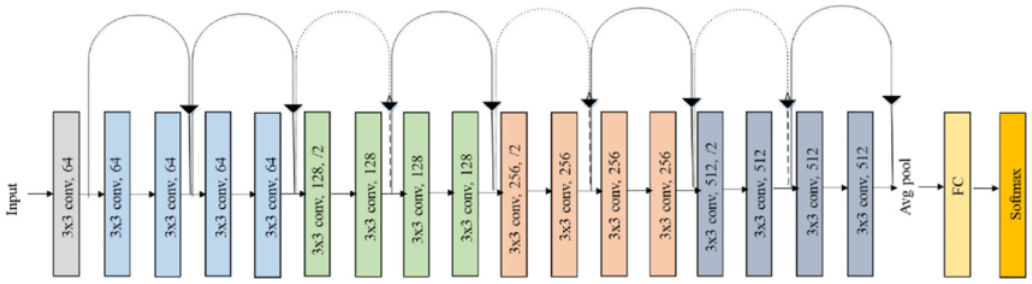

Fig. 2: Architecture of ResNet-18 (13)

2.2.3 ResNet-18 with Transfer Learning

We use the 18-layer version of ResNet once again, except, now we use pre-trained weights. This model is accessed through Torch Vision, which provides ResNet-18 pretrained using ImageNet.

ImageNet has been found to be a popular dataset for transfer learning, providing a wide variety of 'general knowledge' on image classification to models. This model will provide insight into whether preexisting learning of diverse shapes, objects, colours can improve animal classification. (14) We, once again, modify the fully-connected layer to output scores for the 10 labels in our dataset.

2.2.4 Training

All the implementation was completed using Pytorch. Each model was trained and tested on Google Collab Prob which provides a Intel(R) Xeon(R) CPU @ 2.30GHz, NVIDIA T4 Tensor Core GPU with 12 GBs of RAM.

Each of these networks received, as inputs, images of shape 3x224x224 after the preprocessing. They output scores for all 10 labels in ANIMAL-10N. We used a batch size of 32 to load the data. All models were trained for 20 epochs. Stochastic Gradient Descent (SDG) is used as the optimizer with a learning rate of 0.01 and a momentum of 0.1. Cross Entropy Loss was used as the quantitative measure for our three model's abilities to achieve accurate classifications.

$$CrossEntropy(y, \hat{y}) = -\sum_{c=1}^{C} y_c \log(\hat{y}_c)$$

Where y is the ground-truth vector, $\hat{y}$ is the predicted labels calculated by the model, y_c is the ground truth for class C and $\hat{y}_c$ is the prediction for class C.

3 Experiments

3.1 Evaluation Metrics

Our evaluation metrics include top-1 identification accuracy (the ratio of animals that were classified correctly), recall, precision and an F1 score which is a weighted average of precision and recall. The metrics are calculated as follows, where TP is $TruePositives$, TN is $TrueNegatives$, FP is $FalsePositives$ and FN is $FalseNegatives$.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall}$$

Further analysis of the models is conducted by evaluating Confusion Matrices and qualitatively observing Saliency Maps.

3.2 Quantitative Analysis

The Baseline CNN does experience a consistent increase in accuracy but the trajectory is relatively slow, indicated by the flatness of the plot. The accuracy also plateaus around epoch 16, and scheduling learning rate updates at this point may be helpful. The accuracy peaks at 56.09% at the last epoch. Taking a look at the plot of ResNet-18 trained from scratch, it has lower accuracy at the beginning compared to the Baseline CNN. However, this neural network has a faster increase in accuracy than the baseline, with a consistent rate of increase across all epochs. It is possible that ResNet's architecture, which allows gradient flow quickly to deep layers has enabled faster learning and subsequently, a steeper accuracy growth rate. The model has the highest accuracy of 59.03% during the second last epoch. Finally, the pretrained ResNet-18 model proves to be the most accurate. It begins at a high accuracy of 68.78% and grows to reach 87.62% at its highest. However, the accuracy growth seems slow in this model, especially after epoch 16.

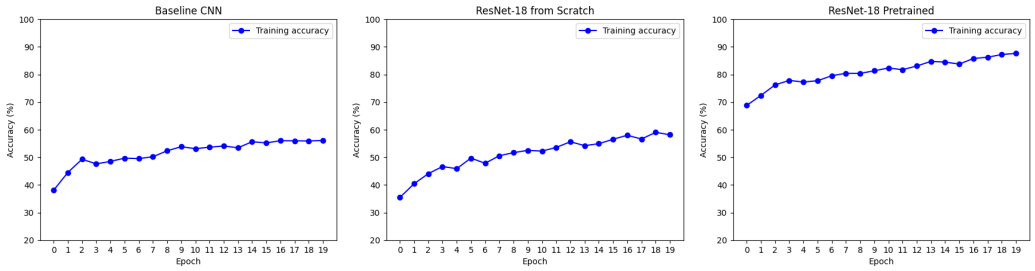

Fig. 3: Training Accuracy vs. Epoch Plots for all three models.

Overall, the training accuracy of all models grows fast up until epoch 3, which can be observed by the steep slopes in Figure 3, indicating that all neural networks are learning animal identification quickly, initially. After epoch 3, all models have a slight drop in accuracy, from where the accuracy growth flattens. Furthermore, none of the networks experience a significant drop in training accuracy within these epochs, which shows that overfit has been prevented. Additionally, ResNet-18 trained from scratch only has a 3% higher training accuracy than the baseline, at their peaks. It seems as though the deeper network of ResNet-18 does not benefit the task of animal identification significantly more. Although, the deeper network does enable faster learning as is indicative by the steeper slop of the ResNet with no transfer learning. Undoubtedly, ResNet-18 pretrained performs the best, with a starting accuracy of 68.78% surpassing the peaks of both models trained from scratch. It is clear that transfer learning of objects, shaped and colours from the ImageNet dataset caused great improvements in animal classification abilities.

Table 2 showcases the various accuracy evaluation metrics of our three neural networks. Pretrained ResNet-18 outperforms the other two models in all metrics by considerable margins, confirming the usefuleness of transfer learning for animal identification. Its top-1 classification training and testing accuracies beat both models trained from scratch by around 30%. ResNet-18 trained from scratch has a precision just 3% higher than our Baseline CNN. This indicates that the ResNet makes fewer incorrect positive predictions. Both of the models without transfer learning have the same recall value, below the halfway value, indicating that they output a high amount of false negatives. The F1 scores of the models without transfer learning are, once again, very close showing that a deeper architecture did not lead to considerable improvement in animal identification.

It is also important to note that the F1, precision and recall values for the Baseline CNN and ResNet-18 trained from scratch do not track closely with their classification accuracies. This is possibly a result of the class imbalance within the ANIMAL-10N dataset. It can be hypothesised that the higher accuracies of the models trained from scratch may be because they can predict majority classes well, but perform poorly on minority classes. This could be avoided by taking the same number of images from each of the 10 classes. On the other hand, all top-1 classification metrics of the pretrained network are close in value, and we can consider this model generally reliable for accuracy across all labels.

Model	Train Acc	Test Acc	F1	Precision	Recall
Baseline CNN	0.56	0.55	0.42	0.39	0.47
ResNet-18 from Scratch	0.58	0.61	0.45	0.43	0.47
ResNet-18 Pretrained	0.87	0.85	0.83	0.83	0.83

Tab. 2: Top-1 Classification values for all three models.

Now, we analyze the Confusion Matrix of our best performing model, pretrained ResNet-18. It can be observed that most diagonal boxes are white or yellow demonstrating that this model makes accurate predictions frequently. Since our dataset, ANIMAL-10N, is loaded with 5 confusing pairs of similar animals, this will give insight into which classifications our model is able to make easily, and which it struggles with. The Confusion Matrix was created using Torchnet and can be viewed in Figure 4, with the classes indexed as follow: 0: lynx, 1: guinea pig, 2: jaguar, 3: cat, 4: hamster, 5: cheetah, 6: coyote, 7: chimpanzee, 8: wolf, 9: orangutan.

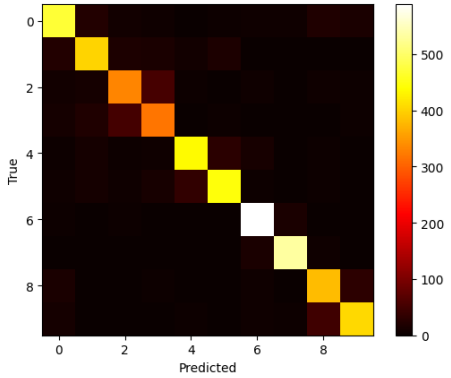

Fig. 4: Pretrained ResNet-18 Confusion Matrix.

The network has the most difficulty in classifying jaguars and cats, and tends to confuse one for another. This is an unexpected behaviour considering that they are not within the same confusing pair. However, it is possible that their similar colours and facial qualities considering that they are both felines may be the features that the model is learning faster than other distinguishing features. It is also noteable that cats have the most images to the label in the dataset comprising 10.93% of ANIMAL-10N, and so it is can be inferred that the increased data for this label is not necessarily introducing a bias. The dataset may not have enough diversity for cats and jaguars to be able to accurately identify them. In addition, it appears that the model is very easily able to identify coyotes, with almost perfect accuracy. This might be because they have easily distinguishing features, the dataset provides a wide variety of coyote images to learn from or overfitting towards coyotes may be occurring.

3.3 Qualitative Analysis

We will now observe visualizations of our best neural network's behaviour to better understand its features and representations. In Figure 5 you can view the Saliency Maps for 3 images from the ANIMAL-10N testing set produced by our model, ResNet-18 with transfer learning. Saliency maps reveal which pixels in the input impact the predicted score the most.

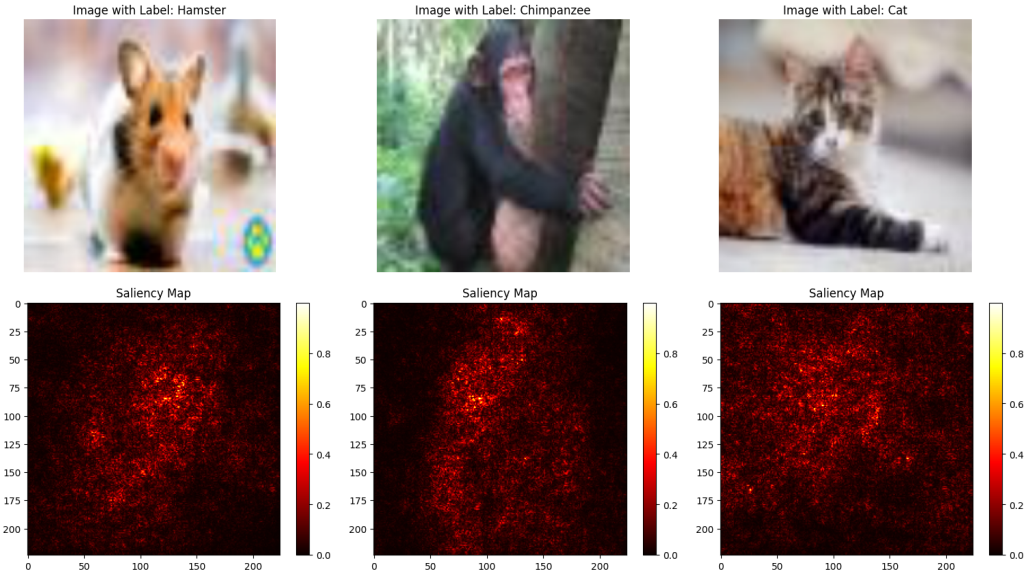

Fig. 5: Saliency Maps of 3 images from ANIMAL-10N Testing Set from Pretrained ResNet-18.

It can be seen that the model is able to isolate the part of the image with the animal quite successfully. This is evident in the image of the hamster and chimpanzee, however, the model seems to confuse

the background with the cats body. This visualization might be an indication of why the poor accuracy of cats was revealed in the Confusion Matrix, previously. Furthermore, the pixels close to the animals face are bright red and can be considered highly influential on the classification. This indicates that, in addition to the shape and body of the animal, the model is utilizing facial features for identification. However, it can be observed that the ears of all animals shown do not contribute much to the classification. From the hamster and cat Saliency maps, it is evident that the area around the animal's eyes and nose are critical to the scores our model outputs. Furthermore, the animals hands and paws do not light up much in the Saliency Map, demonstrating that the neural network does not learn much about the classification from the animal's limbs.

4 Conclusion & Future Work

From the experiments, is it clear the ResNet-18 with transfer learning is superior to other networks that do not use transfer learning. A key takeaway of this project is that transfer learning is a powerful tool to achieve accurate animal classification, and ResNet-18 pretrained specifically can achieve high-levels of accuracy in such image classification tasks even with minimal training. Furthermore, we saw that there wasn't much of a difference in top-1 classification accuracy, precision and recall between the models trained from scratch. Therefore, it can be concluded that the depth of the architecture is not a major contributor to classification capabilites, but transfer learning is extremely impactful. Furthermore, pretrained ResNet-18 has little confusion in classifying animals from this dataset, is able to isolate the animal from the background effectively and uses animal facial features to determine its output scores.

The extent of the research was limited by the amount of computational power that we had access to during the model's training and testing. Future work could use larger animal datasets and train the models for more iterations to attempt to reach the highest accuracy possible. Furthermore, the large difference between the precision and recall with the accuracies of the models trained from scratch are possibly indicative of declassification due to dataset imbalances. Future training could use an equal number of images per label to avoid this issue. Seeing the success that transfer learning had, it would be worth exploring different pretrained architectures to explore which performs the best at animal classification.

References

[1] R. Almond, M. Grooten, and T. Petersen, eds., *Living Planet Report 2020 - Bending the Curve of Biodiversity Loss*. Gland, Switzerland: World Wildlife Fund, 2020.

[2] G. Komarasamy, M. Manish, V. Dheemanth, D. Dhar, and M. Bhattacharjee, "Automation of animal classification using deep learning," in *International Conference on Intelligent Emerging Methods of Artificial Intelligence & Cloud Computing* (F. P. García Márquez, ed.), (Cham), pp. 419–427, Springer International Publishing, 2022.

[3] "Camera traps." `https://www.worldwildlife.org/initiatives/camera-traps`. Accessed: 2023-04-13.

[4] X. Yu, J. Wang, R. Kays, P. A. Jansen, T. Wang, and T. Huang, "Automated identification of animal species in camera trap images," *EURASIP Journal on Image and Video Processing*, vol. 2013, no. 1, p. 52, 2013.

[5] G. Chen, T. Han, Z. He, R. Kays, and T. Forrester, "Deep convolutional neural network based species recognition for wild animal monitoring," *2014 IEEE International Conference on Image Processing, ICIP 2014*, pp. 858–862, 01 2015.

[6] G. Chen, T. Han, Z. He, R. Kays, and T. Forrester, "Deep convolutional neural network based species recognition for wild animal monitoring," *2014 IEEE International Conference on Image Processing, ICIP 2014*, pp. 858–862, 01 2015.

[7] T. Battu and D. S. Reddy Lakshmi, "Animal image identification and classification using deep neural networks techniques," *Measurement: Sensors*, vol. 25, p. 100611, 2023.

[8] R. Parmar, "Demystifying convolutional neural networks," Oct 2018.

[9] N. Viswanathan and Stanford, "Artist identification with convolutional neural networks," 2017.

[10] "Animal (animal10n) dataset." `https://docs.activeloop.ai/v/v2.8.6/datasets/animal-animal10n-dataset`. Accessed: 2023-04-13.

[11] H. Song, M. Kim, and J.-G. Lee, "SELFIE: Refurbishing unclean samples for robust deep learning," in *ICML*, 2019.

[12] Timothy, "[discussion] why normalise according to imagenet mean and std dev for transfer learning?," Mar 2021.

[13] F. Ramzan, M. U. Khan, A. Rehmat, S. Iqbal, T. Saba, A. Rehman, and Z. Mehmood, "A deep learning approach for automated diagnosis and multi-class classification of alzheimer's disease stages using resting-state fmri and residual neural networks," *Journal of Medical Systems*, vol. 44, 12 2019.

[14] M. Huh, P. Agrawal, and A. A. Efros, "What makes imagenet good for transfer learning?," *CoRR*, vol. abs/1608.08614, 2016.