Multi-Class Mammal Image Classification

A Deep Learning Approach for Wildlife Monitoring

Author:

Anahita Nouri

Student ID: VR509464

Course: Computer Vision & Deep Learning
Master's Degree in Artificial Intelligence
University of Verona

Date:

February 2, 2026

Abstract

Biodiversity is declining quickly, and conservation depends on reliable information about where species occur. Camera traps can collect millions of images, but the sheer volume often leads to "data overload." Identifying species such as lions or elephants by hand in these large datasets is not practical at this scale. This project aims to reduce this bottleneck by creating a practical Computer Vision tool for conservationists.

This study examined animal identification using the **"Animal-5-Mammal"** dataset and compared two different methods. First, I built a neural network from scratch to assess its ability to learn from unprocessed data. Second, I applied Transfer Learning with **ResNet50V2** to examine how using pre-trained weights influenced the results. To check that the models were not simply memorizing pixel patterns, I subjected them to "stress testing" using dynamic data augmentation, such as random rotations and flips.

The results were encouraging. The custom model reached an accuracy of **92.55%**, suggesting that lightweight architectures can still perform well. In comparison, the ResNet50V2 model performed best, achieving **96.40%**. These findings indicate that Deep Learning is robust enough to support reliable automation in wildlife monitoring tasks.

Contents

1 Introduction

1.1 Context and Motivation

As global biodiversity loss speeds up, conservation relies on reliable wildlife monitoring to track populations precisely. Camera traps have transformed this area of research by automatically recording millions of images. However, their high data output has created a key problem: manually reviewing millions of photos within a reasonable time frame is not feasible.

Computer-vision-based automated mammal classification provides a fast, consistent, and scalable approach to identify species from images. Rapid image-based species identification supports real-time monitoring and data analysis. This project tackles fine-grained classification tasks where the model must separate biologically similar groups (e.g., different four-legged mammals) under varied environmental conditions.

1.2 Project Objectives

The primary goal of this study is to evaluate the effectiveness of Deep Learning in identifying mammal species. The specific objectives are:

1. **Baseline Development:** Design and train a custom CNN to establish a performance benchmark for the specific dataset.
2. **Transfer Learning Evaluation:** Implement the state-of-the-art ResNet50V2 architecture to assess the benefits of using pre-learned features from the ImageNet domain.
3. **Robustness Analysis:** Investigate the impact of data augmentation and regularization techniques (Early Stopping, Dropout) on model generalization.
4. **Error Analysis:** Perform a qualitative review of misclassified images to understand the limitations of current models in wildlife scenarios.

2 Theoretical Background

2.1 Convolutional Neural Networks (CNNs)

Convolutional Neural Networks have become the de facto standard for image analysis. Unlike traditional machine learning, which relies on handcrafted features, CNNs learn hierarchical feature representations directly from raw pixel data.

- **Convolutional Layers:** These layers apply learnable filters (kernels) to the input image. Early layers typically detect simple structures like edges and textures, while deeper layers combine these to recognize complex shapes like ears, snouts, or fur patterns.
- **Pooling Layers:** Operations like Max Pooling reduce the spatial dimensions of feature maps, providing translation invariance and reducing computational cost.

2.2 Transfer Learning and ResNet

Wildlife monitoring often lacks large species-labeled datasets because manual labeling is slow, costly, and often impractical in the field. Transfer Learning solves this by adapting a model pre-trained on broad, general data to a specified task.

ResNet50V2 is a residual neural network designed to train deep models by adding "skip paths" that keep information and gradients flowing across many layers. Deep models are powerful, but as networks get deeper, gradients may vanish or explode. This instability, known as the vanishing gradient problem, slows learning and lowers accuracy. ResNet addresses this by adding skip links so signals can bypass layers, allowing gradients to pass through the network without being blocked. The "V2" variant uses pre-activation in the weight layers to help keep training stable. By freezing the early layers of ResNet50V2, we reuse its learned features and train only the final classifier for our chosen mammals.

3 Methodology

3.1 Dataset Preparation

The study utilizes the **Animal-5-Mammal** dataset sourced from Kaggle. This dataset is a curated collection of images specifically focused on five mammal classes: **Horse, Lion,**

Dog, Elephant, and Cat.

- **Preprocessing:** All images were resized to 224×224 pixels to match the input requirements of the ResNet architecture. Pixel values were normalized to the range $[0, 1]$ to facilitate gradient descent.
- **Data Splitting:** The dataset was divided into a **Training Set (80%)** for model optimization and a **Validation Set (20%)** for performance evaluation.

3.2 Data Augmentation Strategy

Wild animals appear in diverse poses, lighting, and backgrounds. To simulate this variability and prevent the model from memorizing the training data, we implemented an on-the-fly augmentation pipeline using 'ImageDataGenerator':

- **Rotation:** $\pm 20°$ to account for camera tilt.
- **Shear & Zoom:** Up to 20% intensity to simulate distance variations.
- **Horizontal Flip:** To ensure the model recognizes animals regardless of direction.

3.3 Model Architectures

3.3.1 1. Custom CNN (Baseline)

We designed a lightweight CNN with four convolutional blocks.

- **Structure:** Each block contains a 'Conv2D' layer (filters: 32, 64, 128, 256) followed by 'ReLU' activation and 'MaxPooling'.
- **Classifier:** A Dense layer of 512 units with a Dropout rate of 0.5 to prevent overfitting, followed by a Softmax output layer.
- **Purpose:** To test how well a network can learn mammal features from scratch with limited data.

3.3.2 2. ResNet50V2 (Transfer Learning)

- **Base:** The ResNet50V2 model (pre-trained on ImageNet) was instantiated with the top layers removed.

- **Configuration:** The base layers were set to 'trainable=False' (frozen).
- **Head:** A 'GlobalAveragePooling2D' layer was added to reduce feature maps, followed by a Dense layer (512 units) and the final Softmax classification layer.

3.4 Evaluation Metrics

To rigorously assess the performance of our models, we utilize standard metrics derived from the Confusion Matrix elements: True Positives (TP), False Positives (FP), and False Negatives (FN).

- **Precision:** Quantifies the accuracy of positive predictions. It is crucial when the cost of False Positives is high.

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

- **Recall (Sensitivity):** Measures the model's ability to detect all positive instances. It is vital for monitoring tasks where missing an animal (False Negative) is undesirable.

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

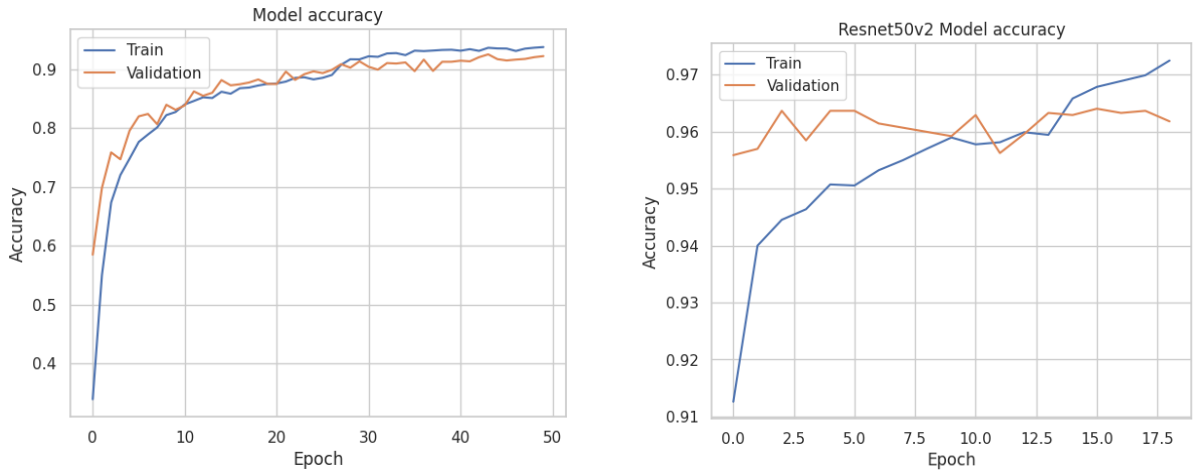
- **F1-Score:** The harmonic mean of Precision and Recall. This metric provides a balanced view of performance, especially when handling class imbalances.

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{3}$$

4 Experimental Results

4.1 Training Dynamics

Both models were trained using the **Adam** optimizer and **Categorical Cross-Entropy** loss. We employed an **Early Stopping** callback (patience=10) to halt training automatically when validation performance plateaued.



(a) Custom CNN Accuracy (b) ResNet50V2 Accuracy

Figure 1: Comparison of training accuracy curves. The ResNet model (right) converges significantly faster.

The learning curves indicate that both models converged effectively. The ResNet50V2 model, benefiting from pre-trained weights, reached its performance plateau significantly faster than the Custom CNN.

4.2 Quantitative Evaluation

The final comparison of the best validation accuracy achieved by each model is presented below:

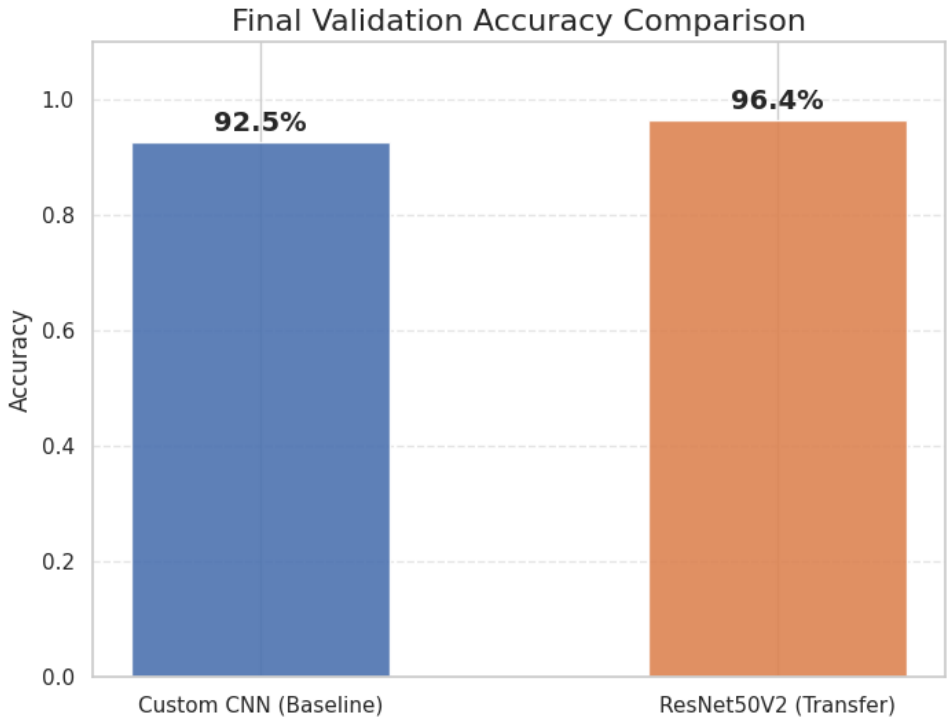


Figure 2: Validation Accuracy: Custom CNN vs ResNet50V2

The results were impressive for both architectures:

- **Custom CNN (Baseline):** Achieved a best accuracy of **92.55%**. This exceptionally high score for a baseline model indicates that the 4-block architecture and data augmentation pipeline were highly effective for this dataset.
- **ResNet50V2 (Transfer):** Achieved a best accuracy of **96.40%**. As hypothesized, the transfer learning model outperformed the baseline, demonstrating superior reliability and feature extraction capabilities.

Detailed Metrics Table:

Model	Accuracy	Precision	Recall	F1-Score
Custom CNN	92.55%	0.93	0.92	0.92
ResNet50V2	**96.40%**	**0.96**	**0.96**	**0.96**

Table 1: Performance comparison across all metrics. The F1-Score confirms that ResNet50V2 offers superior balance and reliability.

4.3 Qualitative Error Analysis

To better understand the limitations of our "Wildlife Monitor," we visualized the specific instances where the model failed.

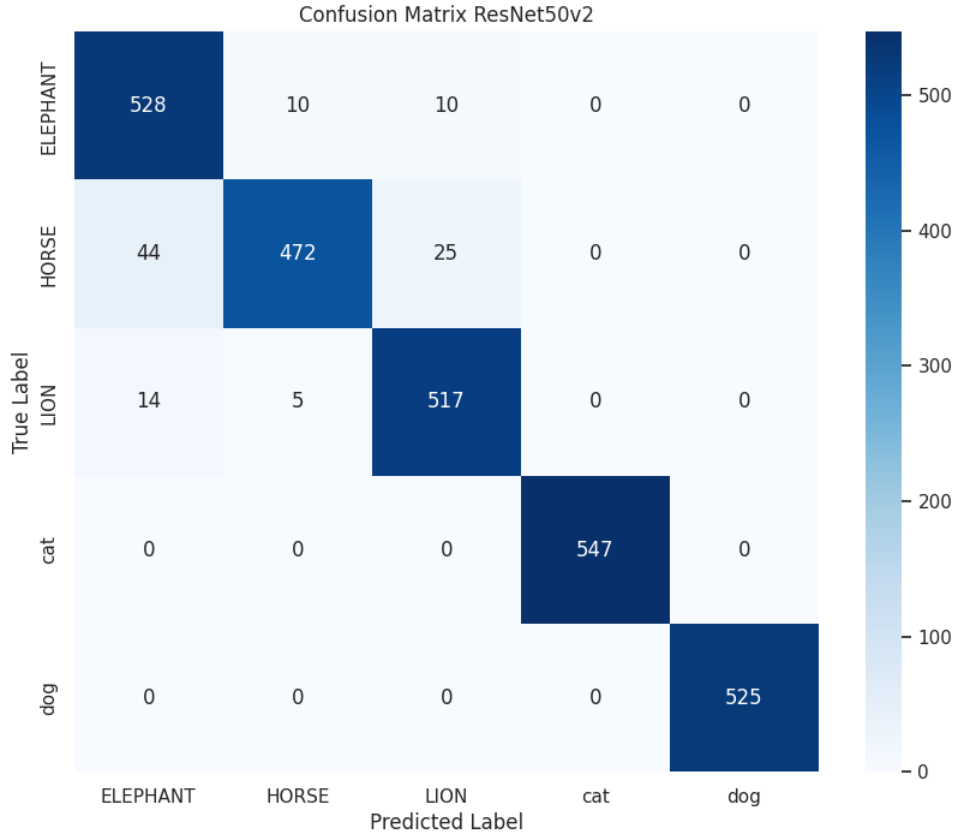


Figure 3: Confusion Matrix for the Best Model (ResNet50V2).

The Confusion Matrix reveals that:

1. **Distinct Silhouettes:** Large animals like *Elephants* were classified with near-perfect accuracy.
2. **Visual Similarity:** The highest error rates occurred between *Dogs* and *Cats*, likely due to similar sizes and domestic backgrounds in the dataset.
3. **Environmental Noise:** Misclassified images often featured cluttered backgrounds (e.g., high grass) that partially occluded the animal.

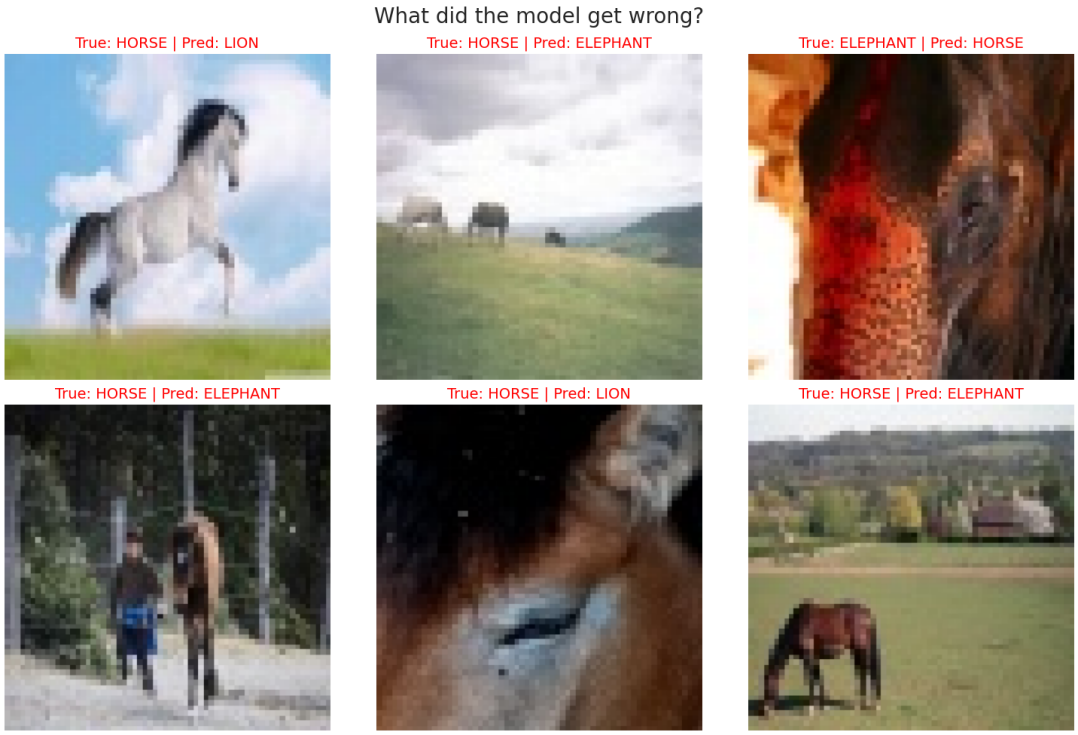


Figure 4: Visualizing Prediction Errors: True Label vs. Predicted Label.

5 Discussion and Conclusion

5.1 Discussion

The experiment yielded stronger results than anticipated. The fact that the Custom CNN reached 92% accuracy suggests that the "Animal-5-Mammal" dataset has distinct, learnable features that do not strictly require deep residual networks to identify. However, the ResNet50V2 model's ability to push performance to 96% confirms the value of Transfer Learning for achieving production-level reliability.

From an ecological perspective, an accuracy of 96% is viable for automated filtering of camera trap data. The remaining errors (mostly due to occlusion or extreme lighting) could be mitigated by implementing an Object Detection stage (e.g., YOLO) to localize and crop the animal before classification.

5.2 Conclusion

This project successfully demonstrated the development of a high-performance automated mammal classification system. By comparing a baseline CNN with a Transfer Learning

approach, we highlighted that while modern CNN architectures are powerful on their own, pre-trained models offer the final edge in accuracy and convergence speed. The final ResNet50V2 model is a robust candidate for deployment in wildlife monitoring applications.

Declaration of Generative AI Use

During the preparation of this work, the author used Large Language Models (Gemini/ChatGPT) to assist in refining the English language, formatting the LaTeX code, and debugging Python scripts. All technical concepts, experimental designs, and final results were verified and authored by the student.

References

[1] He, K., Zhang, X., Ren, S., & Sun, J. (2016). *Deep Residual Learning for Image Recognition*. Proceedings of the IEEE conference on computer vision and pattern recognition.

[2] Simonyan, K., & Zisserman, A. (2015). *Very Deep Convolutional Networks for Large-Scale Image Recognition*. ICLR 2015.

[3] Norouzzadeh, M. S., et al. (2018). *Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning*. Proceedings of the National Academy of Sciences.

[4] Kaggle Dataset. *Animal-5-Mammal*. `https://www.kaggle.com/input/animal-5-mammal`.