Animal Posture Classification with Transfer Learning

Chen Qiao, Johann Verolet

Abstract

This project proposes an Animal Posture Classification System leveraging transfer learning techniques. The core objective is to enhance animal behavior analysis, crucial for wildlife preservation, agricultural health, and pet care, by automating the classification of animal postures. Addressing the challenge of limited available datasets tailored for this purpose, our approach involves refining the ResNet50 and Vision Transformer (ViT) model, pre-trained on ImageNet, with annotated datasets. This project also explores other solutions like Manifold Mixup and the utilization of synthetic data generated by stable diffusion model to overcome data scarcity. This multifaceted strategy aims to yield a more accurate and efficient system for animal posture classification. Experimental results demonstrate that transfer learning significantly outperforms models trained from scratch. Further improvements are achieved through the application of Manifold Mixup techniques and the integration of synthetic data.

1 Introduction

The classification of animal postures plays a pivotal role in understanding animal behavior, which is vital for wildlife conservation, agricultural health, and pet care. Traditional methods for this task often involve manual monitoring, which is not only time-consuming but also resource-intensive. Our project introduces a solution to this challenge: an Animal Posture Classification System utilizing state-of-the-art AI models and transfer learning techniques.

A major challenge in animal posture classification is the small size of available labeled datasets that can be used for training, since most available image datasets focus on animal category classification or keypoint detection. To bridge this gap, we utilize transfer learning strategies, employing models such as ResNet50[1] and ViT[2], both initially pre-trained on the ImageNet dataset. These models are subsequently fine-tuned using available annotated datasets. Our results demonstrate that transfer learning substantially enhances classification accuracy compared to training models from scratch, with improvements of 16.10% for ResNet50 and 22.53% for ViT, respectively.

To further enhance model performance, we incorporate two techniques that are applied concurrently with transfer learning: Manifold Mixup[3] and synthetic data generation via a diffusion model[4]. Manifold Mixup, a data augmentation method, strategically improves model robustness by interpolating features and labels from different training samples. The diffusion model, renowned for its capacity to generate high-fidelity synthetic images, is deployed to augment our training dataset. This strategic addition of synthetic data effectively mitigates the issue of data imbalance and expands the diversity of our dataset. The application of Manifold Mixup increased the classification accuracy by 3.49% for the ResNet50 model and 1.85% for the ViT compared to using only transfer learning. Similarly, the integration of synthetic data improved accuracy by 1.06% for ResNet50 and 1.99% for ViT relative to results achieved with transfer learning alone. These improvements underscore the effectiveness of combining advanced data augmentation and synthetic data techniques to bolster machine learning model performance.

2 Related Work

Recent advances in Transfer Learning have shown significant potential across various domains by leveraging pre-trained models to boost performance on new, but related tasks. Zhuang et al. (2021) provide a comprehensive review[5], emphasizing that transfer learning minimizes reliance on large domain-specific datasets and enhances learning efficiency through knowledge transfer from related source domains. This aligns with our method, where we utilize transfer learning techniques with models such as ResNet50 and ViT, both pre-trained on ImageNet, to address the challenge of limited labeled data in animal posture classification.

The foundational principles of ResNet (He et al., 2015)[1] introduce a residual learning framework to mitigate the degradation problem in deep neural networks. Employing deeper architectures, ResNet demonstrates that increased network depth does not necessarily lead to higher error rates, provided that each layer set learns a residual mapping, a concept we leverage in our work. While the Vision Transformer (ViT) shifts the paradigm by applying self-attention mechanisms across image patches, treating image classification similarly to sequence processing in natural language tasks. This attention-based approach allows ViT to consider the global context of an image, leading to models that excel in understanding complex features without being confined by the sequential locality typical of convolutional neural networks (Dosovitskiy et al., 2020)[2]. In our study, we utilized both ResNet50 and ViT as the model architectures, enabling the utilization of advanced feature extraction techniques essential for accurate posture classification in diverse animal species.

The recent study by Verma et al. (2019)[3] on Manifold Mixup, and the application of the Stable Diffusion Model (Rombach et al., 2022)[5] for synthesizing data, presents significant advancements in the enhancement of deep learning models. Manifold Mixup notably advances model generalization by interpolating between hidden representations, yielding smoother decision boundaries and more conservative predictions. Additionally, the generation of synthetic data via the Stable Diffusion Model, provides a strategic advantage in addressing the challenges of imbalanced and limited training data. The incorporation of such synthetic data not only enlarges the dataset but also brings in a greater degree of variation.

3 Problem definition

The task involves developing a method for classifying the posture of animals using machine learning techniques. The primary challenge is the limited availability of annotated datasets, which are crucial for training and validating the classification models. The goal is to achieve higher accuracy through advanced techniques such as transfer learning, Manifold Mixup, and synthetic data generation. These methods aim to enhance the dataset's diversity and volume, thereby improving the robustness and performance of the classification models.

4 Methodology

4.1 Data Preparation and Cleaning

For this study, we selected the dataset entitled "Animals with Skeleton Key Points and Action Mark"[6] available on Kaggle as our primary data source. This dataset is composed of 967 images, featuring six different types of animals, each annotated with corresponding action classes which serve as posture labels. The posture categories defined within this dataset include lying, sitting, running, and standing/walking. To ensure the integrity and quality of our analysis, we conducted a rigorous data cleaning process. This process entailed a meticulous examination of the dataset, followed by the removal of images that were deemed unsuitable for our study due to quality issues or incorrect labeling. Following this cleaning process, the dataset was reduced to 819 images. The distribution of images across each posture category is detailed in Table 1.

Category	Number of Images
Lying	88
Sitting	67
Running	70
Standing/Walking	592
Total	819

Table 1: Distribution of Images After Data Cleaning

4.2 Model Architectures

In our Animal Posture Classification task, we utilize two distinct models: ResNet50 and the Vision Transformer (ViT), each selected for their unique architectural strengths.

4.2.1 ResNet50

The ResNet50 is chosen for its demonstrated effectiveness in training deeper neural networks[1], which is especially beneficial for our small dataset. Its innovative residual blocks help overcome the vanishing gradient problem, ensuring it performs well even with limited data. As illustrated in Figure 1, The input passes through a series of convolutional layers, each followed by a ReLU activation function. The feature map output from these layers is combined with the input signal, creating a "shortcut" path. Another ReLU activation function follows this skip connection. Figure 2 delineates the architecture of ResNet50. It is segmented into multiple stages, with Stage 0 dedicated to the initial processing of the image for low-level feature extraction. Subsequent stages, 1 through 4, comprise a series of residual blocks designed to incrementally extract higher-level features. This hierarchically structured approach allows for a comprehensive analysis of the visual data, capturing details essential for distinguishing a variety of animal postures. By strategically deepening the network through these stages, the model pays better attention to important features without using too much computing power.

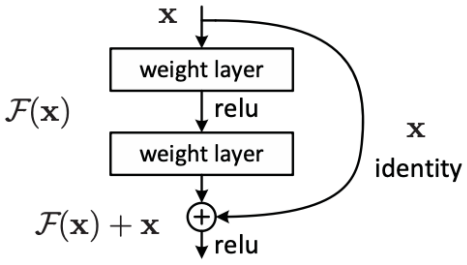

Figure 1: Residual Block

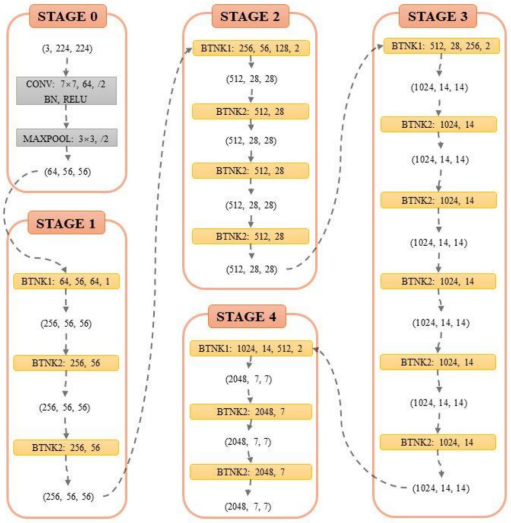

Figure 2: Network Architecture of ResNet50

4.2.2 Vision Transformer

Vision Transformer (ViT) is employed for its pioneering approach to image analysis, which abandons the local processing paradigm of CNNs in favor of global information aggregation using self-attention mechanisms. As detailed in Figure 3, it harnesses a Transformer Encoder that refines image analysis. This Encoder, through its multi-headed self-attention mechanism, attentively assesses

relationships across image patches to grasp complex visual information. Integral to its design are residual connections and normalization layers, which preserve gradient flow and feature scale across deep networks. This facilitates stable learning and nuanced feature discernment, essential for recognizing diverse animal postures with precision. Supplementing this, Figure 4 illustrates the initial stage of ViT's processing pipeline, where an image is divided into patches. These patches are then linearly projected and combined with positional embeddings to form a sequence of vectors, representing both the visual content and positional information. This encoded sequence is what the Transformer Encoder processes. The global contextual understanding gleaned from this methodology is critical, allowing the MLP head at the output to classify animal postures.

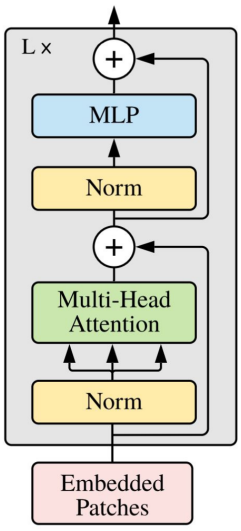

Figure 3: Transformer Encoder

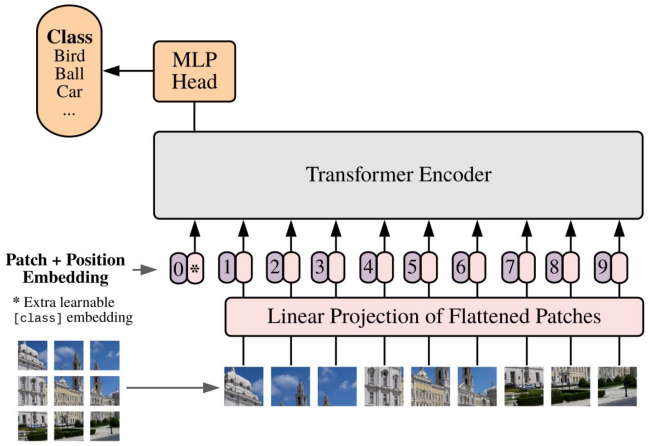

Figure 4: ViT Architecture Overview

4.3 Transfer Learning

We integrate the two models introduced earlier: ResNet50 and the ViT, both pretrained on the ImageNet dataset. This pretraining endows the models with a broad understanding of visual features applicable across various domains. In our experiments, we conduct two types of transfer learning approaches. The first involves freezing the pretrained layers, rendering them non-trainable to preserve the learned features, and allowing the models to function as fixed feature extractors. The second approach allows for fine-tuning, where the pretrained layers are set to be trainable, enabling the model to adapt these features to the nuances of our specific classification task.

To tailor the models to our classification needs, we add a custom classification head comprising a linear layer. The output layer is designed to classify four distinct classes. This additional layer effectively transforms the general feature representations learned in previous layers into predictions specific to our posture classification task.

By employing transfer learning, we capitalize on the knowledge learned from vast amounts of visual data and distill it for our specific task, significantly reducing the need for extensive animal posture datasets and computational resources, while providing higher accuracy in our classification results.

4.4 Data Augmentation through Manifold Mixup

We incorporate the technique of Manifold Mixup as an advanced form of data augmentation because of its ability to smooth the decision boundaries and provide a robust regularization effect, which is achieved by creating border regions in-between different classes within the data representation space. This process also tends to flatten the learned representations, i.e., the network learns to represent the input data in fewer directions of variance[3], leading to a more rapid decline in the singular values. As a result, the network tends to have better generalization ability.

Our implementation of Manifold Mixup follows the technique described in the study by Verma et al. (2019)[3]. During training, we first select a layer K within the network architecture. At this layer, two random mini-batches, denoted as (x, y) and (x', y') resulting in the transformed features $(g(x), y)$ and $(g(x'), y')$ are mixed up. To perform the Mixup, a random interpolation weight λ is generated uniformly from the interval [0, 1]. This weight and its complement $1 - \lambda$ are then used to mix both the latent representations and labels of the two mini-batches. Specifically, the mixed latent representation is computed as

$$Mix\ (g(x),\ g(x')) = \lambda \cdot g(x) + (1 - \lambda) \cdot g(x'),$$

and the mixed labels are computed as

$$Mix\ (y,\ y') = \lambda \cdot y + (1 - \lambda) \cdot y'$$

The resultant mixed representations and labels are subsequently used for further training, thereby encouraging the network to predict not only on the observed data but also on the interpolated data, enhancing the model's generalization capabilities.

By applying Manifold Mixup, we induce our network to predict consistently not only at the data points but also in their vicinity, hence encouraging the learning of more robust features.

4.5 Integrating Synthetic Data from Stable Diffusion Model

To address the challenge of limited and imbalanced dataset, we incorporate synthetic images generated by the pre-trained Stable Diffusion model (version 1.4). Stable Diffusion model is trained to generate images by learning the underlying distribution of vast and diverse datasets. By sampling from this learned distribution, we generate additional images that exhibit varied animal postures, enriching the representational diversity of our training data.

The integration process involves selecting specific prompts that describe a range of animal postures and feeding these into the Stable Diffusion system. We refine our prompts through testing, discovering that clear and specific prompts, such as "a realistic image of the full body of a white cat lying on a sofa," yield good results. The model generates new images from these prompts, effectively creating a multitude of images that help our system learn more generalized features. These synthesized images are subsequently reviewed and cleaned before being introduced into our training set. During this process, we removed unrealistic images, such as a horse with five legs or a cow in an improbable posture. Approximately 30% of the initial images remain after this process. Figure 5 presents a selection of synthetic images. Figure 6 presents pie charts that illustrate the makeup of our dataset before and after incorporating synthetic data. This figure demonstrates how the addition of synthetic data aids in addressing the issue of data imbalance.

Figure 5: Synthetic Data Generated by Stable Diffusion

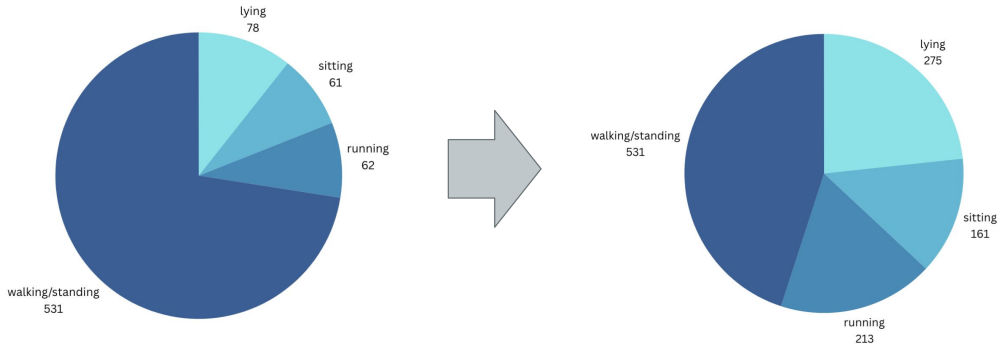

Figure 6: Training Set Before and After Integrating Synthetic Data

5 Experimental Results

5.1 Baseline Evaluation

To establish a baseline, we assessed the performance of models trained from scratch on our cleaned animal posture dataset without the use of transfer learning. We experimented with training the ResNet50 and ViT models, both with and without the implementation of data augmentation techniques. The data augmentation protocol included rotations up to 10 degrees, horizontal and vertical shifts of 10% of the image size, shear transformations that altered image shape by up to 10%, zoom adjustments of 10%, and horizontal flips. Our findings revealed that data augmentation enhanced the performance of the ResNet50 model, while the ViT model achieved higher accuracy without the use of data augmentation. Therefore, we reported the results with data augmentation for ResNet50 and results without data augmentation for ViT. These results are presented in Table 2. The reported accuracy, precision, recall, and F1 score represent the average results across three experimental runs.

	ViT	ResNet50
Lying Acc (%)	10.00	0.00
Sitting Acc (%)	0.00	0.00
Running Acc (%)	0.00	0.00
Stand/Walking Acc (%)	95.08	100.00
Overall Acc (%)	67.82	70.11
Precision (%)	22.16	17.53
Recall (%)	26.27	25.00
F1 Score (%)	24.04	20.61

Table 2: Performance of ResNet50 and ViT Models

5.2 Transfer Learning Impact

We applied transfer learning to the ResNet50 and ViT models using pretrained ImageNet weights, adapting them to our animal posture dataset. To tailor the models for our task, a classification head was added. We explored two methods: freezing all layers and fine-tuning the entire model. ViT showed improved performance with full fine-tuning, while ResNet50 excelled with frozen layers, Consequently, in Table 3, we present the results for ViT(TL) with fully fine-tuned layers and for ResNet50(TL) with frozen layers. ViT was tested without data augmentation, whereas ResNet50 incorporated it, enhancing accuracy. TL stands for Transfer Learning.

Table 3 details the comparison between models trained from scratch and those using transfer learning. Transfer learning significantly improved the models' ability to recognize diverse animal postures beyond just walking or standing. This analysis underscores the effectiveness of transfer learning in refining model performance on our task.

	ViT	ViT (TL)	ResNet50	ResNet50 (TL)
Lying Acc (%)	10.00	90.00	0.00	60.00
Sitting Acc (%)	0.00	87.50	0.00	87.50
Running Acc (%)	0.00	50.00	0.00	62.50
Stand/Walking Acc (%)	95.08	98.36	100.00	93.44
Overall Acc (%)	67.82	91.95	70.11	86.21
Precision (%)	22.16	90.38	17.53	77.53
Recall (%)	26.27	81.47	25.00	75.86
F1 Score (%)	24.04	85.69	20.61	75.38

Table 3: Impact of Transfer Learning, TL stands for Transfer Learning

5.3 Manifold Mixup Impact

Following the transfer learning phase, we explored the application of Manifold Mixup as a data augmentation technique to further enhance model performance. This approach was evaluated by assessing its impact on accuracy, precision, recall, and F1 Score, and by comparing these metrics with the results obtained from the transfer learning implementation. The ResNet50 model included basic data augmentation, as detailed in Section 4.1, which contributed to improved accuracy. In contrast, the ViT model was tested without such augmentations because they did not yield any improvement.

Table 4 presents a detailed comparison between the models employing only transfer learning and those utilizing both transfer learning and Manifold Mixup. Although Manifold Mixup does not significantly address the issue of dataset imbalance—since it synthesizes new data points by mixing up existing ones rather than creating completely new examples—it still managed to enhance overall accuracy, precision, recall, and F1 Score by a decent margin. This demonstrates that while Manifold Mixup may not solve all challenges related to dataset characteristics, it effectively contributes to improving model robustness and performance in our animal posture classification system.

	ViT (TL)	ViT (TL + MM)	ResNet50 (TL)	ResNet50 (TL + MM)
Lying Acc (%)	90.00	100.00	60.00	88.89
Sitting Acc (%)	87.50	87.50	87.50	64.29
Running Acc (%)	50.00	50.00	62.50	57.14
Stand/Walking Acc (%)	98.36	98.36	93.44	96.64
Overall Acc (%)	91.95	93.10	86.21	89.70
Precision (%)	90.38	94.27	77.53	82.80
Recall (%)	81.47	83.97	75.86	76.74
F1 Score (%)	85.69	88.82	75.38	78.89

Table 4: Impact of Manifold Mixup, TF stands for Transfer Learning, MM stands for Manifold Mixup

5.4 Synthetic Data Integration and Comparative Analysis

In our pursuit to address the challenges posed by dataset imbalance, we experimented with the integration of synthetic images generated by stable diffusion model into our training dataset. This approach aimed to enrich the dataset with diverse examples and assess the potential benefits of synthetic data on model performance. We conducted comparative analyses between models trained on the original dataset and those trained on a combined set that included both original real-world data and synthetic images.

Table 5 presents the outcomes of this experiment, comparing model performance following transfer learning against performance after integrating both synthetic data and transfer learning. A detailed examination of the accuracies across different categories revealed benefits. Specifically, the integration of synthetic data effectively mitigated the issue of dataset imbalance by introducing a wider variety of data representations. This result underscores the utility of synthetic data in enhancing model robustness and addressing specific challenges such as class imbalance in training datasets.

	ViT (TL)	ViT (TL + SD)	ResNet50 (TL)	ResNet50 (TL + SD)
Lying Acc (%)	90.00	100.00	60.00	88.89
Sitting Acc (%)	87.50	87.50	87.50	71.43
Running Acc (%)	50.00	87.50	62.50	64.29
Stand/Walking Acc (%)	98.36	96.72	93.44	91.60
Overall Acc (%)	91.95	95.40	86.21	87.27
Precision (%)	90.38	92.29	77.53	79.05
Recall (%)	81.47	92.93	75.86	75.27
F1 Score (%)	85.69	92.61	75.38	76.81

Table 5: Impact of Synthetic Data, TL stands for Transfer learning, SD stands for Synthetic Data

6 Limitation and Future work

While we employed Stable Diffusion version 1.4 to generate synthetic data due to its wide adoption and robust community support, this choice brings inherent limitations. This version, though popular, may not encompass the latest advancements in generative model capabilities, which could limit the diversity and realism of the synthetic data generated.

Future iterations of our project will explore the use of newer versions of the model, such as Stable Diffusion version 2.1, which promises enhanced data generation with potentially greater variability and higher fidelity. By integrating these advancements, we aim to further address the challenges of data imbalance and improve the robustness and accuracy of our animal posture classification models.

7 Conclusion

This study demonstrates the effectiveness of transfer learning in animal posture classification, enhancing model accuracy using advanced architectures like ResNet50 and Vision Transformer, pre-trained on ImageNet. Techniques such as Manifold Mixup and synthetic data generation with Stable Diffusion have effectively mitigated the problem of limited and imbalanced datasets, improving model generalization. Our findings support the potential of integrating cutting-edge AI technologies to advance animal behavior analysis, crucial for applications in wildlife conservation and pet care. Future work will explore newer generative models and expand the dataset diversity, further refining the robustness and applicability of our classification system.

References

[1] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[2] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. & Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

[3] V. Verma, A. Lamb, C. Beckham, A. Najafi, I. Mitliagkas, A. Courville, D. Lopez-Paz, and Y. Bengio. Manifold mixup: Better representations by interpolating hidden states. In *International Conference on Machine Learning*, 2018.

[4] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *Conference on Computer Vision and Pattern Recognition*, 2021.

[5] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He. A comprehensive survey on transfer learning. In *Institute of Electrical and Electronics Engineers*, 2021.

[6] https://www.kaggle.com/datasets/egorovalexeyd/animals-with-skelet-key-points-and-action-mark.