CS 5974: Final Project Report
Animal Data and Weight Estimation

Authors
Amun Kharel (akharel@vt.edu)

Abstract

The weight of an animal serves as a pivotal indicator for assessing its development and overall health, providing crucial insights into growth rates, market weight, diet efficiency, energy balance, and overall health status. Effective monitoring of animal weight is imperative to prevent production losses, optimize feeding efficiency, enhance reproductive performance, and avert adverse health events in livestock.

Existing practices involve collecting animal weight data at specific stages in their productive cycle, such as birth, weaning, and finishing. However, for a comprehensive understanding of animal growth, consistent gathering of weight data throughout all seasons is necessary. This research leverages Deep Learning Techniques, encompassing Convolutional Neural Networks (CNNs), Vision Transformer, and Video Vision Transformer, to achieve accurate animal weight prediction. Various metrics are employed to evaluate and compare the performance of these models.

The proposed deep learning methods offer substantial advantages over previous biometric approaches, eliminating the need for intricate processing and modification of individual images to extract parameters for weight prediction. Results indicate that the Vision Transformer outperforms CNN, boasting a lower RMSE Score of 92.5 compared to CNN's 260.7. However, Video Vision Transformer falls short of expectations due to a limited amount of labeled video data. Our findings provide a comparative analysis of these deep learning methods, offering insights into future directions for advancing animal weight prediction methodologies.

1 Introduction

Variations in animal dimensions, including factors like body weight and conformation, play a crucial role as indicators of animal development and health in livestock production [1, 2, 3, 4]. Monitoring these characteristics can offer valuable insights into factors influencing growth rates, market weight, diet efficiency, energy balance, and the health status of animals. Effective management of weight gain or loss is crucial to avoid production losses, enhance feeding efficiency, improve reproductive performance, and prevent adverse health events in livestock [1, 2, 3, 4, 5, 6].

In practical terms, the current methods of data collection require manual handling of animals. In the case of cattle, for instance, the difficulties in obtaining regular weight measurements have led most operations to weigh animals only at key points in their productive cycle, such as birth, weaning, and finishing [1].

The scarcity of data poses a challenge for animal scientists, hindering a thorough comprehension of the authentic growth curves of animals. This limitation may lead to economic losses for farmers. Termed the 'phenotyping bottleneck,' this issue highlights the constraints in phenotyping activities, preventing a comprehensive characterization of animals at an individual level [1].

Precision livestock farming has witnessed the adoption of sensing technologies in recent years, aimed at capturing biometric changes in the dynamics of animal growth and body composition [1, 2, 3]. This state-of-the-art technology not only enhances output but also ensures long-term viability and welfare. Notably, computer vision technologies play a crucial role in expediting phenotyping efforts through

the provision of non-intrusive structural assessments with high temporal and spatial resolution.
In addition to offering two-dimensional data, depth sensor cameras can measure an animal's depth or height. Typically positioned overhead, these cameras are non-invasive and do not disrupt daily farm operations [1].
In the evaluation of body mass or structural features in cattle, acquiring top-view depth images may become a standard practice on farms. The decreasing cost of 3D depth sensor cameras (RGB-D cameras) provides farmers with a cost-effective alternative for monitoring their animals [1].
In this work, we will be using the top view RGB and depth images of cows to predict the weight of those animals using various deep learning techniques such as CNN, Vision Transformer, and Video Transformer. We will also provide with details on how we processed the images/videos of the cows into the state of arts models while providing the reasoning behind such processing. Finally, we will address limitations of our work and provide with directions to further improve this research. Being able to dynamically predict the weights of animals will aid us in the collection of animal weights during all seasons, which will provide us better modeling of animal growth curves. Such development will improve research on animal development, provide insights into factors affecting growth rates, health status of animals, etc.

2 Related work

2.1 Biometric measurement

In Previous work, the Biometric method uses four parameters (Width, Length, Height, and Volume) to get the weight of the animal [1].
The work [1] utilized OpenCV in Python to extract cows from the depth images. They defined the boundaries in the vertical direction as the fence rails. First, they cropped the image to remove the surrounding area while preserving the walk-through space. Then they converted the cropped image into a hue, saturation, and value (HSV) image. Using the HSV image, they transformed those pictures into black and white using a threshold value. They detected image contours from the thresholded image and retained the largest contour for the final frame result. To fill the empty sections within the retained contour, they applied morphological closing using square structural elements of size 10X10".
From the final transformered image, they measured four parameters, namely Width, Length, Height, and Volume.
They utilized all four parameters as predictors to construct regression models for predicting dairy cow body weights. In this study, they evaluated the performance of the Ordinary Least Squares (OLS) and Random Forest (RF) Regression Models. The model function was represented as y = f(X), where y denotes the predicted body weight in pounds, and X represents a combination of height, width, length, and volume.
They assessed their model's performance through two cross-validation approaches: time series forecasting and leave-several-animal-out. In the time series approach, they partitioned our dataset into training and testing sets using five different ratios based on time points: 90:10, 80:20, 70:30, 60:40, and 50:50. In the latter approach, they excluded several cows as the testing set and employed the remaining cows as training sets. Model evaluation utilized mean square error and Pearson correlation coefficients between the training and testing sets
Although the Biometric approach is a good starting point, each images in this process has to be carefully modified to get the four parameters. Also, this technique is unable to leverage the power of deep learning methods to automate this process in a simplified manner.

2.2 Deep Learning

Deep learning, a prominent subset of machine learning, emulates the functionality of the human brain. Coined from the intricate connections among the vast number of neurons in the human brain, deep learning is adept at executing complex tasks [7]. It facilitates the creation of multiple intricate prediction models and intricate neural networks with multi-hidden layers [8]. A key advantage of the deep learning approach lies in its elimination of the need for feature engineering, a common practice in traditional machine learning, leading to improved accuracy. By automatically identifying and combining important features, it accelerates the learning process [9]. This capability underscores the efficiency of deep learning in reducing the workload and time required to acquire knowledge about a specific problem. Consequently, these algorithms have garnered significant attention for addressing complex challenges in artificial intelligence, including natural language processing, spam detection,

and image classification [9].
In the realm of animal scientific studies, the adoption of deep learning-based computer vision systems emerges as a promising strategy for monitoring animal health and enhancing precise measurements of animal bodies through image analysis [10].

2.2.1 Convolutional Neural Network (CNN)

The Convolutional Neural Network (CNN) stands out as a prominent and widely utilized deep learning network, currently gaining significant attention. Its capability to handle vast amounts of data contributes to the increasing popularity of deep learning. A CNN is a mathematical construct comprising several essential components, including convolution, pooling, rectified linear unit (ReLU), and fully connected layers. Designed to process input images and automatically discern spatial hierarchies of features, CNN employs a filtering process wherein neurons connect only to neurons with identical weights that are in close proximity [11]. This distinctive characteristic sets CNN apart from other neural networks, simplifying the processing and comprehension of complex images.
This state-of-the-art methodology plays a crucial role in segmentation, feature extraction, object detection, and classification [11]. The history of CNN architectures dates back to the 1980s with the neocognitron, followed by LeNet-5 in 1989-1998 for handwritten digit recognition. Subsequent developments include AlexNet in 2012, ZFNet in 2013, VGGNet and GoogLeNet in 2014, and ResNet in 2015, all contributing to advancements in the field [12]. The applications of CNN in livestock have seen significant growth, with various models such as Faster R-CNN, YOLO, FCN, etc. [13].
While recent studies have optimized CNN-based computer vision systems for managing farm animals, there remains a notable gap in the use of RGB images to estimate the actual body weight of cows on the farm. This study aims to address this gap by predicting animal weight through the optimization of a CNN model.

2.2.2 Vision and Video Transformers

An intriguing alternative to Convolutional Neural Networks (CNNs) is the Vision Transformer (ViT), presenting a competitive approach. ViT involves extracting patches from images, utilizing them as input for a transformer model, and transforming them for classification tasks [14]. In recent years, ViT has emerged as a dominant force in image classification compared to CNNs, attributed to its uniform representation across all layers and the inclusion of more global information at lower layers. The original transformer, initially proposed by [15] for scaling natural language processing architectures, has quickly become a promising technique in various fields, including computer vision.
Although introduced relatively recently, ViT has demonstrated considerable success. In 2020, [14] adapted this technique to handle large volumes of data in image classification tasks, showcasing its effectiveness in measuring animal body weight through images captured on the farm.

3 Material and Methods

3.1 Animal Experiments

This study utilized a total of 12 Holstein animals, comprising 10 lactating cows and 2 dry cows, from the Dairy Complex at Kentland Farm (Virginia Tech, Blacksburg, VA). The Holsteins, approximately 2 years old, had an average of 190 ± 111 days in milk and weighed 665 ± 124 kg. The cows were housed in a free-stall barn, milked twice daily (for lactating cows), fed ad libitum once a day, and had free access to water. Data collection took place after cows exited the milking parlor from the 12 AM and 12 PM milking sessions, occurring daily for a consecutive 30 days [1].
For depth data collection, an Intel RealSense D435 depth sensor camera (Intel, Santa Clara, CA, USA) was employed in a 10 12s short video format. The camera provided 87 horizontal and 58 vertical fields of view and used two stereos to determine depth under ideal lighting conditions. Mounted in a heated container to maintain normal operating temperatures, the camera was positioned 2.95 meters above a one-way exit lane between the milking parlor and pen housing. This allowed for top views of cows walking underneath it in an unconstrained manner. The path was narrow, accommodating a single cow at a time, and fitted with a weight-activated door to prevent multiple cows from entering

simultaneously. A laptop connected to the depth-sensing camera using a USB 3.1 cable, and the camera utilized auto-exposure and auto-focus [1].

All videos underwent processing through rs-convert, an open-source program converting video files into images and CSV files per frame. The images included RGB and depth images, with different colors in-depth color images representing varying distances from the object to the camera. The corresponding CSV files contained the meter distance of each pixel [1].

3.2 Convolutional Neural Network

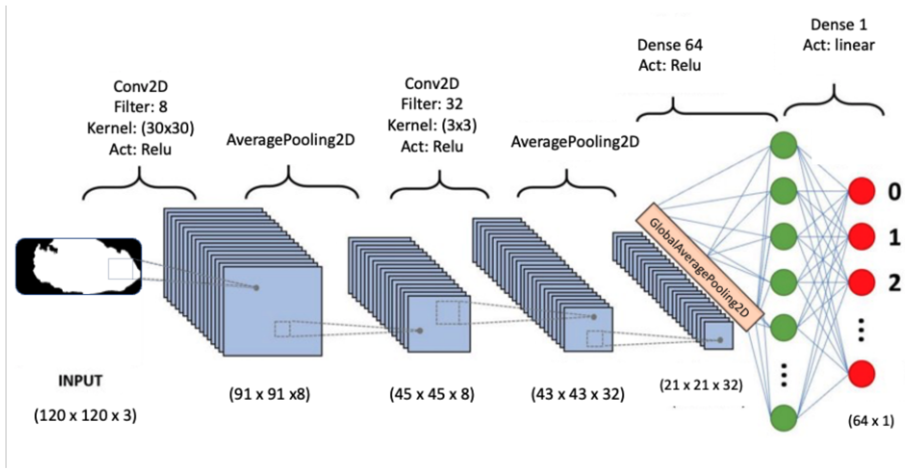

Figure 1: Architecture of CNN [1]

The Architecture of Figure 1 from the paper [1] was used to develop CNN model for this project. We made several improvements on the orginal CNN architecture for this project. In the initial convolution layer, we opted to decrease the number of filters while increasing the kernel size to enhance the encoding of the training image's characteristics. Each image predominantly features a single large focal point—the cow—occupying a significant portion of the image. By employing fewer filters with a larger kernel size, the relationship between the size of the cow and the overall image size is expected to become less abstract, thereby improving our predictive capabilities. The subsequent convolution layer maintained standard settings with a filter number of 32 and a kernel size of 3 [1].

Keras' AveragePooling2D algorithm was employed for both pooling layers instead of the default MaxPooling2D. This choice is deliberate, as our focus is not on concentrating the maximum value from a specific filter, but rather on assessing the average amount of cow body within each filter. The utilization of the AveragePooling2D algorithm aligns better with our desired output, providing a more accurate metric for our objectives [1].

In conclusion, due to the relatively small size of the training set of images, a final fully connected dense layer comprising 64 hidden nodes was employed to consolidate the outputs of the last Average-Pooling2D step into a singular weight estimate for a single dimension.

Note- The original code for CNN and data pre-processing was written by the author of this paper. However, optimization and beautification of code was done by Mr. Keith Myburgh

3.3 Vision Transformer

We used the standard Vision Transformer [14] Architecture as seen in Figure 2 to predict the weight of the animals. We made few modifications to the code. First, we replaced the final layer of ViT with linear layer so that we could predict the weight of the animals. Original architecture classified the pictures into several classes. Secondly, we used the Mean Squared Error (MSE) loss function instead of Cross Entropy loss. In the original paper, they calculated the cross-entropy loss function based on the predicted probabilities assigned to each input image by the model and the true class labels associated with those images.

The Vision Transformer (ViT) [14] is a deep learning architecture specifically crafted for image classification assignments. In contrast to conventional Convolutional Neural Networks (CNNs), ViT diverges from relying on convolutional layers. Instead, ViT processes images by segmenting them

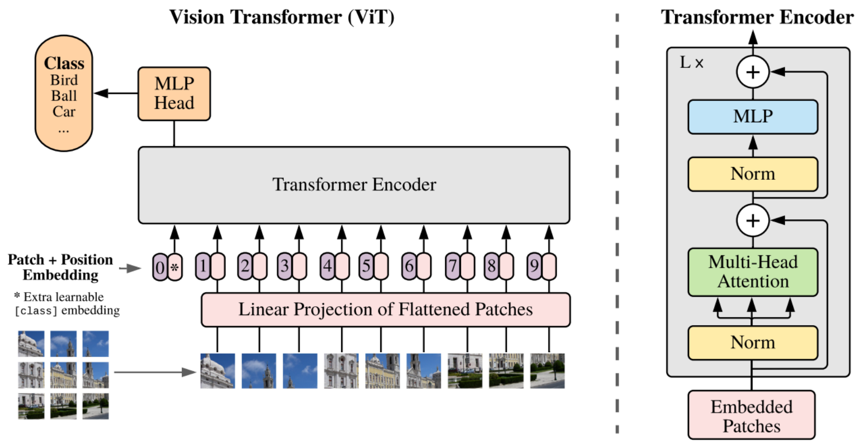

Figure 2: Architecture of Vision Transformer [14]

into patches of a fixed size, which are subsequently linearly embedded. The resultant embeddings are treated as sequences and input into a transformer architecture, originally devised for natural language processing. Finally, we will use fixed number of Transformer Encoders to process the images. The images are passed from a series of encoders to a MLP head so that we can predict the weight of the animals.

3.3.1 Image Pre-processing

All the images were resized to 224x224 pixels, as recommended in [14], and then placed in a single folder. Information about each image was stored in a CSV file with the following rows: File name, Weight, Day, Cow ID, and Time of the day. The images and their respective information were divided into the train loader and test loader. The train loader was utilized for model training, while the test loader was employed to assess the model's loss.

3.4 Video Vision Transformer

3.4.1 Video Pre-Processing

The original sequence of images, when compiled, forms a video because the images were initially extracted as frames from a video. To train the Video Transformer Model, we required vectors with a shape of (16, 224, 224, 3), where 16 represents the number of frames in the video, 224 is the height and width of each image, and 3 denotes the RGB channels for the images.
In the original image source, each folder contained images for each day, time, and Cow ID. To transform the images into the desired shape, we took all the images from the source and converted them into 16-frame videos. For instance, if a folder from a given day and time for a specific cow had 99 images, we padded 13 images from the 99th frame to make it 112 frames, perfectly divisible by 16. We then divided the 112 frames into 7 equal 16-frame videos. This process was applied to all images, resulting in approximately 2500 videos. By using Data Augmentation techniques, we expanded our video count to around 10,000 videos.

3.4.2 ViViT Model 2

The paper [16] introduces Model 2, titled "Factorised Encoder." This encoder presents a transformer-based architecture specifically crafted for video classification. It comprises a Spatial Encoder, a Temporal Encoder, and a Classifier.
The Spatial Encoder handles tokens from the same temporal index, generating representations for each temporal index. These representations potentially encapsulate information regarding the spatial features of each frame in the video. The representations at the frame level are then consolidated into a tensor and fed through a temporal encoder. The resulting output from the temporal encoder serves as the basis for classification. The classification involves utilizing the encoded classification token

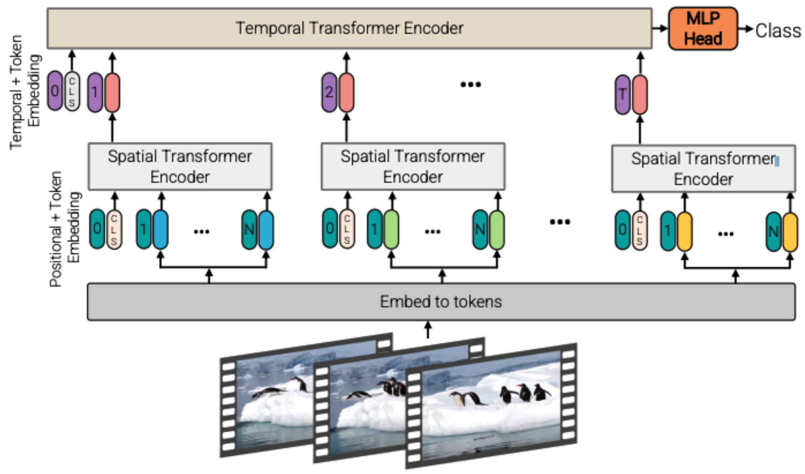

Figure 3: Architecture of Video Vision Transformer Model 2 [16]

derived from the temporal encoder.

In our animal weight estimation research, we opt to substitute the final classification layer with a linear layer to predict the weight of animals. Additionally, we replace the Cross Entropy Loss with Mean Squared Error (MSE) Loss for our project.

3.4.3 ViViT Model 3

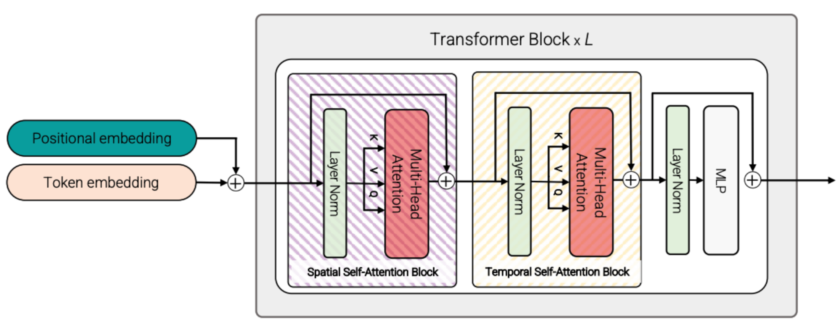

Figure 4: Architecture of Video Vision Transformer Model 3 [16]

The paper [16] introduces Model 3, titled "Factorised Self-Attention," as an additional variant of the transformer-based architecture designed for video classification. In this model, instead of computing multi-head self-attention across all pairs of tokens at a given layer, the self-attention operation is factorized into spatial self-attention and temporal self-attention. The Factorised Self-attention is executed initially in a spatial context (among all the tokens extracted from the same temporal index) and subsequently in a temporal context (among all tokens extracted from the same spatial index). The output derived from temporal attention undergoes processing through an MLP layer, akin to the conventional transformer layer. This model accomplishes a reduction in computational complexity while retaining the capacity to model spatio-temporal interactions in videos.

4 Results

4.1 CNN

The outcomes of the CNN are elucidated in [1]. To assess the prediction weights' quality, both the root mean squared error (RMSE) and the R2 metrics were employed. Furthermore, we present these metrics for predictions made on the entire testing set and a stratified sample of cows, where the strata correspond to the days the predictions were collected. For both error prediction methods, we measured an RMSE of 260.65550 and an R2 of 0.17656. While an RMSE of 260.65550 may be considered favorable, given that the predicted and true weights are measured in the thousands of pounds, an R2 of 0.17656 indicates some weaknesses in the model. The lower R2 suggests that only a small portion of the variance in the predictions can be explained by the model itself. The relationship between the contents of the input images and predictions that could be made without utilizing the model would exhibit comparable performance. This second outcome can be attributed to the nature of the problem at hand. CNNs typically do not excel in regression problems more generally, and additionally, we are employing the CNN architecture in an environment where it is not necessarily intended to be applied. There was little to no object recognition component in the problem structure since we knew beforehand that a cow would be in the input images. Consequently, our model was solely tasked with solving a regression task.

4.2 Vision Transformer

We employed ViT to obtain the MSE loss function, enabling us to predict the RMSE score. Additionally, comparing this RMSE score with that of CNN gives us insights into whether the attention mechanism enhances regression tasks. The best MSE loss for the Vision Transformer model on validation data was 8564.8, while the CNN yielded a best MSE loss of approximately 59,000. The attention mechanism appears to significantly improve regression tasks for images, as it can focus on the crucial parts contributing to the image's weight. The MSE loss also provides an approximate RMSE loss function of 92.5, a notable improvement over the CNN model's 260.7. Given that animal weights are measured in hundreds of pounds, this model serves as a promising initial predictor. Additionally, the CNN requires 74 epochs to converge, whereas the ViT model converges in under 10 epochs. However, after 10 epochs, the ViT model ceases to make significant improvements. This phenomenon may be attributed to the limited amount of labeled image data. For both CNN and ViT, approximately 37,000 images were used, and employing data augmentation techniques could potentially enhance model performance, representing an avenue for future research.

4.3 Video Vision Transformer Model 2

The MSE loss function of the Video Vision Transformer leveled off at around 113,000, a performance significantly inferior to both CNN and Vision Transformer. Notably, it even performed worse than a model randomly predicting the average value of all animal weights. This limitation may be attributed to the insufficient number of labeled videos available for the research, totaling only 2500 labeled videos of 16 frames. Upon applying data augmentation techniques and expanding the labeled video count to 10,000, the MSE loss improved from 113,000 to 100,000. This underscores the necessity for a substantial amount of video data (at least 50,000 videos, as suggested by Deep Learning Practitioners) to achieve satisfactory model performance with Video Vision Transformers. This demand arises from the application of a temporal transformer to multiple spatial transformers, necessitating the fine-tuning of a significant number of parameters compared to regular Vision Transformers.

4.4 Video Vision Transformer Model 3

We utilized the same dataset for Video Vision Transformer Model 3 to verify the absence of bugs in Vision Transformer Model 2. Nonetheless, we obtained similar MSE losses of 113,000 and 100,000 for 2500 and 10,000 labeled videos, respectively. The analysis of these results has already been presented in Section 4.3.

5 Limitation and Future Work

5.1 Data Collection

In terms of data collection, at Kentland Farm, we have a total of 250 lactating cows. Thus, our current dataset of 2500 videos is already substantial. Currently, we collect data by restraining the cows on the scale for 10-12 seconds, twice a day. An alternative approach could involve capturing videos of unrestrained cows as they walk through the weight scales.

5.2 Few Shot Video Regression

Few-shot video regression is a trending research topic where improved performance can be achieved with minimal data. In the realm of few-shot regression, it is essential to have a robust base model that can provide accurate representations of our unique dataset, enabling it to perform subsequent regression tasks effectively. Given that our dataset is distinctive and hasn't been trained by the open-source community, obtaining a suitable base model for accurate representations remains a challenging aspect. Nonetheless, despite the lack of a large dataset, exploring this direction presents an exciting opportunity.

5.3 Camera Angle and 3D models

In the current approach, we capture images of the cow only from the top of the container. This limitation poses a challenge for our model to rely solely on the back of the cow for weight prediction. However, by incorporating various camera angles, we could generate a 3D model of the cow. Utilizing the volume and density information of the cow, we can enhance the accuracy of weight predictions [1].

5.4 Data Balancing

In this project, the images were diverse and covered a wide range of scenarios. We trained the model with as many pictures as possible. However, training the model with different types of cows exhibiting various weights in a systematic manner would prevent bias in the predictions and contribute to a more comprehensive understanding [1].

6 Conclusion

Our research indicates that the attention mechanism proves to be the most effective approach for Image Regression Tasks, suggesting that Video Vision Regression could experience substantial improvement with a larger dataset. The Vision Transformer model could be readily applied in real-world scenarios as a beta model, given its weight prediction closely approximates the actual weight of an animal. Furthermore, fine-tuning these Vision Transformer models with data augmentation could yield an even better RMSE score, allowing the model to selectively focus on relevant parts of the images.

The limitation of having only 37,000 images and 2500 videos (augmented to 10,000) poses a significant challenge in this research. There is a critical need to strategize and collect an exponentially larger volume of labeled images and videos. The scarcity of labeled data also opens up exciting research avenues, such as Few Shot Video Classification.

7 Github code

To visit our code repository, please click Link

References

[1] Kharel, A., Bi, Y., Myburgh, K., Chau, T., & Islam, A. (2022). CS 5824: Final Project Report Animal Data and Weight Estimation.

[2] Dickinson, R. A., Morton, J. M., Beggs, D. S., Anderson, G. A., Pyman, M. F., Mansell, P. D., & Blackwood, C. B. (2013). An automated walk-over weighing system as a tool for measuring liveweight change in lactating dairy cows. Journal of dairy science, 96(7), 4477-4486.

[3] Gomes, R. A., Monteiro, G. R., Assis, G. J. F., Busato, K. C., Ladeira, M. M., & Chizzotti, M. L. (2016). Estimating body weight and body composition of beef cattle trough digital image analysis. Journal of Animal Science, 94(12), 5414-5422.

[4] Ozkaya, S., Neja, W., Krezel-Czopek, S., & Oler, A. (2015). Estimation of bodyweight from body measurements and determination of body measurements on Limousin cattle using digital image analysis. Animal Production Science, 56(12), 2060-2063.

[5] Schofield, C. P., Marchant, J. A., White, R. P., Brandl, N., & Wilson, M. (1999). Monitoring pig growth using a prototype imaging system. Journal of Agricultural Engineering Research, 72(3), 205-210.

[6] Wang, Y., Yang, W., Winter, P., & Walker, L. (2008). Walk-through weighing of pigs using machine vision and an artificial neural network. Biosystems Engineering, 100(1), 117-125.

[7] Goodfellow, I., et al. (2016) Deep Learning. MIT Press, Cambridge, MA.

[8] LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. Nature 521, 436–444 (2015).

[9] Alzubaidi, L., Zhang, J., Humaidi, A.J. et al. Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. J Big Data 8, 53 (2021).

[10] Oliveira e Carmo, L., van den Merkhof, A., Olczak, J., Gordon, M., Jutte, P. C., Jaarsma, R. L., ... & Machine Learning Consortium. (2021). An increasing number of convolutional neural networks for fracture recognition and classification in orthopaedics: are these externally validated and ready for clinical application?. Bone & Joint Open, 2(10), 879-885.

[11] Yamashita, R., Nishio, M., Do, R.K.G. et al. Convolutional neural networks: an overview and application in radiology. Insights Imaging 9, 611–629 (2018).

[12] Kumar, B. (2021, August 31). Convolutional Neural Networks: A brief history of their evolution. Retrieved November 26, 2022, from https://medium.com/appyhigh-technology-blog/convolutional-neural-networks-a-brief-history-of-their-evolution-ee3405568597

[13] Li, G., Huang, Y., Chen, Z., Chesser Jr, G. D., Purswell, J. L., Linhoss, J., & Zhao, Y. (2021). Practices and applications of convolutional neural network-based computer vision systems in animal farming: A review. Sensors, 21(4), 1492.

[14] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

[15] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.

[16] Arnab A, Dehghani M, Heigold G, Sun C, Lučić M, Schmid C (2021) ViViT: A Video Vision Transformer