Cow stall number identification and localization using Object detection models

Haider Ali, Yeshiva Univesity
Hali4@mail.yu.edu

Abstract

In this paper, I defined a system of detection of small numbers using various object detection techniques. The sole purpose of this system is to identify the stall numbers and their location in a certain picture. This system can be used further in places where numbers in the images are tiny, such as Doors and apartment numbers, License numbers, etc.

1. Introduction

The dairy industry is an agricultural sector that produces, processes, and distributes milk and milk-based products. The industry significantly contributes to the global economy and provides employment opportunities for millions worldwide. People of all ages consume dairy products such as milk, cheese, butter, and yogurt which are essential sources of nutrients, including protein, calcium, and vitamins.

The dairy industry has undergone significant changes in recent years, including improvements in technology and changes in consumer preferences. The industry has also faced challenges related to animal welfare, environmental sustainability, and health concerns.

Despite these challenges, the dairy industry remains a critical component of the global food system. It is expected to grow in the coming years as demand for dairy products increases in emerging markets.

Identifying stall images can help automated machines/bots work efficiently, such as milk delivery bots, Cow dung collecting bots, etc. It may also help to solve some of the above problems. The dairy industry has experienced significant growth due to increasing demand for milk and milk products. To maximize milk production and maintain cows' health, managing their health and productivity requires providing a comfortable environment, appropriate feed, and regular health checks. To save time, the authors proposed extracting key teat frames from video images to classify teats and developed a model to recognize cow stall numbers using fine-tuned ResNet34. The model achieved high accuracy in cow stall detection and could provide a valuable tool for dairy farmers to optimize their operations.

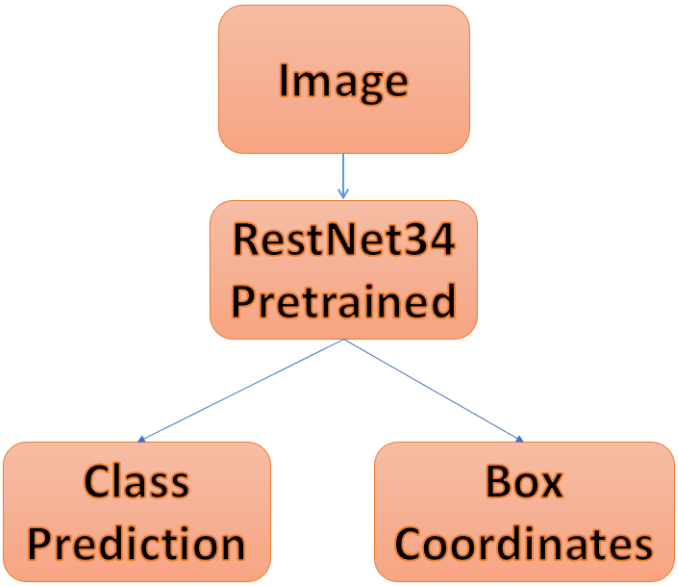

Figure 1. Proposed CNN model: ResNet34 (pre-trained) + Linear

I started my research with a basic CNN model and found that the model loss was getting stuck at the global minimum. To improve accuracy, I started experimenting with the model architectures and building new custom models using some SOTA model concepts such as InceptionNet. To avoid overfitting, as the classes are highly imbalanced, I used FocalLoss and inverse class weights.

2. Related Work

Previously, Dr. Zhang collected and created this dataset using Separable Confident Transductive learning [2]. They used ResNet34 (pre-trained) as a backbone model, achieving 95 percent accuracy in class identification and 40 percent in number localization. He used 0.06 as a learning rate and optimizer Adam with a weight decay of 0.0001 and a learning rate of 0.06.

The research has practical implications for the dairy industry, where the automatic detection of stall numbers can

help improve the efficiency and accuracy of tracking cow health and milk production. Overall, the "CowStallNumbers" dataset and the approach presented in the paper can be useful for researchers and practitioners working on similar tasks in the dairy industry.

"SSD: Single Shot MultiBox Detector" by Wei Liu et al. (2016) [1] introduced a single-shot detector that can detect objects at different scales and aspect ratios with high accuracy and efficiency. Before SSD, most object detection methods used region proposal techniques, such as Selective Search or R-CNN, to generate a set of potential object regions in an image, which were then classified and refined using a separate network.

On the other hand, SSD uses a single neural network that simultaneously predicts object classes and bounding box coordinates for a set of default boxes at multiple feature maps of different resolutions. These default boxes are designed to capture objects of different scales and aspect ratios in the image. Using a multi-scale feature extraction approach, SSD can detect objects at different scales and resolutions in the input image, enabling it to achieve high accuracy while maintaining real-time processing speeds.

In addition to its high accuracy and efficiency, SSD is also highly flexible and can be easily adapted to different datasets and applications. For example, it can be trained on datasets with different numbers of object classes and fine-tuned for specific applications such as pedestrian or vehicle detection.

Since its introduction in 2016, the computer vision community has widely adopted and improved SSD. Several variants of SSD have been proposed, such as SSD with ResNet, which uses a ResNet-based feature extractor for improved feature representation, and SSD with MobileNet, which uses a lightweight MobileNet architecture for real-time processing on mobile devices.

Overall, SSD is a significant advancement in object detection technology that has paved the way for many practical applications in autonomous driving, robotics, surveillance, and more.

3. Methods

I realized that the dataset needs to be balanced using oversampling methods or different strategies to handle an imbalanced dataset. I used inverse class weights as weightage parameters in the Cross entropy and Focal loss to deal with overfitting. Using label smoothing, the model gets stuck at a local minimum.

Below are the Cross-Entropy loss and Focal loss functions. I've used the combined weighted loss as 1.0 * CrossEntropy with weightage loss + 0.1 * box prediction using the SmoothL1Loss + 5.0 * (1 - IOU score).

3.1. Model

Convolutional Neural Networks (CNNs) are popular for image-related tasks because they can learn from image data. In image recognition, ResNet50 is a widely used and highly effective architecture. It has shown outstanding performance in various visual recognition challenges, including object detection and image classification.

ResNet34 is designed to address the problem of vanishing gradients, which can occur in very deep neural networks. The network can better preserve the input information and improve the model's accuracy by using residual connections. Moreover, ResNet34 has a relatively low computational cost compared to other deep neural network architectures, making it a practical choice for real-world applications where efficiency is a concern. Its combination of high accuracy and computational efficiency makes it a suitable choice for detecting stall numbers in images of dairy cows.

In the proposed system for detecting stall numbers in images of dairy cows, ResNet34 is used as the feature extractor. The last layer of the ResNet34 model is removed, and the output is a layer that gives extracted features. The object detection task consists of two stages: finding the coordinates of the object field and identifying the class. Therefore, the model is modified to have two outputs, one for classification and another for the bounding box coordinates.

The proposed system utilizes the ResNet34 model and modifies it for detecting stall numbers in images of dairy cows. Using a widely adopted and effective architecture such as ResNet34 ensures high accuracy and efficiency while modifying to include two outputs makes it suitable for object detection. This system has practical implications for the dairy industry, as it can improve the efficiency and accuracy of tracking cow health and milk production.

$$\text{Cross Entropy Loss} = -\sum_{c=1}^{M} y_{o,c} \log(p_{o,c}), \qquad (1)$$

$$\text{SM L1 Loss} = L_\delta = \begin{cases} \frac{1}{2}(y-\hat{y})^2 & if\ |(y-\hat{y})| < \delta \\ \delta((y-\hat{y}) - \frac{1}{2}\delta) & otherwise \end{cases}, \qquad (2)$$

$$\text{Focal Loss} = -\sum_{c=1}^{M} y_{o,c} \gamma t (1-pt) log(pt), \qquad (3)$$

where Y_i is the original value, $\hat{Y}_i$ is the prediction value. In Focal loss, pt is the cross-entropy loss, and gamma is the weightage parameter.

Set	Class name	Proportion
Training	28	0.032558
	14	0.025581
	3	0.022093
	31	0.022093
	55	0.020930
Testing	0	0.164751
	31	0.030651
	28	0.026820
	51	0.026820
	9	0.022989

Figure 2. Figure: Top classes distributions

4. Dataset

4.1. The dataset

The dataset consists of 1303 images, where 1042 are train images and 261 are test images.

The dataset has 61 classes, with most classes occurring in the dataset being 0, 28, 14, 31, and 3, with the distributions of 0.174664, 0.026871, 0.021113, 0.018234, and 0.018234 showing that the dataset is highly imbalanced towards 0. I've used some techniques to counter the overfitting of the models towards these classes, which are discussed in the methods section.

DatasetsDataset collected and grouped by Zhang [1] were used for training and testing operations. The dataset was previously divided by the author into train and test data. The train part contains 1,043 images in decks corresponding to classes. Test data includes 2443 images. We don't split into train and validation data because we have a small dataset. Train and test datasets have the next cases:

- Image filename - the name of the image file
- Box position 1 - the first point of the box
- Box position 2 - second point of the box
- Box position 3 - third point of the box
- Box position 4 - fourth point of the box
- Class names - the label of the image

4.2. Technological Details

All operations were done in Kaggle. It is a free cloud-based NB runner with a huge dataset collection and competition. The Notebook runner allows you to run Deep learning code with high speed, i.e., on CUDA, an Nvidia GPU in the backend. I used the PyTorch framework (version: 1.13.1, CUDAversion: 11.6) to build and train the Convolutional Neural Network.

5. Results

Here are the losses for the final model

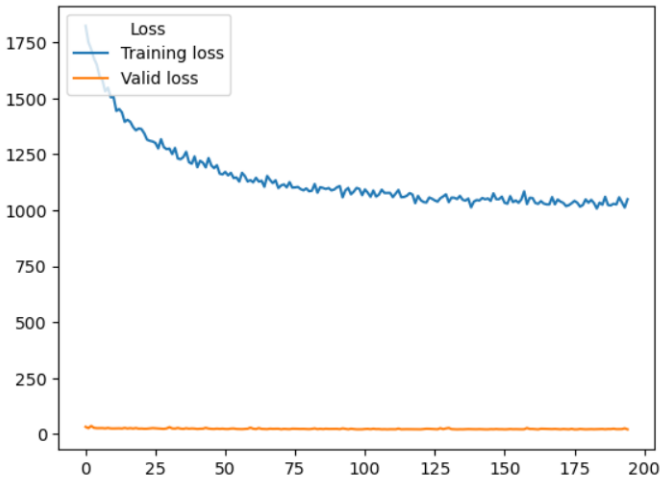

Figure 3. Figure 2. Final model loss plot

Here are the accuracies for the final model

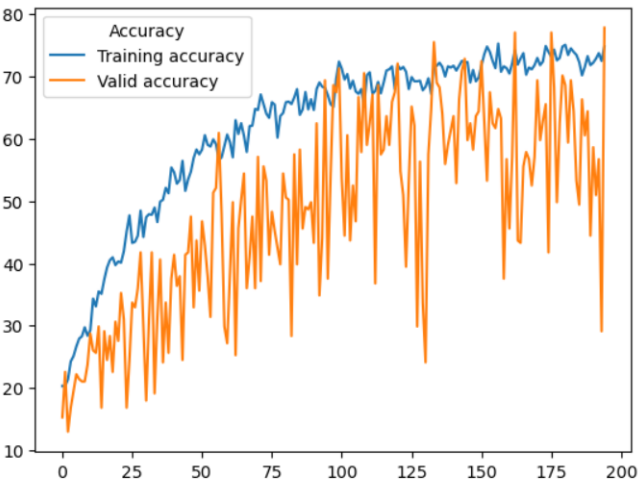

Figure 4. Figure 3: Final model accuracy plot

Table 1. Comparison of models

Model	Train accuracy	Test accuracy
VGG 50	0.30	0.164
ResNet 18	0.22	0.164
ResNet 50	0.20	0.164
ResNet34+Linear	0.81	0.80

6. Discussion

I've reshaped train images to a 256x256 shape and applied random rotation with 20 degrees, Random vertical and horizontal flip with 30 degrees, and Random resized crop for the test images. I didn't apply any transformations other than reshaping the size of the train images.

I've tried VGG19, ResNet18, and ResNet34. I also tried to build a custom model with Inception blocks that didn't give a good result. They only gave me 30 percent accuracy. So my final model was ResNet34 which performed better than all the other models in its size range.

I used Cross Entropy Loss with inverse class weights, SmoothL1Loss, and Focal Loss with gamma values of 2 and 2.5. CEL, MSE, and SmoothL1Loss are used as the final loss.

I've used Adam and SGD as optimizers with different batch sizes and loss functions. I found Smooth L1 loss the best performer, with a batch size of 64, Cross entropy loss, and 1 - IOU score for box accuracy.

Per my experimentations, I can infer that loss functions and transductive learning can greatly boost the model's performance as the model is trained using the test dataset.

All of the models I've run for 250 epochs.

I want to take it to reinforcement learning to learn how a model learns different parameters to predict with just a few images in the dataset correctly.

7. Conclusion

The study "Cow stall number identification and Localization using Object detection models" has practical implications for the dairy industry. The research focuses on the importance of cow stall design and computer vision techniques in automating the detection of cow stall numbers. The study uses a dataset of 1,040 images of cow teats collected from a dairy farm and manually labeled with corresponding stall numbers.

The automatic detection of stall numbers can help improve the efficiency and accuracy of tracking cow health and milk production. By monitoring the behavior of cows with the help of cow stall detection numbers, farmers can identify the best feeding, watering, and milking times for their cows. This can help increase milk production and improve the farm's overall productivity. Moreover, the cow stall detection number generates a large amount of data that can be used to gain valuable insights into the behavior and health of cows. By analyzing this data, farmers can identify patterns and trends and make informed decisions to improve the productivity and profitability of their farms.

This technology can significantly reduce the need for manual labor, saving farmers time and resources while improving the accuracy of their data. Additionally, cow stall detection numbers can provide valuable insights into the behavior and health of cows, allowing farmers to make informed decisions about managing their herds.

Developing efficient and accurate tools for detecting cow stall numbers can significantly affect the dairy industry. The automatic detection of stall numbers can help farmers optimize their herd management practices, leading to improved productivity and profitability. The data generated by cow stall detection numbers can also be used for further research into cow behavior and health, leading to new insights and innovations in the dairy industry.

In conclusion, the study "Cow stall number identification and Localization using Object detection models" highlights the potential of computer vision techniques in the dairy industry. The automatic detection of cow stall numbers can significantly improve the accuracy of tracking cow health and milk production while also generating valuable data that can be used to gain insights into cow behavior and health. Developing efficient and accurate tools for detecting cow stall numbers can help farmers optimize their herd management practices, improving productivity and profitability.

There are lots of scope for improvement. I want to use semi-supervised and reinforcement learning to determine how the models work on this dataset.

I want to use CutMix, and CutNet paper and see if VGG 16 and 19 still overfit the dataset. Creating a synthetic dataset for under-labeled classes to improve accuracy. Applying GoogleNet and Inception models to get more accuracy.

I want to research and expand my vision and find more strategies and methods to increase accuracy and reach 90+ accuracy.

References

[1] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 21–37. Springer, 2016. 2

[2] Youshan Zhang. Stall number detection of cow teats key frames. *arXiv preprint arXiv:2303.10444*, 2023. 1