PICS - Pet Identification and Classification System

Alena Makarova Matthew Pacey Patrick Sullivan
Oregon State University
{makarova, paceym, sullivp2}@oregonstate.edu

Abstract

We introduce a novel, flexible framework for pet identification and breed classification within images, termed PICS - Pet Identification and Classification System. Our approach builds upon the principles of instance segmentation, extending the capabilities of the Mask R-CNN framework to the specific nuances of pet images. By incorporating a branch for high-quality segmentation masks alongside breed classification, PICS efficiently identifies pets and accurately categorizes their breeds within a single, unified model. Trained on the Oxford Pet Dataset, our system demonstrates remarkable precision in delineating pet boundaries and distinguishing among 37 breeds of dogs and cats. Our results indicate that PICS sets a new standard in pet recognition technology, offering substantial improvements over existing methods without significant computational overhead. We envision PICS as a foundational tool for future research and applications in digital media, veterinary services, and beyond, enabling advanced pet management systems and facilitating cross-breed identification.

1. Introduction

In an era marked by rapid advancements in computer vision and deep learning, accurately identifying and classifying objects within images have become increasingly attainable goals. In this context, the recognition and characterization of pets hold significant practical implications across various domains, from digital media to veterinary medicine and beyond. Our project aims to address this challenge by developing a lightweight yet powerful model capable of not only identifying pets within images but also discerning their specific breeds.

Inspired by the versatility and effectiveness of the Mask R-CNN framework in instance segmentation tasks, we extend its capabilities to cater specifically to the domain of pet identification and breed classification. Our tool harnesses the potential of deep learning to provide pixel-level segmentation masks of pets, enabling precise delineation of their boundaries within images. Furthermore, leveraging state-of-the-art classification techniques, our model goes beyond mere identification by accurately categorizing the breeds of the recognized pets.

In this paper, we present the architecture and implementation details of our model, along with experimental results demonstrating its effectiveness and performance in various scenarios. Through rigorous evaluation and analysis, we showcase the practical utility and potential of our approach, laying the groundwork for future research and application development in the exciting intersection of deep learning and pet-related domains.

2. Related Work

2.1. Mask R-CNN Architecture

Mask R-CNN stands as a significant advancement in the field of computer vision. It is uniquely designed to perform two critical tasks simultaneously: identifying the location of objects within an image (object detection) and creating detailed outlines for each of these objects (instance segmentation). This capability builds on the foundation set by Faster R-CNN [1], enhancing it with a specialized component for generating precise segmentation masks for each detected object. This addition is crucial for a deeper understanding and analysis of image contents. As shown in Figure 1, two key innovations at the heart of Mask R-CNN's effectiveness are ROIAlign (Region of Interest Align) and the Feature Pyramid Network (FPN). ROIAlign plays a vital role in accurately capturing the shape and location of objects, especially smaller ones.
The Feature Pyramid Network (FPN) enhances the model's ability to deal with objects of varying sizes. It analyzes the image at multiple levels, each designed to focus on objects of different scales. This multi-layered approach ensures that the model can accurately identify and outline objects regardless of size, making it highly versatile. Mask R-CNN excels at detecting objects and providing highly accurate outlines for each, making it ideal for projects requiring detailed visual understanding.

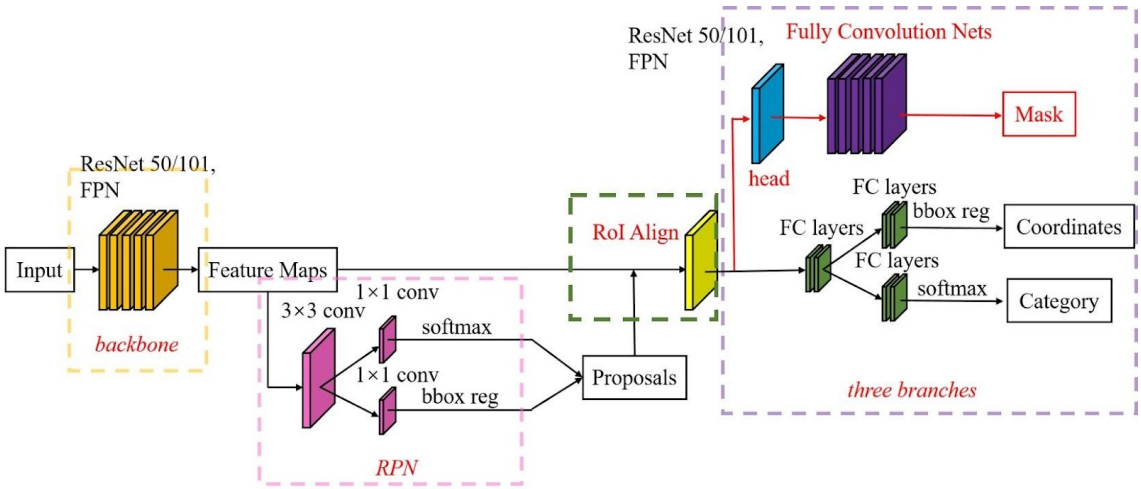

Figure 1. The Mask R-CNN framework for instance segmentation. Source: [4]

2.1.1 Backbone Network - ResNet-50/FPN

Mask R-CNN uses a ResNet-50 backbone that extracts features from input images. ResNet-50's architecture, featuring 50 layers, is preferred for its efficiency in capturing a wide array of features, as its structure has residual connections [2].
FPN enhances the backbone's capability, as shown in the Figure 2 by addressing the challenge of detecting objects at various scales. It builds a multi-scale feature pyramid by integrating features across different resolutions from the ResNet-50 backbone.

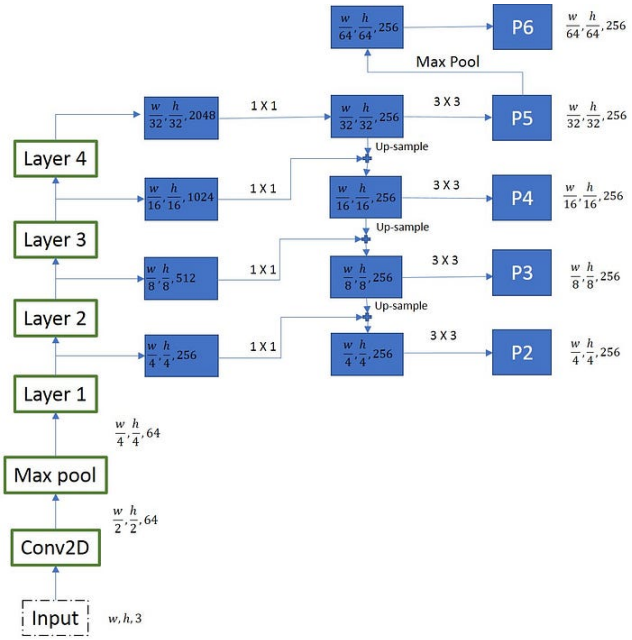

Figure 2. Feature Pyramid Networks (FPN) backbone. Source: [4]

The process involves extracting features using ResNet-50, creating a top-down pathway that fuses high-level semantic information from ResNet-50 with lower-level details, and generating a feature pyramid where each level represents a different scale, enabling effective detection of objects of varying sizes.
The combination of ResNet-50 and FPN equips Mask R-CNN with a robust mechanism for handling the diversity of object sizes in images.

2.1.2 Region Proposal Network (RPN)

The Region Proposal Network (RPN) serves as a critical element of the Mask R-CNN architecture, designed to spotlight areas within an image that are likely to contain objects. By analyzing the feature map generated by the backbone network, the RPN employs predefined anchor boxes of various scales and aspect ratios to identify potential object locations. It provides each anchor with an objectness score to indicate the presence of an object and suggests adjustments for these anchors to better enclose the objects identified. RPN selects the most promising region proposals based on their scores [1], which are then refined and classified in subsequent stages of the framework.

2.1.3 Regions Of Interest (ROIAlign)

The ROIAlign method plays an essential role in the Mask R-CNN model, improving how the model handles regions of interest (ROIs) identified by the RPN. Unlike earlier approaches that roughly estimate the location of features within these regions, leading to potential inaccuracies, ROIAlign uses a more precise technique called bilinear interpolation [1] [4]. This technique calculates the exact feature values for each ROI, preserving the detailed spatial

information needed for accurate object detection and segmentation, as shown in the Figure 3. By ensuring that the features from the backbone network align perfectly with the proposed regions, ROIAlign significantly enhances the model's ability to classify objects and predict their boundaries with greater accuracy. This step is crucial for the overall effectiveness of Mask R-CNN in producing high-quality segmentation results.

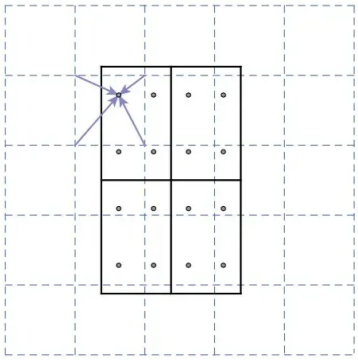

Figure 3. ROIAlign operation. Source: [4]

2.1.4 Mask Head

The Mask Head is a fundamental component of the Mask R-CNN framework, tasked with the precise delineation of object boundaries within the ROIs. Following the accurate alignment of features through the ROIAlign process, the Mask Head applies a series of convolutional layers to these features to generate segmentation masks. Each mask represents a pixel-by-pixel prediction, distinguishing between the object and the background within the ROI. This capability is crucial for achieving detailed instance segmentation, allowing Mask R-CNN to not only identify and classify objects within an image but also to outline their exact shape.

2.2. Dataset

To train our model we are using the Oxford Pet Dataset [3]. This dataset contains 37 classes of various breeds of dogs and cats. Each class contains about 200 labeled images, putting the dataset at about 9,000 images. Each labelled image contains the pet's breed, a bounding box around the head, and a pixel level trimap. The trimap assigns each pixel in the image to one of three categories: object of interest, background, or transition (a bridge between the other two categories). A sample image from the Oxford Pet Dataset can be seen in Figure 4. This shows the original image (left), followed by a bounding box around the animal's head (center), and a tri-map segmentation mask delineating the dog from the background (right).

To prepare our dataset for training (Figure 5), we first standardize all images to a resolution of 224x224 pixels and transform them into tensors (Figure 6). This standardization aligns with the network's expected input format, ensuring that the images are in a consistent form for processing. This step is crucial for maintaining uniformity across our dataset, facilitating more effective learning and performance evaluation.

Figure 4. Annotated data from Oxford Pet Dataset[3]

3. Experiment

3.1. Augmentation

In order to measure the effect of data augmentation for this domain, we measured model accuracy with and without data augmentation. Our base model uses only the original images from the Oxford Pet Dataset. Our augmented model uses the following data augmentations using the Pytorch Transformations (Figure 7):

- **RandomHorizontalFlip** - the image is flipped horizontally about the y-axis; performed 50% of the time
- **RandomVerticalFlip** - the image is flipped vertically about the x-axis; performed 50% of the time
- **RandomCrop** - The original image is randomly cropped to 100 by 100 pixels

3.2. Hyperparameter Selection

To determine the optimal hyperparameters, we conducted a grid search across different combinations of learning rates and batch sizes. We iteratively trained and tested the model for multiple epochs until observing a consistent increase in testing loss, indicating overfitting. Through this process, we identified the most favorable results with a learning rate of 1e-5 and a batch size of 16. In our exploration of optimizers, we compared the Stochastic Gradient Descent (SGD) and the Adaptive Moment Estimation (Adam) algorithms. Our findings indicated that the Adam optimizer outperformed SGD regarding convergence speed and overall model performance on our dataset.

Figure 5. Original image from the train data set

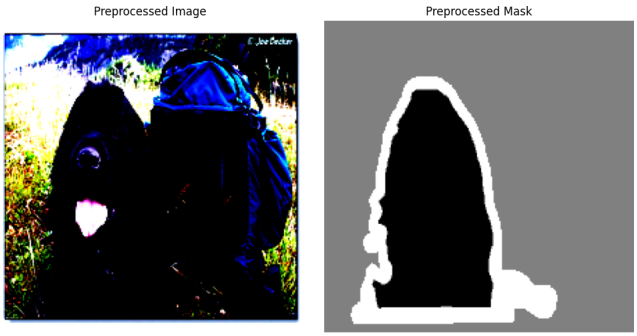

Figure 6. Preprocessed data

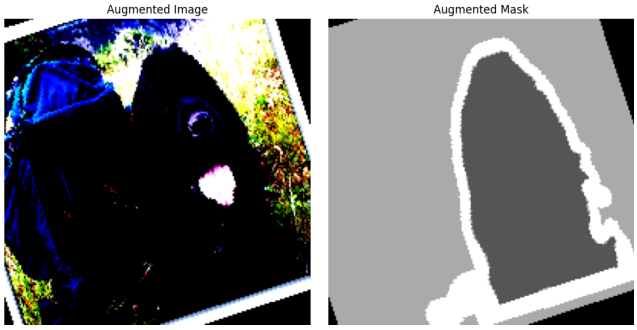

Figure 7. Augmented data

3.3. Evaluation and Results

In our comprehensive evaluation process, we meticulously examined the performance of our model under different conditions—namely, using base data versus augmented data. This evaluation aimed to understand not only the model's overall performance but also the significant benefits derived from data augmentation techniques in terms of training and testing losses, Intersection over Union (IoU) metrics for bounding boxes and masks, and breed classification accuracy as depicted in our confusion matrix.

During training, we rigorously evaluated our model on a dataset specifically held out for testing purposes, which constitutes 20% of the Oxford Pet Dataset. This evaluation thoroughly covers the three crucial outputs generated by our model: bounding boxes, masks, and breed classification. For both bounding boxes and masks, we employed the Intersection over Union (IoU) metric [5] a key measure that calculates the ratio of overlapping regions between the ground truth and model predictions over their combined area. An IoU value of 1 signifies perfect alignment, highlighting our model's precision in segmentation and object localization, whereas lower values indicate discrepancies between the predicted and actual areas. Furthermore, in assessing breed classification accuracy, we not only report accuracy figures but also delve into detailed analysis provided by a confusion matrix (as illustrated in Figure 13). This matrix offers insights into the model's ability to correctly classify breeds correctly, showing a higher concentration of true positives along its diagonal for the augmented data model, indicative of its superior alignment with ground truth data.

Figure 8. Bounding boxes and segmentation masks

Our findings revealed insightful observations. Mainly, augmented data significantly reduced training and testing losses compared to base data, enhancing learning efficiency and model generalization (Figure 12 shows training losses for our augmented and base models.). This was especially notable as while the base dataset's train and test losses began to diverge after 15 epochs—suggesting potential overfitting—the augmented dataset's train and test losses continued to converge even after 100 epochs, showcasing the

Figure 9. Output from the fine tuned Mask R-CNN model using image augmentation. The predicted box is on the leftmost image, the box label is in the middle image and the rightmost image has the predicted mask in green, the label mask in red, and overlap in mustard. Correctly predicts the label "Havanese" in the leftmost image.

Figure 10. Output from the fine tuned Mask R-CNN model using image augmentation. The predicted box is on the leftmost image, the box label is in the middle image and the rightmost image has the predicted mask in green, the label mask in red, and overlap in mustard. Incorrectly predicts "StaffordShire Bull Terrier" [4]

augmented data's contribution to a more robust and generalized model performance.

Similarly, the augmented data led to notable improvements in IoU scores for both bounding boxes and masks, with an increase from 71% to 87% in mask IoU, underscoring the critical role of data augmentation in fine-tuning the model's capability for precise pet image segmentation. The mask and bounding box predictions for our base and augmented models can be seen in Figure 8, 11.

A confusion matrix for our dog breed predictions can be seen in Figure 13. This shows how accurate the model's predictions were (x-axis) vs. the ground truth breed (y-axis). The diagonal line from the top-left to the bottom-right of the matrix represents correct predictions made by the model, indicating its accuracy in classifying dog breeds. Conversely, cells outside this diagonal represent incorrect predictions, with higher or darker values denoting a higher frequency of such errors. As can be seen in the diagnoal, the model predicts breed accuracy with a range of about 30-40%.

These comprehensive results not only underscore our model's effectiveness but also highlight the pivotal role of data augmentation in constructing robust, high-performing models for intricate tasks such as pet identification and breed classification. The insights obtained from our evaluation, particularly the ongoing convergence of training and testing losses with augmented data and the detailed analysis provided by the IoU metric and confusion matrix, guide our continued efforts to refine and expand the capabilities of PICS, advancing the frontier of pet recognition technology.

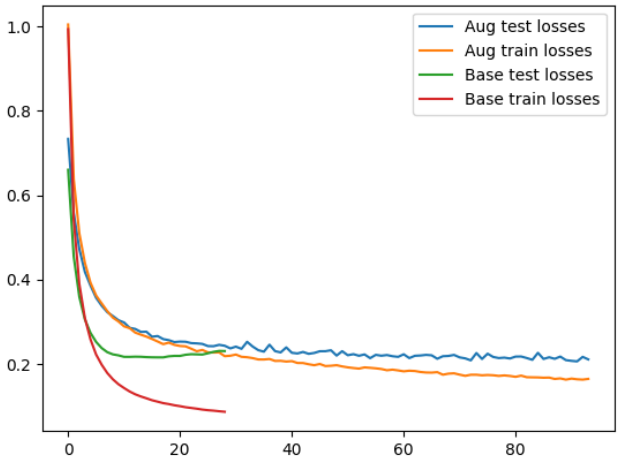

Figure 12. Training Losses

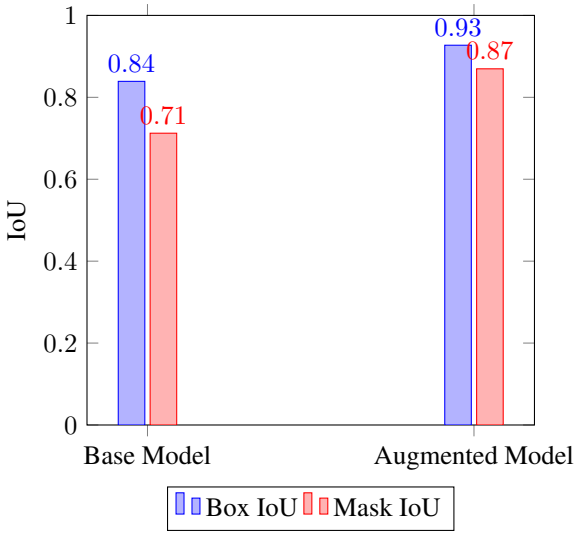

Figure 11. Intersection over Union (IoU) Performance Comparison of Base and Augmented Models

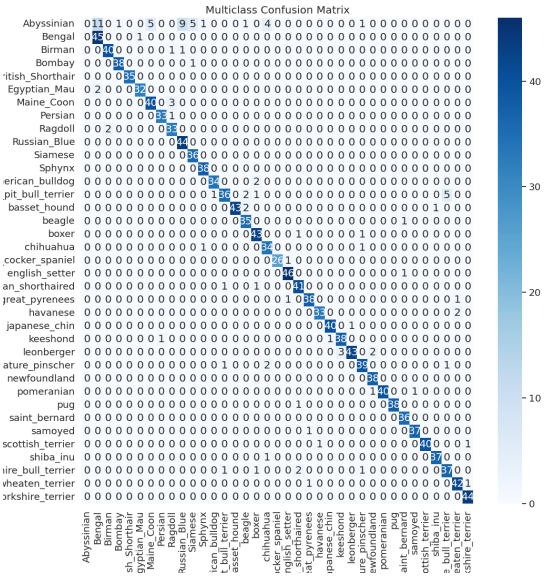

Figure 13. Confusion matrix for all 37 breed categories

4. Conclusion

Our project's journey culminated in successfully developing a model that elevates the precision of pet identification and breed classification within images. By integrating and fine-tuning the Mask R-CNN framework, we achieved substantial improvements in accuracy, particularly highlighted by our model's enhanced performance through data augmentation techniques. The application of random flips and crops proved instrumental, leading to a significant uptick in model accuracy, evidenced by our extensive evaluation process.

Through rigorous testing, including comparisons between base and augmented datasets, our findings underscore the transformative impact of data augmentation on model performance. It reduced training and testing losses and notably increased the Intersection over Union (IoU) metrics for both bounding boxes and masks, signifying a leap from 71% to an impressive 87%in mask IoU. This leap demonstrates our model's refined ability to segment pet images accurately, setting a new standard for pet recognition technologies.

Moreover, the meticulous analysis facilitated by our confusion matrix shed light on the model's breed classification prowess, revealing a higher accuracy rate and a closer alignment with ground truth data when utilizing augmented data. These achievements reflect our commitment to pushing the boundaries of what's possible in pet recognition technology, aiming to meet and exceed current standards.

As we look to the future, the insights gained from this project will continue to guide our endeavors to refine and enhance the capabilities of the PICS system further. Our ambition is to explore new horizons in pet recognition, making strides towards more sophisticated and inclusive applications that cater to a wider array of needs within digital media, veterinary services, and beyond.

In summary, this project represents a significant milestone in the intersection of deep learning and pet-related domains. Our achievements lay the groundwork for future innovations, promising to inspire and inform subsequent efforts to advance pet management systems and facilitate cross-breed identification. The path forward is ripe with potential, and we are eager to continue our exploration and development in this exciting field.

5. Future Work

This project has demonstrated the ability to train a model to detect and identify different pet breeds in images. This was a great learning experience in itself, and could be extended to a number of areas such as:

- **Cross Breed Identification** - This system was trained on pure breeds of pets, but the majority of pets are mixes of multiple breeds. A future development for this system could be to make cross breed predictions rather than single predictions. Instead of predicting the most likely breed, the model's prediction could be tweaked to predict cross-breeds. For example, if the system predicts the top two most likely breeds as Golden Retriever and Poodle, the system could predict Goldendoodle.

- **Pet Management Systems** - This pet identification system could be integrated into a motion activated pet

door designed to allow pets into and out of homes. By adding computer vision to the system, the control of the pet door can be tied to the positive identification of a particular pet or breed. For instance, a dog owner could train a system that would only open her doggie door for her Cocker Spaniel, and not for a stray cat trying to sneak into the house.

- **Animal Tracking System** - This model could be deployed with a trail camera to monitor visits to areas within a dog park. The breed classification system could give park managers a detailed breakdown of not only the counts of dogs visiting the park, but finer grain details about the breeds. For example, if the dog park had a separate areas for small dogs only, this system could be deployed to report instances of "large" breeds of dogs in the "small" breed only area.

References

[1] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. https://arxiv.org/pdf/1703.06870.pdf, 2018.

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun. Deep residual learning for image recognition. https://arxiv.org/pdf/1512.03385.pdf, 2015.

[3] O. M. Parkhi, A. Vedaldi, A. Zisserman, and C. V. Jawahar. Oxford pets dataset. https://www.robots.ox.ac.uk/~vgg/data/pets/, 2012.

[4] P. Potrimba. What is mask r-cnn? the ultimate guide. https://blog.roboflow.com/mask-rcnn/, 2023. Accessed: February 18, 2024.

[5] Wikipedia contributors. Jaccard index. https://en.wikipedia.org/wiki/Jaccard_index. Accessed: March 2024.