BugBot: Using Deep Learning for Household Pest Image Classification

Shirley Fong, Keegan Veazey, Le Ju

January 14, 2025

1 Problem Statement

The issue that our team aims to solve with our project is to be able to identify common household pests. It is important to address this problem because people lack the skills to be able to differentiate bugs from one another, therefore our project aims to fix this issue. Accurate bug classification helps people distinguish between various types of bugs commonly found in the Northeast, potentially protecting their homes from infestations and damage. Currently, there are no existing tools to identify common bugs specifically in the Northeast. Ultimately, our team chose this problem to challenge ourselves by developing a project from start to finish, including data collection and the creation of a functional website. This project has practical value and could serve as a foundation for other applications such as pest control, and public health, demonstrating how data science can address daily challenges.

2 Team and Student Objectives

2.1 Team Objectives

The team aims to accomplish a variety of shared goals throughout this project. The first goal is to successfully collect the data through web scraping images to use and feed into the model we will develop. The next goal is to design and implement a deep learning model to classify common household pests accurately. Lastly, our goal is to collaborate effectively as a team, dividing tasks amongst ourselves so the project can be completed efficiently.

2.2 Individual Learning Objectives

Through this project, Shirley aims to gain hands-on experience with web scraping and the application of deep learning techniques to process and classify images based on their labels. Furthermore, Shirley will contribute by collecting data, helping with model development, and deploying the project using Streamlit. Shirley intends to develop the following skills: web scraping, data augmentation, and building neural networks.

Similarly, Keegan is excited to work on this image classification project and aims to gain experience using PyTorch and to produce an end to end image classifier. She further

intends to learn about and apply techniques to tune and analyze the resulting network for model explainability and transparency. She will contribute by collecting data, model building, and model deployment using Streamlit.

Le aims to strengthen expertise in deep learning and image classification by actively contributing to the development of the machine learning model. Le will focus on optimizing the performance of the model through hyperparameter tuning and advanced techniques such as transfer learning. Additionally, Le plans to assist with data collection, data preprocessing, model evaluation, and ensuring the deployment process is seamless and user-friendly.

3 Dataset Description

The dataset for this project will be a collection of images from Google Images which will be web-scraped by the team. The dataset will be organized into three folders: test, train, and validation. Each folder contains the type of bug, with images stored in JPEG and/or PNG format. The dataset's size will consist of a test set of 20 images per insect, a training set of 100 images per insect, and a validation set of 40 images per insect. The dataset will contain labeled images of bugs, with each image having the following attributes: an image file that represents the bug and a label (e.g., cockroach, bed bug, etc). Although we are collecting the data ourselves, the dataset will require preprocessing such as labeling the images, image resizing, normalization, and data augmentation.

4 Methodology

4.1 Data Collection

Images of household pests will be web scraped from Google Images using Python libraries such as Beautiful Soup and Selenium. The collected images will be categorized into three folders: train, test, and validation, with specified image counts for each category.

4.2 Data Preprocessing

Labeling: Each image will be labeled with the name of the pest it represents.
Image Resizing: All images will be resized to a uniform dimension (e.g., 224x224) to ensure compatibility with deep learning models.
Normalization: Pixel values will be scaled to a range of 0 to 1 for efficient model training.
Data Augmentation: Techniques such as flipping, rotation, and color adjustments will be applied to increase the diversity of the data set and improve the robustness of the model.

4.3 Exploratory Data Analysis

The goal is to understand the characteristics of the data set and ensure data quality using Python libraries such as Matplotlib, Seaborn, and Pandas. The process involves visualizing the distribution of images across pest categories, analyzing image attributes

such as resolution, aspect ratio, and color channels, and identifying any anomalies, such as mislabeled or corrupted images, for correction.

4.4 Modeling

Algorithm: A Convolutional Neural Network (CNN) will be built using PyTorch to perform image classification. The model will include multiple layers for feature extraction and classification. Techniques such as transfer learning (e.g., using a pretrained model like ResNet or EfficientNet) may be utilized to improve performance.
Training: The model will be trained using the training dataset, with the validation dataset being used to fine-tune hyperparameters such as the learning rate, batch size, and number of epochs. Optimization algorithms like Adam and loss functions like Cross-Entropy Loss will be employed.
Evaluation: The test dataset will be used to evaluate the performance of the model using metrics such as accuracy, precision, recall, and F1 score. Additional techniques, such as confusion matrices, will be applied to analyze the predictions of the model and identify areas for improvement.

4.5 Deployment

The goal is to create an interactive and user-friendly application. Streamlit will be used to build and deploy a web application. Users will be able to upload an image of a pest through the app. The app will then classify the pest and display the prediction along with the confidence score.

5 Expected Output

The expected output of this project is a Streamlit app showcasing a household pet machine learning model with a goal accuracy of 90%. Users will be able to upload a photo of a household pest and the app will output the predicted name of the pest based on the image. This project will allow users to gain insight into the pests in their homes. By having an accessible identification tool with high accuracy, the users are able to make informed decisions about how to deal with the pest.

6 Tools and Programming Languages

Languages: Python for web scraping, image analysis and modeling
Libraries: Pandas, NumPy, Scikit-learn, PyTorch, Matplotlib, Seaborn, Selenium, Beautiful Soup
Tools: Jupyter Notebook, PyCharm, GitHub

With this project, we aim to learn new skills, including building neural networks with PyTorch and using Streamlit to deploy and showcase the final classifier.