The Yum or Yuck Dataset of Butterfly Mimics

Keith Pinson[†]

Abstract

This is a dataset of 6 common North American butterflies that birds avoid eating…butterflies that taste bad and butterflies that birds are fooled to think, will taste bad.

We are interested in the question; can an artificial neural network outperform a bird and distinguish between the yummy and yucky butterflies?

Introduction

In the, eat and be eaten world of the animal kingdom, butterflies use a variety of tactics to stay alive long enough to reproduce. Butterflies suffer from predation from birds, insects, and mammals. To combat this, butterflies in all the stages of their life cycle have developed an assortment of defensive measures, including use of toxins, repellents, camouflage, and mimicry.

With this dataset we are presented with images of adult butterflies. Some are toxic and some resemble the toxic butterflies. Birds–who must identify the butterflies often during flight–avoid catching most of them.

While we may not want to identify a butterfly for a quick, tasty meal, accurately identifying butterflies in the wild is still important. Trustworthy surveys help us keep tabs on animal populations in our changing

[†] keith@keithpinson.com

world. Two butterflies in this dataset, the Spicebush Swallowtail and the Monarch, are threatened species. For a human, identifying the butterflies (a visual identification guide is included with the dataset) is easy, though slow and tedious. A machine learning model that can reliably identify butterfly species at speed would be enormously useful.

The Dataset

This 2022 version of the dataset consists of 1028 total images. Each image is a 224x224 pixel jpg showing a single butterfly in the wild. The images are of 6 species of common North American butterflies. The images were gathered from the internet and cropped to a square dimension. Images are for education and research purposes only.

The dataset is split 70/15/15 for training, testing, and a reserved, private holdout.

Additionally, a *"tiny"* dataset of just the Viceroy and Monarch butterflies is included. It is split 83/17 for training and testing. The tiny dataset may be useful for development or classroom work.

Directory Structure

The dataset has the following directory structure:

```
data/ (or tiny/)
├── butterfly_info.csv
⋮
└── butterfly_mimics/
    ├── image_holdouts.csv
    ├── images.csv
    ├── image_holdouts/
    │   ├── ggc1e08cbc.jpg
    │   ├── gh20ab0d9c.jpg
    │   ├── gi31f90cd5.jpg
    │   ⋮
    └── images/
        ├── gh150f104b.jpg
        ├── gh2d5c8c79.jpg
        ├── gh6adf74a4.jpg
        ⋮
```

The full dataset is in the "*data*" folder and the abbreviated dataset is in the *tiny* folder. Both share the same structure.

The Butterflies

Figure 1 The butterfly names starting from the top row: Black, Monarch, Pipevine, Spicebush, Tiger, Viceroy

Name	Common Name	Species	Yum?	Real?	Note
Black	Black Swallowtail	Papilio polyxenes	yum	mimic	mimics pipevine
monarch	Monarch	Danaus plexippus	yuck	real	cardiac glycoside toxins
pipevine	Pipevine Swallowtail	Battus philenor	yuck	real	sequesters aristolochic acid
spicebush	Spicebush Swallowtail	Papilio troilus	yum	mimic	mimics pipevine
tiger	Eastern Tiger Swallowtail	Papilio glaucus	yum	mimic	Females may mimic pipevine
viceroy	Viceroy	Limenitis archippus	yuck	mimic	sequesters salicylic acid

Table 1 From the butterfly_info.csv file

Labels

Four labels are associated with every jpg image:

`Image`, `Name`, `Stage`, and `Side`

For example, `images.csv` begins with:

```
image,name,stage,side
gh150f104b,tiger,adult,both
gh2d5c8c79,monarch,adult,dorsal
gh6adf74a4,pipevine,adult,dorsal
⋮
```

The README files include with the dataset describes the labels in greater detail. It should be noted that every image has a butterfly identified by name and the side of the wing visible: `dorsal`, `ventral`, or `both`.

We call this the *Yum or Yuck Dataset* just for fun.

The dataset is a small and will benefit from augmentation. There is a bias–common to most images gathered from the internet–and that is that the butterflies are posed. Alight, often on a flower, the butterflies are cropped and centered, blurry images are discarded, and only the most eye-catching images are shared. In the wild, trying to snap a picture of a butterfly with one's camera phone, will rarely result in the quality of pictures seen in this dataset.

If using the `"tiny"` dataset, (just for fun) you may wish to classify the Monarch as the Yuck and the Viceroy as the Yum even though this is no longer considered to be their relationship as both are thought to taste bad.

Related Work

This *Yum or Yuck Dataset is* constructed for classification of 6 butterfly species. A Kaggle butterfly dataset (Piosenka, 2021) can be used for classification of 75 butterfly species. This dataset by contributor *gpiosenka* was a source of inspiration. Imagenet (Jia Deng, 2009) contains a few species of labeled butterflies but is best for classification of butterflies in general and not by species. In this role it serves very well.

References

Jia Deng, W. D.-J.-F. (2009). Imagenet: A large-scale hierarchical image database. *CVPR* (pp. 248-255). Ieee.

Piosenka, G. (2021). *Butterfly Image Classification 75 species.* San Francisco: Kaggle. Retrieved 11 4, 2021, from https://www.kaggle.com/datasets/gpiosenka/butterfly-images40-species