2. 데이터 전처리

2021

1. 데이터 전처리 개요

- 데이터프레임 리뷰
- 스케일링과 파티셔닝
- 파일 읽기와 데이터프레임 처리

2. 폴더와 파일 처리

- Os 모듈을 활용
- 파일 및 폴더의 복사/이동/삭제
- 반복문을 활용하는 파일/폴더 처리

3. 엑셀/워드 다루기

- openpyxl과 docx를 활용한 오피스 파일 처리

4. PDF, JSON 다루기 및 기타

PyPDF2와 json 활용

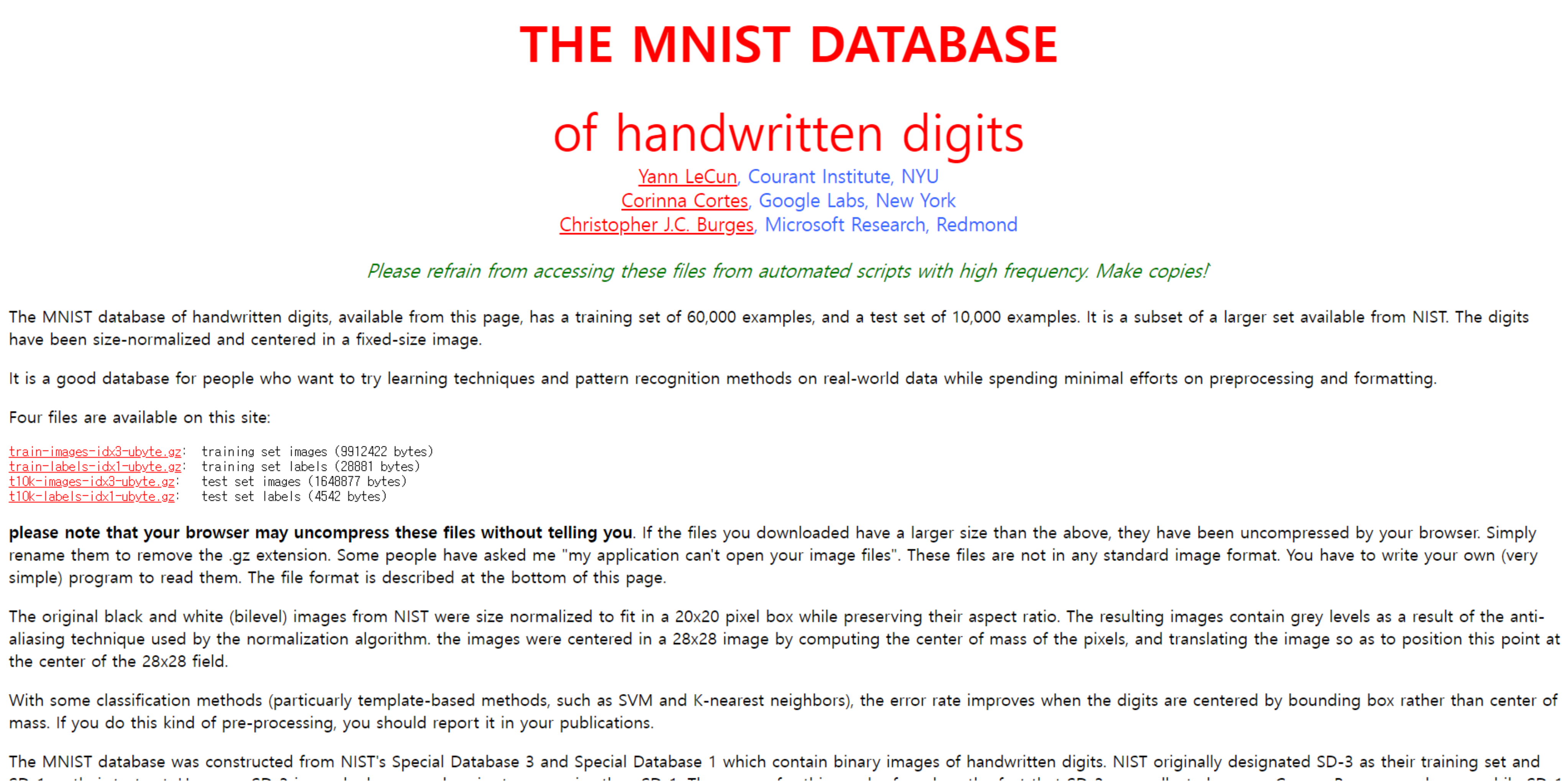
datetime을 활용

5. 이미지 처리 개요

PIL 이용한 이미지 처리 방법 및 실습

6. MNIST 처리하기

MNIST 데이터: 이미지파일로 저장 및 다시 읽어서 데이터프레임 처리



THE MNIST DATABASE

of handwritten digits

Yann LeCun, Courant Institute, NYU
Corinna Cortes, Google Labs, New York
Christopher J.C. Burges, Microsoft Research, Redmond

Please refrain from accessing these files from automated scripts with high frequency. Make copies!

The MNIST database of handwritten digits, available from this page, has a training set of 60,000 examples, and a test set of 10,000 examples. It is a subset of a larger set available from NIST. The digits have been size-normalized and centered in a fixed-size image.

It is a good database for people who want to try learning techniques and pattern recognition methods on real-world data while spending minimal efforts on preprocessing and formatting.

Four files are available on this site:

```
train-images-idx3-ubyte.gz:  training set images (9912422 bytes)
train-labels-idx1-ubyte.gz:  training set labels (28881 bytes)
t10k-images-idx3-ubyte.gz:   test set images (1648877 bytes)
t10k-labels-idx1-ubyte.gz:   test set labels (4542 bytes)
```

please note that your browser may uncompress these files without telling you. If the files you downloaded have a larger size than the above, they have been uncompressed by your browser. Simply rename them to remove the .gz extension. Some people have asked me "my application can't open your image files". These files are not in any standard image format. You have to write your own (very simple) program to read them. The file format is described at the bottom of this page.

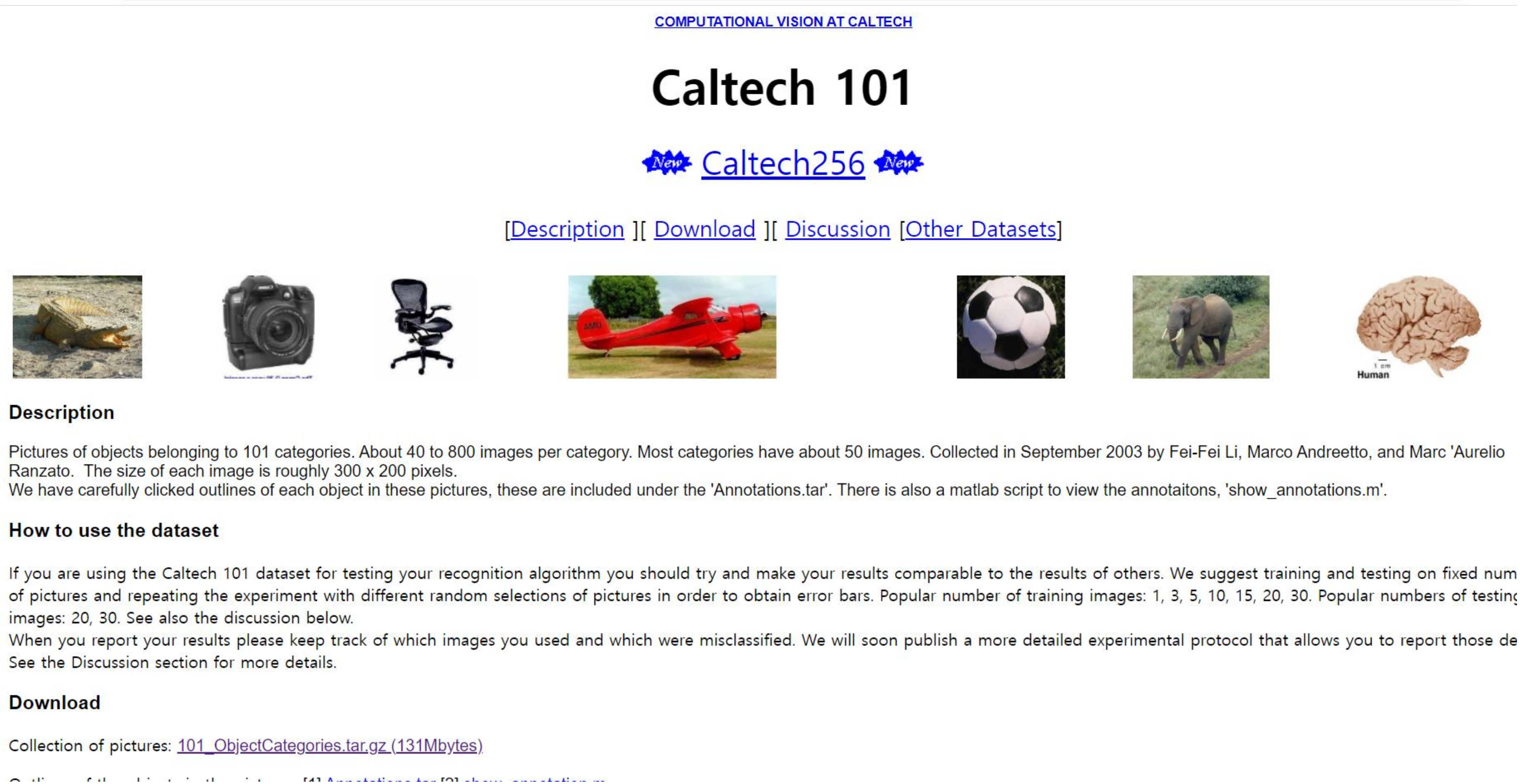
The original black and white (bilevel) images from NIST were size normalized to fit in a 20x20 pixel box while preserving their aspect ratio. The resulting images contain grey levels as a result of the anti-aliasing technique used by the normalization algorithm. the images were centered in a 28x28 image by computing the center of mass of the pixels, and translating the image so as to position this point at the center of the 28x28 field.

With some classification methods (particuarly template-based methods, such as SVM and K-nearest neighbors), the error rate improves when the digits are centered by bounding box rather than center of mass. If you do this kind of pre-processing, you should report it in your publications.

The MNIST database was constructed from NIST's Special Database 3 and Special Database 1 which contain binary images of handwritten digits. NIST originally designated SD-3 as their training set and

7. Caltech101 이미지 처리 연습

Caltech101 이미지 처리



COMPUTATIONAL VISION AT CALTECH

Caltech 101

New Caltech256 New

[Description][Download][Discussion [Other Datasets]

Description

Pictures of objects belonging to 101 categories. About 40 to 800 images per category. Most categories have about 50 images. Collected in September 2003 by Fei-Fei Li, Marco Andreetto, and Marc 'Aurelio Ranzato. The size of each image is roughly 300 x 200 pixels.
We have carefully clicked outlines of each object in these pictures, these are included under the 'Annotations.tar'. There is also a matlab script to view the annotaitons, 'show_annotations.m'.

How to use the dataset

If you are using the Caltech 101 dataset for testing your recognition algorithm you should try and make your results comparable to the results of others. We suggest training and testing on fixed num of pictures and repeating the experiment with different random selections of pictures in order to obtain error bars. Popular number of training images: 1, 3, 5, 10, 15, 20, 30. Popular numbers of testing images: 20, 30. See also the discussion below.
When you report your results please keep track of which images you used and which were misclassified. We will soon publish a more detailed experimental protocol that allows you to report those de See the Discussion section for more details.

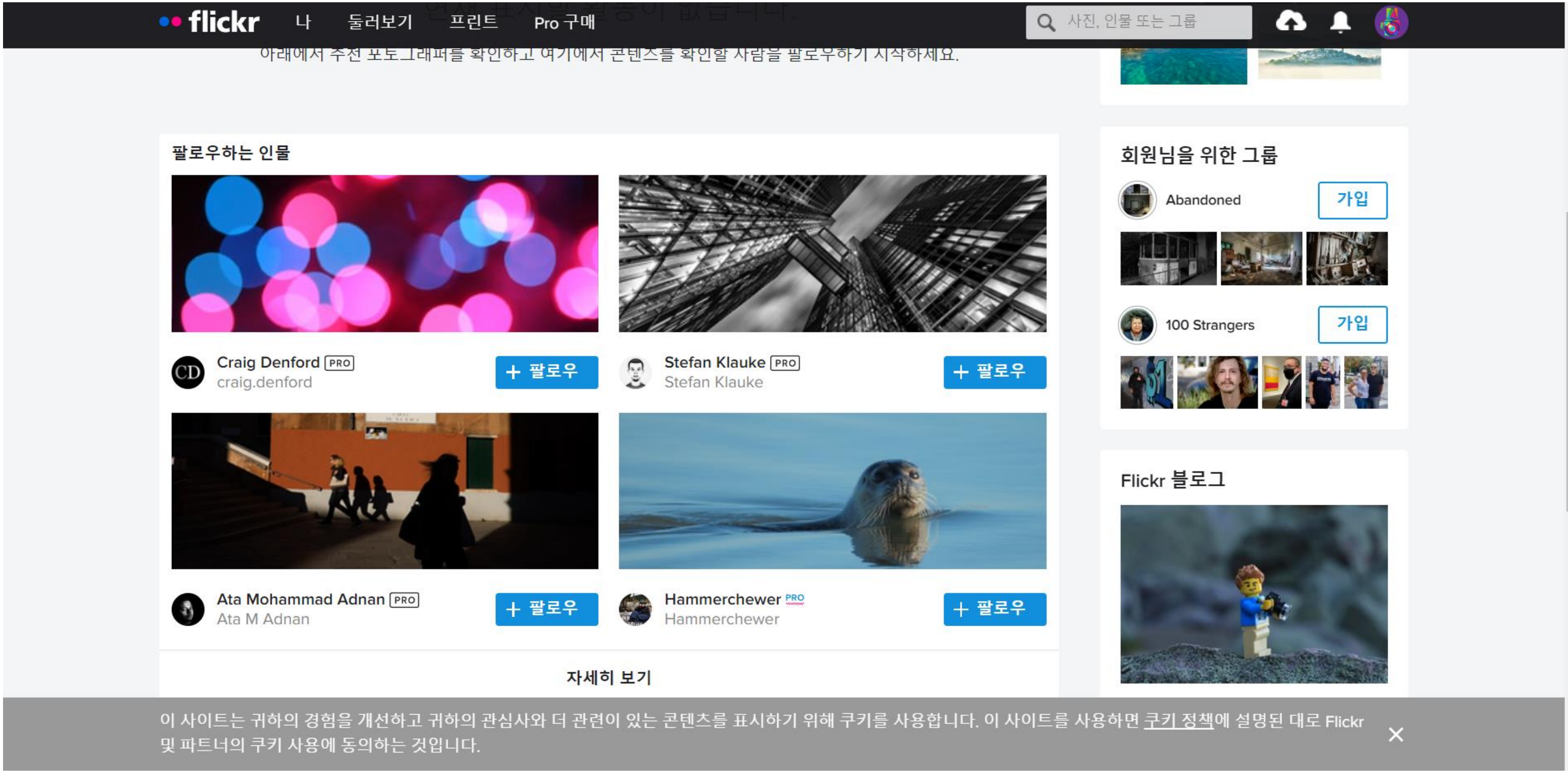
Download

Collection of pictures: 101_ObjectCategories.tar.gz (131Mbytes)

Outlines of the objects in the pictures: [1] Annotations.tar [2] show_annotation.m

8. Flickr 이미지 수집 및 라벨링

Flickr 이용 이미지 수집 및 라벨링



아래에서 추천 포토그래퍼를 확인하고 여기에서 콘텐츠를 확인할 사람을 팔로우하기 시작하세요.

이 사이트는 귀하의 경험을 개선하고 귀하의 관심사와 더 관련이 있는 콘텐츠를 표시하기 위해 쿠키를 사용합니다. 이 사이트를 사용하면 쿠키 정책에 설명된 대로 Flickr 및 파트너의 쿠키 사용에 동의하는 것입니다.

9. 과제

#1. excel폴더 또는 data_politics 폴더의 파일들의 이름을 리스트로 만들어보세요.

#2. keyword_data의 폴더별 파일들을 하나의 새로운 폴더로 이동시키세요. 이동 시에는 각 파일명 앞에, 해당 파일이 속해있던 폴더명을 포함해주세요.

#3. mnist 예제를 전체 이미지에 대해서 처리하고, 예제와 동일한 분류모형을 적용해서 결과를 비교해보세요.

#4. 플릭커 API를 통해서 두 범주 이상의 이미지를 수집한 후 지도학습을 위한 데이터셋 만들어보세요, 만드신 후 분류 모형을 적용해보세요.

#5. Caltech101에서 예제와 다른 카테고리의 이미지를 처리해보세요

Q&A