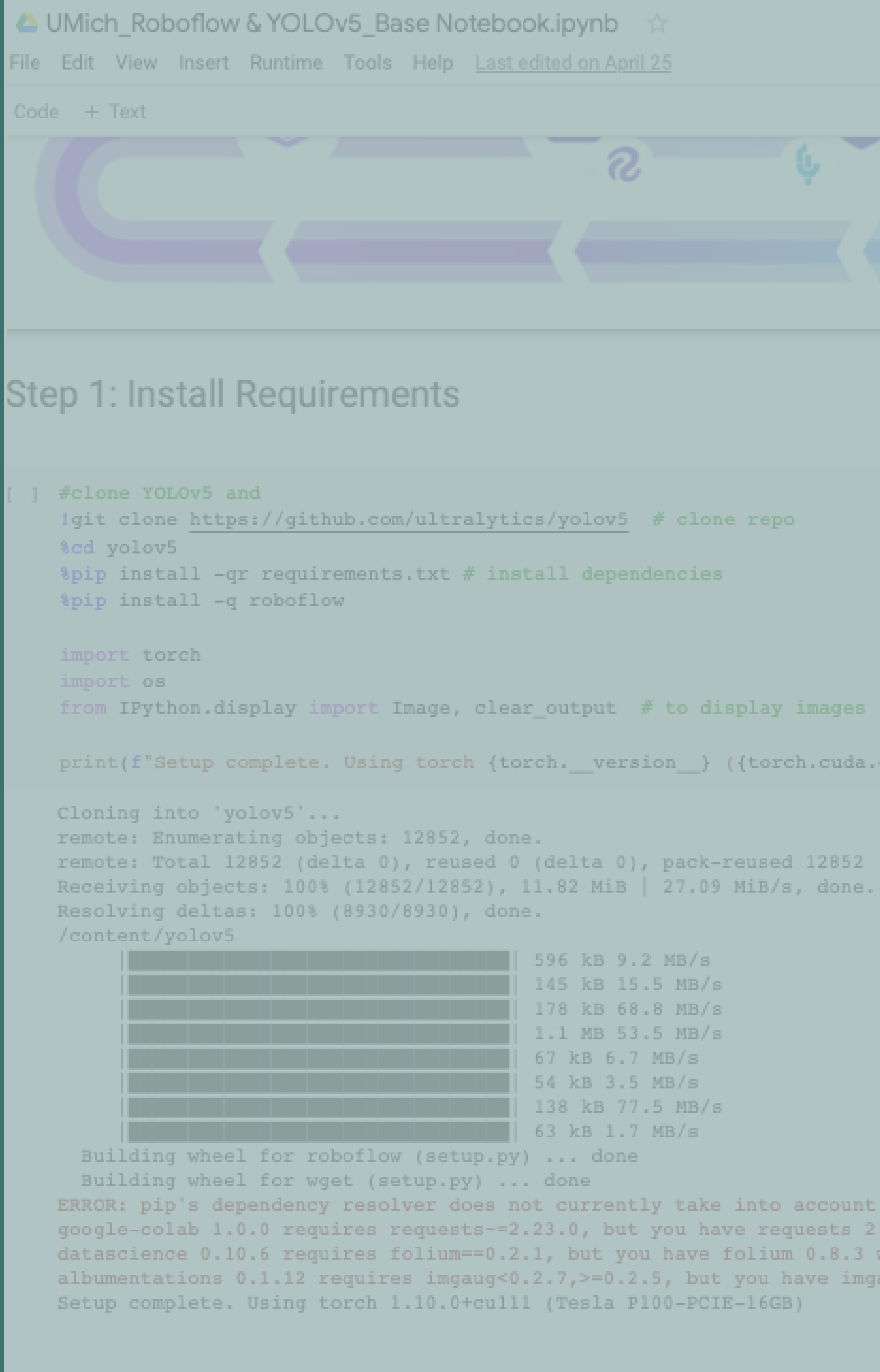
2a

working paper series: Machine Learning 101


UMich_Roboflow & YOLOv5_Base Notebook.ipynb
File Edit View Insert Runtime Tools Help Last edited on April 25
Code + Text

Step 1: Install Requirements

```
#clone YOLOv5 and
!git clone https://github.com/ultralytics/yolov5  # clone repo
%cd yolov5
%pip install -qr requirements.txt # install dependencies
%pip install -q roboflow

import torch
import os
from IPython.display import Image, clear_output  # to display images

print(f"Setup complete. Using torch {torch.__version__} ({torch.cuda.
```

```
Cloning into 'yolov5'...
remote: Enumerating objects: 12852, done.
remote: Total 12852 (delta 0), reused 0 (delta 0), pack-reused 12852
Receiving objects: 100% (12852/12852), 11.82 MiB | 27.09 MiB/s, done.
Resolving deltas: 100% (8930/8930), done.
/content/yolov5
     |████████████████████████████████| 596 kB 9.2 MB/s
     |████████████████████████████████| 145 kB 15.5 MB/s
     |████████████████████████████████| 178 kB 68.8 MB/s
     |████████████████████████████████| 1.1 MB 53.5 MB/s
     |████████████████████████████████| 67 kB 6.7 MB/s
     |████████████████████████████████| 54 kB 3.5 MB/s
     |████████████████████████████████| 138 kB 77.5 MB/s
     |████████████████████████████████| 63 kB 1.7 MB/s
  Building wheel for roboflow (setup.py) ... done
  Building wheel for wget (setup.py) ... done
ERROR: pip's dependency resolver does not currently take into account
google-colab 1.0.0 requires requests~=2.23.0, but you have requests 2
datascience 0.10.6 requires folium==0.2.1, but you have folium 0.8.3 
albumentations 0.1.12 requires imgaug<0.2.7,>=0.2.5, but you have img
Setup complete. Using torch 1.10.0+cu111 (Tesla P100-PCIE-16GB)
```

MACHINE LEARNING 101

2a.1 Introduction

This paper is designed to communicate the foundations of machine learning for entry-level learners and to ease the barriers surrounding machine learning processes for any interested professional. We note, however, that anyone wishing to implement machine learning into their work should also consult external resources to understand the steps and ethical considerations required.

Oriented toward disaster recovery professionals, this paper supplements a tutorial for damage detection using machine learning. The tools to reproduce or recreate our work are linked within this paper and can be accessed through our website.

We present each section as a description of the considerations an entry-level professional should consider at each step of building a machine learning model: collection, annotation, training and application.

2a.2 Data Collection

In machine learning, the first critical step is collecting and labeling pertinent data. With the multitude of options in collecting and annotating, we must first take a step back and ground ourselves in our work's intentions, goals, and methodologies. What will using machine learning do? How will the biases implicit in machine learning impact the work? Can machine learning accomplish the goals of the project? Can other modeling techniques reach these outcomes? Is machine learning enough for this project? The answers to these questions may impact how we approach various choices in the process, or frame how we interpret our data. Once we define these intentions, collection and annotation can begin.

Your research intentionality and the implicit subjectivity in decision-making will directly impact the process of the collection and annotation. Since there is a variety of data used in machine learning, such as text, imagery, and film, the research objective, ideally, is the determinant of the data type you will use in your analysis. Once the type of data is established, the choice of how and from where the data will be collected must be identified. Researchers can seek pre-built datasets of images or text repositories to ensure data integrity. However, relying on a pre-existing data collection method is restricted to the dataset's availability, format, extensiveness, applicability, bias, precision, and reliability. In reality, projects sometimes require the creation of new datasets. Therefore, images must be collected based directly on how the model will be used.

Image 2a.1: Debris Clearance in Southeastern Louisiana. This image demonstrates how perishable data is ripe only in short periods of time. In a few weeks, these debris piles that demonstrate community damage may be cleared, and destruction data eliminated. Machine learning can assit with the capture of data, but researchers should remember the evolution of data prevelance.

KEY TAKEAWAYS

1. Determine the type of data to collect (text, audio, image, or video)
2. Establish spatial and temporal boundaries for data gathering
3. Avoid selection/confirmation bias by collecting diverse data

One notable consideration for dataset creation is the perishability of data. Perishability, as it relates to data collection, is a phenomenon where data and information "becomes less valuable over time as the situation [or research] being predicted may be changing and the predicted event may already have happened." In the context of disaster recovery, the immediate indicators of damage, repair, and recovery evolve. The window for damage assessment narrows as increased recovery speeds and improved distribution of resources alleviates the need for damage-related data. Therefore, timing is strategic consideration in the collection process.

2a.3 Annotation

The next step toward machine learning utilization is to label and describe the data under a coherent, transparent process. Annotation transforms data of distinguishable inputs, such as images or text, and assigns them labels. These labels help machine learning algorithms consume and understand the connections between inputs. Annotations differ based on the machine learning algorithm. Object detection algorithms often require a bounding box to be drawn over detected objects for training purpose, whereas instance segmentation trains on images that have polygon shaping around the detected objects boundaries. (*See Image 2a.2*) Once the machine learning is selected the number of classes–what is to be extracted–can be determined. More specifically, classes are the labels assigned to the data, or parts of the data. With image annotation, one could annotate an image to highlight, extract, and evaluate specific sections of the image. These classes are directly related to the research objectives chosen. For example, segments and classes can identify bicycles on a street from street-level images or the colors of a specific flower within a garden. To summarize, the annotation step establishes a set of classes and corresponding definitions, that the machine learning algorithm will use to extract data. For best results, it is important to allocate an even number of images per class to eliminate sample bias in the model. Sample bias is when a certain class is misrepresented, which may skew a machine learning algorithm to better analyze and detect one class over another.

KEY TAKEAWAYS

1. Choose one or more model: object detection, segmentation, and/or classification
2. Identify and describe the list of classes
3. Organize data with even amounts of images, audios, texts, or videos in each class
4. Select a platform to annotate data

Datasets may be limited to raw data without labels or annotations. There are many online platforms that specialize in annotation for machine learning. Despite the ease in accessing these platforms, there is no standard process for ascribing images or text with labels. As a starting point, you may consider what type of output your analysis requires. There are three major categories of annotation outputs: object detection, segmentation, and classification. These categories can also be used in tandem based on the machine learning algorithm employed. The

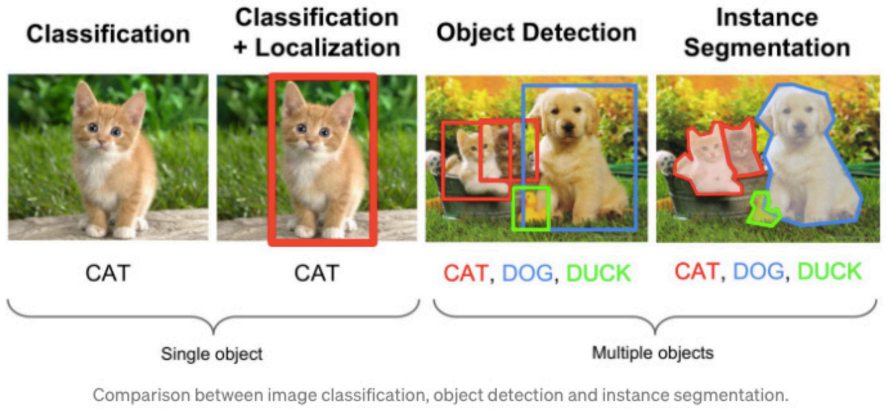


Image 2a.2: Catergories for annotation. This image depicts the differences between different annotation types from object detection to instance segmentation.

Credits to Arthur Ouaknine.

main differences between each category are the timeliness, accessibility/transferability, and accuracy of the outputs. While one object detection model may produce results quickly, it may not be as accurate.

As previously indicated, collection and annotation are far from a perfectly linear process. There are several considerations for each step that influence decision-making and outcomes. The subjectivity of the annotation process creates immense space for error and bias through the labeling process. Researchers must thoroughly ground their decisions with intention and documents in a way that makes their work publicly available for critique.

CASE STUDY

To better understand classification, this example highlights the variety of catergories that contribute to a standard annotation protocol. This model is used in the context of disaster recovery for damage assessments. So, once a natural disaster hits, images of damage are collected to be annotated according to this guidance.

Standardized Annotation Process for FEMA Disaster Assistance

Based on the Federal Emergency Management Agency's (FEMA) Preliminary Damage Assessment guidance, there are four categorical levels of structural damage post-disaster: (1) affected, (2) minor, (3) major, and (4) destroyed (see Figure 2a.3). The FEMA Preliminary Damage Assessment guidelines indicate the differences between level of damage

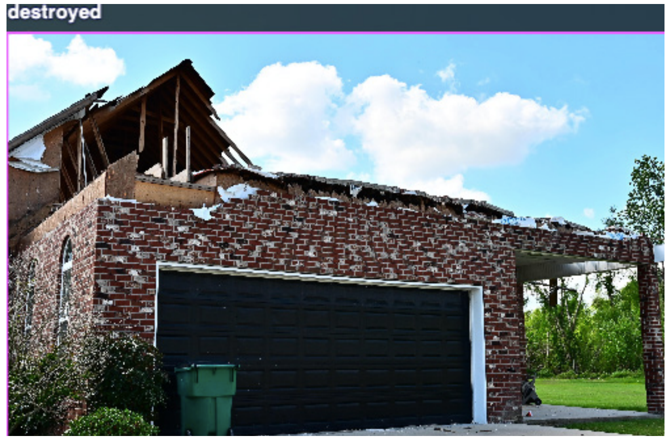


Image 2a.3: Examples of annotations. These images represent annoations on two opposite ends of the spectrum.

AFFECTED	
Description	- Partial missing shingles - Paint discoloration - Broken screens/Gutter damage - Damage to landscaping
Key Phrases	COSMETIC, HABITABLE, NON STRUCTURAL LANDSCAPING

MINOR	
Description	- Nonstructural damage to roof - Damage to chimney - Multiple small vertical cracks in the foundation
Key Phrases	LIVABLE, SMALL, SURFACE LEVEL, EXTERIOR

MAJOR	
Description	- Significant structural damage - Failure or partial failure to structural or walls or foundation - Residences with a water line 18 in above the floor
Key Phrases	BROKEN, FRACTURED, CRUMBLING, DISPLACED, SHIFTED

DESTROYED	
Description	- Total loss - Only foundation remains - Imminent threat of collapse - Completely unlivable
Key Phrases	MISSING, COLLAPSE, UNLIVABLE, DANGER

Image 2a.2: Catergories aligned with FEMA PDA guidelines. This chart depicts the differences between different annotation caterfories for damage detection.

Credits to FEMA Preliminary Damage Assessment.

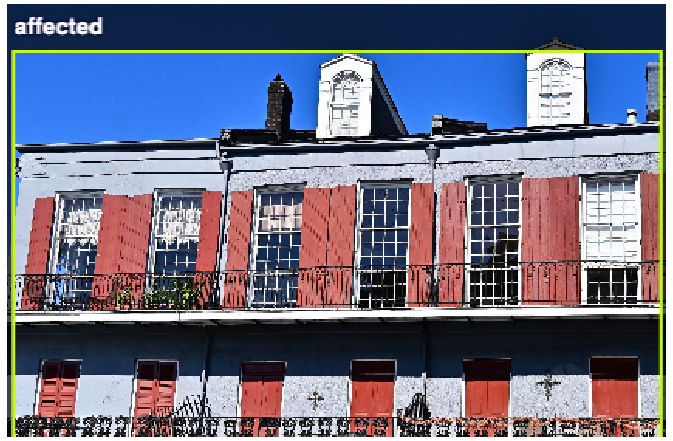


Image 2a.3: Examples of annotations. These images represent annoations on two opposite ends of the spectrum.

2a.4 Training

Once the collection and annotation of data is complete, the actual machine learning training begins. Based on the categories, object detection, classification, or segmentation of the data, there are many algorithms and models that can be deployed, each on multiple platforms and software programs. With many options, and the multitude of technical considerations involved, this process can be confusing for many entry-level practitioners.

There are several fundamental considerations to start. Machine learning algorithms may use one of many programming languages to run and operate their models. Notable languages include Python, and sometimes R. Many pre-built libraries you will find are built using these common languages, but may not be available for all of them.

There are different applications or platforms to type, write, and run programming languages, and not all platforms can run all the languages. When running these analyses on your own, it is important to understand the technical requirements may limit the number of places you can perform the analysis. The number grows even smaller in terms of cloud-based or free platforms. Among the many cloud-based services are Google Colaboratory and Kaggle. These websites understand a variety of languages, have no financial costs, and are fast, user-friendly tools.

The last critical step that is key to running and training an algorithm on your own is knowing how to code at notebook. Since there are many methods, each with coding and technical requirements, it can be confusing to understand where to start and how to implement. Online video tutorials and pre-built, tailored, notebooks can help the process. Following a step-by-step tutorial can make the process more knowable. (This is the reason for our tailored materials for disaster recovery.)

The training of machine learning models can be extremely complex, and is beyond the purposes of this paper. However, to aid understanding for beginners to machine learning, and potential users of these processes, we refer you to the aforementioned tutorials and notebooks.

KEY TAKEAWAYS

1. Find a platform that uses the coding language of selected algorithm
2. Follow custom tutorials or pre-built notebooks for replication

RISING ABOVE THE DELUGE

MACHINE LEARNING BASICS

about this project

This project is a joint effort by students and faculty within the Master of Urban and regional Planning program at the University of Michigan and the National Disaster Preparedness Training Center (NDPTC) as a Capstone project for the Winter 2022 semester.

A key focus of the University of Michigan team is to work in a manner that promotes the values of equity, uplifting local voices, transparency and honesty. As a result, the outcomes of this capstone aim to speak to both our collaborators at the NDPTC and the local communities impacted by disasters across the United States. Our responsibilities as researchers will also include the implementation and/or recommendation of innovative solutions to issues surrounding machine learning, damage assessments, prioritization determinations, and social infrastructure networks.

UNIVERSITY OF MICHIGAN