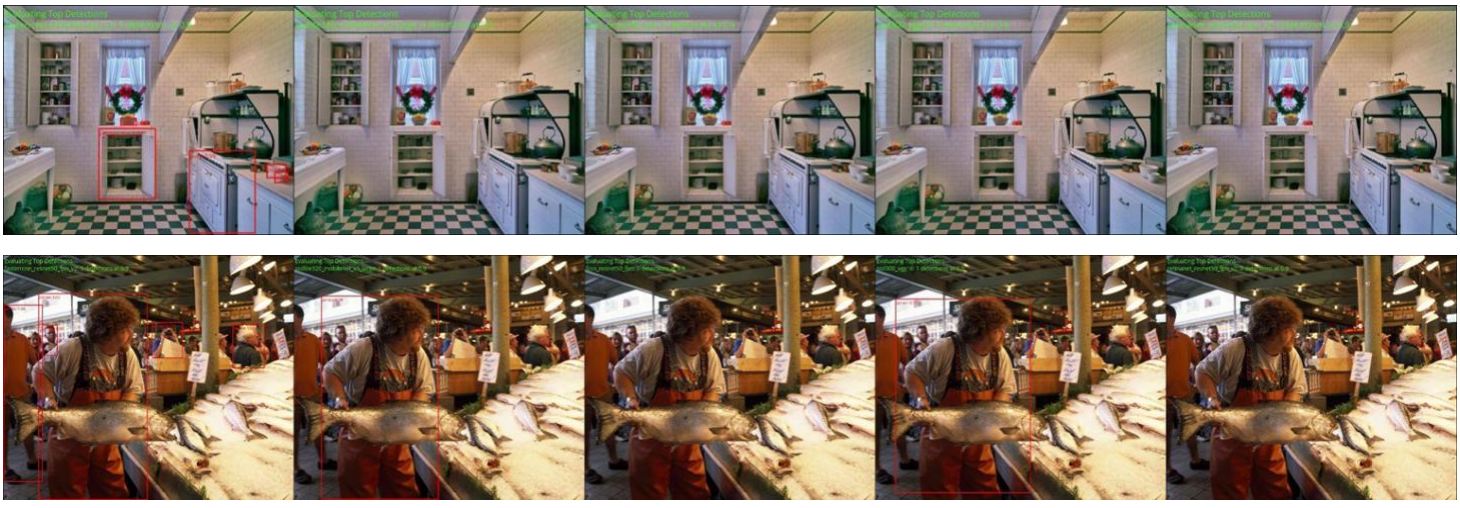
Model Evaluation
LCLabs JFP-24

One of the goals of our collage tool project is to utilize emerging technologies. To fulfill the goal of this project, we have elected to use an object detection model to detect subjects and objects from images and a segmentation model for instance extraction–as a way to simulate the "cut and paste" part of the collaging experience. In this evaluation, we will focus solely on object detection, and review models provided by Pytorch. Further details on the models compared in this evaluation can be found below.

Although varying in architecture, the models from the Torchvision package were pretrained on the Microsoft Common Objects in Context (MS COCO) dataset. The dataset contains 2.5 million labeled instances in 328k images—with 91 common object categories for the model to detect.

To facilitate the model comparison, we selected a confidence threshold of 90% for model predictions, ensuring that the model's predictions were of high standard, and constrained the models' outputs so that it would only label a maximum of five detected objects that it had the highest confidence in. Then, we ran the model on a set of sample images (from the LOC Free to Use and Reuse Data Package) to see what each model would detect and classify correctly.

The example images below compare the performance of each model, given our constraints.
***From left to right: Faster R-CNN, SSDlite, FCOS, SSD, and RetinaNet* (source)**



Based on our comparison grid observations, we can see that the Faster R-CNN model performs the best, due to its ability to detect a variety of objects with the highest confidence. Other models failed to produce such confidence, even in instances where objects in the images were isolated.

However, the Faster R-CNN has its limitations—some inherited from its training dataset. The model is able to isolate and detect people consistently with very high prediction rates; even the

likeness of a person in paintings and statues, but errs in some instances in detecting individuals due to variations in image position or lighting, image quality (for example, objects in stereograph images are often not detected), and general ambiguity in the input image. The quality of the detection is of similar degree in regards to inanimate objects and animals within the scope of MS COCO dataset.

However, for any object categories outside of the training dataset, the model fails to make any accurate prediction or misclassifies the object (e.g. fish identified as bananas). The inability to identify buildings, landmarks, trees and plant life, and accessories could be a challenge for the completeness of our collage tool—as we conceived the use of all of these items as parts of a creative collage to fill out a "city" background. However, this challenge might prove to be less of an issue during the segmentation process, or as we increase the quantity of images the user will have for exploration.

References:

Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., ... & Zitnick, C. L. (2014). Microsoft COCO: Common objects in context. In Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13 (pp. 740-755). Springer International Publishing.

Model Details

Model Alias	**Details**
Faster R-CNN	Faster R-CNN model with a ResNet-50-FPN backbone from Benchmarking Detection Transfer Learning with Vision Transformers paper. ● Model: fasterrcnn_resnet50_fpn_v2
FCOS	FCOS: Fully Convolutional One-stage Object Detection model with a ResNet-50-FPN backbone. ● Model: fcos_resnet50_fpn
RetinaNet	Improved RetinaNet model with a ResNet-50-FPN backbone. ● Model: retinanet_resnet50_fpn_v2
SSD	The SSD300 model is based on the SSD: Single Shot MultiBox Detector paper. ● Model Doc: ssd300_vgg16
SSDlite	SSDlite model architecture with input size 320x320 and a MobileNetV3 Large backbone, as described at Searching for MobileNetV3 and MobileNetV2: Inverted Residuals and Linear Bottlenecks. ● Model: ssdlite320_mobilenet_v3_large