Machine Learning Engineer Nanodegree

Voice Assisted Camera

Seyeon Lee
July 14, 2019

Proposal

Domain Background

Speech Recognition, a field that studies interpretation of human vocal-language system through mathematical algorithms, is recently making impressive strides along with the appearance of Neural Networks. The ability of Neural Networks to recognize patterns shifted the paradigm of Speech Recognition from "what features should be fed to the system" to "machines learning by themselves." Recent studies of Voice Recognition system using recurrent neural networks (RNNs) or Time Delay Neural networks (TDNNs) show great improvement in this technology.

Computer Vision naturally became one of the greatest applications of machine learning. State-of-the-art algorithm for computer vision Convolutional Neural Network, also known as CNN, is an algorithm that was inspired by animals' visual cortex. CNN works with image pixels and the data they inherit. By detecting edges, depth, colors, and many other features of images, CNN can identify shapes and motions. Computer Vision naturally became one of the greatest applications of machine learning. State-of-the-art algorithm for computer vision Convolutional Neural Network, also known as CNN, is an algorithm that was inspired by animals' visual cortex. CNN works with image pixels and the data they inherit. By detecting edges, depth, colors, and many other features of images, CNN can identify shapes and motions.

Problem Statement

The goal is to build a program that intakes users' voice commands and real-time video data from the webcam on a user's computer and do the following tasks:

1. Identify voice commands like "zoom in on" and "zoom out" along with the labels of the objects in the PASCAL VOC (Visual Object Classes) dataset.

2. Find the objects in the video and execute the according command: Zooming in on the object (“zoom in” command), Zooming out

The final program is expected to find objects in my room or on the street through a real-time video via webcam. Object detection, finding objects and locate them in an image, will play a great role in order to complete these tasks. This detection process also has to be fast enough so that it can be operated on a live-cam video.

Datasets and Inputs

There were only limited number of vocabularies in opensource voice command datasets that are available online. Since we need specific commands such as “zoom in on,” “focus on,” or “find,” this project will create its own voice command database; 100 samples for each word. Alternatively, we can use gTTS (Google-Text-To-Speech) python library that uses Google’s Cloud Speech API. Google’s massive data cloud will provide higher accuracy in speech recognition.

For the real-time recognition of humans and objects, PASCAL VOC (Visual Object Classes) 2012 dataset will be in use. This dataset contains large number of images for classification and detection tasks. The dataset has over 500,000 instances. It can classify 20 different objects: bird, cat, cow, dog, horse, sheep, airplane, bicycle, boat, bus, car, motorbike, train, bottle, chair, dining table, potted plant, sofa, tv/monitor, and person.

Solution Statement

I would like to explore Convolutional Neural Networks (CNN) for object detection. I will use transfer learning with CNN, exploring models like ResNet50, VGG, Inception, Xception. By differentiating activation function, layer structure, and optimizer, I will observe how the model gets improved.

Benchmark Model

Regarding CNN object detection models, there exist R-CNN, Fast R-CNN, and Faster R-CNN. When they were trained on PASCAL dataset, 84 hours for R-CNN and 8.75 hours for Fast R-CNN were taken as training time. Test time, operation time taken

per image, was 49 seconds for R-CNN, 2.3 seconds for Fast R-CNN, and 0.2 seconds for Fastest R-CNN. [1]

Evaluation Metrics

Training time and Test time will be used to find how our model is fast in operation compared to the benchmark models. Training time will be determined by how long it takes for our model to train on PASCAL VOC dataset. Test time is the operation time spent per image.

Accuracy score will also be used in this project. How well the program accurately detects objects according to their labels will determine the performance of each model. Each will be tested multiple times, and the accuracy score for each will be evaluated by the following formula.

$$\text{accuracy score} = \frac{the\ number\ of\ correctly\ responsed\ command}{total\ number\ of\ same\ command}$$

Project Design

The project will initially deal with object detection. Importation of datasets would be the first step to be made. The model will be trained on PASCAL VOC 2012, for the convenience in comparing to our benchmark models, and this will be done by downloading the dataset to my AWS GPU instance through this website: http://host.robots.ox.ac.uk/pascal/VOC/ . This is about 2GB. After downloading each training and testing dataset, these pictures will be preprocessed to gray scale images, each image sharing the same size.

For the preliminary analysis, histogram of the number of objects per image will be drawn as below:

[1] /@jonathan_hui. (2019, March 26). Object detection: Speed and accuracy comparison (Faster R-CNN, R-FCN, SSD, FPN, RetinaNet and... Retrieved from https://medium.com/@jonathan_hui/object-detection-speed-and-accuracy-comparison-faster-r-cnn-r-fcn-ssd-and-yolo-5425656ae359

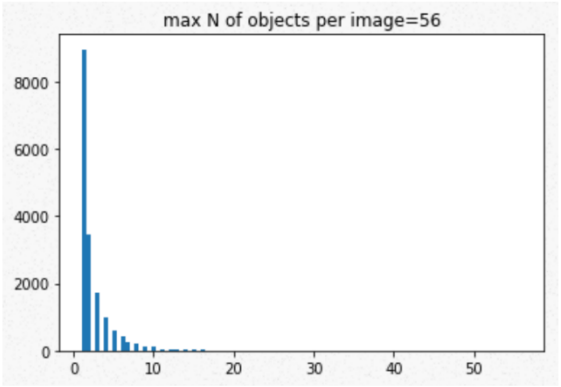

Figure 1 max N of objects per image = 56[2]

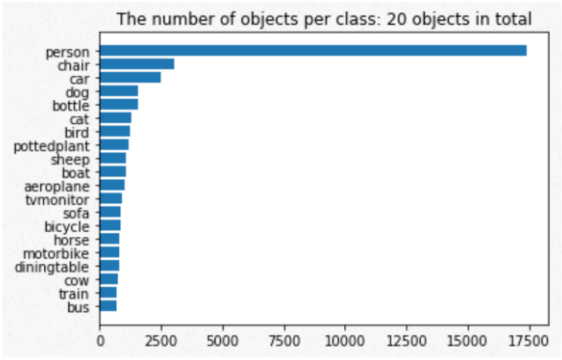

Figure 2 The number of objects per class: 20 objects in total

There are only 20 objects that can be detected in this dataset, and you can see most of the images in this dataset contain "person."

Enabling the voice command will follow the perfect implementation of object detection. This will be either done by my own dataset collected from my friends and I or Google Text to Speech (gTTS) library. For the former case, voice recordings for the command "zoom in" and 3 labels among 20 classifications (person, bottle, and monitor) from PASCAL VOC dataset will be collected, 100 samples for each word, and trained on Recurrent Neural Networks (RNN). Provided by Google, RNN was also used in Sebastian Thrun's Dermatologist AI program. When the voice is recorded, .wav files will be converted into text and be shown on the screen. I will append all the audio files to the same length (~ 2 seconds). The later case, utilizing gTTS, will require some hard coding. The program will understand 4 sentences: "zoom in on the person," "zoom in on the bottle," "zoom in on the monitor," and "zoom out."

If the object detection is fast enough to operate on live-cam videos, voice command will be incorporated to guide the program to only detect specific objects and zoom in on the object by expanding the box drawn around the object. My local computer's webcam will provide real-time video data for the program.

---

[2] Y. (2018, October 20). Part 1 Object Detection using RCNN on Pascal VOC2012 - Data Preparation and Understanding. Retrieved from https://fairyonice.github.io/Object_detection_with_PASCAL_VOC2012_data_preparation_and_understanding.html