Zero shot object detection using a foundation model

학번: 2019320016, 이름: 차주한
지도교수: 김현우 (서명)

2022.3.28

1 Problem (문제 정의)

In order to make life-saving drone, object detection model is needed to find an object and determine whether it is a person or not. Especially in the place like a store, there are many object that can be misclassified to be a person like mannequin, person image in posters or reflected person by glass or mirror. However many existing models are not trained for this kind of tasks, and it is also hard to train a new model from scratch by collecting new data. Making a golden labeled dataset is a cumbersome work and training a model from scratch need lots of resources.

Recently, research has been conducted to pre-train the foundation model and then perform classification through zero-shot transfer to downstream tasks. For example, CLIP(Radford et al. 2021)[1] uses pairs of image and natural language that describes an image during pre-training. It uses natural language as supervision for image to learn image representations and it achieved SOTA in many downstream tasks.

However, for zero-shot transfer to succeed, prompt engineering needs to be also performed. This is because in zero-shot transfer using CLIP, better performance is achieved when labels are given in the form of natural language rather than just words. So in order to perform aforementioned task successfully, prompt engineering study on what form of natural language each label should be given.

In addition, the data provided by drone is not static images, but video data captured in real time, so the speed for real-time processing is also important. The performance difference will increase depending on which vision backbone and which object detection framework is used. For example, according to Ren et al. 2015[2], when using VGG-16 as vision backbone and Selective Search + Fast R-CNN, the processing speed was only 0.5fps. However, changing the object detection framework to RPN + Fast R-CNN increased processing speed up to 5 fps. Furthermore, if ZF net is used as vision backbone, the processing speed incresead to 17 fps. So it is essential to study whhich vision backbone and object detection framework will to use.

2 Approach (아이디어)

References

[1]
[2]