Place of Attention Matters!

An End-to-End Object Detection with Vision Transformation

Saeed Firouzi
Sharif University of Technology
saeedmr881@gmail.com

Abstract

In the object detection task, the purpose is to find the class of object and a bounding box around it. Most of works has focused on just finding the class of object without considering bounding box features properly. We present a new method that focus on relationship between patches of image as a feature for bounding box detector.
Also we combine convolutional neural network as a local feature detector and Transformer network as a long-distance feature detector. We also inspire the method that has been used in Transformer as a relationship between patches in image.
Our implementation can perform in real-time and improve the accuracy of previous works. Training code and pre-training model are available at
https://github.com/saeed5959/object_detection_transformer
to help open source community.

1 Introduction

Most of works has shown that convolutional network are basically consider as a local feature detector. It is quite easy to see if we be familiar with CNN architecture. Using a filter with small dimension like : 3*3 , 5*5 , … , 9*9 can just get a local feature around a pixel value. The strength of using CNN is that we produce high number of these filters to capture features with different styles. some attempts like max-pooling, deformable convolution, dilated casual convolution has been applied to increase receptive fields of a pixel, to look further and to see more around it. But still they have not overcome this issue.

In the other hand, after introducing Transformer, and specifically vision transformer, many works has been tried to apply transformer to image input to get a long-term dependency between pixels or patches of image. The main idea behind the transformer is to get a weighted sum for a vector by point-wise multiplying of vectors as a similarity term.
That means by giving a vector to a transformer, we will get a vector with same dimension as input but some values related to a feature has been increased or decrease based on that features is also exists in other vector or not.

We inspired this point from transform, to use it for bounding box feature to be our major contribution to this work. We call it place-of-object matrix (PO matrix) to be a feature guide for bounding box detection.

Mostly there are two kind of network: two-stage and one-stage networks. In two-stage networks mostly they use a region-proposal-network(RPN) for suggesting some box for objects. These box are predefined and have different ratio and aspect. Ratio are from 20,50,100,200 pixels size and aspects are from 1:1, 1;2, 2;1, … size. In most cases the numbers of these boxes become more than 2000 boxes. This proposal boxes has 2 drawbacks; first it needs a lot computation that makes it impossible for real-time tasks and second, some object can not be fit in these predefined box.

In one-stage network, the main idea is to predict the bounding box coordinate directly. By dividing the input image to patches and predicting bounding box for these patches, it is possible to make the network works in real-time tasks.

In this method, we will use one-stage network, and divide the input image to patches and pass patches to the network that are consisted of CNN and Transformer to get a feature vector for every patch. Then we will calculate PO matrix by multiplying a feature vector of patch at other vectors to get a similarity term as a factor that how much this patch is similar to another patch. And then concatenate this PO matrix with feature vector and then apply a linear layer to get a score for class and another linear layer that outputs 4 values: x of center, y of center, width and length as bounding box parameters.

2 Model Architecture