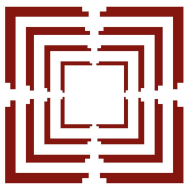

Ahmedabad
University

Machine Learning + Computer Vision Project

Group - 11

Week-5: Progress Report

Project title:

Evaluate performance of various object detection techniques (in case of small objects) on AU Drone dataset.

Group Members:
AU2040014 Jay Patel
AU2040021 Dhanya Mehta
AU2040265 Spandan Shah

Task performed in this week:

Try to understand how to calculate performance matrix accuracy from tfrecord.out file to evaluate the model with different performance matrices. Sir suggested this in the mid-term paper presentation.

The implementation of Synet with the information given in GitHub. In their suggested dataset. In our dataset, we are getting an error that still needs to be solved.

We started implementing the end to end object detection transformer but there are few difficulties in implementing since we didn't go thoroughly through the model architecture but this week we understood the mathematics behind the model and started implementing it.

We have looked through and tried to understand the references provided by the faculty. And to analyze the performance matrix, we have gone through the research papers from the web.

From the performance matrix we have gone through, we have concluded that from many types of it, like MAP, AP, AR, IoU, SVM, PSNR, and FPR. We tried to look into it.

Architecture Analysis of End to End small object DEtection TRansformer:

The input to the model is an image, which is represented as a set of pixels.
The image is first passed through a set of convolutional layers to extract a set of feature maps. These feature maps are then flattened into a sequence of vectors, which are passed through a series of Transformer encoder layers.

Each encoder layer has two sub-layers: multi-head self-attention and feedforward networks.
Let's denote the input feature maps as $X = [x_1, x_2, ..., x_n]$, where each x_i is a d-dimensional vector representing the i-th feature map. The output of the self-attention sub-layer is a sequence of the same length, denoted as
$Z = [z_1, z_2, ..., z_n]$.
Each z_i is also a d-dimensional vector and is computed as follows:

$$z_i = \sum_{j=1}(\alpha_{ij} x_j)$$

where $\alpha_{i,j}$ is the attention weight between feature maps i and j, computed as:

where q_i, k_j are the queries and keys for the self-attention computation and d_k is the dimension of the query and key vectors. These queries and keys are learned parameters.

After the self-attention sub-layer, the output is passed through a feedforward network, which consists of two linear layers with a ReLU activation function in between:

$$FFN(Z)=max(0,ZW_1+b_1)W_2+b2$$

where W_1, W_2, b_1, b_2 are learned parameters.

The output of the last encoder layer is then passed through a set of object detection heads, which predict the location and class of each object in the image. These heads typically consist of a set of fully connected layers followed by a softmax function for classification and a set of regression layers for bounding box prediction.

The task to be performed in the next week:

- Complete all previous models' work that has been left first, and make some conclusions from it.
- AUdrone data callect.

References:

Albaba, B. M., & Ozer, S. (2020). SyNet: An Ensemble Network for Object Detection in UAV Images. **https://arxiv.org/abs/2012.12991**

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020). End-to-End Object Detection with Transformers. **https://arxiv.org/abs/2005.12872**