Project proposal : Topic G
Object Recognition and Computer Vision

Noémie BERGUES
ENS Paris-Saclay
M2 MVA
noemie.bergues@dauphine.eu

Capucine GARÇON
ENS Paris-Saclay
M2 MVA
capucine.garcon@ens-paris-saclay.fr

1. Introduction

We have chosen the topic G : Object Detection and Tracking with DiffusionDet [1]. Our main motivations when choosing this topic are: object detection is a fundamental task in computer vision and can be used for other related recognition tasks such as object tracking; we would like to discover Diffusion models that we have never used before and that are exploited for the first time in object detection. Also, our goal is to extend the use of diffusion models to object tracking.

2. The DiffusionDet model

Diffusion models are generative models that have already been used for text-based image editing, motion planning and more recently in segmentation tasks. Here, for the first time, it enables to detect objects in a generative way by a denoising process from the objects' location hypothesis that are noisy boxes to the object boxes. During inference, we generate hypothesis boxes from a random distribution, then the model refines the boxes from these hypotheses in an iterative way.

The main advantages of the DiffusionDet [1] are the followings: (1) the random boxes can potentially be good candidates in detecting the objects, like anchor boxes for example; (2) there is no need to first learn models' parameters that can give candidates for the boxes like it is the case with region proposal based model, or anchor boxes based model; (3) the model has to be trained only once and then can be used for inference in different settings. The number of random boxes during the inference can be different than for the training. Also, the number of refinement steps can be adapted in order to find a trade-off between speed and accuracy.

Otherwise, in order to adapt this model to object tracking, we are going to leverage the use of the boxes from the previous frame as hypotheses for the actual frame. These boxes are obviously good candidates for the actual frame and it can therefore speed up the inference time because only a few refinement steps are needed.

3. Our experimentations

3.1. Object detection

Firstly, we plan to reproduce two experiments of the paper [1] on the MS-COCO dataset [2]. MS-COCO dataset contains about 118K images in the training set and 5K images in the validation set. The first is to compare performances (AP) between DiffusionDet (with 500 random boxes and one refinement step) and Faster R-CNN.

The second is to compare the performances of the DiffusionDet model with different numbers of random boxes and iteration steps. We will evaluate the average precision of DiffusionDet for 100, 300, and 500 random boxes (for training N_{train} and validation N_{eval}) and iterative steps from 1 to 9. We will match the number of random boxes between the training and the validation step as the authors [1] have shown that the model performs better when $N_{train} = N_{eval}$.

For these two experiments, we will use the following default settings: ResNet50 pre-trained on ImageNet-1K with FPN as backbone, Gaussian Random Distribution concatenating with DDIM as hypothesis boxes and box renewal; and also data augmentation like in the paper [1]. All models are trained with a mini-batch size 16 and 450K iterations.

3.2. Multi-Object tracking

Secondly, we will adapt the DiffusionDet model to solve multi-object tracking problems. This task can be challenging when performed in a crowded and real-world scenario. In our model, we intend to use the boxes of the previous frame as hypotheses for the actual one, therefore, the refinement steps will enable us to track the object between these two frames. We will validate our method on the MOT17 dataset [4]. In the MOT17 there are a training set and a test set, each consisting of 7 sequences and pedestrians annotated with whole body bounding boxes. We will compare our method with the TrackFormer [3]

state-of-the-art method which uses Transformers attention, that jointly performs tracking and detection. For the comparison we will rely on the 2 following metrics: Multiple Object Tracking Accuracy (MOTA) and Identity F1 Score (IDF1).

To train the DiffusionDet model [1], they have used 8 NVDIA A100 GPUs, which are quite powerful. Therefore, if we encounter difficulties in terms of computational cost when training our different models (3.1 and 3.2), we will use less data even if it means having poorer results. Also, we can adapt the number of sample steps (refinement steps) during inference because although increasing this number may significantly improve accuracy, it slows down the detection.

4. Division of the work

To get familiar with the DiffusionDet model, we have planned that each member of the group reproduces one of the experiments detailed in the 3.1. Then we will develop and implement the multi-object tracking model together as it is the most challenging part of the project. Finally, we will share equally the writing of the report.

References

[1] Shoufa Chen, Peize Sun, Yibing Song, and Ping Luo. Diffusiondet: Diffusion model for object detection, 2022. arXiv:2211.09788. 1, 2

[2] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, Pietro Perona James Hays, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2014. arXiv:1405.0312. 1

[3] Tim Meinhardt, Alexander Kirillov, Laura Leal-Taixe, and Christoph Feichtenhofer. Trackformer: Multi-object tracking with transformers, 2022. arXiv:2101.02702. 1

[4] Anton Milan, Laura Leal-Taixe, Ian Reid, Stefan Roth, and Konrad Schindler. Mot16 : A benchmark for multi-object tracking, 2016. arXiv:1603.00831. 1