Segment Anything Also Detect Anything

Rongsheg Wang, Yaofei Duan and Yukun Li

April 11, 2023

Segment anything also Detect anything

Rongsheng Wang
p2213046@mpu.edu.mo
Macao Polytechnic University

Yaofei Duan
p2213964@mpu.edu.mo
Macao Polytechnic University

YuKun Li
p2212990@mpu.edu.mo
Macao Polytechnic University

Abstract

The field of natural language processing (NLP) has been revolutionised by the emergence of large language models (LLMs), which have demonstrated impressive capabilities in zero-shot and few-shot tasks, as well as more complex tasks such as mathematical problem-solving and commonsense reasoning, due to their massive corpus and intensive training computation.The emergence of computer vision macromodels (SAMs) is also transforming computer vision (CV) tasks. In this paper, we propose the use of SAM vision macromodels to guide semi-automated annotation of data in the domain of specific object detection. Also focusing on visual image data augmentation, we propose the High Fine Grain Fill-in Augmentation (HFGFA) method, which can generate false images with higher fineness and greatly improve data imbalance and small object problems. Through early experimental validation, such an approach can improve model generalisation and model generalisation capabilities. Finally, we focus on open world object detection, where the advent of SAM will greatly advance research related to open world object detection.

1 Introduction

Large language models (LLMs) [1, 2, 3, 4, 5, 6] such as ChatGPT, have attracted significant attention from both academia and industry due to their outstanding performance in various natural language processing (NLP) tasks. Based on massive pre-training on text corpora and reinforcement learning from human feedback (RLHF), LLMs can demonstrate exceptional abilities in language understanding, generation, interaction, and reasoning.The empirical trend of Large language models in the field of natural language processing shows that the performance of these models improves with model size, dataset size, and total training computation.

However, despite the great potential of these LLMs in the field of computer vision, their exploration in the field of computer vision is still relatively underdeveloped. Several reasons are included here. First, rich annotated datasets available to support training of LLMs do not exist. Second, the task of computer vision usually requires a more daunting amount of data than natural language processing, and both the large amount of training data and the size of the dataset require higher computational power. Third, it is challenging to design a reasonable promptable model.

Some recent work has led to the development of large models in computer vision. Alexander Kirillov et al. [7] have proposed Segment Anything. The Segment Anything project is an attempt to lift image segmentation into the era of foundation models. Their principal contributions are a new task (promptable segmentation), model (SAM), and dataset (SA-1B) that make this leap possible. While SAM performs well in general, it is not perfect. It can miss fine structures, hallucinates

Preprint. Work in progress.

small disconnected components at times, and does not produce boundaries as crisply as more computationally intensive methods that “zoom-in”.There is still a lot of work to be done here.

The three most important tasks are image classification, object detection, and image segmentation in basic computer vision tasks. Although image classification, object detection and image segmentation are all important tasks in computer vision, their main differences lie in the different problems they solve and in the different objects they operate on to classify the whole image, to localise specific targets or to classify image pixels. Specifically, image classifiers need to learn to extract features from the whole image and predict the probability distribution of each class. Object detector needs to extract features from the entire image and determine the location and category of each object. Image segmenters need to learn to extract features from the entire image and classify each pixel into the correct category. The move from "whole image classification" to "per-pixel classification" means that the difficulty and granularity requirements are gradually increasing. In reality, the fine-grained requirements of most tasks do not amount to "per-pixel classification", but rather to "object classification". Driving image segmentation towards object detection can be better suited to most downstream tasks and can save significant computational and annotation resources.

In this paper, we focus primarily on lower fine-grained tasks, hoping to enable LLMs to contribute to object detection and downstream tasks. Our contributions are as follows:

1. To improve the accuracy of object detection annotation bounding box generation, we suggest utilizing a pixel-level classification model named SAM as a guide.
2. We propose high fine grain fill-in augmentation (HFGFA) using SAM, which reduces image augmentation information redundancy, resolves data imbalance and small object problems.
3. We propose to use SAM to guide the development of the open world object detection (OWOD) task, breaking the closed world assumption.

2 Related Work

Data Annotation. Data annotation refers to operations such as manual or automatic labelling, tagging or classification of data in order to allow computers to better understand the meaning and information of the data, for tasks such as machine learning and deep learning. Data annotation is a very important and common task in areas such as computer vision and natural language processing. Data annotation is a time-consuming and labour-intensive task that requires a large amount of annotation to be done within a certain time frame and without loss of accuracy, and therefore a standardised annotation process must be specified to monitor and manage the quality of the annotation. To address this, a number of companies and organisations specialise in data annotation, providing annotation services to help customers reduce annotation costs, improve annotation quality and speed up model training.

Many automated annotation methods have been proposed, using a small amount of manually annotated data to train a pre-model, using the pre-model to reason about the large amount of unannotated data, and combining this with manual verification to complete the annotation process, such as [8]. While these methods can alleviate the difficulty and cost of data annotation to a certain extent, they still require a small amount of annotation, training and subsequent human correction of the model inference results.

Data Augmentation. Data augmentation is a method of expanding a dataset by generating new and diverse training data using a number of transformations without increasing the amount of data. Data augmentation techniques are commonly used to train machine learning and deep learning models with the aim of improving the robustness and generalisation of the models.Some basic data augmentation methods include: Alex Krizhevsky et al. [9] proposes the use of flip, rotate, crop, colour change, etc. Antreas Antoniou et al. [10] proposes to use DAGAN to generate false images for data augmentation. Some other more advanced data augmentations, such as Mixup [11], Cutout [12], GridMask [13], AutoAugment [14].

Open World Object Detection. Open world object detection refers to the detection and identification of all possible objects and their corresponding locations and bounding boxes from an unknown image or video stream containing unknown or novel objects. Unlike traditional target detection, open world target detection has to deal with larger datasets, more target classes and more complex scenarios. Zhiheng Wu et al. [15] propose two-stage target detectors, including unknown classification and

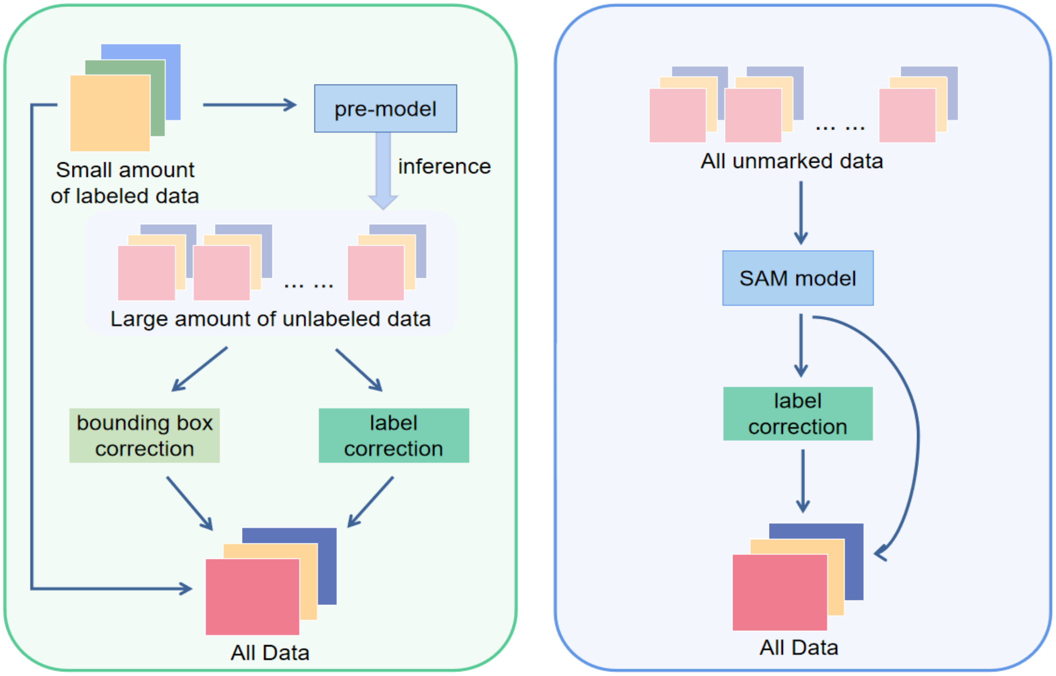

Figure 1: The left side shows the general automated annotation method. On the right is the automated annotation method based on the SAM large model

unknown clustering modules that use unknown label-aware proposals and unknown distinguished classification heads to detect known and unknown objects, and unknown classification and unknown clustering refinement using similarity.Akshita Gupta et al. [16] proposed the use of transformer for pseudotagging, novelty classification and object score calculation to solve the closed world problem.K J Joseph et al. [17] proposed contrast-based clustering and energy-based identification of unknowns.

Unlike these approaches, we propose a range of methods to improve the underlying object detection task based on the SAM large model.

3 Methods

3.1 Guided Data Annotation

Deep learning based automated data annotation methods are methods for directly annotating data from images and videos by means of deep learning models. The basic idea is to train a deep neural network using existing manually annotated data, and use the annotation results from the network for further automated annotation. However, the application of this method is still limited due to the large amount of computational resources and time required for training deep learning models, and the large amount of manually annotated data required for training.

Unlike the above methods, pre-labelling 1 using SAM large models does not require the use of a small amount of manual annotation data and does not require the training of small models to guide annotation generation. It is worth noting that the annotation results generated using SAM are for the image segmentation task, where the minimum positive outer rectangle is computed over the irregular mask of the segmented object to obtain the bounding box of the object for the object detection task, and the minimum outer rectangle can be used to generate the bounding box for rotating object detection.

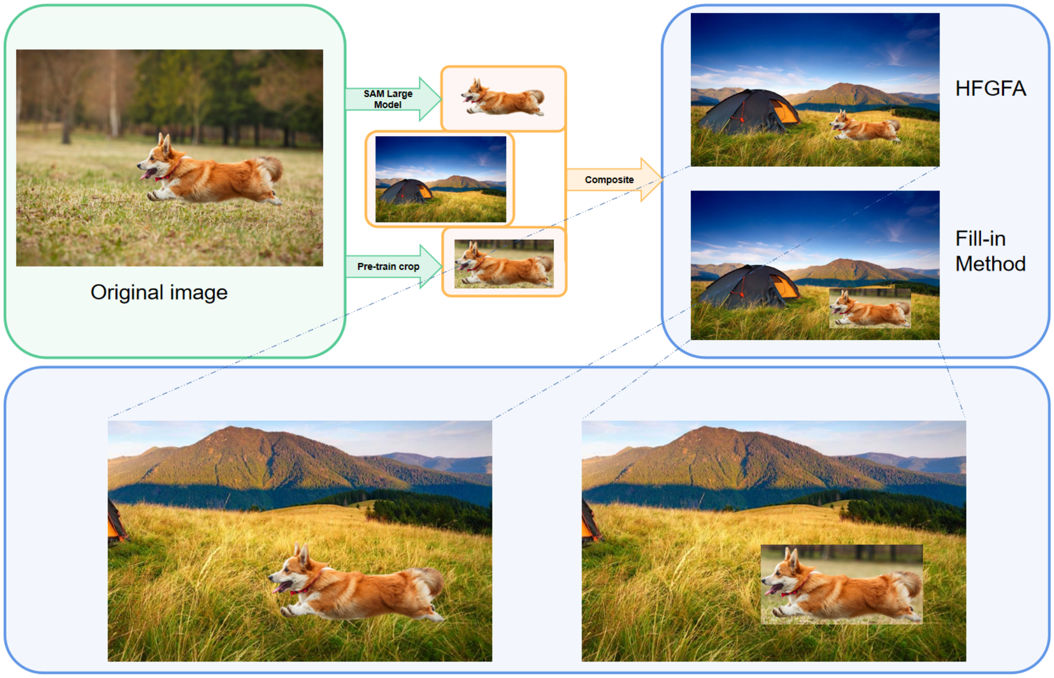

Figure 2: Comparison of general cutting methods and SAM large model cutting data augmentation methods

3.2 High Fine Grain Fill-in Augmentation

Fill-in data augmentation involves training an initial model using a small amount of existing data and then expanding the size and diversity of the dataset by randomly generating a large amount of dummy data and combining it into the original dataset. Instead of collecting and labelling large amounts of real data, fill-in data augmentation expands the dataset by simulating real data. It can greatly alleviate the problem of data deficiency and reduce the amount of data annotation. It also increases the diversity and complexity of the training dataset and improves the generalisation and performance of the model.

High Fine Grain Fill-in Augmentation (HFGFA) 2 no longer requires pre-training for specific task data, but directly puts the high fine-grain segmented objects generated using SAM inference into the new image. Such a high fine-grain cut can reduce data information redundancy, enhance the realism of the generated images and more closely resemble the real environment.

3.3 SAM-based Open World Object Detection

Methods for open world object detection require the identification of unknown instances without explicit supervision and the ability to upgrade knowledge without forgetting earlier instances. the migration of SAM large models to open world object detection can assist in model design.

4 Experiment

Here are some preliminary experimental results. It 3 shows original image, mask label, and bounding box label.These preliminary experimental validations show that high fine-grained segmentation labels can guide the generation of low fine-grained bounding box annotations, thus demonstrating the soundness of our proposed approach.

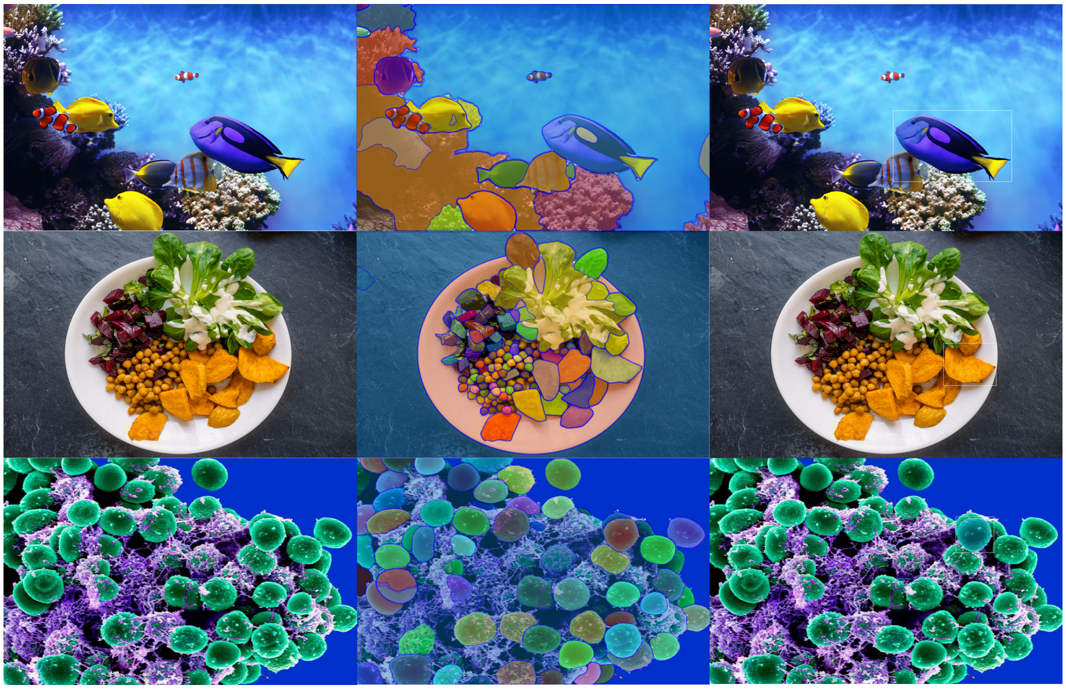

Figure 3: Original image, mask label, and bounding box label visualisation

5 Conclusion

In this paper, we focus on improving the basic object detection task through guidance from large-scale models in the computer vision field. Specifically, we use the high fine-grained segmentation results from SAM to guide the lower fine-grained bounding box and improve object detection. This is essentially embedding the LLMs model into the object detection and downstream tasks. To address the heavy workload involved in annotating object detection tasks, we propose using SAM to generate annotation boxes without the need for training, simplifying the application of object detection models. To solve the issue of information redundancy in image augmentation, we propose using SAM for high fine-grain fill-in augmentation (HFGFA), which addresses the problems of data imbalance and small object detection. Finally, we introduce the SAM model to the open world object detection (OWOD) task. In the future, we hope to extend the proposed methods to actual object detection tasks.

References

[1] Brown T, Mann B, Ryder N, et al. Language models are few-shot learners[J]. Advances in neural information processing systems, 2020, 33: 1877-1901.

[2] Ouyang L, Wu J, Jiang X, et al. Training language models to follow instructions with human feedback[J]. Advances in Neural Information Processing Systems, 2022, 35: 27730-27744.

[3] Chowdhery A, Narang S, Devlin J, et al. Palm: Scaling language modeling with pathways[J]. arXiv preprint arXiv:2204.02311, 2022.

[4] Zhang S, Roller S, Goyal N, et al. Opt: Open pre-trained transformer language models[J]. arXiv preprint arXiv:2205.01068, 2022.

[5] Zeng A, Liu X, Du Z, et al. Glm-130b: An open bilingual pre-trained model[J]. arXiv preprint arXiv:2210.02414, 2022.

[6] Touvron H, Lavril T, Izacard G, et al. Llama: Open and efficient foundation language models[J]. arXiv preprint arXiv:2302.13971, 2023.

[7] Kirillov A, Mintun E, Ravi N, et al.Segment Anything[J]. arXiv preprint arXiv:2304.02643, 2023.

[8] Xie Q, Luong M T, Hovy E, et al. Self-training with noisy student improves imagenet classification[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 10687-10698.

[9] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[J]. Communications of the ACM, 2017, 60(6): 84-90.

[10] ANTONIOU, Antreas; STORKEY, Amos; EDWARDS, Harrison. Data augmentation generative adversarial networks. arXiv preprint arXiv:1711.04340, 2017.

[11] ZHANG, Hongyi, et al. mixup: Beyond empirical risk minimization. arXiv preprint arXiv:1710.09412, 2017.

[12] DEVRIES, Terrance; TAYLOR, Graham W. Improved regularization of convolutional neural networks with cutout. arXiv preprint arXiv:1708.04552, 2017.

[13] CHEN, Pengguang, et al. Gridmask data augmentation. arXiv preprint arXiv:2001.04086, 2020.

[14] CUBUK, Ekin D., et al. Autoaugment: Learning augmentation strategies from data. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019. p. 113-123.

[15] WU, Zhiheng, et al. UC-OWOD: Unknown-Classified Open World Object Detection. In: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part X. Cham: Springer Nature Switzerland, 2022. p. 193-210.

[16] GUPTA, Akshita, et al. Ow-detr: Open-world detection transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022. p. 9235-9244.

[17] JOSEPH, K. J., et al. Towards open world object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021. p. 5830-5840.