Detection of objects and their parts using Transformers

Jiří Vyskočil[1]

1 Introduction

Standard detection and segmentation methods find objects in an image that can often be clearly distinguished from each other. However, there are also tasks, e.g. Visual Question Answering, that require more detailed descriptions, such as attributes or relations with other objects. In such cases, there is already an intermingling, as a more detailed description can belong to several types of objects, e.g. the leg category can be part of the person category, but also the chair category.

In this work, new basic methods for detecting objects and their parts are created. These methods are based on Transformers and the classification layer is created in the same way as in the case of the existing methods of the used dataset. Finally, the methods are compared and evaluated. The best-performing Transformer method is DAB-Deformable-DETR introduced by Liu, et al. (2022), which achieves **35,2 $\mathrm{AP}^{\mathrm{obj}}$** a **16,2 $\mathrm{AP}^{\mathrm{opart}}$**.

2 Dataset

Nowadays, there are many datasets with object hierarchy, e.g. Visual Genome, Open Images, or CityScapes. However, these datasets are often very dirty. They contain incorrect annotations or duplicate labels. In some cases, a tree structure hierarchy is involved (e.g. vehicle → car, train) and all categories can be directly converted to their group/supercategory, so recognizing these groups becomes meaningless. That is why there are datasets that extend the existing annotations, for example by recognizing parts of objects. Unfortunately, these datasets are focused on a certain group of objects, and only some of them get the extension. A recent PACO dataset published by Ramanathan, et al. (2023) aims to the recognition of parts and attributes of common objects. The relationship between objects and their parts is a graph structure, where the hierarchy of the object is taken into account during the evaluation, i.e. what parts the object consists of. The dataset is primarily based on the LVIS dataset and is composed of 456 object-part categories, which were created from selected 75 common objects and 200 parts. Compared to gigantic datasets, it is noticeably cleaner from malicious annotations, and compared to extended datasets, it contains part annotations for every object.

3 Methodology

End-to-end methods have the ability to learn all the steps needed from the input to the output stage and thus do not need manually designed methods (e.g. non-maxima suppression) like Yolo or Faster R-CNN detection networks. Four methods from recent years based on Transformers are selected to detect objects and their parts: Deformable-DETR, DAB-Deformable-

[1] PhD. student, University of West Bohemia, Faculty of Applied Sciences, Department of Cybernetics, e-mail: vyskocj@kky.zcu.cz

Model	Image size	Number of epochs	Mask		Box	
			AP^{obj}	AP^{opart}	AP^{obj}	AP^{opart}
Mask DINO **(new)**	640^2	50	26,3	10,6	19,4	7,8
Def.-DETR **(new)**	640^2	50			24,6	9,3
DETA **(new)**	640^2	50			29,0	12,0
DAB-Def.-DETR **(new)**	640^2	50			29,3	13,2
Mask R-CNN	1024^2	100	31,5	12,3	34,6	16,0
DAB-Def.-DETR **(new)**	1024^2	16			35,2	16,2
Cascade R-CNN	1024^2	100	**32,6**	**12,5**	**37,4**	**16,3**

Table 1: Comparison of methods on the PACO-LVIS test set.

DETR, DETA, and Mask DINO. The classification layer is created in the same way as with existing PACO dataset methods, i.e. direct division into 531 categories. All methods are built on the ResNet-50 architecture and pre-trained on the COCO dataset. Due to the memory requirement of the Mask DINO method during training, all methods are trained on images of size 640×640 for 50 epochs with a total batch size of 16. The initial learning rate is set to 0.0001 and is reduced to 0.00001 after the 40th epoch. The results are shown at the top of Table 1.

The least successful of the compared methods is Mask DINO, which detects objects with 19.4 AP and parts of objects with 7.8 AP. Mask prediction is more accurate - the method is losing 5.2 AP on objects and 1.7 AP on parts of objects compared to Mask R-CNN. DAB-Deformable-DETR achieves the best results with a total of 29.3 AP^{obj} and 13.2 AP^{opart}. Therefore, it was subsequently retrained for 16 epochs with the same batch size and learning constant, which was reduced after the 12th epoch. Augmentations according to the original article were also added with the only difference: the maximum image size is 1024×1024 (compared to 1333×1333). The method trained in this way achieves 35.2 AP^{obj} and 16.2 AP^{opart}, placing it exactly between the baseline methods of Mask and Cascade R-CNN.

4 Conclusion

Even though the methods were trained on 2x nVidia A40 GPUs, the memory requirement of Mask DINO is too high to train 1024×1024 images. However, it still showed decent results in the segmentation task, while DAB-Deformable-DETR achieved comparable results in the detection task. All these methods will serve in the next steps of the research.

Acknowledgement

The work has been supported by the grant of the University of West Bohemia, project No. SGS-2022-017. Computational resources were provided by the e-INFRA CZ project (ID:90140), supported by the Ministry of Education, Youth and Sports of the Czechia.

References

Ramanathan, V., Kalia, A., Petrovic, V., Wen, Y., Zheng, B., Guo, B., ... & Mahajan, D. (2023). PACO: Parts and Attributes of Common Objects. *arXiv preprint arXiv:2301.01795.*

Liu, S., Li, F., Zhang, H., Yang, X., Qi, X., Su, H., ... & Zhang, L. (2022). Dab-detr: Dynamic anchor boxes are better queries for detr. *arXiv preprint arXiv:2201.12329.*