Master-Seminar: Segmenting everything, everywhere, all at once

Sonia Castro Paniello

April 14, 2024

1 Point Transformer [4]

This paper investigates the application of transformer networks to 3D point cloud data. Since scenes can have a high number of points, in order to achieve scalability, self-attention is applied locally using k-nearest neighbors. Another contribution is the use of vector attention to achieve better performance. In this approach, attention weights are vectors that can regulate individual feature channels rather than being shared scalars. Furthermore, the importance of appropriate position encoding is shown.

Point Transformer can be used for classification and semantic segmentation. It demonstrates superior semantic segmentation performance on the S3DIS dataset when compared to earlier models. In fact, the best score to date is achieved by new models that build upon the foundation established in this paper.

The main improvements that could be proposed, some of which have already been implemented, focus on enhancing efficiency to enable greater scalability. For instance, using vector attention significantly increases the number of parameters as the number of channels increases throughout the network; thus, alternative attention types could be explored.

Additionally, the cost of preparing neighbors accounts for up to 70% of the runtime [3]. There have been recent advances in the serialization of point clouds, allowing for the generation of neighborhoods with the same number of points much more efficiently. Furthermore, in the sampling-based pooling procedures employed in the paper, the query sets of points are not spatially aligned due to the uncontrollable information density and overlap between queries. Subsequent work suggests the use of grid pooling to address this issue.

2 OneFormer [1]

The paper proposes OneFormer as a universal image segmentation framework. Its key contribution is the use of a task token to condition the model. This allows training with ground truths of each domain (semantic, instance, and panoptic segmentation) in a single training process. Its architecture design supports multi-task inference as well. Moreover, it includes a contrastive loss between positive pairs of text and queries to establish better inter-task and inter-class distinctions. It outperforms the specialized Mask2Former, trained with three times the resources, across all three segmentation tasks on ADE20k, Cityscapes, and COCO.

Its key strength is being an architecture tailor-made for multi-task joint training. Attempting to apply the same training approach to earlier models does not yield comparable performance. Impressively, OneFormer still outperforms Mask2former even with individual training for a task.

Nevertheless, there's still room for improvement since on COCO, some implementations of OneFormer perform on par or below Mask2former specialized models, indicating potential limitations in specific scenarios. Improvements could be made, as the paper suggests, in the text template space. In the text list created from ground-truth masks (”a photo with a _”), in the padding used (”a/an {task} photo”), and in the task token that is used to initialize the object queries (”the task is {task}”).

It was surprising to learn that using text queries and a contrastative loss worked better than supervising the encoder output with the ground-truth classes. However, I would like clarification on how they get the positive pairs between Q_{text} and Q or if they just base this association on their position. In the second case, wouldn't it be introducing order sensitivity to the model?

3 Mask DINO [2]

A mask prediction branch is added to DINO, which supports all image segmentation tasks (instance, panoptic, and semantic). Its key contribution is the *unified and enhanced query selection* aimed at improving the initialization of the decoder queries. In this approach, three prediction heads are added to the encoder output. To initialize the decoder content queries, top-ranked features are selected based on the classification score. These selected features are then used to regress boxes and masks, supervised by the ground truth. Boxes are derived from the predicted masks to obtain the box initialization for the decoder. Additionally to the query selection, the decoder is trained to predict masks given boxes, treating it as a denoising task. This approach also proposes the removal of box loss and box matching for 'stuff' categories.

Its strength lies in learning supervised proposals in the encoder and letting the decoder refine them with deformable attention. Mask DINO significantly outperformed all existing specialized segmentation methods.

Most papers try to unify semantic, panoptic and instance segmentation as the task of predicting bounding boxes differs slightly. It was interesting to see how in this paper the effectiveness of incorporating this task is shown. Nevertheless, mutual assistance among different segmentation tasks in Mask DINO is hindered, as some individually trained versions of the model continue to outperform the jointly trained one.

Additionally, there is a challenge in terms of memory efficiency, since its mentioned that in large-scale settings the model had to be trained on smaller image sizes and a reduced number of queries which impacted the final performance of object detection. One proposal to address this issue is to, instead of using all scales of features as the encoder input, employ only the larger scale map. The remaining scales can then contribute to cross-attention at different layers. This would mean having two decoders: one for translating features into proposals and another for refining proposals into predictions.

4 X-Decoder [5]

X-Decoder is the first network that can support all types of image segmentation and diverse vision-language (VL) tasks, achieving state-of-the-art results on open-vocabulary. Their main contributions include a novel decoder design that takes two types of queries as input (generic non-semantic queries and textual queries) and produces two types of output (pixel-level masks and token-level semantics), enabling adaptation to a wide range of tasks. Notably, the same text encoder is used for all texts across the different tasks to help the knowledge exchange between them. Additionally, the image and text decoders are decoupled, allowing image-text contrastive learning.

The primary advantage of X-decoder lies in its adaptable architecture, enabling training across diverse tasks and facilitating fine-tuning for additional downstream objectives. Furthermore, it yields good results in open-vocabulary due to the use of contrastive learning when matching a mask to their class and when training image-text retrieval and image captioning. Another notable advantage is the use of a balanced sampling approach for the tasks during pretraining. Despite the larger amount of image-text data in comparison to image-label data, the model sees half of each type at each iteration for balanced learning.

The model currently needs separate pretraining for both image and text encoders. The paper suggests future work in efficiently and effectively pretraining the entire model in a single stage, this could potentially be accomplished in a CLIP-style fashion. Furthermore, the utilization of large-scale web data introduces the risk of inadvertently including private information, inappropriate images/text, or unintentional bias leakage. Also despite the wide range of tasks this model can accomplish, there are still some tasks, such as object detection, that are not integrated into its architecture.

5 SEEM [6]

SEEM is an interactive model aiming to provide a universal segmentation interface. It showcases competitive performance across various segmentation tasks, including interactive segmentation, generic segmentation, referring segmentation, and video object segmentation, all with minimal supervision.

Significant contributions of this paper include introducing a new visual prompt to unify different spatial queries, as well as a memory prompt designed to retain segmentation history. Additionally, it learns a joint visual-semantic space between text and visual prompts (matching them with the class and mask embeddings respectively). Moreover, it has competitive results in open-vocab and zero-shot scenarios since it uses the same text encoder to encode both text queries and mask labels into the same semantic space.

The key advantage of this system is its ability to work well with various tasks and inputs. It's the first interface that can handle different types of user input, like text, points, scribbles, boxes, and images. Another benefit is how efficiently it interacts with users, thanks to memory prompts. By using prompts as input for decoding, the feature extractor only needs to be run once initially, after which the model conducts lightweight decoding for each interaction round.

Finally, its suitability for interactive workflows makes it a good candidate for integration with Large Language Models in interfaces, streamlining the interaction process for the user.

References

[1] Jitesh Jain, Jiachen Li, Mang Tik Chiu, Ali Hassani, Nikita Orlov, and Humphrey Shi. Oneformer: One transformer to rule universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2989–2998, 2023.

[2] Feng Li, Hao Zhang, Huaizhe Xu, Shilong Liu, Lei Zhang, Lionel M Ni, and Heung-Yeung Shum. Mask dino: Towards a unified transformer-based framework for object detection and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3041–3050, 2023.

[3] Zhijian Liu, Xinyu Yang, Haotian Tang, Shang Yang, and Song Han. Flatformer: Flattened window attention for efficient point cloud transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1200–1211, 2023.

[4] Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip HS Torr, and Vladlen Koltun. Point transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 16259–16268, 2021.

[5] Xueyan Zou, Zi-Yi Dou, Jianwei Yang, Zhe Gan, Linjie Li, Chunyuan Li, Xiyang Dai, Harkirat Behl, Jianfeng Wang, Lu Yuan, et al. Generalized decoding for pixel, image, and language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15116–15127, 2023.

[6] Xueyan Zou, Jianwei Yang, Hao Zhang, Feng Li, Linjie Li, Jianfeng Gao, and Yong Jae Lee. Segment everything everywhere all at once. *arXiv preprint arXiv:2304.06718*, 2023.