DETR with Multi-Scale Decoder

Yuan Ma, November 24

Introduction

Object detection has been a cornerstone in computer vision, evolving from traditional methods reliant on hand-crafted features and classifiers to more advanced deep learning techniques. Traditional object detection frameworks, like R-CNN[3] and its variants, often required complex, multi-stage pipelines with extensive parameter tuning. These methods, while effective, faced challenges in terms of speed, accuracy, and generalization.

Detection Transformer [2], known as DETR, emerged as a revolutionary approach, simplifying the detection pipeline by leveraging the transformer architecture, traditionally used in natural language processing. Unlike traditional detection methods, DETR facilitates end-to-end training without the need for intricate, hand-designed components. This novel approach quickly gained attention in the research community for its elegance and effectiveness. However, the baseline DETR is notorious for its low accuracy on small object and unacceptable long training schedule Moreover, its computational efficiency diminishes with high-resolution images, presenting hurdles in practical applications. In response to these challenges, Deformable DETR [12] emerged, introducing deformable multi-head attention. This adaptation enhanced the model's ability to process high-resolution images and improve small object detection. Moreover, thanks to the efficient of deformable attention, it proposed multi-scale deformable attention which makes DETR be capable of processing multi-scale features, which is proved to be very effective in countless publications. Nevertheless, with deformable attention, DETR still requires approximately 150 epochs to converge, and the computational cost is rather high.

On the other hand, instead of adopting attention mechanism, some works focusing on the role of object queries. DETR with dynamic anchor box(DAB-DETR) [7] describe role of object query as finding similarity between keys and queries and find it tend to be multi-pattern which could be harmful to the convergence. So authors give queries a 4 dimension prior information to help queries have better behaviors. Due to DAB-DETR can serve as a baseline of DETR, the authors combined the idea of Deformable DETR with DAB-DETR to get a state-of-art benchmarks in object-detection. At this point, people's exploration to object queries in DETR is well established, but the investigation towards multi-scale structure stands still, which is as same as the original deformable DETR, meaning that the computational cost is still unacceptable to some real time applications. RT-DETR is proposed to solve this issues. In stead of letting attention heads interact with all the multi-scale feature maps, RT-DETR only perform attention on the highest level of features. Then it design a FPN-PAN[6][9] like structure to melting features together to gives us a equally strong multi-scale features in similar segmentations. Although RT-DETR achieve state-of-art accuracy along with short inference time, but the way object queries interact with features maps is still following Deformable DETR, without taking advantage of dynamic queries.

In this work, we aimed to fulfill role of dynamic object queries, as well as design a novel pipeline of multi-scale structure which set the feature maps into vertical manner instead of horizontal one like other works. There are numerous decoder layers in DETR decoder, all the structures are designed in the same manner. In DAB-DETR, authors initialize the object queries by a bounding box, then update it dynamically layers by layers, while the features map stays the same. Intuitively, to locate a object, we have to take a initial guess where the object is, then we can find specific details to gives us an accurate localization. Based on this idea, we propose a vertical arrangement, introducing a unique approach where each layer of the DETR decoder is paired with distinct feature maps, enabling dynamic updates of object queries with varying levels of semantic and positional information: while the object queries is updated layers by layers, we firstly pass features with higher position information to interact with queries, then we pass high semantic information to get a accuracy position of object. Moreover, inspired by RT-DETR [8], we also use a FPN-PAN like network to generate a multi-scale maps, where the top level feature is from DETR encoder, and the others is directly from backbone(ResNet)[10]. Then we feed the features with different level into different decoder layers to dynamic update the object queries. In this way, we can dynamic update the object queries to explore the full potential of dynamic box in a very efficient pipeline.

In summary, our contributions are twofold: we present an innovative approach allowing dynamic interactions between object queries and varying feature maps, providing deeper insights into the role of object queries in DETR. Additionally, we propose a novel multi-scale DETR pipeline with a unique vertical arrangement of feature maps, offering an alternative methodology in the development of multi-scale DETR models.

Related Works

Deformable DETR: Deformable DETR represents a significant evolution in object detection, addressing some of the primary limitations of the standard DETR model. It modifies the standard DETR model by introducing deformable attention modules. The original DETR, while innovative in applying transformers to object detection, faced challenges with lengthy training times and less efficient detection of small objects. Deformable DETR addresses these issues directly. The core advancement in Deformable DETR is the implementation of deformable attention modules. These modules enable the model to focus attention on a set of key sampling points around the reference points, rather than the entire image. This selective attention mechanism is particularly beneficial for detecting small objects, a notable limitation in the original DETR. Another benefit of this modified attention is that it reduces the computation cost of multi-head attention, so that the deformable attention can extended to a multi-scale deformable attention. The multi-scale attention is applied to feature maps that concatenate

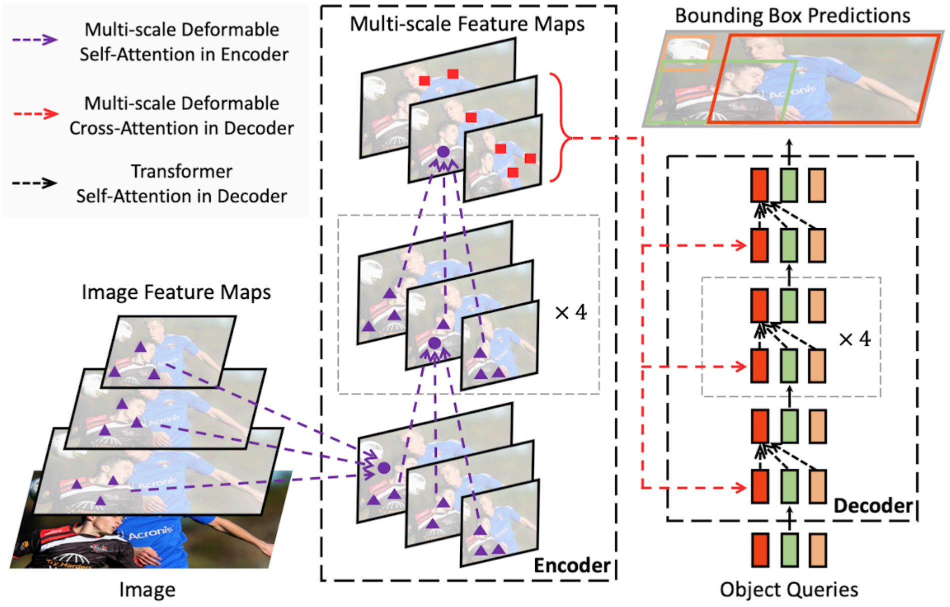
Multi-scale Deformable Self-Attention in Encoder
Multi-scale Deformable Cross-Attention in Decoder
Transformer Self-Attention in Decoder
Multi-scale Feature Maps
Bounding Box Predictions
Image Feature Maps
× 4
× 4
Encoder
Decoder
Image
Object Queries

Fig.1 Multiscale pipeline in deformable DETR

horizontally together, which it’s a huge enhance over single scale based detection. Although it still requires a huge amount of computational resource, the outstanding accuracy makes it a baseline for later multi-scale DETR. Overall, the introduction of deformable attention significantly reduces training time, making Deformable DETR more efficient than its predecessor. Additionally, its enhanced capability to detect small objects boosts overall performance, marking it as a substantial improvement over the standard DETR in practical object detection scenarios.

DAB-DETR (Dynamic Anchor Boxes Are Better Queries For DETR): Several works have proposed modifications to DETR's query design, aiming for faster training and better performance. These include Conditional DETR, which adapts queries based on content features, Efficient DETR and Anchor DETR, which associate queries with specific spatial positions. DAB-DETR is one of representation work of query design, it reveals that the slow convergence of DETR is due to multi-pattern natural of object queries in DETR because of there are multiple objects in one image. Addressing this, the proposed DAB-DETR (Dynamic Anchor Box DETR) introduces a novel approach by using 4D anchor boxes (x, y, w, h) as queries to force queries learn a detection pattern for single object, and updated layer-by-layer. This method provides better spatial priors, considering both the position and size of each box, simplifying implementation and deepening the understanding of queries in DETR. DAB-DETR effectively pools features by modulating the cross-attention map according to the anchor box size. This approach not only accelerates DETR's training but also adapts the Gaussian positional prior to match objects of varying scales, a significant advancement over previous works. Authors also proposed multi-scale deformable attention to testify the effective of DAB-DETR and it also achieve state-of-art results at that time.

Feature Pyramid Network (FPN): Traditional convolutional neural networks (CNNs) for object detection, like Faster R-CNN, often focus on extracting features at a single scale. However, this approach can be less effective for detecting objects of various sizes. The Feature Pyramid Network initially addresses this limitation by creating a pyramid of feature maps at multiple levels of resolution. This allows the network to effectively detect objects at different scales. FPNs combine low-resolution, semantically strong features with high-resolution, semantically weak features through a top-down pathway and lateral connections. **Bottom-up Pathway** is the usual convolutional network that produces a feature hierarchy with feature maps at several scales, with a scaling step of 2. The top-down pathway upsamples spatially coarser feature maps from higher pyramid levels. Lateral connections merge these upsampled maps with the corresponding bottom-up maps, which helps in preserving the fine details. FPNs are widely used in state-of-the-art object detectors like Faster R-CNN, where they improve the detection of objects at different scales. By combining features from different layers, FPN provides a rich multi-scale representation. Particularly effective for detecting small objects. FPN is a flexible architecture and can be integrated with various other deep learning models to improve performance in tasks like object detection, instance segmentation, etc.

RT-DETR (Real-Time Detection Transformer): RT-DETR aims to adapt the DETR model for real-time object detection applications. So the efficiency is at the first place to be considered. To achieve a high accuracy with efficiency, how to use multi-scale features efficiently becomes a central topic. In this paper, author did a very elaborate experiments on multi-scale features. They find out the features fusion appears in Deformable DETR is actually not optimal. The FPN like structure tend to have better accuracy with lower computational cost. Based on this experiment, they proposed a the attention-based Intra-scale Feature Interaction (AIFI) module for high level features and the CNN-based Cross-scale Feature-fusion Module (CCFM) for other features. Along with another queries selected method and hardware&software optimization, the RT-DETR achieves state-of-art performance with lightweight DETR.

Proposed Method

Multi-scale Structure Design: While significant research has been dedicated to Detection Transformers, there's been a limited focus on optimizing the multi-scale structure of DETR. Existing works, including Deformable DETR and multi-scale deformable DAB-DETR, follow similar multi-scale designs. Our work aims to introduce a novel approach in this area. Recently, RT-DETR proposed a very effective pipeline to fusing features and achieve state-art-results with less computational cost, but it still do not change the multi-scale pipeline in transformer decoder side. To design a pipeline with multi-scale features in decoder layers, we proposed a novel multi-scale decoder DETR, which enables dynamic object queries to interact with various multi-scale features across different encoder layers, shown in Fig. In our model, the first two decoder layers engage with lower-level feature maps to capture positional information of objects. The subsequent two layers interact with mid-level feature maps for semantic consistency. Finally, the last two layers, similar to the original DAB-DETR, work with the highest-level feature maps to refine accuracy in object detection.

FPN-PAN structure: The multi-scale features are derived from an FPN-PAN-like structure. This design, inspired by RT-DETR, employs a combination of upsampling and literal addition of low-level backbone features with high-level DETR encoder features. The PAN extends the FPN concept by adding an additional pathway, which in essence, is a reverse of the FPN's top-down approach. This 'bottom-up' pathway in PAN helps in refining the feature maps further. After the FPN completes its top-down enhancement, the PAN takes over. It starts with the lowest level feature map and progressively moves up the pyramid, each time combining the current level's map with a downsampled version from the level below. *The up-side down* unique structure of PAN allows for the reintegration of finer details and higher-level semantic information, resulting in more robust feature maps that can enhance object detection capabilities. By incorporating the PAN, we further refine this process. The PAN's bottom-up pathway helps in aggregating detailed information, enhancing the feature maps' quality for subsequent stages. The primary goal of this fusion strategy is to create multi-scale, semantically rich feature maps that can interact dynamically with the object queries in different

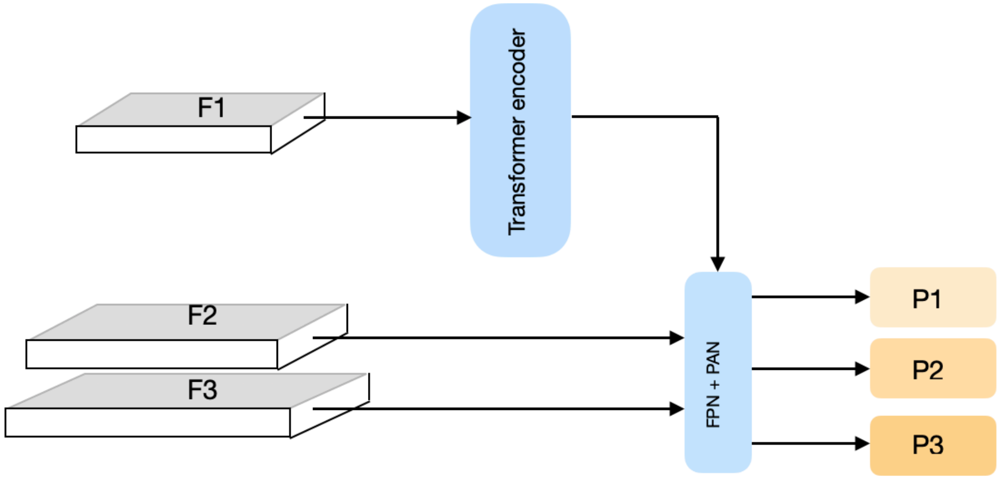

Fig2. Multi-scale feature fusion

decoder layers. This approach aims to maximize the accuracy and efficiency of object detection, especially for varying object sizes and resolutions.

Initial experiments: To testify if object queries can deal with features with features from different scale, we begin with initial experiment which add downsampled high resolution features to low level features. The initial model based is shown in Fig. Following the design of RT-DETR, we let the high level features pass through transformer encoder side, then use a 2*2 Maxpooling to reshape the different scale into same shape, which is followed by a lateral 1*1 ConvNet to shape different scale features into same dimension. Then with these features map which have different semantic information but same shape, we can simply added them in an FPN manner, gives us multi-semantic feature maps, then pass them to different layers in encoder. We can observe a significant improve in first 12 epochs: the accuracy rising from 34% to 38%, which demonstrate our idea is feasible. Furthermore, to test if the accuracy comes from the additive multi-scale features, we conduct another ablation experiments: add multi-scale features together, then let object queries to interact with these queries. The accuracy is dropped to 31%, which showcases that the accuracy improve can not be comes from additive augmented features.

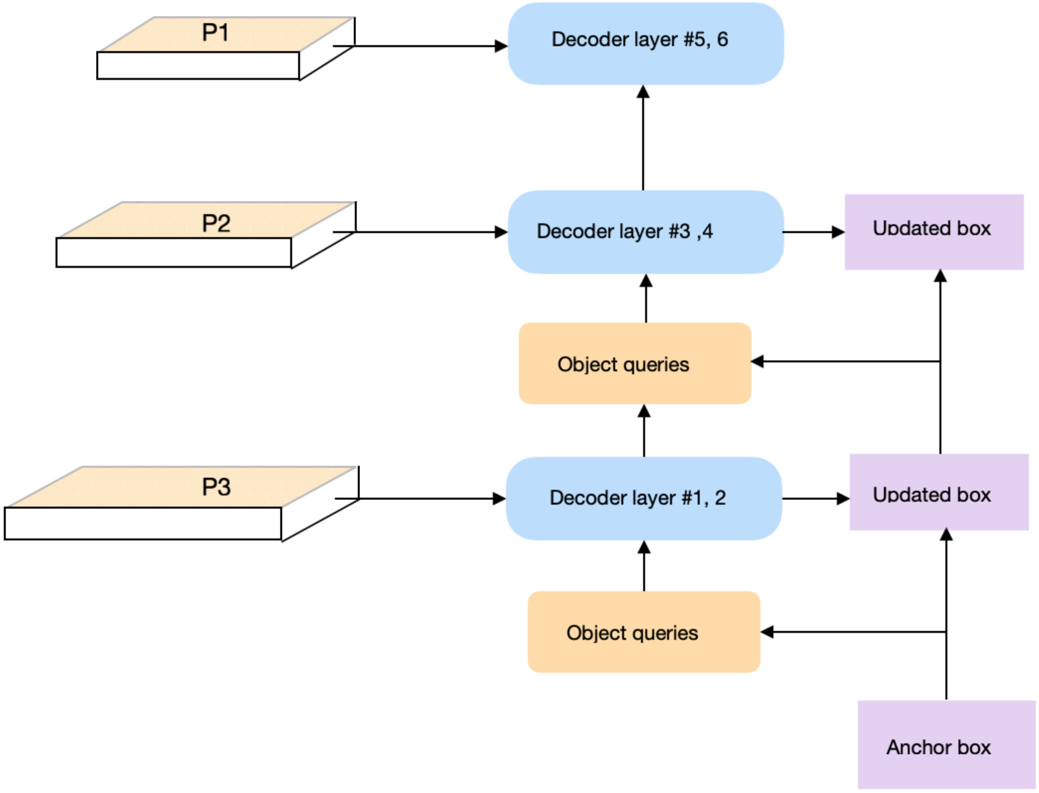

Fig.3 Proposed DETR Encoder

Based on the observation of initial experiments, here we implement our proposed model. As shown in Fig. Firstly, we take multi-scale features from P1, P2, P3 deriving from our FPN-PAN structure. Different from initial experiment which reduce the size of feature maps for simplicity.

Here we keep its pyramid structure. Then feed the features into decoder in sequential manners. At the meanwhile, we keep the design of DAB-DETR, dynamically update the object queries. Additionally, for efficiency, we apply deformable attention in first four encoder layers, combined with a standard attention in the last two layers of encoder. This design has another benefit: the deformable attention will harms model's ability to capture long dependency, while the standard attention in last two layers will solve this issues by standard multi-hand attention.

Experiments settings: We trained our model on COCO2017 dataset and the setting is the same as DAB-DETR for consistency. While we do observe a rapidly convergence in the earlier stage, the final accuracy is slightly lower than the our chosen baseline(DN-DETR). Here It could be few reason to consider: 1. The feature fusion stage is no carefully designed yet. The FPN-PAN structure might works well for features exclusively from Resnet, but it might not work so well for a mixture fusion of attention abstracted features with CNN abstracted features. We could following the feature fusion block in RT-DETR, which is carefully designed and it proved to work for DETR. 2. The downsampled processing is done by Maxpooling which might not be a ideal candidate if we want to achieve a better results, we can try other method such as light-weight CNN or local attention to down-sampled the high resolution features. 3. The size of down sampled features might have it limitations, so instead of implement a down-sampled technique, we can implement a upsample block to let object queries engaging with high resolution features.

Alternative model

Based on the analysis in previous section, we have some alternative implements for further experiment: Instead of designing a top-down multi-scale features for transformer decoder, we should also explore the multi-scale features with different arrangement. For example, we can try upside-down manners or U-shape manners to explore more possibilities and consequently to gain more insights into object queries.

Additionally, we can try different feature fusion instead of FPN-PAN structure. In RT-DETR, authors design a light weight convolutional network based feature extractor to fuse features from different scale. Here we should try similar manner to enhance our multi-scale features.

We can also investigate in leverage multi-scale features in more efficient way. In previous experiments, we testify the feasibility of additive multi-scale features can help improve the performance of DETR. So we extract features from other scale to then to add them together to serve as enhanced features. Specifically, we can use a lightweight CNN whose parameter is initialized by our object queries. Then our dynamic object queries can dynamically update the features that will benefit the model.

Conclusion

In this study, we introduced an innovative multi-scale structure approach for Detection Transformers (DETR), focusing on dynamic interactions between object queries and multi-scale features across different encoder layers. This method addresses key limitations in existing DETR models, particularly in small object detection and computational efficiency. Our proposed model, integrating Feature Pyramid Network (FPN) and Path Aggregation Network (PAN) techniques, demonstrated promising results in early training stages on the COCO2017 dataset. Although the final accuracy was slightly below the DN-DETR baseline, the initial performance improvements and rapid convergence highlight the potential of our approach. This research contributes a new perspective to multi-scale DETR models, laying groundwork for future advancements in object detection. We anticipate that our findings will inspire further exploration into dynamic query interactions and multi-scale feature representations in the field of computer vision.

Reference

[1] Xipeng Cao, Peng Yuan, Bailan Feng2, Kun Niu, CF-DETR: Coarse-to-fine transformers for end-to-end object detection. The Thirty-Sixth AAAI Conference on Artificial Intelligence (AAAI-22)

[2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In ECCV, 2020.

[3] Ross Girshick. Fast r-cnn. In Proceedings of the IEEE international conference on computer vision, pages 1440–1448, 2015

[4] Depu Meng, Xiaokang Chen, Zejia Fan, Gang Zeng, Houqiang Li, Yuhui Yuan, Lei Sun, and Jingdong Wang. Conditional detr for fast training convergence. arXiv preprint arXiv:2108.06152, 2021

[5] Feng Li, Hao Zhang, Shilong Liu, Jian Guo, Lionel M Ni, and Lei Zhang. Dn-detr: Accelerate detr training by introducing query denoising. arXiv preprint arXiv:2203.01305, 2022.

[6] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia, Path aggregation network for instance segmentation. In Pro- ceedings of the IEEE conference on computer vision and pattern recognition, pages 8759–8768, 2018.

[7] Shilong Liu, Feng Li, Hao Zhang, Xiao Yang, Xianbiao Qi, Hang Su, Jun Zhu, and Lei Zhang. DAB-DETR: Dynamic anchor boxes are better queries for DETR. arXiv preprint arXiv:2201.12329, 2022.

[8] Wenyu Lv, Shangliang Xu, Yian Zhao, Guanzhong Wang, Jinman Wei, Cheng Cui Yuning Du, Qingqing Dang, Yi Liu, DETRs beat YOLOs on real-time object detection. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2023

[9] Tsung-Yi Lin, Piotr Doll´ar, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In Proceedings of the IEEE conference on computer vision and pattern recogni- tion, pages 2117–2125, 2017.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 770–778, 2016.

[11] Zhuyu Yao, Jiangbo Ai, Boxun Li, and Chi Zhang. Efficient detr: Improving end-to-end object detector with dense prior. arXiv preprint arXiv:2104.01318, 2021.

[12] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. In ICLR 2021: The Ninth International Conference on Learning Representations, 2021.

[13] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, JunZhu, Lionel Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. In The Eleventh International Conference on Learn- ing Representations, 2022.