**STUDY THE DETECTION TRANSFORMER FOR**

**OBJECT DETECTION IN MEDICAL IMAGES**

|  |  |
| --- | --- |
| **Members** | |
| Đỗ Bá Hoàng Minh | 22BI13278 |
| Đỗ Đức Minh | 22BI13279 |
| Trần Đức Mạnh | 22BI13274 |
| Trần Bình Minh | 22BI13308 |
| Hoàng Đức Nghĩa | 22BI13329 |

**TABLE OF CONTENTS**

**I. INTRODUCTION**

**II. BACKGROUND ON DETR**

**1. General information**

**2. DETR**

**a. Simplify**

**b. DETR Architechture**

**III. METHOD**

**1. Introduction**

**2. Input Resolution**

**3. Characteristics of Medical Image Dataset**

**4. Model complexity and Multiscale Feature Fusion**

**5. Number of Object Queries**

**6. Query Initialization and Iterative Bounding Box Refinement (IBBR)**

**7. Case study setup**

**IV. RESULT**

**V. CONCLUSION**

**VI. FUTURE IMPLICATION**

**1. Potential Impacts of DETR in Clinical Practice**

**2. Further Research and Improvements**

**VII. REFERENCE**

1. **Introduction:**

Object detection entails predicting bounding boundaries and classification labels for objects. Traditional detectors employ regression and classification on a large number of suggestions, anchors, and window centers. However, a new method known as DEtection TRansformer (DETR) streamlines the procedure by anticipating the collection of bounding boxes and category labels. DETR employs a set-based global loss and a transformer encoder-decoder architecture to achieve accuracy equivalent to Faster RCNN. Transformer-based detectors have proven successful with natural pictures, but they confront particular hurdles with medical imaging data. This research assesses the appropriateness of transformer-based design options for certain case study datasets.

1. **Background on DETR**
2. **General information**

The DEtection TRansformer (DETR) is a groundbreaking object detection architecture that simplifies the traditional detection pipeline by formulating object detection as a set prediction problem. Developed by Carion et al. in 2020, it utilizes transformers, originally designed for natural language processing, to detect objects directly from images. DETR offers distinct advantages over conventional models such as Mask-RCNN and YOLO, by learning more expressive features through its transformer-based architecture and benefiting from end-to-end training for optimized performance. However, DETR's slower learning process remains a notable limitation.

1. **DETR**

**Simplify:**

DETR, or DEtection TRansformer, is a model developed for object detection tasks that combines traditional convolutional neural networks (CNNs) with transformer architectures. Introduced by Facebook AI Research, DETR simplifies the object detection pipeline by framing it as a direct set prediction problem.

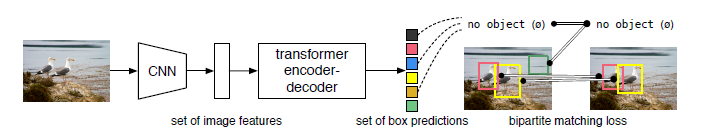
**DETR Architecture:**

The DETR architecture consists of three main components:

* Backbone Network: Produces a low-resolution activation map xs from the input image x.
* Encoder-Decoder Transformer: Processes xs to generate image tokens xf and encoded features xenc.
* Prediction Head: Receives xenc and object queries q, and generates bounding box predictions and class scores.

**Key Components:**

* Object Queries: Learnable embeddings that attend to specific regions of the image and are decoded into bounding box predictions.
* Multi-Head Self-Attention (MHSA): Enables inter-query learning and integrates encoder features.
* Multi-Head Cross-Attention: Integrates encoder features and object queries.



A DETR consists of a CNN backbone to extract a compact feature representation, an encoder-decoder transformer, and a simple feed-forward network (FFN) that makes the final detection prediction. Given an input image x ∈ , the backbone network f produces a low-resolution activation map xs = f(x) ∈ .

This map is further processed by a 1 × 1 convolution to collapse the channel dimension C into a smaller size d, resulting in image tokens xf = conv(xs) ∈ RWH×d.

To preserve spatial information in the original image, each token is paired with a positional encoding, denoted by xp ∈ RWH×d. **The encoder** is a standard attention-based transformer where each layer consists of a **multi-head self-attention module (MHSA)** followed by a **feedforward network (FFN)**. For an in-depth formalization of MHSA, refer to Appendix A.1. Typically, the DETR encoder consists of 6 layers. The encoder preserves the dimension of the input, producing xenc ∈ RWH×d.

**The decoder** receives two inputs, the encoded features xenc and N object queries q ∈ RN×d. Object queries play a central role in DETR architecture. They are learnable embeddings that work as placeholders for the potential objects in an image. Each of them attends to the specific regions of the image and is individually decoded into a bounding box prediction.

Each object query is the sum of two learnable embeddings: **content embeddings qc ∈ RN×d**, initialized as zero vectors, and **the positional embeddings qp ∈ RN×d**, indicating each query’s position. More methods for initializing object queries are discussed in Section Decoder layers consisting of a MHSA, enabling inter-query learning, and multi-head (MH) cross-attention to integrate encoder features, and a FFN.

After the decoder, each object query is independently decoded into bounding box coordinates and class scores through a three-layer FFN and a linear layer respectively.

1. **Methods**
2. **Introduction**

In this section, we outline key design choices of Deformable DETR relevant to the unique characteristics of medical images: input resolution, the number of encoder layers, multi-scale feature fusion, the number of object queries, and two techniques enhancing the decoding process, query initialization and iterative bounding box refinement (IBBR).

1. **Input Resolution**

Downsampling input images is a standard practice in detection models for computational efficiency and to satisfy memory constraints. Natural images can be significantly downsized without losing important features such as edges, shapes, and textures necessary for accurate predictions. Data from many medical imaging modalities, such as X-ray images, CT scans, and whole slide images, are at least an order of magnitude larger. These high-resolution medical images contain fine-grained details such as small lesions or slight changes in tissue density that are crucial for accurate diagnosis.

1. **Characteristics of Medical Image Datasets**

Medical image datasets have three distinct characteristics. Firstly, they are typically smaller compared to natural image datasets due to the limited number of patients with specific conditions. Secondly, medical images within one dataset are usually very homogeneous, depicting a single body part such as the brain, breast, or chest, with uniform textures and grayscale. Furthermore, while natural images contain hundreds or thousands of object classes, medical image datasets usually have much fewer classes.

1. **Model Complexity and Multi-Scale Feature Fusion**

Considering the widely accepted concept that the complexity of a model should align with the complexity of the task, we believe that simpler, less complex architectures may be more suitable for medical imaging datasets. This approach could help reduce overfitting and improve training efficiency. Moreover, objects in medical images typically have more consistent sizes compared to natural images. This observation raises doubts about the need to use multi-scale feature fusion for medical images, which is more advantageous for detecting objects of various sizes.

1. **Number of Object Queries**

In the DETR model, each object query is decoded individually to make a bounding box prediction. The total number of object queries determines how many bounding boxes are predicted per image. Most DETR models are designed for natural image datasets like MS COCO, where a single image can have up to 100 objects. However, a typical medical image usually has fewer than 10 objects, with most containing only 1 or none at all.

1. **Query Initialization and Iterative Bounding Box Refinement (IBBR)**

Object query initialization and iterative bounding box refinement (IBBR) are two techniques used in many DETR models to aid the decoding process of the object queries. These methods have been effective in improving detection performance in natural image datasets. This research aims to assess their applicability to medical data. We experimented with three initialization strategies for the positional and content embeddings of object queries: static queries, pure query selection, and mixed query selection.

1. **Case study setup**

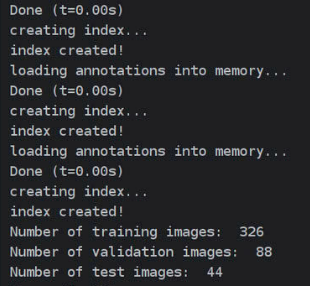
We have come up with the idea of using DETR in detecting broken and fractured bones. The advancements in deep learning have transformed the landscape of medical image analysis, ushering in groundbreaking techniques that have the potential to enhance diagnostic accuracy and patient care. However, in the domain of bone fracture detection, the application of these deep learning methods is still an emerging field. With this in mind, our team has chosen to apply DETR to the task of detecting bone fractures.

Steps involved to apply DETR in bone fracture detection:

* Dataset preparation: The dataset contains X-ray images labeled with COCO-style annotations for bone fracture detection. It is organized into three folders (test, train, and valid), with PNG files and accompanying annotation files.
* Model fine-tuning: We are currently engaged in the fine-tuning of a pre-existing DETR model to suit the requirements of the bone fracture dataset. This entails the adaptation of the model to effectively comprehend and identify fractures in X-ray images.
* Evaluation: In terms of evaluation, the model's performance will be assessed using metrics such as precision, recall, and mAP to ensure a comprehensive and rigorous evaluation process.

1. **Result**

We have developed a model using the COCO dataset, which consists of 458 images. The dataset comprises 326 training images, 88 validation images, and 44 testing images like shown in this image.



The model is functioning properly and is capable of detecting fractures in the provided images, although it has limitations in accurately bounding the fractures. Here are a few test examples.

|  |  |
| --- | --- |
|  |  |

However, the model can still precisely bound the boxes around broken spots in some cases. We’re still working on that to make the model more accurate and faster.

|  |  |
| --- | --- |
|  |  |

1. **Conclusion**

DETR (DEtection TRansformer) has shown promising results in medical image detection, demonstrating high specificity for detecting, localizing, and characterizing abnormalities in medical scans. Its ability to handle complex and high-resolution images makes it a suitable choice for medical imaging applications.

**Advantages of DETR in Medical Image Detection:**

* High accuracy in detecting abnormalities, such as tumors and lesions
* Ability to handle high-resolution images, making it suitable for medical imaging applications
* Can be used for various medical imaging modalities, including X-ray, CT, and MRI scans
* Can be fine-tuned for specific medical imaging tasks, such as breast cancer detection

**Limitations of DETR in Medical Image Detection**

* Requires large amounts of labeled data for training, which can be challenging to obtain in medical imaging
* Requiring significant resources for training and inference
* May not perform well on images with low contrast or poor quality
* The model can run normally, yet the result didn’t meet the requirements.

1. **Future Implications**
2. Potential Impacts of DETR in Clinical Practice

Integrating the DEtection TRansformer (DETR) into clinical workflows has the potential to enhance diagnostic efficiency for radiologists. By leveraging its rapid and precise fracture detection capabilities, DETR can significantly reduce diagnosis times and minimize the risk of human error, particularly in high-demand healthcare environments.

1. Further Research and Improvements

This model is still in development and requires further research and a larger dataset to fully realize its potential. Future work should focus on improving the model's speed and accuracy, optimizing its performance in clinical settings. Additionally, expanding its functionality to include the detection of soft tissue injuries and disabilities could significantly broaden its applicability in medical imaging.

Ongoing efforts in fine-tuning and expanding the dataset will be crucial in refining the model. Researchers may also explore integrating advanced techniques, such as transfer learning, to adapt the model to various medical tasks more efficiently.

1. **Reference**
2. Carion, N., Massa, F., Synnaeve, G., Douze, M., & Ponce, J. (2020). End-to-end object detection with transformers. arXiv preprint arXiv:2005.12872. <https://doi.org/10.48550/arXiv.2005.12872>
3. Liu, W., Anguelov, D., Noh, H., & Fidler, S. (2016). SSD: Single shot multibox detector. European Conference on Computer Vision (ECCV), 21-37. <https://doi.org/10.1007/978-3-319-46448-0_2>
4. Litjens, G., Kooi, T., Bejnordi, B. E., et al. (2017). A survey on deep learning in medical image analysis. Medical Image Analysis, 42, 60-88. <https://doi.org/10.1016/j.media.2017.07.005>
5. Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J., & Zisserman, A. (2010). The PASCAL Visual Object Classes Challenge 2010 (VOC2010) results. Proceedings of the British Machine Vision Conference (BMVC), 1-6. <https://doi.org/10.5244/C.24.61>
6. Esteva, A., Kuprel, B., et al. (2017). Dermatologist-level classification of skin cancer with deep neural networks. Nature, 542(7639), 115-118. <https://doi.org/10.1038/nature21056>