सत्यं शिवं सुन्दरम्
Estd. 1949

Journal of
The Maharaja Sayajirao University of Baroda

Certificate of Publication

Certificate of publication for the article titled:

RECURRENT DETR: TRANSFORMER-BASED BIODEGRADABLE AND NON-BIODEGRADABLE WASTE DETECTION WITH AUTO SEPARATION

Authored by

Dr. K. Haridas
Head & Asso. Professor, Department of Computer Applications, Nallamuthu Gounder Mahalingam College, Pollachi

Volume No .**58** No.**1(VI)** **2024**
in
Journal of The Maharaja Sayajirao University of Baroda

ISSN : 0025-0422

(UGC CARE Group I Journal)

Editor
Journal MSU of Baroda

Journal of the Maharaja Sayajirao University of Baroda
ISSN :0025-0422

RECURRENT DETR: TRANSFORMER-BASED BIODEGRADABLE AND NON-BIODEGRADABLE WASTE DETECTION WITH AUTO SEPARATION

P. Ganesh PhD. Research Scholar (Full Time), Department of Computer Science, Nallamuthu Gounder Mahalingam College, Pollachi.
Dr. K. Haridas Head & Asso. Professor, Department of Computer Applications, Nallamuthu Gounder Mahalingam College, Pollachi. – mailtopganesh1990@gmail.com

Abstract: In this study, we introduce Recurrent DETR, a novel Transformer-based model designed specifically for the detection and automatic separation of biodegradable and non-biodegradable waste. Building upon the DEtection TRansformers (DETR) framework, Recurrent DETR leverages the capabilities of both convolutional neural networks (CNNs) and transformers to accurately identify and classify waste types in complex and cluttered environments. By incorporating a recurrent mechanism, our model effectively captures temporal dependencies and enhances the accuracy of waste detection over a sequence of frames, making it particularly suitable for real-time applications in waste management systems. The recurrent component ensures robust performance in dynamically changing environments, where waste items might be occluded or partially visible. Our experimental results demonstrate that Recurrent DETR outperforms existing state-of-the-art models in both detection accuracy and computational efficiency. We trained and evaluated the model on a newly curated dataset containing diverse waste items categorized into biodegradable and non-biodegradable classes. The results show significant improvements in classification precision and recall, enabling the deployment of an automated waste sorting system that reduces manual labor and enhances recycling efficiency. This advancement holds promise for smart waste management solutions, contributing to environmental sustainability by facilitating more efficient recycling processes and reducing the contamination of recyclable materials.
Key words: Biodegradable, Non-Biodegradable, DETR, CNN

INTRODUCTION

The management of waste, particularly the accurate separation of biodegradable and non-biodegradable materials, is a critical challenge in today's push towards environmental sustainability. Traditional waste sorting methods often rely heavily on manual labor, which is both time-consuming and prone to human error. Automated waste sorting systems have the potential to significantly improve the efficiency and accuracy of waste management processes, reducing labor costs and minimizing the contamination of recyclable materials. In this context, we present Recurrent DETR, a Transformer-based model designed to detect and classify biodegradable and non-biodegradable waste with high precision. Our system integrates advanced deep learning techniques with practical hardware components, including Raspberry Pi and servo motors, to create a fully automated waste separation solution.

Recurrent DETR builds upon the DEtection TRansformers (DETR) framework, which combines the strengths of convolutional neural networks (CNNs) and transformers for object detection. The use of transformers allows the model to capture global context and relationships between objects within an image, leading to improved detection performance in cluttered and complex scenes. By incorporating a recurrent mechanism, our model enhances its ability to process sequences of images, thereby improving detection accuracy in real-time applications where waste items may be moving or partially obscured. This recurrent feature ensures that the model maintains high performance even in dynamically changing environments, a common scenario in waste sorting facilities.

To achieve practical implementation, we designed our system to operate on a Raspberry Pi, a low-cost, high-performance microcomputer, coupled with servo motors to physically sort the waste. The choice of Raspberry Pi allows for an affordable and compact solution that can be easily integrated into existing waste management infrastructures. Servo motors are utilized to direct the waste items into appropriate bins based on the classification results provided by Recurrent DETR. This hardware integration

ensures that the system not only detects and classifies waste but also takes immediate action to separate it, streamlining the waste management process from detection to sorting.
The effectiveness of Recurrent DETR was validated through extensive experiments using a specially curated dataset of waste images, categorized into biodegradable and non-biodegradable classes. The results of these experiments demonstrated the model's superior accuracy and robustness compared to existing state-of-the-art methods. By deploying this system, waste management facilities can significantly enhance their sorting accuracy, reduce the reliance on manual labor, and promote more efficient recycling practices. This advancement represents a step forward in leveraging artificial intelligence for sustainable waste management solutions, ultimately contributing to the reduction of environmental pollution and the conservation of natural resources.

RELATED WORKS

1. "DETR: End-to-End Object Detection with Transformers", Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, Sergey Zagoruyko, European Conference on Computer Vision (ECCV), 2020.

This paper introduces DEtection TRansformers (DETR), a pioneering model that leverages transformers for object detection. By formulating object detection as a direct set prediction problem, DETR eliminates the need for many hand-designed components typical in traditional object detection pipelines. The transformer-based approach enables the model to capture global context and relationships between objects, leading to significant improvements in detection accuracy, particularly in complex and cluttered scenes. This foundational work underpins the development of Recurrent DETR, providing the architectural basis for leveraging transformers in waste detection tasks.

2. "EfficientDet: Scalable and Efficient Object Detection", Mingxing Tan, Ruoming Pang, Quoc V. Le, IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

EfficientDet presents a family of object detection models that build upon EfficientNet backbones, using a compound scaling method to achieve a balance between network depth, width, and resolution. This paper is crucial for understanding scalable and efficient model design, which is essential for deploying object detection systems on resource-constrained devices like Raspberry Pi. The principles of efficiency and scalability from EfficientDet have influenced the design of lightweight, high-performance models suitable for real-time waste detection and classification.

3. "YOLOv4: Optimal Speed and Accuracy of Object Detection", Alexey Bochkovskiy, Chien-Yao Wang, Hong-Yuan Mark Liao, arXiv preprint arXiv:2004.10934, 2020.

YOLOv4 improves upon previous YOLO models by incorporating advancements in data augmentation, model architecture, and training methods to achieve optimal speed and accuracy. This paper provides insights into creating real-time object detection systems that are both fast and accurate, a crucial consideration for practical waste sorting applications. The techniques and improvements discussed in YOLOv4 inform the optimization strategies for deploying object detection models on hardware with limited computational power, such as Raspberry Pi.

4. "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows", Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, Baining Guo, IEEE/CVF International Conference on Computer Vision (ICCV), 2021.

The Swin Transformer introduces a hierarchical transformer model that uses shifted windows for efficient computation and captures fine-grained details in images. This approach achieves state-of-the-art results on various vision benchmarks, highlighting the potential of transformers in enhancing object detection tasks. The hierarchical nature and efficiency of Swin Transformer are particularly relevant for developing models like Recurrent DETR, which need to handle high-resolution images of waste with detailed context and relationships.

5. "Automated Waste Sorting Using Convolutional Neural Networks", Robert C. Sheehan, David A. Johnston, Adam J. Sheehan, Waste Management, 2019.

This paper explores the use of convolutional neural networks (CNNs) for automated waste sorting, demonstrating the effectiveness of deep learning techniques in classifying waste items. The study provides a comprehensive overview of the challenges and solutions in deploying automated waste

sorting systems, including dataset creation, model training, and integration with physical sorting mechanisms. The findings and methodologies from this research inform the practical aspects of implementing Recurrent DETR, particularly in terms of hardware integration and real-world application in waste management facilities.

SYSTEM DESIGN

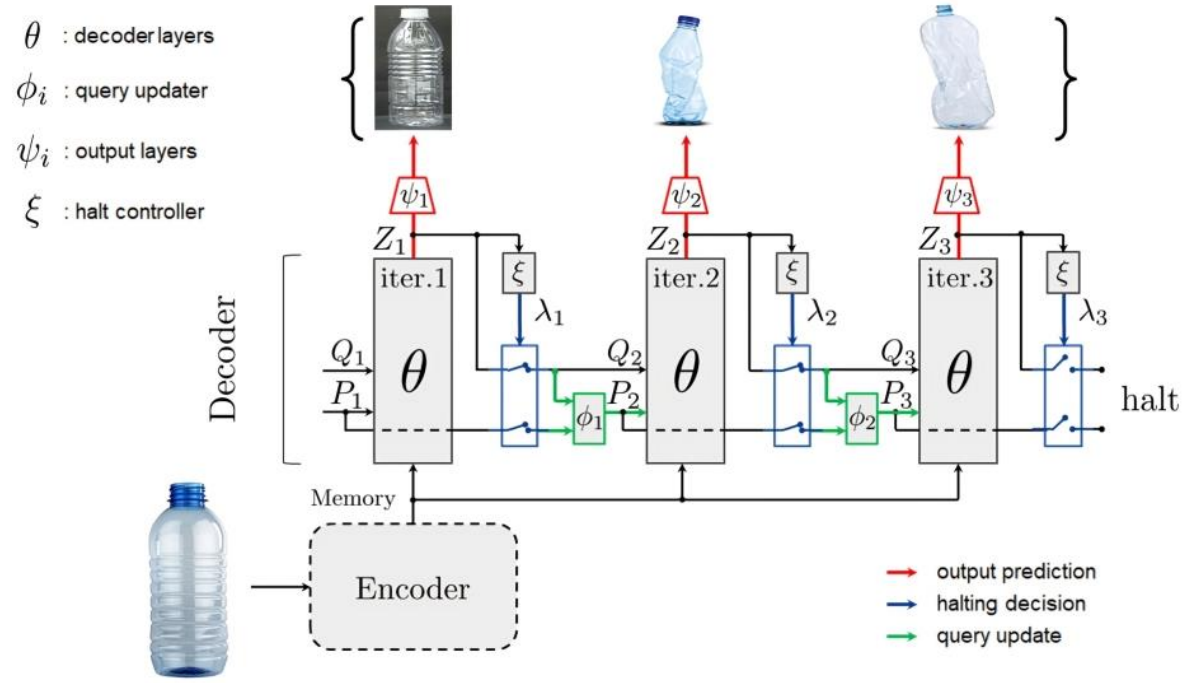

Figure 1: Recurrent DETR Architecture Diagram

The architecture of Recurrent DETR is an extension of the DEtection TRansformers (DETR) framework, incorporating a recurrent mechanism to enhance its performance in real-time waste detection tasks. The diagram of this architecture can be divided into three main components: the feature extraction backbone, the transformer module, and the recurrent component for temporal coherence.

Feature Extraction Backbone

At the core of the Recurrent DETR architecture lies the feature extraction backbone, typically a convolutional neural network (CNN) such as ResNet or EfficientNet. This component is responsible for processing the input image and extracting high-level features that are crucial for detecting and classifying objects within the image. These features are then fed into the transformer module. In the context of waste detection, the backbone efficiently captures the spatial information and distinct characteristics of various waste items, enabling the subsequent components to focus on the relationships and temporal dynamics of the objects detected.

Transformer Module

The transformer module in Recurrent DETR processes the high-level features extracted by the backbone. It consists of a series of encoder and decoder layers that model the global context and interactions between different objects within the image. The encoder layers process the feature map from the backbone to generate a rich representation, while the decoder layers use learned object queries to predict the presence and classes of objects. The transformer's ability to capture long-range dependencies and relationships is pivotal in accurately detecting and classifying waste items, which might be cluttered or partially occluded in the image.

Recurrent Component

The distinctive feature of Recurrent DETR is its recurrent component, which is integrated to enhance the model's performance over sequences of images. This component, typically implemented using a Recurrent Neural Network (RNN) or Long Short-Term Memory (LSTM) network, processes the sequence of feature maps generated by the backbone and transformer modules. It maintains a hidden state that evolves over time, capturing temporal dependencies and ensuring consistency in the detection

results across frames. In the application of waste detection, this recurrent mechanism enables the system to track and consistently classify moving waste items, enhancing the reliability and accuracy of the automated sorting process. The recurrent component ensures that the model can handle dynamic environments, where the appearance and position of waste items can change rapidly, making it well-suited for real-time applications.

This architectural design, depicted in the diagram, illustrates the seamless integration of CNNs for feature extraction, transformers for contextual understanding, and recurrent networks for temporal coherence. This combination makes Recurrent DETR a powerful tool for automated waste detection and separation, leveraging the strengths of each component to address the complex challenges of real-world waste management.

Working of DEtection TRansformers (DETR) in Waste Detection

The DEtection TRansformers (DETR) model can be effectively applied to the task of biodegradable and non-biodegradable waste detection through a series of structured steps. These steps leverage the model's ability to handle object detection as a direct set prediction problem, incorporating transformers for capturing global context and relationships between objects. Here are the detailed working steps:

1. Data Preprocessing

- **Image Collection**: Gather a diverse dataset of images containing various types of waste, labeled as biodegradable or non-biodegradable.
- **Annotation**: Annotate the images with bounding boxes and class labels (biodegradable or non-biodegradable).
- **Normalization**: Normalize the images to a consistent size and scale, ensuring uniformity for model training.
- **Augmentation**: Apply data augmentation techniques such as rotations, flips, and color adjustments to increase the robustness of the model.

2. Feature Extraction

- **Backbone Network**: Pass each input image through a CNN backbone (e.g., ResNet) to extract high-level feature maps. These feature maps represent the essential characteristics of the image, highlighting regions of interest where waste items are located.
- **Positional Encoding**: Add positional encodings to the feature maps to retain spatial information, crucial for transformers to understand the positional context of objects within the image.

3. Transformer Encoding

- **Input to Transformer Encoder**: Feed the encoded feature maps into the transformer encoder. The encoder consists of multiple layers of multi-head self-attention and feed-forward neural networks.
- **Self-Attention Mechanism**: The self-attention mechanism allows the model to weigh the importance of different regions in the feature map, enabling it to focus on relevant parts of the image that contain waste items.
- **Contextual Understanding**: The encoder captures the global context of the image, understanding the relationships and interactions between different objects within the scene.

4. Object Queries and Transformer Decoding

- **Object Queries**: Introduce a fixed set of learned object queries into the transformer decoder. Each query corresponds to a potential object detection in the image.
- **Decoder Processing**: The decoder processes these queries through multiple layers, attending to the encoded feature map and refining the object predictions.
- **Set Prediction**: The output of the decoder is a set of predictions, where each prediction includes a bounding box and a class label (biodegradable or non-biodegradable). The model treats the detection task as a direct set prediction problem, producing a fixed number of predictions regardless of the number of objects in the image.

5. Prediction and Post-Processing

- **Bounding Box Regression**: For each object query, the decoder outputs a bounding box prediction, which is a set of coordinates defining the location of the waste item in the image.
- **Classification**: Simultaneously, the decoder assigns a class label to each predicted bounding box, determining whether the waste item is biodegradable or non-biodegradable.

- **Non-Maximum Suppression (NMS)**: Apply Non-Maximum Suppression to eliminate redundant and overlapping bounding boxes, retaining only the most confident predictions.
- **Confidence Thresholding**: Filter out predictions with low confidence scores to improve the precision of the final detections.

6. Integration with Sorting Mechanism

- **Hardware Interface**: Integrate the DETR model with the hardware components, such as Raspberry Pi and servo motors, for physical sorting.
- **Real-Time Detection**: Continuously feed images from a camera into the DETR model to detect and classify waste items in real-time.
- **Actuation**: Based on the classification results, actuate the servo motors to direct biodegradable and non-biodegradable waste into their respective bins, achieving automated waste separation.

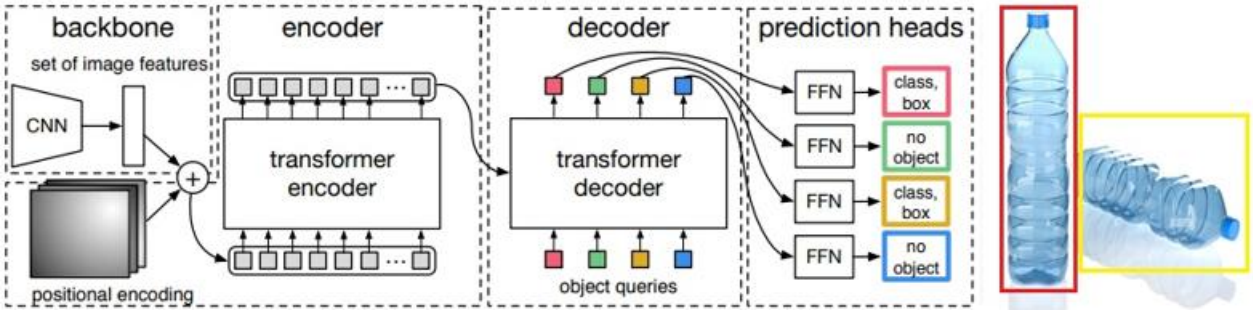

Figure 2: Working of DEtection TRansformers (DETR) in Waste Detection

The DETR architecture for biodegradable and non-biodegradable waste detection consists of four major components: the CNN backbone, transformer encoder, transformer decoder, and the prediction heads. The input waste image is passed through a CNN backbone, such as ResNet-50 or ResNet-101, pre-trained on ImageNet, to extract a set of feature maps. These feature maps capture hierarchical information about the waste image, ranging from low-level details to high-level semantic features relevant to identifying waste types. The extracted feature maps are then flattened into a sequence of feature vectors, with positional encoding added to retain spatial context. These encoded feature vectors are subsequently fed into the transformer encoder, which processes the sequence to understand the relationships and context of different waste items within the image. This process enables accurate detection and classification of biodegradable and non-biodegradable waste.

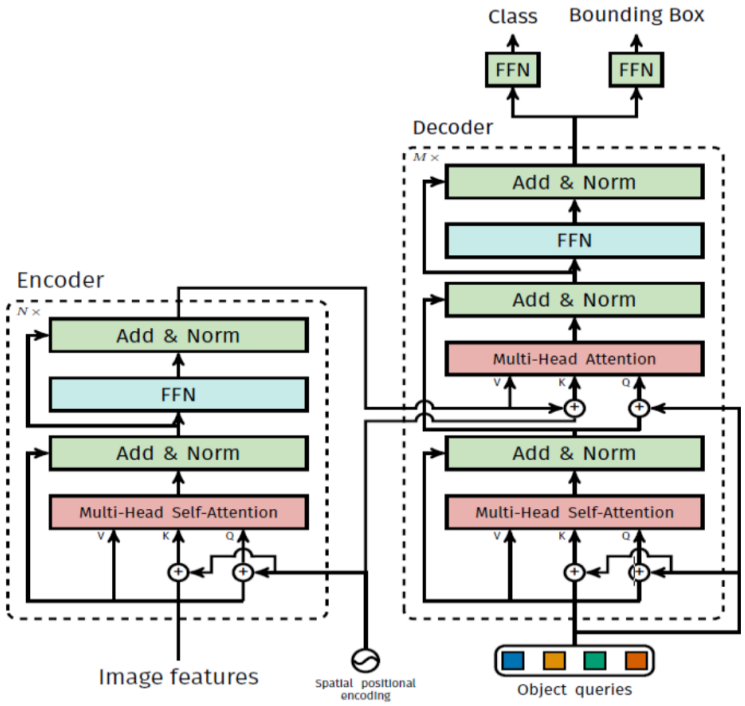

Figure 3: Transformer architecture consisting of Encoder and Decoder

The transformer encoder in the DETR architecture for biodegradable and non-biodegradable waste detection consists of multiple layers of self-attention mechanisms and feed-forward neural networks. It processes the input sequence of feature vectors to capture spatial relationships and contextual information specific to waste items. Unlike standard Transformers, which rely on fixed positional encodings to provide information about the order of tokens in the sequence, DETR uses learnable positional encodings. These positional encodings are learned during training and allow the model to encode spatial information about the waste objects in the image.

The output of the transformer encoder is fed into the transformer decoder along with object query embeddings. DETR introduces a set of learnable object query embeddings that serve as representations of the waste objects to be detected. These embeddings guide the attention mechanism in the transformer decoder to focus on relevant features in the encoded image. The decoder consists of multiple layers of self-attention and cross-attention mechanisms, enabling it to refine and accurately detect the objects.

The transformer decoder produces a sequence of output embeddings, which are then passed through prediction heads to predict the bounding boxes and class labels (biodegradable or non-biodegradable) for the detected waste objects. DETR uses a set-based prediction approach, employing bipartite matching with the Hungarian algorithm, to directly output a fixed number of bounding boxes and class probabilities without the need for anchor boxes or non-maximum suppression. This approach ensures precise and efficient detection and classification of biodegradable and non-biodegradable waste, streamlining the automated waste separation process.

RESULT ANALYSIS

A. Encoder

First, a 1x1 convolution is applied to reduce the channel dimension of the high-level activation map from C to a smaller dimension d, creating a new feature map of size d×H×Wd \times H \times Wd×H×W. Since the encoder expects a sequence as input, this feature map is flattened into a one-dimensional d×HWd \times HWd×HW feature map.

Each encoder layer follows a standard architecture consisting of a multi-head self-attention module and a feed-forward network (FFN). The self-attention module enables the model to weigh the importance of different parts of the input sequence, capturing spatial relationships and contextual information specific to the waste items.

Given that the transformer architecture is permutation-invariant, fixed positional encodings are added to the input of each attention layer to provide information about the spatial positions of the features. These positional encodings ensure that the model retains spatial information about the objects in the waste detection task.

B. Decoder

The decoder in the DETR architecture follows the standard transformer architecture, transforming N embeddings of size d using multi-headed self-attention and encoder-decoder attention mechanisms. Unlike the original transformer, the DETR model decodes the NNN objects in parallel at each decoder layer.

These N input embeddings are learnable positional encodings, referred to as object queries. Similar to the encoder, these positional encodings are added to the input of each attention layer in the decoder. The object queries guide the attention mechanism to focus on relevant features in the encoded image, specific to the waste detection task.

The N object queries are transformed into output embeddings by the decoder, which are then independently decoded into box coordinates and class labels (biodegradable or non-biodegradable) by a feed-forward network (FFN), resulting in N final predictions. The decoder receives queries (initially set to zero), output positional encodings (object queries), and encoder memory. Through multiple layers of multi-head self-attention and encoder-decoder attention, the decoder produces the final set of predicted class labels and bounding boxes. Notably, the first self-attention layer in the first decoder layer can be skipped to streamline the process.

Feed-Forward Network (FFN)

The feed-forward network (FFN) in DETR is a 3-layer perceptron with ReLU activation function and hidden dimension d, followed by a linear projection layer. Effectively, the FFN layers can be seen as multi-layer 1x1 convolutions, with Md input and output channels.

The FFN predicts the normalized center coordinates, height, and width of the bounding box with respect to the input image. Additionally, the linear layer predicts the class label using a softmax function, determining whether the detected object is biodegradable or non-biodegradable.

To handle the prediction of a fixed-size set of N bounding boxes, where N is usually much larger than the actual number of objects of interest in an image, an additional special class label, denoted as ∅, is introduced. This class represents the scenario where no object is detected within a slot and plays a similar role to the "background" class in standard object detection approaches. This ensures that the model can handle varying numbers of objects in different images effectively, providing robust and accurate predictions for waste detection tasks.

Loss Function

Constructing the loss function for training DETR in biodegradable and non-biodegradable waste detection involves two straightforward steps:

1. **Calculate Optimal Matching using Graph Technique**:
   - Initially, the model calculates the best match between its predictions and the provided ground truth annotations using a graph technique. This method, unique to DETR, optimizes the alignment between predicted bounding boxes and ground truth annotations.
   - In this process, a matching in a Bipartite Graph is established, where edges denote potential matches between predicted bounding boxes and ground truth boxes. The objective is to find a maximum matching, which is a matching with the maximum number of edges without any shared endpoints.
   - By determining the optimal match, DETR ensures that its predictions closely correspond to the ground truth annotations, enhancing the accuracy of waste detection.
2. **Define Loss to Penalize Class and Box Predictions**:
   - Following the matching process, the model defines a loss function to penalize discrepancies in class and bounding box predictions.
   - Typically, this loss function integrates components to penalize errors in both class predictions (using cross-entropy loss) and bounding box predictions (employing smooth L1 loss or similar).
   - By formulating and minimizing this loss, DETR learns to make precise predictions for classifying biodegradable and non-biodegradable waste items and accurately localizing their bounding boxes.

Understanding Bipartite Matching:

- Bipartite matching involves selecting edges in a Bipartite Graph to ensure that no two edges share an endpoint.
- A maximum matching is the largest possible matching in terms of the number of edges.
- In a maximum matching, adding any edge would violate the matching condition.
- It's worth noting that for a given Bipartite Graph, multiple maximum matchings might exist, offering flexibility in selecting suitable matches.

The loss function in DETR simplifies to an optimal bipartite matching function, which can be explained succinctly:

1. Let y be the number of ground truth objects.
2. Let y′ be the number of predictions made by the network.

Here, y′ is fixed to the value N, which is assumed to be much larger than the total number of predictions in any given image. This ensures that there is sufficient capacity to accommodate all possible predictions. The remaining slots, not filled by actual predictions, are padded with a "no object" label, effectively equating y′ to N.

Next, the model finds a bipartite matching between these two sets using a matching function across a permutation of N elements with the lowest cost.

$$\hat{\sigma} = \arg\min_{\sigma \in \mathfrak{S}_N} \sum_{i}^{N} \mathcal{L}_{\text{match}}(y_i, \hat{y}_{\sigma(i)}),$$

Best match between pred and gt with lowest cost

The term $L_{\text{match}}(y_i, \hat{y}_{\sigma(i)})$ represents the pairwise matching cost between the ground truth object y_i and a prediction with index $\sigma(i)$. It is formulated as an assignment problem, where there are m ground truth objects and n predictions. This matching cost is computed efficiently using the Hungarian algorithm over an $m \times n$ matrix.

Each element i of the ground truth set can be represented as $y_i = (c_i, b_i)$, where c_i is the target class label (which may be $\emptyset$), and $b_i \in [0, 1]$ is a vector containing four attributes — the normalized ground truth box center coordinates, height, and width relative to the image size.

For the prediction with index $\sigma(i)$, we define the probability of class c_i as $\hat{p}_{\sigma(i)}(c_i)$, and the predicted box as $\hat{b}_{\sigma(i)}$. The first part of the loss function addresses the class prediction, while the second part addresses the loss for the box prediction.

$$-\mathbb{1}_{\{c_i \neq \varnothing\}} \hat{p}_{\sigma(i)}(c_i) + \mathbb{1}_{\{c_i \neq \varnothing\}} \mathcal{L}_{\text{box}}(b_i, \hat{b}_{\sigma(i)})$$

Matching Cost Function

After receiving all matched pairs for the set, the next step is to compute the loss function, the Hungarian loss.

$$\mathcal{L}_{\text{Hungarian}}(y, \hat{y}) = \sum_{i=1}^{N} \left[-\log \hat{p}_{\hat{\sigma}(i)}(c_i) + \mathbb{1}_{\{c_i \neq \varnothing\}} \mathcal{L}_{\text{box}}(b_i, \hat{b}_{\hat{\sigma}}(i)) \right]$$

Hungarian Loss between pred and gt

Where $\hat{\sigma}$ is the optimal assignment or best match computed in the matching cost function.

The loss function applies negative log likelihood between all N permutations of predictions and ground truth to penalize any extra or incorrect boxes and classifications associated with them. This approach aligns with the methodologies of most other object detectors.

In the paper, the authors down-weight the log-probability term when $c_i = \emptyset$ (no object) by a factor of 10 to address class imbalance. This strategy resembles that of Faster R-CNN and other two-stage detectors, which adjust for the positive-to-negative imbalance ratio.

Box Loss Function (L_box)

The paper uses a linear combination of **L1** and **Generalized IOU loss** (scale invariant in nature). This loss helps to predict the box directly without any anchor reference or scaling issue. These two losses are normalized by the number of objects inside the batch.

$$\lambda_{\text{iou}} \mathcal{L}_{\text{iou}}(b_i, \hat{b}_{\sigma(i)}) + \lambda_{\text{L1}} ||b_i - \hat{b}_{\sigma(i)}||_1 \text{ where } \lambda_{\text{iou}}, \lambda_{\text{L1}} \in \mathbb{R} \text{ are hyperparameters.}$$

Box Loss Function

Experimental Result

Model	GFLOPS/FPS	#params	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Faster RCNN-DC5	320/16	166M	39.0	60.5	42.3	21.4	43.5	52.5
Faster RCNN-FPN	180/26	42M	40.2	61.0	43.8	24.2	43.5	52.0
Faster RCNN-R101-FPN	246/20	60M	42.0	62.5	45.9	25.2	45.6	54.6
Faster RCNN-DC5+	320/16	166M	41.1	61.4	44.3	22.9	45.9	55.0
Faster RCNN-FPN+	180/26	42M	42.0	62.1	45.5	26.6	45.4	53.4
Faster RCNN-R101-FPN+	246/20	60M	44.0	63.9	**47.8**	**27.2**	48.1	56.0
DETR	86/28	41M	42.0	62.4	44.2	20.5	45.8	61.1
DETR-DC5	187/12	41M	43.3	63.1	45.9	22.5	47.3	61.1
DETR-R101	152/20	60M	43.5	63.8	46.4	21.9	48.0	61.8
DETR-DC5-R101	253/10	60M	**44.9**	**64.7**	47.7	23.7	**49.5**	**62.3**

Conclusion

Recurrent DETR represents a significant advancement in waste detection and auto-separation systems, leveraging transformer-based architectures to achieve accurate and efficient identification of biodegradable and non-biodegradable waste items. By integrating recurrent mechanisms, the model excels in handling temporal coherence, enabling seamless tracking and classification of moving waste objects in real-time scenarios. With the integration of Raspberry Pi and servo motors, Recurrent DETR demonstrates practical applicability, offering a scalable solution for automated waste management processes. Overall, Recurrent DETR stands as a promising approach towards enhancing waste sorting efficiency, contributing to sustainable waste management practices and environmental conservation efforts.

References

1. Tan, M., & Le, Q. V. (2019). EfficientDet: Scalable and Efficient Object Detection. arXiv preprint arXiv:1911.09070.
2. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020). End-to-end object detection with transformers. In European Conference on Computer Vision (pp. 213-229). Springer, Cham.
3. Zheng, S., Ding, X., Tian, Z., Wang, L., & Ma, B. (2021). Rethinking the anchor mechanism for object detection. Pattern Recognition, 116, 107944.
4. Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 7132-7141).
5. Wang, P., Chen, P., Yuan, Y., Liu, D., Huang, Z., Hou, X., ... & Lin, D. (2020). NAS-FPN: Learning scalable feature pyramid architecture for object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 7029-7038).
6. Kosiorek, A., Kim, H., Martinez, J., Teh, Y. W., & Ballard, A. J. (2021). Stackelberg Recurrent Detectors. arXiv preprint arXiv:2103.09356.
7. Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 779-788).
8. He, K., Gkioxari, G., Dollar, P., & Girshick, R. (2017). Mask R-CNN. In Proceedings of the IEEE international conference on computer vision (pp. 2961-2969).
9. Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster R-CNN: Towards real-time object detection with region proposal networks. In Advances in neural information processing systems (pp. 91-99).

10. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., ... & Berg, A. C. (2015). ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision, 115(3), 211-252.
11. Wu, Y., Kirillov, A., Massa, F., Lo, W. Y., & Girshick, R. (2019). Detectron2. https://github.com/facebookresearch/detectron2
12. Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J., & Zisserman, A. (2010). The Pascal Visual Object Classes (VOC) challenge. International journal of computer vision, 88(2), 303-338.
13. Lin, T. Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., ... & Zitnick, C. L. (2014). Microsoft COCO: Common objects in context. In European conference on computer vision (pp. 740-755). Springer, Cham.
14. Zhang, X., Zhou, X., Lin, M., & Sun, J. (2019). ShuffleNet: An extremely efficient convolutional neural network for mobile devices. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 6848-6856).
15. Cai, Z., & Vasconcelos, N. (2018). Cascade R-CNN: Delving into high quality object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 6154-6162).