INTERNATIONAL JOURNAL FOR INNOVATION EDUCATION AND RESEARCH
ISSN: 2411-2933

Journal URL: https://scholarsjournal.net/index.php/ijier

DOI: https://doi.org/10.31686/ijier.vol12.iss4.4220

Transformer-Based 3D Object Detection

Jiayin Li, Yixin Ma, Jiagu Pan, Xing Xu

School of Electronic and Electrical Engineering, Shanghai University of Engineering Science, Shanghai 201620, China

Abstract

This paper mainly studies object detection methods based on Transformer. Transformer, as a natural language processing technology, is widely used in computer vision tasks such as image classification and object detection. This paper introduces an object detection method based on scale point cloud Transformer, which provides a new research direction for object detection in the future.

Keywords: Transformer; Object Detection; Computer Vision; Point Cloud; Self-Attention Mechanism

1. Introduction

1.1 Importance of Object Detection

Human object detection is very fast and precise. Our brains are designed to recognize elements as a reflection of our visual system. But with the development of computer vision technology, the optical maze of vision has been simplified by technology.

With the rise of emerging technologies such as artificial intelligence, the Internet of Things, and quantum computing, computer scientists have been able to clone our thought processes into computers. This phenomenon is commonly known as object detection, and it has largely eliminated the reliance on humans in various industries.

1.2 Application of Object Detection in the Field of Computer Vision

Object detection is also a key concept behind autonomous driving technology. In addition to using it for image recognition software, many car companies HYPERLINK "https://www.g2.com/categories/image-recognition" are using it to power vehicles equipped with artificial intelligence sensors for safe driving, detecting traffic conditions after the fact, creating 3D maps, and driving in autonomous situations. Navigate down.

2. Introduction to Transformer

2.1 Background of Transformer

In October 2018, Google issued a paper "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". The BERT model was born and swept the best results in 11 tasks in the NLP field! It plays an important role in BERT The structure is Transformer. Later, XLNET, BERT and other models appeared one after another, defeating BERT, but their core has not changed, it is still: Transformer.

2.2 Basic principles of Transformer

In recent years, Transformers have made a splash in the field of computer vision . Transformer was originally proposed by Vaswani et al. in 2017 to solve machine translation tasks. Its core structure consists of two parts: Encoder and Decoder. Each part is stacked by multiple identical layers, each containing a self-attention mechanism and a feed-forward neural network.

The self-attention mechanism allows the model to focus on all other elements while processing each input element, thereby capturing dependencies in the sequence. The feedforward neural network is responsible for further processing the output of the self-attention mechanism.

3. Multi-Scale Point Cloud Transformer

3.1 Point Cloud

Point cloud is a data structure that represents objects in three-dimensional space. It consists of many discrete points. Each point has its own position coordinates and possibly other properties such as color, normal vector, intensity, etc. Point clouds are typically captured by laser scanners, cameras, or other sensors and used to create 3D models, maps, or perform remote sensing analysis. In the field of computer vision and machine learning, point cloud is also widely used in target detection, object recognition, 3D reconstruction and virtual reality.

3.2 Multi-Scale Point Cloud

An important property of multiscale point clouds is that their distribution and density in space vary at different scales. For example, some point clouds may be concentrated on the surface of an object, while other point clouds may cover a larger space but be less dense. In addition, these point clouds can also contain different precision information. For example, high-precision point clouds may contain more detailed structures and details, while low-precision point clouds may only contain rough shape and position information. An important application of multi-scale point clouds is 3D reconstruction and measurement. This data can be used to reconstruct a three-dimensional model of the scene and perform analysis and identification. For example, this data can be used to detect features on object surfaces, perform spatial analysis and navigation, perform machine vision applications, and more. At the same time, multi-scale point clouds are also commonly used in virtual reality and augmented reality applications to provide accurate three-dimensional spatial data for these technologies.

3.3 Self-Attention Mechanism

In the Transformer model, the self-attention mechanism is a key component used to establish the dependence between elements in the input sequence. The self-attention mechanism allows the model to interact with each element with all other elements when processing sequence data to better understand the meaning of each element with the help of contextual information. The advantage of the self-attention mechanism is that it can model the relationship between any two elements in the input sequence without being restricted by sequence distance. This enables the model to better capture long-distance dependencies, thereby

improving sequence modeling capabilities. In addition, the self-attention mechanism can be calculated in parallel and is therefore more efficient in practice.

To sum up, the self-attention mechanism allows the model to dynamically perform weighted aggregation of each element in the input sequence based on the relationship between the elements, thereby achieving better sequence modeling and representation learning. This enables the Transformer model to achieve important breakthroughs in sequence tasks such as natural language processing.

3.4 System Framework Based on Multi-Scale Point Cloud

A 3D object detection system based on multi-scale point cloud Transformer. The framework consists of two stages. The first stage is to better learn the local geometric information of the point cloud, establish attention between different scale information, and obtain the correlation between points. property, a multi-scale neighborhood embedding module based on coordinates is designed. This module first extracts features point by point, then uses the farthest point sampling to downsample the point cloud, and finally uses Euclidean distance to downsample each point in the original KNN search is performed in the point cloud, and finally the local neighborhood geometric features are calculated; the skip connection offset attention module first obtains different levels of semantic information, and global semantic features are obtained after maximum average pooling; then local neighbors are aggregated based on the sampled foreground points. The domain geometric information and global semantic features generate an initial 3D bounding box from near to far; the second stage is to optimize the position, size, orientation and confidence of the bounding box, coordinate transformation of the points in the initial bounding box obtained in the stage, and then through MLP obtains the local features of the area of interest, and combines the learned local features with point cloud depth information, reflection intensity, single point features at this stage, multi-scale local geometric information of the point cloud in the area of interest, and global semantic features to optimize 3D Bounding box information. Compared with some existing methods, this project fully considers the correlation between local points, makes full use of effective features through the attention mechanism, and pays more attention to the detection effect of smaller and farther objects.

4. Future Research Directions for Object Detection Based on Transformer

Transformer-based object detection technology has made significant progress in recent years, but there are still some challenges and potential research directions. Here are some possible research directions:

(1) Small target detection: The Transformer model has certain challenges when dealing with small target detection problems, because small targets may be difficult to be accurately detected by the model due to their small size. Future research can explore how to improve the Transformer model so that it can better handle small target detection problems.

(2) Real-time: Although the Transformer model has achieved good results in object detection, its computational complexity is high, resulting in certain limitations in real-time applications. Future research can try to design a more efficient Transformer structure or optimization algorithm to improve the real-time performance of the model.

(3) Memory consumption: Due to its self-attention mechanism and decoder structure, the Transformer model usually requires larger memory to store intermediate calculation results. Future research can explore how to reduce memory consumption, such as through model pruning, quantization and other techniques.

(4) Multi-modal information fusion: The Transformer model can be combined with other types of models (such as convolutional neural networks) to fuse information from different modalities and improve the performance of object detection. Future research can explore how to effectively fuse multi-modal information to further improve the accuracy and robustness of object detection.

(5) Cross-domain adaptability: Transformer models may have performance differences in different data sets and application scenarios. Future research can explore how to improve the cross-domain adaptability of the model so that it can achieve good performance in different environments and data sets.

(6) Hardware acceleration: In view of the computing and memory requirements of the Transformer model, future research can explore how to use specific hardware (such as GPU, TPU, etc.) for acceleration to improve the performance of the model in practical applications.

In short, Transformer-based object detection technology still has a lot of room for research and improvement in the future. By constantly exploring new methods and algorithms, we can further improve the performance of object detection and meet the needs of more practical applications.

5 Conclusion

5.1 Summary of Thesis

Transformer provides the possibility to implement efficient and accurate target detection networks, and conducts in-depth research on the operating mechanism of deep neural networks. Traditional methods usually use CNN-based point cloud processing processes, but due to the disorder and variable length of point clouds, CNN may have some limitations when processing point clouds. The Transformer model can directly process point cloud data through the self-attention mechanism, model each point of the point cloud, and capture the relationship and contextual information between points. This gives the Transformer model certain advantages in point cloud classification tasks.

5.2 Looking to the Future

As computing power continues to improve, Transformer models are likely to become larger in the future. Larger models can capture more complex patterns and relationships, thereby increasing the expressive power of the model. At the same time, larger data sets will also provide richer information for the Transformer model, promoting the generalization ability and performance improvement of the model. Overall, the Transformer model, as a powerful sequence modeling and representation learning framework, has broad application prospects in various fields. Future development will focus on model scale, richness of data sets, cross-modal and multi-modal modeling, reinforcement learning, and domain-specific optimization, etc., thereby further promoting the application and innovation of Transformer models in the fields of artificial intelligence and machine learning.

6 References

[1] Liu S., Cao Y., Huang W., etc. Radar point cloud segmentation integrating sparse attention and instance enhancement [J]. Chinese Journal of Image and Graphics, 2023, 28(02): 483-494.

[2] Zhou J., Hu Y., Hu C., et al. Weakly perceptual target detection method based on point cloud completion and multi-resolution Transformer [J/OL]. Computer Applications: 1-13 [2023-03-27].

[3] Han L., Gao Y., Shi Z. Radar point cloud three-dimensional target detection based on sparse Transformer [J]. Computer Engineering, 2022, 48(11): 104-110+144.

[4] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., et al. Attention is all you need. In Advances in neural information processing systems, 2017:5998-6008.

[5] Devlin, J., Chang, MW, Lee, K., et al. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019, 1:4171-4186.

[6] Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In International Conference on Learning Representations, 2021.