A Review of Image location Information

Wang Han, Fu Rui

***Abstract*— The absolute position information of the image content has always been a key issue discussed by people. In the image convolution, people have been studying how to extract the absolute position information of the image, and there has been relatively great progress. With the rapid development of Transformer in recent years, people began to compare two different feature extractors. This article discusses the advantages and disadvantages of the two feature extractors starting from the absolute position information between the image features. And introduced the latest developments in the extraction of image position information for these two different networks in recent years**

***Index Terms*— CNN, Transformer**

INTRODUCTION

Before, CNN and absolute position, these two concepts were rarely discussed together. Because there are two reasons: one is that everyone generally believes that: CNN is translation invariant (for classification tasks), or translation is equivalent (for segmentation and detection tasks); second, there is no specific task requirement. For example, the three major object perception tasks of computer vision, classification, segmentation, and detection. Object classification has nothing to do with position, semantic segmentation, as pixel-level semantic classification, does not depend on position. Object detection tasks that are most likely to be related to absolute positions are decoupled from absolute positions by mainstream methods and become relative to anchor boxes or anchors Point to perform local relative position regression. In this way, the network itself does not need to know the absolute position of the object, and the position information is used as an artificial prior in the pre- and post-processing for coordinate conversion. However, absolute position information is very valuable in many tasks. For example, in problems such as instance segmentation, the object and absolute position can uniquely determine the instance. An obvious observation is that the human visual system can easily know the absolute position, such as: "There is a bird in the upper left corner, and it flies to the right again". Moreover, for objects in the image, different instances are essentially distinguished by position and shape.

Aiming at the translation invariant of the convolution operation, a semi-convolutional operation (semi-convolutional) is proposed [1], and a certain theoretical analysis is carried out. At the same time, this semi-convolutional can also be combined with the pose & verify method such as Mask RCNN, and experiments are also done in the article. Finally, the author conducted experiments on artificial images and VOC data sets, and both achieved better results than Mask RCNN. Convolution is the most used operation for extracting features in deep learning, but due to translation invariance, convolution is not sensitive to the spatial location features of the image, which leads to the same target (such as two pedestrians), the features extracted by convolution The separability is poor. Therefore, the convolution operation can still play a good effect for semantic segmentation, but it has some shortcomings in instance segmentation. This article solves this problem by simply modifying the convolution operation and introducing the spatial characteristics of the picture.

The CoordConv structure developed by uber can also solve part of the coordinate problem [2]. The author studied and analyzed a common defect of convolutional neural networks, that is, it cannot convert spatial representations into coordinates in Cartesian space and coordinates in one-hot pixel space. Convolution is equivariant, which means that when each filter is applied to the input, it does not know where each filter is. We can help convolution and let it know where the filter is. This process requires adding two channels to the input, one at the i coordinate and the other at the j coordinate. The target detection model is also looking in the pixel block space, and the output is the bounding box in Cartesian space, so it seems that CoordConv is very suitable. The author also found that on the simple MNIST data set detection, the IOU score of the Faster-RCNN network increased by about 24%.

The original image features learned by the convolutional network can be visualized by CAM for saliency regions. A simple experiment was carried out in this article, a picture was cropped to test the changes in the salient area before and after cropping [3]. Theoretically, in the pictures before and after cropping because the features are the same, the saliency area of the common object should be unchanged, but in practice, it is found that the saliency area has also shifted after cropping. This uses the translation invariance of the convolutional neural network. It is difficult to explain, and it is suspected to be caused by location information. The article is experimented, input any picture (the image is content), and train the network to output location-related pictures. For example, input a noisy picture, and hope that the network will output a horizontal coordinate map (location information). Specifically, in the case of zero-padding, models based on VGG and ResNet can predict relatively reasonable position-related outputs, such as abscissa or ordinate. Without padding, the output will only respond directly to the input content, and location information that has nothing to do with the content cannot be predicted. The location information is implicitly encoded in the classified convolutional neural network structure without any explicit supervision. Through experimental analysis, the author tested the control variable of the padding number in the Position Encoding Module

Manuscript received Nov 13, 2021

Wang Han: Master's degree in TianGong University. Research area: image processing.

Fu Rui: Master's degree in Taiyuan University of Technology University.

network and found that the padding increased, and the location information became more significant. At the same time, the experiment using the VGG network has further confirmed that padding introduces location information. From the visualization in the above figure, we can also see the effect of padding introducing position information, especially the final VGG, whether there is padding or not has a very large impact. Similarly, for object detection and semantic segmentation tasks, experiments have found that the performance drops a lot after removing the padding operation.

The key to the features captured by CNN is the sliding window covered by the convolution kernel, and the features that CNN can capture are all in this sliding window. This is the reason for increasing the depth of the network, capturing long-distance features. The CNN convolution kernel can retain the relative position between the features. The sliding window slides from left to right, and the captured features are also arranged in this order, so it has already recorded the relative position information in the structure. But if the Pooling layer is connected immediately after the convolutional layer, the location information will be thrown away.

But Transform has developed rapidly in recent years. Transformer is not like RNN or CNN, it must explicitly encode position on the input side. Transformer uses position function to encode position. In general, there are mainly two model architectures in the related work of using Transformer in CV. One is a pure transformer structure, and the other is a hybrid structure that combines CNN backbone network and Transformer. Using Transformers for vision tasks became a new research direction for the sake of reducing architecture complexity and exploring scalability and training efficiency.

Vision Transformer (ViT) can achieve excellent results with pure transformer architecture applied directly to a sequence of image patches for classification tasks [4]. It also outperforms the state-of-the-art convolutional networks on many image classification tasks while requiring substantially fewer computational resources (at least 4 times fewer than SOTA CNN) to pre-train. Position embeddings are added to the image patch embeddings to retain spatial/positional information in a global scope with different strategies. In the paper, they tried different ways to encode the spatial information, including no positional information, 1D/2D positional embeddings, and relative positional embeddings. One of the interesting findings is 2D positional embeddings did not bring significant performance gains when compared with 1D positional embeddings.

Detection Transformer (DETR) is the first object detection framework that successfully used Transformer as the main building blocks in the pipeline [5]. It matches the performance of the previous STOA methods (highly optimized Faster R-CNN) with a much simpler and flexible pipeline. Here is the flow: 1.CNN is used to learn 2D representation of an image and extract the features. 2. The output of the CNN is flattened and supplemented with positional encodings to feed into standard Transformer's encoders. 3.The Transformer's decoder pass the output embeddings to a feed forward network (FNN) for predicting the class and bonding box. Transformer's great success in NLP has been explored in the computer vision domain and became a new research direction. Transformer is proved to be a simple and scalable framework for computer vision tasks like image recognition, classification, and segmentation, or just learning the global image representations. It demonstrated significant advantage in training efficiency when compared with traditional methods. In terms of architecture, it can be used in a pure Transformer manner or in a hybrid manner by combining with CNNs. It also faces challenges, like low performance on detecting small objects in DETR and also did not perform well when the pre-training dataset is small in Vision Transformer (ViT). Transformer is becoming a more general framework for learning sequential data, including text, image, and time-series data. This is just an early glimpse, looking forward to seeing the new emerging things with the increasing convergence of the NLP and CV.

REFERENCES

[1] Novotny D, Albanie S, Larlus D, et al. Semi-convolutional operators for instance segmentation[C]//Proceedings of the European Conference on Computer Vision (ECCV). 2018: 86-102.

[2] Liu R, Lehman J, Molino P, et al. An intriguing failing of convolutional neural networks and the coordconv solution[J]. arXiv preprint arXiv:1807.03247, 2018.

[3] Islam M A, Jia S, Bruce N D B. How much position information do convolutional neural networks encode? [J]. arXiv preprint arXiv:2001.08248, 2020.

[4] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale[C]//International Conference on Learning Representations. 2020.

[5] Carion N, Massa F, Synnaeve G, et al. End-to-end object detection with transformers[C]//European Conference on Computer Vision. Springer, Cham, 2020: 213-229.