Advances in Deep Neural Networks for Vision: A Review

Introduction

Deep Neural Networks (DNNs) play a critical role in computer vision. Frameworks like R-CNN, YOLO, and SSD have pushed boundaries in object detection, and models like U-Net, Mask R-CNN, and DeepLab excel in semantic and instance segmentation tasks. In this essay, we will review the development of the R-CNN family while introducing a new framework, Vision Transformers (ViT).

R-CNN Family

R-CNN

R-CNN (Region-Based Convolutional Neural Network) is a family of object detection models designed to locate and classify objects within an image. The original R-CNN combined region proposals with Convolutional Neural Networks (CNNs) and introduced an auxiliary approach to fine-tune pre-trained models to tackle tasks with scarce labeled training data. R-CNN uses an external algorithm (e.g., Selective Search) to generate category-independent region proposals, and uses a large CNN to extract a fixed-length feature vector from each region. It then uses a set of linear Support Vector Machines (SVMs) to classify these features into object categories. Finally, R-CNN applies box regression to improve localization accuracy.

Fast R-CNN

R-CNN is computationally expensive due to separate stages for region proposal, feature extraction, and classification. It also has slow inference speed, as the CNN processes each region proposal independently.
To overcome these drawbacks, Fast R-CNN proposed a Region of Interest (RoI) Pooling Layer, which converts variable-sized region proposals into fixed-size feature maps, enabling end-to-end training, except the region proposal part. The network first processes the whole image with several convolutional and max pooling layers to produce a convolutional feature map. Then, using the RoI pooling layer, it extracts a fixed-length feature vector from the feature map. After that, Fast R-CNN uses a fully connected layer to branch the feature vector into two sibling output layers: one produces

softmax probability to estimate object category, and another layer that outputs the refined bounding box.

Faster R-CNN

Although Fast R-CNN has a significant performance advance compared to R-CNN, it still uses Selective Search for region proposals, which is computationally expensive and not learnable. This issue was addressed by Faster R-CNN, which introduced a Region Proposal Network (RPN). RPN is a fully convolutional network that simultaneously predicts object bounds and objectness scores at each position. The RPN network was merged with Fast R-CNN by sharing their convolutional features using attention mechanisms, effectively telling the unified network where to look. With this architecture, Faster R-CNN truly performs end-to-end training and has a much faster inference speed.

Mask R-CNN is an extension of the R-CNN family, tailored for instance segmentation. It adds a branch to predict a segmentation mask for each detected object, making it suitable for pixel-level object segmentation. It builds on Faster R-CNN, adds a small Fully Convolutional Network (FCN) to predict a binary mask for each object, and replaces RoIPool with RoIAlign to improve spatial alignment of features and mask predictions.

ViT

While the R-CNN family was advancing and expanding, a new approach using "attention" emerged: the Vision Transformer (ViT). It applies the Transformer Architecture, which was originally developed for natural language processing, to the domain of computer vision.

Instead of processing pixels locally using convolutional filters, ViT divides an image into fixed-size, non-overlapping patches that are flattened and linearly embedded into a vector. These image patches are treated as tokens and passed through a Transformer encoder, which consists of multiple layers of self-attention and feed-forward networks. This helps capture global dependencies and relationships between distant parts of the image. Position embeddings are added to the patch tokens to preserve information about their relative positions in the image. The output of the encoder is used by a fully connected layer for image classification.

ViT can learn the global context of the input image and scales well with the size of the dataset and model, outperforming CNN-based models when trained on large datasets. However, it still has some limitations. It requires a large amount of training data to achieve high performance and has high computational and memory requirements.

Conclusion

The evolution of object detection and image segmentation models, from R-CNN to Vision Transformers, demonstrates the rapid progress in computer vision technologies. Each iteration of the R-CNN family addressed previous limitations, progressively improving accuracy, speed, and versatility. The emergence of Vision Transformers represents a paradigm shift, leveraging attention mechanisms to capture global image context. While challenges remain, particularly in computational efficiency and training data requirements, these advancements showcase the incredible potential of deep learning in transforming how machines perceive and understand visual information. As research continues, we can expect further innovations that push the boundaries of computer vision, making artificial visual perception increasingly sophisticated and reliable.