ESILV
ENGINEERING SCHOOL
DE VINCI PARIS

Machine Learning for Computer Vision

TD3: Vision Transformer (ViT) and object detection

Département d'informatique

Ruiwen HE

Course Review:Vision Transformer (ViT) and object detection

Vision Transformer (ViT)

Overview

- Adaptation of the Transformer architecture, originally designed for NLP, to computer vision tasks.
- Divides an image into patches and processes them as a sequence of tokens.

Key Concepts

- **Patch Embeddings:** Images are split into fixed-size patches (e.g., 16×16) and flattened to form patch embeddings.
- **Self-Attention Mechanism:** Models global dependencies across all patches, enabling better understanding of spatial relationships.
- **Positional Encoding:** Maintains spatial information within the patches.

Strengths

- Excellent at learning global features compared to CNNs, which focus on local features.
- Scalable with large datasets (e.g., ImageNet-21k).

Limitations

- Requires a large amount of data for effective training.
- Computationally expensive compared to traditional CNNs.

Applications

- Image classification, segmentation, and other computer vision tasks.

Object Detection

Definition

- Identifying and localizing objects within an image, typically by generating bounding boxes and classifying objects.

Popular Models

- **YOLO (You Only Look Once):**
  - Real-time object detection.
  - Divides images into grids and predicts bounding boxes and class probabilities in one pass.

- **DETR (DEtection TRansformer):**
  - Combines transformer-based attention mechanisms with object detection.
  - Processes images as a sequence, eliminating the need for anchor boxes.

Comparison Between YOLO and DETR

- **YOLO:**
  - Faster inference.
  - Suited for real-time applications.
  - Anchor-based detection with predefined sizes.
- **DETR:**
  - Better at modeling complex spatial relationships.
  - Requires more training time but removes the need for anchors.

Applications

- Autonomous vehicles, surveillance, medical imaging, and retail analytics.

Takeaways

- **ViT** and **DETR** are examples of how transformer-based architectures are reshaping vision tasks, providing flexibility and global context understanding.
- **YOLO** remains a robust choice for speed-critical applications.
- A combination of methods (e.g., ViT for classification, YOLO/DETR for detection) can cater to diverse real-world needs.

Part I: Basics of Image Segmentation and Evaluation Metrics

Introduction

This practical session is divided into three exercises focusing on model comparisons, training object detection models, and evaluating their performance on video data.

Exercise 1: Comparison of ViT and Traditional CNN Models

- **Objective:** Analyze the performance of Vision Transformer (ViT) and traditional CNN models for classification tasks.
- **Tasks:**
  1. Train ViT and CNN models on a specific dataset.
  2. Evaluate their classification performance using metrics such as accuracy, recall, and F1-score.
  3. Compare results obtained with pre-trained models versus models trained from scratch.

Exercise 2: Training and Testing YOLOv5

- **Objective:** Train the YOLOv5 model on a custom dataset and evaluate its performance on video data.
- **Tasks:**
  1. Configure the YOLOv5 model using a dataset obtained via the https://public.roboflow.comRoboflow API.
  2. Adjust model parameters (e.g., number of classes, train/validation split) and train the model.
  3. Validate the model on video data and generate a video showcasing its predictions.

Exercise 3: Training and Testing DETR

- **Objective:** Train and evaluate the DETR (DEtection TRansformer) model on a similar dataset.
- **Tasks:**
  1. Train the DETR model on a dataset comparable to the one used for YOLOv5.
  2. Test the trained DETR model on video data.
  3. Generate a video illustrating the model's predictions.

Bonus Opportunity

If you submit the final test video with the predicted results back to us, additional bonus points will be awarded. This will ensure models are evaluated on a unified dataset, offering a clearer demonstration of their practical effectiveness.