Towards Embodied Intelligence, A Comprehensive review of Computer Vision Architectures

Renan Monteiro Barbosa

2024

Advances in Computer Vision

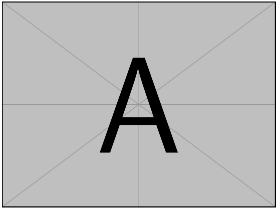

Figure: Title of Image 1

Figure: Bigger Data

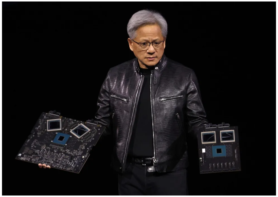

Figure: Faster GPUs

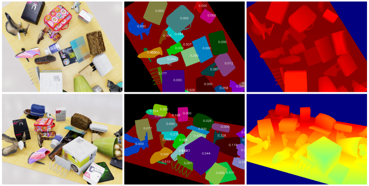

Figure: More powerful models

Real World Data is Imperfect

Challenge: Real world data is imperfect

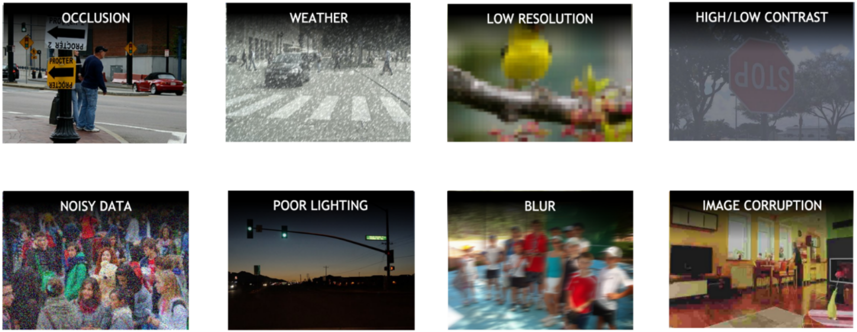

Figure: Different types of imperfect and noisy real-world data create challenges for image analysis.

Introduction

In recent years, computer vision (CV) has made significant strides in performance due to advancements in deep learning models. This presentation covers the evolution of CV models, starting with YOLO (You Only Look Once) and moving towards Vision Transformers (ViT).

YOLO: You Only Look Once

- Introduced by Redmon et al. in 2016, YOLO is a real-time object detection system.
- YOLO performs both object localization and classification in a single forward pass of the network.
- Significantly faster than previous models, making it ideal for real-time applications.
- The architecture is a single convolutional neural network (CNN) that predicts bounding boxes and class probabilities directly from image pixels.

YOLO Architecture

- YOLO divides the input image into a grid.
- Each grid cell predicts bounding boxes, confidence scores, and class probabilities.
- The model outputs a fixed number of bounding boxes, each with a class label and confidence score.
- YOLO's speed and accuracy led to its widespread adoption in applications like self-driving cars, security systems, and robotics.

The Evolution of YOLO

- YOLOv1 (2016): The first version, accurate but struggled with small objects.
- YOLOv2 (2017): Improved performance by using batch normalization and anchor boxes.
- YOLOv3 (2018): Further improvements with multi-scale predictions and better backbone architecture (Darknet-53).
- YOLOv4 (2020): Introduced improvements like CSPDarknet53 and additional optimizations for better accuracy and speed.
- YOLOv5: A community-driven implementation, offering enhanced performance.

The Rise of Transformers in Computer Vision

- Transformers, originally developed for NLP tasks, have revolutionized computer vision in recent years.
- Vision Transformers (ViT) are a new architecture that applies the transformer model to image data.
- ViT treats an image as a sequence of patches, similar to how transformers treat sequences of words in NLP.
- This shift represents a move away from the convolutional layers typically used in traditional CV models.

Vision Transformer (ViT)

- Introduced by Dosovitskiy et al. in 2020.
- The input image is split into non-overlapping patches (e.g., 16x16 pixels).
- Each patch is flattened into a vector and treated as a "token" similar to words in NLP.
- These tokens are fed into the transformer encoder to learn the global context of the image.
- ViT demonstrated superior performance over traditional CNNs when trained on large datasets like ImageNet.

ViT Architecture

- **Patch Embedding:** The image is divided into fixed-size patches.
- **Position Encoding:** Since transformers do not have inherent inductive biases for spatial locality, positional encodings are added to capture spatial relationships.
- **Transformer Encoder:** ViT uses the standard transformer architecture, consisting of multi-head self-attention layers.
- **Classification Head:** A final classification token is used to predict the output class.

Comparing YOLO and ViT

- **YOLO:**
  - Optimized for real-time object detection.
  - Excellent for applications requiring speed and efficiency.
  - Uses CNNs and grid-based predictions for object localization and classification.
- **ViT:**
  - A new paradigm that excels in image classification and recognition.
  - Requires large datasets and computational power for training.
  - Uses transformers to capture global context, offering flexibility and scalability.

Applications and Future Directions

- **YOLO Applications:**
  - Real-time object detection in video streams, surveillance, and autonomous vehicles.
  - Medical image analysis.
- **ViT Applications:**
  - Image classification, segmentation, and more advanced tasks.
  - Future improvements could focus on reducing the need for large datasets.
  - ViT may extend into multimodal applications (e.g., combining vision and language).
- **Future of Computer Vision:**
  - Hybrid models combining CNNs with transformers.
  - Self-supervised learning to reduce reliance on labeled data.
  - Expansion into other domains like 3D vision, robotics, and healthcare.

Conclusion

The development of computer vision models has progressed rapidly, from YOLO's fast and efficient real-time object detection to the powerful and flexible Vision Transformers. Both models represent milestones in the field, with YOLO excelling in speed and ViT pushing the boundaries of performance in image recognition. The future of computer vision is bright, with ongoing research promising even more breakthroughs.