EE6427 Video Signal Processing

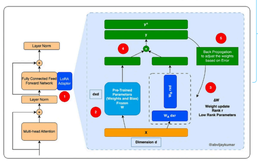
Part1 Emerging Topics in Video Signal Processing

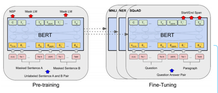
- adaptor tuning
  - adapting / fine-tuning PFMs — Foundation Models (FMs)
    - Low-Rank Adaptation (LoRA)

Figure 1: A detailed illustration of the model architectures of three different adapters: (a) Prefix-Tuning, (b) LoRA, (c) Series Adapter, and (d) Parallel Adapter.

$$W' = W + A \times B$$

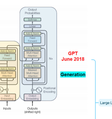
For each selected layer, introduce two low-rank matrices A and B that approximate changes to the original weight matrix W.





- Masked Language Model (MLM)
- Next Sentence Prediction (NSP)
- Bidirectional Encoder Representations from Transformers (BERT)

- B=PER, I=人名实体的起始词
- O=不属于任何命名实体
- ...

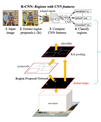


- Generative Pretrained Transformer (GPT)
- Encoder-decoder (e.g. Text-to-Text Transfer Transformer or T5)
- Large Language Models (LLMs)

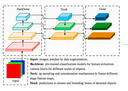
- A prediction is considered as True Positive (TP) if IoU > threshold, and False Positive (FP) if IoU < threshold.
- mean Average Precision (mAP)
  - All for current class the area under the precision-recall curve

$$mAP = \frac{1}{N}\sum_{class} AP_i$$

- performance metrics
  - Measure the total number of floating-point operations required for a single forward pass.
  - Floating Point Operations (FLOPs)
  - Pros: higher accuracy (in the past)
  - Cons: more complex and slower
- Two-Stage Detectors
  - Region-based Convolutional Neural Network (RCNN) Series Detectors

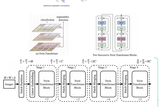


- One stage detectors regress and classify from the feature map directly (no Region Proposal Network (RPN).)
- You Only Look Once (YOLO)



- Lightweight Detectors
  - MobileNet

- Emerging / New Method
  - Swin Transformer
    - W-MSA: Window-based Multi-head Self-Attention
    - SW-MSA: Shifted Window-based Multi-head Self-Attention
    - LN: Layer Normalization, 层归一化
    - MLP: (Multi-Layer Perceptron，多层感知机)



  - Detection Transformer (DETR)
    - End-to-End Object Detection With Transformers
    - positioning and classification

- Object Detection

- Grounding (DINO)
  - step1: Detection: Use a detector to initialize tracks at frame t.
  - step2: Motion prediction: predict the next positions of objects using motion model.
  - step3: Data association: Match predicted positions with detections at frame t+1.
  - Multiple Object Tracking (MOT)
- Object Tracking

- Transformer
- Emerging / New Methods in MOT

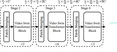
- Regression Method
- Single-Person Pose Estimation
  - Body Part Detection
  - Top-down HPes
  - Bottom-up OpenPose
  - Multi-Person Pose Estimation
- Pose Estimation
  - TransPose
  - MotionBERT — 3D Pose Estimation
- Emerging / New Methods

- Two-Stream Networks

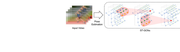
- 3D-CNNs

- Action Recognition Methods
- Video / Human Action Recognition (HAR)

- Temporal Shift Module (TSM)
  - Efficient Video Modeling



- Video Swin Transformer
  - Transformer-based Methods



- Skeleton-based Networks

- Long-Range Dependency Modeling Techniques
- Emerging/New Methods

- Different Deep Neural Network (DNN) Architectures
- Linear Classifiers
  - common regression loss functions

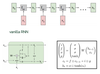
$$R_i = \sum_j (y_i - w_j x_{ij})^2$$

$$\text{where } z_i = f(x_i)$$

$$L = \sum_i L_i(y_i, z_i)$$

  - softmax loss (common classification loss function)
  - activation layer
- Convolutional Neural Network (CNN)

- Recurrent Neural Networks (RNNs)



- Long-Short Term Memory (LSTM)
- Recurrent Neural Network (RNN) & Long-Short Term Memory (LSTM)

- Transformer

- Vision Transformer (ViT)

Part2 Image Compression & JPEG Standard

- Terms and Concept
  - compression ratio: $\frac{B_0}{B_1}$
  - $n = H(S) = \sum_{i=1}^{n} p_i \log_2 \frac{1}{p_i}$
  - $\bar{n} = \sum_{i=1}^{n} p_i l_i$
- Entropy Coding — Huffman coding

  - $l_a = 1$
  - $l_b = 3$
  - $l_c = 3$
  - $l_d = 3$
  - $l_e = 3$
  - $l_f = 4$
  - $l_g = 4$
  - $l_h = 5$
  - $l_i = 5$
- Image & Video Compression Basics
  - 27 Image & Video Compression 的必要性
  - 27 compress冗余数据
    - 27 statistical redundancy
    - 27 psycho-visual redundancy
  - distortion measures / metrics
    - MSE
    - $SNR = 10\log_{10}\frac{\sigma_x^2}{\sigma_d^2}$
    - SNR (dB)
    - $PSNR = 10\log_{10}\frac{S^2}{\sigma_d^2}$
    - PSNR

- Transform-based Coding / Compression
  - $P_n = \sigma(\frac{\sum_{i=1}^{N} x_i w_{i,n}}{\sqrt{\sum_{i=1}^{N} w_{i,n}^2}}), w_{i,n} \sim \mathcal{N}(0,1), x = (x_1, \ldots, x_N)^T$
- Discrete Cosine Transform (DCT)
  - DCT的压缩功能点 — Energy compaction / Redundancy reduction
  - $F(u,v) = \sum_{x=0}^{7}\sum_{y=0}^{7} f(x,y) \cdot \Psi^T$
  - The 2D DCT matrix implementation is simply.
  - $F(i,j) = \sum_{x=0}^{7} f(x,i) \cdot \Psi_{x,j}$
  - $\Psi_{i,j} = \begin{cases} \frac{1}{\sqrt{8}} & \text{if } i = 0 \\ \frac{1}{2}\cos\frac{(2j+1)i\pi}{16} & \text{if } i > 0 \end{cases}$
- JPEG
  - GPCM: Differential Pulse Code Modulation — 连续图像块 DC 系数差异小的特点
  - RLC: Run Length Coding — 用于对经过之字形扫描后得出大量连续零值的 AC 系数进行编码
  - $F(u,v) = \frac{C(u)C(v)}{4}\sum_{x=0}^{7}\sum_{y=0}^{7} f(x,y)\cos\frac{(2x+1)u\pi}{16}\cos\frac{(2y+1)v\pi}{16}$
  - $C(u), C(v) = \begin{cases} \frac{1}{\sqrt{2}} & \text{if } u, v = 0 \\ 1 & \text{otherwise} \end{cases}$

- Video Coding & Standards
  - intraframe compression — JPEG-based compression
  - interframe compression
    - motion vector — reference macroblock -> matched macroblock
    - $SAD(i,j) = \sum_{x=1}^{N}\sum_{y=1}^{N} |C(x,y) - R(x+i,y+j)|$
    - motion estimation
      - full search
        - pros: accurate, best compression ratio
        - cons: slow, computationally intensive
      - three-step search
        - pros: fast
        - cons: less accurate, lower compression ratio

        - 3 steps
      - 2D-log search
        - speed and accuracy between full search and 3 step search

        - 2 steps (including iterative)
      - hierarchical search

        - A Three-level Hierarchical Search
      - diamond search
      - cross search
    - motion compensation — prediction error — PMergeh-PMreference

Sequence
Group of Pictures
Picture
Macroblock
Block
Header
Data
Macroblock
Block
Header
Data
Macroblock
Block
Header
Data