**Part A**

| **Group Number 79**   | **Member Names** | **BITS Id** | **Contribution** | | --- | --- | --- | | M S ANJANA |  | 100 % | | LADDHA SACHIN RAMBILAS | 2023ac05564 | 100 % | | NAKUL V |  | 100 % | | NEGI NUPUR RAVINDERSINGH | 2023ac05812 | 100 % | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Domain** - Human Activity Recognition from videos | | | |

|  | **Paper 1** | **Paper 2** | **Paper 3** |
| --- | --- | --- | --- |
| **Title of the paper** |  |  | ViT-ReT: Vision and Recurrent Transformer Neural Networks for Human Activity Recognition in Videos |
| **Authors** |  |  | James Wensel, Hayat Ullah, and Arslan Munir |
| **Year of publication** |  |  | 2023 |
| **Architecture of Deep Learning** |  |  | ViT-ReT - Vision Transformer - Recurring Transformer Vision Transformer (ViT)  * Purpose: This component replaces traditional Convolutional Neural Networks (CNNs) to extract spatial features from video frames. * How it works:   + Video frames are divided into smaller patches, such as 8×88 \times 88×8 pixel grids. These patches are then embedded into a sequence of vectors.   + A self-attention mechanism processes these vectors to learn the relationships between different patches, effectively capturing spatial features.   + The ViT uses pre-trained weights from large datasets like ImageNet, making it efficient for transfer learning. * Unique Features:   + It is lightweight compared to deep CNNs like ResNet, making it ideal for resource-constrained environments.   + Unlike CNNs, it doesn't rely on convolutional layers; instead, it treats image patches as sequences, akin to words in a sentence.  Recurrent Transformer (ReT)  * Purpose: This component replaces traditional Recurrent Neural Networks (RNNs) to model temporal dependencies in video data. * How it works:   + The ReT processes sequences of video frames with positional encodings to maintain the order of the frames.   + Unlike RNNs, which process sequences step-by-step, the Transformer encoder can parallelize computations, greatly improving efficiency. * Unique Features:   + It avoids the bottleneck associated with RNNs, making it faster and more scalable for long sequences.   By integrating these two components, the ViT-ReT framework efficiently handles the dual challenge of spatial and temporal feature extraction in video-based human activity recognition. |
|
|
|
|
| **How is the network helping the overall task?** |  |  | Human activity recognition involves understanding not just what is visible in individual frames but also how these frames relate over time. For example, recognizing a “high jump” requires analyzing the motion and sequence of the activity.   * Feature Extraction:   + The ViT handles spatial feature extraction, analyzing what is happening in each individual frame, such as identifying objects or body parts. * Temporal Modeling:   + The ReT looks at how these features change over time, making it possible to understand actions like walking, jumping, or waving. * Efficiency:   + This approach significantly reduces the computational cost compared to traditional CNN-RNN models while maintaining or improving accuracy. |
| **Training procedures** |  |  | Data Preprocessing:   * Video frames are resized to 224×224224 \times 224224×224 pixels, and only 20 frames per video are sampled. This reduces computational load without sacrificing key information. * Frames are preprocessed and stored to save time during training.   Optimization Algorithm:   * The Adam optimizer is used for faster and more stable convergence.   Hyperparameters:   * Batch size: 4 (small batches are used due to the computational complexity of Transformers). * Training epochs: 50.   Regularization:   * Dropout and normalization layers are employed to prevent overfitting.   Separate Training for Components:   * The Vision Transformer is pre-trained on ImageNet, and the Recurrent Transformer is trained on extracted features from video datasets. |
| **Evaluation / Performance metric used** |  |  | * Accuracy: The percentage of correctly classified activities. * Loss: Measured using categorical cross-entropy. * Efficiency Metrics:   + Training Time: The time required to train the model.   + Throughput Time: How quickly the model can make predictions.   + Memory Usage: The amount of memory required for inference.   UCF101: A dataset of 13,320 videos across 101 categories (e.g., sports activities, gestures).   * URL: UCF101 Dataset.   Additional Datasets:   * YouTube Action Dataset. * HMDB51 Dataset. * UCF50 Dataset. |
| **Name of Dataset used. If a public dataset, provide the URL.** |  |  | <https://www.kaggle.com/datasets/matthewjansen/ucf101-action-recognition> |