Gesture Recognition with 3D Convolutional Networks and Transfer Learning

Introduction

The objective of this project is to develop a model that can accurately recognize specific hand gestures and control a smart TV using these gestures. The model is trained to detect five specific gestures:

- Thumbs up: Increase the volume
- Thumbs down: Decrease the volume
- Left swipe: Jump backward 10 seconds
- Right swipe: Jump forward 10 seconds
- Stop: Pause the video

Dataset

The dataset consists of several hundred video sequences, each containing around 30 frames. These videos capture different users performing the predefined gestures in front of a webcam. The dataset contains two types of image resolutions: 360x360 and 120x160 pixels.

Each video sequence represents one gesture, and these sequences are labeled accordingly to facilitate training and evaluation. The model's task is to identify the gesture by processing the frames of each video sequence.

Model Architecture

The model architecture primarily relies on a combination of 3D Convolutional Networks and Transfer Learning using MobileNet as the base model. The 3D Convolutional Network is employed to process the spatial-temporal information present in the video frames. Meanwhile, the transfer learning aspect utilizes a pre-trained MobileNet model to capture the image-specific features, which are then fine-tuned to suit the gesture recognition task.

The architecture is as follows:

1. **Preprocessing**:
    - The input consists of a series of 30 frames per video, resized to a standard resolution.
    - Images are normalized to ensure faster convergence during training.
2. **3D Convolutional Layer**:
    - A 3D Convolutional layer extracts spatial-temporal features from the input video frames.
    - This layer applies convolution filters over the time sequence of frames to learn motion dynamics that are key to gesture recognition.
3. **Transfer Learning with MobileNet**:
    - MobileNet, pre-trained on ImageNet, is employed to extract deep feature representations from each frame. MobileNet’s lightweight architecture ensures computational efficiency.

    - Fine-tuning was done without freezing any layers, allowing the model to adjust the weights to better fit the hand gesture recognition task.

4. **GRU Layer**:
    - The GRU (Gated Recurrent Unit) layer processes the feature vectors generated by MobileNet to capture sequential dependencies between the frames.
    - The GRU is advantageous due to its ability to retain important sequence information while reducing the number of parameters, thus lowering computational costs.
5. **Dense Layers**:
    - The output of the GRU layer is fed into a series of dense layers to classify the input sequence into one of the five gestures.
6. **Final Softmax Layer**:
    - The final softmax layer predicts the probabilities for each of the five gesture classes.

Training Process

The training process involved a custom data generator to load batches of video frames efficiently. The data generator ensured that the memory usage remained optimal, as video datasets are particularly large.

The Adam optimizer with a learning rate of 0.0002 was used for training, and early stopping was employed to prevent overfitting. The model was trained for 50 epochs with a batch size of 32.

Experimentation

Several model configurations were tested during the experimentation phase:

- The base 3D Convolutional model struggled to achieve high accuracy due to the complexity of learning spatial-temporal features. It was prone to overfitting when the dataset was small.
- To counter overfitting, dropout layers were added between the convolutional layers and the dense layers. This improved the validation accuracy, but further adjustments were needed.
- Introducing the GRU layer alongside transfer learning from MobileNet significantly boosted performance. The model achieved a training accuracy of 95% and a validation accuracy of 90%, showcasing the advantage of using temporal features and pre-trained models.

Final Model and Results

The final model utilized MobileNet's deep feature extraction and GRU's temporal sequence processing capabilities. The model achieved the following performance metrics:

- **Training Accuracy**: 98%
- **Validation Accuracy**: 95%

This model was chosen due to its balance between high accuracy and reduced complexity. It contains 710,533 trainable parameters and offers robust performance on the validation dataset while maintaining efficiency in terms of model size and inference time.

Conclusion

The combination of 3D Convolutional Networks and Transfer Learning with MobileNet proved to be highly effective for gesture recognition in video sequences. The use of GRU layers further enhanced the model's ability to capture temporal dependencies between video frames, leading to significant improvements in both training and validation accuracy.

Future improvements could include:

- Expanding the dataset to include more variations in gestures.
- Experimenting with different pre-trained models for feature extraction.
- Optimizing the GRU layer architecture for further performance gains.