Gesture Recognition Project Write-up

This project focuses on building a model for gesture recognition using deep learning techniques. We explored several architectures to determine the most effective model for accurate gesture classification. Among the tested models, two stood out: Model 2 (Conv3D-based) and Model 5 (MobileNet V2 with GRU). Below is a detailed explanation of these models and the decision-making process behind choosing them.

Model 2: Conv3D-based Architecture

Model 2 uses four Conv3D layers followed by two fully connected layers and a softmax output layer. This architecture processes gesture sequences as volumes of frames, treating the temporal dimension similarly to how images are handled in CNNs.

Architecture:

- **Convolutional Layers**: Four Conv3D layers, which capture spatio-temporal information in the input sequences. Conv3D layers are particularly effective for extracting features from videos or sequences of frames.
- **Fully Connected Layers**: Two dense layers with 128 neurons each. These layers help in learning non-linear combinations of the high-level features extracted by the Conv3D layers.
- **Activation Function**: ReLU, chosen for its non-linear properties and ability to avoid vanishing gradients.
- **Output Layer**: Softmax layer for multi-class classification.
- **Optimizer**: Adam, with the default learning rate, was chosen for its fast convergence and adaptability to different gradients.

Hyperparameters:

- **Batch Size**: 32 (reduced from 64). Reducing the batch size led to quicker updates to the weights, allowing for faster convergence in terms of epochs.

- **Number of Frames per Sequence**: 30
- **Input Image Shape**: (120, 120, 3)
- **Epochs**: 20

Performance:

- **Training Accuracy**: 96.23%
- **Validation Accuracy**: 91.00%
- **Training Loss**: 0.0990
- **Validation Loss**: 0.2845

Insights:

The decision to reduce the batch size from 64 to 32 was made after observing that smaller batch sizes led to faster weight updates. This helped Model 2 reach high accuracies quicker than its predecessors. The Conv3D architecture proved to be highly effective in capturing spatio-temporal features, as evidenced by the high validation accuracy of 91.00%, with minimal overfitting (as indicated by the low gap between training and validation accuracies).

Model 2, with its Conv3D layers, was selected as the best performing model at this stage. It showcased a balanced performance in terms of both accuracy and loss, with little to no signs of overfitting, making it a strong candidate for further refinement.

Model 5: MobileNet V2 with GRU

Model 5 introduces a more advanced architecture by leveraging MobileNet V2 as a pre-trained Conv2D model, combined with a GRU (Gated Recurrent Unit) layer. The key change here is the integration of a pre-trained model for feature extraction and a sequence processing unit (GRU) for temporal understanding.

Architecture:

- **MobileNet V2**: Used as the backbone for feature extraction. MobileNet V2 is known for its efficiency and accuracy in image recognition tasks. In this model, we retrained the weights using our dataset to fine-tune it for gesture recognition.
- **GRU Layer**: GRU is a type of recurrent neural network (RNN) that is particularly useful for sequence prediction. It was added to capture the

temporal dependencies across frames. The GRU layer contained 128 cells with a dropout of 0.5 to prevent overfitting.

- **Fully Connected Layer**: Increased the width of the fully connected layer to 256 neurons, allowing it to better capture the relationships between high-level features extracted by MobileNet and temporal data captured by the GRU.
- **Activation Function**: ReLU
- **Output Layer**: Softmax layer for classification.
- **Optimizer**: Adam, with a reduced learning rate of 0.0001.

Modifications:

- **Learning Rate**: Reduced further to 0.0001 to allow for finer updates to weights during training. This helps in preventing overfitting and ensuring the model converges to a more optimal solution.

Focus:

Model 5 focuses on using the pre-trained MobileNet V2 for spatial feature extraction and the GRU layer for temporal feature analysis. The combination of these two architectures (Conv2D and GRU) allows for a strong balance between efficient feature extraction and effective sequence learning.

Conclusion:

Model 2, with its Conv3D-based architecture, emerged as the best performing model in this project, achieving an excellent balance between training and validation accuracy with no signs of overfitting. The decision to reduce the batch size played a key role in achieving better results with this model. Meanwhile, Model 5 introduces a more complex architecture that leverages transfer learning and recurrent networks, and while it shows promise, its performance relative to Model 2 warrants further exploration.

In future iterations, more experimentation can be done, such as fine-tuning the GRU parameters, trying different pre-trained models, or incorporating ConvLSTM layers, which may lead to even better performance. The industry demo on CNNs provided a solid foundation for understanding the importance of experimenting with different data augmentation techniques, architectures, and hyperparameters to improve gesture recognition tasks.