Gesture Recognition Model Development

Experimentation on Model Design Hyperparameters

- Out of the standard image dimensions of 128x128, 160x160, 192x192, 224x224, we have chosen **128x128** as the desired image input dimension as it was able to capture the key information from 360x360 sized images and able to capture maximum information from 120x160 images. The model performance validations have also resulted the same.
- We have tested for various epochs and found **25 epochs** to be the optimum model training period for the given data and model designs.
- We tested various batch sizes and used batch size of **32** as it is a standard value and good enough for model training speed and execution.
- Our initial models were developed without dropout which resulted in attaining 100% training accuracy in early stages and validation accuracy being lesser than expected performance resulting in huge overfitting.
- We have experimented with dropout values of 0.2, 0.3, 0.5 and selected **0.2** being the best dropout as it reduced the overfitting significantly and gave maximum validation accuracy for both models.
- To reduce the impact of overfitting observed in 3D CNN model, we have even introduced Batch Normalization but it was unable to attain atleast 50% validation accuracy making the model underperform on new data.
- We have limited to **2 CNN layers** with **32** and **64** filters as best simple architecture for 3D CNN model as the model was unable to perform well on more than 3 layers with high count filters.
- Experimentation was done on various dense layers with high count filters which was simply increasing the trainable parameters but unable to perform better.

Observations

For Model 1: 2D CNN + RNN (using Transfer Learning and GRU with Dropout)

- There are 31,62,885 trainable parameters for the GRU and Dense layers implemented with 32,28,864 non-trainable parameters as part of MobileNet model transfer learning.
- The model has ultimately reached 100% training accuracy in the final stage of training.
- The model has reached its highest validation accuracy of **0.9500** with a validation loss of **0.1858** during 23rd epoch.
- The model had its least validation loss at **0.1662** with a validation accuracy of **0.9400** during 18th epoch.
- From the accuracy and loss plots of the model, it is evident that the validation curves are quite close to the training curves indicating that the model is not overfitting and performs well on new unseen data as well.

For Model 2: 3D CNN (with Dropout)

- There are 5,87,79,013 trainable parameters for the complete model and 0 non-trainable parameters as no transfer learning was used.
- The model has reached 100% training accuracy during initial stages of model training.
- The model has reached its highest validation accuracy of **0.8500** with a validation loss of **0.7641** being the least as well during 9th epoch.
- We can observe from the accuracy and loss plots of the model that the validation curves are quite away from the training curves indicating that the model is quite overfitting and might not perform well on new unseen data.

Conclusion

The model training and performance for the gesture recognition of video frames data has been experimented over various kinds of 2D CNN + RNN and 3D CNN models along with varying design hyperparameters. Ultimately, the 2D CNN + RNN Model designed from MobileNet base model and GRU implementation with dropout is the best model recommended due to its exceptional performance on unseen data with high validation accuracy and low validation loss obtained on the processed image dataset. Whereas the 3D CNN model with dropout was unable to perform well on validation data along with huge number of model parameters leading to another drawback. Considering validation accuracy and loss as primary and secondary constraints respectively, we suggest the model with highest validation accuracy of 95% and low validation loss of 18.6% as the best performance model to capture and recognize the hand gestures of Smart TV users for content streaming controls.