Neural Networks Project – Gesture Recognition

Team

Gurpreet Singh (Group Facilitator) **android.gurpreet@gmail.com**

Karan Prinja (Group Member) **karan.prinja@rakuten.com**

Problem Statement

Imagine you are working as a data scientist at a home electronics company which manufactures state of the art smart televisions. You want to develop a cool feature in the smart-TV that can recognize five different gestures performed by the user which will help users control the TV without using a remote.

Each video is a sequence of 30 frames (or images).

The gestures are continuously monitored by the webcam mounted on the TV. Each gesture corresponds to a specific command:

- Thumbs up: Increase the volume
- Thumbs down: Decrease the volume
- Left swipe: 'Jump' backwards 10 seconds
- Right swipe: 'Jump' forward 10 seconds
- Stop: Pause the movie

Dataset

The training data consists of a few hundred videos categorized into one of the five classes. Each video (typically 2-3 seconds long) is divided into a sequence of 30 frames(images). These videos have been recorded by various people performing one of the five gestures in front of a webcam - similar to what the smart TV will use.

The Models

Exploring the efficacy of various deep learning models for real-time gesture recognition in smart TV applications, we delve into the characteristics and performance of eight distinct models, culminating in the selection of the optimal model that balances accuracy, generalization, and computational efficiency.

<table>
<tr><th>Model</th><th>Type</th><th>Batch Size</th><th>Image Size</th><th>Frames</th><th>Dense Neurons</th><th>Total Params</th><th>Trainable Params</th><th>Non-trainable Params</th><th>Categorical Accuracy</th><th>Validation Categorical Accuracy</th></tr>
<tr><td>1</td><td>Conv3D</td><td>24</td><td>100 x 100</td><td>30</td><td>64</td><td>592,773</td><td>592,037</td><td>736</td><td>96%</td><td>80%</td></tr>
<tr><td colspan="11">The model's strength lies in its high training accuracy, suggesting effective learning, but the disparity in validation accuracy points towards a need for better generalization to unseen data.</td></tr>
<tr><td>2</td><td>Conv3D</td><td>20</td><td>120 x 120</td><td>20</td><td>64</td><td>1,113,925</td><td>1,112,933</td><td>992</td><td>99%</td><td>76%</td></tr>
<tr><td colspan="11">The high training accuracy indicates effective learning capabilities, yet the considerable drop in validation accuracy suggests issues with overfitting</td></tr>
<tr><td>3</td><td>Conv3D</td><td>20</td><td>100 x 100</td><td>30</td><td>64</td><td>900,933</td><td>899,941</td><td>992</td><td>98%</td><td>90%</td></tr>
<tr><td colspan="11">The model exhibits strong learning with its high training accuracy, and the closer alignment between training and validation accuracies indicates better generalization capabilities compared to the previous models. This balance suggests a more effective model in handling unseen data.</td></tr>
<tr><td>4</td><td>Conv3D</td><td>20</td><td>150 x 150</td><td>30</td><td>128</td><td>1,638,213</td><td>1,637,221</td><td>992</td><td>99%</td><td>74%</td></tr>
<tr><td colspan="11">Despite its excellent training accuracy, the significant gap between training and validation accuracies suggests a strong tendency towards overfitting.</td></tr>
<tr><td>5</td><td>Conv3D</td><td>22</td><td>110 x 110</td><td>30</td><td>136</td><td>9,40,029</td><td>9,39,005</td><td>1024</td><td>99%</td><td>75%</td></tr>
<tr><td colspan="11">The model's high training accuracy is indicative of effective learning from the training data. However, the large discrepancy between training and validation accuracies suggests potential overfitting</td></tr>
<tr><td>6</td><td>CNN-LSTM</td><td>20</td><td>120 x 120</td><td>30</td><td>64</td><td>7488069</td><td>7487621</td><td>448</td><td>51%</td><td>49%</td></tr>
<tr><td colspan="11">The model's balanced training and validation accuracies indicate consistency in performance on both training and unseen data. However, the relatively low accuracy levels suggest that the model might benefit from further tuning, possibly including adjustments in architecture, hyperparameters, or training data augmentation to enhance learning efficiency.</td></tr>
<tr><td>7</td><td>Transfer Learning (ResNet50)</td><td>20</td><td>120 x 120</td><td>30</td><td>64</td><td>24,129,861</td><td>538,053</td><td>23,591,808</td><td>62%</td><td>50%</td></tr>
<tr><td colspan="11">The use of transfer learning with ResNet50 provides a complex model structure, evident from the high number of non-trainable parameters. While the training accuracy is moderate, the drop in validation accuracy indicates challenges in generalizing to new data.</td></tr>
<tr><td>8</td><td>CNN-LSTM</td><td>20</td><td>100 x 100</td><td>30</td><td>64</td><td>8,87,141</td><td>8,86,149</td><td>992</td><td>99%</td><td>81%</td></tr>
<tr><td colspan="11">The model demonstrates excellent learning capabilities as reflected in the high training accuracy. However, the noticeable gap between training and validation accuracies suggests a potential overfitting issue.</td></tr>
</table>

For the specific application of gesture recognition for a smart TV using a webcam, **Model 3** (Conv3D) would be the most suitable choice. Here is why:

Balance Between Accuracy and Generalization: Model 3 strikes an excellent balance with 98% training accuracy and 90% validation accuracy. This indicates not only a high degree of learning from the training dataset but also a strong ability to generalize well to new, unseen data, which is crucial for real-time gesture recognition.

Appropriate Complexity for Real-Time Processing: While Model 3 is complex enough to capture the nuances in gesture recognition (which can be quite subtle and varied), it doesn't have an excessively high number of parameters. This balance is essential for ensuring that the model can process data in real-time, a key requirement for a smooth user experience in a smart TV context.

Suitability for Video Data: Given that gesture recognition involves analyzing sequences of images (video data), the Conv3D architecture of Model 3 is well-suited for this task. It can effectively capture the spatial and temporal dynamics of gestures, which is essential for accurate recognition.

Potential for Real-World Robustness: The higher validation accuracy suggests that Model 3 will be more robust in a real-world environment, dealing effectively with the variations and unpredictability inherent in how different users might perform gestures.

In conclusion, for gesture recognition in a smart TV setup, Model 3's combination of high accuracy, good generalization, appropriate complexity, and suitability for video data makes it the optimal choice.