Introduction

Objective:

Using:
Long Term Recurrent Convolutional Network.
- Convolutional Neural Network (CNN)
- Recurrent Neural Network (RNN)
  - LSTM
Our best LRCN model consisted of a pre-trained CNN and a 4-layer LSTM model as the recurrent module, with 44% accuracy.

Dataset
  - Videos shot by amateurs
  - Inconsistent and low quality
  - Contains 4200 videos
  - 244 swimming videos (119 are labeled)
  - Split into > 2600 examples

Features
Video segments:
- 1 second long (30 frames)
- 299 x 299 pixels
- RGB

Models
We implemented a Long-Term Recurrent Convolutional Network (LRCN) model. We used a pre-trained Inception V3 as the visual feature extractor. We experimented multiple LSTM models for the recurrent module. For this model, we used the categorical cross entropy as the loss function:

\[ L = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{C} y_{ij} \log(P_{ij}) \]

Results

<table>
<thead>
<tr>
<th>Name</th>
<th>Visual Feature Extractor</th>
<th>Sequence Learning</th>
<th>Output</th>
<th>Dev Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deep Swim Master</td>
<td>Inception V3</td>
<td>LSTM x4</td>
<td>Dense Softmax</td>
<td>44%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(512, 256, 128, 64)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wide Swim Master</td>
<td>Inception V3</td>
<td>LSTM x1</td>
<td>Dense Softmax</td>
<td>34%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(1024)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Convolutional Baseline</td>
<td>Conv (3x3x5), (5x5x10),</td>
<td>Flatten</td>
<td>Dropout (0.4), Dense Softmax</td>
<td>28%</td>
</tr>
</tbody>
</table>

Discussion
Our best model achieved approximately 44% accuracy. Throughout the project, we discovered the dataset we used was not large enough to build a robust action recognition classifier. The dataset was also imbalanced and skewed to a certain label (i.e., Backstroke), which led our models to overfit the training data. In addition, our experiments indicate that using a deeper LSTM model results in a better performance, as the model is more capable of understanding the temporal features from video data.

Future
Regarding the future work, we would try to increase the model’s accuracy by using:
- Building a more robust data preprocessing pipeline and more videos.
- Modifying the model for multiple object action detection.

References