Introduction

Cerebral palsy (CP) is a motor disorder that causes impaired musculoskeletal development, and affects approximately 3 out of every 1000 children born [1]. There are a range of treatment options from physical therapy to surgery, and quantitative measures calculated from gait analysis such as the Gait Deviation Index (GDI) can be used to help inform treatment decisions. Historically, marker-based motion capture has been used to collect gait analysis data, but these setups are expensive (ranging of thousands of dollars), and require specialized spaces and personnel. Compared to the current standard of one or two visits max to the gait lab for each child, if GDI could be calculated with 3D video, this would allow patients (or their parents) to collect gait data at home, and allow doctors to continuously monitor their progress.

Methods

We started by feeding individual DensePose video frames into a small convolutional network with a classifier predicting individual image GDI in one of seventeen bins. We then used transfer learning on AlexNet with various numbers of feature layers to get a deeper network. We decided to add LSTM layers to account for the time dependence of our video frames. We gathered feature vectors from the output of AlexNet and fed those into the LSTM layers to get a single GDI score for an entire video’s worth of DensePose image frames.

Results

<table>
<thead>
<tr>
<th>Model</th>
<th>Val Acc Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>CNN Baseline</td>
<td>0.595</td>
</tr>
<tr>
<td>Pretrained AlexNet</td>
<td>0.584</td>
</tr>
<tr>
<td>Finetuned AlexNet</td>
<td>0.586</td>
</tr>
<tr>
<td>AlexNet + LSTM</td>
<td>0.605</td>
</tr>
</tbody>
</table>

Our dataset was provided by Gillette Children’s Hospital and consists of ≈ 5000 videos of children with CP walking, along with paired GDI scores calculated from motion capture for each child from that visit. We augmented this data by splitting each video into segments of 124 frames, giving us ≈ 6500 videos. These videos were initially run through the DensePose network, which associates points on a 3D image to a surface-based representation of the human body.

Discussion

Given the nature of our GDI data (i.e., temporal, clustered near middle scores), predicting the output using a purely CNN model was difficult. To train the CNN, it was required that we associate the patient’s GDI with each individual image versus associating a single GDI with a series of images. This removed the temporal aspect from the data, which we believe contributed to our low validation accuracies. In order to combat this, we implemented a sequential LSTM network taking as input the features pulled out from the CNN. Due to the non-representative training of the CNN, along with the lack of data spread across bins, the LSTM still suffered from low validation accuracies (though its correlation value of 0.568 is not vastly different than the correlation of 0.74 in a state-of-the-art network given the same task[6]). Training an end-to-end CNN into LSTM network with more data may solve a lot of these issues and result in higher accuracies.

Future Work

Future work on the model focuses on more personalized training of the model on the provided videos, as well as focus on improving the model accuracy through reduction of bias from the data.

- Consider an end-to-end approach
- Gather more videos
- Preprocess/segment the data with additional information
- Retrain a full AlexNet on images with sequence model
- Increase the size of the network.

References


Acknowledgements

Thanks to Gillette Children’s Hospital for providing the dataset, Laksh Kottipilla for running the videos through DensePose and for project mentorship, and the CS230 teaching team for the guidance provided on this project, especially Auréli Bagda.