PURA Final Report – Biomechanical Data Synthesis for Control System Optimization

Background and Rationale

In the fields of orthotic and prosthesis development, a major bottleneck is in the gathering of novel datasets from relevant subjects, as there is generally a high scarcity level associated with individuals of a specific disability. This results in an unhealthy reliance on datasets gathered from able-bodied subjects and an overreliance on a small collection of datasets gathered from amputees or similarly disabled individuals. Thus, in order to expedite the development stages of assistive mechanisms, it becomes imperative to find a method to artificially expand a small collection of datasets gathered from a relevant group of individuals into a large, yet still novel, collection.

In the current state of biomechanical research, there is a heavy reliance on data science-driven solutions, which generally entail designing and training machine learning algorithms to execute and/or predict motion according to sensor data. To this end, user intent recognition has become a prominent aspect of prosthesis control, resulting in the implementation of feature-based (features being derived from sensor readings) and, presumably, soon to be time series-based models. This transition between basic and deep learning in the field is a main motivator in searching for effective data synthesis solutions since deeper models have been shown to require significantly more training data.

In order to attain additional training data for the development of these models, we can either augment our data with new features derived from those which already exist or we can synthesize entirely new datasets based on trends found in real-life trials. The former option, although likely safe from outputting many outliers which would skew models, has been shown in image processing applications to generally only improve model accuracy marginally. On the other hand, there have been several promising developments in data synthesis both for feature and time series datasets made using generative adversarial networks (GANs) and variational autoencoders (VAEs).

The basic architecture of GANs models consists of one model attempting to replicate novel data from a set of existing data, in this case, the sensor data, and a model attempting to discriminate between real and synthesized data. Through these two models' competition, the synthesizing model becomes progressively more accurate; however, in order to reach a realistic threshold, a great amount of existing data must be provided, which is a potential roadblock in this application. VAEs attempt to encode a signal into a smaller (latent) space and then decode from there to produce a similar signal. This model requires far less training data, but has been shown to be less effective in data synthesis than GANs models. An additional benefit of VAEs is that when encoding the signal, they have the potential to find hidden features that define the gait data. Thus, after decoding, the final synthesized signal may represent these features heavily and thereby inform the model it is eventually training in a more meaningful way than regular (more noisy) signals would.

Results

Dataset and Preprocessing.

In this research, I used an open source dataset developed by Jonathan Camacho and Blair Hu, from which I extracted all biomechanical data channels from each subject, as shown in fig. 1a. However, after preliminary results from the ML models, it became apparent that many strides and channels lacked associated data signals, which was handled by removing these from the datasets. Additionally, it became necessary to normalize both the length of each stride and amplitude of the given sensor channels, at least initially, in order to effectively train the models. Thus, we attain example data signals shown in fig. 1b.

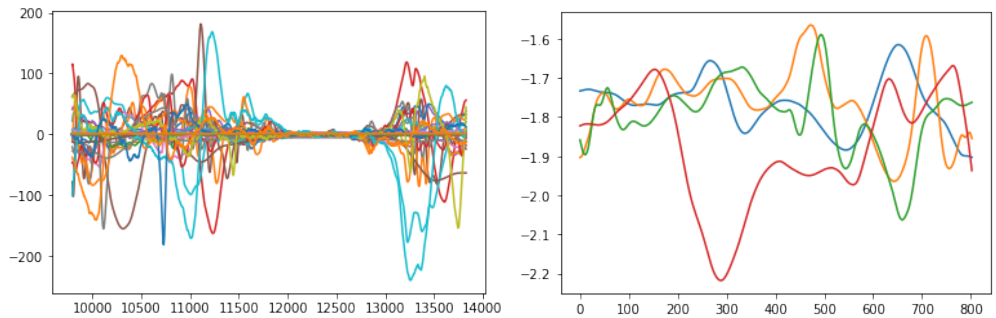

Figure 1a. Initial biomechanical data for example subject. 1b. Preprocessed data signals.

Machine Learning Approaches.

I began with developing the GAN model, as it seemed the most promising through literature review. I defined both the discriminator and generator models using two dense layers, with an Adam optimizer and binary cross entropy loss functions. Using 10000 epochs, the model failed to generate any reasonable results (the training process is shown in fig. 2).

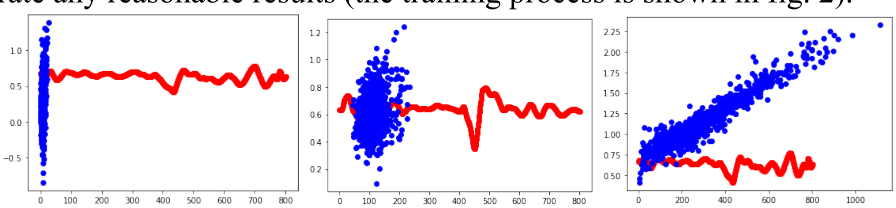

Figure 2. GAN training results at epochs 1000, 5000, and 10000.

My second approach was using a VAE, using a latent space of dimension 10. Unfortunately, the results seemed similar to the GAN model, as shown in fig. 3. I used 128 nodes in the hidden layer and a binary cross entropy loss function. Despite tuning the hyperparameters of the model, it did not improve.

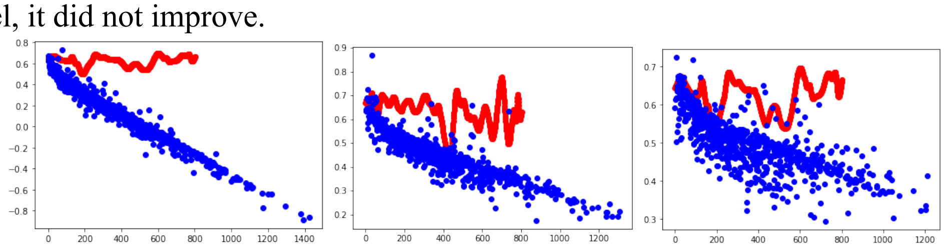

Figure 3. VAE training results at epochs 1000, 5000, and 10000.

Conclusion

Using these two machine learning approaches did not yield reasonable results. I believe this is due to the high dimensionality of biomechanical signals for nearly all joint motion in question and is not able to be captured by ML-based models without copious amounts of training data. As seen in the figures above, the generated data seem to capture the initial linear direction of the signals, supporting the idea that the dimensionality is the issue.