Variational Autoencoders for Biosensor Data Augmentation

Solomon Kim and Rodney Summerscales Ph.D.

Bcn_prkclr md Amknsrgle + Amjjcec md Npmdcqqgmlq* ?lbpcuq Slgtcpqgrw

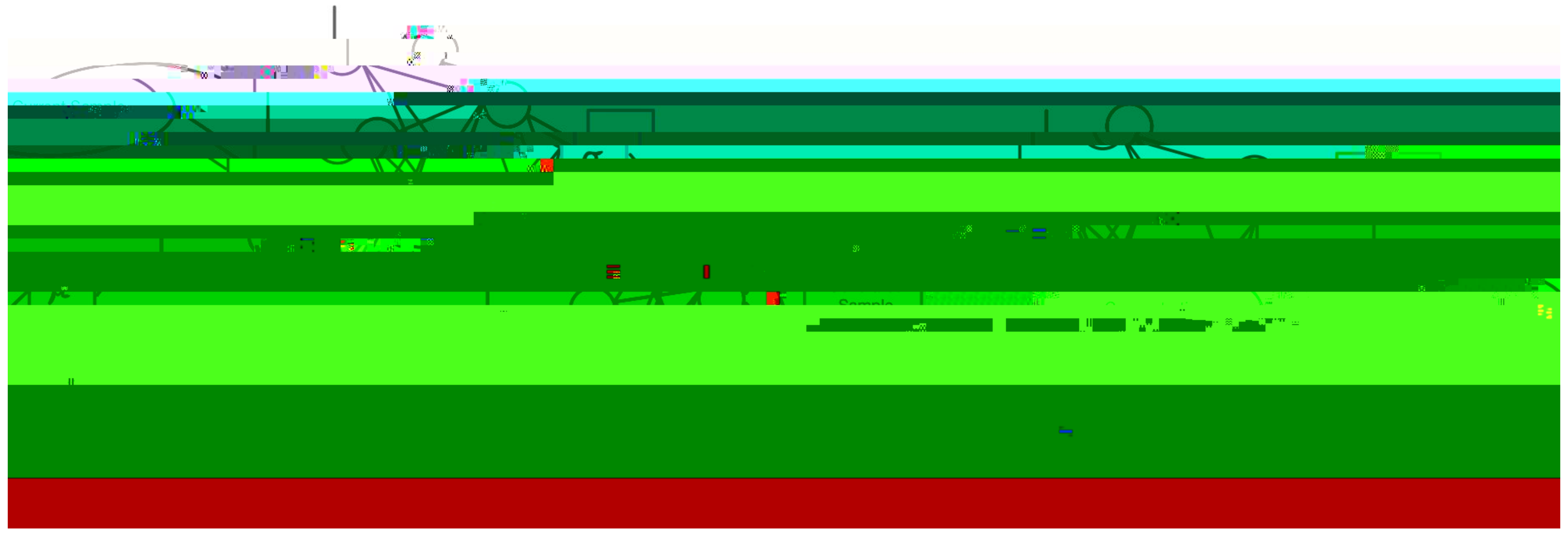
Abstract

Over the past decade machine learning and artificial intelligence's resurgence spawned the desire to mimic human creative ability. Initially attempts to create images, music, and text flooded the community, though little has been learned regarding constrained, one-dimensional data generation. This paper demonstrates a variational autoencoder approach to this problem. By modeling biosensor current and concentration data we aim to augment the existing dataset. In training a multi-layer neural network based encoder and decoder we were able to generate realistic, original samples. These results demonstrate the ability to realistically augment datasets, improving training of machine learning models designed to predict concentration from input signals.

To begin we extracted and cleaned the biosensor current data, organizing by corresponding concentrations. This data was then split into a train and test set. We then created the model. The model was trained for 200 epochs o

Model Overview

The variational autoencoder used differs from most given the unique multi-input to the encoder and decoder. The encoder takes in the current sample as well as the corresponding concentration. Once the encoder produces a mean and variance, the decoder will take that in as a latent input as well as the concentration in order to generate a realistic sample. Most typical variational autoencoders will only take in a sample and produce a sample, not considering outside factors, like the concentration. Within the encoder we have two separate branches to the model, a concentration and current branch that each model the separate inputs. The encoder and decoder interact, as shown in Figure 2, to create the variational autoencoder.



Results and Discussion

In Figure 3, we see the comparison between our generated samples and real samples. The results are promising as there are similarities and a general trend that the VAE captures in generation for different concentrations. However, among these samples we see a couple clear differences between the real and generated samples. For one, the curves that we generate are more shaky and are not smooth like the real curves. Most of these curves are also quite a bit off of real values after the initial peak. These key differences lead us to the need for future work and research. There are two main changes that will be investigated. The model architecture will be changed from a dense neural network to a 1D CNN. This architecture change should allow for the VAE to more consistently sense sequence trends. Another change that will need to be investigated is the amount of the curve we need to predict. The most important part of the curve is generally the initial peak and downslope. If we can limit the amount the model must generate, we can improve the accuracy and make more controlled improvements.